Advanced statistical inference

Suhasini Subba Rao

Email: suhasini.subbarao@stat.tamu.edu

April 26, 2017

Contents

Chapter 1

The Likelihood

In this chapter we review some results that you may have came across previously. We define the likelihood and construct the likelihood in slightly non-standard situations. We derive properties associated with the likelihood, such as the Crámer-Rao bound and sufficiency. Finally we review properties of the exponential family which are an important parametric class of distributions with some elegant properties.

1.1 The likelihood function

Suppose $\underline{x} = \{X_i\}$ is a realized version of the random vector $\underline{X} = \{X_i\}$. Suppose the density f is unknown, however, it is known that the true density belongs to the density class $\mathcal{F}$. For each density in $\mathcal{F}$, $f_{\underline{X}}(\underline{x})$ specifies how the density changes over the sample space of $\underline{X}$. Regions in the sample space where $f_{\underline{X}}(\underline{x})$ is "large" point to events which are more likely than regions where $f_{\underline{X}}(\underline{x})$ is "small". However, we have in our hand $\underline{x}$ and our objective is to determine which distribution the observation $\underline{x}$ may have come from. In this case, it is useful to turn the story around. For a given realisation $\underline{x}$ and each $f \in \mathcal{F}$ one evaluates $f_{\underline{X}}(\underline{x})$. This "measures" the *likelihood* of a particular density in $\mathcal{F}$ based on a realisation $\underline{x}$. The term likelihood was first coined by Fisher.

In most applications, we restrict the class of densities $\mathcal{F}$ to a "parametric" class. That is $\mathcal{F} = \{f(\underline{x};\theta); \theta \in \Theta\}$, where the form of the density $f(\underline{x};\cdot)$ is known but the finite dimensional parameter θ is unknown. Since the aim is to make decisions about θ based on a realisation $\underline{x}$ we often write $L(\theta;\underline{x}) = f(\underline{x};\theta)$ which we call the *likelihood*. For convenience, we will often work with the log-likelihood $\mathcal{L}(\theta;\underline{x}) = \log f(\underline{x};\theta)$. Since the

logarithm is a monotonic transform the maximum of the likelihood and log-likelihood will be the same. This preservation of maximum is very important.

Let us consider the simplest case that $\{X_i\}$ are iid random variables with probability function (or probability density function) $f(x;\theta)$, where f is known but the parameter θ is unknown. The likelihood function of θ based on $\{X_i\}$ is

$$L(\theta;\underline{X}) = \prod_{i=1}^{n} f(X_i;\theta) \tag{1.1}$$

and the log-likelihood turns product into sum

$$\log L(\theta;\underline{X}) = \mathcal{L}(\theta;\underline{X}) = \sum_{i=1}^{n} \log f(X_i;\theta). \tag{1.2}$$

We now consider some simple examples.

Example 1.1.1 *(i) Suppose that $\{X_i\}$ are iid normal random variables with mean μ and variance σ^2 the log likelihood is*

$$\mathcal{L}_n(\mu,\sigma^2;\underline{X}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} - \frac{n}{2}\log 2\pi$$

Observe that the parameters and random variables are "separable".

(ii) Suppose that $\{X_i\}$ are iid binomial random variables $X_i \sim Bin(m,\pi)$. We assume m is known, then the log likelihood for π is

$$\begin{aligned}\mathcal{L}_n(\pi;\underline{X}) &= \sum_{i=1}^{n}\log\binom{m}{X_i} + \sum_{i=1}^{n}\Big(X_i\log\pi + (m-X_i)\log(1-\pi)\Big)\\ &= \sum_{i=1}^{n}\log\binom{m}{X_i} + \sum_{i=1}^{n}\Big(X_i\log\big(\frac{\pi}{1-\pi}\big) + m\log(1-\pi)\Big).\end{aligned}$$

Observe that the parameters and random variables are "separable".

(iii) Suppose that $\{X_i\}$ are independent random variables which give the number of "successes" out of m_i. It seems reasonable to model $X_i \sim Bin(m_i,\pi_i)$. It is believed that the the regressors z_i influence the chance of success π_i. We try to model this influence with the nonlinear transform

$$\pi_i = g(e^{\beta' z_i}) = \frac{e^{\beta' z_i}}{1+e^{\beta' z_i}},$$

where β are the unknown parameters of interest. Then the log likelihood is

$$\mathcal{L}_n(\beta;\underline{X}) = \sum_{i=1}^{n} \log \binom{m_i}{X_i} + \sum_{i=1}^{n} \left(X_i \log \left(\frac{g(\beta' z_i)}{1-g(\beta' z_i)} \right) + m_i \log(1-g(\beta' z_i)) \right).$$

(iv) Modelling categorical data in a contingency table. Suppose a continency table contains C cells, where each cell gives the number for the corresponding event. Let $1 \leq \ell \leq C$, at each "trial" probability of being placed in cell ℓ is π_ℓ. If we do not make any assumptions on the probabilities (except that each trial are iid random variables) then we model the number of counts in each cell using a multinomial distribution. Suppose the total number of counts is n and the number of counts observed in cell ℓ is X_ℓ, then the distribution is $P(X_1 = x_1, \ldots, X_C = x_c) = \binom{n}{x_1,\ldots,x_C} \pi_1^{x_1} \ldots \pi_C^{x_C}$, which has the log-likelihood

$$\begin{aligned} \mathcal{L}_n(\pi_1, \pi_2, \ldots, \pi_{C-1}; X_1, \ldots, X_C) &= \log \binom{n}{X_1, \ldots, X_C} + \sum_{i=1}^{C} X_i \log \pi_i \\ = \log \binom{n}{X_1, \ldots, X_C} &+ \sum_{i=1}^{C-1} X_i \log \frac{\pi_i}{1-\sum_{j=1}^{C-1} \pi_j} + n \log(1 - \sum_{j=1}^{C-1} \pi_j). \end{aligned}$$

Observe that the parameters and random variables are "separable".

(v) Suppose X is a random variable that only takes integer values, however, there is no upper bound on the number of counts. When there is no upper bound on the number of counts, the Poisson distribution is often used as an alternative to the Binomial. If X follows a Poisson distribution then $P(X = k) = \lambda^k \exp(-\lambda)/k!$. The log-likelihood for the iid Poisson random variables $\{X_i\}$ is

$$\mathcal{L}(\lambda;\underline{X}) = \sum_{i=1}^{n} \left(X_i \log \lambda - \lambda - \log X_i! \right).$$

Observe that the parameters and random variables are "separable".

(vi) Suppose that $\{X_i\}$ are independent exponential random variables which have the density $\theta^{-1} \exp(-x/\theta)$. The log-likelihood is

$$\mathcal{L}_n(\theta;\underline{X}) = \sum_{i=1}^{n} \left(-\log \theta - \frac{X_i}{\theta} \right).$$

(vii) *A generalisation of the exponential distribution which gives more flexibility in terms of shape of the distribution is the Weibull. Suppose that* $\{X_i\}$ *are independent Weibull random variables which have the density* $\frac{\alpha x^{\alpha-1}}{\theta^\alpha}\exp(-(x/\theta)^\alpha)$ *where* $\theta, \alpha > 0$ *(in the case that* $\alpha = 0$ *we have the regular exponential) and* x *is defined over the positive real line. The log-likelihood is*

$$\mathcal{L}_n(\alpha, \theta; \underline{X}) = \sum_{i=1}^{n} \left(\log \alpha + (\alpha - 1) \log X_i - \alpha \log \theta - \left(\frac{X_i}{\theta}\right)^\alpha \right).$$

Observe that the parameters and random variables are not *"separable". In the case, that* α *is known, but* θ *is unknown the likelihood is proportional to*

$$\mathcal{L}_n(\theta; \underline{X};) \propto \sum_{i=1}^{n} \left(- \alpha \log \theta - \left(\frac{X_i}{\theta}\right)^\alpha \right),$$

observe the other terms in the distribution are fixed and do not vary, so are omitted. If α *is known, the unknown parameter and random variables are "separable".*

Often I will exchange $\mathcal{L}(\theta; \underline{X}) = \mathcal{L}(\underline{X}; \theta)$, but they are the *same*.

Look closely at the log-likelihood of iid random variables, what does its average

$$\frac{1}{n}\mathcal{L}(\underline{X}; \theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i; \theta) \tag{1.3}$$

converge to as $n \to \infty$?

1.2 Constructing likelihoods

Constructing the likelihood for the examples given in the previous section was straightforward. However, in many real situations, half the battle is finding the correct distribution and likelihood.

Many of the examples we consider below depend on using a dummy/indicator variable that we treat as a Bernoulli random variables. We recall if δ is a Bernoulli random variable that can take either 0 or 1, where $P(\delta = 1) = \pi$ and $P(\delta = 0) = 1 - \pi$, then $P(\delta = x) = (1 - \pi)^{1-x}\pi^x$. We observe that the log-likelihood for π given δ is $(1-\delta)\log(1-\pi) + \delta \log \pi$. Observe after the log transform, that the random variable and the parameter of interest are "separable".

Mixtures of distributions

Suppose Y is a mixture of two subpopulations, with densities $f_0(x;\theta)$ and $f_1(x;\theta)$ respectively. The probability of belonging to density 0 is $1-p$ and probability of belonging to density 1 is p. Based this information, we can represent the random variable $Y = \delta U + (1-\delta)V$, where U, V, δ are independent random variables; U has density f_1, V has density f_0 and $P(\delta = 1) = p$ and $P(\delta = 0) = 1-p$. The density of Y is

$$f_Y(x;\theta) = f_Y(x|\delta = 0, \theta)P(\delta = 0) + f_Y(x|\delta = 1, \theta)P(\delta = 1) = (1-p)f_0(x;\theta) + pf_1(x;\theta).$$

Thus the log likelihood of θ and p given $\{Y_i\}$ is

$$\mathcal{L}(\{Y_i\};\theta,p) = \sum_{i=1}^{n} \log\left[(1-p)f_0(Y_i;\theta) + pf_1(Y_i;\theta)\right].$$

Observe that the random variables and parameters of interest are not separable.

Suppose we not only observe Y but we observe the mixture the individual belongs to. Not only do we have more information about our parameters, but also estimation becomes easier. To obtain the joint likelihood, we require the joint distribution of (Y, δ), which is a mixture of density and point mass. To derive this we note that by using limiting arguments

$$\begin{aligned}
\lim_{\epsilon\to 0} \frac{P(Y \in [y-\epsilon/2, y+\epsilon/2], \delta = x;\theta,p)}{\epsilon} &= \lim_{\epsilon\to 0} \frac{F_x(y+\epsilon/2) - F_x(y-\epsilon/2)}{\epsilon} P(\delta = x;p) \\
&= f_x(y;\theta)P(\delta = x;p) \\
&= f_1(y;\theta)^x f_0(y;\theta)^{1-x} p^x (1-p)^{1-x}.
\end{aligned}$$

Thus the log-likelihood of θ and p given the joint observations $\{Y_i, \delta_i\}$ is

$$\mathcal{L}(Y_i, \delta_i;\theta,p) = \sum_{i=1}^{n} \left\{\delta_i \log f_1(Y_i;\theta) + (1-\delta_i)\log f_0(Y_i;\theta) + \delta_i \log p + (1-\delta_i)\log(1-p)\right\}. \quad (1.4)$$

The parameters and random variables are separable in this likelihood.

Of course in reality, we do not observe δ_i, but we can predict it, by conditioning on what is observed Y_i. This is effectively constructing the expected log-likelihood of $\{Y_i, \delta_i\}$ conditioned on $\{Y_i\}$. This is not a log-likelihood per se. But for reasons that will become clear later in the course, in certain situations it is useful to derive the expected log-likelihood when conditioned on random variables of interest. We now construct the

expected log-likelihood of $\{Y_i, \delta_i\}$ conditioned on $\{Y_i\}$. Using (1.4) and that $\{Y_i, \delta_i\}$ are independent over i we have

$$\mathrm{E}[\mathcal{L}(Y_i, \delta_i; \theta, p)|\{Y_i\}] = \sum_{i=1}^{n} \left\{ \mathrm{E}[\delta_i|Y_i, \theta, p] \left(\log f_1(Y_i;\theta) + \log p\right) + \mathrm{E}[(1-\delta_i)|Y_i] \left(\log f_0(Y_i;\theta) + \log(1-p)\right) \right\}.$$

$\mathrm{E}[\delta_i|Y_i, \theta, p] = P[\delta_i = 1|Y_i, \theta, p]$, hence it measures the probability of the mixture 1 being chosen when Y_i is observed and is

$$P[\delta_i = 1|Y_i, \theta, p] = \frac{P[\delta_i = 1, Y_i, \theta, p]}{P[Y_i, \theta, p]} = \frac{P[Y_i|\delta_i = 1, \theta, p]P(\delta_i = 1, \theta, p)}{P[Y_i, \theta, p]} = \frac{pf_1(Y_i;\theta)}{pf_1(Y_i;\theta) + (1-p)f_0(Y_i;\theta)}.$$

Similarly

$$P[\delta_i = 0|Y_i, \theta, p] = \frac{(1-p)f_0(Y_i;\theta)}{pf_1(Y_i;\theta) + (1-p)f_0(Y_i;\theta)}.$$

Substituting these in the the above gives the expected log-likelihood conditioned on $\{Y_i\}$;

$$\begin{aligned} \mathrm{E}[\mathcal{L}(Y_i, \delta_i; \theta, p)|\{Y_i\}] &= \sum_{i=1}^{n} \Bigg\{ \left(\frac{pf_1(Y_i;\theta)}{pf_1(Y_i;\theta) + (1-p)f_0(Y_i;\theta)} \right) \left(\log f_1(Y_i;\theta) + \log p\right) + \\ &\quad \left(\frac{(1-p)f_0(Y_i;\theta)}{pf_1(Y_i;\theta) + (1-p)f_0(Y_i;\theta)} \right) \left(\log f_0(Y_i;\theta) + \log(1-p)\right) \Bigg\}. \end{aligned}$$

Observe that this is not in terms of δ_i.

The censored exponential distribution

Suppose $X \sim Exp(\theta)$ (density of X is $f(x;\theta) = \theta \exp(-x\theta)$), however X is censored at a known point c and Y is observed where

$$Y = \begin{cases} X & X \leq c \\ c & X > c \end{cases} \tag{1.5}$$

It is known if an observation is censored. We define the censoring variable

$$\delta = \begin{cases} 0 & X \leq c \\ 1 & X > c \end{cases}$$

The only unknown is θ and we observe (Y, δ). Note that δ is a Bernoulli variable (Binomial with $n = 1$) with $P(\delta = 0) = 1 - \exp(-c\theta)$ and $P(\delta = 1) = \exp(-c\theta)$. Thus the likelihood of θ based only δ is $L(\delta;\theta) = (1-\pi)^{1-\delta}\pi^{\delta} = (1 - e^{-c\theta})^{1-\delta}(e^{-c\theta})^{1-\delta}$.

Analogous to the example above, the likelihood of (Y, δ) is a mixture of a density and a point mass. Thus the likelihood θ based on (Y, δ) is

$$\begin{aligned} L(Y, \delta; \theta) &= \begin{cases} f(Y|\delta = 0)P(\delta = 0) & \delta = 0 \\ f(Y|\delta = 1)P(\delta = 1) & \delta = 1 \end{cases} \\ &= [f(Y|\delta = 0)P(\delta = 0)]^{1-\delta}[f(Y|\delta = 1)P(\delta = 1)]^{\delta} \\ &= [\exp(-\theta Y + \log\theta)]^{1-\delta}[\exp(-c\theta)]^{\delta}. \end{aligned}$$

This yields the log-likelihood of θ given $\{Y_i, \delta_i\}$

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \left\{ (1 - \delta_i) \left[-\theta Y_i + \log\theta \right] - \delta_i c\theta \right\}. \tag{1.6}$$

The inflated zero Poisson distribution

The Possion distribution is commonly used to model count data. However, there arises many situations where the proportion of time zero occurs is larger than the proportion one would expect using a Poisson distribution. One often models this "inflation" using a mixture distribution. Let U be a Poission distributed random variable where $P(U = k) = \lambda^k \exp(-\lambda)/k!$. We see that $P(U = 0) = \exp(-\lambda)$. We can boost this chance by defining a new random variable Y, where

$$Y = \delta U$$

and δ is a Bernoulli random variable taking zero or one with $P(\delta = 0) = p$ and $P(\delta = 1) = (1 - p)$. It is clear that

$$\begin{aligned} P(Y = 0) &= P(Y = 0|\delta = 0)P(\delta = 0) + P(Y = 0|\delta = 1)P(\delta = 1) \\ &= 1 \times p + P(U = 0)(1 - p) = p + (1 - p)e^{-\lambda} \geq e^{-\lambda} = P(U = 0). \end{aligned}$$

Thus, in situations where there are a large number of zeros, the inflated zero Poisson seems appropriate. For $k > 1$, we have

$$\begin{aligned} P(Y = k) &= P(Y = k|\delta = 0)P(\delta = 0) + P(Y = k|\delta = 1)P(\delta = 1) \\ &= P(U = k)(1 - p) = (1 - p)\frac{\lambda^k e^{-\lambda}}{k!}. \end{aligned}$$

Thus altogether the distribution of Y is

$$P(Y = k) = \left\{ p + (1 - p)e^{-\lambda} \right\}^{I(k=0)} \left\{ (1 - p)\frac{\lambda^k e^{-\lambda}}{k!} \right\}^{I(k\neq 0)},$$

where $I(\cdot)$ denotes the indicator variable. Thus the log-likelihood of λ, p given $\underline{Y}$ is

$$\mathcal{L}(\underline{Y}; \lambda, p) = \sum_{i=1}^{n} I(Y_i = 0) \log \left(p + (1-p)e^{-\lambda}\right) + \sum_{i=1}^{n} I(Y_i \neq 0) \left(\log(1-p) + \log \frac{\lambda^{Y_i} e^{-\lambda}}{Y_i!}\right).$$

Exercise 1.1 *Let us suppose that X and Z are independent random variables with densities f_X and f_Z respectively. Assume that X is positive.*

(i) Derive the density function of $1/X$.

(ii) Show that the density of XZ is

$$\int \frac{1}{x} f_Z(\frac{y}{x}) f_X(x) dx \tag{1.7}$$

(or equivalently $\int c^{-1} f_Z(cy) f_X(c^{-1}) dc$).

(iii) Consider the linear regression model

$$Y_i = \underline{\alpha}' \underline{x}_i + \sigma_i \varepsilon_i$$

where the regressors x_i is observed, ε_i follows a standard normal distribution (mean zero and variance 1) and σ_i^2 follows a Gamma distribution

$$f(\sigma^2; \lambda) = \frac{\sigma^{2(\kappa-1)} \lambda^{\kappa} \exp(-\lambda \sigma^2)}{\Gamma(\kappa)}, \quad \sigma^2 \geq 0,$$

with $\kappa > 0$.

Derive the log-likelihood of Y_i (assuming the regressors are observed).

Exercise 1.2 *Suppose we want to model the average amount of daily rainfall in a particular region. Empirical evidence suggests that it does not rain on many days in the year. However, if it does rain on a certain day, the amount of rain follows a Gamma distribution.*

(i) Let Y denote the amount of rainfall in a particular day and based on the information above write down a model for Y.

Hint: Use the ideas from the inflated zero Poisson model.

(ii) Suppose that $\{Y_i\}_{i=1}^n$ is the amount of rain observed n consecutive days. Assuming that $\{Y_i\}_{i=1}^n$ are iid random variables with the model given in part (ii), write down the log-likelihood for the unknown parameters.

(iii) Explain why the assumption that $\{Y_i\}_{i=1}^n$ are independent random variables is tenuous.

1.3 Bounds for the variance of an unbiased estimator

So far we have iid observations $\{X_i\}$ with from a known parametric family i.e. the distribution of X_i comes from $\mathcal{F} = \{f(x;\theta); \theta \in \Theta\}$, where θ is a finite dimension parameter however the true θ is unknown. There are an infinite number of estimators of θ based on an infinite number of decision rules. Which estimator do we choose? We should choose the estimator which is "closest" to the true parameter. There are several different distance measures, but the most obvious is the mean square error. As the class of all estimators is "too large" we restrict ourselves to unbiased estimators, $\widetilde{\theta}(\underline{X})$ (where mean of estimator is equal to the true parameter) and show that the mean squared error

$$\mathrm{E}\left(\widetilde{\theta}(\underline{X}) - \theta\right)^2 = \mathrm{var}\left(\widetilde{\theta}(\underline{X})\right) + \left(\mathrm{E}[\widetilde{\theta}(\underline{X})] - \theta\right)^2 = \mathrm{var}\left(\widetilde{\theta}(\underline{X})\right)$$

is bounded below by the inverse of the Fisher information (this is known as the Cramer-Rao bound). To show such a bound we require the regularity assumptions. We state the assumptions and in the case that θ is a scalar, but they can easily be extended to the case that θ is a vector.

Assumption 1.3.1 (Regularity Conditions 1) *Let us suppose that $L_n(\cdot;\theta)$ is the likelihood.*

(i) $\frac{\partial}{\partial\theta}\int L_n(\underline{x};\theta)d\underline{x} = \int \frac{\partial L_n(\underline{x};\theta)}{\partial\theta}d\underline{x} = 0$ *(for iid random variables (rv) this is equivalent to checking if* $\int \frac{\partial f(x;\theta)}{\partial\theta}dx = \frac{\partial}{\partial\theta}\int f(x;\theta)dx$*).*

Observe since a by definition a density integrates to one, then $\frac{\partial}{\partial\theta}\int L_n(\underline{x};\theta)d\underline{x} = 0$*.*

(ii) For any function g not a function of θ, $\frac{\partial}{\partial\theta}\int g(\underline{x})L_n(\underline{x};\theta)d\underline{x} = \int g(\underline{x})\frac{\partial L_n(\underline{x};\theta)}{\partial\theta}d\underline{x}$*.*

(iii) $\mathrm{E}\big(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\big)^2 > 0$*.*

To check Assumption 1.3.1(i,ii) we need to apply Leibniz's rule `https://en.wikipedia.org/wiki/Leibniz_integral_rule`

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} g(x)f(x;\theta)dx = \int_{a(\theta)}^{b(\theta)} g(x)\frac{\partial f(x,\theta)}{\partial\theta}dx + f(b(\theta),\theta)g(b(\theta))b'(\theta) - f(a(\theta),\theta)g(a(\theta))a'(\theta) \quad (1.8)$$

Therefore Assumption 1.3.1(i,ii) holds if $f(b(\theta),\theta)g(b(\theta))b'(\theta) - f(a(\theta),\theta)g(a(\theta))a'(\theta) = 0$.

Example 1.3.1 *(i) If the support of the density does not depend on θ it is clear from (1.8) that Assumption 1.3.1(i,ii) is satisfied.*

(ii) If the density is the uniform distribution $f(x;\theta) = \theta^{-1} I_{[0,\theta]}(x)$ *then the conditions are not satisfied. We know that* $\theta^{-1}\int_0^\theta dx = 1$ *(thus it is independent of* θ*) hence* $\frac{d\theta^{-1}\int_0^\theta dx}{d\theta} = 0$*. However,*

$$\int_0^\theta \frac{d\theta^{-1}}{d\theta} dx = \frac{-1}{\theta} \text{ and } f(b(\theta),\theta)b'(\theta) - f(a(\theta),\theta)a'(\theta) = \theta^{-1}.$$

Thus we see that Assumption 1.3.1(i) is not satisfied. Therefore, the uniform distribution does not satisfy the standard regularity conditions.

(iii) Consider the density

$$f(x;\theta) = \frac{1}{2}(x-\theta)^2 \exp[-(x-\theta)] I_{[\theta,\infty)}(x).$$

The support of this estimator depends on θ*, however, it does satisfy the regularity conditions. This is because* $f(x;\theta) = 0$ *at both* $x = \theta$ *and* $x = \infty$*. This means that for any* θ

$$f(b(\theta),\theta)g(b(\theta))b'(\theta) - f(a(\theta),\theta)g(a(\theta))a'(\theta) = 0.$$

Therefore from the Leibnitz rule we have

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} g(x)f(x;\theta)dx = \int_{a(\theta)}^{b(\theta)} g(z)\frac{\partial f(x,\theta)}{\partial\theta}dx.$$

Thus Assumption 1.3.1 is satisfied.

We now state the Cramer-Rao bound, which gives the minimal attaining variance bound for a large class of estimators. We will use the matrix inequality $A \geq B$ to mean that $A - B$ is a non-negative definite matrix (or equivalently positive semi-definite).

Theorem 1.3.1 (The Cramér-Rao bound) *Suppose the likelihood* $L_n(\underline{X};\theta)$ *satisfies the regularity conditions given in Assumption 1.3.1. Let* $\widetilde{\theta}(\underline{X})$ *be an unbiased estimator of* θ*, then*

$$\text{var}\left[\widetilde{\theta}(\underline{X})\right] \geq \left[\text{E}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right)^2\right]^{-1}.$$

PROOF. We prove the result for the univariate case. Recall that $\widetilde{\theta}(X)$ is an unbiased estimator of θ therefore

$$\int \widetilde{\theta}(\underline{x}) L_n(\underline{x};\theta) d\underline{x} = \theta.$$

Differentiating both sides wrt to θ, and taking the derivative into the integral (allowed under the regularity condition) gives

$$\int \widetilde{\theta}(\underline{x}) \frac{\partial L_n(\underline{x};\theta)}{\partial \theta} d\underline{x} = 1.$$

By Assumption 1.3.1(i) $\frac{d\int L_n(\underline{x};\theta)d\underline{x}}{d\theta} = \int \frac{\partial L_n(\underline{x};\theta)}{\partial \theta} d\underline{x} = 0$. Thus adding $\theta \int \frac{\partial L_n(\underline{x};\theta)}{\partial \theta} d\underline{x}$ to both sides of the above we have

$$\int \left\{\widetilde{\theta}(\underline{x}) - \theta\right\} \frac{\partial L_n(\underline{x};\theta)}{\partial \theta} d\underline{x} = 1.$$

Multiplying and dividing by $L_n(\underline{x};\theta)$ gives

$$\int \left\{\widetilde{\theta}(\underline{x}) - \theta\right\} \frac{1}{L_n(\underline{x};\theta)} \frac{\partial L_n(\underline{x};\theta)}{\partial \theta} L_n(\underline{x};\theta) dx = 1. \tag{1.9}$$

Hence (since $L_n(\underline{x};\theta)$ is the distribution of $\underline{X}$) we have

$$\mathrm{E}\left(\left\{\widetilde{\theta}(\underline{X}) - \theta\right\} \frac{\partial \log L_n(\underline{X};\theta)}{\partial \theta} \right) = 1.$$

Recalling that the Cauchy-Schwartz inequality is $\mathrm{E}(UV) \leq \mathrm{E}(U^2)^{1/2}\mathrm{E}(V^2)^{1/2}$ (where equality only arises if $U = aV + b$ (where a and b are constants)) and applying it to the above we have

$$\mathrm{var}\left[\widetilde{\theta}(\underline{X})\right] \mathrm{E}\left[\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial \theta}\right)^2\right] \geq 1 \qquad \Rightarrow \mathrm{var}\left[\widetilde{\theta}(\underline{X})\right] \geq \mathrm{E}\left[\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial \theta}\right)^2\right]^{-1}.$$

Thus giving the Cramer-Rao inequality. □

Corollary 1.3.1 (Estimators which attain the Cramér-Rao bound) *Suppose Assumption 1.3.1 is satisfied. Then the estimator $\widetilde{\theta}(\underline{X})$ attains the Cramer-Rao bound only if it can be written as*

$$\hat{\theta}(\underline{X}) = a(\theta) + b(\theta) \frac{\partial \log L_n(\underline{X};\theta)}{\partial \theta}$$

for some functions $a(\cdot)$ and $b(\cdot)$ of θ[1].

[1]Of course, in most cases it makes no sense to construct an estimator of θ, which involves θ.

PROOF. The proof is clear and follows from when the Cauchy-Schwartz inequality is an equality in the derivation of the Cramer-Rao bound. □

We next derive an equivalent expression for $\mathrm{E}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right)^2$ (called the Fisher information).

Lemma 1.3.1 *Suppose the likelihood $L_n(\underline{X};\theta)$ satisfies the regularity conditions given in Assumption 1.3.1 and for all $\theta \in \Theta$, $\frac{\partial^2}{\partial\theta^2}\int g(\underline{x})L_n(\underline{x};\theta)d\underline{x} = \int g(\underline{x})\frac{\partial^2 L_n(\underline{x};\theta)}{\partial\theta^2}d\underline{x}$, where g is any function which is not a function of θ (for example the estimator of θ). Then*

$$\mathrm{var}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right) = \mathrm{E}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right)^2 = -\mathrm{E}\left(\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial\theta^2}\right).$$

PROOF. To simplify notation we focus on the case that the dimension of the vector θ is one. To prove this result we use the fact that L_n is a density to obtain

$$\int L_n(\underline{x};\theta)d\underline{x} = 1.$$

Now by differentiating the above with respect to θ gives

$$\frac{\partial}{\partial\theta}\int L_n(\underline{x};\theta)d\underline{x} = 0.$$

By using Assumption 1.3.1(ii) we have

$$\int \frac{\partial L_n(\underline{x};\theta)}{\partial\theta}d\underline{x} = 0 \Rightarrow \quad \int \frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}L_n(\underline{x};\theta)d\underline{x} = 0$$

Differentiating again with respect to θ and taking the derivative inside gives

$$\begin{aligned}
&\int \frac{\partial^2 \log L_n(\underline{x};\theta)}{\partial\theta^2}L_n(\underline{x};\theta)d\underline{x} + \int \frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\frac{\partial L_n(\underline{x};\theta)}{\partial\theta}d\underline{x} = 0 \\
\Rightarrow\ &\int \frac{\partial^2 \log L_n(\underline{x};\theta)}{\partial\theta^2}L_n(\underline{x};\theta)d\underline{x} + \int \frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\frac{1}{L_n(\underline{x};\theta)}\frac{\partial L_n(\underline{x};\theta)}{\partial\theta}L_n(\underline{x};\theta)d\underline{x} = 0 \\
\Rightarrow\ &\int \frac{\partial^2 \log L_n(\underline{x};\theta)}{\partial\theta^2}L_n(\underline{x};\theta)d\underline{x} + \int \left(\frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)d\underline{x} = 0
\end{aligned}$$

Thus

$$-\mathrm{E}\left(\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial\theta^2}\right) = \mathrm{E}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right)^2.$$

The above proof can easily be generalized to parameters θ, with dimension larger than 1. This gives us the required result.

Note in all the derivations we are evaluating the second derivative of the likelihood at the *true parameter*. □

We mention that there exists distributions which do not satisfy Assumption 1.3.1. These are called *non-regular distributions*. The Cramer-Rao lower bound does hold for such distributions.

Definition 1.3.1 (The Fisher information matrix) *The matrix*

$$I(\theta) = \left[\mathrm{E}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial \theta} \right)^2 \right] = -\left[\mathrm{E}\left(\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial \theta^2} \right) \right],$$

whose inverse forms the lower bound of Cramér-Rao bound is called the Fisher information matrix. It plays a critical role in classical inference.

Essentially $I(\theta)$ tells us how much "information" the data $\{X_i\}_{i=1}^n$ contains about the true *parameter θ.*

Remark 1.3.1 *Define the quantity*

$$\begin{aligned} I_{\theta_0}(\theta) &= -\int \left(\frac{\partial^2 \log L_n(\underline{x};\theta)}{\partial \theta^2} \right) L_n(\underline{x};\theta_0) d\underline{x} \\ &= -\left[\mathrm{E}_{\theta_0}\left(\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial \theta^2} \right) \right]. \end{aligned}$$

This quantity evaluates the negative expected second derivative of the log-likelihood over θ, but the expectation is taken with respect to the "true" density $L_n(\underline{x};\theta_0)$. This quantity will not be positive for all θ. However, by the result above we evaluate $I_{\theta_0}(\theta)$ at $\theta = \theta_0$, then

$$I_{\theta_0}(\theta_0) = \mathrm{var}_{\theta_0}\left(\frac{\partial \log L_n(\underline{x};\theta_0)}{\partial \theta} \right).$$

In other words, when the expectation of the negative second derivative of log-likelihood is evaluated at the true *parameter this is the Fisher information which is positive.*

Exercise 1.3 *Suppose $\{X_i\}$ are iid random variables with density $f(x;\theta)$ and the Fisher information for θ based on $\{X_i\}$ is $I(\theta)$.*

Let $Y_i = g(X_i)$ where $g(\cdot)$ is a bijective diffeomorphism (the derivatives of g and its inverse exist). Intuitive when one makes such a transformation no "information" about θ should be lost or gained. Show that the Fisher information matrix of θ based on $\{Y_i\}$ is $I(\theta)$.

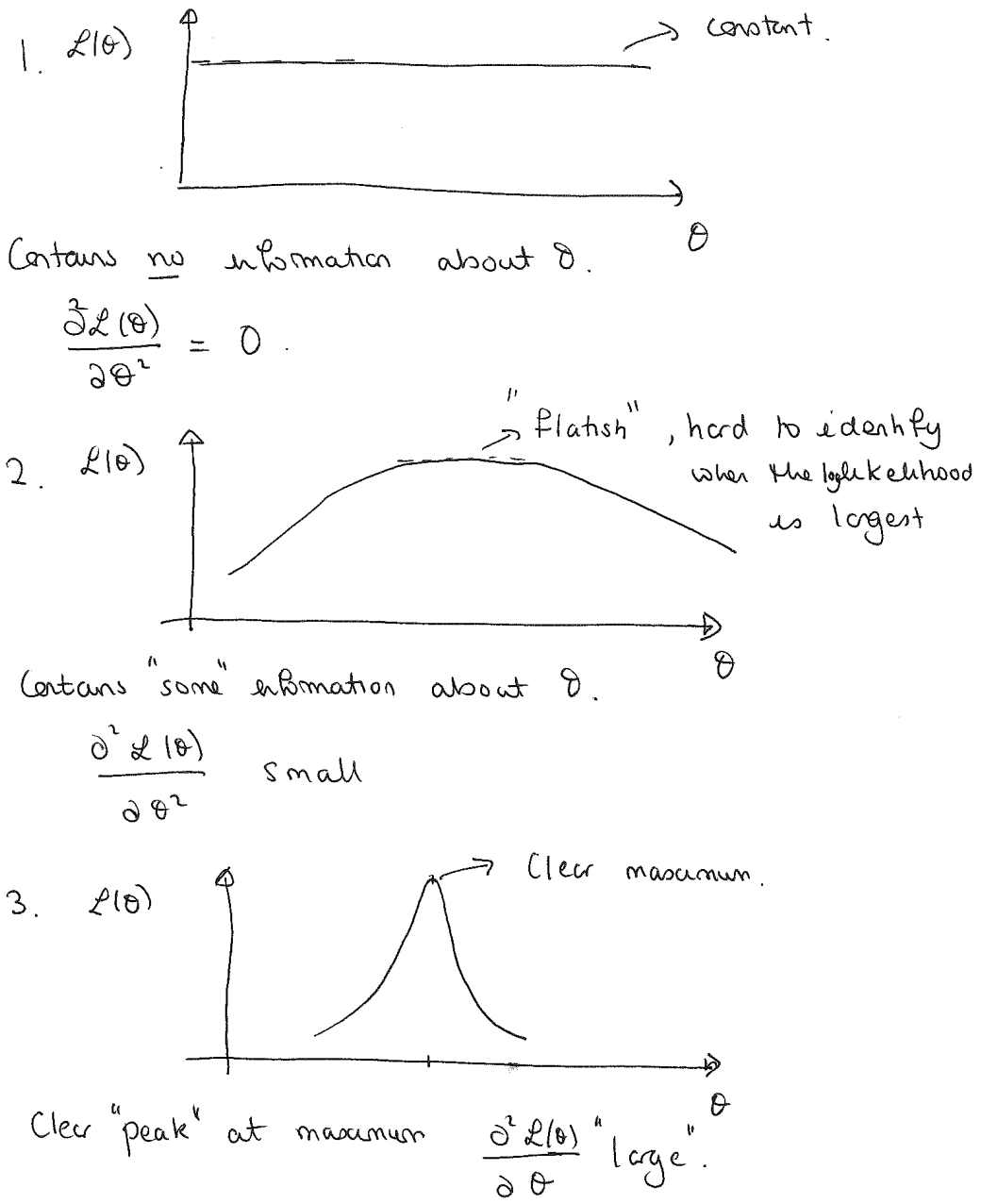

Figure 1.1: Interpretation of the Fisher information

Example 1.3.2 *Consider the example of censored data given in Section 1.2. Both the observations and the censored variables, $\{Y_i\}$ and $\{\delta_i\}$, where*

$$\delta_i = I(Y_i \geq c)$$

contain information about the parameter θ. However, it seems reasonable to suppose that $\{Y_i\}$ contains more information about θ than $\{\delta_i\}$. We articulate what we mean by this in the result below.

Lemma 1.3.2 *Let us suppose that the log-likelihood of $\underline{X}$, $\mathcal{L}_n(\underline{X};\theta)$ satisfies Assumption 1.3.1. Let $\underline{Y} = B(\underline{X})$ be some statistic (of arbitrary dimension) of the original data. Let $\mathcal{L}_{B(\underline{X})}(\underline{Y};\theta)$, and $\mathcal{L}(\underline{X}|\underline{Y};\theta)$ denote the log-likelihood of $\underline{Y} = B(\underline{X})$ and conditional likelihood of $\underline{X}|\underline{Y}$ (we assume these satisfy Assumption 2.6.1, however I think this is automatically true). Then*

$$I_{\underline{X}}(\theta) \geq I_{B(\underline{X})}(\theta)$$

where

$$I_{\underline{X}}(\theta) = \mathrm{E}\left(\frac{\partial \mathcal{L}_{\underline{X}}(\underline{X};\theta)}{\partial\theta}\right)^2 \text{ and } I_{B(\underline{X})}(\theta) = \mathrm{E}\left(\frac{\partial \mathcal{L}_{B(\underline{X})}(\underline{Y};\theta)}{\partial\theta}\right)^2.$$

In other words the original Fisher information contains the most information about the parameter. In general, most transformations of the data will lead to a loss in information. We consider some exceptions in Lemma 1.4.1.

PROOF. Writing the conditional density of $\underline{X}$ given $B(\underline{X})$ as the ratio of a joint density of $\underline{X}, B(\underline{X})$ and marginal density of $B(\underline{X})$ we have

$$f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y}) = \frac{f_{\underline{X},B(\underline{X})}(\underline{x},\underline{y};\theta)}{f_{B(\underline{X})}(\underline{y};\theta)} \Rightarrow f_{\underline{X},B(\underline{X})}(\underline{x},\underline{y};\theta) = f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y}) f_{B(\underline{X})}(\underline{y};\theta),$$

where $f_{\underline{X}|B(\underline{X})}$ denotes the density of $\underline{X}$ conditioned on $B(\underline{X})$ and $f_{\underline{X},B(\underline{X})}$ the joint density of $\underline{X}$ and $B(\underline{X})$. Note that if $B(\underline{x}) = y$, then the joint density $f_{\underline{X},B(\underline{X})}(\underline{x},y;\theta)$ is simply the density of $f_{\underline{X}}(\underline{x};\theta)$ with the constraint that $\underline{y} = B(\underline{x})$ i.e. $f_{\underline{X},B(\underline{X})}(\underline{x},\underline{y};\theta) = f_{\underline{X}}(\underline{x};\theta)\delta(B(\underline{x}) = \underline{y})$, where δ denotes the indicator variable[2]. Thus we have

$$f_{\underline{X}}(\underline{x};\theta)\delta(B(\underline{x}) = \underline{y}) = f_{\underline{X}|B(\underline{X})}(\underline{x}|y,\theta) f_{B(\underline{X})}(\underline{y};\theta).$$

[2]To understand why, consider the joint density of $X, Y = B(X)$ the density ie not defined over $\mathbb{R}^2$ but over the curve $(x, B(x))$ $f_{X,B(X)}(x,y) = f_X(x)\delta(y = B(x))$

Having written the likelihood in this way, the derivative of the log likelihood is

$$\begin{aligned}\frac{\partial \log f_{\underline{X}}(\underline{x};\theta)}{\partial\theta} &= \frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y})f_{B(\underline{X})}(\underline{y};\theta)}{\partial\theta} \\ &= \frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y},\theta)}{\partial\theta} + \frac{\partial \log f_{B(\underline{X})}(\underline{y};\theta)}{\partial\theta}.\end{aligned}$$

Therefore

$$\begin{aligned}I_{\underline{X}}(\theta) = \mathrm{var}\left(\frac{\partial \log f_{\underline{X}}(\underline{X};\theta)}{\partial\theta}\right) &= \mathrm{var}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\right) + \underbrace{\mathrm{var}\left(\frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\right)}_{=I_{B(\underline{X})}(\theta)} + \\ &\quad 2\mathrm{cov}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}, \frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\right). \quad (1.10)\end{aligned}$$

Under the stated regularity conditions, since $f_{B(\underline{X})}$, is a density it is clear that

$$\mathrm{E}\left(\frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\right) = 0$$

and

$$\mathrm{E}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\big|B(\underline{X})\right) = \int \frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y},\theta)}{\partial\theta} f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y},\theta)d\underline{x} = 0. \quad (1.11)$$

Thus using the law of iterated expectation $\mathrm{E}(A) = \mathrm{E}(\mathrm{E}[A|B])$, then $\mathrm{E}[\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}] = 0$. Returning to (1.10), since the mean is zero this implies that

$$I_{\underline{X}}(\theta) = I_{B(\underline{X})}(\theta) + \mathrm{E}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\right)^2 + 2\mathrm{E}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\frac{\partial \log f_{B(\underline{X})}(\underline{Y};\theta)}{\partial\theta}\right).$$

Finally we show that the above covariance is zero. To do so we use that $\mathrm{E}(XY) = \mathrm{E}(X\mathrm{E}[Y|X])$ (by the law of iterated expectation) then by using (1.11) we have

$$\begin{aligned}&\mathrm{E}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\right) \\ = \mathrm{E}&\left(\frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\underbrace{\mathrm{E}\left[\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\Big|\frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\right]}_{=0 \text{ by (1.11)}}\right) = 0.\end{aligned}$$

Thus

$$I_{\underline{X}}(\theta) = I_{B(\underline{X})}(\theta) + \mathrm{E}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\right)^2.$$

As all the terms are positive, this immediately implies that $I_{\underline{X}}(\theta) \geq I_{B(\underline{X})}(\theta)$. □

Definition 1.3.2 (Observed and Expected Fisher Information) *(i) The observed Fisher information matrix is defined as*

$$I(\underline{X};\theta) = -\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial \theta^2}.$$

(ii) The expected Fisher information matrix is defined as

$$I(\theta) = \mathrm{E}\left(-\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial \theta}\right)$$

These will play an important role in inference for parameters.

Often we want to estimate a function of θ, $\tau(\theta)$. The following corollary is a generalization of the Cramer-Rao bound.

Corollary 1.3.2 *Suppose Assumption 1.3.1 is satisfied and $T(\underline{X})$ is an unbiased estimator of $\tau(\theta)$. Then we have*

$$\mathrm{var}\left[T(\underline{X})\right] \geq \frac{\tau'(\theta)^2}{\mathrm{E}\left[\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial \theta}\right)^2\right]}.$$

Exercise 1.4 *Prove the above corollary.*

In this section we have learnt how to quantify the amount of information the data contains about a parameter and show that for the majority of transformations of data (with the exception of bijections) we loose information. In the following section we define a transformation of data, where in some certain situations, will substantially reduce the dimension of the data, but will not result in a loss of information.

1.4 Sufficient statistics

We start with a simple example from introductory statistics.

Example 1.4.1 *Samples of size 10 and 15 are drawn from two different distributions. How to check if the two samples come from the same distribution? The data is given in Figure 1.2. If the distributions are known to come from the Gaussian family of distributions with, for the sake of argument, standard deviation one, then all the information about the unknown parameter, is characteristized in terms of the sample means $\bar{X}_A$ and*

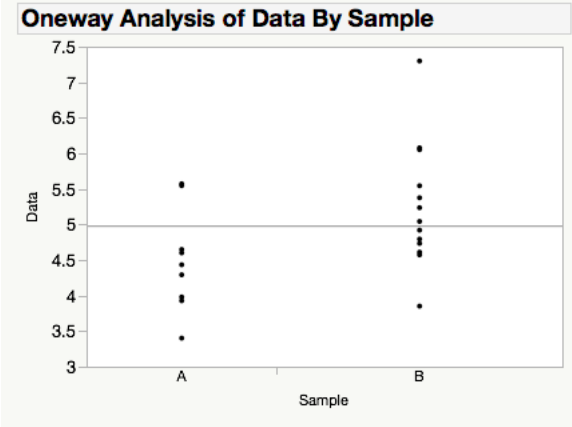

Figure 1.2: Samples from two population

$\bar{X}_B$ *(in this example,* 4.6 *and* 5.2 *respectively). The sample mean is* sufficient *for describing all the information about the unknown mean, more precisely the data conditioned on sample mean is free of any information about* μ.

On the other hand, if the data comes from the Cauchy family of distributions $\{f_\theta(x) = [\pi(1+(x-\theta)^2)]^{-1}\}$ *there does not exist a lower dimensional transformations of the data which contains all the information about* θ. *The observations conditioned on any lower transformation will still contain information about* θ.

This example brings us to a formal definition of sufficiency.

Definition 1.4.1 (Sufficiency) *Suppose that* $\underline{X} = (X_1, \ldots, X_n)$ *is a random vector. A statistic* $s(\underline{X})$ *is said to be sufficient for the family* $\mathcal{F}$ *of distributions, if the conditional density* $f_{\underline{X}|S(\underline{X})}(y|s)$ *is the same for all distributions in* $\mathcal{F}$.

This means in a parametric class of distributions $\mathcal{F} = \{f(\underline{x};\theta); \theta \in \Theta\}$ *the statistic* $s(\underline{X})$ *is sufficient for the parameter* θ, *if the conditional distribution of* $\underline{X}$ *given* $s(\underline{X})$ *is not a function of* θ.

Example 1.4.2 (Order statistics) *Suppose that* $\{X_i\}_{i=1}^n$ *are iid random variables with density* $f(x)$. *Let* $X_{(1)}, \ldots, X_{(n)}$ *denote the ordered statistics (i.e.* $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$*). We will show that the order statistics* $X_{(1)}, \ldots, X_{(n)}$ *is the sufficient statistic over the family of all densities* $\mathcal{F}$.

To see why, note that it can be shown that the joint density of the order statistics

$X_{(1)}, \ldots, X_{(n)}$ *is*

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = \begin{cases} n! \prod_{i=1}^n f(x_i) & x_1 \leq \ldots, \leq x_n \\ 0 & \text{otherwise} \end{cases} \tag{1.12}$$

Clearly the density of the $X_1, \ldots, X_n$ *is* $\prod_{i=1}^n f(X_i)$*. Therefore the density of* $X_1, \ldots, X_n$ *given* $X_{(1)}, \ldots, X_{(n)}$ *is*

$$f_{X_1,\ldots,X_n|X_{(1)},\ldots,X_{(n)}} = \frac{1}{n!},$$

which is simply the chance of selecting the ordering $X_1, \ldots, X_n$ *from a sequence* $X_{(1)}, \ldots, X_{(n)}$*. Clearly this density does not depend on any distribution.*

This example is interesting, but statistically not very useful. In general we would like the number of sufficient statistic to be a lower dimension than the data itself (sufficiency is a form of compression).

Exercise 1.5 *Show (1.12).*

Usually it is extremely difficult to directly obtain a sufficient statistic from its definition. However, the factorisation theorem gives us a way of obtaining the sufficient statistic.

Theorem 1.4.1 (The Fisher-Neyman Factorization Theorem) *A necessary and sufficient condition that* $s(\underline{X})$ *is a sufficient statistic is that the likelihood function,* L *(not log-likelihood), can be factorized as* $L_n(\underline{X}; \theta) = h(\underline{X})g(s(\underline{X}); \theta)$*, where* $h(\underline{X})$ *is not a function of* θ*.*

Example 1.4.3 (The uniform distribution) *Let us suppose that* $\{X_i\}$ *are iid uniformly distributed random variables with density* $f_\theta(x) = \theta^{-1} I_{[0,\theta]}(x)$*. The likelihood is*

$$L_n(\underline{X}; \theta) = \frac{1}{\theta^n} \prod_{i=1}^n I_{[0,\theta]}(X_i) = \frac{1}{\theta^n} I_{[0,\theta]}(\max X_i) = g(\max X_i; \theta)$$

Since $L_n(\underline{X}; \theta)$ *is only a function of* $\max_i X_i$*, it is immediately clear that* $s(\underline{X}) = \max_i X_i$ *is a sufficient.*

Example 1.4.4 (The normal distribution) *Let $\{X_i\}_{i=1}^n$ be iid normal random variables. The likelihood is*

$$\begin{aligned} L_n(\underline{X};\mu,\sigma^2) &= \frac{1}{(2\pi\sigma)^n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n(X_i-\mu)^2\right] = \frac{1}{(2\pi\sigma)^n}\exp\left[-\frac{1}{2\sigma^2}\left(S_{xx}-2S_x\mu+\mu^2\right)\right] \\ &= g(S_x,S_{xx};\mu,\sigma^2) \end{aligned}$$

where $S_x=\sum_{i=1}^n X_i$ and $S_{xx}=\sum_{i=1}^n X_i^2$. We see immediately from the factorisation theorem that the density is a function of two sufficient statistics S_x and S_{xx}. Thus S_x and S_{xx} are the sufficient statistics for μ and σ^2.

Suppose we treat σ^2 as known, then by using

$$\begin{aligned} L_n(\underline{X};\mu,\sigma^2) &= \frac{1}{(2\pi\sigma)^n}\exp\left[-\frac{1}{2\sigma^2}S_{xx}\right]\exp\left[-\frac{S_x\mu}{\sigma^2}+\frac{\mu^2}{\sigma^2}\right] \\ &= g_1(S_{xx};\sigma^2)g_2(S_x;\mu,\sigma^2) \end{aligned}$$

we see that the sufficient statistic for the mean, μ, is $S_x=\sum_{i=1}^n X_i$. I.e. any function of $\{X_i\}$ conditioned on S_x contains no information about the mean μ. This includes the S_{xx}. However, S_x contains information about the both μ and σ^2. We can explicitly see this because $S_x\sim N(n\mu,n\sigma^2)$.

Note that alternative sufficient statistics for the normal distribution are $S_x=\sum_i X_i$ and $S'_{xx}=\sum_i(X_i-n^{-1}S_x)^2$. Sufficient statistics are not unique!

Example 1.4.5 (The exponential family) *The exponential family of distributions, characterized by*

$$f(x;\omega)=\exp\left[s(x)\eta(\omega)-b(\omega)+c(x)\right], \tag{1.13}$$

is broad class of distributions which includes the normal distributions, binomial, exponentials etc. but not the uniform distribution. Suppose that $\{X_i\}_{i=1}^n$ are iid random variables which have the form (1.13) We can write and factorize the likelihood as

$$\begin{aligned} L_n(\underline{X};\omega) &= \exp\left[\eta(\omega)\sum_{i=1}^n s(X_i)-nb(\omega)\right]\exp\left[\sum_{i=1}^n c(X_i)\right] \\ &= g(\sum_{i=1}^n s(X_i);\omega)h(X_1,\ldots,X_n). \end{aligned}$$

We immediately see that $\sum_{i=1}^n s(X_i)$ is a sufficient statistic for ω.

The above example is for the case that the number of parameters is one, however we can generalize the above to the situation that the number of parameters in the family is p

$$f(x;\omega) = \exp\left[\sum_{j=1}^{p} s_j(x)\eta_j(\omega) - b(\omega) + c(x)\right],$$

where $\omega = (\omega_1, \ldots, \omega_p)$. *The sufficient statistics for the p-dimension is* $(\sum_{i=1}^n s_1(X_i), \ldots, \sum_{i=1}^n s_p(X_i))$. *Observe, we have not mentioned, so far, about this being in anyway minimal, that comes later.*

For example, the normal distribution is parameterized by two parameters; mean and variance. Typically the number of sufficient statistics is equal to the the number of unknown parameters. However there can arise situations where the number of sufficient statistics is more than the number of unknown parameters.

Example 1.4.6 *Consider a mixture model, where we know which distribution a mixture comes from. In particular, let* $g_0(\cdot;\theta)$ *and* $g_1(\cdot;\theta)$ *be two different densities with unknown parameter* θ. *Let* δ *be a Bernoulli random variables which takes the values* 0 *or* 1 *and the probability* $P(\delta = 1) = 1/2$. *The random variables* (X, δ) *have the joint "density"*

$$\begin{aligned} f(x,\delta;\theta) &= \begin{cases} \frac{1}{2}g_0(x;\theta) & \delta = 0 \\ \frac{1}{2}g_1(x;\theta) & \delta = 1 \end{cases} \\ &= (1-\delta)\frac{1}{2}g_0(x;\theta) + \delta\frac{1}{2}g_1(x;\theta) = (\frac{1}{2}g_0(x;\theta))^{1-\delta}(\frac{1}{2}g_1(x;\theta))^{\delta}. \end{aligned}$$

Example; the population of males and females where we observe the gender and height of an individual. Both (X, δ) *are the sufficient statistics for* θ. *Observe that* X *by itself is not sufficient because*

$$P(\delta|X = x) = \frac{g_1(x;\theta)}{g_0(x;\theta) + g_1(x;\theta)}.$$

Hence conditioned on just X, *the distribution of* δ *contains information about* θ, *implying* X *by itself is not sufficient.*

Remark 1.4.1 (Ancillary variables) *The above example demonstrates the role of an* ancillary variable. *If we observe only* X, *since the marginal density of* X *is*

$$\frac{1}{2}g_0(x;\theta) + \frac{1}{2}g_1(x;\theta),$$

then X contains information about θ. On the other hand, if we only observe δ, it contains no information about θ (the marginal distribution of δ is half). This means that θ is an ancillary variable (since its marginal distribution contains no information about θ).

Furthermore, since (X, δ) are the sufficient statistics for θ, δ is an ancillary variable and δ in conjuction with X does contain information about θ then δ is called an ancillary complement.

We already came across an ancillary variable. We recall that for the normal distribution one version of the sufficient statistics is $S_x = \sum_i X_i$ and $S'_{xx} = \sum_i (X_i - n^{-1}S_x)^2$. Now we see that $S'_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2 \sim \sigma^2 \chi^2_{n-1}$, hence it is an ancillary variable for the mean, since its marginal distribution does not depend on μ. However, it is not an ancillary complement for S_x since S'_{xx} conditioned on S_x does not depend on μ in fact they are independent! So S'_{xx} conditioned or otherwise contains no information whatsoever about the mean μ.

From the examples above we immediately see that the sufficient statistic is not unique. For example, for the Gaussian family of distributions the order statistics $X_{(1)}, \ldots, X_{(n)}$, S_{xx}, S_x and S_x, S'_{xx} are all sufficient statistics. But it is clear that S_{xx}, S_x or S_x, S'_{xx} is "better" than $X_{(1)}, \ldots, X_{(n)}$, since it "encodes" information the unknown parameters in fewer terms. In other words, it drops the dimension of the data from n to two. This brings us to the notion of minimal sufficiency.

Definition 1.4.2 (Minimal sufficiency) *A statistic $S(\underline{X})$ is minimal sufficient if (a) it is sufficient and (b) if $T(\underline{X})$ is sufficient statistic there exists an f such that $S(\underline{X}) = f(T(\underline{X}))$.*

Note that the minimal sufficient statistic of a family of distributions is not unique.

The minimal sufficient statistic corresponds to the coarsest sufficient partition of sample space, whereas the data generates the finest partition. We show in Lemma 1.6.4 that if a family of distributions belong to the exponential family and the sufficient statistics are linearly independent, then these sufficient statistics are minimally sufficient.

We now show that the minimal sufficient statistics of the exponential class of distributions are quite special.

Theorem 1.4.2 (Pitman-Koopman-Darmois theorem) *Suppose that $\mathcal{F}$ is a parametric class of distributions whose domain* does not depend on the parameter*, this as-*

sumption includes the Cauchy family, Weibull distributions and exponential families of distributions but not *the uniform family. Only in the case that distribution belongs to the exponential family will the number of minimal sufficient statistic* not *depend on sample size.*

The uniform distribution has a finite number of sufficient statistics ($\max X_i$), which does not depend on the sample size and it does not belong the exponential family. However, the Pitman-Koopman-Darmois theorem does not cover the uniform distribution since its domain depends on the parameter θ.

Example 1.4.7 (Number of sufficient statistics is equal to the sample size) *(i) Consider the Cauchy family of distributions*

$$\mathcal{F} = \left\{ f_\theta; f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2} \right\}.$$

the joint distribution of $\{X_i\}_{i=1}^n$ where X_i follow a Cauchy is

$$\prod_{i=1}^{n} \frac{1}{\pi(1+(x_i-\theta)^2)}.$$

We observe that the parameters cannot be separated from any of the variables. Thus we require all the data to characterize the parameter θ.

(ii) The Weibull family of distributions

$$\mathcal{F} = \left\{ f_\theta; f_{\phi,\alpha}(x) = \left(\frac{\alpha}{\phi}\right)\left(\frac{x}{\phi}\right)^{\alpha-1} \exp[-(x/\phi)^\alpha] \right\}$$

Example 1.4.8 (The truncated exponential) *Suppose that X is an exponentially distributed random variable but is truncated at c. That is*

$$f(x;\theta) = \frac{\theta \exp(-\theta x)}{1-e^{-c\theta}} I(x \leq c).$$

However, the truncation point c is the point which cuts the exponential distribution in half, that is $1/2 = e^{-c\theta} = 1 - e^{-c\theta}$. Thus $c = \theta^{-1}\log 2$. Thus the boundary of the distribution depends on the unknown parameter θ (it does not belong to the exponential family).

Suppose $\{X_i\}$ *are iid random variables with distribution* $f(x;\theta) = 2\theta\exp(-x\theta)I(x \leq \theta^{-1}\log 2)$ *where* $\theta \in \Theta = (0,\infty)$. *The likelihood for* θ *is*

$$\begin{aligned} L(\theta;\underline{X}) &= 2^n\theta^n\exp(-\theta\sum_{i=1}^n X_i)\prod_{i=1}^n I_{[0,\theta^{-1}\log 2]}(X_i) \\ &= 2^n\theta^n\exp(-\theta\sum_{i=1}^n X_i)I_{[0,\theta^{-1}\log 2]}(\max X_i), \end{aligned}$$

thus we see there are two sufficient statistics for θ, $s_1(\underline{X}) = \sum_i X_i$ *and* $s_2(\underline{X}) = \max_i X_i$.

We recall that from Lemma 1.3.2 that most transformations in the data will lead to a loss in information about the parameter θ. One important exception are sufficient statistics.

Lemma 1.4.1 (The Fisher information matrix and sufficient statistics) *Suppose Assumption 1.3.1 holds and* $S(\underline{X})$ *is a sufficient statistic for a parametric family of distributions* $\mathcal{F} = \{f_\theta;\theta \in \Theta\}$. *Let* $I_{\underline{X}}(\theta)$ *and* $I_{S(\underline{X})}(\theta)$ *denote the Fisher information of* $\underline{X}$ *and* $S(\underline{X})$ *respectively. Then for all* $\theta \in \Theta$

$$I_{S(\underline{X})}(\theta) = I_{\underline{X}}(\theta).$$

PROOF. From the proof of Lemma 1.3.2 we have

$$I_{\underline{X}}(\theta) = I_{S(\underline{X})}(\theta) + \mathrm{E}\left(\frac{\partial\log f_{\underline{X}|S(\underline{X})}(\underline{X}|S(\underline{X}),\theta)}{\partial\theta}\right)^2. \tag{1.14}$$

By definition of a sufficient statistic

$$f_{\underline{X}|S(\underline{X})}(\underline{x}|\underline{y},\theta)$$

does not depend on θ. This means that $\frac{\partial\log f_{\underline{X}|S(\underline{X})}(\underline{X}|S(\underline{X}),\theta)}{\partial\theta} = 0$, consequently the second term on the right hand side of (1.14) is zero, which gives the required result. □

Remark 1.4.2 *It is often claimed that only transformations of the data which are sufficient statistics have the same information as the original data. This is not necessarily true, sufficiency is not a necessary condition for Lemma 1.4.1 to hold.* `http://arxiv.org/pdf/1107.3797v2.pdf` *gives an example where a statistic that is* not *a sufficient statistic of the data has the same Fisher information as the Fisher information of the data itself.*

1.4.1 The Fisher information and ancillary variables

We defined the notion of ancillary in the previous section. Here we give an application. Indeed we have previously used the idea of an ancillary variable in regression even without thinking about it! I discuss this example below

So let us start with an example. Consider the problem of simple linear regression where $\{Y_i, X_i\}_{i=1}^n$ are iid bivariate Gaussian random variables and

$$Y_i = \beta X_i + \varepsilon_i,$$

where $\mathrm{E}[\varepsilon_i] = 0$, $\mathrm{var}[\varepsilon_i] = 1$ and X_i and ε_i are independent and β is the unknown parameter of interest. We observe $\{Y_i, X_i\}$. Since X_i contains no information about β it seems logical to look at the conditional log-likelihood of Y_i conditioned on X_i

$$\mathcal{L}(\beta; \underline{Y}|\underline{X}) = \frac{-1}{2}\sum_{i=1}^{n}(Y_i - \beta X_i).$$

Using the factorisation theorem we see that sufficient statistics for β are $\sum_{i=1}^n Y_i X_i$ and $\sum_{i=1}^n X_i^2$. We see that the distribution of $\sum_{i=1}^n X_i^2$ contains no information about β. Thus it is an ancillary variable. Furthermore, since the conditional distribution of $\sum_{i=1}^n X_i^2$ conditioned on $\sum_{i=1}^n X_i Y_i$ does depend on β it is an ancillary complement (I have no idea what the distribution is).

Now we calculate the Fisher information matrix. The second derivative of the likelihood is

$$\frac{\partial^2 \mathcal{L}(\beta; \underline{Y}|\underline{X})}{\partial \beta^2} = -\sum_{i=1}^{n} X_i^2 \Rightarrow -\frac{\partial^2 \mathcal{L}(\beta; \underline{Y}|\underline{X})}{\partial \beta^2} = \sum_{i=1}^{n} X_i^2.$$

To evaluate the Fisher information, do we take the expectation with respect to the distribution of $\{X_i\}$ or not? In other words, does it make sense to integrate influence of the observed regressors (which is the ancillary variable) or not? Typically, in regression one does not. We usually write that the variance of the least squares estimator of a simple linear equation with no intercept is $(\sum_{i=1}^n X_i^2)$.

We now generalize this idea. Suppose that (X, A) are sufficient statistics for the parameter θ. However, A is an ancillary variable, thus the marginal distribution contains no information about θ. The joint log-likelihood can be written as

$$\mathcal{L}(\theta; X, A) = \mathcal{L}(\theta; X|A) + \mathcal{L}(A)$$

where $\mathcal{L}(\theta; X|A)$ is the conditional log-likelihood of X conditioned on A and $\mathcal{L}(A)$ is the marginal log distribution of A which does not depend on A. Clearly the second derivative of $\mathcal{L}(\theta; X, A)$ with respect to θ is

$$-\frac{\partial^2 \mathcal{L}(\theta; X, A)}{\partial \theta^2} = -\frac{\partial^2 \mathcal{L}(\theta; X|A)}{\partial \theta^2}.$$

The Fisher information is the expectation of this quantity. But using the reasoning in the example above it would seem reasonable to take the expectation conditioned on the ancillary variable A.

1.5 Sufficiency and estimation

It is clear from the factorisation theorem that the sufficient statistic contains all the "ingredients" about the parameter θ. In the following theorem we show that by projecting any unbiased estimator of a parameter onto its sufficient statistic we reduce its variance (thus improving the estimator).

Theorem 1.5.1 (The Rao-Blackwell Theorem) *Suppose $s(\underline{X})$ is a sufficient statistic and $\widetilde{\theta}(\underline{X})$ is an unbiased estimator of θ then if we define the new unbiased estimator $\mathrm{E}[\widetilde{\theta}(\underline{X})|s(\underline{X})]$, then $\mathrm{E}[\mathrm{E}[\widetilde{\theta}(\underline{X})|s(\underline{X})]] = \theta$ and*

$$\mathrm{var}\left[\mathrm{E}\left(\widetilde{\theta}(\underline{X})|s(\underline{X})\right)\right] \leq \mathrm{var}\left[\widetilde{\theta}(\underline{X})\right].$$

PROOF. Using that the distribution of $\underline{X}$ conditioned on $s(\underline{X})$ *does not* depend on θ, since $s(\underline{X})$ is sufficient (very important, since our aim is to estimate θ) we have

$$\mathrm{E}[\widetilde{\theta}(\underline{X})|s(\underline{X}) = y] = \int \widetilde{\theta}(\underline{x}) f_{\underline{X}|s(\underline{X})=y}(\underline{x}) d\underline{x}$$

is only a function of $s(\underline{X}) = y$ (and not θ).

We know from the theory of conditional expectations that since $\sigma(s(\underline{X})) \subset \sigma(X_1, \ldots, X_n)$, then $\mathrm{E}[\mathrm{E}(X|\mathcal{G})] = \mathrm{E}[X]$ for any sigma-algebra $\mathcal{G}$. Using this we immediately we have $\mathrm{E}[\mathrm{E}[\widetilde{\theta}(\underline{X})|s(\underline{X})]] = \mathrm{E}[\widetilde{\theta}(\underline{X})] = \theta$. Thus $\mathrm{E}[\widetilde{\theta}(\underline{X})|s(\underline{X})]$ is an unbiased estimator.

To evaluate the variance we use the well know equality $\mathrm{var}[X] = \mathrm{var}[\mathrm{E}(X|Y)] + \mathrm{E}[\mathrm{var}(X|Y)]$. Clearly, since all terms are positive $\mathrm{var}[X] \geq \mathrm{var}[\mathrm{E}(X|Y)]$. This immediately gives the Rao-Blackwell bound. □

Example 1.5.1 *Suppose $\{X_i\}_{i=1}^n$ are iid normal random variable with mean μ and variance σ^2. We know that $S_x = \sum_{i=1}^n X_i$ is a sufficient statistic for μ. We also know that X_1 is an unbiased estimator of μ, but it is not sufficient. It is clear that $\mathrm{var}[\widetilde{\theta}] = \mathrm{var}[X_1] = \sigma^2$. To improve the estimator we condition X_1 on S_x, that is define $\widehat{\theta} = \mathrm{E}[X_1|S_x]$, by the Rao-Blackwell theorem this has a smaller variance than X_1. To show that this is true for this example, we use that $X_1, \ldots, X_n$ are jointly normal then $\mathrm{E}[X_1|S_x]$ is the best linear predictor of X_1 given S_x*

$$\mathrm{E}[X_1|S_x] = \frac{\mathrm{cov}[X_1, S_x]}{\mathrm{var}[S_x]} S_x = \frac{\sigma^2}{n\sigma^2} S_x = \bar{X},$$

which is not a surprise.

Is this the best estimator amongst all unbiased estimator? The Lehmann-Scheffe theorem shows that it is.

The Rao-Blackwell theorem tells us that estimators with the smallest variance must be a function of a sufficient statistic. Of course, one can ask is there a unique estimator with the minumum variance. For this we require completeness of the sufficient statistic. Uniqueness immediately follows from the idea of completeness.

Definition 1.5.1 (Completeness) *Let $s(\underline{X})$ be a minimally sufficient statistic for all $\theta \in \Theta$. Suppose $Z(\cdot)$ is a function of $s(\underline{X})$ such that $\mathrm{E}_\theta[Z(s(\underline{X}))] = 0$. $s(\underline{X})$ is a complete sufficient statistic if and only if $\mathrm{E}[Z(s(\underline{X}))] = 0$ implies $Z(t) = 0$ for all t and all $\theta \in \Theta$.*

Example 1.5.2 *If the exponential family has full rank, that is the number of unknown parameters is equal to the dimension of the exponential family (and the parameter space Θ is an open set, as yet I cannot give a good condition for this) then it is complete (see Lehmann (1986), Section 4.3, Theorem 1).*

Examples include the fully parameterized normal distribution, exponential distribution, binomial distribution etc.

Example 1.5.3 (The constrained normal) *Suppose that $X \sim \mathcal{N}(\mu^2, \mu^2)$. Then $S_x = \sum_{i=1}^n X_i$ and $S'_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$ are still the sufficient statistics for μ^2. To see why consider the conditional distribution of $S'_{xx}|S_x$, we know that S'_{xx} and S_x are independent thus it is the marginal distribution of S'_{xx} which is $\mu^2\chi^2_{n-1}$. Clearly this still depends on the parameter μ^2. Hence we cannot reduce the number of sufficient statistics when we constrain the parameters.*

However, S_x and S_{xx} are not complete sufficient statistics for μ^2, since there exists a non-zero function $Z(S_x, S_{xx})$ such that

$$\mathrm{E}\left(Z(S_x, S_{xx})\right) = \mathrm{E}\left(\bar{X} - s^2\right) = \mu^2 - \mu^2 = 0.$$

Example 1.5.4 (The uniform distribution $f(x;\theta) = \theta^{-1} I_{[0,\theta]}(x)$) *Given the random variables $\{X_i\}_{i=1}^n$ we recall that the sufficient statistic is $\max(X_i)$, we now show that it is complete. Since $P(\max(X_i) \leq x) = (x/\theta)^n I_{[0,\theta]}(x)$ the density is $f_{\max}(x) = nx^{n-1}/\theta^n I_{[0,\theta]}(x)$. We now look for functions Z (which do not depend on θ) where*

$$\mathrm{E}_\theta(Z(\max_i X_i)) = \frac{n}{\theta^n}\int_0^\theta Z(x)x^{n-1}dx.$$

It is "clear" that there cannot exist a function X where the above is zero for all $\theta \in (0,\infty)$ (I can't think of a cute mathematical justification). Thus $\max_i(X_i)$ is a complete minimal sufficient statistic for $\{X_i\}$.

Theorem 1.5.2 (Lehmann-Scheffe Theorem) *Suppose that $\{S_1(\underline{X}), \ldots, S_p(\underline{X})\}$ is a complete minimally sufficient statistic for the parametric family $\mathcal{F} = \{f_\theta; \theta \in \Theta\}$ and for all $\theta \in \Theta$ $T(\underline{X})$ is an unbiased estimator estimator of θ then $\widehat{\theta}[\underline{X}] = \mathrm{E}\left[T(\underline{X})|s(\underline{X})\right]$ is the unique minimum variance unbiased estimator (UMVUE) for all $\theta \in \Theta$.*

PROOF. Suppose $\phi[s(\underline{X})]$ is an unbiased estimator of θ with a smaller variance than $\widehat{\theta}[s(\underline{X})]$ then taking differences it is clear by unbiasedness that

$$\mathrm{E}\left(\widehat{\theta}[s(\underline{X})] - \widehat{\phi}[s(\underline{X})]\right) = 0.$$

However, completeness immediately implies that $\widehat{\phi}[s(\underline{x})] - \widehat{\theta}[s(\underline{x})] = 0$ almost surely. Thus proving the result. □

This theorem tells us if the conditions are satisfied, then for every $\theta \in \Theta$, the estimator $T(\underline{X})$ will give the smallest variance amongst all estimators which are unbiased. The condition that the comparison is done over all *unbiased* estimators is very important. If we drop the relax the condition to allow biased estimators then improvements are possible.

Remark 1.5.1 *Consider the example of the truncated exponential in Example 1.4.8. In this example, there are two sufficient statistics, $s_1(\underline{X}) = \sum_{i=1}^n X_i$ and $s_2(\underline{X}) = \max_i X_i$ for the unknown parameter θ, neither are ancillary in the sense that their marginal distributions depend on θ. Thus both sufficient statistics can be used to estimate θ.*

In general if there are two sufficient statistics for one parameter, θ, and neither of the sufficient statistics are ancillary, then usually one can use either sufficient statistic as a means of constructing an estimator of θ.

Exercise 1.6 *In the above remark, calculate the expectation of* $\max_i X_i$ *and* $\sum_i X_i$ *and use this to propose two different estimators for* θ.

Example 1.5.5 *For the curious,* `http://www.tandfonline.com/doi/abs/10.1080/00031305.2015.1100683?journalCode=utas20` *give an example of minimal sufficient statistics which are not complete and use the Rao-Blackwell theorem to improve on the estimators (though the resulting estimator does not have minimum variance for all* θ *in the parameter space).*

1.6 The exponential family of distributions

We now expand a little on the exponential family described in the previous section. In a nutshell the exponential family is where the parameters of interest and the random variables of the log-likelihood are separable. As we shall see below, this property means the number of minimal sufficient statistics will always be finite and estimation relatively straightforward.

1.6.1 The natural/canonical exponential family

We first define the one-dimension natural exponential family

$$f(x;\theta) = \exp\left(s(x)\theta - \kappa(\theta) + c(x)\right), \tag{1.15}$$

where $\kappa(\theta) = \log \int \exp(s(x)\theta + c(x))d\nu(x)$ and $\theta \in \Theta$ (which define below). If the random variable is continuous, then typically $\nu(x)$ is the Lebesgue measure, on the other hand if it is discrete then $\nu(x)$ is the point mass, for example for the Poisson distribution $d\nu(x) = \sum_{k=0}^{\infty} \delta_k(x)dx$.

Example 1.6.1 *We now give an example of a distribution which immediately has this parameterisation. The exponential distribution has the pdf is* $f(x;\lambda) = \lambda \exp(-\lambda x)$*, which can be written as*

$$\log f(x;\lambda) = (-x\lambda + \log \lambda) \qquad \lambda \in (0,\infty)$$

Therefore $s(x) = -x$ *and* $\kappa(\lambda) = -\log\lambda$.

The parameter space for this family is defined as

$$\Theta = \left\{\theta; \int \exp\left(s(x)\theta + c(x)\right) d\nu(x) < \infty\right\},$$

in other words all parameters where this integral is finite and thus gives a well defined density. The role of $\kappa(\theta)$ is as a normaliser and ensures that density integrates to one i.e

$$\int f(x;\theta)d\nu(x) = \int \exp\left(s(x)\theta - \kappa(\theta) + c(x)\right) d\nu(x) = \exp(-\kappa(\theta)) \int \exp\left(s(x)\theta + c(x)\right) d\nu(x) = 1$$

we see that

$$\kappa(\theta) = \log \int \exp\left(s(x)\theta + c(x)\right) d\nu(x)$$

By using the factorisation theorm, we can see that $\sum_{i=1}^{n} s(X)$ is the sufficient statistic for the family $\mathcal{F} = \{f(x;\theta); \theta \in \Theta\}$. The one-dimensional natural exponential is only a function of one-parameter. The p-dimensional natural exponential generalisation is defined as

$$f(x;\theta) = \exp\left[\mathbf{s}(x)'\theta - \kappa(\theta) + c(x)\right]. \tag{1.16}$$

where $\mathbf{s}(x) = (s_1(x), \ldots, s_p(x))$ is a vector which is a function of x and $\theta = \{\theta_1, \ldots, \theta_p\}$ is a p-dimension parameter. The parameter space for this family is defined as

$$\Theta = \left\{\theta; \int \exp\left(\mathbf{s}(x)'\theta + c(x)\right) d\nu(x) < \infty\right\},$$

again $\kappa(\theta)$ is such that

$$\kappa(\theta) = \log \int \exp\left(\sum_{j=1}^{p} s_j(x)\theta_j + c(x)\right) d\nu(x)$$

and ensures that the density integrates to one.

Lemma 1.6.1 *Consider the p-dimension family* $\mathcal{F}$ *of densities where* $\mathcal{F} = \{f(x;\theta); \theta = (\theta_1, \ldots, \theta_p) \in \Theta\}$ *with*

$$f(x;\theta) = \exp\left[\mathbf{s}(x)'\theta - \kappa(\theta) + c(x)\right].$$

By using the Factorisation theorem it can be seen that $\{\sum_{i=1}^{n} s_1(X_i), \ldots, \sum_{i=1}^{n} s_p(X_i)\}$ *are the sufficient statistics for* $\mathcal{F}$.

However, once one goes beyond dimension one, there can arise redundancy in the representation. For example, consider the two-dimensional exponential family defined by

$$\mathcal{F} = \{f(x;\theta_1,\theta_2) = \exp\left(\alpha s(x)\theta_1 + \beta s(x)\theta_2 - \kappa(\theta_1,\theta_2) + c(x)\right); (\theta_1,\theta_2) \in \Theta\},$$

since $f(x;\theta_1,\theta_2)$ is a density, then

$$\kappa(\theta_1,\theta_2) = \log\left(\int \exp\left[(\theta_1\alpha + \theta_2\beta)s(x) + c(x)\right] d\nu(x)\right).$$

We see that $\kappa(\theta_1,\theta_2)$ is the same for all θ_1,θ_2 such that $(\theta_1\alpha + \theta_2\beta)$ is constant. Thus for all parameters

$$(\theta_1,\theta_2) \in \Theta_C = \{(\theta_1,\theta_2); (\theta_1,\theta_2) \in \Theta, (\theta_1\alpha + \theta_2\beta) = C\}$$

the densities $f(x;\theta_1,\theta_2)$ are the same. This means the densities in $\mathcal{F}$ are *not identifiable*.

Definition 1.6.1 *A class of distributions/model $\mathcal{F} = \{f(x;\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ is non-identifiable if there exists a $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ such that $f(x;\boldsymbol{\theta}_1) = f(x;\boldsymbol{\theta}_2)$ for all $x \in \mathbb{R}$.*

Non-identifiability of a model can be hugely problematic in estimation. If you cannot identify the parameter, then a likelihood can have several maximums, the limit of the estimator is no longer well defined (it can be estimating several different estimators).

In the above example, a minimal representation of the above function is the one-dimensional exponential family

$$\mathcal{F} = \{f(x;\theta) = \exp\left[\theta s(x) - \kappa(\theta) + c(x)\right]; \theta \in \Theta\}.$$

Therefore to prevent this over parameterisation and lack of identifiability we assume that the functions $\{s_j(x)\}_{j=1}^p$ in the canonical representation are linear independent i.e. there does not exist constants $\{\alpha_j\}_{j=1}^p$ and C such that

$$\sum_{j=1}^{p} \alpha_j s_j(x) = C$$

for all x in the domain of X. This representation is called minimal. As can be seen from the example above, if there is linear dependence in $\{s_i(x)\}_{i=1}^p$, then it is easy to find an alternative representation which is of a lower dimension and canonical.

Lemma 1.6.2 *If $\{X_i\}_{i=1}^n$ are iid random variables, which belong to the p-dimensional exponential family that has the form*

$$\begin{aligned} \mathcal{F} &= \left\{ f(x;\theta) = \exp\left[\sum_{j=1}^{p} \theta_j s_j(x) - \kappa(\theta) + c(x)\right]; \theta \in \Theta \right\} \\ \text{where } \Theta &= \left\{ \theta; \int \exp\left[\sum_{j=1}^{p} \theta_j s_j(x) + c(x)\right] d\nu(x) < \infty \right\} \end{aligned}$$

and this is a minimal representation. Then the minimal sufficient statistics are $\{\sum_{i=1}^n s_1(X_i), \ldots, \sum_{i=1}^n s_p(X_i)\}$.

If the parameter space Θ is an open set, then the family of distributions $\mathcal{F}$ is called *regular*. The importance of this will become clear in the next chapter. The parameter space Θ is often called the *natural* parameter space. Note that the the natural parameter space is convex. This means if $\theta_1, \theta_2 \in \mathcal{N}$ then for any $0 \leq \alpha \leq 1$ $\alpha\theta_1 + (1-\alpha)\theta_2 \in \mathcal{N}$. This is proved by using Hölder's inequality and that $\kappa(\theta_1), \kappa(\theta_2) < \infty$ and $e^{\kappa(\theta)} = \int \exp(\theta' \mathbf{s}(x) + c(x)) d\nu(x)$.

Remark 1.6.1 *Convexity of the parameter space basically mean if $\theta_1, \theta_2 \in \mathbb{R}^d$ and both of them are such that give a well defined density then for any convex combination (think a line between the two points) will also yield a well defined density.*

1.6.2 Moments of the canonical representation

In this section we derive the moments of the canonical exponential family using some cute tricks. To simplify the exposition we focus on canonical exponential families of dimension one, though the same result holds for higher dimensions.

Definition 1.6.2 (Cumulant generating function) *The cumulant generating function (for a univariate random variable) is defined as $C_X(t) = \log \mathrm{E}[e^{tX}]$. The power series expansion of the cumulant generating function is*

$$C_X(t) = \log \mathrm{E}[e^{tX}] = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!},$$

where $\kappa_n = C_X^{(n)}(0)$ (analogous to the moment generating function). Note that $\kappa_1(X) = \mathrm{E}[X]$, $\kappa_2(X) = \mathrm{var}[X]$ and $\kappa_j = \kappa_j(X, \ldots, X)$. X is a Gaussian random variable iff $\kappa_j = 0$ for $j \geq 3$.

We use the above in the lemma below.

Lemma 1.6.3 *[Moment generating functions] Suppose that X is a random variable with density*

$$f(x;\theta) = \exp\left(s(x)\theta - \kappa(\theta) + c(x)\right), \theta \in \Theta \tag{1.17}$$

where

$$\Theta = \left\{\theta; \int \exp\left(s(x)\theta - \kappa(\theta) + c(x)\right) d\nu(x) < \infty\right\},$$

. If $\theta \in int(\Theta)$ (the interior of θ, to ensure that it is an open set),

(i) Then the moment generating function of $s(X)$ is

$$\mathrm{E}\left[\exp(s(X)t)\right] = M_{s(X)}(t) = \exp\left[\kappa(t+\theta) - \kappa(\theta)\right]$$

(ii) The cumulant generating function is

$$\log \mathrm{E}\left[\exp(s(X)t)\right] = C_{s(X)}(t) = \kappa(t+\theta) - \kappa(\theta).$$

(iii) Furthermore $\mathrm{E}_\theta[s(X)] = \kappa'(\theta) = \mu(\theta)$ and $\mathrm{var}_\theta[s(X)] = \kappa''(\theta)$.

(iv) $\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} = -\kappa''(\theta)$, thus $\log f(x;\theta)$ has a negative definite Hessian.

This result easily generalizes to p-order exponential families.

PROOF. We choose t sufficiently small such that $(\theta + t) \in int(\Theta)$, since $(\theta + t)$ belongs to the parameter space, then $f(y;(\theta + t))$ is a valid density/distribution. The moment generating function of $s(X)$ is

$$M_{s(X)}(t) = \mathrm{E}\left[\exp(ts(X))\right] = \int \exp(ts(x)) \exp(\theta s(x) - \kappa(\theta) + c(x)) d\nu(x).$$

Taking $\exp(-\kappa(\theta))$ out of the integral and adding and subtracting $\exp(\kappa(\theta + t))$ gives

$$\begin{aligned} M_{s(X)}(t) &= \exp(\kappa(\theta + t) - \kappa(\theta)) \int \exp((\theta + t)s(x) - \kappa(\theta + t) + c(x)) d\nu(x) \\ &= \exp(\kappa(\theta + t) - \kappa(\theta)), \end{aligned}$$

since $\int \exp((\theta+t)y - \kappa(\theta+t) + c(y))dy = \int f(y;(\theta+t))dy = 1$. To obtain the moments we recall that the derivatives of the cumulant generating function at zero give the cumulant of the random variable. In particular $C'_{s(X)}(0) = \mathrm{E}[s(X)]$ and $C''_{s(X)}(0) = \mathrm{var}[s(X)]$. Which immediately gives the result. □

1.6.3 Reparameterisations and examples

We recall that a distribution belongs to the exponential family $\mathcal{F}$ if $f \in \mathcal{F}$ can be written as

$$f(x;\omega) = \exp\left(\sum_{j=1}^{p} \phi_j(\omega)s_j(x) - A(\omega) + c(x)\right),$$

where $\omega = (\omega_1, \ldots, \omega_q)$ are the q-dimensional parameters. Since this family of distributions is parameterized by ω and not θ it is not in natural form. With the exponential distribution there are very few distributions which immediately have a canonical/natural exponential representation. However, it can be seen (usually by letting $\theta_j = \phi_j(\omega)$) that all exponential families of distributions can be reparameterized such that it has a canonical/natural representation. Moreover by making sufficient transformations, to ensure the sufficient statistics do not satisfy any linear constraints, the representation will be minimal (see the monograph `http://www.jstor.org/stable/pdf/4355554.pdf?acceptTC=true`, Lawrence Brown (1986), Proposition 1.5, for the precise details). Let $\Phi(\omega) = (\phi_1(\omega), \ldots, \phi_p(\omega))$ and Ω denote the parameter space of ω. Then we see that $\Phi : \Omega \to \Theta$, where Θ is the natural parameter space defined by

$$\Theta = \left\{\theta; \int \exp\left(\sum_{j=1}^{p} \theta_j s_j(x) + c(x)\right) dx < \infty\right\}.$$

Thus Φ is an injection (one-to-one) mapping from Ω to Θ. Often the mapping is a bijection (injective and surjective), in which case $p = q$. In such cases, the exponential family is said to have *full rank* (technically, full rank requires that $\mathcal{N}$ is an open set; when it is closed strange things can happen on the boundary of the set).

If the image of Φ, $\Phi(\Omega)$, is not a linear subset of $\mathcal{N}$, then the exponential family $\mathcal{F}$ is called a curved exponential.

Recall that θ is a function of the d-dimension parameters ω if

(i) If $p = d$ then the exponential family is said to have full rank. In this case the sufficient statistics are complete.

(i) If $p > d$ then the exponential family is said to be a curved exponential family. This means the image $\Phi(\Omega)$ (the parameter space of ω onto θ) is not a linear subset of Θ For curved exponential families there are nonlinear constraints between the unknown parameters.

When the exponential family is curved it is *not complete* (see Exercise 1.6.2). The implication of this is that there is no unique unbiased estimator (in terms of the sufficient statistics), which will give the minimal variance for all parameters in the parameter space. See Brown (1986), Theorem 1.9 (page 13) for details on the above.

Lemma 1.6.4 *If a distribution belongs to the exponential family, and the sufficient statistics are linearly independent then the sufficient statistics are minimally sufficient.*

Example 1.6.2 (The normal distribution) *We recall that S_{xx}, S_x are the sufficient statistics of the normal family of distributions, where $(S_{xx}, S_x) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$. It is clear that (S_{xx}, S_x) are linearly independent (i.e. for no linear combination $\alpha S_{xx} + \beta S_x = 0$ for all S_x and S_{xx}), thus by Lemma 1.6.4 they are minimally sufficient.*

Exercise 1.7 *Suppose that $\{X_i\}_{i=1}^n$ are iid normal random variables where the ratio between mean and standard deviation $\gamma = \sigma/\mu$ is known. What are the minimal sufficient statistics?*

1.6.4 Examples

By making appropriate transformations, we show that the below well known distributions can be written in natural form.

(i) The exponential distribution is already in natural exponential form and the parameter space is $\Theta = (0, \infty)$.

(ii) For the binomial distribution where $X \sim Bin(n, p)$ we note

$$\log f(x; p) = x \log p + (n - x) \log(1 - p) + \log \binom{n}{x}.$$

One natural parameterisation is to let $\theta_1 = \log p$, $\theta_2 = \log(1 - p)$ with sufficient statistics x and $(n - x)$. This a two-dimensional natural exponential representation. However we see that the sufficient statistics are subject to a linear constraint, namely $s_1(x) + s_2(x) = x + (n - x) = n$. Thus this representation is not minimal. Instead we rearrange $\log f(x; p)$

$$\log f(x; p) = x \log \frac{p}{1 - p} + n \log(1 - p) + \log \binom{n}{x}.$$

Let $\theta = \log(\frac{p}{1-p})$, since $\theta(p) = \log(\frac{p}{1-p})$ is invertible this gives the natural representation

$$\log f(x;\theta) = \left[x\theta - n\log\left(1+\exp(\theta)\right) + \log\binom{n}{x} \right].$$

Hence the parameter of interest, $p \in (0,1)$, has been transformed, to $\theta \in (-\infty, \infty)$. The natural parameter space is $\Theta = (-\infty, \infty)$. The sufficient statistic is $\sum_i X_i$. $d\nu(x) = dx$, the Lebesgue measure.

(iii) The normal family of distributions can be written as

$$\log f(x;\mu,\sigma) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log 2\pi. \tag{1.18}$$

In this case the natural exponential parametrisation is $\mathbf{x} = (-\frac{1}{2}x^2, x)$, $\theta = (\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}) = (\theta_1, \theta_2)$ and $\kappa(\theta_1,\theta_2) = \theta_2^2/(2\theta_1) - 1/2\log(\theta_1)$. In this case $\Theta = (0,\infty)\times(-\infty,\infty)$. The sufficient statistics are $\sum_i X_i$ and $\sum_i X_i^2$. $d\nu(x) = dx$, the Lebesgue measure.

(iv) The multinomial distribution can be written as

$$\begin{aligned}\log f(x_1,\ldots,x_p;\pi) &= \sum_{i=1}^{p} x_i \log\pi_i + \log n! - \sum_{i=1}^{p} x_i! \\ &= \sum_{i=1}^{p-1} x_i \log\frac{\pi_i}{\pi_p} + n\log\pi_p + \log n! - \sum_{i=1}^{p} x_i!.\end{aligned}$$

For $1 \le i \le p-1$ let $\theta_i = \log\pi_i/\pi_p$ then the natural representation is

$$\log f(x_1,\ldots,x_p;\pi) = \sum_{i=1}^{p-1}\theta_i x_i - n\log\left(1+\sum_{i=1}^{p-1}\exp(-\theta_i)\right) + \log n! - \sum_{i=1}^{p} x_i!$$

and the parameters space is $\mathbb{R}^{p-1}$. The sufficient statistics are $\sum_i X_{i1},\ldots,\sum_i X_{i,p-1}$. The point measure is $d\nu(x) = \sum_{j_1,j_2,\ldots,j_{p-1}=1}^{n}\delta_{j_1}(x_1)\ldots\delta_{j_{p-1}}(x_{j-1})\delta_{[0,n]}(x_1+\ldots+x_{p-1})dx_1\ldots dx_{p-1}$.

Note that one can also write the multinomial as

$$\log f(x_1,\ldots,x_p;\pi) = \sum_{i=1}^{p}\theta_i x_i + \log n! - \sum_{i=1}^{p} x_i!,$$

where $\theta_i = \log\pi_i$. However this is not in minimal form because $n - \sum_{i=1}^{n} x_i = 0$ for all $\{x_i\}$ in the sample space; thus they are not linearly independent.

(v) The censored exponential distribution. $X \sim Exp(\lambda)$ (density of X is $f(x;\lambda) = \exp[-x\lambda + \log\lambda]$), however X is censored at a known point c and Y is observed where

$$Y = \begin{cases} X & X \leq c \\ c & X > c \end{cases}$$

and c is assumed *known*. Suppose we observe $\{Y_i, \delta_i\}$, using (2.3) we have

$$\mathcal{L}(\lambda) = -\sum_{i=1}^{n}(1-\delta_i)\lambda Y_i + (1-\delta_i)\log\lambda - \delta_i c\lambda.$$

We recall that by definition of Y when $\delta = 1$ we have $Y = c$ thus we can write the above as

$$\mathcal{L}(\lambda) = -\lambda\sum_{i=1}^{n} Y_i - \log\lambda\sum_{i=1}^{n}\delta_i + n\log\lambda.$$

Thus when the sample size is n the sufficient statistics are $s_1(Y,\delta) = \sum_i Y_i$, $s_2(Y,\delta) = \sum_i \delta_i = \sum_i I(Y_i \geq c))$. The natural parameterisation is $\theta_1 = -\lambda$, $\theta_2 = -\log(-\lambda)$ and $\kappa(\theta_1,\theta_2) = \theta_2 = \frac{1}{2}(-\log(-\theta_1) + \theta_2)$ (thus we see that parameters are subject to nonlinear constraints). As $s_1(Y,\delta) = \sum_i Y_i, s_2(Y,\delta) = \sum_i \delta_i$ are not linearly dependent this means that the censored exponential distribution has a 2-dimensional natural exponential representation. The measure is $d\nu(x,\delta) = dx[\delta_0(\delta)d\delta + \delta_1(\delta)d\delta]$

However since the parameter space is not the entire natural parameter space $\mathcal{N} = (-\infty,0)\times(-\infty,0)$ (since $\theta_1(\lambda) = \lambda$ and $\theta_2(\lambda) = \log\lambda$) but a subset of it, then the family is curved and thus the sufficient statistics are not complete. This means that there is no unique unbiased estimator with minimal variance.

(vi) The von-Mises distributions are distributions defined on a sphere. The simplest is the von-Mises distribution defined on a 1-d circle

$$f(x;\kappa,\mu) = \frac{1}{2\pi I_0(\kappa)}\exp\left(\kappa\cos(x-\mu)\right) \qquad x \in [0,2\pi],$$

where I_0 is a Bessel function of order zero ($\kappa > 0$ and $\mu \in \mathbb{R}$). We will show that it has a natural 2-dimensional exponential representation

$$\begin{aligned} \log f(x;\kappa,\mu) &= \kappa\cos(x-\mu) - \log 2\pi I_0(\kappa) \\ &= \kappa\cos(x)\cos(\mu) + \kappa\sin(x)\sin(\mu) - \log 2\pi I_0(\kappa). \end{aligned}$$

Let $s_1(x) = \cos(x)$ and $s_2(x) = \sin(x)$ and we use the parameterisation $\theta_1(\kappa, \mu) = \kappa \cos \mu$, $\theta_2(\kappa, \mu) = \kappa \sin \mu$, $\kappa(\theta_1, \theta_2) = -\log 2\pi I_0(\sqrt{\theta_1^2 + \theta_2^2})$. The sufficient statistics are $\sum_i \cos(X_i)$ and $\sum_i \sin(X_i)$ and $(\theta_1, \theta_2) \in \mathbb{R}^2$ The measure is $d\nu(x) = dx$.

(vii) Consider the inflated zero Poisson distribution which has the log-likelihood

$$\begin{aligned}
& \mathcal{L}(\underline{Y}; \lambda, p) \\
= & \sum_{i=1}^n I(Y_i = 0) \log\left(p + (1-p)e^{-\lambda}\right) + \sum_{i=1}^n I(Y_i \neq 0)\left(\log(1-p) + \log \frac{\lambda^{Y_i} e^{-\lambda}}{Y_i!}\right) \\
= & \sum_{i=1}^n [1 - I(Y_i \neq 0)] \log\left(p + (1-p)e^{-\lambda}\right) + \log\lambda \sum_{i=1}^n I(Y_i \neq 0) Y_i \\
& + (\log(1-p) - \lambda) \sum_{i=1}^n I(Y_i \neq 0) - \sum_{i=1}^n I(Y_i \neq 0) \log Y! \\
= & \left\{-\log\left(p + (1-p)e^{-\lambda}\right) + (\log(1-p) - \lambda)\right\} \sum_{i=1}^n I(Y_i \neq 0) \\
& + \log\lambda \sum_{i=1}^n I(Y_i \neq 0) Y_i + n \underbrace{\log\left(p + (1-p)e^{-\lambda}\right)}_{-\kappa(\cdot)} - \sum_{i=1}^n I(Y_i \neq 0) \log Y!.
\end{aligned}$$

This has a natural 2-dimension exponential representation. Let

$$\begin{aligned}
\theta_1 &= \left\{-\log\left(p + (1-p)e^{-\lambda}\right) + (\log(1-p) - \lambda)\right\} \\
\theta_2 &= \log\lambda
\end{aligned}$$

with sufficient statistics $s_1(\underline{Y}) = \sum_{i=1}^n I(Y_i \neq 0)$, $s_2(\underline{Y}) = \sum_{i=1}^n I(Y_i \neq 0) Y_i$. The parameter space is $(\theta_1, \theta_2) \in (-\infty, 0] \times (-\infty, \infty)$, the 0 end point for θ_1 corresponds to $p = 0$. If we allowed $p < 0$ (which makes no sense), then the parameter space for θ_1 can possibly be greater than 0, but this makes no sense. If calculated correctly

$$\kappa(\theta_1, \theta_2) = -\log\left(\frac{e^{\theta_1} - \theta_2^{-1}}{1 - \theta_2^{-1}}(1 - e^{-e^{\theta_2}}) + e^{-e^{\theta_2}}\right).$$

The measure is the point mass $d\nu(x) = \sum_{j=0}^{\infty} \delta_j(x) dx$.

(viii) Suppose (X_i, Y_i) are iid random variables with densities $\theta \exp(-\theta x)$ and $\theta^{-1} \exp(-\theta^{-1} y)$ respectively. Then the joint density is $f(x, y) = \exp(-\theta x - \theta^{-1} y)$. The slight difference here is that there are two random variables at play. But this not change the analysis. The natural exponential parameterisation is

$$f(x, y; \theta_1, \theta_2) = \exp\left(-\theta_1 x - \theta_2 y\right) \quad \theta_1, \theta_2 > 0$$

subject to tthe constraint $\theta_1\theta_2 = 1$. The log-likelihood is

$$\mathcal{L}_n(\theta) \;=\; -\theta_1 \sum_{i=1}^{n} X_i - \theta_2 \sum_{i=1}^{n} Y_i,$$

thus the minimal sufficient statistics are $s_1(\underline{X}, \underline{Y}) = \sum_i^n X_i$ and $s_2(\underline{X}, \underline{Y}) = \sum_i^n Y_i$. However, the parameter space is $(\theta, 1/\theta)$ which is not a linear subset in $(R^+)^2$, thus it is not complete. This is a curved exponential. The measure is $d\nu(x, y) = dxdy$.

1.6.5 Some additional properties of the exponential family

We first state some definitions which we use later.

Definition 1.6.3 (Concave, convex functions and the Hessian)

- *A function is said to be concave if*

$$f\left(y + \alpha(x-y)\right) = f\left(\alpha x + (1-\alpha)y\right) \geq \alpha f(x) + (1-\alpha)f(y) = f(y) + \alpha\left[f(x) - f(y)\right].$$

and strictly concave if

$$f\left(y + \alpha(x-y)\right) = f\left(\alpha x + (1-\alpha)y\right) > \alpha f(x) + (1-\alpha)f(y) = f(y) + \alpha\left[f(x) - f(y)\right].$$

For $d = 1$ this can be seen as the curve of f lying above the tangent between the points $(x, f(x))$ and $(y, f(y))$. This immediately implies that if $y > x$, then

$$f(y) - f(x) \;<\; \frac{f\left(x + \alpha(y-x)\right) - f(x)}{\alpha} \Rightarrow \frac{f(y) - f(x)}{y - x} < \frac{f\left(x + \alpha(y-x)\right) - f(x)}{\alpha(y-x)}$$

for all $0 < \alpha < 1$. Thus

$$\frac{f(y) - f(x)}{y - x} < f'(x).$$

- *The Hessian of a function of p variables $f : \mathbb{R}^p \to \mathbb{R}$ is its second derivative $\nabla_\theta^2 f(\theta) = \left\{\frac{\partial^2 f(\theta)}{\partial\theta_i \partial\theta_j}; 1 \leq i, j \leq p\right\}$.*
- *Examples of concave functions are $f(x) = -x^2$ (for $x \in (-\infty, \infty)$) and $f(x) = \log x$ for $x \in (0, \infty)$. Observe that $-x^2$ is maximised at $x = 0$, whereas the maximum of $\log x$ lies outside of the interval $(0, \infty)$.*

A function is a concave function if and only if the Hessian, $\nabla_\theta^2 f$, is negative semi-definite.

We now show consider the properties of the log likelihood of the natural exponential.

(i) We now show that second derivative of log-likelihood of a function from the natural exponential family has a negative definite Hessian. It is straightforward to show that the second derivative of the log-likelihood is

$$\nabla_\theta^2 \mathcal{L}_n(\theta) = -\sum_{i=1}^n \nabla_\theta^2 \kappa(\theta) = -n\nabla_\theta^2 \kappa(\theta).$$

From Lemma 1.6.3 we see that for all $\theta \in \Theta$ $\nabla_\theta^2 \kappa(\theta)$ corresponds to the variance of a random variable X_θ with density f_θ. This implies that $\nabla_\theta^2 \kappa(\theta) \geq 0$ for all $\theta \in \Theta$ and thus the Hessian $\nabla_\theta^2 \mathcal{L}_n(\theta)$ is semi-negative definite. We will later show that this means that $\mathcal{L}_n(\underline{X}; \theta)$ can easily be maximised. Thus for the natural exponential family the observed and expected Fisher information are the same.

Examples of different concave likelihoods are given in Figure 1.3. Observe that the maximum may not always lie within the interior of the parameter space.

(ii) We recall that θ is a function of the parameters ω. Therefore the Fisher information for ω is related, but not equal to the Fisher information for θ. More precisely, in the case of the one-dimension exponential family the likelihood is

$$\mathcal{L}_n(\theta(\omega)) = \theta(\omega) \sum_{i=1}^n s(X_i) - n\kappa(\theta(\omega)) + n \sum_{i=1}^n c(X_i).$$

Therefore the second derivative with respect to ω is

$$\frac{\partial^2 \mathcal{L}_n[\theta(\omega)]}{\partial \omega^2} = -n \frac{\partial \theta'}{\partial \omega} \frac{\partial^2 \kappa(\theta)}{\partial \theta^2} \frac{\partial \theta}{\partial \omega} + \left(\sum_{i=1}^n X_i - n \frac{\partial \kappa(\theta)}{\partial \theta} \right) \frac{\partial^2 \theta}{\partial \omega^2}.$$

Recall that $\mathrm{E}[\left(\sum_{i=1}^n X_i - n\frac{\partial \kappa(\theta)}{\partial \theta} \right) = n\mathrm{E}[X_i] - n\kappa'(\theta) = 0$. Using this we have

$$I(\omega) \;\; = \;\; -\mathrm{E}\left(\frac{\partial^2 \mathcal{L}_n[\theta(\omega)]}{\partial \omega^2} \right) = n \frac{\partial \theta'}{\partial \omega} \frac{\partial^2 \kappa(\theta)}{\partial \theta^2} \frac{\partial \theta}{\partial \omega}.$$

In this case the observed and expected Fisher information matrices are not the same.

However, if there is a diffeomorphism between the space of θ and ω, negative definite $\nabla_\theta^2 \mathcal{L}_n(\theta) = \frac{\partial^2 \kappa(\theta)}{\partial \theta^2}$ implies negative definite $\nabla_\omega^2 \mathcal{L}_n(\theta(\omega))$. This is because when there is a diffeomorphism (a continuous invertible mapping between two spaces), the

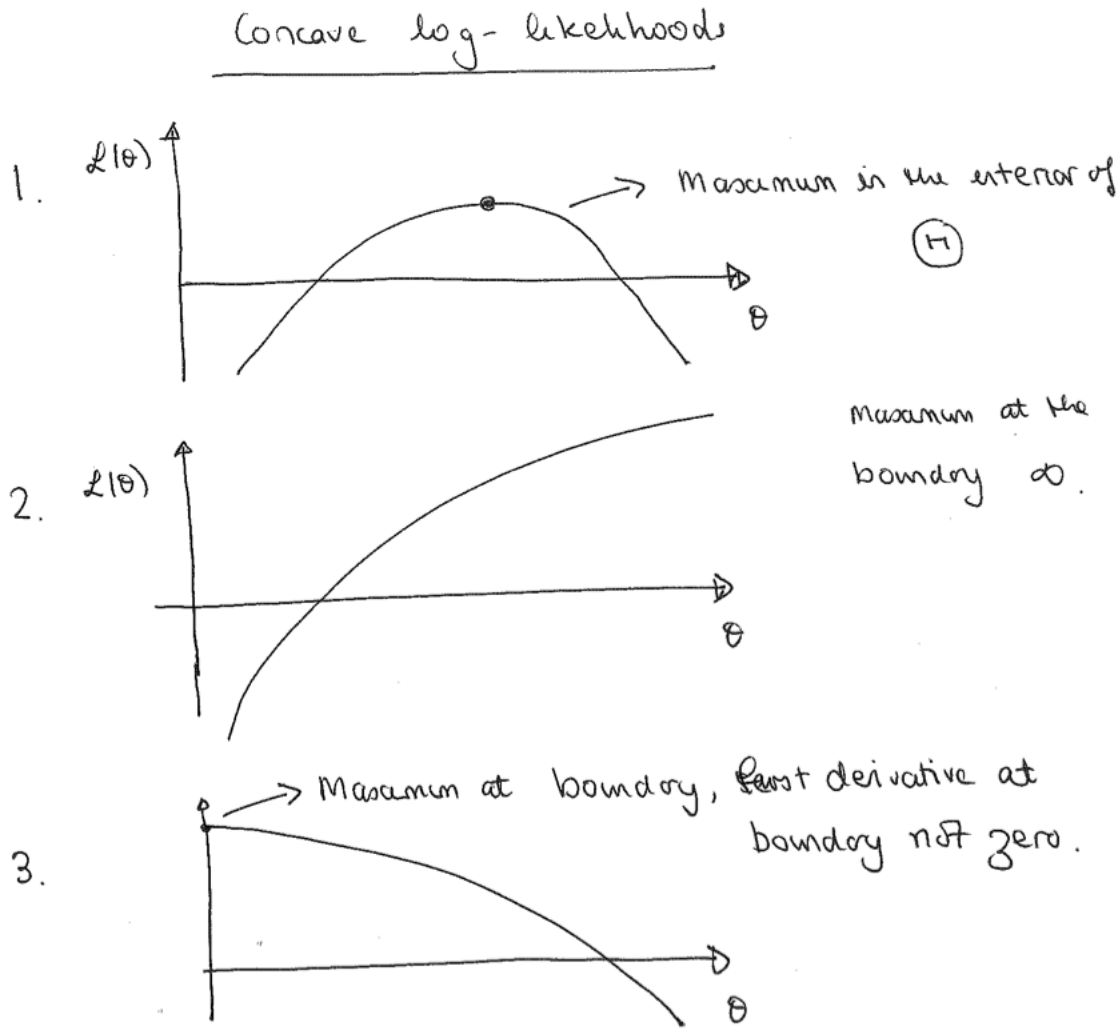

Figure 1.3: Examples of different concave likelihoods

eigen-values of the corresponding Hessian matrices will change, however *the signs will not*. Therefore if $\nabla^2_\theta \mathcal{L}_n(\theta)$ is negative definite then so is its reparametrisation $\nabla^2_\omega \mathcal{L}_n(\theta(\omega))$.

(ii) The natural parameter space $\mathcal{N}$ is convex, this means if $\theta_1, \theta_2 \in \Theta$ then $\alpha\theta_1 + (1-\alpha)\theta_2 \in \Theta$ for $0 \leq \alpha \leq 1$ (easily proved using Hölder's inequality).

(iii) The function $\kappa(\theta)$ is convex (easily proved using that $\kappa(\theta) = \log \int \exp(\theta s(x) + c(x)) d\nu(x)$ and Hölder's inequality).

1.7 The Bayesian Cramer-Rao inequality

The classical Cramér-Rao inequality is useful for assessing the quality of a given estimator. But from the derivation we can clearly see that it only holds if the estimator is unbiased.

As far as I am aware no such inequality exists for the *mean squared error* of estimators that are biased. For example, this can be a problem in nonparametric regression, where estimators in general will be biased. How does one access the estimator in such cases? To answer this question we consider the Bayesian Cramer-Rao inequality. This is similar to the Cramer-Rao inequality but does not require that the estimator is unbiased, so long as we place a prior on the parameter space. This inequality is known as the Bayesian Cramer-Rao or van-Trees inequality (see [**?**] and [**?**]).

Suppose $\{X_i\}_{i=1}^n$ are random variables with distribution function $L_n(\underline{X}; \theta)$. Let $\tilde{\theta}(\underline{X})$ be an estimator of θ. We now Bayesianise the set-up by placing a prior distribution on the parameter space Θ, the density of this prior we denote as λ. Let $\mathrm{E}[g(\underline{x})|\theta] = \int g(\underline{x}) L_n(\underline{x}|\theta) d\underline{x}$ and E_λ denote the expectation over the density of the parameter λ. For example

$$\mathrm{E}_\lambda \mathrm{E}[\widetilde{\theta}(X)|\theta] = \int_a^b \int_{\mathbb{R}^n} \widetilde{\theta}(\underline{x}) L_n(\underline{x}|\theta) d\underline{x} \lambda(\theta) d\theta.$$

Now we place some assumptions on the prior distribution λ.

Assumption 1.7.1 *θ is defined over the compact interval $[a, b]$ and $\lambda(x) \to 0$ as $x \to a$ and $x \to b$ (so $\lambda(a) = \lambda(b) = 0$).*

Theorem 1.7.1 *Suppose Assumptions 1.3.1 and 1.7.1 hold. Let $\widetilde{\theta}(\underline{X})$ be an estimator of θ. Then we have*

$$\mathrm{E}_\lambda\left[\mathrm{E}_\theta\left\{\left(\widetilde{\theta}(\underline{X})-\theta\right)^2\middle|\theta\right\}\right] \geq \left[\mathrm{E}_\lambda[I(\theta)]+I(\lambda)\right]^{-1}$$

where

$$\begin{aligned}\mathrm{E}_\lambda[I(\theta)] &= \int\int\left(\frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta \\ \textit{and}\quad I(\lambda) &= \int\left(\frac{\partial\log\lambda(\theta)}{\partial\theta}\right)^2\lambda(\theta)d\theta.\end{aligned}$$

PROOF. First we derive a few equalities. We note that under Assumption 1.7.1 we have

$$\int_a^b \frac{dL_n(\underline{x};\theta)\lambda(\theta)}{d\theta}d\theta = L_n(\underline{x};\theta)\lambda(\theta)\Big]_a^b = 0,$$

thus

$$\int_{\mathbb{R}^n}\widetilde{\theta}(\underline{x})\int_a^b\frac{\partial L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}d\theta d\underline{x}=0. \tag{1.19}$$

Next consider $\int_{\mathbb{R}^n}\int_a^b\theta\frac{\partial L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}d\theta d\underline{x}$. Using integration by parts we have

$$\begin{aligned}&\int_{\mathbb{R}^n}\int_a^b\theta\frac{dL_n(\underline{x};\theta)\lambda(\theta)}{d\theta}d\theta d\underline{x} = \int_{\mathbb{R}^n}\left(\theta L_n(\underline{x};\theta)\lambda(\theta)\Big]_a^b\right)d\underline{x} - \int_{\mathbb{R}^n}\int_a^b L_n(\underline{x};\theta)\lambda(\theta)d\theta dx \\ =\ &-\int_{\mathbb{R}^n}\int_a^b L_n(\underline{x};\theta)\lambda(\theta)d\theta d\underline{x} = -1.\end{aligned} \tag{1.20}$$

Subtracting (1.20) from (1.19) we have

$$\int_{\mathbb{R}^n}\int_a^b\left(\widetilde{\theta}(\underline{x})-\theta\right)\frac{\partial L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}d\theta d\underline{x} = \int_{\mathbb{R}^n}\int_a^b L_n(\underline{x};\theta)\lambda(\theta)d\theta d\underline{x}=1.$$

Multiplying and dividing the left hand side of the above by $L_n(\underline{x};\theta)\lambda(\theta)$ gives

$$\begin{aligned}&\int_{\mathbb{R}^n}\int_a^b\left(\widetilde{\theta}(\underline{x})-\theta\right)\frac{1}{L_n(\underline{x};\theta)\lambda(\theta)}\frac{dL_n(\underline{x};\theta)\lambda(\theta)}{d\theta}L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta=1. \\ &\Rightarrow\int_{\mathbb{R}^n}\int_a^b\left(\widetilde{\theta}(\underline{x})-\theta\right)\frac{d\log L_n(\underline{x};\theta)\lambda(\theta)}{d\theta}\underbrace{L_n(\underline{x};\theta)\lambda(\theta)}_{\text{measure}}d\underline{x}d\theta=1\end{aligned}$$

Now by using the Cauchy-Schwartz inequality we have

$$1\leq\underbrace{\int_a^b\int_{\mathbb{R}^n}\left(\widetilde{\theta}(\underline{x})-\theta\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta}_{\mathrm{E}_\lambda\left(\mathrm{E}((\widetilde{\theta}(X)-\theta)^2|\theta)\right)}\int_a^b\int_{\mathbb{R}^n}\left(\frac{d\log L_n(\underline{x};\theta)\lambda(\theta)}{d\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta.$$

Rearranging the above gives

$$\mathrm{E}_\lambda\big[\mathrm{E}_\theta(\widetilde{\theta}(X)-\theta)^2\big] \geq \left[\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta\right]^{-1}.$$

Finally we want to show that the denominator of the RHS of the above can equivalently written as the information matrices:

$$\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta = \mathrm{E}_\lambda(I(\theta)) + I(\lambda).$$

We use basic algebra to show this:

$$\begin{aligned}
&\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta \\
=\ &\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta} + \frac{\partial \log \lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta \\
=\ &\underbrace{\left(\frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta}_{\mathrm{E}_\lambda(I(\theta))} + 2\int_a^b \int_{\mathbb{R}^n} \frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\frac{\partial \log \lambda(\theta)}{\partial\theta} L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta \\
&+ \underbrace{\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log \lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta}_{I(\lambda)}.
\end{aligned}$$

We note that

$$\int_a^b \int_{\mathbb{R}^n} \frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\frac{\partial \log \lambda(\theta)}{\partial\theta}dxd\theta = \int \frac{\partial \log \lambda(\theta)}{\partial\theta} \underbrace{\int \frac{\partial L_n(\underline{x};\theta)}{\partial\theta}d\underline{x}}_{=0} d\theta = 0.$$

and $\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log \lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta = \int_a^b \left(\frac{\partial \log \lambda(\theta)}{\partial\theta}\right)^2 \lambda(\theta)d\theta$. Therefore we have

$$\begin{aligned}
&\int\int \left(\frac{\partial \log L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta \\
=\ &\underbrace{\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta}_{\mathrm{E}_\lambda(I(\theta))} + \int_{\mathbb{R}^n} L_n(\underline{x};\theta) \underbrace{\int_a^b \left(\frac{\partial \log \lambda(\theta)}{\partial\theta}\right)^2 \lambda(\theta)d\theta}_{I(\lambda)} d\underline{x}.
\end{aligned}$$

Since $\int_{\mathbb{R}^n} L_n(\underline{x};\theta)d\underline{x} = 1$ we obtain the required result. □

We will consider applications of the Bayesian Cramer-Rao bound in Section **??** for obtaining lower bounds of nonparametric density estimators.

1.8 Some questions

Exercise 1.8 *The distribution function of the random variable X_i is $F(x) = 1-\exp(-\lambda x)$.*

(i) Give a transformation of $\{X_i\}_i$, such that the transformed variable is uniformly distributed on the interval $[0, 1]$.

(ii) Suppose that $\{X_i\}$ are iid random variables. Use your answer in (i), to suggest a method for checking that $\{X_i\}$ has the distribution $F(x) = 1-\exp(-\lambda x)$ (by checking I mean a graphical tool)?

Exercise 1.9 *Find the Fisher information matrix of*

(i) The normal distribution with unknown mean μ and variance σ^2.

(ii) The normal distribution with unknown mean μ and variance μ^2.

(iii) Let $g : \mathbb{R} \to \mathbb{R}$ be density. Show that $\frac{1}{\rho} g\left(\frac{x-\mu}{\rho}\right)$ is a density function. This is known as the location-scale model.

Define the family of distributions

$$\mathcal{F} = \left\{ f(x; \mu, \rho) = \frac{1}{\rho} g\left(\frac{x-\mu}{\rho}\right); \mu \in \mathbb{R}, \rho \in (0, \infty) \right\}.$$

Suppose that μ and ρ is unknown, obtain the corresponding expected Fisher information (make your derivation neat, explaining which terms depend on parameters and which don't); compare your result to (i) when are they similar?

Exercise 1.10 *Construct a distribution which does not belong to the exponential family but has only a finite number of sufficient statistics (the minimal number of sufficient statistics does not grow with n).*

Exercise 1.11 *Suppose that Z is a Weibull random variable with density $f(x; \phi, \alpha) = (\frac{\alpha}{\phi})(\frac{x}{\phi})^{\alpha-1} \exp(-(x/\phi)^{\alpha})$. Show that*

$$\mathrm{E}(Z^r) = \phi^r \Gamma\left(1 + \frac{r}{\alpha}\right).$$

Hint: Use

$$\int x^a \exp(-x^b) dx = \frac{1}{b} \Gamma\left(\frac{a}{b} + \frac{1}{b}\right) \qquad a, b > 0.$$

This result will be useful in some of the examples used later in this course.

Suppose we have two different sampling schemes to estimate a parameter θ, one measure for understanding which method is better able at estimating the parameter is the *relative frequency*. Relative frequency is defined as the ratio between the two corresponding Fisher information matrices. For example, if we have two iid samples from a normal distribution $N(\mu, 1)$ (one of size n and the other of size m), then the relative frequency is $I_n(\mu)/I_m(\mu) = \frac{n}{m}$. Clearly if $n > m$, then $I_n(\mu)/I_m(\mu) = \frac{n}{m} > 1$. Hence the sample of size n contains more information about the parameter μ.

Exercise 1.12 *Consider the censored exponential in equation (1.5), where* $\{(Y_i, \delta_i)\}_{i=1}^n$ *is observed.*

(i) Calculate the expected Fisher information of the censored likelihood.

(ii) Calculate the expected Fisher information of $\{\delta_i\}$.

(iii) Calculate the expected Fisher information when there is no censoring.

By using the notion of relative efficiency comment on which sampling scheme contains the most and least information about the parameter θ.

Chapter 2

The Maximum Likelihood Estimator

We start this chapter with a few "quirky examples", based on estimators we are already familiar with and then we consider classical maximum likelihood estimation.

2.1 Some examples of estimators

Example 1

Let us suppose that $\{X_i\}_{i=1}^n$ are iid normal random variables with mean μ and variance σ^2. The "best" estimators unbiased estimators of the mean and variance are $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ respectively. To see why recall that $\sum_i X_i$ and $\sum_i X_i^2$ are the sufficient statistics of the normal distribution and that $\sum_i X_i$ and $\sum_i X_i^2$ are complete minimal sufficient statistics. Therefore, since $\bar{X}$ and s^2 are functions of these minimally sufficient statistics, by the Lehmann-Scheffe Lemma, these estimators have minimal variance.

Now let us consider the situation where the mean is μ and the variance is μ^2. In this case we have only one unknown parameter μ but the minimally sufficient statistics are $\sum_i X_i$ and $\sum_i X_i^2$. Moreover, it is not complete since both

$$\left(\frac{n}{n+1}\right)\bar{X}^2 \text{ and } s^2 \tag{2.1}$$

are unbiased estimators of μ^2 (to understand why the first estimator is an unbiased estimator we use that $\mathrm{E}[X^2] = 2\mu^2$). Thus violating the conditions of completeness. Furthermore

any convex linear combination of these estimators

$$\alpha\left(\frac{n}{n+1}\right)\bar{X}^2+(1-\alpha)s^2 \qquad 0\leq\alpha\leq 1$$

is an unbiased estimator of μ. Observe that this family of distributions is incomplete, since

$$\mathrm{E}\left[\left(\frac{n}{n+1}\right)\bar{X}^2-s^2\right]=\mu^2-\mu^2,$$

thus there exists a non-zero function $Z(S_x,S_{xx})$ Furthermore

$$\left(\frac{n}{n+1}\right)\bar{X}^2-s^2=\frac{1}{n(n+1)}S_x^2-\frac{1}{n-1}\left(S_{xx}-\frac{1}{n}S_x\right)=Z(S_x,S_{xx}).$$

Thus there exists a non-zero function $Z(\cdot)$ such that $\mathrm{E}[Z(S_x,S_{xx})]=0$, impying the minimal sufficient statistics are not complete.

Thus for all sample sizes and μ, it is not clear which estimator has a minimum variance. We now calculate the variance of both estimators and show that there is no clear winner for all n. To do this we use the normality of the random variables and the identity (which applies only to normal random variables)

$$\begin{aligned}\mathrm{cov}\left[AB,CD\right] &= \mathrm{cov}[A,C]\mathrm{cov}[B,D]+\mathrm{cov}[A,D]\mathrm{cov}[B,C]+\mathrm{cov}[A,C]\mathrm{E}[B]\mathrm{E}[D]+\\ &\quad \mathrm{cov}[A,D]\mathrm{E}[B]\mathrm{E}[C]+\mathrm{E}[A]\mathrm{E}[C]\mathrm{cov}[B,D]+\mathrm{E}[A]\mathrm{E}[D]\mathrm{cov}[B,C]\end{aligned}$$

[12]. Using this result we have

$$\begin{aligned}\mathrm{var}\left[\frac{n}{n+1}\bar{X}^2\right] &= \left(\frac{n}{n+1}\right)^2\mathrm{var}[\bar{X}^2]=\left(\frac{n}{n+1}\right)^2\left\{2\mathrm{var}[\bar{X}]^2+4\mu^2\mathrm{var}[\bar{X}]\right\}\\ &= \left(\frac{n}{n+1}\right)^2\left[\frac{2\mu^4}{n^2}+\frac{4\mu^4}{n}\right]=\frac{2\mu^4}{n}\left(\frac{n}{n+1}\right)^2\left(\frac{1}{n}+4\right).\end{aligned}$$

[1]Observe that this identity comes from the general identity

$$\begin{aligned}&\mathrm{cov}\left[AB,CD\right]\\ =\ &\mathrm{cov}[A,C]\mathrm{cov}[B,D]+\mathrm{cov}[A,D]\mathrm{cov}[B,C]+\mathrm{E}[A]\mathrm{cum}[B,C,D]+\mathrm{E}[B]\mathrm{cum}[A,C,D]\\ &+\mathrm{E}[D]\mathrm{cum}[A,B,C]+\mathrm{E}[C]\mathrm{cum}[A,B,D]+\mathrm{cum}[A,B,C,D]\\ &+\mathrm{cov}[A,C]\mathrm{E}[B]\mathrm{E}[D]+\mathrm{cov}[A,D]\mathrm{E}[B]\mathrm{E}[C]+\mathrm{E}[A]\mathrm{E}[C]\mathrm{cov}[B,D]+\mathrm{E}[A]\mathrm{E}[D]\mathrm{cov}[B,C]\end{aligned}$$

recalling that cum denotes cumulant and are the coefficients of the cumulant generating function (`https://en.wikipedia.org/wiki/Cumulant`), which applies to non-Gaussian random variables too

[2]Note that $\mathrm{cum}(A,B,C)$ is the coefficient of $t_1t_2t_3$ in the series expansion of $\log\mathrm{E}[e^{t_1A+t_2B+t_3B}]$ and can be obtained with $\frac{\partial^3\log\mathrm{E}[e^{t_1A+t_2B+t_3B}]}{\partial t_1\partial t_2\partial t_3}\rfloor_{t_1,t_2,t_3=0}$

On the other hand using that s^2 has a chi-square distribution with $n-1$ degrees of freedom (with variance $2(n-1)^2$) we have

$$\text{var}\left[s^2\right] = \frac{2\mu^4}{(n-1)}.$$

Altogether the variance of these two difference estimators of μ^2 are

$$\text{var}\left[\frac{n}{n+1}\bar{X}^2\right] = \frac{2\mu^4}{n}\left(\frac{n}{n+1}\right)^2\left(4+\frac{1}{n}\right) \text{ and var}\left[s^2\right] = \frac{2\mu^4}{(n-1)}.$$

There is no estimator which clearly does better than the other. And the matter gets worse, since any convex combination is also an estimator! This illustrates that Lehman-Scheffe theorem does not hold in this case; we recall that Lehman-Scheffe theorem states that under completeness any unbiased estimator of a sufficient statistic has minimal variance. In this case we have two different unbiased estimators of sufficient statistics neither estimator is uniformly better than another.

Remark 2.1.1 *Note, to estimate μ one could use $\bar{X}$ or $\sqrt{s^2} \times sign(\bar{X})$ (though it is unclear to me whether the latter is unbiased).*

Exercise 2.1 *Calculate (the best you can)* $\mathrm{E}[\sqrt{s^2} \times sign(\bar{X})]$.

Example 2

Let us return to the censored data example considered in Sections 1.2 and 1.6.4, Example (v). $\{X_i\}_{i=1}^n$ are iid exponential distributed random variables, however we do not observe X_i we observe a censored version $Y_i = \min(X_i, c)$ (c is assumed known) and $\delta_i = 0$ if $Y_i = X_i$ else $\delta_i = 1$.

We recall that the log-likelihood of (Y_i, δ_i) is

$$\begin{aligned}\mathcal{L}_n(\theta) &= \sum_i (1-\delta_i)\left\{-\theta Y_i + \log\theta\right\} - \sum_i \delta_i c\theta \\ &= -\sum_i \theta Y_i - \log\theta \sum_i \delta_i + n\log\theta,\end{aligned}$$

since $Y_i = c$ when $\delta_i = 1$. hence the minimal sufficient statistics for θ are $\sum_i \delta_i$ and $\sum_i Y_i$. This suggests there may be several different estimators for θ.

(i) $\sum_{i=1}^{n} \delta_i$ gives the number of observations which have been censored. We recall that $P(\delta_i = 1) = \exp(-c\theta)$, thus we can use $n^{-1} \sum_{i=1}^{n} \delta_i$ as an estimator of $\exp(-c\theta)$ and solve for θ.

(ii) The non-censored observations also convey information about θ. The likelihood of a non-censored observations is

$$\mathcal{L}_{nC,n}(\theta) = -\theta \sum_{i=1}^{n} (1 - \delta_i) Y_i + \sum_{i=1}^{n} (1 - \delta_i) \left\{ \log \theta - \log(1 - e^{-c\theta}) \right\}.$$

One could maximise this to obtain an estimator of θ

(iii) Or combine the censored and non-censored observations by maximising the likelihood of θ given (Y_i, θ_i) to give the estimator

$$\frac{\sum_{i=1}^{n} (1 - \delta_i)}{\sum_{i=1}^{n} Y_i}.$$

The estimators described above are not unbiased (hard to take the expectation), but they do demonstrate that often there is often no unique best method for estimating a parameter.

Though it is usually difficult to find an estimator which has the smallest variance for all sample sizes, in general the maximum likelihood estimator "asymptotically" (think large sample sizes) usually attains the Cramer-Rao bound. In other words, it is "asymptotically" efficient.

Exercise 2.2 (Two independent samples from a normal distribution) *Suppose that $\{X_i\}_{i=1}^{m}$ are iid normal random variables with mean μ and variance σ_1^2 and $\{Y_i\}_{i=1}^{m}$ are iid normal random variables with mean μ and variance σ_2^2. $\{X_i\}$ and $\{Y_i\}$ are independent, calculate their joint likelihood.*

(i) Calculate their sufficient statistics.

(ii) Propose a class of estimators for μ.

2.2 The Maximum likelihood estimator

There are many different parameter estimation methods. However, if the family of distributions from the which the parameter comes from is known, then the maximum likelihood

estimator of the parameter θ, which is defined as

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} L_n(\underline{X};\theta) = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta),$$

is the most commonly used. Often we find that $\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\widehat{\theta}_n} = 0$, hence the solution can be obtained by solving the derivative of the log likelihood (the derivative of the log-likelihood is often called the *score function*). However, if θ_0 lies on the boundary of the parameter space this will *not* be true. In general, the maximum likelihood estimator will not be an unbiased estimator of the parameter.

We note that the likelihood is invariant to bijective transformations of the data. For example if X has the density $f(\cdot;\theta)$ and we define the transformed random variable $Z = g(X)$, where the function g has an inverse, then it is easy to show that the density of Z is $f(g^{-1}(z);\theta)\frac{\partial g^{-1}(z)}{\partial z}$. Therefore the likelihood of $\{Z_i = g(X_i)\}$ is

$$\prod_{i=1}^{n} f(g^{-1}(Z_i);\theta)\frac{\partial g^{-1}(z)}{\partial z}\rfloor_{z=Z_i} = \prod_{i=1}^{n} f(X_i;\theta)\frac{\partial g^{-1}(z)}{\partial z}\rfloor_{z=Z_i}.$$

Hence it is proportional to the likelihood of $\{X_i\}$ and the maximum of the likelihood in terms of $\{Z_i = g(X_i)\}$ is the same as the maximum of the likelihood in terms of $\{X_i\}$.

Example 2.2.1 (The uniform distribution) *Consider the uniform distribution, which has the density $f(x;\theta) = \theta^{-1}I_{[0,\theta]}(x)$. Given the iid uniform random variables $\{X_i\}$ the likelihood (it is easier to study the likelihood rather than the log-likelihood) is*

$$L_n(\underline{X}_n;\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n} I_{[0,\theta]}(X_i).$$

Using $L_n(\underline{X}_n;\theta)$, the maximum likelihood estimator of θ is $\widehat{\theta}_n = \max_{1\le i\le n} X_i$ (you can see this by making a plot of $L_n(\underline{X}_n;\theta)$ against θ).

To derive the properties of $\max_{1\le i\le n} X_i$ we first obtain its distribution. It is simple to see that

$$P(\max_{1\le i\le n} X_i \le x) \;=\; P(X_1 \le x,\ldots,X_n \le x) = \prod_{i=1}^{n} P(X_i \le x) = \big(\frac{x}{\theta}\big)^n I_{[0,\theta]}(x),$$

and the density of $\max_{1\le i\le n} X_i$ is $f_{\widehat{\theta}_n}(x) = nx^{n-1}/\theta^n$.

Exercise 2.3 *(i) Evaluate the mean and variance of $\widehat{\theta}_n$ defined in the above example.*

(ii) Is the estimator biased? If it is, find an unbiased version of the estimator.

Example 2.2.2 (Weibull with known α) *$\{Y_i\}$ are iid random variables, which follow a Weibull distribution, which has the density*

$$\frac{\alpha y^{\alpha-1}}{\theta^{\alpha}}\exp(-(y/\theta)^{\alpha}) \qquad \theta,\alpha>0.$$

Suppose that α is known, but θ is unknown. Our aim is to fine the MLE of θ.

The log-likelihood is proportional to

$$\begin{aligned}\mathcal{L}_n(\underline{X};\theta) &= \sum_{i=1}^{n}\left(\log\alpha+(\alpha-1)\log Y_i-\alpha\log\theta-\left(\frac{Y_i}{\theta}\right)^{\alpha}\right)\\ &\propto \sum_{i=1}^{n}\left(-\alpha\log\theta-\left(\frac{Y_i}{\theta}\right)^{\alpha}\right).\end{aligned}$$

The derivative of the log-likelihood wrt to θ is

$$\frac{\partial\mathcal{L}_n}{\partial\theta}=-\frac{n\alpha}{\theta}+\frac{\alpha}{\theta^{\alpha+1}}\sum_{i=1}^{n}Y_i^{\alpha}=0.$$

Solving the above gives $\widehat{\theta}_n=(\frac{1}{n}\sum_{i=1}^{n}Y_i^{\alpha})^{1/\alpha}$.

Example 2.2.3 (Weibull with unknown α) *Notice that if α is given, an explicit solution for the maximum of the likelihood, in the above example, can be obtained. Consider instead the case that both α and θ are unknown. Now we need to find α and θ which maximise the likelihood i.e.*

$$\arg\max_{\theta,\alpha}\sum_{i=1}^{n}\left(\log\alpha+(\alpha-1)\log Y_i-\alpha\log\theta-(\frac{Y_i}{\theta})^{\alpha}\right).$$

The derivative of the likelihood is

$$\begin{aligned}\frac{\partial\mathcal{L}_n}{\partial\theta} &= -\frac{n\alpha}{\theta}+\frac{\alpha}{\theta^{\alpha+1}}\sum_{i=1}^{n}Y_i^{\alpha}=0\\ \frac{\partial\mathcal{L}_n}{\partial\alpha} &= \frac{n}{\alpha}-\sum_{i=1}^{n}\log Y_i-n\log\theta-\frac{n\alpha}{\theta}+\sum_{i=1}^{n}\log(\frac{Y_i}{\theta})\times(\frac{Y_i}{\theta})^{\alpha}=0.\end{aligned}$$

It is clear that an explicit expression to the solution of the above does not exist and we need to find alternative methods for finding a solution (later we show how profiling can be used to estimate α).

2.3 Maximum likelihood estimation for the exponential class

Typically when maximising the likelihood we encounter several problems (i) for a given likelihood $\mathcal{L}_n(\theta)$ the maximum may lie on the boundary (even if in the limit of $\mathcal{L}_n$ the maximum lies with in the parameter space) (ii) there are several local maximums (so a numerical routine may not capture the true maximum) (iii) $\mathcal{L}_n$ may not be concave, so even if you are close the maximum the numerical routine just cannot find the maximum (iv) the parameter space may not be convex (ie. $(1-\alpha)\theta_1 + \alpha\theta_2$ may lie outside the parameter space even if θ_1 and θ_2 are in the parameter space) again this will be problematic for numerically maximising over the parameter space. When there is just one unknown parameters these problems are problematic, when the number of unknown parameters is p this becomes a nightmare. However for the full rank exponential class of distributions we now show that everything behaves, in general, very well. First we heuristically obtain its maximum likelihood estimator, and later justify it.

2.3.1 Full rank exponential class of distributions

Suppose that $\{X_i\}$ are iid random variables which has a the natural exponential representation and belongs to the family $\mathcal{F} = \{f(x;\theta) = \exp[\sum_{j=1}^p \theta_j s_j(x) - \kappa(\theta) + c(x)]; \theta \in \Theta\}$ and $\Theta = \{\theta; \kappa(\theta) = \log \int \exp(\sum_{j=1}^p \theta_j s_j(x) + c(x))dx < \infty\}$ (note this condition defines the parameter space, if $\kappa(\theta) = \infty$ the density is no longer defined). Therefore the log likelihood function is

$$\mathcal{L}_n(\underline{X};\theta) = \theta \sum_{i=1}^n \mathbf{s}(X_i) - n\kappa(\theta) + \sum_{i=1}^n c(X_i),$$

where $\sum_{i=1}^n \mathbf{s}(X_i) = (\sum_{i=1}^n s_1(X_i), \ldots, \sum_{i=1}^n s_p(X_i))$ are the sufficient statistics. By the Rao-Blackwell theorem the unbiased estimator with the smallest variance will be a function of $\sum_{i=1}^n \mathbf{s}(X_i)$. We now show that the maximum likelihood estimator of θ is a function of $\sum_{i=1}^n \mathbf{s}(X_i)$ (though there is no guarantee it will be unbiased);

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} \big\{\theta \sum_{i=1}^n \mathbf{s}(X_i) - n\kappa(\theta) + \sum_{i=1}^n c(X_i)\big\}.$$

The natural way to obtain $\widehat{\theta}_n$ is to solve

$$\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta=\widehat{\theta}_n} = 0.$$

However, this equivalence will only hold if the maximum lies *within* the interior of the parameter space (we show below that in general this will be true). Let us suppose this is true, then differentiating $\mathcal{L}_n(\underline{X};\theta)$ gives

$$\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta} = \sum_{i=1}^{n} \mathbf{s}(X_i) - n\kappa'(\theta) = 0.$$

To simplify notation we often write $\kappa'(\theta) = \mu(\theta)$ (since this is the mean of the sufficient statistics). Thus we can invert back to obtain the maximum likelihood estimator

$$\widehat{\theta}_n = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i)\right). \tag{2.2}$$

Because the likelihood is a concave function, it has a unique maximum. But the maximum will only be at $\widehat{\theta}_n = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i)\right)$ *if* $\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i) \in \mu(\Theta)$. If $\mu^{-1}(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i))$ takes us outside the parameter space, then clearly this cannot be an estimator of the parameter[3]. Fortunately, in most cases (specifically, if the model is said to be "steep"), $\mu^{-1}(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i))$ will lie in the interior of the parameter space. In other words,

$$\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i)\right) = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta).$$

In the next section we define steepness and what may happen if this condition is not satisfied. But first we go through a few examples.

Example 2.3.1 (Normal distribution) *For the normal distribution, the log-likelihood is*

$$\mathcal{L}(\underline{X};\sigma^2,\mu) = \frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n} X_i^2 - 2\mu\sum_{i=1}^{n} X_i + n\mu^2\right) - \frac{n}{2}\log\sigma^2,$$

note we have ignored the 2π constant. Differentiating with respect to σ^2 and μ and setting to zero gives

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \widehat{\mu})^2.$$

[3]For example and estimator of the variance which is negative, clearly this estimator has no meaning

This is the only solution, hence it must the maximum of the likelihood.

Notice that $\widehat{\sigma}^2$ is a slightly biased estimator of σ^2.

Example 2.3.2 (Multinomial distribution) *Suppose $\underline{Y} = (Y_1, \ldots, Y_q)$ (with $n = \sum_{i=1}^q Y_i$) has a multinomial distribution where there are q cells. Without any constraints on the parameters the log likelihood is proportional to (we can ignore the term $c(\underline{Y}) = \log \binom{n}{Y_1,\ldots,Y_q}$)*

$$\mathcal{L}_n(\underline{Y}; \pi) = \sum_{j=1}^{q-1} Y_i \log \pi_i + Y_q \log(1 - \sum_{i=1}^{q-1} \pi_i).$$

The partial derivative for each i is

$$\frac{\mathcal{L}(\underline{Y}; \pi)}{\partial \pi_i} = \frac{Y_i}{\pi} - \frac{Y_p}{1 - \sum_{i=1}^{p-1} \pi_i}.$$

Solving the above we get one solution as $\widehat{\pi}_i = Y_i/n$ (check by plugging it in).

Since there is a diffeomorphism between $\{\pi_i\}$ and its natural parameterisation $\theta_i = \log \pi_i/(1 - \sum_{j=1}^{q-1} \pi_i)$ and the Hessian corresponding to the natural paramerisation is negative definite (recall the variance of the sufficient statistics is $\kappa''(\theta)$), this implies that the Hessian of $\mathcal{L}_n(\underline{Y}; \pi)$ is negative definite, thus $\widehat{\pi}_i = Y_i/n$ is the unique maximum of $\mathcal{L}(\underline{Y}; \pi)$.

Example 2.3.3 ($2 \times 2 \times 2$ Contingency tables) *Consider the example where for n individuals three binary variables are recorded; Z =gender (here, we assume two), X =whether they have disease A (yes or no) and Y =whether they have disease B (yes or no). We assume that the outcomes of all n individuals are independent.*

Without any constraint on variables, we model the above with a multinomial distribution with $q = 3$ i.e. $P(X = x, Y = y, Z = z) = \pi_{xyz}$. In this case the likelihood is proportional to

$$\begin{aligned}
\mathcal{L}(\underline{Y}; \pi) &= \sum_{x=0}^{1} \sum_{y=0}^{1} \sum_{z=0}^{1} Y_{xyz} \log \pi_{xyz} \\
&= Y_{000}\pi_{000} + Y_{010}\pi_{010} + Y_{000}\pi_{001} + Y_{100}\pi_{100} + Y_{110}\pi_{110} + Y_{101}\pi_{101} \\
&\quad Y_{011}\pi_{011} + Y_{111}(1 - \pi_{000} - \pi_{010} - \pi_{001} - \pi_{100} - \pi_{110} - \pi_{101} - \pi_{011}).
\end{aligned}$$

Differentiating with respect to each variable and setting to one it is straightfoward to see that the maximum is when $\widehat{\pi}_{xyz} = Y_{xyz}/n$; which is intuitively what we would have used as the estimator.

However, suppose the disease status of X and Y are independent conditioned on gender. i.e. $P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z)$ then $P(X = x, Y = y, Z = z) = \pi_{X=x|Z=z}\pi_{Y=y|Z=z}\pi_{Z=z}$, since these are binary variables we drop the number of unknown parameters from 7 to 5. This is a curved exponential model (though in this case the constrained model is simply a 5-dimensional hyperplane in 7 dimensional space; thus the parameter space is convex). The log likelihood is proportional to

$$\begin{aligned}\mathcal{L}(\underline{Y};\pi) &= \sum_{x=0}^{1}\sum_{y=0}^{1}\sum_{z=0}^{1} Y_{xyz}\log \pi_{x|z}\pi_{y|z}\pi_z \\ &= \sum_{x=0}^{1}\sum_{y=0}^{1}\sum_{z=0}^{1} Y_{xyz}\left(\log \pi_{x|z} + \log \pi_{y|z} + \log \pi_z\right).\end{aligned}$$

Thus we see that the maximum likelihood estimators are

$$\widehat{\pi}_z = \frac{Y_{++z}}{n} \quad \widehat{\pi}_{x|z} = \frac{Y_{x+z}}{Y_{++z}} \quad \widehat{\pi}_{y|z} = \frac{Y_{+yz}}{Y_{++z}}.$$

Where in the above we use the standard notation $Y_{+++} = n$, $Y_{++z} = \sum_{x=0}^{1}\sum_{y=0}^{1} Y_{xyz}$ etc. We observe that these are very natural estimators. For example, it is clear that Y_{x+z}/n is an estimator of the joint distribution of X and Z and Y_{++z}/n is an estimator of the marginal distribution of Z. Thus Y_{x+z}/Y_{++z} is clearly an estimator of X conditioned on Z.

Exercise 2.4 *Evaluate the mean and variance of the numerator and denominator of (2.4). Then use the continuous mapping theorem to evaluate the limit of $\widehat{\theta^{-1}}$ (in probability).*

Example 2.3.4 (The beta distribution) *Consider the family of densities defined by*

$$\mathcal{F} = \left\{f(x;\alpha,\beta) = B(\alpha,\beta)^{-1}x^{\alpha-1}(1-x)^{\beta-1}; \alpha \in (0,\infty), \beta \in (0,\infty)\right\}$$

and $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$. This is called the family of beta distributions.

The log likelihood can be written as

$$\begin{aligned}\mathcal{L}_n(\underline{X};\theta) &= \alpha\sum_{i=1}^{n}\log X_i + \beta\sum_{i=1}^{n}\log(1-X_i) - n\left[\log(\Gamma(\alpha)) + \log(\Gamma(\beta) - \log(\Gamma(\alpha+\beta))\right] - \\ &\quad \sum_{i=1}^{n}[\log X_i - \log(1-X_i)].\end{aligned}$$

Thus $\theta_1 = \alpha$, $\theta_2 = \beta$ and $\kappa(\theta_1, \theta_2) = \log(\theta_1) + \log(\theta_2) - \log(\theta_1 + \theta_2)$.

Taking derivatives and setting to zero gives

$$\frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \log X_i \\ \sum_{i=1}^n \log(1-X_i) \end{pmatrix} = \begin{pmatrix} \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} \\ \frac{\Gamma'(\beta)}{\Gamma(\beta)} - \frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} \end{pmatrix}.$$

To find estimators for α and β we need to numerically solve for the above. But will the solution lie in the parameter space?

Example 2.3.5 (Inverse Gaussian distribution) *Consider the inverse Gaussian distribution defined as*

$$f(x;\theta_1,\theta_2) = \frac{1}{\pi^{1/2}} x^{-3/2} \exp\left(\theta_1 x - \theta_2 x^{-1} + [-2(\theta_1\theta_2)^{1/2} - \frac{1}{2}\log(-\theta_2)]\right),$$

where $x \in (0,\infty)$. Thus we see that $\kappa(\theta_1,\theta_2) = [-2(\theta_1\theta_2)^{1/2} - \frac{1}{2}\log(-\theta_2)]$. In this case we observe that for $\theta_1 = 0$ $\kappa(0,\theta_2) < \infty$ thus the parameter space is not open and $\Theta = (-\infty, 0] \times (-\infty, 0)$. Taking derivatives and setting to zero gives

$$\frac{1}{n}\begin{pmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i^{-1} \end{pmatrix} = \begin{pmatrix} -\left(\frac{\theta_2}{\theta_1}\right)^{1/2} \\ -\frac{\theta_1^{1/2}}{\theta_2^{1/2}} + \frac{1}{\theta_2}. \end{pmatrix}.$$

To find estimators for α and β we need to numerically solve for the above. But will the solution lie in the parameter space?

Example 2.3.6 (The inflated zero Poisson distribution) *Using the natural parameterisation of the infltaed zero Poisson distribution we have*

$$\begin{aligned} \mathcal{L}(\underline{Y};\theta_1,\theta_2) &= \theta_1 \sum_{i=1}^n I(Y_i \neq 0) + \theta_2 \sum_{i=1}^n I(Y_i \neq 0) Y_i \\ &\quad - \log\left(\frac{e^{\theta_1} - \theta_2^{-1}}{1-\theta_2^{-1}}(1 - e^{-e^{\theta_2}}) + e^{-e^{\theta_2}}\right). \end{aligned}$$

where the parameter space is $\Theta = (-\infty, 0] \times (-\infty, \infty)$, which is not open (note that 0 corresponds to the case $p = 0$, which is the usual Poisson distribution with no inflation).

To find estimators for θ and p we need to numerically solve for the above. But will the solution lie in the parameter space?

2.3.2 Steepness and the maximum of the likelihood

The problem is that despite the Hessian $\nabla^2\mathcal{L}(\theta)$ being non-negative definite, it could be that the maximum is at the boundary of the likelihood. We now state some results that show that in most situations, this does not happen and usually (2.2) maximises the likelihood. For details see Chapters 3 and 5 of `http://www.jstor.org/stable/pdf/4355554.pdf?acceptTC=true` (this reference is mathematically quite heavy) for a maths lite review see Davidson (2004) (page 170). Note that Brown and Davidson use the notation $\mathcal{N}$ to denote the parameter space Θ.

Let $\mathcal{X}$ denote the range of the sufficient statistics $\mathbf{s}(X_i)$ (i.e. what values can $s(X)$ take). Using this we define its convex hull as

$$C(\mathcal{X}) = \{\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2; \quad \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, 0 \leq \alpha \leq 1\}.$$

Observe that $\frac{1}{n}\sum_i \mathbf{s}(X_i) \in C(\mathcal{X})$, even when $\frac{1}{n}\sum_i \mathbf{s}(X_i)$ does not belong to the observation space of the sufficient statistic $\mathcal{X}$. For example X_i may be counts from a Binomial distribution $\text{Bin}(m, p)$ but $C(\mathcal{X})$ would be the reals between $[0, m]$.

Example 2.3.7 (Examples of $C(\mathcal{X})$) *(i) The normal distribution*

$$C(\mathcal{X}) = \left\{\alpha(x, x^2) + (1-\alpha)(y, y^2); \quad x, y \in \mathbb{R}, 0 \leq \alpha \leq 1\right\} = (-\infty, \infty)(0, \infty).$$

(ii) The β-distribution

$$C(\mathcal{X}) = \{\alpha(\log x, \log(1-x)) + (1-\alpha)(\log y, \log(1-y)); \quad x, y \in [0,1], 0 \leq \alpha \leq 1\} = (\mathbb{R}^{-1})^2$$

(iii) The exponential with censoring (see 2.3)

$$C(\mathcal{X}) = \{\alpha(y_1, \delta_1) + (1-\alpha)(y_2, \delta_2); \quad y_1 \in [0, c], \delta_1, \delta_2 = \{0, 1\}; 0 \leq \alpha \leq 1\} = \textit{triangle}.$$

(iv) The binomial distribution $Y \sim Bin(n, \pi)$. Then

$$C(\mathcal{X}) = \{\alpha x + (1-\alpha)y; 0 \leq \alpha \leq 1, y = 0, \ldots, m\} = [0, m].$$

Now we give conditions under which $\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right)$ maximises the likelihood within the parameter space Θ. Define the parameter space $\Theta = \{\theta; \kappa(\theta) < \infty\} \subset \mathbb{R}^q$. Let $int(\Theta)$ denote the interior of a set, which is the largest open set in Θ. Next we define the notion of *steep*.

Definition 2.3.1 *Let $\kappa : \mathbb{R}^p \to (-\infty, \infty)$ be a convex function (so $-\kappa$ is concave). κ is called steep if for all $\theta_1 \in B(\Theta)$ and $\theta_0 \in int(\Theta)$, $\lim_{\rho\to\infty}(\theta_1 - \theta_0)\frac{\partial\kappa(\theta)}{\partial\theta}\big\rfloor_{\theta=\theta_0+\rho(\theta_1-\theta_0)} = \infty$. This condition is equivalent to $\lim_{\theta\to B(\Theta)} |\kappa'(\theta)| \to \infty$. Intuitively, steep simply means the function is very steep at the boundary.*

- *Regular exponential family*

  If the parameter space is open (such as $\Theta = (0, 1)$ or $\Theta = (0, \infty)$) meaning the density is not defined on the boundary, then the family of exponentials is called a regular family.

  In the case that Θ is open (the boundary does not belong to Θ), then κ is not defined at the boundary, in which case κ is steep.

  Note, at the boundary $\lim_{\theta\to B(\Theta)} \log f(x;\theta)$ will approach $-\infty$, since $\{\log f(x;\theta)\}$ is convex over θ this means that its maximum will be within the interior of the parameter space (just what we want!).

- *Non-regular exponential family*

  If the parameter space is closed, this means at the boundary the density is defined, then we require that at the boundary of the parameter space $\kappa(\cdot)$ is *steep*. This condition needs to be checked by considering the expectation of the sufficient statistic at the boundary or equivalently calculating $\kappa'(\cdot)$ at the boundary.

If $\kappa(\theta)$ is steep we have the following result. Brown (1986), Theorem 3.6 shows that there is a homeomorphism[4] between $int(\Theta)$ and $int(C(\mathcal{X}))$.

Most *importantly* Brown (1986), Theorem 5.5 shows that if the density of X_i belongs to a full rank exponential family (using the natural parameterisation) $f(x;\theta) = \exp[\sum_{j=1}^p \theta_j s_j(x) - \kappa(\theta) + c(x)]$ with $\theta = (\theta_1, \ldots, \theta_p) \in \Theta$, where $\kappa(\cdot)$ is steep and for a given data set $\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i) \in C(\mathcal{X})$, then

$$\widehat{\theta}_n = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right) = \arg\max_{\theta\in int(\Theta)} \big\{\theta\sum_{i=1}^n \mathbf{x}_i - n\kappa(\theta) + \sum_{i=1}^n c(X_i)\big\}.$$

In most situations the full rank exponential family will have a parameter space which is open and thus steep.

[4] A homeomorphism between two spaces means there is a bijection between two spaces and the f and f^{-1} which maps between the two spaces is continuous.

Example 2.3.8 (Binomial distribution and observations that lie on the boundary) *Suppose that $\{Y_i\}_{i=1}^n$ are iid Binomially distributed random variables $Y_i \sim Bin(m, \pi_i)$. The log likelihood of Y_i is $Y_i \log(\frac{\pi}{1-\pi}) + m(1-\pi)$. Thus the log likelihood of the sample is proportional to*

$$\mathcal{L}_n(\underline{Y}; \pi) = \sum_{i=1}^n Y_i \log \pi + \sum_{i=1}^n (m - Y_i) \log(1-\pi) = \theta \sum_{i=1}^n Y_i - nm \log(1 + e^\theta),$$

where $\theta \in (-\infty, \infty)$. The theory states above that the maximum of the likelihood lies within the interior of $(-\infty, \infty)$ if $\sum_{i=1}^n Y_i$ lies within the interior of $C(\mathcal{Y}) = (0, nm)$.

On the other hand, there is a positive probability that $\sum_{i=1}^n Y_i = 0$ or $\sum_{i=1}^n Y_i = nm$ (i.e. all successes or all failures). In this case, the above result is not informative. However, a plot of the likelihood in this case is very useful (see Figure 2.1). More precisely, if $\sum_i Y_i = 0$, then $\widehat{\theta}_n = -\infty$ (corresponds to $\widehat{p} = 0$), if $\sum_i Y_i = nm$, then $\widehat{\theta}_n = \infty$ (corresponds to $\widehat{p} = 1$). Thus even when the sufficient statistics lie on the boundary of $C(\mathcal{Y})$ we obtain a very natural estimator for θ.

Example 2.3.9 (Inverse Gaussian and steepness) *Consider the log density of the inverse Gaussian, where X_i are iid positive random variables with log likelihood*

$$\mathcal{L}(\underline{X}; \theta) = \theta_1 \sum_{i=1}^n X_i + \theta_2 \sum_{i=1}^n X_i^{-1} - n\kappa(\theta_1, \theta_2) - \frac{3}{2} \sum_{i=1}^n \log X_i - \frac{1}{2} \log \pi,$$

where $\kappa(\theta_1, \theta_2) = -2\sqrt{\theta_1 \theta_2} - \frac{1}{2} \log(-2\theta_2)$. Observe that $\kappa(0, \theta_2) < \infty$ hence $(\theta_1, \theta_2) \in (-\infty, 0] \times (-\infty, 0)$.

However, at the boundary $\frac{\partial \kappa(\theta_1, \theta_2)}{\partial \theta_1} \rfloor_{\theta_1 = 0} = -\infty$. Thus the inverse Gaussian distribution is steep but non-regular. Thus the MLE is $\mu^{-1}(\cdot)$.

Example 2.3.10 (Inflated zero Poisson) *Recall*

$$\begin{aligned} \mathcal{L}(\underline{Y}; \theta_1, \theta_2) &= \theta_1 \sum_{i=1}^n I(Y_i \neq 0) + \theta_2 \sum_{i=1}^n I(Y_i \neq 0) Y_i \\ &\quad - \log \left(\frac{e^{\theta_1} - \theta_2^{-1}}{1 - \theta_2^{-1}} (1 - e^{-e^{\theta_2}}) + e^{-e^{\theta_2}} \right). \end{aligned}$$

where the parameter space is $\Theta = (-\infty, 0] \times (-\infty, \infty)$, which is not open (note that 0 corresponds to the case $p = 0$, which is the usual Poisson distribution with no inflation).

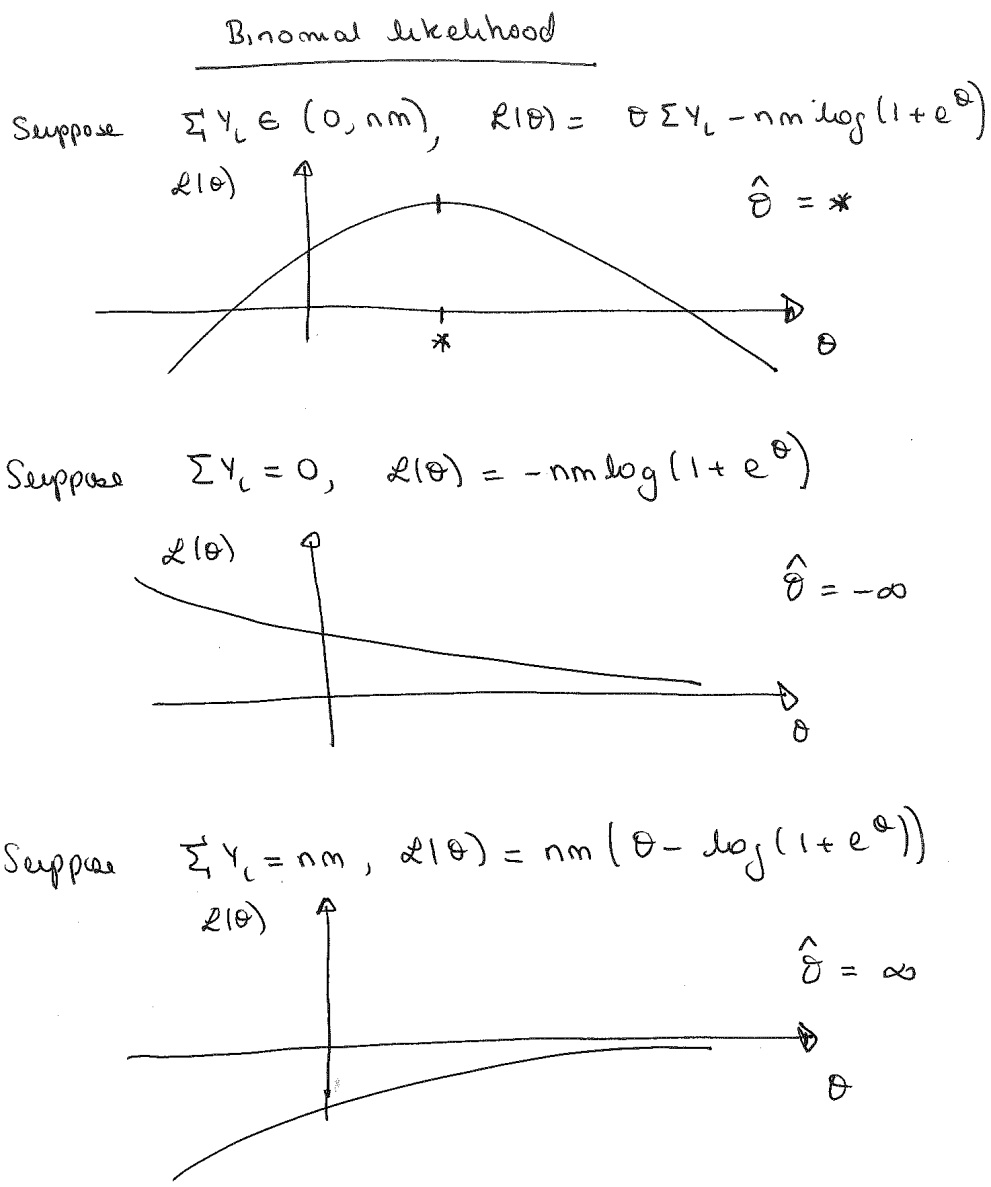

Figure 2.1: Likelihood of Binomial for different scenarios

However, the derivative $\frac{\partial \kappa(\theta_1,\theta_2)}{\partial \theta_1}$ *is finite at* $\theta_1 = 0$ *(for* $\theta_2 \in \mathbb{R}$*). Thus* $\kappa(\cdot)$ *is not steep and care needs to be taken in using* μ^{-1} *as the MLE.*

$\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i))$ *may lie outside the parameter space. For example,* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i))$ *may give an estimator of* θ_1 *which is greater than zero; this corresponds to the probability* $p < 0$*, which makes no sense. If* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i))$ *lies out the parameter space we need to search on the boundary for the maximum.*

Example 2.3.11 (Constraining the parameter space) *If we place an "artifical" constraint on the parameter space then a maximum may not exist within the interior of the parameter space. For example, if we model survival times using the exponential distribution* $f(x;\theta) = \theta\exp(-\theta x)$ *the parameter space is* $(0,\infty)$*, which is open (thus with probability one the likelihood is maximised at* $\widehat{\theta} = \mu^{-1}(\bar{X}) = 1/\bar{X}$*). However, if we constrain the parameter space* $\widetilde{\Theta} = [2,\infty)$*,* $1/\bar{X}$ *may lie outside the parameter space and we need to use* $\widehat{\theta} = 2$*.*

Remark 2.3.1 (Estimating ω**)** *The results above tell us if* $\kappa(\cdot)$ *is steep in the parameter space and* $\mathcal{L}_n(\theta)$ *has a unique maximum and there is a diffeomorphism between* θ *and* ω *(if the exponential family is full rank), then* $\mathcal{L}_n(\theta(\omega))$ *will have a unique maximum. Moreover the Hessian of the likelihood of both parameterisations will be negative definite. Therefore, it does not matter if we maximise over the natural parametrisation or the usual parameterisation*

$$\widehat{\omega}_n = \eta^{-1}\left(\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\right)\right).$$

Remark 2.3.2 (Minimum variance unbiased estimators) *Suppose* X_i *has a distribution in the natural exponential family, then the maximum likelihood estimator is a function of the sufficient statistic* $s(\underline{X})$*. Moreover if the exponential is full and* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i)$ *is an* **unbiased** *estimator of* θ*, then* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i)$ *is the minumum variance unbiased estimator of* θ*. However, in general* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i)$ *will not be an unbiased estimator. However, by invoking the continuous mapping theorem (*`https://en.wikipedia.org/wiki/Continuous_mapping_theorem`*), by the law of large numbers* $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \xrightarrow{a.s.} \mathrm{E}[\mathbf{x}]_i$*, then* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i) \xrightarrow{a.s.} \mu^{-1}(\mathrm{E}(\mathbf{x})) = \mu^{-1}[\kappa'(\theta)] = \theta$*. Thus the maximum likelihood estimator converges to* θ*.*

2.3.3 The likelihood estimator of the curved exponential

Example 2.3.12 (Normal distribution with constraint) *Suppose we place the constraint on the parameter space $\sigma^2 = \mu^2$. The log-likelihood is*

$$\mathcal{L}(\underline{X};\mu) = \frac{-1}{2\mu^2}\sum_{i=1}^{n} X_i^2 + \frac{1}{\mu}\sum_{i=1}^{n} X_i - \frac{n}{2}\log\mu^2.$$

Recall that this belongs to the curved exponential family and in this case the parameter space is not convex. Differentiating with respect to μ gives

$$\frac{\partial\mathcal{L}(\underline{X};\sigma^2,\mu)}{\partial\mu} = \frac{1}{\mu^3}S_{xx} - \frac{S_x}{\mu^2} - \frac{1}{\mu} = 0.$$

Solving for μ leads to the quadratic equation

$$p(\mu) = \mu^2 + S_x\mu - S_{xx} = 0.$$

Clearly there will be two real solutions

$$\frac{-S_x \pm \sqrt{S_x^2 + 4S_{xx}}}{2}.$$

We need to plug them into the log-likelihood to see which one maximises the likelihood.

Observe that in this case the Hessian of the log-likelihood cannot be negative (unlike the full normal). However, we know that a maximum exists since a maximum exists on for the full Gaussian model (see the previous example).

Example 2.3.13 (Censored exponential) *We recall that the likelihood corresponding to the censored exponential is*

$$\mathcal{L}_n(\theta) = -\theta\sum_{i=1}^{n} Y_i - \log\theta\sum_{i=1}^{n}\delta_i + \log\theta. \tag{2.3}$$

We recall that $\delta_i = 1$ if censoring takes place. The maximum likelihood estimator is

$$\widehat{\theta} = \frac{n - \sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n} Y_i} \in (0,\infty)$$

Basic calculations show that the mean of the exponential is $1/\theta$, therefore the estimate of the mean is

$$\widehat{\theta^{-1}} = \frac{\sum_{i=1}^{n} Y_i}{\underbrace{n - \sum_{i=1}^{n}\delta_i}_{\text{no. not censored terms}}}. \tag{2.4}$$

If the exponential distribution is curved (number of unknown parameters is less than the number of minimally sufficient statistics), then the parameter space $\Omega = \{\omega = (\omega_1, \ldots, \omega_d); (\theta_1(\omega), \ldots, \theta_q(\omega)) \in \Theta\} \subset \Theta$ (hence it is a curve on Θ). Therefore, by differentiating the likelihood with respect to ω, a maximum within the parameter space must satisfy

$$\nabla_\omega \mathcal{L}_n(\theta(\omega)) = \frac{\partial \theta(\omega)}{\partial \omega} \left(\sum_{i=1}^n \mathbf{x}_i - n \frac{\partial \kappa(\theta)}{\partial \theta} \rfloor_{\theta=\theta(\omega)} \right) = 0. \tag{2.5}$$

Therefore, either (a) there exists an $\omega \in \Omega$ such that $\theta(\omega)$ is the global maximum of $\{\mathcal{L}_n(\theta); \theta \in \Theta\}$ (in this case $\sum_{i=1}^n \mathbf{x}_i - n \frac{\partial \kappa(\theta)}{\partial \theta} \rfloor_{\theta=\theta(\omega)} = 0$) or (b) there exists an $\omega \in \Omega$ such that $\frac{\partial \theta(\omega)}{\partial \omega}$ and $\sum_{i=1}^n \mathbf{x}_i - n \frac{\partial \kappa(\theta)}{\partial \theta} \rfloor_{\theta=\theta(\omega)}$ are orthogonal. Since $\mathcal{L}_n(\underline{X}, \theta)$ for $\theta \in \Theta$ and $\sum_{i=1}^n \mathbf{x}_i \in int(\mathcal{X})$ has a global maximum a simple illustration this means that $\mathcal{L}_n(\theta(\omega))$ will have a maximum. In general (2.5) will be true. As far as I can see the only case where it may not hold is when $\theta(\omega)$ lies on some contour of $\mathcal{L}_n(\theta)$. This suggests that a solution should in general exist for the curved case, but it may not be unique (you will need to read Brown (1986) for full clarification). Based this I suspect the following is true:

- If Ω is a curve in Θ, then $\frac{\partial \mathcal{L}_n(\omega)}{\partial \omega} = 0$ may have multiple solutions. In this case, we have to try each solution $\mathcal{L}_n(\widehat{\omega})$ and use the solution which maximises it (see Figure 2.2).

Exercise 2.5 *The aim of this question is to investigate the MLE of the inflated zero Poisson parameters λ and p. Simulate from a inflated zero poisson distribution with (i) $p = 0.5$, $p = 0.2$ and $p = 0$ (the class is when there is no inflation), use $n = 50$. Evaluate the MLE (over 200 replications) make a Histogram and QQplot of the parameter estimators (remember if the estimator of p is outside the parameter space you need to locate the maximum on the parameter space).*

Exercise 2.6 *(i) Simulate from the model defined in Example 2.3.12 (using $n = 20$) using R. Calculate and maximise the likelihood over 200 replications. Make a QQplot of the estimators and calculate the mean squared error.*

For one realisation make a plot of the log-likelihood.

(ii) Sample from the inverse Gamma distribution (using $n = 20$) and obtain its maximum likelihood estimator. Do this over 200 replications and make a table summarizing its bias and average squared error. Make a QQplot of the estimators.

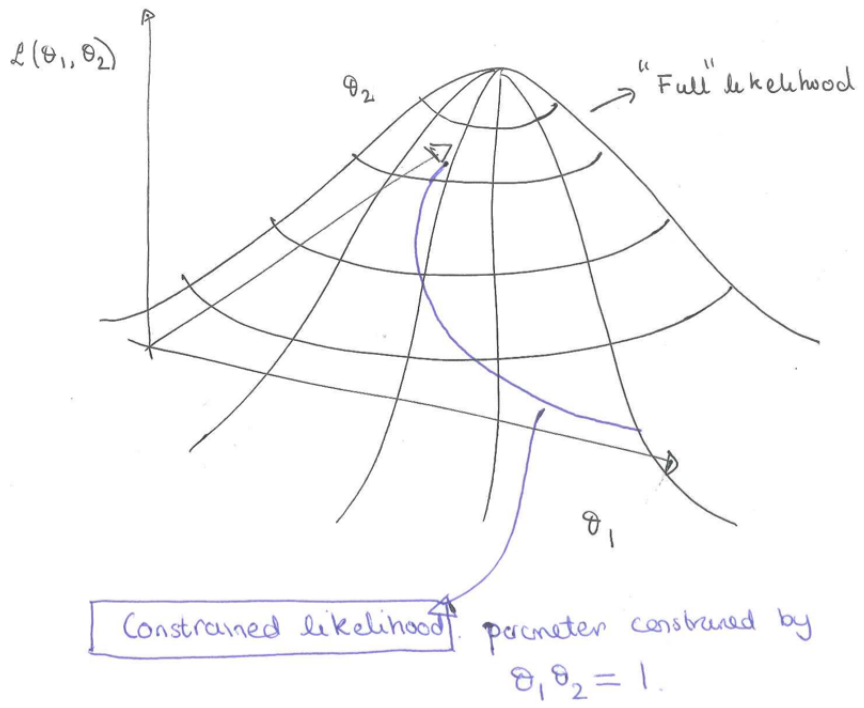

Figure 2.2: Likelihood of 2-dimension curved exponential

(iii) Consider the exponential distribution described in Example 2.3.11 where the parameter space is constrained to $[2, \infty]$*. For samples of size* $n = 50$ *obtain the maximum likelihood estimator (over 200 replications). Simulate using the true parameter*

(a) $\theta = 5$ *(b)* $\theta = 2.5$ *(c)* $\theta = 2$*.*

Summarise your results and make a QQplot (against the normal distribution) and histogram of the estimator.

2.4 The likelihood for dependent data

We mention that the likelihood for dependent data can also be constructed (though often the estimation and the asymptotic properties can be a lot harder to derive). Suppose $\{X_t\}_{t=1}^n$ is a time series (a sequence of observations over time where there could be dependence). Using Bayes rule (ie. $P(A_1, A_2, \ldots, A_n) = P(A_1) \prod_{i=2}^n P(A_i|A_{i-1}, \ldots, A_1)$)

we have

$$L_n(\underline{X};\theta) = f(X_1;\theta)\prod_{t=2}^{n} f(X_t|X_{t-1},\ldots,X_1;\theta).$$

Under certain conditions on $\{X_t\}$ the structure above $\prod_{t=2}^{n} f(X_t|X_{t-1},\ldots,X_1;\theta)$ can be simplified. For example if X_t were Markovian then X_t conditioned on the past on depends only on the recent past, i.e. $f(X_t|X_{t-1},\ldots,X_1;\theta) = f(X_t|X_{t-1};\theta)$ in this case the above likelihood reduces to

$$L_n(\underline{X};\theta) = f(X_1;\theta)\prod_{t=2}^{n} f(X_t|X_{t-1};\theta). \tag{2.6}$$

We apply the above to a very simple time series. Consider the AR(1) time series

$$X_t = \phi X_{t-1} + \varepsilon_t, \qquad t \in \mathbb{Z},$$

where ε_t are iid random variables with mean zero and variance σ^2. In order to ensure that the recurrence is well defined for all $t \in \mathbb{Z}$ we assume that $|\phi| < 1$ in this case the time series is called stationary[5].

We see from the above that the observation X_{t-1} has a linear influence on the next observation and it is Markovian; conditioned on X_{t-1}, X_{t-2} and X_t are independent (the distribution function $P(X_t \leq x|X_{t-1},X_{t-2}) = P(X_t \leq x|X_{t-1})$). Therefore by using (2.6) the likelihood of $\{X_t\}_t$ is

$$L_n(\underline{X};\phi) = f(X_1;\phi)\prod_{t=2}^{n} f_\varepsilon(X_t - \phi X_{t-1}), \tag{2.7}$$

where f_ε is the density of ε and $f(X_1;\phi)$ is the marginal density of X_1. This means the likelihood of $\{X_t\}$ only depends on f_ε and the marginal density of X_t. We use $\widehat{\phi}_n = \arg\max L_n(\underline{X};\phi)$ as the mle estimator of a.

We now derive an explicit expression for the likelihood in the case that ε_t belongs to the exponential family. We focus on the case that $\{\varepsilon_t\}$ is Gaussian; since X_t is the sum of Gaussian random variables $X_t = \sum_{j=0}^{\infty}\phi^j\varepsilon_{t-j}$ (almost surely) X_t is also Gaussian. It can be shown that if $\varepsilon_t \sim \mathcal{N}(0,\sigma^2)$, then $X_t \sim \mathcal{N}(0,\sigma^2/(1-\phi^2))$. Thus the log likelihood

[5]If we start the recursion at some finite time point t_0 then the time series is random walk and is called a unit root process it is not stationary.

for Gaussian "innovations" is

$$\begin{aligned}\mathcal{L}_n(\phi,\sigma^2) &= -\frac{1}{2\sigma^2}\underbrace{\sum_{t=2}^{n} X_t^2}_{=\sum_{t=2}^{n-1} X_t^2 - X_n^2} + \frac{\phi}{\sigma^2}\sum_{t=2}^{n} X_t X_{t-1} - \frac{\phi^2}{2\sigma^2}\underbrace{\sum_{t=2}^{n} X_{t-1}^2}_{=\sum_{t=2}^{n-1} X_t^2 - X_1^2} - \frac{n-1}{2}\log\sigma^2 \\ &\quad -\frac{(1-\phi^2)}{2\sigma^2}X_1^2 - \frac{1}{2}\log\frac{\sigma^2}{1-\phi^2} \\ &= -\frac{1-\phi^2}{2\sigma^2}\sum_{t=1}^{n-1} X_t^2 + \frac{\phi}{\sigma^2}\sum_{t=2}^{n} X_t X_{t-1} - \frac{1}{2\sigma^2}(X_1^2+X_n^2) - \frac{n-1}{2}\log\sigma^2 - \frac{1}{2}\log\frac{\sigma^2}{1-\phi^2},\end{aligned}$$

see Efron (1975), Example 3. Using the factorisation theorem we see that the sufficient statistics, for this example are $\sum_{t=1}^{n-1} X_t^2$, $\sum_{t=2}^{n} X_t X_{t-1}$ and $(X_1^2 + X_n^2)$ (it almost has two sufficient statistics!). Since the data is dependent some caution needs to be applied before ones applies the results on the exponential family to dependent data (see Küchler and Sørensen (1997)). To estimate ϕ and σ^2 we maximise the above with respect to ϕ and σ^2. It is worth noting that the maximum can lie on the boundary -1 or 1.

Often we ignore the term the distribution of X_1 and consider the *conditional log-likelihood*, that is $X_2, \ldots, X_n$ conditioned on X_1. This gives the conditional log likelihood

$$\begin{aligned}Q_n(\phi,\sigma^2;X_1) &= \log\prod_{t=2}^{n} f_\varepsilon(X_t - \phi X_{t-1}) \\ &= -\frac{1}{2\sigma^2}\sum_{t=2}^{n} X_t^2 + \frac{\phi}{\sigma^2}\sum_{t=2}^{n} X_t X_{t-1} - \frac{\phi^2}{2\sigma^2}\sum_{t=2}^{n} X_{t-1}^2 - \frac{n-1}{2}\log\sigma^2, \quad (2.8)\end{aligned}$$

again there are three sufficient statistics. However, it is interesting to note that if the maximum of the likelihood lies within the parameter space $\phi \in [-1,1]$ then $\widehat{\phi}_n = \sum_{t=2}^{n} X_t X_{t-1} / \sum_{t=2}^{n} X_{t-1}^2$ (the usual least squares estimator).

2.5 Evaluating the maximum: Numerical Routines

In an ideal world an explicit closed form expression would exist for the maximum of a (log)-likelihood. In reality this rarely happens.

Usually, we have to use a numerical routine to maximise the likelihood. It is relative straightforward to maximise the likelihood of random variables which belong to the exponential family (since they typically have a negative definite Hessian). However, the story

becomes more complicated if the likelihood does not belong to the exponential family, for example mixtures of exponential family distributions.

Let us suppose that $\{X_i\}$ are iid random variables which follow the classical normal mixture distribution

$$f(y;\theta) = pf_1(y;\theta_1) + (1-p)f_2(y;\theta_2),$$

where f_1 is the density of the normal with mean μ_1 and variance σ_1^2 and f_2 is the density of the normal with mean μ_2 and variance σ_2^2. The log likelihood is

$$\mathcal{L}_n(\underline{Y};\theta) = \sum_{i=1}^{n} \log\left(p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{1}{2\sigma_1^2}(X_i-\mu_1)^2\right] + (1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left[-\frac{1}{2\sigma_2^2}(X_i-\mu_2)^2\right]\right).$$

Studying the above it is clear there does not explicit solution to the maximum. Hence one needs to use a numerical algorithm to maximise the above likelihood.

We discuss a few such methods below.

The Newton Raphson Routine The Newton-Raphson routine is the standard method to numerically maximise the likelihood, this can often be done automatically in `R` by using the `R` functions `optim` or `nlm`. To apply Newton-Raphson, we have to assume that the derivative of the likelihood exists (this is not always the case - think about the ℓ_1-norm based estimators!) and the maximum lies inside the parameter space such that $\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\widehat{\theta}_n} = 0$. We choose an initial value $\theta_n^{(1)}$ and apply the routine

$$\theta_n^{(j)} = \theta_n^{(j-1)} - \left(\frac{\partial^2\mathcal{L}_n(\theta)}{\partial\theta^2}\rfloor_{\theta_n^{(j-1)}}\right)^{-1}\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta_n^{(j-1)}}.$$

This routine can be derived from the Taylor expansion of $\frac{\partial\mathcal{L}_n(\theta_{n-1})}{\partial\theta}$ about θ_0 (see Section 2.6.3). A good description is given in `https://en.wikipedia.org/wiki/Newton%27s_method`. We recall that $-\frac{\partial^2\mathcal{L}_n(\theta)}{\partial\theta^2}\rfloor_{\theta_n^{(j-1)}}$ is the observed Fisher information matrix. If the algorithm does not converge, sometimes we replace $-\frac{\partial^2\mathcal{L}_n(\theta)}{\partial\theta^2}\rfloor_{\theta_n^{(j-1)}}$ with its expectation (the Fisher information matrix); since this is positive definite it may give better results (this is called Fisher scoring).

If the likelihood has just one global maximum and is concave, then it is quite easy to maximise. If on the other hand, the likelihood has a few local maximums and the initial value θ_1 is not chosen close enough to the true maximum, then the

routine may converge to a local maximum. In this case it may be a good idea to do the routine several times for several different initial values θ_n^*. For each candidate value $\widehat{\theta}_n^*$ evaluate the likelihood $\mathcal{L}_n(\widehat{\theta}_n^*)$ and select the value which gives the largest likelihood. It is best to avoid these problems by starting with an informed choice of initial value.

Implementing a Newton-Rapshon routine without much thought can lead to estimators which take an incredibly long time to converge. If one carefully considers the likelihood one can shorten the convergence time by rewriting the likelihood and using faster methods (often based on the Newton-Raphson).

Iterative least squares This is a method that we shall describe later when we consider Generalised linear models. As the name suggests the algorithm has to be interated, however at each step weighted least squares is implemented (see later in the course).

The EM-algorithm This is done by the introduction of dummy variables, which leads to a new 'unobserved' likelihood which can easily be maximised (see later in the course).

2.6 Statistical inference

2.6.1 A quick review of the central limit theorem

In this section we will not prove the central limit theorem. Instead we summarise the CLT and generalisations of it. The purpose of this section is not to lumber you with unnecessary mathematics but to help you understand when an estimator is close to normal (or not).

Lemma 2.6.1 (The famous CLT) *Let us suppose that $\{X_i\}_{i=1}^n$ are iid random variables, let $\mu = \mathrm{E}(X_i) < \infty$ and $\sigma^2 = \mathrm{var}(X_i) < \infty$. Define $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then we have*

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2).$$

Heuristically, we can write $(\bar{X} - \mu) \approx \mathcal{N}(0, \frac{\sigma^2}{n})$.

What this means that if we have a large enough sample size and made a quantile plot against the normal distribution the points should lie roughly on the $x = y$ line (though there will be less matching in the tails).

Remark 2.6.1 *(i) The above lemma appears to be 'restricted' to just averages. However, it can be used in several different contexts. Averages arise in several different situations. It is not just restricted to the average of the observations. By judicious algebraic manipulations, one can show that several estimators can be rewritten as an average (or approximately as an average). At first appearance, the MLE does not look like an average, however, in Section 2.6.3 we show that it can be approximated by a "useable" average.*

(ii) The CLT can be extended in several ways.

(a) To random variables whose variance are not all the same (ie. independent but identically distributed random variables).

(b) Dependent random variables (so long as the dependency 'decays' in some way).

(c) Weighted averages can also be asymptotically normal; so long as the weights are 'distributed evenly' over all the random variables.

* *Suppose that $\{X_i\}$ are iid non-normal random variables, $Y = \sum_{j=0}^{M} \phi^j X_j$ ($|\phi| < 1$) will never be normal (however large M).*
* *However, $Y = \frac{1}{n}\sum_{i=1}^{n} \sin(2\pi i/12) X_i$ is asymptotically normal.*

- *There exists several theorems which one can use to prove normality. But really the take home message is, look at your estimator and see whether asymptotic normality it looks plausible. Always check through simulations (even if asymptically it is normal, it may require a very large sample size for it to be close to normal).*

Example 2.6.1 (Some problem cases) *A necessary condition is that the second moment of X_i should exist. If it does not the CLT will not hold. For example if $\{X_i\}$ follow a t-distribution with 2 degrees of freedom*

$$f(x) = \frac{\Gamma(3/2)}{\sqrt{2\pi}} \left(1 + \frac{x^2}{2}\right)^{-3/2},$$

then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ will not have a normal limit.

We apply can immediately apply the above result to the MLE in the full rank exponential class.

2.6.2 Sampling properties and the full rank exponential family

In Section 2.3.1 we showed that if $\{X_i\}_{i=1}^n$ belonged to the exponential family and the maximum of the likelihood lay inside the parameter space (satisfied if the distribution is "steep") then

$$\widehat{\theta}_n = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right)$$

is the maximum likelihood estimator. Since we have an "explicit" expression for the estimator it is straightforward to derive the sampling properties of $\widehat{\theta}_n$. By using the law of large numbers

$$\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i) \stackrel{\text{a.s.}}{\to} \mathrm{E}[\mathbf{s}(X)] \qquad n \to \infty$$

then by the continuous mapping theorem

$$\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right) \stackrel{\text{a.s.}}{\to} \mu^{-1}\left(\mathrm{E}[\mathbf{s}(X)]\right) = \theta \qquad n \to \infty.$$

Thus the maximum likelihood estimator is a consistent estimator of θ. By using the CLT we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\mathbf{s}(X_i) - \mathrm{E}[\mathbf{s}(X_i)]\right) \stackrel{\mathcal{D}}{\to} \mathcal{N}\left(0, \mathrm{var}[\mathbf{s}(X_i)]\right) \qquad n \to \infty$$

where we recall that $\mathrm{var}[\mathbf{s}(X_i)] = \kappa''(\theta)$. Now by using that

$$\begin{aligned}
\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right) - \mu^{-1}\left(\mu(\theta)\right) &\approx \frac{\partial \mu^{-1}(x)}{\partial x}\rfloor_{x=\theta}\left(\frac{1}{n}\sum_{i=1}^n \left[\mathbf{s}(X_i) - \mathrm{E}(\mathbf{s}(X_i))\right]\right) \\
&= \left(\frac{\partial \mu(x)}{\partial x}\right)\rfloor_{x=\theta}^{-1}\left(\frac{1}{n}\sum_{i=1}^n \left[\mathbf{s}(X_i) - \mathrm{E}(\mathbf{s}(X_i))\right]\right).
\end{aligned}$$

Thus by using the above, the continuous mapping theorem and the CLT for averages we have

$$\begin{aligned}
\sqrt{n}\left[\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right) - \mu^{-1}\left(\mu(\theta)\right)\right] &\stackrel{\mathcal{D}}{\to} \mathcal{N}\left(0, \left(\frac{\partial \mu(x)}{\partial x}\right)\rfloor_{x=\theta}^{-1}\kappa''(\theta)\left(\frac{\partial \mu(x)}{\partial x}\right)\rfloor_{x=\theta}^{-1}\right) \\
&= \mathcal{N}\left(0, \kappa''(\theta)^{-1}\right).
\end{aligned}$$

We recall that $\kappa''(\theta)$ is the Fisher information of θ based on X_1.

Thus we have derived the sampling properties of the maximum likelihood estimator for the exponential class. It is relatively straightfoward to derive. Interesting we see that the limiting variance is the inverse of the Fisher information. So asymptotically the MLE estimator attains the Cramer-Rao lower bound (though it is not really a variance). However, the above derivation apply only to the full exponential class, in the following section we derive a similar result for the general MLE.

2.6.3 The Taylor series expansion

The Taylor series is used all over the place in statistics. It can be used to prove consistency of an estimator, normality (based on the assumption that averages converge to a normal distribution), obtaining the limiting variance of an estimator etc. We start by demonstrating its use for the log likelihood.

We recall that the mean value (in the univariate case) states that

$$f(x) = f(x_0) + (x - x_0)f'(\bar{x}_1) \text{ and } f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2}f''(\bar{x}_2),$$

where $\bar{x}_1 = \alpha x_0 + (1 - \alpha)x$ and $\bar{x}_2 = \beta x + (1 - \beta)x_0$ (for some $0 \leq \alpha, \beta \leq 1$). In the case that $f : \mathbb{R}^q \to \mathbb{R}$ we have

$$\begin{aligned} f(\underline{x}) &= f(\underline{x}_0) + (\underline{x} - \underline{x}_0)\nabla f(\underline{x})\rfloor_{\underline{x}=\underline{\bar{x}}_1} \\ f(\underline{x}) &= f(\underline{x}_0) + (\underline{x} - \underline{x}_0)'\nabla f(\underline{x})\rfloor_{\underline{x}=\underline{x}_0} + \frac{1}{2}(\underline{x} - \underline{x}_0)'\nabla^2 f(\underline{x})\rfloor_{\underline{x}=\underline{\bar{x}}_2}(\underline{x} - \underline{x}_0), \end{aligned}$$

where $\underline{\bar{x}}_1 = \alpha x_0 + (1 - \alpha)x$ and $\underline{\bar{x}}_2 = \beta x + (1 - \beta)x_0$ (for some $0 \leq \alpha, \beta \leq 1$). In the case that $f(\underline{x})$ is a vector, then the mean value theorem does not directly work, i.e. the following *is not true*

$$\underline{f}(\underline{x}) = \underline{f}(\underline{x}_0) + (\underline{x} - \underline{x}_0)'\nabla \underline{f}(\underline{x})\rfloor_{\underline{x}=\underline{\bar{x}}_1},$$

where $\underline{\bar{x}}_1$ lies between $\underline{x}$ and $\underline{x}_0$. However, it is quite straightforward to overcome this inconvience. The mean value theorem does hold pointwise, for every element of the vector $\underline{f}(\underline{x}) = (f_1(\underline{x}), \ldots, f_p(\underline{x}))$, ie. for every $1 \leq j \leq p$ we have

$$f_j(\underline{x}) = f_i(\underline{x}_0) + (\underline{x} - \underline{x}_0)\nabla f_j(\underline{y})\rfloor_{\underline{y}=\alpha\underline{x}+(1-\alpha)\underline{x}_0},$$

where $\bar{\underline{x}}_j$ lies between $\underline{x}$ and $\underline{x}_0$. Thus if $\nabla f_j(\underline{x})\rfloor_{\underline{x}=\bar{\underline{x}}_j} \rightarrow \nabla f_j(\underline{x})\rfloor_{\underline{x}=\underline{x}_0}$, we do have that

$$\underline{f}(\underline{x}) \approx \underline{f}(\underline{x}_0) + (\underline{x} - \underline{x}_0)'\nabla \underline{f}(\underline{x}).$$

We use the above below.

- Application 1: An expression for $\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0)$ in terms of $(\widehat{\theta}_n - \theta_0)$.

  The expansion of $\mathcal{L}_n(\widehat{\theta}_n)$ about θ_0 (the true parameter)

  $$\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0) = -\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\widehat{\theta}_n}(\widehat{\theta}_n - \theta_0) - \frac{1}{2}(\widehat{\theta}_n - \theta_0)'\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n}(\widehat{\theta}_n - \theta_0)$$

  where $\bar{\theta}_n = \alpha\theta_0 + (1-\alpha)\widehat{\theta}_n$. If $\widehat{\theta}_n$ lies in the interior of the parameter space (this is an extremely important assumption here) then $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\widehat{\theta}_n} = 0$. Moreover, if it can be shown that $|\widehat{\theta}_n - \theta_0| \xrightarrow{\mathcal{P}} 0$ and $n^{-1}\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}$ converges uniformly to $\mathrm{E}(n^{-1}\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\theta_0})$ (see Assumption 2.6.1(iv), below), then we have

  $$\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n} \xrightarrow{\mathcal{P}} \mathrm{E}\left(\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\theta_0}\right) = -I_n(\theta_0). \tag{2.9}$$

  This altogether gives

  $$2(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0)) \approx (\widehat{\theta}_n - \theta_0)'I_n(\theta_0)(\widehat{\theta}_n - \theta_0). \tag{2.10}$$

- Application 2: An expression for $(\widehat{\theta}_n - \theta_0)$ in terms of $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0}$

  The expansion of the p-dimension vector $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\widehat{\theta}_n}$ pointwise about θ_0 (the true parameter) gives (for $1 \leq j \leq d$)

  $$\frac{\partial \mathcal{L}_{j,n}(\theta)}{\partial \theta}\rfloor_{\widehat{\theta}_n} = \frac{\partial \mathcal{L}_{j,n}(\theta)}{\partial \theta}\rfloor_{\theta_0} + \frac{\partial^2 \mathcal{L}_{j,n}(\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_{j,n}}(\widehat{\theta}_n - \theta_0),$$

  where $\bar{\theta}_{j,n} = \alpha_j\bar{\theta}_{j,n} + (1-\alpha_j)\theta_0$. Using the same arguments as in Application 1 and equation (2.9) we have

  $$\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0} \approx I_n(\theta_0)(\widehat{\theta}_n - \theta_0).$$

  We mention that $U_n(\theta_0) = \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0}$ is often called the *score or U statistic*. And we see that the asymptotic sampling properties of U_n determine the sampling properties of $(\widehat{\theta}_n - \theta_0)$.

Remark 2.6.2 *(i) In practice $I_n(\theta_0)$ is unknown and it is approximated by the Hessian evaluated at the estimated parameter $\widehat{\theta}_n$, $-\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\widehat{\theta}_n}$. A discussion on the quality of this approximation is given in Efron and Hinkley (1978).*

(ii) Bear in mind that $\nabla^2_\theta \mathcal{L}_n(\theta)$ is not necessarily negative definite, but its limit is the negative Fisher information matrix $-I_n(\theta)$ (non-negative definite over $\theta \in \Theta$). Therefore for "large n $\nabla^2_\theta \mathcal{L}_n(\theta)$ will be negative definite".

(iii) The quality of the approximation (2.9) depends on the the second order efficiency measure $I_n(\widehat{\theta}_n) - I_n(\theta_0)$ (this term was coined by C.R.Rao and discussed in Rao (1961, 1962, 1963)). Efron (1975), equation (1.1) shows this difference depends on the so called curvature of the parameter space.

Example 2.6.2 (The Weibull) *Evaluate the second derivative of the likelihood given in Example 2.2.3, take the expection on this, $I_n(\theta, \alpha) = \mathrm{E}(\nabla^2 \mathcal{L}_n)$ (we use the ∇ to denote the second derivative with respect to the parameters α and θ).*

Application 2 implies that the maximum likelihood estimators $\widehat{\theta}_n$ and $\widehat{\alpha}_n$ (recalling that no explicit expression for them exists) can be written as

$$\begin{pmatrix} \widehat{\theta}_n - \theta \\ \widehat{\alpha}_n - \alpha \end{pmatrix} \approx I_n(\theta, \alpha)^{-1} \begin{pmatrix} \sum_{i=1}^n \left(-\frac{\alpha}{\theta} + \frac{\alpha}{\theta^{\alpha+1}} Y_i^\alpha \right) \\ \sum_{i=1}^n \left(\frac{1}{\alpha} - \log Y_i - \log \theta - \frac{\alpha}{\theta} + \log(\frac{Y_i}{\theta}) \times (\frac{Y_i}{\theta})^\alpha \right) \end{pmatrix}$$

2.6.4 Sampling properties of the maximum likelihood estimator

We have shown that under certain conditions the maximum likelihood estimator can often be the minimum variance unbiased estimator (for example, in the case of the normal distribution). However, in most situations for finite samples the mle may not attain the Cramer-Rao lower bound. Hence for finite sample $\mathrm{var}(\widehat{\theta}_n) > I_n(\theta)^{-1}$. However, it can be shown that asymptotically the "variance" (it is not the true variance) of the mle attains the Cramer-Rao lower bound. In other words, for large samples, the "variance" of the mle is close to the Cramer-Rao bound. We will prove the result in the case that $\mathcal{L}_n$ is the log likelihood of independent, identically distributed random variables. The proof can be generalised to the case of non-identically distributed random variables.

We first state sufficient conditions for this to be true.

Assumption 2.6.1 *Suppose $\{X_i\}$ be iid random variables with density $f(X;\theta)$.*

(i) The conditions in Assumption 1.3.1 hold. In particular:

(a)

$$\mathrm{E}_{\theta_0}\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta=\theta_0}\right) = \int \frac{\partial f(x;\theta)}{\partial\theta}\rfloor_{\theta=\theta_0}dx = \frac{\partial}{\partial\theta}\int \frac{\partial f(x;\theta)}{\partial\theta}dx\rfloor_{\theta=\theta_0} = 0.$$

(b)

$$\mathrm{E}_{\theta_0}\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta=\theta_0}\right)^2\right] = \mathrm{E}_{\theta_0}\left[-\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\rfloor_{\theta=\theta_0}\right].$$

(ii) Almost sure uniform convergence of the likelihood:

$\sup_{\theta\in\Theta}\frac{1}{n}|\mathcal{L}_n(\underline{X};\theta) - \mathrm{E}(\mathcal{L}_n(\underline{X};\theta))| \xrightarrow{a.s.} 0$ *as* $n\to\infty$.

We mention that directly verifying uniform convergence can be difficult. However, it can be established by showing that the parameter space is compact, point wise convergence of the likelihood to its expectation and almost sure equicontinuity in probability.

(iii) Model identifiability:

For every $\theta\in\Theta$, there does not exist another $\tilde{\theta}\in\Theta$ such that $f(x;\theta) = f(x;\tilde{\theta})$ for all x in the sample space.

(iv) Almost sure uniform convergence of the second derivative of the likelihood (using the notation ∇_θ): $\sup_{\theta\in\Theta}\frac{1}{n}|\nabla^2_\theta\mathcal{L}_n(\underline{X};\theta) - \mathrm{E}(\nabla^2_\theta\mathcal{L}_n(\underline{X};\theta))| \xrightarrow{a.s.} 0$ *as* $n\to\infty$.

This can be verified by using the same method described in (ii).

We require Assumption 2.6.1(ii,iii) to show consistency and Assumptions 1.3.1 and 2.6.1(iii-iv) to show asymptotic normality.

Theorem 2.6.1 *Supppose Assumption 2.6.1(ii,iii) holds. Let θ_0 be the true parameter and $\widehat{\theta}_n$ be the mle. Then we have $\widehat{\theta}_n \xrightarrow{a.s.} \theta_0$ (consistency).*

PROOF. First define $\ell(\theta) = \mathrm{E}[\log f(X;\theta)]$ (the limit of the expected log-likelihood). To prove the result we first need to show that the expectation of the maximum likelihood is

maximum at the true parameter and that this is the unique maximum. In other words we need to show that $\mathrm{E}(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta) - \frac{1}{n}\mathcal{L}_n(\underline{X};\theta_0)) \leq 0$ for all $\theta \in \Theta$. To do this, we have

$$\begin{aligned}\ell(\theta) - \ell(\theta_0) = \mathrm{E}\left(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta) - \mathrm{E}(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta_0)\right) &= \int \log \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)dx \\ &= \mathrm{E}\left(\log \frac{f(X;\theta)}{f(X;\theta_0)}\right).\end{aligned}$$

Now by using Jensen's inequality (since log is a concave function) we have

$$\mathrm{E}\left(\log \frac{f(X;\theta)}{f(X;\theta_0)}\right) \leq \log \mathrm{E}\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right) = \log \int \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)dx = \log \int f(x;\theta)dx = 0,$$

since $\theta \in \Theta$ and $\int f(x;\theta)dx = 1$. Thus giving $\mathrm{E}(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta)) - \mathrm{E}(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta_0)) \leq 0$. To prove that $\mathrm{E}(\frac{1}{n}[\mathcal{L}_n(\underline{X};\theta)) - \mathcal{L}_n(\underline{X};\theta_0)]) = 0$ only when θ_0, we use the identifiability condition in Assumption 2.6.1(iii), which means that $f(x;\theta) = f(x;\theta_0)$ for all x *only* when θ_0 and no other function of f gives equality. Hence only when $\theta = \theta_0$ do we have

$$\mathrm{E}\left(\log \frac{f(X;\theta)}{f(X;\theta_0)}\right) = \log \int \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)dx = \log \int f(x;\theta)dx = 0,$$

thus $\mathrm{E}(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta))$ has a unique maximum at θ_0.

Finally, we need to show that $\widehat{\theta}_n \stackrel{\text{a.s.}}{\to} \theta_0$. To simplify notation for the remainder of this proof we assume the likelihood has been standardized by n i.e.

$$\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta).$$

We note that since $\ell(\theta)$ is maximum at θ_0 if $|\mathcal{L}_n(\widehat{\theta}_n) - \ell(\theta_0)| \stackrel{\text{a.s.}}{\to} 0$, then $\widehat{\theta}_n \stackrel{\text{a.s.}}{\to} \theta_0$. Thus we need only prove $|\mathcal{L}_n(\widehat{\theta}_n) - \ell(\theta_0)| \stackrel{\text{a.s.}}{\to} 0$. We do this using a sandwich argument.

First we note for every mle $\widehat{\theta}_n$

$$\mathcal{L}_n(\underline{X};\theta_0) \leq \mathcal{L}_n(\underline{X};\widehat{\theta}_n) \stackrel{\text{a.s.}}{\to} \ell(\widehat{\theta}_n) \leq \ell(\theta_0), \tag{2.11}$$

where we are treating $\widehat{\theta}_n$ as if it were a non-random fixed value in Θ. Returning to $|\mathrm{E}(\mathcal{L}_n(\underline{X};\theta_0)) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)|$ (they swapped round) we note that the difference can be written as

$$\ell(\theta_0) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n) = \{\ell(\theta_0) - \mathcal{L}_n(\underline{X};\theta_0)\} + \left\{\ell(\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)\right\} + \left\{\mathcal{L}_n(\underline{X};\theta_0) - \ell(\widehat{\theta}_n)\right\}.$$

Now by using (2.11) we have

$$\begin{aligned}\ell(\theta_0) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n) &\leq \{\ell(\theta_0) - \mathcal{L}_n(\underline{X};\theta_0)\} + \Big\{\ell(\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)\Big\} + \left\{\underbrace{\mathcal{L}_n(\underline{X};\widehat{\theta}_n)}_{\geq \mathcal{L}_n(\underline{X};\theta_0)} - \ell(\widehat{\theta}_n)\right\} \\ &= \{\ell(\theta_0) - \mathcal{L}_n(\underline{X};\theta_0)\}\end{aligned}$$

and

$$\begin{aligned}\ell(\theta_0) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n) &\geq \{\ell(\theta_0) - \mathcal{L}_n(\underline{X};\theta_0)\} + \Big\{\ell(\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)\Big\} + \left\{\mathcal{L}_n(\underline{X};\theta_0) - \underbrace{\ell(\theta_0)}_{\geq \ell(\widehat{\theta}_n)}\right\} \\ &= \Big\{\ell(\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)\Big\}\end{aligned}$$

Thus

$$\Big\{\ell(\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)\Big\} \leq \ell(\theta_0) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n) \leq \{\ell(\theta_0) - \mathcal{L}_n(\underline{X};\theta_0)\}.$$

The above also immediately follows from (2.11). This is easily seen in Figure 2.3, which Reza suggested. Thus we have sandwiched the difference $\mathrm{E}(\mathcal{L}_n(\underline{X};\theta_0)) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)$. Therefore, under Assumption 2.6.1(ii) we have

$$|\ell(\theta_0) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)| \leq \sup_{\theta\in\Theta} |\ell(\theta) - \mathcal{L}_n(\underline{X};\theta)| \stackrel{\text{a.s.}}{\to} 0.$$

Since $\mathrm{E}[\mathcal{L}_n(\underline{X};\theta)]$ has a unique maximum at $\mathrm{E}[\mathcal{L}_n(\underline{X};\theta_0)]$ this implies $\widehat{\theta}_n \stackrel{\text{a.s.}}{\to} \theta_0$.

□

Hence we have shown consistency of the mle. It is important to note that this proof is not confined to just the likelihood it can also be applied to other contrast functions. We now show asymptotic normality of the MLE.

Theorem 2.6.2 *Suppose Assumption 2.6.1 is satisfied (where θ_0 is the true parameter). Let*

$$I(\theta_0) = \mathrm{E}\left(\left[\frac{\partial \log f(X_i;\theta)}{\partial\theta}\rfloor_{\theta_0}\right]^2\right) = \mathrm{E}\left(-\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}\rfloor_{\theta_0}\right).$$

(i) Then the score statistic is

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial\theta}\rfloor_{\theta_0} \stackrel{\mathcal{D}}{\to} \mathcal{N}\Big(0, I(\theta_0)\Big). \tag{2.12}$$

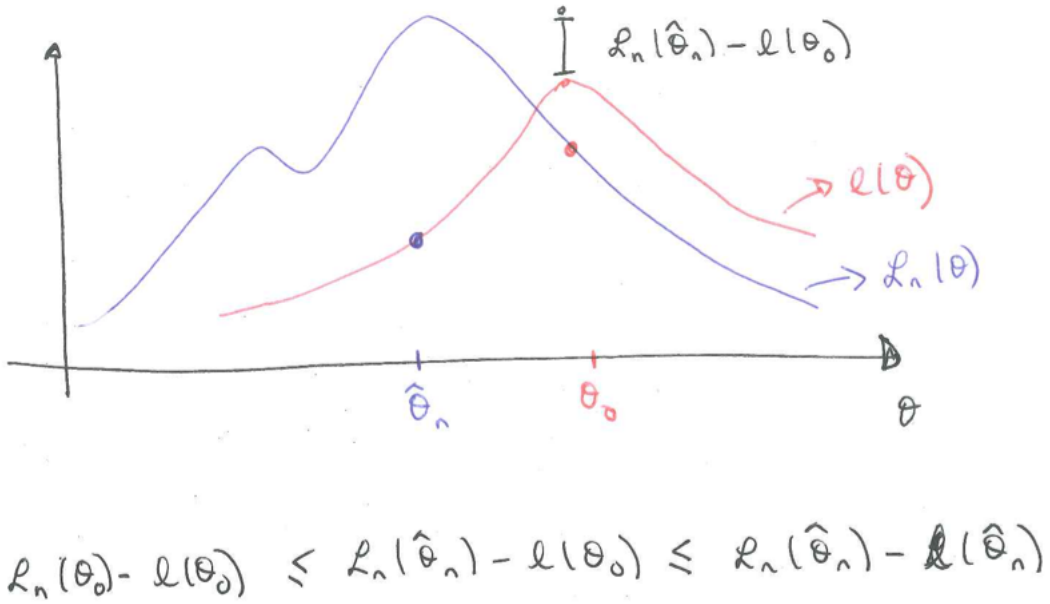

Figure 2.3: Difference between likelihood and expectation.

(ii) Then the mle is

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, I(\theta_0)^{-1}\Big).$$

(iii) The log likelihood ratio is

$$2\Big(\mathcal{L}_n(\underline{X}; \widehat{\theta}_n) - \mathcal{L}_n(\underline{X}; \theta_0)\Big) \xrightarrow{\mathcal{D}} \chi^2_p$$

(iv) The square MLE

$$n(\widehat{\theta}_n - \theta_0)' I_n(\theta_0)(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \chi^2_p.$$

PROOF. First we will prove (i). We recall because $\{X_i\}$ are iid random variables, then

$$\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\underline{X}; \theta)}{\partial \theta} \rfloor_{\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i; \theta)}{\partial \theta} \rfloor_{\theta_0},$$

is the sum of independent random variables. We note that under Assumption 2.6.1(i) we have

$$\mathrm{E}\left(\frac{\partial \log f(X_i; \theta)}{\partial \theta} \rfloor_{\theta_0}\right) = \int \frac{\partial \log f(x; \theta)}{\partial \theta} \rfloor_{\theta_0} f(x; \theta_0) dx = 0,$$

thus $\frac{\partial \log f(X_i;\theta)}{\partial \theta}\rfloor_{\theta_0}$ is a zero mean random variable and its variance is $I(\theta_0)$.

Hence $\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta_0}$ is the sum of iid random variables with mean zero and variance $I(\theta_0)$. Therefore, by the CLT for iid random variables we have (2.12).

We use (i) and Taylor (mean value) theorem to prove (ii). We first note that by the mean value theorem we have

$$\underbrace{\frac{1}{n}\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\widehat{\theta}_n}}_{=0} = \frac{1}{n}\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta_0} + (\widehat{\theta}_n - \theta_0)\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\underline{X};\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n}. \tag{2.13}$$

Using the consistency result in Theorem 2.6.1 ($\widehat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$, thus $\bar{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$) and Assumption 2.6.1(iv) we have

$$\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\underline{X};\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n} \xrightarrow{\text{a.s.}} \frac{1}{n}\mathrm{E}\left(\frac{\partial^2 \mathcal{L}_n(\underline{X};\theta)}{\partial \theta^2}\rfloor_{\theta_0}\right) = \mathrm{E}\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\rfloor_{\theta_0}\right) = -I(\theta_0). \tag{2.14}$$

Substituting the above in (2.15) we have

$$\frac{1}{n}\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta_0} - I(\theta_0)(\widehat{\theta}_n - \theta_0) + \underbrace{\left(\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\underline{X};\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n} - I(\theta_0)\right)(\widehat{\theta}_n - \theta_0)}_{\text{small}} = 0 \tag{2.15}$$

Multiplying the above by $\sqrt{n}$ and rearranging gives

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \;=\; I(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta_0} + o_p(1).$$

[6] Hence by substituting the (2.12) into the above we have (ii).

To prove (iii) we use (2.10), which we recall is

$$2\left(\mathcal{L}_n(\underline{X};\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\theta_0)\right) \approx (\widehat{\theta}_n - \theta_0)' n I(\theta_0)(\widehat{\theta}_n - \theta_0)'.$$

Now by using that $\sqrt{n}I(\theta_0)^{-1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I)$ (see (i)) and substituting this into the above gives (iii).

The proof of (iv) follows immediately from (ii). □

This result tells us that asymptotically the mle attains the Cramer-Rao bound. Furthermore, if $\widehat{\theta}$ is a p-dimension random vector and $I(\theta_0)$ is diagonal, then the elements of

[6] We mention that the proof above is for univariate $\frac{\partial^2 \mathcal{L}_n(\underline{X};\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n}$, but by redo-ing the above steps pointwise it can easily be generalised to the multivariate case too

$\widehat{\theta}$ will be asymptotically independent (for example the sample mean and sample variance estimator for the normal distribution). However if $I(\theta_0)$ is not diagonal, then off-diagonal elements in $I(\theta_0)^{-1}$ measure the degree of correlation between the estimators. See Figure 2.4.

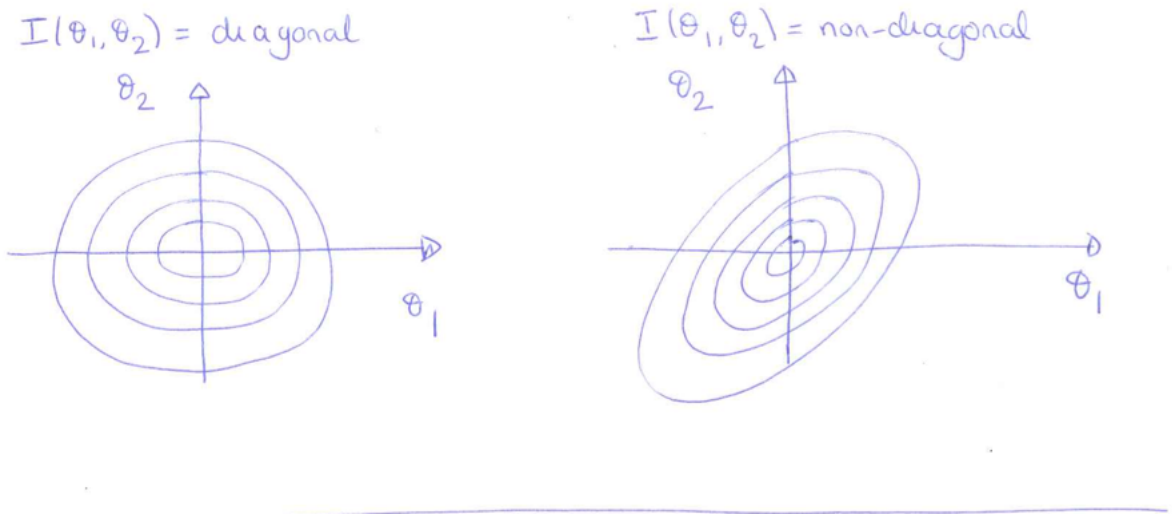

Figure 2.4: Contour plot of two dimensional normal distribution of two parameter estimators with diagonal and non-diagonal information matrix

Example 2.6.3 (The Weibull) *By using Example 2.6.2 we have*

$$\begin{pmatrix} \widehat{\theta}_n - \theta \\ \widehat{\alpha}_n - \alpha \end{pmatrix} \approx I_n(\theta, \alpha)^{-1} \begin{pmatrix} \sum_{i=1}^n \left(-\frac{\alpha}{\theta} + \frac{\alpha}{\theta^{\alpha+1}} Y_i^{\alpha} \right) \\ \sum_{i=1}^n \left(\frac{1}{\alpha} - \log Y_i - \log \theta - \frac{\alpha}{\theta} + \log(\frac{Y_i}{\theta}) \times (\frac{Y_i}{\theta})^{\alpha} \right) \end{pmatrix}.$$

Now we observe that RHS consists of a sum iid random variables (this can be viewed as an average). Since the variance of this exists (you can show that it is $I_n(\theta, \alpha)$), the CLT can be applied and we have that

$$\sqrt{n} \begin{pmatrix} \widehat{\theta}_n - \theta \\ \widehat{\alpha}_n - \alpha \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, I(\theta, \alpha)^{-1}\right),$$

where $I(\theta, \alpha) = \mathrm{E}[(\nabla \log f(X; \theta, \alpha))^2]$.

Remark 2.6.3 *(i) We recall that for iid random variables that the Fisher information for sample size n is*

$$I_n(\theta_0) = \mathrm{E}\Big\{ \frac{\partial \log L_n(X;\theta)}{\partial \theta}\rfloor_{\theta_0} \Big\}^2 = n\mathrm{E}\Big(\frac{\partial \log f(X;\theta)}{\partial \theta}\rfloor_{\theta_0} \Big)^2 = nI(\theta_0)$$

Therefore since

$$\begin{aligned} (\widehat{\theta}_n - \theta_0) &\approx I_n(\theta_0)^{-1} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0} = \Big[\frac{1}{n} I(\theta_0)\Big]^{-1} \frac{1}{n} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0} \\ \Rightarrow \sqrt{n}(\widehat{\theta}_n - \theta_0) &\approx \sqrt{n} I_n(\theta_0)^{-1} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0} = \Big[\frac{1}{n} I(\theta_0)\Big]^{-1} \frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0} \end{aligned}$$

and $\mathrm{var}\big(\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0}\big) = n^{-1}\mathrm{E}\Big[\big(\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0}\big)^2\Big] = I(\theta_0)$*, it can be seen that* $|\widehat{\theta}_n - \theta_0| = O_p(n^{-1/2})$*.*

(ii) Under suitable conditions a similar result holds true for data which is not iid.

(iii) These results only apply when θ_0 *lies* **inside** *the parameter space* Θ*.*

We have shown that under certain regularity conditions the mle will asymptotically attain the Fisher information bound. It is reasonable to ask how one can interprete this bound.

(i) Situation 1. $I_n(\theta_0) = \mathrm{E}\Big(- \frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\theta_0} \Big)$ is large (hence variance of the mle will be small) then it means that the gradient of $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}$ is large. Hence even for small deviations from θ_0, $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}$ is likely to be far from zero. This means the mle $\widehat{\theta}_n$ is likely to be in a close neighbourhood of θ_0.

(ii) Situation 2. $I_n(\theta_0) = \mathrm{E}\Big(-\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\theta_0} \Big)$ is small (hence variance of the mle will large). In this case the gradient of the likelihood $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}$ is flatter and hence $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \approx 0$ for a large neighbourhood about the true parameter θ. Therefore the mle $\widehat{\theta}_n$ can lie in a large neighbourhood of θ_0.

Remark 2.6.4 (Lagrange Multipliers) *Often when maximising the likelihood it has to be done under certain constraints on the parameters. This is often achieved with the use of Lagrange multipliers (a dummy variable), which enforces this constraint.*

For example suppose the parameters in the likelihood must sum to one then we can enforce this constraint by maximising the criterion

$$\mathcal{L}_n(\theta, \lambda) = \underbrace{\mathcal{L}_n(\theta)}_{likelihood} + \lambda[\sum_{j=1}^{q} \theta_j - 1]$$

with respect to θ and *the dummy variable* λ.

2.7 Some questions

Exercise 2.7 *Suppose* $X_1, \ldots, X_n$ *are i.i.d. observations. A student wants to test whether each* X_i *has a distribution in the parametric family* $\{f(x;\alpha) : \alpha \in \Theta\}$ *or the family* $\{g(x;\beta) : \beta \in \Gamma\}$. *To do this he sets up the hypotheses*

$$H_0 : X_i \sim f(\cdot\,;\alpha_0) \quad vs. \quad H_A : X_i \sim g(\cdot\,;\beta_0),$$

where α_0 *and* β_0 *are the unknown true parameter values. He constructs the log-likelihood ratio statistic*

$$L = \max_{\beta\in\Gamma} \mathcal{L}_g(\boldsymbol{X};\beta) - \max_{\alpha\in\Theta} \mathcal{L}_f(\boldsymbol{X};\alpha) = \mathcal{L}_g(\boldsymbol{X};\hat{\beta}) - \mathcal{L}_f(\boldsymbol{X};\hat{\alpha}),$$

where

$$\mathcal{L}_g(\boldsymbol{X};\beta) = \sum_{i=1}^{n} \log g(X_i;\beta), \quad \mathcal{L}_f(\boldsymbol{X};\alpha) = \sum_{i=1}^{n} \log f(X_i;\alpha),$$

$\hat{\alpha} = \arg\max_{\alpha\in\Theta} \mathcal{L}_f(\boldsymbol{X};\alpha)$ *and* $\hat{\beta} = \arg\max_{\beta\in\Gamma} \mathcal{L}_g(\boldsymbol{X};\beta)$. *The student applies what he believe he learned in class to* L *and assumes that the distribution of* L *under the null hypothesis (asymptotically) follows a chi-squared distribution with one-degree of freedom. He does the test at the 5% level using the critical value* $\chi^2 = 3.84$, *rejecting the null in favor of the alternative if* $L > 3.84$.

(a) Using well known results, derive the asymptotic distribution of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \log \frac{g(X_i;\beta_0)}{f(X_i;\alpha_0)}$$

under the null and the alternative.

(b) *Is the distribution of L chi-squared? If not, derive the asymptotic distribution of L.*

Hint: You will need to use your answer from (a).

Note: This part is tough; but fun (do not be disillusioned if it takes time to solve).

(c) *By using your solution to parts (a) and (b), carefully explain what the actual type error I of the student's test will be (you do not need to derive the Type I error, but you should explain how it compares to the 5% level that the student uses).*

(d) *By using your solution to parts (a) and (b), carefully explain what the power of his test will be (you do not have to derive an equation for the power, but you should explain what happens to the power as the sample size grows, giving a precise justification).*

(e) *Run some simulations to illustrate the above.*

Exercise 2.8 *Find applications where likelihoods are maximised with the use of Lagrange multipliers. Describe the model and where the Lagrange multiplier is used.*

2.8 Applications of the log-likelihood theory

We first summarise the results in the previous section (which will be useful in this section). For convenience, we will assume that $\{X_i\}_{i=1}^n$ are iid random variables, whose density is $f(x;\theta_0)$ (though it is relatively simple to see how this can be generalised to general likelihoods - of not necessarily iid rvs). Let us suppose that θ_0 is the true parameter that we wish to estimate. Based on Theorem 2.6.2 we have

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0, I(\theta_0)^{-1}\bigg), \tag{2.16}$$

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n}{\partial \theta}\rfloor_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0, I(\theta_0)\bigg) \tag{2.17}$$

and

$$2\left[\mathcal{L}_n(\hat{\theta}_n) - \mathcal{L}_n(\theta_0)\right] \xrightarrow{\mathcal{D}} \chi^2_p, \tag{2.18}$$

where p are the number of parameters in the vector θ and $I(\theta_0) = \mathrm{E}[(\frac{\partial \log f(X;\theta)}{\partial \theta}\rfloor_{\theta_0})^2] = n^{-1}\mathrm{E}[(\frac{\partial \log \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0})^2]$. It is worth keeping in mind that by using the usual Taylor expansion the log-likelihood ratio statistic is asymptotically equivalent to

$$2\Big[\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}(\theta_0)\Big] \stackrel{\mathcal{D}}{=} ZI(\theta_0)Z,$$

where $Z \sim \mathcal{N}(0, I(\theta_0))$.

Note: There are situations where the finite sampling distributions of the above are known, in which case there is no need to resort to the asymptotic sampling properties.

2.8.1 Constructing confidence sets using the likelihood

One the of main reasons that we show asymptotic normality of an estimator (it is usually not possible to derive normality for finite samples) is to construct confidence intervals/sets and to test.

In the case that θ_0 is a scaler (vector of dimension one), it is easy to use (2.16) to obtain

$$\sqrt{n}I(\theta_0)^{1/2}\big(\hat{\theta}_n - \theta_0\big) \stackrel{\mathcal{D}}{\to} N(0,1). \tag{2.19}$$

Based on the above the 95% CI for θ_0 is

$$\left[\widehat{\theta}_n - \frac{1}{\sqrt{n}}I(\theta_0)z_{\alpha/2}, \hat{\theta}_n + \frac{1}{\sqrt{n}}I(\theta_0)z_{\alpha/2}\right].$$

The above, of course, requires an estimate of the (standardised) expected Fisher information $I(\theta_0)$, typically we use the (standardised) observed Fisher information evaluated at the estimated value $\widehat{\theta}_n$.

The CI constructed above works well if θ is a scalar. But beyond dimension one, constructing a CI based on (2.16) (and the p-dimensional normal) is extremely difficult. More precisely, if θ_0 is a p-dimensional vector then the analogous version of (2.19) is

$$\sqrt{n}I(\theta_0)^{1/2}\big(\hat{\theta}_n - \theta_0\big) \stackrel{\mathcal{D}}{\to} N(0, I_p).$$

However, this does not lead to a simple set construction. One way to construct the confidence interval (or set) is to 'square' $\big(\hat{\theta}_n - \theta_0\big)$ and use

$$n\big(\hat{\theta}_n - \theta_0\big)'I(\theta_0)\big(\hat{\theta}_n - \theta_0\big) \stackrel{\mathcal{D}}{\to} \chi^2_p. \tag{2.20}$$

Based on the above a 95% CI is

$$\Big\{\theta; \big(\hat{\theta}_n - \theta\big)' n\mathrm{E}\Big(\frac{\partial \log f(X;\theta)}{\partial \theta}\rfloor_{\theta_0}\Big)^2 \big(\hat{\theta}_n - \theta\big) \leq \chi^2_p(0.95)\Big\}. \tag{2.21}$$

Note that as in the scalar case, this leads to the interval with the smallest length. A disadvantage of (2.21) is that we have to (a) estimate the information matrix and (b) try to find all θ such the above holds. This can be quite unwieldy. An alternative method, which is asymptotically equivalent to the above but removes the need to estimate the information matrix is to use (2.18). By using (2.18), a $100(1-\alpha)\%$ confidence set for θ_0 is

$$\Big\{\theta; 2\big(\mathcal{L}_n(\hat{\theta}_n) - \mathcal{L}_n(\theta)\big) \leq \chi^2_p(1-\alpha)\Big\}. \tag{2.22}$$

The above is not easy to calculate, but it is feasible.

Example 2.8.1 *In the case that θ_0 is a scalar the 95% CI based on (2.22) is*

$$\Big\{\theta; \mathcal{L}_n(\theta) \geq \mathcal{L}_n(\hat{\theta}_n) - \frac{1}{2}\chi^2_1(0.95)\Big\}.$$

See Figure 2.5 which gives the plot for the confidence interval (joint and disjoint).

Both the 95% confidence sets in (2.21) and (2.22) will be very close for relatively large sample sizes. However one advantage of using (2.22) instead of (2.21) is that it is easier to evaluate - no need to obtain the second derivative of the likelihood etc.

A feature which differentiates (2.21) and (2.22) is that the confidence sets based on (2.21) is symmetric about $\widehat{\theta}_n$ (recall that $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$) is symmetric about $\bar{X}$, whereas the symmetry condition may not hold for sample sizes when constructing a CI for θ_0 using (2.22). Using (2.22) there is no guarantee the confidence sets consist of only one interval (see Figure 2.5). However, if the distribution is exponential with full rank (and is steep) the likelihood will be concave with the maximum in the interior of the parameter space. This will mean the CI constructed using (2.22) will be connected.

If the dimension of θ is large it is difficult to evaluate the confidence set. Indeed for dimensions greater than three it is extremely hard. However in most cases, we are only interested in constructing confidence sets for certain parameters of interest, the other unknown parameters are simply nuisance parameters and confidence sets for them are not of interest. For example, for the normal family of distribution we may only be interested in constructing an interval for the mean, and the variance is simply a nuisance parameter.

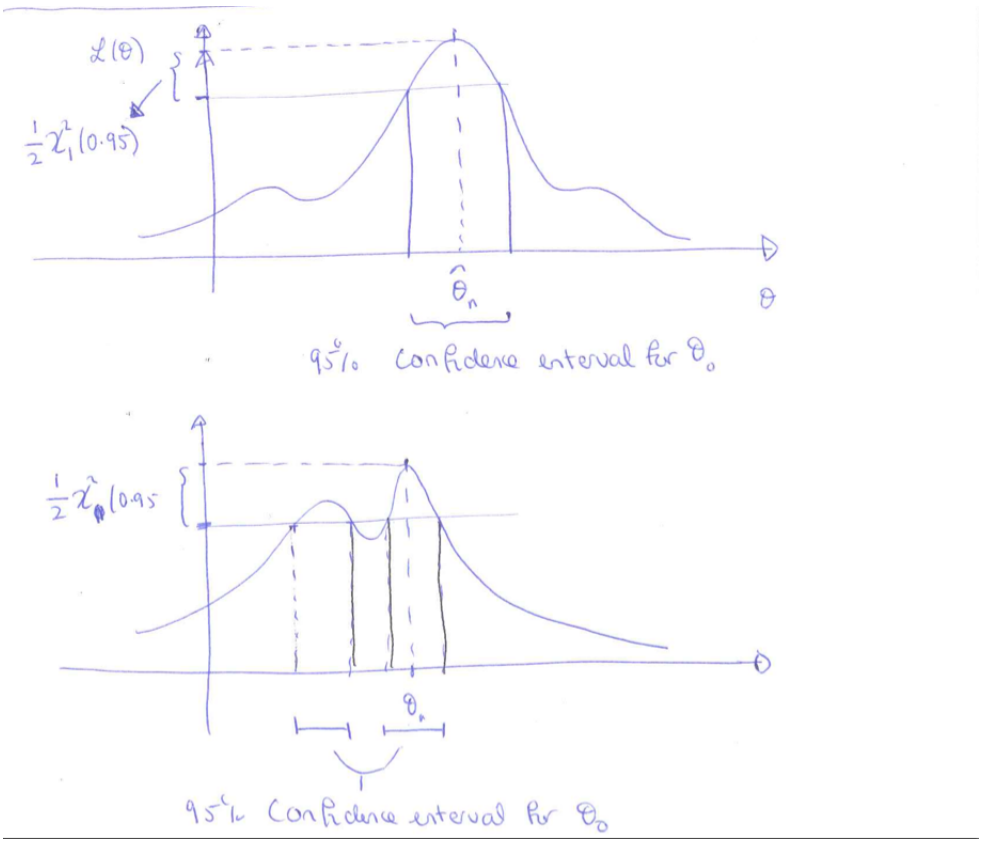

Figure 2.5: Constructing confidence intervals using method (2.22).

2.8.2 Testing using the likelihood

Let us suppose we wish to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_A : \theta \neq \theta_0$. We can use any of the results in (2.16), (2.17) and (2.18) to do the test - they will lead to slightly different p-values, but 'asympototically' they are all equivalent, because they are all based (essentially) on the same derivation.

We now list the three tests that one can use.

The Wald test

The Wald statistic is based on (2.16). We recall from (2.16) that if the null is true, then we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \left\{\mathrm{E}\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\rfloor_{\theta_0}\right)^2\right\}^{-1}\right).$$

Thus we can use as the test statistic

$$T_1 = \sqrt{n} I(\theta_0)^{1/2} \big(\widehat{\theta}_n - \theta_0\big)$$

to test the hypothesis. Under the null

$$T_1 \stackrel{\mathcal{D}}{\to} \mathcal{N}(0,1).$$

We now consider how the test statistics behaves under the alternative $H_A : \theta = \theta_1$. If the alternative were true, then we have

$$\begin{aligned} I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_0) &= I(\theta_0)^{1/2} \Big((\widehat{\theta}_n - \theta_1) + (\theta_1 - \theta_0) \Big) \\ &\approx I(\theta_0)^{1/2} I_n(\theta_1)^{-1} \sum_i \frac{\partial \log f(X_i;\theta_1)}{\partial \theta_1} + I(\theta_0)^{1/2}(\theta_1 - \theta_0) \end{aligned}$$

where $I_n(\theta_1) = \mathrm{E}_{\theta_1}[(\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\theta=\theta_1})^2]$.

Local alternatives and the power function

In the case that the alternative is *fixed* (does not change with sample size), it is clear that the power in the test goes to 100% as $n \to \infty$. To see we write

$$\begin{aligned} \sqrt{n} I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_0) &= \sqrt{n} I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_1) + \sqrt{n} I(\theta_0)^{1/2}(\theta_1 - \theta_0) \\ &\approx I(\theta_0)^{1/2} I(\theta_1)^{-1} \frac{1}{\sqrt{n}} \sum_i \frac{\partial \log f(X_i;\theta_1)}{\partial \theta_1} + \sqrt{n} I(\theta_0)^{1/2}(\theta_1 - \theta_0) \\ &\stackrel{\mathcal{D}}{\to} N\left(0, I(\theta_0)^{1/2} I(\theta_1)^{-1} I(\theta_0)^{1/2}\right) + \sqrt{n} I(\theta_0)^{1/2}(\theta_1 - \theta_0). \end{aligned}$$

Using the above calculation we see that

$$P\left(\text{Reject} | \theta = \theta_1\right) = 1 - \Phi\left(\frac{z_{1-\alpha/2} - \sqrt{n}(\theta_1 - \theta_0) I(\theta_0)^{1/2}}{\sqrt{I(\theta_0)^{1/2} I(\theta_1)^{-1} I(\theta_0)^{1/2}}} \right).$$

Thus, we see that as $n \to \infty$, the power gets closer to 100%. However, this calculation does not really tell us how the test performs for θ_1 close to the θ_0.

To check the effectiveness of a given testing method, one lets the alternative get *closer* to the the null as $n \to \infty$. This allows us to directly different statistical tests (and the factors which drive the power).

How to choose the closeness:

- Suppose that $\theta_1 = \theta_0 + \frac{\phi}{n}$ (for fixed ϕ), then the center of T_1 is

$$\begin{aligned}
\sqrt{n}I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_0) &= \sqrt{n}I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_1) + \sqrt{n}I(\theta_0)^{1/2}(\theta_1 - \theta_0) \\
&\approx I(\theta_0)^{1/2}I(\theta_1)^{-1}\frac{1}{\sqrt{n}}\sum_i \frac{\partial \log f(X_i;\theta_1)}{\partial \theta_1} + \sqrt{n}(\theta_1 - \theta_0) \\
&\xrightarrow{\mathcal{D}} N\left(0, \underbrace{I(\theta_0)^{1/2}I(\theta_1)^{-1}I(\theta_0)^{1/2}}_{\to I}\right) + \underbrace{\frac{I(\theta_0)^{1/2}\phi}{\sqrt{n}}}_{\to 0} \approx N(0, I_p).
\end{aligned}$$

  Thus the alternative is too close to the null for us to discriminate between the null and alternative.

- Suppose that $\theta_1 = \theta_0 + \frac{\phi}{\sqrt{n}}$ (for fixed ϕ), then

$$\begin{aligned}
\sqrt{n}I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_0) &= \sqrt{n}I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_1) + \sqrt{n}I(\theta_0)^{1/2}(\theta_1 - \theta_0) \\
&\approx I(\theta_0)^{1/2}I(\theta_1)^{-1}\frac{1}{\sqrt{n}}\sum_i \frac{\partial \log f(X_i;\theta_1)}{\partial \theta_1} + \sqrt{n}I(\theta_0)^{1/2}(\theta_1 - \theta_0) \\
&\xrightarrow{\mathcal{D}} N\left(0, I(\theta_0)^{1/2}I(\theta_1)^{-1}I(\theta_0)^{1/2}\right) + I(\theta_0)^{1/2}\phi \\
&\approx N\left(I(\theta_0)^{1/2}\phi, I(\theta_0)^{1/2}I(\theta_0 + \phi n^{-1/2})^{-1}I(\theta_0)^{1/2}\right).
\end{aligned}$$

  Therefore, for a given ϕ we can calculate the power at a given level α. Assume for simplicity that $\phi > 0$ and θ is univariate. Then

$$\begin{aligned}
P\left(|T_1| > z_{1-\alpha/2}\right) &\geq P\left(T_1 > z_{1-\alpha/2}\right) = P\left(Z > \frac{z_{1-\alpha/2} - I(\theta_0)^{1/2}\phi}{\sqrt{I(\theta_0)^{1/2}I(\theta_0 + \phi n^{-1/2})^{-1}I(\theta_0)^{1/2}}}\right) \\
&= 1 - \Phi\left(\frac{z_{1-\alpha/2} - I(\theta_0)^{1/2}\phi}{\sqrt{I(\theta_0)^{1/2}I(\theta_0 + \phi n^{-1/2})^{-1}I(\theta_0)^{1/2}}}\right) \\
&\approx 1 - \Phi\left(z_{1-\alpha/2} - \phi I(\theta_0)^{1/2}\right).
\end{aligned}$$

  this gives the power function of the test for a fixed n over ϕ. What we observe is that the power of the test $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$ depends on the size of $I(\theta_0)^{1/2}$. The larger the Fisher information $I(\theta_0)$ the greater the ability of the Wald test to discriminate between the null and the alternative. Based on what we understand about the Fisher information this make sense. The larger the Fisher information the "better" our ability to estimate the true parameter.

In the case that the dimension of θ is $p > 1$, we use the test statistic $\tilde{n}_1 = \big(\hat{\theta}_n - \theta_0\big)\sqrt{n}\mathrm{E}\bigg(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\bigg)^2\big(\hat{\theta}_n - \theta_0\big)$ instead of T_1. Noting that the distribution of T_1 is a chi-squared with p-degrees of freedom.

The Score test

The score test is based on the score. Under the null the distribution of the score is

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n}{\partial\theta}\rfloor_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0, \bigg\{\mathrm{E}\bigg(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\bigg)^2\bigg\}\bigg).$$

Thus we use as the test statistic

$$T_2 = \frac{1}{\sqrt{n}}\bigg\{\mathrm{E}\bigg(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\bigg)^2\bigg\}^{-1/2}\frac{\partial \mathcal{L}_n}{\partial\theta}\rfloor_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

An advantage of this test is that the maximum likelihood estimator (under either the null or alternative) does not have to be calculated.

The log-likelihood ratio test

This test is based on (2.18), and the test statistic is

$$T_3 = 2\big(\max_{\theta\in\Theta}\mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\big) \xrightarrow{\mathcal{D}} \chi^2_p.$$

T_3 is often called Wilk's statistic. An advantage of this test statistic is that it is asymptotically *pivotal*, in the sense that it does not depend on any nuisance parameters (we discuss this in the next chapter). However, using the chi-square distribution will only give the p-value corresponding to a "two-sided" hypothesis. This is because the chi-square distribution is based on the approximation

$$T_3 = 2\big(\mathcal{L}_n(\hat{\theta}_n) - \mathcal{L}_n(\theta_0)\big) \approx n(\widehat{\theta}_n - \theta_0)^2 I(\theta_0),$$

which assumes that $\widehat{\theta}_n = \arg\max_\theta \mathcal{L}_n(\theta)$ and solves $\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\widehat{\theta}_n} = 0$. However, in a one-sided test $H_0 : \mu = \mu_0$ vs $H_A : \mu > \mu_0$ parameter space is restricted to $\mu \geq \mu_0$ (it is on the boundary), this means that T_3 will not have a chi-square (though it will be very close) and p-value will be calculated in a slightly diifferent way. This boundary issue is not a problem for Wald test, since for the Wald test we simply calculate

$$P\left(Z \geq T_1\right) = \Phi\left(\sqrt{n}I(\theta_0)^{1/2}(\widehat{\theta} - \theta_0)\right).$$

Indeed, we show in Chapter 4, that the p-value for the one-sided test using the log-likelihood ratio statistic corresponds to that of p-value of the one-sided tests using the Wald statistic.

Exercise 2.9 *What do the score and log-likelihood ratio test statistics look like under the alternative? Derive the power function for these test statistics.*

You should observe that the power function for all three tests is the same.

Applications of the log-likeihood ratio to the multinomial distribution

We recall that the multinomial distribution is a generalisation of the binomial distribution. In this case at any given trial there can arise m different events (in the Binomial case $m = 2$). Let Z_i denote the outcome of the ith trial and assume $P(Z_i = k) = \pi_i$ ($\pi_1 + \ldots + \pi_m = 1$). Suppose that n trials are conducted and let Y_1 denote the number of times event 1 arises, Y_2 denote the number of times event 2 arises and so on. Then it is straightforward to show that

$$P(Y_1 = k_1, \ldots, Y_m = k_m) = \binom{n}{k_1, \ldots, k_m} \prod_{i=1}^{m} \pi_i^{k_i}.$$

If we do not impose any constraints on the probabilities $\{\pi_i\}$, given $\{Y_i\}_{i=1}^m$ it is straightforward to derive the mle of $\{\pi_i\}$ (it is very intuitive too!). Noting that $\pi_m = 1 - \sum_{i=1}^{m-1} \pi_i$, the log-likelihood of the multinomial is proportional to

$$\mathcal{L}_n(\underline{\pi}) \;=\; \sum_{i=1}^{m-1} y_i \log \pi_i + y_m \log(1 - \sum_{i=1}^{m-1} \pi_i).$$

Differentiating the above with respect to π_i and solving gives the mle estimator $\hat{\pi}_i = Y_i/n$! We observe that though there are m probabilities to estimate due to the constraint $\pi_m = 1 - \sum_{i=1}^{m-1} \pi_i$, we only have to estimate $(m-1)$ probabilities. We mention, that the same estimators can also be obtained by using Lagrange multipliers, that is maximising $\mathcal{L}_n(\underline{\pi})$ subject to the parameter constraint that $\sum_{j=1}^{p} \pi_i = 1$. To enforce this constraint, we normally add an additional term to $\mathcal{L}_n(\underline{\pi})$ and include the dummy variable λ. That is we define the constrained likelihood

$$\tilde{\mathcal{L}}_n(\underline{\pi}, \lambda) \;=\; \sum_{i=1}^{m} y_i \log \pi_i + \lambda(\sum_{i=1}^{m} \pi_i - 1).$$

Now if we maximise $\tilde{\mathcal{L}}_n(\underline{\pi}, \lambda)$ with respect to $\{\pi_i\}_{i=1}^m$ and λ we will obtain the estimators $\hat{\pi}_i = Y_i/n$ (which is the same as the maximum of $\mathcal{L}_n(\underline{\pi})$).

To derive the limiting distribution we note that the second derivative is

$$-\frac{\partial^2 \mathcal{L}_n(\underline{\pi})}{\partial \pi_i \pi_j} = \begin{cases} \frac{y_i}{\pi_i^2} + \frac{y_m}{(1-\sum_{r=1}^{m-1}\pi_r)^2} & i = j \\ \frac{y_m}{(1-\sum_{r=1}^{m-1}\pi_r)^2} & i \neq j \end{cases}$$

Hence taking expectations of the above the information matrix is the $(k-1) \times (k-1)$ matrix

$$I(\pi) = n \begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_m} & \frac{1}{\pi_m} & \cdots & \frac{1}{\pi_m} \\ \frac{1}{\pi_m} & \frac{1}{\pi_2} + \frac{1}{\pi_m} & \cdots & \frac{1}{\pi_m} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\pi_{m-1}} & \cdots & \frac{1}{\pi_{m-1}} + \frac{1}{\pi_m} & \end{pmatrix}.$$

Provided no of π_i is equal to either 0 or 1 (which would drop the dimension of m and make $I(\pi)$) singular, then the asymptotic distribution of the mle the normal with variance $I(\pi)^{-1}$.

Sometimes the probabilities $\{\pi_i\}$ will not be 'free' and will be determined by a parameter θ (where θ is an r-dimensional vector where $r < m$), ie. $\pi_i = \pi_i(\theta)$, in this case the likelihood of the multinomial is

$$\mathcal{L}_n(\underline{\pi}) = \sum_{i=1}^{m-1} y_i \log \pi_i(\theta) + y_m \log(1 - \sum_{i=1}^{m-1} \pi_i(\theta)).$$

By differentiating the above with respect to θ and solving we obtain the mle.

Pearson's goodness of Fit test

We now derive Pearson's goodness of Fit test using the log-likelihood ratio.

Suppose the null is $H_0 : \pi_1 = \tilde{\pi}_1, \ldots, \pi_m = \tilde{\pi}_m$ (where $\{\tilde{\pi}_i\}$ are some pre-set probabilities) and H_A : the probabilities are not the given probabilities. Hence we are testing restricted model (where we do not have to estimate anything) against the full model where we estimate the probabilities using $\pi_i = Y_i/n$.

The log-likelihood ratio in this case is

$$W = 2\big\{ \arg\max_{\pi} \mathcal{L}_n(\pi) - \mathcal{L}_n(\tilde{\pi}) \big\}.$$

Under the null we know that $W = 2\big\{ \arg\max_\pi \mathcal{L}_n(\pi) - \mathcal{L}_n(\tilde{\pi}) \big\} \xrightarrow{\mathcal{D}} \chi^2_{m-1}$ (because we have to estimate $(m-1)$ parameters). We now derive an expression for W and show that the Pearson-statistic is an approximation of this.

$$\begin{aligned} \frac{1}{2}W &= \sum_{i=1}^{m-1} Y_i \log\big(\frac{Y_i}{n}\big) + Y_m \log\frac{Y_m}{n} - \sum_{i=1}^{m-1} Y_i \log\tilde{\pi}_i - Y_m \log\tilde{\pi}_m \\ &= \sum_{i=1}^{m} Y_i \log\big(\frac{Y_i}{n\tilde{\pi}_i}\big). \end{aligned}$$

Recall that Y_i is often called the observed $Y_i = O_i$ and $n\tilde{\pi}_i$ the expected under the null $E_i = n\tilde{\pi}_i$. Then $W = 2\sum_{i=1}^m O_i \log\big(\frac{O_i}{E_i}\big) \xrightarrow{\mathcal{P}} \chi^2_{m-1}$. By making a Taylor expansion of $x\log(xa^{-1})$ about $x = a$ we have $x\log(xa^{-1}) \approx a\log(aa^{-1}) + (x-a) + \frac{1}{2}(x-a)^2/a$. We let $O = x$ and $E = a$, then assuming the null is true and $E_i \approx O_i$ we have

$$W = 2\sum_{i=1}^{m} Y_i \log\big(\frac{Y_i}{n\tilde{\pi}_i}\big) \approx 2\sum_{i=1}^{m} \big((O_i - E_i) + \frac{1}{2}\frac{(O_i-E_i)^2}{E_i}\big).$$

Now we note that $\sum_{i=1}^m E_i = \sum_{i=1}^m O_i = n$ hence the above reduces to

$$W \approx \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{\mathcal{D}} \chi^2_{m-1}.$$

We recall that the above is the Pearson test statistic. Hence this is one methods for deriving the Pearson chi-squared test for goodness of fit.

Remark 2.8.1 (Beyond likelihood) *In several applications in statistics we cannot articulate the question of interest in terms of the parameters of a distribution. However, we can often articulate it in terms of some parameters, ϕ. Indeed, whether ϕ is zero or will tell us something about the data. For example:*

(i) Are some parameters in a linear regression zero?

(ii) Is there correlation between two variables?

(iii) Is there an interaction between two categorical variables in a regression?

(iv) In my own area of research on detecting nonstationarities, transforming the time series can yield more information then the original data. For example, nonstationarities imply correlations in the transformed data. The list goes on.

None of the above requires us to place distributional assumptions on the data. However, we can still test $H_0 : \phi = 0$ against $H_A : \phi \neq 0$. If we can estimate this quantity and obtain its limiting distribution under the null and show that under the alternative it "shifts", using $\widehat{\phi}$ we can construct a test statistic which has some power (though it may not be the most powerful test).

2.9 Some questions

Exercise 2.10 *A parameterisation of a distribution is identifiably if there does not exist another set of parameters which can give the same distribution.* `https://en.wikipedia.org/wiki/Identifiability`. *Recall this assumption was used when deriving the sampling properties of the maximum likelihood estimator.*

Suppose X_i are iid random variables which come from a mixture of distributions. The density of X_i is

$$f(x; \pi, \lambda_1, \lambda_2) = \pi\lambda_1 \exp(-\lambda_1 x) + (1 - \pi)\lambda_2 \exp(-\lambda_2 x)$$

where $x > 0$, $\lambda_1, \lambda_2 > 0$ and $0 \leq \pi \leq 1$.

(i) Are the parameters identifiable?

(ii) Does standard theory apply when using the log-likelihood ratio test to test $H_0 : \pi = 0$ vs $H_A : \pi \neq 0$.

(iii) Does standard theory apply when using the log-likelihood to estimate π when $\lambda_1 = \lambda_2$.

Chapter 3

The Profile Likelihood

3.1 The Profile Likelihood

3.1.1 The method of profiling

Let us suppose that the unknown parameters θ can be partitioned as $\theta' = (\psi', \lambda')$, where ψ are the p-dimensional parameters of interest (eg. mean) and λ are the q-dimensional nuisance parameters (eg. variance). We will need to estimate both ψ and λ, but our interest lies only in the parameter ψ. To achieve this one often profiles out the nuisance parameters. To motivate the profile likelihood, we first describe a method to estimate the parameters (ψ, λ) in two stages and consider some examples.

Let us suppse that $\{X_i\}$ are iid random variables, with density $f(x; \psi, \lambda)$ where our objective is to estimate ψ and λ. In this case the log-likelihood is

$$\mathcal{L}_n(\psi, \lambda) = \sum_{i=1}^{n} \log f(X_i; \psi, \lambda).$$

To estimate ψ and λ one can use $(\hat{\lambda}_n, \hat{\psi}_n) = \arg\max_{\lambda,\psi} \mathcal{L}_n(\psi, \lambda)$. However, this can be difficult to directly maximise. Instead let us consider a different method, which may, sometimes, be easier to evaluate. Suppose, for now, ψ is known, then we rewrite the likelihood as $\mathcal{L}_n(\psi, \lambda) = \mathcal{L}_\psi(\lambda)$ (to show that ψ is fixed but λ varies). To estimate λ we maximise $\mathcal{L}_\psi(\lambda)$ with respect to λ, i.e.

$$\hat{\lambda}_\psi = \arg\max_{\lambda} \mathcal{L}_\psi(\lambda).$$

In reality ψ is unknown, hence for each ψ we can evaluate $\hat{\lambda}_\psi$. Note that for each ψ, we have a new curve $\mathcal{L}_\psi(\lambda)$ over λ. Now to estimate ψ, we evaluate the maximum $\mathcal{L}_\psi(\lambda)$, over λ, and choose the ψ, which is the maximum over all these curves. In other words, we evaluate

$$\hat{\psi}_n = \arg\max_\psi \mathcal{L}_\psi(\hat{\lambda}_\psi) = \arg\max_\psi \mathcal{L}_n(\psi, \hat{\lambda}_\psi).$$

A bit of logical deduction shows that $\hat{\psi}_n$ and $\lambda_{\hat{\psi}_n}$ are the maximum likelihood estimators $(\hat{\lambda}_n, \hat{\psi}_n) = \arg\max_{\psi,\lambda} \mathcal{L}_n(\psi, \lambda)$.

We note that we have *profiled* out nuisance parameter λ, and the likelihood $\mathcal{L}_\psi(\hat{\lambda}_\psi) = \mathcal{L}_n(\psi, \hat{\lambda}_\psi)$ is in terms of the parameter of interest ψ.

The advantage of this procedure is best illustrated through some examples.

Example 3.1.1 (The Weibull distribution) *Let us suppose that $\{X_i\}$ are iid random variables from a Weibull distribution with density $f(x;\alpha,\theta) = \frac{\alpha y^{\alpha-1}}{\theta^\alpha}\exp(-(y/\theta)^\alpha)$. We know from Example 2.2.2, that if α, were known an explicit expression for the MLE can be derived, it is*

$$\begin{aligned}
\hat{\theta}_\alpha &= \arg\max_\theta \mathcal{L}_\alpha(\theta) \\
&= \arg\max_\theta \sum_{i=1}^n \left(\log\alpha + (\alpha-1)\log Y_i - \alpha\log\theta - \big(\frac{Y_i}{\theta}\big)^\alpha \right) \\
&= \arg\max_\theta \sum_{i=1}^n \left(-\alpha\log\theta - \big(\frac{Y_i}{\theta}\big)^\alpha \right) = (\frac{1}{n}\sum_{i=1}^n Y_i^\alpha)^{1/\alpha},
\end{aligned}$$

where $\mathcal{L}_\alpha(\underline{X};\theta) = \sum_{i=1}^n \left(\log\alpha + (\alpha-1)\log Y_i - \alpha\log\theta - \big(\frac{Y_i}{\theta}\big)^\alpha \right)$. Thus for a given α, the maximum likelihood estimator of θ can be derived. The maximum likelihood estimator of α is

$$\hat{\alpha}_n = \arg\max_\alpha \sum_{i=1}^n \left(\log\alpha + (\alpha-1)\log Y_i - \alpha\log(\frac{1}{n}\sum_{i=1}^n Y_i^\alpha)^{1/\alpha} - \big(\frac{Y_i}{(\frac{1}{n}\sum_{i=1}^n Y_i^\alpha)^{1/\alpha}}\big)^\alpha \right).$$

Therefore, the maximum likelihood estimator of θ is $(\frac{1}{n}\sum_{i=1}^n Y_i^{\hat{\alpha}_n})^{1/\hat{\alpha}_n}$. We observe that evaluating $\hat{\alpha}_n$ can be tricky but no worse than maximising the likelihood $\mathcal{L}_n(\alpha,\theta)$ over α and θ.

As we mentioned above, we are not interest in the nuisance parameters λ and are only interesting in testing and constructing CIs for ψ. In this case, we are interested in the limiting distribution of the MLE $\hat{\psi}_n$. Using Theorem 2.6.2(ii) we have

$$\sqrt{n}\begin{pmatrix} \hat{\psi}_n - \psi \\ \hat{\lambda}_n - \lambda \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix}^{-1}\Big).$$

where

$$\begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} \mathrm{E}\big(-\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial\psi^2}\big) & \mathrm{E}\big(-\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial\psi\partial\lambda}\big) \\ \mathrm{E}\big(-\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial\psi\partial\lambda}\big)' & \mathrm{E}\big(-\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial\psi^2}\big) \end{pmatrix}. \tag{3.1}$$

To derive an exact expression for the limiting variance of $\sqrt{n}(\hat{\psi}_n - \psi)$, we use the block inverse matrix identity.

Remark 3.1.1 (Inverse of a block matrix) *Suppose that*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

is a square matrix. Then

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}CB(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}. \tag{3.2}$$

Using (3.2) we have

$$\sqrt{n}(\widehat{\psi}_n - \psi) \xrightarrow{\mathcal{D}} \mathcal{N}(0, (I_{\psi,\psi} - I_{\psi,\lambda}I_{\lambda\lambda}^{-1}I_{\lambda,\psi})^{-1}). \tag{3.3}$$

Thus if ψ is a scalar we can use the above to construct confidence intervals for ψ.

Example 3.1.2 (Block diagonal information matrix) *If*

$$I(\psi, \lambda) = \begin{pmatrix} I_{\psi,\psi} & 0 \\ 0 & I_{\lambda,\lambda} \end{pmatrix},$$

then using (3.3) we have

$$\sqrt{n}(\widehat{\psi}_n - \psi) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_{\psi,\psi}^{-1}).$$

3.1.2 The score and the log-likelihood ratio for the profile likelihood

To ease notation, let us suppose that ψ_0 and λ_0 are the true parameters in the distribution. We now consider the log-likelihood ratio

$$2\Big\{ \max_{\psi,\lambda} \mathcal{L}_n(\psi,\lambda) - \max_{\lambda} \mathcal{L}_n(\psi_0,\lambda)\Big\}, \tag{3.4}$$

where ψ_0 is the true parameter. However, to derive the limiting distribution in this case for this statistic is a little more complicated than the log-likelihood ratio test that does not involve nuisance parameters. This is because directly applying Taylor expansion does not work since this is usually expanded about the true parameters. We observe that

$$\begin{aligned} & 2\Big\{ \max_{\psi,\lambda} \mathcal{L}_n(\psi,\lambda) - \max_{\lambda} \mathcal{L}_n(\psi_0,\lambda)\Big\} \\ = & \underbrace{2\Big\{ \max_{\psi,\lambda} \mathcal{L}_n(\psi,\lambda) - \mathcal{L}_n(\psi_0,\lambda_0)\Big\}}_{\chi^2_{p+q}} - \underbrace{2\Big\{ \max_{\lambda} \mathcal{L}_n(\psi_0,\lambda) - \max_{\lambda} \mathcal{L}_n(\psi_0,\lambda_0)\Big\}}_{\chi^2_q}. \end{aligned}$$

It seems reasonable that the difference may be a χ^2_p but it is really not clear by. Below, we show that by using a few Taylor expansions why this is true.

In the theorem below we will derive the distribution of the score and the nested log-likelihood.

Theorem 3.1.1 *Suppose Assumption 2.6.1 holds. Suppose that (ψ_0,λ_0) are the true parameters. Then we have*

$$\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0} \approx \frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\psi_0,\lambda_0} - \frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \lambda}\rfloor_{\psi_0,\lambda_0} I^{-1}_{\lambda_0\lambda_0} I_{\lambda_0\psi_0} \tag{3.5}$$

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\psi_0,\hat{\lambda}_{\psi_0}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,(I_{\psi_0\psi_0} - I_{\psi_0\lambda_0} I^{-1}_{\lambda_0\lambda_0} I_{\lambda_0,\psi_0})) \tag{3.6}$$

where I is defined as in (3.1) and

$$2\Big\{ \mathcal{L}_n(\hat{\psi}_n,\hat{\lambda}_n) - \mathcal{L}_n(\psi_0,\hat{\lambda}_{\psi_0})\Big\} \xrightarrow{\mathcal{D}} \chi^2_p, \tag{3.7}$$

where p denotes the dimension of ψ. This result is often called Wilks Theorem.

PROOF. We first prove (3.5) which is the basis of the proofs of (3.6). To avoid, notational difficulties we will assume that $\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ and $\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \lambda}\rfloor_{\lambda=\lambda_0,\psi_0}$ are univariate random variables.

Our objective is to find an expression for $\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ in terms of $\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \lambda}\rfloor_{\lambda=\lambda_0,\psi_0}$ and $\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda=\lambda_0,\psi_0}$ which will allow us to obtain its variance and asymptotic distribution.

Making a Taylor expansion of $\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ about $\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0}$ gives

$$\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0} \approx \frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0} + (\hat{\lambda}_{\psi_0}-\lambda_0)\frac{\partial^2 \mathcal{L}_n(\psi,\lambda)}{\partial \lambda \partial \psi}\rfloor_{\lambda_0,\psi_0}.$$

Notice that we have used $\approx$ instead of $=$ because we replace the second derivative with its true parameters. If the sample size is large enough then $\frac{\partial^2 \mathcal{L}_n(\psi,\lambda)}{\partial \lambda \partial \psi}\rfloor_{\lambda_0,\psi_0} \approx \mathrm{E}\big(\frac{\partial^2 \mathcal{L}_n(\psi,\lambda)}{\partial \lambda \partial \psi}\rfloor_{\lambda_0,\psi_0}\big)$; eg. in the iid case we have

$$\begin{aligned}\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\psi,\lambda)}{\partial \lambda \partial \psi}\rfloor_{\lambda_0,\psi_0} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial \lambda \partial \psi}\rfloor_{\lambda_0,\psi_0} \\ &\approx \mathrm{E}\left(\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial \lambda \partial \psi}\rfloor_{\lambda_0,\psi_0}\right) = -I_{\lambda,\psi}\end{aligned}$$

Therefore

$$\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0} \approx \frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0} - n(\hat{\lambda}_{\psi_0}-\lambda_0)I_{\lambda\psi}. \tag{3.8}$$

Next we make a decomposition of $(\hat{\lambda}_{\psi_0}-\lambda_0)$. We recall that since $\mathcal{L}_n(\psi_0,\hat{\lambda}_{\psi_0}) = \arg\max_\lambda \mathcal{L}_n(\psi_0,\lambda)$ then

$$\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \lambda}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0} = 0$$

(if the maximum is not on the boundary). Therefore making a Taylor expansion of $\frac{\partial \mathcal{L}_n(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ about $\frac{\partial \mathcal{L}_n(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0}$ gives

$$\underbrace{\frac{\partial \mathcal{L}_n(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}}_{=0} \approx \frac{\partial \mathcal{L}_n(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0} + \frac{\partial^2 \mathcal{L}_n(\psi_0,\lambda)}{\partial \lambda^2}\rfloor_{\lambda_0,\psi_0}(\hat{\lambda}_{\psi_0}-\lambda_0).$$

Replacing $\frac{\partial^2 \mathcal{L}_n(\psi_0,\lambda)}{\partial \lambda^2}\rfloor_{\lambda_0,\psi_0}$ with $I_{\lambda\lambda}$ gives

$$\frac{\partial \mathcal{L}_n(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0} - nI_{\lambda\lambda}(\hat{\lambda}_{\psi_0}-\lambda_0) \approx 0,$$

and rearranging the above gives

$$(\hat{\lambda}_{\psi_0} - \lambda_0) \approx \frac{I_{\lambda\lambda}^{-1}}{n} \frac{\partial \mathcal{L}_n(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\lambda_0, \psi_0}. \tag{3.9}$$

Therefore substituting (3.9) into (3.8) gives

$$\frac{\partial \mathcal{L}_n(\psi, \lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0}, \psi_0} \approx \frac{\partial \mathcal{L}_n(\psi, \lambda)}{\partial \psi} \rfloor_{\lambda_0, \psi_0} - \frac{\partial \mathcal{L}_n(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\psi_0, \lambda_0} I_{\lambda\lambda}^{-1} I_{\lambda\psi}$$

and thus we have proved (3.5).

To prove (3.6) we note that

$$\frac{\partial \mathcal{L}_n(\psi, \lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0}, \psi_0} \approx \frac{\partial \mathcal{L}_n(\psi_0, \lambda)}{\partial \theta} \rfloor'_{\psi_0, \lambda_0} \left(I, -I_{\lambda\lambda}^{-1} \lambda_{\lambda,\psi} \right)'. \tag{3.10}$$

We recall that the regular score function satisfies

$$\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\psi, \lambda)}{\partial \theta} \rfloor_{\lambda_0, \psi_0} = \frac{1}{\sqrt{n}} \begin{pmatrix} \frac{\partial \mathcal{L}_n(\psi, \lambda)}{\partial \psi} \rfloor_{\lambda_0, \psi_0} \\ \frac{\partial \mathcal{L}_n(\psi, \lambda)}{\partial \lambda} \rfloor_{\psi_0, \lambda_0} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0)).$$

Now by substituting the above into (3.10) and calculating the variance gives (3.6).

Finally to prove (3.7) we apply the Taylor expansion on the decomposition

$$\begin{aligned} 2\Big\{ \mathcal{L}_n(\hat{\psi}_n, \hat{\lambda}_n) - \mathcal{L}_n(\psi_0, \hat{\lambda}_{\psi_0}) \Big\} &= 2\Big\{ \mathcal{L}_n(\hat{\psi}_n, \hat{\lambda}_n) - \mathcal{L}_n(\psi_0, \lambda_0) \Big\} - 2\Big\{ \mathcal{L}_n(\psi_0, \hat{\lambda}_{\psi_0}) - \mathcal{L}_n(\psi_0, \lambda_0) \Big\} \\ &\approx (\hat{\theta}_n - \theta_0)' I(\theta) (\hat{\theta}_n - \theta_0) - (\hat{\lambda}_{\psi_0} - \lambda_0)' I_{\lambda\lambda} (\hat{\lambda}_{\psi_0} - \lambda_0), \end{aligned} \tag{3.11}$$

where $\hat{\theta}_n' = (\hat{\psi}, \hat{\lambda})$ (the mle). We now find an approximation of $(\hat{\lambda}_{\psi_0} - \lambda_0)'$ in terms $(\hat{\theta}_n - \theta_0)$. We recall that $(\widehat{\theta} - \theta) = I(\theta_0)^{-1} \nabla_\theta \mathcal{L}_n(\theta) \rfloor_{\theta=\theta_0}$ therefore

$$\begin{pmatrix} \frac{\partial \mathcal{L}_n(\theta)}{\partial \psi} \\ \frac{\partial \mathcal{L}_n(\theta)}{\partial \lambda} \end{pmatrix} \approx \begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix} \begin{pmatrix} \widehat{\psi}_n - \psi_0 \\ \widehat{\lambda}_n - \lambda_n \end{pmatrix} \tag{3.12}$$

From (3.9) and the expansion of $\frac{\partial \mathcal{L}_n(\theta)}{\partial \lambda}$ given in (3.12) we have

$$\begin{aligned} (\hat{\lambda}_{\psi_0} - \lambda_0) &\approx \frac{I_{\lambda\lambda}^{-1}}{n} \frac{\partial \mathcal{L}_n(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\lambda_0, \psi_0} \approx \frac{I_{\lambda\lambda}^{-1}}{n} \Big(I_{\lambda\psi} (\hat{\psi} - \psi_0) + I_{\lambda\lambda} (\hat{\lambda} - \lambda_0) \Big) \\ &\approx \frac{1}{n} I_{\lambda\lambda}^{-1} I_{\lambda\psi} (\hat{\psi} - \psi_0) + (\hat{\lambda} - \lambda_0) = \frac{1}{n} \left(I_{\lambda\lambda}^{-1} I_{\lambda\psi}, 1 \right) \left(\widehat{\theta}_n - \theta_0 \right). \end{aligned}$$

Substituting the above into (3.11) and making lots of cancellations we have

$$2\Big\{ \mathcal{L}_n(\hat{\psi}_n, \hat{\lambda}_n) - \mathcal{L}_n(\psi_0, \hat{\lambda}_{\psi_0}) \Big\} \approx n(\hat{\psi} - \psi_0)' (I_{\psi\psi} - I_{\psi\lambda} I_{\lambda,\lambda}^{-1} I_{\lambda,\psi}) (\hat{\psi} - \psi_0).$$

Finally, by using (3.3) we substitute $\sqrt{n}(\hat{\psi} - \psi_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, (I_{\psi\psi} - I_{\psi\lambda} I_{\lambda,\lambda}^{-1} I_{\lambda,\psi})^{-1})$, into the above which gives the desired result. □

Remark 3.1.2 *(i) The limiting variance of* $\widehat{\psi} - \psi_0$ *if* λ *were* known *is* $I^{-1}_{\psi,\psi}$*, whereas the the limiting variance of* $\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ *is* $(I_{\psi\psi} - I_{\psi\lambda}I^{-1}_{\lambda,\lambda}I_{\lambda,\psi})$ *and the limiting variance of* $\sqrt{n}(\hat{\psi} - \psi_0)$ *is* $(I_{\psi\psi} - I_{\psi\lambda}I^{-1}_{\lambda,\lambda}I_{\lambda,\psi})^{-1}$*. Therefore if* ψ *and* λ *are scalars and the correlation* $I_{\lambda,\psi}$ *is positive, then the limiting variance of* $\widehat{\psi} - \psi_0$ *is* more *than if* λ *were known. This makes sense, if we have less information the variance grows.*

(ii) Look again at the expression

$$\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0} \approx \frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0} - I_{\psi\lambda}I^{-1}_{\lambda\lambda}\frac{\partial \mathcal{L}_n(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0} \tag{3.13}$$

It is useful to understand where it came from. Consider the problem of linear regression. Suppose X *and* Y *are random variables and we want to construct the best linear predictor of* Y *given* X*. We know that the best linear predictor is* $\hat{Y}(X) = \mathrm{E}(XY)/\mathrm{E}(X^2)X$ *and the residual and mean squared error is*

$$Y - \hat{Y}(X) = Y - \frac{\mathrm{E}(XY)}{\mathrm{E}(X^2)}X \text{ and } \mathrm{E}\bigg(Y - \frac{\mathrm{E}(XY)}{\mathrm{E}(X^2)}X\bigg)^2 = \mathrm{E}(Y^2) - \mathrm{E}(XY)\mathrm{E}(X^2)^{-1}\mathrm{E}(XY).$$

Compare this expression with (3.13). We see that in some sense $\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ *can be treated as the residual (error) of the projection of* $\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0}$ *onto* $\frac{\partial \mathcal{L}_n(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0}$*.*

3.1.3 The log-likelihood ratio statistics in the presence of nuisance parameters

Theorem 3.1.1 can be used to test $H_0 : \psi = \psi_0$ against $H_A : \psi \neq \psi_0$ since

$$2\bigg\{\max_{\psi,\lambda}\mathcal{L}_n(\psi,\lambda) - \max_{\lambda}\mathcal{L}_n(\psi_0,\lambda)\bigg\} \xrightarrow{\mathcal{D}} \chi^2_p.$$

The same quantity can be used in the construction of confidence intervals By using (3.7) we can construct CIs. For example, to construct a 95% CI for ψ we can use the mle $\hat{\theta}_n = (\hat{\psi}_n, \hat{\lambda}_n)$ and the profile likelihood (3.7) to give

$$\bigg\{\psi; 2\bigg\{\mathcal{L}_n(\hat{\psi}_n,\hat{\lambda}_n) - \mathcal{L}_n(\psi,\hat{\lambda}_\psi)\bigg\} \leq \chi^2_p(0.95)\bigg\}.$$

Example 3.1.3 (The normal distribution and confidence intervals) *This example is taken from Davidson (2004), Example 4.31, p129.*

We recall that the log-likelihood for $\{Y_i\}$ which are iid random variables from a normal distribution with mean μ and variance σ^2 is

$$\mathcal{L}_n(\mu,\sigma^2) = \mathcal{L}_\mu(\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i-\mu)^2 - \frac{n}{2}\log\sigma^2.$$

Our aim is to the use the log-likelihood ratio statistic, analogous to Section 2.8.1 to construct a CI for μ. Thus we treat σ^2 as the nuisance parameter.

Keeping μ fixed, the maximum likelihood estimator of σ^2 is $\widehat{\sigma}^2(\mu) = \frac{1}{n}\sum_{i=1}^n (Y_i-\mu)^2$. Rearranging $\widehat{\sigma}^2(\mu)$ gives

$$\widehat{\sigma}^2(\mu) = \frac{n-1}{n}s^2\left(1+\frac{t_n^2(\mu)}{n-1}\right)$$

where $t_n^2(\mu) = n(\bar{Y}-\mu)^2/s^2$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i-\bar{Y})^2$. Substituting $\widehat{\sigma}^2(\mu)$ into $\mathcal{L}_n(\mu,\sigma^2)$ gives the profile likelihood

$$\begin{aligned}\mathcal{L}_n(\mu,\widehat{\sigma}^2(\mu)) &= \underbrace{\frac{-1}{\widehat{\sigma}^2(\mu)}\sum_{i=1}^n (Y_i-\mu)^2}_{=-n/2} - \frac{n}{2}\log\widehat{\sigma}^2(\mu) \\ &= -\frac{n}{2} - \frac{n}{2}\log\left\{\frac{n-1}{n}s^2\left(1+\frac{t_n^2(\mu)}{n-1}\right)\right\}.\end{aligned}$$

It is clear that $\mathcal{L}_n(\mu,\widehat{\sigma}^2(\mu))$ is maximised at $\widehat{\mu} = \bar{Y}$. Hence

$$\mathcal{L}_n(\widehat{\mu},\widehat{\sigma}^2(\widehat{\mu})) = -\frac{n}{2} - \frac{n}{2}\log\left\{\frac{n-1}{n}s^2\right\}.$$

Thus the log-likelihood ratio is

$$W_n(\mu) = 2\left\{\mathcal{L}_n(\widehat{\mu},\widehat{\sigma}^2(\widehat{\mu})) - \mathcal{L}_n(\mu,\widehat{\sigma}^2(\mu))\right\} = \underbrace{n\log\left(1+\frac{t_n^2(\mu)}{n-1}\right)}_{\xrightarrow{\mathcal{D}}\chi_1^2 \text{ for true } \mu}.$$

Therefore, using the same argument to those in Section 2.8.1, the 95% confidence interval for the mean is

$$\begin{aligned}\left\{\mu; 2\left\{\mathcal{L}_n(\widehat{\mu},\widehat{\sigma}^2(\widehat{\mu})) - \mathcal{L}_n(\mu,\widehat{\sigma}^2(\mu))\right\}\right\} &= \left\{\mu; W_n(\mu) \le \chi_1^2(0.95)\right\} \\ &= \left\{\mu; n\log\left(1+\frac{t_n^2(\mu)}{n-1}\right) \le \chi_1^2(0.95)\right\}.\end{aligned}$$

However, this is an asymptotic result. With the normal distribution we can get the exact distribution. We note that since log *is a monotonic function the log-likelihood ratio is equivalent to*

$$\left\{\mu; t_n^2(\mu) \leq C_\alpha\right\},$$

where C_α is an appropriately chosen critical value. We recall that $t_n(\mu)$ is a t-distribution with $n-1$ degrees of freedom. Thus C_α is the critical value corresponding to a Hotelling T^2-distribution.

Exercise 3.1 *Derive the χ^2 test for independence (in the case of two by two tables) using the log-likelihood ratio test. More precisely, derive the asymptotic distribution of*

$$T = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4},$$

under the null that there is no association between the categorical variables C and R, where and $E_1 = n_3 \times n_1/N$, $E_2 = n_4 \times n_1/N$, $E_3 = n_3 \times n_2/N$ and $E_2 = n_4 \times n_2/N$. State

	C_1	C_2	Subtotal
R_1	O_1	O_2	n_1
R_2	O_3	O_4	n_2
Subtotal	n_3	n_4	N

all results you use.

Hint: You may need to use the Taylor approximation $x\log(x/y) \approx (x-y)+\frac{1}{2}(x-y)^2/y$.

Pivotal Quantities

Pivotal quantities are statistics whose distribution does not depend on any parameters. These include the t-ratio $t = \sqrt{n}(\bar{X}-\mu)/s_n \sim t_{n-1}$ (in the case the data is normal) F-test etc.

In many applications it is not possible to obtain a pivotal quantity, but a quantity can be *asymptotically* pivotal. The log-likelihood ratio statistic is one such example (since its distribution is a chi-square).

Pivotal statistics have many advantages. The main is that it avoids the need to estimate extra parameters. However, they are also useful in developing Bootstrap methods etc.

3.1.4 The score statistic in the presence of nuisance parameters

We recall that we used Theorem 3.1.1 to obtain the distribution of $2\big\{\max_{\psi,\lambda}\mathcal{L}_n(\psi,\lambda) - \max_{\lambda}\mathcal{L}_n(\psi_0,\lambda)\big\}$ under the null, we now consider the score test.

We recall that under the null $H_0:\psi=\psi_0$ the derivative $\frac{\partial\mathcal{L}_n(\psi,\lambda)}{\partial\lambda}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}=0$, but the same is not true of $\frac{\partial\mathcal{L}_n(\psi,\lambda)}{\partial\psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$. However, if the null were true we would expect $\hat{\lambda}_{\psi_0}$ to be close to the true λ_0 and for $\frac{\partial\mathcal{L}_n(\psi,\lambda)}{\partial\psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ to be close to zero. Indeed this is what we showed in (3.6), where we showed that under the null

$$n^{-1/2}\frac{\partial\mathcal{L}_n(\psi,\lambda)}{\partial\psi}\rfloor_{\hat{\lambda}_{\psi_0}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_{\psi\psi} - I_{\psi\lambda}I_{\lambda,\lambda}^{-1}I_{\lambda,\psi}), \tag{3.14}$$

where $\lambda_{\psi_0}=\arg\max_{\lambda}\mathcal{L}_n(\psi_0,\lambda)$.

Therefore (3.14) suggests an alternative test for $H_0:\psi=\psi_0$ against $H_A:\psi\neq\psi_0$. We can use $\frac{1}{\sqrt{n}}\frac{\partial\mathcal{L}_n(\psi,\lambda)}{\partial\psi}\rfloor_{\hat{\lambda}_{\psi_0}}$ as the test statistic. This is called the score or LM test.

The log-likelihood ratio test and the score test are asymptotically equivalent. There are advantages and disadvantages of both.

(i) An advantage of the log-likelihood ratio test is that we do not need to calculate the information matrix.

(ii) An advantage of the score test is that we do not have to evaluate the the maximum likelihood estimates under the alternative model.

3.2 Applications

3.2.1 An application of profiling to frequency estimation

Suppose that the observations $\{X_t; t=1,\ldots,n\}$ satisfy the following nonlinear regression model

$$X_t = A\cos(\omega t) + B\sin(\omega t) + \varepsilon_i$$

where $\{\varepsilon_t\}$ are iid standard normal random variables and $0<\omega<\pi$ (thus allowing the case $\omega=\pi/2$, but not the end points $\omega=0$ or π). The parameters A, B, and ω are real and unknown. Full details can be found in the paper `http://www.jstor.org/stable/pdf/2334314.pdf` (Walker, 1971, Biometrika).

(i) Ignoring constants, obtain the log-likelihood of $\{X_t\}$. Denote this likelihood as $\mathcal{L}_n(A, B, \omega)$.

(ii) Let

$$\mathcal{S}_n(A, B, \omega) = \Big(\sum_{t=1}^{n} X_t^2 - 2 \sum_{t=1}^{n} X_t \big(A\cos(\omega t) + B\sin(\omega t) \big) - \frac{1}{2} n(A^2 + B^2) \Big).$$

Show that

$$2\mathcal{L}_n(A, B, \omega) + \mathcal{S}_n(A, B, \omega) = -\frac{(A^2 - B^2)}{2} \sum_{t=1}^{n} \cos(2t\omega) + AB \sum_{t=1}^{n} \sin(2t\omega).$$

Thus show that $|\mathcal{L}_n(A, B, \omega) + \frac{1}{2}\mathcal{S}_n(A, B, \omega)| = O(1)$ (ie. the difference does not grow with n).

Since $\mathcal{L}_n(A, B, \omega)$ and $-\frac{1}{2}\mathcal{S}_n(A, B, \omega)$ are asymptotically equivalent, for the rest of this question, *use* $\frac{-1}{2}\mathcal{S}_n(A, B, \omega)$ instead of the likelihood $\mathcal{L}_n(A, B, \omega)$.

(iii) Obtain the profile likelihood of ω.

(hint: Profile out the parameters A and B, to show that $\hat{\omega}_n = \arg\max_\omega |\sum_{t=1}^{n} X_t \exp(it\omega)|^2$).

Suggest, a graphical method for evaluating $\hat{\omega}_n$?

(iv) By using the identity

$$\sum_{t=1}^{n} \exp(i\Omega t) = \begin{cases} \frac{\exp(\frac{1}{2}i(n+1)\Omega)\sin(\frac{1}{2}n\Omega)}{\sin(\frac{1}{2}\Omega)} & 0 < \Omega < 2\pi \\ n & \Omega = 0 \text{ or } 2\pi. \end{cases} \tag{3.15}$$

show that for $0 < \Omega < 2\pi$ we have

$$\sum_{t=1}^{n} t\cos(\Omega t) = O(n) \quad \sum_{t=1}^{n} t\sin(\Omega t) = O(n)$$
$$\sum_{t=1}^{n} t^2\cos(\Omega t) = O(n^2) \quad \sum_{t=1}^{n} t^2\sin(\Omega t) = O(n^2).$$

(v) By using the results in part (iv) show that the Fisher Information of $\mathcal{L}_n(A, B, \omega)$ (denoted as $I(A, B, \omega)$) is asymptotically equivalent to

$$2I(A, B, \omega) = E\big(\frac{\partial^2 \mathcal{S}_n}{\partial \omega^2}\big) = \begin{pmatrix} \frac{n}{2} & 0 & \frac{n^2}{2}B + O(n) \\ 0 & \frac{n}{2} & -\frac{n^2}{2}A + O(n) \\ \frac{n^2}{2}B + O(n) & -\frac{n^2}{2}A + O(n) & \frac{n^3}{3}(A^2 + B^2) + O(n^2) \end{pmatrix}.$$

(vi) Derive the asymptotic variance of maximum likelihood estimator, $\hat{\omega}_n$, derived in part (iv).

Comment on the rate of convergence of $\hat{\omega}_n$.

Useful information: The following quantities may be useful:

$$\sum_{t=1}^{n} \exp(i\Omega t) = \begin{cases} \frac{\exp(\frac{1}{2}i(n+1)\Omega)\sin(\frac{1}{2}n\Omega)}{\sin(\frac{1}{2}\Omega)} & 0 < \Omega < 2\pi \\ n & \Omega = 0 \text{ or } 2\pi. \end{cases} \tag{3.16}$$

the trignometric identities: $\sin(2\Omega) = 2\sin\Omega\cos\Omega$, $\cos(2\Omega) = 2\cos^2(\Omega) - 1 = 1 - 2\sin^2\Omega$, $\exp(i\Omega) = \cos(\Omega) + i\sin(\Omega)$ and

$$\sum_{t=1}^{n} t = \frac{n(n+1)}{2} \qquad \sum_{t=1}^{n} t^2 = \frac{n(n+1)(2n+1)}{6}.$$

Solution

Since $\{\varepsilon_i\}$ are standard normal iid random variables the likelihood is

$$\mathcal{L}_n(A, B, \omega) = -\frac{1}{2}\sum_{t=1}^{n}(X_t - A\cos(\omega t) - B\sin(\omega t))^2.$$

If the frequency ω were known, then the least squares estimator of A and B would be

$$\begin{pmatrix} \widehat{A} \\ \widehat{B} \end{pmatrix} = \left(n^{-1}\sum_{t=1}^{n} \mathbf{x}_t'\mathbf{x}_t\right)^{-1} \frac{1}{n}\sum_{t=1}^{n} X_t \begin{pmatrix} \cos(\omega t) \\ \sin(\omega t) \end{pmatrix}$$

where $\mathbf{x}_t = (\cos(\omega t), \sin(\omega t))$. However, because the sine and cosine functions are near orthogonal we have that $n^{-1}\sum_{t=1}^{n} \mathbf{x}_t'\mathbf{x}_t \approx I_2$ and

$$\begin{pmatrix} \widehat{A} \\ \widehat{B} \end{pmatrix} \approx \frac{1}{n}\sum_{t=1}^{n} X_t \begin{pmatrix} \cos(\omega t) \\ \sin(\omega t) \end{pmatrix},$$

which is simple to evaluate! The above argument is not very precise. To make it precise we note that

$$\begin{aligned}
& -2\mathcal{L}_n(A,B,\omega) \\
= & \sum_{t=1}^{n} X_t^2 - 2\sum_{t=1}^{n} X_t\big(A\cos(\omega t) + B\sin(\omega t)\big) \\
& + A^2\sum_{t=1}^{n}\cos^2(\omega t) + B^2\sum_{t=1}^{n}\sin^2(\omega t) + 2AB\sum_{t=1}^{n}\sin(\omega t)\cos(\omega t) \\
= & \sum_{t=1}^{n} X_t^2 - 2\sum_{t=1}^{n} X_t\big(A\cos(\omega t) + B\sin(\omega t)\big) + \\
& \frac{A^2}{2}\sum_{t=1}^{n}(1+\cos(2t\omega)) + \frac{B^2}{2}\sum_{t=1}^{n}(1-\cos(2t\omega)) + AB\sum_{t=1}^{n}\sin(2t\omega) \\
= & \sum_{t=1}^{n} X_t^2 - 2\sum_{t=1}^{n} X_t\big(A\cos(\omega t) + B\sin(\omega t)\big) + \frac{n}{2}(A^2+B^2) + \\
& \frac{(A^2-B^2)}{2}\sum_{t=1}^{n}\cos(2t\omega) + AB\sum_{t=1}^{n}\sin(2t\omega) \\
= & \mathcal{S}_n(A,B,\omega) + \frac{(A^2-B^2)}{2}\sum_{t=1}^{n}\cos(2t\omega) + AB\sum_{t=1}^{n}\sin(2t\omega)
\end{aligned}$$

where

$$\mathcal{S}_n(A,B,\omega) = \sum_{t=1}^{n} X_t^2 - 2\sum_{t=1}^{n} X_t\big(A\cos(\omega t) + B\sin(\omega t)\big) + \frac{n}{2}(A^2+B^2).$$

The important point abut the above is that $n^{-1}\mathcal{S}_n(A,B,\omega)$ is bounded away from zero, *however* $n^{-1}\sum_{t=1}^{n}\sin(2\omega t)$ and $n^{-1}\sum_{t=1}^{n}\cos(2\omega t)$ both converge to zero (at the rate n^{-1}, though it is not uniform over ω); use identity (3.16). Thus $\mathcal{S}_n(A,B,\omega)$ is the dominant term in $\mathcal{L}_n(A,B,\omega)$;

$$-2\mathcal{L}_n(A,B,\omega) \;=\; \mathcal{S}_n(A,B,\omega) + O(1).$$

Thus ignoring the $O(1)$ term and differentiating $\mathcal{S}_n(A,B,\omega)$ wrt A and B (keeping ω fixed) gives the estimators

$$\begin{pmatrix} \widehat{A}(\omega) \\ \widehat{B}(\omega) \end{pmatrix} = \frac{1}{n}\sum_{t=1}^{n} X_t \begin{pmatrix} \cos(\omega t) \\ \sin(\omega t) \end{pmatrix}.$$

Thus we have "profiled out" the nuisance parameters A and B.

Using the approximation $\mathcal{S}_n(\widehat{A}_n(\omega), \widehat{B}_n(\omega), \omega)$ we have

$$\mathcal{L}_n(\widehat{A}_n(\omega), \widehat{B}_n(\omega), \omega) = \frac{-1}{2}\mathcal{S}_p(\omega) + O(1),$$

where

$$\begin{aligned}\mathcal{S}_p(\omega) &= \Big(\sum_{t=1}^n X_t^2 - 2\sum_{t=1}^n X_t\big(\widehat{A}_n(\omega)\cos(\omega t) + \widehat{B}(\omega)\sin(\omega t)\big) + \frac{n}{2}(\widehat{A}_n(\omega)^2 + \widehat{B}(\omega)^2)\Big) \\ &= \Big(\sum_{t=1}^n X_t^2 - \frac{n}{2}\Big[\widehat{A}_n(\omega)^2 + \widehat{B}_n(\omega)^2\Big]\Big).\end{aligned}$$

Thus

$$\begin{aligned}\arg\max \mathcal{L}_n(\widehat{A}_n(\omega), \widehat{B}_n(\omega), \omega) &\approx \arg\max \frac{-1}{2}\mathcal{S}_p(\omega) \\ &= \arg\max \Big[\widehat{A}_n(\omega)^2 + \widehat{B}_n(\omega)^2\Big].\end{aligned}$$

Thus

$$\begin{aligned}\widehat{\omega}_n &= \arg\max_\omega(-1/2)\mathcal{S}_p(\omega) = \arg\max_\omega\big(\widehat{A}_n(\omega)^2 + \widehat{B}_n(\omega)^2\big) \\ &= \arg\max_\omega\big|\sum_{t=1}^n X_t\exp(it\omega)\big|^2,\end{aligned}$$

which is easily evaluated (using a basic grid search).

(iv) Differentiating both sides of (3.15) with respect to Ω and considering the real and imaginary terms gives $\sum_{t=1}^n t\cos(\Omega t) = O(n)$ $\sum_{t=1}^n t\sin(\Omega t) = O(n)$. Differentiating both sides of (3.15) twice wrt to Ω gives the second term.

(v) In order to obtain the rate of convergence of the estimators, $\widehat{\omega}, \widehat{A}(\widehat{\omega}), \widehat{B}(\widehat{\omega})$ we evaluate the Fisher information of $\mathcal{L}_n$ (the inverse of which will give us limiting rate of convergence). For convenience rather than take the second derivative of $\mathcal{L}$ we evaluate the second derivative of $\mathcal{S}_n(A, B, \omega)$ (though, you will find the in the limit both the second derivative of $\mathcal{L}_n$ and $\mathcal{S}_n(A, B, \omega)$ are the same).

Differentiating $\mathcal{S}_n(A, B, \omega) = \big(\sum_{t=1}^n X_t^2 - 2\sum_{t=1}^n X_t\big(A\cos(\omega t) + B\sin(\omega t)\big) + \frac{1}{2}n(A^2 +$

$B^2)\big)$ twice wrt to A, B and ω gives

$$\frac{\partial \mathcal{S}_n}{\partial A} = -2\sum_{t=1}^{n} X_t \cos(\omega t) + An$$
$$\frac{\partial \mathcal{S}_n}{\partial B} = -2\sum_{t=1}^{n} X_t \sin(\omega t) + Bn$$
$$\frac{\partial \mathcal{S}_n}{\partial \omega} = 2\sum_{t=1}^{n} AX_t t \sin(\omega t) - 2\sum_{t=1}^{n} BX_t t \cos(\omega t).$$

and $\frac{\partial^2 \mathcal{S}_n}{\partial A^2} = n$, $\frac{\partial^2 \mathcal{S}_n}{\partial B^2} = n$, $\frac{\partial^2 \mathcal{S}_n}{\partial A \partial B} = 0$,

$$\frac{\partial^2 \mathcal{S}_n}{\partial \omega \partial A} = 2\sum_{t=1}^{n} X_t t \sin(\omega t)$$
$$\frac{\partial^2 \mathcal{S}_n}{\partial \omega \partial B} = -2\sum_{t=1}^{n} X_t t \cos(\omega t)$$
$$\frac{\partial^2 \mathcal{S}_n}{\partial \omega^2} = 2\sum_{t=1}^{n} t^2 X_t \big(A\cos(\omega t) + B\sin(\omega t)\big).$$

Now taking expectations of the above and using (v) we have

$$\begin{aligned} E(\frac{\partial^2 \mathcal{S}_n}{\partial \omega \partial A}) &= 2\sum_{t=1}^{n} t\sin(\omega t)\big(A\cos(\omega t) + B\sin(\omega t)\big) \\ &= 2B\sum_{t=1}^{n} t\sin^2(\omega t) + 2\sum_{t=1}^{n} At\sin(\omega t)\cos(\omega t) \\ &= B\sum_{t=1}^{n} t(1-\cos(2\omega t)) + A\sum_{t=1}^{n} t\sin(2\omega t) = \frac{n(n+1)}{2}B + O(n) = B\frac{n^2}{2} + O(n). \end{aligned}$$

Using a similar argument we can show that $E(\frac{\partial^2 \mathcal{S}_n}{\partial \omega \partial B}) = -A\frac{n^2}{2} + O(n)$ and

$$\begin{aligned} E(\frac{\partial^2 \mathcal{S}_n}{\partial \omega^2}) &= 2\sum_{t=1}^{n} t^2 \bigg(A\cos(\omega t) + B\sin(\omega t)\bigg)^2 \\ &= (A^2+B^2)\frac{n(n+1)(2n+1)}{6} + O(n^2) = (A^2+B^2)n^3/3 + O(n^2). \end{aligned}$$

Since $E(-\nabla^2 \mathcal{L}_n) \approx \frac{1}{2}E(\nabla^2 \mathcal{S}_n)$, this gives the required result.

(vi) Noting that the asymptotic variance for the profile likelihood estimator $\hat{\omega}_n$

$$\bigg(I_{\omega,\omega} - I_{\omega,(AB)}I_{A,B}^{-1}I_{(BA),\omega}\bigg)^{-1},$$

by subsituting (vi) into the above we have

$$2\bigg(\frac{A^2+B^2}{6}n^3+O(n^2)\bigg)^{-1} \approx \frac{12}{(A^2+B^2)n^3}$$

Thus we observe that the asymptotic variance of $\hat{\omega}_n$ is $O(n^{-3})$.

Typically estimators have a variance of order $O(n^{-1})$, so we see that the estimator $\hat{\omega}_n$ converges to to the true parameter, far faster than expected. Thus the estimator is extremely good compared with the majority of parameter estimators.

Exercise 3.2 *Run a simulation study to illustrate the above example.*

Evaluate $I_n(\omega)$ for all $\omega_k = \frac{2\pi k}{n}$ using the `fft` *function in R (this evaluates $\{\sum_{t=1}^n Y_t e^{it\frac{2\pi k}{n}}\}_{k=1}^n$), then take the absolute square of it. Find the maximum over the sequence using the function* `which.max`*. This will estimate $\widehat{\omega}_n$. From this, estimate A and B. However, $\widehat{\omega}_n$ will only estimate ω to $O_p(n^{-1})$, since we have discretized the frequencies. To improve on this, one can use one further iteration see* `http://www.jstor.org/stable/pdf/2334314.pdf` *for the details.*

Run the above over several realisations and evaluate the average squared error.

3.2.2 An application of profiling in survival analysis

This application uses some methods from Survival Analysis which is covered later in this course.

Let T_i denote the survival time of an electrical component (we cover survival functions in Chapter 6.1). Often for each survival time, there are known regressors x_i which are believed to influence the survival time T_i. The survival function is defined as

$$P(T_i > t) = \mathcal{F}_i(t) \quad t \geq 0.$$

It is clear from the definition that what defines a survival function is that $\mathcal{F}_i(t)$ is positive, $\mathcal{F}_i(0) = 1$ and $\mathcal{F}_i(\infty) = 0$. The density is easily derived from the survival function taking the negative derivative; $f_i(t) = -\frac{d\mathcal{F}_i(t)}{dt}$.

To model the influence the regressors have on the survival time, the Cox-proportional hazard model is often used with the exponential distribution as the baseline distribution and $\psi(x_i;\beta)$ is a positive "link" function (typically, we use $\psi(x_i;\beta) = \exp(\beta x_i)$ as the link function). More precisely the survival function of T_i is

$$\mathcal{F}_i(t) = \mathcal{F}_0(t)^{\psi(x_i;\beta)},$$

where $\mathcal{F}_0(t) = \exp(-t/\theta)$. Not all the survival times of the electrical components are observed, and there can arise censoring. Hence we observe $Y_i = \min(T_i, c_i)$, where c_i is the (non-random) censoring time and δ_i, where δ_i is the indicator variable, where $\delta_i = 1$ denotes censoring of the ith component and $\delta_i = 0$ denotes that it is not censored. The parameters β and θ are unknown.

(i) Derive the log-likelihood of $\{(Y_i, \delta_i)\}$.

(ii) Compute the profile likelihood of the regression parameters β, profiling out the baseline parameter θ.

Solution

(i) The survivial function and the density are

$$f_i(t) = \psi(x_i;\beta)\big\{\mathcal{F}_0(t)\big\}^{[\psi(x_i;\beta)-1]} f_0(t) \quad \text{and} \quad \mathcal{F}_i(t) = \mathcal{F}_0(t)^{\psi(x_i;\beta)}.$$

Thus for this example, the logarithm of density and survival function is

$$\begin{aligned}
\log f_i(t) &= \log\psi(x_i;\beta) - \big[\psi(x_i;\beta)-1\big]\mathcal{F}_0(t) + \log f_0(t) \\
&= \log\psi(x_i;\beta) - \big[\psi(x_i;\beta)-1\big]\frac{t}{\theta} - \log\theta - \frac{t}{\theta} \\
\log \mathcal{F}_i(t) &= \psi(x_i;\beta)\log\mathcal{F}_0(t) = -\psi(x_i;\beta)\frac{t}{\theta}.
\end{aligned}$$

Since

$$f_i(y_i,\delta_i) = \begin{cases} f_i(y_i) = \psi(x_i;\beta)\big\{\mathcal{F}_0(y_i)\big\}^{[\psi(x_i;\beta)-1} f_0(t) & \delta_i = 0 \\ \mathcal{F}_i(y_i) = \mathcal{F}_0(t)^{\psi(x_i;\beta)} & \delta_i = 1 \end{cases}$$

the log-likelihood of (β,θ) based on (Y_i,δ_i) is

$$\begin{aligned}
\mathcal{L}_n(\beta,\theta) &= \sum_{i=1}^n (1-\delta_i)\big\{\log\psi(x_i;\beta) + \log f_0(Y_i) + (\psi(x_i;\beta)-1)\log\mathcal{F}_0(Y_i)\big\} + \\
&\quad \sum_{i=1}^n \delta_i\big\{\psi(x_i;\beta)\log\mathcal{F}_0(Y_i)\big\} \\
&= \sum_{i=1}^n (1-\delta_i)\left(\log\psi(x_i;\beta) - \log\theta - \frac{Y_i}{\theta} - (\psi(x_i;\beta)-1)\frac{Y_i}{\theta}\right) \\
&\quad - \sum_{i=1}^n \delta_i\psi(x_i;\beta)\frac{Y_i}{\theta} \\
&= \sum_{i=1}^n (1-\delta_i)\big\{\log\psi(x_i;\beta) - \log\theta\big\} - \sum_{i=1}^n \psi(x_i;\beta)\frac{Y_i}{\theta}
\end{aligned}$$

Differentiating the above wrt β and θ gives

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta} &= \sum_{i=1}^{n}(1-\delta_i)\Big\{\frac{\nabla \psi_\beta(x_i;\beta)}{\psi(x_i;\beta)}\Big\} - \sum_{i=1}^{n}\nabla_\beta \psi(x_i;\beta)\frac{Y_i}{\theta} \\ \frac{\partial \mathcal{L}}{\partial \theta} &= \sum_{i=1}^{n}(1-\delta_i)\Big\{-\frac{1}{\theta}\Big\} + \sum_{i=1}^{n}\psi(x_i;\beta)\frac{Y_i}{\theta^2}\end{aligned}$$

which is not simple to solve.

(ii) Instead we keep β fixed and differentiate the likelihood with respect to θ and equate to zero, this gives

$$\frac{\partial \mathcal{L}_n}{\partial \theta} = \sum_{i=1}^{n}(1-\delta_i)\Big\{-\frac{1}{\theta}\Big\} + \sum_{i=1}\psi(x_i;\beta)\frac{Y_i}{\theta^2}$$

and

$$\widehat{\theta}(\beta) = \frac{\sum_{i=1}^{n}\psi(x_i;\beta)Y_i}{\sum_{i=1}^{n}(1-\delta_i)}.$$

This gives us the best estimator of θ for a given β. Next we find the best estimator of β. The profile likelihood (after profiling out θ) is

$$\ell_P(\beta) = \mathcal{L}_n(\beta, \widehat{\theta}(\beta)) = \sum_{i=1}^{n}(1-\delta_i)\Big\{\log\psi(x_i;\beta) - \log\widehat{\theta}(\beta)\Big\} - \sum_{i=1}^{n}\psi(x_i;\beta)\frac{Y_i}{\widehat{\theta}(\beta)}.$$

Hence to obtain the ML estimator of β we maximise the above with respect to β, this gives us $\widehat{\beta}$. Which in turn gives us the MLE $\widehat{\theta}(\widehat{\beta})$.

3.2.3 An application of profiling in semi-parametric regression

Here we apply the profile "likelihood" (we use inverted commas here because we do not use the likelihood, but least squares instead) to semi-parametric regression. Recently this type of method has been used widely in various semi-parametric models. This application requires a little knowledge of nonparametric regression, which is considered later in this course. Suppose we observe (Y_i, U_i, X_i) where

$$Y_i = \beta X_i + \phi(U_i) + \varepsilon_i,$$

$(X_i, U_i, \varepsilon_i)$ are iid random variables and ϕ is an unknown function. Before analyzing the model we summarize some of its interesting properties:

- When a model does not have a parametric form (i.e. a finite number of parameters cannot describe the model), then we cannot usually obtain the usual $O(n^{-1/2})$ rate. We see in the above model that $\phi(\cdot)$ does not have a parametric form thus we cannot expect than an estimator of it $\sqrt{n}$-consistent.

- The model above contains βX_i which does have a parametric form, can we obtain a $\sqrt{n}$-consistent estimator of β?

The Nadaraya-Watson estimator

Suppose

$$Y_i = \phi(U_i) + \varepsilon_i,$$

where U_i, ε_i are iid random variables. A classical method for estimating $\phi(\cdot)$ is to use the Nadarayan-Watson estimator. This is basically a local least squares estimator of $\phi(u)$. The estimator $\widehat{\phi}_n(u)$ is defined as

$$\widehat{\phi}_n(u) \;=\; \arg\min_a \sum_i \frac{1}{b} W\left(\frac{u-U_i}{b}\right)(Y_i - a)^2 = \frac{\sum_i W_b(u-U_i)Y_i}{\sum_i W_b(u-U_i)}$$

where $W(\cdot)$ is a kernel (think local window function) with $\int W(x)dx = 1$ and $W_b(u) = b^{-1}W(u/b)$ with $b \to 0$ as $n \to \infty$; thus the window gets narrower and more localized as the sample size grows. Dividing by $\sum_i W_b(u-U_i)$ "removes" the clustering in the locations $\{U_i\}$.

Note that the above can also be treated as an estimator of

$$\mathrm{E}\,(Y|U=u) = \int_{\mathbb{R}} y f_{Y|U}(y|u)dy \int_{\mathbb{R}} \frac{y f_{Y,U}(y,u)}{f_U(u)}dy = \phi(u),$$

where we replace $f_{Y,U}$ and f_U with

$$\begin{aligned}
\widehat{f}_{Y,U}(u,y) &= \frac{1}{bn}\sum_{i=1}^{n} \delta_{Y_i}(y) W_b\,(u-U_i) \\
\widehat{f}_U(u) &= \frac{1}{bn}\sum_{i=1}^{n} W_b\,(u-U_i)\,,
\end{aligned}$$

with $\delta_Y(y)$ denoting the Dirac-delta function. Note that the above is true because

$$\begin{aligned}\int_{\mathbb{R}} \frac{\widehat{f}_{Y,U}(y,u)}{\widehat{f}_U(u)}dy &= \frac{1}{\widehat{f}_U(u)}\int_{\mathbb{R}} y\widehat{f}_{Y,U}(y,u)dy \\ &= \frac{1}{\widehat{f}_U(u)}\int_{\mathbb{R}} \frac{1}{bn}\sum_{i=1}^{n} y\delta_{Y_i}(y)W_b\left(u-U_i\right)dy \\ &= \frac{1}{\widehat{f}_U(u)}\frac{1}{bn}\sum_{i=1}^{n} W_b\left(u-U_i\right)\underbrace{\int_{\mathbb{R}} y\delta_{Y_i}(y)dy}_{=Y_i} = \frac{\sum_i W_b(u-U_i)Y_i}{\sum_i W_b(u-U_i)}.\end{aligned}$$

The Nadaraya-Watson estimator is a non-parametric estimator and suffers from a far slower rate of convergence to the non-parametric function than parametric estimators. This rates are usually (depending on the smoothness of ϕ and the density of U)

$$|\widehat{\phi}_n(u)-\phi(u)|^2 = O_p\left(\frac{1}{bn}+b^4\right).$$

Since $b\to 0$, $bn\to\infty$ as $n\to\infty$ we see this is far slower than the parametric rate $O_p(n^{-1/2})$. Heuristically, this is because not all n observations are used to estimate $\phi(\cdot)$ at any particular point u (the number is about bn).

Estimating β using the Nadaraya-Watson estimator and profiling

To estimate β, we first profile out $\phi(\cdot)$ (this is the nuisance parameter), which we estimate as if β were known. In other other words, we suppose that β were known and let

$$Y_i(\beta) = Y_i - \beta X_i = \phi(U_i)+\varepsilon_i,$$

We then estimate $\phi(\cdot)$ using the Nadaraya-Watson estimator, in other words the $\phi(\cdot)$ which minimises the criterion

$$\begin{aligned}\hat{\phi}_\beta(u) &= \arg\min_a \sum_i W_b(u-U_i)(Y_i(\beta)-a)^2 = \frac{\sum_i W_b(u-U_i)Y_i(\beta)}{\sum_i W_b(u-U_i)} \\ &= \frac{\sum_i W_b(u-U_i)Y_i}{\sum_i W_b(u-U_i)} - \beta\frac{\sum_i W_b(u-U_i)X_i}{\sum_i W_b(u-U_i)} \\ &:= G_b(u)-\beta H_b(u), \end{aligned} \tag{3.17}$$

where

$$G_b(u) = \frac{\sum_i W_b(u-U_i)Y_i}{\sum_i W_b(u-U_i)} \quad \text{and} \quad H_b(u) = \frac{\sum_i W_b(u-U_i)X_i}{\sum_i W_b(u-U_i)}.$$

Thus, given β, the estimator of ϕ and the residuals ε_i are

$$\hat{\phi}_\beta(u) = G_b(u) - \beta H_b(u)$$

and

$$\widehat{\varepsilon}_\beta = Y_i - \beta X_i - \hat{\phi}_\beta(U_i).$$

Given the estimated residuals $Y_i - \beta X_i - \hat{\phi}_\beta(U_i)$ we can now use least squares to estimate coefficient β. We define the least squares criterion

$$\begin{aligned}
\mathcal{L}_n(\beta) &= \sum_i \big(Y_i - \beta X_i - \hat{\phi}_\beta(U_i)\big)^2 \\
&= \sum_i \big(Y_i - \beta X_i - G_b(U_i) + \beta H_b(U_i)\big)^2 \\
&= \sum_i \big(Y_i - G_b(U_i) - \beta[X_i - H_b(U_i)]\big)^2.
\end{aligned}$$

Therefore, the least squares estimator of β is

$$\hat{\beta}_{b,T} = \frac{\sum_i [Y_i - G_b(U_i)][X_i - H_b(U_i)]}{\sum_i [X_i - H_b(U_i)]^2}.$$

Using $\beta_{b,T}$ we can then estimate (3.18). We observe how we have the used the principle of profiling to estimate the unknown parameters. There is a large literature on this, including Wahba, Speckman, Carroll, Fan etc. In particular it has been shown that under some conditions on b (as $T \to \infty$), the estimator $\hat{\beta}_{b,T}$ has the usual $\sqrt{n}$ rate of convergence.

It should be mentioned that using random regressors U_i is not necessary. It could be that $U_i = \frac{i}{n}$ (observations lie on a on a grid). In this case $n^{-1}\sum_i W_b(u - i/n) = \frac{1}{nb}\sum_{i=1}^n W(\frac{u-i/n}{b}) = b^{-1}\int W(\frac{u-x}{b})dx + O((bn)^{-1}) = 1 + O((bn)^{-1})$ (with a change of variables). This gives

$$\begin{aligned}
\hat{\phi}_\beta(u) &= \arg\min_a \sum_i W_b(u - \frac{i}{n})(Y_i(\beta) - a)^2 = \frac{\sum_i W_b(u - \frac{i}{n})Y_i(\beta)}{\sum_i W_b(u - \frac{i}{n})} \\
&= \sum_i W_b(u - \frac{i}{n})Y_i - \beta \sum_i W_b(u - U_i)X_i \\
&:= G_b(u) - \beta H_b(u), \qquad (3.18)
\end{aligned}$$

where

$$G_b(u) = \sum_i W_b(u - \frac{i}{n})Y_i \quad \text{and} \quad H_b(u) = \sum_i W_b(u - \frac{i}{n})X_i.$$

Using the above estimator of $\phi(\cdot)$ we continue as before.

Chapter 4

Non-standard inference

As we mentioned in Chapter 2 the the log-likelihood ratio statistic is useful in the context of statistical testing because typically it is "pivotal" (does not depend on any nuisance) under the null hypothesis. Typically, the log-likelihood ratio statistic follows a chi-square distribution under the null hypothesis. However, there are realistic situations where the this statistic does not follow a chi-square distribution and the purpose of this chapter is to consider some of these cases.

At the end of this chapter we consider what happens when the "regularity" conditions are not satisfied.

4.1 Detection of change points

This example is given in Davison (2004), pages 141, and will be considered in class. It is not related to the boundary problems discussed below but none the less is very interesting.

4.2 Estimation on the boundary of the parameter space

In this section we consider the distribution of parameters which are estimated on the boundary of the parameter space. We will use results from Chapter 2.

4.2.1 Estimating the mean on the boundary

There are situations where the parameter to be estimated lies on the boundary (or very, very close to it). In such cases the limiting distribution of the the parameter may not be normal (since when we maximise the likelihood we do so over the parameter space and not outside it). This will not impact Wald based tests (by much), but it will have an impact on the log-likelihood ratio test.

To understand the changes involved, we start with a simple example.

Suppose $X_i \sim \mathcal{N}(\mu, 1)$, where the mean μ is unknown. In addition it is known that the mean is non-negative hence the parameter space of the mean is $\Theta = [0, \infty)$. In this case $\bar{X}$ can no longer be the MLE because there will be some instances where $\bar{X} < 0$. Let us relook at the maximum likelihood on the restricted parameter space

$$\hat{\mu}_n = \arg\max_{\mu \in \Theta} \mathcal{L}_n(\mu) = \arg\max_{\mu \in \Theta} \frac{-1}{2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Since $\mathcal{L}_n(\mu)$ is concave over μ, we see that the MLE estimator is

$$\hat{\mu}_n = \begin{cases} \bar{X} & \bar{X} \geq 0 \\ 0 & \bar{X} < 0. \end{cases}$$

Hence in this restricted space it is not necessarily true that $\frac{\partial \mathcal{L}_n(\mu)}{\partial \mu} \rfloor_{\hat{\mu}_n} \neq 0$, and the usual Taylor expansion method cannot be used to derive normality. Indeed we will show that it is not normal.

We recall that $\sqrt{n}(\bar{X} - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\mu)^{-1})$ or equivalently $\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\mu)}{\partial \mu} \rfloor_{\bar{X}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\mu))$. Hence if the true parameter $\mu = 0$, then approximately half the time $\bar{X}$ will be less than zero and the other half it will be greater than zero. This means that half the time $\hat{\mu}_n = 0$ and the other half it will be greater than zero. Therefore the distribution function of $\hat{\mu}_n$ is

$$\begin{aligned} & P(\sqrt{n}\widehat{\mu}_n \leq x) = P(\sqrt{n}\hat{\mu}_n = 0 \text{ or } 0 < \sqrt{n}\hat{\mu}_n \leq x) \\ \approx & \begin{cases} 0 & x \leq 0 \\ 1/2 & x = 0 \\ 1/2 + P(0 < \sqrt{n}\bar{X} \leq x) = \Phi(\sqrt{n}\bar{X} \leq x) & x > 0 \end{cases}, \end{aligned}$$

where Φ denotes the distribution function of the normal distribution. Observe the distribution of $\sqrt{n}\bar{X}$ is a mixture of a point mass and a density. However, this result does not

change our testing methodology based on the sample mean. For example, if we want to test $H_0 : \mu = 0$ vs $H_A : \mu > 0$, the parameter space is $[0, \infty)$, thus we use the estimator $\widehat{\mu}_n$ and the p-value is

$$1 - \Phi\left(\sqrt{n}\widehat{\mu}_n\right),$$

which is the p-value corresponding to the one-sided test (using the normal distribution).

Now we consider using the log-likelihood ratio statistics to test the $H_0 : \mu = 0$ vs $H_A : \mu > 0$ (parameter space is $[0, \infty)$). In this set-up the test statistic is

$$W_n = 2\left\{\arg\max_{\mu\in[0,\infty)} \mathcal{L}_n(\mu) - \mathcal{L}_n(0)\right\} = 2\big\{\mathcal{L}_n(\widehat{\mu}_n) - \mathcal{L}_n(0)\big\}.$$

However, since the derivative of the likelihood at $\widehat{\mu}_n$ is not necessarily zero, means that W will not be a standard chi-square distribtion. To obtain the distribution we note that likelihoods under $\mu \in [0, \infty)$ and $\mu = 0$ can be written as

$$\mathcal{L}_n(\hat{\mu}_n) = -\frac{1}{2}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2 \qquad \mathcal{L}_n(0) = -\frac{1}{2}\sum_{i=1}^{n} X_i^2.$$

Thus we observe that when $\bar{X} \leq 0$ then $\mathcal{L}_n(\hat{\mu}_n) = \mathcal{L}_n(0)$ and

$$2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \;=\; \begin{cases} 0 & \bar{X} \leq 0 \quad P(\bar{X} \leq 0) = 1/2 \\ n|\bar{X}|^2 & \bar{X} > 0 \quad P(\bar{X} > 0) = 1/2 \end{cases},$$

the above probabilities are exact since $\bar{X}$ is normally distributed. Hence we have that

$$\begin{aligned} & P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x) \\ = \;& P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x\big|\bar{X} \leq 0)P(\bar{X} \leq 0) + P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x\big|\bar{X} > 0)P(\bar{X} > 0). \end{aligned}$$

Now using that

$$P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x\big|\bar{X} \leq 0) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases},$$

$$P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x\big|\bar{X} > 0) = P(n\bar{X}^2 \leq x\big|\bar{X} > 0) = \begin{cases} 0 & x < 0 \\ \chi_1^2 & x > 0 \end{cases}$$

and $P(\sqrt{n}\bar{X} < 0) = 1/2$, gives

$$P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x) = \begin{cases} 0 & x \leq 0 \\ 1/2 & x = 0 \\ 1/2 + \frac{1}{2}P(n|\bar{X}|^2 \leq x) & x > 0 \end{cases}$$

Therefore

$$P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x) = \frac{1}{2} + \frac{1}{2}P(\chi^2 \leq x) = \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2,$$

where we use the χ_0^2 notation to denote the point mass at zero. Therefore, suppose we want to test the hypothesis $H_0 : \mu = 0$ against the hypothesis $H_A : \mu > 0$ using the log likelihood ratio test. We would evaluate $W_n = 2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\}$ and find the p such that

$$\frac{1}{2} + \frac{1}{2}P(W_n \leq \chi_1^2) = 1 - p.$$

This is the p-value, which we then use to make the decision on the test.

Remark 4.2.1 *Essentially what has been done is turned the log-likelihood test for the mean, which is a two-sided test, into a one-sided test.*

(i) It is clear that without a boundary testing $H_0 : \mu = 0$ against $H_A : \mu \neq 0$ the LLRT is simply

$$2\big\{\mathcal{L}_n(\bar{X}) - \mathcal{L}_n(0)\big\} = n|\bar{X}|^2 \xrightarrow{\mathcal{D}} \chi_1^2,$$

under the null.

Example, $n = 10$ and $\bar{x} = 0.65$ the p-value for the above hypothesis is

$$\begin{aligned} P\left(W_n > 10 \times (0.65)^2\right) &= P\left(\chi_1^2 > 10 \times (0.65)^2\right) \\ &= 1 - P(\chi_1^2 \leq 4.2) = 1 - 0.96 = 0.04. \end{aligned}$$

The p-value is 4%.

(ii) On the other hand, to test $H_0 : \mu = 0$ against the hypothesis $H_A : \mu > 0$ we use

$$2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \xrightarrow{\mathcal{D}} \frac{1}{2} + \frac{1}{2}\chi_1^2.$$

Example: Using the same data, but the one-sided test we have

$$\begin{aligned} P\left(W_n > 10 \times (0.65)^2\right) &= 1 - P\left(W_n \leq 10 \times (0.65)^2\right) \\ &= 1 - \left(\frac{1}{2} + \frac{1}{2}P\left(\chi_1^2 \leq 10 \times (0.65)^2\right)\right) = \frac{1}{2}\left(1 - P(\chi_1^2 \leq 4.2)\right) = 0.02. \end{aligned}$$

The p-value is 2%. Thus, as we would expect, the result of the one-sided test simply gives half the p-value corresponding to the two-sided test.

Exercise 4.1 *The survivial time of disease A follow an exponential distribution, where the distribution function has the form $f(x) = \lambda^{-1}\exp(-x/\lambda)$. Suppose that it is known that at least one third of all people who have disease A survive for more than 2 years.*

(i) Based on the above information obtain the appropriate parameter space for λ. Let λ_B denote the lower boundary of the parameter space and $\Theta = [\lambda_B, \infty)$ and the corresponding parameter space.

(ii) What is the maximum likelihood estimator of $\widehat{\lambda}_n = \arg\max_{\lambda\in\Theta}\mathcal{L}_n(\lambda)$.

(iii) Derive the sampling properties of maximum likelihood estimator of λ, for the cases $\lambda = \lambda_B$ and $\lambda > \lambda_B$.

(iv) Suppose the true parameter is λ_B derive the distribution of $2[\max_{\theta\in\Theta}\mathcal{L}_n(\lambda) - \mathcal{L}_n(\lambda_B)]$.

4.2.2 General case with parameter on the boundary

It was straightfoward to derive the distributions in the above examples because a closed form expression exists for the estimator. However the same result holds for general maximum likelihood estimators; *so long as certain regularity conditions are satisfied.*

Suppose that the log-likelihood is $\mathcal{L}_n(\theta)$, the parameter space is $[0, \infty)$ and

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta}\mathcal{L}_n(\theta).$$

We consider the case that the true parameter $\theta_0 = 0$. To derive the limiting distribution we extend the parameter space $\widetilde{\Theta}$ such that $\theta_0 = 0$ is an *interior point* of $\widetilde{\Theta}$. Let

$$\widetilde{\theta}_n \in \arg\max_{\theta\in\widetilde{\Theta}}\mathcal{L}_n(\theta),$$

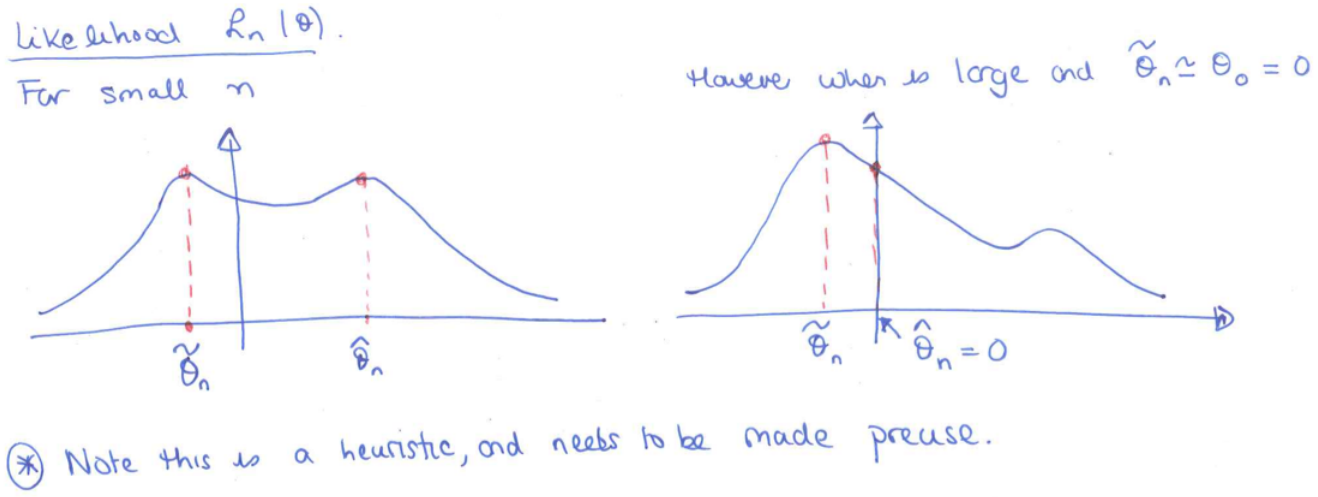

Figure 4.1: A plot of the likelihood for large and small n. For large n, the likelihood tends to be concave about the true parameter, which in this case is zero. This means that if the true parameter is $\theta = 0$, then for large enough n, there is a 50% chance $\widetilde{\theta}_n$ is less than zero and 50% chance $\widehat{\theta}_n$ that greater than zero.

this is the maximum likelihood estimator in the non-constrained parameter space. We assume that for this non-constrained estimator $\sqrt{n}(\widetilde{\theta}_n - 0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(0)^{-1})$ (this needs to be verified and may not always hold). This means that for sufficiently large n, the likelihood will have a maximum close to 0 and that in the neighbourhood of zero, the likelihood is concave (with only one maximum). We use this result to obtain the distribution of the restricted estimator. The log-likelihood ratio involving the restricted estimator is

$$\begin{aligned} W_n &= 2\left(\arg_{\theta\in[0,\infty)} \mathcal{L}_n(\theta) - \mathcal{L}_n(0)\right) \\ &= 2\left(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(0)\right). \end{aligned}$$

Roughly speaking $\widehat{\theta}_n$ can be considered as a "reflection" of $\widetilde{\theta}_n$ i.e. if $\widetilde{\theta}_n < 0$ then $\widehat{\theta}_n = 0$ else $\widehat{\theta}_n = \widetilde{\theta}_n$ (see Figure 4.1) (since for a sufficiently large sample size, if $\widetilde{\theta}_n < 0$, then the maximum within $[0, \infty)$ will lie at zero). We use this principle to obtain the distribution of W_n by conditioning on $\widetilde{\theta}_n$

$$P\left(W_n \leq x\right) = P(W_n \leq x | \widetilde{\theta}_n \leq 0) P(\widetilde{\theta}_n \leq 0) + P(W_n \leq x | \widetilde{\theta}_n > 0) P(\widetilde{\theta}_n > 0).$$

Now using that $\sqrt{n}\widetilde{\theta}_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(0)^{-1})$ and that $\mathcal{L}_n(\theta)$ is close to concave about its maximum thus for $\widetilde{\theta}_n \leq 0$ we have $W_n = 0$, and we have a result analogous to the mean

case

$$P\left(W_n \leq x\right) \;\; = \;\; \frac{1}{2}P(W_n \leq x|\widetilde{\theta}_n \leq 0) + \frac{1}{2}P(W_n \leq x|\widetilde{\theta}_n > 0) = \frac{1}{2} + \frac{1}{2}\chi^2_1.$$

The precise argument for the above uses a result by Chernoff (1954), who shows that

$$W_n \stackrel{\mathcal{D}}{=} \max_{\theta\in[0,\infty)} \left[-(Z-\theta)I(0)(Z-\theta)\right] + ZI(0)Z + o_p(1), \tag{4.1}$$

where $Z \sim \mathcal{N}(0, I(0)^{-1})$ (and is the same for both quadratic forms). Observe that when $Z < 0$ the above is zero, whereas when $Z > 0$ $\max_{\theta\in[0,\infty)} \left[-(Z-\theta)I(0)(Z-\theta)\right] = 0$ and we have the usual chi-square statistic.

To understand the approximation in (4.1) we return to the log-likelihood ratio and add and subtract the maximum likelihood estimator based on the non-restricted parameter space $\widetilde{\Theta}$

$$2\left[\max_{\theta\in\Theta} \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\right] = 2\left[\max_{\theta\in\Theta} \mathcal{L}_n(\theta) - \max_{\theta\in\widetilde{\Theta}} \mathcal{L}_n(\theta)\right] + 2\left[\max_{\theta\in\widetilde{\Theta}} \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\right] \tag{4.2}$$

Now we do the usual Taylor expansion about $\widetilde{\theta}_n$ (which guarantees that the first derivative is zero) for both terms to give

$$\begin{aligned}
& 2\left[\max_{\theta\in\Theta} \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\right] \\
= \;& -n\left(\widetilde{\theta}_n - \widehat{\theta}_n\right) I(\theta_0) \left(\widetilde{\theta}_n - \widehat{\theta}_n\right) + n\left(\widetilde{\theta}_n - \theta_0\right) I(\theta_0) \left(\widetilde{\theta}_n - \theta_0\right) + o_p(1) \\
= \;& -n\left(\left[\widetilde{\theta}_n - \theta_0\right] - \left[\widehat{\theta}_n - \theta_0\right]\right) I(\theta_0) \left(\left[\widetilde{\theta}_n - \theta_0\right] - \left[\widehat{\theta}_n - \theta_0\right]\right) + n\left(\widetilde{\theta}_n - \theta_0\right) I(\theta_0) \left(\widetilde{\theta}_n - \theta_0\right).
\end{aligned}$$

We recall that asymptotically $\sqrt{n}\left(\widetilde{\theta}_n - \theta_0\right) \sim \mathcal{N}(0, I(\theta_0)^{-1})$. Therefore we define the random variable $\sqrt{n}\left(\widetilde{\theta}_n - \theta_0\right) \sim Z \sim \mathcal{N}(0, I(\theta_0)^{-1})$ and replace this in the above to give

$$\begin{aligned}
& 2\left[\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0)\right] \\
\stackrel{\mathcal{D}}{=} \;& -\left(Z - n^{1/2}\left[\widehat{\theta}_n - \theta_0\right]\right) I(\theta_0) \left(Z - n^{1/2}\left[\widehat{\theta}_n - \theta_0\right]\right) + ZI(\theta_0)Z.
\end{aligned}$$

Finally, it can be shown (see, for example, Self and Liang (1987), Theorem 2 or Andrews (1999), Section 4.1) that $\sqrt{n}(\widehat{\theta}_n - \theta_0) \in \Theta - \theta_0 = \Lambda$, where Λ is a convex cone about θ_0 (this is the terminology that is often used); in the case that $\Theta = [0, \infty)$ and $\theta_0 = 0$ then $\sqrt{n}(\widehat{\theta}_n - \theta_0) \in \Lambda = [0, \infty)$ (the difference can never be negative). And that the maximum

likelihood estimator is equivalent to the maximum of the quadratic form over Θ i.e.

$$\begin{aligned} 2\left[\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0)\right] &\stackrel{\mathcal{D}}{=} \max_{\theta\in\Theta} -\left(Z - n^{1/2}\left[\widehat{\theta}_n - \theta_0\right]\right) I(\theta_0)\left(Z - n^{1/2}\left[\widehat{\theta}_n - \theta_0\right]\right) + ZI(\theta_0)Z \\ &= \max_{\theta\in\Theta-\theta_0=[0,\infty)=\Theta} -(Z-\theta)\, I(\theta_0)\,(Z-\theta) + ZI(\theta_0)Z, \end{aligned}$$

which gives (4.1).

Example 4.2.1 (Example 4.39 (page 140) in Davison (2002)) *In this example Davison reparameterises the t-distribution. It is well known that if the number of degrees of freedom of a t-distribution is one, it is the Cauchy distribution, which has extremely thick tails (such that the mean does not exist). At the other extreme, if we let the number of degrees of freedom tend to ∞, then the limit is a normal distribution (where all moments exist). In this example, the t-distribution is reparameterised as*

$$f(y;\mu,\sigma^2,\psi) = \frac{\Gamma\left[\frac{(1+\psi^{-1})}{2}\right]\psi^{1/2}}{(\sigma^2\pi)^{1/2}\Gamma(\frac{1}{2\pi})}\left(1+\frac{\psi(y-\mu)^2}{\sigma^2}\right)^{-(\psi^{-1}+1)/2}$$

It can be shown that $\lim_{\psi\to 1} f(y;\mu,\sigma^2,\psi)$ is a t-distribution with one-degree of freedom and at the other end of the spectrum $\lim_{\psi\to 0} f(y;\mu,\sigma^2,\psi)$ is a normal distribution. Thus $0<\psi\leq 1$, and the above generalisation allows for fractional orders of the t-distribution.

In this example it is assumed that the random variables $\{X_i\}$ have the density $f(y;\mu,\sigma^2,\psi)$, and our objective is to estimate ψ, when $\psi\to 0$, this the true parameter is on the boundary of the parameter space $(0,1]$ (it is just outside it!). Using similar, arguments to those given above, Davison shows that the limiting distribution of the MLE estimator is close to a mixture of distributions (as in the above example).

Testing on the boundary in the presence of independent nuisance parameters

Suppose that the iid random variables come from the distribution $f(x;\theta,\psi)$, where (θ,ψ) are unknown. We will suppose that θ is a univariate random variable and ψ can be multivariate. Suppose we want to test $H_0:\theta=0$ vs $H_A:\theta>0$. In this example we are testing on the boundary in the presence of nuisance parameters ψ.

Example 4.2.2 *Examples include the random coefficient regression model*

$$Y_i = (\alpha+\eta_i)X_i + \varepsilon_i, \tag{4.3}$$

where $\{(Y_i, X_i)\}_{i=1}^n$ *are observed variables.* $\{(\eta_i, \varepsilon_i)\}_{i=1}^n$ *are independent zero mean random vector, where* $var((\eta_i, \varepsilon_i)) = \text{diag}(\sigma_\eta^2, \sigma_\varepsilon^2)$. *We may want to test whether the underlying model is a classical regression model of the type*

$$Y_i = \alpha X_i + \varepsilon_i,$$

vs the random regression model in (4.3). This reduces to testing $H_0 : \sigma_\eta^2 = 0$ *vs* $H_A : \sigma_\eta^2 > 0$.

In this section we will assume that the Fisher information matrix associated for the mle of (θ, ψ) is block diagonal i.e. $\text{diag}(I(\theta), I(\psi))$. In other words, if we did not constrain the parameter space in the maximum likelihood estimation then

$$\sqrt{n} \begin{pmatrix} \widetilde{\theta}_n - \theta \\ \widetilde{\psi}_n - \psi \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \text{diag}(I(\theta), I(\psi))\right).$$

The log-likelihood ratio statistic for testing the hypothesis is

$$W_n = 2\left[\max_{\theta \in [0,\infty), \psi} \mathcal{L}_n(\theta, \psi) - \max_{\psi} \mathcal{L}_n(0, \psi)\right]$$

Now using the heuristics presented in the previous section we have

$$P\left(W_n \leq x\right) = P(W_n \leq x | \widetilde{\theta}_n \leq 0) P(\widetilde{\theta}_n \leq 0) + P(W_n \leq x | \widetilde{\theta}_n > 0) P(\widetilde{\theta}_n > 0).$$

The important observation is that because $(\widetilde{\theta}_n, \widetilde{\psi}_n)$ are asymptotically independent of each other, the estimator of $\widetilde{\theta}_n$ has no influence on the estimate of $\widetilde{\psi}_n$. Thus the estimator of ψ conditional on $\widehat{\theta}_n = 0$ will not change the estimator of ψ thus

$$\begin{aligned} & 2\left[\max_{\theta \in [0,\infty), \psi} \mathcal{L}_n(\theta, \psi) - \max_{\psi} \mathcal{L}_n(0, \psi)\right] \Big| \widetilde{\theta}_n < 0 \\ = \; & 2\left[\max_{\theta \in [0,\infty), \psi} \mathcal{L}_n(\widehat{\theta}, \widehat{\psi}) - \max_{\psi} \mathcal{L}_n(0, \psi)\right] \Big| \widetilde{\theta}_n < 0 \\ = \; & 2\left[\mathcal{L}_n(0, \widetilde{\psi}) - \max_{\psi} \mathcal{L}_n(0, \psi)\right] = 0. \end{aligned}$$

This gives the result

$$P\left(W_n \leq x\right) \;=\; \frac{1}{2} P(W_n \leq x | \widetilde{\theta}_n \leq 0) + \frac{1}{2} P(W_n \leq x | \widetilde{\theta}_n > 0) = \frac{1}{2} + \frac{1}{2}\chi_1^2.$$

4.2.3 Estimation on the boundary with several parameters when the Fisher information is block diagonal

In the following section we summarize some of the results in Self and Liang (1987).

One parameter lies on the boundary and the rest do not

We now generalize the above to estimating the parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_{p+1})$. We start by using an analogous argument to that used in the mean case and then state the precise result from which it comes from.

Suppose the true parameter θ_1 lies on the boundary, say zero, however the other parameters $\theta_2, \ldots, \theta_{p+1}$ lie within the interior of the parameter space and the parameter space is denoted as Θ. Examples include mixture models where θ_1 is the variance (and cannot be negative!). We denote the true parameters as $\theta_0 = (\theta_{10} = 0, \theta_{20}, \ldots, \theta_{p+1,0})$. Let $\mathcal{L}_n(\theta)$ denote the log-likelihood. We make the informal assumption that if we were to extend the parameter space such that $\theta_0 = 0$ were in the interior of this new parameter space $\widetilde{\Theta}$ i.e. $(\theta_{10} = 0, \theta_{20}, \ldots, \theta_{p+1,0}) = (\theta_{10}, \underline{\theta}_p) \in int(\widetilde{\Theta})$, and $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\underline{\theta}}_p) = \arg\max_{\theta \in \widetilde{\Theta}} \mathcal{L}_n(\theta)$ then

$$\sqrt{n}\begin{pmatrix} \widetilde{\theta}_{1n} - \theta_0 \\ \widetilde{\underline{\theta}}_{pn} - \underline{\theta}_{p0} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} I_{11}(\theta_0) & 0 \\ 0 & I_{pp}(\theta_0) \end{pmatrix}^{-1}\right).$$

It is worth noting that the block diagonal nature of the information matrix assumes that the two sets of parameters are asymptotically independent. The asymptotic normality results needs to be checked; it does not always hold.[1]. Let $\widehat{\theta}_n = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta)$ denote the maximum likelihood estimator in the restricted parameter space (with the cut off at zero). Our aim is to derive the distribution of

$$W_n = 2\left(\max_{\theta \in \Theta} \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\right).$$

We use heuristics to obtain the distribution, by conditioning on the unrestricted estimator $\widetilde{\theta}_n$ (we make this a little more precisely later on). Conditioning on $\widetilde{\theta}_{1n}$ we have

$$P\left(W_n \leq x\right) = P(W_n \leq x | \widetilde{\theta}_{1n} \leq 0) P(\widetilde{\theta}_{1n} \leq 0) + P(W_n \leq x | \widetilde{\theta}_{1n} > 0) P(\widetilde{\theta}_{1n} > 0).$$

[1] Sometimes we cannot estimate on the boundary (consider some of the example considered in Chapter 2.9 with regards to the exponential family), sometimes the $\sqrt{n}$-rates and/or the normality result is completely different for parameters which are defined at the boundary (the Dickey-Fuller test is a notable example)

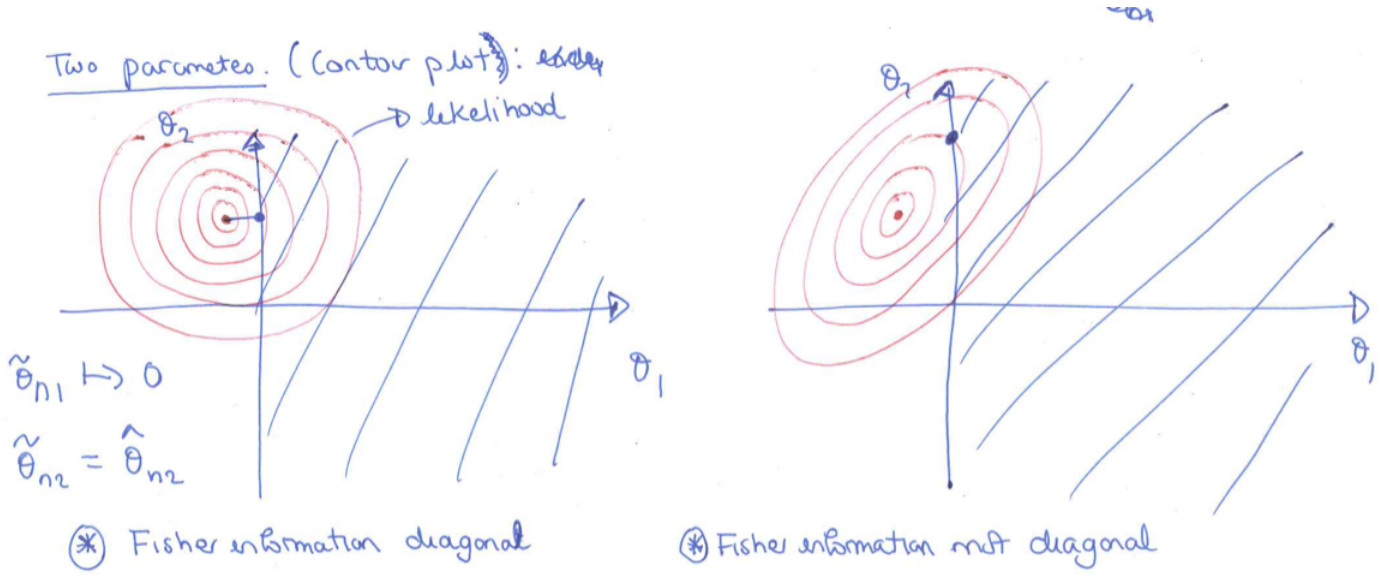

Figure 4.2: The likelihood for diagonal and nondiagonal Fisher information matrices.

Again assuming that for large n, $\mathcal{L}_n(\theta)$ is concave about $\widetilde{\theta}_n$ such that when $\widetilde{\theta}_n < 0$, $\widehat{\theta}_n = 0$. However, asymptotic independence between $\widetilde{\theta}_{n1}$ and $\widetilde{\theta}_{np}$ (since the Fisher information matrix is block diagonal) means that setting $\widehat{\theta}_{n1} = 0$ does not change the estimator of θ_p $\widetilde{\theta}_{np}$ i.e. roughly speaking

$$2[\mathcal{L}_n(\widehat{\theta}_{1n}, \widehat{\theta}_{pn}) - \mathcal{L}_n(0, \theta_p)]|\widetilde{\theta}_{n2} < 0 = \underbrace{2[\mathcal{L}_n(0, \widetilde{\theta}_{pn}) - \mathcal{L}_n(0, \theta_p)]}_{\chi^2_p}$$

If $\widetilde{\theta}_{n1}$ and $\widetilde{\theta}_{np}$ were dependent then the above equality does not hold and it is not a χ^2_p (see Figure 4.2). The above gives

$$\begin{aligned} P(W_n \leq x) &= P(\underbrace{W_n \leq x}_{\chi^2_p}|\widetilde{\theta}_{1n} \leq 0)P(\widetilde{\theta}_{1n} \leq 0) + P(\underbrace{W_n \leq x}_{\chi^2_{p+1}}|\widetilde{\theta}_{1n} > 0)P(\widetilde{\theta}_{1n} > 0) \\ &= \frac{1}{2}\chi^2_p + \frac{1}{2}\chi^2_{p+1}. \end{aligned} \tag{4.4}$$

See Figure 4.3 for a plot of the parameter space and associated probabilities.

The above is a heuristic argument. If one wanted to do it precisely one needs to use the asymptotic equivalent (based on the same derivations given in(4.2)) where (under

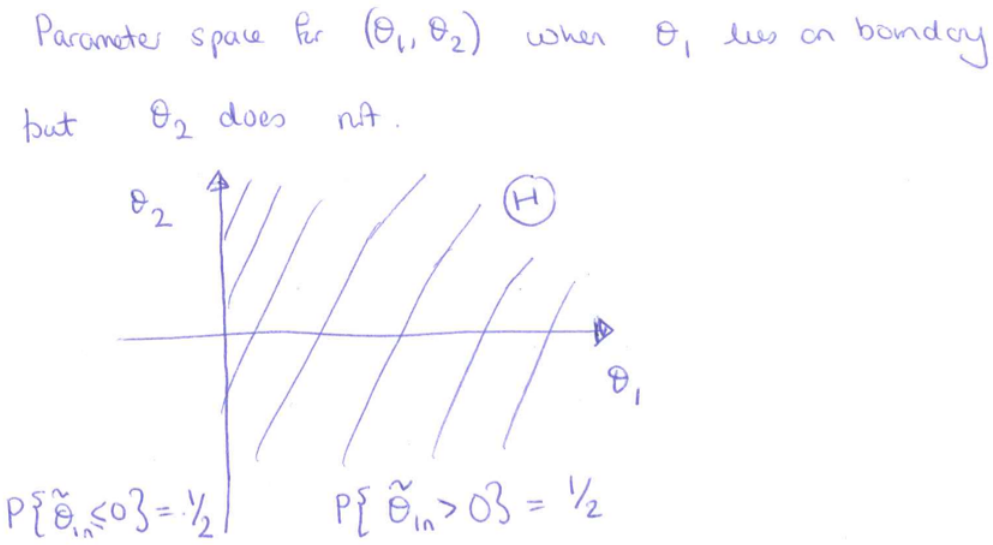

Figure 4.3: Two parameters: one on boundary and one in interior.

certain regularity conditions) we have

$$\begin{aligned} W_n &\stackrel{\mathcal{D}}{=} \max_{\theta \in \Theta} [-(Z-\theta) I(0) (Z-\theta)] + Z I(\theta_0) Z + o_p(1) \\ &= \max_{\theta_1 \in [0,\infty)} [-(Z-\theta_1) I_{11}(\theta_0)(Z-\theta_1)] + Z I_{11}(\theta_0) Z \\ &\quad + \underbrace{\max_{\underline{\theta}_p \in \Theta_p} \left[-(\underline{Z}_p - \underline{\theta}_p) I_{11}(\theta_0)(\underline{Z}_p - \underline{\theta}_p)\right]}_{=0} + \underline{Z}_p I_{pp}(\theta_0) \underline{Z}_p \\ &= \max_{\theta_1 \in [0,\infty)} [-(Z-\theta_1) I_{11}(\theta_0)(Z-\theta_1)] + Z I_{11}(\theta_0) Z \\ &\quad + \underline{Z}_p I_{pp}(\theta_0) \underline{Z}_p \end{aligned}$$

where $Z \sim N(0, I_{11}(\theta_0)^{-1})$ and $\underline{Z}_p \sim N(0, I_{pp}(\theta_0)^{-1})$ (Z and $\underline{Z}_p$ are independent). Using the above we can obtain the same distribution as that given in (4.4)

More than one parameter lies on the boundary

Suppose that the parameter space is $\Theta = [0, \infty) \times [0, \infty)$ and the true parameter $\theta_0 = (\theta_{10}, \theta_{20}) = (0, 0)$ (thus is on the boundary). As before we make the informal assumption that we can extend the parameter space such that θ_0 lies within its interior of $\widetilde{\Theta}$. In this

extended parameter space we have

$$\sqrt{n}\begin{pmatrix} \widetilde{\theta}_1 - \theta_{10} \\ \widetilde{\theta}_2 - \theta_{20} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} I_{11}(\theta_0) & 0 \\ 0 & I_{22}(\theta_0) \end{pmatrix}^{-1}\right).$$

In order to derive the limiting distribution of the log-likelihood ratio statistic

$$W_n = 2\left(\max_{\theta\in\Theta} \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\right)$$

we condition on $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$. This gives

$$\begin{aligned} & P\left(W_n \leq x\right) \\ = \; & P\left(W_n \leq x | \widetilde{\theta}_1 \leq 0, \widetilde{\theta}_2 \leq 0\right) P\left(\widetilde{\theta}_1 \leq 0, \widetilde{\theta}_2 \leq 0\right) + P\left(W_n \leq x | \widetilde{\theta}_1 \leq 0, \widetilde{\theta}_2 > 0\right) P\left(\widetilde{\theta}_1 \leq 0, \widetilde{\theta}_2 > 0\right) + \\ & P\left(W_n \leq x | \widetilde{\theta}_1 > 0, \widetilde{\theta}_2 \leq 0\right) P\left(\widetilde{\theta}_1 > 0, \widetilde{\theta}_2 \leq 0\right) + P\left(W_n \leq x | \widetilde{\theta}_1 > 0, \widetilde{\theta}_2 > 0\right) P\left(\widetilde{\theta}_1 > 0, \widetilde{\theta}_2 > 0\right). \end{aligned}$$

Now by using the asymptotic independence of $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ and for $\widetilde{\theta}_1 > 0, \widetilde{\theta}_2 > 0$ $W_n = 0$ the above is

$$P(W_n \leq x) = \frac{1}{4} + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2.$$

This is easiest seen in Figure 4.4.

Again the above argument can be made precise by using that the distribution of W_n can be approximated with the quadratic form

$$\begin{aligned} W_n & \stackrel{\mathcal{D}}{=} \max_{\theta_1 \in [0,\infty)} \left[-(Z_1 - \theta_1) I_{11}(0)(Z_1 - \theta_1)\right] + Z_1 I_{11}(0) Z_1 \\ & = + \max_{\theta_2 \in [0,\infty)} \left[-(Z_2 - \theta_2) I_{22}(0)(Z_2 - \theta_2)\right] + Z_2 I_{22}(0) Z_2 \end{aligned}$$

where $Z_1 \sim N(0, I_{11}(\theta_0)^{-1})$ and $Z_2 \sim N(0, I_{22}(\theta_0)^{-1})$. This approximation gives the same result.

4.2.4 Estimation on the boundary when the Fisher information is not block diagonal

In the case that the Fisher information matrix is not block diagonal the same procedure can be use, but the results are no longer so clean. In particular, the limiting distribution

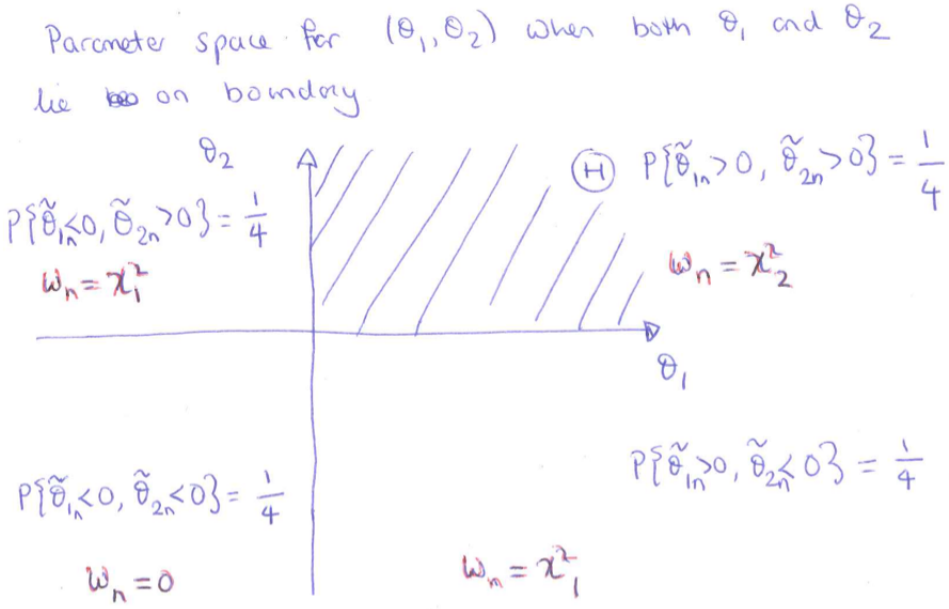

Figure 4.4: Two parameters: both parameters on boundary.

may no longer be a mixture of chi-square distributions and/or the weighting probabilities will depend on the parameter θ (thus the log-likelihood ratio will not be pivotal).

Let us consider the example where one parameter lies on the boundary and the other does not. i.e the parameter space is $[0, \infty) \times (-\infty, \infty)$. The true parameter $\theta_0 = (0, \theta_{20})$ however, unlike the examples considered above the Fisher information matrix is not diagonal. Let $\widehat{\theta}_n = (\widehat{\theta}_1, \widehat{\theta}_2) = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta)$. We can use the conditioning arguments given above however they become ackward because of the dependence between the estimators of $\widehat{\theta}_1$ and $\widehat{\theta}_2$. Instead we use the quadratic form approximation

$$\begin{aligned} W_n &= 2\left(\max_{\theta \in \Theta} \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\right) \\ &\stackrel{\mathcal{D}}{=} \max_{\theta \in \Theta} [-(Z-\theta) I(\theta_0)(Z-\theta)] + Z' I(\theta_0) Z + o_p(1) \end{aligned}$$

where $Z \sim \mathcal{N}(0, I(\theta_0)^{-1})$. To simplify the derivation we let $\overline{Z} \sim \mathcal{N}(0, I_2)$. Then the above

can be written as

$$\begin{aligned} W_n &= 2\left(\max_{\theta\in\Theta}\mathcal{L}_n(\theta)-\mathcal{L}_n(\theta_0)\right) \\ &\stackrel{\mathcal{D}}{=} \max_{\theta\in\Theta}\Big[-\left\{I(\theta_0)^{-1/2}\overline{Z}-I(\theta_0)^{-1/2}I(\theta_0)^{1/2}\theta\right\}' I(\theta_0)\left\{I(\theta_0)^{-1/2}\overline{Z}-I(\theta_0)^{-1/2}I(\theta_0)^{1/2}\theta\right\}\Big] \\ &\quad +\left\{I(\theta_0)^{-1/2}\overline{Z}\right\}' I(\theta_0)\left\{I(\theta_0)^{-1/2}\overline{Z}\right\}+o_p(1) \\ &= \max_{\overline{\theta}\in\overline{\Theta}}\left[-(\overline{Z}-\overline{\theta})'(\overline{Z}-\overline{\theta})\right]+\overline{Z}'\overline{Z}+o_p(1) \end{aligned}$$

where $\overline{\Theta}=\{\overline{\theta}=I(\theta_0)^{1/2}\theta;\theta\in\Theta\}$. This orthogonalisation simplifies the calculations. Using the spectral decomposition of $I(\theta)=P\Lambda P'$ where $P=(\underline{p}_1,\underline{p}_2)$ (thus $I(\theta)^{1/2}=P\Lambda^{1/2}P'$) we see that the half plane (which defines Θ) turns into the rotated half plane $\overline{\Theta}$ which is determined by the eigenvectors $\underline{p}_1$ and $\underline{p}_2$ (which rotates the line $\alpha(0,1)$ into

$$L = \alpha[\lambda_1^{1/2}\langle\underline{p}_1,(0,1)\rangle\underline{p}_1+\lambda_2^{1/2}\langle\underline{p}_2,(0,1)\rangle\underline{p}_2]=\alpha[\lambda_1^{1/2}\langle\underline{p}_1,\underline{1}\rangle\underline{p}_1+\lambda_2^{1/2}\langle\underline{p}_2,\underline{1}\rangle\underline{p}_2]$$

where $\underline{1}=(0,1)$. We observe that

$$W_n=\begin{cases}\underbrace{\overline{Z}'\overline{Z}}_{\chi_2^2} & \overline{Z}\in\overline{\Theta} \\ -[\overline{Z}-P_{\overline{\Theta}}(\overline{Z})]'[\overline{Z}-P_{\overline{\Theta}}(\overline{Z})]+\overline{Z}'\overline{Z} & \overline{Z}\in\overline{\Theta}^c.\end{cases}$$

We note that $P_{\overline{\Theta}}(\overline{Z})$ is the nearest closest point on the line L, thus with some effort one can calculate the distribution of $-[\overline{Z}-P_{\overline{\Theta}}(\overline{Z})]'[\overline{Z}-P_{\overline{\Theta}}(\overline{Z})]+\overline{Z}'\overline{Z}$ (it will be some weighted chi-square), noting that $P(\overline{Z}\in\overline{\Theta})=1/2$ and $P(\overline{Z}\in\overline{\Theta}^c)=1/2$ (since they are both in half a plane). Thus we observe that the above is a mixture of distributions, but they are not as simple (or useful) as when the information matrix has a block diagonal structure.

The precise details can be found in Chernoff (1954), Moran (1971), Chant (1974), Self and Liang (1987) and Andrews (1999). For the Bayesian case see, for example, Botchkina and Green (2014).

Exercise 4.2 *The parameter space of θ is $[0,\infty)\times[0,\infty)$. The Fisher information matrix corresponding to the distribution is*

$$I(\theta)=\begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta)\end{pmatrix}.$$

Suppose that the true parameter is $\theta=(0,0)$ obtain (to the best you can) the limiting distribution of the log-likelihood ratio statistic $2(\arg\max_{\theta\in\Theta}\mathcal{L}_n(\theta)-\mathcal{L}_n(0,0))$.

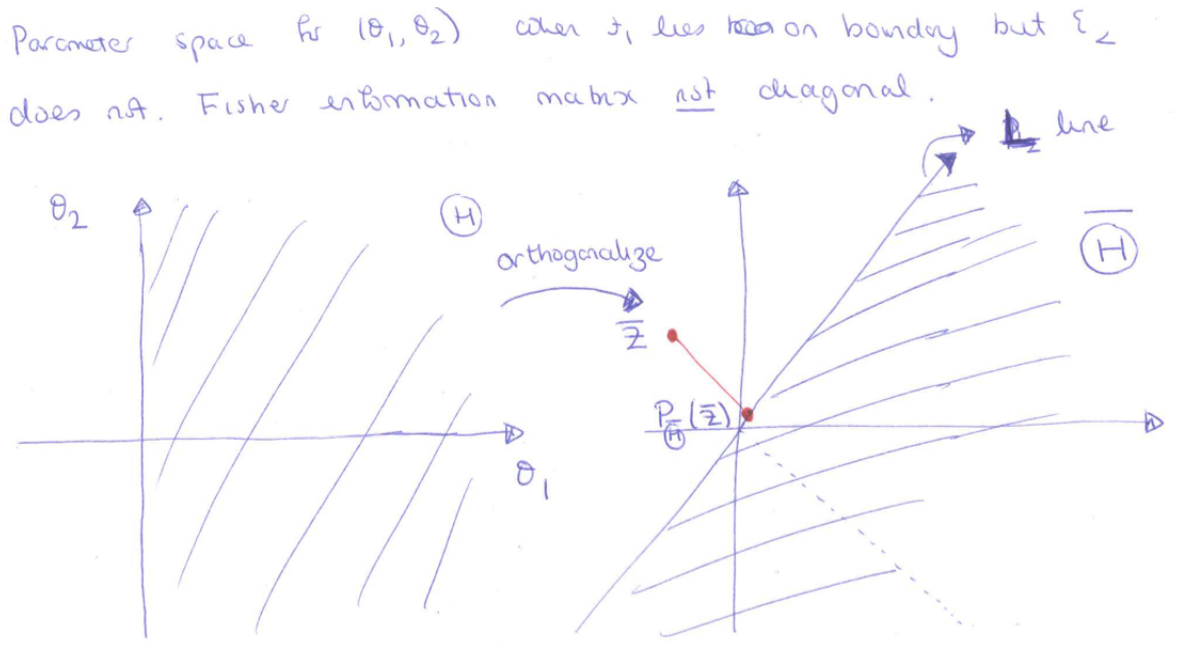

Figure 4.5: Two parameters: one parameter on boundary and the other in interior.

4.3 Regularity conditions which are not satisfied

In this section we consider another aspect of nonstandard inference. Namely, deriving the asymptotic sampling properties of estimators (mainly MLEs) when the usual regularity conditions are not satisfied, thus the results in Chapter 2 do not hold. Some of this material was covered or touched on previously. Here, for completeness, we have collected the results together.

The uniform distribution

The standard example where the regularity conditions (mainly Assumption 1.3.1(ii)) are not satisfied is the uniform distribution

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

We can see that the likelihood in this case is

$$L_n(\underline{X};\theta) = \prod_{i=1}^{n} \theta^{-1} I(0 < X_i < \theta).$$

In this case the the derivative of $L_n(\underline{X};\theta)$ is not well defined, hence we cannot solve for the derivative. Instead, to obtain the mle we try to reason what the maximum is. We should plot $L_n(\underline{X};\theta)$ against θ and place X_i on the θ axis. We can see that if $\theta < X_i$, then L_n is zero. Let $X_{(i)}$ denote the ordered data $X_{(1)} \leq X_{(2)}, \ldots \leq X_{(T)}$. We see that for $\theta = X_{(T)}$, we have $L_n(\underline{X};\theta) = (X_{(T)})^{-T}$, then beyond this point $L_n(\underline{X};\theta)$ decays ie. $L_n(\underline{X};\theta) = \theta^{-T}$ for $\theta \geq X_{(T)}$. Hence the maximum of the likelihood is $\hat{\theta}_n = \max_{1\leq t\leq T} X_i$. The sampling properties of $\widehat{\theta}_n$ were calculated in Exercise 2.3.

The shifted exponential

Let us consider the shifted exponential distribution

$$f(x;\theta,\phi) = \frac{1}{\theta}\exp\left(-\frac{(x-\phi)}{\theta}\right) \quad x \geq \phi,$$

which is only well defined for $\theta, \phi > 0$. We first observe when $\phi = 0$ we have the usual exponential function, ϕ is simply a shift parameter. It can be shown that the usual regularity conditions (Assumption 1.3.1) will not be satisfied. This means the Cramer-Rao bound does not hold in this case and the limiting variance of the mle estimators will not be the inverse of the Fisher information matrix.

The likelihood for this example is

$$L_n(\underline{X};\theta,\phi) = \frac{1}{\theta^n}\prod_{i=1}^{n}\exp\left(-\frac{(X_i-\phi)}{\theta}\right) I(\phi \leq X_i).$$

We see that we cannot obtain the maximum of $L_n(\underline{X};\theta,\phi)$ by differentiating. Instead let us consider what happens to $L_n(\underline{X};\theta,\phi)$ for different values of ϕ. We see that for $\phi > X_i$ for any t, the likelihood is zero. But at $\phi = X_{(1)}$ (smallest value), the likelihood is $\frac{1}{\theta^n}\prod_{i=1}^{n}\exp(-\frac{(X_{(t)}-X_{(1)})}{\theta}$. But for $\phi < X_{(1)}$, $L_n(\underline{X};\theta,\phi)$ starts to decrease because $(X_{(t)} - \phi) > (X_{(t)} - X_{(1)})$, hence the likelihood decreases. Thus the MLE for ϕ is $\widehat{\phi}_n = X_{(1)}$, notice that this estimator is completely independent of θ. To obtain the mle of θ, differentiate and solve $\frac{\partial L_n(\underline{X};\theta,\phi)}{d\theta}\rfloor_{\widehat{\phi}_n = X_{(1)}} = 0$. We obtain $\widehat{\theta}_n = \bar{X} - \widehat{\phi}_n$. For a reality check, we recall that when $\phi = 0$ then the MLE of θ is $\widehat{\theta}_n = \bar{X}$.

We now derive the distribution of $\widehat{\phi}_n - \phi = X_{(1)} - \phi$ (in this case we can actually obtain the finite sample distribution). To make the calculation easier we observe that X_i can be rewritten as $X_i = \phi + E_i$, where $\{E_i\}$ are iid random variables with the

standard exponential distribution starting at zero: $f(x;\theta,0) = \theta^{-1}\exp(-x/\theta)$. Therefore the distribution function of $\widehat{\phi}_n - \phi = \min_i E_i$

$$\begin{aligned} P(\widehat{\phi}_n - \phi \leq x) &= P(\min_i(E_i) \leq x) = 1 - P(\min_i(E_i) > x) \\ &= 1 - [\exp(-x/\theta)]^n. \end{aligned}$$

Therefore the density of $\widehat{\phi}_n - \phi$ is $\frac{\theta}{n}\exp(-nx/\theta)$, which is an exponential with parameter n/θ. Using this, we observe that the mean of $\widehat{\phi}_n - \phi$ is θ/n and the variance is θ^2/n^2. In this case when we standardize $(\widehat{\phi}_n - \phi)$ we need to do so with n (and not the classical $\sqrt{n}$). When we do this we observe that the distribution of $n(\widehat{\phi}_n - \phi)$ is exponential with parameter θ^{-1} (since the sum of n iid exponentials with parameter θ^{-1} is exponential with parameter $n\theta^{-1}$).

In summary, we observe that $\widehat{\phi}_n$ is a biased estimator of ϕ, but the bias decreases as $n \to \infty$. Morover, the variance is quite amazing. Unlike standard estimators where the variance decreases at the rate $1/n$, the variance of $\widehat{\phi}_n$ decreases at the rate $1/n^2$.

Even though $\widehat{\phi}_n$ behaves in a nonstandard way, the estimator $\widehat{\theta}_n$ is completely standard. If ϕ were known then the regularity conditions are satisfied. Furthermore, since $[\widehat{\phi}_n - \phi] = O_p(n^{-1})$ then the difference between the likelihoods with known and estimated ϕ are almost the same; i.e. $\mathcal{L}_n(\theta,\phi) \approx \mathcal{L}_n(\theta,\widehat{\phi}_n)$. Therefore the sampling properties of $\widehat{\theta}_n$ are asymptotically equivalent to the sampling properties of the MLE if ϕ were known.

See Davison (2002), page 145, example 4.43 for more details.

Note that in many problems in inference one replaces the observed likelihood with the unobserved likelihood and show that the difference is "asymptotically negligible". If this can be shown then the sampling properties of estimators involving the observed and unobserved likelihoods are asymptotically equivalent.

Example 4.3.1 *Let us suppose that $\{X_i\}$ are iid exponentially distributed random variables with density $f(x) = \frac{1}{\lambda}\exp(-x/\lambda)$. Suppose that we only observe $\{X_i\}$, if $X_i > c$ (else X_i is not observed).*

(i) Show that the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is a biased estimator of λ.

(ii) Suppose that λ and c are unknown, obtain the log-likelihood of $\{X_i\}_{i=1}^n$ and the maximum likelihood estimators of λ and c.

Solution

(i) It is easy to see that $\mathrm{E}(\bar{X}) = \mathrm{E}(X_i|X_i > c)$*, thus*

$$\begin{aligned}\mathrm{E}(X_i|X_i > c) &= \int_0^\infty x\frac{f(x)I(X \geq c)}{P(X > c)}dx \\ &= \int_c^\infty x\frac{f(x)I(X \geq c)}{P(X > c)}dx = \frac{1}{e^{-c/\lambda}}\int_c^\infty xf(x)dx \\ &= \frac{\lambda e^{-c/\lambda}(\frac{c}{\lambda}+1)}{e^{-c/\lambda}} = \lambda + c.\end{aligned}$$

Thus $\mathrm{E}(\bar{X}) = \lambda + c$ *and not the desired* λ*.*

(ii) We observe that the density of X_i *given* $X_i > c$ *is* $f(x|X_i > c) = \frac{f(x)I(X>c)}{P(X>c)} = \lambda^{-1}\exp(-1/\lambda(X-c))I(X \geq c)$*; this is close to a shifted exponential and the density does not satisfy the regularity conditions.*

Based on this the log-likelihood $\{X_i\}$ *is*

$$\begin{aligned}\mathcal{L}_n(\lambda) &= \sum_{i=1}^n \big\{ \log f(X_i) + \log I(X_i \geq c) - \log P(X_i > c)\big\} \\ &= \sum_{i=1}^n \big\{ -\log\lambda - \frac{1}{\lambda}(X_i - c) + \log I(X_i \geq c)\big\}.\end{aligned}$$

Hence we want to find the λ *and* c *which maximises the above. Here we can use the idea of profiling to estimate the parameters - it does not matter which parameter we profile out. Suppose we fix,* λ*, and maximise the above with respect to* c*, in this case it is easier to maximise the actual likelihood:*

$$L_\lambda(c) = \prod_{i=1}^n \frac{1}{\lambda}\exp(-(X_i - c)/\lambda)I(X_i > c).$$

By drawing L *with respect to* c*, we can see that it is maximum at* $\min X_{(i)}$ *(for all* λ*), thus the MLE of* c *is* $\widehat{c} = \min_i X_i$*. Now we can estimate* λ*. Putting* $\widehat{c}$ *back into the log-likelihood gives*

$$\sum_{i=1}^n \big\{ -\log\lambda - \frac{1}{\lambda}(X_i - \widehat{c}) + \log I(X_i \geq \widehat{c})\big\}.$$

Differentiating the above with respect to λ *gives* $\sum_{i=1}^n (X_i - \widehat{c}) = \lambda n$*. Thus* $\widehat{\lambda}_n = \frac{1}{n}\sum_{i=1}^n X_i - \widehat{c}$*. Thus* $\widehat{c} = \min_i X_i$ $\widehat{\lambda}_n = \frac{1}{n}\sum_{i=1}^n X_i - \widehat{c}_n$*, are the MLE estimators of* c *and* λ *respectively.*

Chapter 5

Misspecification, the Kullbach Leibler Criterion and model selection

5.1 Assessing model fit

The Kullbach Leibler criterion is a method for measuring the "distance" between two densities. Rather than define it here it will come naturally from the discussion below on model misspecification.

5.1.1 Model misspecification

Until now we have assumed that the model we are fitting to the data is the correct model and our objective is to estimate the parameter θ. In reality the model we are fitting will not be the correct model (which is usually unknown). In this situation a natural question to ask is what are we estimating?

Let us suppose that $\{X_i\}$ are iid random variables which have the density $g(x)$. However, we fit the incorrect family of densities $\{f(x;\theta);\theta \in \Theta\}$ to the data using the MLE and estimate θ. The misspecified log likelihood is

$$\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta).$$

To understand what the MLE is actually estimating we use tthe LLN (law of large num-

bers) to obtain the limit of $\mathcal{L}_n(\theta)$

$$\frac{1}{n}\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta) \stackrel{\text{a.s.}}{\to} \mathrm{E}_g\big(\log f(X_i;\theta)\big) = \int \log f(x;\theta) g(x) dx. \tag{5.1}$$

Therefore it is clear that $\widehat{\theta}_n = \arg\max \mathcal{L}_n(\theta)$ is an estimator of

$$\theta_g = \arg\max \bigg(\int \log f(x;\theta) g(x) dx \bigg).$$

Hence $\widehat{\theta}_n$ is an estimator of the parameter which best fits the model in the specified family of models. Of course, one would like to know what the limit distribution of $(\widehat{\theta}_n - \theta_g)$ is (it will not be the same as the correctly specified case). Under the regularity conditions given in Theorem 5.1.1 and Assumption 2.6.1 (adapted to the misspecified case; these need to be checked) we can use the same proof as that given in Theorem 2.6.1 to show that $\widehat{\theta}_n \stackrel{\mathcal{P}}{\to} \theta_g$ (thus we have "consistency" of the misspecified MLE). We will assume in this section that this result is holds.

To obtain the limit distribution we again use the Taylor expansion of $\mathcal{L}_n(\theta)$ and the approximation

$$\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\theta_g} \approx \frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\widehat{\theta}_n} + I(\theta_g)\sqrt{n}(\widehat{\theta}_n - \theta_g), \tag{5.2}$$

where $I(\theta_g) = \mathrm{E}(-\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2} \rfloor_{\theta_g})$.

Theorem 5.1.1 *Suppose that $\{X_i\}$ are iid random variables with density g. However, we fit the incorrect family of densities $\{f(x;\theta); \theta \in \Theta\}$ to the data using the MLE and estimate θ, using $\widehat{\theta}_g = \arg\max \mathcal{L}_n(\theta)$ where*

$$\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta).$$

We assume

$$\frac{\partial \int_{\mathbb{R}} \log f(x;\theta) g(x) dx}{\partial \theta} \rfloor_{\theta=\theta_g} = \int_{\mathbb{R}} \frac{\log f(x;\theta)}{\partial \theta} \rfloor_{\theta=\theta_g} g(x) dx = 0 \tag{5.3}$$

and the usual regularity conditions are satisfied (exchanging derivative and integral is allowed and the third order derivative exists). Then we have

$$\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\theta_g} \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, J(\theta_g)), \tag{5.4}$$

$$\sqrt{n}(\widehat{\theta}_n - \theta_g) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, I(\theta_g)^{-1} J(\theta_g) I(\theta_g)^{-1}\big). \tag{5.5}$$

and

$$2\left(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_g)\right) \xrightarrow{\mathcal{D}} \sum_{j=1}^{p} \lambda_j Z_j^2 \tag{5.6}$$

where

$$\begin{aligned} I(\theta_g) &= \mathrm{E}\Big(- \frac{\partial^2 \log f(X;\theta)}{\partial \theta^2} \rfloor_{\theta_g} \Big) = - \int \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} g(x) dx \\ J(\theta_g) &= \mathrm{var}\Big(\frac{\partial \log f(X;\theta)}{\partial \theta} \rfloor_{\theta=\theta_g} \Big) = \mathrm{E}\Big(\frac{\partial \log f(X;\theta)}{\partial \theta} \rfloor_{\theta=\theta_g} \Big)^2 = \int \left(\frac{\partial \log f(x;\theta)}{\partial \theta} \right)^2 g(x) dx \end{aligned}$$

and $\{\lambda_j\}$ *are the eigenvalues of the matrix* $I(\theta_g)^{-1/2} J(\theta_g) I(\theta_g)^{-1/2}$.

PROOF. First the basics. Under assumption (5.3) $\frac{\partial f(X_i;\theta)}{\partial \theta} \rfloor_{\theta_g}$ are zero mean iid random variables. Therefore by using the CLT we have

$$\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\theta_g} \xrightarrow{\mathcal{D}} \mathcal{N}(0, J(\theta_g)). \tag{5.7}$$

If (5.3) is satisfied, then for large enough n we have $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\widehat{\theta}_n} = 0$, using the same ideas as those in Section 2.6.3 we have

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\theta_g} &\approx I(\theta_g) \sqrt{n}(\widehat{\theta}_n - \theta_g) \\ \Rightarrow \sqrt{n}(\widehat{\theta}_n - \theta_g) &\approx I(\theta_g)^{-1} \underbrace{\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\theta_g}}_{\text{term that determines normality}} . \end{aligned} \tag{5.8}$$

Hence asymptotic normality of $\sqrt{n}(\widehat{\theta}_n - \theta_g)$ follows from asymptotic normality of $\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\theta_g}$. Substituting (5.7) into (5.8) we have

$$\sqrt{n}(\widehat{\theta}_n - \theta_g) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, I(\theta_g)^{-1} J(\theta_g) I(\theta_g)^{-1}\big). \tag{5.9}$$

This gives (5.8).

To prove (5.6) we make the usual Taylor expansion

$$2\left(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_g)\right) \approx n \left(\widehat{\theta}_n - \theta_g\right)' I(\theta_g) \left(\widehat{\theta}_n - \theta_g\right) \tag{5.10}$$

Now we recall that since

$$\sqrt{n}(\widehat{\theta}_n - \theta_g) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, I(\theta_g)^{-1} J(\theta_g) I(\theta_g)^{-1}\big), \tag{5.11}$$

then asymptotically the distribution of $\sqrt{n}(\widehat{\theta}_n-\theta_g)$ is $\sqrt{n}(\widehat{\theta}_n-\theta_g) \stackrel{D}{=} I(\theta_g)^{-1/2}J(\theta_g)^{1/2}I(\theta_g)^{-1/2}\underline{Z}$ where $\underline{Z}$ is a p-dimension standard normal random variable. Thus we have

$$\begin{aligned} & 2\Big(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_g)\Big) \\ \stackrel{D}{=} & n\underline{Z}'I(\theta_g)^{-1/2}J(\theta_g)^{1/2}I(\theta_g)^{-1/2}I(\theta_g)I(\theta_g)^{-1/2}J(\theta_g)^{1/2}I(\theta_g)^{-1/2}\underline{Z} \\ = & \underline{Z}'I(\theta_g)^{-1/2}J(\theta_g)I(\theta_g)^{-1/2}\underline{Z} \end{aligned}$$

Let $P\Lambda P$ denote the spectral decomposition of the matrix $I(\theta_g)^{-1/2}J(\theta_g)I(\theta_g)^{-1/2}$. We observe that $P\underline{Z} \sim \mathcal{N}(0, I_p)$, thus we have

$$2\Big(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_g)\Big) = \sum_{j=1}^{p} \lambda_j Z_j^2$$

where λ_j are the eigenvalues of Λ and $I(\theta_g)^{-1/2}J(\theta_g)I(\theta_g)^{-1/2}$ and $\{Z_j\}$ are iid Gaussian random variables. Thus we have shown (5.6). □

An important feature is that in the misspecified case $I(\theta_g) \neq J(\theta_g)$. Hence whereas in the correctly specified case we have $\sqrt{n}(\widehat{\theta}_n - \theta_0) \stackrel{\mathcal{D}}{\to} \mathcal{N}\big(0, I(\theta_0)^{-1}\big)$ in the misspecified case it is $\sqrt{n}(\widehat{\theta}_n - \theta_g) \stackrel{\mathcal{D}}{\to} \mathcal{N}\big(0, I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}\big)$.

Recall that in the case the distributions are correctly specified we can estimate the information criterion with either the observed Fisher information

$$\widehat{I}_n(\widehat{\theta}_n) = \frac{-1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}\rfloor_{\theta=\widehat{\theta}_n}$$

or

$$\widehat{J}_n(\widehat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial \log f(X_i;\theta)}{\partial\theta}\rfloor_{\theta=\widehat{\theta}_n}\right)^2.$$

In the misspecified case we need to use *both* $\widehat{I}_n(\widehat{\theta}_n)$ and $\widehat{J}_n(\widehat{\theta}_n)$, which are are estimators of $I(\theta_g)$ and $J(\theta_g)$ respectively. Hence using this and Theorem 5.1.1 we can construct CIs for θ_g. To use the log-likelihood ratio statistic, the eigenvalues in the distribution need to calculated using $\widehat{I}_n(\widehat{\theta}_n)^{-1/2}\widehat{J}_n(\widehat{\theta}_n)\widehat{I}_n(\widehat{\theta}_n)^{-1/2}$. The log-likelihood ratio statistic is no longer pivotal.

Example 5.1.1 (Misspecifying the mean) *Let us suppose that $\{X_i\}_i$ are indepedendent random variables which satisfy the model $X_i = g(\frac{i}{n}) + \varepsilon_i$, where $\{\varepsilon_i\}$ are iid random*

variables which follow a t-distribution with 6-degrees of freedom (the variance of ε_i is finite). Thus, as n gets large we observe a corrupted version of $g(\cdot)$ on a finer grid.

The function $g(\cdot)$ is unknown, instead a line is fitted to the data. It is believed that the noise is Gaussian, and the slope $\widehat{a}_n$ maximises

$$\mathcal{L}_n(a) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i - a\cdot\frac{i}{n}\right)^2,$$

where $\sigma^2 = \mathrm{var}(\varepsilon_t)$ (note the role of σ^2 is meaningless in the minimisation).

Question: (i) What is $\widehat{a}_n$ estimating? (ii) What is the limiting distribution of $\widehat{a}_n$?

Solution:

(i) Rewriting $\mathcal{L}_n(a)$ we observe that

$$\begin{aligned}\frac{1}{n}\mathcal{L}_n(a) &= \frac{-1}{2\sigma^2 n}\sum_{i=1}^{n}\left(g(\frac{i}{n}) + \varepsilon_i - a\cdot\frac{i}{n}\right)^2 \\ &= \frac{-1}{2\sigma^2 n}\sum_{i=1}^{n}\left(g(\frac{i}{n}) - a\cdot\frac{i}{n}\right)^2 + \frac{-1}{2\sigma^2 n}\sum_{i=1}^{n}\varepsilon_i^2 + \frac{2}{2\sigma^2 n}\sum_{i=1}^{n}\left(g(\frac{i}{n}) - a\cdot\frac{i}{n}\right)\varepsilon_i \\ &\xrightarrow{\mathcal{P}} \frac{-1}{2\sigma^2 n}\int_0^1 (g(u)-au)^2 - \frac{1}{2}.\end{aligned}$$

Thus we observe $\widehat{a}_n$ is an estimator of the line which best fits the curve $g(\cdot)$ according to the ℓ_2-distance

$$a_g = \arg\min \int_0^1 \big(g(u) - au\big)^2 du.$$

If you draw a picture, this seems logical.

(ii) Now we derive the distribution of $\sqrt{n}(\widehat{a}_n - a_g)$. We assume (and it can be shown) that all the regularity conditions are satisfied. Thus we proceed to derive the derivatives of the "likelihoods"

$$\frac{1}{n}\frac{\partial\mathcal{L}_n(a)}{\partial a}\rfloor_{a_g} = \frac{1}{n\sigma^2}\sum_{i=1}^{n}\big(X_i - a_g\cdot\frac{i}{n}\big)\frac{i}{n} \qquad \frac{1}{n}\frac{\partial^2\mathcal{L}_n(a)}{\partial a^2}\rfloor_{a_g} = -\frac{1}{n\sigma^2}\sum_{i=1}^{n}\big(\frac{i}{n}\big)^2.$$

Note that $\frac{1}{n}\frac{\partial\mathcal{L}_n(a)}{\partial a}\rfloor_{a_g}$ are not iid random variables with mean zero. However, "globally" the mean will the close to zero, and $\frac{\partial\mathcal{L}_n(a)}{\partial a}\rfloor_{a_g}$ is the sum of independent X_i thus asymptotic normality holds i.e

$$\frac{1}{\sqrt{n}}\frac{\partial\mathcal{L}_n(a)}{\partial a}\rfloor_{a_g} = \frac{1}{\sqrt{n}\sigma^2}\sum_{i=1}^{n}\left(X_i - a_g\cdot\frac{i}{n}\right)\frac{i}{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, J(a_g)).$$

Evaluating the variance of the first derivative and expectation of the negative second derivative (and using the definition of the Reimann integral)

$$\begin{aligned} J(a_g) &= \frac{1}{n}\mathrm{var}\left(\frac{\partial \mathcal{L}_n(a)}{\partial a}\rfloor_{a_g}\right) = \frac{1}{n\sigma^4}\sum_{i=1}^{n}\mathrm{var}(X_i)\left(\frac{i}{n}\right)^2 \rightarrow \frac{1}{\sigma^2}\int_0^1 u^2 du = \frac{1}{3\sigma^2} \\ I(a_g) &= \frac{1}{n}\mathrm{E}\left(-\frac{\partial^2 \mathcal{L}_n(a)}{\partial a^2}\rfloor_{a_g}\right) = \frac{1}{n\sigma^2}\sum_{i=1}^{n}\left(\frac{i}{n}\right)^2 \rightarrow \frac{1}{\sigma^2}\int_0^1 u^2 du = \frac{1}{3\sigma^2}. \end{aligned}$$

We observe that in this case despite the mean and the distribution being misspecified we have that $I(a_g) \approx J(a_g)$. Altogether, this gives the limiting distribution

$$\sqrt{n}\big(\widehat{a}_n - a_g\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 3\sigma^2).$$

We observe that had we fitted a Double Laplacian to the data (which has the distribution $f_i(x) = \frac{1}{2b}\exp(-\frac{|x-\mu_i|}{b})$), the limit of the estimator would be different, and the limiting distribution would also be different.

5.2 The Kullbach-Leibler information criterion

The discussion above, in particular (5.1), motivates the definition of the Kullbach-Liebler criterion. We recall that the parameter which best fits the model using the maximum likelihood is an estimator of

$$\theta_g = \arg\max\left(\int \log f(x;\theta) g(x) dx\right).$$

θ_g can be viewed as the parameter which best fits the distribution out of all distributions in the misspecified parametric family. Of course the word 'best' is not particularly precise. It is best according to the criterion $\int \log f(x;\theta)g(x)dx$. To determine how well this fits the distribution we compare it to the limit likelihood using the correct distribution, which is

$$\int \log g(x) g(x) dx \quad \text{(limit of of likelihood of correct distribution)}.$$

In other words, the closer the difference

$$\int \log f(x;\theta_g) g(x) dx - \int \log g(x) g(x) dx = \int \log \frac{f(x;\theta_g)}{g(x)} g(x) dx$$

is to zero, the better the parameter θ_g fits the distribution g, using this criterion. Using Jenson's inequality we have

$$\int \log \frac{f(x;\theta)}{g(x)} g(x)dx = \mathrm{E}_g\left(\log \frac{f(X_i;\theta)}{g(X_i)}\right) \leq \log \mathrm{E}_g\left(\frac{f(X_i;\theta)}{g(X_i)}\right) \log \int f(x)dx \leq 0. \quad (5.12)$$

where equality arises only if $f(x;\theta) = g(x)$.

Therefore an alternative, but equivalent interpretation of θ_g, is the parameter which minimises the 'distance' between g and f_θ which is defined as

$$D(g, f_\theta) = \int \log f(x;\theta_g) g(x)dx - \int \log g(x) g(x)dx = \int \log \frac{f(x;\theta_g)}{g(x)} g(x)dx,$$

i.e. $\theta_g = \arg\max_{\theta\in\Theta} D(g, f_\theta)$. $D(g, f_\theta)$ is called the Kullbach-Leibler criterion. It can be considered as a measure of fit between the two distributions, the closer these two quantities are to zero the better the fit. We note that $D(g, f_\theta)$ is technically not a distance since $D(g, f_\theta) \neq D(f_\theta, g)$ (though it can be symmetrified). The Kullbach-Leibler criterion arises in many different contexts. We will use it in the section on model selection.

Often when comparing the model fit of different families of distributions our aim is to compare $\max_{\theta\in\Theta} D(g, f_\theta)$ with $\max_{\omega\in\Omega} D(g, h_\omega)$ where $\{f_\theta; \theta \in \Omega\}$ and $\{h_\omega; \omega \in \Omega\}$. In practice these distances cannot be obtained since the density g is unknown. Instead we estimate the maximum likelihood for both densities (but we need to keep all the constants, which are usually ignored in estimation) and compare these; i.e. compare $\max_{\theta\in\Theta} \mathcal{L}_f(\theta)$ with $\max_{\omega\in\Omega} \mathcal{L}_h(\omega)$. However, a direct comparison of log-likelihoods is problematic since the log-likelihood is a *biased* estimator of the K-L criterion. The bias can lead to overfitting of the model and a correction needs to be made (this we pursue in the next section).

We observe that $\theta_g = \arg\max_{\theta\in\Theta} D(g, f_\theta)$, hence $f(x;\theta_g)$ is the best fitting distribution using the K-L criterion. This does not mean it is the best fitting distribution according to another criterion. Indeed if we used a different distance measure, we are likely to obtain a different best fitting distribution. There are many different information criterions. The motivation for the K-L criterion comes from the likelihood. However, in the model misspecification set-up there are alternative methods, to likelihood methods, to finding the best fitting distribution (alternative methods may be more robust - for example the Renyi information criterion).

5.2.1 Examples

Example 5.2.1 *An example of misspecification is when we fit the exponential distribution* $\{f(x;\theta) = \theta^{-1}\exp(-x/\theta); \theta > 0\}$ *to the observations which come from the Weibull distribution. Suppose the data follows the Weibull distribution*

$$g(x) = \left(\frac{\alpha}{\phi}\right)\left(\frac{x}{\phi}\right)^{\alpha-1}\exp\left(-(x/\phi)^{\alpha}\right); \qquad \alpha,\phi > 0, \quad x > 0.$$

but we fit the exponential with the likelihood

$$\frac{1}{n}\mathcal{L}_n(\theta) = \frac{-1}{n}\sum_{i=1}^{n}\left(\log\theta + \frac{X_i}{\theta}\right) \stackrel{a.s.}{\to} -\log\theta - \mathrm{E}\big(\frac{X_i}{\theta}\big) = -\int\big(\log\theta + \frac{x}{\theta}\big)g(x)dx.$$

Let $\widehat{\theta}_n = \arg\max\mathcal{L}_n(\theta) = \bar{X}$. *Then we can see that* $\widehat{\theta}_n$ *is an estimator of*

$$\theta_g = \arg\max\{-\big(\log\theta + \mathrm{E}(X_i/\theta)\big)\} = \phi\Gamma(1+\alpha^{-1}) = \mathrm{E}(X_i) \tag{5.13}$$

Therefore by using Theorem 5.1.1 (or just the regular central limit theorem for iid random variables) we have

$$\sqrt{n}\big(\widehat{\theta}_n - \phi\Gamma(1+\alpha^{-1})\big) \stackrel{\mathcal{P}}{\to} \mathcal{N}\big(0, \underbrace{I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}}_{=\mathrm{var}(X_i)}\big)$$

where

$$\begin{aligned} I(\theta_g) &= \mathrm{E}\bigg(-\big(\theta^{-2} - 2X\theta^{-3}\big)\bigg)\rfloor_{\theta=\mathrm{E}(X)} = [\mathrm{E}(X)]^{-2} \\ J(\theta_g) &= \mathrm{E}\bigg(\big(-\theta^{-1} + X\theta^{-2}\big)^2\bigg)\rfloor_{\theta=\mathrm{E}(X)} = \frac{\mathrm{E}(X^2)}{[\mathrm{E}(X)]^4} - \frac{1}{[\mathrm{E}(X)]^2} = \frac{1}{\mathrm{E}[X^2]}\left(\frac{\mathrm{E}[X^2]}{\mathrm{E}[X]^2} - 1\right). \end{aligned}$$

Thus it is straightforward to see that $I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1} = var[X]$. *We note that for the Weibull distribution* $\mathrm{E}(X) = \phi\Gamma(1+\alpha^{-1})$ *and* $\mathrm{E}(X^2) = \phi^2\Gamma(1+2\alpha^{-1})$.

To check how well the best fitting exponential fits the Weibull distribution for different values of ϕ *and* α *we use the K-L information criterion;*

$$\begin{aligned} D(g, f_{\theta_g}) &= \int\log\left(\frac{\theta_g^{-1}\exp(-\theta_g^{-1}x)}{\frac{\alpha}{\phi}(\frac{x}{\phi})^{\alpha-1}\exp(-(\frac{x}{\phi})^{\alpha}}\right)\frac{\alpha}{\phi}(\frac{x}{\phi})^{\alpha-1}\exp(-(\frac{x}{\phi})^{\alpha})dx \\ &= \int\log\left(\frac{\phi\Gamma(1+\alpha^{-1})^{-1}\exp(-\phi\Gamma(1+\alpha^{-1})^{-1}x)}{\frac{\alpha}{\phi}(\frac{x}{\phi})^{\alpha-1}\exp(-(\frac{x}{\phi})^{\alpha}}\right)\frac{\alpha}{\phi}(\frac{x}{\phi})^{\alpha-1}\exp(-(\frac{x}{\phi})^{\alpha})dx. \end{aligned} \tag{5.14}$$

We note that by using (5.14), we see that $D(g, f_{\theta_g})$ *should be close to zero when* $\alpha = 1$ *(since then the Weibull is a close an exponential), and we conjecture that this difference should grow the further* α *is from one.*

Example 5.2.2 *Suppose $\{X_i\}_{i=1}^n$ are independent, identically distributed normal random variables with distribution $\mathcal{N}(\mu, \sigma^2)$, where $\mu > 0$. Suppose that μ and σ^2 are unknown.*

A non-central t-distribution with 11 degrees of freedom

$$f(x;a) = C(11)\left(1 + \frac{(x-a)^2}{11}\right)^{-(11+1)/2},$$

where $C(\nu)$ is a finite constant which only depends on the degrees of freedom, is mistakenly fitted to the observations. *[8]*

(i) Suppose we construct the likelihood using the t-distribution with 11 degrees of freedom, to estimate a. In reality, what is this MLE actually estimating?

(ii) Denote the above ML estimator as $\widehat{a}_n$. Assuming that standard regularity conditions are satisfied, what is the approximate distribution of $\widehat{a}_n$?

Solution

(i) The MLE seeks to estimate the maximum of $\mathrm{E}(\log f(X;a))$ wrt a.

Thus for this example $\widehat{a}_n$ is estimating

$$a_g = \arg\max_a \mathrm{E}\big(-6\log(1 + \frac{(X-a)^2}{11})\big) = \arg\min \int \log(1 + \frac{(x-a)^2}{11}))d\Phi\big(\frac{x-\mu}{\sigma}\big)dx.$$

(ii) Let a_g be defined a above. Then we have

$$\sqrt{n}(\widehat{a}_n - a_g) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, J^{-1}(a_g)I(a_g)J^{-1}(a_g)\big),$$

where

$$\begin{aligned} I(a_g) &= -C(11)6\mathrm{E}\big(\frac{d\log(1 + (X-a)^2/11)}{da}\rfloor_{a=a_g}\big)^2 \\ J(a_g) &= -C(11)6\mathrm{E}\big(\frac{d^2\log(1 + (X-a)^2/11)}{da^2}\rfloor_{a=a_g}\big). \end{aligned}$$

5.2.2 Some questions

Exercise 5.1 *The iid random variables $\{X_i\}_i$ follow a geometric distribution $\pi(1-\pi)^{k-1}$. However, a Poisson distribution with $P(X=k) = \frac{\theta^k \exp(-\theta)}{k!}$ is fitted to the data/*

(i) What quantity is the misspecified maximum likelihood estimator actually estimating?

(ii) How well does the best fitting Poisson distribution approximate the geometric distribution?

(iii) Given the data, suggest a method the researcher can use to check whether the Poisson distribution is an appropriate choice of distribution.

Exercise 5.2 *Let us suppose that the random variable X is a mixture of Weibull distributions*

$$f(x;\theta) = p(\frac{\alpha_1}{\phi_1})(\frac{x}{\phi_1})^{\alpha_1-1}\exp(-(x/\phi_1)^{\alpha_1}) + (1-p)(\frac{\alpha_2}{\phi_2})(\frac{x}{\phi_2})^{\alpha_2-1}\exp(-(x/\phi_2)^{\alpha_2}).$$

(i) Derive the mean and variance of X.

(ii) Obtain the exponential distribution which best fits the above mixture of Weibulls according to the Kullbach-Lieber criterion (recall that the exponential is $g(x;\lambda) = \frac{1}{\lambda}\exp(-x/\lambda)$).

Exercise 5.3 *Let us suppose that we observe the response variable and regressor (Y_i, X_i). Y_i and X_i are related through the model*

$$Y_i = g(X_i) + \varepsilon_i$$

where ε_i are iid Gaussian random variables (with mean zero and variance σ^2) which are independent of the regressors X_i. X_i are independent random variables, and the density of X_i is f. Suppose that it is wrongly assumed that Y_i satisfies the model $Y_i = \beta X_i + \varepsilon_i$, where ε_i are iid Gaussian random variables (with mean zero and variance σ^2, which can be assumed known).

(i) Given $\{(Y_i, X_i)\}_{i=1}^n$, what is the maximum likelihood estimator of β?

(ii) Derive an expression for the limit of this estimator (ie. what is the misspecified likelihood estimator actually estimating).

(iii) Derive an expression for the Kullbach-Leibler information between the true model and the best fitting misspecified model (that you derived in part (ii)).

5.3 Model selection

Over the past 30 years there have been several different methods for selecting the 'best' model out of a class of models. For example, the regressors $\{x_{i,j}\}$ are believed to influence the response Y_i with the model

$$Y_i = \sum_{j=1}^{p} a_j x_{i,j} + \varepsilon_i.$$

The natural question to ask is how many regressors should be included in the model. Without checking, we are prone to 'overfitting' the model.

There are various ways to approach this problem. One of the classical methods is to use an information criterion (for example the AIC). There are different methods for motivating the information criterion. Here we motivate it through the Kullbach-Leibler criterion. The main features of any criterion is that it can be split into two parts, the first part measures the model fit the second part measures the increased variance which is due to the inclusion of several parameters in the model.

To simplify the approach we will assume that $\{X_i\}$ are iid random variables with unknown distribution $g(x)$. We fit the family of distributions $\{f(x;\theta); \theta \in \Theta\}$ and want to select the best fitting distribution. Let

$$\begin{aligned} I(\theta_g) &= \mathrm{E}\Big(-\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2} \rfloor_{\theta_g} \Big) = -\int \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} g(x)dx \rfloor_{\theta_g} \\ J(\theta_g) &= \mathrm{E}\Big(\frac{\partial \log f(X;\theta)}{\partial \theta} \rfloor_{\theta=\theta_g} \Big)^2 = \int \Big(\frac{\partial \log f(x;\theta)}{\partial \theta} \Big)^2 g(x)dx \rfloor_{\theta_g}. \end{aligned}$$

Given the observations $\{X_i\}$ we use the mle to estimate the parameter

$$\widehat{\theta}_n(\underline{X}) = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\underline{X};\theta) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log f(X_i;\theta),$$

we have included $\underline{X}$ in $\widehat{\theta}$ to show that the mle depends on it. We will use the result

$$\sqrt{n}\left(\widehat{\theta}(\underline{X}) - \theta_g\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, I(\theta_g)^{-1} J(\theta_g) I(\theta_g)^{-1}\right).$$

Example 5.3.1 *Suppose we fit a Weibull distribution to the iid random variables $\{X_i\}_{i=1}^n$, and the best fitting parameter according to the K-L criterion is $\theta = \theta_g$ and $\alpha = 1$ (thus the parameters of an exponential), then*

$$\sqrt{n}\begin{pmatrix} \widehat{\theta}_n - \theta_g \\ \widehat{\alpha}_n - 1 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, I(\theta_g)^{-1} J(\theta_g) I(\theta_g)^{-1}\right).$$

Of course in practice, $\widehat{\alpha}_n \neq 1$. Thus we would like a model selection criterion to penalize the "larger" Weibull distribution in favour of the exponential distribution.

We cannot measure "fit" of an estimator by simply plugging the MLE back into the *same* likelihood (which gave the MLE)

$$\mathcal{L}_n(\widehat{\theta}_n(\underline{X});\underline{X}) = -\sum_{i=1}^{n} \log f(X_i;\widehat{\theta}_n(\underline{X})),$$

because $\widehat{\theta}(\underline{X})$ is finding the best fitting parameter for *the* data set $\underline{X}$. For example, suppose $\{X_i\}$ are iid random variables coming from a Cauchy distribution

$$c(x;\theta) = \frac{1}{\pi\left(1+(x-\theta)^2\right)}.$$

Let $\mathcal{L}_C(\theta;\underline{X})$ and $\widehat{\theta}(\underline{X})$ correspond to the log-likelihood and corresponding MLE. Suppose we also fit a Gaussian distribution to the same data set, let $\mathcal{L}_G(\mu,\sigma;\underline{X})$ and $\widehat{\mu}(\underline{X})$ and $\sigma^2(\underline{X})$ correspond to the log-likelihood and corresponding MLE. Even though the Gaussian distribution is the incorrect distribution, because it has the flexibility of two parameters rather than one, it is likely that

$$\mathcal{L}_G[\widehat{\mu}(\underline{X}),\widehat{\sigma}^2(\underline{X});\underline{X}] > \mathcal{L}_C[\widehat{\theta}(\underline{X});\underline{X}].$$

Which suggests the Gaussian likelihood better fits the data than the Cauchy, when its simply that there are more parameters in the Gaussian likelihood. This is the reason that validation data sets are often used. This is a data set $\underline{Y}$, which is independent of $\underline{X}$, but where $\underline{Y}$ and $\underline{X}$ have the same distribution. The quantity

$$\mathcal{L}_n(\widehat{\theta}_n(\underline{X});\underline{Y}) = -\sum_{i=1}^{n} \log f(Y_i;\widehat{\theta}_n(\underline{X}))$$

measures how well $\widehat{\theta}(\underline{X})$ fits *another* equivalent data. In this case, if $\{X_i\}$ and $\{Y_i\}$ are iid random variables from a Cauchy distribution it is highly *unlikely*

$$\mathcal{L}_G[\widehat{\mu}(\underline{X}),\widehat{\sigma}^2(\underline{X});\underline{Y}] > \mathcal{L}_C[\widehat{\theta}(\underline{X});\underline{Y}].$$

Since $\underline{Y}$ is random and we want to replace highly unlikely to definitely will not happen, we consider the limit and measure how well $f(y;\widehat{\theta}_n(\underline{X}))$ fits the expectation

$$\mathrm{E}_{\underline{Y}}\left[\frac{1}{n}\mathcal{L}_n(\widehat{\theta}_n(\underline{X});\underline{Y})\right] = \int \log f(y;\widehat{\theta}_n(\underline{X}))g(y)dy.$$

The better the fit, the larger the above will be. Note that if we subtract $\int \log g(y)g(y)dy$ from the above we have the K-L criterion. As a matter of convention we define the negative of the above

$$\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right] = -\int \log f(y;\widehat{\theta}_n(\underline{X}))g(y)dy.$$

The better the fit, the smaller $\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]$ will be. We observe that $\widetilde{D}[g, f_{\widehat{\theta}_n(\underline{X})}]$ depends on the sample $\underline{X}$. Therefore, a more sensible criterion is to consider the expectation of the above over all random samples $\underline{X}$

$$\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} = -\mathrm{E}_{\underline{X}}\left\{\mathrm{E}_Y\left[\log f(Y;\widehat{\theta}_n(\underline{X})\right]\right\}.$$

$\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\}$ is the information criterion that we aim to estimate. First we show that $\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\}$ penalizes models which are over fitted (which $n^{-1}\mathcal{L}(\widehat{\theta}(\underline{X});\underline{X})$) is unable to do). Making a Taylor expansion of $\mathrm{E}_{\underline{X}}\big(\mathrm{E}_Y(\log f(Y;\widehat{\theta}_n(\underline{X}))\big)$ about θ_g gives

$$\begin{aligned}
&\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} = -\mathrm{E}_{\underline{X}}\left\{\mathrm{E}_Y\left[\log f(Y;\widehat{\theta}_n(\underline{X})\right]\right\} \\
\approx\ & -\mathrm{E}_{\underline{X}}\left\{\mathrm{E}_Y\left[\log f(Y;\theta_g)\right]\right\} - \mathrm{E}_{\underline{X}}\left\{\left[\widehat{\theta}(X)-\theta_g\right]\underbrace{\mathrm{E}_Y\left[\frac{\partial \log f(Y;\theta)}{\partial\theta}{}_{\rfloor\theta=\theta_g}\right]}_{=0}\right\} \\
& -\mathrm{E}_{\underline{X}}\left\{\left[\widehat{\theta}(X)-\theta_g\right]\mathrm{E}_Y\left[\frac{\partial^2 \log f(Y;\theta)}{\partial\theta^2}{}_{\rfloor\theta=\theta_g}\right]\left[\widehat{\theta}(X)-\theta_g\right]\right\} \\
\approx\ & -\frac{1}{2}\mathrm{E}_Y[\log f(Y;\theta_g)] + \frac{1}{2}\mathrm{E}_{\underline{X}}\Big((\widehat{\theta}_n(\underline{X})-\theta_g)'I(\theta_g)(\widehat{\theta}_n(\underline{X})-\theta_g)\Big).
\end{aligned}$$

The second term on the right of the above grows as the number of parameters grow (recall it has a χ^2-distribution where the number of degrees of freedom is equal to the number of parameters). Hence $\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\}$ penalises unnecessary parameters making it an ideal criterion. For example, we may be fitting a Weibull distribution to the data, however, the best fitting distribution turns out to be an exponential distribution, the additional term will penalize the over fit.

However, in practise $\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\}$ is unknown and needs to estimated. Many information criterions are based on estimating $\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\}$ (including the AIC and corrected AIC, usually denoted as AICc). Below we give a derivation of the AIC based on approximating $\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\}$.

We recall that $\widehat{\theta}_n(\underline{X})$ is an estimator of θ_g hence we start by replacing $\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\}$ with $\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\theta_g})\right]\right\} = \widetilde{D}[g, f_{\theta_g}]$ to give

$$\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} = \widetilde{D}[g, f_{\theta_g}] + \left(\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} - \widetilde{D}\left[g, f_{\theta_g}\right]\right).$$

We first focus on the first term $\widetilde{D}[g, f_{\theta_g}]$. Since $\mathrm{E}_{\underline{X}}\big(\widetilde{D}(g, f_{\theta_g})\big)$ is unknown we replace it by its sample average

$$\widetilde{D}[g, f_{\theta_g}] = -\int f(y;\theta_g)g(y)dy \approx -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g).$$

Hence we have

$$\begin{aligned}\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} &\approx -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g) + \left(\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} - \mathrm{E}_{\underline{X}}\left\{\widetilde{D}[g, f_{\theta_g}]\right\}\right)\\ &= -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g) + I_1.\end{aligned}$$

Of course, θ_g is unknown so this is replaced by $\widehat{\theta}_n(\underline{X})$ to give

$$\mathrm{E}_{\underline{X}}\big(\widetilde{D}(g, f_{\widehat{\theta}_n(\underline{X})})\big) \approx -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\widehat{\theta}_n(\underline{X})) + I_1 + I_2 \tag{5.15}$$

where

$$I_2 = \left(\frac{1}{n}\sum_{i=1}^{n}\log f\left(X_i;\widehat{\theta}_n(\underline{X})\right) - \frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g)\right).$$

We now find approximations for I_1 and I_2. We observe that the terms I_1 and I_2 are both positive; this is because $\theta_g = \arg\min\big(\widetilde{D}(g, f_\theta)\big)$ (recall that $\widetilde{D}$ is the expectation of the *negative* likelihood) and $\widehat{\theta}_n = \arg\max\sum_{i=1}^{n}\log f(X_i;\theta)$. This implies that

$$\begin{aligned}\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} &\geq \mathrm{E}_{\underline{X}}\left\{\widetilde{D}[g, f_{\theta_g}]\right\}\\ \text{and } \frac{1}{n}\sum_{i=1}^{n}\log f\left(X_i;\widehat{\theta}_n(\underline{X})\right) &\geq \frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g).\end{aligned}$$

Thus if θ_g are the parameters of a Weibull distribution, when the best fitting distribution is an exponential (i.e. a Weibull with $\alpha = 1$), the additional terms I_1 and I_2 will penalize this.

We bound I_1 and I_2 by making Taylor expansions. By using the Taylor expansion (and the assumption that $\mathrm{E}(\frac{\partial \log f(x;\theta)}{\partial\theta}\rfloor_{\theta=\theta_g}) = 0$) we have

$$\begin{aligned}
& \mathrm{E}_{\underline{X}}\left[\widetilde{D}(g, f_{\widehat{\theta}_n(\underline{X})}) - \widetilde{D}(g, f_{\theta_g})\right] \\
= & -\mathrm{E}_{\underline{X}}\mathrm{E}_{\underline{Y}}\left(\frac{1}{n}\sum_{i=1}^{n}\left\{\log f(Y_i;\widehat{\theta}_n(\underline{X})) - \log f(Y_i;\theta_g)\right\}\right) \\
= & -\frac{1}{n}\mathrm{E}_{\underline{X}}\mathrm{E}_{\underline{Y}}\left(\mathcal{L}_n(\underline{Y},\widehat{\theta}_n(\underline{X})) - \mathcal{L}_n(\underline{Y},\theta_g)\right) \\
= & -\frac{1}{n}\mathrm{E}_{\underline{X}}\underbrace{\mathrm{E}_{\underline{Y}}\left(\frac{\partial\mathcal{L}_n(\underline{Y},\theta)}{\partial\theta}\rfloor_{\theta_g}(\widehat{\theta}_n(\underline{X}) - \theta_g)\right)}_{=0} - \frac{1}{2n}\mathrm{E}_{\underline{Y}}\mathrm{E}_{\underline{X}}\left((\widehat{\theta}_n(\underline{X}) - \theta_g)'\frac{\partial^2\mathcal{L}_n(\underline{Y},\theta)}{\partial\theta}\rfloor_{\bar{\theta}(\underline{X})}(\widehat{\theta}_n(\underline{X}) - \theta_g)\right) \\
= & -\frac{1}{2n}\mathrm{E}_{\underline{Y}}\mathrm{E}_{\underline{X}}\left((\widehat{\theta}_n(\underline{X}) - \theta_g)'\frac{\partial^2\mathcal{L}_n(\underline{Y},\theta)}{\partial\theta}\rfloor_{\bar{\theta}(\underline{X})}(\widehat{\theta}_n(\underline{X}) - \theta_g)\right),
\end{aligned}$$

where $\bar{\theta}(\underline{X}) = \alpha\theta(\underline{X}) + (1-\alpha)\theta_g$ for some $0 \leq \alpha \leq 1$. Now we note that

$$-\frac{1}{n}\frac{\partial^2\mathcal{L}_n(\underline{Y},\theta)}{\partial\theta^2}\rfloor_{\bar{\theta}(\underline{X})} \approx -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f(X_i,\theta)}{\partial\theta^2}\rfloor_{\theta=\theta_g} \xrightarrow{\mathcal{P}} I(\theta_g),$$

which (using a hand wavey argument) gives

$$I_1 = \mathrm{E}_{\underline{X}}\big(\widetilde{D}(g, f_{\widehat{\theta}_n(\underline{X})}) - \widetilde{D}(g, f_{\theta_g})\big) \approx \frac{1}{2}\mathrm{E}_{\underline{X}}\left((\widehat{\theta}_n(\underline{X}) - \theta_g)'I(\theta_g)(\widehat{\theta}_n(\underline{X}) - \theta_g)\right) \quad (5.16)$$

We now obtain an estimator of I_2 in (5.15). To do this we make the usual Taylor expansion (noting that $\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\widehat{\theta}_n} = 0$)

$$\begin{aligned}
I_2 &= \left(\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g) - \frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\widehat{\theta}_n(\underline{X}))\right) \\
&\approx \frac{1}{2}(\widehat{\theta}_n(\underline{X}) - \theta_g)'I(\theta_g)(\widehat{\theta}_n(\underline{X}) - \theta_g). \qquad (5.17)
\end{aligned}$$

To obtain the final approximations for (5.16) and (5.17) we use (5.11) where

$$\sqrt{n}(\widehat{\theta}_n - \theta_g) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}\big).$$

Now by using the above and the relationship that if $\underline{Z} \sim \mathcal{N}(0,\Sigma)$ then $\mathrm{E}(\underline{Z}'A\underline{Z}) =$

trace$\big\{A\Sigma\big\}$ (check your linear models notes). Therefore by using the above we have

$$\begin{aligned} I_2 &= -\bigg(\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g) - \frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\widehat{\theta}_n(\underline{X}))\bigg) \\ &\approx \frac{1}{2}(\widehat{\theta}_n(\underline{X})-\theta_g)'I(\theta_g)(\widehat{\theta}_n(\underline{X})-\theta_g) \\ &\approx \frac{1}{2}\mathrm{E}\left(\underbrace{(\widehat{\theta}_n(\underline{X})-\theta_g)'}_{\approx\mathcal{N}(0,I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}/n)} I(\theta_g)(\widehat{\theta}_n(\underline{X})-\theta_g)\right) \\ &\approx \frac{1}{2n}\mathrm{trace}\Big(I(\theta_g)^{-1}J(\theta_g)\Big) \end{aligned}$$

and by the same reasoning we have

$$\begin{aligned} I_1 &= \mathrm{E}_{\underline{X}}\big(\widetilde{D}(g,f_{\widehat{\theta}_n(\underline{X})}) - \widetilde{D}(g,f_{\theta_g})\big) \approx \frac{1}{2}\mathrm{E}_{\underline{X}}\Big((\widehat{\theta}_n(\underline{X})-\theta_g)'I(\theta_g)(\widehat{\theta}_n(\underline{X})-\theta_g)\Big) \\ &\approx \frac{1}{2n}\mathrm{trace}\Big(I(\theta_g)^{-1}J(\theta_g)\Big). \end{aligned}$$

Simplifying the above and substituting into (5.15) gives

$$\begin{aligned} \mathrm{E}_{\underline{X}}\big\{\widetilde{D}[g,f_{\widehat{\theta}_n(\underline{X})}]\big\} &\approx -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\widehat{\theta}_n(\underline{X})) + \frac{1}{n}\mathrm{trace}\Big(J(\theta_g)I(\theta_g)^{-1}\Big) \\ &= -\frac{1}{n}\mathcal{L}_n(\underline{X};\widehat{\theta}_n(\underline{X})) + \frac{1}{n}\mathrm{trace}\Big(J(\theta_g)I(\theta_g)^{-1}\Big). \end{aligned}$$

Altogether one approximation of $\mathrm{E}_{\underline{X}}\Big\{\widetilde{D}[g,f_{\widehat{\theta}_n(\underline{X})}]\Big\}$ is

$$\mathrm{E}_{\underline{X}}\big(\widetilde{D}(g,f_{\widehat{\theta}_n(\underline{X})})\big) \approx -\frac{1}{n}\mathcal{L}_n(\underline{X};\widehat{\theta}_n(\underline{X})) + \frac{1}{n}\mathrm{trace}\Big(J(\theta_g)I(\theta_g)^{-1}\Big). \tag{5.18}$$

This approximation of the $K-L$ information is called the AIC (Akaike Information Criterion). In the case that $J(\theta_g)=I(\theta_g)$ the AIC reduces to

$$AIC(p) = -\frac{1}{n}\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n}) + \frac{p}{n},$$

and we observe that it penalises the number of parameters (this is the classical AIC). This is one of the first information criterions.

We apply the above to the setting of model selection. The idea is that we have a set of candidate models we want to fit to the data, and we want to select the best model.

- Suppose there are N different candidate family of models. Let $\{f_p(x;\theta_p);\theta_p \in \Theta_p\}$ denote the pth family.

- Let

$$\mathcal{L}_{p,n}(\underline{X};\theta_p) = \sum_{i=1}^{n} \log f(X_i;\theta_p)$$

  denote the likelihood associated with the pth family. Let $\widehat{\theta}_{p,n} = \arg\max_{\theta_p \in \Theta_p} \mathcal{L}_{p,n}(\underline{X};\theta_p)$ denote the maximum likelihood estimator of the pth family.

- In an ideal world we would compare the different families by selecting the family of distributions $\{f_p(x;\theta_p);\theta_p \in \Theta_p\}$ which minimise the criterion $\mathrm{E}_{\underline{X}}\big(\widetilde{D}(g, f_{p,\widehat{\theta}_{p,n}(\underline{X})})\big)$. However, we do not know $\mathrm{E}_{\underline{X}}\big(\widetilde{D}(g, f_{p,\widehat{\theta}_{p,n}(\underline{X})})\big)$ hence we consider an estimator of it given in (5.18).

  This requires estimators of $J(\theta_{p,g})$ and $I(\theta_{p,g})$, this we can be easily be obtained from the data and we denote this as $\widehat{J}_p$ and $\widehat{I}_p$.

- We then choose the the family of distributions which minimise

$$\min_{1\leq p\leq N}\Big(-\frac{1}{n}\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n}) + \frac{1}{n}\mathrm{trace}\big(\widehat{J}_p\widehat{I}_p^{-1}\big)\Big) \tag{5.19}$$

  In other words, the order we select is $\widehat{p}$ where

$$\widehat{p} = \arg\min_{1\leq p\leq N}\Big(-\frac{1}{n}\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n}) + \frac{1}{n}\mathrm{trace}\big(\widehat{J}_p\widehat{I}_p^{-1}\big)\Big)$$

Often (but not always) in model selection we assume that the true distribution is nested in the many candidate model. For example, the 'true' model $Y_i = \alpha_0 + \alpha_1 x_{i,1} + \varepsilon_i$ belongs to the set of families defined by

$$Y_{i,p} = \alpha_0 + \sum_{j=1}^{p} \alpha_j x_{i,j} + \varepsilon_i \qquad p > 1.$$

In this case $\{\alpha_0 + \sum_{j=1}^{p} \alpha_j x_{i,j} + \varepsilon_i; \alpha_j \in \mathbb{R}^{p+1}\}$ denotes the pth family of models. Since the true model is nested in most of the candidate model we are in the specified case. Hence we have $J(\theta_g) = I(\theta_g)$, in this case $\mathrm{trace}\big(J(\theta_g)I(\theta_g)^{-1}\big) = \mathrm{trace}\big(I(\theta_g)I(\theta_g)^{-1}\big) = p$. In this case (5.19) reduces to selecting the family which minimises

$$AIC(p) = \min_{1\leq p\leq N}\Big(-\frac{1}{n}\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n}) + \frac{p}{n}\Big).$$

There is a bewildering array of other criterions (including BIC etc), but most are similar in principle and usually take the form

$$-\frac{1}{n}\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n}) + \text{pen}_n(p),$$

where $\text{pen}_n(p)$ denotes a penality term (there are many including Bayes Information criterion etc.).

Remark 5.3.1

- *Usually the AIC is defined as*

$$AIC(p) = -2\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n}) + 2p,$$

  this is more a matter of preference (whether we include the factor $2n$ or not).

- *We observe that as the sample size grows, the weight of penalisation relative to the likelihood declines (since $\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n}) = O(n)$).*

  This fact can mean that the AIC can be problematic; it means that the AIC can easily overfit, and select a model with a larger number of parameters than is necessary (see Lemma 5.3.1).

- *Another information criterion is the BIC (this can be obtained using a different reasoning), and is defined as*

$$BIC(p) = -2\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n}) + p\log n.$$

- *The AIC does not place as much weight on the number of parameters, whereas the BIC the does place a large weight on the parameters. It can be shown that the BIC is a consistent estimator of the model (so long as the true model is in the class of candidate models). However, it does have a tendency of underfitting (selecting a model with too few parameters).*

- *However, in the case that the the true model does not belong to any the families, the AIC can be a more suitable criterion than other criterions.*

Note that "estimators" such as the AIC (or even change point detection methods, where the aim is to detect the location of a change point) are different to classical estimators in the sense that the estimator is "discrete valued". In such cases, often the intention is to show that the estimator is consistent, in the sense that

$$P(\widehat{p}_n = p) \xrightarrow{\mathcal{P}} 1$$

as $n \to \infty$ (where $\widehat{p}$ denotes the estimator and p the true parameter). There does exist some paper which try to construct confidence intervals for such discrete valued estimators, but they tend to be rarer.

Lemma 5.3.1 (Inconsistency of the AIC) *Suppose that we are in the specified case and θ_p is the true model. Hence the true model has order p. Then for any $q > 0$ we have that*

$$\lim_{n\to\infty} P\big(\arg \min_{1\leq m\leq p+q} AIC(m) > p\big) > 0,$$

moreover

$$\lim_{n\to\infty} P\big(\arg \min_{1\leq m\leq p+q} AIC(m) = p\big) \neq 1.$$

In other words, the AIC will with a positive probability choose the larger order model, and is more likely to select large models, as the the order q increases.

PROOF. To prove the result we note that $(p+q)$-order model will be selected over p-order in the AIC if $-\mathcal{L}_{p+q,T} + (p+q) < -\mathcal{L}_{p,n} + p$, in other words we select $(p+q)$ if

$$\mathcal{L}_{p+q,n} - \mathcal{L}_{p,n} > q.$$

Hence

$$\begin{aligned} & P\big(\arg \min_{1\leq m\leq p+q} AIC(m) > p\big) = P\big(\arg \min_{p\leq m\leq p+q} AIC(m) < AIC(p)\big) \\ \geq \; & P\big(AIC(p+q) < AIC(p)\big) \geq P(2(\mathcal{L}_{p+q,n} - \mathcal{L}_{p,n}) > 2q). \end{aligned}$$

But we recall that $\mathcal{L}_{p+q,n}$ and $\mathcal{L}_{p,n}$ are both log-likelihoods and under the null that the pth order model is the true model we have $2(\mathcal{L}_{p+q,n} - \mathcal{L}_{p,n}) \xrightarrow{\mathcal{D}} \chi^2_q$. Since $\mathrm{E}(\chi^2_q) = q$ and $\mathrm{var}[\chi^2_q] = 2q$, we have for any $q > 0$ that

$$P\big(\arg \min_{1\leq m\leq p+q} AIC(m) > p\big) \geq P(2(\mathcal{L}_{p+q,n} - \mathcal{L}_{p,n}) > 2q) > 0.$$

Hence with a positive probability the AIC will choose the larger model.

This means as the sample size n grows, with a positive probability we will not necessarily select the correct order p, hence the AIC is inconsistent and

$$\lim_{n\to\infty} P\big(\arg \min_{1\leq m\leq p+q} AIC(m) = p\big) \neq 1.$$

□

Remark 5.3.2 (The corrected AIC) *In order to correct for the bias in the AIC the corrected AIC was proposed in Sugiura (1978) and Hurvich and Tsai (1989). This gives a more subtle approximation of* $\mathrm{E}_{\underline{X}}\{\widetilde{D}[g, f_{\widehat{\theta}_n(\underline{X})}]\}$ *which results in an additional penalisation term being added to the AIC. It can be shown that for linear models the AICc consistently estimates the order of the model.*

Remark 5.3.3 *The AIC is one example of penalised model that take the form*

$$-\mathcal{L}_{p,n}(\underline{X};\theta) + \lambda\|\theta\|_\alpha,$$

where $\|\theta\|_\alpha$ *is a "norm" on* θ*. In the case of the AIC the* ℓ_0*-norm* $\|\theta\|_0 = \sum_{i=1}^p I(\theta_i \neq 0)$ *(where* $\theta = (\theta_1, \ldots, \theta_p)$ *and* I *denotes the indicator variable). However, minimisation of this model over all subsets of* $\theta = (\theta_1, \ldots, \theta_p)$ *is computationally prohibitive if* p *is large. Thus norms where* $\alpha \geq 1$ *are often seeked (such as the LASSO etc).*

Regardless of the norm used, if the number of non-zero parameter is finite, with a positive probability we will over estimate the number of non-zero parameters in the model.

5.3.1 Examples

This example considers model selection for logistic regression, which is covered later in this course.

Example 5.3.2 *Example: Suppose that* $\{Y_i\}$ *are independent binomial random variables where* $Y_i \sim B(n_i, p_i)$*. The regressors* $x_{1,i}, \ldots, x_{k,i}$ *are believed to influence the probability* p_i *through the logistic link function*

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1,i} + \beta_p x_{p,i} + \beta_{p+1} x_{p+1,i} + \ldots + \beta_q x_{q,i},$$

where $p < q$*.*

(a) Suppose that we wish to test the hypothesis

$$H_0 : \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1,i} + \beta_p x_{p,i}$$

against the alternative

$$H_0 : \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1,i} + \beta_p x_{p,i} + \beta_{p+1} x_{p+1,i} + \ldots + \beta_q x_{q,i}.$$

State the log-likelihood ratio test statistic that one would use to test this hypothesis. If the null is true, state the limiting distribution of the test statistic.

(b) Define the model selection criterion

$$M_n(d) = 2\mathcal{L}_n(\widehat{\beta}_d) - 2Cd$$

where C is a finite constant,

$$\mathcal{L}_{i,d}(\beta_d) = \sum_{i=1}^{n} \left(Y_i \beta_d' \underline{x}_{id} - n_i \log(1 + \exp(\beta_d' \underline{x}_{id}) + \binom{n_i}{Y_i} \right),$$

$\underline{x}_{id} = (x_{1,i}, \ldots, x_{d,i})$ and $\widehat{\beta}_d = \arg\max_{\beta_d} \mathcal{L}_{i,d}(\beta_d)$. We use $\widehat{d} = \arg\max_d M_n(d)$ as an estimator of the order of the model.

Suppose that H_0 defined in part (2a) is true, use your answer in (2a) to explain whether the model selection criterion $M_n(d)$ consistently estimates the order of model.

Solution:

(a) The likelihood for both hypothesis is

$$\mathcal{L}_{i,d}(\beta_d) = \sum_{i=1}^{n} \left(Y_i \beta_d' \underline{x}_{id} - n_i \log(1 + \exp(\beta_d' \underline{x}_{id}) + \binom{n_i}{Y_i} \right).$$

Thus the log-likelihood ratio test is

$$\begin{aligned} \rangle_n &= 2\big(\mathcal{L}_{n,q}(\widehat{\beta}_q) - \mathcal{L}_{i,p}(\widehat{\beta}_p\big) \\ &= 2\sum_{i=1}^{n} \left(Y_i[\widehat{\beta}_A' - \widehat{\beta}_0']\underline{x}_i - n_i[\log(1 + \exp(\widehat{\beta}_A' \underline{x}_i) - \log(1 + \exp(\widehat{\beta}_0' \underline{x}_i)] \right) \end{aligned}$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_A$ are the maximum likelihood estimators under the null and alternative respectively.

If the null is true, then $\rangle_n \xrightarrow{\mathcal{D}} \chi^2_{q-p}$ as $T \to \infty$.

(b) Under the null we have that $\rangle_n = 2\big(\mathcal{L}_{n,q}(\widehat{\beta}_q) - \mathcal{L}_{n,p}(\widehat{\beta}_p\big) \xrightarrow{\mathcal{D}} \chi^2_{q-p}$. Therefore, by definition, if $\widehat{d} = \arg\max_d M_n(d)$, then we have

$$\big(\mathcal{L}_{\widehat{d}}(\widehat{\beta}_d) - 2C\widehat{d}\big) - \big(\mathcal{L}_p(\widehat{\beta}_p) - 2Cp\big) > 0.$$

Suppose $q > p$, then the model selection criterion would select q over p if

$$2\big[\mathcal{L}_{\widehat{d}}(\widehat{\beta}_q) - \mathcal{L}_p(\widehat{\beta}_p)\big] > 2C(q - p).$$

Now the LLRT test states that under the null $2\big[\mathcal{L}_q(\widehat{\beta}_q)-\mathcal{L}_p(\widehat{\beta}_p) \xrightarrow{\mathcal{D}} \chi^2_{q-p}$*, thus roughly speaking we can say that*

$$P\bigg[\mathcal{L}_q(\widehat{\beta}_q) - (\mathcal{L}_p(\widehat{\beta}_p) > 2C(\widehat{d}-p)\bigg] \approx P(\chi^2_{q-p} > 2C(q-p)).$$

As the above is a positive probability, this means that the model selection criterion will select model q *over the true smaller model with a positive probability. This argument holds for all* $q > p$*, thus the model selection criterion* $M_n(d)$ *does not consistently estimate* d*.*

5.3.2 Recent model selection methods

The AIC and its relatives have been extensively in statistics over the past 30 years because it is easy to evaluate. There are however problems in the case that p is large (more so when p is large with respect to the sample size n, often called the large p small n problem). For example, in the situation where the linear regression model takes the form

$$Y_i = \sum_{j=1}^{p} a_j x_{i,j} + \varepsilon_i,$$

where the number of possible regressors $\{x_{i,j}\}$ is extremely large. In this case, evaluating the mle for all the p different candidate models, and then making a comparisoon can take a huge amount of computational time. In the past 10 years there has been a lot of work on alternative methods of model selection. One such method is called the LASSO, this is where rather than estimating all model individually parameter estimation is done on the large model using a penalised version of the MLE

$$\mathcal{L}_n(\theta) + \lambda \sum_{i=1}^{p} |\theta_i|.$$

The hope is by including the $\lambda \sum_{i=1}^{p} |\theta_i|$ in the likelihood many of coefficients of the regressors would be set to zero (or near zero). Since the introduction of the LASSO in 1996 many variants of the LASSO have been proposed and also the LASSO has been applied to several different situations.

Chapter 6

Survival Analysis

6.1 An introduction to survival analysis

6.1.1 What is survival data?

Data where a set of 'individuals' are observed and the failure time or lifetime of that individual is recordered is usually called survival data. We note that individual does not necessarily need to be a person but can be an electrical component etc. Examples include:

- Lifetime of a machine component.
- Time until a patient's cure, remission, passing.
- Time for a subject to perform a task.
- Duration of an economic cycle.
- Also it may not be 'time' we are interested in but:
  - Length of run of a particle.
  - Amount of usage of a machine, eg. amount of petrol used etc.

In the case that we do not observe any regressors (explanatory variables) which influence the survival time (such as gender/age of a patient etc), we can model the survival times as iid random variables. If the survival times are believed to have the density $f(x;\theta_0)$, where $f(x;\theta)$ is known but θ_0 is unknown, then the maximum likelihood can be used to

estimate θ. The standard results discussed in Section 2.2 can be easily applied to this type of data.

6.1.2 Definition: The survival, hazard and cumulative hazard functions

Let T denote the survival time of an individual, which has density f. The density f and the distribution function $F(x) = \int_0^x f(u)du$ are not particularly informative about the chance of survival at a given time point. Instead, the survival, hazard and cumlative hazard functions, which are functions of the density and distribution function, are used instead.

- **The survival function**.

  This is $\mathcal{F}(x) = 1 - F(x)$. It is straightforward to see that $\mathcal{F}(x) = P(T > x)$ (observe that the strictly greater than sign is necessary). Therefore, $\mathcal{F}(x)$ is the probability of survival over x.

- **The hazard function**

  The hazard function is defined as

$$\begin{aligned} h(x) &= \lim_{\delta x \to 0} \frac{P(x < T \leq x + \delta x | T > x)}{\delta x} = \lim_{\delta x \to 0} \frac{P(x < T \leq x + \delta x)}{\delta x P(T > x)} \\ &= \frac{1}{\mathcal{F}(x)} \lim_{\delta x \to 0} \frac{F(x + \delta x) - F(x)}{\delta x} = \frac{f(x)}{\mathcal{F}(x)} = -\frac{d \log \mathcal{F}(x)}{dx}. \end{aligned}$$

  We can see from the definition the hazard function is the 'chance' of failure (though it is a normalised probability, not a probability) at time x, given that the individual has survived until time x.

  We see that the hazard function is similar to the density in the sense that it is a positive function. However it does not integrate to one. Indeed, it is not integrable.

- **The cumulative Hazard function**

  This is defined as

$$H(x) = \int_{-\infty}^{x} h(u)du.$$

It is straightforward to see that

$$H(x) = \int_{-\infty}^{x} -\frac{d \log \mathcal{F}(x)}{dx} \rfloor_{x=u} du = -\log \mathcal{F}(x).$$

This is just the analogue of the distribution function, however we observe that unlike the distribution function, $H(x)$ is unbounded.

It is straightforward to show that $f(x) = h(x)\exp(-H(x))$ and $\mathcal{F}(x) = \exp(-H(x))$.

It is useful to know that given any one of $f(x)$, $F(x)$, $H(x)$ and $h(x)$, uniquely defines the other functions. Hence there is a one-to-one correspondence between all these functions.

Example 6.1.1 • **The Exponential distribution**

Suppose that $f(x) = \frac{1}{\theta}\exp(-x/\theta)$.

Then the distribution function is $F(x) = 1 - \exp(-x/\theta)$. $\mathcal{F}(x) = \exp(-x/\theta)$, $h(x) = \frac{1}{\theta}$ *and* $H(x) = x/\theta$.

The exponential distribution is widely used. However, it is not very flexible. We observe that the hazard function is constant over time. This is the well known memoryless property of the exponential distribution. In terms of modelling it means that the chance of failure in the next instant does not depend on on how old the individual is. The exponential distribution cannot model 'aging'.

- **The Weibull distribution**

We recall that this is a generalisation of the exponential distribution, where

$$f(x) = \left(\frac{\alpha}{\theta}\right)\left(\frac{x}{\theta}\right)^{\alpha-1} \exp(-(x/\theta)^{\alpha}); \alpha, \theta > 0, \quad x > 0.$$

For the Weibull distribution

$$\begin{aligned} F(x) &= 1 - \exp\big(-(x/\theta)^{\alpha}\big) \qquad \mathcal{F}(x) = \exp\big(-(x/\theta)^{\alpha}\big) \\ h(x) &= (\alpha/\theta)(x/\theta)^{\alpha-1} \qquad H(x) = (x/\theta)^{\alpha}. \end{aligned}$$

Compared to the exponential distribution the Weibull has a lot more flexibility. Depending on the value of α, *the hazard function* $h(x)$ *can either increase over time or decay over time.*

- **The shortest lifetime model**

  Suppose that $Y_1, \ldots, Y_k$ are independent life times and we are interested in the shortest survival time (for example this could be the shortest survival time of k sibling mice in a lab when given some disease). Let g_i, $\mathcal{G}_i$ H_i and h_i denote the density, survival function, cumalitive hazard and hazard function respectively of Y_i (we do not assume they have the same distribution) and $T = \min(Y_i)$. Then the survival function is

$$\mathcal{F}_T(x) = P(T > x) = \prod_{i=1}^{k} P(Y_i > x) = \prod_{i=1}^{k} \mathcal{G}_i(x).$$

  Since the cumulative hazard function satisfies $H_i(x) = -\log \mathcal{G}_i(x)$, the cumulative hazard function of T is

$$H_T(x) = -\sum_{i=1}^{k} \log \mathcal{G}_i(x) = \sum_{i=1}^{k} H_i(x)$$

  and the hazard function is

$$h_T(x) = \sum_{i=1}^{k} \frac{d(-\log \mathcal{G}_i(x))}{dx} = \sum_{i=1}^{k} h_i(x)$$

- **Survival function with regressors** *See Section 3.2.2.*

Remark 6.1.1 (Discrete Data) *Let us suppose that the survival time are not continuous random variables, but discrete random variables. In other words, T can take any of the values $\{t_i\}_{i=1}^{\infty}$ where $0 \leq t_1 < t_2 < \ldots$. Examples include the first time an individual visits a hospital post operation, in this case it is unlikely that the exact time of visit is known, but the date of visit may be recorded.*

Let $P(T = t_i) = p_i$, using this we can define the survival function, hazard and cumulative hazard function.

(i) **Survival function** *The survival function is*

$$\mathcal{F}_i = P(T > t_i) = \sum_{j=i+1}^{\infty} P(T = t_j) = \sum_{j=i+1}^{\infty} p_j.$$

(ii) **Hazard function** *The hazard function is*

$$\begin{aligned} h_i &= P(t_{i-1} < T \leq t_i | T > t_{i-1}) = \frac{P(T = t_i)}{P(T > T_{i-1})} \\ &= \frac{p_i}{\mathcal{F}_{i-1}} = \frac{\mathcal{F}_{i-1} - \mathcal{F}_i}{\mathcal{F}_{i-1}} = 1 - \frac{\mathcal{F}_i}{\mathcal{F}_{i-1}}. \end{aligned} \tag{6.1}$$

Now by using the above we have the following useful representation of the survival function in terms of hazard function

$$\mathcal{F}_i = \prod_{j=2}^{i} \frac{\mathcal{F}_i}{\mathcal{F}_{i-1}} = \prod_{j=2}^{i} \big(1 - h_j\big) = \prod_{j=1}^{i} \big(1 - h_j\big), \tag{6.2}$$

since $h_1 = 0$ *and* $\mathcal{F}_1 = 1$.

(iii) **Cumulative hazard function** *The cumulative hazard function is* $H_i = \sum_{j=1}^{i} h_j$.

These expression will be very useful when we consider nonparametric estimators of the survival function $\mathcal{F}$.

6.1.3 Censoring and the maximum likelihood

One main feature about survival data which distinguishes, is that often it is "incomplete". This means that there are situations where the random variable (survival time) is not completely observed (this is often called incomplete data). Usually, the incompleteness will take the form as *censoring* (this will be the type of incompleteness we will consider here).

There are many type of censoring, the type of censoring we will consider in this chapter is right censoring. This is where the time of "failure", may not be observed if it "survives" beyond a certain time point. For example, is an individual (independent of its survival time) chooses to leave the study. In this case, we would only know that the individual survived beyond a certain time point. This is called right censoring. Left censoring arises when the start (or birth) of an individual is unknown (hence it is known when an individual passes away, but the individuals year of birth is unknown), we will not consider this problem here.

Let us suppose that T_i is the survival time, which may not be observed and we observe instead $Y_i = \min(T_i, c_i)$, where c_i is the potential censoring time. We *do know* if the data

has been censored, and together with Y_i we observe the indicator variable

$$\delta_i = \begin{cases} 1 & T_i \leq c_i \quad \text{(uncensored)} \\ 0 & T_i > c_i \quad \text{(censored)} \end{cases}.$$

Hence, in survival analysis we typically observe $\{(Y_i, \delta_i)\}_{i=1}^n$. We use the observations $\{(Y_i, \delta_i)\}_{i=1}^n$ to make inference about unknown parameters in the model.

Let us suppose that T_i has the distribution $f(x; \theta_0)$, where f is known but θ_0 is unknown.

Naive approaches to likelihood construction

There are two naive approaches for estimating θ_0. One method is to ignore the fact that the observations are censored and use time of censoring as if the were failure times. Hence define the likelihood

$$\mathcal{L}_{1,n}(\theta) = \sum_{i=1}^n \log f(Y_i; \theta),$$

and use as the parameter estimator $\hat{\theta}_{1,n} = \arg\max_{\theta \in \Theta} \mathcal{L}_{1,n}(\theta)$. The fundemental problem with this approach is that it will be biased. To see this consider the expectation of $n^{-1}\mathcal{L}_{1,n}(\theta)$ (for convenience let $c_i = c$). Since

$$Y_i = T_i I(T_i \leq c) + cI(T_i > c) \Rightarrow \log f(Y_i; \theta) = [\log f(T_i; \theta)] I(T_i \leq c) + [\log f(c; \theta)] I(T_i > c)$$

this gives the likelihood

$$\mathrm{E}(\log f(Y_i; \theta)) = \int_0^c \log f(x; \theta) f(x; \theta_0) dx + \underbrace{\mathcal{F}(c; \theta_0)}_{\text{probability of censoring}} \log f(c; \theta).$$

There is no reason to believe that θ_0 maximises the above. For example, suppose f is the exponential distribution, using the $\mathcal{L}_{1,n}(\theta)$ leads to the estimator $\widehat{\theta}_{1,n} = n^{-1} \sum_{i=1}^n Y_i$, which is clearly a biased estimator of θ_0. Hence this approach should be avoided since the resulting estimator is biased.

Another method is to construct the likelihood function by ignoring the censored data. In other words use the log-likelihood function

$$\mathcal{L}_{2,n}(\theta) = \sum_{i=1}^n \delta_i \log f(Y_i; \theta),$$

and let $\hat{\theta}_{2,n} = \arg\max_{\theta\in\Theta} \mathcal{L}_{2,n}(\theta)$ be an estimator of θ. It can be shown that if a fixed censor value is used, i.e. $Y_i = \min(T_i, c)$, then this estimator is not a consistent estimator of θ, it is also biased. As above, consider the expectation of $n^{-1}\mathcal{L}_{2,n}(\theta)$, which is

$$\mathrm{E}(\delta_i \log f(Y_i;\theta)) = \int_0^c \log f(x;\theta) f(x;\theta_0) dx.$$

It can be shown that θ_0 does not maximise the above. Of course, the problem with the above "likelihood" is that it is not the correct likelihood (if it were then Theorem 2.6.1 tells us that the parameter will maximise the expected likelihood). The correct likelihood conditions on the non-censored data being less than c to give

$$\mathcal{L}_{2,n}(\theta) = \sum_{i=1}^{n} \delta_i \left(\log f(Y_i;\theta) - \log(1 - \mathcal{F}(c;\theta))\right).$$

This likelihood gives a consistent estimator of the θ; this see why consider its expectation

$$\mathrm{E}\left(n^{-1}\mathcal{L}_{2,n}(\theta)\right) = \mathrm{E}\left(\log \frac{f(Y_i;\theta)}{\mathcal{F}(c;\theta)} \big| T_i < c\right) = \int_0^c \log \frac{f(x;\theta)}{\mathcal{F}(c;\theta)} \log \frac{f(c;\theta_0)}{\mathcal{F}(c;\theta_0)} dx.$$

Define the "new" density $g(x;\theta) = \frac{f(x;\theta)}{\mathcal{F}(c;\theta)}$ for $0 \leq x < c$. Now by using Theorem 2.6.1 we immediately see that the $\mathrm{E}\left(n^{-1}\mathcal{L}_{2,n}(\theta)\right)$ is maximised at $\theta = \theta_0$. However, since we have not used all the data we have lost "information" and the variance will be larger than a likelihood that includes the censored data.

The likelihood under censoring (review of Section 1.2)

The likelihood under censoring can be constructed using both the density and distribution functions or the hazard and cumulative hazard functions. Both are equivalent. The log-likelihood will be a mixture of probabilities and densities, depending on whether the observation was censored or not. We observe (Y_i, δ_i) where $Y_i = \min(T_i, c_i)$ and δ_i is the indicator variable. In this section we treat c_i as if they were deterministic, we consider the case that they are random later.

We first observe that if $\delta_i = 1$, then the log-likelihood of the individual observation Y_i is $\log f(Y_i;\theta)$, since

$$P(Y_i = x|\delta_i = 1) = P(T_i = x|T_i \leq c_i) = \frac{f(x;\theta)}{1 - \mathcal{F}(c_i;\theta)} dx = \frac{h(y;\theta)\mathcal{F}(x;\theta)}{1 - \mathcal{F}(c_i;\theta)} dx. \tag{6.3}$$

On the other hand, if $\delta_i = 0$, the log likelihood of the individual observation $Y_i = c|\delta_i = 0$ is simply one, since if $\delta_i = 0$, then $Y_i = c_i$ (it is given). Of course it is clear that

$P(\delta_i = 1) = 1 - \mathcal{F}(c_i;\theta)$ and $P(\delta_i = 0) = \mathcal{F}(c_i;\theta)$. Thus altogether the joint density of $\{Y_i, \delta_i\}$ is

$$\left(\frac{f(x;\theta)}{1-\mathcal{F}(c_i;\theta)} \times (1-\mathcal{F}(c_i;\theta))\right)^{\delta_i}\left(1 \times \mathcal{F}(c_i;\theta)\right)^{1-\delta_i} = f(x;\theta)^{\delta_i}\mathcal{F}(c_i;\theta)^{1-\delta_i}.$$

Therefore by using $f(Y_i;\theta) = h(Y_i;\theta)\mathcal{F}(Y_i;\theta)$, and $H(Y_i;\theta) = -\log \mathcal{F}(Y_i;\theta)$, the joint log-likelihood of $\{(Y_i,\delta_i)\}_{i=1}^n$ is

$$\begin{aligned}\mathcal{L}_n(\theta) &= \sum_{i=1}^n \Big(\delta_i \log f(Y_i;\theta) + (1-\delta_i)\log\big(1-F(Y_i;\theta)\big)\Big) \\ &= \sum_{i=1}^n \delta_i\big(\log h(T_i;\theta) - H(T_i;\theta)\big) - \sum_{i=1}^n (1-\delta_i)H(c_i;\theta) \\ &= \sum_{i=1}^n \delta_i \log h(Y_i;\theta) - \sum_{i=1}^n H(Y_i;\theta). \end{aligned} \tag{6.4}$$

You may see the last representation in papers on survival data. Hence we use as the maximum likelihood estimator $\hat{\theta}_n = \arg\max \mathcal{L}_n(\theta)$.

Example 6.1.2 The exponential distribution *Suppose that the density of T_i is $f(x;\theta) = \theta^{-1}\exp(x/\theta)$, then by using (6.4) the likelihood is*

$$\mathcal{L}_n(\theta) = \sum_{i=1}^n \Big(\delta_i\big(-\log\theta - \theta^{-1}Y_i\big) - (1-\delta_i)\theta^{-1}Y_i\Big).$$

By differentiating the above it is straightforward to show that the maximum likelihood estimator is

$$\hat{\theta}_n = \frac{\sum_{i=1}^n \delta_i T_i + \sum_{i=1}^n (1-\delta_i)c_i}{\sum_{i=1}^n \delta_i}.$$

6.1.4 Types of censoring and consistency of the mle

It can be shown that under certain censoring regimes the estimator converges to the true parameter and is asymptotically normal. More precisely the aim is to show that

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0)^{-1}), \tag{6.5}$$

where

$$I(\theta) = -\mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^n \delta_i \frac{\partial^2 \log f(Y_i;\theta)}{\partial\theta^2} + \frac{1}{n}\sum_{i=1}^n (1-\delta_i)\frac{\partial^2 \log \mathcal{F}(c_i;\theta)}{\partial\theta^2}\Big).$$

Note that typically we replace the Fisher information with the observed Fisher information

$$\widetilde{I}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\partial^2 \log f(Y_i;\theta)}{\partial\theta^2} + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{\partial^2 \log \mathcal{F}(c_i;\theta)}{\partial\theta^2}.$$

We discuss the behaviour of the likelihood estimator for different censoring regimes.

Non-random censoring

Let us suppose that $Y_i = \min(T_i, c)$, where c is some deterministic censoring point (for example the number of years cancer patients are observed). We first show that the expectation of the likelihood is maximum at the true parameter (this under certain conditions means that the mle defined in (6.4) will converge to the true parameter). Taking expectation of $\mathcal{L}_n(\theta)$ gives

$$\begin{aligned}\mathrm{E}\Big(n^{-1}\mathcal{L}_n(\theta)\Big) &= \mathrm{E}\Big(\delta_i \log f(T_i;\theta) + (1-\delta_i)\log \mathcal{F}(T_i;\theta)\Big)\\ &= \int_0^c \log f(x;\theta) f(x;\theta_0)dx + \mathcal{F}(c;\theta_0)\log \mathcal{F}(c;\theta).\end{aligned}$$

To show that the above is maximum at θ (assuming no restrictions on the parameter space) we differentiate $\mathrm{E}\big(\mathcal{L}_n(\theta)\big)$ with respect to θ and show that it is zero at θ_0. The derivative at θ_0 is

$$\begin{aligned}\frac{\partial \mathrm{E}\big(n^{-1}\mathcal{L}_n(\theta)\big)}{\partial\theta}\rfloor_{\theta=\theta_0} &= \frac{\partial}{\partial\theta}\int_0^c f(x;\theta)dx\rfloor_{\theta=\theta_0} + \frac{\partial \mathcal{F}(c;\theta)}{\partial\theta}\rfloor_{\theta=\theta_0}\\ &= \frac{\partial(1-\mathcal{F}(c;\theta))}{\partial\theta}\rfloor_{\theta=\theta_0} + \frac{\partial \mathcal{F}(c;\theta)}{\partial\theta}\rfloor_{\theta=\theta_0} = 0.\end{aligned}$$

This proves that the expectation of the likelihood is maximum at zero (which we would expect, since this all fall under the classical likelihood framework). Now assuming that the standard regularity conditions are satisfied then (6.5) holds where the Fisher information matrix is

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2}\Big(\int_0^c f(x;\theta_0)\log f(x;\theta)dx + \mathcal{F}(c;\theta_0)\log \mathcal{F}(c;\theta)\Big).$$

We observe that when $c = 0$ (thus all the times are censored), the Fisher information is zero, thus the asymptotic variance of the mle estimator, $\hat{\theta}_n$ is not finite (which is consistent with out understanding of the Fisher information matrix). It is worth noting that under this censoring regime the estimator is consistent, but the variance of the estimator will

be larger than when there is no censoring (just compare the Fisher informations for both cases).

In the above, we assume the censoring time c was common for all individuals, such data arises in several studies. For example, a study where life expectancy was followed for up to 5 years after a procedure. However, there also arises data where the censoring time varies over individuals, for example an individual, i, may pull out of a study at time c_i. In this case, the Fisher information matrix is

$$\begin{aligned} I_n(\theta) &= -\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \Big(\int_0^{c_i} f(x;\theta_0) \log f(x;\theta) dx + \mathcal{F}(c_i;\theta_0) \log \mathcal{F}(c_i;\theta) \Big) \\ &= \sum_{i=1}^{n} I(\theta; c_i). \end{aligned} \tag{6.6}$$

However, if there is a lot of variability between the censoring times, one can "model these as if they were random". I.e. that c_i are independent realisations from the random variable C. Within this model (6.6) can be viewed as the Fisher information matrix conditioned on the censoring time $C_i = c_i$. However, it is clear that as $n \to \infty$ a limit can be achieved (which cannot be when the censoring in treated as deterministic) and

$$\frac{1}{n} \sum_{i=1}^{n} I(\theta; c_i) \stackrel{\text{a.s.}}{\to} \int_{\mathbb{R}} I(\theta; c) k(c) dc, \tag{6.7}$$

where $k(c)$ denotes the censoring density. The advantage of treating the censoring as random, is that it allows one to understand how the different censoring times influences the limiting variance of the estimator. In the section below we formally incorporate random censoring in the model and consider the conditions required such that the above is the Fisher information matrix.

Random censoring

In the above we have treated the censoring times as fixed. However, they can also be treated as if they were random i.e. the censoring times $\{c_i = C_i\}$ are random. Usually it is assumed that $\{C_i\}$ are iid random variables which are *independent* of the survival times. Furthermore, it is assumed that the distribution of C does not depend on the unknown parameter θ.

Let k and K denote the density and distribution function of $\{C_i\}$. By using the arguments given in (6.3) the likelihood of the joint distribution of $\{(Y_i, \delta_i)\}_{i=1}^n$ can be

obtained. We recall that the probability of $(Y_i \in [y - \frac{h}{2}, y + \frac{h}{2}], \delta_i = 1)$ is

$$\begin{aligned} & P\left(Y_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right], \delta_i = 1\right) \\ = \; & P\left(Y_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right] \big| \delta_i = 1\right) P\left(\delta_i = 1\right) \\ = \; & P\left(T_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right] \big| \delta_i = 1\right) P\left(\delta_i = 1\right). \end{aligned}$$

Thus

$$\begin{aligned} & P\left(Y_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right], \delta_i = 1\right) = P\left(T_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right], \delta_i = 1\right) \\ = \; & P\left(\delta_i = 1 | T_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right]\right) P\left(T_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right]\right) \\ \approx \; & P\left(\delta_i = 1 | T_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right]\right) f_{T_i}(y) h \\ = \; & P(T_i \leq C_i | T_i = y) f_{T_i}(y) h \\ = \; & P(y \leq C_i) f_{T_i}(y) h = f_{T_i}(y) \left(1 - K(y)\right) h. \end{aligned}$$

It is very important to note that the last line $P(T_i \leq C_i | T_i = y) = P(y \leq C_i)$ is due to *independence* between T_i and C_i, if this does not hold the expression would involve the joint distribution of Y_i and C_i.

Thus the likelihood of $(Y_i, \delta_i = 1)$ is $f_{T_i}(y)\left(1 - K(y)\right)$. Using a similar argument the probability of $(Y_i \in [y - \frac{h}{2}, y + \frac{h}{2}], \delta_i = 0)$ is

$$\begin{aligned} & P\left(Y_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right], \delta_i = 0\right) = P\left(C_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right], \delta_i = 0\right) \\ = \; & P\left(\delta_i = 0 | C_i = \left[y - \frac{h}{2}, y + \frac{y}{2}\right]\right) f_{C_i}(y) h \\ = \; & P\left(C_i < T_i | C_i = \left[y - \frac{h}{2}, y + \frac{h}{2}\right]\right) f_{C_i}(y) h = k(C_i) \mathcal{F}(C_i; \theta) h. \end{aligned}$$

Thus the likelihood of (Y_i, δ_i) is

$$\left[f_{T_i}(Y_i)\left(1 - K(Y_i)\right)\right]^{\delta_i} \left[k(Y_i)\mathcal{F}(Y_i; \theta)\right]^{1 - \delta_i}.$$

This gives the log-likelihood

$$\begin{aligned}
\mathcal{L}_{n,R}(\theta) &= \sum_{i=1}^{n}\Big(\delta_i\big[\log f(Y_i;\theta)+\log(1-K(Y_i))\big]+(1-\delta_i)\big[\log\big(1-F(C_i;\theta)\big)+\log k(C_i)\big]\Big)\\
&= \underbrace{\sum_{i=1}^{n}\Big(\delta_i\log f(Y_i;\theta)+(1-\delta_i)\log\big(1-F(C_i;\theta)\big)\Big)}_{=\mathcal{L}_n(\theta)}\\
&\quad+\sum_{i=1}^{n}\Big(\delta_i\log(1-K(Y_i))+(1-\delta_i)\log k(C_i)\Big)\\
&= \mathcal{L}_n(\theta)+\sum_{i=1}^{n}\Big(\delta_i\log(1-K(Y_i))+(1-\delta_i)\log k(C_i)\Big). \qquad (6.8)
\end{aligned}$$

The interesting aspect of the above likelihood is that if the censoring density $k(y)$ *does not depend on* θ, then the maximum likelihood estimator of θ_0 is identical to the maximum likelihood estimator using the non-random likelihood (or, equivalently, the likelihood conditioned on C_i) (see (6.3)). In other words

$$\widehat{\theta}_n = \arg\max \mathcal{L}_n(\theta) = \arg\max \mathcal{L}_{n,R}(\theta).$$

Hence the estimators using the two likelihoods are the same. The only difference is the limiting distribution of $\widehat{\theta}_n$.

We now examine what $\widehat{\theta}_n$ is actually estimating in the case of random censoring. To ease notation let us suppose that the censoring times follow an exponential distribution $k(x) = \beta\exp(-\beta x)$ and $K(x) = 1-\exp(-\beta x)$. To see whether $\hat{\theta}_n$ is biased we evaluate the derivative of the likelihood. As both the full likelihood and the conditional yield the same estimators, we consider the expectation of the conditional log-likelihood. This is

$$\begin{aligned}
\mathrm{E}\big(\mathcal{L}_n(\theta)\big) &= n\mathrm{E}\Big(\delta_i\log f(T_i;\theta)\Big)+n\mathrm{E}\Big((1-\delta_i)\log\mathcal{F}(C_i;\theta)\Big)\\
&= n\mathrm{E}\Big(\log f(T_i;\theta)\,\underbrace{\mathrm{E}(\delta_i|T_i)}_{=\exp(-\beta T_i)}\Big)+n\mathrm{E}\Big(\log\mathcal{F}(C_i;\theta)\,\underbrace{\mathrm{E}\big(1-\delta_i|C_i\big)}_{=\mathcal{F}(C_i;\theta_0)}\Big),
\end{aligned}$$

where the above is due to $\mathrm{E}(\delta_i|T_i) = P(C_i > T_i|T_i) = \exp(-\beta T_i)$ and $\mathrm{E}(1-\delta_i|C_i) = P(T_i > C_i|C_i) = \mathcal{F}(C_i;\theta_0)$. Therefore

$$\mathrm{E}\big(\mathcal{L}_n(\theta)\big) = n\Big(\int_0^\infty \exp(-\beta x)\log f(x;\theta)f(x;\theta_0)dx+\int_0^\infty \mathcal{F}(c;\theta_0)\beta\exp(-\beta c)\log\mathcal{F}(c;\theta)dc\Big).$$

It is not immediately obvious that the true parameter θ_0 maximises $\mathrm{E}\big(\mathcal{L}_n(\theta)\big)$, however by using (6.8) the expectation of the true likelihood is

$$\mathrm{E}\left(\mathcal{L}_{R,n}(\theta)\right) = \mathrm{E}\left(\mathcal{L}_n(\theta)\right) + nK.$$

Thus the parameter which maximises the true likelihood also maximises $\mathrm{E}\big(\mathcal{L}_n(\theta)\big)$. Thus by using Theorem 2.6.1, we can show that $\widehat{\theta}_n = \arg\max \mathcal{L}_{R,n}(\theta)$ is a consistent estimator of θ_0. Note that

$$\sqrt{n}\left(\hat{\theta}_{n,R} - \theta_0\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0)^{-1})$$

where

$$\begin{aligned} I(\theta) &= n\int_0^\infty \exp(-\beta x)\left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta_0)^{-1}dx - \\ & n\int_0^\infty \exp(-\beta x)\frac{\partial^2 f(x;\theta)}{\partial\theta^2}dx \\ & +n\int_0^\infty \beta\exp(-\beta c)\left(\frac{\partial \mathcal{F}(c;\theta)}{\partial\theta}\right)^2 \mathcal{F}(c;\theta)^{-1}dc \\ & -n\int_0^\infty \beta\exp(-\beta c)\frac{\partial^2 \mathcal{F}(c;\theta)}{\partial\theta^2}dc. \end{aligned}$$

Thus we see that the random censoring does have an influence on the limiting variance of $\widehat{\theta}_{n,R}$.

Remark 6.1.2 *In the case that the censoring time C depends on the survival time T it is tempting to still use (6.8) as the "likelihood", and use the parameter estimator the parameter which maximises this likelihood. However, care needs to be taken. The likelihood in (6.8) is constructed under the assumption T and C, thus it is not the true likelihood and we cannot use Theorem 2.6.1 to show consistency of the estimator, in fact it is likely to be biased.*

In general, given a data set, it is very difficult to check for dependency between survival and censoring times.

Example 6.1.3 *In the case that T_i is an exponential, see Example 6.1.2, the MLE is*

$$\hat{\theta}_n = \frac{\sum_{i=1}^n \delta_i T_i + \sum_{i=1}^n (1-\delta_i)C_i}{\sum_{i=1}^n \delta_i}.$$

Now suppose that C_i is random, then it is possible to calculate the limit of the above. Since the numerator and denominator are random it is not easy to calculate the expectation. However under certain conditions (the denominator does not converge to zero) we have by Slutsky's theorem that

$$\widehat{\theta}_n \xrightarrow{\mathcal{P}} \frac{\sum_{i=1}^n \mathrm{E}(\delta_i T_i + (1-\delta_i)C_i)}{\sum_{i=1}^n \mathrm{E}(\delta_i)} = \frac{\mathrm{E}(\min(T_i, C_i))}{P(T_i < C_i)}.$$

Definition: Type I and Type II censoring

- *Type I sampling* In this case, there is an upper bound on the observation time. In other words, if $T_i \leq c$ we observe the survival time but if $T_i > c$ we do not observe the survival time. This situation can arise, for example, when a study (audit) ends and there are still individuals who are alive. This is a special case of non-random sampling with $c_i = c$.

- *Type II sampling* We observe the first r failure times, $T_{(1)}, \ldots, T_{(r)}$, but do not observe the $(n-r)$ failure times, whose survival time is greater than $T_{(r)}$ (we have used the ordering notation $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(n)}$).

6.1.5 The likelihood for censored discrete data

Recall the discrete survival data considered in Remark 6.1.1, where the failures can occur at $\{t_s\}$ where $0 \leq t_1 < t_2 < \ldots$. We will suppose that the censoring of an individual can occur only at the times $\{t_s\}$. We will suppose that the survival time probabilities satisfy $P(T = t_s) = p_s(\theta)$, where the parameter θ is unknown but the function p_s is known, and we want to estimate θ.

Example 6.1.4

(i) The geometric distribution $P(X = k) = p(1-p)^{k-1}$ for $k \geq 1$ (p is the unknown parameter).

(ii) The Poisson distribution $P(X = k) = \lambda^k \exp(-\lambda)/k!$ for $k \geq 0$ (λ is the unknown parameter).

As in the continuous case let Y_i denote the failure time or the time of censoring of the *ith* individual and let δ_i denote whether the *ith* individual is censored or not. Hence, we

observe $\{(Y_i, \delta_i)\}$. To simplify the exposition let us define

$$\begin{aligned} d_s &= \text{number of failures at time } t_s \qquad q_s = \text{number censored at time } t_s \\ N_s &= \sum_{i=s}^{\infty} (d_i + q_i)\,. \end{aligned}$$

So there data would look like this:

Time	No. Failures at time t_i	No. censored at time t_i	Total Number
t_1	d_1	q_1	$N_1 = \sum_{i=1}^{\infty}(d_i + q_i)$
t_2	d_2	q_2	$N_2 = \sum_{i=2}^{\infty}(d_i + q_i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Thus we observe from the table above that

$$N_s - d_s = \text{Number survived just before time } t_{s+1}.$$

Hence at any given time t_s, there are d_s "failures" and $N_s - d_s$ "survivals".

Hence, since the data is discrete observing $\{(Y_i, \delta_i)\}$ is equivalent to observing $\{(d_s, q_s)\}$ (i.e. the number of failures and censors at time t_s), in terms of likelihood construction (this leads to equivalent likelihoods). Using $\{(d_s, q_s)\}$ and Remark 6.1.1 we now construct the likelihood. We shall start with the usual (not log) likelihood. Let $P(T = t_s|\theta) = p_s(\theta)$ and $P(T \geq t_s|\theta) = \mathcal{F}_s(\theta)$. Using this notation observe that the probability of (d_s, q_s) is $p_s(\theta)^{d_s} P(T \geq t_s)^{q_s} = p_s(\theta)^{d_s} \mathcal{F}_s(\theta)^{q_s}$, hence the likelihood is

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} p_{Y_i}(\theta)^{\delta_i} \mathcal{F}_{Y_i}(\theta)^{1-\delta_i} = \prod_{s=1}^{\infty} p_s(\theta)^{d_s} \mathcal{F}_s(\theta)^{q_s} \\ &= \prod_{s=1}^{\infty} p_s(\theta)^{d_s} [\sum_{j=s}^{\infty} p_j(\theta)]^{q_s}. \end{aligned}$$

For most parametric inference the above likelihood is relatively straightforward to maximise. However, in the case that our objective is to do nonparametric estimation (where we do not assume a parametric model and directly estimate the probabilities without restricting them to a parametric family), then rewriting the likelihood in terms of the hazard function greatly simplies matters. By using some algebraic manipulations and Remark 6.1.1 we now rewrite the likelihood in terms of the hazard functions. Using that $p_s(\theta) = h_s(\theta)\mathcal{F}_{s-1}(\theta)$ (see equation (6.1)) we have

$$L_n(\theta) = \prod_{s=1} h_s(\theta)^{d_s} \mathcal{F}_s(\theta)^{q_s} \mathcal{F}_{s-1}(\theta)^{d_s} = \prod_{s=1} \underbrace{h_s(\theta)^{d_s} \mathcal{F}_s(\theta)^{q_s + d_{s+1}}}_{\text{realigning the } s} \text{ (since } \mathcal{F}_0(\theta) = 1).$$

Now, substituting $F_s(\theta) = \prod_{j=1}^{s}(1-h_j(\theta))$ (see equation (6.2)) into the above gives

$$\begin{aligned} L_n(\theta) &= \prod_{s=1}^{n} h_s(\theta)^{d_s}\left[\prod_{j=1}^{s}(1-h_j(\theta))\right]^{q_s+d_{s+1}} \\ &= \prod_{s=1}^{n} h_s(\theta)^{d_s}\prod_{j=1}^{s}(1-h_j(\theta))^{q_s+d_{s+1}} \end{aligned}$$

Rearranging the multiplication we see that $h_1(\theta)$ is multiplied by $(1-h_1(\theta))^{\sum_{i=1}(q_i+d_{i+1})}$, $h_2(\theta)$ is multiplied by $(1-h_1(\theta))^{\sum_{i=2}(q_i+d_{i+1})}$ and so forth. Thus

$$L_n(\theta) = \prod_{s=1}^{n} h_s(\theta)^{d_s}\,(1-h_s(\theta))^{\sum_{m=s}^{n}(q_m+d_{m+1})}\,.$$

Recall $N_s = \sum_{m=s}^{n}(q_m+d_m)$. Thus $\sum_{m=s}^{\infty}(q_m+d_{m+1}) = N_s - d_s$ (number survived just before time t_{s+1}) the likelihood can be rewritten as

$$\begin{aligned} L_n(\theta) &= \prod_{s=1}^{\infty} p_s(\theta)^{d_s}[\sum_{j=s}^{\infty} p_j(\theta)]^{q_s} \\ &= \prod_{s=1}^{\infty} h_s(\theta)^{d_s}(1-h_s(\theta))^{N_s-d_s}. \end{aligned}$$

The corresponding log-likelihood is

$$\begin{aligned} \mathcal{L}_n(\theta) &= \sum_{s=1}^{\infty}\left\{d_s\log p_s(\theta)^{d_s} + \log\left[\sum_{j=s}^{\infty} p_j(\theta)]^{q_s}\right]\right\} \\ &= \sum_{s=1}^{\infty}\Big(d_s\log h_s(\theta) + (N_s-d_s)\log(1-h_s(\theta))\Big). \end{aligned} \tag{6.9}$$

Remark 6.1.3 *At time t_s the number of "failures" is d_s and the number of survivors is $N_s - d_s$. The probability of "failure" and "success" is*

$$h_s(\theta) = P(T=s|T\geq s) = \frac{p_s(\theta)}{\sum_{i=s}^{\infty} p_s(\theta)} \qquad 1-h_s(\theta) = P(T>s|T\geq s) = \frac{\sum_{i=s+1}^{\infty} p_i(\theta)}{\sum_{i=s}^{\infty} p_s(\theta)}.$$

Thus $h_s(\theta)^{d_s}(1-h_s(\theta))^{N_s-d_s}$ can be viewed as the probability of d_s failures and $N_s - d_s$ successes at time t_s.

Thus for the discrete time case the mle of θ is the parameter which maximises the above likelihood.

6.2 Nonparametric estimators of the hazard function - the Kaplan-Meier estimator

Let us suppose that $\{T_i\}$ are iid random variables with distribution function F and survival function $\mathcal{F}$. However, we do not know the class of functions from which F or $\mathcal{F}$ may come from. Instead, we want to estimate $\mathcal{F}$ nonparametrically, in order to obtain a good idea of the 'shape' of the survival function. Once we have some idea of its shape, we can conjecture the parametric family which may best fit its shape. See `https://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator` for some plots.

If the survival times have *not* been censored the 'best' nonparametric estimator of the cumulative distribution function F is the empirical likelihood

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x \leq T_i).$$

Using the above the empirical survival function $\mathcal{F}(x)$ is

$$\hat{\mathcal{F}}_n(x) = 1 - \hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > x),$$

observe this is a left continuous function (meaning that the limit $\lim_{0<\delta\to 0} F_n(x-\delta)$ exists). We use the notation

$$\begin{aligned} d_s &= \text{number of failures at time } t_s \\ N_s &= n - \sum_{i=1}^{s-1} d_i = \sum_{i=s}^{\infty} d_i = N_{s-1} - d_{s-1} \text{ (corresponds to number of survivals just before } t_s). \end{aligned}$$

If $t_s < x \leq t_{s+1}$, then the empirical survival function can be rewritten as

$$\hat{\mathcal{F}}_n(x) = \frac{N_s - d_s}{n} = \prod_{i=1}^{s}\left(\frac{N_i - d_i}{N_i}\right) = \prod_{i=1}^{s}\left(1 - \frac{d_i}{N_i}\right) \qquad x \in (t_i, t_{i+1}]$$

where $N_1 = n$ are the total in the group. Since the survival times usually come from continuous random variable, $d_i = \{0, 1\}$, the above reduces to

$$\hat{\mathcal{F}}_n(x) = \prod_{i=1}^{s}\left(1 - \frac{1}{N_i}\right)^{d_i} \qquad x \in (t_i, t_{i+1}].$$

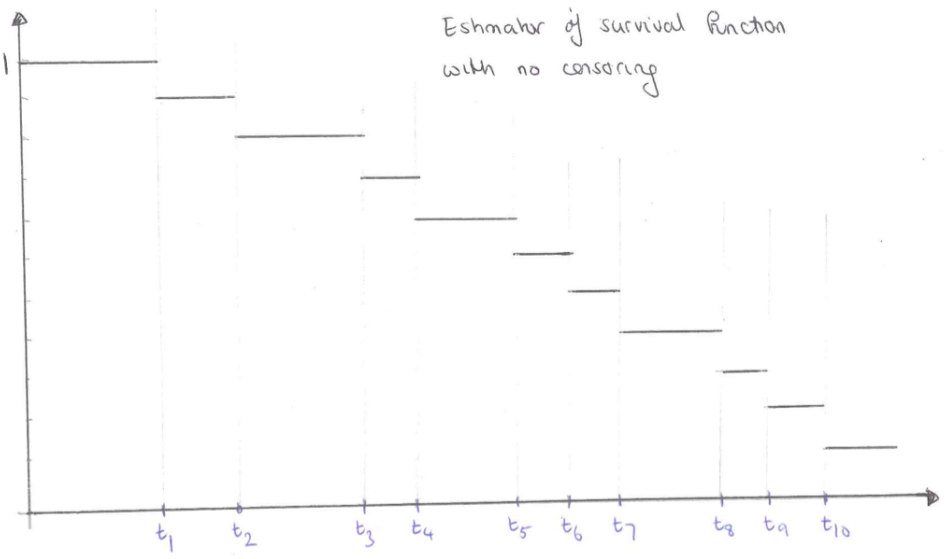

Figure 6.1: The noparametric estimator of the survival function based on the empirical distribution function (with no censoring).

However, in the case, that the survival data is censored and we observe $\{Y_i, \delta_i\}$, then some adjustments have to made to $\hat{\mathcal{F}}_n(x)$ to ensure it is a consistent estimator of the survival function. This leads to the Kaplan-Meier estimator, which is a nonparametric estimator of the survival function $\mathcal{F}$ that takes into account censoring. We will now derive the Kaplan-Meier estimator for discrete data. A typical data set looks like this:

Time	No. Failures at time t_i	No. censored at time t_i	Total Number
t_1	0	0	$N_1 = \sum_{i=1}^{\infty} (d_i + q_i)$
t_2	1	0	$N_2 = \sum_{i=2}^{\infty} (d_i + q_i)$
t_3	0	1	$N_3 = \sum_{i=3}^{\infty} (d_i + q_i)$
t_4	0	1	$N_4 = \sum_{i=4}^{\infty} (d_i + q_i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

It is important to note that because these are usually observations from a continuous random variable and observation which as been censored at time t_{s-1} may not have survived up to time t_s. This means that we cannot say that the total number of survivors at time $t_s - \varepsilon$ is $N_s + q_{s-1}$, all we know for sure is that the number of survivors at $t_s - \varepsilon$ is N_s.

The Kaplan-Meier estimator of the hazard function $h_s = P(T = t_s)/P(T > t_s - \varepsilon)$ is

$$\widehat{h}_s = \frac{d_s}{N_s},$$

where d_s are the number of failures at time t_s and N_s are the number of survivors just before time t_s (think $t_s - \varepsilon$). The corresponding estimator of the survival function $P(T > t_s) = \mathcal{F}(t_s)$ is

$$\hat{\mathcal{F}}(t_s) = \prod_{j=1}^{s} \left(1 - \frac{d_j}{N_j}\right).$$

We show below that this estimator maximises the likelihood and in many respects, this is a rather intuitive estimator of the hazard function. For example, if there is *no censoring* then it can be shown that maximum likelihood estimator of the hazard function is

$$\widehat{h}_s = \frac{d_s}{\sum_{i=s}^{\infty} d_s} = \frac{\text{number of failures at time } s}{\text{number who survive just before time } s},$$

which is a very natural estimator (and is equivalent to the nonparametric MLE estimator discussed in Section **??**).

For continuous random variables, $d_j \in \{0, 1\}$ (as it is unlikely two or more survival times are identical), the Kaplan-Meier estimator can be extended to give

$$\hat{\mathcal{F}}(t) = \prod_{j;t>Y_j} \left(1 - \frac{1}{N_j}\right)^{d_j},$$

where Y_j is the time of an event (either failure or censor) and d_j is an indicator on whether it is a failure. One way of interpreting the above is that only the failures are recorded in the product, the censored times simply appear in the number N_j. Most statistical software packages will plot of the survival function estimator. A plot of the estimator is given in Figure 6.2.

We observe that in the case that the survival data is not censored then $N_j = \sum_{s=j}^{m} d_s$, and the Kaplan-Meier estimator reduces to

$$\hat{\mathcal{F}}(t) = \prod_{j;t>Y_j} \left(1 - \frac{1}{N_j}\right).$$

Comparing the estimator of the survival function with and without censoring (compare Figures 6.1 and 6.2) we see that one major difference is the difference between step sizes. In the case there is no censoring the difference between steps in the step function is always n^{-1} whereas when censoring arises the step differences change according to the censoring.

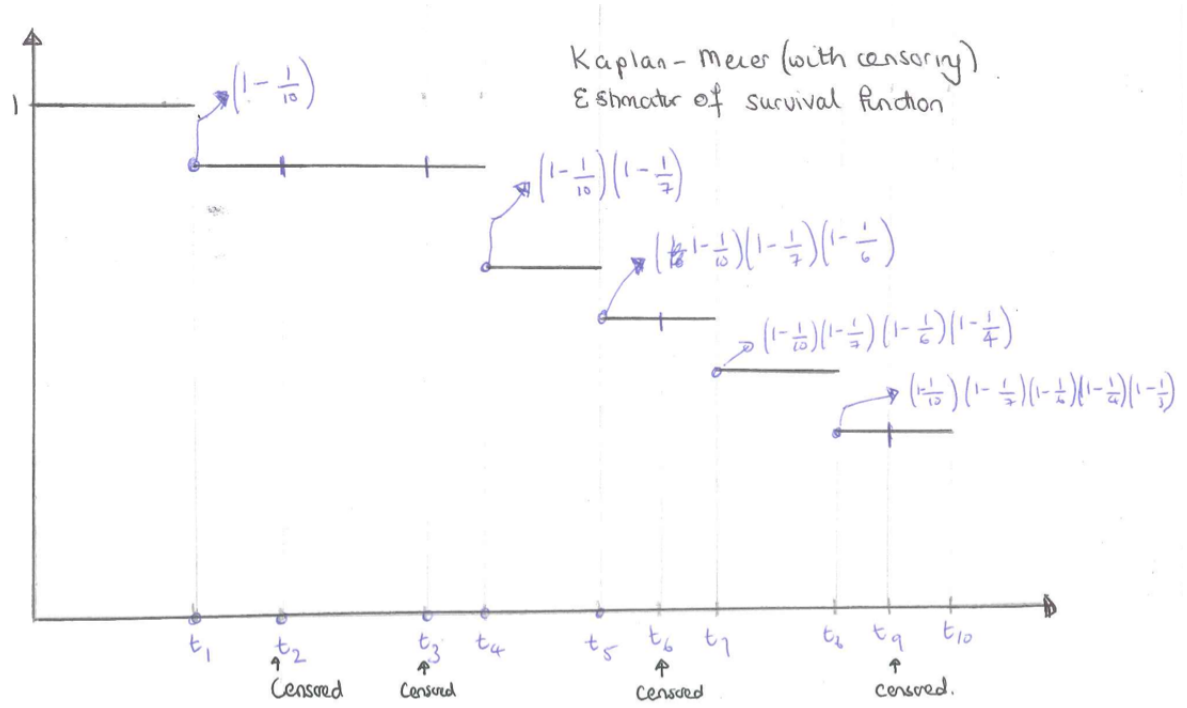

Figure 6.2: An example of the Kaplan-Meier estimator with censoring. The small vertical lines in the plot correspond to censored times.

Derivation of the Kaplan-Meier estimator

We now show that the Kaplan-Meier estimator is the maximum likelihood estimator in the case of censoring. In Section **??** we showed that the empirical distribution is the maximimum of the likelihood for non-censored data. We now show that the Kaplan-Meier estimator is the maximum likelihood estimator when the data is censored. We recall in Section 6.1.5 that the discrete log-likelihood for censored data is

$$\begin{aligned}\mathcal{L}_n(\theta) &= \sum_{s=1}\left(d_s \log p_s(\theta)^{d_s} + q_s \log[\sum_{j=s}^{\infty} p_j(\theta)]\right) \\ &= \sum_{s=1}^{\infty}\left(d_s \log h_s(\theta) + (N_s - d_s)\log(1-h_s(\theta)\right),\end{aligned}$$

where $P(T = t_s) = p_s(\theta)$, d_s are the number of failures at time t_s, q_s are the number of individuals censored at time t_s and $N_s = \sum_{m=s}^{\infty}(q_m + d_m)$. Now the above likelihood is constructed under the assumption that the distribution has a parametric form and the

only unknown is θ. Let us suppose that the probabilities p_s do not have a parametric form. In this case the likelihood is

$$\mathcal{L}_n(p_1, p_2, \ldots) = \sum_{s=1} \left(d_s \log p_s + q_s \log[\sum_{j=s}^{\infty} p_j] \right)$$

subject to the condition that $\sum p_j = 1$. However, it is quite difficult to directly maximise the above. Instead we use the likelihood rewritten in terms of the hazard function (recall equation (6.9))

$$\mathcal{L}_n(h_1, h_2, \ldots) = \sum_{s=1} \Big(d_s \log h_s + (N_s - d_s) \log(1 - h_s) \Big),$$

and maximise this. The derivative of the above with respect to h_s is

$$\frac{\partial \mathcal{L}_n}{\partial h_s} = \frac{d_s}{h_s} - \frac{(N_s - d_s)}{1 - h_s}.$$

Hence by setting the above to zero and solving for h_s gives

$$\hat{h}_s = \frac{d_s}{N_s}.$$

If we recall that d_s = number of failures at time t_s and N_s = number of alive just before time t_s. Hence the non-parametric estimator of the hazard function is rather logical (since the hazard function is the chance of failure at time t, given that no failure has yet occured, ie. $h(t_i) = P(t_i \leq T < t_{i+1} | T \geq t_i)$). Now recalling (6.2) and substituting $\hat{h}_s$ into (6.2) gives the survival function estimator

$$\hat{\mathcal{F}}_s = \prod_{j=1}^{s} \big(1 - \hat{h}_j\big).$$

Rewriting the above, we have the Kaplan-Meier estimator

$$\hat{\mathcal{F}}(t_s) = \prod_{j=1}^{s} \big(1 - \frac{d_j}{N_j}\big).$$

For continuous random variables, $d_j \in \{0, 1\}$ (as it is unlikely two or more survival times are identical), the Kaplan-Meier estimator cab be extended to give

$$\hat{\mathcal{F}}(t) = \prod_{j; Y_j \leq t} \big(1 - \frac{1}{N_j}\big)^{d_j}.$$

Of course given an estimator it is useful to approximate its variance. Some useful approximations are given in Davison (2002), page 197.

6.3 Problems

6.3.1 Some worked problems

Problem: Survival times and random censoring

Example 6.3.1 *Question*

Let us suppose that T and C are exponentially distributed random variables, where the density of T is $\frac{1}{\lambda}\exp(-t/\lambda)$ and the density of C is $\frac{1}{\mu}\exp(-c/\mu)$.

(i) Evaluate the probability $P(T-C<x)$, where x is some finite constant.

(ii) Let us suppose that $\{T_i\}_i$ and $\{C_i\}_i$ are iid survival and censoring times respectively (T_i and C_i are independent of each other), where the densities of T_i and C_i are $f_T(t;\lambda)=\frac{1}{\lambda}\exp(-t/\lambda)$ and $f_C(c;\mu)=\frac{1}{\mu}\exp(-c/\mu)$ respectively. Let $Y_i=\min(T_i,C_i)$ and $\delta_i=1$ if $Y_i=T_i$ and zero otherwise. Suppose λ and μ are unknown. We use the following "likelihood" to estimate λ

$$\mathcal{L}_n(\lambda)=\sum_{i=1}^{n}\delta_i\log f_T(Y_i;\lambda)+\sum_{i=1}^{n}(1-\delta_i)\log\mathcal{F}_T(Y_i;\lambda),$$

where $\mathcal{F}_T$ denotes is the survival function.

Let $\hat{\lambda}_n=\arg\max\mathcal{L}_n(\lambda)$. Show that $\hat{\lambda}_n$ is an asymptotically, unbiased estimator of λ (you can assume that $\hat{\lambda}_n$ converges to some constant).

(iii) Obtain the Fisher information matrix of λ.

(iv) Suppose that $\mu=\lambda$, what can we say about the estimator derived in (ii).

Solutions

(i) $P(T>x)=\exp(-x/\lambda)$ and $P(C>c)=\exp(-c/\mu)$, thus

$$\begin{aligned}P(T<C+x) &= \int \underbrace{P(T<C+x|C=c)}_{\text{use independence}} f_C(c)dc\\ &= \int P(T<c+x)f_C(c)dc\\ &= \int_0^\infty\left[1-\exp(-\frac{c+x}{\lambda})\right]\frac{1}{\mu}\exp(-\frac{c}{\mu})dc = 1-\exp(-x/\lambda)\frac{\lambda}{\lambda+\mu}.\end{aligned}$$

(ii) Differentiating the likelihood

$$\frac{\partial \mathcal{L}_n(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} \delta_i \frac{\partial \log f_T(T_i;\lambda)}{\partial \lambda} + \sum_{i=1}^{n} (1-\delta_i) \frac{\partial \log \mathcal{F}_T(C_i;\lambda)}{\partial \lambda},$$

substituting $f(x;\lambda) = \lambda^{-1}\exp(-x/\lambda)$ *and* $\mathcal{F}(x;\lambda) = \exp(-x/\lambda)$ *into the above and equating to zero gives the solution*

$$\hat{\lambda}_n = \frac{\sum_{i=1} \delta_i T_i + \sum_i (1-\delta_i) C_i}{\sum_i \delta_i}.$$

Now we evaluate the expectaton of the numerator and the denominator.

$$\begin{aligned}
\mathrm{E}(\delta_i T_i) &= \mathrm{E}(T_i I(T_i < C_i)) = \mathrm{E}\left(T_i \mathrm{E}(I(C_i > T_i|T_i))\right) \\
&= \mathrm{E}\left(T_i P(C_i > T_i|T_i)\right) = \mathrm{E}\left(T_i P(C_i - T_i > 0|T_i)\right) \\
&= \mathrm{E}(T_i \exp(-T_i/\mu)) = \int t \exp(-t/\mu) \frac{1}{\lambda} \exp(-t/\lambda) dt \\
&= \frac{1}{\lambda} \times \big(\frac{\mu\lambda}{\mu+\lambda}\big)^2 = \frac{\mu^2\lambda}{(\mu+\lambda)^2}
\end{aligned}$$

Similarly we can show that

$$\mathrm{E}((1-\delta_i)C_i) = P\left(C_i P(T_i > C_i|C_i)\right) = \frac{\mu\lambda^2}{(\mu+\lambda)^2}.$$

Finally, we evaluate the denominator $\mathrm{E}(\delta_i) = P(T < C) = 1 - \frac{\lambda}{\mu+\lambda} = \frac{\mu}{\mu+\lambda}$. *Therefore by Slutsky's theorem we have*

$$\hat{\lambda}_n \xrightarrow{\mathcal{P}} \frac{\frac{\mu\lambda^2}{(\mu+\lambda)^2} + \frac{\mu^2\lambda}{(\mu+\lambda)^2}}{\frac{\mu}{\mu+\lambda}} = \lambda.$$

Thus $\hat{\lambda}_n$ *converges in probability to* λ.

(iii) Since the censoring time does not depend on λ *the Fisher information of* λ *is*

$$\begin{aligned}
I(\lambda) &= n\mathrm{E}\left(-\frac{\partial^2 \mathcal{L}_{n,R}(\lambda)}{\partial \lambda^2}\right) = n\mathrm{E}\left(-\frac{\partial^2 \mathcal{L}_n(\lambda)}{\partial \lambda^2}\right) = \frac{n}{\lambda^2}\mathrm{E}[\delta_i] = \frac{n}{\lambda^2} P(T < C) \\
&= \frac{n}{\lambda^2}\frac{\mu}{\lambda+\mu},
\end{aligned}$$

where $\mathcal{L}_{N,R}$ *is defined in (6.8). Thus we observe, the larger the average censoring time* μ *the more information the data contains about* λ.

(iv) It is surprising, but the calculations in (ii) show that even when $\mu = \lambda$ (but we require that T and C are independent), the estimator defined in (ii) is still a consistent estimator of λ. However, because we did not use $\mathcal{L}_{n,R}$ to construct the maximum likelihood estimator and the maximum of $\mathcal{L}_n(\lambda)$ and $\mathcal{L}_{n,R}(\lambda)$ are not necessarily the same, the estimator will not have optimal (smallest) variance.

Problem: survival times and fixed censoring

Example 6.3.2 *Question*

Let us suppose that $\{T_i\}_{i=1}^n$ are survival times which are assumed to be iid (independent, identically distributed) random variables which follow an exponential distribution with density $f(x;\lambda) = \frac{1}{\lambda}\exp(-x/\lambda)$, where the parameter λ is unknown. The survival times may be censored, and we observe $Y_i = \min(T_i, c)$ and the dummy variable $\delta_i = 1$, if $Y_i = T_i$ (no censoring) and $\delta_i = 0$, if $Y_i = c$ (if the survival time is censored, thus c is known).

(a) State the censored log-likelihood for this data set, and show that the estimator of λ is

$$\hat{\lambda}_n = \frac{\sum_{i=1}^n \delta_i T_i + \sum_{i=1}^n (1-\delta_i)c}{\sum_{i=1}^n \delta_i}.$$

(b) By using the above show that when $c > 0$, $\hat{\lambda}_n$ is a consistent of the the parameter λ.

(c) Derive the (expected) information matrix for this estimator and comment on how the information matrix behaves for various values of c.

Solution

(1a) Since $P(Y_i \geq c) = \exp(-c\lambda)$, the log likelihood is

$$\mathcal{L}_n(\lambda) = \sum_{i=1}^n \Big(\delta_i \log\lambda - \delta_i \lambda Y_i - (1-\delta_i)c\lambda \Big).$$

Thus differentiating the above wrt λ and equating to zero gives the mle

$$\hat{\lambda}_n = \frac{\sum_{i=1}^n \delta_i T_i + \sum_{i=1}^n (1-\delta_i)c}{\sum_{i=1}^n \delta_i}.$$

(b) To show that the above estimator is consistent, we use Slutsky's lemma to obtain

$$\hat{\lambda}_n \xrightarrow{\mathcal{P}} \frac{\mathrm{E}\big[\delta T + (1-\delta)c\big]}{\mathrm{E}(\delta)}$$

To show tha $\lambda = \frac{\mathrm{E}[\delta T + (1-\delta)c]}{\mathrm{E}(\delta)}$ *we calculate each of the expectations:*

$$\begin{aligned}
\mathrm{E}(\delta T) &= \int_0^c y \frac{1}{\lambda} \exp(-\lambda y) dy = c\exp(-c/\lambda) - \frac{1}{\lambda}\exp(-c/\lambda) + \lambda \\
\mathrm{E}((1-\delta)c) &= cP(Y > c) = c\exp(-c/\lambda) \\
\mathrm{E}(\delta) &= P(Y \leq c) = 1 - \exp(-c/\lambda).
\end{aligned}$$

Substituting the above into gives $\hat{\lambda}_n \xrightarrow{\mathcal{P}} \lambda$ *as* $n \to \infty$.

(iii) To obtain the expected information matrix we differentiate the likelihood twice and take expections to obtain

$$I(\lambda) = -n\mathrm{E}\left(\delta_i \lambda^{-2}\right) = -\frac{1-\exp(-c/\lambda)}{\lambda^2}.$$

Note that it can be shown that for the censored likelihood $\mathrm{E}(\frac{\partial \mathcal{L}_n(\lambda)}{\partial})^2 = -\mathrm{E}(\frac{\partial^2 \mathcal{L}_n(\lambda)}{\partial \lambda^2})$. *We observe that the larger* c*, the larger the information matrix, thus the smaller the limiting variance.*

6.3.2 Exercises

Exercise 6.1 *If* $\{\mathcal{F}_i\}_{i=1}^n$ *are the survival functions of independent random variables and* $\beta_1 > 0, \ldots, \beta_n > 0$ *show that* $\prod_{i=1}^n \mathcal{F}_i(x)^{\beta_i}$ *is also a survival function and find the corresponding hazard and cumulative hazard functions.*

Exercise 6.2 *Let* $\{Y_i\}_{i=1}^n$ *be iid random variables with hazard function* $h(x) = \lambda$ *subject to type I censoring at time* c*.*

Show that the observed information for λ *is* m/λ^2 *where* m *is the number of* Y_i *that are non-censored and show that the expected information is* $I(\lambda|c) = n[1 - e^{-\lambda c}]/\lambda^2$*.*

Suppose that the censoring time c *is a reailsation from a random variable* C *whose density is*

$$f(c) = \frac{(\lambda\alpha)^\nu c^\nu}{\Gamma(\nu)} \exp(-c\lambda\alpha) \qquad c > 0, \alpha, \nu > 0.$$

Show that the expected information for λ after averaging over c is

$$I(\lambda) = n\left[1 - (1 + 1/\alpha)^{-\nu}\right]/\lambda^2.$$

Consider what happens when

(i) $\alpha \to 0$

(ii) $\alpha \to \infty$

(iii) $\alpha = 1$ *and* $\nu = 1$

(iv) $\nu \to \infty$ *but such that* $\mu = \nu/\alpha$ *is kept fixed.*

In each case explain quantitively the behaviour of $I(\lambda)$.

Exercise 6.3 *Let us suppose that* $\{T_i\}_i$ *are the survival times of lightbulbs. We will assume that* $\{T_i\}$ *are iid random variables with the density* $f(\cdot;\theta_0)$ *and survival function* $\mathcal{F}(\cdot;\theta_0)$*, where* θ_0 *is unknown. The survival times are censored, and* $Y_i = \min(T_i, c)$ *and* δ_i *are observed* ($c > 0$)*, where* $\delta_i = 1$ *if* $Y_i = T_i$ *and is zero otherwise.*

(a) *(i) State the log-likelihood of* $\{(Y_i, \delta_i)\}_i$.

(ii) We denote the above log-likelihood as $\mathcal{L}_n(\theta)$*. Show that*

$$-\mathrm{E}\left(\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\theta=\theta_0}\right) = \mathrm{E}\left(\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta=\theta_0}\right)^2,$$

stating any important assumptions that you may use.

(b) Let us suppose that the above survival times satisfy a Weibull distribution $f(x;\phi,\alpha) = (\frac{\alpha}{\phi})(\frac{x}{\phi})^{\alpha-1}\exp(-(x/\phi)^\alpha)$ *and as in part (a) we observe and* $Y_i = \min(T_i, c)$ *and* δ_i*, where* $c > 0$.

(i) Using your answer in part 2a(i), give the log-likelihood of $\{(Y_i, \delta_i)\}_i$ *for this particular distribution (we denote this as* $\mathcal{L}_n(\alpha, \phi)$*) and derive the profile likelihood of* α *(profile out the nusiance parameter* ϕ*).*

Suppose you wish to test $H_0 : \alpha = 1$ *against* $H_A : \alpha \neq 1$ *using the log-likelihood ratio test, what is the limiting distribution of the test statistic under the null?*

(ii) Let $\hat{\phi}_n, \hat{\alpha}_n = \arg\max \mathcal{L}_n(\alpha, \phi)$ *(maximum likelihood estimators involving the censored likelihood). Do the estimators* $\hat{\phi}_n$ *and* $\hat{\alpha}_n$ *converge to the true parameters* ϕ *and* α *(you can assume that* $\hat{\phi}_n$ *and* $\hat{\alpha}_n$ *converge to some parameters, and your objective is to find whether these parameters are* ϕ *and* α*).*

(iii) Obtain the (expected) Fisher information matrix of maximum likelihood estimators.

(iv) Using your answer in part 2b(iii) derive the limiting variance of the maximum likelihood estimator of $\hat{\alpha}_n$*.*

Exercise 6.4 *Let* T_i *denote the survival time of an electrical component. It is known that the regressors* x_i *influence the survival time* T_i*. To model the influence the regressors have on the survival time the Cox-proportional hazard model is used with the exponential distribution as the baseline distribution and* $\psi(x_i; \beta) = \exp(\beta x_i)$ *as the link function. More precisely the survival function of* T_i *is*

$$\mathcal{F}_i(t) = \mathcal{F}_0(t)^{\psi(x_i;\beta)},$$

where $\mathcal{F}_0(t) = \exp(-t/\theta)$*. Not all the survival times of the electrical components are observed, and there can arise censoring. Hence we observe* $Y_i = \min(T_i, c_i)$*, where* c_i *is the censoring time and* δ_i*, where* δ_i *is the indicator variable, where* $\delta_i = 0$ *denotes censoring of the* i*th component and* $\delta_i = 1$ *denotes that it is not censored. The parameters* β *and* θ *are unknown.*

(i) Derive the log-likelihood of $\{(Y_i, \delta_i)\}$*.*

(ii) Compute the profile likelihood of the regression parameters β*, profiling out the baseline parameter* θ*.*

Chapter 7

The Expectation-Maximisation Algorithm

7.1 The EM algorithm - a method for maximising the likelihood

Let us suppose that we observe $\underline{Y} = \{Y_i\}_{i=1}^n$. The joint density of $\underline{Y}$ is $f(\underline{Y}; \theta_0)$, and θ_0 is an unknown parameter. Our objective is to estimate θ_0. The log-likelihood of $\underline{Y}$ is

$$\mathcal{L}_n(\underline{Y}; \theta) = \log f(\underline{Y}; \theta),$$

Observe, that we have not specified that $\{Y_i\}$ are iid random variables. This is because the procedure that we will describe below is very general and the observations do not need to be either independent or identically distributed (indeed an interesting extension of this procedure, is to time series with missing data first proposed in Shumway and Stoffer (1982) and Engle and Watson (1982)). Our objective is to estimate θ_0, in the situation where either evaluating the log-likelihood $\mathcal{L}_n$ or maximising $\mathcal{L}_n$ is difficult. Hence an alternative means of maximising $\mathcal{L}_n$ is required. Often, there may exist unobserved data $\{\underline{U} = \{U_i\}_{i=1}^m\}$, where the likelihood of $(\underline{Y}, \underline{U})$ can be 'easily' evaluated. It is through these unobserved data that we find an alternative method for maximising $\mathcal{L}_n$.

The EM-algorithm was specified in its current form in Dempster, Laird and Rubin (1977)(`https://www.jstor.org/stable/pdf/2984875.pdf`) however it was applied previously to several specific models.

Example 7.1.1 *(i) Suppose that* $\{f_j(\cdot;\theta);\theta\}_{j=1}^m$ *are a sequence of densities from* m *exponential classes of densities. In Sections 1.6 and 1.6.5 we showed that it was straightforward to maximise each of these densities. However, let us suppose that each* $f_j(\cdot;\theta)$ *corresponds to one subpopulation. All the populations are pooled together and given an observation* X_i *it is unknown which population it comes from. Let* δ_i *denote the subpopulation the individual* X_i *comes from i.e.* $\delta_i \in \{1,\ldots,m\}$ *where* $P(\delta_i = j) = p_j$.

The density of all these mixtures of distribution is

$$f(x;\theta) = \sum_{j=1}^m f(X_i = x|\delta_i = j)P(\delta_i = j) = \sum_{j=1}^m p_j f_j(x;\theta)$$

where $\sum_{j=1}^m p_j = 1$. *Thus the log-likelihood of* $\{X_i\}$ *is*

$$\sum_{i=1}^n \log\left(\sum_{j=1}^m p_j f_j(X_i;\theta)\right).$$

Of course we require that $\sum_{j=1}^m p_j = 1$, *thus we include a lagrange multiplier to the likelihood to ensure this holds*

$$\sum_{i=1}^n \log\left(\sum_{j=1}^m p_j f_j(X_i;\theta)\right) + \lambda\left(\sum_{j=1}^m p_j - 1\right).$$

It is straightforward to maximise the likelihood for each individual subpopulation, however, it is extremely difficult to maximise the likelihood of this mixture of distributions.

The data $\{X_i\}$ *can be treated as missing, since the information* $\{\delta_i\}$ *about the which population each individual belongs to is not there. If* δ_i *were known the likelihood of* $\{X_i,\delta_i\}$ *is*

$$\prod_{j=1}^m\prod_{i=1}^n (p_j f_j(X_i;\theta))^{I(\delta_i=j)} = \prod_{i=1}^n p_{\delta_i} f_{\delta_i}(X_i;\theta)$$

which leads to the log-likelihood of $\{X_i,\delta_i\}$ *which is*

$$\sum_{i=1}^n \log p_{\delta_i} f_{\delta_i}(X_i;\theta) = \sum_{i=1}^n (\log p_{\delta_i} + \log f_{\delta_i}(X_i;\theta))$$

which is far easier to maximise. Again to ensure that $\sum_{j=1}^{m} p_j = 1$ *we include a Lagrange multiplier*

$$\sum_{i=1}^{n} \log p_{\delta_i} f_{\delta_i}(X_i;\theta) = \sum_{i=1}^{n} (\log p_{\delta_i} + \log f_{\delta_i}(X_i;\theta)) + \lambda \left(\sum_{j=1}^{m} p_j - 1 \right).$$

It is easy to show that $\widehat{p}_j = n^{-1} \sum_{i=1}^{n} I(\delta_i = j)$.

(ii) Let us suppose that $\{T_i\}_{i=1}^{n+m}$ *are iid survival times, with density* $f(x;\underline{\theta}_0)$. *Some of these times are censored and we observe* $\{Y_i\}_{i=1}^{n+m}$, *where* $Y_i = \min(T_i, c)$. *To simplify notation we will suppose that* $\{Y_i = T_i\}_{i=1}^{n}$, *hence the survival time for* $1 \leq i \leq n$, *is observed but* $Y_i = c$ *for* $n+1 \leq i \leq n+m$. *Using the results in Section the log-likelihood of* $\underline{Y}$ *is*

$$\mathcal{L}_n(\underline{Y};\theta) = \left(\sum_{i=1}^{n} \log f(Y_i;\theta) \right) + \left(\sum_{i=n+1}^{n+m} \log \mathcal{F}(Y_i;\theta) \right).$$

The observations $\{Y_i\}_{i=n+1}^{n+m}$ *can be treated as if they were missing. Define the 'complete' observations* $\underline{U} = \{T_i\}_{i=n+1}^{n+m}$, *hence* $\underline{U}$ *contains the unobserved survival times. Then the likelihood of* $(\underline{Y}, \underline{U})$ *is*

$$\mathcal{L}_n(\underline{Y}, \underline{U};\theta) = \sum_{i=1}^{n+m} \log f(T_i;\theta).$$

If no analytic express exists for the survival function $\mathcal{F}$, *it is easier to maximise* $\mathcal{L}_n(\underline{Y}, \underline{U})$ *than* $\mathcal{L}_n(\underline{Y})$.

We now formally describe the EM-algorithm. As mentioned in the discussion above it is often easier to maximise the joint likelihood of $(\underline{Y}, \underline{U})$ than with the likelihood of $\underline{Y}$ itself. the EM-algorithm is based on maximising an approximation of $(\underline{Y}, \underline{U})$ based on the data that is observed $\underline{Y}$.

Let us suppose that the joint likelihood of $(\underline{Y}, \underline{U})$ is

$$\mathcal{L}_n(\underline{Y}, \underline{U};\theta) = \log f(\underline{Y}, \underline{U};\theta).$$

This likelihood is often called the complete likelihood, we will assume that if $\underline{U}$ were known, then this likelihood would be easy to obtain and differentiate. We will assume

that the density $f(\underline{U}|\underline{Y};\theta)$ is also known and is easy to evaluate. By using Bayes theorem it is straightforward to show that

$$\begin{aligned} \log f(\underline{Y},\underline{U};\theta) &= \log f(\underline{Y};\theta) + \log f(\underline{U}|\underline{Y};\theta) \\ \Rightarrow \mathcal{L}_n(\underline{Y},\underline{U};\theta) &= \mathcal{L}_n(\underline{Y};\theta) + \log f(\underline{U}|\underline{Y};\theta). \end{aligned} \tag{7.1}$$

Of course, in reality $\log f(\underline{Y},\underline{U};\theta)$ is unknown, because $\underline{U}$ is unobserved. However, let us consider the expected value of $\log f(\underline{Y},\underline{U};\theta)$ given what we observe $\underline{Y}$. That is

$$Q(\theta_0,\theta) = \mathrm{E}\bigg(\log f(\underline{Y},\underline{U};\theta)\big|\underline{Y},\theta_0\bigg) = \int \bigg(\log f(\underline{Y},\underline{u};\theta)\bigg) f(\underline{u}|\underline{Y},\underline{\theta}_0)d\underline{u}, \tag{7.2}$$

where $f(\underline{u}|\underline{Y},\underline{\theta}_0)$ is the conditional distribution of $\underline{U}$ given $\underline{Y}$ and the unknown parameter θ_0. Hence if $f(\underline{u}|\underline{Y},\underline{\theta}_0)$ were known, then $Q(\theta_0,\theta)$ can be evaluated.

Remark 7.1.1 *It is worth noting that* $Q(\theta_0,\theta) = \mathrm{E}\big(\log f(\underline{Y},\underline{U};\theta)\big|\underline{Y},\theta_0\big)$ *can be viewed as the best predictor of the complete likelihood (involving both observed and unobserved data -* $(\underline{Y},\underline{U})$*) given what is observed* $\underline{Y}$*. We recall that the conditional expectation is the best predictor of* U *in terms of mean squared error, that is the function of* Y *which minimises the mean squared error:* $\mathrm{E}(U|Y) = \arg\min_g \mathrm{E}(U-g(Y))^2$.

The EM algorithm is based on iterating $Q(\cdot)$ in such a way that at each step we obtaining an estimator which gives a larger value of $Q(\cdot)$ (and as we will show later, this gives a larger $\mathcal{L}_n(\underline{Y};\theta)$). We describe the EM-algorithm below.

The EM-algorithm:

(i) Define an initial value $\theta_1 \in \Theta$. Let $\theta_* = \theta_1$.

(ii) **The expectation step (The (k+1)-step)**,

For a fixed θ_* evaluate

$$Q(\theta_*,\theta) = \mathrm{E}\bigg(\log f(\underline{Y},\underline{U};\theta)\bigg|\underline{Y},\theta_*\bigg) = \int \big(\log f(\underline{Y},\underline{u};\theta)\big) f(\underline{u}|\underline{Y},\underline{\theta}_*)du,$$

for all $\theta \in \Theta$.

(iii) **The maximisation step**

Evaluate $\theta_{k+1} = \arg\max_{\theta\in\Theta} Q(\theta_*,\theta)$.

We note that the maximisation can be done by finding the solution of

$$\mathrm{E}\left(\frac{\partial \log f(\underline{Y},\underline{U};\theta)}{\partial\theta}\Big|\underline{Y},\theta_*\right)=0.$$

(iv) If θ_k and θ_{k+1} are sufficiently close to each other stop the algorithm and set $\widehat{\theta}_n=\theta_{k+1}$. Else set $\theta_*=\theta_{k+1}$, go back and repeat steps (ii) and (iii) again.

We use $\hat{\theta}_n$ as an estimator of θ_0. To understand why this iteration is connected to the maximising of $\mathcal{L}_n(\underline{Y};\theta)$ and, under certain conditions, gives a good estimator of θ_0 (in the sense that $\hat{\theta}_n$ is close to the parameter which maximises $\mathcal{L}_n$) let us return to (7.1). Taking the expectation of $\log f(\underline{Y},\underline{U};\theta)$, conditioned on $\underline{Y}$ we have

$$\begin{aligned} Q(\theta_*,\theta) &= \mathrm{E}\left(\log f(\underline{Y},\underline{U};\theta)\big|\underline{Y},\theta_*\right) \\ &= \mathrm{E}\left[\log f(\underline{Y};\theta)+\log f(\underline{U}|\underline{Y};\theta)\big|\underline{Y},\theta_*\right] \\ &= \log f(\underline{Y};\theta)+\mathrm{E}\left[\log f(\underline{U}|\underline{Y};\theta)\big|\underline{Y},\theta_*\right]. \end{aligned} \tag{7.3}$$

Define

$$D(\theta_*,\theta) = \mathrm{E}\left(\log f(\underline{U}|\underline{Y};\theta)\big|\underline{Y},\theta_*\right)=\int\left[\log f(\underline{u}|\underline{Y};\theta)\right]f(\underline{u}|\underline{Y},\theta_*)d\underline{u}.$$

Substituting $D(\theta_*,\theta)$ into (7.3) gives

$$Q(\theta_*,\theta)=\mathcal{L}_n(\theta)+D(\theta_*,\theta), \tag{7.4}$$

we use this in expression in the proof below to show that $\mathcal{L}_n(\theta_{k+1})>\mathcal{L}_n(\theta_k)$. First we that at the $(k+1)$th step iteration of the EM-algorithm, θ_{k+1} maximises $Q(\theta_k,\theta)$ over all $\theta\in\Theta$, hence $Q(\theta_k,\theta_{k+1})\geq Q(\theta_k,\theta_k)$ (which will also be used in the proof).

In the lemma below we show that $\mathcal{L}_n(\theta_{k+1})\geq\mathcal{L}_n(\theta_k)$, hence at each iteration of the EM-algorthm we are obtaining a θ_{k+1} which increases the likelihood over the previous iteration.

Lemma 7.1.1 *When running the EM-algorithm the inequality $\mathcal{L}_n(\theta_{k+1})\geq\mathcal{L}_n(\theta_k)$ always holds.*

Furthermore, if $\theta_k\to\widehat{\theta}$ and for every iteration $\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})}=0$, then $\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\widehat{\theta}}=0$ (this point can be a saddle point, a local maximum or the sought after global maximum).

PROOF. From (7.4) it is clear that

$$Q(\theta_k,\theta_{k+1})-Q(\theta_k,\theta_k)=\big[\mathcal{L}_n(\theta_{k+1})-\mathcal{L}_n(\theta_k)\big]+\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big], \qquad (7.5)$$

where we recall

$$D(\theta_1,\theta) \;=\; \mathrm{E}\bigg(\log f(\underline{U}|\underline{Y};\theta)\big|\underline{Y},\theta_1\bigg)=\int\big[\log f(\underline{u}|\underline{Y};\theta)\big]f(\underline{u}|\underline{Y},\theta_1)d\underline{u}.$$

We will show that $\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big]\leq 0$, the result follows from this. We observe that

$$\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big]=\int\log\frac{f(\underline{u}|\underline{Y},\theta_{k+1})}{f(\underline{u}|\underline{Y},\theta_k)}f(\underline{u}|\underline{Y},\theta_k)d\underline{u}.$$

By using the Jenson's inequality (which we have used several times previously)

$$\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big]\leq\log\int f(\underline{u}|\underline{Y},\theta_{k+1})d\underline{u}=0.$$

Therefore, $\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big]\leq 0$. Note that if θ uniquely identifies the distribution $f(\underline{u}|\underline{Y},\theta)$ then equality only happens when $\theta_{k+1}=\theta_k$. Since $\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big]\leq 0$ by (7.5) we have

$$\big[\mathcal{L}_n(\theta_{k+1})-\mathcal{L}_n(\theta_k)\big]\geq Q(\theta_k,\theta_{k+1})-Q(\theta_k,\theta_k)\geq 0.$$

and we obtain the desired result ($\mathcal{L}_n(\theta_{k+1})\geq\mathcal{L}_n(\theta_k)$).

To prove the second part of the result we will use that for all $\theta\in\Theta$

$$\frac{\partial D(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)}=\int\frac{\partial\log f(\underline{u}|\underline{Y};\theta)}{\partial\theta}f(\underline{u}|\underline{Y},\theta)d\underline{u}=\frac{\partial}{\partial\theta}\int f(\underline{u}|\underline{Y},\theta)d\underline{u}=0. \quad (7.6)$$

We will to show that the derivative of the likelihood is zero at θ_* i.e. $\frac{\partial\mathcal{L}}{\partial\theta}\rfloor_{\theta=\theta_*}=0$. To show this we use the identity

$$\mathcal{L}_n(\theta_{k+1})=Q(\theta_k,\theta_{k+1})-D(\theta_k,\theta_{k+1}).$$

Taking derivatives with respect to θ_{k+1} gives

$$\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\theta_{k+1}}=\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})}-\frac{\partial D(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})}.$$

By definition of θ_{k+1}, $\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})}=0$, thus we have

$$\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\theta_{k+1}}=-\frac{\partial D(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})}.$$

Furthermore, since by assumption $\theta_k \to \widehat{\theta}$ this implies that as $k \to \infty$ we have

$$\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta=\widehat{\theta}} = -\frac{\partial D(\theta_1,\theta_2)}{\partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} = 0,$$

which follows from (7.6), thus giving the required result. □

Further information on convergence can be found in Boyles (1983) (`http://www.jstor.org/stable/pdf/2345622.pdf?_=1460485744796`) and Wu (1983) (`https://www.jstor.org/stable/pdf/2240463.pdf?_=1460409579185`).

Remark 7.1.2 *Note that the EM algorithm will converge to a $\widehat{\theta}$ where $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$. The reason can be seen from the identity*

$$Q(\theta_*,\theta) = \mathcal{L}_n(\theta) + D(\theta_*,\theta).$$

The derivative of the above with respect to θ is

$$\frac{\partial Q(\theta_*,\theta)}{\partial \theta} = \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} + \frac{\partial D(\theta_*,\theta)}{\partial \theta}. \tag{7.7}$$

Observe that $D(\theta_*,\theta)$ is maximum only when $\theta = \theta_*$ (for all θ_*, this is clear from the proof above), thus $\frac{\partial D(\theta_*,\theta)}{\partial \theta}_{\theta=\theta_*}$ which for $\theta_* = \widehat{\theta}$ implies $\frac{\partial D(\widehat{\theta},\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$. Furthermore, by definition $\frac{\partial Q(\widehat{\theta},\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$.

Since $\frac{\partial Q(\widehat{\theta},\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$ and $\frac{\partial D(\widehat{\theta},\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$ by using (7.7) this implies $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$.

In order to prove the results in the following section we use the following identities. Since

$$\begin{aligned}
& Q(\theta_1,\theta_2) = \mathcal{L}(\theta_2) + D(\theta_1,\theta_2) \\
\Rightarrow \quad & \frac{\partial^2 Q(\theta_1,\theta_2)}{\partial \theta_2^2} = \frac{\partial^2 \mathcal{L}(\theta_2)}{\partial \theta^2} + \frac{\partial^2 D(\theta_1,\theta_2)}{\partial \theta^2} \\
\Rightarrow \quad & \frac{\partial^2 Q(\theta_1,\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} = \frac{\partial^2 \mathcal{L}(\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} + \frac{\partial^2 D(\theta_1,\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} \\
\Rightarrow \quad & -\frac{\partial^2 \mathcal{L}(\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} = -\frac{\partial^2 Q(\theta_1,\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} + \frac{\partial^2 D(\theta_1,\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)}
\end{aligned} \tag{7.8}$$

We observe that the LHS of the above is the observed Fisher information matrix $I(\theta|\underline{Y})$,

$$-\frac{\partial^2 Q(\theta_1,\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} = I_C(\theta|\underline{Y}) = -\int \frac{\partial^2 \log f(\underline{u},\underline{Y};\theta)}{\partial \theta^2} f(\underline{u}|\underline{Y},\theta)d\underline{u} \tag{7.9}$$

is the complete Fisher information *conditioned* on what is observed and

$$-\frac{\partial^2 D(\theta_1,\theta_2)}{\partial\theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} = I_M(\theta|\underline{Y}) = -\int \frac{\partial^2 \log f(\underline{u}|\underline{Y};\theta)}{\partial\theta^2} f(\underline{u}|\underline{Y},\theta)d\underline{u} \qquad (7.10)$$

is the Fisher information matrix of the unobserved data conditioned on what is observed. Thus

$$I(\theta|\underline{Y}) = I_C(\theta|\underline{Y}) - I_M(\theta|\underline{Y}).$$

7.1.1 Speed of convergence of θ_k to a stable point

When analyzing an algorithm it is instructive to understand how fast it takes to converge to the limiting point. In the case of the EM-algorithm, this means what factors determine the rate at which θ_k converges to a stable point $\widehat{\theta}$ (note this has *nothing* to do with the rate of convergence of an estimator to the true parameter, and it is important to understand this distinction).

The rate of convergence of an algorithm is usually measured by the ratio of the current iteration with the previous iteration:

$$R = \lim_{k\to\infty}\left(\frac{\theta_{k+1}-\widehat{\theta}}{\theta_k-\widehat{\theta}}\right),$$

if the algorithm converges to a limit in a finite number of iterations we place the above limit to zero. Thus the smaller R the faster the rate of convergence (for example if (i) $\theta_k - \widehat{\theta} = k^{-1}$ then $R = 1$ if (ii) $\theta_k - \widehat{\theta} = \rho^k$ then $R = \rho$, assuming $|\rho| < 1$). Note that since $(\theta_{k+1} - \widehat{\theta}) = \prod_{j=1}^{k}\left(\frac{\theta_{j+1}-\widehat{\theta}}{\theta_j-\widehat{\theta}}\right)$, then typically $|R| \leq 1$.

To obtain an approximation of R we will make a Taylor expansion of $\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}$ around the limit $(\theta_1,\theta_2) = (\widehat{\theta},\widehat{\theta})$. To do this we recall that for a bivariate function $f : \mathbb{R}^2 \to \mathbb{R}$ for (x_0, y_0) "close" to (x, y) we have the Taylor expansion

$$f(x,y) = f(x_0,y_0) + (x-x_0)\frac{\partial f(x,y)}{\partial x}\rfloor_{(x,y)=(x_0,y_0)} + (y-y_0)\frac{\partial f(x,y)}{\partial y}\rfloor_{(x,y)=(x_0,y_0)} + \text{lower order terms.}$$

Applying the above to $\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}$ gives

$$\begin{aligned}
&\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})} \\
\approx\ &\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} + (\theta_{k+1}-\widehat{\theta})\frac{\partial^2 Q(\theta_1,\theta_2)}{\partial\theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} + (\theta_k-\widehat{\theta})\frac{\partial^2 Q(\theta_1,\theta_2)}{\partial\theta_1\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}.
\end{aligned}$$

Since θ_{k+1} maximises $Q(\theta_k, \theta)$ and $\widehat{\theta}$ maximises $Q(\widehat{\theta}, \theta)$ within the interior of the parameter space then

$$\frac{\partial Q(\theta_1, \theta_2)}{\partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})} = 0 \text{ and } \frac{\partial Q(\theta_1, \theta_2)}{\partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} = 0$$

This implies that

$$(\theta_{k+1} - \widehat{\theta})\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} + (\theta_k - \widehat{\theta})\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} = 0.$$

Thus

$$\lim_{k\to\infty}\left(\frac{\theta_{k+1} - \widehat{\theta}}{\theta_k - \widehat{\theta}}\right) = -\left(\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}\right)^{-1}\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}. \tag{7.11}$$

This result shows that the rate of convergence depends on the ratio of gradients of $Q(\theta_1, \theta_2)$ around $(\theta_1, \theta_2) = (\widehat{\theta}, \widehat{\theta})$. Some further simplifications can be made by noting that

$$Q(\theta_1, \theta_2) \;=\; \mathcal{L}_n(\theta_2) + D(\theta_1, \theta_2) \quad \Rightarrow \quad \frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2} = \frac{\partial^2 D(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}.$$

Substituting this into (7.11) gives

$$\lim_{k\to\infty}\left(\frac{\theta_{k+1} - \widehat{\theta}}{\theta_k - \widehat{\theta}}\right) = -\left(\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}\right)^{-1}\frac{\partial^2 D(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}. \tag{7.12}$$

To make one further simplification, we note that

$$\begin{aligned}
\frac{\partial^2 D(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} &= \int \frac{1}{f(\underline{u}|\underline{Y}, \theta_2)}\frac{\partial f(\underline{u}|\underline{Y}, \theta_2)}{\partial \theta_2}\frac{\partial f(\underline{u}|\underline{Y}, \theta_1)}{\partial \theta_1}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} d\underline{u} \\
&= \int \frac{1}{f(\underline{u}|\underline{Y}, \theta)}\left(\frac{\partial f(\underline{u}|\underline{Y}, \theta)}{\partial \theta}\right)^2\rfloor_{\theta=\widehat{\theta}} d\underline{u} \\
&= -\int \frac{\partial^2 \log f(\underline{u}|\underline{Y}, \theta)}{\partial \theta_2} f(\underline{u}|\underline{Y}, \theta)\rfloor_{\theta=\widehat{\theta}} d\underline{u}
\end{aligned} \tag{7.13}$$

were the last line of the above follows from the identity

$$\int \frac{1}{f(\underline{x}; \theta)}\left(\frac{\partial f(\underline{x}; \theta)}{\partial \theta}\right)^2 d\underline{x} + \int \frac{\partial^2 \log f(\underline{x}; \theta)}{\partial \theta_2} f(\underline{x}; \theta) d\underline{x} = 0$$

(see the proof of Corollary 1.3.1). Substituting (7.12) into (7.13) gives

$$\lim_{k\to\infty}\left(\frac{\theta_{k+1} - \widehat{\theta}}{\theta_k - \widehat{\theta}}\right) = \left(\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}\right)^{-1}\frac{\partial^2 D(\theta_1, \theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}. \tag{7.14}$$

Substituting (7.9) and (7.10) into the above gives

$$\lim_{k\to\infty}\left(\frac{\theta_{k+1}-\widehat{\theta}}{\theta_k-\widehat{\theta}}\right)=I_C(\theta|\underline{Y})^{-1}I_M(\theta|\underline{Y}) \tag{7.15}$$

Hence the rate of convergence of the algorithm depends on on the ratio $I_C(\theta|\underline{Y})^{-1}I_M(\theta|\underline{Y})$. The closer the largest eigenvalue of $I_C(\theta|\underline{Y})^{-1}I_M(\theta|\underline{Y})$ to one, the slower the rate of convergence, and a larger number of iterations are required. The heuristic of this result is that if the missing information is a large proportion of the complete or total information than this ratio will be large.

Further details can be found in Dempster et. al. (1977) pages 9-10 and Meng and Rubin (1994) (`http://www.sciencedirect.com/science/article/pii/0024379594903638`).

7.2 Applications of the EM algorithm

7.2.1 Censored data

Let us return to the example at the start of this section, and construct the EM-algorithm for censored data. We recall that the log-likelihoods for censored data and complete data are

$$\mathcal{L}_n(\underline{Y};\theta)=\Big(\sum_{i=1}^{n}\log f(Y_i;\theta)\Big)+\Big(\sum_{i=n+1}^{n+m}\log\mathcal{F}(Y_i;\theta)\Big).$$

and

$$\mathcal{L}_n(\underline{Y},\underline{U};\theta)=\Big(\sum_{i=1}^{n}\log f(Y_i;\theta)\Big)+\Big(\sum_{i=n+1}^{n+m}\log f(T_i;\theta)\Big).$$

To implement the EM-algorithm we need to evaluate the expectation step $Q(\theta_*,\theta)$. It is easy to see that

$$Q(\theta_*,\theta)=\mathrm{E}\Big(\mathcal{L}_n(\underline{Y},\underline{U};\theta)\big|\underline{Y},\theta_*\Big) \;=\; \Big(\sum_{i=1}^{n}\log f(Y_i;\theta)\Big)+\Big(\sum_{i=n+1}^{n+m}\mathrm{E}\big(\log f(T_i;\theta)|\underline{Y},\theta_*\big)\Big).$$

To obtain $\mathrm{E}\big(\log f(T_i;\theta)|\underline{Y},\theta_*\big)$ $(i\geq n+1)$ we note that

$$\begin{aligned}\mathrm{E}\big(\log f(T_i;\theta)|\underline{Y},\theta_*\big) &= \mathrm{E}(\log f(T_i;\theta)|T_i\geq c)\\ &= \frac{1}{\mathcal{F}(c;\theta)}\int_c^{\infty}\big[\log f(T_i;\theta)\big]f(u;\theta_*)du.\end{aligned}$$

Therefore we have

$$Q(\theta_*,\theta) = \Big(\sum_{i=1}^{n}\log f(Y_i;\theta)\Big) + \frac{m}{\mathcal{F}(c;\theta_*)}\int_c^\infty \big[\log f(T_i;\theta)\big] f(u;\theta_*)du.$$

We also note that the derivative of $Q(\theta_*,\theta)$ with respect to θ is

$$\frac{\partial Q(\theta_*,\theta)}{\partial\theta} = \Big(\sum_{i=1}^{n}\frac{1}{f(Y_i;\theta)}\frac{\partial f(Y_i;\theta)}{\partial\theta}\Big) + \frac{m}{\mathcal{F}(c;\theta_*)}\int_c^\infty \frac{1}{f(u;\theta)}\frac{\partial f(u;\theta)}{\partial\theta} f(u;\theta_*)du.$$

Hence for this example, the EM-algorithm is

(i) Define an initial value $\theta_1 \in \Theta$. Let $\theta_* = \theta_1$.

(ii) **The expectation step**:

For a fixed θ_* evaluate

$$\frac{\partial Q(\theta_*,\theta)}{\partial\theta} = \Big(\sum_{i=1}^{n}\frac{1}{f(Y_i;\theta)}\frac{\partial f(Y_i;\theta)}{\partial\theta}\Big) + \frac{m}{\mathcal{F}(c;\theta_*)}\int_c^\infty \frac{1}{f(u;\theta)}\frac{\partial f(u;\theta)}{\partial\theta} f(u;\theta_*)du.$$

(iii) **The maximisation step**:

Solve for $\frac{\partial Q(\theta_*,\theta)}{\partial\theta}$. Let θ_{k+1} be such that $\frac{\partial Q(\theta_*,\theta)}{\partial\theta}\rfloor_{\theta=\theta_k} = 0$.

(iv) If θ_k and θ_{k+1} are sufficiently close to each other stop the algorithm and set $\hat{\theta}_n = \theta_{k+1}$. Else set $\theta_* = \theta_{k+1}$, go back and repeat steps (ii) and (iii) again.

7.2.2 Mixture distributions

We now consider a useful application of the EM-algorithm, to the estimation of parameters in mixture distributions. Let us suppose that $\{Y_i\}_{i=1}^n$ are iid random variables with density

$$f(y;\theta) = pf_1(y;\theta_1) + (1-p)f_2(y;\theta_2),$$

where $\theta = (p,\theta_1,\theta_2)$ are unknown parameters. For the purpose of identifiability we will suppose that $\theta_1 \neq \theta_2$, $p \neq 1$ and $p \neq 0$. The log-likelihood of $\{Y_i\}$ is

$$\mathcal{L}_n(\underline{Y};\theta) = \sum_{i=1}^{n}\log\big(pf_1(Y_i;\theta_1) + (1-p)f_2(Y_i;\theta_2)\big). \tag{7.16}$$

Now maximising the above can be extremely difficult. As an illustration consider the example below.

Example 7.2.1 *Let us suppose that $f_1(y;\theta_1)$ and $f_2(y;\theta_1)$ are normal densities, then the log likelihood is*

$$\mathcal{L}_n(\underline{Y};\theta) = \sum_{i=1}^{n} \log\left(p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp(-\frac{1}{2\sigma_1^2}(Y_i-\mu_1)^2) + (1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp(-\frac{1}{2\sigma_2^2}(Y_i-\mu_2)^2)\right).$$

We observe this is extremely difficult to maximise. On the other hand if Y_i were simply normally distributed then the log-likelihood is extremely simple

$$\mathcal{L}_n(\underline{Y};\theta) \propto -\sum_{i=1}^{n}\left(\log\sigma_1^2 + \frac{1}{2\sigma_1^2}(Y_i-\mu_1)^2)\right). \tag{7.17}$$

In other words, the simplicity of maximising the log-likelihood of the exponential family of distributions (see Section 1.6) is lost for mixtures of distributions.

We use the EM-algorithm as an indirect but simple method of maximising (7.17). In this example, it is not clear what observations are missing. However, let us consider one possible intepretation of the mixture distribution. Let us define the random variables δ_i and Y_i, where $\delta_i \in \{1,2\}$,

$$P(\delta_i = 1) = p \text{ and } P(\delta_i = 2) = (1-p)$$

and the density of $Y_i|\delta_i = 1$ is f_1 and the density of $Y_i|\delta_i = 1$ is f_2. Based on this definition, it is clear from the above that the density of Y_i is

$$f(y;\theta) = f(y|\delta=1,\theta)P(\delta=1) + f(y|\delta=2,\theta)P(\delta=2) = pf_1(y;\theta_1) + (1-p)f_2(y;\theta_2).$$

Hence, one interpretation of the mixture model is that there is a hidden unobserved random variable which determines the state or distribution of Y_i. A simple example, is that Y_i is the height of an individual and δ_i is the gender. However, δ_i is unobserved and only the height is observed. Often a mixture distribution has a physical interpretation, similar to the height example, but sometimes it can be used to parametrically model a wide class of densities.

Based on the discussion above, $\underline{U} = \{\delta_i\}$ can be treated as the missing observations. The likelihood of (Y_i, U_i) is

$$\big\{p_1 f_1(Y_i;\theta_1)\big\}^{I(\delta_i=1)}\big\{p_2 f_2(Y_i;\theta_2)\big\}^{I(\delta_i=2)} = p_{\delta_i} f_{\delta_i}(Y_i;\theta_{\delta_i}).$$

where we set $p_2 = 1 - p$. Therefore the log likelihood of $\{(Y_i, \delta_i)\}$ is

$$\mathcal{L}_n(\underline{Y}, \underline{U}; \theta) = \sum_{i=1}^{n} \left(\log p_{\delta_i} + \log f_{\delta_i}(Y_i; \theta_{\delta_i})\right).$$

We now need to evaluate

$$Q(\theta_*, \theta) = \mathrm{E}\big(\mathcal{L}_n(\underline{Y}, \underline{U}; \theta)|\underline{Y}, \theta_*\big) = \sum_{i=1}^{n} \left[\mathrm{E}\big(\log p_{\delta_i}|Y_i, \theta_*\big) + \mathrm{E}\big(\log f_{\delta_i}(Y_i; \theta_{\delta_i})|Y_i, \theta_*\big)\right].$$

We see that the above expectation is taken with respect the distribution of δ_i conditioned on Y_i and the parameter θ_* Thus, in general,

$$\mathrm{E}\left(A(Y, \delta)|Y, \theta^*\right) = \sum_j A(Y, \delta = j) P(\delta = j|Y_i, \theta^*),$$

which we apply to $Q(\theta_*, \theta)$ to give

$$Q(\theta_*, \theta) = \sum_j \sum_{i=1}^{n} \left[\log p_{\delta_i = j} + \log f_{\delta_i = j}(Y_i; \theta)\right] P(\delta_i = j|Y_i, \theta^*).$$

Therefore we need to obtain $P(\delta_i = j|Y_i, \theta^*)$. By using conditioning arguments it is easy to see that [1]

$$\begin{aligned}
P(\delta_i = 1|Y_i = y, \theta_*) &= \frac{P(\delta_i = 1, Y_i = y; \theta_*)}{P(Y_i = y; \theta_*)} = \frac{p_* f_1(y, \theta_{1,*})}{p_* f_1(y, \theta_{1,*}) + (1 - p_*) f_2(y, \theta_{2,*})} \\
&:= w_1(\theta_*, y) \\
P(\delta_i = 2|Y_i = y, \theta_*) &= \frac{p_* f_2(y, \theta_{2,*})}{p_* f_1(y, \theta_{1,*}) + (1 - p_*) f_2(y, \theta_{2,*})} \\
&:= w_2(\theta_*, y) = 1 - w_1(\theta_*, y).
\end{aligned}$$

Therefore

$$Q(\theta_*, \theta) = \sum_{i=1}^{n} \Big(\log p + \log f_1(Y_i; \theta_1)\Big) w_1(\theta_*, Y_i) + \sum_{i=1}^{n} \Big(\log(1 - p) + \log f_2(Y_i; \theta_2)\Big) w_2(\theta_*, Y_i).$$

Now maximising the above with respect to p, θ_1 and θ_2 in general will be much easier than maximising $\mathcal{L}_n(\underline{Y}; \theta)$. For this example the EM algorithm is

(i) Define an initial value $\theta_1 \in \Theta$. Let $\theta_* = \theta_1$.

[1] To see why note that $P(\delta_i = 1 \text{ and } Y_i \in [y - h/2, y + h/2]|\theta^*) = h p_* f_1(y)$ and $P(Y_i \in [y - h/2, y + h/2]|\theta^*) = h\left(p_* f_1(y) + (1 - p_*) f_2(y)\right)$.

(ii) **The expectation step**:

For a fixed θ_* evaluate

$$Q(\theta_*,\theta) = \sum_{i=1}^{n}\Big(\log p + \log f_1(Y_i;\theta_1)\Big)w_1(\theta_*,Y_i) + \sum_{i=1}^{n}\Big(\log(1-p) + \log f_2(Y_i;\theta_2)\Big)w_2(\theta_*,Y_i).$$

(iii) **The maximisation step**:

Evaluate $\theta_{k+1} = \arg\max_{\theta\in\Theta} Q(\theta_*,\theta)$ by differentiating $Q(\theta_*,\theta)$ wrt to θ and equating to zero. Since the parameters p and θ_1,θ_2 are in separate subfunctions, they can be maximised separately.

(iv) If θ_k and θ_{k+1} are sufficiently close to each other stop the algorithm and set $\hat{\theta}_n = \theta_{k+1}$. Else set $\theta_* = \theta_{k+1}$, go back and repeat steps (ii) and (iii) again.

Example 7.2.2 (Normal mixtures and mixtures from the exponential family) *(i)* *We briefly outline the algorithm in the case of a mixture two normal distributions. In this case*

$$Q(\theta_*,\theta) = -\frac{1}{2}\sum_{j=1}^{2}\sum_{i=1}^{n} w_j(\theta_*,Y_i)\left(\sigma_j^{-2}(Y_i-\mu_j)^2 + \log\sigma_j^2\right) + \sum_{i=1}^{n} w_j(\theta_*,Y_i)\left(\log p + \log(1-p)\right).$$

By differentiating the above wrt to μ_j, σ_j^2 (for $j=1$ and 2) and p it is straightforward to see that the μ_j, σ_j^2 and p which maximises the above is

$$\widehat{\mu}_j = \frac{\sum_{i=1}^{n} w_j(\theta_*,Y_i)Y_i}{\sum_{i=1}^{n} w_j(\theta_*,Y_i)} \text{ and } \widehat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} w_j(\theta_*,Y_i)(Y_i-\widehat{\mu}_j)^2}{\sum_{i=1}^{n} w_j(\theta_*,Y_i)}$$

and

$$\widehat{p} = \frac{\sum_{i=1}^{n} w_1(\theta_*,Y_i)}{n}.$$

Once these estimators are obtained we let $\theta_* = (\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma}_1^2,\widehat{\sigma}_2^2,\widehat{p})$. The quantities $w_j(\theta_*,Y_i)$ are re-evaluated and $Q(\theta_*,\theta)$ maximised with respect to the new weights.

(ii) In general if Y is a mixture from the exponential family with density

$$f(y;\theta) = \sum_{j=1}^{m} p_j \exp\left(y\theta_j - \kappa_j(\theta_j) + c_j(y)\right)$$

the corresponding $Q(\theta_*, \theta)$ *is*

$$Q(\theta_*, \theta) = \sum_{j=1}^{m} \sum_{i=1}^{n} w_j(\theta_*, Y_i) \left[Y_i\theta_j - \kappa_j(\theta_j) + c_j(Y_i) + \log p_j \right],$$

where

$$w_j(\theta_*, Y_i) = \frac{p_j^* \exp\left(Y_i\theta_j^* - \kappa_j(\theta_j^*) + c_j(Y_i)\right)}{\sum_{k=1}^{m} p_k^* \exp\left(Y_i\theta_k^* - \kappa_k(\theta_k^*) + c_k(Y_i)\right)}$$

subject to the constraint that $\sum_{j=1}^{m} p_j = 1$. *Thus for* $1 \leq j \leq m$, $Q(\theta_*, \theta)$ *is maximised for*

$$\widehat{\theta}_j = \mu_j^{-1} \left(\frac{\sum_{i=1}^{n} w_j(\theta_*, Y_i) Y_i}{\sum_{i=1}^{n} w_j(\theta_*, Y_i)} \right)$$

where $\mu_j = \kappa_j'$ *(we assume all parameter for each exponential mixture is open) and*

$$\widehat{p}_j = \frac{\sum_{i=1}^{n} w_j(\theta_*, Y_i)}{n}.$$

Thus we set $\theta_* = (\{\widehat{\theta}_j, \widehat{p}_j\}_{j=1}^{m})$ *and re-evaluate the weights.*

Remark 7.2.1 *Once the algorithm is terminated, we can calculate the chance that any given observation* Y_i *is in subpopulation* j *since*

$$\widehat{P}(\delta_i = j | Y_i) = \frac{\widehat{p}_j f_j(Y; \widehat{\theta})}{\sum_{j=1}^{2} \widehat{p}_j f_j(Y; \widehat{\theta})}.$$

This allows us to obtain a classifier for each observation Y_i.

It is straightforward to see that the arguments above can be generalised to the case that the density of Y_i is a mixture of m different densities. However, we observe that the selection of m can be quite adhoc. There are methods for choosing m, these include the reversible jump MCMC methods.

7.2.3 Problems

Example 7.2.3 *Question: Suppose that the regressors* x_t *are believed to influence the response variable* Y_t. *The distribution of* Y_t *is*

$$P(Y_t = y) = p\frac{\lambda_{t1}^y \exp(-\lambda_{t1} y)}{y!} + (1-p)\frac{\lambda_{t2}^y \exp(-\lambda_{t2} y)}{y!},$$

where $\lambda_{t1} = \exp(\beta_1' x_t)$ *and* $\lambda_{t2} = \exp(\beta_2' x_t)$.

(i) State minimum conditions on the parameters, for the above model to be identifiable?

(ii) Carefully explain (giving details of $Q(\theta^*, \theta)$ and the EM stages) how the EM-algorithm can be used to obtain estimators of β_1, β_2 and p.

(iii) Derive the derivative of $Q(\theta^*, \theta)$, and explain how the derivative may be useful in the maximisation stage of the EM-algorithm.

(iv) Given an initial value, will the EM-algorithm always find the maximum of the likelihood?

Explain how one can check whether the parameter which maximises the EM-algorithm, maximises the likelihood.

Solution

(i) $0 < p < 1$ and $\beta_1 \neq \beta_2$ (these are minimum assumptions, there could be more which is hard to account for given the regressors x_t).

(ii) We first observe that $P(Y_t = y)$ is a mixture of two Poisson distributions where each has the canonical link function. Define the unobserved variables, $\{U_t\}$, which are iid and where $P(U_t = 1) = p$ and $P(U_t = 2) = (1-p)$ and $P(Y = y|U_i = 1) = \frac{\lambda_{t1}^y \exp(-\lambda_{t1} y)}{y!}$ and $P(Y = y)|U_i = 2) = \frac{\lambda_{t2}^y \exp(-\lambda_{t2} y)}{y!}$. Therefore, we have

$$\log f(Y_t, U_t, \theta) = \bigg(Y_t \beta'_{u_t} x_t - \exp(\beta'_{u_t} x_t) + \log Y_t! + \log p \bigg),$$

where $\theta = (\beta_1, \beta_2, p)$. Thus, $E(\log f(Y_t, U_t, \theta)|Y_t, \theta_*)$ is

$$\begin{aligned} E(\log f(Y_t, U_t, \theta)|Y_t, \theta_*) &= \bigg(Y_t \beta'_1 x_t - \exp(\beta'_1 x_t) + \log Y_t! + \log p \bigg) \pi(\theta_*, Y_t) \\ &\quad + \bigg(Y_t \beta'_2 x_t - \exp(\beta'_2 x_t) + \log Y_t! + \log p \bigg) (1 - \pi(\theta_*, Y_t)). \end{aligned}$$

where $P(U_i|Y_t, \theta^*)$ is evaluated as

$$P(U_i = 1|Y_t, \theta^*) = \pi(\theta_*, Y_t) = \frac{p f_1(Y_t, \theta_*)}{p f_1(Y_t, \theta_*) + (1-p) f_2(Y_t, \theta_*)},$$

with

$$f_1(Y_t, \theta_*) = \frac{\exp(\beta'_{*1} x_t Y_t) \exp(-Y_t \exp(\beta'_{*1} x_t))}{Y_t!} \quad f_1(Y_t, \theta_*) = \frac{\exp(\beta'_{*1} x_t Y_t) \exp(-Y_t \exp(\beta'_{*1} x_t)}{Y_t!}.$$

Thus $Q(\theta_*, \theta)$ is

$$\begin{aligned} Q(\theta_*, \theta) &= \sum_{t=1}^{T} \Big(Y_t \beta_1' x_t - \exp(\beta_1' x_t) + \log Y_t! + \log p \Big) \pi(\theta_*, Y_t) \\ &\quad + \Big(Y_t \beta_2' x_t - \exp(\beta_2' x_t) + \log Y_t! + \log(1-p) \Big) (1 - \pi(\theta_*, Y_t)). \end{aligned}$$

Using the above, the EM algorithm is the following:

(a) Start with an initial value which is an estimator of β_1, β_2 and p, denote this as θ_*.

(b) For every θ evaluate $Q(\theta_*, \theta)$.

(c) Evaluate $\arg\max_\theta Q(\theta_*, \theta)$. Denote the maximum as θ_* and return to step (b).

(d) Keep iterating until the maximums are sufficiently close.

(iii) The derivative of $Q(\theta_*, \theta)$ is

$$\begin{aligned} \frac{\partial Q(\theta_*, \theta)}{\partial \beta_1} &= \sum_{t=1}^{T} \Big(Y_t - \exp(\beta_1' x_t) \Big) x_t \pi(\theta_*, Y_t) \\ \frac{\partial Q(\theta_*, \theta)}{\partial \beta_2} &= \sum_{t=1}^{T} \Big(Y_t - \exp(\beta_2' x_t) \Big) x_t (1 - \pi(\theta_*, Y_t)) \\ \frac{\partial Q(\theta_*, \theta)}{\partial p} &= \sum_{t=1}^{T} \Big(\frac{1}{p} \pi(\theta_*, Y_t) - \frac{1}{1-p} (1 - \pi(\theta_*, Y_t) \Big). \end{aligned}$$

Thus maximisation of $Q(\theta_*, \theta)$ can be achieved by solving for the above equations using iterative weighted least squares.

(iv) Depending on the initial value, the EM-algorithm may only locate a local maximum.

To check whether we have found the global maximum, we can start the EM-algorithm with several different initial values and check where they converge.

Example 7.2.4 *Question*

(2) Let us suppose that $\mathcal{F}_1(t)$ and $\mathcal{F}_2(t)$ are two survival functions. Let x denote a univariate regressor.

(i) Show that $\mathcal{F}(t; x) = p\mathcal{F}_1(t)^{\exp(\beta_1 x)} + (1-p)\mathcal{F}_2(t)^{\exp(\beta_2 x)}$ is a valid survival function and obtain the corresponding density function.

(ii) Suppose that T_i are survival times and x_i is a univariate regressor which exerts an influence an T_i. Let $Y_i = \min(T_i, c)$, where c is a common censoring time. $\{T_i\}$ are independent random variables with survival function $\mathcal{F}(t; x_i) = p\mathcal{F}_1(t)^{\exp(\beta_1 x_i)} + (1-p)\mathcal{F}_2(t)^{\exp(\beta_2 x_i)}$, where both $\mathcal{F}_1$ and $\mathcal{F}_2$ are known, but p, β_1 and β_2 are unknown.

State the censored likelihood and show that the EM-algorithm together with iterative least squares in the maximisation step can be used to maximise this likelihood (sufficient details need to be given such that your algorithm can be easily coded).

Solution

i) Since $\mathcal{F}_1$ and $\mathcal{F}_2$ are monotonically decreasing positive functions where $\mathcal{F}_1(0) = \mathcal{F}_2(0) = 1$ and $\mathcal{F}_1(\infty) = \mathcal{F}_2(\infty) = 0$, then it immediately follows that

$$\mathcal{F}(t, x) = p\mathcal{F}_1(t)^{e^{\beta_1 x}} + (1-p)\mathcal{F}_2(t)^{e^{\beta_2 x}}$$

satisfies the same conditions. To obtain the density we differential wrt x

$$\begin{aligned} \frac{\partial \mathcal{F}(t,x)}{\partial t} &= -pe^{\beta_1 x} f_1(t)\mathcal{F}_1(t)^{e^{\beta_1 x}-1} - (1-p)e^{\beta_2 x} f_2(t)\mathcal{F}_2(t)^{e^{\beta_2 x}-1} \\ \Rightarrow f(t;x) &= pe^{\beta_1 x} f_1(t)\mathcal{F}_1(t)^{e^{\beta_1 x-1}} + (1-p)e^{\beta_2 x} f_2(t)\mathcal{F}_2(t)^{e^{\beta_2 x-1}}, \end{aligned}$$

where we use that $\frac{dF_(t)}{dt} = -f_(t)$.

ii) The censored log likelihood is

$$\mathcal{L}_n(\beta_1, \beta_2, p) = \sum_{i=1}^{n} [\delta_i \log f(Y_i; \beta_1, \beta_2, p) + (1-\delta_i) \log \mathcal{F}(Y_i; \beta_1, \beta_2, p)].$$

Clearly, directly maximizing the above is extremely difficult. Thus we look for an alternative method via the EM algorithm.

We first define the indicator variable (which corresponds to the missing variables) which denotes the state 1 or 2

$$I_i = \begin{cases} 1 & \text{with } P(I_i = 1) = p = p_1 \\ 2 & \text{with } P(I_i = 2) = (1-p) = p_2. \end{cases}$$

Then the joint density of (Y_i, δ_i, I_i) is

$$p_{I_i}\left(e^{\beta_{I_i}x_i}f_{I_i}(t)\mathcal{F}_{I_i}(t)^{e^{\beta_{I_i}x_i}-1}\right)\left(\mathcal{F}_{I_i}(t)^{e^{\beta_{I_i}x_i}}\right)^{1-\delta_i}$$

which gives the log-density

$$\delta_i\left\{\log p_{I_i}+\beta_{I_i}x_i+\log f_{I_i}(Y_i)+(e^{\beta_{I_i}x_i}-1)\log F_{I_i}(Y_i)\right\}+(1-\delta_i)\left\{\log p_{I_i}+(e^{\beta_{I_i}x_i})\log F_{I_i}(Y_i)\right\}.$$

Thus the complete log likelihood of (Y_i, δ_i, I_i) is

$$\begin{aligned}\mathcal{L}_n(\underline{Y},\underline{\delta},I_i;\beta_1,\beta_2,p)=\sum_{i=1}^{n}\{\delta_i[&\log p_{I_i}+\beta_{I_i}x_i+\log f_{I_i}(Y_i)+(e^{\beta_{I_i}x_i}-1)\log\mathcal{F}_{I_i}(Y_i)\\&+(1-\delta_i)[\log p_{I_i}+(e^{\beta_{I_i}x_i})\log\mathcal{F}_{I_i}(Y_i)]\}\end{aligned}$$

Next we need to calculate $P(I_i = 1|Y_i, \delta_i, \theta^*)$ and $P(I_i = 2|Y_i, \delta_i, \theta^*)$;

$$\begin{aligned}\omega_i^{\delta_i=1}(1) &= P(I_i=1|Y_i,\delta_i=1,p^*,\beta_1^*,\beta_2^*)\\ &= \frac{p^*e^{\beta_1^*x_i}f_1(Y_i)\mathcal{F}_1(Y_i)^{e^{\beta_1^*x_i}-1}}{p^*e^{\beta_1^*x_i}f_1(Y_i)\mathcal{F}_1(Y_i)^{e^{\beta_1^*x_i}-1}+(1-p^*)e^{\beta_2^*x_i}f_2(Y_i)\mathcal{F}_2(Y_i)^{e^{\beta_2^*x_i}-1}}\\ \omega_i^{\delta_i=0}(1) &= P(I_i=1|Y_i,\delta_i=0,p^*,\beta_1^*,\beta_2^*)\\ &= \frac{p^*\mathcal{F}_1(Y_i)^{e^{\beta_1^*x_i}}}{p^*\mathcal{F}_1(Y_i)^{e^{\beta_1^*x_i}}+(1-p^*)\mathcal{F}_2(Y_i)^{e^{\beta_2^*x_i}}}\end{aligned}$$

and $\omega_i^{\delta_i=1}(2) = 1-\omega_i^{\delta_i=1}(1)$ and $\omega_i^{\delta_i=0}(2) = 1-\omega_i^{\delta_i=0}(1)$. Let $p_1 = p$ and $p_2 = 1-p$. Therefore the complete likelihood conditioned on what we observe is

$$\begin{aligned}Q(\theta_*,\theta) &= \sum_{s=1}^{2}\sum_{i=1}^{n}\{\delta_i\omega_i^{\delta_i=1}(s)[\log p_s+\beta_1x_i+\log f_s(Y_i)+(e^{\beta_sx_i}-1)\log\mathcal{F}_s(Y_i)]\\ &\quad+(1-\delta_i)\omega_i^{\delta_i=0}(s)\left[\log p_s+e^{\beta_sx_i}\log\mathcal{F}_s(Y_i)\right]\}\\ &= \sum_{s=1}^{2}\sum_{i=1}^{n}\Big\{\{\delta_i\omega_i^{\delta_i=1}(s)[\beta_1x_i+\log f_s(Y_i)+(e^{\beta_sx_i}-1)\log\mathcal{F}_s(Y_i)]\\ &\quad+e^{\beta_sx_i}\log\mathcal{F}_s(Y_i)\Big\}\\ &\quad+\sum_{s=1}^{2}\sum_{i=1}^{n}\left\{\delta_i\omega_i^{\delta_i=1}(s)\log p_s+(1-\delta_i)\omega_i^{\delta_i=0}(s)\log p_s\right\}\\ &= Q(\theta_*,\beta_1)+Q(\theta_*,\beta_2)+Q(\theta_*,p_1,p_2)\end{aligned}$$

The conditional likelihood, above, looks unwieldy. However, the parameter estimators can to be separated. First, differentiating with respect to p gives

$$\begin{aligned}\frac{\partial Q}{\partial p} &= \frac{\partial Q(\theta_*, p, 1-p)}{\partial p} \\ &= \sum_{i=1}^{n} \delta_i \omega_i^{\delta_i=1}(1)\frac{1}{p} + \sum_{i=1}^{n} \omega_i^{\delta_i=0}(1)(1-\delta_i)\frac{1}{p} - \\ &\quad \sum_{i=1}^{n} \delta_i \omega_i^{\delta_i=1}(2)\frac{1}{1-p} - \sum_{i=1}^{n} \omega_i^{\delta_i=0}(2)(1-\delta_i)\frac{1}{1-p}.\end{aligned}$$

Equating the above to zero we have the estimator $\hat{p} = \frac{a}{a+b}$, where

$$\begin{aligned}a &= \sum_{i=1}^{n} \delta_i \omega_i^{\delta_i=1}(1) + \sum_{i=1}^{n} \omega_i^{\delta_i=0}(1)(1-\delta_i) \\ b &= \sum_{i=1}^{n} \delta_i \omega_i^{\delta_i=1}(2) + \sum_{i=1}^{n} \omega_i^{\delta_i=0}(2)(1-\delta_i).\end{aligned}$$

Next we consider the estimates of β_1 and β_2 at the ith iteration step. Differentiating Q wrt to β_1 and β_2 gives for $s = 1, 2$

$$\begin{aligned}\frac{\partial Q}{\partial \beta_s} &= \frac{\partial Q_s(\theta_*, \beta_s)}{\partial \beta_s} \\ &= \sum_{i=1}^{n} \left\{ \delta_i \omega_i^{\delta_i=1}(s) \left[1 + e^{\beta_s x_i} \log F_s(Y_i)\right] + (1-\delta_i)\omega_i^{\delta_i=0}(s) e^{\beta_s x_i} \log F_s(Y_i) \right\} x_i \\ \frac{\partial^2 Q(\theta_*, \theta)}{\partial \beta_s^2} &= \frac{\partial^2 Q_s(\theta^*, \beta_s)}{\partial \beta_s} \\ &= \sum_{i=1}^{n} \left\{ \delta_i \omega_i^{\delta_i=1}(s) e^{\beta_s x_i} \log F_s(Y_i) + (1-\delta_i)\omega_i^{\delta_i=0}(s) e^{\beta_s x_i} \log F_s(Y_i) \right\} x_i^2 \\ \frac{\partial^2 Q(\theta_*; \theta)}{\partial \beta_1 \partial \beta_2} &= 0.\end{aligned}$$

Observe that setting the first derivative to zero, we cannot obtain an explicit expression for the estimators at each iteration. Thus we need to use the Newton-Rapshon scheme but in a very simply set-up. To estimate (β_1, β_2) at the jth iteration we use

$$\begin{bmatrix} \beta_1^{(j)} \\ \beta_2^{(j)} \end{bmatrix} = \begin{bmatrix} \beta_1^{(j-1)} \\ \beta_2^{(j-1)} \end{bmatrix} + \begin{bmatrix} \frac{\partial^2 Q}{\partial \beta_1^2} & 0 \\ 0 & \frac{\partial^2 Q}{\partial \beta_1^2} \end{bmatrix}^{-1}_{\underline{\beta}^{(j-1)}} \begin{bmatrix} \frac{\partial Q}{\partial \beta_1} \\ \frac{\partial Q}{\partial \beta_2} \end{bmatrix}_{\underline{\beta}^{(j-1)}}$$

Thus for $s = 1, 2$ we have $\beta_s^{(j)} = \beta_s^{(j-1)} + \left(\frac{\partial^2 Q}{\partial \beta_s^2}\right)^{-1} \frac{\partial Q}{\partial \beta_s} \rfloor_{\underline{\beta}^{(j-1)}}$.

We can rewrite the above Newton Raphson scheme as something that resembles weighted least squares. We recall the weighted least squares estimator are the parameters α which minimise the weighted least squares criterion

$$\sum_{i=1}^{n} W_{ii} \left(Y_i - \underline{x}_i' \underline{\alpha}\right)^2.$$

The α which minimises the above is

$$\widehat{\underline{\alpha}} = (\underline{X}' W \underline{X})^{-1} \underline{X}' W \underline{Y}.$$

The Newtons-Raphson scheme can be written as

$$\begin{aligned} \beta_s^{(j)} &= \beta_s^{(j-1)} - \left(\frac{\partial Q^2}{\partial \beta_s^2}\right)^{-1} \frac{\partial Q}{\partial \beta_s} \rfloor_{\underline{\beta}^{(j-1)}} \\ &= \beta_s^{(j-1)} - (\underline{X}' W_s^{(j-1)} \underline{X})^{-1} \underline{X} \underline{S}_s^{(j-1)} \end{aligned}$$

where

$$\begin{aligned} \underline{X}' &= (x_1, x_2, \ldots, x_n), \\ W_s^{(j-1)} &= \mathrm{diag}[\omega_1^{(j-1)}(s), \ldots, \omega_n^{(j-1)}(s)], \\ \underline{S}_s^{(j-1)} &= \begin{bmatrix} S_{s1}^{(j-1)} \\ \vdots \\ S_{sn}^{(j-1)} \end{bmatrix}, \end{aligned}$$

where the elements of the above are

$$\begin{aligned} \omega_{si}^{(j-1)} &= \delta_i \omega_i^{\delta_i = 1} e^{\beta_s^{(j-1)}} \log \mathcal{F}_s(Y_i) + (1 - \delta_i) \omega_i^{\delta_i = 0} e^{\beta_s^{(j-1)} x_i} \log \mathcal{F}_s(Y_i) \\ S_{si}^{(j-1)} &= \delta_i \omega_i^{\delta_i = 1} [1 + e^{\beta_s^{(j-1)} x_i} \log \mathcal{F}_s(Y_i)] + (1 - \delta_i) \omega_i^{\delta_i = 0} e^{\beta_s^{(j-1)} x_i} \log \mathcal{F}_s(Y_i)]. \end{aligned}$$

By using algebraic manipulations we can rewrite the iteration as an iterated weighted least squared algorithm

$$\begin{aligned} \beta_s^{(j)} &= \beta_s^{(j-1)} - \left(\frac{\partial Q^2}{\partial \beta_s^2}\right)^{-1} \frac{\partial Q}{\partial \beta_s} \rfloor_{\underline{\beta}^{(j-1)}} \\ &= \beta_s^{(j-1)} - (\underline{X}' \omega_s^{(j-1)} \underline{X})^{-1} \underline{X}' \underline{S}_s^{(j-1)} \\ &= (\underline{X}' W_s^{(j-1)} \underline{X})^{-1} (\underline{X}' W_s^{(j-1)} \underline{X}) \beta_s^{(j-1)} - (\underline{X}' W_s^{(j-1)} \underline{X})^{-1} \underline{X} \underline{S}_s^{(j-1)} \\ &= (\underline{X}' W_s^{(j-1)} \underline{X})^{-1} \underline{X}' W_s^{(j-1)} \underline{X} \beta_s^{(j-1)} - (\underline{X}' W_s^{(j-1)} \underline{X})^{-1} \underline{X} W_s^{(j-1)} [W_s^{(j-1)}]^{-1} \underline{S}_s^{(j-1)} \end{aligned}$$

Now we rewrite the above in weighted least squares form. Define

$$\underline{Z}_s^{(j-1)} = \underline{X}\beta_s^{(j-1)} - [W_s^{(j-1)}]^{-1}\underline{S}_s^{(j-1)}$$

this "acts" as our pseudo y-variable. Using this notation we have

$$\beta_s^{(j)} \;=\; (\underline{X}'W_s^{(j-1)}\underline{X})^{-1}\underline{X}'W_s^{(j-1)}\underline{Z}_s^{(j-1)}.$$

Thus at each step of the Newton-Raphson iteration we minimise the weighted least equation

$$\sum_{i=1}^{n}\omega_{si}^{(j-1)}\left(Z_s^{(j-1)} - \beta x_i\right)^2 \text{ for } s = 1, 2.$$

Thus altogether in the EM-algorithm we have:
Start with initial value $\beta_1^0, \beta_2^0, p^0$

Step 1 Set $(\beta_{1,r-1}, \beta_{2,r-1}, p_{r-1}) = (\beta_1^*, \beta_2^*, p^*)$. Evaluate $\omega_i^{\delta_i}$ and $\omega_i^{1-\delta_i}$ (these probabilies/weights stay the same throughout the iterative least squares).

Step 2 Maximize $Q(\theta_*, \theta)$ by using the algorithm $p_r = \frac{a_r}{a_r+b_r}$ where a_r, b_r are defined previously. Now evaluate for $s = 1, 2$

$$\beta_s^{(j)} = (\underline{X}'W_s^{(j-1)}\underline{X})^{-1}\underline{X}'W_s^{(j-1)}Z_s^{(j-1)}.$$

Iterate until convergence of the parameters.

Step 3 Go back to step 1 until convergence of the EM algorithm.

7.2.4 Exercises

Exercise 7.1 *Consider the linear regression model*

$$Y_i = \underline{\alpha}'\underline{x}_i + \sigma_i\varepsilon_i$$

where ε_i follows a standard normal distribution (mean zero and variance 1) and σ_i^2 follows a Gamma distribution

$$f(\sigma^2; \lambda) = \frac{\sigma^{2(\kappa-1)}\lambda^\kappa \exp(-\lambda\sigma^2)}{\Gamma(\kappa)}, \quad \sigma^2 \geq 0,$$

with $\kappa > 0$.

Let us suppose that $\underline{\alpha}$ and λ are unknown parameters but κ is a known parameter. We showed in Exercise 1.1 that directly maximising the log-likelihood was extremely difficult.

Derive the EM-algorithm for estimating the parameters in this model. In your derivation explain what quantities will have to be evaluated numerically.

Exercise 7.2 *Consider the following shifted exponential mixture distribution*

$$f(x;\lambda_1,\lambda_2,p,a) = p\frac{1}{\lambda_1}\exp(-x/\lambda_1)I(x \geq 0) + (1-p)\frac{1}{\lambda_2}\exp(-(x-a)/\lambda_2)I(x \geq a),$$

where p, λ_1, λ_2 and a are unknown.

(i) Make a plot of the above mixture density.

Considering the cases $x \geq a$ and $x < a$ separately, calculate the probability of belonging to each of the mixtures, given the observation X_i (i.e. Define the variable δ_i, where $P(\delta_i = 1) = p$ and $f(x|\delta_i = 1) = \frac{1}{\lambda_1}\exp(-x/\lambda_1)$ and calculate $P(\delta_i|X_i)$).

(ii) Show how the EM-algorithm can be used to estimate $a, p, \lambda_1, \lambda_2$. At each iteration you should be able to obtain explicit solutions for most of the parameters, give as many details as you can.

Hint: It may be beneficial for you to use profiling too.

(iii) From your knowledge of estimation of these parameters, what do you conjecture the rates of convergence to be? Will they all be the same, or possibly different?

Exercise 7.3 *Suppose $\{Z_i\}_{i=1}^n$ are independent random variables, where Z_i has the density*

$$f_Z(z;\beta_0,\beta_1,\mu,\alpha,u_i) = ph(z;\beta_0,\beta_1,u_i) + (1-p)g(z;\alpha,\mu),$$

$g(x;\alpha,\mu) = (\frac{\alpha}{\mu})(\frac{x}{\mu})^{\alpha-1}\exp(-(x/\mu)^\alpha)I_{(0,\infty)}(x)$ (the Weibull distribution) and $h(x;\beta_0,\beta_1,u_i) = \frac{1}{\lambda_i}\exp(-x/\lambda_i)I_{(0,\infty)}(x)$ (the exponential distribution), with $\lambda_i = \beta_0\exp(\beta_1 u_i)$ and $\{u_i\}_{i=1}^n$ are observed regressors.

The parameters p, β_0, β_1, μ and α are unknown and our objective in this question is to estimate them.

(a) What is the log-likelihood of $\{Z_i\}$? (Assume we also observe the deterministic regressors $\{u_i\}$.)

(b) By defining the correct dummy variable δ_i derive the steps of the EM-algorithm to estimate the parameters $p, \beta_0, \beta_1, \mu, \alpha$ (using the method of profiling if necessary).

7.3 Hidden Markov Models

Finally, we consider applications of the EM-algorithm to parameter estimation in Hidden Markov Models (HMM). This is a model where the EM-algorithm pretty much surpasses any other likelihood maximisation methodology. It is worth mentioning that the EM-algorithm in this setting is often called the *Baum-Welch algorithm*.

Hidden Markov models are a generalisation of mixture distributions, however unlike mixture distibutions it is difficult to derive an explicit expression for the likelihood of a Hidden Markov Models. HMM are a general class of models which are widely used in several applications (including speech recongition), and can easily be generalised to the Bayesian set-up. A nice description of them can be found on Wikipedia.

In this section we will only briefly cover how the EM-algorithm can be used for HMM. We do not attempt to address any of the issues surrounding how the maximisation is done, interested readers should refer to the extensive literature on the subject.

The general HMM is described as follows. Let us suppose that we observe $\{Y_t\}$, where the rvs Y_t satisfy the Markov property $P(Y_t|Y_{t-1}, Y_{t-1}, \ldots) = P(Y_t|Y_{t-1})$. In addition to $\{Y_t\}$ there exists a 'hidden' unobserved discrete random variables $\{U_t\}$, where $\{U_t\}$ satisfies the Markov property $P(U_t|U_{t-1}, U_{t-2}, \ldots) = P(U_t|U_{t-1})$ and 'drives' the dependence in $\{Y_t\}$. In other words $P(Y_t|U_t, Y_{t-1}, U_{t-1}, \ldots) = P(Y_t|U_t)$. To summarise, the HMM is described by the following properties:

(i) We observe $\{Y_t\}$ (which can be either continuous or discrete random variables) but do not observe the hidden discrete random variables $\{U_t\}$.

(ii) Both $\{Y_t\}$ and $\{U_t\}$ are time-homogenuous Markov random variables that is $P(Y_t|Y_{t-1}, Y_{t-1}, \ldots) = P(Y_t|Y_{t-1})$ and $P(U_t|U_{t-1}, U_{t-1}, \ldots) = P(U_t|U_{t-1})$. The distributions of $P(Y_t)$, $P(Y_t|Y_{t-1})$, $P(U_t)$ and $P(U_t|U_{t-1})$ do not depend on t.

(iii) The dependence between $\{Y_t\}$ is driven by $\{U_t\}$, that is $P(Y_t|U_t, Y_{t-1}, U_{t-1}, \ldots) = P(Y_t|U_t)$.

There are several examples of HMM, but to have a clear intepretation of them, in this section we shall only consider one classical example of a HMM. Let us suppose that the hidden random variable U_t can take N possible values $\{1, \ldots, N\}$ and let $p_i = P(U_t = i)$ and $p_{ij} = P(U_t = i|U_{t-1} = j)$. Moreover, let us suppose that Y_t are continuous random variables where $(Y_t|U_t = i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and the conditional random variables $Y_t|U_t$ and $Y_\tau|U_\tau$ are independent of each other. Our objective is to estimate the parameters $\theta = \{p_i, p_{ij}, \mu_i, \sigma_i^2\}$ given $\{Y_i\}$. Let $f_i(\cdot;\theta)$ denote the normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$.

Remark 7.3.1 (HMM and mixture models) *Mixture models (described in the above section) are a particular example of HMM. In this case the unobserved variables $\{U_t\}$ are iid, where $p_i = P(U_t = i|U_{t-1} = j) = P(U_t = i)$ for all i and j.*

Let us denote the log-likelihood of $\{Y_t\}$ as $\mathcal{L}_T(\underline{Y};\theta)$ (this is the observed likelihood). It is clear that constructing an explicit expression for $\mathcal{L}_T$ is difficult, thus maximising the likelihood is near impossible. In the remark below we derive the observed likelihood.

Remark 7.3.2 *The likelihood of $\underline{Y} = (Y_1, \ldots, Y_T)$ is*

$$\begin{aligned} L_T(\underline{Y};\theta) &= f(Y_T|Y_{T-1}, Y_{T-2}, \ldots;\theta) \ldots f(Y_2|Y_1;\theta)P(Y_1;\theta) \\ &= f(Y_T|Y_{T-1};\theta) \ldots f(Y_2|Y_1;\theta)f(Y_1;\theta). \end{aligned}$$

Thus the log-likelihood is

$$\mathcal{L}_T(\underline{Y};\theta) = \sum_{t=2}^{T} \log f(Y_t|Y_{t-1};\theta) + f(Y_1;\theta).$$

The distribution of $f(Y_1;\theta)$ is simply the mixture distribution

$$f(Y_1;\theta) = p_1 f(Y_1;\theta_1) + \ldots + p_N f(Y_1;\theta_N),$$

where $p_i = P(U_t = i)$. The conditional $f(Y_t|Y_{t-1})$ is more tricky. We start with

$$f(Y_t|Y_{t-1};\theta) = \frac{f(Y_t, Y_{t-1};\theta)}{f(Y_{t-1};\theta)}.$$

An expression for $f(Y_t;\theta)$ is given above. To evaluate $f(Y_t, Y_{t-1};\theta)$ we condition on

U_t, U_{t-1} to give (using the Markov and conditional independent propery)

$$\begin{aligned} f(Y_t, Y_{t-1}; \theta) &= \sum_{i,j} f(Y_t, Y_{t-1}|U_t = i, U_{t-1} = j) P(U_t = i, U_{t-1} = j) \\ &= \sum_{i,j} f(Y_t|U_t = i) P(Y_{t-1}|U_{t-1} = j) P(U_t = i|U_{t-1} = j) P(U_{t-1} = i) \\ &= \sum_{i,j} f_i(Y_t; \theta_i) f_j(Y_{t-1}; \theta_j) p_{ij} p_i. \end{aligned}$$

Thus we have

$$f(Y_t|Y_{t-1}; \theta) = \frac{\sum_{i,j} f_i(Y_t; \theta_i) f_j(Y_{t-1}; \theta_j) p_{ij} p_i}{\sum_i p_i f(Y_{t-1}; \theta_i)}.$$

We substitute the above into $\mathcal{L}_T(\underline{Y}; \theta)$ to give the expression

$$\mathcal{L}_T(\underline{Y}; \theta) = \sum_{t=2}^{T} \log \left(\frac{\sum_{i,j} f_i(Y_t; \theta_i) f_j(Y_{t-1}; \theta_j) p_{ij} p_i}{\sum_i p_i f(Y_{t-1}; \theta_i)} \right) + \log \left(\sum_{i=1}^{N} p_i f(Y_1; \theta_i) \right).$$

Clearly, this is extremely difficult to maximise.

Instead we seek an indirect method for maximising the likelihood. By using the EM algorithm we can maximise a likelihood which is a lot easier to evaluate. Let us suppose that we observe $\{Y_t, U_t\}$. Since $P(\underline{Y}|\underline{U}) = P(Y_T|Y_{T-1}, \ldots, Y_1, \underline{U}) P(Y_{T-1}|Y_{T-2}, \ldots, Y_1, \underline{U}) \ldots P(Y_1|\underline{U}) = \prod_{t=1}^{T} P(Y_t|U_t)$, and the distribution of $Y_t|U_t$ is $\mathcal{N}(\mu_{U_t}, \sigma^2_{U_t})$, then the complete likelihood of $\{Y_t, U_t\}$ is

$$\left(\prod_{t=1}^{T} f(Y_t|U_t; \theta) \right) p_{U_1} \prod_{t=2}^{T} p_{U_t|U_{t-1}}.$$

Thus the log-likelihood of the complete observations $\{Y_t, U_t\}$ is

$$\mathcal{L}_T(\underline{Y}, \underline{U}; \theta) = \sum_{t=1}^{T} \log f(Y_t|U_t; \theta) + \sum_{t=2}^{T} \log p_{U_t|U_{t-1}} + \log p_{U_1}.$$

Of course, we do not observe the complete likelihood, but the above can be used in order to define the function $Q(\theta_*, \theta)$ which is maximised in the EM-algorithm. It is worth mentioning that given the transition probabilities of a discrete Markov chain (that is $\{p_{i,j}\}_{ij}$) the marginal/stationary probabilities $\{p_i\}$ can be obtained by solving $\pi = \pi P$, where P is the transition matrix. Thus it is not necessary to estimate the marginal

probabilities $\{p_i\}$ (note that the exclusion of $\{p_i\}$ in the log-likelihood, above, gives the conditional complete log-likelihood).

We recall that to maximise the observed likelihood $\mathcal{L}_T(\underline{Y};\theta)$ using the EM algorithm involves evaluating $Q(\theta_*,\theta)$, where

$$\begin{aligned} Q(\theta_*,\theta) &= \mathrm{E}\Big(\sum_{t=1}^{T}\log f(Y_t|U_t;\theta)+\sum_{t=2}^{T}\log p_{U_t|U_{t-1}}+\log p_{U_1}\Big|\underline{Y},\theta_*\Big) \\ &= \sum_{\underline{U}\in\{1,\ldots,N\}^T}\Big(\sum_{t=1}^{T}\log f(Y_t|U_t;\theta)+\sum_{t=2}^{T}\log p_{U_t|U_{t-1}}+\log p_{U_1}\Big)p(\underline{U}|\underline{Y},\theta_*). \end{aligned}$$

Note that each step in the algorithm the probability $p(\underline{U}|\underline{Y},\theta_*)$ needs to be evaluated. This is done by using conditioning

$$\begin{aligned} p(\underline{U}|\underline{Y},\theta_*) &= p(U_1|\underline{Y},\theta_*)\prod_{t=2}^{T}P(U_t|U_{t-1},\ldots,U_1\underline{Y};\theta_*) \\ &= p(U_1|\underline{Y},\theta_*)\prod_{t=2}^{T}P(U_t|U_{t-1},\underline{Y};\theta_*) \text{ (using the Markov property).} \end{aligned}$$

Evaluation of the above is not simple (mainly because one is estimating the probability of being in state U_t based on U_{t-1} *and* the observation information Y_t in the past, present and future). This is usually done using the so called forward backward algorithm (and is related to the idea of Kalman filtering), see `https://en.wikipedia.org/wiki/Forward/backward_algorithm`.

For this example the EM algorithm is

(i) Define an initial value $\theta_1\in\Theta$. Let $\theta_*=\theta_1$.

(ii) **The expectation step**,

For a fixed θ_* evaluate $P(U_t,\underline{Y},\theta_*)$, $P(U_t|U_{t-1},\underline{Y},\theta_*)$ and $Q(\theta_*,\theta)$.

(iii) **The maximisation step**

Evaluate $\theta_{k+1}=\arg\max_{\theta\in\Theta}Q(\theta_*,\theta)$ by differentiating $Q(\theta_*,\theta)$ wrt to θ and equating to zero.

(iv) If θ_k and θ_{k+1} are sufficiently close to each other stop the algorithm and set $\hat{\theta}_n=\theta_{k+1}$. Else set $\theta_*=\theta_{k+1}$, go back and repeat steps (ii) and (iii) again.

Since $P(U_1|\underline{Y},\theta_*) = P(U_1,\underline{Y},\theta_*)/P(\underline{Y},\theta_*)$ and $P(U_t,U_{t-1}|\underline{Y},\theta_*) = P(U_t,U_{t-1},\underline{Y},\theta_*)/P(\underline{Y},\theta_*)$; $P(\underline{Y},\theta_*)$ is common to all $\underline{U}$ in $\{1,\ldots,N\}^T$ and is independent of θ_*, Thus rather than maximising $Q(\theta_*,\theta)$ one can equvalently maximise

$$\widetilde{Q}(\theta_*,\theta) = \sum_{\underline{U}\in\{1,\ldots,N\}^T} \left(\sum_{t=1}^{T} \log f(Y_t|U_t;\theta) + \sum_{t=2}^{T} \log p_{U_t|U_{t-1}} + \log p_{U_1} \right) p(\underline{U},\underline{Y},\theta_*),$$

noting that $\widetilde{Q}(\theta_*,\theta) \propto Q(\theta_*,\theta)$ and the maximum of $\tilde{Q}(\theta_*,\theta)$ with respect to θ is the same as the maximum of $Q(\theta_*,\theta)$.

Chapter 8

Non-likelihood methods

8.1 Loss functions

Up until now our main focus has been on parameter estimating via the maximum likelihood. However, the negative maximum likelihood is simply one member of loss criterions. Loss functions are usually distances, such as the ℓ_1 and ℓ_2 distance. Typically we estimate a parameter by minimising the loss function, and using as the estimator the parameter which minimises the loss. Usually (but not always) the way to solve the loss function is to differentiate it and equate it to zero. Below we give examples of loss functions whose formal derivative does not exist.

8.1.1 L_1-loss functions

The Laplacian

Consider the Laplacian (also known as the double exponential), which is defined as

$$f(y;\theta,\rho) = \frac{1}{2\rho}\exp\left(-\frac{|y-\theta|}{\rho}\right) = \begin{cases} \frac{1}{2\rho}\exp\left(\frac{y-\theta}{\rho}\right) & y<\theta \\ \frac{1}{2\rho}\exp\left(\frac{\theta-y}{\rho}\right) & y\geq\theta. \end{cases}$$

We observe $\{Y_i\}$ and our objective is to estimate the location parameter θ, for now the scale parameter ρ is not of interest. The log-likelihood is

$$\mathcal{L}_n(\theta,\rho) = -n\log 2\rho - \rho^{-1}\underbrace{\frac{1}{2}\sum_{i=1}^{n}|Y_i-\theta|}_{=L_n(\theta)}.$$

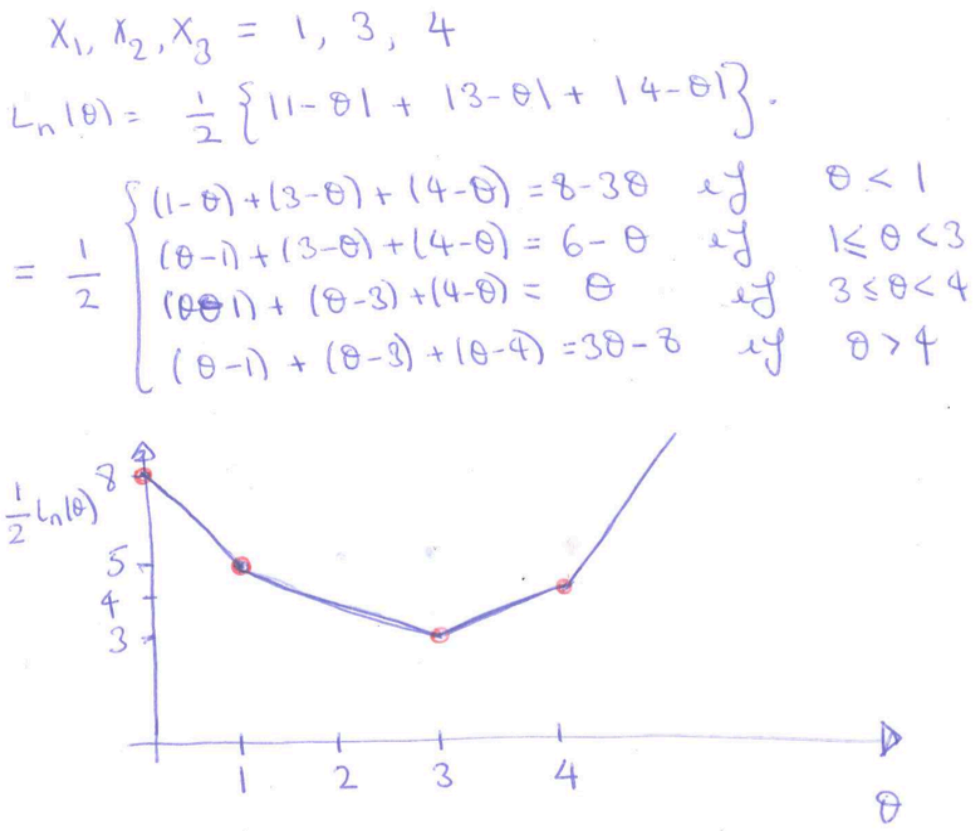

Figure 8.1: Plot of $L1$-norm

Since the θ which maximises the above does not depend on ρ we can simply focus on the component which maximises θ. We see that this is equivalent to minimising the loss function

$$L_n(\theta) = \frac{1}{2}\sum_{i=1}^{n}|Y_i - \theta| = \sum_{Y_{(i)}>\theta}\frac{1}{2}(Y_{(i)} - \theta) + \sum_{Y_{(i)}\le\theta}\frac{1}{2}(\theta - Y_{(i)}).$$

If we make a plot of L_n over θ, and consider how L_n, behaves at the ordered observations $\{Y_{(i)}\}$, we see that it is piecewise continuous (that is it is a piecewise continuous function, with joints at $Y_{(i)}$). On closer inspection (if n is odd) we see that L_n has its minimum at $\theta = Y_{(n/2)}$, which is the sample median (see Figure 8.1 for an illustration).

In summary, the normal distribution gives rise to the ℓ_2-loss function and the sample mean. In contrast the Laplacian gives rise to the ℓ_1-loss function and the sample median.

The asymmetric Laplacian

Consider the generalisation of the Laplacian, usually called the assymmetric Laplacian, which is defined as

$$f(y;\theta,\rho) = \begin{cases} \frac{p}{\rho}\exp\left(p\frac{y-\theta}{\rho}\right) & y < \theta \\ \frac{(1-p)}{\rho}\exp\left(-(1-p)\frac{y-\theta}{\rho}\right) & y \geq \theta. \end{cases}$$

where $0 < p < 1$. The corresponding negative likelihood to estimate θ is

$$L_n(\theta) = \sum_{Y_{(i)}>\theta} (1-p)(Y_i - \theta) + \sum_{Y_{(i)}\leq\theta} p(\theta - Y_i).$$

Using similar arguments to those in part (i), it can be shown that the minimum of L_n is approximately the pth quantile.

8.2 Estimating Functions

8.2.1 Motivation

Estimating functions are a unification and generalisation of the maximum likelihood methods and the method of moments. It should be noted that it is a close cousin of the *generalised method of moments* and *generalised estimating equation*. We first consider a few examples and will later describe a feature common to all these examples.

Example 8.2.1 *(i) Let us suppose that $\{Y_i\}$ are iid random variables with $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. The log-likelihood in proportional to*

$$\mathcal{L}_n(\mu,\sigma^2) = -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2.$$

We know that to estimate μ and σ^2 we use the μ and σ^2 which are the solution of

$$\frac{-1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2 = 0 \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0. \tag{8.1}$$

(ii) In general suppose $\{Y_i\}$ are iid random variables with $Y_i \sim f(\cdot;\theta)$. The log-likelihood is $\mathcal{L}_n(\theta) = \sum_{i=1}^{n}\log f(\theta;Y_i)$. If the regularity conditions are satisfied then to estimate θ we use the solution of

$$\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta} = 0. \tag{8.2}$$

(iii) Let us suppose that $\{X_i\}$ *are iid random variables with a Weibull distribution* $f(x;\theta) = (\frac{\alpha}{\phi})(\frac{x}{\phi})^{\alpha}\exp(-(x/\phi)^{\alpha})$*, where* $\alpha, \phi > 0$.

We know that $\mathrm{E}(X) = \phi\Gamma(1+\alpha^{-1})$ *and* $\mathrm{E}(X^2) = \phi^2\Gamma(1+2\alpha^{-1})$*. Therefore* $\mathrm{E}(X) - \phi\Gamma(1+\alpha^{-1}) = 0$ *and* $\mathrm{E}(X^2) - \phi^2\Gamma(1+2\alpha^{-1}) = 0$*. Hence by solving*

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \phi\Gamma(1+\alpha^{-1}) = 0 \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \phi^2\Gamma(1+2\alpha^{-1}) = 0, \tag{8.3}$$

we obtain estimators of α *and* Γ*. This is essentially a method of moments estimator of the parameters in a Weibull distribution.*

(iv) We can generalise the above. It can be shown that $\mathrm{E}(X^r) = \phi^r\Gamma(1+r\alpha^{-1})$*. Therefore, for any distinct* s *and* r *we can estimate* α *and* Γ *using the solution of*

$$\frac{1}{n}\sum_{i=1}^{n} X_i^r - \phi^r\Gamma(1+r\alpha^{-1}) = 0 \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^s - \phi^s\Gamma(1+s\alpha^{-1}) = 0. \tag{8.4}$$

(v) Consider the simple linear regression $Y_i = \alpha x_i + \varepsilon_i$*, with* $\mathrm{E}(\varepsilon_i) = 0$ *and* $\mathrm{var}(\varepsilon_i) = 1$*, the least squares estimator of* α *is the solution of*

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - ax_i)x_i = 0. \tag{8.5}$$

We observe that all the above estimators can be written as the solution of a homogenous equations - see equations (8.1), (8.2), (8.3), (8.4) and (8.5). In other words, for each case we can define a random function $G_n(\theta)$, such that the above estimators are the solutions of $G_n(\tilde{\theta}_n) = 0$. In the case that $\{Y_i\}$ are iid then $G_n(\theta) = \sum_{i=1}^{n} g(Y_i;\theta)$, for some function $g(Y_i;\theta)$. The function $G_n(\tilde{\theta})$ is called an estimating function. All the function G_n, defined above, satisfy the unbiased property which we define below.

Definition 8.2.1 (Estimating function) *An estimating function* G_n *is called unbiased if at the true parameter* θ_0 $G_n(\cdot)$ *satisfies*

$$\mathrm{E}\left[G_n(\theta_0)\right] = 0.$$

If there are p *unknown parameters and* p *estimating equations, the estimation equation estimator is the* θ *which solves* $G_n(\theta) = 0$.

Hence the estimating function is an alternative way of viewing parameter estimating. Until now, parameter estimators have been defined in terms of the maximum of the likelihood. However, an alternative method for defining an estimator is as the solution of a function. For example, suppose that $\{Y_i\}$ are random variables, whose distribution depends in some way on the parameter θ_0. We want to estimate θ_0, and we know that there exists a function such that $G(\theta_0) = 0$. Therefore using the data $\{Y_i\}$ we can define a random function, G_n where $\mathrm{E}(G_n(\theta)) = G(\theta)$ and use the parameter $\tilde{\theta}_n$, which satisfies $G_n(\tilde{\theta}) = 0$, as an estimator of θ. We observe that such estimators include most maximum likelihood estimators and method of moment estimators.

Example 8.2.2 *Based on the examples above we see that*

(i) The estimating function is

$$G_n(\mu, \sigma) = \begin{pmatrix} \frac{-1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (X_i - \mu)^2 \\ \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu) \end{pmatrix}.$$

(ii) The estimating function is $G_n(\theta) = \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}$.

(iii) The estimating function is

$$G_n(\alpha, \phi) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i - \phi\Gamma(1 + \alpha^{-1}) \\ \frac{1}{n}\sum_{i=1}^n X_i^2 - \phi^2\Gamma(1 + 2\alpha^{-1}) \end{pmatrix}.$$

(iv) The estimating function is

$$G_n(\alpha, \phi) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^s - \phi^s\Gamma(1 + s\alpha^{-1}) \\ \frac{1}{n}\sum_{i=1}^n X_i^r - \phi^r\Gamma(1 + r\alpha^{-1}) \end{pmatrix}.$$

(v) The estimating function is

$$G_n(a) = \frac{1}{n}\sum_{i=1}^n (Y_i - a x_i) x_i.$$

Observe that regardless of the distribution of the errors (or dependency between $\{Y_i\}$)

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n (Y_i - \alpha x_i) x_i\right) = 0, \tag{8.6}$$

is true regardless of the distribution of Y_i ($\{\varepsilon_i\}$) and is also true if there $\{Y_i\}$ are dependent random variables (see Rao (1973), Linear Statistical Inference and its applications).

The advantage of this approach is that sometimes the solution of an estimating equation will have a smaller finite sample variance than the MLE. Even though asymptotically (under certain conditions) the MLE will asymptotically attain the Cramer-Rao bound (which is the smallest variance). Moreover, MLE estimators are based on the assumption that the distribution is known (else the estimator is misspecified - see Section 5.1.1), however sometimes an estimating equation can be free of such assumptions.

Example 8.2.3 *In many statistical situations it is relatively straightforward to find a suitable estimating function rather than find the likelihood. Consider the time series* $\{X_t\}$ *which is "stationary" (moments are invariant to shift i.e* $\mathrm{E}[X_tX_{t+r}] = \mathrm{E}[X_0X_r]$*) which satisfies*

$$X_t = a_1X_{t-1} + a_2X_{t-2} + \sigma\varepsilon_t,$$

where $\{\varepsilon_t\}$ *are iid zero mean random variables (zero mean ensures that* $\mathrm{E}[X_t] = 0$*). We do not know the distribution of* ε_t*, but under certain conditions on* a_1 *and* a_2 *(causality conditions)* ε_t *is independent of* X_{t-1} *and* X_{t-2}*. Thus by multiplying the above equation by* X_{t-1} *or* X_{t-2} *and taking expections we have*

$$\begin{aligned}\mathrm{E}(X_tX_{t-1}) &= a_1\mathrm{E}(X_{t-1}^2) + a_2\mathrm{E}(X_{t-1}X_{t-2})\\ \mathrm{E}(X_tX_{t-2}) &= a_1\mathrm{E}(X_{t-1}X_{t-2}) + a_2\mathrm{E}(X_{t-2}^2).\end{aligned}$$

Since the above time series is 'stationary' (we have not formally defined this - but basically it means the properties of $\{X_t\}$ *do not "evolve" over time), the above reduces to*

$$\begin{aligned}c(1) &= a_1c(0) + a_2c(1)\\ c(2) &= a_1c(1) + a_2c(0),\end{aligned}$$

where $\mathrm{E}[X_tX_{t+r}] = c(r)$*. Given* $\{X_t\}_{t=1}^n$*, it can be shown that* $\widehat{c}_n(r) = n^{-1}\sum_{t=|r|+1}^n X_tX_{t-|r|}$ *is an estimator of* $c(r)$ *and that for small* r $\mathrm{E}[\widehat{c}_n(r)] \approx c(r)$ *(and is consistent). Hence replacing the above with its estimators we obtain the estimating equations*

$$G_1(a_1, a_2) = \begin{pmatrix} \widehat{c}_n(1) - a_1\widehat{c}_n(0) - a_2\widehat{c}_n(1) \\ \widehat{c}_n(2) - a_1\widehat{c}_n(1) - a_2\widehat{c}_n(0) \end{pmatrix}$$

8.2.2 The sampling properties

We now show that under certain conditions $\tilde{\theta}_n$ is a consistent estimator of θ.

Theorem 8.2.1 *Suppose that $G_n(\theta)$ is an unbiased estimating function, where $G_n(\tilde{\theta}_n) = 0$ and $\mathrm{E}(G_n(\theta_0)) = 0$.*

(i) If θ is a scalar, for every n $G_n(\theta)$ is a continuous monotonically decreasing function in θ and for all θ $G_n(\theta) \xrightarrow{\mathcal{P}} \mathrm{E}(G_n(\theta))$ (notice that we do require an equicontinuous assumption), then we have $\tilde{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$.

(ii) If we can show that $\sup_\theta |G_n(\theta) - \mathrm{E}(G_n(\theta))| \xrightarrow{\mathcal{P}} 0$ and $\mathrm{E}(G_n(\theta))$ is uniquely zero at θ_0 then we have $\tilde{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$.

PROOF. The proof of case (i) is relatively straightforward (see also page 318 in Davison (2002)). The idea is to exploit the monotonicity property of $G_n(\cdot)$ to show for every $\varepsilon > 0$ $P(\widetilde{\theta}_n < \theta_0 - \varepsilon$ or $\widetilde{\theta}_n > \theta_0 + \varepsilon) \to 0$ as $n \to \infty$. The proof is best understand by making a plot of $G_n(\theta)$ with $\tilde{\theta}_n < \theta_0 - \varepsilon < \theta_0$ (see Figure 8.2). We first note that since $\mathrm{E}[G_n(\theta_0)] = 0$, then for any fixed $\varepsilon > 0$

$$G_n(\theta_0 - \varepsilon) \xrightarrow{\mathcal{P}} \mathrm{E}\big[G_n(\theta_0 - \varepsilon)\big] > 0, \tag{8.7}$$

since G_n is monotonically decreasing for all n. Now, since $G_n(\theta)$ is monotonically decreasing we see that $\tilde{\theta}_n < (\theta_0 - \varepsilon)$ implies $G_n(\tilde{\theta}_n) - G_n(\theta_0 - \varepsilon) > 0$ (and visa-versa) hence

$$P\big(\tilde{\theta}_n - (\theta_0 - \varepsilon) \leq 0\big) = P\big(G_n(\tilde{\theta}_n) - G_n(\theta_0 - \varepsilon) > 0\big).$$

But we have from (8.7) that $\mathrm{E}(G_n(\theta_0 - \varepsilon)) \xrightarrow{\mathcal{P}} \mathrm{E}(G_n(\theta_0 - \varepsilon)) > 0$. Thus $P\big(G_n(\tilde{\theta}_n) - G_n(\theta_0 - \varepsilon) > 0\big) \xrightarrow{\mathcal{P}} 0$ and

$$P\big(\tilde{\theta}_n - (\theta_0 - \varepsilon) < 0\big) \xrightarrow{\mathcal{P}} 0 \quad \text{as} \quad n \to \infty.$$

A similar argument can be used to show that that $P\big(\tilde{\theta}_n - (\theta_0 + \varepsilon) > 0\big) \xrightarrow{\mathcal{P}} 0$ as $n \to \infty$. As the above is true for all ε, together they imply that $\tilde{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$ as $n \to \infty$.

The proof of (ii) is more involved, but essentially follows the lines of the proof of Theorem 2.6.1. □

We now show normality, which will give us the variance of the limiting distribution of $\widetilde{\theta}_n$.

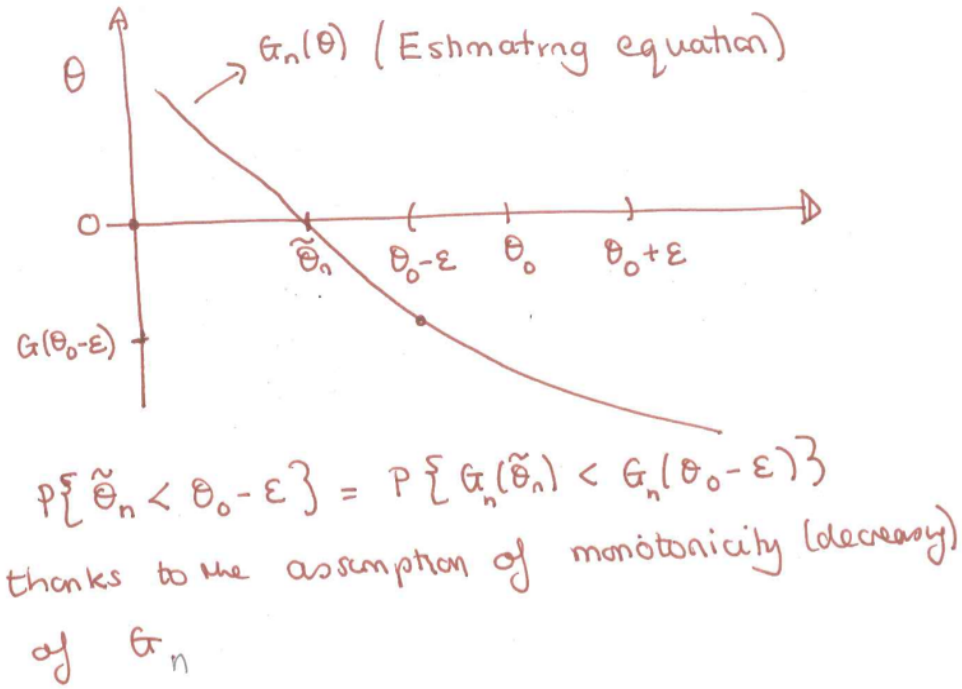

Figure 8.2: Plot of $G_n(\cdot)$

Theorem 8.2.2 *Let us suppose that $\{Y_i\}$ are iid random variables, where $\mathrm{E}[g(Y_i,\theta)]=0$. Define the estimating equation $G_n(\theta)=\frac{1}{n}\sum_{i=1}^n g(Y_i;\theta)$ and suppose $G_n(\widetilde{\theta}_n)=0$.*

Suppose that $\widetilde{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$ and the first and second derivatives of $g(Y_i,\cdot)$ have a finite expectation (we will assume that θ is a scalar to simplify notation). Then we have

$$\sqrt{n}\big(\widetilde{\theta}_n-\theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0,\frac{\mathrm{var}(g(Y_i;\theta_0))}{\big[\mathrm{E}\big(\frac{\partial g(Y_i;\theta)}{\partial\theta}\rfloor_{\theta_0}\big)\big]^2}\bigg),$$

as $n\to\infty$.

Suppose $\{Y_i\}$ are independent but not identically distributed random variables, where for all i $\mathrm{E}[g_i(Y_i,\theta)]=0$. Define the estimating equation $G_n(\theta)=\frac{1}{n}\sum_{i=1}^n g(Y_i;\theta)$ and suppose $G_n(\widetilde{\theta}_n)=0$. Suppose that $\widetilde{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$ and the first and second derivatives of $g_i(Y_i,\cdot)$ have a finite, uniformly bounded expectation (we will assume that θ is a scalar to simplify notation). Then we have

$$\sqrt{n}\big(\widetilde{\theta}_n-\theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\frac{n^{-1}\sum_{i=1}^n \mathrm{var}(g_i(Y_i;\theta_0))}{\big[\mathrm{E}\big(n^{-1}\sum_{i=1}^n \frac{\partial g(Y_i;\theta)}{\partial\theta}\rfloor_{\theta_0}\big)\big]^2}\right), \tag{8.8}$$

as $n\to\infty$.

PROOF. We use the standard Taylor expansion to prove the result (which you should be expert in by now). Using a Taylor expansion and that $\widetilde{\theta}$

$$\begin{aligned} G_n(\tilde{\theta}_n) &= G_n(\theta_0)+(\tilde{\theta}_n-\theta_0)\frac{\partial G_n(\theta)}{\partial\theta}\rfloor_{\bar{\theta}_n} \qquad (8.9)\\ \Rightarrow (\widetilde{\theta}_n-\theta_0) &= \left(\mathrm{E}\left(-\frac{\partial g_n(\theta)}{\partial\theta}\rfloor_{\theta_0}\right)\right)^{-1} G_n(\theta_0), \end{aligned}$$

where the above is due to $\frac{\partial G_n(\theta)}{\partial\theta}\rfloor_{\bar{\theta}_n} \xrightarrow{\mathcal{P}} \mathrm{E}(\frac{\partial g_n(\theta)}{\partial\theta}\rfloor_{\theta_0})$ as $n\to\infty$. Now, since $G_n(\theta_0)=\frac{1}{n}\sum_i g(Y_i;\theta)$ is the sum of iid random variables we have

$$\sqrt{n}G_n(\theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0,\underbrace{\mathrm{var}(G_n(\theta_0))}_{=var[g(Y_i;\theta_0)]}\big), \tag{8.10}$$

(since $\mathrm{E}(g(Y_i;\theta_0))=0$). Therefore (8.9) and (8.10) together give

$$\sqrt{n}\big(\tilde{\theta}_n-\theta_0\big) \xrightarrow{\mathcal{P}} \mathcal{N}\bigg(0,\frac{\mathrm{var}(g(Y_i;\theta_0))}{\big[\mathrm{E}\big(\frac{-\partial g(Y_i;\theta)}{\partial\theta}\rfloor_{\theta_0}\big)\big]^2}\bigg),$$

as required. □

In most cases $\frac{\mathrm{var}(g(Y_i;\theta_0))}{\big[\mathrm{E}\big(\frac{-\partial g(Y_i;\theta)}{\partial\theta}\rfloor_{\theta_0}\big)\big]^2} \geq I(\theta_0)^{-1}$ (where $I(\theta)$ is the Fisher information).

Example 8.2.4 (The Huber estimator) *We describe the Huber estimator which is a well known estimator of the mean which is robust to outliers. The estimator can be written as an estimating function.*

Let us suppose that $\{Y_i\}$ are iid random variables with mean θ, and density function which is symmetric about the mean θ. So that outliers do not effect the mean, a robust method of estimation is to truncate the outliers and define the function

$$g_{(c)}(Y_i;\theta) = \begin{cases} -c & Y_i < -c+\theta \\ Y_i - c & -c+\theta \leq Y_i \leq c+\theta \\ c & Y_i > c+\theta \end{cases}.$$

The estimating equation is

$$G_{c,n}(\theta) = \sum_{i=1}^{n} g_{(c)}(Y_i;\theta).$$

And we use as an estimator of θ, the $\tilde{\theta}_n$ which solves $G_{c,n}(\tilde{\theta}_n) = 0$.

(i) In the case that $c = \infty$, then we observe that $G_{\infty,n}(\theta) = \sum_{i=1}^{n}(Y_i - \theta)$, and the estimator is $\tilde{\theta}_n = \bar{Y}$. Hence without truncation, the estimator of the mean is the sample mean.

(ii) In the case that c is small, then we have truncated many observations.

Definition 8.2.2 (Generalized method of moments) *We observe from Example 8.2.1(iii,iv) that there are several estimating equations which can be used to estimate a finite number of parameters (number of estimating equations is more than the number of parameters). In this case, we can use M estimating equations to construct the estimator by minimising the L_2 criterion*

$$L_n(\alpha,\phi) = \sum_{r=1}^{M}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^r - \phi^r\Gamma(1+r\alpha^{-1})\right)^2.$$

This is an example of the generalized method of moments, which generalizes the ideas of solving estimating equations to obtain parameter estimators.

8.2.3 A worked problem

(1) Let us suppose we observe the response Y_i and regressor X_i. We assume they satisfy the random coefficient regression model

$$Y_i = (\phi + \xi_i) X_i + \varepsilon_i,$$

where $\{\xi_i\}_i$ and $\{\varepsilon_i\}_i$ are zero mean iid random variables which are independent of each other, with $\sigma_\xi^2 = \text{var}[\xi_i]$ and $\sigma_\varepsilon^2 = \text{var}[\varepsilon_i]$. In this question we will consider how to estimate ϕ, ξ_i and ε_i based on the observations $\{Y_i, X_i\}$.

(a) What is the Expectation of Y_i given (conditioned on) X_i?

(b) What is the variance of Y_i given (conditioned on) X_i?

(c) Use your answer in part (a) and least squares to obtain an explicit expression for estimating ϕ.

(d) Use your answer in part (c) to define the 'residual'.

(e) Use your answer in part (b) and (d) and least squares to obtain an explicit expression for estimating σ_ξ^2 and σ_ε^2.

(f) By conditioning on the regressors $\{X_i\}_{i=1}^n$, obtain the negative log-likelihood of $\{Y_i\}_{i=1}^n$ under the assumption of Gaussianity of ξ_i and ε_i. Explain the role that (c) and (e) plays in your maximisation algorithm.

(g) Assume that the regressors, $\{X_i\}$, are iid random variables that are independent of ε_i and ξ_i.

Show that the expectation of the *negative* log-likelihood is minimised at the true parameters ϕ, σ_ξ^2 and σ_ε^2 *even* when ξ_i and ε_i are not Gaussian.

Hint: You may need to use that $-\log x + x$ is minimum at $x = 1$.

Solution:

(a) *What is the Expectation of Y_i given (conditioned on) X_i?*

$\text{E}[Y_i|X_i] = \phi X_i$.

(b) *What is the variance of Y_i given (conditioned on) X_i?*

$\text{var}[Y_i|X_i] = \text{E}[(\xi_i X_i + \varepsilon_i)^2|X_i] = \sigma_\xi^2 X_i^2 + \sigma_\varepsilon^2$

(c) *Use your answer in part (a) and least squares to obtain an explicit expression for estimating ϕ.*

$\widehat{\phi} = \arg\min_\phi \sum_{i=1}^n (Y_i - \phi X_i)^2 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}$

(d) *Use your answer in part (c) to define the 'residual'.*

For $1 \leq i \leq n$, $\widehat{r}_i = Y_i - \widehat{\phi} X_i$

(e) *Use your answer in part (b) and (d) and least squares to obtain an explicit expression for estimating σ_ξ^2 and σ_ε^2.*

Let

$$r_i = Y_i - \mathrm{E}[Y_i] = Y_i - \phi X_i = \xi_i X_i + \varepsilon_i.$$

From (b) it is clear that $\mathrm{E}[r_i|X_i] = 0$ and $\mathrm{E}[r_i^2|X_i] = \sigma_\xi^2 X_i^2 + \sigma_\varepsilon^2$, thus we can write

$$r_i^2 = \sigma_\xi^2 X_i^2 + \sigma_\varepsilon^2 + \epsilon_i$$

where $\epsilon_i = r_i^2 - \mathrm{E}[r_i^2|X_i]$ hence $\mathrm{E}[\epsilon_i] = 0$, resembles a simple linear equation (with hetero errors). Since $\widehat{r}_i$ is an estimator of r_i we can use least squares to estimate σ_ξ^2 and σ_ε^2, where we replace r_i with $\widehat{r}_i$ and minimise

$$\sum_{i=1}^n \left(\widehat{r}_i^2 - \sigma_\xi^2 X_i^2 - \sigma_\varepsilon^2\right)^2$$

with respect to σ_ξ^2 and σ_ε^2. These gives use explicit estimators.

(f) *By conditioning on the regressors $\{X_i\}_{i=1}^n$, obtain the negative log-likelihood of $\{Y_i\}_{i=1}^n$ under the assumption of Gaussianity of ξ_t and ε_t. Explain the role that (c) and (e) plays in your maximisation algorithm.*

The log-likelihood is equal to

$$\sum_{i=1}^n \log f(Y_i|X_i;\theta).$$

We recall from (a) and (b) that $\mathrm{E}[Y_i|X_i] = \phi X_i$ and $\mathrm{var}[Y_i|X_i] = \sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2$. Therefore $Y_i|X_i \sim \mathcal{N}(\phi X_i, \sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2)$. Thus the *negative* log likelihood is proportional to

$$L(\theta; \underline{Y}_n) = \sum_{i=1}^n \left(\log[\sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2] + \frac{(Y_i - \phi X_i)^2}{\sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2} \right).$$

We choose the parameters which *minimise* $L(\theta; \underline{Y}_n)$. We note that this means we need to take the derivative of $L(\theta; \underline{Y}_n)$ with respect to the three parameters and solve using the Newton Raphson scheme. However, the estimators obtained in (c) and (d) can be used as initial values in the scheme.

(g) *Let us assume that the regressors are iid random variables. Show that the expectation of the negative log-likelihood is minimised at the true parameters ϕ, σ_ξ^2 and σ_ε^2 even when ξ_t and ε_t are not Gaussian.*

Hint: You may need to use that $-\log x + x$ is minimum at $x = 1$.

Since $\{X_i\}$ are iid random variables, $\{Y_i\}$ are iid random variables the expectation of $\frac{1}{n}L(\theta; \underline{Y}_n)$ is

$$L(\theta) = \mathrm{E}\left(\frac{1}{n}L(\theta; \underline{Y}_n)\right) = \frac{1}{n}\sum_{i=1}^{n} L_i(\theta)$$

where

$$\begin{aligned} L_i(\theta) &= \mathrm{E}\log[\sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2] + \mathrm{E}\left[\frac{(Y_i - \phi X_i)^2}{\sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2}\right] \\ &= \log[\sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2] + \frac{1}{\sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2}\mathrm{E}\left[(Y_i - \phi X_i)^2\right] \end{aligned}$$

Let θ_0 denote the true parameter in the model. Our aim is to show that

$$L(\theta) - L(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\left(L_i(\theta) - L_i(\theta_0)\right) \geq 0,$$

where equality to zero arises when $\theta = \theta_0$. Taking differences we have

$$\begin{aligned} & L_i(\theta) - L_i(\theta_0) \\ = & \log\frac{[\sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2]}{[\sigma_{0,\varepsilon}^2 + \sigma_{0,\xi}^2 X_i^2]} + \mathrm{E}\left[\frac{(Y_i - \phi X_i)^2}{\sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2}\right] - \left[\frac{(Y_i - \phi_0 X_i)^2}{\sigma_{0,\varepsilon}^2 + \sigma_{0,\xi}^2 X_i^2}\right] \\ = & -\log\frac{[\sigma_{0,\varepsilon}^2 + \sigma_{0,\xi}^2 X_i^2]}{[\sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2]} + \mathrm{E}\left[\frac{(Y_i - \phi X_i)^2}{\sigma_\varepsilon^2 + \sigma_\xi^2 X_i^2}\right] - \mathrm{E}\left[\frac{(Y_i - \phi_0 X_i)^2}{\sigma_{0,\varepsilon}^2 + \sigma_{0,\xi}^2 X_i^2}\right] \end{aligned}$$

We will show that $L_i(\theta) - L_i(\theta_0)$ is non-negative for all θ and zero when $\theta = \theta_0$. This immediately implies that θ_0 minimises the negative pseudo (pseudo because we do not assume Gaussianity) likelihood.

Our aim is to place the difference in the form $-logx + x$ plus an additional positive term (it is similar in idea to completing the square), but requires a lot of algebraic manipulation. Let

$$L_i(\theta) - L_i(\theta_0) = A_i(\theta) + B_i(\theta)$$

where

$$\begin{aligned} A_i(\theta) &= -\left(\log \frac{[\sigma^2_{0,\varepsilon} + \sigma^2_{0,\xi} X_i^2]}{[\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2]}\right) \\ B_i(\theta) &= \mathrm{E}\left[\frac{(Y_i - \phi X_i)^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}\right] - \mathrm{E}\left[\frac{(Y_i - \phi_0 X_i)^2}{\sigma^2_{0,\varepsilon} + \sigma^2_{0,\xi} X_i^2}\right]. \end{aligned}$$

First consider the difference

$$\begin{aligned} B_i(\theta) &= \mathrm{E}\left[\frac{(Y_i - \phi X_i)^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}\right] - \underbrace{\mathrm{E}\left[\frac{(Y_i - \phi_0 X_i)^2}{\sigma^2_{0,\varepsilon} + \sigma^2_{0,\xi} X_i^2}\right]}_{=(\sigma^2_{0,\varepsilon} + \sigma^2_{0,\xi} X_i^2)^{-1}\mathrm{var}(Y_i)=1} \\ &= \mathrm{E}\left[\frac{(Y_i - \phi X_i)^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}\right] - 1. \end{aligned}$$

Now replace ϕ by ϕ_0

$$\begin{aligned} B_i(\theta) &= \mathrm{E}\left[\frac{(Y_i - \phi_0 X_i)^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}\right] + \mathrm{E}\left[\frac{(Y_i - \phi X_i)^2 - (Y_i - \phi_0 X_i)^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}\right] - 1 \\ &= \mathrm{E}\left[\frac{(\varepsilon_t + \xi_i X_i)^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}\right] + \mathrm{E}\left[\frac{2(\phi - \phi_0)(Y_i - \phi_0 X_i) X_i}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}\right] + \\ &\quad \mathrm{E}\left[\frac{(\phi - \phi_0)^2 X_i^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}\right] - 1 \\ &= \frac{\mathrm{E}\left[(\varepsilon_t + \xi_i X_i)^2\right]}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2} + \mathrm{E}\left[\frac{2(\phi - \phi_0)(Y_i - \phi_0 X_i) X_i}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}\right] + \\ &\quad \frac{(\phi - \phi_0)^2 X_i^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2} - 1 \\ &= \frac{\sigma^2_{0,\varepsilon} + \sigma^2_{0,\xi} X_i^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2} + \frac{(\phi - \phi_0)^2 X_i^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2} - 1. \end{aligned}$$

Therefore, substituting this into $L_i(\theta) - L_i(\theta_0)$ we have

$$\begin{aligned} & L_i(\theta) - L_i(\theta_0) \\ = & -\log \frac{[\sigma^2_{0,\varepsilon} + \sigma^2_{0,\xi} X_i^2]}{[\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2]} + \frac{\sigma^2_{0,\varepsilon} + \sigma^2_{0,\xi} X_i^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2} + (\phi - \phi_0)^2 \frac{X_i^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2} - 1. \end{aligned}$$

Let

$$x = \frac{\sigma^2_{0,\varepsilon} + \sigma^2_{0,\xi} X_i^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}.$$

Hence

$$L_i(\theta) - L_i(\theta_0) \quad = \quad -\log x + x - 1 + (\phi - \phi_0)^2 \frac{X_i^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}.$$

Since $-\log x + x$ is minimum at $x = 1$ where it is 1, we can see that $L_i(\theta) - L_i(\theta_0)$ is non-negative and zero at $\theta = \theta_0$. As this is true for all i we have that

$$L(\theta) - L(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} (L_i(\theta) - L_i(\theta_0)) \geq 0,$$

where equality to zero arises when $\theta = \theta_0$.

This example, illustrates the versatility of the models based on the assumption of Gaussianity. Even if the Gaussian assumption does not hold, often we can obtain reasonable (consistent) estimators of the known parameters by treating the errors as if they were Gaussian.

8.3 Optimal estimating functions

As illustrated in Example 8.2.2(iii,iv) there are several different estimators of the same parameters. But which estimator does one use?

Suppose that $\{Y_i\}$ are independent random variables with mean $\{\mu_i(\theta_0)\}$ and variance $\{V_i(\theta_0)\}$, where the parametric form of $\{\mu_i(\cdot)\}$ and $\{V_i(\cdot)\}$ are known, but θ_0 is unknown. One possible estimating equation is

$$G_{1,n}(\theta) = \sum_{i=1}^{n} [Y_i - \mu_i(\theta)],$$

which is motivated by the observation $\mathrm{E}(G_{1,n}(\theta_0)) = 0$. Another estimating equation comes from the least squares criterion

$$\sum_{i=1}^{n} \left[Y_i - \mu_i(\theta)\right]^2,$$

which leads to the estimating equation

$$G_{2,n}(\theta) = \sum_{i=1}^{n} \frac{\mu_i(\theta)}{\partial\theta} \left[Y_i - \mu_i(\theta)\right],$$

again it can be seen that $\mathrm{E}(G_{2,n}(\theta_0)) = 0$. Based on the above examples, we see that by simply weighting $[Y_i - \mu_i(\theta)]$ we obtain a valid estimating equation

$$G_n^{(W)}(\theta) = \sum_{i=1}^{n} w_i(\theta) \left[Y_i - \mu_i(\theta)\right].$$

We observe that $\mathrm{E}(G_n^{(W)}(\theta_0)) = 0$, thus giving a valid estimating equation. But we need to select the weights $w_i(\theta)$. It seems reasonable to select the weights which minimise the asymptotic "variance"

$$\mathrm{var}\big(\widetilde{\theta}_n\big) \approx \frac{\sum_{i=1}^{n} \mathrm{var}(g_i(Y_i;\theta_0))}{\big[\mathrm{E}\big(\sum_{i=1}^{n} \frac{\partial g(Y_i;\theta)}{\partial\theta}\rfloor_{\theta_0}\big)\big]^2}. \tag{8.11}$$

Note the above comes from (8.8) (observe the n^{-1} has been removed, since we have not standardized $\widetilde{\theta}_n$). Since $\{Y_i\}$ are independent we observe that

$$\begin{aligned}
\mathrm{var}(G_n^{(W)}(\theta_0)) &= n^{-1} \sum_{i=1}^{n} \mathrm{var}(g_i(Y_i;\theta_0) = \sum_{i=1}^{n} w_i(\theta_0)^2 V_i(\theta_0) \\
\mathrm{E}\left(\frac{\partial G_n^{(W)}(\theta)}{\partial\theta}\rfloor_{\theta_0}\right) &= \mathrm{E}\left(\sum_{i=1}^{n} \frac{\partial g(Y_i;\theta)}{\partial\theta}\rfloor_{\theta_0}\right) \\
&= \mathrm{E}\left(\sum_{i=1}^{n} w_i'(\theta_0)\left[Y_i - \mu_i(\theta_0)\right] - \sum_{i=1}^{n} w_i(\theta_0)\mu_i'(\theta_0)\right) = -\sum_{i=1}^{n} w_i(\theta_0)\mu_i'(\theta_0).
\end{aligned}$$

Substituting the above into (8.11) gives

$$\mathrm{var}\big(\tilde{\theta}_n\big) \approx \frac{\sum_{i=1}^{n} w_i(\theta_0)^2 V_i(\theta_0)}{(\sum_{i=1}^{n} w_i(\theta_0)\mu_i'(\theta_0))^2}.$$

Now we want to choose the weights, thus the estimation function, which has the smallest variance. Therefore we look for weights which minimise the above. Since the above is a

ratio, and we observe that a small $w_i(\theta)$ leads to a large denominator but a small numerator. To resolve this, we include a Lagrangian multipler (this, essentially, minimises the numerator by controlling the magnitude of the denominator). We constrain the numerator to equal one; $(\sum_{i=1}^n w_i(\theta)\mu'(\theta))^2 = 1$ and minimise under this constraint

$$\sum_{i=1}^n w_i(\theta_0)^2 V_i(\theta) + \lambda\Big[\sum_{i=1}^n w_i(\theta)\mu_i'(\theta) - 1\Big],$$

with respect to $\{w_i(\theta)\}$ and λ. Partially differentiating the above with respect to $\{w_i(\theta)\}$ and λ and setting to zero gives for all i

$$2w_i(\theta)V_i(\theta) + \mu_i'(\theta) = 0 \text{ subject to } \sum_{i=1}^n w_i(\theta)\mu_i'(\theta) = 1.$$

Thus we choose

$$w_i(\theta) = -\frac{\mu_i'(\theta)}{2V_i(\theta)}$$

but standardize to ensure $\sum_{i=1}^n w_i(\theta)\mu_i'(\theta) = 1$;

$$w_i(\theta) = \left(\sum_{j=1}^n V_j(\theta)^{-1}\mu_j'(\theta)\right)^{-1} \frac{\mu_i'(\theta)}{V_i(\theta)}.$$

Since $\left(\sum_{j=1}^n V_j(\theta)^{-1}\mu_j'(\theta)\right)^{-1}$ is common for all weights $w_i(\theta)$ it can be ignored, thus leading to the optimal estimating function is

$$G_n^{(\mu' V^{-1})}(\theta) = \sum_{i=1}^n \frac{\mu_i'(\theta)}{V_i(\theta)}\left(Y_i - \mu_i(\theta)\right). \tag{8.12}$$

The interesting point about the optimal estimating equation, is that even if the *variance* has been mispecified, the estimating equation can still be used to consistently estimate θ (it just will not be optimal).

Example 8.3.1 *(i) Consider the case where* $\{Y_i\}$ *is such that* $\mathrm{E}[Y_i] = \mu_i(\beta) = \exp(\beta' x_i)$ *and* $\mathrm{var}(Y_i) = V_i(\beta) = \exp(\beta' x_i)$. *Then,* $\frac{d\mu(\beta' x_i)}{d\beta} = \exp(\beta' x_i)x_i$. *Substituting this yields the optimal estimating equation*

$$\sum_{i=1}^n (Y_i - e^{\beta' x_i})x_i = 0.$$

In general if $\mathrm{E}[Y_i] = \mathrm{var}[Y_i] = \mu(\beta' x_i)$, *the optimal estimating equation is*

$$\sum_{i=1}^{n} \frac{[Y_i - \mu(\beta' x_i)]}{\mu(\beta' x_i)} \mu'(\beta' x_i) x_i = 0,$$

where we use the notation $\mu'(\theta) = \frac{d\mu(\theta)}{d\theta}$. *But it is interesting to note that when* Y_i *comes from a Poisson distribution (where the main feature is that the mean and variance are equal), the above estimating equation corresponds to the score of the likelihood.*

(ii) Suppose $\{Y_i\}$ *are independent random variables where* $\mathrm{E}[Y_i] = \mu_i(\beta)$ *and* $\mathrm{var}[Y_i] = \mu_i(\beta)(1 - \mu_i(\beta))$ *(thus* $0 < \mu_i(\beta) < 1$*). Then the optimal estimating equation corresponds to*

$$\sum_{i=1}^{n} \frac{[Y_i - \mu(\beta' x_i)]}{\mu(\beta' x_i)[1 - \mu(\beta' x_i)]} \mu'(\beta' x_i) x_i = 0,$$

where we use the notation $\mu'(\theta) = \frac{d\mu(\theta)}{d\theta}$. *This corresponds to the score function of binary random variables. More of this in the next chapter!*

Example 8.3.2 *Suppose that* $Y_i = \sigma_i Z_i$ *where* σ_i *and* Z_i *are positive,* $\{Z_i\}$ *are iid random variables and the regressors* x_i *influence* σ_i *through the relation* $\sigma_i = \exp(\beta_0 + \beta_1' x_i)$. *To estimate* β_0 *and* β_1 *we can simply take logarithms of* Y_i

$$\log Y_i = \beta_0 + \beta_1' x_i + \log Z_i.$$

Least squares can be used to estimate β_0 *and* β_1. *However, care needs to be taken since in general* $\mathrm{E}[\log Z_i] \neq 0$, *this will mean the least squares estimator of the intercept* β_0 *will be biased, as it estimates* $\beta_0 + \mathrm{E}[\log Z_i]$.

Examples where the above model can arise is $Y_i = \lambda_i Z_i$ *where* $\{Z_i\}$ *are iid with exponential density* $f(z) = \exp(-z)$. *Observe this means that* Y_i *is also exponential with density* $\lambda_i^{-1} \exp(-y/\lambda_i)$.

Remark 8.3.1 (Weighted least squares) *Suppose that* $\mathrm{E}[Y_i] = \mu_i(\theta)$ *and* $\mathrm{var}[Y_i] = V_i(\theta)$, *motivated by the normal distribution, we can construct the weighted least squared criterion*

$$\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \left[\frac{1}{V_i(\theta)} (Y_i - \mu_i(\theta))^2 + \log V_i(\theta) \right].$$

Taking derivatives, we see that this corresponds to the estimating equation

$$\begin{aligned} G_n(\theta) &= \sum_{i=1}^{n} \left[-\frac{2}{V_i(\theta)} \{Y_i - \mu_i(\theta)\} \frac{d\mu_i(\theta)}{d\theta} - \frac{1}{V_i(\theta)^2} \{Y_i - \mu_i(\theta)\}^2 \frac{dV_i(\theta)}{d\theta} + \frac{1}{V_i(\theta)} \frac{dV_i(\theta)}{d\theta} \right] \\ &= G_{1,n}(\theta) + G_{2,n}(\theta) \end{aligned}$$

where

$$\begin{aligned} G_{1,n}(\theta) &= -2 \sum_{i=1}^{n} \frac{1}{V_i(\theta)} \{Y_i - \mu_i(\theta)\} \frac{d\mu_i(\theta)}{d\theta} \\ G_{2,n}(\theta) &= -\sum_{i=1}^{n} \left[\frac{1}{V_i(\theta)^2} \{Y_i - \mu_i(\theta)\}^2 \frac{dV_i(\theta)}{d\theta} - \frac{1}{V_i(\theta)} \frac{dV_i(\theta)}{d\theta} \right]. \end{aligned}$$

Observe that $\mathrm{E}[G_{1,n}(\theta_0)] = 0$ *and* $\mathrm{E}[G_{2,n}(\theta_0)] = 0$*, which implies that* $\mathrm{E}[G_n(\theta_0)] = 0$*. This proves that the true parameter* θ_0 *corresponds to either a local minimum or saddle point of the weighted least squares criterion* $\mathcal{L}_n(\theta)$*. To show that it is the global minimum one must use an argument similar to that given in Section 8.2.3.*

Remark 8.3.2 *We conclude this section by mentioning that one generalisation of estimating equations is the generalised method of moments. We observe the random vectors* $\{Y_i\}$ *and it is known that there exist a function* $g(\cdot;\theta)$ *such that* $\mathrm{E}(g(Y_i;\theta_0)) = 0$*. To estimate* θ_0*, rather than find the solution of* $\frac{1}{n}\sum_{i=1}^{n} g(Y_i;\theta)$*, a matrix* M_n *is defined and the parameter which mimimises*

$$\left(\frac{1}{n} \sum_{i=1}^{n} g(Y_i;\theta) \right)' M_n \left(\frac{1}{n} \sum_{i=1}^{n} g(Y_i;\theta) \right)$$

is used as an estimator of θ*.*

Chapter 9

Generalised Linear Models

To motivate the GLM approach let us briefly overview linear models.

9.1 An overview of linear models

Let us consider the two competing linear nested models

$$\begin{aligned} \text{Restricted model:} \qquad & Y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{i,j} + \varepsilon_i, \\ \text{Full model:} \qquad & Y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{i,j} + \sum_{j=q+1}^{p} \beta_j x_{i,j} + \varepsilon_i, \end{aligned} \tag{9.1}$$

where $\{\varepsilon_i\}$ are iid random variables with mean zero and variance σ^2. Let us suppose that we observe $\{(Y_i, x_{i,j})\}_{i=1}^n$, where $\{Y_i\}$ are normal. The classical method for testing H_0 : Model 0 against H_A : Model 1 is to use the F-test (ANOVA). That is, let $\hat{\sigma}_R^2$ be the residual sum of squares under the null and $\hat{\sigma}_F^2$ be the residual sum of squares under the alternative. Then the F-statistic is

$$F = \frac{\left(S_R^2 - S_F^2\right)/(p-q)}{\hat{\sigma}_F^2},$$

where

$$\begin{aligned} S_F^2 &= \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} \hat{\beta}_j^F x_{i,j})^2 \quad S_R^2 = \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{q} \hat{\beta}_j^R x_{i,j})^2 \\ \sigma_F^2 &= \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} \hat{\beta}_j^F x_{i,j})^2. \end{aligned}$$

and under the null $F \sim F_{p-q,n-p}$. Moreover, if the sample size is large $(p-q)F \xrightarrow{\mathcal{D}} \chi^2_{p-q}$. We recall that the residuals of the full model are $r_i = Y_i - \hat{\beta}_0 - \sum_{j=1}^{q} \hat{\beta}_j x_{i,j} - \sum_{j=q+1}^{p} \hat{\beta}_j x_{i,j}$ and the residual sum of squares S_F^2 is used to measure how well the linear model fits the data (see STAT612 notes).

The F-test and ANOVA are designed specifically for linear models. In this chapter the aim is to generalise

- Model specification.
- Estimation
- Testing.
- Residuals.

to a larger class of models.

To generalise we will be in using a log-likelihood framework. To see how this fits in with the linear regression, let us now see how ANOVA and the log-likelihood ratio test are related. Suppose that σ^2 is known, then the log-likelihood ratio test for the above hypothesis is

$$\frac{1}{\sigma^2}\big(S_R^2 - S_F^2\big) \sim \chi^2_{p-q},$$

where we note that since $\{\varepsilon_i\}$ is Gaussian, this is the exact distribution and not an asymptotic result. In the case that σ^2 is unknown and has to be replaced by its estimator $\hat{\sigma}_F^2$, then we can either use the approximation

$$\frac{1}{\hat{\sigma}_F^2}\big(S_R^2 - S_F^2\big) \xrightarrow{\mathcal{D}} \chi^2_{p-q}, \quad n \to \infty,$$

or the exact distribution

$$\frac{\big(S_R^2 - S_F^2\big)/(p-q)}{\hat{\sigma}_F^2} \sim F_{p-q,n-p},$$

which returns us to the F-statistic.

On the other hand, if the variance σ^2 is unknown we return to the log-likelihood ratio statistic. In this case, the log-likelihood ratio statistic is

$$\log \frac{S_R^2}{S_F^2} = \log\left(1 + \frac{(S_F^2 - S_R^2)}{\hat{\sigma}_F^2}\right) \xrightarrow{\mathcal{D}} \chi^2_{p-q},$$

recalling that $\frac{1}{\widehat{\sigma}}\sum_{i=1}^{n}(Y_i - \widehat{\beta}x_i) = n$. We recall that by using the expansion $\log(1+x) = x + O(x^2)$ we obtain

$$\begin{aligned}\log\frac{S_R^2}{S_F^2} &= \log\left(1 + \frac{(S_R^2 - S_F^2)}{S_F^2}\right) \\ &= \frac{S_R^2 - S_F^2}{S_F^2} + o_p(1).\end{aligned}$$

Now we know the above is approximately χ^2_{p-q}. But it is straightforward to see that by dividing by $(p-q)$ and multiplying by $(n-p)$ we have

$$\begin{aligned}\frac{(n-p)}{(p-q)}\log\frac{S_R^2}{S_F^2} &= \frac{(n-p)}{(p-q)}\log\left(1 + \frac{(S_R^2 - S_F^2)}{S_F^2}\right) \\ &= \frac{(S_R^2 - S_F^2)/(p-q)}{\hat{\sigma}_F^2} + o_p(1) = F + o_p(1).\end{aligned}$$

Hence we have transformed the log-likelihood ratio test into the F-test, which we discussed at the start of this section. The ANOVA and log-likelihood methods are asymptotically equivalent.

In the case that $\{\varepsilon_i\}$ are non-Gaussian, but the model is linear with iid random variables, the above results also hold. However, in the case that the regressors have a nonlinear influence on the response and/or the response is not normal we need to take an alternative approach. Through out this section we will encounter such models. We will start by focussing on the following two problems:

(i) How to model the relationship between the response and the regressors when the reponse is non-Gaussian, and the model is nonlinear.

(ii) Generalise ANOVA for nonlinear models.

9.2 Motivation

Let us suppose $\{Y_i\}$ are independent random variables where it is believed that the regressors x_i (x_i is a p-dimensional vector) has an influence on $\{Y_i\}$. Let us suppose that Y_i is a binary random variable taking either zero or one and $\mathrm{E}(Y_i) = P(Y_i = 1) = \pi_i$.

How to model the relationship between Y_i and x_i? A simple approach, is to use a linear model, ie. let $\mathrm{E}(Y_i) = \beta' x_i$, But a major problem with this approach is that $\mathrm{E}(Y_i)$,

is a probability, and for many values of β, $\beta' x_i$ will lie outside the unit interval - hence a linear model is not meaningful. However, we can make a nonlinear transformation which transforms the a linear combination of the regressors to the unit interval. Such a meaningful transformation forms an important component in statistical modelling. For example let

$$\mathrm{E}(Y_i) = \pi_i = \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} = \mu(\beta' x_i),$$

this transformation lies between zero and one. Hence we could just use nonlinear regression to estimate the parameters. That is rewrite the model as

$$Y_i = \mu(\beta' x_i) + \underbrace{\varepsilon_i}_{Y_i - \mu(\beta' x_i)}$$

and use the estimator $\widehat{\beta}_i$, where

$$\widehat{\beta}_n = \arg\min_{\beta} \sum_{i=1}^{n} \Big(Y_i - \mu(\beta' x_i) \Big)^2, \tag{9.2}$$

as an estimator of β. This method consistently estimates the parameter β, but there are drawbacks. We observe that Y_i are not iid random variables and

$$Y_i = \mu(\beta' x_i) + \sigma_i \epsilon_i$$

where $\{\epsilon_i = \frac{Y_i - \mu(\beta' x_i)}{\sqrt{Y_i}}\}$ are iid random variables and $\sigma_i = \sqrt{\mathrm{var} Y_i}$. Hence Y_i has a heterogeneous variance. However, the estimator in (9.2) gives each observation the same weight, without taking into account the variability between observations (which will result in a large variance in the estimator). To account for this one can use the weighted leasts squares estimator

$$\widehat{\beta}_n = \arg\min_{\beta} \sum_{i=1}^{n} \frac{(Y_i - \mu(\beta' x_i))^2}{\mu(\beta' x_i)(1 - \mu(\beta' x_i))}, \tag{9.3}$$

but there is no guarantee that such an estimator is even consistent (the only way to be sure is to investigate the corresponding estimating equation).

An alternative approach is to use directly use estimating equations (refer to Section 8.2). The the simplest one solves

$$\sum_{i=1}^{n} (Y_i - \mu(\beta' x_i)) = 0,$$

where $\mu(\beta' x_i)$. However, this solution does not lead to an estimator with the smallest "variance". Instead we can use the "optimal estimation equation" given in Section 8.3 (see equation 8.12). Using (8.12) the optimal estimating equation is

$$\begin{aligned} & \sum_{i=1}^{n} \frac{\mu_i'(\theta)}{V_i(\theta)} (Y_i - \mu_i(\theta)) \\ = & \sum_{i=1}^{n} \frac{(Y_i - \mu(\beta' x_i))}{\mu(\beta' x_i)[1-\mu(\beta' x_i)]} \frac{\partial \mu(\beta' x_i)}{\partial \beta} = \sum_{i=1}^{n} \frac{(Y_i - \mu(\beta' x_i))}{\mu(\beta' x_i)[1-\mu(\beta' x_i)]} \mu'(\beta' x_i) x_i = 0, \end{aligned}$$

where we use the notation $\mu'(\theta) = \frac{d\mu(\theta)}{d\theta}$ (recall $var[Y_i] = \mu(\beta' x_i)(1-\mu(\beta' x_i))$). We show below (using the GLM machinery) that this corresponds to the score function of the log-likelihood function.

The GLM approach is a general framework for a wide class of distributions. We recall that in Section 1.6 we considered maximum likelihood estimation for iid random variables which come from the natural exponential family. Distributions in this family include the normal, binary, binomial and Poisson, amongst others. We recall that the natural exponential family has the form

$$f(y;\theta) = \exp\Big(y\theta - \kappa(\theta) + c(y)\Big),$$

where $\kappa(\theta) = b(\eta^{-1}(\theta))$. To be a little more general we will suppose that the distribution can be written as

$$f(y;\theta) = \exp\left(\frac{y\theta - \kappa(\theta)}{\phi} + c(y,\phi)\right), \tag{9.4}$$

where ϕ is a nuisance parameter (called the disperson parameter, it plays the role of the variance in linear models) and θ is the parameter of interest. We recall that examples of exponential models include

(i) The exponential distribution is already in natural exponential form with $\theta = \lambda$ and $\phi = 1$. The log density is

$$\log f(y;\theta) = -\lambda y + \log \lambda.$$

(ii) For the binomial distribution we let $\theta = \log(\frac{\pi}{1-\pi})$ and $\phi = 1$, since $\log(\frac{\pi}{1-\pi})$ is invertible this gives

$$\log f(y;\theta) = \log f(y; \log\frac{\pi}{1-\pi}) = \big(y\theta - n\log\big(\frac{\exp(\theta)}{1+\exp(\theta)}\big) + \log\binom{n}{y}\big).$$

(iii) For the normal distribution we have that

$$\begin{aligned}\log f(y;\mu,\sigma^2) &= \left(-\frac{(y-\mu)^2}{2\sigma^2}-\frac{1}{2}\log\sigma^2-\frac{1}{2}\log(2\pi)\right)\\ &= \frac{-y^2+2\mu y-\mu^2}{2\sigma^2}-\frac{1}{2}\log\sigma^2-\frac{1}{2}\log(2\pi).\end{aligned}$$

Suppose $\mu=\mu(\beta' x_i)$, whereas the variance σ^2 is constant for all i, then σ^2 is the scale parameter and we can rewrite the above as

$$\log f(y;\mu,\sigma^2) = \frac{\underbrace{\mu}_{\theta}y-\underbrace{\mu^2/2}_{\kappa(\theta)}}{\sigma^2}-\underbrace{\left(-\frac{y^2}{2\sigma^2}-\frac{1}{2}\log\sigma^2-\frac{1}{2}\log(2\pi)\right)}_{=c(y,\phi)}.$$

(iv) The Poisson log distribution can be written as

$$\log f(y;\mu)=y\log\mu-\mu+\log y!,$$

Hence $\theta=\log\mu$, $\kappa(\theta)=-\exp(\theta)$ and $c(y)=\log y!$.

(v) Other members in this family include, Gamma, Beta, Multinomial and inverse Gaussian to name but a few.

Remark 9.2.1 (Properties of the exponential family (see Chapter 1 for details)) *(i) Using Lemma 1.6.3 (see Section 1.6) we have* $\mathrm{E}(Y)=\kappa'(\theta)$ *and* $\mathrm{var}(Y)=\kappa''(\theta)\phi$.

(ii) If the distribution has a "full rank parameter space" (number of parameters is equal to the number of sufficient statistics) and $\theta(\eta)$ *(where* η *is the parameter of interest) is a diffeomorphism then the second derivative of the log-likelihood is non-negative. To see why we recall for a one-dimensional exponential family distribution of the form*

$$f(y;\theta)=\exp\left(y\theta-\kappa(\theta)+c(y)\right),$$

the second derivative of the log-likelihood is

$$\frac{\partial^2\log f(y;\theta)}{\partial\theta^2}=-\kappa''(\theta)=-\mathrm{var}[Y].$$

If we reparameterize the likelihood in terms of η, such that $\theta(\eta)$ then

$$\frac{\partial^2 \log f(y;\theta(\eta))}{\partial \eta^2} = y\theta''(\eta) - \kappa'(\theta)\theta''(\eta) - \kappa''(\theta)[\theta'(\eta)]^2.$$

Since $\theta(\eta)$ is a diffeomorphism between the space spanned by η to the space spanned by θ, $\log f(y;\theta(\eta))$ will be a deformed version of $\log f(y;\theta)$ but it will retain properties such concavity of the likelihood with respect to η.

GLM is a method which generalises the methods in linear models to the exponential family (recall that the normal model is a subclass of the exponential family). In the GLM setting it is usually assumed that the response variables $\{Y_i\}$ are independent random variables (but not identically distributed) with log density

$$\log f(y_i;\theta_i) = \left(\frac{y_i\theta_i - \kappa(\theta_i)}{\phi} + c(y_i,\phi)\right), \tag{9.5}$$

where the parameter θ_i depends on the regressors. The regressors influence the response through a linear predictor $\eta_i = \beta' x_i$ and a link function, which connects $\beta' x_i$ to the mean $\mathrm{E}(Y_i) = \mu(\theta_i) = \kappa'(\theta_i)$.

Remark 9.2.2 (Modelling the mean) *The main "philosophy/insight" of GLM is connecting the mean $\mu(\theta_i)$ of the random variable (or sufficient statistic) to a linear transform of the regressor $\beta' x_i$. The "link" function g is a monotonic (bijection) such that $\mu(\theta_i) = g^{-1}(\beta' x_i)$, and usually needs to be selected. The main features of the link function depends on the distribution. For example*

(i) If Y_i are positive then the link function g^{-1} should be positive (since the mean is positive).

(i) If Y_i take binary values the link function g^{-1} should lie between zero and one (it should be probability).

Let $g : \mathbb{R} \to \mathbb{R}$ be a bijection such that $g(\mu(\theta_i)) = \eta_i = \beta' x_i$. If we ignore the scale parameter, then by using Lemma 1.6.3 (which relates the mean and variance of sufficient statistics to $\kappa(\theta_i)$) we have

$$\begin{aligned} \frac{d\kappa(\theta_i)}{d\theta_i} &= g^{-1}(\eta_i) = \mu(\theta_i) = \mathrm{E}(Y_i) \\ \theta_i &= \mu^{-1}(g^{-1}(\eta_i)) = \theta(\eta_i) \\ \mathrm{var}(Y_i) &= \frac{d^2\kappa(\theta_i)}{d\theta_i^2} = \frac{d\mu(\theta_i)}{d\theta_i}. \end{aligned} \tag{9.6}$$

Based on the above and (9.5) the log likelihood function of $\{Y_i\}$ is

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \left(\frac{Y_i\theta(\eta_i) - \kappa(\theta(\eta_i))}{\phi} + c(Y_i, \phi) \right).$$

Remark 9.2.3 (Concavity of the likelihood with regressors) *We mentioned in Remark 9.2.1 that natural exponential family has full rank and $\theta(\eta)$ is a reparameterisation in terms of η, then $\frac{\partial^2 log f(y;\eta)}{\partial \eta^2}$ is non-positive definite, thus $\log f(y;\theta)$ is a concave function. We now show that the likelihood in the presence of regressors in also concave.*

We recall that

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \Big(Y_i\theta(\eta_i) - \kappa(\theta(\eta_i)) + c(Y_i) \Big).$$

where $\eta_i = \beta' x_i$. Differentiating twice with respect to β gives

$$\nabla^2_\beta \mathcal{L}_n(\beta) = \mathbf{X}' \sum_{i=1}^{n} \frac{\partial^2 \log f(Y_i; \theta(\eta_i))}{\partial \eta_i^2} \mathbf{X},$$

where $\mathbf{X}$ is the design matrix corresponding to the regressors. We mentioned above that $\frac{\partial^2 \log f(Y_i;\eta_i)}{\partial \eta_i^2}$ is non-positive definite for all i which in turn implies that its sum is non-positive definite. Thus $\mathcal{L}_n(\beta)$ is concave in terms of β, hence it is simple to maximise.

Example: Suppose the link function is in canonical form i.e. $\theta(\eta_i) = \beta' x_i$ (see the following example), the log-likelihood is

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \Big(Y_i\beta' x_i - \kappa(\beta' x_i) + c(Y_i) \Big).$$

which has second derivative

$$\nabla^2_\beta \mathcal{L}_n(\beta) = -\mathbf{X}' \sum_{i=1}^{n} \kappa''(\beta' x_i)\mathbf{X}$$

which is clearly non-positive definite.

The choice of link function is rather subjective. One of the most popular is the canonical link which we define below.

Definition 9.2.1 (The canonical link function) *Every distribution within the exponential family has a* canonical link *function, this is where $\eta_i = \theta_i$. This immediately implies that $\mu_i = \kappa'(\eta_i)$ and $g(\kappa'(\theta_i)) = g(\kappa'(\eta_i)) = \eta_i$ (hence g is inverse function of κ').*

The canonical link is often used because it make the calculations simple (it also saves one from "choosing a link function"). We observe with the canonical link the log-likelihood of $\{Y_i\}$ is

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \Big(\frac{Y_i \beta' x_i - \kappa(\beta' x_i)}{\phi} + c(Y_i, \phi) \Big).$$

Example 9.2.1 (The log-likelihood and canonical link function)

(i) The canonical link for the exponential $f(y_i; \lambda_i) = \lambda_i \exp(-\lambda_i y_i)$ *is* $\theta_i = -\lambda_i = \beta' x_i$*, and* $\lambda = -\beta' x_i$*. The log-likelihood is*

$$\sum_{i=1}^{n} \Big(Y_i \beta' x_i - \log(\beta' x_i) \Big).$$

(ii) The canonical link for the binomial is $\theta_i = \beta' x_i = \log(\frac{\pi_i}{1-\pi_i})$*, hence* $\pi_i = \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)}$*. The log-likelihood is*

$$\sum_{i=1}^{n} \Big(Y_i \beta' x_i + n_i \log\big(\frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)}\big) + \log \binom{n_i}{Y_i} \Big).$$

(iii) The canonical link for the normal is $\theta_i = \beta' x_i = \mu_i$*. The log-likelihood is*

$$\Big(- \frac{(Y_i - \beta' x_i)^2}{2\sigma^2} + \frac{1}{2} \log \sigma^2 + \frac{1}{2} \log(2\pi) \Big),$$

which is the usual least squared criterion. If the canonical link were not used, we would be in the nonlinear least squares setting, with log-likelihood

$$\Big(- \frac{(Y_i - g^{-1}(\beta' x_i))^2}{2\sigma^2} + \frac{1}{2} \log \sigma^2 + \frac{1}{2} \log(2\pi) \Big),$$

(iv) The canonical link for the Poisson is $\theta_i = \beta' x_i = \log \lambda_i$*, hence* $\lambda_i = \exp(\beta' x_i)$*. The log-likelihood is*

$$\sum_{i=1}^{n} \Big(Y_i \beta' x_i - \exp(\beta' x_i) + \log Y_i! \Big).$$

However, as mentioned above, the canonical link is simply used for its mathematical simplicity. There exists other links, which can often be more suitable.

Remark 9.2.4 (Link functions for the Binomial) *We recall that the link function is defined as a monotonic function g, where $\eta_i = \beta' x_i = g(\mu_i)$. The choice of link function is up to the practitioner. For the binomial distribution it is common to let g^{-1} = a well known distribution function. The motivation for this is that for the Binomial distribution $\mu_i = n_i \pi_i$ (where π_i is the probability of a 'success'). Clearly $0 \leq \pi_i \leq 1$, hence using g^{-1} = distribution function (or survival function) makes sense. Examples include*

(i) The Logistic link, this is the canonical link function, where $\beta' x_i = g(\mu_i) = \log(\frac{\pi_i}{1-\pi_i}) = \log(\frac{\mu_i}{n_i - \mu_i})$.

(i) The Probit link, where $\pi_i = \Phi(\beta' x_i)$, Φ is the standard normal distribution function and the link function is $\beta' x_i = g(\mu_i) = \Phi^{-1}(\mu_i/n_i)$.

(ii) The extreme value link, where the distribution function is $F(x) = 1 - \exp(-exp(x))$. Hence in this case the link function is $\beta' x_i = g(\mu_i) = \log(-\log(1 - \mu_i/n_i))$.

Remark 9.2.5 *GLM is the motivation behind single index models where $\mathrm{E}[Y_i|X_i] = \mu(\sum_{j=1}^{p} \beta_j x_{ij})$, where both the parameters $\{\beta_j\}$ and the link function $\mu(\cdot)$ is unknown.*

9.3 Estimating the parameters in a GLM

9.3.1 The score function for GLM

The score function for generalised linear models has a very interesting form, which we will now derive.

From now on, we will suppose that $\phi_i \equiv \phi$ for all t, and that ϕ is known. Much of the theory remains true without this restriction, but this makes the derivations a bit cleaner, and is enough for all the models we will encounter.

With this substitution, recall that the log-likelihood is

$$\mathcal{L}_n(\beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i - \kappa(\theta_i)}{\phi} + c(Y_i, \phi) \right\} = \sum_{i=1}^{n} \ell_i(\beta, \phi),$$

where

$$\ell_i(\beta, \phi) = \left\{ \frac{Y_i \theta_i - \kappa(\theta_i)}{\phi} + c(Y_i, \phi) \right\}$$

and $\theta_i = \theta(\eta_i)$.

For the MLE of β, we need to solve the *likelihood equations*

$$\frac{\partial \mathcal{L}_n}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0 \quad \text{for } j = 1, \ldots, p.$$

Observe that

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j} = \frac{(Y_i - \kappa'(\theta_i))}{\phi} \theta'(\eta_i) x_{ij}.$$

Thus the score equation is

$$\frac{\partial \mathcal{L}_n}{\partial \beta_j} = \sum_{i=1}^{n} \frac{[Y_i - \kappa'(\theta_i)]}{\phi} \theta'(\eta_i) x_{ij} = 0 \quad \text{for } j = 1, \ldots, p. \tag{9.7}$$

Remark 9.3.1 (Connection to optimal estimating equations) *Recall from (8.12) the optimal estimating equation is*

$$G_n(\beta) = \sum_{i=1}^{n} \frac{1}{V_i(\beta)} \left(Y_i - \mu_i(\beta)\right) \frac{\partial}{\partial \beta_j} \mu_i(\beta), \tag{9.8}$$

we now show this is equivalent to (9.7). Using classical results on the exponential family (see chapter 1) we have

$$\begin{aligned} \mathrm{E}[Y_i] &= \kappa'(\theta) = \mu_i(\beta) \\ \mathrm{var}[Y_i] &= \kappa''(\theta) = V_i(\beta). \end{aligned}$$

We observe that

$$\frac{\partial}{\partial \beta_j} \mu_i(\beta) = \underbrace{\frac{\partial \mu(\theta_i)}{\partial \theta_i}}_{=V_i(\beta)} \frac{\partial \theta_i}{\partial \beta_j} = V_i(\beta) \theta'(\eta_i) x_{ij},$$

substituting this into (9.8) gives

$$\frac{\partial \mathcal{L}_n}{\partial \beta_j} = \sum_{i=1}^{n} \frac{[Y_i - \kappa'(\theta_i)]}{\phi} \theta'(\eta_i) x_{ij} = 0$$

which we see corresponds to the score of the likelihood.

To obtain an interesting expression for the above, recall that

$$\mathrm{var}(Y_i) = \phi \mu'(\theta_i) \text{ and } \eta_i = g(\mu_i),$$

and let $\mu'(\theta_i) = V(\mu_i)$. Since $V(\mu_i) = \frac{d\mu_i}{d\theta_i}$, inverting the derivative we have $\frac{d\theta_i}{d\mu_i} = 1/V(\mu_i)$. Furthermore, since $\frac{d\eta_i}{d\mu_i} = g'(\mu_i)$, inverting the derivative we have $\frac{d\mu_i}{d\eta_i} = 1/g'(\mu_i)$. By the chain rule for differentiation and using the above we have

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \beta_j} &= \frac{d\ell_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{d\ell_i}{d\eta_i}\frac{\partial \eta_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} \\
&= \frac{d\ell_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} \\
&= \frac{d\ell_i}{d\theta_i}\left(\frac{d\mu_i}{d\theta_i}\right)^{-1}\left(\frac{d\eta_i}{d\mu_i}\right)^{-1}\frac{\partial \eta_i}{\partial \beta_j} \\
&= \frac{(Y_i - \kappa'(\theta_i))}{\phi}(\kappa''(\theta_i))^{-1}(g'(\mu_i))^{-1}x_{ij} \\
&= \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)}
\end{aligned} \tag{9.9}$$

Thus the likelihood equations we have to solve for the MLE of β are

$$\sum_{i=1}^{n}\frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} = \sum_{i=1}^{n}\frac{(Y_i - g^{-1}(\beta' x_i))x_{ij}}{\phi V(g^{-1}(\beta' x_i))g'(\mu_i)} = 0, \qquad 1 \le j \le p, \tag{9.10}$$

(since $\mu_i = g^{-1}(\beta' x_i)$).

(9.10) has a very similar structure to the Normal equations arising in ordinary least squares.

Example 9.3.1 *(i) Normal $\{Y_i\}$ with mean $\mu_i = \beta' x_i$.*

Here, we have $g(\mu_i) = \mu_i = \beta' x_i$ so $g'(\mu_i) = \frac{dg(\mu_i)}{d\mu_i} \equiv 1$; also $V(\mu_i) \equiv 1$, $\phi = \sigma^2$, so the equations become

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \beta' x_i)x_{ij} = 0.$$

Ignoring the factor σ^2, the LHS is the jth element of the vector $X^T(Y - X\beta')$, so the equations reduce to the Normal equations of least squares:

$$X^T(Y - X\beta') = 0 \quad \text{or equivalently} \quad X^T X\beta' = X^T Y.$$

(ii) Poisson $\{Y_i\}$ with log-link function, hence mean $\mu_i = \exp(\beta' x_i)$ (hence $g(\mu_i) = \log \mu_i$). This time, $g'(\mu_i) = 1/\mu_i$, $\mathrm{var}(Y_i) = V(\mu_i) = \mu_i$ and $\phi = 1$. Substituting $\mu_i = \exp(\beta' x_i)$, into (9.10) gives

$$\sum_{i=1}^{n}(Y_i - e^{\beta' x_i})x_{ij} = 0.$$

9.3.2 The GLM score function and weighted least squares

The GLM score has a very interesting relationship with weighted least squares. First recall that the GLM takes the form

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} = \sum_{i=1}^{n} \frac{(Y_i - g^{-1}(\beta' x_i))x_{ij}}{\phi V_i g'(\mu_i)} = 0, \qquad 1 \leq j \leq p. \tag{9.11}$$

Next let us construct the weighted least squares criterion. Since $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = \phi V_i$, the weighted least squares criterion corresponding to $\{Y_i\}$ is

$$S_i(\beta) = \sum_{i=1}^{n} \frac{(Y_i - \mu(\theta_i))^2}{\phi V_i} = \sum_{i=1}^{n} \frac{(Y_i - g^{-1}(\beta' x_i))^2}{\phi V_i}.$$

The weighted least squares criterion $\mathcal{S}_i$ is independent of the underlying distribution and has been constructed using the first two moments of the random variable. Returning to the weighted least squares estimator, we observe that this is the solution of

$$\frac{\partial S_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} + \sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial V_i}\frac{\partial V_i}{\partial \beta_j} = 0 \quad 1 \leq j \leq p,$$

where $s_i(\beta) = \frac{(Y_i - \mu(\theta_i))^2}{\phi V_i}$. Now let us compare $\frac{\partial S_i}{\partial \beta_j}$ with the estimating equations corresponding to the GLM (those in (9.11)). We observe that (9.11) and the first part of the RHS of the above are the same, that is

$$\sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} = 0.$$

In other words, the GLM estimating equations corresponding to the exponential family and the weighted least squares estimating equations are closely related (as are the corresponding estimators). However, it is simpler to solve $\sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial \mu_i}\frac{\partial \mu_i}{\beta_j} = 0$ than $\frac{\partial S_i}{\partial \beta_j} = 0$.

As an aside, note that since at the true β the derivatives are

$$\mathrm{E}\Big(\sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\Big) = 0 \text{ and } \mathrm{E}\Big(\frac{\partial S_i}{\partial \beta_j}\Big) = 0,$$

then this implies that the other quantity in the partial sum, $\mathrm{E}\big(\frac{\partial S_i}{\partial \beta_j}\big)$ is also zero, i.e.

$$\mathrm{E}\Big(\sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial V_i}\frac{\partial V_i}{\partial \beta_j}\Big) = 0.$$

9.3.3 Numerical schemes

The Newton-Raphson scheme

It is clear from the examples above that usually there does not exist a simple solution for the likelihood estimator of β. However, we can use the Newton-Raphson scheme to estimate β (and thanks to the concavity of the likelihood it is guaranteed to converge to the maximum). We will derive an interesting expression for the iterative scheme. Other than the expression being useful for implementation, it also highlights the estimators connection to weighted least squares.

We recall that the Newton Raphson scheme is

$$(\beta^{(m+1)})' = (\beta^{(m)})' - (H^{(m)})^{-1}u^{(m)}$$

where the $p \times 1$ gradient vector $u^{(m)}$ is

$$u^{(m)} = \left(\frac{\partial \mathcal{L}_n}{\partial \beta_1}, \ldots, \frac{\partial \mathcal{L}_n}{\partial \beta_p}\right)' \rfloor_{\beta=\beta^{(m)}}$$

and the $p \times p$ Hessian matrix $H^{(m)}$ is given by

$$H_{jk}^{(m)} = \frac{\partial^2 \mathcal{L}_n(\beta)}{\partial \beta_j \partial \beta_k} \rfloor_{\beta=\beta^{(m)}},$$

for $j, k = 1, 2, \ldots, p$, both $u^{(m)}$ and $H^{(m)}$ being evaluated at the current estimate $\beta^{(m)}$.

By using (9.9), the score function at the mth iteration is

$$\begin{aligned} u_j^{(m)} &= \frac{\partial \mathcal{L}_n(\beta)}{\partial \beta_j} \rfloor_{\beta=\beta^{(m)}} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} \rfloor_{\beta=\beta^{(m)}} \\ &= \sum_{i=1}^{n} \frac{d\ell_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} \rfloor_{\beta=\beta^{(m)}} = \sum_{i=1}^{n} \frac{d\ell_i}{d\eta_i} \rfloor_{\beta=\beta^{(m)}} x_{ij}. \end{aligned}$$

The Hessian at the ith iteration is

$$\begin{aligned} H_{jk}^{(m)} &= \frac{\partial^2 \mathcal{L}_i(\beta)}{\partial\beta_j \partial\beta_k}\rfloor_{\beta=\beta^{(m)}} = \sum_{i=1}^{n} \frac{\partial^2 \ell_i}{\partial\beta_j \partial\beta_k}\rfloor_{\beta=\beta^{(m)}} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\beta_k}\left(\frac{\partial\ell_i}{\partial\beta_j}\right)\rfloor_{\beta=\beta^{(m)}} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\beta_k}\left(\frac{\partial\ell_i}{\partial\eta_i}x_{ij}\right)\rfloor_{\beta=\beta^{(m)}} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\eta_i}\left(\frac{\partial\ell_i}{\partial\eta_i}x_{ij}\right)x_{ik} \\ &= \sum_{i=1}^{n} \frac{\partial^2\ell_i}{\partial\eta_i^2}\rfloor_{\beta=\beta^{(m)}} x_{ij}x_{ik} \end{aligned} \tag{9.12}$$

Let $s^{(m)}$ be an $n \times 1$ vector with

$$s_i^{(m)} = \frac{\partial\ell_i}{\partial\eta_i}\rfloor_{\beta=\beta^{(m)}}$$

and define the $n \times n$ diagonal matrix $\widetilde{W}^{(m)}$ with entries

$$\widetilde{W}_{ii} = -\frac{d^2\ell_i}{d\eta_i^2}.$$

Then we have $u^{(m)} = X^T s^{(m)}$ and $H = -X^T\widetilde{W}^{(m)}X$ and the Newton-Raphson iteration can succinctly be written as

$$\begin{aligned} (\beta^{(m+1)})' &= (\beta^{(m)})' - (H^{(m)})^{-1}u^{(m)} \\ &= (\beta^{(m)})' + (X^T\widetilde{W}^{(m)}X)^{-1}X^T s^{(m)}. \end{aligned}$$

Fisher scoring for GLMs

Typically, partly for reasons of tradition, we use a modification of this in fitting statistical models. The matrix $\widetilde{W}$ is replaced by W, another diagonal $n \times n$ matrix with

$$W_{ii}^{(m)} = \mathrm{E}(\widetilde{W}_{ii}^{(m)}|\beta^{(m)}) = \mathrm{E}\left(-\frac{d^2\ell_i}{d\eta_i^2}|\beta^{(m)}\right).$$

Using the results in Section 1.6 we have

$$W_{ii}^{(m)} = \mathrm{E}\left(-\frac{d^2\ell_i}{d\eta_i^2}|\beta^{(m)}\right) = \mathrm{var}\left(\frac{d\ell_i}{d\eta_i}|\beta^{(m)}\right)$$

so that $W = \text{var}(s^{(m)}|\beta^{(m)})$, and the matrix is non-negative-definite.

Using the Fisher score function the iteration becomes

$$(\beta^{(i+1)})' = (\beta^{(m)})' + (X^T W^{(m)} X)^{-1} X^T s^{(m)}.$$

Iteratively reweighted least squares

The iteration

$$(\beta^{(i+1)})' = (\beta^{(m)})' + (X^T W^{(m)} X)^{-1} X^T s^{(m)} \tag{9.13}$$

is similar to the solution for least squares estimates in linear models

$$\beta = (X^T X)^{-1} X^T Y$$

or more particularly the related *weighted least squares* estimates:

$$\beta = (X^T W X)^{-1} X^T W Y$$

In fact, (9.13) can be re-arranged to have exactly this form. Algebraic manipulation gives

$$\begin{aligned} (\beta^{(m)})' &= (X^T W^{(m)} X)^{-1} X^T W^{(m)} X (\beta^{(m)})' \\ (X^T W^{(m)} X)^{-1} X^T s^{(m)} &= (X^T W^{(m)} X)^{-1} X^T W^{(m)} (W^{(m)})^{-1} s^{(m)}. \end{aligned}$$

Therefore substituting the above into (9.13) gives

$$\begin{aligned} (\beta^{(m+1)})' &= (X^T W^{(m)} X)^{-1} X^T W^{(m)} X (\beta^{(m)})' + (X^T W^{(m)} X)^{-1} X^T W^{(m)} (W^{(m)})^{-1} s^{(m)} \\ &= (X^T W^{(m)} X)^{-1} X^T W^{(m)} \big(X (\beta^{(m)})' + (W^{(m)})^{-1} s^{(m)} \big) \\ &:= (X^T W^{(m)} X)^{-1} X^T W^{(m)} Z^{(m)}. \end{aligned}$$

One reason that the above equation is of interest is that it has the 'form' of weighted least squares. More precisely, it has the form of a weighted least squares regression of $Z^{(m)}$ on X with the diagonal weight matrix $W^{(m)}$. That is let $z_i^{(m)}$ denote the ith element of the vector $Z^{(m)}$, then $\beta^{(m+1)}$ minimises the following weighted least squares criterion

$$\sum_{i=1}^{n} W_i^{(m)} \big(z_i^{(m)} - \beta' x_i \big)^2.$$

Of course, in reality $W_i^{(m)}$ and $z_i^{(m)}$ are functions of $\beta^{(m)}$, hence the above is often called *iteratively reweighted least squares.*

9.3.4 Estimating of the dispersion parameter ϕ

Recall that in the linear model case, the variance σ^2 did not affect the estimation of β.

In the general GLM case, continuing to assume that $\phi_i = \phi$, we have

$$s_i = \frac{d\ell_i}{d\eta_i} = \frac{d\ell_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i} = \frac{Y_i - \mu_i}{\phi V(\mu_i) g'(\mu_i)}$$

and

$$\begin{aligned} W_{ii} &= \text{var}(s_i) = \frac{\text{var}(Y_i)}{\{\phi V(\mu_i) g'(\mu_i)\}^2} = \frac{\phi V(\mu_i)}{\{\phi V(\mu_i) g'(\mu_i)\}^2} \\ &= \frac{1}{\phi V(\mu_i)(g'(\mu_i))^2} \end{aligned}$$

so that $1/\phi$ appears as a scale factor in W and s, but otherwise does not appear in the estimating equations or iteration for $\widehat{\beta}$. Hence ϕ does not play a role in the estimation of β.

As in the Normal/linear case, (a) we are less interested in ϕ, and (b) we see that ϕ can be separately estimated from β.

Recall that $\text{var}(Y_i) = \phi V(\mu_i)$, thus

$$\frac{E((Y_i - \mu_i)^2)}{V(\mu_i)} = \phi$$

We can use this to suggest a simple estimator for ϕ:

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} = \frac{1}{n-p}\sum_{i=1}^{T} \frac{(Y_i - \widehat{\mu}_i)^2}{\mu'(\widehat{\theta}_i)}$$

where $\widehat{\mu}_i = g^{-1}(\widehat{\beta}' x_i)$ and $\widehat{\theta}_i = \mu^{-1} g^{-1}(\widehat{\beta}' x_i)$. Recall that the above resembles estimators of the residual variance. Indeed, it can be argued that the distribution of the above is close to χ^2_{n-p}.

Remark 9.3.2 *We mention that a slight generalisation of the above is when the dispersion parameter satisfies $\phi_i = a_i \phi$, where a_i is known. In this case, an estimator of the ϕ is*

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{a_i V(\widehat{\mu}_i)} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{a_i \mu'(\widehat{\theta}_i)}$$

9.3.5 Deviance, scaled deviance and residual deviance

Scaled deviance

Instead of *minimising* the sum of squares (which is done for linear models) we have been *maximising* a log-likelihood $\mathcal{L}_i(\beta)$. Furthermore, we recall

$$S(\hat{\beta}) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \big(Y_i - \hat{\beta}_0 - \sum_{j=1}^{q} \hat{\beta}_j x_{i,j} - \sum_{j=q+1}^{p} \hat{\beta}_j x_{i,j}\big)^2$$

is a numerical summary of how well the linear model fitted, $S(\widehat{\beta}) = 0$ means a perfect fit. A perfect fit corresponds to the Gaussian log-likelihood $-\frac{n}{2}\log\sigma^2$ (the likelihood cannot be larger than this).

In this section we will define the equivalent of residuals and square residuals for GLM.

What is the *best* we can do in fitting a GLM? Recall

$$\ell_i = \frac{Y_i\theta_i - \kappa(\theta_i)}{\phi} + c(Y_i, \phi)$$

so

$$\frac{d\ell_i}{d\theta_i} = 0 \iff Y_i - \kappa'(\theta_i) = 0$$

A model that achieves this equality for all i is called *saturated* (the same terminology is used for linear models). In other words, will need one free parameter for each observation. Denote the corresponding θ_i by $\widetilde{\theta}_i$, i.e. the solution of $\kappa'(\widetilde{\theta}_i) = Y_i$.

Consider the differences

$$\begin{aligned} 2\{\ell_i(\widetilde{\theta}_i) - \ell_i(\theta_i)\} &= \frac{2}{\phi}\{Y_i(\widetilde{\theta}_i - \theta_t) - \kappa(\widetilde{\theta}_i) + \kappa(\theta_i)\} \geq 0 \\ \text{and } 2\sum_{i=1}^{n}\Big\{\ell_i(\widetilde{\theta}_i) - \ell_i(\theta_i)\Big\} &= \frac{2}{\phi}\{Y_i(\widetilde{\theta}_i - \theta_t) - \kappa(\widetilde{\theta}_i) + \kappa(\theta_i)\}. \end{aligned}$$

Maximising the likelihood is the same as *minimising* the above quantity, which is always non-negative, and is 0 only if there is a perfect fit for all the i^{th} observations. This is analogous to linear models, where maximising the normal likelihood is the same as minimising least squares criterion (which is zero when the fit is best). Thus $\mathcal{L}_n(\widetilde{\theta}) = \sum_{i=1}^{n}\ell_i(\tilde{\theta}_i)$ provides a baseline value for the log-likelihood in much the same way that $-\frac{n}{2}\log\sigma^2$ provides a baseline in least squares (Gaussian set-up).

Example 9.3.2 (The normal linear model) $\kappa(\theta_i) = \frac{1}{2}\theta_i^2$, $\kappa'(\theta_i) = \theta_i = \mu_i$, $\tilde{\theta}_i = Y_t$ *and* $\phi = \sigma^2$ *so*

$$2\{\ell_i(\widetilde{\theta}_n) - \ell_i(\theta_i)\} = \frac{2}{\sigma^2}\{Y_i(Y_i - \mu_i) - \frac{1}{2}Y_i^2 + \frac{1}{2}\mu_i^2\} = (Y_i - \mu_i)^2/\sigma^2.$$

Hence for Gaussian observations $2\{\ell_i(\widetilde{\theta}_i) - \ell_i(\theta_i)\}$ *corresponds to the classical residual squared.* □

In general, let

$$D_i = 2\{Y_i(\widetilde{\theta}_i - \hat{\theta}_i) - \kappa(\widetilde{\theta}_i) + \kappa(\hat{\theta}_i)\}$$

We call $D = \sum_{i=1}^n D_i$ the *deviance* of the model. If ϕ is present, let

$$\frac{D}{\phi} = 2\{\mathcal{L}_n(\widetilde{\theta}) - \mathcal{L}_n(\widehat{\theta})\}.$$

$\phi^{-1}D$ is the *scaled deviance*. Thus the residual deviance plays the same role for GLM's as does the residual sum of squares for linear models.

Interpreting D_i

We will now show that

$$D_i \;\; = \;\; 2\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - \kappa(\tilde{\theta}_i) + \kappa(\hat{\theta}_i)\} \approx \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

To show the above we require expression for $Y_i(\tilde{\theta}_i - \hat{\theta}_i)$ and $\kappa(\tilde{\theta}_i) - \kappa(\hat{\theta}_i)$. We use Taylor's theorem to expand κ and κ' about $\hat{\theta}_i$ to obtain

$$\kappa(\tilde{\theta}_i) \approx \kappa(\hat{\theta}_i) + (\tilde{\theta}_i - \hat{\theta}_i)\kappa'(\hat{\theta}_i) + \frac{1}{2}(\tilde{\theta}_i - \hat{\theta}_i)^2\kappa''(\hat{\theta}_i) \tag{9.14}$$

and

$$\kappa'(\tilde{\theta}_i) \approx \kappa'(\hat{\theta}_i) + (\tilde{\theta}_i - \hat{\theta}_i)\kappa''(\hat{\theta}_i) \tag{9.15}$$

But $\kappa'(\tilde{\theta}_i) = Y_i$, $\kappa'(\hat{\theta}_i) = \widehat{\mu}_i$ and $\kappa''(\hat{\theta}_i) = V(\widehat{\mu}_i)$, so (9.14) becomes

$$\begin{aligned} \kappa(\tilde{\theta}_i) &\approx \kappa(\hat{\theta}_i) + (\tilde{\theta}_i - \hat{\theta}_i)\widehat{\mu}_i + \frac{1}{2}(\tilde{\theta}_i - \hat{\theta}_i)^2 V(\widehat{\mu}_i) \\ \Rightarrow \kappa(\tilde{\theta}_i) - \kappa(\hat{\theta}_i) &\approx (\tilde{\theta}_i - \hat{\theta}_i)\widehat{\mu}_i + \frac{1}{2}(\tilde{\theta}_i - \hat{\theta}_i)^2 V(\widehat{\mu}_i), \end{aligned} \tag{9.16}$$

and (9.15) becomes

$$\begin{aligned} Y_i &\approx \widehat{\mu}_i + (\tilde{\theta}_i - \hat{\theta}_i)V(\widehat{\mu}_i) \\ \Rightarrow Y_i - \widehat{\mu}_i &\approx (\tilde{\theta}_i - \hat{\theta}_i)V(\widehat{\mu}_i) \end{aligned} \tag{9.17}$$

Now substituting (9.16) and (9.17) into D_i gives

$$\begin{aligned} D_i &= 2\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - \kappa(\tilde{\theta}_i) + \kappa(\hat{\theta}_i)\} \\ &\approx 2\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - (\tilde{\theta}_i - \hat{\theta}_i)\widehat{\mu}_i - \frac{1}{2}(\tilde{\theta}_i - \hat{\theta}_i)^2 V(\widehat{\mu}_i)\} \\ &\approx (\tilde{\theta}_i - \hat{\theta}_i)^2 V(\widehat{\mu}_i) \approx \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}. \end{aligned}$$

Recalling that $\text{var}(Y_i) = \phi V(\mu_i)$ and $\text{E}(Y_i) = \mu_i$, $\phi^{-1}D_i$ behaves like a standardised squared residual. The signed square root of this approximation is called the *Pearson residual*. In other words

$$\text{sign}(Y_i - \widehat{\mu}_i) \times \sqrt{\frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}} \tag{9.18}$$

is called a Pearson residual. The distribution theory for this is very approximate, but a rule of thumb is that *if the model fits*, the scaled deviance $\phi^{-1}D$ (or in practice $\widehat{\phi}^{-1}D$) $\approx \chi^2_{n-p}$.

Deviance residuals

The analogy with the normal example can be taken further. The square roots of the individual terms in the residual sum of squares are the residuals, $Y_i - \beta' x_i$.

We use the square roots of the individual terms in the deviances residual in the same way. However, the classical residuals can be both negative and positive, and the deviances residuals should behave in a similar way. But what sign should be used? The most obvious solution is to use

$$r_i = \begin{cases} -\sqrt{D_i} & \text{if } Y_i - \widehat{\mu}_i < 0 \\ \sqrt{D_i} & \text{if } Y_i - \widehat{\mu}_i \geq 0 \end{cases}$$

Thus we call the quantities $\{r_i\}$ the *deviance residuals*. Observe that the deviance residuals and Pearson residuals (defined in (9.18)) are the same up to the standardisation $\sqrt{V(\widehat{\mu}_i)}$.

Diagnostic plots

We recall that for linear models we would often plot the residuals against the regressors to check to see whether a linear model is appropriate or not. One can make similar diagnostics plots which have exactly the same form as linear models, except that deviance residuals are used instead of ordinary residuals, and linear predictor values instead of fitted values.

9.4 Limiting distributions and standard errors of estimators

In the majority of examples we have considered in the previous sections (see, for example, Section 2.2) we observed iid $\{Y_i\}$ with distribution $f(\cdot;\theta)$. We showed that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \approx \mathcal{N}\left(0, I(\theta)^{-1}\right),$$

where $I(\theta) = \int -\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta)dx$ ($I(\theta)$ is Fisher's information). However this result was based on the observations being iid. In the more general setting where $\{Y_i\}$ are independent but not identically distributed it can be shown that

$$\left(\widehat{\beta} - \beta\right) \approx \mathcal{N}_p(0, (I(\beta))^{-1})$$

where now $I(\beta)$ is a $p \times p$ matrix (of the entire sample), where (using equation (9.12)) we have

$$(I(\beta))_{jk} = E\left(-\frac{\partial^2 \mathcal{L}_n(\beta)}{\partial \beta_j \partial \beta_k}\right) = \mathrm{E}\left(-\sum_{i=1}^{n} \frac{d^2\ell_i}{d\eta_i^2} x_{ij}x_{ik}\right) = (X^T W X)_{jk}.$$

Thus for large samples we have

$$\left(\widehat{\beta} - \beta\right) \approx \mathcal{N}_p(0, (X^T W X)^{-1}),$$

where W is evaluated at the MLE $\widehat{\beta}$.

Analysis of deviance

How can we test hypotheses about models, and in particular decide which explanatory variables to include? The two close related methods we will consider below are the log-likelihood ratio test and an analogue of the analysis of variance (ANOVA), called the analysis of deviance.

Let us concentrate on the simplest case, of testing a full vs. a reduced model. Partition the model matrix X and the parameter vector β as

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

where X_1 is $n \times q$ and β_1 is $q \times 1$ (this is analogous to equation (9.1) for linear models). The full model is $\eta = X\beta' = X_1\beta_1 + X_2\beta_2$ and the reduced model is $\eta = X_1\beta_1'$. We wish to test $H_0 : \beta_2 = 0$, i.e. that the reduced model is adequate for the data.

Define the rescaled deviances for the full and reduced models

$$\frac{D_R}{\phi} = 2\{\mathcal{L}_n(\tilde{\theta}) - \sup_{\beta_2=0,\beta_1} \mathcal{L}_n(\theta)\}$$

and

$$\frac{D_F}{\phi} = 2\{\mathcal{L}_n(\tilde{\theta}) - \sup_{\beta_1,\beta_2} \mathcal{L}_n(\beta)\}$$

where we recall that $\mathcal{L}_n(\tilde{\theta}) = \sum_{i=1}^{T} \ell_t(\tilde{\theta}_i)$ is likelihood of the saturated model defined in Section 9.3.5. Taking differences we have

$$\frac{D_R - D_F}{\phi} = 2\Big\{ \sup_{\beta_1,\beta_2} \mathcal{L}_n(\beta) - \sup_{\beta_2=0,\beta_1} \mathcal{L}_n(\theta)\Big\}$$

which is the likelihood ratio statistic.

The results in Theorem 3.1.1, equation (3.7) (the log likelihood ratio test for composite hypothesis) also hold for observations which are not identically distributed. Hence using a generalised version of Theorem 3.1.1 we have

$$\frac{D_R - D_F}{\phi} = 2\Big\{ \sup_{\beta_1,\beta_2} \mathcal{L}_n(\beta) - \sup_{\beta_2=0,\beta_1} \mathcal{L}_n(\theta)\Big\} \xrightarrow{\mathcal{D}} \chi^2_{p-q}.$$

So we can conduct a test of the adequacy of the reduced model $\frac{D_R - D_F}{\phi}$ by referring to a χ^2_{p-q}, and rejecting H_0 if the statistic is too large (p-value too small). If ϕ is not present in the model, then we are good to go.

If ϕ is present (and unknown), we estimate ϕ with

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_t)} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\mu'(\widehat{\theta}_n)}$$

(see Section 9.3.4). Now we consider $\frac{D_R - D_F}{\widehat{\phi}}$, we can then continue to use the χ^2_{p-q} distribution, but since we are estimating ϕ we can use the statistic

$$\frac{D_R - D_F}{p-q} \div \frac{D_F}{n-p} \quad \text{against} \quad F_{(p-q),(n-p)},$$

as in the normal case (compare with Section 9.1).

9.5 Examples

Example 9.5.1 *Question Suppose that $\{Y_i\}$ are independent random variables with the canonical exponential family, whose logarithm satisfies*

$$\log f(y;\theta_i) = \frac{y\theta_i - \kappa(\theta_i)}{\phi} + c(y;\phi),$$

where ϕ is the dispersion parameter. Let $E(Y_i) = \mu_i$. Let $\eta_i = \beta' x_i = \theta_i$ (hence the canonical link is used), where x_i are regressors which influence Y_i. *[14]*

(a) (m) Obtain the log-likelihood of $\{(Y_i, x_i)\}_{i=1}^n$.

(ii) Denote the log-likelihood of $\{(Y_i, x_i)\}_{i=1}^n$ as $\mathcal{L}_n(\beta)$. Show that

$$\frac{\partial \mathcal{L}_n}{\partial \beta_j} = \sum_{i=1}^n \frac{(Y_i - \mu_i)x_{i,j}}{\phi} \quad \text{and} \quad \frac{\partial^2 \mathcal{L}_n}{\partial \beta_k \partial \beta_j} = -\sum_{i=1}^n \frac{\kappa''(\theta_i)x_{i,j}x_{i,k}}{\phi}.$$

(b) Let Y_i have Gamma distribution, where the log density has the form

$$\log f(Y_i;\mu_i) = \frac{-Y_i/\mu_i - \log \mu_i}{\nu^{-1}} + \left\{ -\frac{1}{\nu^{-1}} \log \nu^{-1} + \log \Gamma(\nu^{-1}) \right\} + \left\{ \nu^{-1} - 1 \right\} \log Y_i$$

$E(Y_i) = \mu_i$, $var(Y_i) = \mu_i^2/\nu$ and $\nu_i = \beta' x_i = g(\mu_i)$.

(m) What is the canonical link function for the Gamma distribution and write down the corresponding likelihood of $\{(Y_i, x_i)\}_{i=1}^n$.

(ii) Suppose that $\eta_i = \beta' x_i = \beta_0 + \beta_1 x_{i,1}$. Denote the likelihood as $\mathcal{L}_n(\beta_0, \beta_1)$. What are the first and second derivatives of $\mathcal{L}_n(\beta_0, \beta_1)$?

(iii) Evaluate the Fisher information matrix at β_0 and $\beta_1 = 0$.

(iv) Using your answers in (ii,iii) and the mle of β_0 with $\beta_1 = 0$, derive the score test for testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$.

Solution

(a) (m) The general log-likelihood for $\{(Y_i, x_i)\}$ with the canonical link function is

$$\mathcal{L}_n(\beta, \phi) = \sum_{i=1}^n \left(\frac{Y_i(\beta' x_i - \kappa(\beta' x_i))}{\phi} + c(Y_i, \phi) \right).$$

(ii) In the differentiation use that $\kappa'(\theta_i) = \kappa'(\beta' x_i) = \mu_i$.

(b) (m) For the gamma distribution the canonical link is $\theta_i = \eta_i = -1/\mu_i = -1/beta' x_i$. *Thus the log-likelihood is*

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \frac{1}{\nu}\Big(Y_i(\beta' x_i) - \log(-1/\beta' x_i)\Big) + c(\nu_1, Y_i),$$

where $c(\cdot)$ *can be evaluated.*

(ii) Use part (ii) above to give

$$\begin{aligned} \frac{\partial \mathcal{L}_n(\beta)}{\partial \beta_j} &= \nu^{-1} \sum_{i=1}^{n} \Big(Y_i + 1/(\beta' x_i)\Big) x_{i,j} \\ \frac{\partial \mathcal{L}_n(\beta)}{\partial \beta_i \partial \beta_j} &= -\nu^{-1} \sum_{i=1}^{n} \frac{1}{(\beta' x_i)}\Big) x_{i,i} x_{i,j} \end{aligned}$$

(iii) Take the expectation of the above at a general β_0 *and* $\beta_1 = 0$.

(iv) Using the above information, use the Wald test, Score test or log-likelihood ratio test.

Example 9.5.2 *Question: It is a belief amongst farmers that the age of a hen has a negative influence on the number of eggs she lays and the quality the eggs. To investigate this, m hens were randomly sampled. On a given day, the total number of eggs and the number of bad eggs that each of the hens lays is recorded. Let* N_i *denote the total number of eggs hen i lays,* Y_i *denote the number of bad eggs the hen lays and* x_i *denote the age of hen i.*

It is known that the number of eggs a hen lays follows a Poisson distribution and the quality (whether it is good or bad) of a given egg is an independent event (independent of the other eggs).

Let N_i *be a Poisson random variable with mean* λ_i, *where we model* $\lambda_i = \exp(\alpha_0 + \gamma_1 x_i)$ *and* π_i *denote the probability that hen i lays a bad egg, where we model* π_i *with*

$$\pi_i = \frac{\exp(\beta_0 + \gamma_1 x_i)}{1 + \exp(\beta_0 + \gamma_1 x_i)}.$$

Suppose that $(\alpha_0, \beta_0, \gamma_1)$ *are unknown parameters.*

(a) Obtain the likelihood of $\{(N_i, Y_i)\}_{i=1}^{m}$.

(b) Obtain the estimating function (score) of the likelihood and the Information matrix.

(c) Obtain an iterative algorithm for estimating the unknown parameters.

(d) For a given $\alpha_0, \beta_0, \gamma_1$, evaluate the average number of bad eggs a 4 year old hen will lay in one day.

(e) Describe in detail a method for testing $H_0 : \gamma_1 = 0$ against $H_A : \gamma_1 \neq 0$.

Solution

(a) Since the canonical links are being used the log-likelihood function is

$$\begin{aligned}
\mathcal{L}_m(\alpha_0, \beta_0, \gamma_1) &= \mathcal{L}_m(\underline{Y}|\underline{N}) + \mathcal{L}_m(\underline{N}) \\
&= \sum_{i=1}^{m} \left(Y_i \underline{\beta} \underline{x}_i - N_i \log(1 + \exp(\underline{\beta} \underline{x}_i)) + N_i \underline{\alpha} \underline{x}_i - \underline{\alpha} \underline{x}_i + \log \binom{N_i}{Y_i} + \log N_i! \right) \\
&\propto \sum_{i=1}^{m} \left(Y_i \underline{\beta} \underline{x}_i - N_i \log(1 + \exp(\underline{\beta} \underline{x}_i)) + N_i \underline{\alpha} \underline{x}_i - \underline{\alpha} \underline{x}_i \right).
\end{aligned}$$

where $\underline{\alpha} = (\alpha_0, \gamma_1)'$, $\underline{\beta} = (\beta_0, \gamma_1)'$ and $\underline{x}_i = (1, x_i)$.

(b) We know that if the canonical link is used the score is

$$\nabla \mathcal{L} = \sum_{i=1}^{m} \phi^{-1} \big(Y_i - \kappa'(\beta' x_i) \big) = \sum_{i=1}^{m} \big(Y_i - \mu_i \big)$$

and the second derivative is

$$\nabla^2 \mathcal{L} = -\sum_{i=1}^{m} \phi^{-1} \kappa''(\beta' x_i) = -\sum_{i=1}^{m} \mathrm{var}(Y_i).$$

Using the above we have for this question the score is

$$\begin{aligned}
\frac{\partial \mathcal{L}_m}{\partial \alpha_0} &= \sum_{i=1}^{m} \big(N_i - \lambda_i \big) \\
\frac{\partial \mathcal{L}_m}{\partial \beta_0} &= \sum_{i=1}^{m} \big(Y_i - N_i \pi_i \big) \\
\frac{\partial \mathcal{L}_m}{\partial \gamma_1} &= \sum_{i=1}^{m} \Big(\big(N_i - \lambda_i \big) + \big(Y_i - N_i \pi_i \big) \Big) x_i.
\end{aligned}$$

The second derivative is

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}_m}{\partial \alpha_0^2} &= -\sum_{i=1}^{m} \lambda_i \quad \frac{\partial^2 \mathcal{L}_m}{\partial \alpha_0 \partial \gamma_1} = -\sum_{i=1}^{m} \lambda_i x_i \\
\frac{\partial^2 \mathcal{L}_m}{\partial \beta_0^2} &= -\sum_{i=1}^{m} N_i \pi_i (1-\pi_i) \quad \frac{\partial^2 \mathcal{L}_m}{\partial \beta_0 \partial \gamma_1} = -\sum_{i=1}^{m} N_i \pi_i (1-\pi_i) x_i \\
\frac{\partial^2 \mathcal{L}_m}{\partial \gamma_1 2} &= -\sum_{i=1}^{m} \Big(\lambda_i + N_i \pi_i (1-\pi_i) \Big) x_i^2.
\end{aligned}$$

Observing that $E(N_i) = \lambda_i$ *the information matrix is*

$$I(\theta) = \begin{pmatrix}
\sum_{i=1}^{m} \lambda_i & 0 & \sum_{i=1}^{m} \lambda_i \pi_i (1-\pi_i) x_i \\
0 & \sum_{i=1}^{m} \lambda_i \pi_i (1-\pi_i) & \sum_{i=1}^{m} \lambda_i \pi_i (1-\pi_i) x_i \\
\sum_{i=1}^{m} \lambda_i \pi_i (1-\pi_i) x_i & \sum_{i=1}^{m} \lambda_i \pi_i (1-\pi_i) x_i & \sum_{i=1}^{m} \Big(\lambda_i + \lambda_i \pi_i (1-\pi_i) \Big) x_i^2
\end{pmatrix}.$$

(c) We can estimate $\theta_0 = (\alpha_0, \beta_0, \gamma_1)$ *using Newton-Raphson with Fisher scoring, that is*

$$\theta_i = \theta_i + I(\theta_i)^{-1} S_{i-1}$$

where

$$S_{i-1} = \begin{pmatrix}
\sum_{i=1}^{m} \big(N_i - \lambda_i \big) \\
\sum_{i=1}^{m} \big(Y_i - N_i \pi_i \big) \\
\sum_{i=1}^{m} \Big(\big(N_i - \lambda_i \big) + \big(Y_i - N_i \pi_i \big) \Big) x_i.
\end{pmatrix}.$$

(d) We note that given the regressor $x_i = 4$*, the average number of bad eggs will be*

$$\begin{aligned}
E(Y_i) &= E(E(Y_i|N_i)) = E(N_i \pi_i) = \lambda_i \pi_i \\
&= \frac{\exp(\alpha_0 + \gamma_1 x_i) \exp(\beta_0 + \gamma_1 x_i)}{1 + \exp(\beta_0 + \gamma_1 x_i)}.
\end{aligned}$$

(e) Give either the log likelihood ratio test, score test or Wald test.

Chapter 10

Count Data

In the previous chapter we generalised the linear model framework to the exponential family. GLM is often used for modelling count data, in these cases usually the Binomial, Poisson or Multinomial distributions are used.

Types of data and the distribution:

Distribution	Regressors	Response variables
Binomial	x_i	$\mathbf{Y}_i = (Y_i, N - Y_i) = (Y_{i,1}, Y_{i,2})$
Poission	x_i	$\mathbf{Y}_i = Y_i$
Multinomial	x_i	$\mathbf{Y_i} = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,m})$ $(\sum_j Y_{i,j} = N)$

Distribution	Probabilities
Binomial	$P(Y_{i,1} = k, Y_{i,2} = N - k) = \binom{N}{k}(1 - \pi(\beta' x_i))^{N-k}\pi(\beta' x_i)^k$
Poission	$P(Y_i = k) = \frac{\lambda(\beta' x_i)^k \exp(-\beta' x_i)}{k!}$
Multinomial	$P(Y_{i,1} = k_1, \ldots, Y_{i,m} = k_m) = \binom{N}{k_1,\ldots,k_m}\pi_1(\beta' x_i)^{k_1} \ldots \pi_m(\beta' x_i)^{k_m}$

In this section we will be mainly dealing with count data where the regressors tend to be ordinal (not continuous regressors). This type of data normally comes in the form of a contingency table. One of the most common type of contingency table is the two by two table, and we will consider this in the Section below.

Towards the end of this chapter we use estimating equations to estimate the parameters in overdispersed models.

10.1 Two by Two Tables

Consider the following 2×2 contingency table

	Male	Female	Total
Blue	25	35	60
Pink	15	25	40
Total	40	60	100

Given the above table, one can ask if there is an association between gender and colour preference. The standard method is test for independence. However, we could also pose question in a different way: are proportion of females who like blue the same as the proportion of males who like blue. In this case we can (equivalently) test for equality of proportions (this equivalance usually only holds for 2 by 2 tables).

There are various methods for testing the above hypothesis

- The log-likelihood ratio test.
- The Score test
- The Wald test.
- Through Pearson residuals (which is the main motivation of the chi-squared test for independence).

There can be so many tests for doing the same thing. But recall from Section 2.8.2 that asymptotically all of these tests are equivalent; for a large enough sample size their p-values are nearly the same.

We go through some examples in the following section.

10.1.1 Tests for independence

Approach 1: Pearson and log-likelihood ratio test

The chi-square test for independence is based upon the Pearson residuals:

$$T_1 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

where $O_{i,j}$ are the observed numbers and E_{ij} are the expected numbers under independence. We recall that by modelling the counts are a multinomial distribution we can show that the test statistic T_1 is asymptotically equivalent to the a log-likelihood ratio test.

Approach 2: Score test

Let us consider the altenative approach, testing for equality of proportions. Let π_M denote the proportion of males who prefer pink over blue and π_F the proportion of females who prefer pink over blue. Suppose we want to test that $H_0 : \pi_F = \pi_M$ against $H_0 : \pi_F \neq \pi_M$. On method for testing the above hypothesis is to use the test for equality of proportions using the Wald test, which gives the test statistic

$$T_2 = \frac{\hat{\pi}_F - \hat{\pi}_M}{I(\pi)^{-1/2}} = \frac{\hat{\pi}_F - \hat{\pi}_M}{\sqrt{\widehat{\pi}\left(\frac{1}{N_F} + \frac{1}{N_M}\right)}},$$

where

$$\widehat{\pi} = \frac{N_{\mathrm{M,P}} + N_{F,P}}{N_M + N_F}$$

and N_M N_F correspond to the number of males and females and $N_{M,P}$ and $N_{F,P}$ the number of males and females who prefer pink.

Approach 3: modelling

An alternative route for conducting the test, is to parameterise π_M and π_F and do a test based on the parametrisation. For example, without loss of generality we can rewrite π_M and π_F as

$$\pi_F = \frac{\exp(\gamma)}{1 + \exp(\gamma)} \qquad \pi_M = \frac{\exp(\gamma + \delta)}{1 + \exp(\gamma + \delta)}.$$

Hence using this parameterisation, the above test is equivalent to testing $H_0 : \delta = 0$ against $H_A : \delta \neq 0$. We can then use the log likelihood ratio test to do the test.

10.2 General contingency tables

Consider the following experiment. Suppose we want to know whether ethnicity plays a role in the number of children a females has. We interview a sample of women, where we

	1	2	3
Background A	20	23	28
Background B	14	27	23

determine her ethnicity and the number of children. The data is collected below in the form of a 3×2 contingency table.

How can such data arise? There are several ways this data could have been collected, and this influences the model we choose to fit to this data. Consider the general $R \times C$ table, with cells indexed by (i, j). Note that in the above example $R = 2$ and $C = 3$.

(a) The subjects arise at random, the study continues until a fixed time elapses. Each subject is categorised according to two variables. Suppose the number in cell (i, j) is Y_{ij}, then it is reasonable to assume $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$ for some $\{\lambda_{ij}\}$, which will be the focus of study. In this case the distribution is

$$P(Y = y) = \prod_{i=1}^{C} \prod_{j=1}^{R} \frac{\lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij})}{y_{ij}!}$$

(b) The total number of subjects is fixed at N, say. The numbers in cells follow a multinomial distribution: $(Y_{ij}) \sim M(N; (\pi_{ij}))$:

$$P(Y = y) = \frac{N!}{\prod_{i=1}^{C} \prod_{j=1}^{R} y_{ij}!} \prod_{i=1}^{C} \prod_{j=1}^{R} \pi_{ij}^{y_{ij}}$$

if $\sum_i \sum_j y_{ij} = N$.

(c) One margin is fixed: say $\{y_{+j} = \sum_{i=1}^{C} y_{ij}\}$ for each $j = 1, 2, \ldots, R$. In each column, we have an independent multinomial sample

$$P(Y = y) = \prod_{j=1}^{R} \left(\frac{y_{+j}!}{\prod_{i=1}^{C} y_{ij}!} \prod_{i=1}^{C} \rho_{ij}^{y_{ij}} \right)$$

where ρ_{ij} is the probability that a column-j individual is in row i (so $\rho_{+j} = \sum_{i=1}^{C} \rho_{ij} = 1$).

Of course, the problem is without knowledge of how the data was collected it is not possible to know which model to use. However, we now show that all the models are

closely related, and with a suitable choice of link functions, diffferent models can lead to the same conclusions. We will only show the equivalence between cases (a) and (b), a similar argument can be extended to case (c).

We start by show that if π_{ij} and λ_{ij} are related in a certain way, then the log-likelihoods of both the poisson and the multinomial are effectively the same. Define the following log-likelihoods for the Poisson, Multinomial and the sum of independent Poissons as follows

$$\begin{aligned}
\mathcal{L}_P(\lambda) &= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\lambda_{ij} - \lambda_{ij} - \log y_{ij}!\Big) \\
\mathcal{L}_M(\pi) &= \log\frac{N!}{\prod_{i=1}^{C}\prod_{j=1}^{R} y_{ij}!} + \sum_{i=1}^{C}\sum_{j=1}^{R} y_{ij}\log\pi_{ij} \\
\mathcal{L}_F(\lambda_{++}) &= N\log\lambda_{++} - \lambda_{++} - \log N!.
\end{aligned}$$

We observe that $\mathcal{L}_P$ is the log distribution of $\{y_{i,j}\}$ under Poisson sampling, $\mathcal{L}_M$ is the log distribution of $\{y_{i,j}\}$ under multinomial sampling, and $\mathcal{L}_F$ is the distribution of $\sum_{ij} Y_{ij}$, where Y_{ij} are independent Poisson distributions each with mean λ_{ij}, $N = \sum_{ij} Y_{ij}$ and $\lambda_{++} = \sum_{ij}\lambda_{ij}$.

Theorem 10.2.1 *Let $\mathcal{L}_P, \mathcal{L}_M$ and $\mathcal{L}_F$ be defined as above. If λ and π are related through*

$$\pi_{ij} = \frac{\lambda_{ij}}{\sum_{s,t}\lambda_{st}} \quad \lambda_{ij} = \lambda_{++}\pi_{ij},$$

where λ_{++} is independent of (i,j). Then we have that

$$\mathcal{L}_P(\lambda) = \mathcal{L}_M(\pi) + \mathcal{L}_F(\lambda_{++}).$$

PROOF. The proof is straightforward. Consider the log-likelihood of the Poisson

$$\begin{aligned}
\mathcal{L}_P(\lambda) &= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\lambda_{ij} - \lambda_{ij} - \log y_{ij}!\Big) \\
&= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\lambda_{++}\pi_{ij} - \lambda_{++}\pi_{ij} - \log y_{ij}!\Big) \\
&= \Big[\sum_{i=1}^{C}\sum_{j=1}^{R} y_{ij}\log\pi_{ij} + \log N! - \sum_{i=1}^{C}\sum_{j=1}^{R}\log y_{ij}!\Big] + \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\lambda_{++} - \lambda_{++} - \log N!\Big) \\
&= \mathcal{L}_M(\pi) + \mathcal{L}_F(\lambda_{++}).
\end{aligned}$$

Which leads to the required result. □

Remark 10.2.1 *The above result means that the likelihood of the independent Poisson conditioned on the total number of participants is N, is equal to the likelihood of the multinomial distribution where the relationship between probabilities and means are given above.*

By connecting the probabilities and mean through the relation

$$\pi_{ij} = \frac{\lambda_{ij}}{\sum_{s,t} \lambda_{st}} \quad \text{and} \quad \lambda_{ij} = \lambda_{++} p_{ij},$$

it does not matter whether the multinomial distribution or Possion distribution is used to do the estimation. We consider a few models which are commonly used in categorical data.

Example 10.2.1 *Let us consider suitable models for the number of children and ethnicity data. Let us start by fitting a multinomial distribution using the logistic link. We start y modelling $\beta' x_i$. One possible model is*

$$\beta' x = \eta + \alpha_1 \delta_1 + \alpha_2 \delta_2 + \alpha_3 \delta_3 + \beta_1 \delta_1^* + \beta_2 \delta_2^*,$$

where $\delta_i = 1$ if the female has i children and zero otherwise, $\delta_1^* = 1$ if female belongs to ethnic group A and zero otherwise, $\delta_2^* = 1$ if female belongs to ethnic group B and zero otherwise. The regressors in this example are $x = (1, \delta_1, \ldots, \delta_2^*)$. Hence for a given cell (i, j) we have

$$\beta' x_{ij} = \eta_{ij} = \eta + \alpha_i + \beta_j.$$

One condition that we usually impose when doing the estimation is that $\sum_{i=1}^{3} \alpha_i = 0$ and $\beta_1 + \beta_2 = 0$. These conditions mean the system is identifiable. Without these conditions you can observe that there exists another $\{\tilde{\alpha}_i\}$, $\{\tilde{\beta}_i\}$ and $\tilde{\eta}$, such that $\eta_{ij} = \eta + \alpha_i + \beta_j = \tilde{\eta} + \tilde{\alpha}_i + \tilde{\beta}_j$.

Now let understand what the above linear model means in terms of probabilities. Using the logistic link we have

$$\pi_{ij} = g^{-1}(\beta' x_{ij}) = \frac{\exp(\eta + \alpha_i + \beta_j)}{\sum_{s,t} \exp(\eta + \alpha_s + \beta_t)} = \frac{\exp(\alpha_i)}{\sum_s \exp(\alpha_s)} \times \frac{\exp(\beta_j)}{\sum_t \exp(\beta_t)},$$

where π_{ij} denotes the probability of having i children and belonging to ethnic group j and x_{ij} is a vector with ones in the appropriate places. What we observe is that the above

model is multiplicative, that is

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

where $\pi_{i+} = \sum_j \pi_{ij}$ *and* $\pi_{+j} = \sum_i \pi_{i+}$. *This means by fitting the above model we are assuming independence between ethnicity and number of children. To model dependence we would use an interaction term in the model*

$$\beta' x = \eta + \alpha_1\delta_1 + \alpha_t\delta_2 + \alpha_3\delta_3 + \beta_1\delta_1^* + \beta_2\delta_1^* + \sum_{i,j}\gamma_{ij}\delta_i\delta_j^*,$$

hence

$$\eta_{ij} = \eta + \alpha_i + \beta_j + \gamma_{ij}.$$

However, for $R \times C$ *tables an interaction term means the model is saturated (i.e. the MLE estimator of the probability* π_{ij} *is simply* y_{ij}/N*). But for* $R \times C \times L$*, we can model interactions without the model becoming saturated i.e.*

$$\eta_{ijk} = \eta + \alpha_i + \beta_j + \epsilon_k + \gamma_{ij}$$

models an interaction between R *and* C *but independence from* L*. These interactions may have interesting interpretations about the dependence structure between two variables. By using the log-likelihood ratio test (or analysis of deviance), we can test whether certain interaction terms are significant.*

We transform the above probabilities into Poisson means using $\lambda_{ij} = \gamma\pi_{ij}$*. In the case there is no-interaction the mean of Poisson at cell* (i, j) *is* $\lambda_{ij} = \gamma \exp(\eta + \alpha_i + \beta_j)$.

In the above we have considered various methods for modelling the probabilities in a mulitnomial and Poisson distributions. In the theorem we show that so long as the probabilities and Poisson means are linked in a specific way, the estimators of β will be identical.

Theorem 10.2.2 (Equivalence of estimators) *Let us suppose that* π_{ij} *and* μ_{ij} *are defined by*

$$\pi_{ij} = \pi_{ij}(\beta) \quad \lambda_{ij} = \gamma\pi_{ij}(\beta),$$

where γ *and* $\beta = \{\alpha_i, \beta_j\}$ *are unknown and* $C(\beta)$ *is a known function of* β *(such as* $\sum_{i,j} \exp(\alpha_i + \beta_j)$ *or* 1*). Let*

$$\begin{aligned}
\mathcal{L}_P(\beta, \gamma) &= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\gamma\pi_{ij}(\beta) - \gamma\pi_{ij}(\beta)\Big) \\
\mathcal{L}_M(\beta) &= \sum_{i=1}^{C}\sum_{j=1}^{R} y_{ij}\log\pi_{ij}(\beta) \\
\mathcal{L}_F(\beta, \gamma) &= N\log\gamma - \gamma,
\end{aligned}$$

which is the log-likelihoods for the Multinomial and Poisson distributions without unnecessary constants (such as $y_{ij}!$*). Define*

$$\begin{aligned}
(\hat{\beta}_P, \hat{\gamma}_P) &= \arg\max \mathcal{L}_P(\beta, \gamma) \\
\hat{\beta}_B &= \arg\max \mathcal{L}_M(\beta) \quad \hat{\gamma}_F = \arg\max \mathcal{L}_F(\beta, \gamma).
\end{aligned}$$

Then $\widehat{\beta}_P = \widehat{\beta}_M$ *and* $\widehat{\gamma}_P = \hat{\gamma}_M = N/C(\widehat{\beta}_M)$*.*

PROOF. We first consider $\mathcal{L}_P(\beta, \gamma)$. Since $\sum_{i,j} p_{i,j}(\beta) = 1$ and $\sum_{i,j} y_{i,j} = 1$ we have

$$\begin{aligned}
\mathcal{L}_P(\beta, \gamma) &= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\gamma C(\beta)\pi_{ij}(\beta) + \gamma C(\beta)\pi_{ij}(\beta)\Big) \\
&= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\pi_{ij}(\beta)\Big) + N\log\gamma C(\beta) - C(\beta)\gamma.
\end{aligned}$$

Now we consider the partial derivatives of $\mathcal{L}_P$ to obtain

$$\begin{aligned}
\frac{\partial\mathcal{L}_P}{\partial\beta} &= \frac{\partial\mathcal{L}_M}{\partial\beta} + \gamma\frac{\partial C(\beta)}{\partial\beta}\left(\frac{N}{\gamma C(\beta)} - 1\right) = 0 \\
\frac{\partial\mathcal{L}_P}{\partial\gamma} &= \left(\frac{N}{\gamma} - C(\beta)\right) = 0.
\end{aligned}$$

Solving the above we have that $\hat{\beta}_P$ and $\hat{\gamma}_P$ satisfy

$$\hat{\gamma}_P = \frac{N}{\widehat{C}(\beta)} \qquad \frac{\partial\mathcal{L}_M}{\partial\beta}\rfloor_{\beta=\hat{\beta}_P} = 0. \tag{10.1}$$

Now we consider the partial derivatives of $\mathcal{L}_M$ and $\mathcal{L}_C$

$$\frac{\partial\mathcal{L}_M}{\partial\beta} = 0 \qquad \frac{\partial\mathcal{L}_F}{\partial\gamma} = \left(\frac{N}{\gamma} - C(\beta)\right) = 0. \tag{10.2}$$

Comparing the estimators in (10.1) and (10.2) it is clear that the maximum likelihood estimators of β based on the Poisson and the Binomial distributions are the same. □

Example 10.2.2 *Let us consider fitting the Poisson and the multinomial distributions to the data in a contingency table where π_{ij} and λ_{ij} satisfy*

$$\lambda_{ij} = \exp(\eta + \beta' x_{ij}) \text{ and } \pi_{ij} = \frac{\exp(\beta' x_{ij})}{\sum_{s,t} \exp(\beta' x_{s,t})}.$$

Making a comparison with $\lambda_{ij}(\beta) = \gamma C(\beta)\pi_{ij}(\beta)$ we see that $\gamma = \exp(\eta)$ and $C(\beta) = \sum_{s,t} \exp(\beta' x_{s,t})$. Then it by using the above theorem the estimator of β is the parameter which maximises

$$\sum_{i=1}^{C}\sum_{j=1}^{R}\left(y_{ij} \log \frac{\exp(\beta' x_{ij})}{\sum_{s,t} \exp(\beta' x_{s,t})}\right),$$

and the estimator of γ is the parameter which maximises

$$N \log \exp(\eta) C(\hat{\beta}) - \exp(\eta) C(\hat{\beta}),$$

which is $\eta = \log N - \log(\sum_{s,t} \exp(\hat{\beta}' x_{s,t}))$.

10.3 Overdispersion

The binomial and Poisson distributions have the disadvantage that they are determined by only one parameter (π in the case of Binomial and λ in the case of Possion). This can be a disadvantage when it comes to modelling certain types of behaviour in the data. A type of common behaviour in count data is overdispersed, in the sense that the variance appears to be larger than the model variance.

Checking for overdispersion

- First fit a Poisson model to the data.
- Extract the Pearson residuals from the data (see Section 9.3.5), for the Poisson it is

  $$r_i = \frac{(Y_i - \hat{\mu}_i)}{\phi^{1/2} V(\mu_i)^{1/2}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\mu_i}}.$$

  If the model is correct, the residuals $\{r_i\}$ should be 'close' to a standard normal distribution. However, in the case of overdispersion it is likely that the estimated variance of r_i will be greater than one.
- Plot r_i against μ_i.

10.3.1 Modelling overdispersion

Modelling overdispersion can be done in various ways. Below we focus on Poisson-type models.

Zero inflated models

The number of zeros in count data can sometimes be more (inflated) than Poisson or binomial distributions are capable of modelling (for example, if we model the number of times a child visits the dentist, we may observe that there is large probability the child will not visit the dentist). To model this type of behaviour we can use the inflated zero Poisson model, where

$$P(Y=k) = \begin{cases} (1-p)(1-\exp(-\lambda)) = 1-p+p\exp(-\lambda) & k=0 \\ p\frac{\exp(-\lambda)\lambda^k}{k!} & k>0 \end{cases}.$$

We observe that the above is effectively a mixture model. It is straightforward to show that $\mathrm{E}(Y) = p\lambda$ and $\mathrm{var}(Y) = p\lambda(1+\lambda(1-p))$, hence

$$\frac{\mathrm{var}(Y)}{\mathrm{E}(Y)} = (1+\lambda(1-p)).$$

We observe that there is more dispersion here than classical Poisson where $\mathrm{var}(Y)/\mathrm{E}(Y) = 1$.

Modelling overdispersion through moments

One can introduce overdispersion by simply modelling the moments. That is define a psuedo Poisson model in terms of its moments, where $\mathrm{E}(Y) = \lambda$ and $\mathrm{var}(Y) = \lambda(1+\delta)$ ($\delta \geq 0$). This method does not specify the distribution, it simply places conditions on the moments.

Modelling overdispersion with another distribution (latent variable)

Another method for introducing overdispersion into a model is to include a 'latent' (unobserved) parameter ε. Let us assume that ε is a positive random variable where $\mathrm{E}(\varepsilon) = 1$ and $\mathrm{var}(\varepsilon) = \xi$. We suppose that the distribution of Y conditioned on ε is Poisson, i.e. $P(Y=k|\varepsilon) = \frac{(\lambda\varepsilon)^k \exp(-\lambda\varepsilon)}{k!}$. The introduction of latent variables allows one to generalize

several models in various directions. It is a powerful tool in modelling. For example, if one wanted to introduce dependence between theY_is one can do this by conditioning on a latent variable which is dependent (eg. the latent variable can be a time series).

To obtain the moments of Y we note that for any random variable Y we have

$$\begin{aligned}\mathrm{var}(Y) &= \mathrm{E}(Y^2) - \mathrm{E}(Y)^2 = \mathrm{E}\Big(\mathrm{E}(Y^2|\varepsilon) - \mathrm{E}(Y|\varepsilon)^2\Big) + \mathrm{E}\big(\mathrm{E}(Y|\varepsilon)^2\big) - \mathrm{E}\big(\mathrm{E}(Y|\varepsilon)\big)^2 \\ &= \mathrm{E}\Big(\mathrm{var}(Y|\varepsilon)\Big) + \mathrm{var}(\mathrm{E}(Y|\varepsilon)),\end{aligned}$$

where we note that $\mathrm{var}(Y|\varepsilon) = \sum_{k=0}^{\infty} k^2 P(Y=k|\varepsilon) - (\sum_{k=0}^{\infty} kP(Y=k|\varepsilon))^2$ and $\mathrm{E}(Y|\varepsilon) = \sum_{k=0}^{\infty} kP(Y=k|\varepsilon)$. Applying the above to the conditional Poisson we have

$$\begin{aligned}\mathrm{var}(Y) &= \mathrm{E}(2(\lambda\varepsilon) - (\lambda\varepsilon)) + \mathrm{var}(\lambda\varepsilon) \\ &= \lambda + \lambda^2\xi = \lambda(1+\lambda\xi) \\ \text{and } \mathrm{E}(Y) &= \mathrm{E}(\mathrm{E}(Y|\varepsilon)) = \lambda.\end{aligned}$$

The above gives an expression in terms of moments. If we want to derive the distribution of Y, we require the distribution of ε. This is normally hard in practice to verify, but for reasons of simple interpretation we often let ε have a Gamma distribution $f(\varepsilon;\nu,\kappa) = \frac{\nu^\kappa \varepsilon^{k-1}}{\Gamma(\kappa)}\exp(-\nu\varepsilon)$, where $\nu=\kappa$, hence $\mathrm{E}(\varepsilon)=1$ and $\mathrm{var}(\varepsilon) = 1/\nu(=\xi)$. Therefore in the case that ε is a Gamma distribution with density $f(\varepsilon;\nu,\nu) = \frac{\nu^\nu \varepsilon^{\nu-1}}{\Gamma(\nu)}\exp(-\nu\varepsilon)$ the distribution of Y is

$$\begin{aligned}P(Y=k) &= \int P(Y=k|\varepsilon) f(\varepsilon;\nu,\nu)d\varepsilon \\ &= \int \frac{(\lambda\varepsilon)^k \exp(-\lambda\varepsilon)}{k!}\frac{\nu^\nu\varepsilon^{k-1}}{\Gamma(\nu)}\exp(-\nu\varepsilon)d\varepsilon \\ &= \frac{\Gamma(k+\nu)}{\Gamma(\nu)k!}\frac{\nu^\nu\lambda^k}{(\nu+\lambda)^{\nu+\lambda}}.\end{aligned}$$

This is called a negative Binomial (because in the case that ν is an integer it resembles a regular Binomial but can take infinite different outcomes). The negative binomial only belongs to the exponential family if ν is known (and does not need to be estimated). Not all distributions on ε lead to explicit distributions of Y. The Gamma is popular because it leads to an explicit distribution for Y (often it is called the conjugate distribution).

A similar model can also be defined to model overdispersion in proportion data, using a random variable whose conditional distribution is Binomial (see page 512, Davison (2002)).

Remark 10.3.1 (Using latent variables to model dependence) *Suppose Y_j conditioned on $\{\varepsilon_j\}$ follows a Poisson distribution where $P(Y_j = k|\varepsilon_j) = \frac{(\lambda\varepsilon_j)^k \exp(-\lambda\varepsilon_j)}{k!}$ and $Y_i|\varepsilon_i$ and $Y_j|\varepsilon_j$ are conditionally independent. We assume that $\{\varepsilon_j\}$ are positive continuous random variables with correlation $\mathrm{cov}[\varepsilon_i, \varepsilon_j] = \rho_{i,j}$. The correlations in ε_j induce a correlation between Y_j through the relation*

$$\begin{aligned}\mathrm{cov}\left[Y_i, Y_j\right] &= \mathrm{E}\left(\underbrace{\mathrm{cov}\left[Y_i, Y_j|\varepsilon_i, \varepsilon_j\right]}_{=0(a.s.)}\right) + \mathrm{cov}\left(\underbrace{\mathrm{E}[Y_i|\varepsilon_i]}_{=\lambda\varepsilon_i}, \underbrace{\mathrm{E}[Y_j|\varepsilon_j]}_{=\lambda\varepsilon_j}\right) \\ &= \lambda^2 \mathrm{cov}\left(\varepsilon_i, \varepsilon_j\right) = \lambda^2\rho_{ij}.\end{aligned}$$

10.3.2 Parameter estimation using estimating equations

We now consider various methods for estimating the parameters. Some of the methods described below will be based on the Estimating functions and derivations from Section 9.3.1, equation (9.10).

Let us suppose that $\{Y_i\}$ are overdispersed random variables with regressors $\{x_i\}$ and $\mathrm{E}(Y_i) = \mu_i$ with $g(\mu_i) = \beta' x_i$. The natural way to estimate the parameters β is to use a likelihood method. However, the moment based modelling of the overdispersion does not have a model attached (so it is not possible to use a likelihood method), and the modelling of the overdispersion using, say, a Gamma distribution, is based on a assumption that is hard in practice to verify (that the latent variable is Gaussian). An alternative approach is to use moment based/estimating function methods which are more robust to misspecification than likelihood methods. In the estimation we discuss below we will focus on the Poisson case, though it can easily be generalised to the non-Poisson case.

Let us return to equation (9.10)

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i) g'(\mu_i)} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)} \frac{d\mu_i}{d\eta_i} = 0 \qquad 1 \leq j \leq p. \tag{10.3}$$

In the case of the Poisson distribution, with the log link the above is

$$\sum_{i=1}^{n} (Y_i - \exp(\beta' x_i))x_{ij} = 0 \qquad 1 \leq j \leq p. \tag{10.4}$$

We recall if $\{Y_i\}$ are Possion random variables with mean $\exp(\beta' x_i)$, then variance of the limiting distribution of β is

$$(\widehat{\beta} - \beta) \approx N_p(0, (X^T W X)^{-1}),$$

since the Fisher information matrix can be written as

$$(I(\beta))_{jk} = E\left(-\frac{\partial^2 \mathcal{L}_n(\beta)}{\partial\beta_j\partial\beta_k}\right) = \mathrm{E}\left(-\sum_{i=1}^{n}\frac{d^2\ell_i}{d\eta_i^2}x_{ij}x_{ik}\right) = (X^TWX)_{jk}.$$

where

$$\begin{aligned} W &= \mathrm{diag}\left(\mathrm{E}(-\frac{\partial^2\ell_1(\eta_1)}{\partial\eta_1^2}),\ldots,\mathrm{E}(-\frac{\partial^2\ell_n(\eta_n)}{\partial\eta_n^2})\right) \\ &= \mathrm{diag}\left(\exp(\beta' x_1),\ldots,\exp(\beta' x_n)\right). \end{aligned}$$

However, as we mentioned in Section 9.3.1, equations (10.3) and (10.4) do not have to be treated as derivatives of a likelihood. Equations (10.3) and (10.4) can be viewed as estimating equation, since they only use the first and second order moments of $\{Y_i\}$. Hence they can be used as the basis of the estimation scheme even if they are not as efficient as the likelihood. In the overdispersion literature the estimating equations (functions) are often called the Quasi-likelihood.

Example 10.3.1 *Let us suppose that $\{Y_i\}$ are independent random variables with mean $\exp(\beta' x_i)$. We use the solution of the estimating function*

$$\sum_{i=1}^{n} g(Y_i;\beta) = \sum_{i=1}^{n}(Y_i - \exp(\beta' x_i))x_{ij} \;=\; 0 \qquad 1 \le j \le p.$$

to estimate β. Using Theorem 8.2.2 we derive the asymptotic variance for two models:

(i) $\mathrm{E}(Y_i) = \exp(\beta' x_i)$ **and** $\mathrm{var}(Y_i) = (1+\delta)\exp(\beta' x_i)$ **($\delta \ge 0$).**

Let us suppose that $\mathrm{E}(Y_i) = \exp(\beta' x_i)$ *and* $\mathrm{var}(Y_i) = (1+\delta)\exp(\beta' x_i)$ *($\delta \ge 0$). Then if the regularity conditions are satisfied we can use Theorem 8.2.2 to obtain the limiting variance. Since*

$$\begin{aligned} \mathrm{E}\left(\frac{-\partial\sum_{i=1}^{n} g(Y_i;\beta)}{\partial\beta}\right) &= X^T\mathrm{diag}\left(\exp(\beta' x_1),\ldots,\exp(\beta' x_n)\right)X \\ \mathrm{var}\left(\sum_{i=1}^{n} g(Y_i;\beta)\right) &= (1+\delta)X^T\mathrm{diag}\big(\exp(\beta' x_1),\ldots,\exp(\beta' x_n)\big)X, \end{aligned}$$

the limiting variance is

$$(1+\delta)(X^TWX)^{-1} = (1+\delta)(X^T\mathrm{diag}\left(\exp(\beta' x_1),\ldots,\exp(\beta' x_n)\right)X)^{-1}.$$

Therefore, in the case that the variance is $(1+\delta)\exp(\beta' x_i)$*, the variance of the estimator using the estimating equations* $\sum_{i=1}^n g(Y_i;\beta)$*, is larger than for the regular Poisson model. If* δ *is quite small, the difference is also small. To estimate* δ *we can use*

$$\sum_{i=1}^{n} \frac{(Y_i - \exp(\widehat{\beta}' x_i))^2}{\exp(\widehat{\beta}' x_i)}.$$

(ii) $\mathrm{E}(Y_i) = \exp(\beta' x_i)$ **and** $\mathrm{var}(Y_i) = \exp(\beta' x_i)(1+\xi\exp(\beta' x_i))$.

In this case we have

$$\mathrm{E}\left(\frac{-\partial \sum_{i=1}^n g(Y_i;\beta)}{\partial \beta}\right) = X^T W X \quad \textit{and} \quad \mathrm{var}(\sum_{i=1}^{n} g(Y_i;\beta)) = X^T \tilde{W} X,$$

where

$$\begin{aligned} W &= \mathrm{diag}\Big(\exp(\beta' x_1), \ldots, \exp(\beta' x_n)\Big) \\ \tilde{W} &= \mathrm{diag}\Big(\exp(\beta' x_1)(1+\xi\exp(\beta' x_1)), \ldots, \exp(\beta' x_n)(1+\xi\exp(\beta' x_n))\Big). \end{aligned}$$

Hence the limiting variance is

$$(X^T W X)^{-1}(X^T \tilde{W} X)(X^T W X)^{-1}.$$

We mention that the estimating equation can be adapted to take into count the overdispersion in this case. In other words we can use as an estimator of β*, the* β *which solves*

$$\sum_{i=1}^{n} \frac{(Y_i - \exp(\beta' x_i))}{(1+\xi\exp(\beta' x_i))} x_{ij} = 0 \qquad 1 \leq j \leq p.$$

Though we mention that we probably have to also estimate ξ *when estimating* β*.*

10.4 A worked problem

(1) (a) Suppose that U is a Poisson distributed random variable with mean λ. Then for $k \geq 0$,

$$P(U=k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{10.5}$$

(i) Let us suppose $U_1, \ldots, U_n$ are independent, identically distributed random variables from a Poisson distribution. What is the maximum likelihood estimator of λ?

(ii) For several count data sets it has been observed that there is an excessive number of zeros. To model 'inflated-zero' count data, the zero-inflated Poisson distribution model was proposed, where the observations are modelled as

$$Y = \delta U,$$

where δ and U are independent random variables, δ takes on value either zero or one with $P(\delta = 0) = p$, $P(\delta = 1) = (1 - p)$, and U has a Poisson distribution as defined as in (10.5).

Briefly explain why this model can account for an excessive number of zeros.

(iii) Show that the estimator defined in (i) is a *biased* estimator of λ when the observations come from a zero-inflated Poisson distribution.

(b) In this part of the question we consider the zero-inflated Poisson regression model, proposed in Lambert (1992), which is defined as

$$Y_j = \delta_j U_j,$$

where δ_j and U_j are independent random variables, $P(\delta_j = 0) = p$, $P(\delta_j = 1) = (1 - p)$, and U_j has a Poisson distribution with mean $\lambda_j = e^{\beta x_j}$ and x_j is a *fixed* covariate value. Our objective is to first construct crude estimators for p and β and to use these estimates as the initial values in an iterative scheme to obtain the maximum likelihood estimator.

(i) *Estimation of* β. What is the distribution of Y_j conditioned on $Y_j > 0$ and x_j?

Argue that, for each $k = 1, 2, \ldots$,

$$P(Y_j = k | Y_j > 0) = \frac{e^{-\lambda_j} \lambda_j^k / k!}{(1 - e^{-\lambda_j})}. \tag{10.6}$$

Let $\boldsymbol{Y}^+$ be the vector of all the non-zero Y_js. Use result (10.6) to define a *conditional* log-likelihood for $\boldsymbol{Y}^+$ given that all the Y_js in $\boldsymbol{Y}^+$ are positive.

Determine the derivative of this conditional log-likelihood, and explain how it can be used to determine an estimate of β. Denote this estimator as $\widehat{\beta}$.

(ii) *Estimation of p.* Define the dummy variable

$$Z_j = \begin{cases} 0 & \text{if } Y_j = 0 \\ 1 & \text{if } Y_j > 0. \end{cases}$$

Use $Z_1, \ldots, Z_n$ to obtain an explicit estimator of p in terms of $Y_1, \ldots, Y_n$, $x_1, \ldots, x_n$ and $\widehat{\beta}$.

Hint: One possibility is to use estimating equations.

(iii) We may regard each δ_j as a missing observation or latent variable. What is the full log-likelihood of (Y_j, δ_j), $j = 1, \ldots, n$, given the regressors $x_1, \ldots, x_n$?

(iv) Evaluate the conditional expectations $E[\delta_j | Y_j = k]$, $k = 0, 1, 2, \ldots$.

(v) Use your answers in part (iii) and (iv) to show how the EM-algorithm can be used to estimate β and p. (You need to state the criterion that needs to be maximised and the steps of the algorithm).

(vi) Explain why for the EM-algorithm it is important to use good initial values.

Reference: Zero-inflated Poisson Regression, with an application to defects in manufactoring. **Diane Lambert**, *Technometrics*, vol 34, 1992.

Solution

(1) (a) Suppose that U is a Poisson distributed random variable, then for $k \geq 0$,

$$P(U = k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{10.7}$$

(i) Let us suppose $\{U_j\}$ independent, identically distributed random variables from a Poisson distribution. What is the maximum likelihood estimator of λ?

It is clear that it is the sample mean, $\widehat{\lambda} = \frac{1}{n}\sum_{j=1}^{n} U_j$

(ii) For several count data sets it has been observed that there is an excessive number of zeros. To model 'inflated-zero' count data, the zero-inflated

Poisson distribution model was proposed where the observations are modelled as

$$Y = \delta U,$$

δ and U are random variables which are independent of each other, where δ is random variable taking either zero or one, $P(\delta = 0) = p$, $P(\delta = 1) = (1 - p)$ and U is a Poisson random variable as defined as in (10.7) Briefly explain why this model can is able to model excessive number of zeros.

The probability of zero is $P(Y = 0) = p + (1 - p)e^{-\lambda}$. Thus if p is sufficiently large, the chance of zeros is larger than the usual Poisson distribution (for a given λ).

(ii) Show that the estimator defined in (i) is a *biased* estimator of λ when the observations come from an zero-inflated Poisson distribution.

$\mathrm{E}[\widehat{\lambda}] = (1 - p)\lambda$**, thus when** $p > 0$**,** $\widehat{\lambda}$ **underestimates** λ**.**

(b) In this part of the question we consider the zero-inflated poisson regression model, proposed in Lambert (1992), which is defined as

$$Y_j = \delta_j U_j$$

where δ_j and U_j are random variables which are independent of each other, δ_j is an indicator variable, where $P(\delta_j = 0) = p$ and $P(\delta_j = 1) = (1 - p)$ and U_j has a Poisson regression distribution with

$$P(U_j = k|x_j) = \frac{\lambda_j^k e^{-\lambda_j}}{k!}$$

where $\lambda_j = e^{\beta x_j}$ and x_j is an observed regressor. Our objective is to first construct initial-value estimators for p and β and then use this to estimate as the initial values in when obtaining the maximum likelihood estimator.

(i) *Estimation of* β First obtain the distribution of Y_j conditioned on $Y_j > 0$ and x_j.

We note that $P(Y_j > 0) = P(\delta_j = 1, U_j > 0) = P(\delta_j = 1)P(U_j > 0) = (1 - p)(1 - e^{\lambda_j})$**. Similarly** $P(Y_j = k, Y_j > 0) = P(U_j = k, \delta_j = 1) = (1 - p)P(U_j = k)$**. Thus**

$$P(Y_j = k|Y_j > 0) = \frac{\lambda_j^k \exp(-\lambda_j)}{(1 - e^{\lambda_j})k!}$$

Let $\underline{Y}^+ = \{Y_j > 0\}$ (all the non-zero Y_j). Obtain the conditional log-likelihood of $\underline{Y}_+$ *conditioned* on $Y_j > 0$ and $\underline{x} = (x_1, \ldots, x_n)$. Derive the score equation and explain how β can be estimated from here. Denote this estimator as $\widehat{\beta}$.

The log-conditional likelihood is proportional to

$$\begin{aligned}\mathcal{L}_C(\beta) &= \sum_{Y_j>0} \left[Y_j \log \lambda_j - \lambda_j - \log(1 - e^{\lambda_j})\right] \\ &= \sum_{Y_j>0} \left\{\beta Y_j x_j - e^{\beta x_j} - \log(1 - e^{\beta x_j})\right\}.\end{aligned}$$

Thus to estimate β we differentiate the above wrt β (giving the score) and numerically solve the folling equation wrt β

$$\sum_{Y_j>0} Y_j x_j = \sum_{Y_j>0} x_j e^{\beta x_j} \left\{1 - \frac{1}{1 - e^{\beta} x_j}\right\}.$$

(ii) **Estimation of p** Define the dummy variable

$$Z_j = \begin{cases} 0 & \text{if } Y_j = 0 \\ 1 & \text{if } Y_j > 0. \end{cases}$$

Use $\{Z_j\}$ to obtain an explicit estimator of p in terms of $\underline{Y}$, $\underline{x}$ and $\widehat{\beta}$.
Hint: One possibility is to use estimating equations.

We solve the estimating equation

$$\sum_{j=1}^{n} [Z_j - \mathrm{E}(Z_j)] = 0,$$

wrt p. It is clear that $\mathrm{E}(Z_j) = P(Z_j = 1) = (1 - P(Z_j = 0)) = (1 - p)(1 - e^{-\lambda_j})$**. Thus the estimating equation is**

$$\sum_{j=1}^{n} \left[Z_j - (1 - p)(1 - e^{-\lambda_j})\right] = 0.$$

Replacing λ_j with $\hat{\lambda}_j = e^{\hat{\beta} x_j}$ and solving for p yields the estimator

$$\hat{p} = 1 - \frac{\sum_{j=1}^{n} Z_j}{\sum_{j=1}^{n} [1 - \exp(-e^{\hat{\beta} x_j})]}.$$

(iii) What is the complete log-likelihood of $\{Y_j, \delta_j; j = 1, \ldots, n\}$ (acting as if the variable δ_j is observed) given the regressors $\{x_j\}$?

The distribution of (Y_j, δ_j) is

$$P(Y_j = k, \delta_j) = [P(U_j = k)P(\delta_j = 1)]^{\delta_j} [P(\delta_j = 0)]^{1-\delta_j}.$$

Thus the log-likelihood of $\{Y_j, \delta_j; j = 1, \ldots, n\}$ is

$$\mathcal{L}_F(p, \beta) = \sum_{j=1}^{n} \delta_j [Y_j \log \lambda_j - \lambda_j + \log(1-p)] + \sum_{j=1}^{n} (1-\delta_j) \log p.$$

(iv) Evaluate the conditional expectations $\mathrm{E}[\delta_j | Y_j > 0]$ and $\mathrm{E}[\delta_j | Y_j = 0]$.

$\mathrm{E}[\delta_j | Y_j > 0] = 1$ **(since if $Y_j > 0$ then the only choice is $\delta_j = 1$),**

$$\mathrm{E}[\delta_j | Y_j = 0] = P(\delta_j = 1 | Y_j = 0) = \frac{(1-p)e^{-\lambda_j}}{p + (1-p)e^{-\lambda_j}}$$

and

$$\mathrm{E}[1 - \delta_j | Y_j = 0] = P(\delta_j = 0 | Y_j = 0) = \frac{p}{p + (1-p)e^{-\lambda_j}}$$

(v) Use your answers in part (iii) and (iv) to show how the EM-algorithm can be used to estimate β and p (you need to state the criterion that needs to be maximise and the steps of the algorithm).

Splitting the sum $\sum_{j=1}^{n}$ into $\sum_{Y_j>0}$ and $\sum_{Y_j=0}$, and taking expectations of $\mathcal{L}_F$ with respect to $\underline{Y}$ gives

$$\begin{aligned} Q(\theta; \theta^*) &= \sum_{Y_j>0} [Y_j \log \lambda_j - \lambda_j + \log(1-p)] \\ &\quad + \sum_{Y_j=0} \left(\frac{(1-p^*)e^{-\lambda_j^*}}{p^* + (1-p^*)e^{-\lambda_j^*}} \right) [\lambda_j + \log(1-p)] \\ &\quad + \sum_{Y_j=0} \left(\frac{p^*}{p^* + (1-p^*)e^{-\lambda_j^*}} \right) \log p \\ &= Q_1(\beta; \theta^*) + Q_2(p; \theta^*), \end{aligned}$$

where $\lambda_j^* = \exp(\beta^* x_j)$, $\theta = (p, \beta)$, $\theta^* = (p^*, \beta^*)$,

$$\begin{aligned} Q_1(\beta; \theta^*) &= \sum_{Y_j>0} [Y_j \log \lambda_j - \lambda_j] + \sum_{Y_j=0} (1-\pi^*)\lambda_j \\ Q_2(p; \theta^*) &= \sum_{Y_j>0} \log(1-p) + \sum_{Y_j=0} \left[(1-\pi_j^*) \log(1-p) + \pi_j^* \log p \right] \end{aligned}$$

and

$$\pi_j^* = \frac{p^*}{p^* + (1-p^*)e^{-\lambda_j^*}}.$$

Using $Q(\theta;\theta^*)$ we can then implement the EM-algorithm:

1. **Let $p^* = \hat{p}$ and $\beta^* = \hat{\beta}$. Then evaluate $Q_1(\beta;\theta^*)$ and $Q_2(p;\theta^*)$.**
2. **Differentiate $Q_1(\beta;\theta^*)$ wrt β and $Q_2(p;\theta^*)$ wrt p (keeping θ^* fixed) and solve for p and θ (needs to be done numerically). Set the solution $\theta^* = \hat{\theta}$.**
3. **Evaluate $Q_1(\beta;\theta^*)$ and $Q_2(p;\theta^*)$ with respect to the new θ^* and go back to (2).**
4. **Keep iterating until convergence.**

(vi) Explain why in the EM-algorithm it is important to use good initial values.
The EM algorithm is an iterative scheme which successively maximises the likelihood. However, if it climbs to a local maximum it will stay at that point. By using initial values, which are consistent, thus relatively close to the global maximum we can be reasonably sure that the EM-algorithm converged to a global maximum (rather than a local one).

Chapter 11

Survival Analysis with explanatory variables

11.1 Survival analysis and explanatory variables

In this section we build on the introduction to survival analysis given in Section 12. Here we consider the case that some explanatory variables (such as gender, age etc.) may have an influence on survival times. See also Section 10.8, Davison (2002).

We recall that in Section 12, the survivial times $\{T_i\}$ were iid random variables, which may or may not be observed. We observe $Y_i = \min(c_i, T_i)$ and the indicator variable δ_i which tells us whether the individual is censored or not, that is $\delta_i = 1$ if $Y_i = T_i$ (ie. the ith individual was not censored) and other zero otherwise. In this case, we showed that the likelihood (with the censoring times $\{c_i\}$ treated as deterministic) is given in (6.4) as

$$\begin{aligned} \mathcal{L}_n(\theta) &= \sum_{i=1}^{n} \Big(\delta_i \log f(Y_i;\theta) + (1-\delta_i) \log \big(1 - F(Y_i;\theta)\big) \Big) \\ &= \sum_{i=1}^{n} \delta_i \log h(T_i;\theta) - \sum_{i=1}^{n} H(Y_i;\theta), \end{aligned}$$

where $f(t;\theta)$, $\mathcal{F}(t;\theta) = P(T_i \geq t)$, $h(t;\theta) = f(t;\theta)/\mathcal{F}(t;\theta)$ and $H(t;\theta) = \int_0^t h(y;\theta)dy = -\log \mathcal{F}(t;\theta)$ denote the density, survival function, hazard function and cumulative hazard function respectively.

We now consider the case that the survival times $\{T_i\}$ are not identically distributed but determined by some regressors $\{x_i\}$. Furthermore, the survival times could be cen-

sored, hence we observe $\{(Y_i, \delta_i, x_i)\}$, where $Y_i = \min(T_i, c_i)$. Let us suppose that T_i has the distribution specified by the hazard function $h(t; x_i, \beta)$ (hence the hazard depends on both parameters and explanatory variables x_i, and we want to analyse the dependency on x_i. It is straightforward to see that the log-likelihood of $\{(Y_i, \delta_i, x_i)\}$ is

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \delta_i \log h(T_i; x_i, \beta) - \sum_{i=1}^{n} H(Y_i; x_i, \beta).$$

There are two main approaches for modelling the hazard function h:

- Proportional hazards (PH). The effect of x_i is to scale up or down the hazard function
- Accelerated life (AL). The effect of x_i is to speed up or slow down time.

We recall that from the hazard function, we can obtain the density of T_i, though for survival data, the hazard function is usually more descriptive.

In the sections below we define the proportional hazard and accelerated life hazard function and consider methods for estimating β.

11.1.1 The proportional hazards model

Proportional hazard functions are used widely in medical applications.

Suppose the effect of x is summarised by a one-dimensional non-negative *hazard ratio* function $\psi(x; \beta)$ (sometimes called the risk score). That is

$$h(t; x, \beta) = \psi(x; \beta) h_0(t),$$

where $h_0(t)$ is a fully-specified *baseline* hazard. We choose the scale of measurement for x so that $\psi(0) = 1$, i.e. $h_0(t) = h(t; 0, \beta)$. It follows that

$$\begin{aligned} H(t; x, \beta) &= \psi(x; \beta) H_0(t) \\ \mathcal{F}(t; x, \beta) &= \mathcal{F}_0(t)^{\psi(x;\beta)} \\ f(t; x, \beta) &= \psi(x)(\mathcal{F}_0(t))^{\psi(x;\beta)-1} f_0(t). \end{aligned}$$

Recall that in question 5.4.5 (HW5), we showed that if $\mathcal{F}(x)$ was a survival function, then $\mathcal{F}(x)^\gamma$ also defines a survival function, hence it corresponded to a well defined density. The

same is true of the proportional hazards function. By defining $h(t; x, \beta) = \psi(x; \beta)h_0(t)$, where h_0 is a hazard function, we have that $h(t; x, \beta)$ is also a viable hazard function.

A common choice is $\psi(x; \beta) = \exp(\beta' x)$, with β to be estimated. This is called the exponential hazard ratio.

MLE for the PH model with exponential hazard ratio

The likelihood equations corresponding to $h(t; x, \beta)$ and $H(t; x, \beta)$ are

$$\begin{aligned}\mathcal{L}_n(\beta) &= \sum_{i=1}^{n} \delta_i \log \exp(\beta' x_i) h_0(Y_i) - \sum_{i=1}^{n} \exp(\beta' x_i) H_0(Y_i) \\ &= \sum_{i=1}^{n} \delta_i \big[\beta' x_i) + \log h_0(Y_i)\big] - \sum_{i=1}^{n} \big[\exp(\beta' x_i)) H_0(Y_i)\big],\end{aligned}$$

where the baseline hazard h_0 and H_0 is assumed known. The derivative of the above likelihood is

$$\frac{\partial \mathcal{L}_n(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i x_{ij} - \sum_{i=1}^{n} x_{ij} e^{\beta' x_i} H_0(Y_i) = 0 \qquad 1 \le j \le p.$$

In general ther is no explicit solution for $\hat{\beta}$, but there is in some special cases. For example, suppose the observations fall into k disjoint groups with $x_{ij} = 1$ if i is in group j, 0 otherwise. Let m_j be the number of uncensored observations in group j, that is $m_j = \sum_i \delta_i x_{ij}$. Then the likelihood equations become

$$\frac{\partial \mathcal{L}_T(\beta)}{\partial \beta_j} = m_j - \sum_i \delta_i e^{\beta_j} H_0(Y_i) = 0$$

hence the mle estimator of β_j is

$$\widehat{\beta}_j = \log \big[m_j / \sum_i \delta_i H_0(Y_i)\big].$$

Another case that can be solved explicitly is where there is a single explanatory variable x that takes only non-negative integer values. Then $\frac{\partial \mathcal{L}_n}{\partial \beta}$ is just a polynomial in e^{β} and may be solvable.

But in general, we need to use numerical methods. The numerical methods can be simplified by rewriting the likelihood as a GLM log-likelihood, plus an additional term

which plays no role in the estimation. This means, we can easily estimate β using existing statistical software. We observe that log-likelihood can be written as

$$\begin{aligned}\mathcal{L}_n(\beta) &= \sum_{i=1}^{n} \delta_i\big[\beta' x_i) + \log h_0(Y_i)\big] - \sum_{i=1}^{n} \big[\exp(\beta' x_i)) H_0(Y_i)\big] \\ &= \sum_{i=1}^{n} \delta_i\big[\beta' x_i) + \log H_0(Y_i)\big] - \sum_{i=1}^{n} \exp(\beta' x_i) H_0(Y_i) + \sum_{i=1}^{n} \delta_i \log \frac{h_0(Y_i)}{H_0(Y_i)}.\end{aligned}$$

Hence the parameter which maximises $\mathcal{L}_n(\beta)$ also maximises $\tilde{\mathcal{L}}_n(\beta)$, where

$$\tilde{\mathcal{L}}_n(\beta) = \sum_{i=1}^{n} \delta_i\big[\beta' x_i) + \log H_0(Y_i)\big] - \sum_{i=1}^{n} \exp(\beta' x_i) H_0(Y_i).$$

In other words the likelihoods $\mathcal{L}_n(\beta)$ and $\tilde{\mathcal{L}}_n(\beta)$ lead to the same estimators. This means that we can use $\tilde{\mathcal{L}}_n(\beta)$ as a means of estimating β. The interesting feature about $\tilde{\mathcal{L}}_n(\beta)$ is that it is the log-likelihood of the Poisson distribution where δ_i is the variable (though in our case it only takes zero and one) with mean $\lambda_i = \exp(\beta' x_i) H_0(Y_i)$. Hence we can do the estimation of β within the GLM framework, where we use a Poisson log-likelihood (δ_i, x_i) as the observations and regressors, and model the mean λ_i as $\exp(\beta' x_i) H_0(Y_i)$.

It is worth mentioning that the above estimation method is based on the assumption that the baseline hazard h_0 is known. This will not always be the case, and we may want to estimate β without placing any distributional assumptions on h. This is possible using a Kaplan Meier (semiparametric) type likelihood. The reader is referred to a text book on survival analysis for further details.

11.1.2 Accelerated life model

An alternative method for modelling the influence of explanatory variables (regressors) on the response is to use the accelerated life model. An individual with explanatory variables x is assumed to experience time speeded up by a non-negative factor $\xi(x)$, where we suppose $\xi(0) = 1$, i.e. $x = 0$ represents the baseline again. Thus:

$$\begin{aligned}\mathcal{F}(t; x) &= \mathcal{F}_0(\xi(x)t) \\ f(t; x) &= \xi(x) f_0(\xi(x)t) \\ h(t; x) &= \xi(x) h_0(\xi(x)t).\end{aligned}$$

If there were only a small number of possible values for $\xi(x)$, either through x being very discrete (ordinal), or because of the assumed form for ξ, we could just take the unique values of $\xi(x)$ as parameters, and estimate these (the same can be done in the PH case).

Except in the case mentioned above, we usually assume a parametric form for ξ and estimate the parameters. As with the PH model, a natural choice is $\xi(x) = e^{\beta' x}$.

Popular choices for the baseline $\mathcal{F}_0$ is exponential, gamma, Weibull, log-normal and log-logistic.

MLE for the AL model with exponential speed-up

In this section we will assume that $\xi(x) = \exp(\beta' x)$. Hence $\mathcal{F}(t;x) = \mathcal{F}_0(\exp(\beta' x)t)$. There are various methods we can use to estimate β. One possibility is to go the likeihood route

$$\mathcal{L}_n(\beta) \;=\; \sum_{i=1}^{n} \delta_i \log h_0\big(\exp(\beta' x_i) Y_i\big) - \sum_{i=1}^{n} \exp(\beta' x_i) H_0(\exp(\beta' x_i) Y_i),$$

where the baseline hazard function h_0 is known. But this would mean numerically maximising the likelihood through brute force. To use such a method, we would require a good initial value for β. To obtain a good initial value, we now consider an alternative method for estimating β.

Let us define the transformed random variable $W = \log T + \beta' x$. The distribution function of W is

$$\begin{aligned} P\{W < w\} &= P\{\log T < w - \beta' x\} \\ &= 1 - P\{T \geq \exp(w - \beta' x)\} \\ &= 1 - \mathcal{F}(e^{w - \beta' x}) = 1 - \mathcal{F}_0(e^w) \end{aligned}$$

Thus, W has a distribution that is independent of x, and indeed is completely known if we assume the baseline is fully specified. Hence $\log T$ satisfies the linear model

$$\log T_i = \mu_0 - \beta' x_i + \varepsilon_i,$$

where $\mathrm{E}(W) = \mu_0$ and ε_i are iid random variables with mean zero.

Hence if the observations have not been censored we can estimate $\beta' x$, by minimising the log-likelihood

$$\sum_{i=1}^{n} \big(\beta' x_i + \log f_0(\log T_i + \beta' x_i)\big).$$

However, an even simplier method is to use classical least squares to estimate β. In other words use the μ and β which minimise

$$\sum_{i=1}^{n} \left(\log T_i - \mu - \beta' x_i \right)^2$$

as estimators of μ_0 and β respectively. Hence, this gives us the best Minimum Variance Linear Unbiased Estimator (MVLUE) of β. But it is worth mentioning that a likelihood based estimator gives a smaller asymptotic variance. If there is censoring, there are more complicated algorithms for censored linear models, or use Newton-Raphson for solving the likelihood equations.

Unlike, the proportional hazard models, there is no connection between parameter estimation in accelerated life models and GLM.

11.1.3 The relationship between the PH and AL models

The survivor functions under the two models are PH with hazard ratio function ψ: $\mathcal{F}(t;x) = (\mathcal{F}_0(t))^{\psi(x)}$ and AL with speed-up function ξ: $\mathcal{F}(t;x) = \mathcal{F}_0(\xi(x)t)$. Let us suppose the baseline survival distribution in both cases is the Weibull with

$$\mathcal{F}_0(t) = \exp\left\{ -\left(\frac{t}{\theta}\right)^{\alpha} \right\}$$

Hence using this distribution the proportional hazards and accelerated life survival functions are

$$\mathcal{F}_{\mathrm{PH}}(t;x) = \exp\left\{ -\left(\frac{t}{\theta}\right)^{\alpha} \psi(x) \right\} \text{ and } \mathcal{F}_{\mathrm{AL}}(t;x) = \exp\left\{ -\left(\frac{t\xi(x)}{\theta}\right)^{\alpha} \right\}.$$

Comparing the above survival functions we see that if $\xi(x) \equiv (\psi(x))^{1/\alpha}$, then we have $\mathcal{F}_{\mathrm{PH}}(t;x) = \mathcal{F}_{\mathrm{AL}}(t;x)$.

In fact, it is quite easy to show that this is the *only* case where the two models coincide.

11.1.4 Goodness of fit

As in most cases of statistical modelling we want to verify whether a model is appropriate for a certain data set. In the case of linear models, we do this by considering the residual sum of squares and for GLM we consider the deviance (see Section 9.3.5). The notion of 'residual' can be extended to survival data. We recall that 'residuals' in general should be

pivotal (or asymptotically pivotal) in the sense that their distribution should not depend on the unknown parameters. We now make a tranformation of the survival data which is close to pivotal if the survival distribution and model are correctly specified.

Let us first consider the case that the data is not censored. Let T_i denote the survival time, which has the survival and cumulative hazard functions $\mathcal{F}_i$ and $H_i(t) = -\log \mathcal{F}_i(t)$ (later we will introduce its dependence on the explanatory variable x_i). Let us consider the distribution of $H_i(T_i)$. The distribution function of $H_i(T_i)$ is

$$\begin{aligned} P(H_i(T_i) \leq y) &= P(\log \mathcal{F}_i(T_i) > -y) \\ &= P(\mathcal{F}_i(T_i) > \exp(-y)) \\ &= P(1 - F_i(T_i) \geq \exp(-y)) = P(F(T_i) \leq 1 - \exp(-y)) \\ &= P(T_i \leq F^{-1}(1 - \exp(-y))) = F(F^{-1}(1 - \exp(-y))) = 1 - \exp(-y). \end{aligned}$$

Hence the distribution of $H_i(T_i)$ is a exponential with mean one, in other words it does not depend on any unknown parameters.

Therefore in the case of uncensored data, to check for adequency of the model we can fit the survival models $\{\mathcal{F}(t, x_i; \beta)\}$ to the observations $\{T_i\}$ and check whether the transformed data $\{H(T_i, x_i; \hat{\beta})\}$ are close to iid exponentials. These are called the Cox-Snell residuals, they can be modified in the case of censoring.