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 numbers) to obtain the limit of $\mathcal{L}_n(\theta)$

$$\frac{1}{n}\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta) \overset{\text{a.s.}}{\to} \mathrm{E}_g\big(\log f(X_i;\theta)\big) = \int \log f(x;\theta)g(x)dx. \qquad (5.1)$$

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 $\widehat{\theta}_n \overset{\mathcal{P}}{\to} \theta_g$ (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

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta_g} \approx \frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\widehat{\theta}_n} + I(\theta_g)\sqrt{n}(\widehat{\theta}_n - \theta_g), \qquad (5.2)$$

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

$$\frac{\partial \int_{\mathbb{R}}\log f(x;\theta)g(x)dx}{\partial\theta}\rfloor_{\theta=\theta_g} = \int_{\mathbb{R}}\frac{\log f(x;\theta)}{\partial\theta}\rfloor_{\theta=\theta_g}g(x)dx = 0 \qquad (5.3)$$

and the usual regularity conditions are satisfied (exchanging derivative and integral is allowed and the third order derivative exists). Then we have

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta_g} \overset{\mathcal{D}}{\to} \mathcal{N}(0, J(\theta_g)), \qquad (5.4)$$

$$\sqrt{n}(\widehat{\theta}_n - \theta_g) \overset{\mathcal{D}}{\to} \mathcal{N}\big(0, I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}\big). \tag{5.5}$$

and

$$2\left(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_g)\right) \overset{\mathcal{D}}{\to} \sum_{j=1}^{p} \lambda_j Z_j^2 \tag{5.6}$$

where

$$\begin{aligned} I(\theta_g) &= \mathrm{E}\bigg(- \frac{\partial^2 \log f(X;\theta)}{\partial \theta^2} \rfloor_{\theta_g} \bigg) = -\int \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} g(x)dx \\ J(\theta_g) &= \mathrm{var}\bigg(\frac{\partial \log f(X;\theta)}{\partial \theta} \rfloor_{\theta=\theta_g} \bigg) = \mathrm{E}\bigg(\frac{\partial \log f(X;\theta)}{\partial \theta} \rfloor_{\theta=\theta_g} \bigg)^2 = \int \left(\frac{\partial \log f(x;\theta)}{\partial \theta} \right)^2 g(x)dx \end{aligned}$$

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

$$\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\theta_g} \overset{\mathcal{D}}{\to} \mathcal{N}(0, J(\theta_g)). \tag{5.7}$$

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

$$\sqrt{n}(\widehat{\theta}_n - \theta_g) \overset{\mathcal{D}}{\to} \mathcal{N}\big(0, I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}\big). \tag{5.9}$$

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

$$\sqrt{n}(\widehat{\theta}_n - \theta_g) \overset{\mathcal{D}}{\to} \mathcal{N}\big(0, I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}\big), \tag{5.11}$$

then asymptotically the distribution of $\sqrt{n}(\widehat{\theta}_n-\theta_g)$ is $\sqrt{n}(\widehat{\theta}_n-\theta_g) \overset{D}{=} I(\theta_g)^{-1/2}J(\theta_g)^{1/2}I(\theta_g)^{-1/2}\underline{Z}$ where $\underline{Z}$ is a p-dimension standard normal random variable. Thus we have

$$\begin{aligned} & 2\left(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_g)\right) \\ \overset{D}{=} \;& n\underline{Z}'I(\theta_g)^{-1/2}J(\theta_g)^{1/2}I(\theta_g)^{-1/2}I(\theta_g)I(\theta_g)^{-1/2}J(\theta_g)^{1/2}I(\theta_g)^{-1/2}\underline{Z} \\ = \;& \underline{Z}'I(\theta_g)^{-1/2}J(\theta_g)I(\theta_g)^{-1/2}\underline{Z} \end{aligned}$$

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

$$2\left(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_g)\right) = \sum_{j=1}^{p} \lambda_j Z_j^2$$

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 $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, I(\theta_0)^{-1}\big)$ in the misspecified case it is $\sqrt{n}(\widehat{\theta}_n - \theta_g) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}\big)$.

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)

$$\begin{aligned} J(a_g) &= \frac{1}{n}\mathrm{var}\left(\frac{\partial \mathcal{L}_n(a)}{\partial a}\rfloor_{a_g}\right) = \frac{1}{n\sigma^4}\sum_{i=1}^{n}\mathrm{var}(X_i)\left(\frac{i}{n}\right)^2 \to \frac{1}{\sigma^2}\int_0^1 u^2 du = \frac{1}{3\sigma^2} \\ I(a_g) &= \frac{1}{n}\mathrm{E}\left(-\frac{\partial^2 \mathcal{L}_n(a)}{\partial a^2}\rfloor_{a_g}\right) = \frac{1}{n\sigma^2}\sum_{i=1}^{n}\left(\frac{i}{n}\right)^2 \to \frac{1}{\sigma^2}\int_0^1 u^2 du = \frac{1}{3\sigma^2}. \end{aligned}$$

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

$$\int \log \frac{f(x;\theta)}{g(x)} g(x)dx = \mathrm{E}_g\left(\log \frac{f(X_i;\theta)}{g(X_i)}\right) \leq \log \mathrm{E}_g\left(\frac{f(X_i;\theta)}{g(X_i)}\right) \log \int f(x)dx \leq 0.(5.12)$$

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

$$\frac{1}{n}\mathcal{L}_n(\theta) = \frac{-1}{n}\sum_{i=1}^{n}\left(\log\theta + \frac{X_i}{\theta}\right) \overset{a.s.}{\to} -\log\theta - \mathrm{E}\big(\frac{X_i}{\theta}\big) = -\int\big(\log\theta + \frac{x}{\theta}\big)g(x)dx.$$

Let $\widehat{\theta}_n = \arg\max\mathcal{L}_n(\theta) = \bar{X}$*. Then we can see that* $\widehat{\theta}_n$ *is an estimator of*

$$\theta_g = \arg\max\{-\big(\log\theta + \mathrm{E}(X_i/\theta)\big)\} = \phi\Gamma(1+\alpha^{-1}) = \mathrm{E}(X_i) \qquad (5.13)$$

Therefore by using Theorem 5.1.1 (or just the regular central limit theorem for iid random variables) we have

$$\sqrt{n}\big(\widehat{\theta}_n - \phi\Gamma(1+\alpha^{-1})\big) \overset{\mathcal{P}}{\to} \mathcal{N}\big(0, \underbrace{I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}}_{=\mathrm{var}(X_i)}\big)$$

where

$$\begin{aligned} I(\theta_g) &= \mathrm{E}\Big(-\big(\theta^{-2} - 2X\theta^{-3}\big)\Big)\rfloor_{\theta=\mathrm{E}(X)} = [\mathrm{E}(X)]^{-2} \\ J(\theta_g) &= \mathrm{E}\Big(\big(-\theta^{-1} + X\theta^{-2}\big)^2\Big)\rfloor_{\theta=\mathrm{E}(X)} = \frac{\mathrm{E}(X^2)}{[\mathrm{E}(X)]^4} - \frac{1}{[\mathrm{E}(X)]^2} = \frac{1}{\mathrm{E}[X^2]}\left(\frac{\mathrm{E}[X^2]}{\mathrm{E}[X]^2} - 1\right). \end{aligned}$$

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;*

$$\begin{aligned} D(g, f_{\theta_g}) &= \int\log\left(\frac{\theta_g^{-1}\exp(-\theta_g^{-1}x)}{\frac{\alpha}{\phi}(\frac{x}{\phi})^{\alpha-1}\exp(-(\frac{x}{\phi})^{\alpha}}\right)\frac{\alpha}{\phi}(\frac{x}{\phi})^{\alpha-1}\exp(-(\frac{x}{\phi})^{\alpha})dx \\ &= \int\log\left(\frac{\phi\Gamma(1+\alpha^{-1})^{-1}\exp(-\phi\Gamma(1+\alpha^{-1})^{-1}x)}{\frac{\alpha}{\phi}(\frac{x}{\phi})^{\alpha-1}\exp(-(\frac{x}{\phi})^{\alpha}}\right)\frac{\alpha}{\phi}(\frac{x}{\phi})^{\alpha-1}\exp(-(\frac{x}{\phi})^{\alpha})dx. \quad (5.14) \end{aligned}$$

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

$$a_g = \arg\max_a \mathrm{E}\big(-6\log(1 + \frac{(X-a)^2}{11})\big) = \arg\min \int \log(1 + \frac{(x-a)^2}{11})\big)d\Phi\big(\frac{x-\mu}{\sigma}\big)dx.$$

(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

$$\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} = \widetilde{D}[g, f_{\theta_g}] + \Big(\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} - \widetilde{D}\left[g, f_{\theta_g}\right]\Big).$$

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

$$\begin{aligned}\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} &\approx -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g) + \Big(\mathrm{E}_{\underline{X}}\left\{\widetilde{D}\left[g, f_{\widehat{\theta}_n(\underline{X})}\right]\right\} - \mathrm{E}_{\underline{X}}\left\{\widetilde{D}[g, f_{\theta_g}]\right\}\Big)\\ &= -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g) + I_1.\end{aligned}$$

Of course, θ_g is unknown so this is replaced by $\widehat{\theta}_n(\underline{X})$ to give

$$\mathrm{E}_{\underline{X}}\big(\widetilde{D}(g, f_{\widehat{\theta}_n(\underline{X})})\big) \approx -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\widehat{\theta}_n(\underline{X})) + I_1 + I_2 \qquad (5.15)$$

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

$$\begin{aligned}
& \mathrm{E}_{\underline{X}}\Big[\widetilde{D}(g, f_{\widehat{\theta}_n(\underline{X})}) - \widetilde{D}(g, f_{\theta_g})\Big] \\
= \;& -\mathrm{E}_{\underline{X}}\mathrm{E}_{\underline{Y}}\bigg(\frac{1}{n}\sum_{i=1}^{n}\Big\{\log f(Y_i;\widehat{\theta}_n(\underline{X})) - \log f(Y_i;\theta_g)\Big\}\bigg) \\
= \;& -\frac{1}{n}\mathrm{E}_{\underline{X}}\mathrm{E}_{\underline{Y}}\bigg(\mathcal{L}_n(\underline{Y}, \widehat{\theta}_n(\underline{X})) - \mathcal{L}_n(\underline{Y}, \theta_g)\bigg) \\
= \;& -\frac{1}{n}\mathrm{E}_{\underline{X}}\underbrace{\mathrm{E}_{\underline{Y}}\bigg(\frac{\partial \mathcal{L}_n(\underline{Y}, \theta)}{\partial \theta}\rfloor_{\theta_g}(\widehat{\theta}_n(\underline{X}) - \theta_g)\bigg)}_{=0} - \frac{1}{2n}\mathrm{E}_{\underline{Y}}\mathrm{E}_{\underline{X}}\bigg((\widehat{\theta}_n(\underline{X}) - \theta_g)'\frac{\partial^2 \mathcal{L}_n(\underline{Y}, \theta)}{\partial \theta}\rfloor_{\bar{\theta}(\underline{X})}(\widehat{\theta}_n(\underline{X}) - \theta_g)\bigg) \\
= \;& -\frac{1}{2n}\mathrm{E}_{\underline{Y}}\mathrm{E}_{\underline{X}}\bigg((\widehat{\theta}_n(\underline{X}) - \theta_g)'\frac{\partial^2 \mathcal{L}_n(\underline{Y}, \theta)}{\partial \theta}\rfloor_{\bar{\theta}(\underline{X})}(\widehat{\theta}_n(\underline{X}) - \theta_g)\bigg),
\end{aligned}$$

where $\bar{\theta}(\underline{X}) = \alpha\theta(\underline{X}) + (1-\alpha)\theta_g$ for some $0 \leq \alpha \leq 1$. Now we note that

$$-\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\underline{Y}, \theta)}{\partial \theta^2}\rfloor_{\bar{\theta}(\underline{X})} \approx -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i, \theta)}{\partial \theta^2}\rfloor_{\theta=\theta_g} \overset{\mathcal{P}}{\to} I(\theta_g),$$

which (using a hand wavey argument) gives

$$I_1 \;=\; \mathrm{E}_{\underline{X}}\big(\widetilde{D}(g, f_{\widehat{\theta}_n(\underline{X})}) - \widetilde{D}(g, f_{\theta_g})\big) \approx \frac{1}{2}\mathrm{E}_{\underline{X}}\bigg((\widehat{\theta}_n(\underline{X}) - \theta_g)'I(\theta_g)(\widehat{\theta}_n(\underline{X}) - \theta_g)\bigg). \qquad (5.16)$$

We now obtain an estimator of I_2 in (5.15). To do this we make the usual Taylor expansion (noting that $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta=\widehat{\theta}_n} = 0$)

$$\begin{aligned}
I_2 \;&=\; \bigg(\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g) - \frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\widehat{\theta}_n(\underline{X}))\bigg) \\
&\approx\; \frac{1}{2}(\widehat{\theta}_n(\underline{X}) - \theta_g)'I(\theta_g)(\widehat{\theta}_n(\underline{X}) - \theta_g). \qquad (5.17)
\end{aligned}$$

To obtain the final approximations for (5.16) and (5.17) we use (5.11) where

$$\sqrt{n}(\widehat{\theta}_n - \theta_g) \overset{\mathcal{D}}{\to} \mathcal{N}\big(0, I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}\big).$$

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

$$\begin{aligned}
I_2 &= -\Big(\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta_g) - \frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\widehat{\theta}_n(\underline{X}))\Big) \\
&\approx \frac{1}{2}(\widehat{\theta}_n(\underline{X})-\theta_g)'I(\theta_g)(\widehat{\theta}_n(\underline{X})-\theta_g) \\
&\approx \frac{1}{2}\mathrm{E}\left(\underbrace{(\widehat{\theta}_n(\underline{X})-\theta_g)'}_{\approx\mathcal{N}(0,I(\theta_g)^{-1}J(\theta_g)I(\theta_g)^{-1}/n)} \quad I(\theta_g)(\widehat{\theta}_n(\underline{X})-\theta_g)\right) \\
&\approx \frac{1}{2n}\mathrm{trace}\Big(I(\theta_g)^{-1}J(\theta_g)\Big)
\end{aligned}$$

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
$$\min_{1\leq p\leq N}\bigg(-\frac{1}{n}\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n})+\frac{1}{n}\mathrm{trace}\big(\widehat{J}_p\widehat{I}_p^{-1}\big)\bigg) \tag{5.19}$$
In other words, the order we select is $\widehat{p}$ where
$$\widehat{p}=\arg\min_{1\leq p\leq N}\bigg(-\frac{1}{n}\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n})+\frac{1}{n}\mathrm{trace}\big(\widehat{J}_p\widehat{I}_p^{-1}\big)\bigg)$$

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 $Y_i=\alpha_0+\alpha_1x_{i,1}+\varepsilon_i$ belongs to the set of families defined by
$$Y_{i,p}=\alpha_0+\sum_{j=1}^{p}\alpha_jx_{i,j}+\varepsilon_i \qquad p>1.$$
In this case $\{\alpha_0+\sum_{j=1}^{p}\alpha_jx_{i,j}+\varepsilon_i;\alpha_j\in\mathbb{R}^{p+1}\}$ 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
$$AIC(p)=\min_{1\leq p\leq N}\bigg(-\frac{1}{n}\mathcal{L}_{p,n}(\underline{X};\widehat{\theta}_{p,n})+\frac{p}{n}\bigg).$$

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 $2(\mathcal{L}_{p+q,n} - \mathcal{L}_{p,n}) \overset{\mathcal{D}}{\to} \chi^2_q$. 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* $\mathrm{E}_{\underline{X}}\big\{\widetilde{D}[g, f_{\widehat{\theta}_n(\underline{X})}]\big\}$ *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

$$\begin{aligned}\rangle_n &= 2\big(\mathcal{L}_{n,q}(\widehat{\beta}_q) - \mathcal{L}_{i,p}(\widehat{\beta}_p\big) \\ &= 2\sum_{i=1}^{n}\Big(Y_i[\widehat{\beta}_A' - \widehat{\beta}_0']\underline{x}_i - n_i[\log(1+\exp(\widehat{\beta}_A'\underline{x}_i) - \log(1+\exp(\widehat{\beta}_0'\underline{x}_i)]\Big)\end{aligned}$$

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 $2\big[\mathcal{L}_q(\widehat{\beta}_q) - \mathcal{L}_p(\widehat{\beta}_p) \overset{\mathcal{D}}{\to} \chi^2_{q-p}$*, 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.