# 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 $\ell_1$ and $\ell_2$ 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 $\theta$, for now the scale parameter $\rho$ is not of interest. The log-likelihood is

$$\mathcal{L}_n(\theta,\rho) = -n\log 2\rho - \rho^{-1}\underbrace{\frac{1}{2}\sum_{i=1}^{n}|Y_i-\theta|}_{=L_n(\theta)}.$$

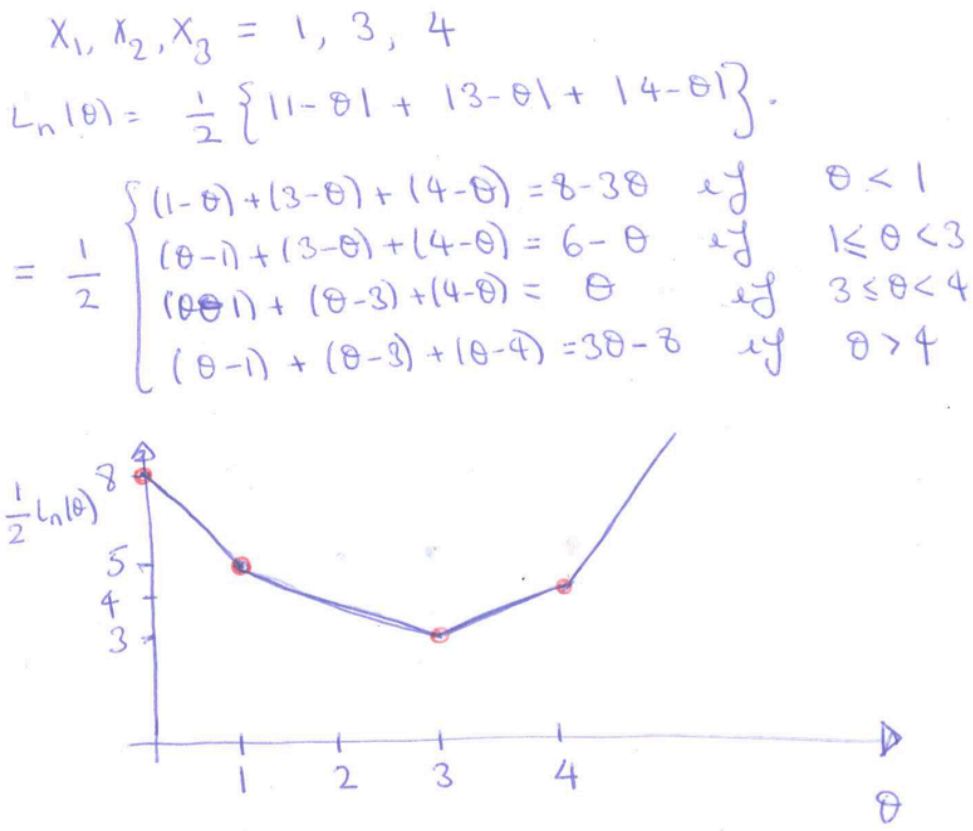


Figure 8.1: Plot of $L1$-norm

Since the $\theta$ which maximises the above does not depend on $\rho$ we can simply focus on the component which maximises $\theta$. We see that this is equivalent to minimising the loss function

$$L_n(\theta) = \frac{1}{2}\sum_{i=1}^{n}|Y_i - \theta| = \sum_{Y_{(i)}>\theta}\frac{1}{2}(Y_{(i)} - \theta) + \sum_{Y_{(i)}\leq\theta}\frac{1}{2}(\theta - Y_{(i)}).$$

If we make a plot of $L_n$ over $\theta$, 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 $\ell_2$-loss function and the sample mean. In contrast the Laplacian gives rise to the $\ell_1$-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 $\theta$ 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 $p$th 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 $\mu$ and $\sigma^2$ we use the $\mu$ and $\sigma^2$ 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 $\theta$ 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* $\alpha$ *and* $\Gamma$*. 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* $\alpha$ *and* $\Gamma$ *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* $\alpha$ *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* $\theta_0$ $G_n(\cdot)$ *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* $\theta$ *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 $\theta_0$. We want to estimate $\theta_0$, 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 $\theta$. 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* $\varepsilon_t$*, but under certain conditions on* $a_1$ *and* $a_2$ *(causality conditions)* $\varepsilon_t$ *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 $\theta$.

**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 $\theta$ is a scalar, for every $n$ $G_n(\theta)$ is a continuous monotonically decreasing function in $\theta$ and for all $\theta$ $G_n(\theta) \xrightarrow{\mathcal{P}} \mathrm{E}(G_n(\theta))$ (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 $\theta_0$ 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 $\varepsilon > 0$ $P(\widetilde{\theta}_n < \theta_0 - \varepsilon \text{ or } \widetilde{\theta}_n > \theta_0 + \varepsilon) \to 0$ 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 $\varepsilon$, 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 $\widetilde{\theta}_n$.

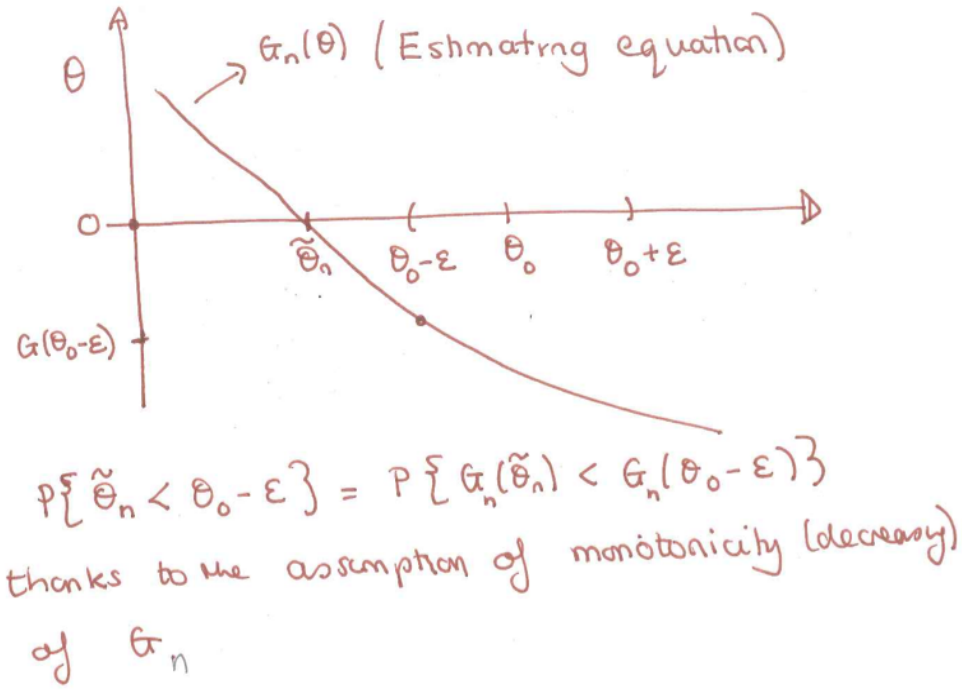


Figure 8.2: Plot of $G_n(\cdot)$

**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 $\theta$ 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 $\theta$ 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 $\theta$, and density function which is symmetric about the mean $\theta$. 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 $\theta$, 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 $\phi$, $\xi_i$ and $\varepsilon_i$ 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 $\phi$.

(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 $\sigma_\xi^2$ and $\sigma_\varepsilon^2$.

(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 $\xi_i$ and $\varepsilon_i$. 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 $\varepsilon_i$ and $\xi_i$.

Show that the expectation of the *negative* log-likelihood is minimised at the true parameters $\phi$, $\sigma_\xi^2$ and $\sigma_\varepsilon^2$ *even* when $\xi_i$ and $\varepsilon_i$ 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 $\phi$.*

$\widehat{\phi} = \arg\min_\phi \sum_{i=1}^n (Y_i - \phi X_i)^2 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}$

(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 $\sigma_\xi^2$ and $\sigma_\varepsilon^2$.*

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 $\sigma_\xi^2$ and $\sigma_\varepsilon^2$, 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 $\sigma_\xi^2$ and $\sigma_\varepsilon^2$. 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 $\xi_t$ and $\varepsilon_t$. 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 $\phi$, $\sigma_\xi^2$ and $\sigma_\varepsilon^2$ even when $\xi_t$ and $\varepsilon_t$ 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 $\theta_0$ 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 $\theta$ and zero when $\theta = \theta_0$. This immediately implies that $\theta_0$ 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 $\phi$ by $\phi_0$

$$\begin{aligned} B_i(\theta) &= \mathrm{E}\left[\frac{(Y_i - \phi_0 X_i)^2}{\sigma^2_\varepsilon + \sigma^2_\xi X_i^2}\right] + \mathrm{E}\left[\frac{(Y_i - \phi X_i)^2 - (Y_i - \phi_0 X_i)^2}{\sigma^2_\varepsilon + \sigma^2_\xi X_i^2}\right] - 1 \\ &= \mathrm{E}\left[\frac{(\varepsilon_t + \xi_i X_i)^2}{\sigma^2_\varepsilon + \sigma^2_\xi X_i^2}\right] + \mathrm{E}\left[\frac{2(\phi - \phi_0)(Y_i - \phi_0 X_i) X_i}{\sigma^2_\varepsilon + \sigma^2_\xi X_i^2}\right] + \\ &\quad \mathrm{E}\left[\frac{(\phi - \phi_0)^2 X_i^2}{\sigma^2_\varepsilon + \sigma^2_\xi X_i^2}\right] - 1 \\ &= \frac{\mathrm{E}\left[(\varepsilon_t + \xi_i X_i)^2\right]}{\sigma^2_\varepsilon + \sigma^2_\xi X_i^2} + \mathrm{E}\left[\frac{2(\phi - \phi_0)(Y_i - \phi_0 X_i) X_i}{\sigma^2_\varepsilon + \sigma^2_\xi X_i^2}\right] + \\ &\quad \frac{(\phi - \phi_0)^2 X_i^2}{\sigma^2_\varepsilon + \sigma^2_\xi X_i^2} - 1 \\ &= \frac{\sigma^2_{0,\varepsilon} + \sigma^2_{0,\xi} X_i^2}{\sigma^2_\varepsilon + \sigma^2_\xi X_i^2} + \frac{(\phi - \phi_0)^2 X_i^2}{\sigma^2_\varepsilon + \sigma^2_\xi X_i^2} - 1. \end{aligned}$$

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

$$L_i(\theta) - L_i(\theta_0) \;\; = \;\; -\log x + x - 1 + (\phi - \phi_0)^2 \frac{X_i^2}{\sigma^2_{\varepsilon} + \sigma^2_{\xi} X_i^2}.$$

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

$$L(\theta) - L(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \left( L_i(\theta) - L_i(\theta_0) \right) \geq 0,$$

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 $\theta_0$ is unknown. One possible estimating equation is

$$G_{1,n}(\theta) = \sum_{i=1}^{n} \left[ Y_i - \mu_i(\theta) \right],$$

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 $\lambda$. Partially differentiating the above with respect to $\{w_i(\theta)\}$ and $\lambda$ 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 $\theta$ (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* $\sigma_i$ *and* $Z_i$ *are positive,* $\{Z_i\}$ *are iid random variables and the regressors* $x_i$ *influence* $\sigma_i$ *through the relation* $\sigma_i = \exp(\beta_0 + \beta_1' x_i)$*. To estimate* $\beta_0$ *and* $\beta_1$ *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* $\beta_0$ *and* $\beta_1$*. 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* $\beta_0$ *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* $\theta_0$ *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* $\theta_0$*, 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* $\theta$*.*