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*

$$\begin{aligned}
\hat{\theta}_{\alpha} &= \arg\max_{\theta} \mathcal{L}_{\alpha}(\theta) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \bigg(\log\alpha + (\alpha-1)\log Y_i - \alpha\log\theta - (\frac{Y_i}{\theta})^{\alpha} \bigg) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \bigg(-\alpha\log\theta - (\frac{Y_i}{\theta})^{\alpha} \bigg) = (\frac{1}{n}\sum_{i=1}^{n} Y_i^{\alpha})^{1/\alpha},
\end{aligned}$$

where $\mathcal{L}_{\alpha}(\underline{X};\theta) = \sum_{i=1}^{n} \bigg(\log\alpha + (\alpha-1)\log Y_i - \alpha\log\theta - \big(\frac{Y_i}{\theta}\big)^{\alpha} \bigg)$. Thus for a given α, the maximum likelihood estimator of θ can be derived. The maximum likelihood estimator of α is

$$\hat{\alpha}_n = \arg\max_{\alpha} \sum_{i=1}^{n} \bigg(\log\alpha + (\alpha-1)\log Y_i - \alpha\log(\frac{1}{n}\sum_{i=1}^{n} Y_i^{\alpha})^{1/\alpha} - (\frac{Y_i}{(\frac{1}{n}\sum_{i=1}^{n} Y_i^{\alpha})^{1/\alpha}})^{\alpha} \bigg).$$

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

$$\sqrt{n}\left(\begin{array}{c} \hat{\psi}_n - \psi \\ \hat{\lambda}_n - \lambda \end{array}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \left(\begin{array}{cc} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{array}\right)^{-1}\Big).$$

where

$$\left(\begin{array}{cc} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{array}\right) = \left(\begin{array}{cc} \mathrm{E}\big(-\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial\psi^2}\big) & \mathrm{E}\big(-\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial\psi\partial\lambda}\big) \\ \mathrm{E}\big(-\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial\psi\partial\lambda}\big)' & \mathrm{E}\big(-\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial\psi^2}\big) \end{array}\right). \tag{3.1}$$

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*

$$\left(\begin{array}{cc} A & B \\ C & D \end{array}\right)$$

is a square matrix. Then

$$\left(\begin{array}{cc} A & B \\ C & D \end{array}\right)^{-1} = \left(\begin{array}{cc} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}CB(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{array}\right). \tag{3.2}$$

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*

$$I(\psi,\lambda) = \left(\begin{array}{cc} I_{\psi,\psi} & 0 \\ 0 & I_{\lambda,\lambda} \end{array}\right),$$

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

$$\begin{aligned} & 2\Big\{\max_{\psi,\lambda}\mathcal{L}_n(\psi,\lambda) - \max_{\lambda}\mathcal{L}_n(\psi_0,\lambda)\Big\} \\ = \; & \underbrace{2\left\{\max_{\psi,\lambda}\mathcal{L}_n(\psi,\lambda) - \mathcal{L}_n(\psi_0,\lambda_0)\right\}}_{\chi^2_{p+q}} - \underbrace{2\left\{\max_{\lambda}\mathcal{L}_n(\psi_0,\lambda) - \max_{\lambda}\mathcal{L}_n(\psi_0,\lambda_0)\right\}}_{\chi^2_q}. \end{aligned}$$

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

$$\begin{aligned} \frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\psi,\lambda)}{\partial \lambda \partial \psi}\rfloor_{\lambda_0,\psi_0} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial \lambda \partial \psi}\rfloor_{\lambda_0,\psi_0} \\ &\approx \mathrm{E}\bigg(\frac{\partial^2 \log f(X_i;\psi,\lambda)}{\partial \lambda \partial \psi}\rfloor_{\lambda_0,\psi_0}\bigg) = -I_{\lambda,\psi} \end{aligned}$$

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

$$\begin{aligned}(\hat{\lambda}_{\psi_0}-\lambda_0) &\approx \frac{I_{\lambda\lambda}^{-1}}{n}\frac{\partial\mathcal{L}_n(\psi_0,\lambda)}{\partial\lambda}\rfloor_{\lambda_0,\psi_0}\approx\frac{I_{\lambda\lambda}^{-1}}{n}\Big(I_{\lambda\psi}(\hat{\psi}-\psi_0)+I_{\lambda\lambda}(\hat{\lambda}-\lambda_0)\Big)\\ &\approx I_{\lambda\lambda}^{-1}I_{\lambda\psi}(\hat{\psi}-\psi_0)+(\hat{\lambda}-\lambda_0)=\left(I_{\lambda\lambda}^{-1}I_{\lambda\psi},1\right)\left(\widehat{\theta}_n-\theta_0\right).\end{aligned}$$

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

$$\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial\psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0} \approx \frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial\psi}\rfloor_{\lambda_0,\psi_0} - I_{\psi\lambda}I^{-1}_{\lambda\lambda}\frac{\partial \mathcal{L}_n(\psi_0,\lambda)}{\partial\lambda}\rfloor_{\lambda_0,\psi_0} \qquad (3.13)$$

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 $\hat{Y}(X)=\mathrm{E}(XY)/\mathrm{E}(Y^2)X$ and the residual and mean squared error is

$$Y-\hat{Y}(X) = Y-\frac{\mathrm{E}(XY)}{\mathrm{E}(Y^2)}X \text{ and } \mathrm{E}\bigg(Y-\frac{\mathrm{E}(XY)}{\mathrm{E}(Y^2)}X\bigg)^2 = \mathrm{E}(Y^2)-\mathrm{E}(XY)\mathrm{E}(Y^2)^{-1}\mathrm{E}(XY).$$

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

$$\begin{aligned}\left\{\mu; 2\left\{\mathcal{L}_n(\widehat{\mu},\widehat{\sigma}^2(\widehat{\mu})) - \mathcal{L}_n(\mu,\widehat{\sigma}^2(\mu))\right\}\right\} &= \left\{\mu; W_n(\mu) \leq \chi_1^2(0.95)\right\} \\ &= \left\{\mu; n\log\left(1+\frac{t_n^2(\mu)}{n-1}\right) \leq \chi_1^2(0.95)\right\}.\end{aligned}$$

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 all results you use.

	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

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

$$n^{-1/2}\frac{\partial \mathcal{L}_n(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0}} \overset{\mathcal{D}}{\to} \mathcal{N}(0, I_{\psi\psi} - I_{\psi\lambda}I_{\lambda,\lambda}^{-1}I_{\lambda,\psi}), \tag{3.14}$$

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

$$\sum_{t=1}^{n}\exp(i\Omega t) = \begin{cases} \frac{\exp(\frac{1}{2}i(n+1)\Omega)\sin(\frac{1}{2}n\Omega)}{\sin(\frac{1}{2}\Omega)} & 0<\Omega<2\pi \\ n & \Omega = 0 \text{ or } 2\pi. \end{cases} \qquad (3.15)$$

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 $\hat{\omega}_n$.

<u>Useful information:</u> 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

$$\mathcal{L}_n(\widehat{A}_n(\omega), \widehat{B}_n(\omega), \omega) \quad = \quad \frac{-1}{2}\mathcal{S}_p(\omega) + O(1),$$

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 $\sum_{t=1}^{n} t\cos(\Omega t) = O(n) \quad \sum_{t=1}^{n} t\sin(\Omega t) = O(n)$. 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

$$2\left(\frac{A^2+B^2}{6}n^3+O(n^2)\right)^{-1} \approx \frac{12}{(A^2+B^2)n^3}$$

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

$$f_i(y_i, \delta_i) = \left\{ \begin{array}{ll} f_i(y_i) = \psi(x_i;\beta)\big\{\mathcal{F}_0(y_i)\big\}^{[\psi(x_i;\beta)-1} f_0(t) & \delta_i = 0 \\ \mathcal{F}_i(y_i) = \mathcal{F}_0(t)^{\psi(x_i;\beta)} & \delta_i = 1 \end{array}\right.$$

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

$$\mathrm{E}\left(Y|U=u\right) = \int_{\mathbb{R}} y f_{Y|U}(y|u)dy \int_{\mathbb{R}} \frac{y f_{Y,U}(y,u)}{f_U(u)}dy = \phi(u),$$

where we replace $f_{Y,U}$ and f_U with

$$\begin{aligned}
\widehat{f}_{Y,U}(u,y) &= \frac{1}{bn}\sum_{i=1}^{n} \delta_{Y_i}(y) W_b\left(u-U_i\right) \\
\widehat{f}_U(u) &= \frac{1}{bn}\sum_{i=1}^{n} W_b\left(u-U_i\right),
\end{aligned}$$

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

$$\begin{aligned}\hat{\phi}_\beta(u) &= \arg\min_a \sum_i W_b(u-U_i)(Y_i(\beta)-a)^2 = \frac{\sum_i W_b(u-U_i)Y_i(\beta)}{\sum_i W_b(u-U_i)} \\ &= \frac{\sum_i W_b(u-U_i)Y_i}{\sum_i W_b(u-U_i)} - \beta\frac{\sum_i W_b(u-U_i)X_i}{\sum_i W_b(u-U_i)} \\ &:= G_b(u) - \beta H_b(u), \qquad (3.17)\end{aligned}$$

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.