Chapter 6

The Profile Likelihood

6.1 The Profile Likelihood

See also Section 4.5.2, Davison (2002).

6.1.1 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 is in testing only the parameter ψ (without any information on λ) and construction confidence intervals for ψ (without constructing unnecessary confidence intervals for λ - confidence intervals for a large number of parameters are wider than those for a few parameters). To achieve this one often uses the profile likelihood. To motivate the profile likelihood, we first describe a method to estimate the parameters (ψ, ψ) in two stages and consider some examples.

Let us suppose that $\{X_t\}$ 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

$$\mathcal{L}_T(\psi, \lambda) = \sum_{t=1}^T \log f(X_t; \psi, \lambda).$$

To estimate ψ and λ one can use $(\hat{\lambda}_T, \hat{\psi}_T) = \arg \max_{\lambda, \psi} \mathcal{L}_T(\psi, \lambda)$. However, this can quite difficult, and lead to expression which are hard to 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 $\mathcal{L}_T(\psi, \lambda) = \mathcal{L}_{\psi}(\lambda)$ (to show that ψ is fixed but λ varies). To estimate λ we maximise $\mathcal{L}_{\psi}(\lambda)$ with respect to λ , ie.

$$\hat{\lambda}_{\psi} = \arg \max_{\lambda} \mathcal{L}_{\psi}(\lambda).$$

In reality ψ this 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

$$\hat{\psi}_T = \arg\max_{\psi} \mathcal{L}_{\psi}(\hat{\lambda}_{\psi}) = \arg\max_{\psi} \mathcal{L}_T(\psi, \hat{\lambda}_{\psi}).$$

A bit of logical deduction shows that $\hat{\psi}_T$ and $\lambda_{\hat{\psi}_T}$ are the maximum likelihood estimators $(\hat{\lambda}_T, \hat{\psi}_T) = \arg \max_{\psi, \lambda} \mathcal{L}_T(\psi, \lambda)$.

We note that we have *profiled* out nuisance parameter λ , and the likelihood $\mathcal{L}_{\psi}(\hat{\lambda}_{\psi}) = \mathcal{L}_{T}(\psi, \hat{\lambda}_{\psi})$ is completely in terms of the parameter of interest ψ .

The advantage of this best illustrated through some examples.

Example 6.1.1 Let us suppose that $\{X_t\}$ 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 4.1.2, that if α , were known an explicit expression for the MLE can be derived, it is

$$\begin{split} \hat{\theta}_{\alpha} &= \arg \max_{\theta} \mathcal{L}_{\alpha}(\theta) \\ &= \arg \max_{\theta} \sum_{t=1}^{T} \left(\log \alpha + (\alpha - 1) \log Y_{t} - \alpha \log \theta - \left(\frac{Y_{t}}{\theta} \right)^{\alpha} \right) \\ &= \arg \max_{\theta} \sum_{t=1}^{T} \left(-\alpha \log \theta - \left(\frac{Y_{t}}{\theta} \right)^{\alpha} \right) = \left(\frac{1}{T} \sum_{t=1}^{T} Y_{t}^{\alpha} \right)^{1/\alpha}, \end{split}$$

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

$$\hat{\alpha}_T = \arg\max_{\alpha} \sum_{t=1}^T \left(\log\alpha + (\alpha - 1)\log Y_t - \alpha \log(\frac{1}{T} \sum_{t=1}^T Y_t^{\alpha})^{1/\alpha} - \left(\frac{Y_t}{(\frac{1}{T} \sum_{t=1}^T Y_t^{\alpha})^{1/\alpha}} \right)^{\alpha} \right).$$

Therefore, the maximum likelihood estimator of θ is $(\frac{1}{T}\sum_{t=1}^{T}Y_{t}^{\hat{\alpha}_{T}})^{1/\hat{\alpha}_{T}}$. We observe that evaluating $\hat{\alpha}_{T}$ can be tricky but no worse than maximising the likelihood $\mathcal{L}_{T}(\alpha, \theta)$ over α and θ .

As we mentioned above, we often do not have any interest in the nuisance parameters λ and are only interesting in testing and constructing CIs for α . In this case, we are interested in the limiting distribution of the MLE $\hat{\alpha}_T$. This can easily be derived by observing that

$$\sqrt{T} \begin{pmatrix} \hat{\psi}_T - \psi \\ \hat{\lambda}_T - \lambda \end{pmatrix} \stackrel{\mathcal{D}}{\to} \mathcal{N} \left(0, \begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix}^{-1} \right).$$

where

$$\begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left(-\frac{\partial^2 \log f(X_t;\psi,\lambda)}{\partial \psi^2}\right) & \mathbb{E}\left(-\frac{\partial^2 \log f(X_t;\psi,\lambda)}{\partial \psi \partial \lambda}\right) \\ \mathbb{E}\left(-\frac{\partial^2 \log f(X_t;\psi,\lambda)}{\partial \psi \partial \lambda}\right)' & \mathbb{E}\left(-\frac{\partial^2 \log f(X_t;\psi,\lambda)}{\partial \psi^2}\right) \end{pmatrix}.$$
(6.1)

To derive an exact expression for the limiting variance of $\sqrt{T}(\hat{\psi}_T - \psi)$, we note that the inverse of a block matrix is

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

Thus the above implies that

$$\sqrt{T}(\hat{\psi}_T - \psi) \xrightarrow{\mathcal{D}} \mathcal{N}(0, (I_{\psi,\psi} - I_{\psi,\lambda}I_{\lambda\lambda}^{-1}I_{\lambda,\psi})^{-1}).$$

Thus if ψ is a scalar we can easily use the above to construct confidence intervals for ψ .

Exercise: How to estimate $I_{\psi,\psi} - I_{\psi,\lambda}I_{\lambda\lambda}^{-1}I_{\lambda,\psi}$?

6.1.2 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. The above gives us the limiting distribution of $(\hat{\psi}_T - \psi_0)$, this allows us to test ψ , however the test ignores any dependency that may exist with the nusiance estimator parameter $\hat{\lambda}_T$. An alternative test, which circumvents this issue is to do a log-likelihood ratio test of the type

$$2\left\{\max_{\psi,\lambda} \mathcal{L}_T(\psi,\lambda) - \max_{\lambda} \mathcal{L}_T(\psi_0,\lambda)\right\}. \tag{6.2}$$

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 nusiance parameters. This is because a direct Taylor expansion does not work. However we observe that

$$2\left\{\max_{\psi,\lambda} \mathcal{L}_T(\psi,\lambda) - \max_{\lambda} \mathcal{L}_T(\psi_0,\lambda)\right\} = 2\left\{\max_{\psi,\lambda} \mathcal{L}_T(\psi,\lambda) - \mathcal{L}_T(\psi_0,\lambda_0)\right\} - 2\left\{\max_{\lambda} \mathcal{L}_T(\psi_0,\lambda) - \max_{\lambda} \mathcal{L}_T(\psi_0,\lambda_0)\right\}$$

now we will show below that by using a few Taylor expansions we can derive the limiting distribution of (6.2).

In the theorem below we will derive the distribution of the score and the nested-loglikelihood. Please note you do not have to learn this proof.

Theorem 6.1.1 Suppose Assumption 4.1.1 holds. Suppose that (ψ_0, λ_0) are the true parameters. Then we have

$$\frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0}, \psi_0} \approx \frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\lambda_0, \psi_0} - I_{\psi_0 \lambda_0} I_{\lambda_0 \lambda_0}^{-1} \frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\psi_0, \lambda_0}$$
(6.3)

$$\frac{1}{\sqrt{T}} \frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0}, \psi_0} \xrightarrow{\mathcal{D}} \mathcal{N}(0, (I_{\psi_0 \psi_0} - I_{\psi_0 \lambda_0} I_{\lambda_0 \lambda_0}^{-1} I_{\lambda, \psi}))$$

$$\tag{6.4}$$

and

$$2\left\{\mathcal{L}_T(\hat{\psi}_T, \hat{\lambda}_T) - \mathcal{L}_T(\psi_0, \hat{\lambda}_{\psi_0})\right\} \xrightarrow{\mathcal{D}} \chi_p^2$$
(6.5)

where I is defined as in (6.1).

PROOF. We first prove (6.3) which is the basis of the proofs of (6.4) and (6.5) - in the remark below we try to interprete (6.3). To avoid, notationally difficulties by considering the elements of the vector $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ and $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \lambda} \rfloor_{\lambda=\lambda_0,\psi_0}$ (as discussed in Section 4.1.3) we will suppose that these are univariate random variables.

Our objective is to find an expression for $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ in terms of $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \lambda} \rfloor_{\lambda=\lambda_0,\psi_0}$ and $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi} \rfloor_{\lambda=\lambda_0,\psi_0}$ which will allow us to obtain its variance and asymptotic distribution easily.

Now making a Taylor expansion of $\frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0}, \psi_0}$ about $\frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\lambda_0, \psi_0}$ gives

$$\frac{\partial \mathcal{L}_{T}(\psi, \lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_{0}}, \psi_{0}} \approx \frac{\partial \mathcal{L}_{T}(\psi, \lambda)}{\partial \psi} \rfloor_{\lambda_{0}, \psi_{0}} + \frac{\partial^{2} \mathcal{L}_{T}(\psi, \lambda)}{\partial \lambda \partial \psi} \rfloor_{\lambda_{0}, \psi_{0}} (\hat{\lambda}_{\psi_{0}} - \lambda_{0}).$$

Notice that we have used \approx instead of = because we replace the second derivative with its true parameters. Now if the sample size is large enough then we can say that $\frac{\partial^2 \mathcal{L}_T(\psi,\lambda)}{\partial \lambda \partial \psi} \rfloor_{\lambda_0,\psi_0} \approx \mathbb{E}\left(\frac{\partial^2 \mathcal{L}_T(\psi,\lambda)}{\partial \lambda \partial \psi} \rfloor_{\lambda_0,\psi_0}\right)$. To see why this is true consider the case that of iid random variables then

$$\frac{1}{T} \frac{\partial^2 \mathcal{L}_T(\psi, \lambda)}{\partial \lambda \partial \psi} \rfloor_{\lambda_0, \psi_0} = \frac{1}{T} \sum_{t=1}^T \frac{\partial^2 \log f(X_t; \psi, \lambda)}{\partial \lambda \partial \psi} \rfloor_{\lambda_0, \psi_0} \\
\approx \mathbb{E} \left(\frac{\partial^2 \log f(X_t; \psi, \lambda)}{\partial \lambda \partial \psi} \rfloor_{\lambda_0, \psi_0} \right).$$

Therefore we have that

$$\frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0}} \approx \frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\lambda_0, \psi_0} + T \cdot I_{\lambda \psi} (\hat{\lambda}_{\psi_0} - \lambda_0)$$
(6.6)

Hence we have the first part of the decomposition of $\frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \psi} |_{\hat{\lambda}_{\psi_0}}$ into the distribution which is known, now we need find a decomposition of $(\hat{\lambda}_{\psi_0} - \lambda_0)$ into known distributions. We first recall that since $\mathcal{L}_T(\psi_0, \hat{\lambda}_{\psi_0}) = \arg\max_{\lambda} \mathcal{L}_T(\psi_0, \lambda)$ then

$$\frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\hat{\lambda}_{\psi_0}} = 0$$

(as long as the parameter space is large enough and the maximum is not on the boundary). Therefore making a Taylor expansion of $\frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} |_{\hat{\lambda}_{\psi_0}}$ about $\frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} |_{\lambda_0, \psi = \psi_0}$ gives

$$\frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\hat{\lambda}_{\psi_0}} \approx \frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\lambda_0, \psi_0} + \frac{\partial^2 \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda^2} \rfloor_{\lambda_0, \psi_0} (\hat{\lambda}_{\psi_0} - \lambda_0).$$

Again using the same trick as in (6.6) we have

$$\frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\hat{\lambda}_{\psi_0}} \approx \frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\lambda_0, \psi_0} + T \cdot I_{\lambda\lambda}(\hat{\lambda}_{\psi_0} - \lambda_0) = 0.$$

Therefore

$$(\hat{\lambda}_{\psi_0} - \lambda_0) = -\frac{I_{\lambda\lambda}^{-1}}{T} \frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \Big|_{\lambda_0, \psi_0}. \tag{6.7}$$

Therefore substituting (6.6) into (6.7) gives

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

and (6.3).

To prove (6.4) (ie. obtain the asymptototic distribution and limiting variance of $\frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0}}$), we recall that the regular score function satisfies

$$\frac{1}{\sqrt{T}} \frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\lambda_0, \psi_0} = \frac{1}{\sqrt{T}} \left(\begin{array}{c} \frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\lambda_0, \psi_0} \\ \frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\psi_0, \lambda_0} \end{array} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0)).$$

Now by substituting the above into (6.4) we immediately obtain (6.4).

Finally to prove (6.5) we the following decomposition, Taylor expansions and the trick in (6.6) to obtain

$$2\left\{\mathcal{L}_{T}(\hat{\psi}_{T}, \hat{\lambda}_{T}) - \mathcal{L}_{T}(\psi_{0}, \hat{\lambda}_{\psi_{0}})\right\} = 2\left\{\mathcal{L}_{T}(\hat{\psi}_{T}, \hat{\lambda}_{T}) - \mathcal{L}_{T}(\psi_{0}, \lambda_{0})\right\} - 2\left\{\mathcal{L}_{T}(\psi_{0}, \hat{\lambda}_{\psi_{0}}) - \mathcal{L}_{T}(\psi_{0}, \lambda_{0})\right\}$$

$$\approx (\hat{\theta}_{T} - \theta_{0})'I(\theta)(\hat{\theta}_{T} - \theta_{0}) - (\hat{\lambda}_{\psi_{0}} - \lambda_{0})'I_{\lambda\lambda}(\hat{\lambda}_{\psi_{0}} - \lambda_{0}), \tag{6.8}$$

where $\hat{\theta}_T' = (\hat{\psi}, \hat{\lambda})$ (the mle). Now we want to rewrite $(\hat{\lambda}_{\psi_0} - \lambda_0)'$ in terms of $(\hat{\theta}_T - \theta_0)$. We start by recalling that from (6.6) we have

$$(\hat{\lambda}_{\psi_0} - \lambda_0) = -\frac{I_{\lambda\lambda}^{-1}}{T} \frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\lambda_0, \psi_0}.$$

Now we will rewrite $\frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\lambda_0, \psi_0}$ in terms of $(\hat{\theta}_T - \theta_0)$ by using

$$\begin{split} &\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta} \rfloor_{\hat{\theta}_T} \approx \frac{\partial \mathcal{L}_T(\theta)}{\partial \theta} \rfloor_{\theta_0} + T \cdot I(\theta) (\hat{\theta}_T - \theta_0) \\ \Rightarrow &\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta} \rfloor_{\theta_0} \approx -I(\theta) (\hat{\theta}_T - \theta_0). \end{split}$$

Therefore concentrating on the subvector $\frac{\partial \mathcal{L}_T(\theta)}{\partial \lambda} \rfloor_{\psi_0,\lambda_0}$ we see that

$$\frac{\partial \mathcal{L}_T(\theta)}{\partial \lambda} \rfloor_{\psi_0, \lambda_0} \approx I_{\lambda \psi}(\hat{\psi} - \psi_0) + I_{\lambda \lambda}(\hat{\lambda} - \lambda_0). \tag{6.9}$$

Substituting (6.9) into (6.7) gives

$$(\hat{\lambda}_{\psi_0} - \lambda_0) \approx -I_{\lambda\lambda}^{-1} I_{\lambda\psi} (\hat{\psi} - \psi_0) + (\hat{\lambda} - \lambda_0).$$

Finally substituting the above into (6.8) and making lots of cancellations we have

$$2\bigg\{\mathcal{L}_T(\hat{\psi}_T, \hat{\lambda}_T) - \mathcal{L}_T(\psi_0, \hat{\lambda}_{\psi_0})\bigg\} \approx T(\hat{\psi} - \psi_0)'(I_{\psi\psi} - I_{\psi\lambda}I_{\lambda,\lambda}^{-1}I_{\lambda,\psi})(\hat{\psi} - \psi_0).$$

Finally, since

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, I(\theta)^{-1}),$$

by using inversion formulas for block matrices we have that $\sqrt{T}(\hat{\psi} - \psi_0) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, (I_{\psi\psi} - I_{\psi\lambda}I_{\lambda,\lambda}^{-1}I_{\lambda,\psi})^{-1})$, which gives the desired result.

- **Remark 6.1.1** (i) We first make the rather interesting observation. The limiting variance of $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi} \rfloor_{\psi_0,\lambda_0}$ is $I_{\psi\psi}$, whereas the limiting variance of $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ is $(I_{\psi\psi} I_{\psi\lambda}I_{\lambda,\lambda}^{-1}I_{\lambda,\psi})$ and the limiting variance of $\sqrt{T}(\hat{\psi} \psi_0)$ is $(I_{\psi\psi} I_{\psi\lambda}I_{\lambda,\lambda}^{-1}I_{\lambda,\psi})^{-1}$.
- (ii) Look again at the expression

$$\frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0}, \psi_0} \approx \frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\lambda_0, \psi_0} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} \frac{\partial \mathcal{L}_T(\psi_0, \lambda)}{\partial \lambda} \rfloor_{\lambda_0, \psi_0}$$
(6.10)

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

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

Compare this expression with (6.10). We see that in some sense $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ can be treated as the residual (error) of the projection of $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi} \rfloor_{\lambda_0,\psi_0}$ onto $\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda} \rfloor_{lambda_0,\psi_0}$.

This is quite surprising!

We now aim to use the above result. It is immediately clear that (6.5) can be used for both constructing likelihoods and testing. For example, to construct a 95% CI for ψ we can use the mle $\hat{\theta}_T = (\hat{\psi}_T, \hat{\lambda}_T)$ and the profile likelihood and use the 95% CI

$$\left\{\psi; 2\left\{\mathcal{L}_T(\hat{\psi}_T, \hat{\lambda}_T) - \mathcal{L}_T(\psi, \hat{\lambda}_\psi)\right\} \le \chi_p^2(0.95)\right\}.$$

As you can see by profiling out the parameter λ , we have avoided the need to also construct a CI for λ too. This has many advantages, from a practical perspective it reduced the dimension of the parameters.

The log-likelihood ratio test in the presence of nuisance parameters

An application of Theorem 6.1.1 is for nested hypothesis testing, as stated at the beginning of this section. (6.5) can be used to test $H_0: \psi = \psi_0$ against $H_A: \psi \neq \psi_0$ since

$$2\left\{\max_{\psi,\lambda} \mathcal{L}_T(\psi,\lambda) - \max_{\lambda} \mathcal{L}_T(\psi_0,\lambda)\right\} \stackrel{\mathcal{D}}{\to} \chi_p^2.$$

Example 6.1.2 (χ^2 -test for independence) Now it is worth noting that using the Profile likelihood one can derive the chi-squared test for independence (in much the same way that the Pearson goodness of fit test was derived using the log-likelihood ratio test).

Do this as an exercise (see Davison, Example 4.37, page 135).

The score test in the presence of nuisance parameters

We recall that we used Theorem 6.1.1 to obtain the distribution of $2\{\max_{\psi,\lambda} \mathcal{L}_T(\psi,\lambda) - \max_{\lambda} \mathcal{L}_T(\psi_0,\lambda)\}$ under the null, we now motivate an alternative test to test the same hypothesis (which uses the same Theorem). We recall that under the null $H_0: \psi = \psi_0$ the derivative $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \lambda} |_{\hat{\lambda}_{\psi_0},\psi_0} = 0$, but the same is not true of $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi} |_{\hat{\lambda}_{\psi_0},\psi_0}$. However, if the null is true we would expect if $\hat{\lambda}_{\psi_0}$ to be close to the true λ_0 and for $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi} |_{\hat{\lambda}_{\psi_0},\psi_0}$ to be close to zero. Indeed this is what we showed in (6.4), where we showed that under the null

$$\frac{\partial \frac{1}{\sqrt{T}} \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \Big|_{\hat{\lambda}_{\psi_0}} \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, I_{\psi\psi} - I_{\psi\lambda} I_{\lambda, \lambda}^{-1} I_{\lambda, \psi}), \tag{6.11}$$

where $\lambda_{\psi_0} = \arg \max_{\lambda} \mathcal{L}_T(\psi_0, \lambda)$.

Therefore (6.11) suggests an alternative test for $H_0: \psi = \psi_0$ against $H_A: \psi \neq \psi_0$. We can use $\frac{1}{\sqrt{T}} \frac{\partial \mathcal{L}_T(\psi, \lambda)}{\partial \psi} \rfloor_{\hat{\lambda}_{\psi_0}}$ 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.

6.1.3 Examples

Example: An application of profiling to frequency estimation

Question

Suppose that the observations $\{X_t; t = 1, ..., T\}$ satisfy the following nonlinear regression model

$$X_t = A\cos(\omega t) + B\sin(\omega t) + \varepsilon_t$$

where $\{\varepsilon_t\}$ are iid standard normal random variables and $0 < \omega < \pi$. The parameters A, B, and ω are real and unknown.

Some useful identities are given at the end of the question.

- (i) Ignoring constants, obtain the log-likelihood of $\{X_t\}$. Denote this likelihood as $\mathcal{L}_T(A, B, \omega)$.
- (ii) Let

$$S_T(A, B, \omega) = \left(\sum_{t=1}^T X_t^2 - 2\sum_{t=1}^T X_t (A\cos(\omega t) + B\sin(\omega t)) + \frac{1}{2}T(A^2 + B^2)\right).$$

Show that

$$2\mathcal{L}_{T}(A, B, \omega) + \mathcal{S}_{T}(A, B, \omega) = \frac{(A^{2} - B^{2})}{2} \sum_{t=1}^{T} \cos(2\omega) + AB \sum_{t=1}^{T} \sin(2\omega).$$

Thus show that $|\mathcal{L}_T(A, B, \omega) + \frac{1}{2}\mathcal{S}_T(A, B, \omega)| = O(1)$ (ie. the difference does not grow with T).

Since $\mathcal{L}_T(A, B, \omega)$ and $-\frac{1}{2}\mathcal{S}_T(A, B, \omega)$ are asymptotically equivalent, for the rest of this question, $use \frac{-1}{2}\mathcal{S}_T(A, B, \omega)$ instead of the likelihood $\mathcal{L}_T(A, B, \omega)$.

(iii) Obtain the profile likelihood of ω .

(hint: Profile out the parameters A and B, to show that $\hat{\omega}_T = \arg \max_{\omega} |\sum_{t=1}^T X_t \exp(it\omega)|^2$). Suggest, a graphical method for evaluating $\hat{\omega}_T$?

(iv) By using the identity

$$\sum_{t=1}^{T} \exp(i\Omega t) = \begin{cases} \frac{\exp(\frac{1}{2}i(T+1)\Omega)\sin(\frac{1}{2}T\Omega)}{\sin(\frac{1}{2}\Omega)} & 0 < \Omega < 2\pi\\ T & \Omega = 0 \text{ or } 2\pi. \end{cases}$$
(6.12)

show that for $0 < \Omega < 2\pi$ we have

$$\sum_{t=1}^{T} t \cos(\Omega t) = O(T) \qquad \sum_{t=1}^{T} t \sin(\Omega t) = O(T)$$
$$\sum_{t=1}^{T} t^{2} \cos(\Omega t) = O(T^{2}) \qquad \sum_{t=1}^{T} t^{2} \sin(\Omega t) = O(T^{2}).$$

(v) By using the results in part (iv) show that the Fisher Information of $\mathcal{L}_T(A, B, \omega)$ (denoted as $I(A, B, \omega)$) is asymptotically equivalent to

$$2I(A,B,\omega) = E\left(\frac{\partial^2 \mathcal{S}_T}{\partial \omega^2}\right) = \begin{pmatrix} \frac{T}{2} & 0 & \frac{T^2}{2}B + O(T) \\ 0 & \frac{T}{2} & -\frac{T^2}{2}A + O(T) \\ \frac{T^2}{2}B + O(T) & -\frac{T^2}{2}A + O(T) & \frac{T^3}{3}(A^2 + B^2) + O(T^2) \end{pmatrix}.$$

(vi) Derive the asymptotic variance of maximum likelihood estimator, $\hat{\omega}_T$, derived in part (iv).

Comment on the rate of convergence of $\hat{\omega}_T$.

Useful information: In this question the following quantities may be useful:

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

$$(6.13)$$

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

$$\sum_{t=1}^{T} t = \frac{T(T+1)}{2} \qquad \sum_{t=1}^{T} t^2 = \frac{T(T+1)(2T+1)}{6}.$$

Solution

(i) Since $\{\varepsilon_t\}$ are standard normal iid random variables the likelihood is

$$\mathcal{L}_T(A, B, \omega) = -\frac{1}{2} \sum_{t=1}^T (X_t - A\cos(\omega t) - B\sin(\omega t))^2.$$

(ii) It is straightforward to show that

$$\begin{aligned} &-2\mathcal{L}_{T}(A,B,\omega) \\ &= \sum_{t=1}^{T} X_{t}^{2} - 2\sum_{t=1}^{T} X_{t} \left(A\cos(\omega t) + B\sin(\omega t) \right) \\ &+ A^{2} \sum_{t=1}^{T} \cos^{2}(\omega t) + B^{2} \sum_{t=1}^{T} \sin^{2}(\omega t) + 2AB \sum_{t=1}^{T} \sin(\omega t) \cos(\omega t) \\ &= \sum_{t=1}^{T} X_{t}^{2} - 2\sum_{t=1}^{T} X_{t} \left(A\cos(\omega t) + B\sin(\omega t) \right) + \\ &\frac{A^{2}}{2} \sum_{t=1}^{T} (1 + \cos(2\omega)) + \frac{B^{2}}{2} \sum_{t=1}^{T} (1 - \cos(2\omega)) + AB \sum_{t=1}^{T} \sin(2\omega) \\ &= \sum_{t=1}^{T} X_{t}^{2} - 2\sum_{t=1}^{T} X_{t} \left(A\cos(\omega t) + B\sin(\omega t) \right) + \frac{T}{2} (A^{2} + B^{2}) + \\ &\frac{(A^{2} - B^{2})}{2} \sum_{t=1}^{T} \cos(2\omega) + AB \sum_{t=1}^{T} \sin(2\omega) \\ &= \mathcal{S}_{T}(A, B, \omega) + \frac{(A^{2} - B^{2})}{2} \sum_{t=1}^{T} \cos(2\omega) + AB \sum_{t=1}^{T} \sin(2\omega) \end{aligned}$$

Now by using (6.13) we have

$$-2\mathcal{L}_T(A, B, \omega) = \mathcal{S}_T(A, B, \omega) + O(1),$$

as required.

(iii) To obtain the profile likelihood, let us suppose that ω is known, Then the mle of A and B (using $\frac{-1}{2}S_T$) is

$$\hat{A}_T(\omega) = \frac{2}{T} \sum_{t=1}^T X_t \cos(\omega t) \quad \hat{B}_T(\omega) = \frac{2}{T} \sum_{t=1}^T X_t \sin(\omega t).$$

Thus the profile likelihood (using the approximation S_T) is

$$-\frac{1}{2}S_{p}(\omega) = \frac{-1}{2} \left(\sum_{t=1}^{T} X_{t}^{2} - 2 \sum_{t=1}^{T} X_{t} (\hat{A}_{T}(\omega) \cos(\omega t) + \hat{B}(\omega) \sin(\omega t)) + \frac{T}{2} (\hat{A}_{T}(\omega)^{2} + \hat{B}(\omega)^{2}) \right)$$

$$= \frac{-1}{2} \left(\sum_{t=1}^{T} X_{t}^{2} - \frac{T}{2} [\hat{A}_{T}(\omega)^{2} + \hat{B}_{T}(\omega)^{2}] \right).$$

Thus the ω which maximises $-\frac{1}{2}S_p(\omega)$ is the parameter that maximises $\hat{A}_T(\omega)^2 + \hat{B}_T(\omega)^2$. Since $\hat{A}_T(\omega)^2 + \hat{B}_T(\omega)^2 = \frac{1}{2T}|\sum_{t=1}^T X_t \exp(it\omega)|$, we have

$$\hat{\omega}_{T} = \arg \max_{\omega} (-1/2) \mathcal{S}_{p}(\omega) = \arg \max_{\omega} \left(\hat{A}_{T}(\omega)^{2} + \hat{B}_{T}(\omega)^{2} \right)$$
$$= \arg \max_{\omega} \left| \sum_{t=1}^{T} X_{t} \exp(it\omega) \right|^{2},$$

as required.

- (iv) Differentiating both sides of (6.12) with respect to Ω and considering the real and imaginary terms gives $\sum_{t=1}^{T} t \cos(\Omega t) = O(T)$ $\sum_{t=1}^{T} t \sin(\Omega t) = O(T)$. Differentiating both sides of (6.12) twice wrt to Ω gives the second term.
- (v) Differentiating $S_T(A, B, \omega) = \left(\sum_{t=1}^T X_t^2 2\sum_{t=1}^T X_t \left(A\cos(\omega t) + B\sin(\omega t)\right) + \frac{1}{2}T(A^2 + B^2)\right)$ twice wrt to A, B and ω gives

$$\begin{split} \frac{\partial \mathcal{S}_T}{\partial A} &= -2\sum_{t=1}^T X_t \cos(\omega t) + AT \\ \frac{\partial \mathcal{S}_T}{\partial B} &= -2\sum_{t=1}^T X_t \sin(\omega t) + BT \\ \frac{\partial \mathcal{S}_T}{\partial \omega} &= 2\sum_{t=1}^T AX_t t \sin(\omega t) - 2\sum_{t=1}^T BX_t t \cos(\omega t). \end{split}$$

and
$$\frac{\partial^2 S_T}{\partial A^2} = T$$
, $\frac{\partial^2 S_T}{\partial B^2} = T$, $\frac{\partial^2 S_T}{\partial A \partial B} = 0$,
$$\frac{\partial^2 S_T}{\partial \omega \partial A} = 2 \sum_{t=1}^T X_t t \sin(\omega t)$$
$$\frac{\partial^2 S_T}{\partial \omega \partial B} = -2 \sum_{t=1}^T X_t t \cos(\omega t)$$
$$\frac{\partial^2 S_T}{\partial \omega^2} = 2 \sum_{t=1}^T t^2 X_t \left(A \cos(\omega t) + B \sin(\omega t) \right).$$

Now taking expectations of the above and using (v) we have

$$E(\frac{\partial^2 \mathcal{S}_T}{\partial \omega \partial A}) = 2\sum_{t=1}^T t \sin(\omega t) \left(A \cos(\omega t) + B \sin(\omega t) \right)$$

$$= 2B\sum_{t=1}^T t \sin^2(\omega t) + 2\sum_{t=1}^T At \sin(\omega t) \cos(\omega t)$$

$$= B\sum_{t=1}^T t (1 - \cos(2\omega t)) + A\sum_{t=1}^T t \sin(2\omega t) = \frac{T(T+1)}{2}B + O(T) = B\frac{T^2}{2} + O(T).$$

Using a similar argument we can show that $E(\frac{\partial^2 S_T}{\partial \omega \partial B}) = -A \frac{T^2}{2} + O(T)$ and

$$E(\frac{\partial^2 \mathcal{S}_T}{\partial \omega^2}) = 2\sum_{t=1}^T t^2 \left(A\cos(\omega t) + B\sin(\omega t) \right)^2$$
$$= (A^2 + B^2) \frac{T(T+1)(2T+1)}{6} + O(T^2) = (A^2 + B^2)T^3/3 + O(T^2).$$

Since $E(-\nabla^2 \mathcal{L}_T) \approx \frac{1}{2} E(\nabla^2 \mathcal{S}_T)$, this gives the required result.

(vi) Noting that the asymptotic variance for the profile likelihood estimator $\hat{\omega}_T$

$$\left(I_{\omega,\omega} - I_{\omega,(AB)}I_{A,B}^{-1}I_{(BA),\omega}\right)^{-1},$$

by substituting (vi) into the above we have

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

Thus we observe that the asymptotic variance of $\hat{\omega}_T$ is $O(T^{-3})$.

Typically estimators have a variance of order $O(T^{-1})$, so we see that the estimator $\hat{\omega}_T$ variance which converges to zero, much faster. Thus the estimator is extremely good compared with the majority of parameter estimators.

Example: An application of profiling in survival analysis

Question (This question also uses some methods from Survival Analysis which is covered later in this course - see Sections 13.1 and 19.1).

Let T_i denote the survival time of an electrical component. It is known that the regressors x_i influence the survival time T_i . To model the influence the regressors have on the survival time the Cox-proportional hazard model is used with the exponential distribution as the baseline distribution and $\psi(x_i;\beta) = \exp(\beta x_i)$ as the link function. More precisely the survival function of T_i is

$$\mathcal{F}_i(t) = \mathcal{F}_0(t)^{\psi(x_i;\beta)},$$

where $\mathcal{F}_0(t) = \exp(-t/\theta)$. Not all the survival times of the electrical components are observed, and there can arise censoring. Hence we observe $Y_i = \min(T_i, c_i)$, where c_i is the censoring time and δ_i , where δ_i is the indicator variable, where $\delta_i = 0$ denotes censoring of the *i*th component and $\delta_i = 1$ denotes that it is not censored. The parameters β and θ are unknown.

- (i) Derive the log-likelihood of $\{(Y_i, \delta_i)\}$.
- (ii) Compute the profile likelihood of the regression parameters β , profiling out the baseline parameter θ .

Solution

(i) The survivial function and the density are

$$f_i(t) = \psi(x_i; \beta) \left\{ \mathcal{F}_0(t) \right\}^{[\psi(x_i; \beta) - 1]} f_0(t)$$
 and $\mathcal{F}_i(t) = \mathcal{F}_0(t)^{\psi(x_i; \beta)}$.

Hence for this example we have

$$\log f_i(t) = \log \psi(x_i; \beta) - \left[\psi(x_i; \beta) - 1\right] \frac{t}{\theta} - \log \theta - \frac{t}{\theta}$$
$$\log \mathcal{F}_i(t) = -\psi(x_i; \beta) \frac{t}{\theta}.$$

Therefore, the likelihood is

$$\mathcal{L}_{n}(\beta, \theta) = \sum_{i=1}^{n} \delta_{i} \left\{ \log \psi(x_{i}; \beta) + \log f_{0}(T_{i}) + (\psi(x_{i}; \beta) - 1) \log \mathcal{F}_{0}(t) \right\} +$$

$$\sum_{i=1}^{n} (1 - \delta_{i}) \left\{ \psi(x_{i}; \beta) \log \mathcal{F}_{0}(t) \right\}$$

$$= \sum_{i=1}^{n} \delta_{i} \left\{ \log \psi(x_{i}; \beta) - \log \theta \right\} - \sum_{i=1}^{n} \psi(x_{i}; \beta) \frac{T_{i}}{\theta}$$

(ii) Keeping β fixed and differentiating the above with respect to θ and equating to zero gives

$$\frac{\partial \mathcal{L}_n}{\partial \theta} = \sum_{i=1}^n \delta_i \left\{ -\frac{1}{\theta} \right\} + \sum_{i=1} \psi(x_i; \beta) \frac{T_i}{\theta^2}$$

and

$$\hat{\theta}(\beta) = \frac{\sum_{i=1}^{n} \psi(x_i; \beta) T_i}{\sum_{i=1}^{n} \delta_i}.$$

Hence the profile likelihood is

$$\ell_P(\beta) = \sum_{i=1}^n \delta_i \left\{ \log \psi(x_i; \beta) - \log \hat{\theta}(\beta) \right\} - \sum_{i=1}^n \psi(x_i; \beta) \frac{T_i}{\hat{\theta}(\beta)}.$$

Hence to obtain an estimator of β we maximise the above with respect to β .

An application of profiling in semi-parametric regression

We now consider how the profile 'likelihood' (we use inverted commas here because we do not use the likelihood, but least squares instead) can be used in semi-parametric regression. Recently this type of method has been used widely in various semi-parametric models. This section needs a little knowledge of nonparametric regression, which is considered later in this course. Suppose we observe (Y_t, U_t, X_t) where

$$Y_t = \beta X_t + \phi(U_t) + \varepsilon_t,$$

 (Y_t, X_t, U_t) are iid random variables and ϕ is an unknown function. To estimate β , we first profile out $\phi(\cdot)$, which we estimate as if β were known. In other other words, we suppose that β is known and let $Y_t(\beta) = Y_t - \beta X_t$. We then estimate $\phi(\cdot)$ using the classic local least estimator, in other words the $\phi(\cdot)$ which minimises the criterion

$$\hat{\phi}_{\beta}(u) = \arg \min_{a} \sum_{t} W_{b}(u - U_{t})(Y_{t}(\beta) - a)^{2} = \frac{\sum_{t} W_{b}(u - U_{t})Y_{t}(\beta)}{\sum_{t} W_{b}(u - U_{t})}$$

$$= \frac{\sum_{t} W_{b}(u - U_{t})Y_{t}}{\sum_{t} W_{b}(u - U_{t})} - \beta \frac{\sum_{t} W_{b}(u - U_{t})X_{t}}{\sum_{t} W_{b}(u - U_{t})}$$

$$:= G_{b}(u) - \beta H_{b}(u), \qquad (6.14)$$

where

$$G_b(u) = \frac{\sum_t W_b(u - U_t) Y_t}{\sum_t W_b(u - U_t)}$$
 and $H_b(u) = \frac{\sum_t W_b(u - U_t) X_t}{\sum_t W_b(u - U_t)}$.

Thus, given β the estimator of ϕ and the residuals ε_t are $\hat{\phi}_{\beta}(u) = G_b(u) - \beta H_b(u)$ and $Y_t - \beta X_t - \hat{\phi}_{\beta}(U_t)$. Given the estimated residuals $Y_t - \beta X_t - \hat{\phi}_{\beta}(U_t)$ we can now use least squares to estimate coefficient β , where

$$\mathcal{L}_{T}(\beta) = \sum_{t} (Y_{t} - \beta X_{t} - \hat{\phi}_{\beta}(U_{t}))^{2}$$

$$= \sum_{t} (Y_{t} - \beta X_{t} - G_{b}(U_{t}) + \beta H_{b}(U_{t}))^{2}$$

$$= \sum_{t} (Y_{t} - G_{b}(U_{t}) - \beta [X_{t} - H_{b}(U_{t})])^{2}.$$

Therefore, the least squares estimator of β is

$$\hat{\beta}_{b,T} \ = \ \frac{\sum_t [Y_t - G_b(U_t)] [X_t - H_b(U_t)]}{\sum_t [X_t - H_b(U_t)]^2}.$$

Using $\beta_{b,T}$ we can then estimate (6.15). 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{T} rate of convergence.

It should be mentioned that using random regressors U_t are not necessary. It could be that $U_t = \frac{t}{T}$ (on a grid). In this case

$$\hat{\phi}_{\beta}(u) = \arg\min_{a} \sum_{t} W_{b}(u - \frac{t}{T})(Y_{t}(\beta) - a)^{2} = \frac{\sum_{t} W_{b}(u - \frac{t}{T})Y_{t}(\beta)}{\sum_{t} W_{b}(u - \frac{t}{T})}$$

$$= \sum_{t} W_{b}(u - \frac{t}{T})Y_{t} - \beta \sum_{t} W_{b}(u - U_{t})X_{t}$$

$$:= G_{b}(u) - \beta H_{b}(u), \tag{6.15}$$

where

$$G_b(u) = \sum_t W_b(u - \frac{t}{T})Y_t$$
 and $H_b(u) = \sum_t W_b(u - \frac{t}{T})X_t$.

Using the above estimator of $\phi(\cdot)$ we continue as before.