8 Non-regular models

8.1 Estimating the mean on the boundary

There are situations where the parameter to be estimated in a model is exactly on the boundary. In such cases the limiting distribution of the parameter may not be normal (with variance the inverse Fisher information). Even in the case that the parameter is very close to the boundary very large sample sizes are required for normality to hold. In this case alternative techniques are required.

I illustrate one such method for the example below, though it is worth noting that it may be hard to use such methods for more complex situations.

Suppose that $X_i \sim \mathcal{N}(\mu, 1)$, where the mean $\mu$ is unknown. Suppose in addition that it is known that the mean is non-negative hence the parameter space of the mean is $\Theta = [0, \infty)$. In this case $\bar{X}$ can no longer be the MLE because there will be some instances where $\bar{X} < 0$. But it makes no sense estimating $\mu$ with a negative value when $\bar{X}$ is negative. Let us look at the likelihood for this restricted space

$$
\hat{\mu}_T = \arg \max_{\mu \in \Theta} L_T(\mu) = \arg \max_{\mu \in \Theta} \frac{1}{2} \sum_{t=1}^{T} (X_t - \mu)^2.
$$

Due to the convexity of $L_T(\mu)$ with respect to $\mu$ we see that the MLE estimator is

$$\hat{\mu}_T = \begin{cases} 
\bar{X} & \bar{X} \geq 0 \\
0 & \bar{X} < 0.
\end{cases}$$

Hence in this restricted space $\frac{\partial L_T(\mu)}{\partial \mu} |_{\hat{\mu}_T} \neq 0$, and the usual Taylor expansion method cannot be used to derive normality. Indeed we will show that it is not normal.

We recall that $\sqrt{T}(\bar{X} - \mu) \xrightarrow{D} \mathcal{N}(0, (T^{-1}I(\mu))^{-1})$ or equivalently $\frac{1}{\sqrt{T}} \frac{\partial L_T(\mu)}{\partial \mu} |_{\bar{X}} \xrightarrow{D} \mathcal{N}(0, I(\mu))$. Hence if the true parameter $\mu_0 = 0$, then approximately half the time $\bar{X}$ will be less than zero and the other half it will be greater than zero. This means that half the time $\hat{\mu}_T = 0$ and the other half it will be greater than zero. Therefore the distribution function of $\hat{\mu}_T$ is

$$P(\sqrt{T}\hat{\mu}_T \leq x) = P(\hat{\mu}_T = 0 \text{ or } 0 < \sqrt{T}\hat{\mu}_T \leq x) = \begin{cases} 
0 & x \leq 0 \\
1/2 & x = 0 \\
1/2 + P(0 < \sqrt{T}\bar{X} \leq x) & x > 0
\end{cases}.
$$

Now we may want to test the following hypothesis $H_0 : \mu = 0$ against the hypothesis $H_A : \mu > 0$. We would use the log-likelihood ratio $W = 2\{L_T(\hat{\mu}_T) - L_T(0)\}$, but now it is unlikely to be a chi-squared. So we need to derive this distribution.
It can be argued that under the null half the time the likelihood takes the correct value, hence we have

$$2\{\mathcal{L}_T(\hat{\mu}_T) - \mathcal{L}_T(0)\} \xrightarrow{D} \frac{1}{2} + \frac{1}{2} \chi^2_1.$$ 

I am not a big fan of this argument. I prefer to use the following argument

$$\mathcal{L}_T(\hat{\mu}_T) = \frac{-1}{2} \sum_{t=1}^{T} (X_t - \hat{\mu}_T)^2 \quad \mathcal{L}_T(\hat{\mu}_T) = \frac{-1}{2} \sum_{t=1}^{T} X_t^2.$$ 

Therefore

$$2\{\mathcal{L}_T(\hat{\mu}_T) - \mathcal{L}_T(0)\} = \sum_{t=1}^{T} (2\hat{\mu}_T \bar{X} - \hat{\mu}^2_T)$$

Hence we have that

$$P(2\{\mathcal{L}_T(\hat{\mu}_T) - \mathcal{L}_T(0)\} \leq x) = \begin{cases} 0 & x \leq 0 \\ 1/2 & x = 0 \\ 1/2 + 1/2 P(T|\bar{X}|^2 \leq x) & x > 0 \end{cases}$$

Therefore

$$P(2\{\mathcal{L}_T(\hat{\mu}_T) - \mathcal{L}_T(0)\} \leq x) = \frac{1}{2} + \frac{1}{2} P(\chi^2 \leq x).$$

Therefore, suppose I wanted to test the hypothesis $H_0 : \mu = 0$ against the hypothesis $H_A : \mu > 0$ then I would use the above log likelihood ratio test. In other words, evaluate $W = 2\{\mathcal{L}_T(\hat{\mu}_T) - \mathcal{L}_T(0)\}$ and find the $p$ such that

$$\frac{1}{2} + \frac{1}{2} P(\chi^2 \leq W) = 1 - p.$$ 

This is the $p$-value, depending on it we can see whether we are able to reject the null.

**Example 8.1** Question: The survival time of disease $A$ follow an exponential distribution, where the distribution function has the form $f(x) = \lambda^{-1} \exp(-x/\lambda)$. Suppose that it is known that at least one third of all people who have disease $A$ survive for more than 2 years.

(i) Based on the above information derive a lower bound for $\lambda$.

(ii) Suppose that it is known that $\lambda \geq \lambda_0$. What is the maximum likelihood estimator of $\lambda$. 

51
(iii) Derive the sampling properties of maximum likelihood estimator of $\lambda$, for the cases $\lambda = \lambda_0$ and $\lambda > \lambda_0$.

Solution

(i) $P(X > x) = \exp(-x/\lambda)$. Hence we have that $P(X > 2) = \exp(-2/\lambda) > 1/3$, thus $\lambda > 2/\log 3$.

(ii) The likelihood is

$$L_T(\lambda) = - \sum_{t=1}^{T} \left( \log \lambda + \frac{X_t}{\lambda} \right).$$

Thus $\hat{\lambda}_T = \arg \max_{\lambda_0, \infty} L_T(\lambda)$; noting that of the parameter space is not constrained than the maximum arises when $\hat{\lambda}_T = \frac{1}{T} \sum_{t=1}^{T} X_t$. In the constrained space

$$\hat{\lambda}_T = \begin{cases} 
\lambda_0 & \text{if } \bar{X} \leq \lambda_0 \\
\bar{X} & \text{if } \bar{X} > \lambda_0.
\end{cases}$$

(iii) If $\lambda > \lambda_0$, then the true parameter does not lie on the boundary of the parameter space and for a large enough sample we have $\sqrt{T}(\hat{\lambda}_T - \lambda) \xrightarrow{D} \mathcal{N}(0, \text{var}(X_t))$.

On the other hand if $\lambda = \lambda_0$, then we are on the boundary. We know that $\sqrt{T}(\bar{X} - \lambda_0) \xrightarrow{D} \mathcal{N}(0, \text{var}(X_t))$. From this result, we see that for large samples, about there is a 50% chance that $\bar{X} < \lambda_0$ and 50% chance $\bar{X} \geq \lambda_0$. Based on this, it can be argued that the limiting distribution of $\hat{\lambda}_T$ is

$$\sqrt{T}(\hat{\lambda}_T - \lambda_0) \xrightarrow{D} \begin{cases} 
1/2 & \text{for } \hat{\lambda}_T = \lambda_0 \\
(1/2)\mathcal{N}(0, \text{var}(X_t)) & \text{if } \hat{\lambda}_T > \lambda_0.
\end{cases}$$

Example 8.2 (Example 4.39 (page 140) in Davison (2002)) In this example Davison reparameterises the $t$-distribution. It is well known that if the number of degrees of freedom of a $t$-distribution is one, it is the Cauchy distribution, which has extremely thick tails (such that the mean does not exist). At the other extreme, if we let the number of degrees of freedom tend to $\infty$, then the limit is a normal distribution (where all moments exist). In this example, the $t$-distribution is reparameterised as

$$f(y; \mu, \sigma^2, \psi) = \frac{\Gamma\left(\frac{1+\psi^{-1}}{2}\right)\psi^{1/2}}{(\sigma^2 \pi)^{1/2} \Gamma\left(\frac{1}{2}\right)} \left(1 + \frac{\psi(y - \mu)^2}{\sigma^2}\right)^{-(\psi^{-1}+1)/2}$$

It can be shown that $\lim_{\psi \to 1} f(y; \mu, \sigma^2, \psi)$ is a $t$-distribution with one-degree of freedom and at the other end of the spectrum $\lim_{\psi \to 0} f(y; \mu, \sigma^2, \psi)$ is a normal distribution. Thus $0 < \psi \leq 1$, and the above generalisation allows for fractional orders of the $t$-distribution.
In this example it is assumed that the random variables \( \{X_t\} \) have the density \( f(y; \mu, \sigma^2, \psi) \), and our objective is to estimate \( \psi \), when \( \psi \to 0 \), this the true parameter is on the boundary of the parameter space \((0,1]\) (it is just outside it!). Using similar, arguments to those given above, Davison shows that the limiting distribution of the MLE estimator is close to a mixture of distributions (as in the above example).

8.2 Regularity conditions which are not satisfied

8.2.1 The uniform distribution

The standard example where the regularity conditions (mainly Assumption 1.1(ii)) are not satisfied is the uniform distribution

\[
f(x; \theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}
\]

We can see that the likelihood in this case is

\[
L_T(X; \theta) = \prod_{t=1}^{T} \theta^{-1} I(0 < X_t < \theta).
\]

In this case the derivative of \( L_T(X; \theta) \) is not well defined, hence we cannot solve for the derivative. Instead, to obtain the mle we try to reason what the maximum is. We should plot \( L_T(X; \theta) \) against \( \theta \) and place \( X_i \) on the \( \theta \) axis. We can see that if \( \theta < X_i \), then \( L_T \) is zero. Let \( X_{(i)} \) denote the ordered data \( X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(T)} \). We see that for \( \theta = X_{(T)} \), we have \( L_T(X; \theta) = (X_T)^T \), then beyond this point \( L_T(X; \theta) \) decays i.e. \( L_T(X; \theta) = \theta^{-T} \) for \( \theta \geq X_{(T)} \). Hence the maximum of the likelihood is \( \hat{\theta}_T = \max_{1 \leq t \leq T} X_t \).

To investigate the limiting behaviour, we need to consider the likelihood. However, since \( X_t \) are iid we need only consider the density \( f(x; \theta) \). We observe that Assumption 1.1(ii), is not satisfied, since we cannot exchange the integral and derivative

\[
\frac{d}{d\theta} \int_0^{\theta} \frac{1}{\theta} dx \neq \int 0 \cdot dI(0 \leq x \leq \theta) \frac{d}{d\theta} dx,
\]

hence the Cramer-Rao bound no longer necessarily holds etc. And the limit distribution does not necessarily converge to a normal with the inverse of the Fisher information. In fact you cannot use the standard methods of differentiating the likelihood to obtain the sampling properties of the estimator because the derivative at the true value is not well defined. But you can calculate the limiting distribution of \( \hat{\theta}_T = \max_{1 \leq t \leq T} X_t \) (try it).
8.2.2 The shifted exponential

Let us consider the shifted exponential distribution

\[ f(x; \theta, \phi) = \frac{1}{\theta} \exp\left( -\frac{x - \phi}{\theta} \right) \quad x \geq \phi, \theta, \phi > 0. \]

We first observe when \( \phi = 0 \) we have the usual exponential function, \( \phi \) is simply a shift parameter. It is clear that since the support of the distribution function involves the parameter \( \phi \) that the regularity condition (Assumption 1.1(ii)) will not be satisfied (try it and see). This means the Cramer-Rao bound does not exist in this case and the distribution of the mle estimators of the parameters will not be normal with the inverse of the Fisher information as its variance.

The likelihood for this example is

\[ L_T(X; \theta, \phi) = \frac{1}{\theta^T} \prod_{t=1}^T \exp\left( -\frac{(X_t - \phi)}{\theta} \right)I(\phi \leq X_t). \]

We see that we cannot obtain the maximum of \( L_T(X; \theta, \phi) \) by differentiating. Instead let us consider what happens to \( L_T(X; \theta, \phi) \) for different values of \( \phi \). We see that for \( \phi > X_{(1)} \) (smallest value), the likelihood is zero. But at \( \phi = X_{(1)} \), \( L_T(X; \theta, \phi) \) starts to decrease because \( (X_{(t)} - \phi) > (X_{(t)} - X_{(1)}) \), hence the likelihood decreases. Thus the MLE for \( \phi \) is \( \hat{\phi}_T = X_{(1)} \), notice that this estimator is completely independent of \( \theta \). To obtain the mle of \( \theta \), differentiate \( \frac{\partial L_T(X; \theta, \phi)}{\partial \theta} \bigg|_{\hat{\phi}_T = X_{(1)}} \) and equate to zero.

We obtain \( \hat{\theta}_T = \bar{X} - \hat{\phi}_T \). This makes sense because we recall that when \( \phi = 0 \), then the MLE of \( \theta \) is \( \hat{\theta}_T = \bar{X} \).

We now obtain the distribution of \( \hat{\phi}_T - \phi = X_{(1)} - \phi \). To make the calculation easier we observe that \( X_t \) can be rewritten as \( X_t = \phi + E_t \), where \( \{E_t\} \) is a random variables which as a standard exponential distribution \( f(x; 0) = \theta^{-1} \exp(-x/\theta) \). Therefore the distribution function of \( \hat{\phi}_T - \phi = \min E_t \)

\[ P(\hat{\phi}_T - \phi \leq x) = P(\min E_t \leq x) = 1 - P(\min E_t > x) \]

\[ = 1 - [\exp(-x/\theta)]^T. \]

Therefore the density of \( \hat{\phi}_T - \phi \) is \( \frac{1}{T} \exp(-Tx/\theta) \), or in other words an exponential with parameter \( T/\theta \). Hence the mean of \( \hat{\phi}_T - \phi \) is \( \theta/T \) (notices it goes to zero as \( T \to \infty \)) and the variance is \( \theta^2/T^2 \). Standardising we see that the distribution of \( T(\hat{\phi}_T - \phi) \) is exponential with parameter \( \theta^{-1} \) (since the sum of \( T \) iid exponentials with parameter \( \theta^{-1} \) is exponential with parameter \( T\theta^{-1} \)).

Hence we observe that \( \hat{\phi}_T \) is a biased estimator of \( \phi \), but the bias decreases as \( T \to \infty \). Moreover, the variance is quite amazing. Unlike standard estimators where the variance decreases at the rate \( 1/T \), the variance of \( \hat{\phi}_T \) decreases at the rate \( 1/T^2 \).
Example 8.3 Let us suppose that \{X_t\} are iid exponentially distributed random variables with density \( f(x) = \frac{1}{\lambda} \exp(-x/\lambda) \). Suppose that we only observe \{X_t\}, if \( X_t > c \) (else \( X_t \) is not observed).

(i) Show that the sample mean \( \bar{X} = \frac{1}{T} \sum_{t=1}^{T} X_t \) is a biased estimator of \( \lambda \).

(ii) Suppose that \( \lambda \) and \( c \) are unknown, derive the log-likelihood of \( \{X_t\} \) and the maximum likelihood estimators of \( \lambda \) and \( c \).

Solution

(i) The observations are biased and \( \mathbb{E}(\bar{X}) = \mathbb{E}(X_t|X_t > c) \), thus

\[
\mathbb{E}(X_t|X_t > c) = \int_0^{\infty} x f(x)I(X \geq c) \frac{P(X > c)}{P(X > c)} \, dx
\]

\[
= \int_c^{\infty} x f(x)I(X \geq c) \frac{1}{P(X > c)} \, dx = \frac{1}{e^{-c/\lambda}} \int_c^{\infty} x f(x) \, dx
\]

\[
= \frac{\lambda e^{-c/\lambda}(\frac{c}{\lambda} + 1)}{e^{-c/\lambda}} = \lambda + c.
\]

Note that this is quite logical.

(ii) We observe that \( P(X_t = x|X_t > c) = \frac{f(x)I(X \geq c)}{P(X > c)} = \lambda^{-1} \exp(-1/\lambda(X - c))I(X \geq c) \) (this is a shifted exponential).

Based on this the log-likelihood is

\[
\sum_{t=1}^{T} \{ \log f(X_t) + \log I(X_t \geq c) - \log P(X > c) \}
\]

\[
= \sum_{t=1}^{T} \{ -\log \lambda - \frac{1}{\lambda} (X_t - c) + \log I(X_t \geq c) \}.
\]

Hence we want to find the \( \lambda \) and \( c \) which maximises the above.

Differentiating the above with respect to \( \lambda \) gives \( \sum_{t=1}^{T} (X_t - c) = \lambda T \). Thus \( \hat{\lambda}_T = \frac{1}{T} \sum_{t=1}^{T} X_t - c \).

Maximising the above with respect to \( c \) (this cannot be done by differentiation as the derivative of the likelihood with respect to \( c \) does not exist everywhere) gives \( \hat{c}_T = \inf_t X_t \). Substituting this into the estimator of \( \lambda_T \) gives \( \hat{\lambda}_T = \frac{1}{T} \sum_{t=1}^{T} X_t - \hat{c}_T \), thus giving us the MLE estimators.