10 Parametric Inference II: Maximum Likelihood

The most common method for estimating parameters in a parametric model is the *maximum likelihood method.* Let $X_1, \ldots, X_n$ be iid with pdf $f(x;\theta)$. The **likelihood function** is defined by

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

The likelihood function is just the joint density of the data, except that we treat it is a function of the parameter θ.

The *maximum likelihood estimator (mle)*, denoted by $\widehat{\theta}_n$, is the value of θ that maximizes $\mathcal{L}_n(\theta)$. The function $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$ is called the *log-likelihood function.* The maximum of $\ell_n(\theta)$ occurs at the same place as the maximum of $\mathcal{L}_n(\theta)$, so maximizing the log-likelihood leads to the same answer as maximizing the likelihood. Often, it is easier to work with the log-likelihood. Why (or rather, when) is the mle a good estimator? We defer this question until later.

REMARK 10.1 *If we multiplied $\mathcal{L}_n(\theta)$ by any positive constant c (not depending on θ) then this will not change the mle. Hence, we shall often be sloppy about dropping constants in front of the likelihood.*

EXAMPLE 10.1 *Consider flipping a coin n times resulting in data $X_1, \ldots, X_n$ where $X_i \in \{0,1\}$. The probability function for a single toss is $f(x;\theta) = \theta^x(1-\theta)^{1-x}$ for $x = 0,1$. The unknown parameter is θ. Then,*

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\,\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} = \theta^S(1-\theta)^{n-S}$$

where $S = \sum_i X_i$. Hence,

$$\ell_n(\theta) = S\log\theta + (n-S)\log(1-\theta).$$

Take the derivative of $\ell_n(\theta)$, set it equal to 0 to find that $\widehat{\theta}_n = S/n$.

EXAMPLE 10.2 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ so that $\theta = (\mu, \sigma)$ represent the unknown parameters. The likelihood function is*

$$\begin{aligned} \mathcal{L}_n(\mu, \sigma) &= \prod_i \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(X_i - \mu)^2 \right\} \\ &= \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right\} \\ &= \sigma^{-n} \exp\left\{ -\frac{nS^2}{2\sigma^2} \right\} \exp\left\{ -\frac{n(\overline{X} - \mu)^2}{2\sigma^2} \right\} \end{aligned}$$

where $\overline{X} = n^{-1} \sum_i X_i$ is the sample mean and $S^2 = n^{-1} \sum_i (X_i - \overline{X})^2$ is the sample variance. The last equality above follows from the fact that $\sum_i (X_i - \mu)^2 = nS^2 + n(\overline{X} - \mu)^2$ which can be verified by writing $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \overline{X} + \overline{X} - \mu)^2$ and then expanding the square. Thus, the log-likelihood is

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X} - \mu)^2}{2\sigma^2}.$$

Solving the equations

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0$$

we conclude that $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = S$. (It can be verified that these are indeed global maxima of the likelihood.)

EXAMPLE 10.3 *Suppose that $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and that $p \in \{.1, .2, .6\} \equiv A$. Let $S = \sum_i X_i$. The likelihood is $\mathcal{L}(p) = p^S (1-p)^{n-S}$ for $p \in A$ and 0 otherwise. The mle $\widehat{p}$ is the element $p_j \in A$ closest to S/n.*

EXAMPLE 10.4 (A HARD EXAMPLE.) *Here is the one that confuses everyone. Let $X_1, \ldots, X_n \sim Unif(0, \theta)$. Let us find $\mathcal{L}_n(\theta)$ and then $\widehat{\theta}_n$. First, note that*

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

Consider a fixed value of θ. Suppose there were some X_i such that $\theta < X_i$. Then, $f(X_i; \theta) = 0$ and hence $\mathcal{L}_n(\theta) = \prod_i f(X_i; \theta) = 0$. It follows that $\mathcal{L}_n(\theta) = 0$ if any $X_i > \theta$. In other words, $\mathcal{L}_n(\theta) = 0$ if $\theta < X_{\max}$ where $X_{\max} = \max\{X_1, \ldots, X_n\}$. Now consider a $\theta \ge X_{\max}$. For every X_i we then *have that* $f(X_i;\theta) = 1/\theta$ *so that* $\mathcal{L}_n(\theta) = \prod_i f(X_i;\theta) = \theta^{-n}$. *In conclusion we see that*

$$\mathcal{L}_n(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \theta \geq X_{\max} \\ 0 & \theta < X_{\max}. \end{cases}$$

Note that $\mathcal{L}_n(\theta)$ *is strictly decreasing over the interval* $[X_{\max}, \infty)$. *Hence,* $\widehat{\theta}_n = X_{\max}$.

10.1 Properties of Maximum Likelihood Estimators.

The maximum likelihood estimator (mle) $\widehat{\theta}_n$ possesses many properties that make it an appealing choice of estimator. We will now list these properties although, at this point, they will sound mysterious. The properties of the mle are:

(1) It is **consistent**: $\widehat{\theta}_n \xrightarrow{p} \theta_0$ where θ_0 denotes the true value of the parameter θ;

(2) It is **equivariant**: if $\widehat{\theta}_n$ is the mle of θ then $g(\widehat{\theta}_n)$ is the mle of $g(\theta)$;

(3) It is **asymptotically Normal:** $\sqrt{n}(\widehat{\theta} - \theta_0)/\widehat{se} \xrightarrow{d} N(0,1)$ where $\widehat{se}$ can be computed analytically;

(4) It is **asymptotically optimal:** roughly, this means that among all well behaved estimators, the mle has the smallest variance.

(5) The mle is approximately the Bayes estimator. (To be explained later.)

We will spend some time explaining what these properties mean and why they are good things. In sufficiently complicated problems, these properties will no longer hold and the mle will no longer be a good estimator. But for now we focus on the simpler situations where the mle works well. The properties we discuss only hold if the model satisfies certain "regularity conditions." These are essentially smoothness conditions on $f(x;\theta)$. We shall tacitly assume that these conditions hold.

10.2 Consistency of Maximum Likelihood Estimators.

Consistency means that the mle converges in probability to the true value. To prove consistency, we need a definition. If f and g are pdf's, define the

Kullback-Leibler distance between f and g to be

$$D(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx.$$[9]

It can be shown that $D(f,g) \geq 0$ and $D(f,f) = 0$. For any $\theta, \psi \in \Theta$ write $D(\theta, \psi)$ to mean $D(f(x;\,\theta), f(x;\,\psi))$. Assume that $\theta \neq \psi$ implies that $D(\theta, \psi) > 0$.

Let θ_0 denote the true value of θ. The log-likelihood function is $\ell_n(\theta) = \sum_i \log f(X_i;\,\theta)$. Maximizing $\ell_n(\theta)$ is equivalent to maximizing

$$M_n(\theta) = \frac{1}{n}\sum_i \log\frac{f(X_i;\,\theta)}{f(X_i;\,\theta_0)}.$$

By the law of large numbers, $M_n(\theta)$ converges to

$$\begin{aligned}
E_{\theta_0} \log\frac{f(X_i;\,\theta)}{f(X_i;\,\theta_0)} &= \int \log\left(\frac{f(x;\,\theta)}{f(x;\,\theta_0)}\right) f(x;\,\theta_0)dx \\
&= -\int \log\left(\frac{f(x;\,\theta_0)}{f(x;\,\theta)}\right) f(x;\,\theta_0)dx \\
&= -D(\theta_0, \theta).
\end{aligned}$$

Hence, $M_n(\theta) \approx -D(\theta_0, \theta)$ which is maximized at θ_0 since $-D(\theta_0, \theta_0) = 0$ and $-D(\theta_0, \theta_0) < 0$ for $\theta \neq \theta_0$. Hence, we expect that the maximizer will tend to θ_0. To prove this formally, we need more than $M_n(\theta) \xrightarrow{p} -D(\theta_0, \theta)$. We need this convergence to be uniform over θ. We also have to make sure that the function $D(\theta_0, \theta)$ is well behaved. Here are the formal details.

THEOREM 10.1 *Let θ_0 denote the true value of θ and define*

$$M_n(\theta) = \frac{1}{n}\sum_i \log\frac{f(X_i;\,\theta)}{f(X_i;\,\theta_0)}$$

and let $M(\theta) = -D(\theta_0, \theta)$. Suppose that

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0 \tag{6}$$

and that, for every $\epsilon > 0$,

$$\sup_{\theta : |\theta - \theta_0| \geq \epsilon} M(\theta) < M(\theta_0). \tag{7}$$

Let $\widehat{\theta}_n$ denote the mle. Then $\widehat{\theta}_n \xrightarrow{p} \theta_0$.

[9] This is not a distance in the formal sense.

PROOF. See appendix.

10.3 Equivariance of the mle

Let $\tau = g(\theta)$ be a one-to-one function of θ with inverse h so that $\theta = h(\tau)$. Let $\widehat{\theta}$ be the mle of θ so that $\mathcal{L}_n(\widehat{\theta}) > \mathcal{L}_n(\theta)$ for any $\theta \neq \widehat{\theta}$. Let $\widehat{\tau} = g(\widehat{\theta})$. Hence, $\widehat{\theta} = h(\widehat{\tau})$. Let $L(\tau)$ denote the likelihood function for τ. Then, $L(\tau) = \prod_i f(x_i; h(\tau)) = \prod_i f(x_i;\theta) = \mathcal{L}(\theta)$ where $\theta = h(\tau)$. Let τ be any value not equal to $\widehat{\tau}$. Then, $L(\widehat{\tau}) = \mathcal{L}(\widehat{\theta}) > \mathcal{L}(\theta) = L(\tau)$. Therefore, $\widehat{\tau}$ is the mle for τ.

EXAMPLE 10.5 *Let $X_1, \ldots, X_n \sim N(\theta, 1)$. The mle for θ is $\widehat{\theta}_n = \overline{X}_n$. Let $\tau = e^\theta$. Then, the mle for τ is $\widehat{\tau} = e^{\widehat{\theta}} = e^{\overline{X}}$.*

10.4 Asymptotic Normality

It turns out that $\widehat{\theta}_n$ is approximately Normal and we can compute its variance analytically. We need a few more definitions. $s(X;\theta) = \partial \log f(X;\theta)/\partial\theta$. We call $s(X;\theta)$ the *score function*. The *Fisher information* is defined by $I(\theta) = E_\theta s^2(X;\theta) = \int s^2(x;\theta) f(x;\theta)dx$. Here is an important property of the Fisher information which makes it easier to compute.

LEMMA 10.1 *The Fisher information satisfies:*

$$I(\theta) = E\left(-\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right).$$

THEOREM 10.2 (Asymptotic Normality of the MLE.) *Under appropriate regularity conditions,*

$$\frac{(\widehat{\theta}_n - \theta)}{\widehat{se}} \xrightarrow{d} N(0,1)$$

where

$$\widehat{se} = \frac{1}{\sqrt{nI(\widehat{\theta}_n)}}$$

is the estimated standard error.

The proof is in the appendix.

REMARK 10.2 *We could define the Fisher information in terms of the joint distribution* $f(x_1, \ldots, x_n;\ \theta) = \prod_i f(x_i;\ \theta)$. *Denote the Fisher information defined this way as* $I_n(\theta)$. *The standard error in this framework turns out to be* $\sqrt{1/I_n(\theta)}$. *It is easy to show that* $I_n(\theta) = nI(\theta)$. *Hence, we end up with the same formula for the standard error.*

Informally, the theorem says that the distribution of the mle can be approximated with $N(\theta, \widehat{se}^2)$. The theorem allows us to construct an (asymptotic) confidence interval. Let

$$C_n = [\widehat{\theta}_n - z_{\alpha/2}\, \widehat{se},\ \widehat{\theta}_n + z_{\alpha/2}\, \widehat{se}].$$

Then, it follows from the Theorem that

$$P_\theta(\theta \in C_n) \to 1 - \alpha.$$

To see this, let Z denote a standard normal random variable. Then,

$$\begin{aligned} P_\theta(\theta \in C_n) &= P_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\widehat{se} \le \theta \le \widehat{\theta}_n + z_{\alpha/2}\widehat{se}\right) \\ &= P_\theta\left(-z_{\alpha/2} \le \frac{\widehat{\theta}_n - \theta}{\widehat{se}} \le z_{\alpha/2}\right) \\ &\approx P(z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha. \end{aligned}$$

For $\alpha = .05$, $z_{\alpha/2} = 1.96 \approx 2$, so

$$\widehat{\theta}_n \pm 2\, \widehat{se}$$

is an approximate 95 per cent confidence interval.

EXAMPLE 10.6 *Let* $X_1, \ldots, X_n \sim Ber(p)$. *The mle is* $\widehat{p} = \sum_i X_i/n$ *and* $f = p^x(1-p)^{1-x}$, $\log f = x \log p + (1-x)\log(1-p)$, $s = (x/p) - (1-x)/(1-p)$, *and* $-s' = (x/p^2) + (1-x)/(1-p)^2$. *Thus,*

$$I(p) = E(-s') = \frac{p}{p^2} + \frac{(1-p)}{(1-p)^2} = \frac{1}{p(1-p)}.$$

Hence,

$$\widehat{se} = \frac{1}{\sqrt{nI(\widehat{p})}} = \left\{\frac{\widehat{p}(1-\widehat{p})}{n}\right\}^{1/2}.$$

So, $\widehat{p} \approx N(p_0, \widehat{se}^2)$. An approximate 95 per cent confidence interval is

$$\widehat{p} \pm 2 \left\{ \frac{\widehat{p}(1-\widehat{p})}{n} \right\}^{1/2}.$$

Compare this with the Hoeffding interval.

EXAMPLE 10.7 *Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ where σ^2 is known. The score function is $s(X;\theta) = (X-\theta)/\sigma^2$ and $s' = -1/\sigma^2$ so that $I(\theta) = 1/\sigma^2$. The mle is $\widehat{\theta} = \overline{X}$. According to the Theorem, $\overline{X} \approx N(\theta, \sigma^2/n)$. In fact, in this case, the distribution is exact.*

EXAMPLE 10.8 *Let $X_1, \ldots, X_n \sim Poisson(\lambda)$. Then $\widehat{\lambda} = \overline{X}$. In the homework you will calculate that $I(\lambda) = 1/\lambda$, so*

$$\widehat{se} = \frac{1}{\sqrt{nI(\widehat{\lambda})}} = \left\{ \frac{\widehat{\lambda}}{n} \right\}^{1/2}.$$

So, $\widehat{\lambda} = \overline{X} \approx N(\lambda, se^2)$.

So far, we have confined attention to models with a single parameter. More realistically, the model will have several parameters $\theta = (\theta_1, \ldots, \theta_p)$. The above theory extends to this case. We will discuss the extension in the second term.

10.5 Efficiency.

Throughout the course, we will meet many estimators besides maximum likelihood estimators. We will need a way to compare estimators. One way to compare estimators is to use "decision theory" which we will discuss second term. Here, we will compare estimators based on the idea of efficiency.

Let us start with an example. Suppose that $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. The mle is $\widehat{\theta}_n = \overline{X}$. Another reasonable estimator is the sample median $\widetilde{\theta}_n$ which is defined to be the middle of the data points after sorting the numbers. In case n is even, the median is the average of the two middle points. The mle satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

It can be proved that the median satisfies

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \xrightarrow{d} N\left(0, \sigma^2 \frac{\pi}{2}\right).$$

This means that the median converges to the right value but has a larger variance than the mle.

More generally, consider two estimators T_n and U_n and suppose that

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, t^2)$$

and that

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, u^2).$$

We define the asymptotic relative efficiency of U to T by $ARE(U, T) = t/u$. In the Normal example, $ARE(\widetilde{\theta}_n, \widehat{\theta}_n) = \sqrt{2/\pi} = .798$. The interpretation is that if you use the median, you are only using about 80 per cent of the available data.

Under appropriate conditions it can be shown that if $\widehat{\theta}_n$ is the mle and $\widetilde{\theta}_n$ is any other estimator then $ARE(\widetilde{\theta}_n, \widehat{\theta}_n) \leq 1$. This leads us to say that mle is **efficient**. It has the smallest (asymptotic) variance.

WARNING: The above statements are predicated upon the assumed model being correct. If the data are not quite Normal, then the median may be a much better estimator than the mean.

10.6 The Delta Method.

Recall that the mle satisfies

$$\widehat{\theta} \approx N\left(\theta, \frac{1}{nI(\theta)}\right).$$

Suppose that $\tau = g(\theta)$ where g is a smooth function. The maximum likelihood estimator of τ is $\widehat{\tau} = g(\widehat{\theta})$. Now we address the following question: what is the distribution of $\widehat{\tau}$?

To answer this question, write

$$\widehat{\tau} = g(\widehat{\theta}) \approx g(\theta) + (\widehat{\theta} - \theta)g'(\theta) = \tau + (\widehat{\theta} - \theta)g'(\theta)$$

where $\tau = g(\theta)$. Thus,

$$\sqrt{n}(\widehat{\tau} - \tau) \approx \sqrt{n}(\widehat{\theta} - \theta)g'(\theta)$$

and hence

$$\frac{\sqrt{nI(\theta)}(\widehat{\tau} - \tau)}{g'(\theta)} \approx \sqrt{nI(\theta)}(\widehat{\theta} - \theta).$$

But the mle Theorem tells us that the right hand side tends in distribution to a N(0,1). Hence,

$$\frac{\sqrt{nI(\theta)}(\widehat{\tau} - \tau)}{g'(\theta)} \xrightarrow{d} N(0,1)$$

or, in other words,

$$\widehat{\tau} \approx N\left(\tau, \frac{(g'(\theta))^2}{nI(\theta)}\right).$$

The result remains true if we substitute $\widehat{\theta}$ for θ so

$$\widehat{\tau} \approx N\left(\tau, (g'(\widehat{\theta}))^2 se^2(\widehat{\theta})\right).$$

SUMMARY. If $\widehat{\tau} = g(\widehat{\theta})$ then

$$\widehat{\tau} \approx N\left(\tau, se^2(\widehat{\tau})\right)$$

where $se(\widehat{\tau}) = |g'(\widehat{\theta})| se(\widehat{\theta})$.

EXAMPLE 10.9 *Let $X_1, \ldots, X_n \sim Ber(p)$. Let $\psi = \log(p/(1-p))$. Find an approximate 95 per cent confidence interval for ψ. The mle for p is $\widehat{p} = n^{-1}\sum_i X_i$. The Fisher information function is $I(p) = 1/(p(1-p))$ so the standard error of the mle is $se = \{\widehat{p}(1-\widehat{p})/n\}^{1/2}$. Let $\psi = g(p) = \log p/(1-p)$. The mle is $\widehat{\psi} = \log \widehat{p}/(1-\widehat{p})$. Then, $g'(p) = 1/(p/(1-p))$. So, according to the delta method*

$$se(\widehat{\psi}) = |g'(\widehat{p})| se(\widehat{p}) = \frac{1}{\sqrt{n\widehat{p}(1-\widehat{p})}}.$$

An approximate 95 per cent confidence interval is

$$\widehat{\psi} \pm \frac{2}{\sqrt{n\widehat{p}(1-\widehat{p})}}.$$

EXAMPLE 10.10 *Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Suppose that μ is known, σ is unknown and that we want to estimate $\psi = \log \sigma$. The log-likelihood is $\ell(\sigma) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2$. Differentiate and set equal to 0 and conclude that*

$$\widehat{\sigma} = \left\{ \frac{\sum_i (X_i - \mu)^2}{n} \right\}^{1/2}.$$

To get the standard error we need the Fisher information. First,

$$\log f(X; \sigma) = -\log \sigma - \frac{(X-\mu)^2}{2\sigma^2}$$

with second derivative

$$\frac{1}{\sigma^2} - \frac{3(X-\mu)^2}{\sigma^4}$$

and hence

$$I(\sigma) = -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} = \frac{2}{\sigma^2}.$$

Hence,

$$se = \frac{\widehat{\sigma}}{\sqrt{2n}}.$$

Let $\psi = g(\sigma) = \log(\sigma)$. Then, $\widehat{\psi} = \log \widehat{\sigma}$. Since, $g' = 1/\sigma$,

$$se(\widehat{\psi}) = \frac{1}{\widehat{\sigma}} \frac{\widehat{\sigma}}{\sqrt{2n}} = \frac{1}{\sqrt{2n}}$$

and an approximate 95 per cent confidence interval is

$$\widehat{\psi} \pm \frac{2}{\sqrt{2n}}.$$

10.7 Multiparameter Models

These ideas can directly be extended to models with several parameters. Here we describe the extension to two parameters; the extension to more parameters will then be obvious. Let $\theta = (\theta_1, \theta_2)$ and let $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2)$ be the mle. To get the standard errors of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ we proceed as follows. Define the *Fisher Information Matrix* by

$$I(\theta) = \begin{bmatrix} -E_\theta\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta_1^2}\right) & -E_\theta\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta_1 \partial \theta_2}\right) \\ -E_\theta\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta_2 \partial \theta_1}\right) & -E_\theta\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta_2^2}\right) \end{bmatrix}.$$

Let $J(\theta) = I^{-1}(\theta)$ be the inverse of I. The estimated standard errors of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are

$$\widehat{se}(\widehat{\theta}_1) = \sqrt{\frac{J_{11}(\widehat{\theta})}{n}}$$

and

$$\widehat{se}(\widehat{\theta}_2) = \sqrt{\frac{J_{22}(\widehat{\theta})}{n}}$$

where J_{11} is the upper left entry of J and where J_{22} is the lower right entry of J. The estimated covariance between $\widehat{\theta}_1$ and $\widehat{\theta}_2$ is

$$\widehat{cov}(\widehat{\theta}_1, \widehat{\theta}_2) = \sqrt{\frac{J_{12}(\widehat{\theta})}{n}}.$$

Now let $\tau = g(\theta_1, \theta_2)$ be a function of both parameters. Let

$$\nabla g = \begin{pmatrix} \frac{\partial g}{\partial \theta_1} \\ \frac{\partial g}{\partial \theta_2} \end{pmatrix}$$

be the gradient of g. The *multiparameter delta method* says that the estimated standard error of $\widehat{\tau} = g(\widehat{\theta})$ is

$$\widehat{se}(\widehat{\tau}) = \sqrt{\frac{(\nabla g)^T I^{-1} (\nabla g)}{n}}$$

where it is understood that the expression is evaluated at $\theta = \widehat{\theta}$.

EXAMPLE 10.11 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Let $\tau = g(\mu, \sigma) = \sigma/\mu$. In the homework, you will show that*

$$I(\mu, \sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

and

$$I^{-1}(\mu, \sigma) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix}.$$

Now we compute the gradient of g:

$$\nabla g = \begin{pmatrix} -\frac{\sigma}{\mu^2} \\ \frac{1}{\mu} \end{pmatrix}.$$

Thus,

$$\sqrt{\frac{(\nabla g)^T I^{-1} (\nabla g)}{n}} = \frac{1}{\sqrt{n}} \left\{ \frac{1}{\widehat{\mu}^4} + \frac{\widehat{\sigma}^2}{2\widehat{\mu}^2} \right\}^{1/2}.$$

10.8 Parametric Bootstrap

Standard errors and confidence intervals may also be estimated using the bootstrap. There is only one change. In the nonparametric bootstrap, we sampled $X_1^*, \ldots, X_n^*$ from the empirical distribution $\widehat{F}_n$. In the parametric bootstrap we sample instead from $f(x;\ \widehat{\theta})$.

EXAMPLE 10.12 *Recall example 10.11. To get to bootstrap standard error, simulate* $X_1, \ldots, X_n^* \sim N(\widehat{\mu}, \widehat{\sigma}^2)$*, compute* $\widehat{\mu}^* = n^{-1}\sum_i X_i^*$ *and* $\widehat{\sigma}^{2*} = n^{-1}\sum_i (X_i^* - \widehat{\mu}^*)^2$*. Then compute* $\widehat{\tau}^* = g(\widehat{\mu}^*, \widehat{\sigma}^*) = \widehat{\sigma}^*/\widehat{\mu}^*$*. Repeat this many times and use the standard deviation of the* $\widehat{\tau}^*$*'s to estimate* $se(\widehat{\tau})$*. This is much easier than doing the delta method. But the delta method has the advantage that it gives a nice closed form expression for the standard error.*

10.9 Appendix

Proof of Theorem 10.1.

Since $\widehat{\theta}_n$ maximizes $M_n(\theta)$, we have $M_n(\widehat{\theta}_n) \geq M_n(\theta_0)$. Hence,

$$\begin{aligned}
M(\theta_0) - M(\widehat{\theta}_n) &= M_n(\theta_0) - M(\widehat{\theta}_n) + M(\theta_0) - M_n(\theta_0) \\
&\leq M_n(\widehat{\theta}) - M(\widehat{\theta}_n) + M(\theta_0) - M_n(\theta_0) \\
&\leq \sup_\theta |M_n(\theta) - M(\theta)| + M(\theta_0) - M_n(\theta_0) \\
&\xrightarrow{p} 0
\end{aligned}$$

where the last line follows from (6). It follows that, for any $\delta > 0$,

$$\mathrm{P}\left(M(\widehat{\theta}_n) < M(\theta_0) - \delta\right) \to 0.$$

Pick any $\epsilon > 0$. By (7), there exists $\delta > 0$ such that $|\theta - \theta_0| \geq \epsilon$ implies that $M(\theta) < M(\theta_0) - \delta$. Hence,

$$\mathrm{P}(|\widehat{\theta}_n - \theta_0| > \epsilon) \leq \mathrm{P}\left(M(\widehat{\theta}_n) < M(\theta_0) - \delta\right) \to 0.$$

Proof of Theorem 10.2.

LEMMA 10.2 *The score function satisfies*

$$E_\theta\left[s(X;\theta)\right] = 0.$$

PROOF. Note that $1 = \int f(x;\theta)dx$. Differentiate both sides of this equation to conclude that

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int f(x;\theta)dx \\
&= \int \frac{\partial}{\partial\theta} f(x;\theta)dx \\
&= \int \frac{\frac{\partial f(x;\theta)}{\partial\theta}}{f(x;\theta)} f(x;\theta)dx \\
&= \int \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx \\
&= \int s(x;\theta) f(x;\theta)dx = E_\theta s(X;\theta).
\end{aligned}$$

EXAMPLE 10.13 *Let $X \sim N(\theta, 1)$. Then $s(X;\theta) = X - \theta$ and $E_\theta(s) = E_\theta(X) - \theta = \theta - \theta = 0$.*

EXAMPLE 10.14 *Let $X \sim Ber(p)$. Then $\lambda(X;p) = X\log p + (1 - X)\log(1-p)$ so that $s(X;p) = (X/p) - (1-X)/(1-p)$. Again we see that $E(s) = (p/p) - (1-p)/(1-p) = 0$.*

PROOF OF THEOREM. Let $\ell(\theta) = \log \mathcal{L}(\theta)$. Then,

$$0 = \ell'(\widehat{\theta}) \approx \ell'(\theta) + (\widehat{\theta} - \theta)\ell''(\theta).$$

Rearrange the above equation to get $\widehat{\theta} - \theta = -\ell'(\theta)/\ell''(\theta)$ or, in other words,

$$\sqrt{n}(\widehat{\theta} - \theta) = \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} \equiv \frac{\text{TOP}}{\text{BOTTOM}}.$$

Let $Y_i = \partial \log f(X_i;\theta)/\partial\theta$. Recall that $E(Y_i) = 0$ from the property of the score function and also $Var(Y_i) = I(\theta)$. Hence,

$$\text{TOP} = n^{-1/2}\sum_i Y_i = \sqrt{n}\overline{Y} = \sqrt{n}(\overline{Y} - 0) \xrightarrow{d} W$$

where $W \sim N(0, I(\theta))$, by the central limit theorem. Let $A_i = -\partial^2 \log f(X_i;\theta)/\partial\theta^2$. Thus $E(A_i) = I(\theta)$. Then

$$\text{BOTTOM} = \overline{A} \xrightarrow{p} I(\theta)$$

by the law of large numbers. Apply Slutzky's theorem to conclude that

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \frac{W}{I(\theta)} \stackrel{d}{=} N\left(0, \frac{1}{I(\theta)}\right).$$

Assuming that $I(\theta)$ is a continuous function of θ, it follows that $I(\widehat{\theta}_n) \xrightarrow{p} I(\theta)$. Now

$$\begin{aligned} \frac{\widehat{\theta}_n - \theta}{\widehat{se}} &= \sqrt{n} I^{1/2}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta) \\ &= \sqrt{n} I^{1/2}(\theta)(\widehat{\theta}_n - \theta)\left\{\frac{I(\widehat{\theta}_n)}{I(\theta)}\right\}^{1/2}. \end{aligned}$$

The first terms tends in distribution to N(0,1) from the proof of Version 2. The second term tends in probability to 1. The result follows from Slutzky's Theorem.