

Section 8.5. Breakdown of assumptions

- Non-Existence of the MLE
- Multiple Solutions to Maximization Problem
- Multiple Solutions to Score Equations
- Number of Parameters Increase with the Sample Size
- Support of $p(x; \theta)$ depends on θ
- Non-I.I.D. Data

Non-Existence of the MLE

The non-existence of the MLE may occur for all values of \mathbf{x}_n or for only some of them. In general, this is due either to the fact that the parameter space is not compact or that the log-likelihood is discontinuous in θ .

Example 8.1: Suppose that $X \sim \text{Bernoulli}(1/(1 + \exp(\theta)))$, where $\Theta = \mathcal{R}$. If we observe $x = 1$, then $L(\theta; 1) = 1/(1 + \exp(\theta))$. The likelihood function is a decreasing function of θ and the maximum is not attained on Θ . If Θ were closed, i.e., $\Theta = \bar{\mathcal{R}}$, the MLE would be $-\infty$.

Example 8.2: Suppose that $X \sim \text{Normal}(\mu, \sigma^2)$. So, $\theta = (\mu, \sigma^2)$ and $\Theta = \mathcal{R} \times \mathcal{R}^+$. Now, $l(\theta; x) \propto -\log \sigma - \frac{1}{2\sigma^2}(x - \mu)^2$. Take $\mu = x$. Then as $\sigma \rightarrow 0$, $l(\theta; x) \rightarrow +\infty$. So, the MLE does not exist.

Multiple Solutions

One reason for multiple solutions to the maximization problem is non-identification of the parameter θ .

Example 8.3: Suppose that $Y \sim \text{Normal}(X\theta, I)$, where X is an $n \times k$ matrix with rank smaller than k and $\theta \in \Theta \subset R^k$. The density function is

$$p(y; \theta) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(y - X\theta)'(y - X\theta)\right)$$

Since X is not full rank, there exists an infinite number of solutions to $X\theta = 0$. That means that there exists an infinite number of θ 's that generate the same density function. So, θ is not identified.

Furthermore, note that the likelihood is maximized at all values of θ satisfying $X'X\theta = X'y$.

Multiple Roots to the Score Equations

Even though the score equations may have multiple roots for fixed n , we can still use our theorems to show consistency and asymptotic normality. This will work provided that as n gets large there is a unique maximum with large probability.

Example 8.4: Suppose that $\mathbf{X}_n = (X_1, \dots, X_n)$, where the X_i 's are i.i.d. $Cauchy(\theta, 1)$. We assume that θ_0 lies in the interior of a compact set $\Theta \subset R$. So,

$$p(x; \theta) = \frac{1}{\pi(1 + (x - \theta)^2)}$$

So, the log-likelihood for the full sample is

$$l(\theta; \mathbf{x}) = -n \log \pi - \sum_{i=1}^n \log(1 + (x_i - \theta)^2)$$

Note that as $\theta \rightarrow \pm\infty$, $l(\theta; \mathbf{x}) \rightarrow -\infty$.

The score for θ is given by

$$\frac{dl(\theta; \mathbf{x})}{d\theta} = \sum_{i=1}^n \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}$$

As the picture below demonstrates, there can be multiple roots to the score equations.

We can verify the conditions of Theorem 8.2 to show that the MLE is consistent. First, we know that $Q_0(\theta)$ is uniquely maximized at θ_0 since we can show that θ_0 is identified. Does there exist $\theta \neq \theta_0$ so that $p(x; \theta) = p(x; \theta_0)$? If so, then it must be the case that $(x - \theta)^2 = (x - \theta_0)^2$ for all x . This can only happen if $\theta = \theta_0$. Thus, θ_0 is identified. By assumption, we know that Θ is compact. To show continuity of $Q_0(\theta)$ and uniform convergence in probability of $Q(\theta; \mathbf{X}_n)$ to $Q_0(\theta)$, we appeal to the conditions of Lemma 8.3. We have to show that $\log p(x; \theta)$ is continuous in θ for $\theta \in \Theta$ and all $x \in \mathcal{X}$. This function clearly satisfies this continuity condition. Finally, we have to show that there exists a function $d(x)$ such that $|\log p(x; \theta)| \leq d(x)$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$ and $E_{\theta_0}[d(X)] < \infty$.

Note that there exist positive constants $C_1, C_2 > 1$ and C_3 so that

$$\begin{aligned}
 |\log p(x; \theta)| &= |-\log \pi - \log(1 + (x - \theta)^2)| \\
 &= \log \pi + \log(1 + (x - \theta)^2) \\
 &\leq C_1 + \log(C_2 + C_3 x^2) = d(x)
 \end{aligned}$$

It remains to show that $E_{\theta_0}[d(X)] < \infty$. Note that

$$\begin{aligned}
 E_{\theta_0}[d(X)] &= \int_{-\infty}^{\infty} \{C_1 + \log(C_2 + C_3 x^2)\} \frac{1}{\pi(1 + (x - \theta_0)^2)} dx \\
 &= C_1 + \int_{-\infty}^{\infty} \log(C_2 + C_3 x^2) \frac{1}{\pi(1 + (x - \theta_0)^2)} dx \\
 &= C_1 + \int_{-\infty}^{\infty} \log(C_2 + C_3 (x + \theta_0)^2) \frac{1}{\pi(1 + x^2)} dx \\
 &= C_1 + \int_{-\infty}^{-\theta_0} \log(C_2 + C_3 (x + \theta_0)^2) \frac{1}{\pi(1 + x^2)} dx + \\
 &\quad \int_{-\theta_0}^{\infty} \log(C_2 + C_3 (x + \theta_0)^2) \frac{1}{\pi(1 + x^2)} dx
 \end{aligned}$$

Now, $\int_{-\infty}^{-\theta_0} \log(C_2 + C_3(x + \theta_0)^2) \frac{1}{\pi(1+x^2)} dx$ is equal to

$$\int_{x(\theta_0)}^{-\theta_0} \log(C_2 + C_3(x + \theta_0)^2) \frac{1}{\pi(1+x^2)} dx + \int_{-\infty}^{x(\theta_0)} \log(C_2 + C_3(x + \theta_0)^2) \frac{1}{\pi(1+x^2)} dx$$

which is less than

$$\int_{x(\theta_0)}^{-\theta_0} \log(C_2 + C_3(x + \theta_0)^2) \frac{1}{\pi(1+x^2)} dx + \int_{-\infty}^{x(\theta_0)} \frac{\sqrt{|x|}}{\pi(1+x^2)} dx$$

for $x(\theta_0)$ small enough. Both of the integrals in the sum are bounded. Similar arguments can be made for the

$\int_{\theta_0}^{\infty} \log(C_2 + C_3(x + \theta_0)^2) \frac{1}{\pi(1+x^2)} dx$. Thus, we know that $E_{\theta_0}[d(X)] < \infty$.

Number of Parameters Increase with the Sample Size

Up to now, we have implicitly assumed that the number of parameters is equal to a fixed constant k . In some cases the number of parameters increases naturally with the number of observations. In such cases, the MLE may

- i. no longer converge
- ii. may converge to a parameter value different than θ_0
- iii. may still converge to θ_0 .

In general, the outcome depends on the importance of the number of parameters relative to the number of observations.

Example 8.5: (Neyman-Scott, Econometrika, 1948)

Suppose that $\mathbf{X}_n = (X_1, \dots, X_n)$, where the X_i 's are independent with $X_i = (X_{i1}, X_{i2})$, X_{i1} independent of X_{i2} and $X_{ip} \sim N(\mu_i, \sigma^2)$ for $p = 1, 2$. We are interested in estimating the μ_i 's and σ^2 . In this problem, we have $n + 1$ parameters. The likelihood function is

$$L(\mu_1, \dots, \mu_n, \sigma^2; \mathbf{x}_n) = \prod_{i=1}^n \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2} \sum_{p=1}^2 (X_{ip} - \mu_i)^2\right)$$

It is easy to show that the MLE's are

$$\begin{aligned} \hat{\mu}_i &= \frac{1}{2}(X_{i1} + X_{i2}) \text{ for } i = 1, \dots, n \\ \hat{\sigma}^2 &= \frac{1}{2n} \sum_{i=1}^n \sum_{p=1}^2 (X_{ip} - \hat{\mu}_i)^2 \end{aligned}$$

Note that $\hat{\mu}_i$ doesn't converge to μ_i and we can show that $\hat{\sigma}^2$ converges in probability to $\sigma^2/2$. To show this latter fact, note that we can express $\hat{\sigma}^2$ as $\frac{1}{4n} \sum_{i=1}^n (X_{i1} - X_{i2})^2$. Let $Z_i = \frac{X_{i1} - X_{i2}}{\sqrt{2}\sigma}$. Then $Z_i \sim N(0, 1)$ and Z_i^2 is χ_1^2 . Since we have an i.i.d. sample of Z_i^2 's, we can employ the WLLN to show that $\frac{1}{n} \sum_{i=1}^n Z_i^2 \xrightarrow{P} 1$. This implies that

$$\hat{\sigma}^2 = \frac{\sigma^2}{2} \cdot \frac{1}{n} \sum_{i=1}^n Z_i^2 \xrightarrow{P} \frac{\sigma^2}{2}$$

Example 8.6: Suppose that $\mathbf{X}_n = (X_1, \dots, X_n)$, where the X_i 's are independent with $X_i = (X_{i1}, X_{i2}, \dots, X_{in})$, X_{ip} 's are independent $N(\mu_i, \sigma^2)$ random variables for $p = 1, 2, \dots, n$. We are interested in estimating the μ_i 's and σ^2 . Again, we have $n + 1$ parameters. The likelihood function is

$$L(\mu_1, \dots, \mu_n, \sigma^2; \mathbf{x}_n) = \prod_{i=1}^n \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2} \sum_{p=1}^n (X_{ip} - \mu_i)^2\right)$$

It is easy to show that the MLE's are

$$\begin{aligned} \hat{\mu}_i &= \frac{1}{n} \sum_{p=1}^n X_{ip} \text{ for } i = 1, \dots, n \\ \hat{\sigma}^2 &= \frac{1}{n^2} \sum_{i=1}^n \sum_{p=1}^n (X_{ip} - \hat{\mu}_i)^2 \end{aligned}$$

By the WLLN, we know that $\hat{\mu}_i$ converges in probability to μ_i and

we can also show that $\hat{\sigma}^2$ converges in probability to σ^2 .

Support of $p(x; \theta)$ depends on θ

In this case, the MLE is frequently consistent, but not asymptotically normal.

Example 8.7: Suppose $\mathbf{X}_n = (X_1, \dots, X_n)$, where the X_i 's are i.i.d. from a shifted exponential. That is,

$$p(x; \theta) = \exp(-(x - \theta))I(x \geq \theta)$$

Then, the likelihood for the full sample is

$$L(\theta; \mathbf{x}_n) = \exp\left(-\sum_{i=1}^n (x_i - \theta)\right)I(\min x_i \geq \theta)$$

From the plot of likelihood function above, we see that the MLE is $\min X_i$ or the first order statistic $X_{(1)}$. Note that the likelihood is not differentiable at the MLE. This violates condition (iv) of Theorem 8.6.

We can show that the MLE is consistent.

$$\begin{aligned} P_{\theta_0}[|X_{(1)} - \theta_0| > \epsilon] &= P_{\theta_0}[X_{(1)} - \theta_0 > \epsilon] + P_{\theta_0}[X_{(1)} - \theta_0 < -\epsilon] \\ &= P_{\theta_0}[X_{(1)} > \theta_0 + \epsilon] + P_{\theta_0}[X_{(1)} < \theta_0 - \epsilon] \\ &= \prod_{i=1}^n P_{\theta_0}[X_i > \theta_0 + \epsilon] \\ &= \exp(-n\epsilon) \rightarrow 0 \end{aligned}$$

It is obvious that $\sqrt{n}(X_{(1)} - \theta_0)$ cannot be centered at zero since $X_{(1)}$ is always greater than θ_0 . We can show that $n(X_{(1)} - \theta_0) \xrightarrow{D} \text{Exponential}(1)$. To see this, note that

$$\begin{aligned} P_{\theta_0}[n(X_{(1)} - \theta_0) \geq a] &= P_{\theta_0}[X_{(1)} \geq a/n + \theta_0] \\ &= P_{\theta_0}[X_i \geq a/n + \theta_0]^n = \exp(-a) \end{aligned}$$

Here the rate of convergence is n instead of \sqrt{n} .

Non-I.I.D. Data

Example 8.8: Consider independent random variables $Y_i \sim \text{Normal}(\theta x_i, 1)$, where the x_i 's are given constants. The MLE of θ is

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \sim \text{Normal}\left(\theta, 1 / \sum_{i=1}^n x_i^2\right)$$

This estimator may not be consistent. Suppose that $\sum_{i=1}^n x_i^2 \rightarrow 1$. Then, we know that $\hat{\theta} \xrightarrow{D(\theta_0)} N(\theta_0, 1)$, which is not θ_0 .

If $\sum_{i=1}^n x_i^2 \rightarrow \infty$, then $\hat{\theta}$ is consistent. To see this, note that $\hat{\theta}$ is unbiased and its variance goes to zero.

What about the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$? We know that

$$\sqrt{\sum_{i=1}^n x_i^2} (\hat{\theta} - \theta_0) \xrightarrow{D} N(0, 1)$$

If $\sqrt{n}/\sqrt{\sum_{i=1}^n x_i^2} \rightarrow 1$, then $\hat{\theta}$ converges at \sqrt{n} rates. In general, it converges at $\sqrt{\sum_{i=1}^n x_i^2}$ rates.