Advanced Econometrics

Chapter 4: Finite-Sample Properties of the LSE

In Choi

Sogang University

Finite–Sample Properties of the LSE

Reading: Chapter 4 of Greene

- Finite-sample theory : n is assumed to be fixed, normal distribution assumed
- Large-sample theory: n is sent to ∞, general distributions assumed
- It is hard to tell which approach is better. It depends on the situation.

Unbiasedness

- Write

$$\begin{aligned} b &= \left(X'X\right)^{-1}X'y = \left(X'X\right)^{-1}X'\left(X\beta+\varepsilon\right) \\ &= \beta+\left(X'X\right)^{-1}X'\varepsilon. \end{aligned}$$

Then

$$\begin{aligned} E\left(b|X\right) &= \beta+E\left[\left(X'X\right)^{-1}X'\varepsilon|X\right] \\ &= \beta+\left(X'X\right)^{-1}XE\left(\varepsilon|X\right) \\ &= \beta. \end{aligned}$$

Therefore

$$\underset{\substack{\shortparallel \\ \text{center of the} \\ \text{distribution of } b}}{E\left(b\right)} = E\left(E(b|X)\right) = E(\beta) = \underset{\substack{\shortparallel \\ \text{true parameter} \\ \text{vector}}}{\beta.}$$

The variance of the LSE and the Gauss–Markov theorem

- The OLS estimator of β is

$$b = \left(X'X\right)^{-1} X'y.$$

$\left(X'X\right)^{-1} X'$ is an $K \times n$ vector. Thus, each element of b can be written as a linear combination of $y_1, \cdots, y_n$. We call b a linear estimator for this reason.

The variance of the LSE and the Gauss–Markov theorem

- The conditional variance-covariance matrix of b is

$$
\begin{aligned}
Var\left(b|X\right) &= E\left[\left(b-\beta\right)\left(b-\beta\right)'|X\right] \\
&= E\left[\left(X'X\right)^{-1}X'\varepsilon\varepsilon'X\left(X'X\right)^{-1}|X\right] \\
&= \left(X'X\right)^{-1}X'E\left(\varepsilon\varepsilon'|X\right)X\left(X'X\right)^{-1} \\
&= \left(X'X\right)^{-1}X'\left(\sigma^2 I\right)X\left(X'X\right)^{-1} \\
&= \sigma^2\left(X'X\right)^{-1}.
\end{aligned}
$$

The variance of the LSE and the Gauss–Markov theorem

- Consider an arbitrary linear estimator of β, $b_0 = Cy$ where C is a $K \times n$ matrix. For b_0 to be unbiased, we should have

$$\begin{aligned} E(Cy|X) &= E(CX\beta + C\varepsilon|X) \\ &= \beta. \end{aligned}$$

For this to hold, it is required that

$$CX = I,$$

and that C is a function of X.

The variance of the LSE and the Gauss–Markov theorem

- The variance-covariance matrix of b_0 is

$$\begin{aligned} Var\left[b_0|X\right] &= E\left[(b_0-\beta)(b_0-\beta)'|X\right] \\ &= E\left[(Cy-\beta)(Cy-\beta)'|X\right] \\ &= E\left[(CX\beta+C\varepsilon-\beta)(CX\beta+C\varepsilon-\beta)'|X\right] \\ &= E\left[(C\varepsilon)(C\varepsilon)'|X\right] \\ &= E\left[C\varepsilon\varepsilon'C'|X\right] \\ &= \sigma^2 CC'. \end{aligned}$$

The variance of the LSE and the Gauss–Markov theorem

- Now let $D = C - (X'X)^{-1} X'$. Since $CX = I$,

$$\begin{aligned} DX &= CX - (X'X)^{-1} X'X \\ &= CX - I \\ &= 0. \end{aligned}$$

Using this gives

$$\begin{aligned} Var[b_0|X] &= \sigma^2 \left(D + (X'X)^{-1} X'\right) \left(D + (X'X)^{-1} X'\right)' \\ &= \sigma^2 (X'X)^{-1} + \sigma^2 DD' \\ &= Var[b|X] + \sigma^2 DD'. \end{aligned}$$

The variance of the LSE and the Gauss–Markov theorem

- Since DD' is a nonnegative definite matrix,

$$Var\left[b_0|X\right] \geq Var\left[b|X\right]. \tag{1}$$

  That is, for any vector $\mathbf{a} \in R^K - \{0\}$,

$$\mathbf{a}' Var\left[b_0|X\right] \mathbf{a} \geq \mathbf{a}' Var\left[b|X\right] \mathbf{a}.$$

  This is the Gauss–Markov theorem given X.
- Since (1) holds for every particular X,

$$Var\left(b\right) \leq Var\left(b_0\right).$$

  This is the unconditional version of the Gauss–Markov theorem.

- Note that

$$\begin{aligned} Var[b] &= E[Var(b|X)] + Var[E(b|X)] \\ &= E[Var(b|X)] \\ &= E\left[\sigma^2 (X'X)^{-1}\right] \\ &= \sigma^2 E\left[(X'X)^{-1}\right]. \end{aligned}$$

For the first equality, consider the definitional relation

$$Var(y \mid x) = E(y^2 \mid x) - [E(y \mid x)]^2 .$$

- Taking expectations of both sides, we obtain

$$
\begin{aligned}
E\left[Var(y \mid x)\right] &= E(y^2) - E\left[\{E(y \mid x)\}^2\right] \\
&= E(y^2) - \mu_y^2 - \left(E\left[\{E(y \mid x)\}^2\right] - \mu_y^2\right) \\
&= Var(y) - Var\left[E(y \mid x)\right].
\end{aligned}
$$

Estimating the variance of the least squares estimator

- Since $\sigma^2 = E\left(\varepsilon_i^2\right)$, a natural estimator of σ^2 based on the weak law of large numbers is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}e'e.$$

But this estimator is biased as discussed now.
Since $e = My = M\left(X\beta + \varepsilon\right) = M\varepsilon$,

$$e'e = \varepsilon' M\varepsilon.$$

Estimating the variance of the least squares estimator

- Thus

$$\begin{aligned} E\left[e'e|X\right] &= E\left[\varepsilon' M\varepsilon|X\right] \\ &= E\left[tr\left(\varepsilon' M\varepsilon\right)|X\right] \\ &= E\left[tr\left(M\varepsilon\varepsilon'\right)|X\right] \\ &= tr\left(ME\left(\varepsilon\varepsilon'|X\right)\right) \\ &= tr\left(M\sigma^2 I\right) \\ &= \sigma^2 tr\left(M\right). \end{aligned}$$

But

$$\begin{aligned} tr\left(M\right) &= tr\left[I_n - X\left(X'X\right)^{-1}X'\right] \\ &= tr\left(I_n\right) - tr\left(\left(X'X\right)^{-1}X'X\right) \\ &= tr\left(I_n\right) - tr\left(I_K\right) \\ &= n - K. \end{aligned}$$

Estimating the variance of the least squares estimator

- Therefore,

$$E\left[e'e|X\right] = (n-K)\,\sigma^2.$$

and an unbiased estimator of σ^2 is

$$s^2 = \frac{e'e}{n-K},$$

not $\hat{\sigma}^2$.

- The estimator s^2 is also unbiased unconditionally, because

$$E\left[s^2\right] = E\left\{E\left[s^2|X\right]\right\} = E\left(\sigma^2\right) = \sigma^2.$$

Estimating the variance of the least squares estimator

- Using s^2, we obtain an estimator of $Var\left[b|X\right]$

$$\widehat{Var}\left[b|X\right] = s^2\left(X'X\right)^{-1}.$$

The standard error of the estimator b_k is

$$\sqrt{\left[s^2\left(X'X\right)^{-1}\right]_{kk}}$$

Multivariate normal distribution

(i) An $m \times 1$ random vector X is said to have an m-variate normal distribution if, for every $\alpha \in R^m$, the distribution of $\alpha' X$ is univariate normal.

(ii) If $X \sim N(\mu, \Sigma)$, $AX \sim N(A\mu, A\Sigma A')$ and $X + \delta \sim N(\mu + \delta, \Sigma)$.

(iii) If $X \sim N(\mu, \Sigma)$ and X is an $m \times 1$ random vector, its pdf is given as
$f(x) = (2\pi)^{-m/2} (\det \Sigma)^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]$.

- Assume $\varepsilon \sim N\left(0, \sigma^2 I\right)$. Then

$$\begin{aligned} b|X &= \beta + \left(X'X\right)^{-1} X'\varepsilon|X \\ &\sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1} X'X \left(X'X\right)^{-1}\right)_{|X} \\ &= N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)_{|X}. \end{aligned}$$

This implies that each element of $b|X$ is normally distributed. That is,

$$b_k|X \sim N\left(\beta_k, \sigma^2 \left(X'X\right)^{-1}_{kk}\right)_{|X}.$$

Student's t-distribution

$$\frac{N(0,1)}{\sqrt{\chi^2(k)/k}} \sim t(k)$$

when $N(0,1)$ and $\chi^2(k)$ are independent.

- Consider the null hypothesis

$$H_0 : \beta_k = \beta_k^0.$$

The t-test for this null hypothesis is defined by

$$t(b_k) = \frac{b_k - \beta_k^0}{\sqrt{s^2 \left(X'X\right)_{kk}^{-1}}}.$$

This can be written as

$$t(b_k) = \frac{\left(b_k - \beta_k^0\right) / \sqrt{\sigma^2 \left(X'X\right)_{kk}^{-1}}}{\sqrt{(n-K)\frac{s^2}{\sigma^2} / (n-K)}}.$$

- Under the normality assumption,

$$\frac{b_k - \beta_k^0}{\sqrt{\sigma^2 \left(X'X\right)_{kk}^{-1}}} \sim N\left(0,1\right).$$

Note that this is an unconditional distribution.

- In addition

$$\begin{aligned}\frac{\left(n-K\right)s^2}{\sigma^2} &= \frac{e'e}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right) \\ &\sim \chi^2 (tr\left(M\right)) \\ &= \chi^2 (n-K). \end{aligned} \tag{2}$$

- Furthermore,

$$\frac{b-\beta}{\sigma} = \left(X'X\right)^{-1} X' \left(\frac{\varepsilon}{\sigma}\right)$$

is independent of

$$\frac{\left(n-K\right)s^2}{\sigma^2}.$$

- This follows because

$$
\begin{aligned}
Cov\left(e, b|X\right) &= E\left(e\left(b-\beta\right)'|X\right) \\
&= E\left[\left(I-P\right)\varepsilon\varepsilon' X\left(X'X\right)^{-1}|X\right] \\
&= \sigma^2\left(I-P\right)X\left(X'X\right)^{-1} \\
&= 0
\end{aligned}
$$

which implies

$$Cov\left(e, b\right) = 0.$$

Thus, e and b are independent which implies independence of s^2 and b since s^2 is a function of e. (See also Theorem $B-12$). Therefore, $t(b_k)$ follows the Student's t-distribution.

Independence and covariance When X and Y are independent, they have zero covariance. However, zero covariance does not mean independence in general. When X and Y have a multivariate normal distribution, zero covariance implies independence.

- We deduce from the distribution of t_k

$$P\left(b_k - t_{\alpha/2}s_{b_k} \leq \beta_k \leq b_k + t_{\alpha/2}s_{b_k}\right) = 1 - \alpha$$

where $s_{b_k} = \sqrt{s^2\left(X'X\right)^{-1}_{kk}}$ and $t_{\alpha/2}$ is the critical value from the $t-$distribution with $(n-K)$ degrees of freedom. The 100(1-α)% confidence interval for β_k is

$$\left[b_k - t_{\alpha/2}s_{b_k}, b_k + t_{\alpha/2}s_{b_k}\right].$$

- If $X \overset{d}{=} N(0,1)$, $Y \overset{d}{=} \chi^2_k$, then

$$Z = \frac{X+\delta}{\sqrt{Y/k}}$$

has the singly noncentral $t-$distribution with parameters k and δ. The pdf of Z is

$$\begin{aligned} pdf(Z) &= \frac{e^{-\delta^2/2}}{\sqrt{\pi k}\Gamma(k/2)} \left(\frac{k}{k+Z^2}\right)^{(k+1)/2} \sum_{j=0}^{\infty} \frac{\Gamma((k+j+1)/2)}{j!} \\ &\quad \times \left[\frac{\delta Z\sqrt{2}}{\sqrt{k+Z^2}}\right]^j. \end{aligned}$$

- Let the value of β_k under the alternative γ_k and let $\mu = \gamma_k - \beta_k^0$. The power function of t is given as

$$\begin{aligned} Pow\left(\mu, n-K\right) &= prob\left[\left|t\right| > C_\alpha\right] \text{ under } H_A \\ &= 1 - \int_{-C_\alpha}^{C_\alpha} pdf\left(Z\right) dZ. \end{aligned}$$

The power function of t is an increasing function of μ for given $n-K$.

F-distribution

$$\frac{\chi^2(k_1)/k_1}{\chi^2(k_2)/k_2} \sim F(k_1, k_2)$$

where $\chi^2(k_1)$ and $\chi^2(k_2)$ are independent.

- Consider the null hypothesis

$$H_0 : R\beta = r$$

where the $J \times K$ matrix R has full row rank. The F-test for this null is defined as

$$F = (Rb - r)' \left[s^2 R \left(X'X \right)^{-1} R' \right]^{-1} (Rb - r) / J$$

The null distribution of F is $F\left(J, n - K\right)$.

Example

Let $K = 2$ and $H_0 : \beta_1 - \beta_2 = 0$.
Taking $R = \begin{pmatrix} 1 & -1 \end{pmatrix}$ and $r = 0$, we have

$$\begin{aligned} F &= (b_1 - b_2)\left[s^2 \begin{pmatrix} 1 & -1 \end{pmatrix} (X'X)^{-1} \begin{pmatrix} 1 \\ -1 \end{pmatrix}\right]^{-1} (b_1 - b_2) \\ &\sim F(1, n-2). \end{aligned}$$

Example

If $H_0 : \beta_2 = 0, ..., \beta_K = 0$, the F-test is distributed as $F(K-1, n-K)$.
In this case

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}. \tag{3}$$

- Write

$$F = \frac{(Rb - r)' \left[\sigma^2 R \left(X'X\right)^{-1} R'\right]^{-1} (Rb - r) / J}{\left(\frac{(n-K)s^2}{\sigma^2}\right) / (n - K)}.$$

The null distribution follows because

1.
$$Rb - r \sim N\left(0, \sigma^2 R\left(X'X\right)^{-1} R'\right)$$

or

$$\left[R\left(X'X\right)^{-1} R'\right]^{-1/2} (Rb - r) \sim N\left(0, \sigma^2 I_J\right)$$

which implies $(Rb - r)' \left[\sigma^2 R\left(X'X\right)^{-1} R'\right]^{-1} (Rb - r) \sim \chi^2(J)$.
(Note: For a positive-definite $m \times m$ matrix A, there exists a positive-definite $m \times m$ matrix, written as $A^{1/2}$ such that $A^{1/2}A^{1/2} = A$. The inverse of $A^{1/2}$ is written as $A^{-1/2}$. Obviously, $A^{-1/2}AA^{-1/2} = A^{-1/2}A^{1/2}A^{1/2}A^{-1/2} = I_m$.)

2.
$$\frac{(n-K)\,s^2}{\sigma^2} \sim \chi^2_{n-K}.$$

3. b and s^2 are independent.

- If $X \overset{d}{=} N(\mu, I_n)$, $Z = X'X$ has the noncentral χ^2 distribution with n degrees of freedom (written as $\chi^2_n(\delta)$, $\delta = \mu'\mu$). The pdf of Z is written as

$$pdf(z) = e^{-\delta/2} {}_0F_1\left(\frac{1}{2}n; \frac{1}{4}\delta z\right) \frac{1}{2^{n/2}\Gamma(n/2)} e^{-z/2} z^{n/2-1}, \ (z > 0),$$

where

$${}_0F_1(x; y) = \sum_{k=0}^{\infty} \frac{y^k}{(x)_k \, k!},$$

$$(x)_k = x(x+1)\cdots(x+k-1).$$

- If $Z_1 \overset{d}{=} \chi^2_{n_1}(\delta)$, $Z_2 \overset{d}{=} \chi^2_{n_2}$, and Z_1 and Z_2 are independent,

$$G = \frac{Z_1 / n_1}{Z_2 / n_2}$$

has the noncentral F distribution with n_1 and n_2 d.f. and noncentrality parameter δ (denoted as $F_{n_1, n_2}(\delta)$).

- The pdf of G is

$$pdf_G(g) = e^{-\delta/2}{}_1F_1\left(\frac{1}{2}(n_1+n_2);\frac{1}{2}n_1;\frac{\frac{1}{2}\frac{n_1}{n_2}\delta g}{1+\frac{n_1}{n_2}g}\right) \times \frac{\Gamma\left[\frac{1}{2}(n_1+n_2)\right]}{\Gamma\left(\frac{1}{2}n_1\right)\Gamma\left(\frac{1}{2}n_2\right)}\cdot\frac{g^{n_1/2-1}\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\left(1+\frac{n_1}{n_2}g\right)^{(n_1+n_2)/2}},\quad (g>0),$$

where

$${}_1F_1(p;q;r) = \sum_{k=0}^{\infty}\frac{(p)_k}{(q)_k}\frac{r^k}{k!}.$$

- Under the alternative

$$H_A : R\beta = s \ \ (r \neq s),$$

$$\begin{aligned} R\hat{\beta} - r &= R\hat{\beta} - s + (s - r) \\ &\overset{d}{=} N\left(\mu, \sigma^2 \left(X'X\right)^{-1}\right), \ (\mu = s - r) \end{aligned}$$

and hence

$$\left(R\hat{\beta} - r\right)' \left[R\left(X'X\right)^{-1} R'\right]^{-1} \left(R\hat{\beta} - r\right) \equiv \chi^2_q \left(\mu'\mu\right).$$

Thus, it follows that

$$F \overset{d}{=} F_{J,n-K}\left(\mu'\mu\right)$$

under the alternative hypothesis. Using this, we may obtain the power function of F as follows:

$$\begin{aligned} Pow\left(\mu, q, n-K\right) &= prob\left[F > C_\alpha\right] \text{ under } H_A \\ &= 1 - \int_0^{C_\alpha} pdf_G\left(g\right) dg. \end{aligned}$$

We find that the power of the test depends on J, $n-K$ and μ. The power function is an increasing function of $\mu'\mu = \left(s-r\right)'\left(s-r\right)$ for given n, K and J.

Multicollinearity

- If the regressors are perfectly or nearly correlated, the variance of the LSE becomes high to the extent that the regression results look unreliable. This is called the multicollinearity problem. When the regression model is subject to multicollinearity,

1. Small change in the data $\rightarrow$ wide swings in the parameter estimates
2. High standard errors and high R^2.
3. Wrong coefficients signs or implausible magnitudes.

Multicollinearity

- A useful fact related to multicollinearity is

$$Var(b_k \mid X) = \frac{\sigma^2}{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2 (1 - R_k^2)} \tag{4}$$

  where R_k^2 is the R^2 from regressing x_k on all other regressors.
- This relation implies that a larger R_k^2 brings a larger conditional variance for b_k. That is, if the regressor x_k is highly correlated with other regressors, the conditional variance of b_k will be small.
- Multicollinearity is the extreme case where $R_k^2 = 1$. The relation (4) also implies that a larger $\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2$ brings a smaller conditional variance of $\hat{\beta}_k$.

Maximum likelihood estimation

Assume X is a nonstochastic matrix. The likelihood function is written as

$$L\left(\beta, \sigma^2 | X, y\right) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left[-\frac{\left(y - X\beta\right)'\left(y - X\beta\right)}{2\sigma^2}\right].$$

Hence, the log likelihood is

$$\ell\left(\beta, \sigma^2 | X, y\right) = C - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(y - X\beta\right)'\left(y - X\beta\right),$$

where C is a constant.

Maximum likelihood estimation

Minimizing $\ell\left(\beta, \sigma^2 | X, y\right)$ w.r.t. β and σ^2 gives

$$\hat{\beta}_{MLE} = \left(X'X\right)^{-1} X'y$$

and

$$\hat{\sigma}^2_{MLE} = \frac{\left(y - X\hat{\beta}_{MLE}\right)' \left(y - X\hat{\beta}_{MLE}\right)}{n}$$

Note: $\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \left(y - X\beta\right)' \left(y - X\beta\right)$

Properties of MLE

1.

$$\hat{\beta}_{MLE} = \beta + \left(X'X\right)^{-1} X'\varepsilon$$
$$\equiv N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)$$

2. $\hat{\beta}_{MLE}$ and $\hat{\sigma}^2_{MLE}$ are statistically independent, because

$$Cov\left(\hat{\beta}_{MLE}, e\right) = \left(X'X\right)^{-1} X' Cov\left(y, y\right)\left(I - P_X\right) = 0.$$

Properties of MLE

3. Since

$$\frac{\varepsilon' (I - P_X) \varepsilon}{\sigma^2} \sim \chi^2_{(n-K)},$$

it follows that

$$\begin{aligned} E\hat{\sigma}^2_{MLE} &= \frac{1}{n} E\left[\varepsilon' (I - P_X) \varepsilon\right] \\ &= \frac{\sigma^2}{n} E\left[\frac{\varepsilon' (I - P_X) \varepsilon}{\sigma^2}\right] \\ &= \frac{\sigma^2}{n} E\left[\chi^2_{(n-K)}\right] \\ &= \frac{n-K}{n}\sigma^2 \\ &\neq \sigma^2 \text{ (biased)}, \end{aligned}$$

and that

$$Var\left(\hat{\sigma}^2_{MLE}\right) = \frac{2\sigma^4\left(n-K\right)}{n^2}. \tag{5}$$

Properties of MLE

For $s^2 = e'e/(n-K)$, we have

$$Var\left(s^2\right) = \frac{2\sigma^4}{n-K}. \tag{6}$$

Comparing $Var\left(\hat{\sigma}^2_{MLE}\right)$ and $Var\left(s^2\right)$, we find that $Var\left(\hat{\sigma}^2_{MLE}\right) < Var\left(s^2\right)$. But note once again that $\hat{\sigma}^2_{MLE}$ is biased.

Properties of MLE

An estimator of σ^2 which is best in the sense of *MSE* is

$$\tilde{\sigma}^2 = \frac{e'e}{n-K+2}.$$

As $n \longrightarrow \infty$, s^2, $\hat{\sigma}^2_{MLE}$, $\tilde{\sigma}^2 \xrightarrow{p} \sigma^2$, and hence these estimators are essentially the same asymptotically.

Cramér–Rao inequality

The Cramér–Rao inequality gives a lower bound for the variance of an estimator. The estimator for the C–R inequality need not be unbiased, but in most cases we are interested in the C–R inequality for unbiased estimators.

$$
\begin{aligned}
Z: &\quad \text{an } n \times 1 \text{ vector of random variables} \\
L(z,\theta): &\quad \text{the joint density of } Z \text{ where } \theta \text{ is a } K \times 1 \text{ vector of parameters} \\
&\quad \text{(likelihood function for } \theta) \\
\tilde{\theta}: &\quad \text{an unbiased estimator of } \theta
\end{aligned}
$$

Assume

(i) $\int \frac{\partial L}{\partial \theta} dz = 0$; (ii) $\int \frac{\partial^2 L}{\partial \theta \partial \theta'} dz = 0$; (iii) $E\left(\frac{\partial \log L}{\partial \theta} \frac{\partial \log L}{\partial \theta'}\right)_{|\theta=\theta^0} > 0$ (θ^0 : the true parameter vector) and (iv) $\int \frac{\partial L}{\partial \theta} \tilde{\theta}' dz = I$. Then

$$
Var\left(\tilde{\theta}\right) \geq \left[-E\left(\frac{\partial^2 \log L}{\partial \theta \partial \theta'}_{|\theta=\theta^0}\right)\right]^{-1}.
$$

Cramér–Rao inequality

- Condition (i) is equivalent to $E\frac{\partial \log L}{\partial\theta} = 0$, since

$$E\frac{\partial \log L}{\partial\theta} = E\left[\frac{\partial L}{\partial\theta}\frac{1}{L}\right] = \int \frac{\partial L}{\partial\theta}\frac{1}{L}\cdot L dz = 0.$$

- Condition (ii) is also equivalent to $E\frac{\partial \log L}{\partial\theta\partial\theta'} = -E\frac{\partial \log L}{\partial\theta}\frac{\partial \log L}{\partial\theta'}$, since

$$\begin{aligned} E\frac{\partial^2 \log L}{\partial\theta\partial\theta'} &= E\left[\frac{\partial}{\partial\theta}\left(\frac{1}{L}\frac{\partial L}{\partial\theta'}\right)\right] \\ &= E\left(-\frac{1}{L^2}\frac{\partial L}{\partial\theta}\frac{\partial L}{\partial\theta'}\right) + E\left(\frac{1}{L}\frac{\partial^2 L}{\partial\theta\partial\theta'}\right) \\ &= -E\left(\frac{\partial \log L}{\partial\theta}\frac{\partial \log L}{\partial\theta'}\right) + \int \frac{\partial^2 L}{\partial\theta\partial\theta'} dz \end{aligned}$$

Cramér–Rao inequality

- If the integral of the derivative is equal to the derivative of integral, (i) and (ii) follow from $\int L dz = 1$. Hence, the regularity conditions amount to allowing the interchange of integral and derivative signs.
- $\frac{\partial \log L}{\partial \theta}$ is called the score vector, and $E\left(\frac{\partial \log L}{\partial \theta}\frac{\partial \log L}{\partial \theta'}\right)$ the information matrix.

Cramér–Rao inequality

We apply the C–R inequality to the linear regression model

$$y = X\beta + \varepsilon, \;\; \varepsilon \sim N\left(0, \sigma^2 I\right).$$

We find

$$\begin{aligned}
\frac{\partial \log L}{\partial \beta} &= \frac{1}{\sigma^2}\left(X'y - X'X\beta\right) = \frac{1}{\sigma^2}X'\varepsilon \\
\frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\left(y - X\beta\right)'\left(y - X\beta\right) \\
\frac{\partial^2 \log L}{\partial \beta \partial \beta'} &= -\frac{1}{\sigma^2}X'X \\
\frac{\partial^2 \log L}{\left(\partial \sigma^2\right)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\left(y - X\beta\right)'\left(y - X\beta\right) \\
\frac{\partial^2 \log L}{\partial \beta \partial \sigma^2} &= -\frac{1}{\sigma^4}\left(X'y - X'X\beta\right) = -\frac{1}{\sigma^4}X'\varepsilon.
\end{aligned}$$

Cramér–Rao inequality

These give

$$E\left[\frac{\partial \log L}{\partial \theta}\right] = 0, \ \left(\theta = \left[\beta', \sigma^2\right]'\right),$$

$$E\frac{\partial^2 \log L}{\partial \theta \partial \theta'} = -E\frac{\partial \log L}{\partial \theta}\frac{\partial \log L}{\partial \theta'},$$

and

$$I\left(\theta\right) > 0.$$

Cramér–Rao inequality

For condition (iv), note that

$$
\begin{aligned}
\int \frac{\partial L}{\partial \beta} b' dy &= \int \frac{\partial \log L}{\partial \beta} \cdot b' L dy \\
&= \int \frac{(X'y - X'X\beta)}{\sigma^2} \cdot b' L dy \\
&= \frac{1}{\sigma^2} \int (X'y)\, y'X \left(X'X\right)^{-1} L dy - \frac{1}{\sigma^2} X'X\beta \int b' L dy \\
&= \frac{1}{\sigma^2} X' E yy' X \left(X'X\right)^{-1} - \frac{1}{\sigma^2} X'X\beta\beta' \\
&= \frac{1}{\sigma^2} \left[X'X\beta\beta'X'X \left(X'X\right)^{-1} + X'\sigma^2 I X \left(X'X\right)^{-1} \right] \\
&\quad - \frac{1}{\sigma^2} X'X\beta\beta' \\
&= I.
\end{aligned}
$$

Cramér–Rao inequality

In the same way, we may show that

$$\int \frac{\partial L}{\partial \sigma^2} \hat{\sigma}^2 dy = 1. \tag{7}$$

Thus, the lower bound for $Var\left(\hat{\theta}\right)$ is given as

$$Var\left(\hat{\theta}\right) \geq I\left(\theta^0\right)^{-1} = \begin{bmatrix} \sigma^2 \left(X'X\right)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}.$$

$b = \left(X'X\right)^{-1} X'y$ attains its lower bound, but NOT $s^2 = e'e / \left(n - K\right)$.

Exercises

1. Prove relation (2).[1]
2. Prove relation (3).
3. Let $K = 3$ and consider the null hypothesis $H_0 : \beta_2 = \beta_3$. How is the F-statistic for the null hypothesis is formulated? What is its distribution under the normality assumption?
4. Prove relation (4).

[1] Hint: 1. If A is a real, symmetric matrix with eigenvalues $\lambda_1, ..., \lambda_m$, there exists an orthogonal matrix U (i.e., $U'U = UU' = I$) such that $U'AU = diag(\lambda_1, ..., \lambda_m)$. 2. $I - P$ has $n - K$ eigenvalues equal to one and the rest equal to zero.

Exercises

5. Suppose that the regression model is

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

where ε_i are independent and identically distributed given $\{x_i\}$ with its pdf given as $f(\varepsilon_i) = (1/\lambda)\exp(-\lambda\varepsilon_i),\ \varepsilon_i \geq 0$. Note that $E(\varepsilon_i \mid x_1, ...x_n\} = \lambda$. Show that the OLS estimator of β is unbiased and that that of α is biased.

6. The true data generating process (DGP) is $y_i = \beta x_i + \varepsilon_i$, where $\{\varepsilon_i\}$ satisfy standard assumptions of the classical linear regression model. But the model used for estimation is $y_i = \alpha + \beta x_i + \varepsilon_i$. Show that the variance of the OLS estimator of the slope coefficient increases compared to the case where the true model is used.

7. Prove relation (5).

8. Prove relation (6).

9. Prove relation (7).

10. Consider the linear regression model

$$y_t = \alpha + \beta t + \varepsilon_t, \varepsilon_t \sim iid(0, \sigma^2), (t = 1, ..., T).$$

An estimator of β is

$$b = \frac{y_T - y_1}{T - 1}.$$

a. Is estimator b linear and unbiased?
b. Calculate the variance of b. Is the variance of estimator b shrink to zeros as T gets larger?
c. Which estimator should be preferred between b and the corresponding OLS estimator?

Exercises

11. Consider the estimator of β

$$\hat{\beta} = \frac{\sum_{t=1}^{n} x_t y_t}{\frac{\sigma^2}{\beta^2} + \sum_{t=1}^{n} x_t^2}$$

for the standard linear regression model $y_t = \beta x_t + \varepsilon_t$. Show that this estimator is biased and that the mean squared error is

$$E(\hat{\beta} - \beta) = \frac{\sigma^2}{\frac{\sigma^2}{\beta^2} + \sum_{t=1}^{n} x_t^2}.$$

Compare the mean squred error with that of the OLS estimator of β.

12. We are interested in estimating the simple regression model

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i,\ \ \varepsilon_i \sim iid\ N(0, \sigma^2),\ \ i = 1, ..., n.$$

To estimate parameter β_2, the data were divided into two groups $(x_i, y_i)_{i=1}^{n_1}$ and $(x_i, y_i)_{i=n_1+1}^{n}$, where $x_i \leq x_j$ for all $i = 1, ..., n_1$ and all $j = n_1 + 1, ..., n$. Denoting the sample means of these two groups as $(\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_2, \bar{y}_2)$, a researcher wants to estimate the parameter by

$$\bar{\beta}_2 = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}.$$

a. Find the small-sample distribution of $\bar{\beta}_2$.
b. Is $\bar{\beta}_2$ unbiased?
c. Derive the variance of $\bar{\beta}_2$ and compare it to that of the OLS estimator.
d. Suppose that we use an estimator of β_1 defined by $\bar{y} - \bar{\beta}_2 \bar{x}$. How can we formulate the t-ratio for $H_0 : \beta_2 = 0$?