8th Lecture: Maximum Likelihood (ML)

- ML estimation of the model

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

- Define the Likelihood function

$$L(\beta, \sigma^2; y) = \prod_{i=1}^{n} N(y_i | x_i'\beta, \sigma^2) =$$

$$= (2\pi)^{-n/2} (\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\{(y-X\beta)'(y-X\beta)\}}$$

- Usually work with log-Likelihood:

$$\ell(\beta, \sigma^2; y) = \log L(\beta, \sigma^2; y)$$

$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)$$

- Use FOCs to derive MLEs

$$\left.\begin{aligned} \frac{\partial \ell}{\partial \beta} &= 0 \\ \frac{\partial \ell}{\partial \sigma^2} &= 0 \end{aligned}\right\} \Rightarrow \begin{aligned} \hat{\beta}_{ML} &= (X'X)^{-1}X'y \\ \hat{\sigma}^2_{ML} &= \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n} \end{aligned}$$

(Notice that $\hat{\sigma}^2_{ML} \neq \hat{\sigma}$, i.e. $\hat{\sigma}^2_{ML}$ is biased)

Score function: The vector of gradients, i.e, for $\theta = (\beta, \sigma^2)$ the score is the vecto $\frac{\partial \log L}{\partial \theta}$.

Proposition: The expected value of the score function is 0, that is $E\left(\frac{\partial \log L}{\partial \theta}\right) = 0$

Proof

$$E\left(\frac{\partial \log L}{\partial \beta}\right) = E\left(E\left(\frac{\partial \log L}{\partial \beta} \middle| X\right)\right) =$$

$$= E\left(E\left(\frac{1}{2\sigma^2} 2X'\varepsilon \middle| X\right)\right) =$$

$$= E\left(\frac{1}{\sigma^2} X' \underbrace{E(\varepsilon|X)}_{0}\right) = 0$$

$$E\left(\frac{\partial \log L}{\partial \sigma^2}\right) = E\left(E\left(-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \varepsilon_i^2 \middle| X\right)\right) =$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \underbrace{E(\varepsilon_i^2|X)}_{\sigma^2} =$$

$$= -\frac{n}{2\sigma^2} + \frac{n\sigma^2}{2\sigma^4} = 0$$

Information matrix:

$$I(\theta) = -E\left(\frac{\partial^2 \log L}{\partial\theta\,\partial\theta'}\right)$$

* If θ one dimensional

• For $\theta = (\beta, \sigma^2)$ we expect that $I(\theta)$ is 4×4 matrix.

$$I(\theta) = \begin{pmatrix} \frac{\partial^2 \log L}{\partial\theta_1^2} & \frac{\partial^2 \log L}{\partial\theta_1\partial\theta_2} \\ \frac{\partial^2 \log L}{\partial\theta_2\partial\theta_1} & \left(\frac{\partial^2 \log L}{\partial\theta_2^2}\right) \end{pmatrix}$$

i) $$\frac{\partial \log L}{\partial\beta\,\partial\beta'} = \frac{\partial}{\partial\beta'}\left(\frac{\partial \log L}{\partial\beta}\right) =$$

$$= \frac{\partial}{\partial\beta'}\left(\frac{1}{2\sigma^2} 2X'(y - X\beta)\right)$$

$$= -\frac{X'X}{\sigma^2} \Rightarrow$$

$$\Rightarrow E\left(\frac{\partial \log L}{\partial\beta\,\partial\beta}\right) = E\left(E\left(\frac{X'X}{\sigma^2}\,|\,X\right)\right)$$

$$= \frac{X'X}{\sigma^2}$$

ii) $$\frac{\partial^2 \log L}{\partial\beta\,\partial\sigma^2} = \frac{\partial}{\partial\sigma^2}\left(\frac{\partial \log L}{\partial\beta}\right) =$$

$$= \frac{\partial}{\partial\sigma^2}\left(\frac{1}{\sigma^2} X'\varepsilon\right) =$$

$$= -\frac{X'\varepsilon}{\sigma^4}$$

Therefore,

$$E\left(-\frac{\partial^2 \log L}{\partial \beta \partial \sigma^2} \middle| X\right) = \frac{E(X'\varepsilon|X)}{\sigma^4} = \frac{X' E(\varepsilon|X)}{\sigma^4} = 0$$

iii)

$$\frac{\partial^2 \log L}{(\partial \sigma^2)^2} = \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\varepsilon'\varepsilon\right) =$$

$$= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\varepsilon'\varepsilon$$

$$E\left(-\frac{\partial^2 \log L}{(\partial \sigma^2)^2} \middle| X\right) = -E\left(\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\varepsilon'\varepsilon \middle| X\right)$$

$$= -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} E\left(\sum_{i=1}^{n} \varepsilon_i^2 \middle| X\right)$$

$$= -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{i=1}^{n} E(\varepsilon_i^2|X)$$

$$= -\frac{n}{2\sigma^4} + \frac{n\sigma^2}{\sigma^6} =$$

$$= -\frac{n}{2\sigma^4} + \frac{n}{\sigma^4}$$

$$= -\frac{n}{2\sigma^4} + \frac{2n}{2\sigma^4} = \frac{n}{2\sigma^4}$$

Thus,

$$\underset{(k+1)\times(k+1)}{I(\theta)} = \begin{pmatrix} \frac{X'X}{\sigma^2} & 0_{k\times 1} \\ 0_{1\times k} & \frac{n}{2\sigma^4} \end{pmatrix}$$

Variance of Score function:

$$\mathrm{Var}\left(\frac{\partial \log L}{\partial \theta}\right) = E\left[\frac{\partial \log L}{\partial \theta}\left(\frac{\partial \log L}{\partial \theta}\right)'\right] - E\left(\frac{\partial \log L}{\partial \theta}\right) E\left(\frac{\partial \log L}{\partial \theta'}\right)'$$

(* $\mathrm{Var}(x) = E(x-Ex)(x-Ex)' = E\,xx' - Ex\,(Ex)'$)

So $$\mathrm{Var}\,\frac{\partial \log L}{\partial \theta} = -E\left(\frac{\partial^2 \log L}{\partial \theta \partial \theta'}\right) = I(\theta)$$

Remark: We have that $\mathrm{Var}^{(*)}\left(\frac{\partial \log L}{\partial \theta}\right) = -E^{(**)}\left(\frac{\partial^2 \log L}{\partial \theta \partial \theta'}\right)$

because we assume that $\varepsilon \sim N(0, \sigma^2 I)$.

To check this assumption we can measure the diff. between (*) and (**)

- Cramer - Rao Lower Bound (CRLB)

Theorem (Statistics): If $E(\hat{\theta}) = \theta$ then $Var(\hat{\theta}) \geq I(\theta)^{-1}$ and $I(\theta)^{-1}$ is called the CRLB.

- Remark 1: $I(\theta)^{-1} = \begin{pmatrix} \sigma^2 (X'X)^{-1} & 0_{k\times 1} \\ 0_{1\times k} & \frac{2\sigma^4}{n} \end{pmatrix}$

- Remark 2:
Since $\hat{\beta}_{ML} = \hat{\beta}_{OLS}$ we have that $\hat{\beta}_{OLS}$ attains the CRLB. This is stronger than the GM theorem (BLUE) since we remove the linear assumption for the estimator. This is due to the normality assumption.

- Properties of the MLEs $\hat{\theta}$

(* An estimator $\hat{\theta}$ is called asymptotically efficient if it is consistent and asymptotically normal.)

- Let $\hat{\theta}$ the MLE of the unknown θ for a well-behaved likelihood function $L(y;\theta)$ then:

i) $\hat{\theta}$ is consistent $\operatorname{plim}_{n\to\infty} \hat{\theta} = \theta$

ii) $\hat{\theta}$ is asymptotically normal

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \xrightarrow{d} N\left(0, \left(\frac{I(\theta)}{n}\right)^{-1}\right)$$

where $I(\theta) = -E\left(\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right)$

which in practice implies the for large n:

$$\hat{\theta} \overset{a}{\sim} N\left(\theta, \left(I(\theta)\right)^{-1}\right)$$

iii) $\hat{\theta}$ attains the CRLB and is asympt. efficient

iv) Invariance of MLE: the MLE of $g(\theta)$ is $\hat{\theta} = g(\hat{\theta})$.

- Hypothesis testing

Let $\hat{\theta}$ MLE of θ and we wish to test $H_0: g(\theta) = 0$.

In ML approach this test is conducted by using the lik. Ratio test:

We test if the diff $\log L_U - \log L_R$ is large or not.

(*The analogous in OLS estimation is the F-test based on SSR_U and SSR_R → Wald Test)

Alternative approach: Lagrange Multiplier or Rao's Score test:

If H_0 is true then the restricted estimator should be near to the point that max. the log-likelihood

→ Theorem: Let $H_0: g(\theta) = C$ Vs $H_1: g(\theta) \neq C$

and set $\lambda = \frac{L_R}{L_U}$ then

$$-2 \log \lambda = 2(\ln L_U - \ln L_R) \xrightarrow[H_0]{d} \chi^2_{(J)}$$

J number of restrictions.

(Reminder: We have shown the F-test $\sim \chi^2_J$).

(Remarkably) Same asymptotic result for Wald's test can be shown as well as for the Langrage test

---

Generalized least Squares

=

Assume the model

$$y = X\beta + \varepsilon$$ but

$$E(\varepsilon\varepsilon' | X) = \sigma^2 V(X) \neq \sigma^2 I_n$$

↙ symmetric and pos. definite

- Examples for V

i) $V = \begin{pmatrix} h_1 & & \\ & \ddots & \\ & & h_n \end{pmatrix}$ (i.e only diff. variance ⇒ ⇒ no autocorrelated errors

ii) $V = \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ & \ddots & & \\ \rho^{n-1} & \cdots & \rho & 1 \end{pmatrix}$ → autocorrelated errors

- Then $\hat{\beta}_{OLS}$ is not BLUE:

Since $Var(\hat{\beta}_{OLS}|X) =$

$$= E\left((\hat{\beta} - E(\hat{\beta}|X)(..)'\right)$$

$$= E\left((\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X\right)$$

$$= E\left((X'X)^{-1}X'\varepsilon)(X'X)^{-1}X'\varepsilon)|X\right)$$

$$= \sigma^2 (X'X)^{-1} X'VX (X'X)^{-1}$$

(Notice that for $V = I_n \Rightarrow Var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$

Conseq. of not being BLUE:

We cannot assume that

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 (X'X)^{-1}_{jj}} \Rightarrow$$

→ The t-test does not follow the t-distribution and the F-test does not follow the F-distribution. as well as for the Wald test.

→ Solution: By noting the GM theorem requires spherical errors we make suitable transformations to go back to spherical errors.

In particular, let C an $n \times n$ matrix s.t $V^{-1} = C'C$. (e.g. Cholesky factor)

↳ Remark: $CVC' = I_n$

the make the following (linear) transformation:

$$y \to Cy = \tilde{y}$$
$$X \to CX = \tilde{X}$$
$$\varepsilon \to C\varepsilon = \tilde{\varepsilon}$$

$(C')^{-1} V^{-1} = C$
$(C')^{-1} V^{-1} C^{-1} = I_n$
$(CVC')^{-1} = I_n$

That is we now work with the model

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}$$

then, $E(\tilde{\varepsilon}\tilde{\varepsilon}') = \sigma^2 CVC' = \sigma^2 I_n$.

(Also $E(\tilde{\varepsilon}|X) = 0$ still true since $E(\tilde{\varepsilon}|X) = C\,E(\varepsilon|X) = 0$)

- $\hat{\beta}_{GLS}$ is BLUE since the transformed model follows GM conditions

and $\hat{\beta}_{GLS} = (\tilde{X}\tilde{X})^{-1}\tilde{X}'\tilde{y} =$

$= (X'V^{-1}X)^{-1}X'V^{-1}y$

We weight differently each observation to account for the fact that the variance is not constant across observations

Easier to understand when

$$V = \begin{pmatrix} h_1 & & \\ & \ddots & \\ & & h_n \end{pmatrix}$$

Where $\hat{\beta}_{GLS}$ is known as Weighted Least Squares estimator

- $Var(\hat{\beta}_{GLS}|X) = \sigma^2(\tilde{X}'\tilde{X})^{-1} =$

$= \sigma^2(X'V^{-1}X)^{-1}$

In case we had time:

- FOCs for GLS
- MLEs
- Finite Sample Properties
- Asymptotic Properties
- Hypothesis testing
- Explore different forms of Heteroskedasticity.