Lecture Note 1

2014-03-18

1.1. Classical Linear Regression

Definition $\boldsymbol{y}$ is a dependent variable.

$$\boldsymbol{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{y} \in \mathbb{R}^n,\ y_i \in \mathbb{R}\ \ {}^{\forall} i \in \{1,2,\cdots,n\}$$

Definition $\boldsymbol{X}$ is an independent variable.

$$\begin{aligned} \boldsymbol{X} &:= \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix} \qquad \boldsymbol{X} \in \mathbb{R}^{n \times k},\ x_{ij} \in \mathbb{R}\ \ {}^{\forall} j \in \{1,2,\cdots,k\} \\ &= \begin{pmatrix} \mathbf{1} & \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_k \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{x}_{,1} \\ \boldsymbol{x}_{,2} \\ \vdots \\ \boldsymbol{x}_{,n} \end{pmatrix} \end{aligned}$$

Definition $\boldsymbol{\beta}$ is a parameter.

$$\boldsymbol{\beta} := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \boldsymbol{\beta} \in \mathbb{R}^{k+1}$$

Definition $\tilde{\boldsymbol{\varepsilon}}$ is a residual.

$$\begin{aligned} \tilde{\boldsymbol{\varepsilon}} &:= \begin{pmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \\ \vdots \\ \tilde{\varepsilon}_n \end{pmatrix} \\ \Rightarrow \tilde{\boldsymbol{y}} &= \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}} \end{aligned}$$

Since $\tilde{\boldsymbol{\varepsilon}}$ is stochastic, $\boldsymbol{y}$ is also stochastic. However, $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are not random.

1.1.1. Least Squares

However, in reality, $\boldsymbol{y}$ and $\boldsymbol{X}$ are given, while $\boldsymbol{\beta}$ is unknown. Therefore, we should estimate it based on several assumptions; *Least Squares Estimation* is one way to estimate it.

Objective function of the estimation can be denoted as below. By minimizing the function, we can estimate $\boldsymbol{\beta}$, which is indeed unknown in reality.

$$\begin{aligned} & \min_{\tilde{\boldsymbol{\beta}}} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}}\right)^2 \\ = \quad & \min_{\tilde{\boldsymbol{\beta}}} \tilde{\boldsymbol{\varepsilon}}' \tilde{\boldsymbol{\varepsilon}} \end{aligned}$$

Classical linear regression equation can be depicted as below.

$$\begin{aligned} \tilde{\boldsymbol{y}} &= \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}} \\ \Rightarrow \; E(\tilde{\boldsymbol{y}}|\boldsymbol{X}) &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \qquad \text{(conditional regression } \boldsymbol{y} \text{ on } \boldsymbol{X}\text{)} \end{aligned}$$

Assumption 1 $\quad E(\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X}) = \mathbf{0} \;\Leftrightarrow\; E(\tilde{\varepsilon}_i|\boldsymbol{X}) = 0 \;\; {}^{\forall} i$

Assumption 2 $\quad E(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}'|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n \;\Leftrightarrow\; E(\tilde{\varepsilon}_i^2|\boldsymbol{X}) = \sigma^2$ and $E(\tilde{\varepsilon}_i \tilde{\varepsilon}_j|\boldsymbol{X}) = 0 \;\; {}^{\forall} i \neq j$

$$E(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}'|\boldsymbol{X}) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

Assumption 2 is called as *Homoskedasticity* assumption, which is frequently violated in reality. Hypothesis testing based on OLS estimator will not be valid if this assumption is violated.

Assumption 3 $\quad \operatorname{rank}(\boldsymbol{X}) = k+1$

In terms of econometrics, following statements are equivalent. $\boldsymbol{X}$ is full rank; determinants of $\boldsymbol{X}$ are not zero; $\boldsymbol{X}$ is non-singular; ${}^{\exists} \boldsymbol{X}^{-1}$; columns of $\boldsymbol{X}$ are linearly independent.

Assumption 4 $\quad \tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$

Assumption 4 is called as *Normality* assumption, which is not critical, but convenient; hence, we will use this assumption.

1.1.2. Ordinary Least Squares Estimation

Then, we can estimate $\boldsymbol{\beta}$ by solving the minimization problem above.

$$\hat{\boldsymbol{\beta}} := \operatorname*{argmin}_{\tilde{\boldsymbol{\beta}}} \tilde{\boldsymbol{\varepsilon}}' \tilde{\boldsymbol{\varepsilon}}$$

The problem can be solved by using differentiation.

$$\tilde{\boldsymbol{\varepsilon}}' \tilde{\boldsymbol{\varepsilon}} = (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})' (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$$

$$\begin{aligned} &= \boldsymbol{y}'\boldsymbol{y} - \tilde{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\tilde{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \tilde{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} \end{aligned}$$

$$\begin{aligned} \Rightarrow \frac{\partial \tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}}}{\partial \tilde{\boldsymbol{\beta}}} &= -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{0} \end{aligned}$$

$$\Rightarrow \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$\begin{aligned} \therefore \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}) \\ &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{\varepsilon}} \end{aligned}$$

Note that the term *estimator* means function, while the term *estimate* means numeric.

1.1.3. Properties of OLS Estimator

Property 1 $E(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta}$; $\hat{\boldsymbol{\beta}}$ is unbiased.

Proof

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{\varepsilon}}$$

$$\begin{aligned} \therefore E(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= E[\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X}] \\ &= E(\boldsymbol{\beta}|\boldsymbol{X}) + E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X}] \\ &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X}) \qquad \text{Assumption 1} \\ &= \boldsymbol{\beta} \qquad \blacksquare \end{aligned}$$

Property 2 $\hat{\boldsymbol{\beta}}$ is efficient.

In terms of econometrics, following statements are equivalent. $\hat{\boldsymbol{\beta}}$ is the *best* estimator; $\hat{\boldsymbol{\beta}}$ has the smallest variance; $\hat{\boldsymbol{\beta}}$ is the most efficient estimator.

If an estimator has large variance, we cannot say whether the estimator is significant or not. With Assumption 4, it is well-known that the OLS estimator $\hat{\boldsymbol{\beta}}$ is equal to the ML estimator $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, which has the smallest variance. Thus, $\hat{\boldsymbol{\beta}}$ is efficient.

Since the proof requires too tedious mathematics, we will skip the proof.

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\boldsymbol{X}] \\ &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}] \qquad \text{Assumption 2} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

Property 3 $\operatorname{plim}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$; $\hat{\boldsymbol{\beta}}$ is consistent.

Proof

$$\begin{aligned}
\frac{\boldsymbol{X}'\boldsymbol{X}}{n} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' \\
&= \hat{E}(\boldsymbol{x}_i\boldsymbol{x}_i') \qquad \text{as } n\to\infty
\end{aligned}$$

$$\begin{aligned}
\frac{\boldsymbol{X}'\tilde{\boldsymbol{\varepsilon}}}{n} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\tilde{\varepsilon}_i \\
&= \hat{E}(\boldsymbol{x}_i\tilde{\varepsilon}_i) \qquad \text{as } n\to\infty \\
&= \boldsymbol{x}_i\hat{E}(\tilde{\varepsilon}_i) \\
&= \boldsymbol{0}
\end{aligned}$$

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{\varepsilon}} \\
&= \boldsymbol{\beta}+\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\frac{\boldsymbol{X}'\tilde{\boldsymbol{\varepsilon}}}{n} \\
&= \boldsymbol{\beta} \qquad \text{as } n\to\infty
\end{aligned}$$

$$\begin{aligned}
Var(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \frac{\sigma^2}{n}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1} \\
&= 0\times\hat{E}(\boldsymbol{x}_i\boldsymbol{x}_i') \qquad \text{as } n\to\infty \\
&= \boldsymbol{0} \qquad \blacksquare
\end{aligned}$$

Based on these properties, OLS estimator is BLUE; *Best*, *Linear*, *Unbiased* and *Efficient*. By Gauss-Markov Theorem, it can be proved that OLS estimator is BLUE. Note that Lucas, one of the prominent economist, said, "OLS is everything!"

1.2. Hypothesis Testing

1.2.1. Test Statistic

We can conduct *Hypothesis Testing* with the estimator already estimated above. The null hypothesis H_0 can be denoted as below.

$$H_0:\ \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{c} \qquad \boldsymbol{R}\in\mathbb{R}^{q\times(k+1)},\ \boldsymbol{c}\in\mathbb{R}^q$$

In above expression, q is the number of parameters to be tested. Then, the test statistic for the null hypothesis can be written as below.

$$t = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c}}{\sqrt{Var(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c})}}$$

Above t will be utilized in the process of hypothesis testing. Under the null hypothesis,

above t statistic will follow t distribution, which is similar to Normal distribution. Thus, if the absolute value of calculated t statistic is large, then we can reject the null hypothesis under a given significance level; we can reject the null hypothesis if the absolute value of $\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c}$ is significantly big or if $\sqrt{Var(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c})}$ is significantly small.

$$\begin{aligned}
\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c} &\sim N[0, Var(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{c})] \qquad \text{under } H_0 \\
\because\ \boldsymbol{R}\boldsymbol{\beta} &= \boldsymbol{c} \\
\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\tilde{\boldsymbol{\varepsilon}} \\
\Rightarrow\ \hat{\boldsymbol{\beta}} &\sim N[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}] \qquad \text{under } H_0 \\
\therefore\ Var(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c}) &= Var[\boldsymbol{R}(\boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\tilde{\boldsymbol{\varepsilon}})-\boldsymbol{c}] \\
&= Var[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{\varepsilon}}] \\
&= \sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' \qquad \text{under } H_0
\end{aligned}$$

The problem is that σ^2 is unknown in reality. Instead of σ^2, we adopt $\hat{\sigma}^2$ as an alternative.

1.2.2. Estimation of Residual Variance

$$\begin{aligned}
\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} &= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
\therefore\ \boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} \qquad \text{where } \boldsymbol{M}_{\boldsymbol{X}} = [\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'] \\
&= \boldsymbol{M}_{\boldsymbol{X}}\tilde{\boldsymbol{\varepsilon}}
\end{aligned}$$

By Spectral Decomposition, $\boldsymbol{M}_{\boldsymbol{X}}$ can be decomposed as below.

$$\begin{aligned}
\boldsymbol{M}_{\boldsymbol{X}} &= \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}' \qquad \ni\ \boldsymbol{V}'\boldsymbol{V}=\boldsymbol{V}\boldsymbol{V}'=\boldsymbol{I}_n \\
\boldsymbol{V} &= \begin{pmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_n \end{pmatrix} \\
\boldsymbol{\Lambda} &= \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{I}_{n-(k+1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}
\end{aligned}$$

Therefore, we can find a property of its *Eigenvalue*.

$$\begin{aligned}
\lambda\boldsymbol{V} &= \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{V} \\
&= \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{V} \\
&= \boldsymbol{M}_{\boldsymbol{X}}(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{V})
\end{aligned}$$

$$\begin{aligned} &= \boldsymbol{M_X}(\lambda \boldsymbol{V}) \\ &= \lambda \boldsymbol{M_X V} \\ &= \lambda^2 \boldsymbol{V} \end{aligned}$$

$$\therefore \ \lambda = 0 \text{ or } 1$$

Hereafter, we will examine the distributional characteristics of $\hat{\boldsymbol{\varepsilon}}$ with the spectral decomposition of $\boldsymbol{M_X}$ mentioned above.

$$\begin{aligned} (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) &= \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\ &= \left(\boldsymbol{M_X}\tilde{\boldsymbol{\varepsilon}}\right)'\boldsymbol{M_X}\tilde{\boldsymbol{\varepsilon}} \\ &= \tilde{\boldsymbol{\varepsilon}}'\boldsymbol{M_X}\tilde{\boldsymbol{\varepsilon}} \\ &= \tilde{\boldsymbol{\varepsilon}}'\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'\tilde{\boldsymbol{\varepsilon}} \\ &= \tilde{\boldsymbol{\varepsilon}}_*'\boldsymbol{\Lambda}\tilde{\boldsymbol{\varepsilon}}_* \qquad \text{where } \tilde{\boldsymbol{\varepsilon}}_* = \boldsymbol{V}'\tilde{\boldsymbol{\varepsilon}} \end{aligned}$$

$$\begin{aligned} \therefore \ E\left(\tilde{\boldsymbol{\varepsilon}}_*\right) &= E\left(\boldsymbol{V}'\tilde{\boldsymbol{\varepsilon}}\right) \\ &= \boldsymbol{0} \end{aligned}$$

$$\begin{aligned} \therefore \ Var\left(\tilde{\boldsymbol{\varepsilon}}_*\right) &= E\left(\tilde{\boldsymbol{\varepsilon}}_*\tilde{\boldsymbol{\varepsilon}}_*'\right) \\ &= E\left(\boldsymbol{V}'\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}'\boldsymbol{V}\right) \\ &= \sigma^2\boldsymbol{V}'\boldsymbol{V} \\ &= \sigma^2\boldsymbol{I}_n \end{aligned}$$

$$\therefore \ \tilde{\boldsymbol{\varepsilon}}_* \sim N\left(\boldsymbol{0},\sigma^2\boldsymbol{I}_n\right)$$

$$\Leftrightarrow \ \frac{\tilde{\boldsymbol{\varepsilon}}_*}{\sigma} \sim N\left(\boldsymbol{0},\boldsymbol{I}_n\right)$$

$$\Leftrightarrow \ \frac{\tilde{\boldsymbol{\varepsilon}}_*}{\sigma} \sim N(0,1) \ \ ^{\forall} i$$

$$\therefore \ \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{\varepsilon}}_*'\boldsymbol{\Lambda}\tilde{\boldsymbol{\varepsilon}}_*$$

$$\begin{aligned} \Rightarrow \ \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} &= \frac{\tilde{\boldsymbol{\varepsilon}}_*'}{\sigma}\boldsymbol{\Lambda}\frac{\tilde{\boldsymbol{\varepsilon}}_*}{\sigma} \\ &= \frac{1}{\sigma^2}\tilde{\boldsymbol{\varepsilon}}_*'\boldsymbol{\Lambda}\tilde{\boldsymbol{\varepsilon}}_* \\ &= \frac{1}{\sigma^2}\left(\tilde{\varepsilon}_{*1}^2+\cdots+\tilde{\varepsilon}_{*n-k-1}^2\right) \\ &= \sum_{j=1}^{n-k-1}\left(\frac{\tilde{\varepsilon}_{*j}}{\sigma}\right)^2 \sim \chi^2(n-k-1) \end{aligned}$$

Note that if $z_i \sim N(0,1)$, then $\sum_{i=1}^{n} z_i^2 \sim \chi^2(n)$.

$$\therefore \ \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \sigma^2\chi^2(n-k-1)$$

Therefore, if $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$, then $E(\hat{\sigma}^2) = (n-k-1)\sigma^2 \neq \sigma^2$ and hence biased. Futhermore, if $z_i \sim \chi^2(n)$, then $E(z_i) = n$ and $Var(z_i) = n^2$

$$\begin{aligned} \therefore \ \hat{\sigma}^2 &= \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k-1} \\ &\sim \frac{\sigma^2}{n-k-1}\chi^2(n-k-1) \end{aligned}$$

$$\Rightarrow \ \frac{n-k-1}{\sigma^2}\hat{\sigma}^2 \sim \chi^2(n-k-1)$$

1.2.3. Construction of Test Statistic

With these distributional characteristics, recall the test statistic for the null hypothesis mentioned above.

$$\begin{aligned} t &= \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c}}{\sqrt{Var(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c})}} \\ &= \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c}}{\sigma\sqrt{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}} \\ &\sim N(0,1) \qquad \text{under } H_0 \end{aligned}$$

Note that if $\tilde{z}$ and $\tilde{x}$ are independent and follow Standard Normal distribution and Chi-square distribution respectively, then the statement below can be derived.

$$\frac{\tilde{z}}{\sqrt{\tilde{x}/n}} \sim t(n)$$

Since we can construct one Normal variable and the other Chi-square variable under the null hypothesis, we can also derive another variable which follows t distribution.

$$\begin{aligned} \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c}}{\sigma\sqrt{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}} &\sim N(0,1) \\ \frac{\sqrt{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/\sigma^2}}{n-k-1} &\sim \chi^2(n-k-1) \end{aligned}$$

$$\therefore \quad \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c}}{\sigma\sqrt{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}} \Big/ \frac{\sqrt{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/\sigma^2}}{n-k-1} = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c}}{\sqrt{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}} \Big/ \sqrt{\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k-1}}$$
$$= \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c}}{\hat{\sigma}\sqrt{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}}$$
$$\sim \quad t(n-k-1)$$

Through the t statistic above, we can test the null hypothesis correctly if the assumptions are not violated. And this process can also be used in many cases by manipulating $\boldsymbol{R}$ and $\boldsymbol{c}$ of the null hypothesis.

For instance, in the given regression equation $y=\beta_0+\beta_1 x+\varepsilon$ and the null hypothesis is stated as H_0: $\beta_1=0$, then $\boldsymbol{R}$ and $\boldsymbol{c}$ for the test can be designed respectively.

$$\boldsymbol{R} = (0\,1)$$
$$\boldsymbol{c} = 0$$

$$\Rightarrow \quad \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{c}$$

$$\therefore \quad (0\,1)\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} = 0$$

Lecture Note 2

2014-03-25

2.1. Wald Test

2.1.1. Construction of Test Statistic

$$H_0: \ \boldsymbol{R\beta} = \boldsymbol{c}$$

Note Wald criterion

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$$

$$\Rightarrow \ \boldsymbol{\Sigma}^{-1/2}\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$$

$$\therefore \ \boldsymbol{u}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u} \sim \chi^2(q)$$

Or equivalently,

$$u_i \overset{i.i.d.}{\sim} N(0,1)$$

$$\Rightarrow \ \sum_{i=1}^{n} u_i^2 \sim \chi^2(n)$$

For $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, hereafter $\hat{\boldsymbol{\beta}}$, we already know the fact that

$$\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}') \qquad \text{under } H_0$$

$$\therefore \ (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'[\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \sim \chi^2(q)$$

Recall that

$$\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k-1)$$

Remark For $x_1 \sim \chi^2(n_1)$ and $x_2 \sim \chi^2(n_2)$

$$\Rightarrow \ \frac{x_1/n_1}{x_2/n_2} \sim F(n_1, n_2)$$

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'[\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c})/q}{(n-k-1)\hat{\sigma}^2/[(n-k-1)\sigma^2]} = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'[\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c})/q}{\hat{\sigma}^2/\sigma^2}$$

$$\sim \quad F(q, n-k-1)$$

Therefore,

$$\frac{1}{q\hat{\sigma}^2}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c})'\left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c}) \quad \sim \quad F(q, n-k-1) \qquad \text{under } H_0$$

This is Wald test statistic.

2.1.2. Relation between t Test and F Test

Note The mathematical relationship between t distribution with n degrees of freedom and F distribution with 1 and n degrees of freedom can be depicted as below.

$$\begin{aligned} x &\sim t(n) \\ \Rightarrow x^2 &\sim F(1, n) \end{aligned}$$

Not rigorously but roughly, if $q=1$, then $t^2 = F$; this distributional characteristic is similar to the relationship between Standard Normal distribution and Chi-square distribution.

$$\begin{aligned} y &\sim N(0,1) \\ \Rightarrow y^2 &\sim \chi^2(1) \end{aligned}$$

2.2. Heteroskedasticity

Recall t test statistic above.

$$t = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c}}{\hat{\sigma}\sqrt{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}}$$

Traditional Econometrics course often teaches the way to test the null hypothesis based on several assumptions, such as Homoskedasticity assumption; unfortunately, these binding assumptions are frequently violated in reality and hence Econometric course also teaches various helpful remedies whereby the test can be correctly conducted. This is a brief sketch of an undergraduate level Econometrics course.

Heteroskedasticity, which is the violation of Assumption 2, can be sketched as below; for a classical linear regression equation below,

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

Heteroskedasticity literally means that the size of variance for each residual term is heterogeneous, which could mathematically be denoted as below.

$$E\left(\tilde{\varepsilon}_i^2|\boldsymbol{X}\right) = \sigma_i^2 \qquad \sigma_i^2 \neq \sigma_j^2 \ \ ^\forall i \neq j$$

Or compactly,

$$\begin{aligned} E(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}'|\boldsymbol{X}) &= \boldsymbol{\Omega} \\ &= \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \\ &= \sigma^2 \underbrace{\begin{pmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_n \end{pmatrix}}_{\boldsymbol{\Sigma}} \qquad \boldsymbol{\Sigma} \in \mathbb{R}^{n \times n},\ \Sigma_i \in \mathbb{R},\ \sigma_i^2 = \sigma^2 \Sigma_i \\ &= \sigma^2 \boldsymbol{\Sigma} \end{aligned}$$

Although there exists Heteroskedasticity problem in the residual, $\hat{\boldsymbol{\beta}}$ is still unbiased since Assumption 1 is not violated; i.e. $E(\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X}) = \boldsymbol{0}$. However, $\hat{\boldsymbol{\beta}}$ is not efficient anymore.

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= Var(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{X}) \\ &= Var\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X}\right] \\ &= E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}\right] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' E(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}'|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &\neq \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

Because of this inefficiency, other econometric techniques to calculate the size of variance are required to test the null hypothesis precisely.

Case 1 If $\boldsymbol{\Sigma}$ is known and diagonal, then we will use *Weighted Least Squares*. We don't know σ^2 in this case. Hence, we will again use Spectral Decomposition, which is also known as Cholesky Decomposition or Eigenvalue Decomposition.

Theorem Positive semi-definite matrix can always be decomposed as below.

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{P}\boldsymbol{P}'$$

Instead of introducing a tedious proof of above Spectral Theorem, we will just use this theorem hereafter.

$$\begin{aligned} \Rightarrow \boldsymbol{\Sigma} &= (\boldsymbol{P}\boldsymbol{P}')^{-1} \\ &= \boldsymbol{P}'^{-1}\boldsymbol{P}^{-1} \end{aligned}$$

$$\Rightarrow \boldsymbol{P}'\boldsymbol{\Sigma}\boldsymbol{P} = \boldsymbol{I}_n$$

$$\Rightarrow \boldsymbol{P} := \begin{pmatrix} \Sigma_1^{-1/2} & 0 & \cdots & 0 \\ 0 & \Sigma_2^{-1/2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_n^{-1/2} \end{pmatrix}$$

$$\begin{aligned} \therefore \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \Rightarrow \boldsymbol{P}'\boldsymbol{y} &= \boldsymbol{P}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}'\boldsymbol{\varepsilon} \\ \Rightarrow \boldsymbol{y}_* &= \boldsymbol{X}_*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_* \end{aligned}$$

$$\begin{aligned} \Rightarrow Var(\boldsymbol{\varepsilon}_*|\boldsymbol{X}) &= Var(\boldsymbol{P}'\boldsymbol{\varepsilon}|\boldsymbol{X}) \\ &= E(\boldsymbol{P}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{P}|\boldsymbol{X}) \\ &= \sigma^2\boldsymbol{P}'\boldsymbol{\Sigma}\boldsymbol{P} \\ &= \sigma^2\boldsymbol{I}_n \end{aligned}$$

$$\begin{aligned} \therefore \hat{\boldsymbol{\beta}}_{\mathrm{WLS}} &= (\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*'\boldsymbol{y}_* \\ &= (\boldsymbol{X}'\boldsymbol{P}\boldsymbol{P}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{P}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \end{aligned}$$

Therefore, above $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}$ is BLUE.

$$\begin{aligned} \therefore Var(\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}|\boldsymbol{X}) &= \sigma^2(\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \end{aligned}$$

However, there is no answer for the question that "How do we know the structure of $\boldsymbol{\Sigma}$ in reality?" If there is no justification for this question, WLS is no more valid; *Inference Manipulation*. Various papers have adopted WLS without any notification. This bad attitude should be avoided.

Case 2 If $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Omega}$ equivalently) is unknown, then we should estimate $\hat{\boldsymbol{\Omega}}$ firstly and then conduct remain procedures secondly; this is called as *Feasible Generalized Least Squares*.

Indeed, it is impossible to estimate $\hat{\boldsymbol{\Omega}}$ since it contains n^2 entries in it and we cannot estimate it based on n observations. Instead, we can successfully estimate $\hat{E}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}) = \widehat{\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}}$, which includes k^2 elements in it; usually, n is bigger than k. Note below.

$$\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X} = \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_{,i}\boldsymbol{x}_{,i}'$$

$$\because \boldsymbol{x}_{,i} = \begin{pmatrix} 1 \\ x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix}$$

2.2.1. White's Robust Variance

Definition White (1980,Econometrica) suggests *Heteroskedasticity Consistent Variance Estimator*, which is also known as *HC Estimator* or *White Standard Error*.

$$\begin{aligned} \hat{E}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}) &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \boldsymbol{x}_{,i}\boldsymbol{x}_{,i}' \qquad \text{for } \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_{,i}\boldsymbol{x}_{,i}' \\ \therefore \; Var\left(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}|\boldsymbol{X}\right) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{E}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \left(\sum_{i=1}^{n} \boldsymbol{x}_{,i}\boldsymbol{x}_{,i}'\right)^{-1}\left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \boldsymbol{x}_{,i}\boldsymbol{x}_{,i}'\right)\left(\sum_{i=1}^{n} \boldsymbol{x}_{,i}\boldsymbol{x}_{,i}'\right)^{-1} \\ \hat{\varepsilon}_i &= y_i - \boldsymbol{x}_{,i}'\hat{\boldsymbol{\beta}} \end{aligned}$$

Fortunately, White $Var\left(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}|\boldsymbol{X}\right)$ is consistent with Heteroskedasticity.

2.2.2. Newey-West Robust Variance

However, this estimator is still problematic if there exists not only Heteroskedasticity, but also Autocorrelation, simultaneously.

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}$$

Or equivalently,

$$\Rightarrow E(\varepsilon_i\varepsilon_j|\boldsymbol{X}) = \begin{cases} \sigma_i^2 & \forall\, i = j \\ \sigma_{ij} & \forall\, i \neq j \end{cases}$$

$$\therefore \; \boldsymbol{X}'\begin{pmatrix} 0 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & 0 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & 0 \end{pmatrix}\boldsymbol{X} = \sum_{j=1}^{n-1}\sum_{i=j+1}^{n} E\left(\varepsilon_i\varepsilon_{i-j}|\boldsymbol{X}\right)\left(\boldsymbol{x}_{,i}\boldsymbol{x}_{,i-j}' + \boldsymbol{x}_{,i-j}\boldsymbol{x}_{,i}'\right)$$

The term j of above equation means the distance between two observations, which should be determined decreasingly in increasing distance. *Bartlett Kernel* is frequently applied as a weighting scheme such that above relationships can be reflected.

$$w_j = \begin{cases} 1 - \dfrac{j}{J+1} & \forall\, 1 \leq j \leq J \\ 0 & \forall\, J < j \leq n \end{cases}$$

Definition Newey and West (1987,Econometrica) suggests *Heteroskedasticity Autocorrelation Consistent Variance Estimator*, which is also known as *HAC Estimator* or *Newey-West Standard Error*.

$$\begin{aligned}\hat{E}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}) &= \sum_{i=1}^{n}\hat{\varepsilon}_i^2\boldsymbol{x}_{,i}\boldsymbol{x}_{,i}' + \sum_{j=1}^{J}w_j\sum_{i=j+1}^{n}\hat{\varepsilon}_i\hat{\varepsilon}_{i-j}\left(\boldsymbol{x}_{,i}\boldsymbol{x}_{,i-j}' + \boldsymbol{x}_{,i-j}\boldsymbol{x}_{,i}'\right)\\ \therefore\ Var\left(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}|\boldsymbol{X}\right) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{E}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}\end{aligned}$$

Newey-West $Var\left(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}|\boldsymbol{X}\right)$ is consistent with Heteroskedasticity and Autocorrelation.

2.3. Seemingly Unrelated Regression

Suppose that there are m regression equations.

$$\begin{aligned}\boldsymbol{y}^1 &= \boldsymbol{X}^1\boldsymbol{\beta}^1 + \boldsymbol{u}^1\\ \boldsymbol{y}^2 &= \boldsymbol{X}^2\boldsymbol{\beta}^2 + \boldsymbol{u}^2\\ &\vdots\\ \boldsymbol{y}^m &= \boldsymbol{X}^m\boldsymbol{\beta}^m + \boldsymbol{u}^m\end{aligned}$$

Note that $\boldsymbol{\beta}^1$, $\boldsymbol{\beta}^2$, $\cdots$, $\boldsymbol{\beta}^m$ can have different dimensions, respectively.

$$\begin{aligned}\Rightarrow \begin{pmatrix}\boldsymbol{y}^1\\ \boldsymbol{y}^2\\ \vdots\\ \boldsymbol{y}^m\end{pmatrix} &= \begin{pmatrix}\boldsymbol{X}^1 & \boldsymbol{0} & \cdots & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{X}^2 & \cdots & \boldsymbol{0}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}^m\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}^1\\ \boldsymbol{\beta}^2\\ \vdots\\ \boldsymbol{\beta}^m\end{pmatrix} + \begin{pmatrix}\boldsymbol{u}^1\\ \boldsymbol{u}^2\\ \vdots\\ \boldsymbol{u}^m\end{pmatrix}\\ \Rightarrow\ \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}\end{aligned}$$

If the assumption below is satisfied, then there is no problem at all.

$$\begin{aligned}E(\boldsymbol{u}^i|\boldsymbol{X}) &= \boldsymbol{0}\quad {}^{\forall}i \in \{1,2,\cdots,m\}\\ E(\boldsymbol{u}^i\boldsymbol{u}^{i\prime}|\boldsymbol{X}) &= \sigma^{i2}\boldsymbol{I}_n\end{aligned}$$

Above means i.e. each regression equation is Classical Linear Regression. However,

$$E(\boldsymbol{u}^j\boldsymbol{u}^{l\prime}|\boldsymbol{X}) = \sigma^{jl}\boldsymbol{I}_n$$

i.e.

$$\begin{aligned}E\left(u_t^j u_t^l|\boldsymbol{X}\right) &= \sigma^{jl}\quad {}^{\forall}j \neq l\\ E\left(u_t^j u_s^j|\boldsymbol{X}\right) &= 0\quad {}^{\forall}t \neq s\end{aligned}$$

This relationship cannot be found if there is only one regression equation; thus, it is

usually viewed 'seemingly unrelated.' If we use GLS, we can estimate these estimators, which are more efficient than OLS estimator.

$$\Rightarrow \begin{pmatrix} \boldsymbol{y}^1 \\ \boldsymbol{y}^2 \\ \vdots \\ \boldsymbol{y}^m \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}^2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}^m \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^1 \\ \boldsymbol{\beta}^2 \\ \vdots \\ \boldsymbol{\beta}^m \end{pmatrix} + \begin{pmatrix} \boldsymbol{u}^1 \\ \boldsymbol{u}^2 \\ \vdots \\ \boldsymbol{u}^m \end{pmatrix}$$

$$\Rightarrow \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \qquad \boldsymbol{y} \in \mathbb{R}^{mn},\ \boldsymbol{X} \in \mathbb{R}^{mn \times \sum k_i},\ \boldsymbol{\beta} \in \mathbb{R}^{\sum k_i},\ \boldsymbol{u} \in \mathbb{R}^{mn}$$

Note that if $\boldsymbol{y}^i$ have same dimensions for all $i \in \{1,2,\cdots,m\}$, the data can be called as a balanced panel. If not, it can be called as an unbalanced panel and using an unbalanced panel requires careful deliberation.

SAS is not good to analyze these *Seemingly Unrelated Regressions* with unbalanced panel data because it truncates its observation to make the data balanced. Thus, SAS can be recommended only if the data is balanced in terms of panel analysis. On the other hand, STATA provides more attractive way to analyze these regressions. In addition, EViews can do this, but it is not as powerful as STATA.

2.3.1. Generalized Least Squares Estimation for SUR

These non-standard error structure can be written as below.

$$E(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}) \neq \sigma^2 \boldsymbol{I}_{mn}$$

Therefore, GLS technique is required for this case.

$$\begin{aligned} E(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}) &= \boldsymbol{\Omega} \\ &= \begin{pmatrix} E(\boldsymbol{u}^1\boldsymbol{u}^{1\prime}|\boldsymbol{X}) & E(\boldsymbol{u}^1\boldsymbol{u}^{2\prime}|\boldsymbol{X}) & \cdots & E(\boldsymbol{u}^1\boldsymbol{u}^{m\prime}|\boldsymbol{X}) \\ E(\boldsymbol{u}^2\boldsymbol{u}^{1\prime}|\boldsymbol{X}) & E(\boldsymbol{u}^2\boldsymbol{u}^{2\prime}|\boldsymbol{X}) & \cdots & E(\boldsymbol{u}^2\boldsymbol{u}^{m\prime}|\boldsymbol{X}) \\ \vdots & \vdots & \ddots & \vdots \\ E(\boldsymbol{u}^m\boldsymbol{u}^{1\prime}|\boldsymbol{X}) & E(\boldsymbol{u}^m\boldsymbol{u}^{2\prime}|\boldsymbol{X}) & \cdots & E(\boldsymbol{u}^m\boldsymbol{u}^{m\prime}|\boldsymbol{X}) \end{pmatrix} \\ &= \begin{pmatrix} \sigma^{11}\boldsymbol{I}_n & \sigma^{12}\boldsymbol{I}_n & \cdots & \sigma^{1m}\boldsymbol{I}_n \\ \sigma^{21}\boldsymbol{I}_n & \sigma^{22}\boldsymbol{I}_n & \cdots & \sigma^{2m}\boldsymbol{I}_n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^{m1}\boldsymbol{I}_n & \sigma^{m2}\boldsymbol{I}_n & \cdots & \sigma^{mm}\boldsymbol{I}_n \end{pmatrix} \\ &= \begin{pmatrix} \sigma^{11} & \sigma^{12} & \cdots & \sigma^{1m} \\ \sigma^{21} & \sigma^{22} & \cdots & \sigma^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^{m1} & \sigma^{m2} & \cdots & \sigma^{mm} \end{pmatrix} \otimes \boldsymbol{I}_n \\ &= \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n \qquad \boldsymbol{\Sigma} \in \mathbb{R}^{m \times m} \end{aligned}$$

Remark *Kronecker* product $\otimes$ can be expressed as below; for two matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times q}$,

$$\underset{(m\times n)}{\boldsymbol{A}}\otimes\underset{(p\times q)}{\boldsymbol{B}} = \underbrace{\begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B}\cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B}\cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & \ddots \quad \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B}\cdots & a_{mn}\boldsymbol{B} \end{pmatrix}}_{(mp\times nq)}$$

Property The inverse of a Kronecker product has below characteristic.

$$\begin{aligned} (\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) &= \boldsymbol{AC}\otimes\boldsymbol{BD} \\ \Rightarrow\ (\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}) &= \boldsymbol{AA}^{-1}\otimes\boldsymbol{BB}^{-1} \\ &= \boldsymbol{I}_{mn} \\ \therefore\ (\boldsymbol{A}\otimes\boldsymbol{B})^{-1} &= \boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1} \end{aligned}$$

Thus,

$$\begin{aligned} \boldsymbol{\Omega}^{-1} &= \boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_n \\ \Rightarrow\ \hat{\boldsymbol{\beta}}_{\mathrm{GLS}} &= \left[\begin{pmatrix} \boldsymbol{X}^{1\prime} & \boldsymbol{0}' & \cdots & \boldsymbol{0}' \\ \boldsymbol{0}' & \boldsymbol{X}^{2\prime} & \cdots & \boldsymbol{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0}' & \boldsymbol{0}' & \cdots & \boldsymbol{X}^{m\prime} \end{pmatrix}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_n)\begin{pmatrix} \boldsymbol{X}^{1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}^{2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}^{m} \end{pmatrix}\right]^{-1}\begin{pmatrix} \boldsymbol{X}^{1\prime} & \boldsymbol{0}' & \cdots & \boldsymbol{0}' \\ \boldsymbol{0}' & \boldsymbol{X}^{2\prime} & \cdots & \boldsymbol{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0}' & \boldsymbol{0}' & \cdots & \boldsymbol{X}^{m\prime} \end{pmatrix}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_n)\begin{pmatrix} \boldsymbol{y}^{1} \\ \boldsymbol{y}^{2} \\ \vdots \\ \boldsymbol{y} \end{pmatrix} \end{aligned}$$

Let

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \sigma_{11} & \sigma_{12}\cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22}\cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots \quad \vdots \\ \sigma_{m1} & \sigma_{m2}\cdots & \sigma_{mm} \end{pmatrix}$$

$$\Rightarrow\ \hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \begin{pmatrix} \sigma_{11}\boldsymbol{X}^{1\prime}\boldsymbol{X}^{1} & \sigma_{12}\boldsymbol{X}^{1\prime}\boldsymbol{X}^{2} & \cdots & \sigma_{1m}\boldsymbol{X}^{1\prime}\boldsymbol{X}^{m} \\ \sigma_{21}\boldsymbol{X}^{2\prime}\boldsymbol{X}^{1} & \sigma_{22}\boldsymbol{X}^{2\prime}\boldsymbol{X}^{2} & \cdots & \sigma_{2m}\boldsymbol{X}^{2\prime}\boldsymbol{X}^{m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}\boldsymbol{X}^{m\prime}\boldsymbol{X}^{1} & \sigma_{m2}\boldsymbol{X}^{m\prime}\boldsymbol{X}^{2} & \cdots & \sigma_{mm}\boldsymbol{X}^{m\prime}\boldsymbol{X}^{m} \end{pmatrix}^{-1}\begin{pmatrix} \sum_{j=1}^{m}\sigma_{1j}\boldsymbol{X}^{1\prime}\boldsymbol{y}^{j} \\ \sum_{j=1}^{m}\sigma_{2j}\boldsymbol{X}^{2\prime}\boldsymbol{y}^{j} \\ \vdots \\ \sum_{j=1}^{m}\sigma_{mj}\boldsymbol{X}^{m\prime}\boldsymbol{y}^{j} \end{pmatrix}$$

2.3.2. Relationship between GLS and OLS Estimators

When can we obtain same $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$?

Case 1 $\sigma^{ij}=0\ {}^{\forall}i\neq j\ \Rightarrow\ \sigma_{ij}=0\ {}^{\forall}i\neq j$; then, $\sigma^{ij}=0\ \Leftrightarrow\ \sigma_{ij}=0$; i.e. no correlation between observations.

$$
\begin{aligned}
\Rightarrow\ \hat{\boldsymbol{\beta}}_{\mathrm{GLS}} &= \begin{pmatrix} \sigma_{11}\boldsymbol{X}^{1\prime}\boldsymbol{X}^{1} & \sigma_{12}\boldsymbol{X}^{1\prime}\boldsymbol{X}^{2} & \cdots & \sigma_{1m}\boldsymbol{X}^{1\prime}\boldsymbol{X}^{m} \\ \sigma_{21}\boldsymbol{X}^{2\prime}\boldsymbol{X}^{1} & \sigma_{22}\boldsymbol{X}^{2\prime}\boldsymbol{X}^{2} & \cdots & \sigma_{2m}\boldsymbol{X}^{2\prime}\boldsymbol{X}^{m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}\boldsymbol{X}^{m\prime}\boldsymbol{X}^{1} & \sigma_{m2}\boldsymbol{X}^{m\prime}\boldsymbol{X}^{2} & \cdots & \sigma_{mm}\boldsymbol{X}^{m\prime}\boldsymbol{X}^{m} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^{m}\sigma_{1j}\boldsymbol{X}^{1\prime}\boldsymbol{y}^{j} \\ \sum_{j=1}^{m}\sigma_{2j}\boldsymbol{X}^{2\prime}\boldsymbol{y}^{j} \\ \vdots \\ \sum_{j=1}^{m}\sigma_{mj}\boldsymbol{X}^{m\prime}\boldsymbol{y}^{j} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11}\boldsymbol{X}^{1\prime}\boldsymbol{X}^{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_{22}\boldsymbol{X}^{2\prime}\boldsymbol{X}^{2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sigma_{mm}\boldsymbol{X}^{m\prime}\boldsymbol{X}^{m} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{11}\boldsymbol{X}^{1\prime}\boldsymbol{y}^{1} \\ \sigma_{22}\boldsymbol{X}^{2\prime}\boldsymbol{y}^{2} \\ \vdots \\ \sigma_{mm}\boldsymbol{X}^{m\prime}\boldsymbol{y}^{m} \end{pmatrix} \\
&= \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}
\end{aligned}
$$

In this case, GLS estimator is directly equivalent to OLS estimator.

Case 2 $\boldsymbol{X}^{1} = \boldsymbol{X}^{2} = \cdots = \boldsymbol{X}^{m}\ (\equiv \boldsymbol{X}^{C})$

Above means that every regressions exploit exactly same independent variables. In this case, also, it is natural that GLS estimator and OLS estimator are equivalent.

Lecture Note 3

2014-04-02

3.1. Binary Choice Model

Definition I_i is a dependent variable. $(^{\forall} i \in \{1,2,\cdots,n\})$ Where

$$I_i := \begin{cases} 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p \end{cases}$$

Definition $\boldsymbol{x}_i$ is an independent variable.

With these given dependent and independent variables, what we want to do can be denoted as below.

$$I_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

Figure 3.1 Ordinary Dependent Variable

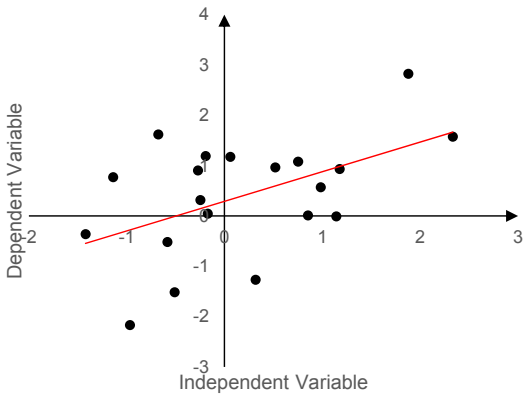

Figure 3.2 Binomial Dependent Variable

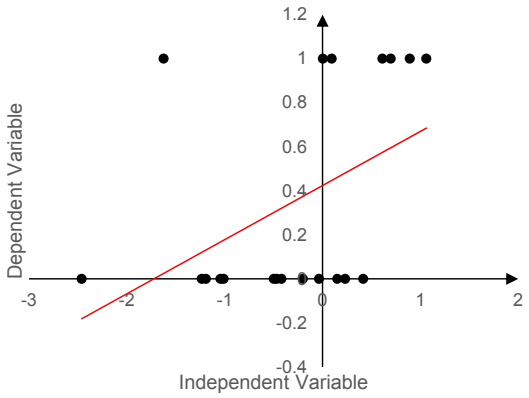

Figure 3.3 Logit Model (or Probit Model)

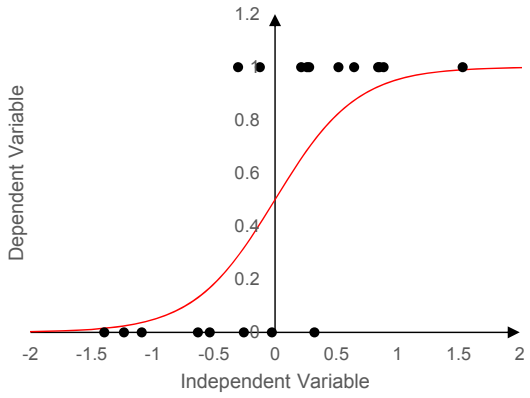

Figure 3.1 shows classical linear regression line. However, if the dependent variable is

binomially distributed, classical linear regression will not be good. Figure 3.2 graphically shows that linear regression can be wrong for this case. Fortunately, we can solve this problem based on Logit model (or Probit model if possible) that is usually adopted in binomial dependent variable case.

$$I_i = F(\boldsymbol{x}_i'\boldsymbol{\beta}) + \varepsilon_i$$

Where F is a cumulative distribution function, which has below properties.

$$\begin{aligned} F &\in [0,1] \\ \frac{dF(z)}{dz} &\geq 0 \end{aligned}$$

Note If I_i follows Bernoulli distribution as below, then

$$I_i := \begin{cases} 0 & f_1 = p \\ 1 & f_2 = 1-p \end{cases}$$

$$\begin{aligned} \Rightarrow E(I_i) &= p \\ Var(I_i) &= p(1-p) \end{aligned}$$

$$\begin{aligned} \therefore E(I_i|\boldsymbol{x}_i) &= F(\boldsymbol{x}_i'\boldsymbol{\beta}) \\ &= P(I_i = 1|\boldsymbol{x}_i) \end{aligned}$$

If F is Standard Normal, then Probit model can be used.

$$F = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$$

If F is a logistic distribution function, then Logit model can be used.

$$F(\boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}$$

Note F is bounded with 0 and 1.

If $\boldsymbol{x}_i > \mathbf{0}$, then

$$\begin{aligned} \lim_{\boldsymbol{\beta}\to\infty} F(\boldsymbol{x}_i'\boldsymbol{\beta}) &= \lim_{\boldsymbol{\beta}\to\infty} \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} \\ &= \lim_{\boldsymbol{\beta}\to\infty} \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} \qquad \text{by l'Hôspital's rule} \\ &= 1 \end{aligned}$$

Similarly,

$$\begin{aligned}\lim_{\beta \to -\infty} F(\boldsymbol{x}_i'\boldsymbol{\beta}) &= \lim_{\beta \to -\infty} \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} \\ &= \frac{0}{1+0} \\ &= 0\end{aligned}$$

Figure 3.4 The Use of Logit and Probit Model