Stata course in advanced econometrics I

Meiting Wang*

October 22, 2021

1 Simple regression model

- The population model:

$$y = \beta_0 + \beta_1 x + u, \quad \mathbb{E}(u \mid x) = 0$$

- Estimate the population parameters(β_0, β_1):

$$\begin{aligned}
& \mathbb{E}(u \mid x) = 0 \\
\Longrightarrow & \begin{cases} \mathbb{E}(u) = 0 \\ \mathbb{E}(xu) = 0 \end{cases} \\
\Longrightarrow & \begin{cases} \mathbb{E}(y - \beta_0 - \beta_1 x) = 0 \\ \mathbb{E}\left[x(y - \beta_0 - \beta_1 x)\right] = 0 \end{cases} \\
\Longrightarrow & \begin{cases} n^{-1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \\ n^{-1} \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \end{cases} \\
\Longrightarrow & \begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u} \end{cases}
\end{aligned}$$

- The sum of squared residuals(SSR) after estimation:

$$\mathrm{SSR} = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

- Algebraic Properties of OLS Statistics:

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

*Meiting Wang, Email: wangmeiting92@gmail.com

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$
$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$$
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i = \hat{y}_i + \hat{u}_i$$
$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{\hat{y}}$$

- Goodness-of-Fit:

$$\begin{aligned}
\mathrm{SST} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} [(y_i - \hat{y}_i) - (\hat{y}_i - \bar{y})]^2 \\
&= \sum_{i=1}^{n} [\hat{u}_i - (\hat{y}_i - \bar{y})]^2 \\
&= \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{\mathrm{SSE}} + \underbrace{\sum_{i=1}^{n} \hat{u}_i^2}_{\mathrm{SSR}} \\
\implies \quad R^2 &= \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} \in [0, 1]
\end{aligned}$$

- Assumption for the simple linear regression(SLR):

SLR.1 Linear in parameters:

$$y = \beta_0 + \beta_1 x + u$$

SLR.2 Random sampling

SLR.3 Some sample variation in the x_i

SLR.4 Zero conditional mean:

$$\mathbb{E}(u \mid x) = 0$$

SLR.5 Homoskedasticity:

$$\mathrm{Var}(u \mid x) = \sigma^2$$

- Unbiasedness of the OLS estimators(under assumptions SLR.1 through SLR.4):

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\end{aligned}$$

$$
\begin{aligned}
&= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\hat{\beta}_0 &= \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u} \\
\Longrightarrow \mathbb{E}(\hat{\beta}_1) &= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,\mathbb{E}(u_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 \\
\Longrightarrow \mathbb{E}(\hat{\beta}_0) &= \beta_0 + \left[\beta_1 - \mathbb{E}(\hat{\beta}_1)\right]\bar{x} + \mathbb{E}(\bar{u}) = \beta_0
\end{aligned}
$$

- Variance of the OLS estimators(under assumptions SLR.1 through SLR.5):

$$
\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\hat{\beta}_0 &= \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u} \\
\Longrightarrow \mathrm{Var}(\hat{\beta}_1) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\mathrm{Var}(u_i)}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\Longrightarrow \mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}(\bar{u} - \hat{\beta}_1\bar{x}) \\
&= \mathrm{Var}(\bar{u}) + \bar{x}^2\mathrm{Var}(\hat{\beta}_1) - 2\bar{x}\underbrace{\mathrm{Cov}(\bar{u},\hat{\beta}_1)}_{=0} \\
&= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sigma^2 n^{-1}\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}
$$

- Unbiased estimator of σ^2(under assumptions SLR.1 through SLR.5):

$$
\begin{aligned}
& & \hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = u_i + (\beta_1 - \hat{\beta}_1)x_i + (\beta_0 - \hat{\beta}_0) \\
&\Longrightarrow & 0 &= \bar{u} + (\beta_1 - \hat{\beta}_1)\bar{x} + (\beta_0 - \hat{\beta}_0) \\
&\Longrightarrow & \hat{u}_i &= (u_i - \bar{u}) + (\beta_1 - \hat{\beta}_1)(x_i - \bar{x}) \\
&\Longrightarrow & \hat{u}_i^2 &= (u_i - \bar{u})^2 + (\beta_1 - \hat{\beta}_1)^2(x_i - \bar{x})^2 + 2(\beta_1 - \hat{\beta}_1)(x_i - \bar{x})(u_i - \bar{u}) \\
&\Longrightarrow \mathbb{E}\sum_{i=1}^{n}\hat{u}_i^2 & &= \underbrace{\mathbb{E}\sum_{i=1}^{n}(u_i - \bar{u})^2}_{=(n-1)\sigma^2} + \underbrace{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\mathbb{E}(\beta_1 - \hat{\beta}_1)^2}_{=\sigma^2} + \underbrace{2\mathbb{E}\left[(\beta_1 - \hat{\beta}_1)\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})\right]}_{=-2\sigma^2} \\
& & &= (n-2)\sigma^2 \\
&\Longrightarrow & \hat{\sigma}^2 &= \frac{\sum_{i=1}^{n}\hat{u}_i^2}{n-2} = \frac{\mathrm{SSR}}{n-2}
\end{aligned}
$$

- Standard error:
  - Standard error of the regression(also called root mean squared error in Stata):

$$
\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\mathrm{SSR}}{n-2}}
$$

– Standard error of the OLS estimators:

$$\text{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\text{se}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{n^{-1}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

2 Multiple regression model

- Assumption for the multiple linear regression(MLR):

  MLR.1 Linear in parameters:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

  MLR.2 Random sampling

  MLR.3 No perfect collinearity

  MLR.4 Zero conditional mean:

  $$\mathbb{E}\left(u \mid x_1, x_2, \ldots, x_k\right) = 0$$

  MLR.5 Homoskedasticity:

  $$\text{Var}\left(u \mid x_1, x_2, \cdots, x_k\right) = \sigma^2$$

  MLR.6 Normality: The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2, i.e. $u \sim \mathcal{N}\left(0, \sigma^2\right)$.

  Assumptions MLR.1 through MLR.5 are called the Gauss Markov assumptions, and assumptions MLR.1 to MLR.6 are called the classical linear model(CLM) assumptions.

- Write the regression model in the form of matrices:

  $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

  where

  $$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ 1 & x_{2,1} & \cdots & x_{2,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

- Estimate the population parameters:

  $$\begin{aligned} & \min_{\boldsymbol{\beta}} Q = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \Longrightarrow \quad & \mathbf{0} = \frac{\partial Q}{\partial \boldsymbol{\beta}} = \frac{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'}{\partial \boldsymbol{\beta}} \frac{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})} \\ & \quad = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \Longrightarrow \quad & \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \\ \Longrightarrow \quad & \hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \end{aligned}$$

- Unbiasedness of the OLS estimators(under assumptions MLR.1 through MLR.4):

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}) \\ &= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\ \Longrightarrow \mathbb{E}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbb{E}\left(\mathbf{u}\right) \\ &= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{0} \\ &= \boldsymbol{\beta} \end{aligned}$$

- Omitted variable bias. Suppose the population model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

  In the multiple regression,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

  In the simple regression,

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

  In the auxilliary regression,

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$$

  Then, there will be

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$$

- Variance of the OLS estimators(under assumptions MLR.1 through MLR.5):
  - Formula I:

$$\operatorname{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\mathrm{SST}_j\left(1-\mathrm{R}_j^2\right)} = \frac{\sigma^2}{\sum_{i=1}^n \hat{r}_{ij}^2}, \quad j = 1, 2, \cdots, k$$

    where $\mathrm{SST}_j = \sum_{i=1}^n \left(x_{ij} - \bar{x}_j\right)^2$; R_j^2 and $\hat{r}_{ij}$ are the R-squared and residual from regressing x_j on all other independent variables (including an intercept), respectively.
  - Formula II:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\ \Longrightarrow \operatorname{Var}(\hat{\boldsymbol{\beta}}) &= \operatorname{Var}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\right] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\operatorname{Var}\left(\mathbf{u}\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \end{aligned}$$

- Unbiased estimator of σ^2(under assumptions MLR.1 through MLR.5)[1]:

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{u}$$

[1] Define $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

$$
\begin{aligned}
\Longrightarrow \quad & \hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u} \\
\Longrightarrow \mathbb{E}\left(\hat{\mathbf{u}}'\hat{\mathbf{u}}\right) &= \mathbb{E}\left(\mathbf{u}'\mathbf{M}\mathbf{u}\right) = \mathbb{E}\left[\operatorname{tr}\left(\mathbf{u}'\mathbf{M}\mathbf{u}\right)\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left(\mathbf{M}\mathbf{u}\mathbf{u}'\right)\right] = \operatorname{tr}\left[\mathbb{E}\left(\mathbf{M}\mathbf{u}\mathbf{u}'\right)\right] \\
&= \operatorname{tr}\left[\mathbf{M}\mathbb{E}\left(\mathbf{u}\mathbf{u}'\right)\right] = \sigma^2 \operatorname{tr}\left(\mathbf{M}\right) \\
&= \sigma^2\left[\operatorname{tr}\left(\mathbf{I}_n\right) - \operatorname{tr}\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\right] \\
&= \sigma^2\left[\operatorname{tr}\left(\mathbf{I}_n\right) - \operatorname{tr}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right)\right] \\
&= (n-k-1)\sigma^2 \\
\Longrightarrow \quad & \hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k-1} = \frac{\text{SSR}}{n-k-1}
\end{aligned}
$$

- Standard error:
  - Standard error of the regression(also called root mean squared error in Stata):

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\text{SSR}}{n-k-1}}$$

  - Standard error of the OLS estimators(Formula I):

$$\text{se}\left(\hat{\beta}_j\right) = \frac{\hat{\sigma}}{\sqrt{\text{SST}_j\left(1-\text{R}_j^2\right)}} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n \hat{r}_{ij}^2}}, \quad j = 1, 2, \cdots, k$$

  - Standard error of the OLS estimators(Formula II):

$$\text{se}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma} \cdot \text{sqrt}\left[\text{diagonal}\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)\right]$$

- Gauss-Markov Theorem: Under assumptions MLR.1 through MLR.5, $\hat{\beta}_0$, $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are the best linear unbiased estimators (BLUEs) of $\beta_0, \beta_1, \ldots,$ β_k, respectively.

- Related distribution(under assumptions MLR.1 through MLR.6)[2]:

$$
\begin{aligned}
\hat{\beta}_j &\sim \mathcal{N}\left[\beta_j, \text{Var}(\hat{\beta}_j)\right] \\
\frac{\hat{\beta}_j - \beta_j}{\text{sd}(\hat{\beta}_j)} &\sim \mathcal{N}(0,1) \\
\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} &\sim \chi^2_{n-k-1} \\
\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} &\sim t_{n-k-1} \\
\frac{(\text{SSR}_r - \text{SSR}_{ur})/q}{\text{SSR}_{ur}/(n-k-1)} = \frac{(R^2_{ur} - R^2_r)/q}{(1-R^2_{ur})/(n-k-1)} &\sim F_{q,n-k-1}
\end{aligned}
$$

[2] $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j)$ and $\hat{\sigma}^2/\sigma^2$ can be shown to be independent.