Review of Ordinary Least Square
February 4, 2020

Teaching Assistant: Jeffrey Je-uei Kuo **Email:** jeffkuo@gwu.edu

1. The Causality, Selection Bias, and Randomized Control Trial (RCT)

The example of the Randomized Control Trial is excerpted from Chapter 1 to 3 from Mostly Harmless Econometrics, authored by J. Angrist, and J-S. Pischke.

In social sciences, the main goal of the research is to identify the causality between the events. However, due to the essence of social science, it is hard (or even impossible) to conduct experiments of interests by running the standard laboratory trial. For example, if we would like to see whether a new medicine is useful, we can not have the same patient received two different trials. Once the patients receive a specific medical treatment, we are not able to reverse the process and rerun the other treatment on the same person. In some cases, it might be possible to do it without involving ethical problems, but it usually costs a lot.

Hence, most of the economic data we have seen are not laboratory data. Instead, the data represents the observed outcomes after the treatment, i.g. The income survey after one enrolled in a particular training program or the income survey after the students have finished their degree.

Here is the example that Angrist and Pischke use in the text. There is an indicator representing the health status of the patients. One is the worst and five is the best. The researcher would like to know if the admission in the hospital ("hospitalized") improves the health status.

The following table shows the result after surveying the people.

Group	Sample Size	Mean Health Status	Std. Error
Hospital	7,774	**3.21**	0.014
No Hospital	90,049	**3.93**	0.003

Does "Being Hospitalized" jeopardize the health?

$E(Y_i|D_i=1) - E(Y_i|D_i=0)$
$= 3.21 - 3.93 = -0.72$

Assume there are two potential outcomes for the health status for each individual i, Y_{1i} is the health status if they had been hospitalized ($D_i = 1$), Y_{0i} is the health status if they had not been hospitalized ($D_i = 0$).

$$\text{Potential Outcome for } i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

We have seen is so-called the Average Treatment Effect, which is $E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0)$
However, the term capturing the causal effect of the hospitalization should be
Average Treatment on the Treated, which is $E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1)$, or $E(Y_{1i} - Y_{0i}|D_i = 1)$
What are the relation between the two terms?

Observed Treatment Effects (or Average Treatment Effect).

$= E(Y_i|D_i=1) - E(Y_i|D_i=0)$.

$= E(Y_{1i}|D_i=1) - E(Y_{0i}|D_i=0)$ ← because treatment depends on outcome.

$= \underbrace{E(Y_{1i}|D_i=1) - E(Y_{0i}|D_i=1)}_{\text{Average Treatment Effect on the Treated}} + \underbrace{E(Y_{0i}|D_i=1) - E(Y_{0i}|D_i=0)}_{\text{Selection Bias}}$ ← mathematical manipulation.

★ Average Treatment Effect on the Treated
Causality!

↳ Selection Bias
In the case, this term is negative.
means. People get treated have worse health status than people do NOT get treated.

So If treatment is independent with outcome., randomly admit patients
$E(Y_i|D_i=1) - E(Y_i|D_i=0) = E(Y_{1i}|D_i=1) - E(Y_{0i}|D_i=0) \xrightarrow{\text{RCT}} E(Y_{1i}|D_i=1) - E(Y_{0i}|D_i=1)$

2. Review of Simple Linear Regression

2.1 Definition

"Regression" is a functional relationship between two or more correlated variables that is often empirically determined from data and is used to predict values of one variable when given values of the others. Two words can be used to describe regression analysis: "empirical" and "informative."

2.2 Simple Linear Regression

There is only one independent variable / regressors/ predictor variable X in the model. The proposed (population) regression model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where β_0 is the intercept coefficient, β_1 is the slope coefficient, and ε_i is the model random error term, which takes into account all unpredictable and unknown factors that are not included in the model.

2.3 Estimation with Ordinary Least Square (OLS)

Firstly, lets define the model residual $e_i = y_i - \hat{y}_i$, where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the predicted value of y_i for a given value of x_i. The OLS chooses the prediction equation that minimizes the residual sum of squares for all sample observations, that is:

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The goal of OLS is to find the estimator b_0, b_1, such that minimize the objective function.

$$b_0, b_1 = \arg\min S = \sum_{i=1}^{n} (Y_i - \hat{B}_0 - \hat{B}_1 X_i)^2$$

FOC $\frac{\partial S}{\partial \hat{B}_0} = 0 \Rightarrow 2\sum_{i=1}^{n} (Y_i - \hat{B}_0 - \hat{B}_1 X_i)(-1) = 0 \Rightarrow \sum_{i=1}^{n} (Y_i - \hat{B}_0 - \hat{B}_1 X_i) = 0$

$\frac{\partial S}{\partial \hat{B}_1} = 0 \Rightarrow \sum_{i=1}^{n} (Y_i - \hat{B}_0 - \hat{B}_1 X_i)(-X_i) = 0$

$\Rightarrow n\bar{Y} - n\hat{B}_0 - \hat{B}_1 n\bar{X} = 0$

$\nearrow \sum Y_i - n\hat{B}_0 - \hat{B}_1 \sum X_i = 0$

3. The Role of R^2. The Goodness of Fit of the Model.

SST: Sum of Squares Total= $\sum_{i=1}^{n} (y_i - \bar{y})^2$

SSR: Sum of Squared Regression= $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

SSE: Sum of Squared Errors = $\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Coefficient of Determinants: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ [SST = SSR + SSE, but why?]

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} \left([Y_i - \hat{Y}_i]^2 + 2(Y_i - \hat{Y}_i) + (Y_i - \bar{Y})^2 \right]$$

Note The names of the terms are confusing here. Some Statistics texts refer Sum of Squared Regression as Explained Sum of Squares (ESS), and refer Sum of Square Errors as Residual Sum of Squares (RSS). And often Sum of Squares Total will be called as Total Sum of Squares (TSS). One should pay more attention to the conventions of notation usages before reading through the text.

✡ Moment conditions (FOCs) of OLS estimation

$$\min S = \sum_{i=1}^{n} (Y_i - \hat{\beta_0} - \hat{\beta_1} X_i)^2$$

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n} 2(Y_i - \hat{\beta_0} - \hat{\beta_1} X_i)(-1) = 0 \Rightarrow \sum_{i=1}^{n} (Y_i - \hat{\beta_0} - \hat{\beta_1} X_i) = 0 \quad -①$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{n} 2(Y_i - \hat{\beta_0} - \hat{\beta_1} X_i)(-X_i) = 0 \Rightarrow \sum_{i=1}^{n} (Y_i - \hat{\beta_0} - \hat{\beta_1} X_i) X_i = 0 \quad -②$$

① $\sum_{i=1}^{n} Y_i - n\hat{\beta_0} - \hat{\beta_1} \sum_{i=1}^{n} X_i = 0$ (using $\sum_{i=1}^{n} Y_i = n\bar{Y}$, $\sum_{i=1}^{n} X_i = n\bar{X}$)

$$n\bar{Y} - n\hat{\beta_0} - \hat{\beta_1} n\bar{X} = 0 \Rightarrow \bar{Y} - \hat{\beta_0} = \hat{\beta_1}\bar{X} \Rightarrow \bar{Y} = \hat{\beta_0} + \hat{\beta_1}\bar{X} \quad -①'$$

$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X}$$

② $\sum_{i=1}^{n} (Y_i - \hat{\beta_0} - \hat{\beta_1} X_i) X_i = 0$

$$\Rightarrow \sum_{i=1}^{n} X_i Y_i - \hat{\beta_0} \sum_{i=1}^{n} X_i - \hat{\beta_1} \sum_{i=1}^{n} X_i^2 = 0 \quad \text{plug in } ①'$$

$$\sum_{i=1}^{n} X_i Y_i - (\bar{Y} - \hat{\beta_1}\bar{X}) \sum_{i=1}^{n} X_i - \hat{\beta_1} \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i + \hat{\beta_1} \underbrace{\sum_{i=1}^{n} X_i}_{n} \sum_{i=1}^{n} X_i - \hat{\beta_1} \sum_{i=1}^{n} X_i^2 = 0 \quad \text{sub } \sum_{i=1}^{n} X_i = n\bar{X}$$

$$\sum_{i=1}^{n} Y_i = n\bar{Y}$$

$$\sum_{i=1}^{n} X_i Y_i - \bar{Y} \cdot n\bar{X} + \hat{\beta_1}(n\bar{X} \cdot n\bar{X} - \sum_{i=1}^{n} X_i^2) = 0$$

$$\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} = \hat{\beta_1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$$

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2} \quad \text{if } n \to \infty$$

$$\to \frac{E(X_i Y_i) - E(X_i)E(Y_i)}{E(X_i^2) - [E(X_i)]^2} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

✓

✡ Fit of the Model. Role of the R^2.

$SST = SSR + SSE$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2.$$

substitute $y_i - \hat{y}_i = \hat{e}_i$

$$= \sum_{i=1}^{n} (\hat{e}_i + \hat{y}_i - \bar{y})^2.$$

$$= \sum_{i=1}^{n} (\hat{e}_i^2 + 2 \cdot \hat{e}_i (\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2)$$

$$= \sum_{i=1}^{n} \hat{e}_i^2 + 2 \sum_{i=1}^{n} \hat{e}_i (\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$= SSE + 2 \sum_{i=1}^{n} \hat{e}_i \hat{y}_i - 2 \sum_{i=1}^{n} \hat{e}_i \bar{y} + SSR.$$

$\Downarrow$

b/c ② prediction & residual are orthogonal.

$\sum_{i=1}^{n} (\hat{e}_i) = 0$ Sum of residual $= 0$

So $SST = SSE + SSR$ is shown. ✡

Appendix: Taxonomy of Data, Cross-sectional v.s. Panel

Econometrics textbooks usually start with the introduction of cross-sectional data since the methodology, estimation, and analysis could be straight-forwardly constructed based on the Gauss-Markov assumptions. In the previous lectures, we had already gone through the main analysis methods, Ordinary Least Square (OLS) and learned how to make the estimators become unbiased and consistent. The specification (i.e. econometric model) is usually written as following, based on the assumptions that the data was **randomly selected** from the population **at the same time**. i = 1, 2,..., N could be states, cities, countries, individuals, etc.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + e_i$$

However, most of the institutions in charge of collecting the macroeconomic data usually keep track of their observations. For example, the World Bank would like to know each country's annual GDP, CPI, or currency exchange rate; and the Federal Reserve pays attention to the indices of the financial market such as interest rates and Yields to the Treasury Bonds day-by-day. Hence, following an identical group of sample creates another dimension of the existing variables in the OLS model, which is the time. Furthermore, usually, we will put another subscript t underneath the variables in the original OLS equation, and the finite time index is usually denoted by $t = 1, 2, ..., T$.

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + ... + \beta_k X_{kit} + e_{it}$$

However, what is wrong with the OLS? Aren't they just different in notations? Aren't they the same mathematically? What will happen if we run the Ordinary Least Square on panel data or applied the Weighted Least Square by using White's heteroskedasticity robustness correction?

The answer to the above is one of the key Gauss-Markov assumptions is no longer hold. Specifically, because the nature of panel data, the condition $\text{Cov}(e_i, e_j) = 0, i \neq j$ usually will be violated. If we use the notation in the panel data equation, $\text{Cov}(e_{it}, e_{is}) \neq 0$, and it is also the fundamental reason why we need to learn another method as well as some additional assumption on the error terms for the panel data or time-series data.

Let me use a simple reduced example to walk you through the idea. Assume we would like to know should the flash flood affects the economic activities in the DC area. For some reason, we could only get access to the GDP and RainFall data of the State of Virginia since 1989. Now, we should be able to construct a model that is similar to the following equation. $i = 1, t = 1989, 1990, ..., 2019$ ($N = 1$, $T = 20$). In other words, for the State of Virginia,

$$GDP_t = \beta_0 + \beta_1 RainFall_t + e_t$$

where

$$t = 1989(1), 1990(2), 1991(3), ..., 2018(20)$$

Then we can think about it. $RainFall_t$ for each year should be independent with each other, but what about the GDP_t? We have learned Real Business Cycle theory in Macroeconomics, so we should presume $\text{Cov}(GDP_t, GDP_s) \neq 0$, for $t \neq s$ or $t \to s$. Then if we plug in two data points of dependent variables $GDP_t = \beta_0 + \beta_1 RainFall_t + e_t$ and $GDP_s = \beta_0 + \beta_1 RainFall_s + e_s$.

$$\text{Cov}(GDP_t, GDP_s) \neq 0$$

$$\Rightarrow \text{Cov}(\beta_0 + \beta_1 RainFall_t + e_t, \beta_0 + \beta_1 RainFall_s + e_s) \neq 0$$

$$\Rightarrow \text{Cov}(e_t, \beta_0 + \beta_1 RainFall_s + e_s) \neq 0$$

$$\Rightarrow \text{Cov}(e_t, e_s) \neq 0$$

So now, you should be able to see what happened if we apply OLS to this data set. Since one of the important assumption of OLS is violated, the estimator of the coefficient will NOT be unbiased toward to the parameters, no matter how large the sample size is. [i.e. $E(b_1) \neq \beta_1$ $E(b_0) \neq \beta_0$] The whole estimation process becomes problematic and meaningless.