# Recitation 2: Ordinary least squares


Seung-hun Lee

Columbia University
Undergraduate Introduction to Econometrics Recitation


September 22nd, 2022

# Econometrics and RCT

# What is econometrics trying to achieve?

- Econometrics is a field in economics that tries to answer real life questions.
  - Ultimately, it is about making a **quantitative** statement about two or more random events
  - We can make **correlational** statements, but we want to identify *causal relationships*
- In order to achieve this goal, we collect data from a suitably defined population and use various methods to estimate a parameter that implies correlational/causal relationship.
- To fully understand what econometrics is trying to achieve, we need to ask ourselves these three questions
  - What is the difference between correlational and causal relation?
  - What is the suitably defined population?
  - What are the methods that we need to use in econometrics?

# Correlation vs Causation

- Suppose you have two random variables $X$ and $Y$. You want to identify if $X$ *causes* $Y$
- A **correlation** between $X$ and $Y$ is a statistical measure that describes how the two variables move together
  - It captures *any* type of statistical dependence that moves the two variables together: Causation, but also others too!
  - Not causal 1: $Y$ can cause $X$
  - Not causal 2: $X$ and $Y$ are jointly moving because there is $Z$ that affects both
- A **causal** relationship: *Cleanly (exogenously)* changing variables $X$ leads to changes in $Y$
  - Much more difficult: Changes in $X$ may be a combination of many things
  - Changes from $X$ alone and changes from other factors that may indirectly affect $X$
  - RCT: Isolates clean changes in $X$ that can help us tell whether changes in $X$ affects $Y$, and by how much
  - Econometrics: We can express concisely the relationship between $X$ and $Y$ variables in a single equation.

# Suitably defined population?

- When we say we are interested in the relationship between schooling and wages, whose effects are we interested in?
  - The entire US population, high school graduates, or college graduates?
  - Determines sampling methods we use to obtain a representative and comparable sample
  - Complete randomization, stratified randomization, or cluster randomization

- Note: We are almost surely never going to get the data from the entire population.
  - Gathering data from the entire population is logistically (and maybe ethically) difficult.
  - The estimate we are obtaining through any econometric exercise is thus a *sample analogue* of the actual value we are trying to get
  - We will do diagnostic tests to see if they can be reasonably close to the true value.

# What methods?

- In econometrics, we will use many estimation methods to obtain the sample analogue of the parameter of interest.
  - Ordinary least squares (OLS): Suitable for randomized control trials or in any case where the treatment assignment is as good as random.
  - Panel estimation: If data has multiple individuals and multiple time periods.
  - Instrumental variable methods: When we have proxy variables relevant to variable of interest and is reasonably exogenous
  - Difference-in-differences: When we study 'before & after' events with multiple entities
  - Regression discontinuity: In treatment with a cutoff determining treatment assignment
  - Time series: When we observe one entity over multiple periods

- Depending on the type of variables we use in our exercise, we have:
  - Univariate regression: One variable (besides an overall constant) controlled for
  - Multivariate regression: Multiple variables (besides an overall constant) controlled for
  - Nonlinear regression: Binary dependent variables
  - Big data methods

# Understanding RCTs

- In **randomized control trials**, we randomly categorize some individuals under treatment group and controlled group and run various tests
  - Benchmark for good program evaluation

- Potential outcomes framework
  - $Y_i$ : Observed outcome for individual $i \in \{1, ..., N\}$
  - $i$: Either in treatment or control group (not both)$\rightarrow$ $W_i = 1$ if $i$ is treated, 0 if otherwise
  - **W**: an $N$-tuple vector of treatment assignment for all individuals
  - Key assumption: Others' treatment assignment has no effect on my treatment (stable unit treatment value assumption (SUTVA))
  - Potential outcome $Y_i(w)$: Outcome for treated ($Y_i(1)$) and the untreated individual ($Y_i(0)$)
    - ★ Fundamental problem of missing data: Individual $i$ cannot have both $Y_i(1)$ and $Y_i(0)$ - at most one of them

# Potential vs observed outcome

- We can bridge the two with this relation

$$
\begin{aligned}
Y_i &= Y_i(1)W_i + Y_i(0)(1 - W_i) \\
&= Y_i(0) + W_i(Y_i(1) - Y_i(0))
\end{aligned}
$$

  - We know $W_i$ and $Y_i$ for everyone regardless of treatment assignment
  - We cannot see $Y_i(0)$ for the treated group and $Y_i(1)$ for the untreated group

# Treatment effect

- If we want to see if the treatment has any effect, we would ideally see

$$Y_i(1) - Y_i(0)$$

- But they cannot be obtained b/c fundamental problem of missing data
    - Alternative is average treatment effect or average treatment effect on the treated

$$\text{ATE} = E[Y_i(1) - Y_i(0)]$$
$$\text{ATT} = E[Y_i(1) - Y_i(0)|W_i = 1]$$

    - We also need assumptions about our treatment: Randomized assignment is one of them

$$E[Y_i(1)] = E[Y_i(1)|W_i = 1] = E[Y_i(1)|W_i = 0]$$
$$E[Y_i(0)] = E[Y_i(0)|W_i = 1] = E[Y_i(0)|W_i = 0]$$

# Obtaining ATE

- With this assumption and the definition of potential outcomes framework

$$\begin{aligned} E[Y_i|W_i] &= E[Y_i(1)W_i + Y_i(0)(1-W_i)|W_i] \\ &= E[Y_i(1)|W_i]W_i + E[Y_i(0)|W_i](1-W_i) \\ &= E[Y_i(1)]W_i + E[Y_i(0)](1-W_i) \end{aligned}$$

  - $W_i = 1$: Get $E[Y_i(1)] = E[Y_i|W_i = 1]$.
  - $W_i = 0$: Get $E[Y_i(0)] = E[Y_i|W_i = 0]$.

- Under random assignment, we can identify the average treatment effect as

$$\text{ATE} = E[Y_i(1)] - E[Y_i(0)] = E[Y_i|W_i = 1] - E[Y_i|W_i = 0]$$

- Econometrically: OLS with $W_i$ as independent variable (Problem set 2!)

# Ordinary least squares

# Ordinary Least Squares: Population vs sample linear models

- Suppose that the **population linear regression model** (also known as data generating process in some books) is
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
- However, we do not know the true values of the population parameters - $\beta_0$ and $\beta_1$
- An alternative way to approach the problem is to use the **sample linear regression model** (or just model)
$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$
where $\hat{\beta}_0, \hat{\beta}_1$ are estimates of $\beta_0, \beta_1$

# Ordinary Least Squares: Definition

- The ideal estimator minimizes the squared sum of residuals.
- Mathematically, this can be obtained by solving the following minimization problem and the first order conditions

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$$[\hat{\beta}_0] : -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$[\hat{\beta}_1] : -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

The resulting **least squares estimators** are

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

# Ordinary Least Squares: Main assumptions

- For OLS to be unbiased, consistent, efficient, and asymptotic normal, the following assumptions must be made

## Assumptions

**A0** Linearity: The regression is assumed to be linear in parameters.

$$\text{Okay: } Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i, \text{ Not: } Y_i = \beta_0 + \beta_1 X_i + \beta_2^2 X_i + u_i$$

**A1** $E(u_i|X_i) = 0$: Conditional on letting $X_i$ take a certain value, we are not making any systematical error in the linear regression. This is required for the OLS to be unbiased. (or $cov(X_i, u_i) = 0$)

**A2** i.i.d. (random sampling): $(X_i, Y_i)$ is assumed to be from independent, identical distribution

**A3** No Outliers: Outlier has no impact on the regression results. ($E(X_i^4), E(Y_i^4) < \infty$)

**A4** Homoskedasticity: $var(u_i) = \sigma_u$ (variance of $u_i$ does not depend on $X_i$). $\leftrightarrow$ *heteroskedasticity*

**A5** No Autocorrelation (Serial Correlation): For $i \neq j$, $cov(u_i, u_j) = 0$. Error at the previous period does not have any impact on the current period. This is usually broken in time series settings

# Ordinary Least Squares: Useful alternative expression for $\hat{\beta}_1$

- OLS estimate that we are getting is a random variable - getting different estimates depending on sample we work with.
- $\hat{\beta}_1$: Recall that we can write

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Now, replace $Y_i$ an $\bar{Y}$ with

$$Y_i = \beta_0 + \beta X_i + u_i, \ \bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u},$$

which allows us to write

$$(Y_i - \bar{Y}) = (\beta_1(X_i - \bar{X}) + (u_i - \bar{u}))$$

and get

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

# Ordinary Least Squares: Unbiasedness of $\hat{\beta}_1$

- $E[\hat{\beta}_1]$: It can be written as

$$
\begin{aligned}
E[\hat{\beta}_1] &= E\left[\beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]
\end{aligned}
$$

$\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})$ can be written to something simpler.

$$
\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n} X_i u_i - \bar{u}\sum_{i=1}^{n} X_i - \bar{X}\sum_{i=1}^{n} u_i + n\bar{X}\bar{u} = \sum_{i=1}^{n}(X_i - \bar{X})u_i
$$

→ Since $\bar{X}$ is a sample mean of $X$, $\sum_{i=1}^{n} X_i = n\bar{X}$.

→ The assumption that conditional mean is zero and $(X_i, u_i)$ are uncorrelated means that the term on the left hand side is zero.

→ Therefore, UNDER CLASSICAL ASSUMPTIONS, $E[\hat{\beta}_1] = \beta_1$.

# Ordinary Least Squares: Unbiasedness of $\hat{\beta}_0$

- $\hat{\beta}_0$: The formula for $\hat{\beta}_0$ is $\bar{Y} - \hat{\beta}_1 \bar{X}$. By changing $\bar{Y}$, we can get

$$\begin{aligned}\hat{\beta}_0 &= (\beta_0 + \beta_1 \bar{X} + \bar{u}) - \hat{\beta}_1 \bar{X} \\ &= \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{X} + \bar{u}\end{aligned}$$

  Then we can say the following about the sampling distribution

- $E[\hat{\beta}_0]$: We can write

$$E[\hat{\beta}_0] = \beta_0 + E[(\beta_1 - \hat{\beta}_1)\bar{X}] + E[\bar{u}] = \beta_0$$

  since $\hat{\beta}_1$ is unbiased and conditional expectation of $u_i$ is zero.

$\rightarrow$ Thus, under our current assumptions, $\hat{\beta}_0$ is unbiased.

# Ordinary Least Squares: Variances of $\hat{\beta}_0$ and $\hat{\beta}_1$

- Might take bit of a work, but when you follow the notes, you get

$$var(\hat{\beta}_0) = \frac{\sigma_u^2}{n} \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2}, \; var(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- At the end of the day, we can say

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$
$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma_u^2}{n} \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

- The importance of this is that now we can conduct a hypothesis test and create a test statistic based on this distribution

# Ordinary Least Squares: How well does the model capture the data?

Measure of fitness

- These numbers tell us how informative the sample linear regression we used is in telling us about the population data
- $R^2$: It is defined as a fraction of total variation which is explained by the model. Mathematically, this is

$$Y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_i}_{\hat{Y}_i} + u_i, \ \bar{Y} = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 \bar{X}}_{\bar{\hat{Y}}} + \bar{u},$$

$$\implies Y_i - \bar{Y} = (\hat{Y}_i - \bar{\hat{Y}}) - (u_i - \bar{u})$$

$$\implies \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}})^2 + \sum_{i=1}^{n} (u_i - \bar{u})^2 - 2 \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}})(u_i - \bar{u})$$

# Ordinary Least Squares: Getting to $R^2$

- Note that
$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})(u_i - \bar{u}) = \sum_{i=1}^{n}\hat{Y}_i u_i - \bar{\hat{Y}}\sum_{i=1}^{n}u_i - \bar{u}\sum_{i=1}^{n}\hat{Y}_i + n\bar{u}\bar{\hat{Y}}$$
- Since $\sum_{i=1}^{n} u_i = n\bar{u}$, $\sum_{i=1}^{n} \hat{Y}_i = n\bar{\hat{Y}}$ and $\sum_{i=1}^{n} \hat{Y}_i u_i = n\bar{u}\bar{\hat{Y}}$, all terms cancel each other out.
- So we are left with

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{TSS} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}_{ESS} + \underbrace{\sum_{i=1}^{n}(u_i - \bar{u})^2}_{RSS}$$
$$\implies 1 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} + \frac{\sum_{i=1}^{n}(u_i - \bar{u})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

# Ordinary Least Squares: Getting to $R^2$

- Thus, the $R^2$ can be found as

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- Intuitively, higher $R^2$ implies that the model explains more of the total variance, which implies that the regression fits the data well.