Lecture 2:Simple OLS Regression Estimation

Introduction to Econometrics

Zhaopeng Qu

Nanjing University Business School

March 08 2023

1. Review the previous lecture

2. Making Comparison Make Sense

3. OLS Estimation: Simple Regression

4. The Least Squares Assumptions

5. Properties of the OLS Estimators

6. Simple OLS and RCT

Review the previous lecture

Causal Inference and RCT

- **Causality** is our main goal in the studies of empirical social science.
- The existence of **selection bias** makes social science more difficult than science.
- Although RCTs is a powerful tool for economists, every project or topic can NOT be carried on by it.
- This is the reason why modern econometrics exists and develops. The main job of econometrics is using **non-experimental** data to *making convincing causal inference.*

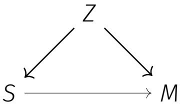
Furious Seven Weapons（七种武器）

- To build a *reasonable counterfactual world* or to find a *proper control group* is the core of econometric methods.
  1. Random Trials(**随机试验**)
  2. Regression(**回归**)
  3. Matching and Propensity Score（**匹配与倾向得分**）
  4. Decomposition（**分解**）
  5. Instrumental Variable（**工具变量**）
  6. Regression Discontinuity（**断点回归**）
  7. Panel Data and Difference in Differences（**双差分或倍差法**)
- The most basic of these tools is **regression**, which compares treatment and control subjects who have the same **observable** characteristics.
- Regression concepts are foundational, paving the way for the more elaborate tools used in the class that follow.
- *Let's start our exciting journey from it.*

Making Comparison Make Sense

Case: Smoke and Mortality

- Criticisms from **Ronald A. Fisher**
  - No experimental evidence to incriminate smoking as a cause of lung cancer or other serious disease.
  - Correlation between smoking and mortality may be spurious due to **biased selection** of subjects.



- **Confounder**, Z, creates backdoor path between smoking and mortality

Case: Smoke and Mortality(Cochran 1968)

Table 1: Death rates(死亡率) per 1,000 person-years

Smoking group	Canada	U.K.	U.S.
Non-smokers(不吸烟)	20.2	11.3	13.5
Cigarettes(香烟)	20.5	14.1	13.5
Cigars/pipes(雪茄/烟斗)	35.5	20.7	17.4

- It seems that taking cigars is more hazardous to the health?

Case: Smoke and Mortality(Cochran 1968)

Table 2: Non-smokers and smokers differ in age

Smoking group	Canada	U.K.	U.S.
Non-smokers(**不吸烟**)	54.9	49.1	57.0
Cigarettes(**香烟**)	50.5	49.8	53.2
Cigars/pipes(**雪茄**/**烟斗**)	65.9	55.7	59.7

- Older people die at a higher rate, and for reasons other than just smoking cigars.
- Maybe cigar smokers higher observed death rates is because **they're older on average**.

Case: Smoke and Mortality(Cochran 1968)

- The problem is that the age are *not balanced*, thus their mean values differ for treatment and control group.
- let's try to **balance** them, which means to compare mortality rates across the different smoking groups *within* age groups so as to neutralize age imbalances in the observed sample.
- It naturally relates to the concept of **Conditional Expectation Function**.

Case: Smoke and Mortality(Cochran 1968)

How to balance?

1. Divide the smoking group samples into age groups.
2. For each of the smoking group samples, calculate the mortality rates for the age group.
3. Construct probability weights for each age group as the proportion of the sample with a given age.
4. Compute the **weighted averages** of the age groups mortality rates for each smoking group using the probability weights.

Case: Smoke and Mortality(Cochran 1968)

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	0.15	11	29
Age 50-70	0.35	13	9
Age +70	0.5	16	2
Total		40	40

- **Question**: What is the average death rate for pipe smokers?

$$0.15 \cdot \left(\frac{11}{40}\right) + 0.35 \cdot \left(\frac{13}{40}\right) + 0.5 \cdot \left(\frac{16}{40}\right) = 0.355$$

Case: Smoke and Mortality(Cochran 1968)

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	0.15	11	29
Age 50-70	0.35	13	9
Age +70	0.5	16	2
Total		40	40

- **Question**: What would the average mortality rate be for pipe smokers if *they had the same age distribution as the non-smokers*?

$$0.15 \cdot \left(\frac{29}{40}\right) + 0.35 \cdot \left(\frac{9}{40}\right) + 0.5 \cdot \left(\frac{2}{40}\right) = 0.212$$

Case: Smoke and Mortality(Cochran 1968)

Table 3: Non-smokers and smokers differ in mortality and age

Smoking group	Canada	U.K.	U.S.
Non-smokers(**不吸烟**)	20.2	11.3	13.5
Cigarettes(**香烟**)	28.3	12.8	17.7
Cigars/pipes(**雪茄**/**烟斗**)	21.2	12.0	14.2

- **Conclusion**: It seems that taking cigarettes is most hazardous, and taking pipes is not different from non-smoking.

Formalization: Covariates

Definition: Covariates

Variable X is predetermined with respect to the treatment D if for each individual i, $X_i^0 = X_i^1$, i.e., the value of X_i does not depend on the value of D_i. Such characteristics are called *covariates*.

- Covariates are often time invariant (e.g., sex, race), but time invariance is not a necessary condition.

Identification under Independence

- Recall that randomization in RCTs implies

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

and therefore:

$$\begin{aligned}
E[Y|D=1] - E[Y|D=0] &= \underbrace{E[Y^1|D=1] - E[Y^0|D=0]}_{\text{by the switching equation}} \\
&= \underbrace{E[Y^1|D=1] - E[Y^0|D=1]}_{\text{by independence}} \\
&= \underbrace{E[Y^1 - Y^0|D=1]}_{\text{ATT}} \\
&= \underbrace{E[Y^1 - Y^0]}_{\text{ATE}}
\end{aligned}$$

Identification under Conditional Independence

- **Conditional Independence Assumption(CIA)**: which means that if we can “balance” covariates X then we can take the treatment D as randomized, thus

$$(Y^1, Y^0) \perp\!\!\!\perp D|X$$

- Now as $(Y^1, Y^0) \perp\!\!\!\perp D|X \not\Leftrightarrow (Y^1, Y^0) \perp\!\!\!\perp D$,

$$E[Y^1|D=1] - E[Y^0|D=0] \neq E[Y^1|D=1] - E[Y^0|D=1]$$

Identification under Conditional Independence(CIA)

- But using the CIA assumption, then

$$
\begin{aligned}
\underbrace{E[Y^1|D=1]-E[Y^0|D=0]}_{\text{association}} &= \underbrace{E[Y^1|D=1,X]-E[Y^0|D=0,X]}_{\text{conditional on covariates}} \\
&= \underbrace{E[Y^1|D=1,X]-E[Y^0|D=1,X]}_{\text{conditional independence}} \\
&= \underbrace{E[Y^1-Y^0|D=1,X]}_{\text{conditional ATT}} \\
&= \underbrace{E[Y^1-Y^0|X]}_{\text{conditional ATE}}
\end{aligned}
$$

Curse of Multiple Dimensionality

- Sub-classification in one or two dimensions as Cochran(1968) did in the case of *Smoke and Mortality* is feasible.
- But as the number of covariates we would like to balance grows(like many personal characteristics such as age, gender,education,working experience,married,industries,income,...), then method become less feasible.
- Assume we have k covariates and we divide each into 3 coarse categories (e.g., age: young, middle age, old; income: low,medium, high, etc.)
- The number of cells(or groups)is 3^K.
    - If $k = 10$ then $3^{10} = 59049$

Make Comparison Make Sense

- Selection on Observables
    - Regression
    - Matching
- Selection on Unobservables
    - IV,RD,DID,FE and SCM.
- The most basic of these tools is **regression**, which compares treatment and control subjects who have the same **observable** characteristics.
- Regression concepts is foundational, paving the way for the more elaborate tools used in the class that follow.

OLS Estimation: Simple Regression

Question: Class Size and Student's Performance

- **Specific Question**:

  What is the effect on district **test scores** if we would increase district average **class size** by 1 student (or one unit of Student-Teacher's Ratio)

- If we could know the full relationship between two variables which can be summarized by a real value function,$f()$

$$Testscore = f(ClassSize)$$

- Unfortunately, the function form is always unknown.

Question: Class Size and Student's Performance

- Two basic methods to describe the function.
  - **non-parametric**: we don't care the specific form of the function, unless we know all the values of two variables, which actually are the *whole distributions* of class size and test scores.
  - **parametric**: we have to suppose the basic form of the function, then to find values of some *unknown parameters* to determine the specific function form.
- Both methods need to use **samples** to inference **populations** in our random and unknown world.

Question: Class Size and Student's Performance

- Suppose we choose *parametric* method, then we just need to know the real value of a **parameter** β_1 to describe the relationship between Class Size and Test Scores

$$\beta_1 = \frac{\Delta Testscore}{\Delta ClassSize}$$

- Next step, we have to suppose specific forms of the function$f()$, still two categories: linear and non-linear
- And we start to use a *simplest* function form: a **linear** equation, which is graphically a straight line, to summarize the relationship between two variables.

$$Test\ score = \beta_0 + \beta_1 \times Class\ size$$

where β_1 is actually the **the slope** and β_0 is the **intercept** of the straight line.

Class Size and Student's Performance

- BUT the average test score in district *i* does not **only** depend on the average class size
- It also depends on **other factors** such as
  - Student background
  - Quality of the teachers
  - School's facilitates
  - Quality of text books
  - Random deviation......
- So the equation describing the linear relation between Test score and Class size is **better** written as

$$Test\,score_i = \beta_0 + \beta_1 \times Class\,size_i + u_i$$

where u_i lumps together all **other factors** that affect average test scores.

Terminology for Simple Regression Model

- The linear regression model with one regressor is denoted by

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Where
  - Y_i is the **dependent variable**(Test Score)
  - X_i is the **independent variable** or regressor(Class Size or Student-Teacher Ratio)
  - $\beta_0 + \beta_1 X_i$ is the **population regression line** or the **population regression function**

Population Regression: relationship in average

- The linear regression model with one regressor is denoted by

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Both side to conditional on X, then

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i + E[u_i|X_i]$$

- Suppose $E[u_i|X_i] = 0$ then

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

- Population regression function is the relationship that holds between Y and X **on average over the population**.

Terminology for Simple Regression Model

- The intercept β_0 and the slope β_1 are the **coefficients** of the **population regression line**, also known as the **parameters** of the population regression line.
- u_i is the **error term** which contains all the other factors *besides* X that determine the value of the dependent variable, Y, for a specific observation, i.

Graphics for Simple Regression Model

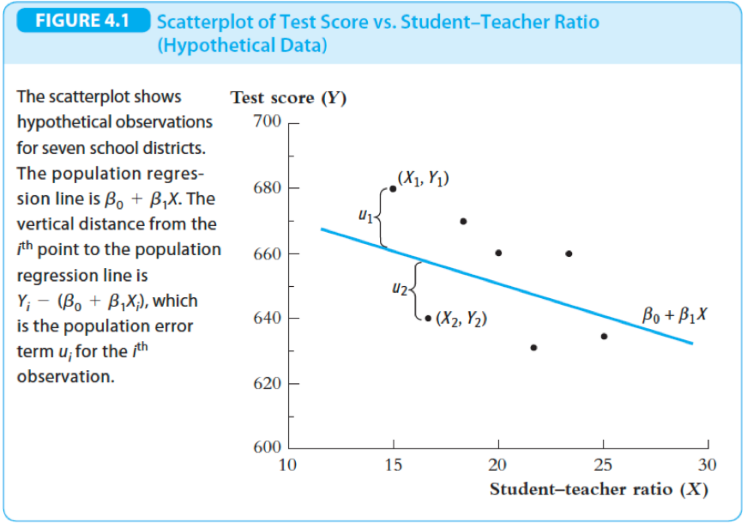


FIGURE 4.1 Scatterplot of Test Score vs. Student–Teacher Ratio (Hypothetical Data)

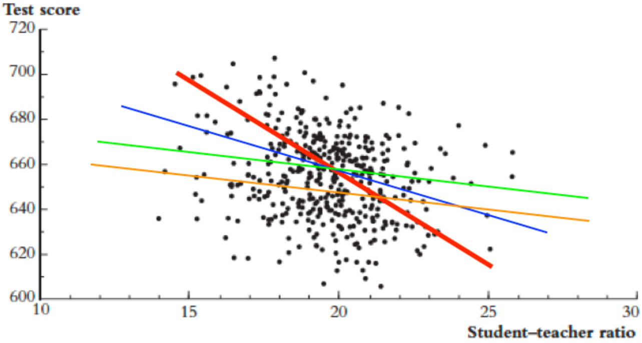
The scatterplot shows hypothetical observations for seven school districts. The population regression line is $\beta_0 + \beta_1 X$. The vertical distance from the i^{th} point to the population regression line is $Y_i - (\beta_0 + \beta_1 X_i)$, which is the population error term u_i for the i^{th} observation.

How to find the “best” fitting line?

- In general we don’t know β_0 and β_1 which are parameters of *population regression function*.We have to calculate them using a bunch of data: **the sample**.

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.



The Ordinary Least Squares Estimator (OLS)

The OLS estimator

- Chooses the **best** regression coefficients so that the estimated regression line is as close as possible to the observed data, where closeness is measured by *the sum of the squared mistakes* made in predicting Y given X.
- Let b_0 and b_1 be estimators of β_0 and β_1,thus $b_0 \equiv \hat{\beta}_0$,$b_1 \equiv \hat{\beta}_1$
- The predicted value of Y_i given X_i using these estimators is $b_0 + b_1 X_i$, or $\hat{\beta}_0 + \hat{\beta}_1 X_i$ formally denotes as $\hat{Y}_i$

The Ordinary Least Squares Estimator (OLS)

The OLS estimator

- The prediction mistake is the difference between Y_i and $\hat{Y}_i$,which denotes as $\hat{u}_i$

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$$

- The estimators of the slope and intercept that *minimize the sum of the squares* of $\hat{u}_i$,thus

$$\underset{b_0, b_1}{arg\ min} \sum_{i=1}^{n} \hat{u}_i^2 = \underset{b_0, b_1}{min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

are called the **ordinary least squares (OLS) estimators** of β_0 and β_1.

The Ordinary Least Squares Estimator (OLS)

- OLS minimizes sum of squared prediction mistakes:

$$\underset{b_0,b_1}{min}\sum_{i=1}^{n}\hat{u}_i^2=\sum_{i=1}^{n}(Y_i-b_0-b_1X_i)^2$$

- Solve the problem by **F.O.C**(the first order condition)
  - Step 1 for β_0:

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i-b_0-b_1X_i)^2=0$$

  - Step 2 for β_1:

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i-b_0-b_1X_i)^2=0$$

Step 1: OLS estimator of β_0

- Recall the sample mean of Y_i is

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

- Optimization

$$\begin{aligned}
\frac{\partial}{\partial b_0} \sum_{i=1}^{n} \hat{u}_i^2 &= -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0 \\
&\Rightarrow \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1 X_i = 0 \\
&\Rightarrow \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{1}{n} \sum_{i=1}^{n} b_0 - b_1 \frac{1}{n} \sum_{i=1}^{n} X_i = 0 \\
&\Rightarrow \overline{Y} - b_0 - b_1 \overline{X} = 0
\end{aligned}$$

Step 1: OLS estimator of β_0

OLS estimator of β_0:

$$b_0 \equiv \hat{\beta}_0 = \overline{Y} - b_1\overline{X}$$

Step 2: OLS estimator of β_1

$$
\begin{aligned}
\frac{\partial}{\partial b_1}\sum_{i=1}^{n}\hat{u}_i^2 &= -2\sum_{i=1}^{n}X_i\big(Y_i - b_0 - b_1X_i\big) = 0 \\
&\Rightarrow \sum_{i=1}^{n}X_i\big[Y_i - \big(\overline{Y} - b_1\overline{X}\big) - b_1X_i\big] = 0 \\
&\Rightarrow \sum_{i=1}^{n}X_i\big[\big(Y_i - \overline{Y}\big) - b_1\big(X_i - \overline{X}\big)\big] = 0 \\
&\Rightarrow \sum_{i=1}^{n}X_i\big(Y_i - \overline{Y}\big) - b_1\sum_{i=1}^{n}X_i\big(X_i - \overline{X}\big) = 0
\end{aligned}
$$

Step 2: OLS estimator of β_1

- Some Algebraic Facts

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y}) &= \sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}X_i\overline{Y}-\sum_{i=1}^{n}\overline{X}Y_i+\sum_{i=1}^{n}\overline{XY}\\ &= \sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}X_i\overline{Y}-n\overline{X}(\frac{1}{n}\sum_{i=1}^{n}Y_i)+n\overline{XY}\\ &= \sum_{i=1}^{n}X_i(Y_i-\overline{Y})\end{aligned}$$

- By a similar reasoning, we could obtain

$$\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})=\sum_{i=1}^{n}X_i(X_i-\overline{X})=\sum_{i=1}^{n}(X_i-\overline{X})X_i$$

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

Step 2: OLS estimator of β_1

- Thus

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y}) - b_1\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X}) = 0$$

OLS estimator of β_1:

$$b_1 \equiv \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})}$$

Some Algebraic of $\hat{u}_i$

- Recall the F.O.C

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = 0$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = 0$$

- We obtain two intermediate formulas

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) = 0$$

$$\sum_{i=1}^{n}X_i(Y_i - b_0 - b_1X_i) = 0$$

Some Algebraic of $\hat{u}_i$

- Recall the OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$
$$\hat{u}_i = Y_i - \hat{Y}_i$$

- Then we have(*prove them by yourself,Appendix 4.3 in SW,pp184-185*)

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

The Estimated Regression Line

- Obtain the values of OLS estimator for a certain data,

$$\hat{\beta}_1 = -2.28 \ and \ \hat{\beta}_0 = 698.9$$

- Then the regression line is

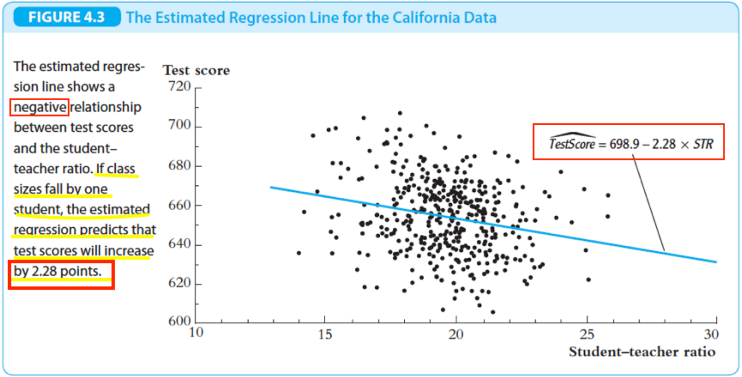


FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student–teacher ratio. If class sizes fall by one student, the estimated regression predicts that test scores will increase by 2.28 points.

Measures of Fit: The R^2

- Because the variation of Y can be summarized by a statistic: **Variance**,so the total variation of Y_i, which are also called as the **total sum of squares** (TSS), is:

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- Because Y_i can be decomposed into the fitted value plus the residual: $Y_i = \hat{Y}_i + \hat{u}_i$,then likewise Y_i, we can obtain
  - The **explained sum of squares** (ESS): $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$
  - The **sum of squared residuals** (SSR): $\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 = \sum_{i=1}^{n}\hat{u}_i^2$
- And more importantly, the variation of Y_i should be a sum of the variations of $\hat{Y}_i$ and $\hat{u}_i$, thus

$$TSS = ESS + SSR$$

Measures of Fit: The R^2

R^2 or the coefficient of determination

R^2 or the coefficient of determination, is the fraction of the sample variance of Y_i explained/predicted by X_i

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- So $0 \leq R^2 \leq 1$, it measures that how much can the variations of Y be explained by the variations of X_i in share.
- **NOTICE**: It seems that *R-squares is bigger, the regression is better*, which is wrong in most cases. Because we **DON'T** care much about R^2 when we make *causal inference* about two variables.

The Standard Error of the Regression

- We would also like to know some characteristics of u_i,but u_i are totally unobserved. We have to use the sample statistic to inference the population.
- The **standard error of the regression** (SER) is an *estimator* of the standard deviation of the regression error u_i.
- The **SER** is computed using their sample counterparts, the OLS residuals $\hat{u}_i$,thus

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2}$$

where $s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$

- Think about it: why the denominator is $n - 2$?

The Least Squares Assumptions

Assumption of the Linear regression model

- In order to investigate the statistical properties of OLS, we need to make some statistical assumptions

Linear Regression Model

The observations, (Y_i, X_i) come from a random sample(i.i.d) and satisfy the linear regression equation,

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

and $E[u_i \mid X_i] = 0$

Assumption 1: Conditional Mean is Zero

Assumption 1: Zero conditional mean of the errors given X

The error,u_i has expected value of 0 given any value of the independent variable

$$E[u_i \mid X_i = x] = 0$$

- An *weaker* condition that u_i and X_i are uncorrelated:

$$Cov[u_i, X_i] = E[u_i X_i] = 0$$

- if both are correlated, then Assumption 1 is violated.
- Equivalently, the population regression line is the conditional mean of Y_i given X_i , thus

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

Assumption 1: Conditional Mean is Zero

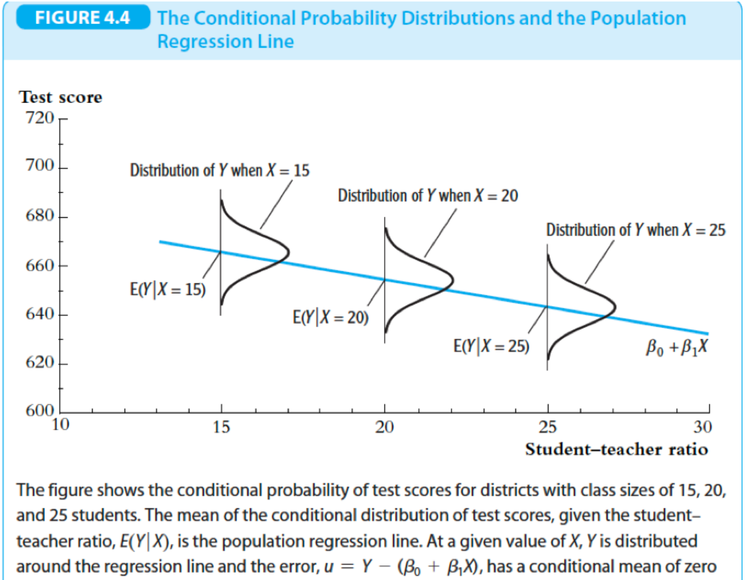


FIGURE 4.4 The Conditional Probability Distributions and the Population Regression Line

The figure shows the conditional probability of test scores for districts with class sizes of 15, 20, and 25 students. The mean of the conditional distribution of test scores, given the student–teacher ratio, $E(Y|X)$, is the population regression line. At a given value of X, Y is distributed around the regression line and the error, $u = Y - (\beta_0 + \beta_1 X)$, has a conditional mean of zero

Assumption 2: Random Sample

Assumption 2: Random Sample

We have a i.i.d random sample of size , $\{(X_i, Y_i), i = 1, ..., n\}$ from the population regression model above.

- This is an implication of random sampling. Then we have such as

$$Cov(X_i, X_j) = 0$$
$$Cov(Y_i, X_j) = 0$$
$$Cov(u_i, X_j) = 0$$

- And it generally won't hold in other data structures.
  - time-series, cluster samples and spatial data.

Assumption 3: Large outliers are unlikely

Assumption 3: Large outliers are unlikely

It states that observations with values of X_i, Y_i or both that are far outside the usual range of the data(Outlier) are unlikely. Mathematically, it assume that X and Y have nonzero finite fourth moments.

- Large outliers can make OLS regression results misleading.
- One source of large outliers is data entry errors, such as a typographical error or incorrectly using different units for different observations.
- Data entry errors aside, the assumption of finite kurtosis is a plausible one in many applications with economic data.

Assumption 3: Large outliers are unlikely

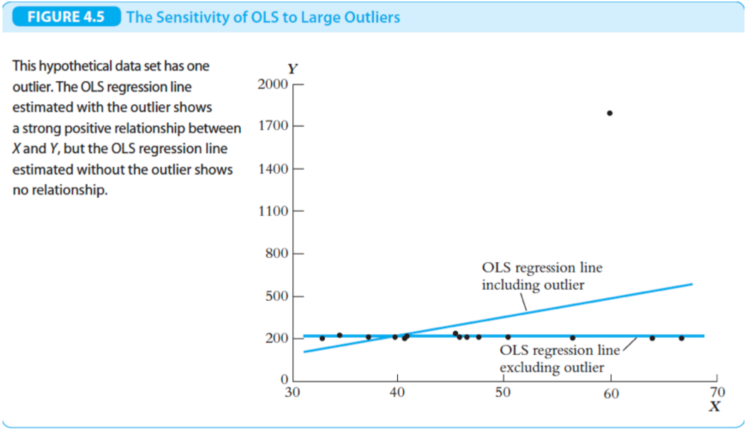


FIGURE 4.5 The Sensitivity of OLS to Large Outliers

This hypothetical data set has one outlier. The OLS regression line estimated with the outlier shows a strong positive relationship between *X* and *Y*, but the OLS regression line estimated without the outlier shows no relationship.

Underlying Assumptions of OLS

- The OLS estimator is **unbiased**, **consistent** and has **asymptotically normal sampling distribution** if
  1. Random sampling.
  2. Large outliers are unlikely.
  3. The conditional mean of u_i given X_i is zero

Underlying assumptions of OLS

- OLS is an **estimator**: it's a machine that we plug data into and we get out estimates.
- It has a **sampling distribution**, with a sampling variance/standard error, etc. like the sample mean, sample difference in means, or the sample variance.
- Let's discuss these characteristics of OLS in the next section.

Properties of the OLS Estimators

The OLS estimators

- Question of interest: What is the effect of a change in X_i(Class Size) on Y_i(Test Score)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- We derived the OLS estimators of β_0 andβ_1:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})(X_i - \bar{X})}$$

Least Squares Assumptions

1. Assumption 1: Conditional Mean is Zero
2. Assumption 2: Random Sample
3. Assumption 3: Large outliers are unlikely

- If the 3 least squares assumptions hold the OLS estimators will be
    - **unbiased**
    - **consistent**
    - **normal sampling distribution**

Properties of the OLS estimator: unbiasedness

- Recall:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- take expectation to β_0 :

$$E[\hat{\beta}_0] = \overline{Y} - E[\hat{\beta}_1]\overline{X}$$

- Then we have: if β_1 is unbiased, then β_0 is also unbiased.

Properties of the OLS estimator: unbiasedness

- Remind we have

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\overline{Y} = \beta_0 + \beta_1 \overline{X} + \overline{u}$$

- So take expectation to β_1:

$$E[\hat{\beta}_1] = E\left[\frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})(X_i - \overline{X})}\right]$$

Properties of the OLS estimator: unbiasedness

- Continued

$$
\begin{aligned}
E[\hat{\beta}_1] &= E\left[\frac{\sum(X_i - \bar{X})(\beta_0 + \beta_1 X_i + u_i - (\beta_0 + \beta_1 \bar{X} + \bar{u}))}{\sum(X_i - \bar{X})(X_i - \bar{X})}\right] \\
&= E\left[\frac{\sum(X_i - \bar{X})(\beta_1(X_i - \bar{X}) + (u_i - \bar{u}))}{\sum(X_i - \bar{X})(X_i - \bar{X})}\right] \\
&= E\left[\frac{\beta_1 \sum(X_i - \bar{X})(X_i - \bar{X})}{\sum(X_i - \bar{X})(X_i - \bar{X})}\right] + E\left[\frac{(X_i - \bar{X}) + (u_i - \bar{u})}{\sum(X_i - \bar{X})(X_i - \bar{X})}\right] \\
&= \beta_1 + E\left[\frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})(X_i - \bar{X})}\right]
\end{aligned}
$$

Review: Conditional Expectation Function(CEF)

- Expectation(for a continuous r.v.)

$$E(y) = \int yf(y)dy$$

- Conditional probability density function

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

- Conditional Expectation Function: the Expectation of Y conditional on X is

$$E(y|x) = \int yf_{Y|X}(y|x)dy$$

Review: Properties of CEF

Let X, Y, Z are random variables; $a, b \in \mathbb{R}$; $g(\cdot)$ is a real valued function, then we have

- $E[a \mid Y] = a$
- $E[(aX + bZ) \mid Y] = aE[X \mid Y] + bE[Z \mid Y]$
- If X and Y are independent, then $E[Y \mid X] = E[X]$
- $E[Yg(X) \mid X] = g(X)E[Y \mid X]$. In particular, $E[g(Y) \mid Y] = g(Y)$

Review: the Law of Iterated Expectations(LIE)

the Law of Iterated Expectations

It states that an unconditional expectation can be written as the unconditional average of conditional expectation function.

$$E(Y_i) = E[E(Y_i|X_i)]$$

and it can easily extend to

$$E(g(X_i)Y_i) = E[E(g(X_i)Y_i|X_i)] = E[g(X_i)E(Y_i|X_i)]$$

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

$$\begin{aligned}
E[E(Y|X)] &= \int E(Y|X=u)f_X(u)du \\
&= \int \Big[\int tf_Y(t|X=u)dt\Big] f_X(u)du \\
&= \int\int tf_Y(t|X=u)f_X(u)dtdu \\
&= \int t\Big[\int f_Y(t|X=u)f_X(u)du\Big]dt \\
&= \int t\Big[\int f_{XY}(u,t)du\Big]dt \\
&= \int tf_y(t)dt \\
&= E(Y)
\end{aligned}$$

Properties of the OLS estimator: unbiasedness

- Because $\sum(X_i - \overline{X})(u_i - \overline{u}) = \sum(X_i - \overline{X})u_i$, so

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1 + E\left[\frac{\sum(X_i - \overline{X})u_i}{\sum(X_i - \overline{X})(X_i - \overline{X})}\right] \\ &= \beta_1 + E\Big[\sum g(X_i)u_i\Big] \\ &= \beta_1 + \sum E\big[g(X_i)u_i\big] \\ &= \beta_1 + \sum E\big[g(X_i)E[u_i|X_i]\big] \end{aligned}$$

where $g(X_i) = \frac{(X_i - \overline{X})}{\sum(X_i - \overline{X})(X_i - \overline{X})}$

- Then we can obtain

$$E\big[\hat{\beta}_1\big] = \beta_1 \; if \; E[u_i|X_i] = 0$$

- Both β_0 and β_1 are **unbiased** on the condition of **Assumption 1**.

Properties of the OLS estimator: Consistency

- **Notation**: $\hat{\beta}_1 \xrightarrow{p} \beta_1$ or $plim \hat{\beta}_1 = \beta_1$, so

$$plim \hat{\beta}_1 = plim \left[\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})(X_i - \bar{X})} \right]$$

- Then we could obtain

$$plim \hat{\beta}_1 = plim \left[\frac{\frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1} \sum (X_i - \bar{X})(X_i - \bar{X})} \right] = plim \left(\frac{s_{xy}}{s_x^2} \right)$$

where s_{xy} and s_x^2 are sample covariance and sample variance.

Math Review: Continuous Mapping Theorem

- **Continuous Mapping Theorem**: For every continuous function $g(t)$ and random variable X:

$$plim(g(X)) = g(plim(X))$$

- Example:

$$plim(X + Y) = plim(X) + plim(Y)$$

$$plim(\frac{X}{Y}) = \frac{plim(X)}{plim(Y)} \ if \ plim(Y) \neq 0$$

Properties of the OLS estimator: Consistency

- Base on L.L.N(the law of large numbers) and random sample(i.i.d)

$$s_X^2 \xrightarrow{p} = \sigma_X^2 = Var(X)$$

$$s_{xy} \xrightarrow{p} \sigma_{XY} = Cov(X, Y)$$

- Combining with Continuous Mapping Theorem,then we obtain the OLS estimator $\hat{\beta}_1$,when $n \longrightarrow \infty$

$$plim \hat{\beta}_1 = plim\left(\frac{s_{xy}}{s_x^2}\right) = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

Properties of the OLS estimator: Consistency

$$
\begin{aligned}
plim\hat{\beta}_1 &= \frac{Cov(X_i, Y_i)}{Var(X_i)} \\
&= \frac{Cov(X_i, (\beta_0 + \beta_1 X_i + u_i))}{Var(X_i)} \\
&= \frac{Cov(X_i, \beta_0) + \beta_1 Cov(X_i, X_i) + Cov(X_i, u_i)}{Var(X_i)} \\
&= \beta_1 + \frac{Cov(X_i, u_i)}{Var(X_i)}
\end{aligned}
$$

- Then we could obtain

$$plim\hat{\beta}_1 = \beta_1 \; if \; E[u_i|X_i] = 0$$

Wrap Up: Unbiasedness vs Consistency

- **Unbiasedness** & **Consistency** both rely on $E[u_i|X_i] = 0$
- **Unbiasedness** implies that $E\left[\hat{\beta}_1\right] = \beta_1$ for a certain sample size n.("small sample")
- **Consistency** implies that the distribution of $\hat{\beta}_1$ becomes more and more _tightly distributed around β_1 if the sample size n becomes larger and larger.("large sample"")
- Additionally,you could prove that $\hat{\beta}_0$ is likewise **Unbiased** and **Consistent** on the condition of **Assumption 1**.

Sampling Distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$: Recalll of $\overline{Y}$

- Firstly, Let's recall: Sampling Distribution of $\overline{Y}$
- Because $Y_1, \ldots, Y_n$ are i.i.d. and μ_Y is the mean of the population,then for L.L.N,we have

$$E(\overline{Y}) = \mu_Y$$

- Based on the Central Limit theorem(C.L.T) and the σ_Y^2 is the variance of the population, the sample distribution in a large sample can *approximates to a normal distribution*, thus

$$\overline{Y} \sim N(\mu_Y, \frac{\sigma_Y^2}{n})$$

- Therefore, the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ could have similar sample distributions *when three least squares assumptions hold.*

Sampling Distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$: Expectation

- Unbiasedness of the OLS estimators implies that

$$E[\hat{\beta}_1] = \beta_1 \textit{ and } E[\hat{\beta}_0] = \beta_0$$

- Likewise as $\overline{Y}$,the sample distribution of β_1 in a large sample can also *approximates to a normal distribution* based on the **Central Limit theorem(C.L.T)**, thus

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$
$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0})$$

- Where it can be shown that

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{Var[(X_i - \mu_x)u_i]}{[Var(X_i)]^2})$$
$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{Var(H_i u_i)}{(E[H_i^2])^2})$$

Sampling Distribution of $\hat{\beta}_1$

- $\hat{\beta}_1$ in terms of regression and errors in following equation

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

$$= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})}$$

Sampling Distribution of $\hat{\beta}_1$:the numerator

- The numerator: $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u})$
- Because $\overline{X}$ is consistent, thus $\overline{X} \xrightarrow{p} \mu_x$,then combine with Continuous Mapping Theorem

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

$$\Longrightarrow \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}) \xrightarrow{p} \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_x)u_i$$

Sampling Distribution of $\hat{\beta}_1$:the numerator

- Let $v_i = (X_i - \mu_x)u_i$
  - Based on **Assumption 1**, then $E(v_i) = 0$
  - Based on **Assumption 2**, $\sigma_v^2 = Var[(X_i - \mu_x)u_i]$
- Then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_x)u_i = \frac{1}{n}\sum_{i=1}^{n} v_i = \bar{v}$$

Sampling Distribution of $\hat{\beta}_1$:the numerator

- Recall: $\overline{Y}$ to Y_i and based on C.L.T,

$$\frac{\overline{Y}-0}{\sigma_{\overline{Y}}} \xrightarrow{d} N(0.1) \; or \; \overline{Y} \xrightarrow{d} N(0,\frac{\sigma_Y^2}{n})$$

- The $\overline{v}$ is the *sample mean* of v_i,based on **C.L.T**,

$$\frac{\overline{v}-0}{\sigma_{\overline{v}}} \xrightarrow{d} N(0.1) \; or \; \overline{v} \xrightarrow{d} N(0,\frac{\sigma_v^2}{n})$$

Sampling Distribution of $\hat{\beta}_1$:the denominator

- Recall the sample variance of X_i is $s^2_{X_i}$

$$s^2_{X_i} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- Then the denominator,is a variation of **sample variance** of X (except dividing by n rather than $n-1$, which is *inconsequential if n is large*)

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})$$

- Based on discussion of the *sample variance* is a **consistent** estimator of the *population variance*,thus

$$s^2_{X_i} \xrightarrow{p} Var[X_i] = \sigma^2_{X_i}$$

Sampling Distribution of $\hat{\beta}_1$

- $\hat{\beta}_1$ in terms of regression and errors

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

- the numerator is $\overline{v}$ and $\overline{v} \xrightarrow{d} N(0, \frac{\sigma_v^2}{n})$
- the denominator is

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X}) \xrightarrow{p} Var[X_i] = \sigma_{X_i}^2$$

- Combining these two results, we have that, *in large samples*

$$\hat{\beta}_1 - \beta_1 \xrightarrow{p} \frac{\overline{v}}{Var[X_i]}$$

Slutsky's Theorem

- It combines consistency and convergence in distribution.

Slutsky's Theorem

Suppose that $a_n \xrightarrow{p} a$,where a is a constant,and $S_n \xrightarrow{d} S$.Then

$$a_n + S_n \xrightarrow{d} a + S$$

$$a_n S_n \xrightarrow{d} aS$$

$$\frac{S_n}{a_n} \xrightarrow{d} \frac{S}{a} \; if \; a \neq 0$$

Sampling Distribution of $\hat{\beta}_1$

- Based on $\bar{v}$ follow a normal distribution, in large samples,thus

$$\bar{v} \xrightarrow{d} N(0, \frac{\sigma_v^2}{n})$$

$$\Rightarrow \frac{\bar{v}}{Var[X_i]} \xrightarrow{d} N\left(0, \frac{\sigma_v^2}{n[Var(X_i)]^2}\right)$$

$$\Rightarrow \hat{\beta}_1 - \beta_1 \xrightarrow{d} N\left(0, \frac{\sigma_v^2}{n[Var(X_i)]^2}\right)$$

- Then the sampling distribution of $\hat{\beta}_1$ is

$$\hat{\beta}_1 \xrightarrow{d} N(\beta_1, \sigma_{\hat{\beta}_1}^2)$$

where

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma_v^2}{n[Var(X_i)]^2} = \frac{Var[(X_i - \mu_x)u_i]}{n[Var(X_i)]^2}$$

Sampling Distribution $\hat{\beta}_1$ in large-sample

- We have shown that

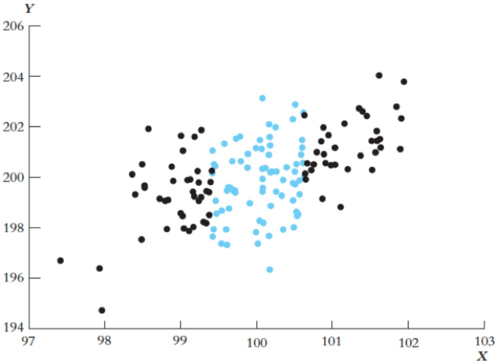
$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{Var[(X_i-\mu_x)u_i]}{[Var(X_i)]^2})$$

- An intuition： The **variation** of X_i is very important.
  - Because if $Var(X_i)$ is *small*, it is difficult to obtain an accurate estimate of the effect of X on Y which implies that $Var(\hat{\beta}_1)$ is *large*.

Variation of X

FIGURE 4.6 The Variance of $\hat{\beta}_1$ and the Variance of X

The colored dots represent a set of X_i's with a small variance. The black dots represent a set of X_i's with a large variance. The regression line can be estimated more accurately with the black dots than with the colored dots.



- When more **variation** in X_i, then there is more information in the data that you can use to fit the regression line.

In a Summary

Under 3 least squares assumptions, the OLS estimators will be

- **unbiased**
- **consistent**
- **normal sampling distribution**
- *more variation in X, more accurate estimation*

Simple OLS and RCT

OLS Regression and RCT

- We learned RCT is the **“golden standard”** for causal inference.Because it can naturally eliminate **selection bias**.
- So far, we did not discuss the relationship between RCT and OLS regression, which means that we can not be sure that the result from an OLS regression can be explained as “causal”.
- Instead of using a continuous regressor X, the regression where D_i is a binary variable, a so-called **dummy variable**, will help us to unveil the relationship between RCT and OLS regression.

Regression when X is a Binary Variable

- For example, we may define D_i as follows:

$$D_i = \begin{cases} 1 \ \text{ if } STR \text{ in } i^{th} \text{ school district} < 20 \\ 0 \ \text{ if } STR \text{ in } i^{th} \text{ school district} \geq 20 \end{cases} \tag{4.2}$$

- The regression can be written as

$$Y_i = \beta_0 + \beta_1 D_i + u_i \tag{4.1}$$

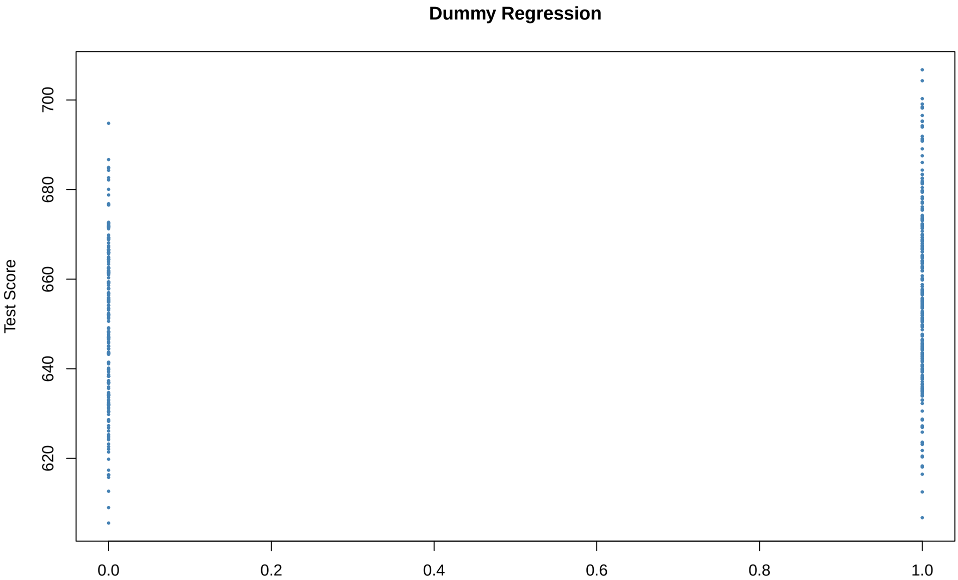
Regression when X is a Binary Variable

- More precisely, the regression model now is

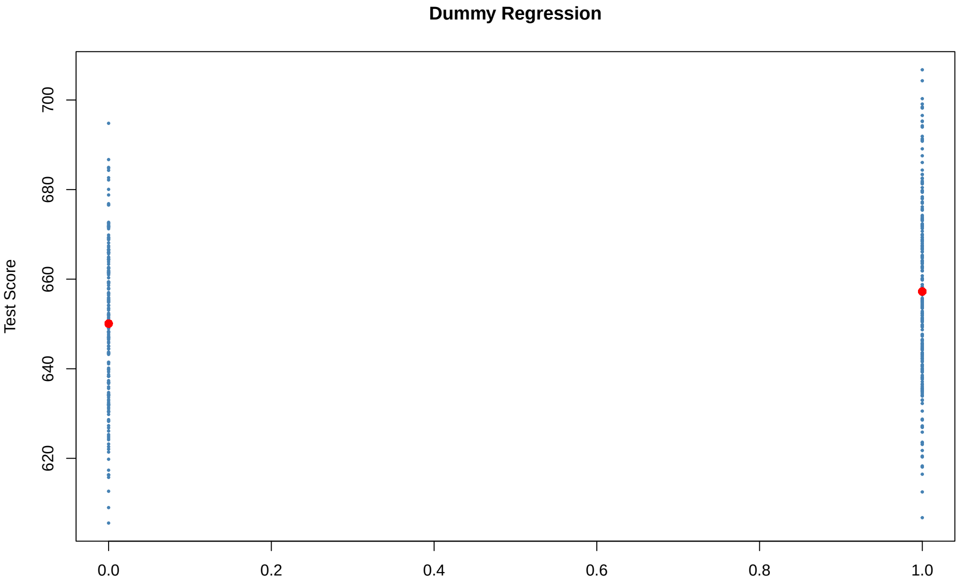
$$TestScore_i = \beta_0 + \beta_1 D_i + u_i \tag{4.3}$$

  - With D as the regressor, it is not useful to think of β_1 as a slope parameter.
  - Since $D_i \in \{0, 1\}$, i.e., we only observe two discrete values instead of a continuum of regressor values.
- There is no continuous line depicting the conditional expectation function $E\left(TestScore_i \middle| D_i\right)$ since this function is solely defined for x-positions 0 and 1.

Class Size and STR



Class Size and STR



Regression when X is a Binary Variable

- Therefore, the interpretation of the coefficients in this regression model is as follows:
  - $E(Y_i|D_i = 0) = \beta_0$, so β_0 is the expected test score in districts where $D_i = 0$ where STR is below 20.
  - $E(Y_i|D_i = 1) = \beta_0 + \beta_1$ where STR is above 20
- Thus, β_1 is **the difference in group specific expectations**, i.e., the difference in expected test score between districts with $STR < 20$ and those with $STR \geq 20$,

$$\beta_1 = E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$$

.

Causality and OLS

- Let us recall, the individual treatment effect

$$ICE = Y_{1i} - Y_{0i} = \rho \quad \forall i$$

- then we can rewrite

$$\mathrm{Y}_i = \mathrm{Y}_{0i} + \mathrm{D}_i \left(\mathrm{Y}_{1i} - \mathrm{Y}_{0i}\right)$$

Causality and OLS

- Regression function is

$$\mathsf{Y}_i = \alpha + \mathsf{D}_i \rho + \eta_i$$

- Further

$$\mathsf{Y}_i = \underbrace{\alpha}_{E[\mathsf{Y}_{0i}]} + \mathsf{D}_i \underbrace{\rho}_{\mathsf{Y}_{1i} - \mathsf{Y}_{0i}} + \underbrace{\eta_i}_{\mathsf{Y}_{0i} - E[\mathsf{Y}_{0i}]}$$

Causality and OLS

- Now write out the conditional expectation of Y_i for both levels of D_i

$$E\left[\mathrm{Y}_i \mid \mathrm{D}_i = 1\right] = E\left[\alpha + \rho + \eta_i \mid \mathrm{D}_i = 1\right] = \alpha + \rho + E\left[\eta_i | \mathrm{D}_i = 1\right]$$

$$E\left[\mathrm{Y}_i \mid \mathrm{D}_i = 0\right] = E\left[\alpha + \eta_i \mid \mathrm{D}_i = 0\right] = \alpha + E\left[\eta_i \mid \mathrm{D}_i = 0\right]$$

- Take the difference

$$E\left[\mathrm{Y}_i \mid \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i \mid \mathrm{D}_i = 0\right] = \rho + \underbrace{E\left[\eta_i | \mathrm{D}_i = 1\right] - E\left[\eta_i \mid \mathrm{D}_i = 0\right]}_{\text{Selection bias}}$$

Causality and OLS

- Again, our estimate of the **treatment effect** (ρ) is only going to be as good as our ability to shut down the **selection bias**.
- *Selection bias in regression model:* $E\left[\eta_i | D_i = 1\right] - E\left[\eta_i \mid D_i = 0\right]$
- There is something in our disturbance η_i that is affecting Y_i and is also correlated with D_i.

Simple OLS Regression v.s. RCT

- In a simple regression model, OLS estimators are just a generalizing continuous version of RCT when least squares assumptions are hold.
- Ideally,regression is a way to control observable confounding factors, which assume the source of selection bias is only from the difference in observed characteristics.

Simple OLS Regression v.s. RCT

- But in contrast to RCT, in observational studies, researchers cannot control the assignment of treatment into a treatment group versus a control group,which means that the two groups are **incomparable**.
- To make two groups comparable, we need to keep treatment and control group “**other thing equal**”in observed characteristics and unobserved characteristics.
- OLS regression is valid only when least squares assumptions are hold.
- In most cases,it is not easy to obtain. We have to know how to make a convincing causal inference when these assumptions are not hold.

Extending Reading

- Ho, Chong and Xia(2017),“Yellow taxis have fewer accidents than blue taxis because yellow is more visible than blue”,PNAS,Vol.114(12),pp3074-3078.