Quantitative Social Science in the Age of Big Data and AI

Lecture 5: Multiple OLS Regression

Zhaopeng Qu

Hopkins-Nanjing Center

April 16 2025

Review of the Last Lecture

Simple OLS Formula

- 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)
    - 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.

The OLS Estimator

- 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 OLS Estimator

- 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.

OLS estimator of β_1:

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

Least Squares Assumptions

- Under 3 least squares assumptions,
    1. Assumption 1: ZERO Conditional Mean
    2. Assumption 2: i.i.d. Samples or random sampling
    3. Assumption 3: Without large outliers
- The OLS estimators will be
    1. **unbiased**
    2. **consistent**
    3. **normal sampling distribution**

Simple OLS Regression v.s. RCT

- A simple OLS regression model is a generalizing continuous version of RCT assuming three least squares assumptions are held.
- In most observational studies, OLS regression suffers from **selection bias**, which violates the assumption of $E(u_i|X_i) = 0$.
- In such cases, OLS estimators are **biased** and **inconsistent**. Therefore the **causal effect** of X on Y cannot be identified by simple OLS regression.
- To address the selection bias problem, we have to **extend** the *simple OLS regression* model in more general settings.

Make Comparison Make Sense

Case: Smoke and Mortality

- Criticisms from **Ronald A. Fisher**
    - There is no experimental evidence to suggest that smoking is a cause of lung cancer or other serious diseases.
    - Correlation between smoking and mortality may be spurious due to **biased selection** of subjects.

Case: Smoke and Mortality

- Criticisms from **Ronald A. Fisher**
  - There is no experimental evidence to suggest that smoking is a cause of lung cancer or other serious diseases.
  - Correlation between smoking and mortality may be spurious due to **biased selection** of subjects.

$$
\begin{tikzcd}
& Z \arrow[dl] \arrow[dr] & \\
S \arrow[rr] & & M
\end{tikzcd}
$$

- **Confounder**, Z, some other factors, affect on smoking and mortality simultaneously.

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

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 than others 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

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.
- Perhaps the higher observed death rates among cigar smokers are because **they're older on average**.

Case: Smoke and Mortality(Cochran 1968)

- The issue is that the ages are not balanced; there is a difference in the age distribution between the treatment and control groups.
- 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?

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**?

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

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 W is predetermined with respect to the treatment D if for each individual i, $W_{0i} = W_{1i}$, 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

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D$$

and therefore:

$$E[Y|D=1] - E[Y|D=0] = \underbrace{E[Y_{1i}|D=1] - E[Y_{0i}|D=0]}_{\text{by the switching equation}}$$

Identification under Independence

- Recall that randomization in RCTs implies

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D$$

and therefore:

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

Identification under Independence

- Recall that randomization in RCTs implies

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D$$

and therefore:

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

Identification under Independence

- Recall that randomization in RCTs implies

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D$$

and therefore:

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

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

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D|X$$

- **NOTE**: Because $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D|X \nLeftrightarrow (Y_{1i}, Y_{0i}) \perp\!\!\!\perp D$,then

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

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D|X$$

- **NOTE**: Because $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D|X \not\Leftrightarrow (Y_{1i}, Y_{0i}) \perp\!\!\!\perp D$,then

$$E[Y_{1i}|D=1] - E[Y_{0i}|D=0] \neq E[Y_{1i}|D=1] - E[Y_{0i}|D=1]$$

Identification under Conditional Independence(CIA)

- **But using the CIA assumption, then**

$$\underbrace{E[Y_{1i}|D=1]-E[Y_{0i}|D=0]}_{\text{association}} = \underbrace{E[Y_{1i}|D=1,X]-E[Y_{0i}|D=0,X]}_{\text{conditional on covariates}}$$

Identification under Conditional Independence(CIA)

- **But using the CIA assumption, then**

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

Identification under Conditional Independence(CIA)

- But using the CIA assumption, then

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

Identification under Conditional Independence(CIA)

- But using the CIA assumption, then

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

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 the 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 $3^{10} = 59049$
  - Even if $k = 6$, then $3^6 = 729$.Assume that we have 1000 observations, then the average number of observations in each cell is less than 2.
- Sub-classification is not a feasible method to balance covariates in high-dimensional space.

Making Comparison Make Sense

- **Question**: How to make comparison make sense in the presence of covariates?
- *Selection on Observables*
  - Regression
  - Matching
- *Selection on Unobservables*
  - IV,RD,DID,FE and SCM.
- The most fundamental tool among them is **multiple regression**, which compares treatment and control subjects who have the same **observable** characteristics **in a generalized manner**.

Multiple OLS Regression: Introduction

Violation of the 1st Least Squares Assumption

- Recall simple OLS regression equation

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

- **Question**: What does u_i represent?
  - Answer: contains **all other factors(variables)** which potentially affect Y_i.
- **Assumption 1**

$$E(u_i|X_i) = 0$$

  - It states that u_i are unrelated to X_i in the sense that,given a value of X_i,the mean of these other factors equals **zero**.
  - But what if u_i is **correlated** with X_i?

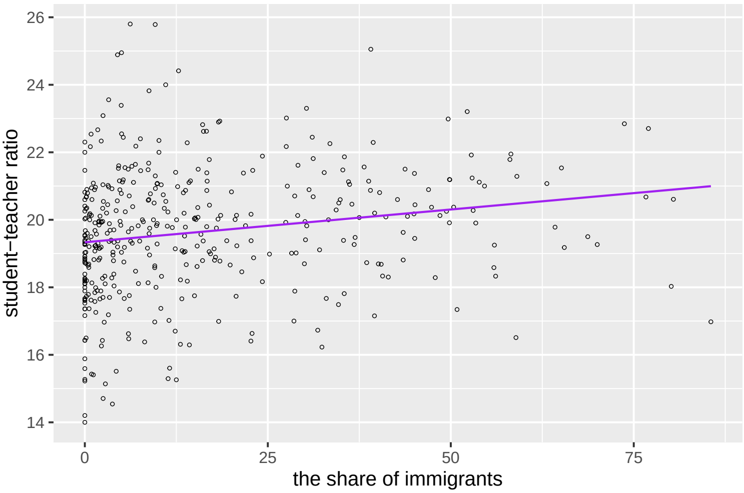
Example: Class Size and Test Score

- Many other factors can affect student's performance in the school.
- One of factors is **the share of immigrants** in the class. Because immigrant children may have different backgrounds from native children, such as
    - parents' education level
    - family income and wealth
    - parenting style
    - traditional culture

Scatter Plot: The share of immigrants and STR

Scatter Plot: The share of immigrants and STR

Scatter Plot: The share of immigrants and STR



- **higher share of immigrants, bigger class size**

Scatter Plot: The share of immigrants and STR

Scatter Plot: The share of immigrants and STR

- **higher share of immigrants, lower testscore**

The share of immigrants as an Omitted Variable

- Class size may be related to percentage of English learners and students who are still learning English likely have lower test scores.
  - In other words, the effect of class size on scores we had obtained in simple OLS may contain *an effect of immigrants on scores*.
- It implies that percentage of English learners is contained in u_i, in turn that **Assumption 1 is violated**.
  - More precisely,the estimates of $\hat{\beta}_1$ and $\hat{\beta}_0$ are **biased** and **inconsistent**.

Omitted Variable Bias: Introduction

- As before, X_i and Y_i represent **STR** and **Test Score**,repectively.
- Besides, W_i is the variable which represents **the share of english learners**.
- Suppose that we have no information about it for some reasons, then we have to omit in the regression.
- Thus we have two regressions in mind:
  - **True model**(the Long regression):

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

    where $E(u_i|X_i) = 0$
  - **OVB model**(the Short regression):

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

    where $v_i = \gamma W_i + u_i$

Omitted Variable Bias(OVB): inconsistency

- **Recall: simple OLS is consistency when n is large, thus**

$$plim\hat{\beta}_1 = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

Omitted Variable Bias(OVB): inconsistency

- Recall: simple OLS is consistency when n is large, thus

$$plim\hat{\beta}_1 = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

$$plim\hat{\beta}_1 = \frac{Cov(X_i, Y_i)}{VarX_i}$$

Omitted Variable Bias(OVB): inconsistency

- Recall: simple OLS is consistency when n is large, thus

$$plim\hat{\beta}_1 = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

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

Omitted Variable Bias(OVB): inconsistency

- **Recall: simple OLS is consistency when n is large, thus**

$$plim\hat{\beta}_1 = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

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

Omitted Variable Bias(OVB): inconsistency

- **Recall: simple OLS is consistency when n is large, thus**

$$plim\hat{\beta}_1 = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

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

Omitted Variable Bias(OVB): inconsistency

- **Recall: simple OLS is consistency when n is large, thus**

$$plim\hat{\beta}_1 = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

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

Omitted Variable Bias(OVB): Inconsistency

- Thus we obtain

$$plim\hat{\beta}_1 = \beta_1 + \gamma\frac{Cov(X_i, W_i)}{VarX_i}$$

- $\hat{\beta}_1$ is still **consistent**
  - if W_i is unrelated to X, thus $Cov(X_i, W_i) = 0$
  - if W_i has no effect on Y_i, thus $\gamma = 0$
- Only if **both two conditions** above are violated *simultaneously*}, then $\hat{\beta}_1$ is **inconsistent**.

Omitted Variable Bias(OVB):Directions

- If OVB can be possible in our regressions,then we should guess the **directions** of the bias, in case that we can't eliminate it.
- A summary of the directions of the OVB bias

Omitted Variable Bias(OVB):Directions

- If OVB can be possible in our regressions,then we should guess the **directions** of the bias, in case that we can't eliminate it.
- A summary of the directions of the OVB bias

	$Cov(X_i, W_i) > 0$	$Cov(X_i, W_i) < 0$
$\gamma > 0$		

Omitted Variable Bias(OVB):Directions

- If OVB can be possible in our regressions,then we should guess the **directions** of the bias, in case that we can't eliminate it.
- A summary of the directions of the OVB bias

	$Cov(X_i, W_i) > 0$	$Cov(X_i, W_i) < 0$
$\gamma > 0$	Positive bias	

Omitted Variable Bias(OVB):Directions

- If OVB can be possible in our regressions,then we should guess the **directions** of the bias, in case that we can't eliminate it.
- A summary of the directions of the OVB bias

	$Cov(X_i, W_i) > 0$	$Cov(X_i, W_i) < 0$
$\gamma > 0$		
	Positive bias	Negative bias
$\gamma < 0$		

Omitted Variable Bias(OVB):Directions

- If OVB can be possible in our regressions,then we should guess the **directions** of the bias, in case that we can't eliminate it.
- A summary of the directions of the OVB bias

	$Cov(X_i, W_i) > 0$	$Cov(X_i, W_i) < 0$
$\gamma > 0$	Positive bias	Negative bias
$\gamma < 0$	Negative bias	

Omitted Variable Bias(OVB):Directions

- If OVB can be possible in our regressions,then we should guess the **directions** of the bias, in case that we can't eliminate it.
- A summary of the directions of the OVB bias

	$Cov(X_i, W_i) > 0$	$Cov(X_i, W_i) < 0$
$\gamma > 0$	Positive bias	Negative bias
$\gamma < 0$	Negative bias	Positive bias

Omitted Variable Bias: Examples

- **Question**: If we omit following variables, then what are the directions of these biases? and why?
  1. Time of day of the test[suppose morning(8:00-12:00am) is better,afternoon(13:00-17:00pm) is worse]
  2. The number of dormitories
  3. Teachers' salary
  4. Family income
  5. Percentage of English learners(the share of immigrants)

Omitted Variable Bias: Examples in R

- Regress *Testscore* on *Class size*

```
#>
#> Call:
#> lm(formula = testscr ~ str, data = ca)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -47.727 -14.251   0.483  12.822  48.540
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
#> str          -2.2798     0.4798  -4.751 2.78e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 18.58 on 418 degrees of freedom
#> Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
#> F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Omitted Variable Bias: Examples in R

- **Regress *Testscore* on *Class size* and *the percentage of English learners***

```
#>
#> Call:
#> lm(formula = testscr ~ str + el_pct, data = ca)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -48.845 -10.240  -0.308   9.815  43.461
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
#> str          -1.10130    0.38028  -2.896  0.00398 **
#> el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 14.46 on 417 degrees of freedom
#> Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237
```

Omitted Variable Bias: Examples in R

	Dependent variable:	
	testscr	
	(1)	(2)
str	-2.280^{***}	-1.101^{***}
	(0.480)	(0.380)
el_pct		-0.650^{***}
		(0.039)
Constant	698.933^{***}	686.032^{***}
	(9.467)	(7.411)
Observations	420	420
R^2	0.051	0.426
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

Warp Up

- **OVB** is **the most common** bias when we run OLS regressions using non-experimental data.
  - It means that there are some variables which should have been included in the regression but actually was not.
- Then the simplest way to overcome OVB: *Putting omitted variables into the right side of the regression*, which means our regression model should be

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

- This strategy can be denoted as **controlling** informally, which introduces the more general regression model: **Multiple OLS Regression**.

Multiple OLS Regression: Estimation

Multiple regression model with k regressors

- The multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i, i = 1, ..., n \tag{4.1}$$

where

- Y_i is the **dependent variable**
- $X_1, X_2, ...X_k$ are the **independent variables(includes one is our of interest and some control variables)**
- $\beta_i, j = 1...k$ are slope coefficients on X_i corresponding.
- β_0 is the estimate *intercept*, the value of Y when all $X_j = 0, j = 1...k$
- u_i is the regression *error term*, still all other factors affect outcomes.

Interpretation of coefficients $\beta_i, j = 1...k$

- β_j is **partial (marginal) effect** of X_j on Y.

$$\beta_j = \frac{\partial Y_i}{\partial X_{j,i}}$$

- β_j is also partial (marginal) effect of $E[Y_i|X_1..X_k]$.

$$\beta_j = \frac{\partial E[Y_i|X_1, ..., X_k]}{\partial X_{j,i}}$$

- it does mean that we are estimate the effect of X on Y when **"other things equal"**, thus the concept of **ceteris paribus**.

OLS Estimation in Multiple Regressors

- As in a **Simple OLS Regression**, the estimators of **Multiple OLS Regression** is just a minimize the following question

OLS Estimation in Multiple Regressors

- As in a **Simple OLS Regression**, the estimators of **Multiple OLS Regression** is just a minimize the following question

$$argmin \sum_{b_0,b_1,...,b_k} (Y_i - b_0 - b_1 X_{1,i} - ... - b_k X_{k,i})^2$$

where $b_0 = \hat{\beta}_1, b_1 = \hat{\beta}_2, ..., b_k = \hat{\beta}_k$ are estimators.

OLS Estimation in Multiple Regressors

- **Similarly in Simple OLS, based on F.O.C,the multiple OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are obtained by solving the following system of normal equations**

OLS Estimation in Multiple Regressors

- **Similarly in Simple OLS, based on F.O.C,the multiple OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are obtained by solving the following system of normal equations**

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} \hat{u}_i^2 = \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - ... - \hat{\beta}_k X_{k,i} \right) = 0$$

OLS Estimation in Multiple Regressors

- Similarly in Simple OLS, based on **F.O.C**,the multiple OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are obtained by solving the following **system of normal equations**

$$
\begin{aligned}
\frac{\partial}{\partial b_0} \sum_{i=1}^{n} \hat{u}_i^2 = \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - ... - \hat{\beta}_k X_{k,i} \right) \quad &= 0 \\
\frac{\partial}{\partial b_1} \sum_{i=1}^{n} \hat{u}_i^2 = \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - ... - \hat{\beta}_k X_{k,i} \right) X_{1,i} \quad &= 0
\end{aligned}
$$

OLS Estimation in Multiple Regressors

- **Similarly in Simple OLS, based on F.O.C,the multiple OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are obtained by solving the following system of normal equations**

$$
\begin{aligned}
\frac{\partial}{\partial b_0}\sum_{i=1}^{n}\hat{u}_i^2 &= \sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - ... - \hat{\beta}_k X_{k,i}\right) &= 0\\
\frac{\partial}{\partial b_1}\sum_{i=1}^{n}\hat{u}_i^2 &= \sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - ... - \hat{\beta}_k X_{k,i}\right)X_{1,i} &= 0\\
\vdots &= \vdots &= \vdots\\
\frac{\partial}{\partial b_k}\sum_{i=1}^{n}\hat{u}_i^2 &= \sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - ... - \hat{\beta}_k X_{k,i}\right)X_{k,i} &= 0
\end{aligned}
$$

OLS Estimation in Multiple Regressors

- Similar to in Simple OLS, the fitted residuals are

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - ... - \hat{\beta}_k X_{k,i}$$

- Therefore, the normal equations also can be written as

$$\begin{aligned}
\sum \hat{u}_i &= 0 \\
\sum \hat{u}_i X_{1,i} &= 0 \\
\vdots &= \vdots \\
\sum \hat{u}_i X_{k,i} &= 0
\end{aligned}$$

- While it is convenient to transform equations above using **matrix algebra** to compute these estimators, we can use **partitioned regression** to obtain the formula of estimators without using matrix algebra.

Measures of Fit in Multiple Regression

Recall: Measures of Fit: The R^2

- Decompose Y_i into the fitted value plus the residual $Y_i = \hat{Y}_i + \hat{u}_i$
- The **total sum of squares** (TSS): $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$
- 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

$$TSS = ESS + SSR$$

- The regression R^2 is the fraction of the sample variance of Y_i explained by (or predicted by) the regressors.

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

Measures of Fit in Multiple Regression

- When you put more variables into the regression, then R^2 **always increases** when you *add another regressor*. Because in general the SSR will decrease.
- **the Adjusted** R^2,is a modified version of the R^2 that does not necessarily increase when a new regressor is added.

$$\overline{R^2} = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}$$

  - because $\frac{n-1}{n-k-1}$ is always greater than 1, so $\overline{R^2} < R^2$
  - adding a regressor has two opposite effects on the $\overline{R^2}$.
  - $\overline{R^2}$ can be negative.
- **Remind**: *neither* R^2 *nor* $\overline{R^2}$ *is NOT the golden criterion for good or bad OLS estimation.*

Example: Test scores and Student Teacher Ratios

```
1 . reg testscr str el_pct
```

Source	SS	df	MS			
Model	64864.3011	2	32432.1506	Number of obs	=	420
Residual	87245.2925	417	209.221325	F(2, 417)	=	155.01
				Prob > F	=	0.0000
Total	152109.594	419	363.030056	R-squared	=	0.4264
				Adj R-squared	=	0.4237
				Root MSE	=	14.464

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	-1.101296	.3802783	-2.90	0.004	-1.848797	-.3537945
el_pct	-.6497768	.0393425	-16.52	0.000	-.7271112	-.5724423
_cons	686.0322	7.411312	92.57	0.000	671.4641	700.6004

Multiple Regression: Assumption

Multiple Regression: Assumption

- Assumption 1: The conditional distribution of u_i given $X_{1i}, ..., X_{ki}$ has mean zero,thus

$$E[u_i|X_{1i}, ..., X_{ki}] = 0$$

  - which is a very strong assumption, which means u_i is uncorrelated with all the independent variables.(we will discuss this later)
- Assumption 2: $(Y_i, X_{1i}, ..., X_{ki})$ are i.i.d.
- Assumption 3: Large outliers are unlikely.
- At last, we have to add one more assumption for multiple regression.
  - **Assumption 4**: No perfect multicollinearity.

Perfect Multicollinearity

- **Perfect multicollinearity** arises when one of the regressors is a **perfect** linear combination of the other regressors.
- If you include a full set of binary variables (a complete and mutually exclusive categorization) and an intercept in the regression, you will have perfect multicollinearity.
  - eg. female and male = 1-female
- This is called the **dummy variable trap**.
- Solutions to the dummy variable trap:
  - Omit one of the groups or the intercept

Categoried Variable as Independent Variables

- Recall if X is a dummy variable, then we can put it into regression equation straightly.
- What if X is a categorical variable?
  - **Question**: What is a categorical variable?
- For example, we may define D_i as follows:

Categoried Variable as Independent Variables

- Recall if X is a dummy variable, then we can put it into regression equation straightly.
- What if X is a categorical variable?
  - **Question**: What is a categorical variable?
- For example, we may define D_i as follows:

$$D_i = \begin{cases} 1 & \text{small-size class if } STR \text{ in } i^{th} \text{ school district} < 18 \\ 2 & \text{middle-size class if } 18 \leq STR \text{ in } i^{th} \text{ school district} < 22 \\ 3 & \text{large-size class if } STR \text{ in } i^{th} \text{ school district} \geq 22 \end{cases} \tag{4.5}$$

A Special Case: Categorical Variable as X

- Naive Solution: a simple OLS regression model

$$TestScore_i = \beta_0 + \beta_1 D_i + u_i$$

- **Question**: Can you explain the meanning of estimate coefficient β_1?
- **Answer**: It doese not make sense that the coefficient of β_1 can be explained as continuous variables.

A Special Case: Categorical Variables as X

- The first step: turn a categorical variable(D_i) into multiple dummy variables(D_{1i}, D_{2i}, D_{3i})

A Special Case: Categorical Variables as X

- The first step: turn a categorical variable(D_i) into multiple dummy variables(D_{1i}, D_{2i}, D_{3i})

$$D_{1i} = \begin{cases} 1 & \text{small-sized class if } STR \text{ in } i^{th} \text{ school district} < 18 \\ 0 & \text{middle-sized class or large-sized class if not} \end{cases}$$

A Special Case: Categorical Variables as X

- The first step: turn a categorical variable(D_i) into multiple dummy variables(D_{1i}, D_{2i}, D_{3i})

$$D_{1i} = \begin{cases} 1 & \text{small-sized class if } STR \text{ in } i^{th} \text{ school district < 18} \\ 0 & \text{middle-sized class or large-sized class if not} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{middle-sized class if } 18 \leq STR \text{ in } i^{th} \text{ school district < 22} \\ 0 & \text{large-sized class or small-sized class if not} \end{cases}$$

A Special Case: Categorical Variables as X

- The first step: turn a categorical variable(D_i) into multiple dummy variables(D_{1i}, D_{2i}, D_{3i})

$$D_{1i} = \begin{cases} 1 & \text{small-sized class if } STR \text{ in } i^{th} \text{ school district} < 18 \\ 0 & \text{middle-sized class or large-sized class if not} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{middle-sized class if } 18 \leq STR \text{ in } i^{th} \text{ school district} < 22 \\ 0 & \text{large-sized class or small-sized class if not} \end{cases}$$

$$D_{3i} = \begin{cases} 1 & \text{large-sized class if } STR \text{ in } i^{th} \text{ school district} \geq 22 \\ 0 & \text{middle-sized class or small-sized class if not} \end{cases}$$

A Special Case: Categorical Variables as X

- The first step: turn a categorical variable(D_i) into multiple dummy variables(D_{1i}, D_{2i}, D_{3i})

$$D_{1i} = \begin{cases} 1 & \text{small-sized class if } STR \text{ in } i^{th} \text{ school district} < 18 \\ 0 & \text{middle-sized class or large-sized class if not} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{middle-sized class if } 18 \leq STR \text{ in } i^{th} \text{ school district} < 22 \\ 0 & \text{large-sized class or small-sized class if not} \end{cases}$$

$$D_{3i} = \begin{cases} 1 & \text{large-sized class if } STR \text{ in } i^{th} \text{ school district} \geq 22 \\ 0 & \text{middle-sized class or small-sized class if not} \end{cases}$$

A Special Case: Categorical Variables as X

- We put these dummies into a multiple regression

$$TestScore_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \tag{4.6}$$

- Then as a dummy variable as the independent variable in a simple regression The coefficients $(\beta_1, \beta_2, \beta_3)$ represent the effect of every categorical class on $testscore$ respectively.

A Special Case: Categorical Variables as X

- In practice, we can't put all dummies into the regression, but only have $n-1$ dummies unless we will suffer **perfect multi-collinearity**.
- The regression may be like as

$$TestScore_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i \tag{4.6}$$

- The default intercept term, β_0,represents the large-sized class.Then, the coefficients (β_1, β_2) represent $testscore$ gaps between small_sized, middle-sized class and large-sized class,respectively.

- **regress** *Testscore* **on** *Class size* **and** *the percentage of English learners*

```
#>
#> Call:
#> lm(formula = testscr ~ str + el_pct, data = ca)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -48.845 -10.240  -0.308   9.815  43.461
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
#> str          -1.10130    0.38028  -2.896  0.00398 **
#> el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

- **add a new variable nel=1-el_pct into the regression**

```
#>
#> Call:
#> lm(formula = testscr ~ str + nel_pct + el_pct, data = ca)
#>
#> Residuals:
#>      Min       1Q  Median       3Q      Max
#> -48.845 -10.240  -0.308   9.815  43.461
#>
#> Coefficients: (1 not defined because of singularities)
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 685.38247    7.41556  92.425  < 2e-16 ***
#> str          -1.10130    0.38028  -2.896  0.00398 **
#> nel_pct       0.64978    0.03934  16.516  < 2e-16 ***
#> el_pct             NA         NA      NA       NA
#>
```

Perfect Multicollinearity

Table 5: Class Size and Test Score

	Dependent variable:	
	testscr	
	(1)	(2)
str	-1.101^{***}	-1.101^{***}
	(0.380)	(0.380)
nel_pct		0.650^{***}
		(0.039)
el_pct	-0.650^{***}	
	(0.039)	
Constant	686.032^{***}	685.382^{***}
	(7.411)	(7.416)
Observations	420	420
Adjusted R^2	0.424	0.424
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

Multiple OLS Regression and Causality

Independent Variable v.s Control Variables

- Generally, we would like to pay more attention to **only one** independent variable(thus we would like to call it **treatment variable**), though there could be many independent variables.
- Because β_j is **partial (marginal) effect** of X_j on Y.

$$\beta_j = \frac{\partial Y_i}{\partial X_{j,i}}$$

which means that we are estimate the effect of X on Y when **"other things equal"**, thus the concept of **ceteris paribus**.

- Therefore,other variables in the right hand of equation, we call them **control variables**, which we would like to explicitly **hold fixed** when studying the effect of X_1 or D on Y.

Independent Variable v.s Control Variables

- In a multiple regression, OLS is a way to control observable confounding factors, which assume the source of selection bias is only from the difference in observed characteristics(Selection-on-Observables)
- If the multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i, i = 1, ..., n$$

- Generally, we would like to pay more attention to **only one** independent variable(thus we would like to call it **treatment variable**), though there could be many independent variables.
- Other variables in the right hand of equation, we call them **control variables**, which we would like to explicitly hold fixed when studying the effect of X_1 on Y.

OLS Regression, Covariates and RCT

- More specifically,our multiple regression model turns into

$$Y_i = \beta_0 + \beta_1 D_i + \gamma_2 C_{2,i} + ... + \gamma_k C_{k,i} + u_i, i = 1, ..., n$$

- We could transform it into as follows

$$Y_i = \alpha + \rho D_i + C_i' \Gamma + u_i$$

where $\alpha = \beta_0, \rho = \beta_1, \Gamma = (\gamma_2, ..., \gamma_k), C_i = (C_{2i}, ..., C_{ki})$

OLS Regression, Covariates and RCT

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

$$
\begin{aligned}
E\left[\mathbf{Y}_i \mid \mathbf{D}_i = 1, C\right] &= E\left[\alpha + \rho + C'\Gamma + u_i \mid \mathbf{D}_i = 1, C\right] \\
&= \alpha + \rho + C'\Gamma + E\left[u_i | \mathbf{D}_i = 1, C\right]
\end{aligned}
$$

OLS Regression, Covariates and RCT

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

$$
\begin{aligned}
E\left[\mathbf{Y}_i \mid \mathbf{D}_i = 1, C\right] &= E\left[\alpha + \rho + C'\Gamma + u_i \mid \mathbf{D}_i = 1, C\right] \\
&= \alpha + \rho + C'\Gamma + E\left[u_i | \mathbf{D}_i = 1, C\right] \\
E\left[\mathbf{Y}_i \mid \mathbf{D}_i = 0, C\right] &= E\left[\alpha + C'\Gamma + u_i \mid \mathbf{D}_i = 0, C\right] \\
&= \alpha + C'\Gamma + E\left[u_i \mid \mathbf{D}_i = 0, C\right]
\end{aligned}
$$

OLS Regression, Covariates and RCT

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

$$
\begin{aligned}
E\left[\mathbf{Y}_i \mid \mathbf{D}_i=1, C\right] &= E\left[\alpha+\rho+C'\Gamma+u_i \mid \mathbf{D}_i=1, C\right] \\
&= \alpha+\rho+C'\Gamma+E\left[u_i \mid \mathbf{D}_i=1, C\right] \\
E\left[\mathbf{Y}_i \mid \mathbf{D}_i=0, C\right] &= E\left[\alpha+C'\Gamma+u_i \mid \mathbf{D}_i=0, C\right] \\
&= \alpha+C'\Gamma+E\left[u_i \mid \mathbf{D}_i=0, C\right]
\end{aligned}
$$

- Taking the difference

$$
\begin{aligned}
& E\left[\mathbf{Y}_i \mid \mathbf{D}_i=1, C\right]-E\left[\mathbf{Y}_i \mid \mathbf{D}_i=0, C\right] \\
= & \rho+\underbrace{E\left[u_i \mid \mathbf{D}_i=1, C\right]-E\left[u_i \mid \mathbf{D}_i=0, C\right]}_{\text{Selection bias}}
\end{aligned}
$$

OLS Regression, Covariates and RCT

- Again, our estimate of the **treatment effect** (ρ) is only going to be as good as our ability to eliminate the **selection bias**,thus

$$E\left[u_{1i}|\mathbf{D}_i = 1, C\right] - E\left[u_{0i} \mid \mathbf{D}_i = 0, C\right] \neq 0$$

Conditional Independence Assumption(CIA)

Balancing or controlling covariates C then we can take the treatment D as randomized, thus

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

OLS Regression, Covariates and RCT

- This is the equivalence of the **CIA** assumption, which is also equivalent to the **1st assumption** of Multiple OLS

$$E\left[u_{1i}|\mathbf{D}_i = 1, C\right] - E\left[u_{0i} \mid \mathbf{D}_i = 0, C\right] = E\left[u_{1i}|C\right] - E\left[u_{0i}|C\right]$$

- Then we can eliminate the **selection bias**, thus making

$$E\left[u_{1i}|\mathbf{D}_i = 1, C\right] = E\left[u_{0i} \mid \mathbf{D}_i = 0, C\right]$$

- Thus

$$E\left[\mathbf{Y}_i \mid \mathbf{D}_i = 1, C\right] - E\left[\mathbf{Y}_i \mid \mathbf{D}_i = 0, C\right] = \rho$$

Wrap up

- OLS regression is valid or can obtain a causal explanation only when least squares assumptions are held.
- The most critical assumption is the **Conditional Independence Assumption(CIA)**, which can be loose to

$$E(u_i|D, C) = E(u_i|C)$$

- This means that not all coefficients in the regression need to be **causal** (unbiased or consistent).
  - Only the coefficient of the treatment variable (D) need to be causal in the regression. which is the interest of the study.
  - If the coefficients of **control variables** (C) are *biased* or *inconsistent*, it does not affect the causal interpretation of the treatment effect.

Picking Control Variables in Ge

- **Questions**: Are "more controls" always better (or at least never worse)?
- **Answer**: It depends on.
  - **Irrelevant controls** are variables which have a ZERO partial effect on the outcome, thus the coefficient in the population regression function is zero.
  - **Relevant controls** are variables which have a NONZERO partial effect on the dependent variable.
    - Non-Omitted Variables
    - Omitted Variables
  - **Highly-correlated Variables**
    - Multicollinearity