Applied Micro-Econometrics,Fall 2023

Lecture 2: Regression

Zhaopeng Qu

Nanjing University Business School

October 12 2023

1. Review the previous lecture
2. OLS Estimation: Simple Regression
3. Properties of the OLS Estimators
4. Simple OLS and RCT
5. Make Comparison Make Sense
6. **Multiple OLS Regression: Introduction**
7. **Multiple OLS Regression: Estimation**
8. Multiple OLS Regression and Causality

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*.

Question: Class Size and Student's Performance

- If fortunately, we can run a RCT, then how to answer the question quantitatively in a standard model?
- Draw schools (n = 420) randomly from all school in California
- Variables:
  - 5th grade test scores (Stanford-9 achievement test, combined math and reading), district average
  - Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teachers

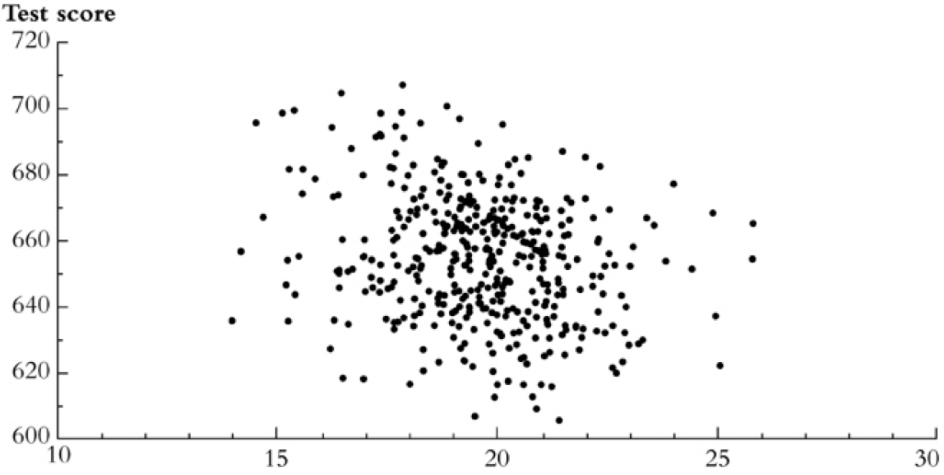
Descriptive Statistics

TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

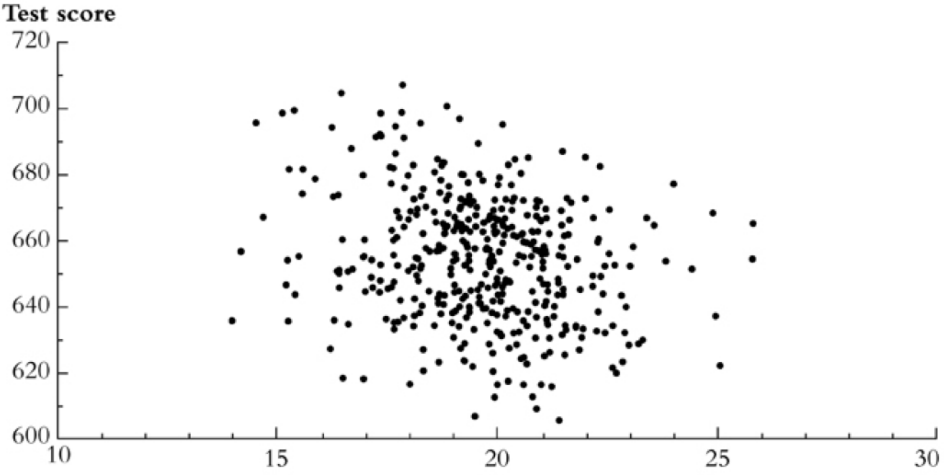
			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	665.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

- Does this table tell us anything about the relationship between test scores and the STR?

Scatterplot: test score v. student-teacher ratio



Scatterplot: test score v. student-teacher ratio



- What does this figure show? and it may suggest...?

The California Test Score

- We need to get some numerical evidence on whether districts with low STRs have higher test scores.
- But how?

1. Compare average test scores in districts with low STRs to those with high STRs ("estimation")
2. Test the "null" hypothesis that the mean test scores in the two types of districts are the same, against the "alternative" hypothesis that they differ ("hypothesis testing")
3. Estimate an interval for the difference in the mean test scores, high v. low STR districts ("confidence interval")

The California Test Score

- Compare districts with “small” and “large” class sizes:
- Block: Small v.s. Large

1. Estimation of $\Delta = difference between group means$
2. Test the hypothesis that $\Delta = 0$
3. Construct a confidence interval for Δ

Comparing Means from Different Populations

- In an RCT, we would like to estimate the average causal effects over the population

$$ATE = ATT = E\{Y_i(1) - Y_i(0)\}$$

- We only have random samples and random assignment to treatment, then what we can estimate instead

$$\text{difference in mean} = \overline{Y}_{treated} - \overline{Y}_{control}$$

- Under randomization, difference-in-means is a good estimate for the ATE.

Hypothesis Tests for the Difference

- To illustrate a test for the difference between two means, let μ_s be the mean scores in the population of small classes and let μ_l be the population mean scores for the large classes.
- Then the null hypothesis and the two-sided alternative hypothesis are

$$H_0 : \mu_s = \mu_l \text{ and } H_1 : \mu_s \neq \mu_l$$

- Consider the null hypothesis that mean scores for these two populations differ by a certain amount, say d_0. The null hypothesis that large classes and small classes have the same mean scores corresponds to $H_0 : d_0 = \mu_s - \mu_l = 0$
- Suppose we have samples of n_s classes and n_l classes drawn at random from the population of CA. Let the sample average scores be $\overline{Y}_s$ for the small and $\overline{Y}_l$ for the large. Then an estimator of $\mu_s - \mu_l$ is $\overline{Y}_s - \overline{Y}_l$.

The Difference Between Two Means

- Let us discuss the distribution of $\overline{Y}_s - \overline{Y}_l$.
- Recall $\overline{Y}_s$ is approximately distributed $N(\mu_s, \frac{\sigma_s^2}{n_s})$ and $\overline{Y}_l$ is approximately distributed $N(\mu_l, \frac{\sigma_l^2}{n_l})$ according to the C.L.T.
- Then $\overline{Y}_s - \overline{Y}_l$ is distributed as

$$\sim N(\mu_s - \mu_l, \frac{\sigma_s^2}{n_s} + \frac{\sigma_l^2}{n_l})$$

The Difference Between Two Means

- If σ_s^2 and σ_l^2 are known, then the this approximate normal distribution can be used to compute p-values for the test of the null hypothesis.
- In practice, however, these population variances are typically unknown so they must be estimated using the variance of the sample mean.
- Thus the standard error of $\overline{Y}_s - \overline{Y}_l$ is

$$SE(\overline{Y}_s - \overline{Y}_l) = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$$

The Difference Between Two Means

- The t-statistic for testing the null hypothesis is constructed analogously to the t-statistic for testing a hypothesis about a single population mean, thus a simplest t-statistic for comparing two means is

$$t_{act} = \frac{\overline{Y}_s - \overline{Y}_l - d_0}{SE(\overline{Y}_s - \overline{Y}_l)}$$

- If both $n_s$and n_l are large, then this t-statistic has a standard normal distribution when the null hypothesis is true,thus

$$\overline{Y}_s - \overline{Y}_l = 0$$

.

Confidence Intervals for the Difference

- the 95% two-sided confidence interval for d consists of those values of d within ± 1.96 standard errors of $\overline{Y}_s - \overline{Y}_l$, thus $d = \mu_s - \mu_l$ is

$$(\overline{Y}_s - \overline{Y}_l) \pm 1.96 SE(\overline{Y}_s - \overline{Y}_l)$$

- Reject the null hypothesis if

$$\mid t^{act} \mid = \mid \frac{\overline{Y}_l - \overline{Y}_s - d_0}{SE(\overline{Y}_l - \overline{Y}_s)} \mid > \textit{critical value}$$

- or if

$$p - \textit{value} < \textit{significance level}$$

Causal Inference and RCT

- However,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. Regression(**回归**)
  2. Matching and Propensity Score（**匹配与倾向得分**）
  3. Decomposition（**分解**）
  4. Instrumental Variable（**工具变量**）
  5. Regression Discontinuity（**断点回归**）
  6. Panel Data and Difference in Differences （**双差分或倍差法**）
  7. Sythetic Control Metnods(**合成控制法**)
- 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.

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, $f(\cdot)$

$$Testscore = f(ClassSize)$$

- 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**

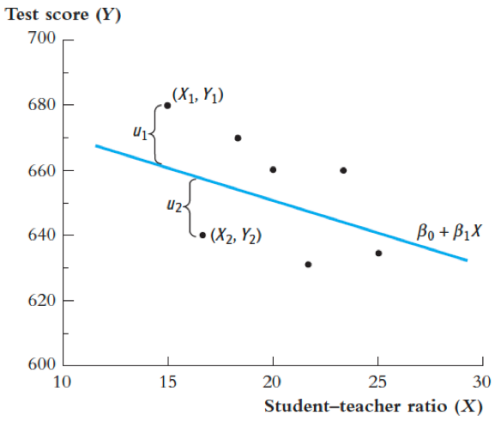
Terminology for Simple Regression Model

- 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

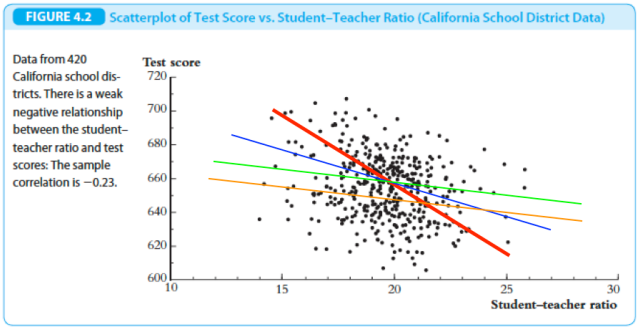
FIGURE 4.1 Scatterplot of Test Score vs. Student–Teacher Ratio (Hypothetical Data)

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 reality, we don’t know how to draw the straight line graphically, unless we know the values of β_0 and β_1. Thus the parameters of *population regression function*.



- In general we have to calculate them using a bunch of data: **the sample**,which is called the estimation.

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:

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

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

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

  - Step 2 for β_1:

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

OLS estimator of β_0 and β_1

OLS estimator of β_0 and β_1:

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

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

The Estimated Regression Line

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

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

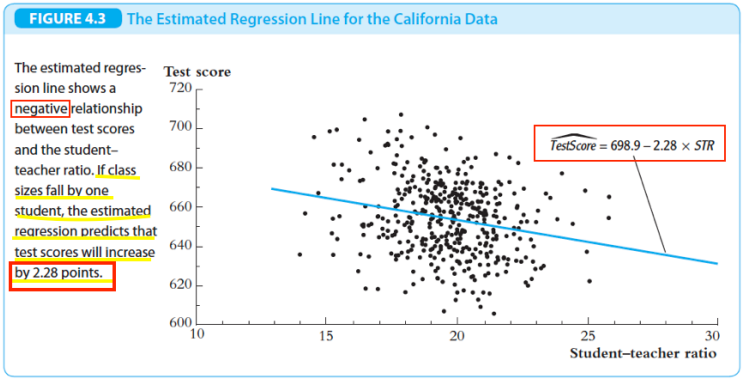
- Then the regression line is

The Estimated Regression Line

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

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

- Then the regression line is



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

$$TSS = ESS + SSR$$

The Least Squares Assumptions

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 $E[Y \mid X] = E[Y]$
- $E[Yg(X) \mid X] = g(X)E[Y \mid X]$. 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.

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

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

Proof

$$E[E(Y|X)] =$$

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

$$E[E(Y|X)] = \int E(Y|X = u) f_X(u) du$$

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Conditional Expectation and Covariance

- Please prove if $E(Y|X) = 0 \Rightarrow Cov(X, Y) = 0$

Proof

$$Cov(XY) = E(XY) - E(X)E(Y)$$

Conditional Expectation and Covariance

- Please prove if $E(Y|X) = 0 \Rightarrow Cov(X, Y) = 0$

Proof

$$\begin{aligned} Cov(XY) &= E(XY) - E(X)E(Y) \\ &= E[E(XY|X)] - E(X)E[E(Y|X)] \end{aligned}$$

Conditional Expectation and Covariance

- Please prove if $E(Y|X) = 0 \Rightarrow Cov(X, Y) = 0$

Proof

$$\begin{aligned} Cov(XY) &= E(XY) - E(X)E(Y) \\ &= E[E(XY|X)] - E(X)E[E(Y|X)] \\ &= E[XE(Y|X)] \end{aligned}$$

Conditional Expectation and Covariance

- Please prove if $E(Y|X) = 0 \Rightarrow Cov(X, Y) = 0$

Proof

$$\begin{aligned} Cov(XY) &= E(XY) - E(X)E(Y) \\ &= E[E(XY|X)] - E(X)E[E(Y|X)] \\ &= E[XE(Y|X)] \\ &= 0 \end{aligned}$$

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

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

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:

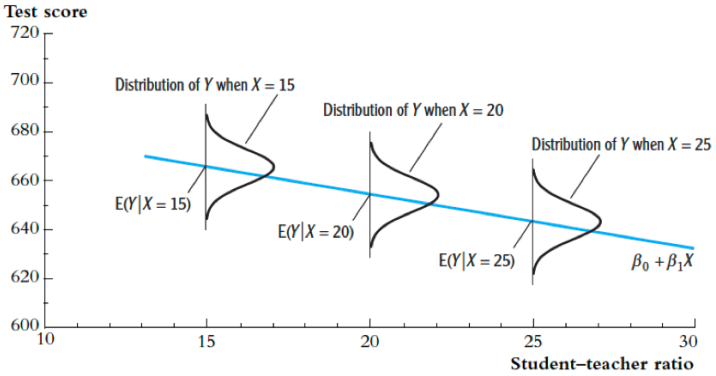
$$Cov[u_i, X_i] = E[u_i X_i] = 0$$

- 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.

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

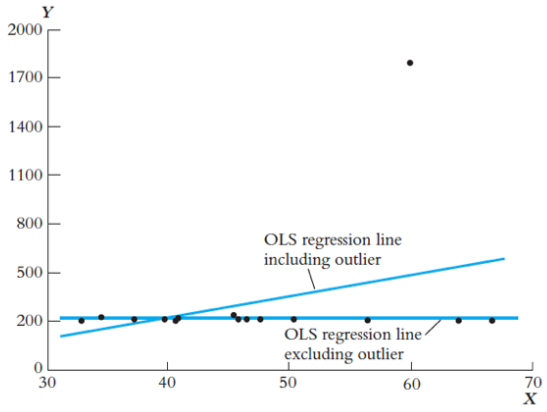
It states that observations with values of X_i, Y_i or both that are far outside the usual range of the data(Outlier) are unlikely. Mathematically, it assume that X and Y have nonzero finite fourth moments.

- 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
- 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.

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:

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

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: Consistency

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

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

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

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

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

- 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

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

Properties of the OLS estimator: Consistency

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

Properties of the OLS estimator: Consistency

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

Properties of the OLS estimator: Consistency

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

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

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

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

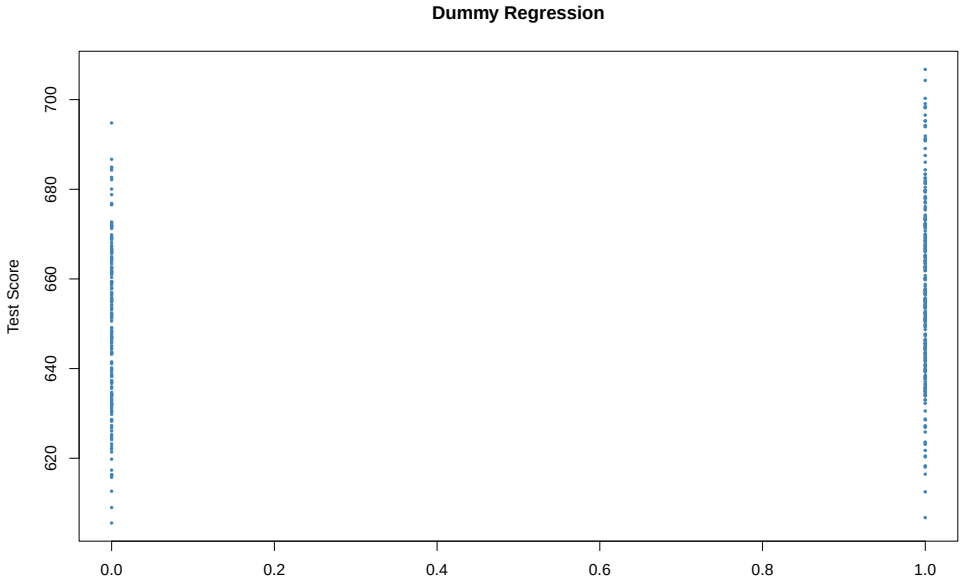
- More precisely, the regression model now is

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

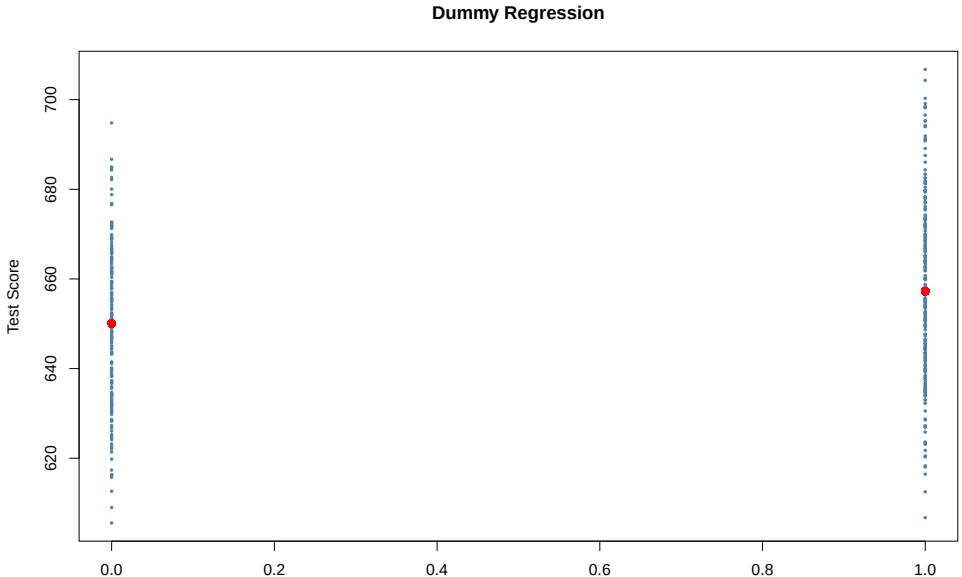
  - With D as the regressor, it is not useful to think of β_1 as a slope parameter.
  - Since $D_i \in \{0, 1\}$, i.e., we only observe two discrete values instead of a continuum of regressor values.

- There is no continuous line depicting the conditional expectation function $E(TestScore_i | D_i)$ 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

$$Y_i = Y_{0i} + D_i \left(Y_{1i} - Y_{0i}\right)$$

Causality and OLS

- Regression function is

$$Y_i = \alpha + D_i\rho + \eta_i$$

- Further

$$Y_i = \underbrace{\alpha}_{E[Y_{0i}]} + D_i \underbrace{\rho}_{Y_{1i}-Y_{0i}} + \underbrace{\eta_i}_{Y_{0i}-E[Y_{0i}]}$$

Causality and OLS

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

$$E\left[Y_i \mid D_i = 1\right] = E\left[\alpha + \rho + \eta_i \mid D_i = 1\right] = \alpha + \rho + E\left[\eta_i | D_i = 1\right]$$
$$E\left[Y_i \mid D_i = 0\right] = E\left[\alpha + \eta_i \mid D_i = 0\right] = \alpha + E\left[\eta_i \mid D_i = 0\right]$$

- Take the difference

$$E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right] = \rho + \underbrace{E\left[\eta_i | D_i = 1\right] - E\left[\eta_i \mid D_i = 0\right]}_{\text{Selection bias}}$$

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:* $E\left[\eta_i | \mathrm{D}_i = 1\right] - E\left[\eta_i \mid \mathrm{D}_i = 0\right]$
- 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.

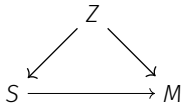
Make Comparison Make Sense

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.

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

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

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?

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 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:

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

Identification under Independence

- Recall that randomization in RCTs implies

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

and therefore:

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

Identification under Independence

- Recall that randomization in RCTs implies

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

and therefore:

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

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

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

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

Identification under Conditional Independence(CIA)

- But using the CIA assumption, then

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

Identification under Conditional Independence(CIA)

- But using the CIA assumption, then

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

$$= \underbrace{E[Y^1|D=1,X] - E[Y^0|D=1,X]}_{\text{conditional independence}}$$

Identification under Conditional Independence(CIA)

- But using the CIA assumption, then

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

$$= \underbrace{E[Y^1|D=1,X] - E[Y^0|D=1,X]}_{\text{conditional independence}}$$

$$= \underbrace{E[Y^1 - Y^0|D=1,X]}_{\text{conditional ATT}}$$

Identification under Conditional Independence(CIA)

- But using the CIA assumption, then

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

$$= \underbrace{E[Y^1|D=1,X] - E[Y^0|D=1,X]}_{\text{conditional independence}}$$

$$= \underbrace{E[Y^1 - Y^0|D=1,X]}_{\text{conditional ATT}}$$

$$= \underbrace{E[Y^1 - Y^0|X]}_{\text{conditional ATE}}$$

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 $3^{10} = 59049$

Making Comparison Make Sense

- *Selection on Observables*
  - Regression
  - Matching
- *Selection on Unobservables*
  - IV,RD,DID,FE and SCM.
- Simple Regression have to extend to Multiple OLS.

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**

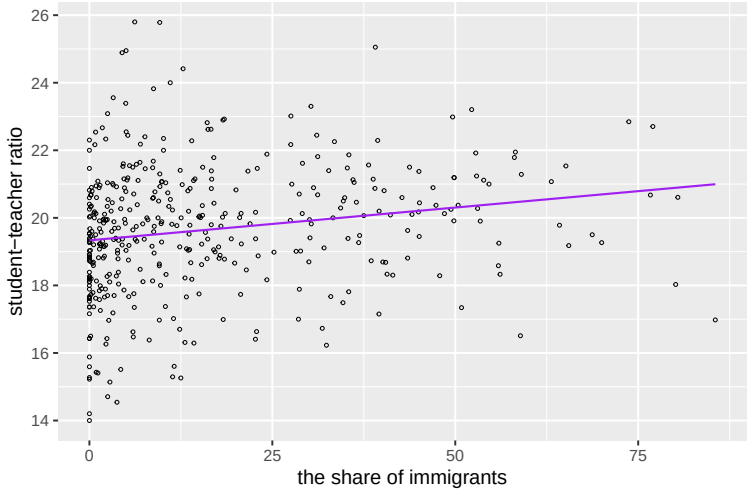
$$E(u_i|X_i) = 0$$

  - 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 they (or at least one) are *correlated* with X_i?

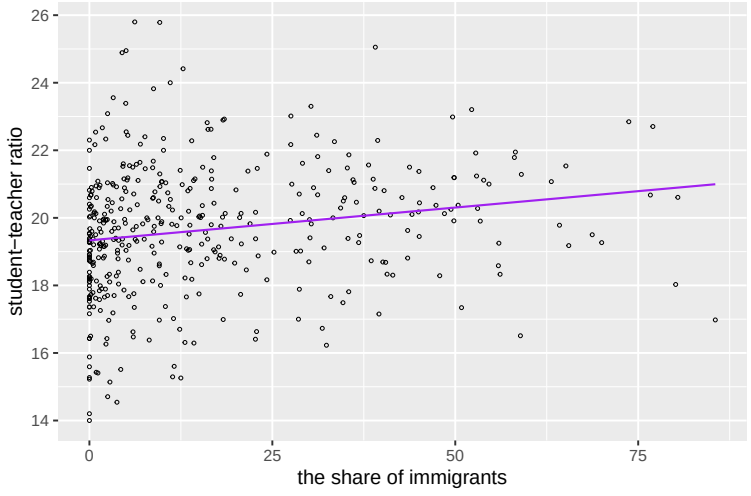
Example: Class Size and Test Score

- 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

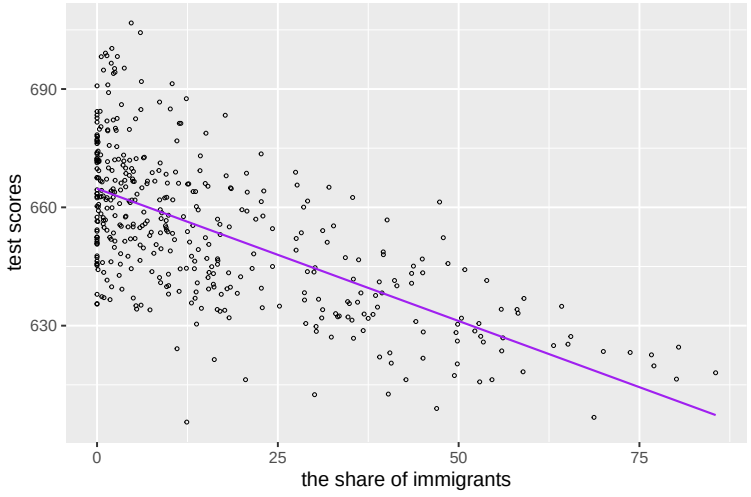


Scatter Plot: The share of immigrants and STR

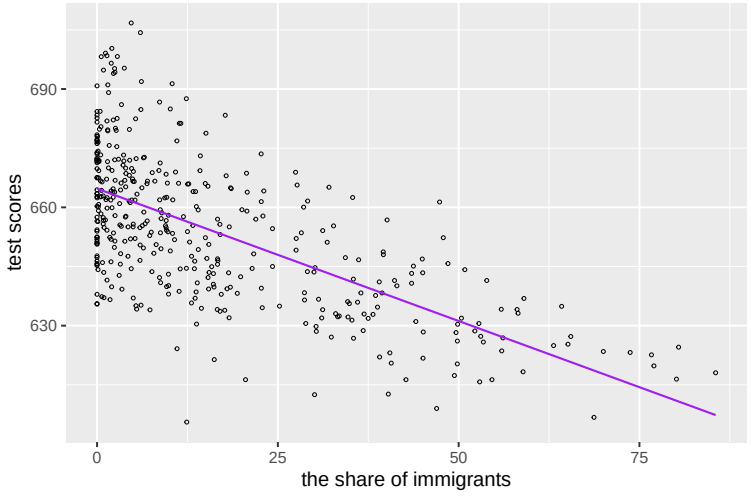


- **higher** share of immigrants, **bigger** class size

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

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

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
  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
```

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 ***
```

Omitted Variable Bias: Examples in R

Table 5: Class Size and Test Score

	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 nonexperimental data.
- OVB 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: *Put omitted the variable into the right side of the regression*, which means our regression model should be

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

- The 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 $E\left[Y_i | X_1..X_k\right]$.

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

- 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

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

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

OLS Estimation in Multiple Regressors

- Similarly in Simple OLS, based on **F.O.C**,the multiple OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ 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**

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

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**

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

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

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**

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

A transformation of FWL theorem

Regression anatomy theorem

The multiple regression model is

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

Then estimator of $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ can be expressed as following

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{j,i} Y_i}{\sum_{i=1}^{n} \tilde{X}_{j,i}^2} \text{ for } j = 1, 2, .., k$$

where $\tilde{X}_{j,i}$ is the fitted OLS residual of the regression X_j on the other Xs.

Test Scores and Student-Teacher Ratios

- Now we put one additional control variables into our OLS regression model

$$Testscore = \beta_0 + \beta_1 STR + \beta_2 elpct + u_i$$

- `elpct`: the share of English learners as an indicator for immigrants

Test Scores and Student-Teacher Ratios(2)

```
tilde.str <- residuals(lm(str ~ el_pct, data=ca))
mean(tilde.str) # should be zero
```

```
#> [1] -1.0111e-16
```

```
sum(tilde.str)  # also is zero
```

```
#> [1] -4.240358e-14
```

Test Scores and Student-Teacher Ratios(3)

- Multiple OLS estimator

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{j,i} Y_i}{\sum_{i=1}^{n} \tilde{X}_{j,i}^2} \text{ for } j = 1, 2, .., k$$

Test Scores and Student-Teacher Ratios(3)

- Multiple OLS estimator

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{j,i} Y_i}{\sum_{i=1}^{n} \tilde{X}_{j,i}^2} \text{ for } j = 1, 2, .., k$$

```
sum(tilde.str*ca$testscr)/sum(tilde.str^2)
```

```
#> [1] -1.101296
```

Test Scores and Student-Teacher Ratios(4)

```
reg3 <- lm(testscr ~ tilde.str,data = ca)
summary(reg3)
```

```
#>
#> Call:
#> lm(formula = testscr ~ tilde.str, data = ca)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -48.693 -14.124   0.988  13.209  50.872
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 654.1565     0.9254 706.864   <2e-16 ***
#> tilde.str    -1.1013     0.4986  -2.209   0.0277 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test Scores and Student-Teacher Ratios(5)

```
reg4 <- lm(testscr ~ str+el_pct,data = ca)
summary(reg4)
```

```
#>
#> 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 ***
#> ---
```

Table 6: Class Size and Test Score

	Dependent variable:	
	testscr	
	(1)	(2)
tilde.str	−1.101**	
	(0.499)	
str		−1.101***
		(0.380)
el_pct		−0.650***
		(0.039)
Constant	654.157***	686.032***
	(0.925)	(7.411)
Observations	420	420
R^2	0.012	0.426

The intuition of partitioned regression

Partialling Out

- First, we regress X_j against the rest of the regressors (and a constant) and keep $\tilde{X}_j$ which is the "part" of X_j that is **uncorrelated**
- Then, to obtain $\hat{\beta}_j$, we regress Y against $\tilde{X}_j$ which is **"clean"** from correlation with other regressors.
- $\hat{\beta}_j$ measures the effect of X_1 after the effects of $X_2, \ldots, X_k$ have been *partialled out or netted out*.

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,regression model turns into

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

- transform it into

$$Y_i = \beta_0 + \beta_1 D_i + \gamma_{2...k} C'_{2...k,i} + u_i, i = 1, ..., n$$

- It turns out

$$Y_i = \alpha + \rho D_i + \gamma C' + u_i$$

OLS Regression, Covariates and RCT

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

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

- Taking the difference

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

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

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

Conditional Independence Assumption(CIA)

"balance" 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

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

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

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

- Thus

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

Controls

Picking Control Variables

- **Questions**: Are "more controls" always better (or at least never worse)?
- **Answer**: It depends on.
- **Irrelevant Variables**: the variables have a **ZERO** partial effect on the dependent variable, thus the coefficient in the population equation is zero.
- **Relevant Variables**: the variables have a **NONZERO** partial effect on the dependent variable.
  - **Non-Omitted Variables**: W is not correlated with X, thus

$$Cov(X_i, W_i) = 0$$

  - **Omitted Variables**: W is correlated with X.

$$Cov(X_i, W_i) \neq 0$$

- **Highly-correlated Variables**: Multicollinearity

Recall: the Standard Error of $\hat{\beta}$

- Our multiple OLS regression model is

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

- Under 4 *basic assumptions*,we can prove the unbiasedness of $\hat{\beta}_j$. Based on the content in multiple OLS and *partitioned regressions*,we have

$$\hat{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j + \frac{\left(\sum_{i=1}^{n} \tilde{X}_{ij} u_i\right)}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)}$$

  - Where $\tilde{X}_{ij}$ is the residual of a regression of X_j on all others X_s

- For simplicity, under the 5th assumption of multiple OLS regression: homoskedastic variance, thus

$$Var(u_i|X_{1i}, X_{2i}..., X_{ki}) = Var(u_i|\boldsymbol{X}) = Var(u_i) = \sigma_u^2$$

  - where $\boldsymbol{X} = X_{1i}X_{2i}..., X_{ki}$

Recall: the Standard Error of $\hat{\beta}$

- Then we have

$$
\begin{aligned}
Var(\hat{\beta}_j) &= Var\left(\boldsymbol{\beta}_j + \frac{\left(\sum_{i=1}^{n} \tilde{X}_{ij} u_i\right)}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)}\right) \\
&= \frac{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2 Var(u_i)\right)}{\left(\sum_{i=1}^{n} \tilde{X}_{i1}^2\right)^2} \\
&= \frac{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2 \sigma_u^2\right)}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)^2} \\
&= \frac{\sigma_u^2}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)}
\end{aligned}
$$

Recall: the Standard Error of $\hat{\beta}$

- **Do not forget**: The $\tilde{X}_{ij}$ is obtained from a multiple OLS regression model

$$X_{ij} = \hat{\delta}_0 + \hat{\delta}_1 X_{1i} + \hat{\delta}_2 X_{2i} + ... + \hat{\delta}_{j-1} X_{j-1,i} + \hat{\delta}_{j+1} X_{j+1,i} + ... \hat{\delta}_k X_{ki} + \tilde{X}_{ji}$$

- The **R-Squared** of this regression is

$$R_j^2 = 1 - \frac{SSR_j}{TSS_j}$$

Recall: the Standard Error of $\hat{\beta}$

- **Do not forget**: The $\tilde{X}_{ij}$ is obtained from a multiple OLS regression model

$$X_{ij} = \hat{\delta}_0 + \hat{\delta}_1 X_{1i} + \hat{\delta}_2 X_{2i} + ... + \hat{\delta}_{j-1} X_{j-1,i} + \hat{\delta}_{j+1} X_{j+1,i} + ... \hat{\delta}_k X_{ki} + \tilde{X}_{ji}$$

- The **R-Squared** of this regression is

$$R_j^2 = 1 - \frac{SSR_j}{TSS_j}$$

$$\Rightarrow SSR_j = TSS_j \times (1 - R_j^2)$$

Recall: the Standard Error of $\hat{\beta}$

- **Do not forget**: The $\tilde{X}_{ij}$ is obtained from a multiple OLS regression model

$$X_{ij} = \hat{\delta}_0 + \hat{\delta}_1 X_{1i} + \hat{\delta}_2 X_{2i} + ... + \hat{\delta}_{j-1} X_{j-1,i} + \hat{\delta}_{j+1} X_{j+1,i} + ... \hat{\delta}_k X_{ki} + \tilde{X}_{ji}$$

- The **R-Squared** of this regression is

$$R_j^2 = 1 - \frac{SSR_j}{TSS_j}$$

$$\Rightarrow SSR_j = TSS_j \times (1 - R_j^2)$$

$$\Rightarrow \tilde{X}_{ij}^2 = \sum_{i=1}^{n} (X_{ji} - \bar{X}_j)^2 (1 - R_j^2)$$

  - where R_j^2 is the **R-squared** from the regression of X_j on all other Xs.

Recall: the Standard Error of $\hat{\beta}$

- Then under 4 basic assumptions and homoskedastic variance of u_i,the **variance** of the OLS estimators β_j simplify to

$$Var\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j} = \frac{\sigma^2_u}{\sum_{i=1}^{n}(X_{ij} - \bar{X})^2(1 - R^2_j)}$$

- Under 3 basic assumptions and homoskedastic variance of u_i, the **variance** of the OLS estimators β_1 simplify to

$$Var\left(\hat{\beta}_1\right) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{\sum(X_i - \bar{X})^2}$$

Irrelevant Variables: Models

- **Irrelevant Variables**: the variables have a ZERO partial effect on the dependent variable, thus the coefficient in the population equation is zero.
- Assume that our model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{8.1}$$

  - Where X_1 is the variable of interest or **treatment variable**.
  - X_2 is a **control variable**, which should be balanced or controlled.
  - X_3 is **irrelevant variable**, thus

$$\beta_3 = 0$$

- The model excluding irrelevant variable is

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_{2i} + v_i \tag{8.2}$$

Irrelevant Variables: Estimate

- Then based on the OVB formula,we have

$$plim\hat{\beta}_1 = \beta_1 + \beta_3 \frac{Cov(\tilde{X}_{12,i}, X_{3i})}{Var\tilde{X}_{12,i}} = \beta_1$$

- the OLS estimator $\hat{\beta}_1$ is still **consistent**.

Irrelevant Variables: Variance

- The variance of $\hat{\beta}_1$ in 8.1 is

$$Var(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_{1i} - \bar{X})^2(1 - R_{23}^2)} \tag{8.3}$$

  Where R_{23}^2 is the R-Squared of the regression of X_1 on X_2 and X_3

- The variance of $\hat{\tilde{\beta}}_1$ in 8.2 is

$$Var(\hat{\tilde{\beta}}_1) = \frac{\sigma_v^2}{\sum^{n} (X_{1i} - \bar{X})^2(1 - R^2)} \tag{8.4}$$

Irrelevant Variables: Variance

- Based on 8.1 and 8.2, we have

$$u_i = Y - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$
$$v_i = Y - \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i} + \tilde{\beta}_2 X_{2i}$$

- Because $\beta_3 = 0$ then $Var(u_i) = Var(v_i) \Rightarrow \sigma_u^2 = \sigma_v^2$
- Because $R_2^2 \leq R_{23}^2$ then we have

$$Var(\hat{\beta}_1) \geq Var(\hat{\tilde{\beta}}_1)$$

- It means controlling an irrelevant variable will only enlarge the variance of the estimator, in other words, make our estimate less precise.

Irrelevant Variables: Wrap up

- The OLS estimator is still unbiased and consistent.
- It increase the variance of estimator, in other words,it will make the estimate **less precise**.
- **Conclusion**: *we should avoid to put irrelevant variables into our regression.*

Relevant Variables: Non-Omitted

- What about *Relevant* but *Non-Omitted* variables? Our regression models is still 8.1 and 8.2, but X_3 now is not an irrelevant variable but a **Non-omitted variable**, thus

$$Cov(X_{1i}, X_{3i}) = 0$$
$$Cov(X_{2i}, X_{3i}) = 0$$

- Then based on the OVB formula,we have

$$plim\hat{\beta}_1 = \beta_1 + \beta_3 \frac{Cov(\tilde{X}_{12,i}, X_{3i})}{Var\tilde{X}_{12,i}} = \beta_1$$

- the OLS estimator $\hat{\beta}_1$ is still **consistent**.

Relevant Variables: Non-Omitted

- Because $Cov(X_{1i}, X_{3i}) = 0$ and $Cov(X_{2i}, X_{3i}) = 0$, then we also have

$$R_2^2 = R_{23}^2$$

- Then the variance of $\hat{\beta}_1$ and $\hat{\tilde{\beta}}_1$ are following respectively.

$$Var(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_{1i} - \bar{X})^2(1 - R_2^2)}$$

$$Var(\hat{\tilde{\beta}}_1) = \frac{\sigma_v^2}{\sum_{i=1}^{n}(X_{1i} - \bar{X})^2(1 - R_2^2)}$$

Revevant Variables: Non-Omitted

- Because $\beta_3 \neq 0$ and $Cov(X_{1i}, X_{3i}) = 0$ and $Cov(X_{2i}, X_{3i}) = 0$,then

$$Var(u_i) \leq Var(v_i) \Rightarrow \sigma_u^2 \leq \sigma_v^2$$

- then we have

$$Var(\hat{\beta}_1) \leq Var(\hat{\tilde{\beta}}_1)$$

- It decrease the variance of estimator, in other words,it will make the estimate **more precise**.
- **Conclusion**: *we should always put Relevant but Non-Omitted Variables into our regression.*

Bad Controls v.s Omitted Variable Bias

- It seems that controlling for more covariates always increases the likelihood that regression estimates have a causal interpretation.
  - often true, but not always.
- eg. Some researchers regressing earnings(Y_i) on schooling(S_i) (and experience) include controls for occupation(O_i). Thus our regression model is

$$Y_i = \beta_0 + \beta_1 S_i + \gamma O_i + u_i$$

  where β_1 is the most of interest coefficient.
- Clearly we can also think of schooling(S_i) affecting the access to higher level occupations(O_i),
  - e.g. you need a Ph.D. to become a university professor. thus

$$O_i = \lambda_0 + \lambda_1 S_i + e_i$$

Bad Controls v.s Omitted Variable Bias

- Assume that the true relation is a two equation system: a simultaneous equations system

$$Y_i = \beta_0 + \beta_1 S_i + \gamma O_i + e_i$$
$$O_i = \lambda_0 + \lambda_1 S_i + u_i$$

- In the case, Occupation O_i is an *endogenous variable*.
- As a result, you could not necessarily estimate the first equation by OLS, which means that the estimation of β_1 is not *unbiased* and *consistent*, because of controlling Occupation(O_i).

Bad Controls: Occupation

- Let us come back to the wage premium of college graduation: the conditional expectation.But now we have additional control variable-*occupations*: *white-color* and *blue-olor*
- Two reasonable assumptions:
  1. white-collar jobs, on average, pay more than blue-collar jobs.
  2. graduating college increases the likelihood of a white-collar job.
- **Question**: Is occupation an omitted variable in the regression of college degree on wage?
- However, should we control for occupation type when considering the effect of college graduation on wages?

Bad Controls: Occupation

- Assume that college degrees are randomly assigned, then we just need to compare the wage difference between workers with college degrees and those without degrees.
- Now we **control** the occupation, which means when we do as follows conditional on occupation:
  - compare degree-earners who chose blue-collar jobs to non-degree-earners who chose blue-collar jobs.
  - or compare degree-earners who chose white-collar jobs to non-degree-earners who chose white-collar jobs.
- Note: the assumption of random degrees says nothing about random job selection.

Bad Controls: Occupation

More formally,

- Y_i denotes i's earnings
- W_i is also a dummy for whether individual i has a white-collar job
- D_i a dummy variable, refers to i's college-graduation status which is randomly assigned, which indicates

$$\{Y_1, Y_0 \perp D\} \text{ and } \{W_1, W_0 \perp D\}$$

- Then

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$
$$W_i = D_i W_{1i} + (1 - D_i) W_{0i}$$

Bad Controls: Occupation

- Because we've assumed D_i is randomly assigned, differences in means yield causal estimates, *i.e.*

$$E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right] = E\left[Y_{1i} - Y_{0i}\right]$$
$$E\left[W_i \mid D_i = 1\right] - E\left[W_i \mid D_i = 0\right] = E\left[W_{1i} - W_{0i}\right]$$

Bad Controls: Occupation

- What happens when we estimate the wage-effect of college graduation for white-collar jobs by controlling occupations?

Bad Controls: Occupation

- What happens when we estimate the wage-effect of college graduation for white-collar jobs by controlling occupations?

$$E\left[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{D}_i = 0\right]$$

Bad Controls: Occupation

- What happens when we estimate the wage-effect of college graduation for white-collar jobs by controlling occupations?

$$
\begin{aligned}
&E\left[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{D}_i = 0\right] \\
&= E\left[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1,\ \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1,\ \mathrm{D}_i = 0\right]
\end{aligned}
$$

Bad Controls: Occupation

- What happens when we estimate the wage-effect of college graduation for white-collar jobs by controlling occupations?

$$
\begin{aligned}
& E\left[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{D}_i = 0\right] \\
&= E\left[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1,\ \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1,\ \mathrm{D}_i = 0\right] \\
&= E\left[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1\right]
\end{aligned}
$$

Bad Controls: Occupation

- What happens when we estimate the wage-effect of college graduation for white-collar jobs by controlling occupations?

$$
\begin{aligned}
& E\left[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{D}_i = 0\right] \\
&= E\left[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1,\ \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1,\ \mathrm{D}_i = 0\right] \\
&= E\left[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1\right] \\
&= E\left[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1\right] + E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1\right]
\end{aligned}
$$

Bad Controls: Occupation

- What happens when we estimate the wage-effect of college graduation for white-collar jobs by controlling occupations?

$$
\begin{aligned}
& E\left[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{D}_i = 0\right] \\
&= E\left[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1,\ \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1,\ \mathrm{D}_i = 0\right] \\
&= E\left[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1\right] \\
&= E\left[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1\right] + E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1\right] \\
&= \underbrace{E\left[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1\right]}_{\text{ATT on white-collar workers}} + \underbrace{E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1\right] - E\left[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1\right]}_{\text{Selection bias}}
\end{aligned}
$$

- By introducing a *bad control*, we introduced **selection bias** into a setting that did not have selection bias without controls.

Bad Controls: Occupation

- Specifically,

$$\underbrace{E\left[Y_{1i} - Y_{0i} \mid W_{1i} = 1\right]}_{\text{ATT on white-collar workers}} + \underbrace{E\left[Y_{0i} \mid W_{1i} = 1\right] - E\left[Y_{0i} \mid W_{0i} = 1\right]}_{\text{Selection bias}}$$

- The First term: Expected potential non-college earnings, given that potential white collar status associated with college education is equal to 1.
- If the occupational choice between white-collor and blue-collor is randomly assigned, then

$$E\left[Y_{0i} \mid W_{1i} = 1\right] = E\left[Y_{0i} \mid W_{0i} = 1\right]$$

- It describes how college graduation changes the composition of the pool of white-collar workers, which in turn change the wage premium between college and high school graduates.
- Even if the true wage causal effect is zero, this selection bias need not be zero.

Bad Controls v.s Omitted Variable Bias

- Putting a bunch of “control” variables might actually be a really **bad idea**: when these variables are themselves **outcomes** of the X variable of interest(another Y).
- But if you don’t control more variables,you may suffer **Omitted Variable Bias**, which also lead a unbiased and inconsistent estimate.
- How to deal with bad control and omitted variable bias, “one of the hard questions in the social sciences” by King(2010).
- Traditionally, economists believe that **Good** control variables should be **fixed** characteristics or **pre-determined** by the time of treatment(Angrist and Pischke,2008).
- A more elaborate way to examine the control variables by logic, we may need new tools
  - Directed Acylic Graphs(DAGs) by Pearl(2009)

Wrap up?

- Which variables belong on the right hand side of a regression equation?
  - Relevant and Omitted Variables : variables determining the treatment and correlated with the outcome.
    - in general these variables will be fixed characteristics or pre-determined by the time of treatment.(Not bad controls)
  - Relevant but Non-omitted Variables: Variables uncorrelated with the treatment but correlated with the outcome.
    - these variables may help reducing standard errors.
- Which variables should NOT be included in the right hand side of the equation?
  - Variables which are outcomes of the treatment itself. These are bad controls.
  - Variables are irrelevant.
  - Variables are highly correlated.