大数据时代的管理决策 (2024 年春)

Lecture 2: OLS Regression Estimation and Inference(I)

Zhaopeng Qu

Nanjing University Business School

March 15 2024

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.
- Based on Rubin Causal Model, **potential outcomes** are the key to causal inference. And RCTs is the golden standard for causal inference.
- 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. Randomized controlled trial(RCTs)
  2. Regression(回归)
  3. Matching and Propensity Score(匹配与倾向得分)
  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. Synthetic Control Method（合成控制法）
- The most fundamental of these tools is **regression**. It compares treatment and control subjects with the same observable characteristics **in a generalized manner**.
- It paves the way for the more elaborate tools used in the class that follow.
- **Let's start our exciting journey from OLS Regression**.

OLS Estimation: Simple Regression

Class Size and Students's Performance

- Recall in the last lecture, we discussed how to find the relationship between class size and students' performance.
- More specifically, we random divide the students into two groups, one with small class size and the other with large class size.
- Then we compare the average test scores of the two groups.
- If the average test scores of the small class size group is higher than the large class size group **significantly**, we can conclude that small class size is better for students' performance.
- However, the answer is really what we want originally?

Question: Class Size and Student's Performance

- **More Quantitative 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 $f(\cdot)$, still two categories: **linear** and **non-linear**
- And we start to use the *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

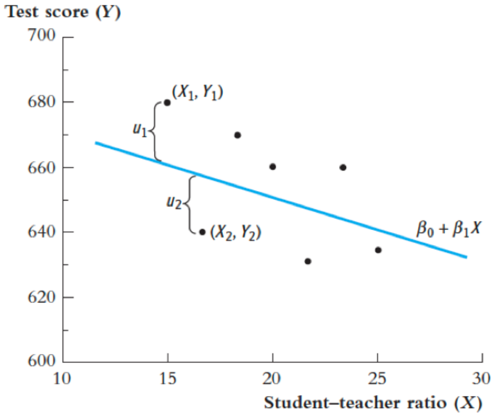
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

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

Graphics for Simple Regression Model

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

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

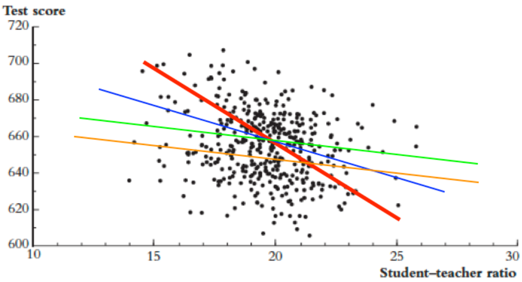


How to find the "best" fitting line?

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

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

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



- So how to find the line that fits the data best?

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 $\hat{Y}_i$, thus

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

The Ordinary Least Squares Estimator (OLS)

The OLS estimator

- The prediction mistake is the resudial, thus 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

OLS estimator of β:

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

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

The Estimated Regression Line

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

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

- Then the regression line is

The Estimated Regression Line

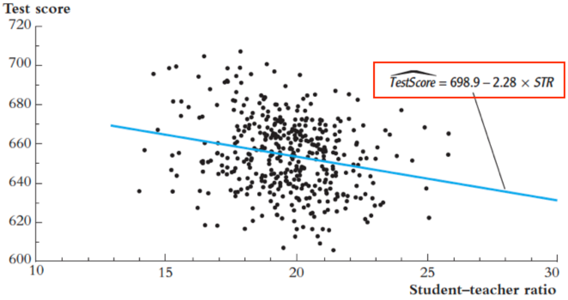
- Obtain the values of OLS estimator for a certain data,

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

- Then the regression line is

FIGURE 4.3 The Estimated Regression Line for the California Data

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



Measures of Fit: The R^2

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

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

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

$$TSS = ESS + SSR$$

Measures of Fit: The R^2

R^2 or the coefficient of determination

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

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

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

The Least Squares Assumptions

The Linear Regression Model

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

The Linear Regression Model

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

Linear Regression Model

Two random variables Y_i and X_i, their relationship can satisfy the linear regression equation, thus

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

- This is not a required assumption. We will extend the model to be nonlinear later on.

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

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

Implications of Assumption 1

With the Iterated Expectation Law, we can obtain an extra implicit assumption about u_i, thus

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

- It seems that the assumption is too strong, but given that the linear regression model have a intercept β_0, which means that we could always make the assumption true by redefining the intercept.

Assumption 1: Conditional Mean is Zero

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

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

Covariance and Conditional Mean

Although $Cov[u_i, X_i] = 0 \nRightarrow E[Y_i|X_i]$, we have

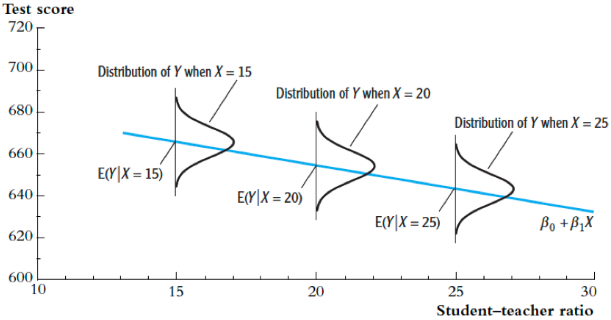
$$Cov[u_i, X_i] \neq 0 \Rightarrow E[u_i|X_i] \neq 0$$

- if u_i and X_i 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 for all values of X.

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

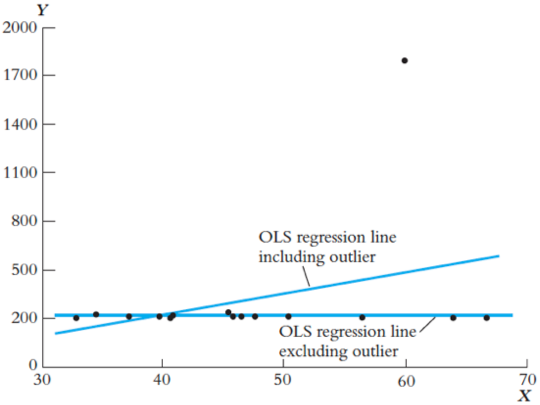
It states that observations with values of X_i, Y_i or both that are far outside the usual range of the data(Outlier) are unlikely. Mathematically, it assume that X and Y have nonzero finite fourth moments.

- 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.

Underlying assumptions of OLS

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

Properties of the OLS Estimators

The OLS estimators

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

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

- We derived the OLS estimators of β_0 andβ_1:

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

The OLS estimators

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

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

- We derived the OLS estimators of β_0 andβ_1:

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

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

Least Squares Assumptions

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

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

Properties of the OLS estimator: unbiasedness

- Skipped the proof of unbiasedness of OLS estimator, but we can show that

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

Review: Conditional Expectation Function(CEF)

- Expectation(for a continuous r.v.)

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

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

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

Review: Properties of CEF

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

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

- where $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ is the conditional probability density function of Y given X.
- 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

$$E[E(Y|X)] = \int E(Y|X=u) f_X(u) du$$
$$= \int \left[\int t f_Y(t|X=u) dt \right] f_X(u) du$$

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

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

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

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

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

Properties of the OLS estimator: Consistency

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

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

- **Then we could obtain**

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

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

Math Review: Continuous Mapping Theorem

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

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

- Example:

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

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

Properties of the OLS estimator: Consistency

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

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

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

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

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

Properties of the OLS estimator: Consistency

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

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**

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

Wrap Up: Unbiasedness vs Consistency

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

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

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

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

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

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

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

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

- Unbiasedness of the OLS estimators implies that

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

- Likewise as $\bar{Y}$,the sample distribution of β_1 or β_0 in a large sample can also *approximates to a normal distribution* based on the **Central Limit theorem(C.L.T)**

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

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

- Where it can be shown that

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

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

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

- We have shown that

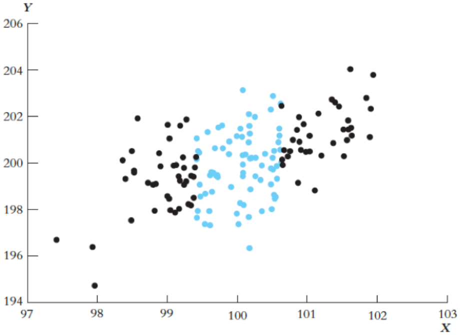
$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{Var[(X_i - \mu_x)u_i]}{[Var(X_i)]^2})$$

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

Variation of X

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

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



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

In a Summary

Under 3 least squares assumptions, the OLS estimators will be

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

Simple OLS and RCT

OLS Regression and RCT

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

Regression when X is a Binary Variable

- For example, we may define D_i as follows:

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

- The regression can be written as

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

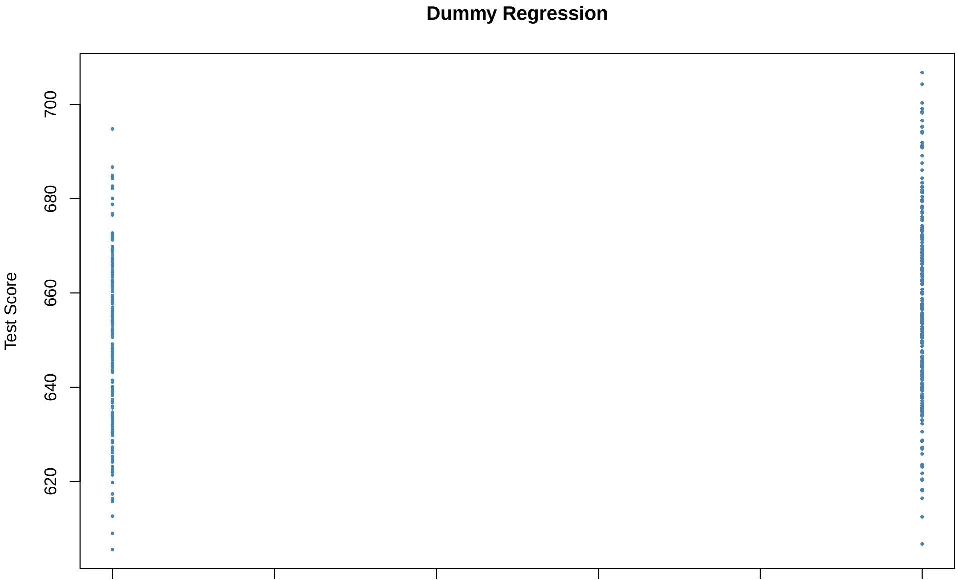
Regression when X is a Binary Variable

- More precisely, the regression model now is

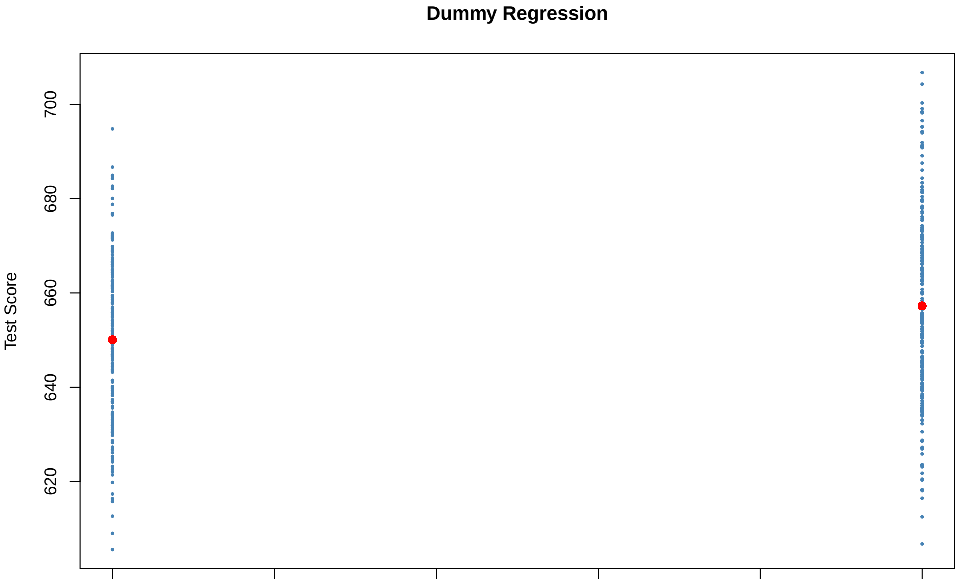
$$TestScore_i = \beta_0 + \beta_1 D_i + u_i \tag{4.3}$$

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

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

- The ATE is the average of the ICE and ATT is the average of the ICE for the treated group.

$$\rho = E(\delta_i) \; or \; \rho = E(\delta_i | D = 1)$$

- Either way, the treatment effect is a constant, i.e., it does not depend on the individual.
- Our OLS regression function is to estimate a constant treatment effect ρ, thus

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

Causality and OLS

- Now write out the conditional expectation of $\mathbf{Y}_i$ for both levels of $\mathbf{D}_i$

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

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

- Take the difference

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

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:* $E\left[\eta_i|\mathbf{D}_i = 1\right] - E\left[\eta_i \mid \mathbf{D}_i = 0\right]$
- 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.
- 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.
- However,it is not easy to obtain in most cases. We have to know how to make a convincing causal inference when these assumptions are not hold.

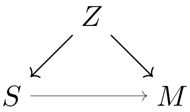
Make Comparison Make Sense

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 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.
- 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

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

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

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

- **Now as** $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D|X \nLeftrightarrow (Y_{1i}, Y_{0i}) \perp\!\!\!\perp D$,

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

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

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

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

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**

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

Identification under Conditional Independence(CIA)

- But using the CIA assumption, then

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

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

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

Identification under Conditional Independence(CIA)

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

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

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

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

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

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.
- The most fundamental tool among them is **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**

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

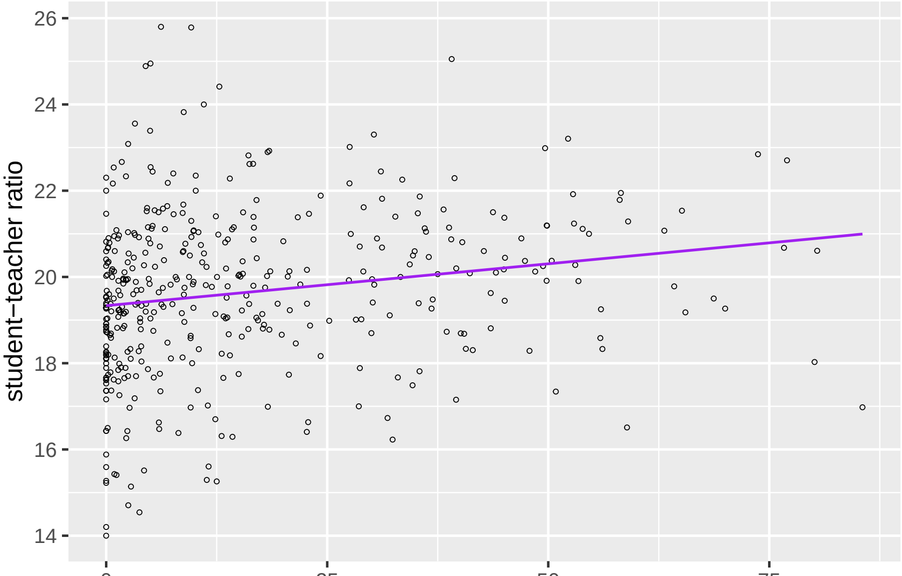
  - It states that u_i are unrelated to X_i in the sense that,given a value of X_i,the mean of these other factors equals **zero**.
  - But what if they (or at least one) are *correlated* with X_i?

Example: Class Size and Test Score

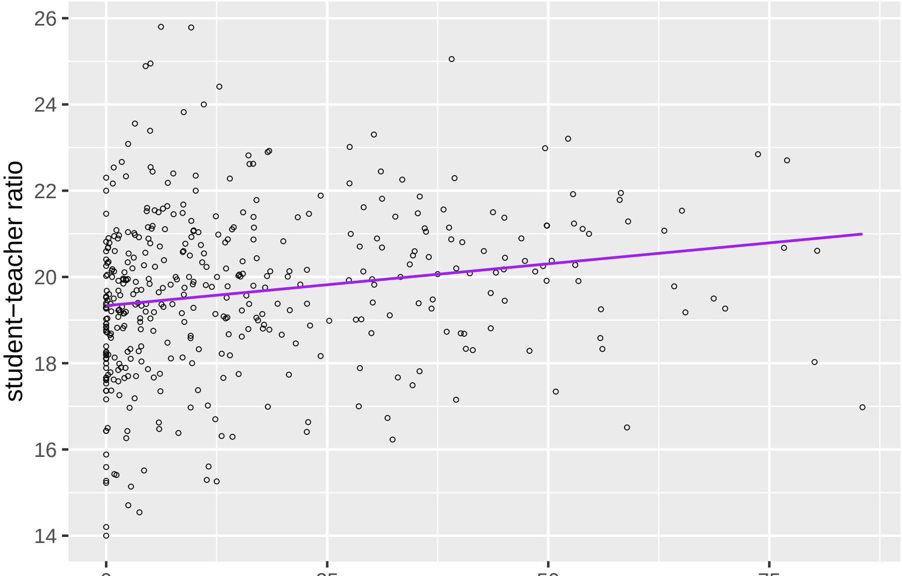
- Many other factors can affect student's performance in the school.
- One of factors is **the share of immigrants** in the class. Because immigrant children may have different backgrounds from native children, such as
  - parents' education level
  - family income and wealth
  - parenting style
  -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

The share of immigrants and STR

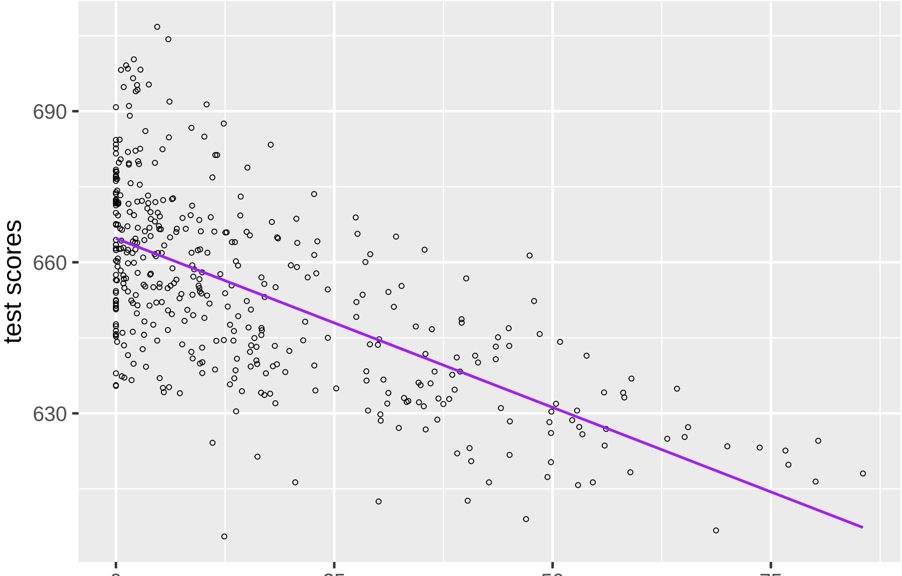
The share of immigrants and STR



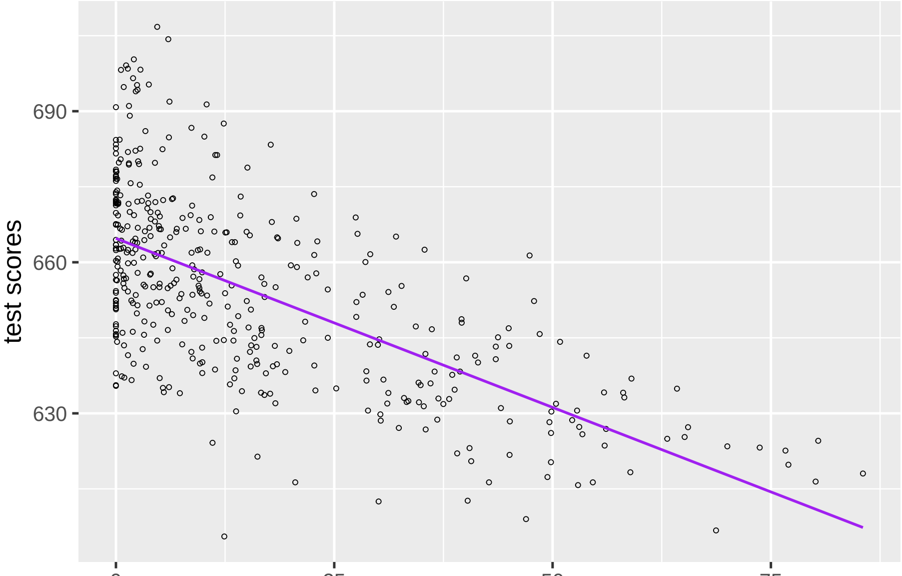
The share of immigrants and STR



The share of immigrants and STR



The share of immigrants and STR



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
#>
#> 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

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 $E[Y_i|X_1..X_k]$.

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

- 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**

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

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**

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

OLS Estimation in Multiple Regressors

- Similar to in Simple OLS, the fitted residuals are

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

- 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 matrices.

Partitioned Regression: OLS Estimators in Multiple Regression

Partitioned regression: OLS estimators

- A useful representation of $\hat{\beta}_j$ could be obtained by the **partitioned regression**, which computed OLS estimators of $\beta_j;\ j = 1, 2...k$ in following 3 steps.
  1. Regress X_j on $X_1, X_2, ...X_{j-1}, X_{j+1}, X_k$, thus

$$X_{j,i} = \gamma_0 + \gamma_1 X_{1i} + ... + \gamma_{j-1} X_{j-1,i} + \gamma_{j+1} X_{j+1,i} ... + \gamma_k X_{k,i} + v_{ji}$$

  2. Obtain the **residuals** from the regression above,denoted as $\tilde{X}_{j,i} = \hat{v}_{ji}$
  3. Regress Y on $\tilde{X}_{j,i}$

- The last step implies that the OLS estimator of β_j can be expressed as follows

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} (\tilde{X}_{ji} - \overline{\tilde{X}}_{ji})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (\tilde{X}_{ji} - \overline{\tilde{X}}_{ji})^2} = \frac{\sum_{i=1}^{n} \tilde{X}_{ji} Y_i}{\sum_{i=1}^{n} \tilde{X}_{ji}^2}$$

Partitioned regression: OLS estimators

- Suppose we want to obtain an expression for $\hat{\beta}_1$.
- Then the first step: regress $X_{1,i}$ on other regressors, thus

Partitioned regression: OLS estimators

- Suppose we want to obtain an expression for $\hat{\beta}_1$.
- Then the first step: regress $X_{1,i}$ on other regressors, thus

$$X_{1,i} = \gamma_0 + \gamma_2 X_{2,i} + \dots + \gamma_k X_{k,i} + v_i$$

Partitioned regression: OLS estimators

- Suppose we want to obtain an expression for $\hat{\beta}_1$.
- Then the first step: regress $X_{1,i}$ on other regressors, thus

$$X_{1,i} = \gamma_0 + \gamma_2 X_{2,i} + \ldots + \gamma_k X_{k,i} + v_i$$

- Then, we can obtain

$$X_{1,i} = \hat{\gamma}_0 + \hat{\gamma}_2 X_{2,i} + \ldots + \hat{\gamma}_k X_{k,i} + \tilde{X}_{1,i}$$

  where $\tilde{X}_{1,i}$ is the fitted OLS residual,thus $\tilde{X}_{j,i} = \hat{v}_{1i}$
- Then we could prove that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \tilde{X}_{1,i} Y_i}{\sum_{i=1}^{n} \tilde{X}_{1,i}^2}$$

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

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

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

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, ..., X_k have been *partialled out or netted out*.

Measures of Fit in Multiple Regression

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.

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

- 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.

Measures of Fit: The Adjusted R^2

- **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.*

Categoried Variable as independent variables in Regression

A Special Case: Categorical Variable as X

- 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:

A Special Case: Categorical Variable as X

- 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:

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

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

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

- 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.

Multiple Regression: Assumption

Multiple Regression: Assumption

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

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

- Assumption 2: $(Y_i, X_{1i}, ..., X_{ki})$ are i.i.d.
- Assumption 3: Large outliers are unlikely.
- Assumption 4: No perfect multicollinearity.

Perfect multicollinearity

Perfect multicollinearity arises when one of the regressors is a **perfect** linear combination of the other regressors.

- Binary variables are sometimes referred to as **dummy variables**
- 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
  - eg. West, Central and East China
- This is called the **dummy variable trap**.
- Solutions to the dummy variable trap: Omit one of the groups or the intercept

Perfect multicollinearity

- 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
#> F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

Perfect multicollinearity

- **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
#> ---
#> 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
#> F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

Perfect multicollinearity

Table 6: Class Size and Test Score

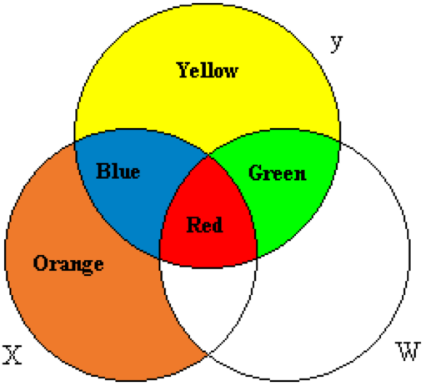
	Dependent variable:	
	testscr	
	(1)	(2)
str	-1.101^{***}	-1.101^{***}
	(0.380)	(0.380)
nel_pct		0.650^{***}
		(0.039)
el_pct	-0.650^{***}	
	(0.039)	
Constant	686.032^{***}	685.382^{***}
	(7.411)	(7.416)

Multicollinearity

Multicollinearity means that two or more regressors are **highly** correlated, but one regressor is **NOT** a perfect linear function of one or more of the other regressors.

- **multicollinearity** is **NOT** a violation of OLS assumptions.
  - It does not impose theoretical problem for the calculation of OLS estimators.
- But if two regressors are highly correlated, then the the coefficient on at least one of the regressors is imprecisely estimated (high variance).
- To what extent two correlated variables can be seen as “highly correlated”?
  - **rule of thumb**: correlation coefficient is over **0.8**.

Venn Diagrams for Multiple Regression Model



- In a simple model (y on X), OLS uses 'Blue' + 'Red' to estimate β.
- When y is regressed on X and W: OLS throws away the red area and just uses blue to estimate β.
- Idea: Red area is contaminated(we do not know if the movements in y are due to X or to W).

Venn Diagrams for Multicollinearity

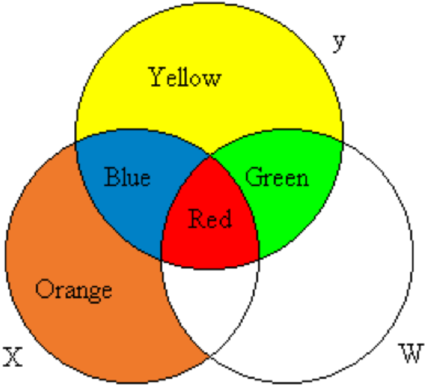


Figure 3a Modest collinearity

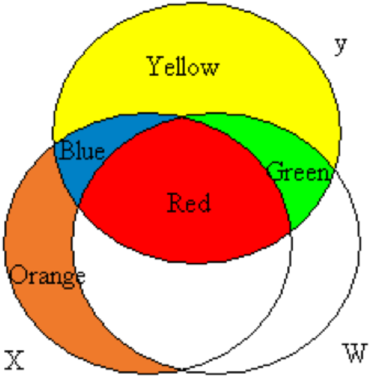


Figure 3b Considerable collinearity

Venn Diagrams for Multicollinearity

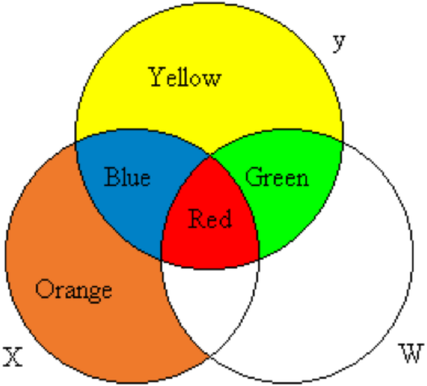


Figure 3a Modest collinearity

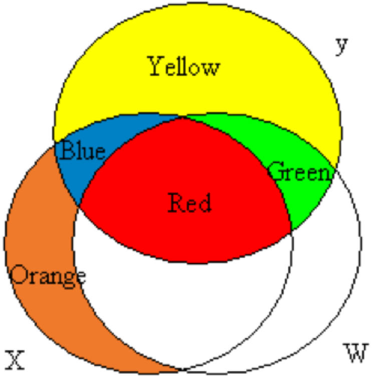


Figure 3b Considerable collinearity

- Less information (compare the Blue and Green areas in both figures) is used, the estimation is less precise.

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.

Picking Control Variables

- **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
- We will come back soon to discuss this topic again in Lecture 8 in details.

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

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

- Taking the difference

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

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)

”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

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

- 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 important assumption is

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

or

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

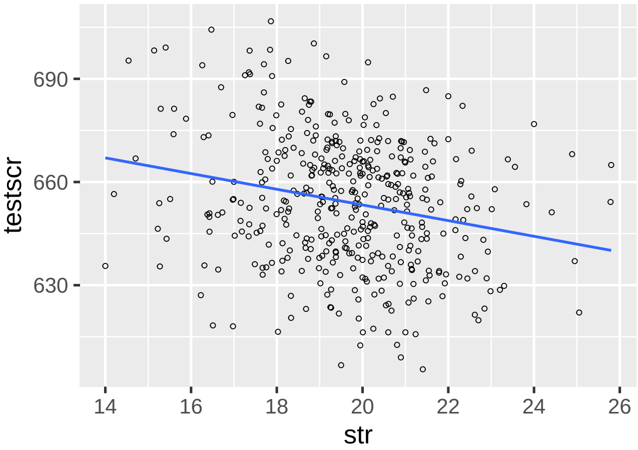
- In most cases,it does not satisfy it when using nonexperimental data. Therefore,how to make a convincing causal inference when these assumptions are not held is the key question.

Hypothesis Testing

Introduction: Class size and Test Score

Recall our simple OLS regression mode is

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



Class Size and Test Score

Then we got the result of a simple OLS regression

$$\widehat{TestScore} = 698.9 - 2.28 \times STR,\ R^2 = 0.051, SER = 18.6$$

- **Don't forget**: the result are not obtained from the population **but from the sample**.
- How can you be sure about the result? In other words, *how confident* you can believe the result from the sample inferring to the population?
- If someone believes that cutting the class size will not help boost test scores. Can you reject the claim based your *scientific evidence-based* data analysis?
- This is the work of **Hypothesis Testing** in OLS regressions.

Review: Hypothesis Testing

- A hypothesis is (usually) an **assertion** or **statement** about **unknown population parameters** like θ.
- Suppose we want to test whether it is significantly different from a certain value μ_0
- Then **null hypothesis** is

$$H_0 : \theta = \mu_0$$

- The alternative hypothesis(two-sided) is

$$H_1 : \theta \neq \mu_0$$

- If the value μ_0 does not lie within the calculated confidence interval, then we *reject the null hypothesis*.
- If the value μ_0 lie within the calculated confidence interval, then we *fail to reject the null hypothesis*.

Review: Hypothesis Testing

- Most countries follow the rule of criminal trials:
  innocent until proven guilty(疑罪从无)
  - The jury or judge starts with the "null hypothesis" that the accused person is innocent.
  - The prosecutor wants to prove their hypothesis that the accused person is guilty.
  - In other words, they have to show strong evidence to make the jury or judge reject the "null hypothesis".
- Likewise, our rule in econometrics is
  presumption of insignificance until proven.
  - At first researchers have to assume that there is **zero** impact of independent variable on dependent variable.
  - In order to prove the relationship between the independent variable and dependent variable, we must provide strong enough evidence to convince readers or policy makers to "reject" the assumption of a **zero** effect.

Review: Two Type Errors(两种错误)

- In both cases, there is a certain risk that our conclusion is wrong

	H_0 is true	H_A is true
Fail to reject H_O	Correct	Type II error
Reject H_O	Type I error	Correct

- Type I and Type II errors can not happen at the same time
- There is a trade-off between Type I and Type II errors

Review: Two Type Errors(两种错误)

- Question: Determine whether each situation belongs to Type I error or Type II error.
  - “宁可错杀一千，不能放过一个”
  - “宁可放过一千，不能错杀一个”

The Significance level(显著性水平)

- The significance level or size of a test, α, is the **maximum probability** of the Type I Error we tolerate.

$$P(Type\ I\ error) = P(reject\ H_0 \mid H_0\ is\ true) = \alpha$$

- In social science, the usual significance level is set at 5%. A less rigorous standard is 10%, whereas a more stringent one is 1%.

The Power of the Test

- The power of a test, is $1 - \beta$, where β is the **probability** of the Type II Error

$$1 - P(Type\ II\ error) = 1 - P(reject\ H_0 \mid H_1\ is\ true) = 1 - \beta$$

- Typically, we desire power to be 0.80 or greater, which alternatively equal to minimize $\beta \leq 0.2$.

Review: Hypothesis Testing of Population Mean

- Let $\mu_{Y,c}$ is a specific value to which the population mean equals(thus we suppose)
  - **the null hypothesis**:

$$H_0 : E(Y) = \mu_{Y,c}$$

  - **the alternative hypothesis(two-sided)**:

$$H_1 : E(Y) \neq \mu_{Y,c}$$

Review: Hypothesis Testing of Population Mean

- Step 1 Compute the *sample mean* $\overline{Y}$
- Step 2 Compute the *standard error* of $\overline{Y}$, recall

$$SE(\overline{Y}) = \frac{s_Y}{\sqrt{n}}$$

- Step 3 Compute the *t-statistic* actually computed

$$t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,c}}{SE(\bar{Y})}$$

- Step 4 Compute the p-value(optional)

$$\text{p-value} = 2\Phi(-|t^{act}|)$$

- Step 5 See if we can **Reject the null hypothesis** at a certain significance level α,like 5%, or p-value is less than significance level.

$$|t^{act}| > critical\ value \text{ or } p - value < significance\ level$$

Simple OLS: Hypotheses Testing

- A Simple OLS regression

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

- This is the population regression equation and the key **unknown population parameters** is β_1.
- Then we would like to test whether β_1 equals to a specific value $\beta_{1,s}$ or not
  - **the null hypothesis**:

$$H_0 : \beta_1 = \beta_{1,s}$$

  - **the alternative hypothesis**:

$$H_1 : \beta_1 \neq \beta_{1,s}$$

A Simple OLS: Hypotheses Testing

- **Step1: Estimate $Y_i = \beta_0 + \beta_1 X_i + u_i$ by OLS to obtain $\hat{\beta}_1$**
- **Step2: Compute the *standard error* of $\hat{\beta}_1$**
- **Step3: Construct the *t-statistic***

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,c}}{SE\left(\hat{\beta}_1\right)}$$

- **Step4: *Reject the null hypothesis* if**

$$| \, t^{act} \, | > critical \ value$$

$$or \quad p - value < significance \ level$$

Recall: General Form of the t-statistics

$$t = \frac{estimator - hypothesized\ value}{standard\ error\ of\ the\ estimator}$$

- Now the key unknown statistic is the **standard error**(S.E).

The Standard Error of $\hat{\beta}_1$

- **Recall** from the **Simple OLS Regression**
  - if the least squares assumptions hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a joint normal sampling distribution,thus $\hat{\beta}_1$

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

  - We also derived the form of the variance of the normal distribution, $\sigma^2_{\hat{\beta}_1}$ is

$$\sigma_{\hat{\beta}_1} = \sqrt{\frac{1}{n}\frac{Var[(X_i - \mu_X)u_i]}{[Var(X_i)]^2}} \tag{4.21}$$

- The **standard error** of $\hat{\beta}_1$ is an **estimator** of the standard deviation of the sampling distribution $\sigma_{\hat{\beta}_1}$, thus

$$SE\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum(X_i - \bar{X})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]^2}} \tag{5.4}$$

Application to Test Score and Class Size

```
. regress test_score class_size, robust

Linear regression                               Number of obs     =        420
                                                  F(1, 418)         =      19.26
                                                  Prob > F          =     0.0000
                                                  R-squared         =     0.0512
                                                  Root MSE          =     18.581
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

- **the OLS regression line**

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{22.8} \times STR, \; R^2 = 0.051, SER = 18.6$$

Testing a two-sided hypothesis concerning β_1

- **the null hypothesis** $H_0 : \beta_1 = 0$
  - It means that the class size will not affect the performance of students.
- **the alternative hypothesis** $H_1 : \beta_1 \neq 0$
  - It means that the class size do affect the performance of students (whatever positive or negative)
- Our primary goal is to **Reject the null**, and then say make a conclusion:
  - Class Size **does matter** for the performance of students.

Testing a two-sided hypothesis concerning β_1

- **Step1: Estimate** $\hat{\beta}_1 = -2.28$
- **Step2: Compute the standard error:** $SE(\hat{\beta}_1) = 0.52$
- **Step3: Compute the *t-statistic***

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,c}}{SE\left(\hat{\beta}_1\right)} = \frac{-2.28 - 0}{0.52} = -4.39$$

- **Step4: Reject the null hypothesis if**
  - $\mid t^{act} \mid = \mid -4.39 \mid > critical\ value = 1.96$
  - $p - value = 0 < significance\ level = 0.05$

Application to Test Score and Class Size

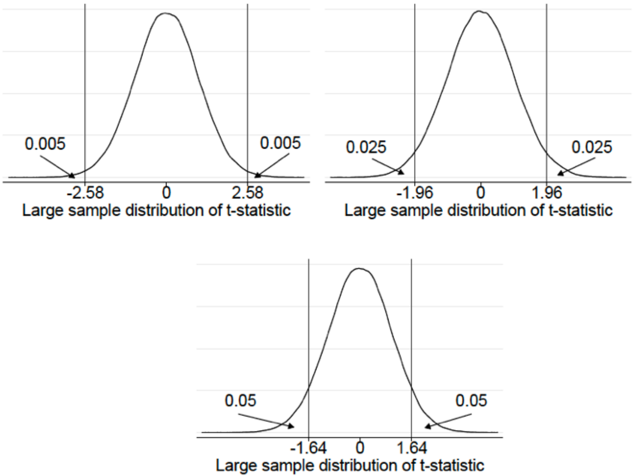
```
. regress test_score class_size, robust

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      19.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0512
                                                Root MSE          =     18.581
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

- **We can reject the null hypothesis that $H_0 : \beta_1 = 0$, which means $\beta_1 \neq 0$ with a high probability(over 95%).**
- It suggests that Class size **matters** the students' performance in a very high chance.

Critical Values of the t-statistic

The critical value of t-statistic depends on significance level α



1% and 10% significant levels

- **Step4: Reject the null hypothesis at a 10% significance level**
  - $\mid t^{act} \mid = \mid -4.39 \mid > critical\ value = 1.64$
  - $p - value = 0.00 < significance\ level = 0.1$
- **Step4: Reject the null hypothesis at a 1% significance level**
  - $\mid t^{act} \mid = \mid -4.39 \mid > critical\ value = 2.58$
  - $p - value = 0.00 < significance\ level = 0.01$

Two-Sided Hypotheses: β_1 in a certain value

- **Step1: Estimate** $\hat{\beta}_1 = -2.28$
- **Step2: Compute the standard error:** $SE(\hat{\beta}_1) = 0.52$
- **Step3: Compute the t-statistic**

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,c}}{SE\left(\hat{\beta}_1\right)} = \frac{-2.28 - (-2)}{0.52} = -0.54$$

- **Step4: can't reject the null hypothesis at 5% significant level because**
  - $\mid t^{act} \mid = \mid -0.54 \mid < critical\ value = 1.96$
  - $p - value = 0.59 > significance\ level = 0.05$

Two-Sided Hypotheses : β_1 in a certain value

```
. lincom class_size-(-2)

 ( 1)  class_size = -2
```

test_score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	-.2798083	.5194892	-0.54	0.590	-1.300945	.7413286

- We cannot reject the null hypothesis that $H_0 : \beta_1 = -2$.
- It suggests that *there is no enough evidence* to support the statement:
  - cutting class size in one unit will boost the test score in 2 points.

One-sided Hypotheses Concerning β_1

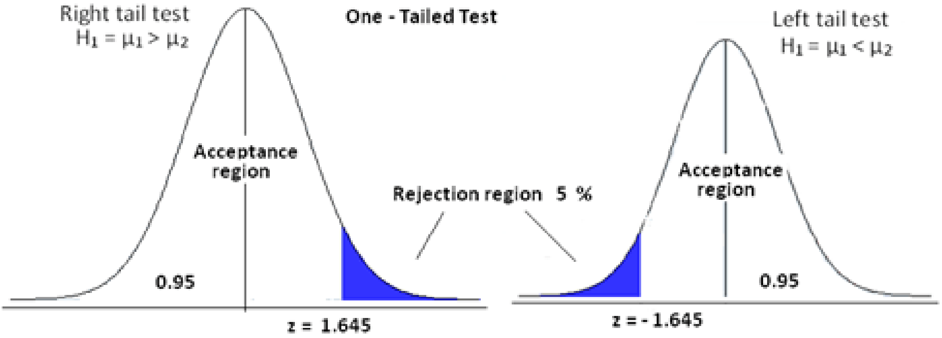
- Sometimes, we want to do a *one-sided Hypothesis testing*
- the null hypothesis is still unchanged $H_0 : \beta_1 = -2$
- **the alternative hypothesis** is $H_1 : \beta_1 < -2$
  - The statement is that reducing(or inversely increasing) class size will boost(or lower) student's performance.
  - More specifically,cutting class size in one unit will increase the test score in 2 points **at least**.
- Because the null hypothesis is the same for a one- and a two-sided hypothesis test, the construction of the t-statistic is the same.
- The difference between the two is the critical value and p-value.

One-sided Hypotheses Concerning β_1

- **Step1: Estimate** $\hat{\beta}_1 = -2.28$
- **Step2: Compute the standard error:** $SE(\hat{\beta}_1) = 0.52$
- **Step3: Compute the t-statistic**

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE\left(\hat{\beta}_1\right)} = \frac{-2.28 - (-2)}{0.52} = -0.54$$

One-sided Hypotheses Concerning β_1



One-sided Hypotheses Concerning β_1

- Step4: under the circumstance, the critical value is not the -1.96 but -1.645 at 5% significant level.
- We can't reject the null hypothesis because

$$t^{act} = -0.54 > critical\ value = -1.645$$

- The p-value is not the $2\Phi(-|t^{act}|)$ now but $Pr(Z < t^{act}) = \Phi(t^{act})$.
- It suggests that *there is NO enough evidence* to support the statement:cutting class size in one unit will increase the test score in **2 points at least**.

One-sided Hypotheses Concerning β_1

- One-sided alternative hypotheses should be used only when there is a clear reason for doing so.
- This reason could come from economic theory, prior empirical evidence, or both.
- However, even if it initially seems that the relevant alternative is one-sided, upon reflection this might not necessarily be so.
- In practice, one-sided test is used much less than two-sided test.

Wrap up

- Hypothesis tests are useful if you have a specific null hypothesis in mind (as did our angry taxpayer).
- Being able to accept or reject this null hypothesis based on the statistical evidence provides a powerful tool for coping with the uncertainty inherent in using a sample to learn about the population.
- Yet, there are many times that no single hypothesis about a regression coefficient is dominant, and instead one would like to know a range of values of the coefficient that are consistent with the data.
- This calls for constructing a **confidence interval**.

Confidence Intervals

Introduction

- Because any statistical estimate of the slope β_1 necessarily has sampling uncertainty, we cannot determine the true value of β_1 exactly from a sample of data.
- It is possible, however, to use the OLS estimators and its standard error to construct a confidence interval for the slope β_1

CI for β_1

- Method for constructing a confidence interval for a population mean can be easily extended to constructing a confidence interval for a regression coefficient.
- Using a two-sided test, a hypothesized value for β_1 will be rejected at 5% significance level if

$$| t^{act} | > critical\ value = 1.96$$

- So $\hat{\beta}_1$ will be in the *confidence set* if $| t^{act} | \leq critical\ value = 1.96$
- Thus the 95% confidence interval for β_1 are within ± 1.96 standard errors of $\hat{\beta}_1$

$$\hat{\beta}_1 \pm 1.96 \cdot SE\left(\hat{\beta}_1\right)$$

CI for $\beta_{ClassSize}$

```
. regress test_score class_size, robust

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      19.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0512
                                                Root MSE          =     18.581
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf. Interval]	
class_size	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

- **Thus the 95% confidence interval for β_1 are within ± 1.96 standard errors of $\hat{\beta}_1$**

$$\hat{\beta}_1 \pm 1.96 \cdot SE\left(\hat{\beta}_1\right) = -2.28 \pm (1.96 \times 0.519) = [-3.3, -1.26]$$

Gauss-Markov theorem and Heteroskedasticity

Introduction

- Recall we discussed the properties of $\bar{Y}$ in Chapter 2.
  - an **unbiased** estimator of μ_Y
  - a **consistent** estimator of μ_Y
  - an **approximate normal sampling distribution** for large n

The Efficiency of $\bar{Y}$

- the fourth properties of $\bar{Y}$ in Chapter 3.
- the **Best Linear Unbiased Estimator(BLUE)**: $\bar{Y}$ is the most efficient estimator of μ_Y among all unbiased estimators that are weighted averages of $Y_1, ..., Y_n$, presented by $\hat{\mu}_Y = \frac{1}{n}\sum a_i Y_i$,thus,

$$Var(\overline{Y}) < Var(\hat{\mu}_Y)$$

Unnecessary Assumption for Simple OLS

- Three Simple OLS Regression Assumptions
  - Assumption 1
  - Assumption 2
  - Assumption 3
- Assumption 4: The error terms are **homoskedastic**

$$Var(u_i \mid X_i) = \sigma_u^2$$

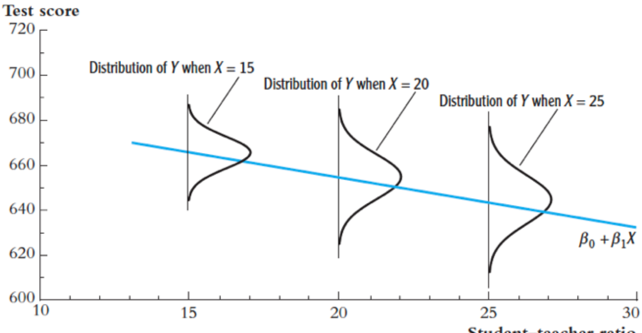
- Then $\hat{\beta}^{OLS}$ is the **Best Linear Unbiased Estimator(BLUE)**: it is the most efficient estimator of β_1 among all conditional unbiased estimators that are a linear function of $Y_1, Y_2, ..., Y_n$.

Heteroskedasticity & homoskedasticity

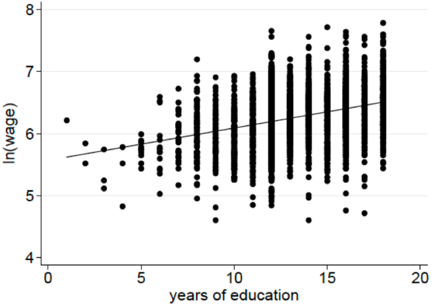
- The error term u_i is **homoskedastic** if the variance of the conditional distribution of u_i given X_i is constant for $i = 1, \ldots n$, in particular does not depend on X_i.
- Otherwise, the error term is **heteroskedastic**.

FIGURE 5.2 An Example of Heteroskedasticity

Like Figure 4.4, this shows the conditional distribution of test scores for three different class sizes. Unlike Figure 4.4, these distributions become more spread out (have a larger variance) for larger class sizes. Because the variance of the distribution of u given X, var($u|X$), depends on X, u is



An Actual Example: the returns to schooling



- The spread of the dots around the line is clearly increasing with years of education X_i.
- Variation in (log) wages is higher at higher levels of education.
- This implies that

$$Var(u_i \mid X_i) \neq \sigma_u^2$$

Homoskedasticity: S.E.

- However,in many applications homoskedasticity is **NOT a plausible assumption**.
- If the error terms are *heteroskedastic*, then you use the *homoskedastic* assumption to compute the S.E. of $\hat{\beta}_1$. It will leads to
  - The standard errors are wrong (often too small)
  - The t-statistic does NOT have a $N(0,1)$ distribution (also not in large samples).
  - But the estimating coefficients in OLS regression will not *change*.

Heteroskedasticity & homoskedasticity

- If the error terms are **heteroskedastic**, we should use the original equation of S.E.

$$SE_{Heter}\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum(X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]^2}}$$

- It is called as *heteroskedasticity robust-standard errors*,also referred to as Eicker-Huber-White standard errors,simply **Robust-Standard Errors**
- In the case, it is not difficult to find that *homoskedasticity* **is just a special case of *heteroskedasticity***.

Heteroskedasticity & homoskedasticity

- Since homoskedasticity is a special case of heteroskedasticity, these heteroskedasticity robust formulas are also **valid** if *the error terms are homoskedastic*.
- Hypothesis tests and confidence intervals based on above SE's are *valid* both in case of homoskedasticity and heteroskedasticity.
- In reality, since in many applications homoskedasticity is not a plausible assumption, *it is best to use heteroskedasticity robust standard errors*. Using **robust standard errors** rather than **standard errors with homoskedasticity** will lead us **lose nothing**.

Heteroskedasticity & homoskedasticity

- It can be quite cumbersome to do this calculation by hand.Luckily,computer can help us do the job.
  - In `Stata`, the default option of regression is to assume homoskedasticity, to obtain heteroskedasticity robust standard errors use the option "robust":

$$regress\ y\ x\ ,\ robust$$

  - In `R`, many ways can finish the job. A convenient function named `vcovHC()` is part of the package `sandwich`.

Test Scores and Class Size

```
. regress test_score class_size

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------

. regress test_score class_size, robust

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      19.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0512
                                                Root MSE          =     18.581

------------------------------------------------------------------------------
             |               Robust
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

Test Scores and Class Size

```
. regress test_score class_size

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------

. regress test_score class_size, robust

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      19.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0512
                                                Root MSE          =     18.581

------------------------------------------------------------------------------
             |               Robust
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

Wrap up: Heteroskedasticity in a Simple OLS

- If the error terms are heteroskedastic
  - The fourth simple OLS assumption is violated.
  - The Gauss-Markov conditions do not hold.
  - The OLS estimator is not BLUE (not most efficient).
- But (given that the other OLS assumptions hold)
  - The OLS estimators are still *unbiased*.
  - The OLS estimators are still *consistent*.
  - The OLS estimators are *normally distributed* in large samples

OLS with Multiple Regressors: Hypotheses tests

Recall: the Multiple OLS Regression

- The multiple regression model is

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

- Four Basic Assumptions
  - Assumption 1 : $E[u_i \mid X_{1i}, X_{2i} ..., X_{ki}] = 0$
  - Assumption 2 : i.i.d sample
  - Assumption 3 : Large outliers are unlikely.
  - Assumption 4 : No perfect multicollinearity.
- The Sampling Distribution: the OLS estimators $\hat{\beta}_j$ for $j = 1, ..., k$ are approximately normally distributed in large samples.

Standard Errors for the Multiple OLS Estimators

- There is *nothing* conceptually different between the single- or multiple-regressor cases.
  - Standard Errors for a Simple OLS estimator β_1

$$SE\left(\hat{\beta}_1\right) = \hat{\sigma}_{\hat{\beta}_1}$$

  - Standard Errors for Mutiple OLS Regression estimators β_j

$$SE\left(\hat{\beta}_j\right) = \hat{\sigma}_{\hat{\beta}_j}$$

- Remind: since now the joint distribution is not only for (Y_i, X_i),but also for (X_{ij}, X_{ik}).
- The formula for the *standard errors* in Multiple OLS regression are related with a *matrix* named **Variance-Covariance matrix**

Hypothesis Tests for a Single Coefficient

- the *t-statistic* in Simple OLS Regression

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,c}}{SE\left(\hat{\beta}_1\right)} \sim N(0,1)$$

- the *t-statistic* in Multiple OLS Regression

$$t = \frac{\hat{\beta}_j - \beta_{j,c}}{SE\left(\hat{\beta}_j\right)} \sim N(0,1)$$

Hypothesis testing for single coefficient

- $H_0 : \beta_j = \beta_{j,c}$ $H_1 : \beta_1 \neq \beta_{j,c}$
- **Step1: Estimate $\hat{\beta}_j$, by run a multiple OLS regression**

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

- **Step2: Compute the standard error of $\hat{\beta}_j$ (*requires matrix algebra*)**
- **Step3: Compute the t-statistic**

$$t^{act} = \frac{\hat{\beta}_j - \beta_{j,c}}{SE\left(\hat{\beta}_j\right)}$$

- **Step4: Reject the null hypothesis if**
  - $\mid t^{act} \mid > critical\ value$
  - **or if** $p - value < significance\ level$

Confidence Intervals for a single coefficient

- **Also the same as in a simple OLS Regression.**
- $\hat{\beta}_j$ **will be in the confidence set if** $|\ t^{act}\ | \leq critical\ value = 1.96$ **at the 95% confidence level.**
- **Thus the 95% confidence interval for** β_j **are within** ± 1.96 **standard errors of** $\hat{\beta}_j$

$$\hat{\beta}_j \pm 1.96 \cdot SE\left(\hat{\beta}_j\right)$$

Test Scores and Class Size

```
. regress test_score class_size el_pct,robust

Linear regression                               Number of obs     =        420
                                                F(2, 417)         =     223.82
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4264
                                                Root MSE          =     14.464
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-1.101296	.4328472	-2.54	0.011	-1.95213	-.2504616
el_pct	-.6497768	.0310318	-20.94	0.000	-.710775	-.5887786
_cons	686.0322	8.728224	78.60	0.000	668.8754	703.189

Case: Class Size and Test scores

- Does changing class size, while holding the percentage of English learners constant, have a statistically significant effect on test scores? (using a 5% significance level)
- $H_0 : \beta_{ClassSize} = 0 \; H_1 : \beta_{ClassSize} \neq 0$
- Step1: Estimate $\hat{\beta}_1 = -1.10$
- Step2: Compute the standard error: $SE(\hat{\beta}_1) = 0.43$
- Step3: Compute the t-statistic

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,c}}{SE\left(\hat{\beta}_1\right)} = \frac{-1.10 - 0}{0.43} = -2.54$$

- Step4: Reject the null hypothesis if
  - $\mid t^{act} \mid = \mid -2.54 \mid > critical\ value.1.96$
  - $p - value = 0.011 < significance\ level = 0.05$

Tests of Joint Hypotheses: on 2 or more coefficients

- **Question**: Can we just test more than one individual coefficient at a time?
- Suppose the angry taxpayer hypothesizes that neither the *student–teacher ratio* nor *expenditures per pupil* have an effect on test scores, once we control for the *percentage of English learners*.
- Therefore, we have to test a **joint null hypothesis** that both the coefficient on **student–teacher ratio** and the coefficient on **expenditures per pupil** are zero?

$$H_0 : \beta_{str} = 0 \;\&\; \beta_{expn} = 0,$$

$$H_1 : \beta_{str} \neq 0 \; and/or \; \beta_{expn} \neq 0$$

Testing 1 hypothesis on 2 or more coefficients

- Suppose we want to test

$$H_0 : \beta_1 = 0 \ \& \ \beta_2 = 0 \quad H_1 : \beta_1 \neq 0 \ and/or \ \beta_2 \neq 0$$

- Then the *F-statistic* can also combine the two *t-statistics* t_1 and t_2 as follows

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1 t_2} t_1 t_2}{1 - \hat{\rho}_{t_1 t_2}^2}\right)$$

where $\hat{\rho}_{t_1 t_2}$ is an estimator of the correlation between the two t-statistics.

Testing 1 hypothesis on 2 or more coefficients

- In general, a joint hypothesis is a hypothesis that imposes two or more restrictions on the regression coefficients.

$$H_0 : \beta_j = \beta_{j,c}, \beta_k = \beta_{k,c}, ..., for\ a\ total\ of\ q\ restrictions$$
$$H_1 : one\ or\ more\ of\ q\ restrictions\ under\ H_0\ does\ not\ hold$$

  - where $\beta_j, \beta_k, ...$ refer to different regression coefficients.
- When the regressors are highly correlated, single **t-statistics** can be misleading.Instead, we use the **F-statistic** for testing joint hypotheses.

Unrestricted v.s Restricted model

- **The unrestricted model**: the model without any of the restrictions imposed. It contains all the variables.
- **The restricted model**: the model on which the restrictions have been imposed.
- And we want to test that $H_0 : \beta_1 = 0 \; and \; \beta_2 = 0$,then $H_1 : \beta_1 \neq 0 \; and/or \; \beta_2 \neq 0$ for the regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i, i = 1, ..., n$$

- Then restricted model is

$$Y_i = \beta_0 + \beta_3 X_{3,i} + u_i$$

The F-statistic with q restrictions

- The F-statistic is computed using a simple formula based on the sum of squared residuals from two regressions.

$$F = \frac{(SSR_{\text{restricted}} - SSR_{\text{unrestricted}})/q}{SSR_{\text{unrestricted}}/(n-k-1)}$$

- $SSR_{restricted}$ is the sum of squared residuals from the **restricted** regression.
- $SSR_{unrestricted}$ is the sum of squared residuals from the **full** model.
- q is the number of restrictions under the null.
- k is the number of regressors in the unrestricted regression.

The heteroskedasticity-robust F-statistic

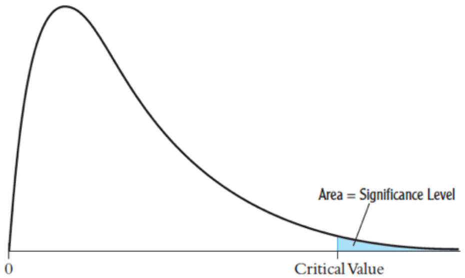
- Using matrix to show the form of the heteroskedasticity-robust F-statistic which is **beyond the scope of our class**.
- While,under the null hypothesis,regardless of whether the errors are homoskedastic or heteroskedastic, the F-statistic with q has a sampling distribution in large samples,

$$F - statistic \sim F_{q,\infty}$$

  - where q is the number of restrictions
- Then we can compute the F-statistic, the critical values from the table of the $F_{q,\infty}$ and obtain the p-value.

F-Distribution

TABLE 4 Critical Values for the $F_{m,\infty}$ Distribution



	Significance Level		
Degrees of Freedom	10%	5%	1%
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78

Testing joint hypothesis with q restrictions

- $H_0 : \beta_j = \beta_{j,0}, ..., \beta_m = \beta_{m,0}$ for a total of q restrictions.
- H_1:at least one of q restrictions under H_0 does not hold.
- Step1: Estimate

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

  by OLS
- Step2: Compute the **F-statistic**
- Step3 : Reject the null hypothesis if

$$F - Statistic > F^{act}_{q,\infty}$$

  or

$$p - value = Pr[F_{q,\infty} > F^{act}] <= significant\ level$$

Case: Class Size and Test Scores

- **We want to test hypothesis that both the coefficient on *student–teacher ratio* and the coefficient on *expenditures per pupil* are zero?**
  - $H_0 : \beta_{str} = 0 \ \& \beta_{expn} = 0$
  - $H_1 : \beta_{str} \neq 0 \ and/or \ \beta_{expn} \neq 0$
- **The null hypothesis consists of two restrictions** $q = 2$

Case: Class Size and Test Scores

```
. regress test_score class_size expn_stu el_pct,robust

Linear regression                               Number of obs     =        420
                                                F(3, 416)         =     147.20
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4366
                                                Root MSE          =     14.353

------------------------------------------------------------------------------
             |               Robust
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -.2863992   .4820728    -0.59   0.553    -1.234002     .661203
    expn_stu |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
      el_pct |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
------------------------------------------------------------------------------

. test class_size expn_stu

 ( 1)  class_size = 0
 ( 2)  expn_stu = 0

       F(  2,   416) =    5.43
            Prob > F =    0.0047
```

- F-statistic with two restrictions has an approximate $F_{2,\infty}$ distribution in large samples

$$F_{act} = 5.43 > F_{2,\infty} = 4.61 \ at \ 1\% \ significant \ level$$

- This implies that we reject H_0 at a 1% significance level.

The "overall" regression F-statistic

- The "overall" F-statistic test the joint hypothesis that all the k slope coefficients are zero
  - $H_0 : \beta_j = \beta_{j,0}, ..., \beta_m = \beta_{m,0}$ for a total of $q = k$ restrictions.
  - H_1: at least one of $q = k$ restrictions under H_0 does not hold.

The "overall" regression F-statistic

```
. regress test_score class_size expn_stu el_pct,robust

Linear regression                               Number of obs     =        420
                                                F(3, 416)         =     147.20
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4366
                                                Root MSE          =     14.353

------------------------------------------------------------------------------
             |               Robust
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -.2863992   .4820728    -0.59   0.553    -1.234002     .661203
    expn_stu |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
      el_pct |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
------------------------------------------------------------------------------

. test class_size expn_stu el_pct

 ( 1)  class_size = 0
 ( 2)  expn_stu = 0
 ( 3)  el_pct = 0

       F(  3,   416) =  147.20
            Prob > F =    0.0000
```

- The overall $F - Statistics = 147.2$ which indicates at least one coefficient can not be **ZERO**.

Case: Analysis of the Test Score Data Set

Introduction

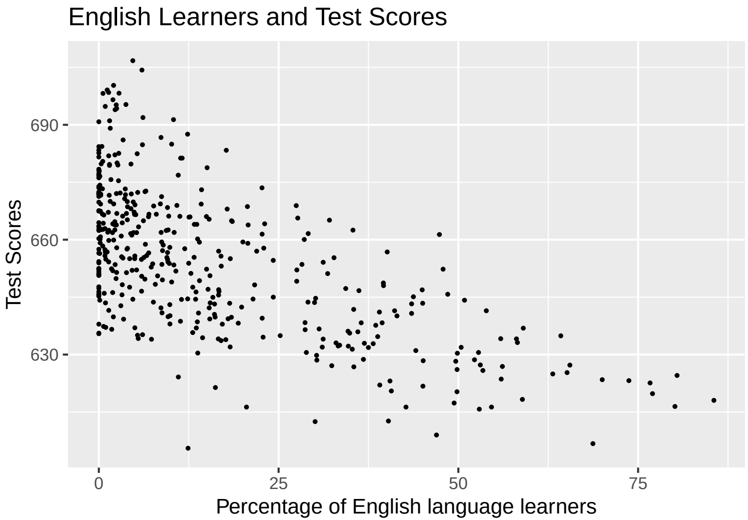
- How to use multiple regression in order to alleviate omitted variable bias and demonstrate how to report results.
- So far we have considered two variables that control for unobservable student characteristics which correlate with the student-teacher ratio *and* are assumed to have an impact on test scores:
  - $English$, the percentage of English learning students
  - $lunch$, the share of students that qualify for a subsidized or even a free lunch at school
  - $calworks$,the percentage of students that qualify for a income assistance program

Five different model equations:

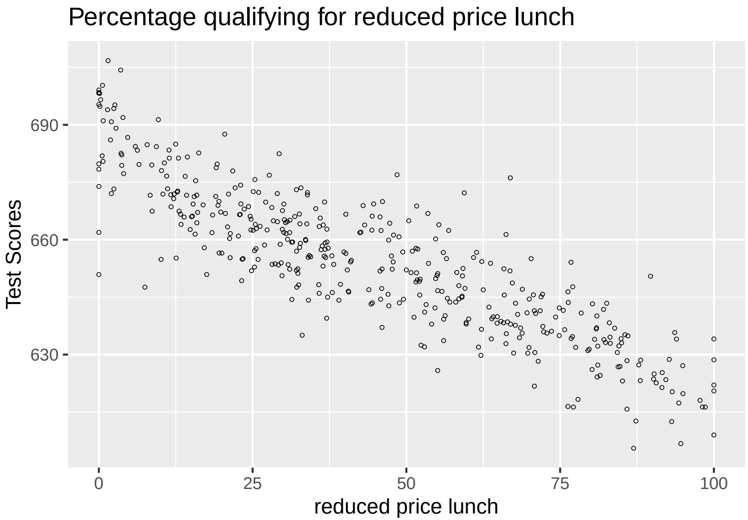
- **We shall consider five different model equations:**

$$
\begin{aligned}
(1) \quad & TestScore = \beta_0 + \beta_1 STR + u, \\
(2) \quad & TestScore = \beta_0 + \beta_1 STR + \beta_2 english + u, \\
(3) \quad & TestScore = \beta_0 + \beta_1 STR + \beta_2 english + \beta_3 lunch + u, \\
(4) \quad & TestScore = \beta_0 + \beta_1 STR + \beta_2 english + \beta_4 calworks + u, \\
(5) \quad & TestScore = \beta_0 + \beta_1 STR + \beta_2 english + \beta_3 lunch + \beta_4 calworks + u
\end{aligned}
$$

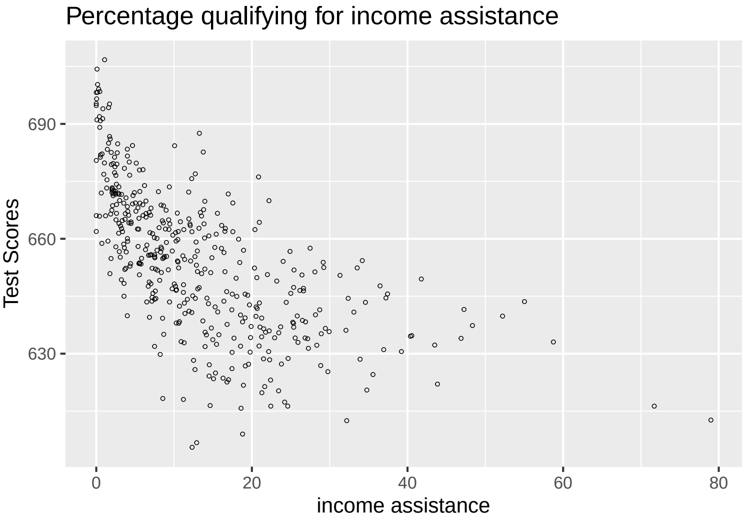
Scatter Plot: English learners and Test Scores



Scatter Plot: Free lunch and Test Scores



Scatter Plot: Income assistant and Test Scores



Correlations between Variables

- **The correlation coefficients are as followed:**

```
# estimate correlation between student characteristics and tes
cor(CASchools$testscr, CASchools$el_pct)
```

```
#> [1] -0.6441237
```

```
cor(CASchools$testscr, CASchools$meal_pct)
```

```
#> [1] -0.868772
```

```
cor(CASchools$testscr, CASchools$calw_pct)
```

```
#> [1] -0.6268534
```

```
cor(CASchools$meal_pct, CASchools$calw_pct)
```

Table 8

	Dependent Variable: Test Score	
	(1)	(2)
str	-2.280^{***}	-1.101^{**}
	(0.519)	(0.433)
el_pct		-0.650^{***}
		(0.031)
Constant	698.933^{***}	686.032^{***}
	(10.364)	(8.728)
Observations	420	420
R^2	0.051	0.426
Adjusted R^2	0.049	0.424
F Statistic	22.575^{***}	155.014^{***}
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$
		Robust S.E. are shown in the parentheses

Table 9

	Dependent Variable: Test Score			
	(1)	(2)	(3)	(4)
str	-2.280^{***}	-1.101^{**}	-0.998^{***}	-1.308^{***}
	(0.519)	(0.433)	(0.270)	(0.339)
el_pct		-0.650^{***}	-0.122^{***}	-0.488^{***}
		(0.031)	(0.033)	(0.030)
meal_pct			-0.547^{***}	
			(0.024)	
calw_pct				-0.790^{***}
				(0.068)
Constant	698.933^{***}	686.032^{***}	700.150^{***}	697.999^{***}
	(10.364)	(8.728)	(5.568)	(6.920)
Observations	420	420	420	420
R^2	0.051	0.426	0.775	0.629
Adjusted R^2	0.049	0.424	0.773	0.626

Table 10

	Dependent Variable: Test Score				
	(1)	(2)	(3)	(4)	(5)
str	-2.280^{***}	-1.101^{**}	-0.998^{***}	-1.308^{***}	-1.014^{***}
	(0.519)	(0.433)	(0.270)	(0.339)	(0.269)
el_pct		-0.650^{***}	-0.122^{***}	-0.488^{***}	-0.130^{***}
		(0.031)	(0.033)	(0.030)	(0.036)
meal_pct			-0.547^{***}		-0.529^{***}
			(0.024)		(0.038)
calw_pct				-0.790^{***}	-0.048
				(0.068)	(0.059)
Constant	698.933^{***}	686.032^{***}	700.150^{***}	697.999^{***}	700.392^{***}
	(10.364)	(8.728)	(5.568)	(6.920)	(5.537)
Observations	420	420	420	420	420
R^2	0.051	0.426	0.775	0.629	0.775
Adjusted R^2	0.049	0.424	0.773	0.626	0.773

The “Star War” and Regression Table

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31* (0.34)	−1.01* (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547* (0.024)		−0.529* (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\overline{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix (4.1). Heteroskedasticity-robust standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

Warp Up

- OLS regression is the most fundamental and important tool in econometricians toolbox.
- The OLS estimators is **unbiased**, **consistent** and approximated **normal distributions** if four key assumptions are satisfied.
- Using the hypothesis testing and confidence interval in OLS regression, we could make a more reliable judgment about the relationship between the treatment and the outcomes.
- Under several reasonable but strong assumptions(CIA), OLS regression can be seen as a continuous version of generalizing continuous version of RCT.
- The OLS regression can be used to estimate the causal effect of the treatment on the outcomes, and the results can be interpreted as the average treatment effect on the treated.