Lec2: Regression Review

Applied MicroEconometrics,Fall 2021

Zhaopeng Qu

Nanjing University Business School

October 08 2021

Review the previous lecture

Causal Inference and RCT

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

Furious Seven Weapons（七种武器）

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

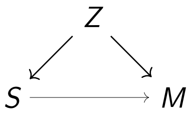
Make Comparison Make Sense

Case: Smoke and Mortality

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

Case: Smoke and Mortality

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



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

Case: Smoke and Mortality(Cochran 1968)

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

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

Case: Smoke and Mortality(Cochran 1968)

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

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

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

Case: Smoke and Mortality(Cochran 1968)

Table 2: Non-smokers and smokers differ in age

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

Case: Smoke and Mortality(Cochran 1968)

Table 2: Non-smokers and smokers differ in age

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

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

Case: Smoke and Mortality(Cochran 1968)

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

Case: Smoke and Mortality(Cochran 1968)

How to balance?

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

Case: Smoke and Mortality(Cochran 1968)

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

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

Case: Smoke and Mortality(Cochran 1968)

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

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

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

Case: Smoke and Mortality(Cochran 1968)

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

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

Case: Smoke and Mortality(Cochran 1968)

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

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

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

Case: Smoke and Mortality(Cochran 1968)

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

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

Case: Smoke and Mortality(Cochran 1968)

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

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

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

Formalization: Covariates

Definition: Covariates

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

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

Identification under independence

- Recall that randomization in RCTs implies

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

and therefore:

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

Identification under independence

- Recall that randomization in RCTs implies

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

and therefore:

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

Identification under independence

- Recall that randomization in RCTs implies

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

and therefore:

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

Identification under independence

- Recall that randomization in RCTs implies

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

and therefore:

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

Identification under Conditional Independence

Conditional Independence Assumption(CIA)

which means that if we can "balance" covariates X then we can take the treatment D as randomized, thus

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

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

Identification under Conditional Independence

Conditional Independence Assumption(CIA)

which means that if we can "balance" covariates X then we can take the treatment D as randomized, thus

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

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

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

Identification under conditional independence(CIA)

- But using the CIA assumption, then

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

Identification under conditional independence(CIA)

- But using the CIA assumption, then

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

Identification under conditional independence(CIA)

- But using the CIA assumption, then

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

Identification under conditional independence(CIA)

- But using the CIA assumption, then

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

Curse of Multiple Dimensionality

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

Make Comparison Make Sense

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

Simple OLS Regression

Question: Class Size and Student's Performance

- **Specific Question**:
  What is the effect on district **test scores** if we would increase district average **class size** by 1 student (or one unit of Student-Teacher's Ratio)
- If we could know the full relationship between two variables which can be summarized by a real value function, $f()$

$$Testscore = f(ClassSize)$$

- Unfortunately, the function form is always unknown.

Question: Class Size and Student's Performance

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

Question: Class Size and Student's Performance

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

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

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

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

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

Class Size and Student's Performance

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

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

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

Terminology for Simple Regression Model

- The linear regression model with one regressor is denoted by

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

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

Population Regression: relationship in average

- The linear regression model with one regressor is denoted by

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

- Both side to conditional on X, then

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

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

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

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

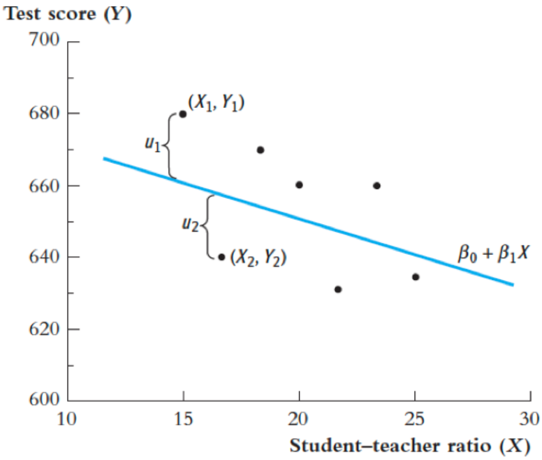
Terminology for Simple Regression Model

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

Graphics for Simple Regression Model

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

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

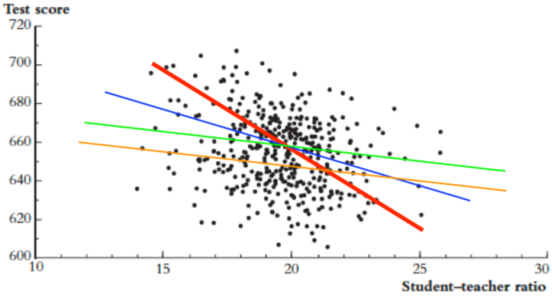


How to find the “best” fitting line?

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

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

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



- 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 $\hat{Y}_i$

The Ordinary Least Squares Estimator (OLS)

The Simple OLS estimator

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

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

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

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

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

The Ordinary Least Squares Estimator (OLS)

The Simple OLS estimator

OLS estimator of β_1 and β_0:

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

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

Assumption of the Linear regression model

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

Assumption of the Linear regression model

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

Linear Regression Model

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

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

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

Assumption 1: Conditional Mean is Zero

Assumption 1: Zero conditional mean of the errors given X

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

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

Assumption 1: Conditional Mean is Zero

Assumption 1: Zero conditional mean of the errors given X

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

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

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

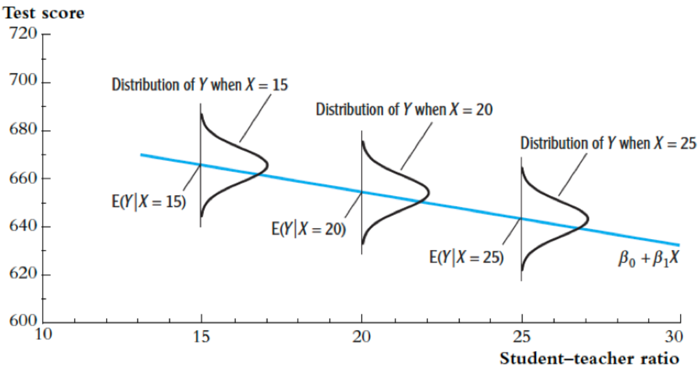
$$Cov[u_i, X_i] = E[u_i X_i] = 0$$

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

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

Assumption 1: Conditional Mean is Zero

FIGURE 4.4 The Conditional Probability Distributions and the Population Regression Line



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

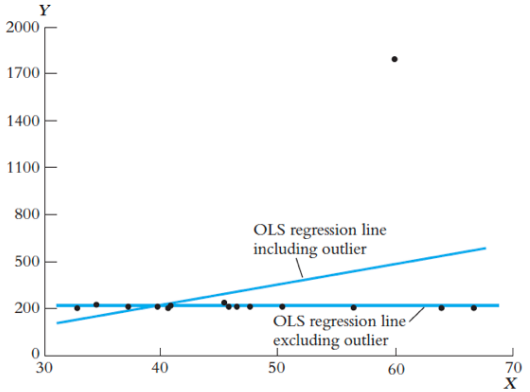
It states that observations with values of X_i, Y_i or both that are far outside the usual range of the data(Outlier) are unlikely. Mathematically, it assume that X and Y have nonzero finite fourth moments.

- 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.



Least Squares Assumptions

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

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

Properties of the OLS estimator: Consistency

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

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

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.

Properties of the OLS estimator: Continuous Mapping Theorem

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

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

- Example:

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

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

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., then we have

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

- Based on the Central Limit theorem(C.L.T), the sample distribution in a large sample can *approximates to a normal distribution*, thus

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

- 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

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

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

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

- Where it can be shown that

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

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

- We have shown that

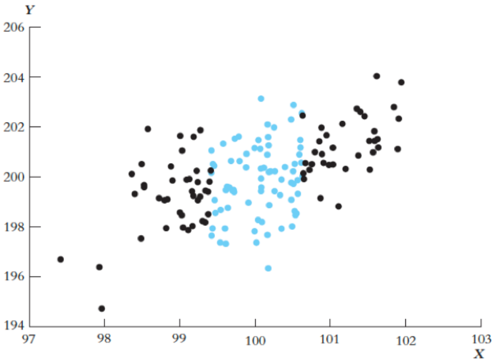
$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{Var[(X_i - \mu_x)u_i]}{[Var(X_i)]^2})$$

- 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*

Multiple OLS Regression

Violation of the first 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**

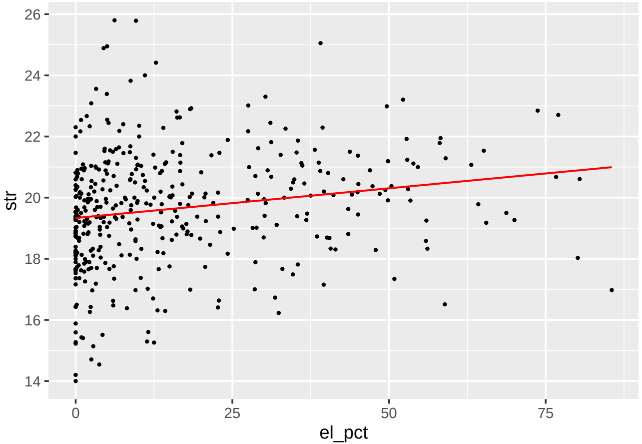
$$E(u_i|X_i) = 0$$

  - 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

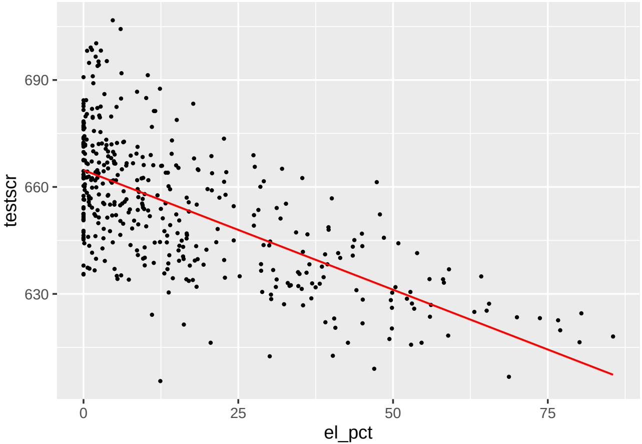
- Many other factors can affect student's performance in the school.
- One of factors is **the share of immigrants** in the class(school, district). Because immigrant children may have different backgrounds from native children, such as
    - parents' education level
    - family income and wealth
 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

Scatter Plot: English learners and STR



- higher share of English learner, bigger class size

Scatter Plot: English learners and testscr



- higher share of English learner, lower testscore

English learner 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.
- It implies that percentage of English learners is contained in u_i, in turn that Assumption 1 is violated.
- It means that the estimates of $\hat{\beta}_1$ and $\hat{\beta}_0$ are **biased** and **inconsistent**.

English Learners as an Omitted Variable

- As before, X_i and Y_i represent STR and Test Score.
- 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.
- Then we have two regression:
  - **True model**(Long regression):

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

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

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

    where $v_i = \gamma W_i + u_i$

Omitted Variable Bias: Biasedness

- Let us see what is the consequence of OVB

$$E[\hat{\beta}_1] = E\left[\frac{\sum(X_i - \bar{X})(\beta_0 + \beta_1 X_i + v_i - (\beta_0 + \beta_1 \overline{X} + \bar{v}))}{\sum(X_i - \bar{X})(X_i - \bar{X})}\right]$$

$$= E\left[\frac{\sum(X_i - \bar{X})(\beta_0 + \beta_1 X_i + \gamma W_i + u_i - (\beta_0 + \beta_1 \overline{X} + \gamma \overline{W} + \overline{u}))}{\sum(X_i - \bar{X})(X_i - \bar{X})}\right]$$

- Skip Several steps in algebra which is very **similar** to procedures for proving unbiasedness of β
- At last, we get (**Please prove it by yourself**)

$$E[\hat{\beta}_1] = \beta_1 + \gamma E\left[\frac{\sum(X_i - \bar{X})(W_i - \bar{W})}{\sum(X_i - \bar{X})(X_i - \bar{X})}\right]$$

Omitted Variable Bias(OVB): inconsistency

- Recall: consistency when n is large, thus
- OLS with on OVB

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

Omitted Variable Bias(OVB): inconsistency

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

Omitted Variable Bias(OVB): inconsistency

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

Omitted Variable Bias(OVB): inconsistency

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

Omitted Variable Bias(OVB): inconsistency

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

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 $\gamma = 0$
- if both two conditions above hold *simultaneously*, then $\hat{\beta}_1$ is **inconsistent**.

Omitted Variable Bias(OVB):Directions

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

Omitted Variable Bias(OVB):Directions

- If OVB can be possible in our regression,then we should guess the **directions** of the bias, in case that we can't eliminate it.
- Summary of the bias when w_i is omitted in estimating equation

	$Cov(X_i, W_i) > 0$	$Cov(X_i, W_i) < 0$
$\gamma > 0$	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. Parking lot space per pupil
  3. Teachers' Salary
  4. Family income
  5. Percentage of English learners

Omitted Variable Bias: Examples

- 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 '
#>
```

Omitted Variable Bias: Examples

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

```
#>
#> Call:
#> lm(formula = testscr ~ str + el_pct, data = ca)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -48.845 -10.240  -0.308   9.815  43.461
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
#> str          -1.10130    0.38028  -2.896  0.00398 **
#> el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
#> ---
```

Omitted Variable Bias: Examples

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 bias is the most possible bias when we run OLS regression using nonexperimental data.
- The simplest way to overcome OVB: **control them**, which means putting them into the regression model.

Multiple regression model with k regressors

- The multiple regression model is

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

where

- Y_i is the *dependent variable*
- $X_1, X_2, ...X_k$ are the *independent variables(includes 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.

Interpretation of coefficients

- 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 "other things equal", thus the concept of **ceteris paribus**

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.
- 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.
- 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

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

OLS Estimation in Multiple Regressors

- As in simple OLS, the estimator multiple Regression is just a minimize the following question

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

OLS Estimation in Multiple Regressors

- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are obtained by solving the following **system of normal equations**

$$\sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - ... - \hat{\beta}_k X_{k,i} \right) = 0$$

$$\sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - ... - \hat{\beta}_k X_{k,i} \right) X_{1,i} = 0$$

$$\vdots = \vdots$$

$$\sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - ... - \hat{\beta}_k X_{k,i} \right) X_{k,i} = 0$$

OLS Estimation in Multiple Regressors

- Since the fitted residuals are

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

- the normal equations can be written as

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

Introduction: Partitioned Regression

If the four least squares assumptions in the multiple regression model hold:

- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 ... \hat{\beta}_k$ are unbiased.
- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 ... \hat{\beta}_k$ are consistent.
- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 ... \hat{\beta}_k$ are normally distributed in large samples.
- Formal proofs need to use the knowledge of **linear algebra**, thus **the matrix**. We only prove them in a simple case.

Partitioned regression: OLS estimators

- A useful representation of $\hat{\beta}_j$ could be obtained by the **partitioned regression**.
- Suppose we want to obtain an expression for $\hat{\beta}_1$.
- Regress $X_{1,i}$ on other regressors,thus

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

where $\tilde{X}_{1,i}$ is the fitted OLS residual(just a variation of u_i)

Partitioned regression: OLS estimators

- Then we could prove that

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

- Identical argument works for $j = 2, 3, ..., k$, thus

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

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*.

Example: Test scores and Student Teacher Ratios(1)

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

```
#> [1] 1.305121e-17
```

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

```
#> [1] 5.412337e-15
```

```
cov(tilde.str,ca$avginc)# should be zero too
```

```
#> [1] 3.650126e-16
```

Example: Test scores and Student Teacher Ratios(2)

```
tilde.str_str <- tilde.str*ca$str # uX
tilde.strstr <- tilde.str^2
sum(tilde.str_str) # sum(uX)=sum(u^2)
```

```
#> [1] 1396.348
```

```
sum(tilde.strstr)# should be equal the result above.
```

```
#> [1] 1396.348
```

Example: Test scores and Student Teacher Ratios(3)

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

```
#> [1] -0.06877552
```

Example: Test scores and Student Teacher Ratios(4)

```
#>
#> Call:
#> lm(formula = testscr ~ tilde.str, data = ca)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -48.50 -14.16   0.39  12.57  52.57
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 654.15655    0.93080 702.790   <2e-16 ***
#> tilde.str    -0.06878    0.51049  -0.135    0.893
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 19.08 on 418 degrees of freedom
#> Multiple R-squared:  4.342e-05,  Adjusted R-squared:  -0.002349
#> F-statistic: 0.01815 on 1 and 418 DF,  p-value: 0.8929
```

Example: Test scores and Student Teacher Ratios(5)

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

```
#>
#> Call:
#> lm(formula = testscr ~ str + el_pct + avginc, data = ca)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -42.800  -6.862   0.275   6.586  31.199
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 640.31550    5.77489 110.879   <2e-16 ***
#> str          -0.06878    0.27691  -0.248    0.804
#> el_pct       -0.48827    0.02928 -16.674   <2e-16 ***
#> avginc        1.49452    0.07483  19.971   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
```

Standard Error of the Regression

- Recall: SER(Standard Error of the Regression)
  - SER is an **estimator** of the standard deviation of the u_i, which are measures of the spread of the Y's around the regression line.
  - Because the regression errors are unobserved, the SER is computed using their sample counterparts, the OLS residuals $\hat{u}_i$

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

where $s_{\hat{u}}^2 = \frac{1}{n-k-1} \sum \hat{u^2}_i = \frac{SSR}{n-k-1}$

- $n-k-1$ because we have $k+1$ stricted conditions in the F.O.C.In another word,in order to construct $\hat{u^2}_i$, we have to estimate $k+1$ parameters,thus $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$

Measures of Fit in Multiple Regression

- Actual = Predicted+residual: $Y_i = \hat{Y}_i + \hat{u}_i$
- The regression R^2 is the fraction of the sample variance of Y_i explained by (or predicted by) the regressors.

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

- 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.*

A Special Case: Categoried Variables as X

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

A Special Case: Categoried Variables as X

- Recall if X is a dummy variable, then we can put it into regression equation straightly.
- What if X is a categoried vriable?
  - **Question**: What is a categoried 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: Categoried Variables as X

- Naive Solution: a simple OLS regression model

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

- 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: Categoried Variables as X

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

A Special Case: Categoried Variables as X

- The first step: turn a categried 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: Categoried Variables as X

- The first step: turn a categried 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: Categoried Variables as X

- The first step: turn a categried 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: Categoried 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 categoried class on *testscore* respectively.

A Special Case: Categoried Variables as X

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

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

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

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 ***
#> ---
```

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
```

Perfect multicollinearity

Table 6: Class Size and Test Score

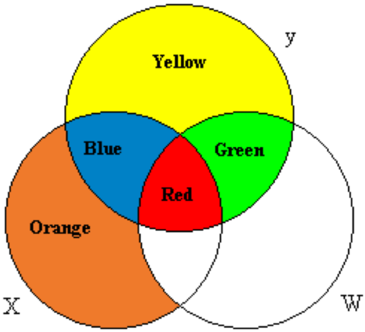
	Dependent variable:	
	testscr	
	(1)	(2)
str	−1.101***	−1.101***
	(0.380)	(0.380)
nel_pct		0.650***
		(0.039)
el_pct	−0.650***	
	(0.039)	
Constant	686.032***	685.382***
	(7.411)	(7.416)
Observations	420	420
R^2	0.426	0.426

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



1) In a simple model (y on X), OLS uses 'Blue' + 'Red' to estimate β. 2) When y is regressed on X and W: OLS throws away the red area and just uses blue to estimate β. 3) 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

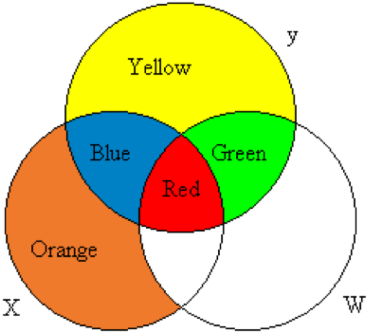


Figure 3a Modest collinearity

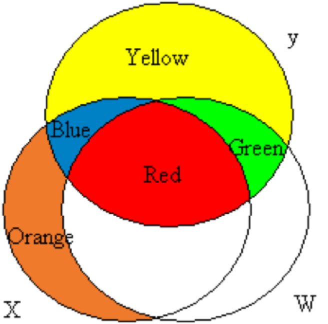


Figure 3b Considerable collinearity

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

Multiple regression model: class size example

Table 7: Class Size and Test Score

	testscr		
	(1)	(2)	(3)
str	−2.280***	−1.101***	−0.069
	(0.480)	(0.380)	(0.277)
el_pct		−0.650***	−0.488***
		(0.039)	(0.029)
avginc			1.495***
			(0.075)
Constant	698.933***	686.032***	640.315***
	(9.467)	(7.411)	(5.775)
N	420	420	420
R^2	0.051	0.426	0.707
Adjusted R^2	0.049	0.424	0.705

Notes: ***Significant at the 1 percent level.

The Distribution of the OLS Estimators

- In addition, in large samples, the sampling distribution of $\hat{\beta}_1$ and $\hat{\beta}_0$ is well approximated by a bivariate normal distribution.
- Under the least squares assumptions,the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_0$, are unbiased and consistent estimators of β_1 and β_0.
- The OLS estimators are averages of the randomly sampled data, and if the sample size is sufficiently large, the sampling distribution of those averages becomes normal. Because the multivariate normal distribution is best handled mathematically using matrix algebra, the expressions for the joint distribution of the OLS estimators are deferred to **Chapter 18**(SW textbook).
- If the least squares assumptions hold, then in large samples the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are jointly normally distributed and each

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j}) \ , j = 0, ..., k$$

Multiple Regression: Assumptions

If the four least squares assumptions in the multiple regression model hold:

- 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.

Then

- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 ... \hat{\beta}_k$ are *unbiased*.
- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 ... \hat{\beta}_k$ are *consistent*.
- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 ... \hat{\beta}_k$ are *normally distributed* in large samples.

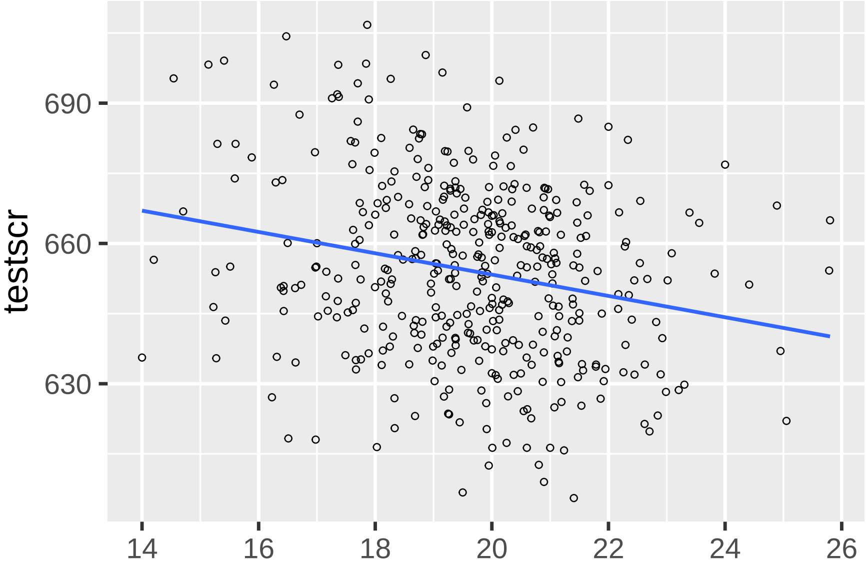
Hypothesis Testing

Introduction: Class size and Test Score

Recall our simple OLS regression mode is

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

Introduction: Class Size and Test Score



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 make the result from the sample infering to **the population**?
- If someone believes that cutting the class size will not help boost test scores. Can you reject the claim based your *scientifical evidence-based* data analysis?
- This is the work of **Hypothesis Testing** in OLS regression.

Review: Hypothesis Testing:

- A hypothesis is (usually) an *assertion* or *statement* about **unknown population parameters**.
- Using the data, we want to determine whether an assertion is **true or false** by a *probability law*.
- Let $\mu_{Y,0}$ is a specific value to which the population mean equals(we suppose)
  - **the null hypothesis**:
  $$H_0 : E(Y) = \mu_{Y,0}$$
  ,
  - **the alternative hypothesis(two-sided)**:
  $$H_1 : E(Y) \neq \mu_{Y,c}$$

Review: Testing a hypothesis 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

$$t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}$$

- Step 4 See if we can **Reject the null hypothesis** at a certain significance levle α,like 5%, or p-value is less than significance level.

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

$$p - value < significance\ level$$

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 woule 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 if the least squares assumptions hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a joint normal sampling distribution.

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

- 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 value of $\sigma_{\hat{\beta}_1}$ is unknown and can not be obtained *directly* by the data.
  - $Var[(X_i - \mu_X)u_i]$ and $[Var(X_i)]^2$ are both unknown.

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

- Because $Var(X) = EX^2 - (EX)^2$, then the *nummerator* in the square root in (4.21) is

$$Var[(X_i - \mu_X)u_i] = E[(X_i - \mu_X)u_i]^2 - (E[(X_i - \mu_X)u_i])^2$$

- Based on the Law of Iterated Expectation(L.I.E), we have

$$E[(X_i - \mu_X)u_i = E(E[(X_i - \mu_X)u_i]|X_i)$$

- Again by the 1st OLS assumption, thus $E(u_i|X_i) = 0$,

$$E[(X_i - \mu_X)u_i] = 0$$

- Then the second term in the equation above

$$Var[(X_i - \mu_X)u_i] = E[(X_i - \mu_X)u_i]^2$$

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

- Because $plime(\overline{X}) = \mu_X$, then we use $\overline{X}$ and $\hat{\mu}_i$ to replace μ_X and μ_i in (4.21)(in large sample), then

$$
\begin{aligned}
Var[(X_i - \mu_X)u_i] =& E[(X_i - \mu_X)u_i]^2 \\
=& E[(X_i - \mu_X)^2 u_i^2] \\
=& plim\Big(\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \hat{u}^2\Big)
\end{aligned}
$$

where $n-2$ is the *freedom of degree*.

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

- Because $plim(s_x) = \sigma_x^2 = Var(X_i)$, then

$$\begin{aligned} Var(X_i) &= \sigma_x^2 \\ &= plim(s_x) \\ &= plim(\frac{n-1}{n}(s_x)) \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{aligned}$$

- Then the *denominator* in the square root in (4.21) is

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

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

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

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

- Everthing in the equation (5.4) are known now or can be obtained by calculation.
- Then we can construct a *t-statistic* and then make a hypothesis test

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

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 safy 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

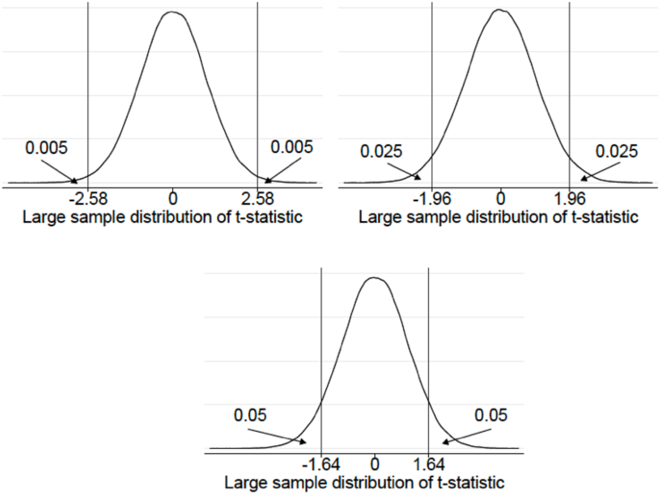
```
. regress test_score class_size, robust

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

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

- 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 $\hat{\beta}_1$

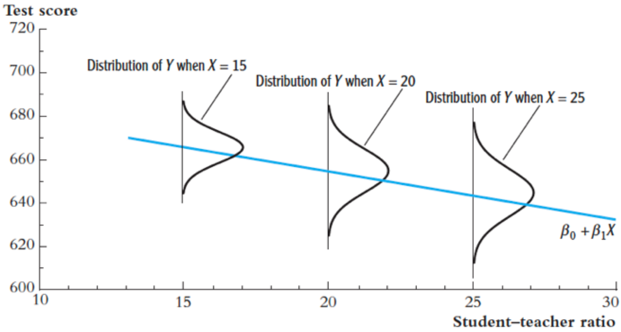
$$\hat{\beta}_1 \pm 1.96 \cdot SE\left(\hat{\beta}_1\right) = -2.28 \pm (1.96 \times 0.519) = [-3.3, -1.26]$$

Heteroskedasticity & homoskedasticity

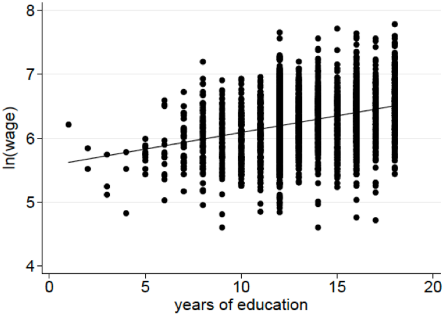
- The error term u_i is **homoskedastic** if the variance of the conditional distribution of u_i given X_i is constant for $i = 1, \dots n$, in particular does not depend on X_i.
- Otherwise, the error term is **heteroskedastic**.

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 heteroskedastic.



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.

- Recall the standard deviation of β_1, $\sigma^2_{\hat{\beta}_1}$, 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 *nummerator* in the square root in (4.21) can be transformed into

$$\begin{aligned} Var[(X_i - \mu_X)u_i] &= E[(X_i - \mu_X)u_i]^2 - (E[(X_i - \mu_X)u_i])^2 \\ &= E[(X_i - \mu_X)u_i]^2 \\ &= E[(X_i - \mu_X)^2 E(u_i^2|X_i)] \\ &= E[(X_i - \mu_X)^2 Var(u_i|X_i)] \end{aligned}$$

Homoskedasticity: S.E.

- So if we assume that the error terms are **homoskedastic**, then the **standard errors** of the OLS estimators β_1 simplify to

$$SE_{Homo}\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \sqrt{\frac{s^2_{\hat{u}}}{\sum(X_i - \bar{X})^2}}$$

- 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 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 stilll *consistent*.
  - The OLS estimators are *normally distributed* in large samples

OLS with Multiple Regressors: Hypotheses tests

- 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 Distrubution: 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*

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

- Can we just test individual coefficients one 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

- If either t_{str} or t_{expn} exceeds 1.96, should we reject the null hypothesis?
- We have to assume that t_{str} and t_{expn} are *uncorrelated* at first:

$$
\begin{aligned}
& Pr(|t_{str}| > 1.96 \; and/or \; |t_{expn}| > 1.96) \\
& = 1 - Pr(|t_{str}| \leq 1.96 \; and \; |t_{expn}| \leq 1.96) \\
& = 1 - Pr(|t_{str}| \leq 1.96) * Pr|t_{expn}| \leq 1.96) \\
& = 1 - 0.95 \times 0.95 \\
& = 0.0975 > 0.05
\end{aligned}
$$

- This "one at a time" method rejects the null too often.

Testing 1 hypothesis on 2 or more coefficients

- If t_{str} and t_{expn} are correlated, then *it is more complicated*. So simple t-statistic is not enough for hypothesis testing in Multiple OLS.
- In general, a joint hypothesis is a hypothesis that imposes two or more restrictions on the regression coefficients.

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

- where $\beta_j, \beta_k, ...$ refer to different regression coefficients.
- There is another approach to testing joint hypotheses that is more powerful, especially when the regressors are highly correlated. That approach is based on the **F-statistic**.

Testing 1 hypothesis on 2 or more coefficients

- If we want to test joint hypotheses that involves multiple coefficients we need to use an **F-test** based on the **F-statistic**
- F-Statistic with $q = 2$: when testing the following hypothesis

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

- Then the *F-statistic* combines the two *t-statisticst*$_1$ 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.

The F-statistic with q restrictions.

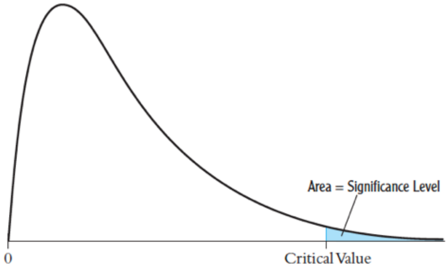
- That is, in large samples, under the null hypothesis,

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

- here q is the number of restrictions
- then we can compute
  - the heteroskedasticity-robust F-statistic
  - the p-value using the F-statistic

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

General procedure for 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 $p - value = Pr[F_{q,\infty} > F^{act}]$

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
```

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 $q = 2$
- It can be shown that the 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

The overall $F - Statistics = 147.2$

```
. 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
```

test_score	Coef.	Robust 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
```

Case: Analysis of the Test Score Data Set

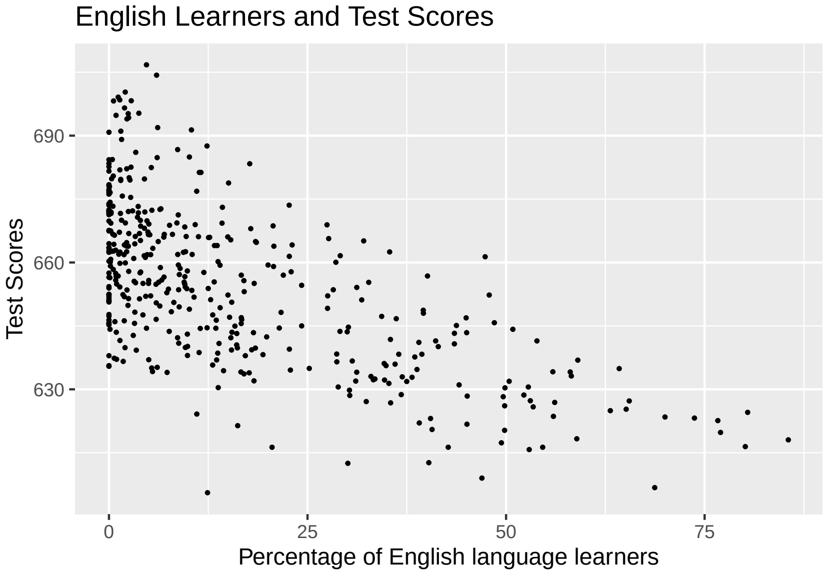
- How to use multiple regression in order to alleviate omitted variable bias and demonstrate how to report results.
- So far we have considered two variables that control for unobservable student characteristics which correlate with the student-teacher ratio *and* are assumed to have an impact on test 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:

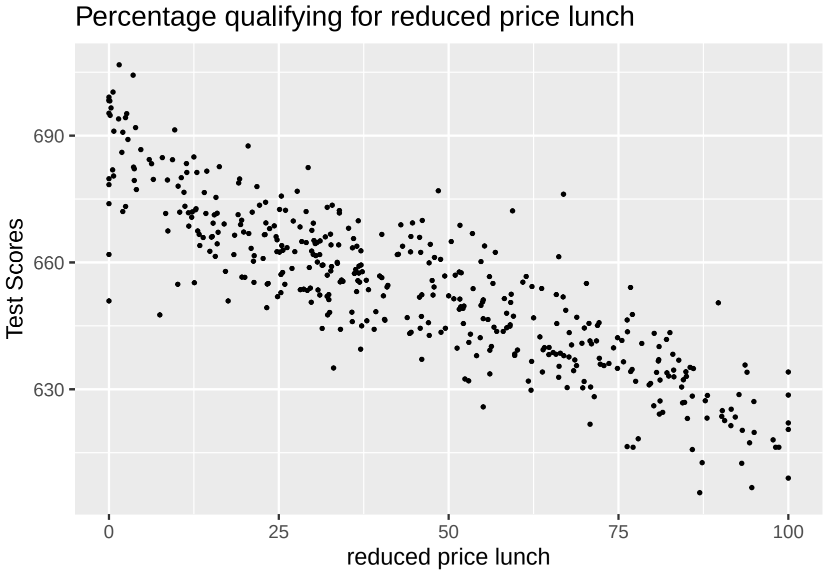
- We shall consider five different model equations:

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

Scatter Plot: English learners and Test Scores



Scatter Plot: Free lunch and Test Scores



Scatter Plot: Income assistant and Test Scores

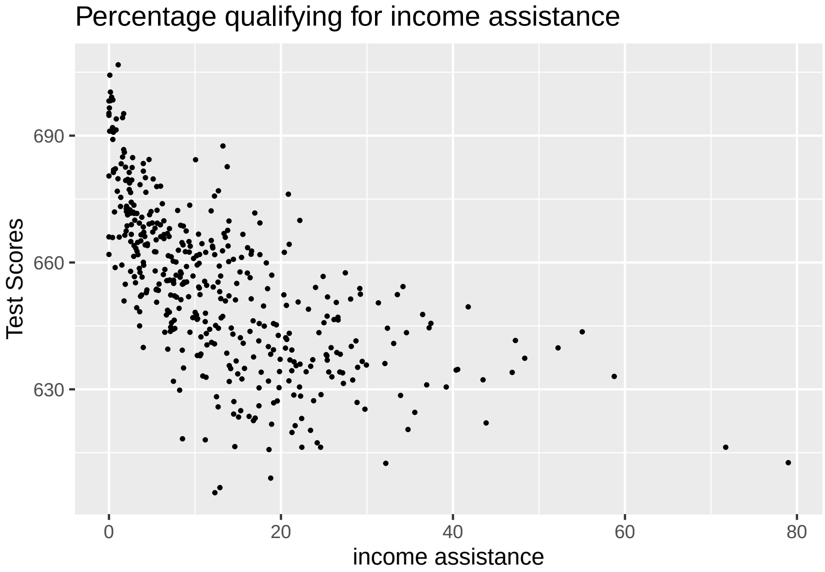


Table 8

	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
Adjusted R^2	0.049	0.424	0.773	0.626	0.773
Residual Std. Error	18.581	14.464	9.080	11.654	9.084
F Statistic	22.575^{***}	155.014^{***}	476.306^{***}	234.638^{***}	357.054^{***}

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)	(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
Adjusted R^2	0.049	0.424	0.773	0.626	0.773
Residual Std. Error	18.581	14.464	9.080	11.654	9.084
F Statistic	22.575***	155.014***	476.306***	234.638***	357.054***

Note: *p<0.1; **p<0.05; ***p<0.01
Robust S.E. are shown in the parentheses

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
Adjusted R^2	0.049	0.424	0.773	0.626	0.773
Residual Std. Error	18.581	14.464	9.080	11.654	9.084
F Statistic	22.575***	155.014***	476.306***	234.638***	357.054***

Note: *p<0.1; **p<0.05; ***p<0.01
Robust S.E. are shown in the parentheses

Table 11

	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
Adjusted R^2	0.049	0.424	0.773	0.626	0.773
Residual Std. Error	18.581	14.464	9.080	11.654	9.084
F Statistic	22.575^{***}	155.014^{***}	476.306^{***}	234.638^{***}	357.054^{***}

Note: *p<0.1; **p<0.05; ***p<0.01
Robust S.E. are shown in the parentheses

Table 12

	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
Adjusted R^2	0.049	0.424	0.773	0.626	0.773
Residual Std. Error	18.581	14.464	9.080	11.654	9.084
F Statistic	22.575***	155.014***	476.306***	234.638***	357.054***
Note:	*p<0.1; **p<0.05; ***p<0.01 Robust S.E. are shown in the parentheses				

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 is the most basic and important tool in econometricians' toolbox.
- The OLS estimators is unbiased, consistent and normal distributions under key assumptions.
- 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.

Regression and Conditional Expectation Function

Case: Education and Earnings

- Most of what we want to do in the social science is learn about how two variables are related, such as *Education and Earnings*.
- On average, people with more schooling earn more than people with less schooling.
  - The connection between schooling and average earnings has considerable predictive power, in spite of the enormous variation in individual circumstances.
  - The fact that more educated people earn more than less educated people does not mean that schooling causes earnings to increase.
  - However, it's clear that education predicts earnings in a narrow statistical sense.
- This predictive power is compellingly summarized by the **Conditional Expectation Function**.

Review: Conditional Expectation Function(CEF)

- Both X and Y are r.v., then conditional on X, Y's probability density function is

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

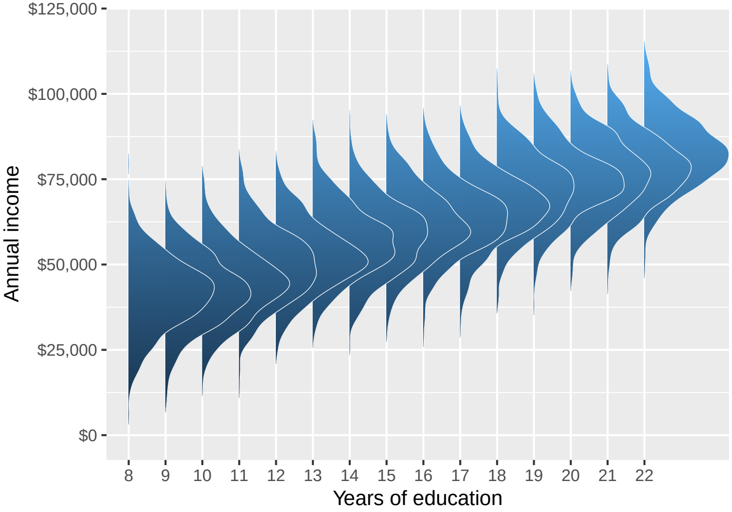
- Conditional on X, Y's expectation is

$$E(Y|X) = \int_Y y f_{Y|X}(y|x)dy = \int_Y y \frac{f(x,y)}{f(x)} dy$$

- So Conditional Expectation Function(CEF) is a function of x, since x is a random variable, so CEF is also a random variable
- 直观理解：期望就是求平均值，而条件期望就是“分组取平均”或“在...条件下的均值”。

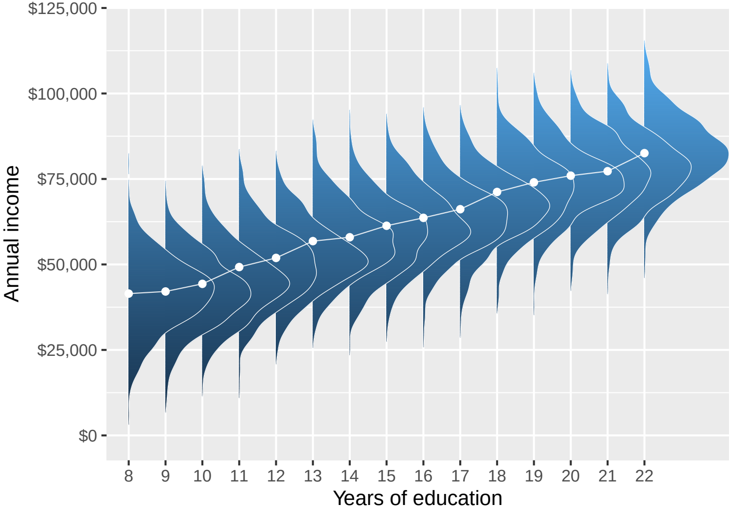
The CEF: Education and Earnings

The conditional distributions of Y_i for $X_i = x$ in 8, ..., 22.



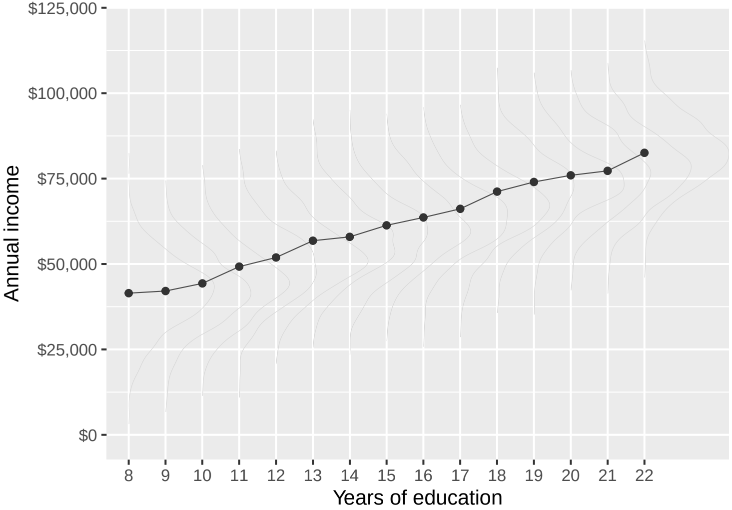
The CEF: Education and Earnings

The CEF, $E[Y_i \mid X_i]$, connects these conditional distributions' means.



The CEF: Education and Earnings

- Focusing in on the CEF, $E[Y_i \mid X_i]$...



Review:Expectation Function(CEF)

1. Additivity : expectation of sums are sums of expectations

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

2. Homogeneity: Suppose that a and b are constants. Then

$$E[(aX+b)|Z]=aE[X|Z]+b$$

3. If X is a r.v,then any function of X, g(X), we have

$$E[g(X)\mid X]=g(X)$$

4. If X and Y are **independent** r.v.s, then

$$E[Y\mid X=x]=E[Y]$$

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

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

where $g(X_i)$ is a continuous function of X_i

- 直观理解：分组平均值 (CEF) 再取平均，应该等于无条件均值。

Review: Expectation

- Expectation(for a continuous r.v.)

$$E(x) = \int xf(x)dx$$

- Conditional probability density function

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

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

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

Proof: the Law of Iterated Expectation(LIE)

- Prove it by a continuous variable way

Proof

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

The CEF Decomposition Property

- **Theorem**: Every random variable such as Y_i can be written as

$$Y_i = E[Y_i \mid X_i] + \varepsilon_i$$

where ε_i is mean-independent of X_i, i.e., $E[\varepsilon_i \mid X_i] = 0$. and therefore ε_i is uncorrelated with any function of X_i.

- This theorem says that any random variable, Y_i, can be decomposed into two parts
  – a piece that's "explained by X_i", i.e. the CEF,
  – a piece left over which is orthogonal to (i.e. uncorrelated with) any function of X_i.

The CEF Decomposition Property

Proof.

$$\varepsilon_i = Y - E[Y_i \mid X_i]$$

The CEF Decomposition Property

Proof.

$$\varepsilon_i = Y - E[Y_i \mid X_i]$$
$$\Rightarrow E[\varepsilon_i \mid X_i] = E\left[Y_i - E[Y_i \mid X_i] \mid X_i\right]$$

The CEF Decomposition Property

Proof.

$$\begin{aligned}\varepsilon_i &= Y - E[Y_i \mid X_i] \\ \Rightarrow E[\varepsilon_i \mid X_i] &= E\left[Y_i - E[Y_i \mid X_i] \mid X_i\right] \\ &= E[Y_i \mid X_i] - E[E[Y_i \mid X_i] \mid X_i] \\ &= 0\end{aligned}$$

□

- We also have

$$E[h(X_i)\varepsilon_i] = E\Big[E[h(X_i)\varepsilon_i] \mid X_i\Big]$$

The CEF Decomposition Property

Proof.

$$\begin{aligned} \varepsilon_i &= Y - E[Y_i \mid X_i] \\ \Rightarrow E[\varepsilon_i \mid X_i] &= E\,[Y_i - E[Y_i \mid X_i] \mid X_i] \\ &= E[Y_i \mid X_i] - E[E[Y_i \mid X_i] \mid X_i] \\ &= 0 \end{aligned}$$

□

- We also have

$$\begin{aligned} E[h(X_i)\varepsilon_i] &= E\Big[E[h(X_i)\varepsilon_i] \mid X_i\Big] \\ &= E\Big[h(X_i)E[\varepsilon_i \mid X_i]\Big] \\ &= 0 \end{aligned}$$

The CEF Prediction Property

Theorem

Let $m(X_i)$ be any function of X_i. The CEF is the Minimum Mean Squared Error (MMSE) predictor of Y_i given X_i. Thus

$$E[Y_i \mid X_i] = \underset{m(X_i)}{argmin} E\left[[Y_i - M(X_i)]^2\right]$$

- $m(X_i)$ can be any class of functions use to predict Y_i

The CEF Prediction Property

Proof.

$$
\begin{aligned}
(Y - m(X_i))^2 &= [(Y_i - E[Y_i \mid X_i]) + (E[Y_i \mid X_i] - m(X_i))]^2 \\
&= (Y_i - E[Y_i \mid X_i)^2 + \\
&\quad 2\,(Y_i - E[Y_i \mid X_i])\,(E[Y_i \mid X_i] - m(X_i)) + \\
&\quad (E[Y_i \mid X_i] - m(X_i))^2
\end{aligned}
$$

□

- Only The last term matters with $m(X_i$, then the function value is minimized at zero when $m(X_i)$ is the CEF.

The CEF Prediction Property

- Suppose we are interested in predicting Y using some function $m(X_i)$, the optimal predictor under the **MMSE (Minimized Mean Squared Error)** criterion is CEF.
- Therefore ,CEF is a natural summary of the relationship between Y and X under MMSE.
- It means that if we can know CEF, then we can describe the relationship of Y and X.

The CEF and Regression

The CEF and Regression

- So far,We have already learned CEF is a natural summary of the relationships which we would like to know it.
- But CEF is an unknown functional form, so the next question is
  - How to model CEF, $E(Y \mid X)$?
- Answer: Two basic approaches
  - Nonparametric(Matching, Kernel Density etc.)
  - Parametric(OLS,NLS,MLE)
- Regression estimates provides a valuable baseline for almost all empirical research because Regression is tightly linked to CEF.

Population Regression: What is a Regression?

- **population regression** as the solution to the population least squares problem. Specifically, the K»1 regression coefficient vector β is defined by solving

$$\beta = \arg\min_b E\left[(Y_i - X_i'b)^2\right]$$

- Using the first order condition

$$E\left[X_i(Y_i - X_i'b)\right] = 0$$

- The solution for b can be written

$$\beta = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i'\right]$$

- Regression is a feature of data: just like expectation, correlation, etc. It's a parametric linear function of population second moments to model $m(X_i)$.

Population Regression: What is a Regression?

- Our "new" result: $\beta = E\left[X_i X_i'\right]^{-1} E[X_i Y_i]$
- In **simple linear regression** (an intercept and one regressor x_i),

$$\beta_1 = \frac{\operatorname{Cov}(Y_i, x_i)}{\operatorname{Var}(x_i)} \quad \beta_0 = E[Y_i] - \beta_1 E[x_i]$$

- For **multivariate regression**, the coefficient on the k^{th} regressor x_{ki} is

$$\beta_k = \frac{\operatorname{Cov}(Y_i, \widetilde{x}_{ki})}{\operatorname{Var}(\widetilde{x}_{ki})}$$

where $\widetilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates.

Linear Regression and the CEF: Why Regress?

- There are three reasons (three justifications) why the vector of population regression coefficient might be of interest.
  1. The Best Linear Predictor Theorem
  2. The Linear CEF Theorem
  3. The Regression-CEF Theorem

Regression Justification I

- **The Best Linear Predictor Theorem**
  - Regression solves the population least squares problem and is therefore the Best Linear Predictor(BLP) of Y_i given X_i.
- **Proof**. By definition of regression.
- In other words, just as CEF, which is the best predictor of Y_i given X_i in the class of all functions of X_i, the population regression function is the best we can do in the class of linear functions.

Regression Justification II

Theorem

The Linear CEF Theorem Suppose the CEF is linear. Then the Regression function is it.

- **Proof**: Suppose $E(Y_i|X_i) = X_i'\beta^*$ for a K»1 vector of coefficients. By the CEF decomposition property, we have

$$E\left[X_i(Y_i - E[Y_i \mid X_i])\right] = 0$$

- Then substitute using $E(Y_i|X_i) = X_i'\beta^*$
- At last find that

$$\beta^* = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right] = \beta$$

Regression Justification II(cont.)

- If the CEF is linear,then the population regression is the CEF.
- Linearity can be a strong assumption. When might we expect linearity?
  1. Situations in which (Y_i, X_i)follows a multivariate normal distribution.
  2. Saturated regression models: the most easy case is a model with two binary indicators and their interaction.

Regression Justification III

Theorem

The Regression-CEF Theorem The population regression function $X_i'\beta$ provides the MMSE linear approximation to $E(Y_i|X_i)$, thus

$$\beta = \arg\min_b E\left[\left(E[Y_i|X_i] - X_i'b\right)^2\right]$$

Regression Justification III

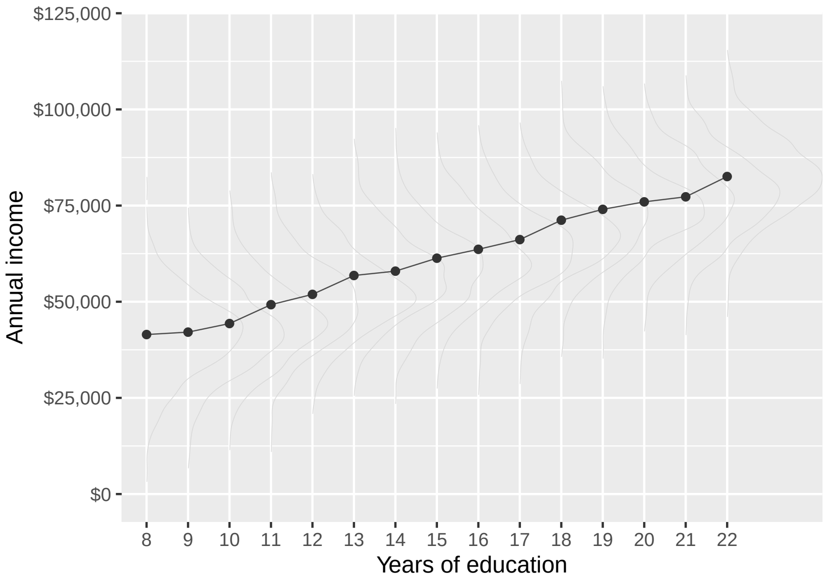
Proof.

$$\begin{aligned}(Y_i - X_i'b)^2 &= [(Y_i - E[Y_i \mid X_i]) + (E[Y_i \mid X_i] - X_ib)]^2 \\ &= (Y_i - E[Y_i \mid X_i])^2 + (E[Y_i \mid X_i] - X_ib)^2 + \\ &\quad 2\,(Y_i - E[Y_i \mid X_i])\,(E[Y_i \mid X_i] - X_ib)\end{aligned}$$

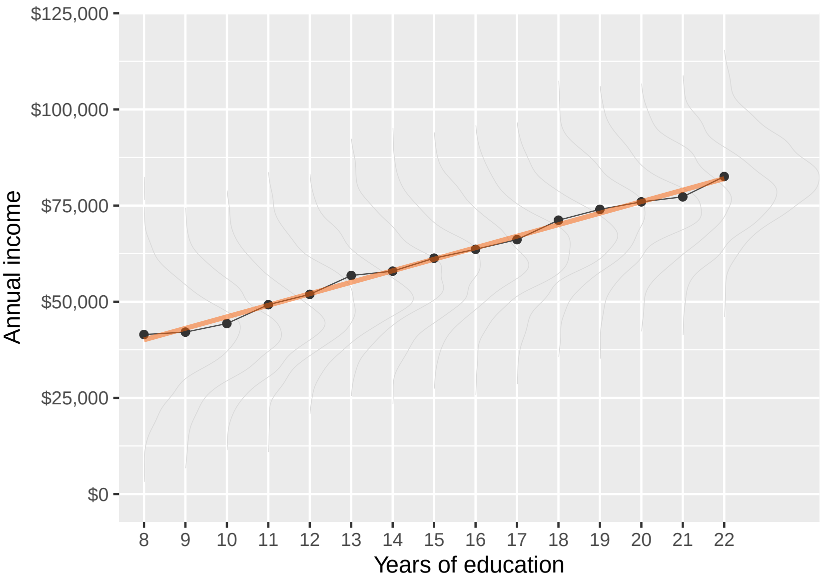
□

- The first term has no b and the last term by the CEF-decomposition property. Therefore the minimized problem has the same solution as the regression least squares problems.

The CEF Function



The CEF and Regression



Warp up: Regression and the CEF

- Model CEF to describe the relationship of Y and X.
  - Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable.
  - When the CEF is linear, the regression function is the CEF.
  - When the CEF is nonlinear, we can still use regression because regression provides the best linear approximation of the CEF.
- Actually, The regression-CEF theorem is our favorite way to motivate regression. The statement that regression approximates the CEF lines up with our view of empirical work as an effort to describe the essential features of statistical relationships, without necessarily trying to pin them down exactly.
- We are not really interested in predicting individual Y_i; it's the distribution of Y_i that we care about.