OLS Regression:Estimation

大数据时代的管理决策,Spring 2019

Zhaopeng Qu

Nanjing University

4/20/2019

1. Review the last lecture
2. OLS Estimation: Simple Regression
3. The Least Squares Assumptions
4. Properties of the OLS estimator
5. Multiple OLS Regression
6. **Multiple OLS Regression: Introduction**
7. **Multiple OLS Regression: Estimation**

Review the last lecture

RCT: Some basic elements

- An Exercise example: Class Size and Student's Performance
- Comparison between treatment and control group: large vs small
  - Calculate the difference of the mean
  - Hypothesis
  - Statistical Test
- Final Conclusion

OLS Estimation: Simple Regression

Question: Class Size and Student's Performance

- A more elegant Question:

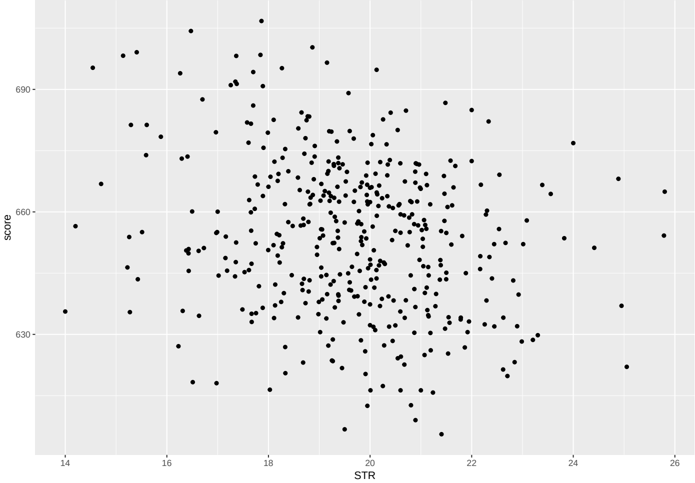
  – *What is the effect on district test scores if we would increase district average class size by 1 student (or one unit of Student-Teacher's Ratio)*

- Technically,we would like to know the real value of a parameter β_1,

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

Question: Class Size and Student's Performance

- Wildly guess the relationship? **Start from a plotting figure!**



A Model of Class Size and Student's Performance

- **Almost Nothing** unless we have a model.
- Generally, we can say that the relationship is describe by a function,

$$Test\,score = f(Class\,size)$$

  - where $f()$ is a real-value function.
- Therefore, at first our question can be transform to
  - What is the specific form of the $f()$?

A Model of Class Size and Student's Performance

- Recall our 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)*
- Technically,we would like to know the real value of a parameter β_1,

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

- Starting from *the simplest way* between testscore and class size: a **linear relationship**,thus

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

- β_1 is the definition of **the slope** and β_0 is **the intercept** of the *straight line* relating test scores and class size.

A Model of Class Size and Student's Performance

- BUT the average test score in district i does not only depend on the average class size.(学生的成绩不仅仅取决于班级规模)
- It also depends on **other factors** such as
  - Student's family background （家庭背景）
  - Quality of the teachers （教师水平）
  - School's facilitates （学校设施）
  - Quality of text books （教材质量）
  - Student's innate ability(学生的先天能力)
- 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 district characteristics** 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**
    - This is the relationship that holds between Y and X on average over the population.

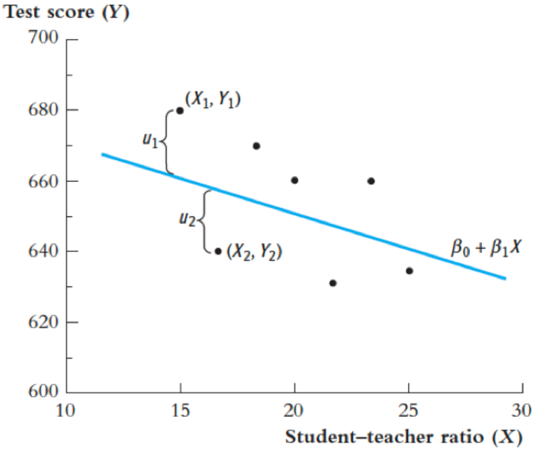
Terminology for Simple Regression Model

- 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.(总体方程或总体回归线)
- 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.

A Graph of u_i

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.

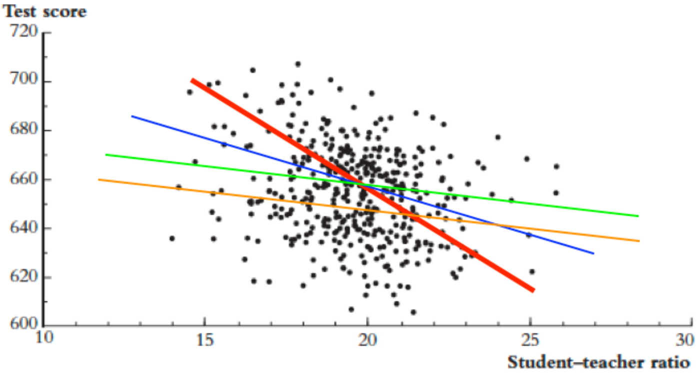


How to find the “best” fitting line?

- In general we don’t know β_0 and β_1 which are parameters of population regression function.

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.



How to find the "best" fitting line? the Population v.s the Sample

- So how to find the line that fits the data **best**?
- If we have **the population** data, we just need calculate the values of parameters.
- Normally, we have no population but a **random sample**. Then we need *estimate* the values of parameters and make an *inference* for the population.

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 $\hat{Y}_i$

Recall the Graph of u_i

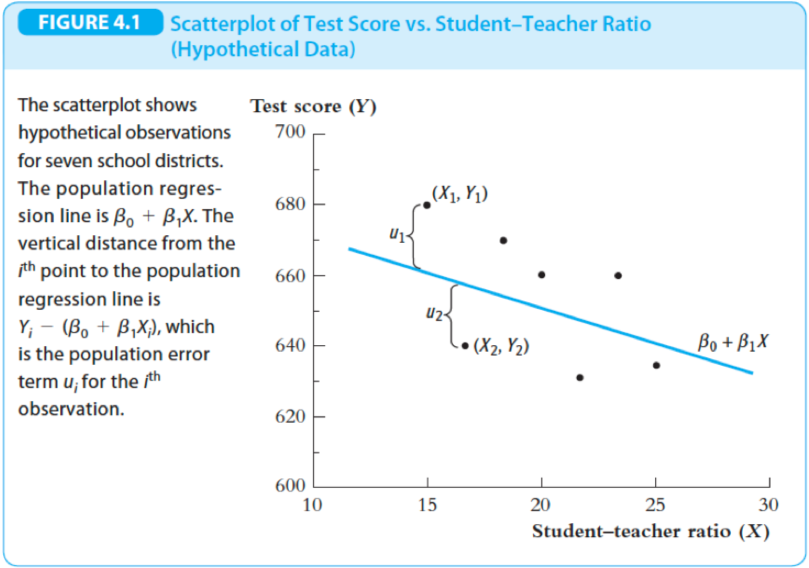


FIGURE 4.1 Scatterplot of Test Score vs. Student–Teacher Ratio (Hypothetical Data)

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.

The Ordinary Least Squares Estimator (OLS)

The OLS estimator

- The prediction mistake is *the difference* between Y_i and $\hat{Y}_i$

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

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

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

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

The Ordinary Least Squares Estimator (OLS)

- OLS minimizes sum of squared prediction mistakes:

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

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

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

    - Step 2 for β_1:

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

OLS estimators of β_0 and β_1

OLS estimator of β_0:

$$b_0 = \overline{Y} - b_1\overline{X}\; or\; \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$$

OLS estimator of β_1:

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

Some Algebraic of $\hat{u}_i$

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

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

$$\hat{u}_i = Y_i - \hat{Y}_i$$

- Then we have

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

- And

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

The Estimated Regression Line

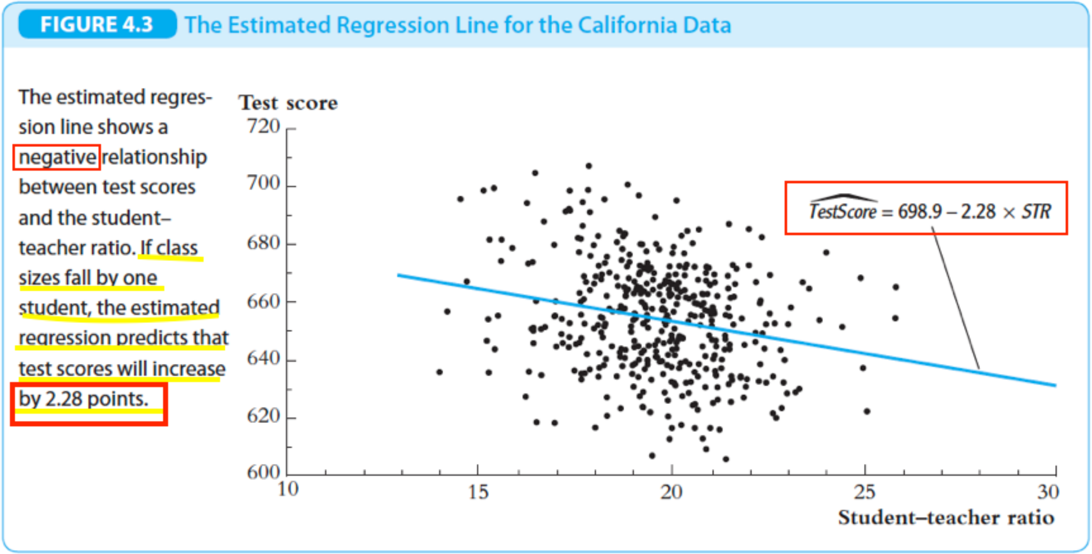


FIGURE 4.3 The Estimated Regression Line for the California Data

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

$\widehat{TestScore} = 698.9 - 2.28 \times STR$

Measures of Fit: The R^2

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

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

- The **explained sum of squares** (ESS):

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

- The **sum of squared residuals** (SSR):

$$\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 = \sum_{i=1}^{n}\hat{u}_i^2$$

-

$$TSS = ESS + SSR$$

Measures of Fit: The R^2(拟合优度)

- R^2 or the coefficient of determination, is the fraction of the sample variance of Y_i explained/predicted by X_i
$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$
- So $0 \leq R^2 \leq 1$
- It seems that **R-squares** is bigger, the regression is better.
- But actually we don't care much about R^2 in modern econometrics.

The Standard Error of the Regression

- The standard error of the regression (SER) is an estimator of the standard deviation of the regression error u_i
- Because the regression errors u_i 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 $s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}^2$

The Least Squares Assumptions

Assumption of the Linear regression model

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

Linear Regression Model

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

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

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

Assumption 1: Conditional Mean is Zero

Assumption 1: Zero conditional mean of the errors given X

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

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

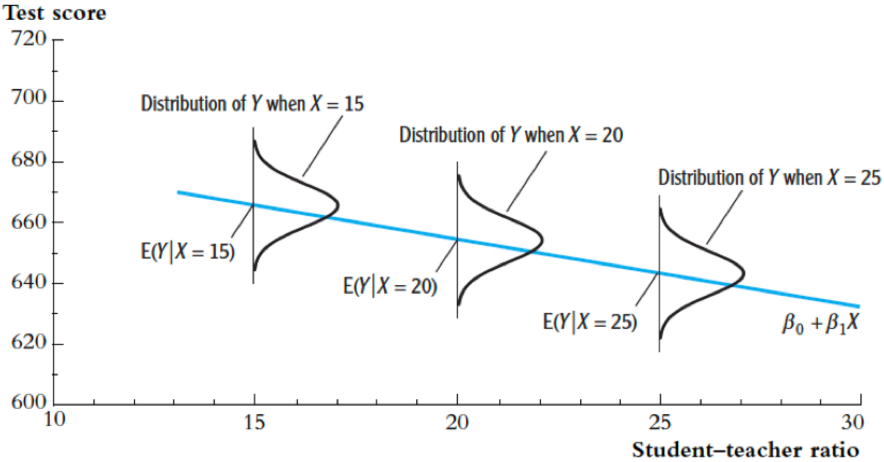
- An *weaker* condition that u_i and X_i are uncorrelated:

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

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

Assumption 1: Conditional Mean is Zero

FIGURE 4.4 The Conditional Probability Distributions and the Population Regression Line



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.
- And it generally won't hold in other data structures.
  - – Violations: time-series, cluster samples.

Assumption 3: Large outliers are unlikely

Assumption 3: Large outliers are unlikely

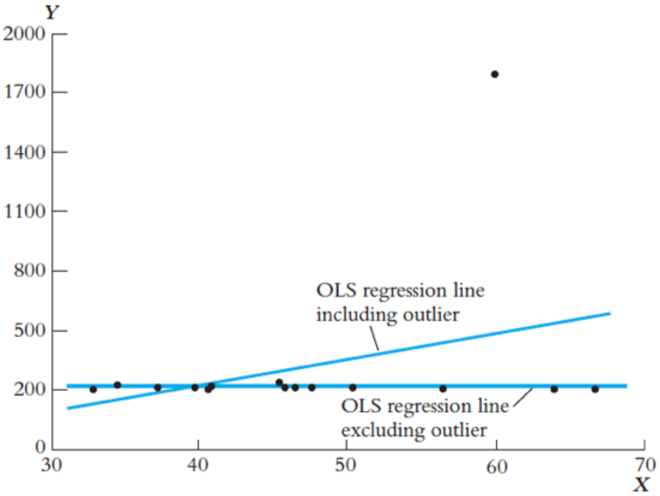
It states that observations with values of X_i, Y_i or both that are far outside the usual range of the data(Outlier)-are unlikely. Mathematically, it assume that X and Y have nonzero finite fourth moments.

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

Assumption 3: Large outliers are unlikely

FIGURE 4.5 The Sensitivity of OLS to Large Outliers

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



Underlying assumptions of OLS

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

Underlying assumptions of OLS

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

Properties of the OLS estimator

The OLS estimators

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

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

- We derived the OLS estimators of β_0 andβ_1:

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

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

Least Squares Assumptions

1. Assumption 1:
2. Assumption 2:
3. Assumption 3:

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

Properties of the OLS estimator: unbiasedness

- Recall:

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

- take expectation to β_0 :

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

- if β_1 is unbiased, then β_0 is also unbiased.

Properties of the OLS estimator: unbiasedness

- Remind we have

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

- So take expectation to β_1:

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

Properties of the OLS estimator: unbiasedness

- Continued

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

Properties of the OLS estimator: unbiasedness

- Continued

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

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

Properties of the OLS estimator: unbiasedness

- Continued

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

Properties of the OLS estimator: unbiasedness

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

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

$$= \beta_1 + E\left[\frac{\sum(X_i - \overline{X})E(u_i|X_1, ..., X_n)}{\sum(X_i - \overline{X})(X_i - \overline{X})}\right]$$

- **the Law of Iterated Expectation(LIE)**

$$E(E(Y|X)) = E(Y) \ and \ E(E(g(X)Y|X) = E(g(X)Y)$$

Properties of the OLS estimator: unbiasedness

- then then we could obtain

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

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

Properties of the OLS estimator: Consistency

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

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

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

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

Properties of the OLS estimator: Consistency

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

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

- Example:

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

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

Properties of the OLS estimator: Consistency

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

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

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

- then we obtain OLS estimator when $n \longrightarrow \infty$

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

Properties of the OLS estimator: Consistency

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

- then then we could obtain

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

- both $\hat{\beta}_0$ and $\hat{\beta}_1$ are **Consistent** on the condition of **Assumption 1**.

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"")

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

- Recall: Sampling Distribution of $\overline{Y}$
- Because Y1,...,Yn 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 $\hat{\beta}_1$

- Unbiasedness of the OLS estimators implies that

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

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

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

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

Sampling Distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ in large-sample

- Recall: Sampling Distribution of $\bar{Y}$, based on the *Central Limit theorem*(C.L.T), the sample distribution in a large sample can approximates to a normal distribution.

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

- So 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 $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$
- Where it can be shown that

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

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

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

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

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

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

- The numerator: $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u})$
- Because $\bar{X}$ is consistent, thus $X \xrightarrow{p} \mu_x$.
- And we know that $\sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}) = \sum_{i=1}^{n}(X_i - \overline{X})u_i$ and we let $v_i = (X_i - \mu_x)u_i$, then

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

- **Assumption 1**, then $E(v_i) = 0$, and **Assumption 2**, $\sigma_v^2 = Var[(X_i - \mu_x)u_i]$
- Then $\bar{v}$ is the sample mean of v_i, based on C.L.T,

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

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

- the expression in the denominator,

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

- this is the *sample variance* of X (except dividing by n rather than $n-1$, which is inconsequential if n is large)
- As discussed in Section 3.2 [Equation (3.8)], the *sample variance* is a **consistent** estimator of the *population variance*,thus

$$\sigma_x^2 \xrightarrow{p} Var[X_i]$$

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

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

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

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

$$\hat{\beta}_1 - \beta_1 \cong \frac{\bar{v}}{Var[X_i]}$$

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

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

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

- Then

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

- So

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

where

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

Sampling Distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ in large-sample

- We have shown that

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

where

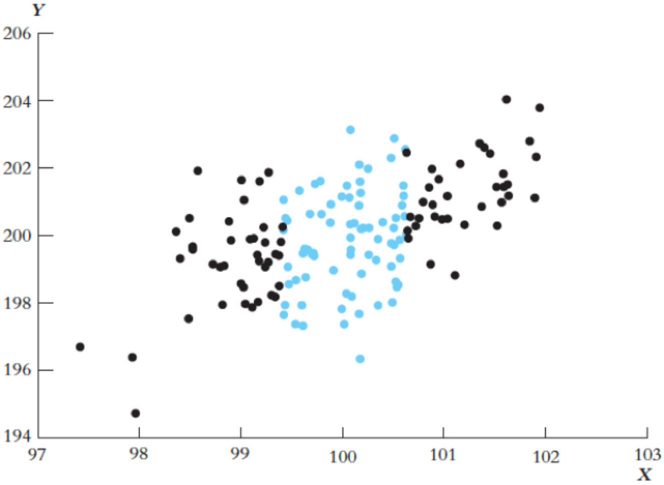
$$H_i = 1 - \left(\frac{\mu_x}{E[X_i^2]}\right)X_i$$

- 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.



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

Simple OLS formula

- The linear regression model with one regressor is denoted by

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

- Where
  - Y_i is the **dependent variable**(Test Score)
  - X_i is the **independent variable** or regressor(Class Size or Student-Teacher Ratio)
  - u_i is the **error term** which contains all the other factors *besides* X that determine the value of the dependent variable, Y, for a specific observation, i.

The OLS Estimator

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

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

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

OLS estimator of β_1:

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

Least Squares Assumptions

Under 3 least squares assumptions,

1. Assumption 1
2. Assumption 2
3. Assumption 3

the OLS estimators will be

- **unbiased**
- **consistent**
- **normal sampling distribution**

Multiple OLS Regression: Introduction

Simple OLS Regression v.s. RCT

- Regression is a way to control observable confounding factors, Which assume the source of selection bias is only from the difference in observed characteristics.
- In a simple regression model, OLS estimators are just a generalizing continuous version of RCT when least squares assumptions are hold.
- But in contrast to RCT, in observational studies, researchers cannot control the assignment of treatment into a treatment group versus a control group.
- To make two groups comparable, we need to keep treatment and control group "**other thing equal**"in observed characteristics and unobserved characteristics.

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**

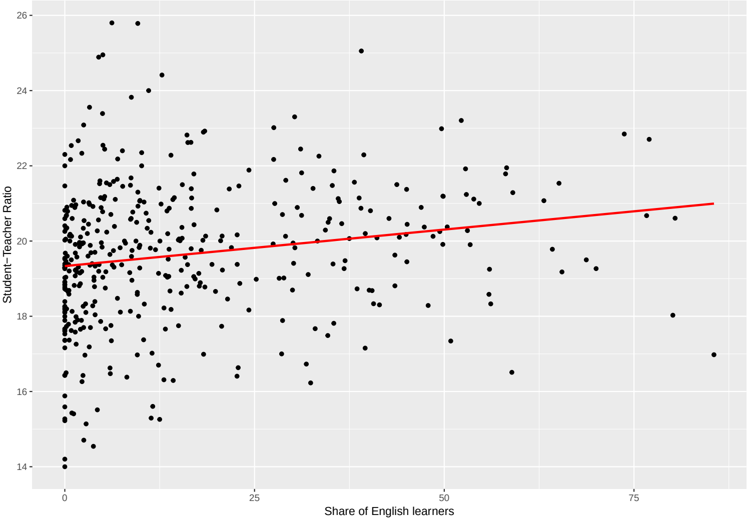
$$E(u_i|X_i) = 0$$

  - 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

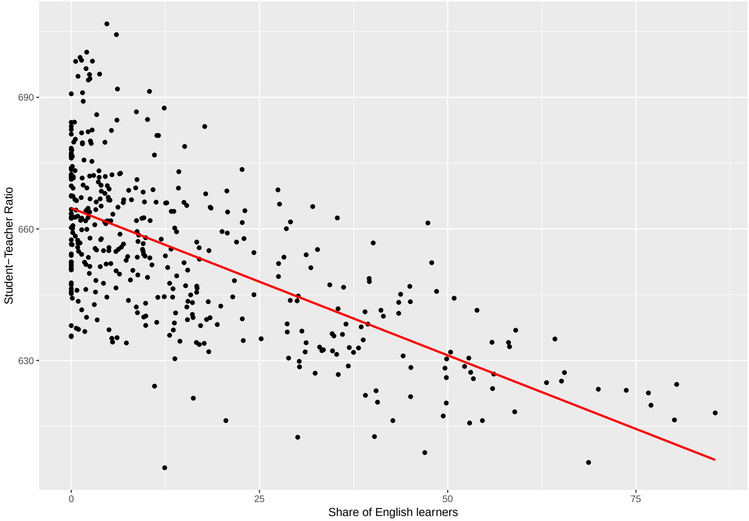
- Many other factors can affect student's performance in the school.
- One of other factors is the share of immigrants(外来移民) in the class(school,district). Because immigrant children may have different backgrouds from native children, such as
    - parents'education level
    - family income and wealth
    - preparenting style
    - traditonal 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 learner as an **Omitted Variable**

- As before, X_i and Y_i represent **STR** and **Test Score**,respectively.
- 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 consequece of OVB
- At last, we can obtain

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

Omitted Variable Bias: Biasedness

- As proving unbiasedness of $\hat{\beta}_1$,we can know
- If W_i is unrelated to X_i,then $E[\hat{\beta}_1] = \beta_1$, because

$$E\left[\frac{\sum(X_i - \bar{X})(W_i - \bar{W})}{\sum(X_i - \bar{X})(X_i - \bar{X})}\right] = 0$$

- If W_i is no determinant of Y_i, which means that

$$\gamma = 0$$

  , then $E[\hat{\beta}_1] = \beta_1$, too,

- Only if **both two conditions** above are violated *simultaneously*, then $\hat{\beta}_1$ is **biased**, which is normally called **Omitted Variable Bias**.

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.
- 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

Omiited 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

Omiited Variable Bias: Examples

- Regress *Testscore* on *Class size*

$$TestScore = \beta_0 + \beta_1 \times STR + u$$

- Regress *Testscore* on *Class size* and *English learner*

$$TestScore = \beta_0 + \beta_1 \times STR + \beta_2 \times PctEL + u$$

Omiited Variable Bias: Examples

```
##
## Call:
## lm(formula = testscr ~ str, data = ca)
##
## Coefficients:
## (Intercept)          str
##      698.93        -2.28
```

```
##
## Call:
## lm(formula = testscr ~ str + el_pct, data = ca)
##
## Coefficients:
## (Intercept)          str       el_pct
##    686.0322      -1.1013      -0.6498
```

Warp Up

- OVB bias is the most possible bias when we run OLS regression using **nonexperiemental** data.
- How to overcome OVB bias?
  - Run a RCT(randomized controlled experiment)
- Or if we can observe it, then the simplest way to overcome OVB(conditionally): **control it**.

Multiple OLS Regression: Estimation

Multiple regression model with k regressors

- The multiple regression model is

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

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

- Suppose the multiple regression model is

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

- Consider changing X_1 by ΔX_1 while holding X_2 constant: then **Population regression line is

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$$

- Make a difference

$$\Delta Y = \beta_1 X_1$$

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

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

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**

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

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**

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

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

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**

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

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**

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

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

$$\sum \hat{u}_i = 0$$

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

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

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

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

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

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

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

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

R Example : Test scores and Student Teacher Ratios

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

$$Testscore = \beta_0 + \beta_1 STR + \beta_2 elpct + \beta_3 avginc + u_i$$

- elpct: the share of english learners as an indictor for immigrants
- avginc: average income of the district as an indictor for family backgrouds

Example: Test scores and Student Teacher Ratios

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

```
##
## Call:
## lm(formula = testscr ~ str + el_pct + avginc, data = ca)
##
## Coefficients:
## (Intercept)          str       el_pct       avginc
##   640.31550     -0.06878     -0.48827      1.49452
```

Multiple regression: Assumpition

Multiple regression: Assumpition

- 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*

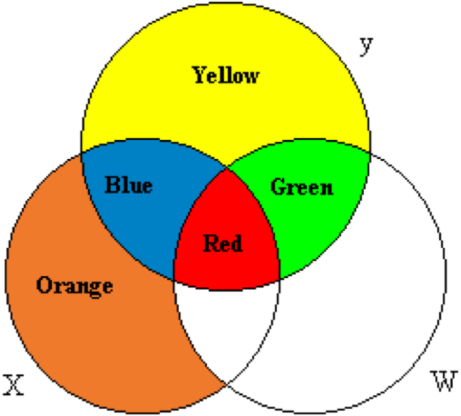
```
##
## Call:
## lm(formula = testscr ~ str + el_pct, data = ca)
##
## Coefficients:
## (Intercept)          str       el_pct
##    686.0322      -1.1013      -0.6498

##
## Call:
## lm(formula = testscr ~ str + nel_pct + el_pct, data = ca)
##
## Coefficients:
## (Intercept)          str      nel_pct       el_pct
##    685.3825      -1.1013       0.6498           NA
```

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

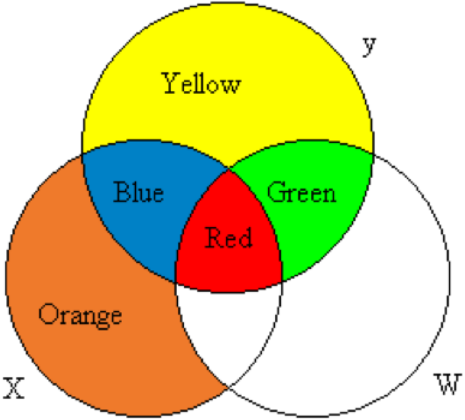


Figure 3a Modest collinearity

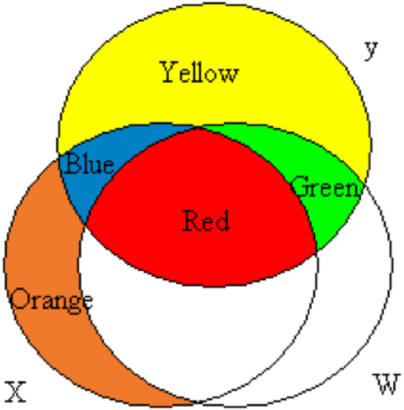


Figure 3b Considerable collinearity

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

Properties of OLS estimator in Multiple Regression

Assumption of OLS estimator in Multiple Regression

- **Assumption #1**: The Conditional Distribution of u_i Given $X_{1i}, X_{2i}, ..., X_{ki}$ has a Mean of Zero.
- **Assumption #2**: $(X_{1i}, X_{2i}, ..., X_{ki}, Y_i), i = 1, 2, ..., n$ are **i.i.d**
- **Assumption #3**: Large Outliers Are Unlikely
- **Assumption #4**: No **Perfect** Multicollinearity
- If the least squares assumptions above hold, then in large samples the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are jointly normally distributed and each

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