Simple OLS Regression:Estimation

Introduction to Econometrics,Fall 2017

Zhaopeng Qu

Nanjing University

9/29/2018

1. Review the last lecture

2. OLS Estimation: Simple Regression

3. The Least Squares Assumptions

4. Properties of the OLS estimator

Review the last lecture

CEF(conditional expectation function): a key concept in Econometrics

- CEF is a natural summary of the relationship between Y and X. If we can know CEF, then we can describe the relationship of Y and X.
- Regression estimates provides a valuable baseline for almost all empirical research because Regression is tightly linked to CEF
    - if CEF is linear, then OLS regression is it.
    - if CEF is nonlinear, then OLS regression provides a best linear approximation to it under MMSE condition.

OLS Estimation: Simple Regression

Question: Class Size and Student's Performance

- Specific Question:

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

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

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

- And β_1 is actually the definition of **the slope** of a straight line relating test scores and class size. Thus

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

where β_0 is the intercept of the straight line.

Question: 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 .....
- 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. (be familiar with? *Recall the concept of CEF*)

Terminology for Simple Regression Model

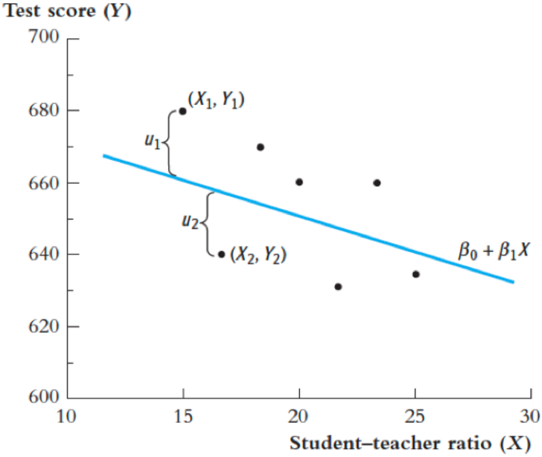
– The intercept β_0 and the slope β_1 are the **coefficients** of the **population regression line**, also known as the **parameters** of the population regression line.

– 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.

Terminology 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.

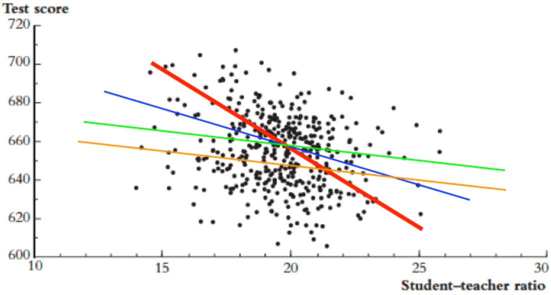


How to find the “best” fitting line?

- 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 OLS Estimator

- **The OLS estimator**

  – Chooses the **best** regression coefficients so that the estimated regression line is *as close as possible* to the observed data, where closeness is measured by *the sum of the squared mistakes* made in predicting Y given X. – Let b_0 and b_1 be estimators of β_0 and β_1,thus $b_0 \equiv \hat{\beta}_0$,$b_1 \equiv \hat{\beta}_1$ – The predicted value of Y_i given X_i using these estimators is $b_0 + b_1 X_i$, or $\hat{\beta}_0 + \hat{\beta}_1 X_i$ formally denotes as $\hat{Y}_i$

The Ordinary Least Squares Estimator (OLS)

The OLS estimator

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

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

– 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:

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

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

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i-b_0-b_1X_i)^2=0$$

  - Step 2 for β_1:

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i-b_0-b_1X_i)^2=0$$

Step 1: OLS estimator of β_0

- Optimization

$$
\begin{aligned}
\frac{\partial}{\partial b_0}\sum_{i=1}^{n}\hat{u}_i^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0 \\
\Rightarrow \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1 X_i = 0 \\
\Rightarrow \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} b_0 - b_1 \frac{1}{n}\sum_{i=1}^{n} X_i = 0 \\
\Rightarrow \overline{Y} - b_0 - b_1\overline{X} = 0
\end{aligned}
$$

Step 1: OLS estimator of β_0

OLS estimator of β_0:

$$\mathbf{b_0 = \overline{Y} - b_1\overline{X}\,or\hat{}\,_0 = \overline{Y} -\hat{}\,_1\overline{X}}$$

Step 2: OLS estimator of β_1

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}\hat{u}_i^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} X_i[Y_i - (\overline{Y} - b_1\overline{X}) - b_1X_i] = 0$$

$$\Rightarrow \sum_{i=1}^{n} X_i[(Y_i - \overline{Y}) - b_1(X_i - \overline{X})] = 0$$

$$\Rightarrow \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - b_1\sum_{i=1}^{n} X_i(X_i - \overline{X}) = 0$$

Step 2: OLS estimator of β_1

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y}) &= \sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}X_i\overline{Y}-\sum_{i=1}^{n}\overline{X}Y_i+\sum_{i=1}^{n}\overline{XY}\\ &= \sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}X_i\overline{Y}-n\overline{X}(\frac{1}{n}\sum_{i=1}^{n}Y_i)+n\overline{XY}\\ &= \sum_{i=1}^{n}X_i(Y_i-\overline{Y})\end{aligned}$$

- By a similar reasoning, we could obtain

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

Step 2: OLS estimator of β_1

- Thus

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}\hat{u}_i^2=\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})-b_1\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})=0$$

OLS estimator of β_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

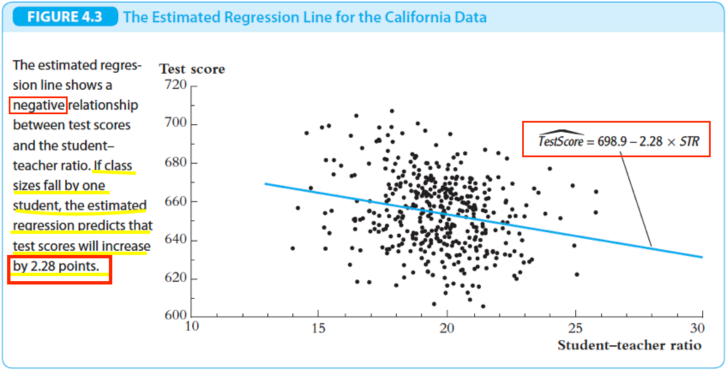


FIGURE 4.3 The Estimated Regression Line for the California Data

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 (SST) = the explained sum of squares (SSE) + the sum of squared residuals (SSR):

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

- 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 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

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

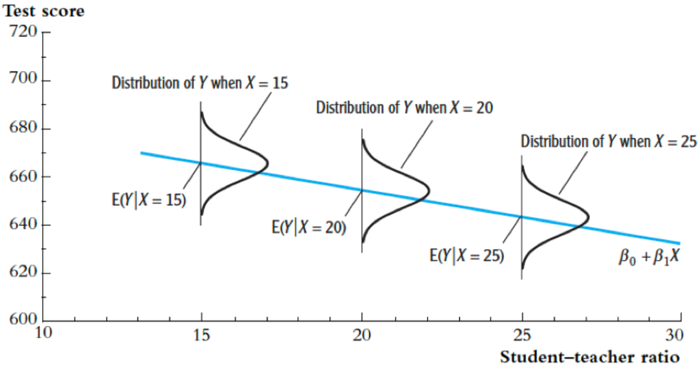
- An *weaker* condition that u_i and X_i are uncorrelated:

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

- 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



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.

- 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

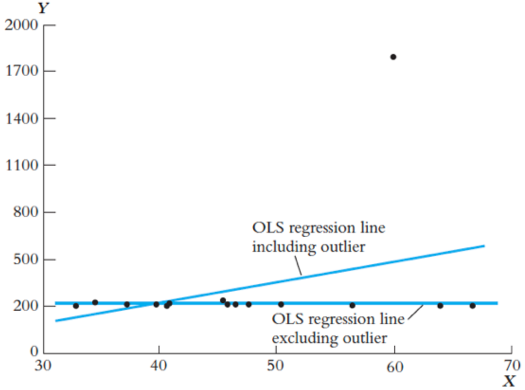
It states that observations with values of X_i, Y_i or both that are far outside the usual range of the data(Outlier)-are unlikely. Mathematically, it assume that X and Y have nonzero finite fourth moments.

- 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

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

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

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

Properties of the OLS estimator: unbiasedness

- Continued
- Because $\sum \overline{u} = 0$ and $\sum \overline{u} X_i = 0$, so

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

Properties of the OLS estimator: unbiasedness

- Continued

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

- 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

- where it can be shown that

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

  where

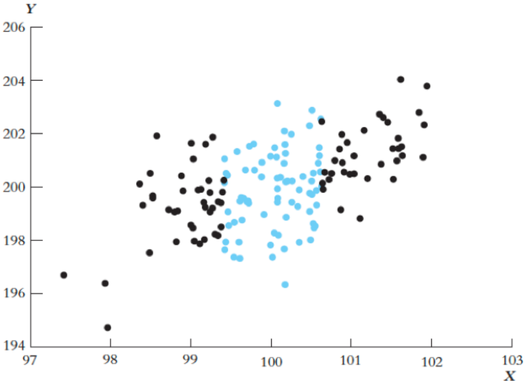
$$H_i = 1 - \left(\frac{\mu_x}{E[X_i^2]}\right)X_i$$

- 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.



- When more **variation** in X, 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*