Linear Regression with One Regressor

Ming-Ching Luoh

2022.2.15.

Introduction

Empirical problem:

Class size and educational output

- Policy question:
  What is the effect of reducing class size by one student per class? by 8 students/class?
- What is the right output (performance) measure?
  - parent satisfaction.
  - student personal development.
  - future adult earnings.
  - performance on standardized tests.

What do data say about class sizes and test scores?

The California Test Score Data Set
All K-6 and K-8 California school districts ($n = 420$)

Variables:

- 5th grade test scores (Stanford-9 achievement test, combined math and reading), district average.
- Student-teacher ratio (STR)
  = number of students in the district divided by number of full-time equivalent teachers.

An initial look at the California test score data

TABLE 4.1 Summary of the Distribution of Student-Teacher Ratios and Fifth-Grade Test Scores for 420 K-8 Districts in California in 1999

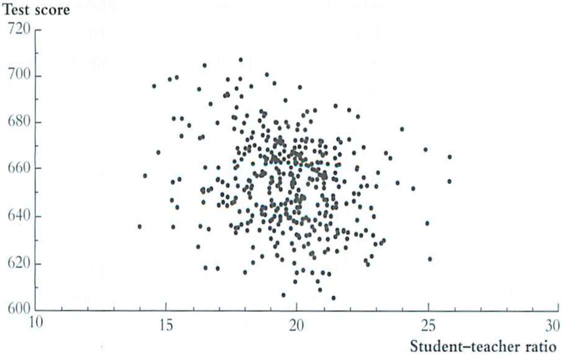
			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

Question:

Do districts with smaller classes (lower STR) have higher test scores? And by how much?

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.



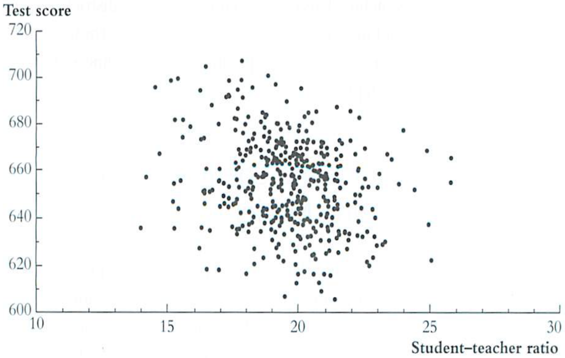
The class size/test score policy question:

- What is the effect of reducing STR by one student/teacher on test scores ?
- Object of policy interest: $\frac{\Delta \text{Test Score}}{\Delta STR}$.
- This is the *slope* of the line relating test score and STR.

This suggests that we want to draw a line through the *Test Score* v.s. *STR* scatterplot.

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.



But how?

Linear Regression: Some Notation and Terminology

The *population regression line* is

$$\begin{aligned}
Test\ Score &= \beta_0 + \beta_1 \cdot STR \\
\beta_1 &= \text{slope of population regression line} \\
&= \frac{\Delta \text{Test Score}}{\Delta STR} \\
&= \text{change in test score for a} \\
&\quad\ \text{unit change in STR}
\end{aligned}$$

$$Test\ Score = \beta_0 + \beta_1 \cdot STR$$

- β_0 and β_1 are "population" parameters.
- We would like to know the population value of β_1.
- We don't know β_1, so we must estimate it using data.

The Population Linear Regression Model— general notation

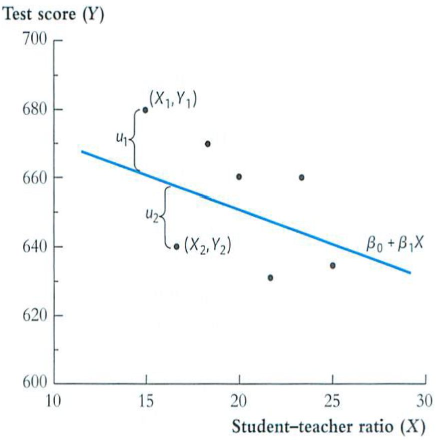
$$Y_i = \beta_0 + \beta_1 X_i + u_i, \; i = 1, \cdots n$$

- X is the **independent variable** or **regressor**.
- Y is the **dependent variable**.
- β_0 = **intercept**.
- β_1 = **slope**.

Figure 4.1 Scatter Plot of Test Score vs. Student-Teacher Ratio (Hypothetical Data)

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.



- u_i = the regression **error**.
- The regression error u_i consists of omitted factors, or possibly measurement error in the measurement of Y. In general, these omitted factors are other factors that influence Y, other than the variable X.

The Ordinary Least Squares Estimator

How can we estimate β_0 and β_1 from data?

We will focus on the least squares (“ordinary least squares" or “OLS") estimator of the unknown parameters β_0 and β_1, which solves

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right)^2$$

The OLS estimator solves:

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right)^2$$

- The OLS estimator minimizes the sum of squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem can be solved.
- The result is the OLS estimators of β_0 and β_1.

Why use OLS, rather than some other estimator?

- The OLS estimator has some desirable properties. Under **certain** assumptions, it is unbiased (that is, $E(\hat{\beta}_1) = \beta_1$), and it has a **tighter** sampling distribution than some other candidate estimators of β_1.

- This is what everyone uses— the common "language" of linear regression.

Derivation of the OLS Estimators

$$
\begin{aligned}
\min_{b_0, b_1} S &\equiv \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 \\
\frac{\partial S}{\partial b_0} &= -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0 \qquad (1) \\
\frac{\partial S}{\partial b_1} &= -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) X_i = 0 \qquad (2)
\end{aligned}
$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the values of b_0 and b_1 that solve the above two normal equations.

From equations (1) and (2), and divide each term by n, we have

$$\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0 \tag{3}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \bar{X} - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 0 \tag{4}$$

From (3), $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, substitute $\hat{\beta}_0$ in (4) and collect terms, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - (\bar{Y} - \hat{\beta}_1 \bar{X})\bar{X} - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right)\hat{\beta}_1$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right)\hat{\beta}_1$$

Therefore,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2} \\ &= \frac{\sum_{i=1}^{n} X_i {\color{red}Y_i} - n\bar{X}{\color{red}\bar{Y}}}{\sum_{i=1}^{n} X_i {\color{red}X_i} - n\bar{X}{\color{red}\bar{X}}}\end{aligned}$$

The numerator can be rewritten as

$$
\begin{aligned}
& \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} - n\bar{Y}\bar{X} + n\bar{X}\bar{Y} \\
= \; & \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \bar{Y} - \sum_{i=1}^{n} Y_i \bar{X} + \sum_{i=1}^{n} \bar{X}\bar{Y} \\
= \; & \sum_{i=1}^{n} \left(X_i Y_i - X_i \bar{Y} - Y_i \bar{X} + \bar{X}\bar{Y} \right) \\
= \; & \sum_{i=1}^{n} \left(X_i - \bar{X} \right)\left(Y_i - \bar{Y} \right)
\end{aligned}
$$

Similarily, the denominator can be written as

$$\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X}) = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

Therefore,

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}
\end{aligned}$$

KEY CONCEPT 4.2

The OLS Estimator, Predicted Values, and Residuals

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.5}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{4.6}$$

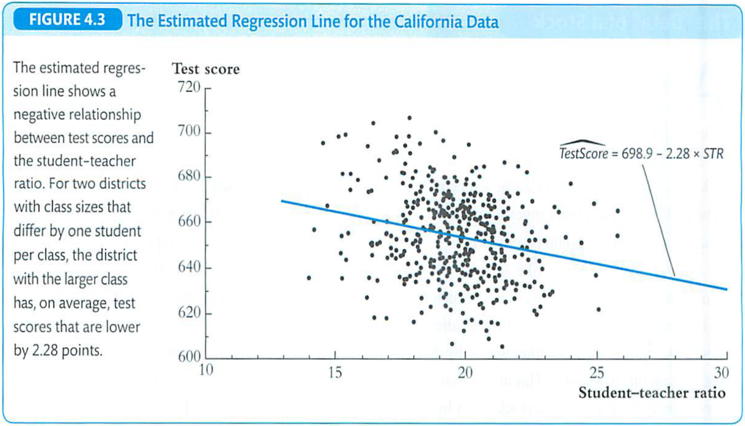
The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n \tag{4.7}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n. \tag{4.8}$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{u}_i$) are computed from a sample of n observations of X_i and Y_i, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

Application to the California Test Score-Class Size data



Estimated slope = $\hat{\beta}_1$ = - 2.28

Estimated intercept = $\hat{\beta}_0$ = 698.9

Estimated regression line: $\widehat{\text{Score}}$ = 698.9 - 2.28 STR

Interpretation of the estimated slope and intercept

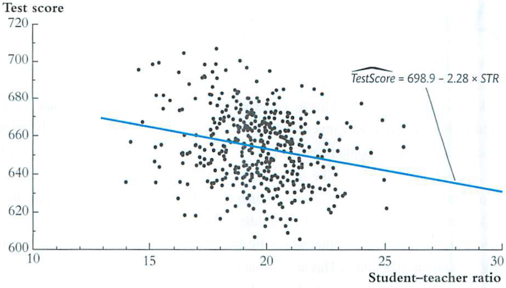
$$\widehat{\text{Test Score}} = 698.9 - 2.28\ STR$$

- Districts with one more student per teacher on average have test scores that are 2.28 points lower.
- That is, $\frac{\Delta \text{Test Score}}{\Delta STR} = -2.28$.
- The intercept (taken literally) means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9.
- This interpretation of the intercept makes no sense - it extrapolates the line outside the range of the data - in this application, the intercept is not itself economically meaningful.

Predicted values and residuals:

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. For two districts with class sizes that differ by one student per class, the district with the larger class has, on average, test scores that are lower by 2.28 points.



One of the districts in the data set is Antelope, CA, for which $STR = 19.33$ and Score = 657.8

$$
\begin{aligned}
\text{predicted value}: \; \hat{Y}_{Antelope} &= 698.9 - 2.28 \times 19.33 \\
&= 654.8 \\
\text{residual}: \; \hat{u}_{Antelope} &= 657.8 - 654.8 = 3.0
\end{aligned}
$$

OLS regression: STATA output

```
regress testscr str, robust

Regression with robust standard errors               Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581

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

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

Measures of Fit

A natural question is how well the regression line "fits" or explains the data. There are two regression statistics that provide complementary measures of the quality of fit.

- The *regression* R^2 measures the fraction of the variance of Y that is explained by X; it is unitless and ranges between zero (no fit) and one (perfect fit).
- The *standard error of the regression (SER)* measures the magnitude of a typical regression residual in the units of Y.

The R^2:

- The regression R^2 is the fraction of the sample variance of Y_i "explained" by the regression.
-

$$
\begin{aligned}
& TSS \\
\equiv \quad & \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}\right)^2 \\
= \quad & \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}\hat{u}_i(\hat{Y}_i - \bar{Y}) \\
= \quad & \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \equiv SSR + ESS
\end{aligned}
$$

where $\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = \sum_{i=1}^{n}\hat{u}_i(\hat{\beta}_0 + \hat{\beta}_1 X_i) = 0$ and $\sum_{i=1}^{n}\hat{u}_i\bar{Y} = 0$, becasue $\sum_{i=1}^{n}\hat{u}_i = 0$ and $\sum_{i=1}^{n}\hat{u}_i X_i = 0$ from equations (1) and (2).

Definition of R^2:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $R^2 = 0$ means $ESS = 0$.
- $R^2 = 1$ means $ESS = TSS$.
- $0 \leq R^2 \leq 1$.
- For regression with a single X, R^2= the square of the correlation coefficient between X and Y. (Exercise 4.12)

The Standard Error of the Regression (SER)

The **SER** measures the spread of the distribution of u. The SER is (almost) the sample standard deviation of the OLS residuals:

$$\begin{aligned} SER &= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2} \\ &= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2} \end{aligned}$$

The second equality holds bacause $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$.

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

The SER:

- has the units of u, which are the units of Y.
- measures the average "size" of the OLS residual (the average "mistake" made by the OLS regression line)
- The **root mean squared error** (RMSE) is closely related to the SER:

$$RSME = \; = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

This measures the same thing as the SER— the minor difference is division by 1/n instead of 1/(n-2).

Technical note: why divide by $n-2$ instead of $n-1$?

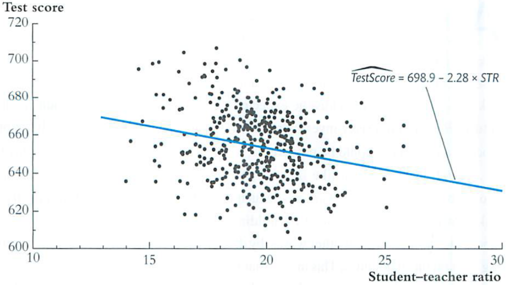
$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

- Division by n-2 is a “degrees of freedom" correction— just like division by n-1 in s_Y^2, except that for the SER, two parameters have been estimated (β_0 and β_1, by $\hat{\beta}_0$ and $\hat{\beta}_1$), whereas in s_Y^2 only one has been estimated (μ_Y, by $\bar{Y}$).
- When n is large, it makes negligible difference whether n, $n-1$, or $n-2$ are used— although the conventional formula uses $n-2$ when there is a single regressor.

Example of the R^2 and the SER

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. For two districts with class sizes that differ by one student per class, the district with the larger class has, on average, test scores that are lower by 2.28 points.



- $R^2 = 0.05$, $SER = 18.6$ STR explains only a small fraction of the variation in test scores.
- Does this make sense? Does this mean the STR is unimportant in a policy sense? No.

The Least Squares Assumptions

- What, in a precise sense, are the properties of the OLS estimator? We would like it to be unbiased, and to have a small variance. Does it? Under what conditions is it an unbiased estimator of the true population parameters?
- To answer these questions, we need to make some assumptions about how Y and X are related to each other, and about how they are collected (the sampling scheme).
- These assumptions— there are three— are known as the Least Squares Assumptions.

The Least Squares Assumptions

- The conditional distribution of u given X has mean zero, that is, $\mathrm{E}(u|X=x)=0$. This implies that $\hat{\beta}_1$ is unbiased.
- (X_i, Y_i), $i=1,\cdots,n$, are $i.i.d.$
    - This is true if X, Y are collected by simple random sampling.
    - This delivers the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Large outliers in X and/or Y are rare.
    - Technically, X and u have four moments, that is: $\mathrm{E}(X^4)<\infty$ and $\mathrm{E}(u^4)<\infty$.
    - Outliers can result in meaningless values of $\hat{\beta}_1$.

Least squares assumption #1: $E(u|X = x) = 0$.

For any given value of X, the mean of u is zero. This implies that X_i and u_i are uncorrelated, or $\text{Corr}(X_i, u_i) = 0$.

$Test\ Score_i = \beta_0 + \beta_1 STR_i + u_i$, u_i = other factors

"Other factors" include

- parental involvement
- outside learning opportunities (extra math class,..)
- home environment
- family income is a useful proxy for many such factors

So, $E(u|X = x) = 0$ means $E(Family\ Income|STR)$ = constant (which implies that family income and STR are uncorrelated).

Least squares assumption #2:

$(X_i, Y_i), i = 1, \cdots, n$ are $i.i.d.$

- This arises automatically if the entity (individual, district) is sampled by *simple random sampling*.
- The entity is selected then, for that entity, X and Y are observed (recorded).

Least squares assumption #3: Large outliers are rare.

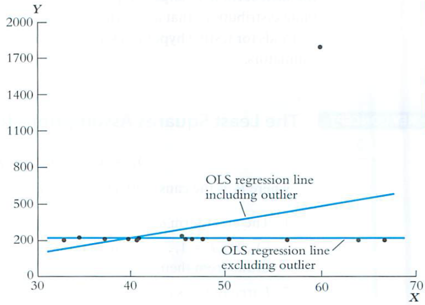
Technical statement: $E(X^4) < \infty$ and $E(u^4) < \infty$.

- A large outlier is an extreme value of X or Y.
- On a technical level, if X and Y are bounded, then they have finite fourth moments. (Standardized test scores automatically satisfy this; STR, family income, etc. satisfy this too).
- However, the substance of this assumption is that a large outlier can strongly influence the results.

OLS can be sensitive to an outlier

FIGURE 4.4 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.



- Is the lone point an outlier in X or Y?
- In practice, outliers often are data glitches (coding/recording problems)— so check your data for outliers! The easiest way is to produce a scatterplot.

Sampling Distribution of the OLS Estimators

The OLS estimator is computed from a sample of data; a different sample gives a different value of $\hat{\beta}_1$. This is the source of the "sampling uncertainty" of $\hat{\beta}_1$.

We want to:

- quantify the sampling uncertainty associated with $\hat{\beta}_1$.
- use $\hat{\beta}_1$ to test hypotheses such as $H_0 : \beta_1 = 0$.
- construct a confidence interval for β_1.

All these require figuring out the sampling distribution of the OLS estimator.

Probability Framework for Linear Regression

The Probability framework for linear regression is summarized by the three least squares assumption.

- **Population**
  population of interest (ex: all possible school districts)
- **Random variables:** Y, X (ex: $Test\ Score$, STR)
- **Joint distribution of** (Y, X)
  The population regression function is linear.
  $\mathrm{E}(u|X) = 0$
  X, Y have finite fourth moments.
- **Data collection by simple random sampling**
  $\{(X_i, Y_i)\}, i = 1, \cdots, n$ are $i.i.d.$

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

Like $\bar{Y}$, $\hat{\beta}_1$ has a sampling distribution.

- What is $E(\hat{\beta}_1)$? (where is it centered?)
- What is $\mathrm{Var}(\hat{\beta}_1)$? (measure of sampling uncertainty)
- What is its sampling distribution in small samples?
- What is its sampling distribution in large samples?

The mean and variance of the sampling distribution of $\hat{\beta}_1$

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + u_i \\
\bar{Y} &= \beta_0 + \beta_1 \bar{X} + \bar{u} \\
\text{so } Y_i - \bar{Y} &= \beta_1 (X_i - \bar{X}) + (u_i - \bar{u})
\end{aligned}
$$

Thus

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n} (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n} (X_i - \bar{X}) u_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}
\end{aligned}
$$

because $\sum_{i=1}^{n} (X_i - \bar{X}) \bar{u} = \bar{u} \sum_{i=1}^{n} (X_i - \bar{X}) = 0$.

$$
\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
\mathrm{E}(\hat{\beta}_1) &= \beta_1 + \mathrm{E}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \\
&= \beta_1 + \mathrm{E}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})\mathrm{E}(u_i|X_1, \cdots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \\
&= \beta_1
\end{aligned}
$$

$\hat{\beta}_1$ is unbiased.

Law of Iterated Expectations: $\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y|X))$.

Calculate the variance of $\hat{\beta}_1$.

$$
\begin{aligned}
\hat{\beta}_1 - \beta_1 &= \frac{\frac{1}{n}\sum_{i=1}^{n} \color{red}{v_i}}{\left(\frac{n-1}{n}\right) s_x^2} \\
\text{where } v_i &\equiv (X_i - \bar{X})u_i \\
s_x^2 &\equiv \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(X_i - \bar{X})^2
\end{aligned}
$$

The calculation is simplified by supposing that n is large (so that s_x^2 can be replaced by σ_x^2), the result is

$$
\mathrm{Var}(\hat{\beta}_1) = \frac{\mathrm{Var}(v)}{n(\sigma_x^2)^2}
$$

- *The central limit theorem.*

  If $Y_1, \cdots, Y_n$ are $i.i.d.$ and $0 < \sigma_Y^2 < \infty$, then

$$\sqrt{n}(\bar{Y} - \mu_Y) \xrightarrow{d} N(0, \sigma_Y^2)$$
$$\bar{Y} \xrightarrow{d} N(\mu_Y, \frac{\sigma_Y^2}{n})$$

  Or, the asymptotic distribution of

$$\sqrt{n}\frac{\bar{Y} - \mu_Y}{\sigma_Y} = \frac{\bar{Y} - \mu_Y}{\frac{\sigma_Y}{\sqrt{n}}} = \frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}}$$

  is $N(0, 1)$.

Because

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right) s_x^2}$$

when n is large

- $v_i = (X_i - \bar{X})u_i$ is $i.i.d.$ and has two moments. That is $\mathrm{Var}(v_i) < \infty$. Thus $\frac{1}{n}\sum_{i=1}^{n} v_i$ is distributed $N(0, \frac{\mathrm{Var}(v)}{n})$ when n is large. (central limit theorem)
- s_x^2 is approximately equal to σ_x^2 when n is large.
- $\frac{n-1}{n} = 1 - \frac{1}{n} \to 1$ when n is large.

Putting these together we have:

Large-n approximation to the distribution of $\hat{\beta}_1$:

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right) s_X^2} \simeq \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2},$$

which is approximately distributed $N(0, \frac{\sigma_v^2}{n(\sigma_X^2)^2})$.

Because $v_i = (X_i - \bar{X})u_i$, we can write this as:
$\hat{\beta}_1$ is approximately distributed $N\left(\beta_1, \frac{\mathrm{Var}[(X_i - \mu_X)u_i]}{n\sigma_X^4}\right)$.

Fact:

The larger the variance of X, the smaller the variance of $\hat{\beta}_1$

The math:

$$\mathrm{Var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\mathrm{Var}\left[(X_i - \mu_X)u_i\right]}{\sigma_x^4}$$

where $\sigma_X^2 = \mathrm{Var}(X_i)$. The variance of X appears in its square in the denominator— so increasing the spread of X decreases the variance of β_1.

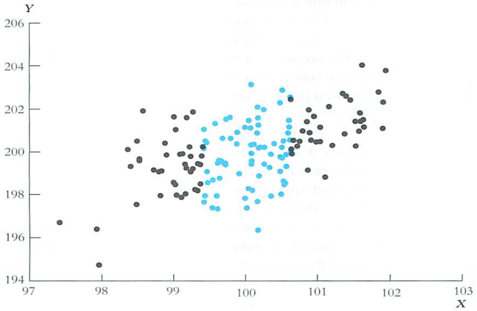
The intuition

If there is more variation in X, then there is more information in the data that you can use to fit the regression line. This is most easily seen in a figure.

The larger the variance of X, the smaller the variance of $\hat{\beta}_1$

FIGURE 4.5 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.



There are the same number of black and blue dots— using which would you get a more accurate regression line?

Another apporach to obtain an estimator: Apply Law of Large Number

- Under certain conditions on $Y_1, \cdots, Y_n$, the sample average $\bar{Y}$ converges in probability to the population mean.

  If $Y_1, \cdots, Y_n$ are $i.i.d.$, $\mathrm{E}(Y_i) = \mu_Y$, and $\mathrm{Var}(Y_i) < \infty$, then $\bar{Y} \xrightarrow{p} \mu_Y$.

- The least square assumption #1 $\mathrm{E}(u_i|1, X_i) = 0$ implies

$$
\begin{aligned}
\mathrm{E}(u_i \cdot 1) &= 0 \\
\mathrm{E}(u_i \cdot X_i) &= 0
\end{aligned}
$$

Apply Law of Large Nnumber, we have

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(u_i \cdot 1) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) \cdot 1 \\
&\xrightarrow{p} \mathrm{E}(u_i \cdot 1) = 0 \\
\frac{1}{n}\sum_{i=1}^{n}(u_i \cdot X_i) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) \cdot X_i \\
&\xrightarrow{p} \mathrm{E}(u_i X_i) = 0
\end{aligned}
$$

- Replacing the population mean with sample average is called the *analogy principle*.
- This leads to the two *normal equations* in the bivariate least squares regression.

Summary for the OLS estimator $\hat{\beta}_1$:

Under the three Least Squares Assumptions,

- The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has mean β_1 ($\hat{\beta}_1$ is an unbiased estmator of β_1), and $\mathrm{Var}(\hat{\beta}_1)$ is inversely proportional to n.
- Other than its mean and variance, the exact distribution of $\hat{\beta}_1$ is complicated and depends on the distribution of (X, u).
- $\hat{\beta}_1 \xrightarrow{p} \beta_1$. (law of large numbers)
- $\frac{\hat{\beta}_1 - \mathrm{E}(\beta_1)}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}}$ is approximately distributed $N(0, 1)$. (CLT)

KEY CONCEPT 4.4

Large-Sample Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

If the least squares assumptions in Key Concept 4.3 hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a jointly normal sampling distribution. The large-sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, where the variance of this distribution, $\sigma^2_{\hat{\beta}_1}$, is

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\operatorname{var}[(X_i - \mu_X)u_i]}{[\operatorname{var}(X_i)]^2}. \tag{4.19}$$

The large-sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, where

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\operatorname{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right]X_i. \tag{4.20}$$