Linear Regression with One Regressor (SW Ch. 4)

Ercan Karadas

Econometrics (ECON 3112)

Belk College of Business, UNCC

September 19, 2018

Outline

Overview of Linear Models

- Some questions:
  - A state implements tough new penalties on drunk drivers:
    $\rightarrow$ What is the effect on highway fatalities?
  - A school district cuts the size of its elementary school classes:
    $\rightarrow$ What is the effect on its students standardized test scores?
  - You successfully complete one more year of college classes:
    $\rightarrow$ What is the effect on your future earnings?
- In all these examples, we are interested in **quantifying** the relationship between two variables, say X and Y:

X	Y
New Penalties	Highway Fatalities
Class Size	Test Scores
Year of Education	Earnings

- Suppose that we think the relationship between X and Y is linear.

  $\implies$ If X changes by one unit, then Y changes by some fixed amount.

- Linear regression model lets us estimate this fixed quantity.
- This fixed quantity will correspond to the slope of the population regression line as we will see.
- Ultimately our aim is to estimate the **causal effect** on Y of a unit change in X - but for now, just think of the problem of fitting a straight line to data on two variables, Y and X.

The Population Linear Regression Model

- The **Population Linear Regression Model:**

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, 2, \ldots, n$$

- Terminology of the Linear Regression Model:
  - $\beta_0 + \beta_1 X_i$: **Population Regression Line** $(= E[Y_i|X_i])$
  - X: **independent variable** or **regressor**
  - Y: **dependent variable**
  - β_0 : **intercept** term
  - β_1 : **slope** term
  - u_i : **regression error**
  - $\{(X_i, Y_i)\}_{i=1}^n$: a **sample** of n observations
- The slope of the population regression line is the **expected (or average)** effect on Y of a unit change in X.
- The regression error u_i consists of **omitted factors**, factors other than X that influence Y.

Linear Regression Model for the Leading Example

- The Population Linear Regression Model:

$$\text{Test Score}_i = \beta_0 + \beta_1 \text{STR}_i + u_i, \quad i = 1, 2, \ldots, n$$

- The Population Linear Regression Line:

$$E\left[\text{Test Score}_i | \text{STR}_i\right] = \beta_0 + \beta_1 \text{STR}_i \quad i = 1, 2, \ldots, n$$

-

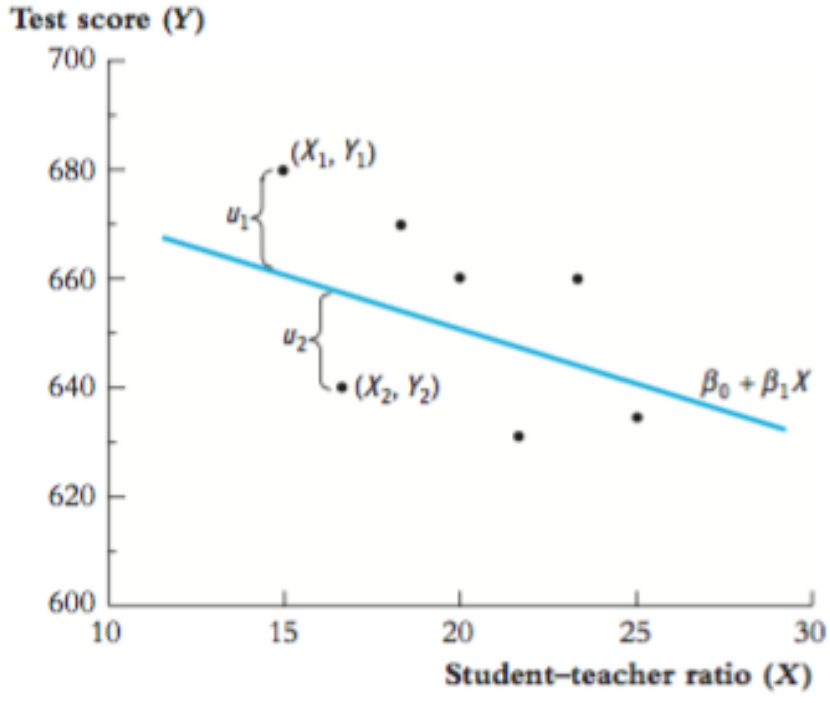
$$\begin{aligned}\beta_1 &= \text{Slope of population regression line} \\ &= \frac{\Delta \text{ Expected Test Score}}{\Delta \text{STR}} \\ &= \text{Change in expected Test Score for a unit change in STR}\end{aligned}$$

- The regression errors u_i consists of factors other than STR that influence Test Score, such as family income, parents' education, etc.

Scatter Plot of Test Score (Y) vs STR (X) and the Population Regression Line

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.



- Why are β_0 and β_1 "population" parameters?
- We would like to know the (unknown) population value of β_1 (why unknown?)
- We will use an estimator $\widehat{\beta}_1$, which will be a function of sample data, to estimate β_1
- The problem of statistical inference for β_1, at a general level, the same as for estimation of the population mean μ_Y by $\overline{Y}$
- Statistical inference about β_1 entails:
  - **Estimation:**
    - How should we draw a line through the data to estimate the population slope?
  - **Hypothesis testing:**
    - How to test if the slope is zero?
  - **Confidence intervals:**
    - How to construct a confidence interval for the slope?

The Ordinary Least Squares Estimator

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, 2, \ldots, n$$

- 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.
- The OLS estimator solves,

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

  - The OLS estimator minimizes the average 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 using calculus (App. 4.2).
- The result is the OLS estimators of β_0 and β_1.

The OLS Estimator, Predicted Values, and Residuals

- The **OLS estimator of the slope** β_1

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

- The **OLS estimator of the intercept** β_0

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

- The **OLS predicted values** $\widehat{Y}_i$ are

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i, \quad i = 1, 2, \ldots, n$$

- The **OLS residuals** $\widehat{u}_i$

$$\widehat{u}_i = Y_i - \widehat{Y}_i, \quad i = 1, 2, \ldots, n$$

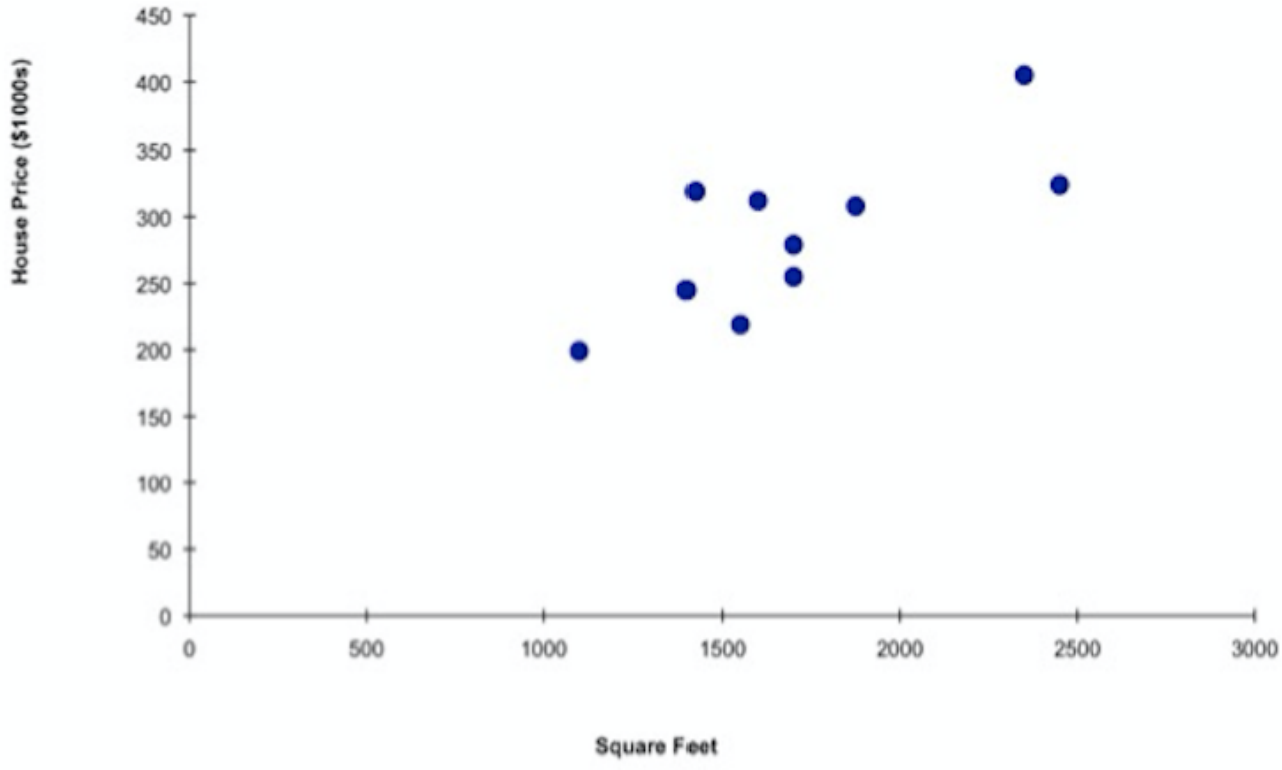
Example: House Price vs House Size

- Suppose a real estate agent wishes to examine the relationship between the selling price of a home and its size.
- As data, a random sample of 10 houses is selected in NODA, CLT
  - Dependent variable (Y): house price in \$1000s
  - Independent variable (X): house size in square feeet

House Price Model: Sample Data

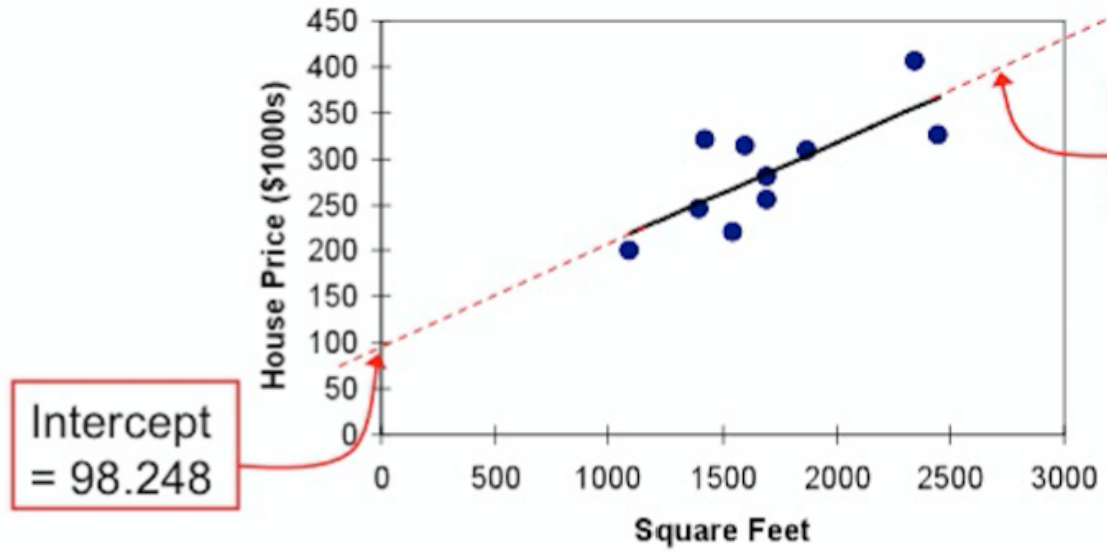
House Price in \$1000s (Y_i)	House Size in Square Feet (X_i)
405	2350
312	1600
279	1700
245	1400
308	1875
219	1550
199	1100
319	1425
324	2450
255	1700

House Price Model: Scatter Plot



House Price Model: Graphical Presentation of Regression

Scatter Plot and Sample Regression Line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

Interpretation of $\widehat{\beta}_0$

$$\widehat{\text{House Price}} = \mathbf{98.25} + 0.11 \times \text{Size (Square Feet)}$$

- $\widehat{\beta}_0$ is the estimated average value of Y when the value of X is zero
- However, this interpretation makes sense only if $X = 0$ is in the range of observed X values.
- Here, no houses had 0 square feet, so $\widehat{\beta}_0 = 98.25$ just indicates that, for houses within the range of sizes observed, \$98,250.00 is the portion of the house price not explained by square feet
- Or another interpretation would be the following: \$98,250.00 is the base price for a house that should be related to factors other than its size.

Interpretation of $\widehat{\beta}_1$

$$\widehat{\text{House Price}} = 98.25 + \mathbf{0.11} \times \text{Size (Square Feet)}$$

- $\widehat{\beta}_1$ measures the estimated change in the average value of Y as a result of a unit change in X
- Here, $\widehat{\beta}_1 = 0.11$ tells us that the average value of a house increases by $0.11 \times \$1000 = \$110$, on average, for each additional one square foot of size

Example: Test Score - Class Size data

- **Estimated Slope:** $\widehat{\beta}_1 = -2.28$
- **Estimated Intercept:** $\widehat{\beta}_0 = 698.9$
- **Estimated Regression Line:**

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

Interpretation of the estimated slope and intercept

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

- Districts with one more student per teacher on average have test scores that are 2.28 points lower.
- That is, $\frac{\Delta \text{Test Score}}{\Delta \text{STR}} = -2.28$
- In this example, 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.
- But this interpretation of the intercept makes no sense - it extrapolates the line outside the range of the data - here, the intercept is not economically meaningful.

Predicted Values & Residuals

- One of the districts in the data set is Antelope, CA, for which

$$\text{STR} = 19.33 \text{ and Test Score} = 657.8$$

- **Predicted Value:**

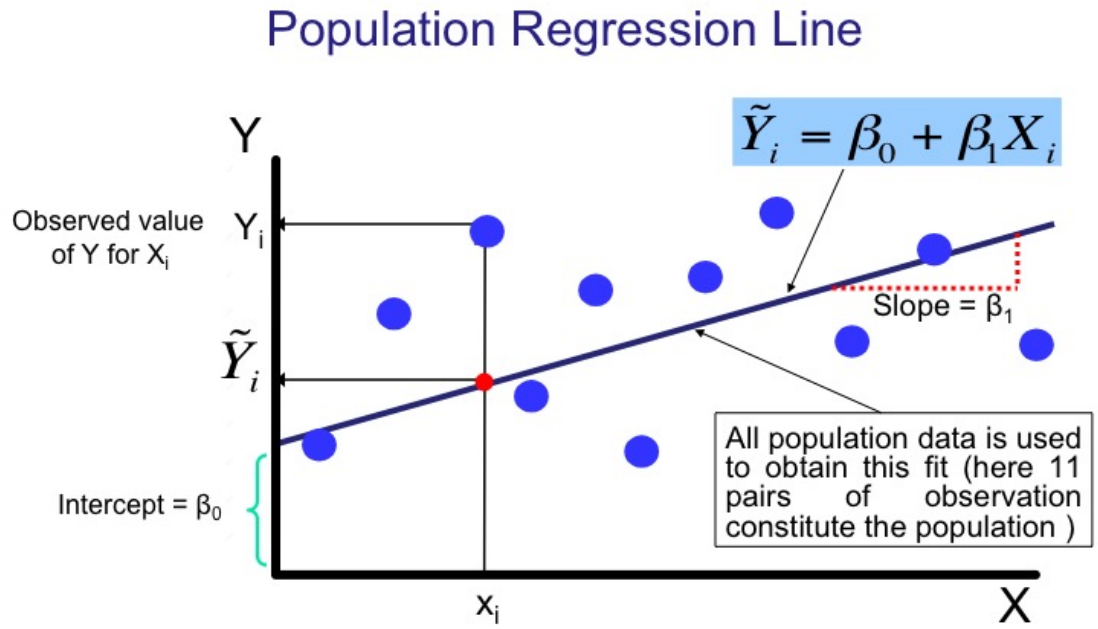
$$\widehat{Y}_{\text{Antelope}} = 698.9 - 2.28 \times 19.33 = 654.8$$

- **Residual:**

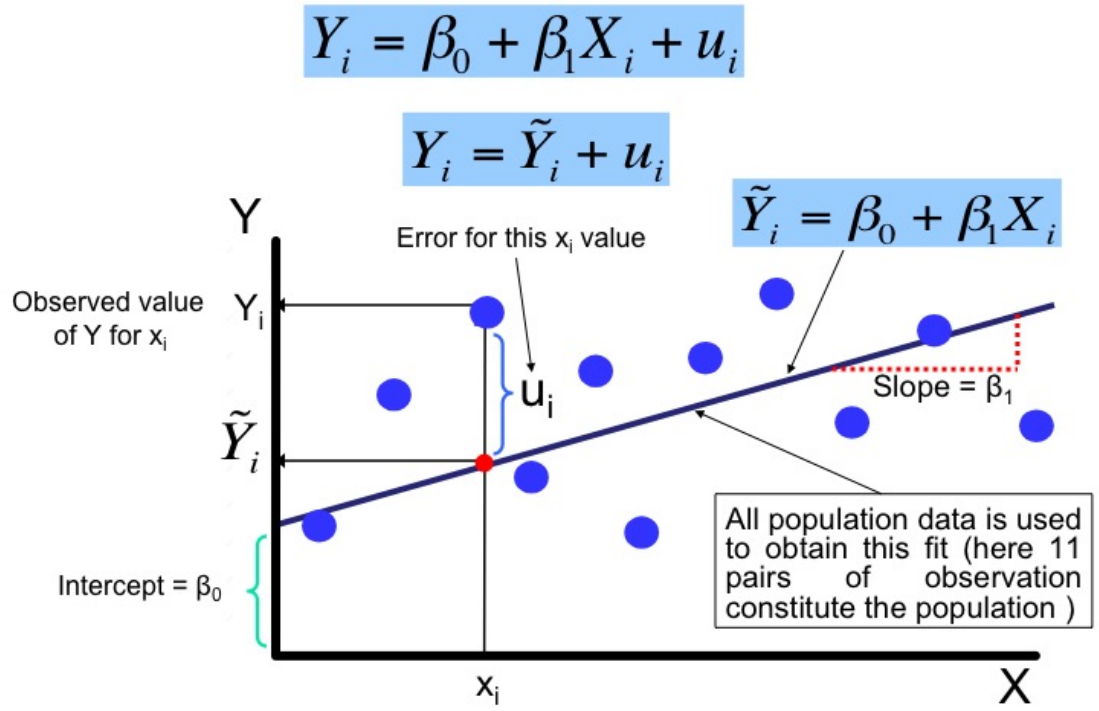
$$\widehat{u}_{\text{Antelope}} = 657.8 - 654.8 = 3.0$$

A Picturesque Story of Regression Analysis

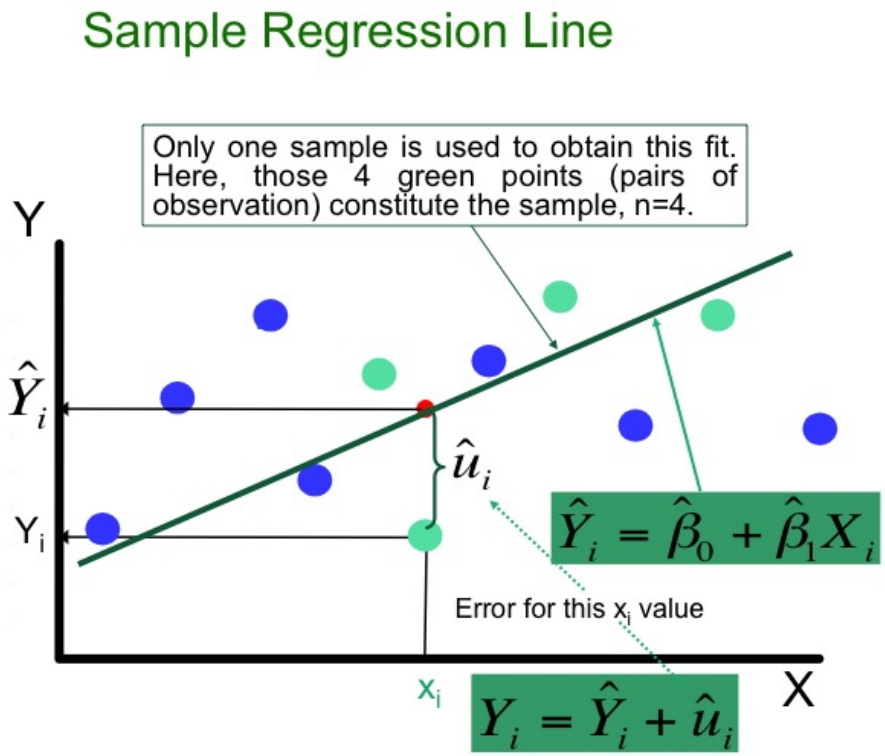
- If we had the population data (blue points below) we could compute β_0 and β_1
- And this would be the best (linear) model we could hope to work with...



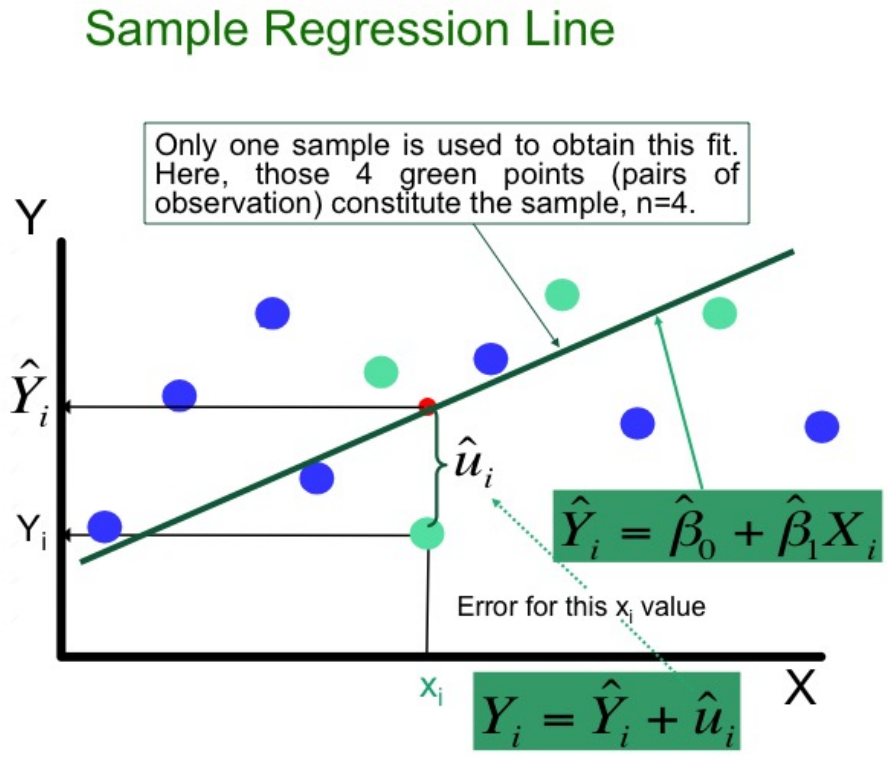
- Even in this case our population regression line, $\tilde{Y}_i$, would not match all the observations, Y_i.
- For each X_i, the model prediction, $\tilde{Y}_i$, contains some error, u_i, and therefore we have $Y_i = \tilde{Y}_i + u_i$
- But this is due to the linear structure of the model that we imposed.
- A much more serious problem is that in practice we never have the population data!

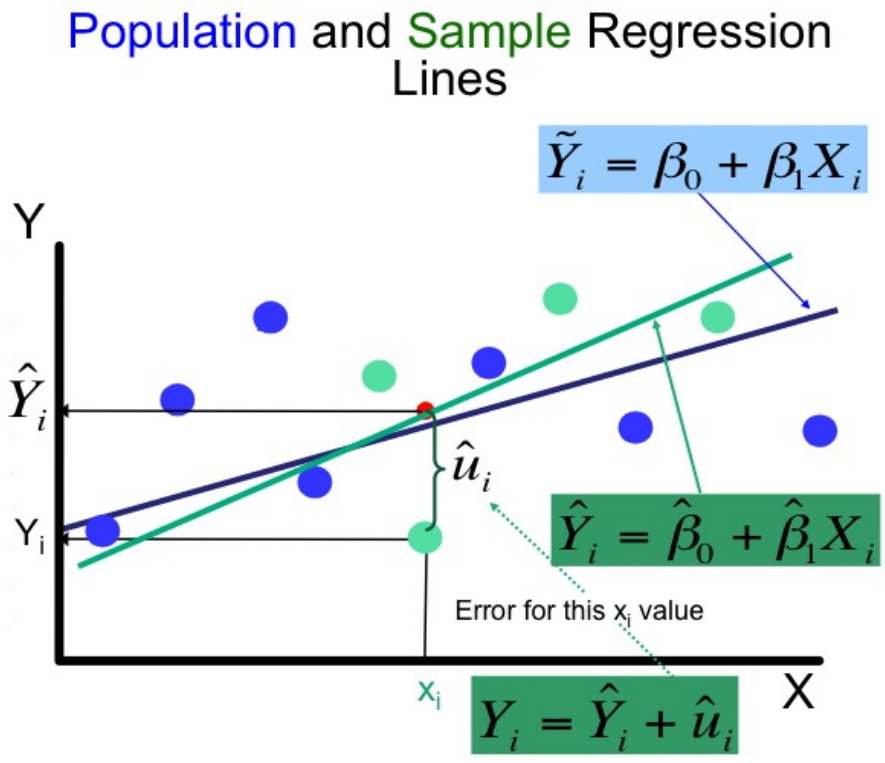


- In another word, in practice, we do not observe all of the blue points in the figure.
- Instead, suppose that we decided to work with a sample of size 4 and a random sample turned out to be the green points in the figure below
- A linear fit just taking into account these four points is also depicted in the figure

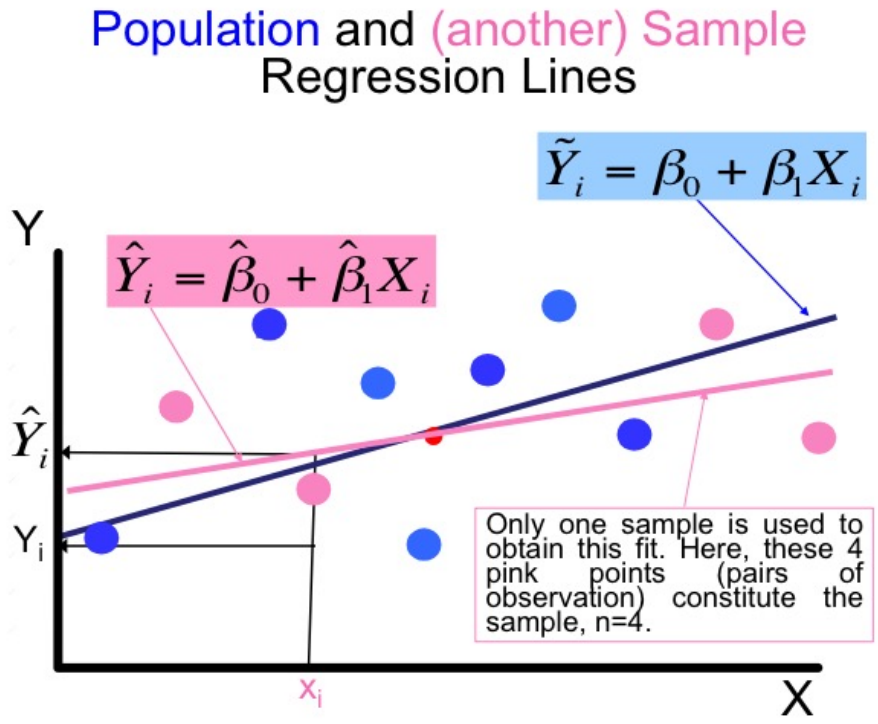


- In another words, in practice, we do not observe all of the blue points in the figure.
- Instead, suppose that we decided to work with a sample of size 4 which turned out to be the green points in the figure below
- A linear fit just taking into account these four points is also depicted in the figure



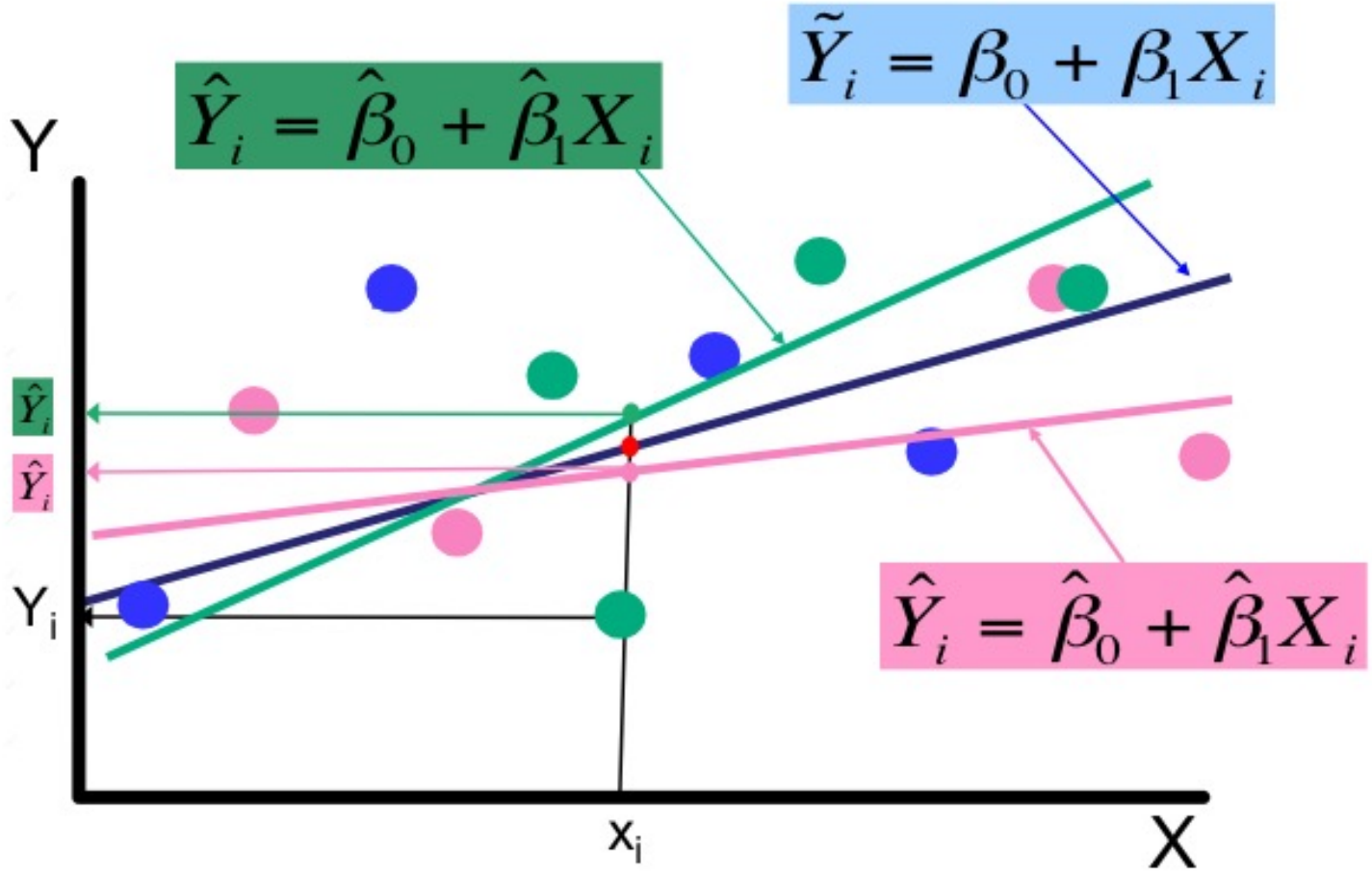


- Note that the population regression line $\tilde{Y}_i$, and the sample regression line $\hat{Y}_i$ are different.



- For another sample, say pink points above, we would fit a different sample regression line, i.e. $\widehat{\beta}_0$ and $\widehat{\beta}_1$ would change as we change the sample from green to pink.

Population and two Sample Regression Lines



Main take aways from this graphical presentation are the following:

- The population regression line is **unique**, i.e. β_0 and β_1 are fixed numbers, but **unobserved** as we don't have the population data
- On the other hand, we can compute the estimators, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, for a particular sample
- But they can change from sample to sample
- Therefore, in the regression analysis, we will study
  - Based on the sample results, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, what we can say about the unobserved population parameters β_0 and β_1
  - And what kind of potential problems we might encounter in interpreting the resultant estimates, $\widehat{\beta}_0$ and $\widehat{\beta}_1$
  - In particular, when can we interpret $\widehat{\beta}_1$ as the causal effect on Y of a unit change in X?

Measures of Fit of the Sample Regression

Question: How can we asses the performance of a model?

- For example, a good model for house prices should give us a satisfactory answer for the following questions: What is your best guess for the price of a randomly chosen house in the city? And how confident are you in your prediction?
- **Without a model.** Since we already have the price data for n houses in the city, $\{Y_1, \ldots, Y_n\}$, our best would be $\overline{Y}$.
  $\rightarrow$ If it turns out that actual price of the house is Y_i, our (prediction) error would be
  $\Longrightarrow Y_i - \overline{Y}$ **(crude prediction error)**
- **With a model.** Our prediction would be $\hat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$, where X_i is the size of the house that is assumed to be known.
  $\rightarrow$ And therefore our prediction error in this case would be
  $\Longrightarrow Y_i - \hat{Y}_i$ **(sophisticated prediction error)**
- Therefore, we can asses performance of the model by looking at how much the sophisticated errors improve the initial crude prediction errors.

Measures of Fit of the Sample Regression

Two regression statistics provide complementary measures of how well the regression line "fits" or "explains" the data:

- The **regression $\mathbf{R^2}$** measures the fraction of the variance of Y that is explained by X;
  - it is unitless
  - 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 Regression $\mathbf{R^2}$

- **The Regression $\mathbf{R^2}$** is the fraction of the sample variance of Y_i explained by the regression.

$$Y_i = \widehat{Y}_i + \widehat{u}_i = \text{OLS Prediction} + \text{OLS Residual}$$

$$\implies \text{sample var}(Y) = \text{sample var}(\widehat{Y}_i) + \text{sample var}(\widehat{u}_i)$$

$$\implies \text{Total SS} = \text{Explained SS} + \text{Residual SS}$$

$$\text{or } \mathbf{TSS} = \mathbf{ESS} + \mathbf{SSR}$$

- Definition of R^2

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}, \quad \left(= \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{\widehat{Y}}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}\right)$$

  - $R^2 = 0$ means $ESS = 0$
  - $R^2 = 1$ means $ESS = TSS$
  - $0 \leq R^2 \leq 1$

Analysis of Variance

- The total variation is made up of two parts:

$$\begin{array}{ccccc} \text{Total SS} & = & \text{Explained SS} & + & \text{Residual SS} \\ (\mathbf{TSS}) & & (\mathbf{ESS}) & & (\mathbf{SSR}) \end{array}$$

$$\sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\mathbf{Y_i} - \overline{\mathbf{Y}}\right)^{\mathbf{2}} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\widehat{\mathbf{Y}}_\mathbf{i} - \overline{\mathbf{Y}}\right)^{\mathbf{2}} + \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\mathbf{Y_i} - \widehat{\mathbf{Y}}_\mathbf{i}\right)^{\mathbf{2}}$$

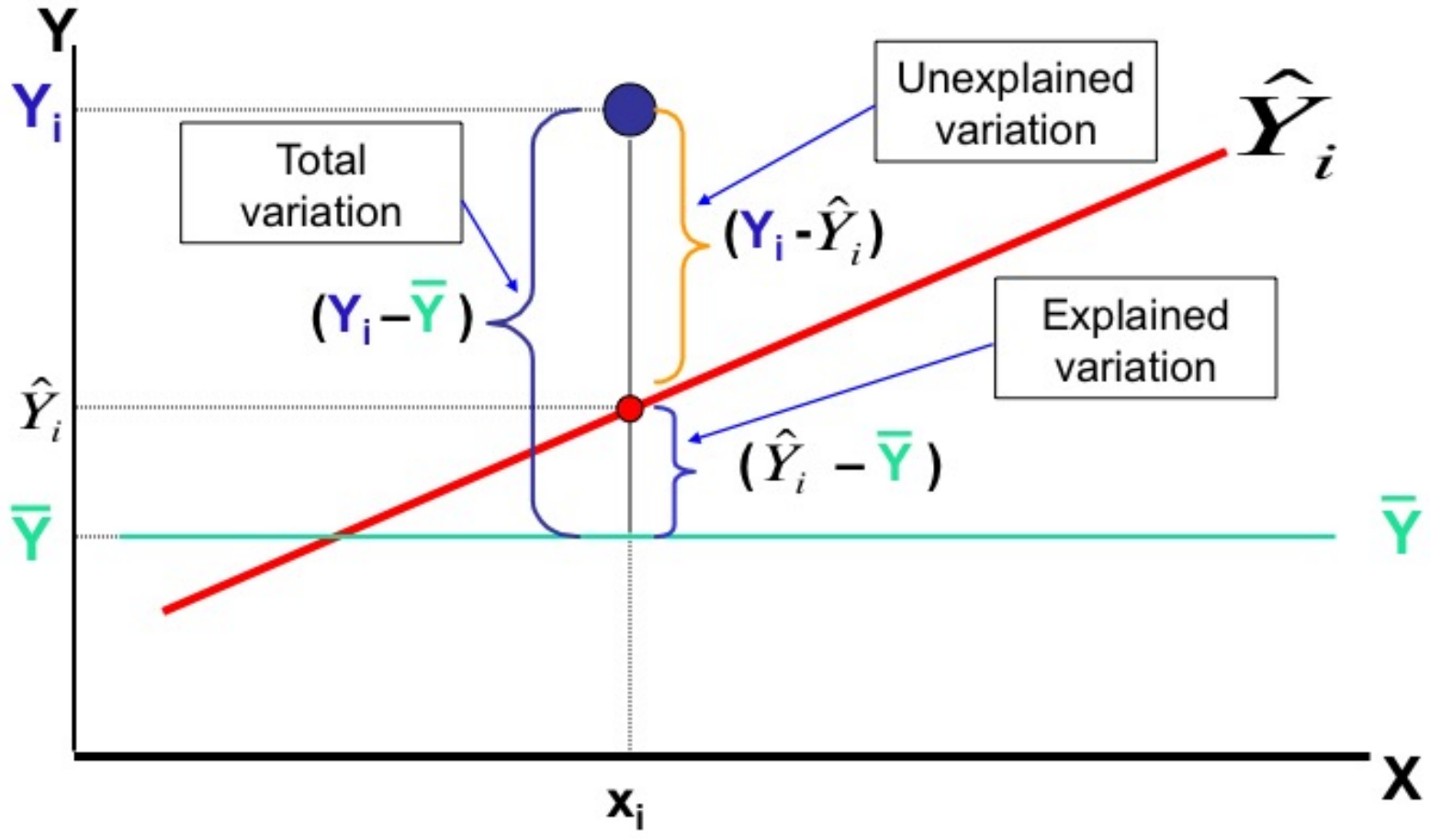
  where
  - Y_i: Observed (actual) values of the dependent variable
  - $\overline{Y}$: Average value of the dependent variable
  - $\widehat{Y}_i$: Predicted value of the dependent variable
- This decomposition (of TSS between ESS and SSR) is known as the analysis of variance.
- Note that using this decomposition ($TSS = ESS + SSR$), we can express R^2 alternatively as

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

Analysis of Variance

- TSS: Total Sum of Squares
  - Measures the variation of the Y_i values around their mean, $\overline{Y}$
- ESS: Explained Sum of Squares
  - Explained variation attributable to the linear relationship assumed in the population regression model between X and Y
- SSR: Residual Sum of Squares
  - Variation attributable to factors other than X

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \quad \text{(error decomposition)}$$



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:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(\widehat{u}_i - \overline{\widehat{u}}\right)^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2} = \sqrt{\frac{1}{n-2}SSR}$$

  where the second equality holds because $\overline{\widehat{u}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$.

- 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 Standard Error of the Regression (SER)

- Why divide by $n-2$ instead of $n-1$?
  - 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 $\widehat{\beta}_0$ and $\widehat{\beta}_1$), whereas in s_Y^2 only one has been estimated (μ_Y by $\overline{Y}$).
  - When n is large, it doesn't matter whether n, $n-1$, or $n-2$ are used - although the conventional formula uses $n-2$ when there is a single regressor.
- The Root Mean Squared Error (RMSE) is closely related to the SER:

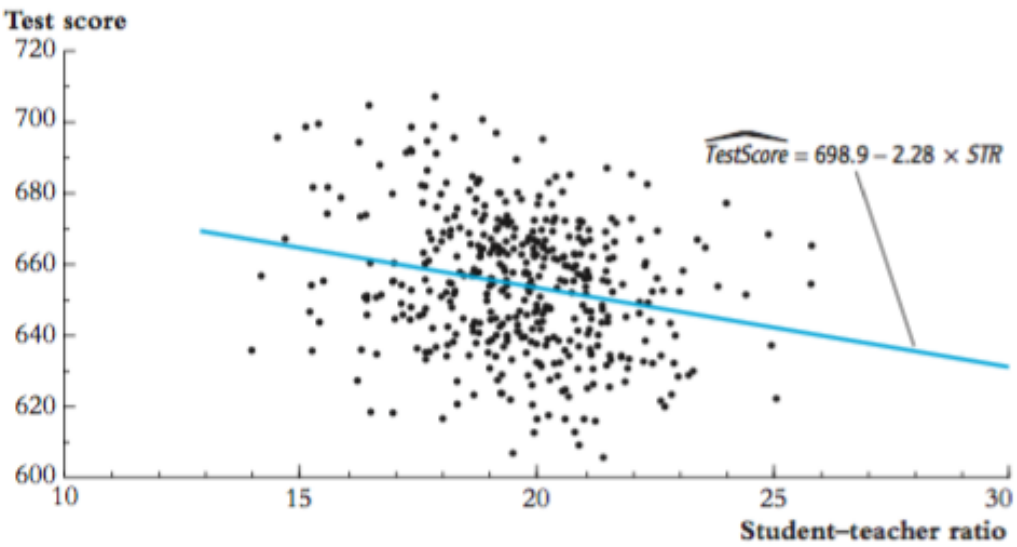
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2}$$

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

Example: Test Score - Class Size data (Cont'd)

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.



- **Regression Output:**

$$\widehat{\text{Test Score}} = 698.9 - 2.28 \times STR, \quad R^2 = 0.05, \quad SER = 18.6$$

Example: Test Score - Class Size data (Cont'd)

- **Regression Output:**

$$\widehat{\text{Test Score}} = 698.9 - 2.28 \times STR, \quad R^2 = 0.05, \quad SER = 18.6$$

- The R^2 of 0.051 means that the regressor STR explains 5.1% of the variance of the dependent variable Test Score.
- The SER of 18.6 means that standard deviation of the regression residuals is 18.6, where the units are points on the standardized test.
- The fact that the R^2 of this regression is low (and the SER is large) does not, by itself, imply that this regression is either "good" or "bad".
- What the low R^2 does tell us is that other important factors influence test scores.

The Least Squares Assumptions

- What, in a precise sense, are the properties of the sampling distribution of the OLS estimator? When will $\widehat{\beta}_1$ be unbiased? What is its variance?
- 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 Population Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, 2, \ldots, n$$

A1. For any given value of X, the mean of u is zero: $E[u|X] = 0$
- This implies that $\widehat{\beta}_1$ is unbiased

A2. (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d.
- This is true if (X, Y) are collected by simple random sampling
- This delivers the sampling distribution of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (Why?)

A3. Large outliers in X and/or Y are rare.
- Technically, X and Y have finite fourth moments
- Outliers can result in meaningless values of $\widehat{\beta}_1$

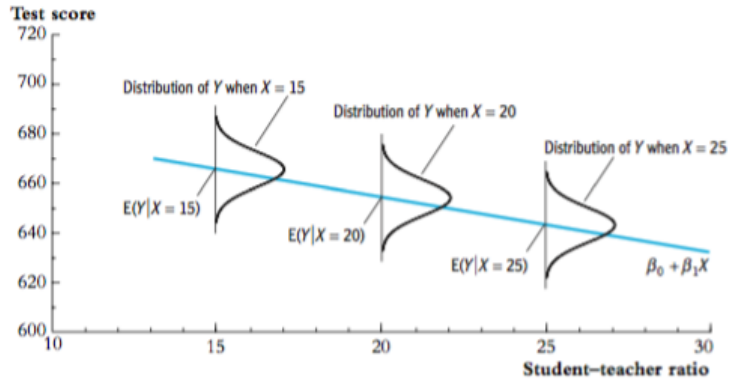
The Least Squares Assumption # 1: $E[u|X] = 0$

A1. For any given value of X, the mean of u is zero: $E[u|X] = 0$

- In our test score example

$$\text{Test Score} = \beta_0 + \beta_1 \text{STR} + u_i, \quad i = 1, 2, \ldots, n$$

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.

- What are some of these "other factors"?
- Is $E[u|X] = 0$ plausible for these other factors?

The Least Squares Assumption # 1: $E[u|X] = 0$

A benchmark for thinking about this assumption is to consider an ideal randomized controlled experiment:

- X is randomly assigned to people (students randomly assigned to different size classes; patients randomly assigned to medical treatments). Randomization is done by computer - using no information about the individual.
- Because X is assigned randomly, all other individual characteristics - the things that make up u - are distributed independently of X, so u and X are independent
- Thus, in an ideal randomized controlled experiment, $E[u|X] = 0$ (that is (A1) holds)
- In actual experiments, or with observational data, we will need to think hard about whether $E[u|X] = 0$ holds.

The Least Squares Assumption # 2: (X_i, Y_i)'s are i.i.d.

- This arises automatically if the entity (individual, district) is sampled by simple random sampling:
- The entities are selected from the same population, so (X_i, Y_i) are identically distributed for all $i = 1, \ldots, n$.
- The entities are selected at random, so the values of (X_i, Y_i) for different entities are independently distributed.
- The main place we will encounter non-i.i.d. sampling is when data are recorded over time for the same entity (panel data and time series data) - we will deal with that complication in later chapters.
- This assumption is the key to apply LLN and CLT to obtain the sampling distribution of our estimators

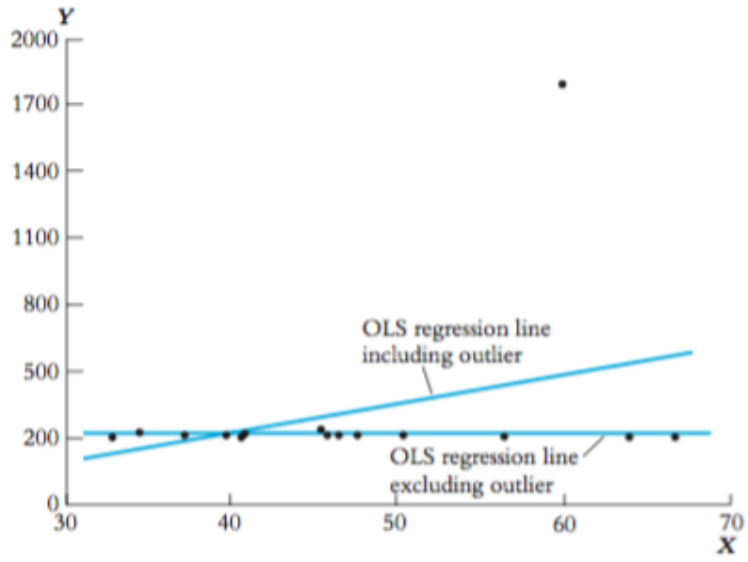
The Least Squares Assumption # 3: Outliers are Rare

- A large outlier is an extreme value of X or Y
- On a technical level, if X and Y are bounded, then this assumption is met (Technically, this implies X and Y have finite fourth moments, i.e. $E\left(X^4\right) < \infty$ and $E\left(Y^4\right) < \infty$)
- Standardized test scores automatically satisfy this; STR, family income, etc. satisfy this too.
- The substance of this assumption is that a large outlier can strongly influence the results - so we need to rule out large outliers.
- Look at your data! If you have a large outlier, is it a typo? Does it belong in your data set? Why is it an outlier?

OLS can be Sensitive to Outliers

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.



- Is the lone point an outlier in X or Y?
- In practice, outliers are often data glitches (coding or recording problems).
- Sometimes they are observations that really shouldnt be in your data set. Plot your data!

The Sampling Distribution of the OLS Estimator

- The OLS estimator is computed from a sample of data. A different sample yields a different value of $\widehat{\beta}_1$. This is the source of the sampling uncertainty of $\widehat{\beta}_1$.
- We want to:
  - quantify the sampling uncertainty associated with $\widehat{\beta}_1$
  - use $\widehat{\beta}_1$ 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. Two steps to get there:
  - Probability framework for linear regression model
  - Distribution of the OLS estimator

Probability Framework for Linear Regression Model

- The probability framework for linear regression is summarized by the following four assumptions:
  - The population regression line is linear
  - $E[u|X] = 0$ (L.S.A # 1)
  - (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d. (L.S.A # 2)
  - Large outliers in X and/or Y are rare (L.S.A # 3)
- These assumptions let us to use LLN and CLT to derive the sampling distribution of

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

$\widehat{\beta}_1$ has a sampling distribution, just like $\overline{Y}$ has one (answer each of the following questions for $\overline{Y}$ to see the analogy between the two cases)

(i) $E(\widehat{\beta}_1) = ?$

- If $E(\widehat{\beta}_1) = \beta_1$, then $\widehat{\beta}_1$ is unbiased estimator of β_1- a desirable property.

(ii) $V(\widehat{\beta}_1) = ?$ (a measure of sampling uncertainty)

- We need to derive a formula so we can compute the standard error of $\widehat{\beta}_1$.

(iii) What is the distribution of $\widehat{\beta}_1$ in *large* samples?

- In large samples, $\widehat{\beta}_1$ is approximately normally distributed (CLT)

(iv) What is the distribution of $\widehat{\beta}_1$ in *small* samples?

- Depends on the parent distribution and it is very complicated in general.
- We will skip this for now.

(i) Computation of $E(\widehat{\beta}_1)$

We will use the formula for $\widehat{\beta}_1$ but first some preliminary algebra:

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

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \beta_0 + \sum_{i=1}^{n} \beta_1 X_i + \sum_{i=1}^{n} u_i$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n} n\beta_0 + \beta_1 \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} u_i$$

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

$$\implies Y_i - \overline{Y} = \beta_1 \left(X_i - \overline{X}\right) + \left(u_i - \overline{u}\right)$$

Thus

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \\
&= \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left[\beta_1 \left(X_i - \overline{X}\right) + \left(u_i - \overline{u}\right)\right]}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \\
&= \beta_1 \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} + \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(u_i - \overline{u}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}
\end{aligned}$$

So we obtain

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(u_i - \overline{u}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

Now

$$\begin{aligned}
\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(u_i - \overline{u}\right) &= \sum_{i=1}^{n} \left(X_i - \overline{X}\right) u_i - \overline{u} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) \\
&= \sum_{i=1}^{n} \left(X_i - \overline{X}\right) u_i - \overline{u} \left[\left(\sum_{i=1}^{n} \left(X_i\right) - n\overline{X} \right) \right] \\
&= \sum_{i=1}^{n} \left(X_i - \overline{X}\right) u_i - \overline{u} \left[\left(\sum_{i=1}^{n} X_i \right) - n\overline{X} \right] \\
&= \sum_{i=1}^{n} \left(X_i - \overline{X}\right) u_i
\end{aligned}$$

Substituting this into the expression for $\widehat{\beta}_1$ we obtain our key expression in this section:

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right) u_i}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \qquad (\star)$$

Now computing the expected value of this expression

$$\begin{aligned} E[\widehat{\beta}_1] &= E\left[\beta_1 + \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right) u_i}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}\right] \\ &= \beta_1 + \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right) E[u_i]}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \\ &= \beta_1 \qquad \text{(because } E[u_i|X] = 0 \text{ by LSA \#1)} \end{aligned}$$

$\Longrightarrow$ Thus LSA #1 implies that

$$E[\widehat{\beta}_1] = \beta_1$$

that is, $\widehat{\beta}_1$ is an unbiased estimator of β_1.

(ii) Computation of $V(\widehat{\beta}_1)$

We will again use our key equation $(\star)$ to compute the variance of $\widehat{\beta}_1$, but first we will simplify it by (i) defining $v_i = \left(X_i - \overline{X}\right) u_i$ in the numerator, (ii) using $s_X^2 = \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ in the denominator, and (iii) multiplying both denominator and numerator by $1/n$ to obtain:

$$\widehat{\beta}_1 = \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n v_i}{\left(\frac{n-1}{n}\right) s_X^2} \qquad (\star\star)$$

Now we can compute the variance:

$$\begin{aligned} V[\widehat{\beta}_1] &= V\left[\beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n v_i}{\left(\frac{n-1}{n}\right) s_X^2}\right] \\ &= V\left[\frac{\frac{1}{n} \sum_{i=1}^n v_i}{\left(\frac{n-1}{n}\right) s_X^2}\right] \\ &= \left(\frac{n}{n-1}\right)^2 \frac{1}{\left(s_X^2\right)^2} V\left[\frac{1}{n} \sum_{i=1}^n v_i\right] \\ \implies V[\widehat{\beta}_1] &= \left(\frac{n}{n-1}\right)^2 \frac{\sigma_v^2/n}{\left(s_X^2\right)^2} \end{aligned}$$

where $\sigma_v^2 = V[v_i] = V\left[\left(X_i - \overline{X}\right) u_i\right]$ and we used the fact that v_i's are i.i.d.

(iii) Large Sample Distribution of $\widehat{\beta}_1$

We will again use our key equation $(\star\star)$ to derive the large sample distribution of $\widehat{\beta}_1$

$$\widehat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right) s_X^2} \qquad (\star\star)$$

The rest of the derivation involves the following observations:

- As $n \to \infty$:
  - $v_i = \left(X_i - \overline{X}\right) u_i \longrightarrow \left(X_i - \mu_X\right) u_i$ as $\overline{X} \longrightarrow \mu_X$
  - $\left(\frac{n-1}{n}\right) \longrightarrow 1$ and $s_X^2 \longrightarrow \sigma_X^2$
- So we obtain $\widehat{\beta}_1 \longrightarrow \beta_1 + \frac{\overline{v}}{\sigma_X^2}$
- As β_1 is a constant it just affects the location of the distribution not its shape and σ_X^2 is also a constant so we can focus on the large sample distribution of $\overline{v}$
- Since v_i's are i.i.d., CLT applies to $\overline{v}$: the distribution of $\overline{v}$ is approximately Normal with the mean $E(\overline{v}) = 0$ and the variance $V(\overline{v}) = \sigma_v^2/n$, where $\sigma_v^2 = V[v_i] = V\left[\left(X_i - \mu_X\right) u_i\right]$
- Combining all these results we obtain the large sample distribution of $\widehat{\beta}_1$ as

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2/n}{\left(\sigma_X^2\right)^2}\right)$$

Summary of the Sampling Distribution of $\widehat{\beta}_1$

- Under the three Least Squares Assumptions we have the following results regardless of the sample size
  - $E(\widehat{\beta}_1) = \beta_1$
  - $V(\widehat{\beta}_1) = \left(\frac{n}{n-1}\right)^2 \frac{\sigma_v^2/n}{\left(s_X^2\right)^2}$
- When n is **small**, other than its mean and variance, the sampling distribution of $\widehat{\beta}_1$ is complicated and depends on the joint distribution of (X, u)
- When n is **large**, the sampling distribution of $\widehat{\beta}_1$ is approx. Normal

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2/n}{\left(\sigma_X^2\right)^2}\right), \qquad \left(\sigma_v^2 = V\left[\left(X_i - \mu_X\right) u_i\right]\right)$$

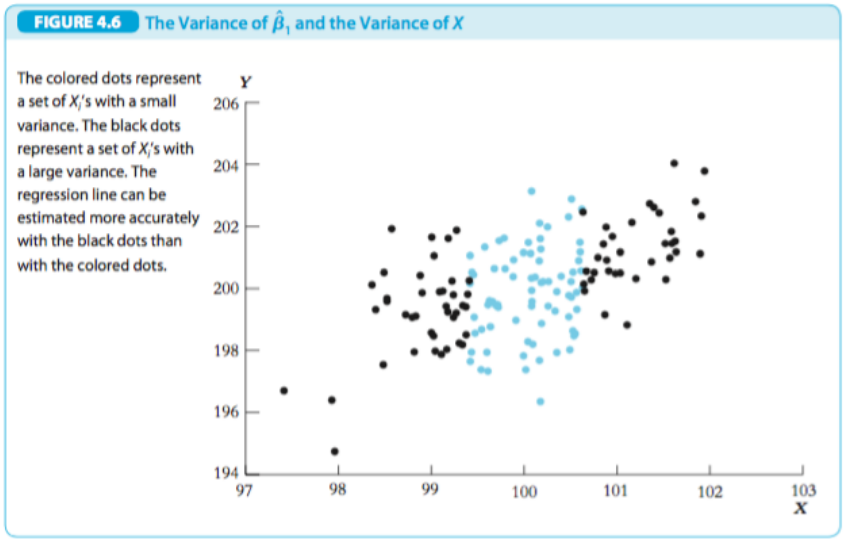
- This parallels the sampling distribution of $\overline{Y}$.

The larger the variance of X, the better $\widehat{\beta}_1$

- Recall that the variance of $\widehat{\beta}_1$ was

$$V[\widehat{\beta}_1] = \left(\frac{n}{n-1}\right)^2 \frac{\sigma_v^2/n}{(s_X^2)^2}$$

- Thus the larger the variance of X, the smaller the variance of $\widehat{\beta}_1$, i.e. it becomes a better (more efficient) estimator of β_1
- 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:



Are We Ready to Make Inferences about β_1?

- The results so far describe the sampling distribution of the OLS estimator when n is large:

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2/n}{(\sigma_X^2)^2}\right), \qquad \left(\sigma_v^2 = V\left[(X_i - \mu_X)\, u_i\right]\right)$$

- By themselves, however, these results are not sufficient to test a hypothesis about the value of β_1 or to construct a confidence interval for β_1.
- Doing so requires an estimator of the standard deviation of the sampling distribution- that is, the standard error of the OLS estimator.
- In another word, we need to estimate (unobserved) σ_v^2 using our sample data.
- This step -moving from the sampling distribution of β_1 to its standard error, hypothesis tests, and confidence intervals- is taken in the next chapter.