ECON 640: Univariate Regression Model

Yang Liang

Department of Economics
San Diego State University

Relating Two Variables

- Econometrics is concerned with understanding relationships between variables that we as economists care about.
  - Education and wages, investment and innovation, advertising and sales, class size and test scores...

Relating Two Variables

- Econometrics is concerned with understanding relationships between variables that we as economists care about.
  - Education and wages, investment and innovation, advertising and sales, class size and test scores...
- But given what we know so far, all we can do to study the relationship between two (or more) variables is to use covariance and correlation.

Covariance and Correlation

- Covariance measures how 2 variables move together
  - $Cov\left(X, Y\right) = E\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right]$

Covariance and Correlation

- Covariance measures how 2 variables move together
    - $Cov\left(X, Y\right) = E\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right]$
- But how can we *estimate* this? Suppose $\left(X_i, Y_i\right) \sim iid$ (pairs of observations are *iid*)
    - Then we can use $s_{XY} = \frac{1}{n-1}\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$
    - Furthermore, we can show that $s_{XY} \xrightarrow{p} \sigma_{XY}$

Covariance and Correlation

- Covariance measures how 2 variables move together
  - $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$
- But how can we *estimate* this? Suppose $(X_i, Y_i) \sim iid$ (pairs of observations are *iid*)
  - Then we can use $s_{XY} = \frac{1}{n-1} \sum (X_i - \overline{X})(Y_i - \overline{Y})$
  - Furthermore, we can show that $s_{XY} \xrightarrow{p} \sigma_{XY}$
- Correlation also measures how two variables move together
  - In particular, $r_{XY} = \frac{s_{XY}}{s_X s_Y}$(it's also true that $r_{XY} \xrightarrow{p} \rho_{XY}$)

Correlation and Causation

- But does $r_{XY} > 0$ mean that high values of X **cause** the values of Y to be high?

Correlation and Causation

- But does $r_{XY} > 0$ mean that high values of X **cause** the values of Y to be high?
- No, correlation is not causation. Examples
  - Umbrellas & rain
  - Police & crime
  - Years of schooling & wage

Correlation and Causation

- But does $r_{XY} > 0$ mean that high values of X **cause** the values of Y to be high?
- No, correlation is not causation. Examples
  - Umbrellas & rain
  - Police & crime
  - Years of schooling & wage
- To get at causation, we often need to control for confounding factors.
- Regression analysis will allow us to do so.

Regression Analysis

- Furthermore, we usually care about more than just correlation.

Regression Analysis

- Furthermore, we usually care about more than just correlation.
- In many cases we want to know
  - If we increase X by a certain amount, what is the expected effect on Y?

Regression Analysis

- Furthermore, we usually care about more than just correlation.
- In many cases we want to know
  - If we increase X by a certain amount, what is the expected effect on Y?
- Are averages enough to answer this question?

Regression Analysis

- Furthermore, we usually care about more than just correlation.
- In many cases we want to know
  - If we increase X by a certain amount, what is the expected effect on Y?
- Are averages enough to answer this question?
- Let's start with a case where X is discrete and compare $E\left(Y \mid X\right)$ for two values of X.

Regression Analysis

Table 3.1 has data on average earnings for men and women.

TABLE 3.1 Hourly Earnings in the United States of Working College Graduates, Aged 25–34: Selected Statistics from the Current Population Survey, in 1998 Dollars

	Men			Women			Difference, Men vs. Women		
Year	$\overline{Y}_m$	s_m	n_m	$\overline{Y}_w$	s_w	n_w	$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	95% Confidence Interval for d
1992	17.57	7.50	1591	15.22	5.97	1371	2.35**	0.25	1.87–2.84
1994	16.93	7.39	1598	15.01	6.41	1358	1.92**	0.25	1.42–2.42
1996	16.88	7.29	1374	14.42	6.07	1235	2.46**	0.26	1.94–2.97
1998	17.94	7.86	1393	15.49	6.80	1210	2.45**	0.29	1.89–3.02

These estimates are computed using data on all full-time workers aged 25–34 from the CPS for the indicated years. The difference is significantly different from zero at the *5% or **1% significance level.

Is there a significant gender gap?

Regression Analysis

Table 3.1 has data on average earnings for men and women.

TABLE 3.1 **Hourly Earnings in the United States of Working College Graduates, Aged 25–34: Selected Statistics from the Current Population Survey, in 1998 Dollars**

	Men			Women			Difference, Men vs. Women		
Year	$\overline{Y}_m$	s_m	n_m	$\overline{Y}_w$	s_w	n_w	$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	95% Confidence Interval for d
1992	17.57	7.50	1591	15.22	5.97	1371	2.35**	0.25	1.87–2.84
1994	16.93	7.39	1598	15.01	6.41	1358	1.92**	0.25	1.42–2.42
1996	16.88	7.29	1374	14.42	6.07	1235	2.46**	0.26	1.94–2.97
1998	17.94	7.86	1393	15.49	6.80	1210	2.45**	0.29	1.89–3.02

These estimates are computed using data on all full-time workers aged 25–34 from the CPS for the indicated years. The difference is significantly different from zero at the *5% or **1% significance level.

Is there a significant gender gap?

Looking at 1998, the wage gap $\left(\overline{Y}_m - \overline{Y}_w\right)$ is \$2.45 per hour.

The standard error $SE\left(\overline{Y}_m - \overline{Y}_w\right) = .29$ so the t-stat for $H_0 : \overline{Y}_m - \overline{Y}_w = 0$ is $\frac{2.45-0}{.29} = 8.45$, which has a p-value that's very close to 0 $(2\Phi(-8.45) \approx 0)$.

Indeed, a 99% CI for the wage gap is $2.45 \pm 2.58 \cdot .29 = (1.7, 3.2)$

Regression Analysis

TABLE 3.1 Hourly Earnings in the United States of Working College Graduates, Aged 25–34: Selected Statistics from the Current Population Survey, in 1998 Dollars

	Men			Women			Difference, Men vs. Women		
Year	$\bar{Y}_m$	s_m	n_m	$\bar{Y}_w$	s_w	n_w	$\bar{Y}_m - \bar{Y}_w$	$SE(\bar{Y}_m - \bar{Y}_w)$	95% Confidence Interval for d
1992	17.57	7.50	1591	15.22	5.97	1371	2.35**	0.25	1.87–2.84
1994	16.93	7.39	1598	15.01	6.41	1358	1.92**	0.25	1.42–2.42
1996	16.88	7.29	1374	14.42	6.07	1235	2.46**	0.26	1.94–2.97
1998	17.94	7.86	1393	15.49	6.80	1210	2.45**	0.29	1.89–3.02

These estimates are computed using data on all full-time workers aged 25–34 from the CPS for the indicated years. The difference is significantly different from zero at the *5% or **1% significance level.

- So there is a gender gap and it's statistically significant.
- But is this a result of discrimination?

Regression Analysis

TABLE 3.1 Hourly Earnings in the United States of Working College Graduates, Aged 25–34: Selected Statistics from the Current Population Survey, in 1998 Dollars

	Men			Women			Difference, Men vs. Women		
Year	$\overline{Y}_m$	s_m	n_m	$\overline{Y}_w$	s_w	n_w	$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	95% Confidence Interval for d
1992	17.57	7.50	1591	15.22	5.97	1371	2.35**	0.25	1.87–2.84
1994	16.93	7.39	1598	15.01	6.41	1358	1.92**	0.25	1.42–2.42
1996	16.88	7.29	1374	14.42	6.07	1235	2.46**	0.26	1.94–2.97
1998	17.94	7.86	1393	15.49	6.80	1210	2.45**	0.29	1.89–3.02

These estimates are computed using data on all full-time workers aged 25–34 from the CPS for the indicated years. The difference is significantly different from zero at the *5% or **1% significance level.

- So there is a gender gap and it's statistically significant.
- But is this a result of discrimination?
- Quite possibly. But why might it not be?

Regression Analysis

TABLE 3.1 Hourly Earnings in the United States of Working College Graduates, Aged 25–34: Selected Statistics from the Current Population Survey, in 1998 Dollars

	Men			Women			Difference, Men vs. Women		
Year	$\overline{Y}_m$	s_m	n_m	$\overline{Y}_w$	s_w	n_w	$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	95% Confidence Interval for d
1992	17.57	7.50	1591	15.22	5.97	1371	2.35**	0.25	1.87–2.84
1994	16.93	7.39	1598	15.01	6.41	1358	1.92**	0.25	1.42–2.42
1996	16.88	7.29	1374	14.42	6.07	1235	2.46**	0.26	1.94–2.97
1998	17.94	7.86	1393	15.49	6.80	1210	2.45**	0.29	1.89–3.02

These estimates are computed using data on all full-time workers aged 25–34 from the CPS for the indicated years. The difference is significantly different from zero at the *5% or **1% significance level.

- So there is a gender gap and it's statistically significant.
- But is this a result of discrimination?
- Quite possibly. But why might it not be?
- Some "other factor" could be driving the relationship (experience, education).

Regression Analysis

- To establish gender bias we need to keep “everything else” constant which means that instead of looking at
$$E(earnings \mid gender)$$

Regression Analysis

- To establish gender bias we need to keep "everything else" constant which means that instead of looking at

$$E(earnings \mid gender)$$

we should be concerned with

$$E(earnings \mid gender, age, experience, education, etc.)$$

which is clearly too complicated to carry out with a simple table.

Regression Analysis

- To establish gender bias we need to keep "everything else" constant which means that instead of looking at
$$E(earnings \mid gender)$$
we should be concerned with
$$E(earnings \mid gender, age, experience, education, etc.)$$
which is clearly too complicated to carry out with a simple table.
- Furthermore, how could we use a difference in means analysis to analyze even $E(earnings \mid age)$?

Regression Analysis

- To establish gender bias we need to keep “everything else” constant which means that instead of looking at
$$E(earnings \mid gender)$$
we should be concerned with
$$E(earnings \mid gender, age, experience, education, etc.)$$
which is clearly too complicated to carry out with a simple table.
- Furthermore, how could we use a difference in means analysis to analyze even $E(earnings \mid age)$?
- It turns out that we can do both using regression analysis.

Univariate regression

- Let's keep it simple in the beginning and start with $E(Y \mid X)$.
- Of course, in most cases a univariate regression will be inadequate.

Univariate regression

- Let's keep it simple in the beginning and start with $E\left(Y \mid X\right)$.
- Of course, in most cases a univariate regression will be inadequate.
- Consider an example. Does
$$E\left(TestScore \mid ClassSize\right)$$
really capture the causal effect of class size on test scores?

Univariate regression

- Let's keep it simple in the beginning and start with $E\left(Y \mid X\right)$.
- Of course, in most cases a univariate regression will be inadequate.
- Consider an example. Does

$$E\left(TestScore \mid ClassSize\right)$$

  really capture the causal effect of class size on test scores?
- Aren't other variables also important (and perhaps driving the relationship)?
  - Teacher quality, parents' income, neighborhood....

Univariate regression

- Let's keep it simple in the beginning and start with $E\left(Y \mid X\right)$.
- Of course, in most cases a univariate regression will be inadequate.
- Consider an example. Does
$$E\left(\mathit{TestScore} \mid \mathit{ClassSize}\right)$$
really capture the causal effect of class size on test scores?
- Aren't other variables also important (and perhaps driving the relationship)?
  - Teacher quality, parents' income, neighborhood....
  - Can we identify the impact of class size on test scores without controlling for these other factors?
    - Probably not. But let's pretend for now...

Univariate regression

- For now we'll just stick to one X (and attribute these other factors to random variation).
- Adding more X's turns out to be pretty simple and will allow us to account for these additional factors explicitly.

Univariate regression

- For now we'll just stick to one X (and attribute these other factors to random variation).
- Adding more X's turns out to be pretty simple and will allow us to account for these additional factors explicitly.

Some Notation

- Y is the dependent variable.
- X is the independent variable or (better) regressor or covariate.

Univariate regression

- For now we'll just stick to one X (and attribute these other factors to random variation).
- Adding more X's turns out to be pretty simple and will allow us to account for these additional factors explicitly.

Some Notation

- Y is the dependent variable.
- X is the independent variable or (better) regressor or covariate.
- We say that we 'regress y on x'

Univariate regression

- We know that $\mathbb{E}\left(Y \mid X\right)$ is a function of X, but what function?

Univariate regression

- We know that $\mathbb{E}(Y \mid X)$ is a function of X, but what function?
- Let's start by assuming it's linear - since it's easy (but we'll relax this assumption later).
- Suppose $\mathbb{E}(Y \mid X)$ is linear in X

$$\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X$$

- In words, this is saying that if we know X, the expected value of Y is a linear function of X.
- $\beta_0 + \beta_1 X$ is then called the *population regression line* (the relationship that holds between Y and X on average).

Univariate regression

- So what do β_0 & β_1 represent? Consider the impact on Y of a one unit change in X.

$$\mathbb{E}\left(Y \mid X = x\right) \quad = \quad \beta_0 + \beta_1 x$$

Univariate regression

- So what do β_0 & β_1 represent? Consider the impact on Y of a one unit change in X.

$$
\begin{aligned}
\mathbb{E}\left(Y \mid X = x\right) &= \beta_0 + \beta_1 x \\
\mathbb{E}\left(Y \mid X = (x+1)\right) &= \beta_0 + \beta_1 \left(x+1\right)
\end{aligned}
$$

Univariate regression

- So what do β_0 & β_1 represent? Consider the impact on Y of a one unit change in X.

$$
\begin{aligned}
\mathbb{E}(Y \mid X = x) &= \beta_0 + \beta_1 x \\
\mathbb{E}(Y \mid X = (x+1)) &= \beta_0 + \beta_1 (x+1) \\
\mathbb{E}(Y \mid x+1) - \mathbb{E}(Y \mid x) &= \beta_0 + \beta_1 (x+1) - \beta_0 - \beta_1 x = \beta_1
\end{aligned}
$$

Univariate regression

- So what do β_0 & β_1 represent? Consider the impact on Y of a one unit change in X.

$$\begin{aligned}
\mathbb{E}\left(Y \mid X = x\right) &= \beta_0 + \beta_1 x \\
\mathbb{E}\left(Y \mid X = (x+1)\right) &= \beta_0 + \beta_1\left(x+1\right) \\
\mathbb{E}\left(Y \mid x+1\right) - \mathbb{E}\left(Y \mid x\right) &= \beta_0 + \beta_1\left(x+1\right) - \beta_0 - \beta_1 x = \beta_1
\end{aligned}$$

- So β_1 is the expected change in Y associated with a one unit change in X (i.e. the slope: $\beta_1 = \frac{\Delta Y}{\Delta X}$).

Univariate regression

- So what do β_0 & β_1 represent? Consider the impact on Y of a one unit change in X.

$$
\begin{aligned}
\mathbb{E}\left(Y \mid X = x\right) &= \beta_0 + \beta_1 x \\
\mathbb{E}\left(Y \mid X = (x+1)\right) &= \beta_0 + \beta_1 \left(x+1\right) \\
\mathbb{E}\left(Y \mid x+1\right) - \mathbb{E}\left(Y \mid x\right) &= \beta_0 + \beta_1 \left(x+1\right) - \beta_0 - \beta_1 x = \beta_1
\end{aligned}
$$

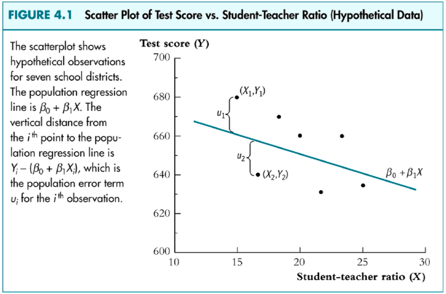
- So β_1 is the expected change in Y associated with a one unit change in X (i.e. the slope: $\beta_1 = \frac{\Delta Y}{\Delta X}$).
- β_0 is the intercept: the expected value of Y when $X = 0$.
    - The intercept is simply the point at which the population regression line intersects the Y axis.
    - Note that in applications where X cannot equal 0, the intercept has no "real world" meaning.

Univariate regression

- But $\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X$ doesn't mean that the data will all lie on the same line does it?

Univariate regression

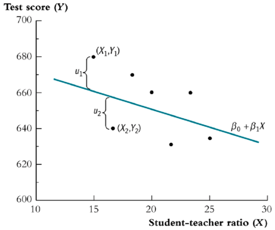
- But $\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X$ doesn't mean that the data will all lie on the same line does it?
- Notice that we **didn't** write $Y_i = \beta_0 + \beta_1 X_i$, but wrote $\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X$ instead.



Univariate regression

FIGURE 4.1 Scatter Plot 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.



- $\mathbb{E}(Y \mid X)$ is an expectation, the actual observations will be scattered around the population regression line:

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

- u_i represents all the other factors besides X_i that determine the value of Y_i for a particular observation i

Estimation

- Given that we've *assumed* there's a linear relationship between $\mathbb{E}(Y \mid X)$ and X, how do we estimate it?

Estimation

- Given that we've *assumed* there's a linear relationship between $\mathbb{E}\left(Y \mid X\right)$ and X, how do we estimate it?
- Intuitively, we want to estimate $\widehat{\mathbb{E}}\left(Y \mid X\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X$, where $\widehat{\beta}_0$ & $\widehat{\beta}_1$ are estimates of the population parameters β_0 & β_1
  - Just like $\overline{X}$ is an estimate of μ...

Estimation

- Given that we've *assumed* there's a linear relationship between $\mathbb{E}(Y \mid X)$ and X, how do we estimate it?
- Intuitively, we want to estimate $\widehat{\mathbb{E}}(Y \mid X) = \widehat{\beta}_0 + \widehat{\beta}_1 X$, where $\widehat{\beta}_0$ & $\widehat{\beta}_1$ are estimates of the population parameters β_0 & β_1
    - Just like $\overline{X}$ is an estimate of μ...
- So how do we find $\widehat{\beta}_0$ & $\widehat{\beta}_1$?

Estimation

- Given that we've *assumed* there's a linear relationship between $\mathbb{E}\left(Y \mid X\right)$ and X, how do we estimate it?
- Intuitively, we want to estimate $\widehat{\mathbb{E}}\left(Y \mid X\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X$, where $\widehat{\beta}_0$ & $\widehat{\beta}_1$ are estimates of the population parameters β_0 & β_1
    - Just like $\overline{X}$ is an estimate of μ...
- So how do we find $\widehat{\beta}_0$ & $\widehat{\beta}_1$?
    - By minimizing the prediction error.

Estimation

- Given that we've *assumed* there's a linear relationship between $\mathbb{E}\left(Y \mid X\right)$ and X, how do we estimate it?
- Intuitively, we want to estimate $\widehat{\mathbb{E}}\left(Y \mid X\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X$, where $\widehat{\beta}_0$ & $\widehat{\beta}_1$ are estimates of the population parameters β_0 & β_1
    - Just like $\overline{X}$ is an estimate of μ...
- So how do we find $\widehat{\beta}_0$ & $\widehat{\beta}_1$?
    - By minimizing the prediction error.
- Our estimates $\widehat{\beta}_0$ & $\widehat{\beta}_1$ will then give us the predicted value of Y conditional on X
    - The predicted values are $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$

Estimation

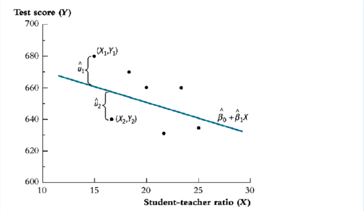
- Although we expect our estimates of β_0 & β_1 to be correct on average, for any particular observation i, we are likely to make a *prediction error*.

Estimation

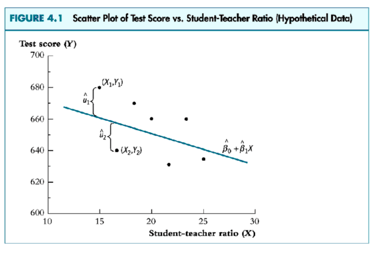
- Although we expect our estimates of β_0 & β_1 to be correct on average, for any particular observation i, we are likely to make a *prediction error*.
- The error made in predicting the i^{th} observation is given by

$$\widehat{u}_i \equiv Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$$

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



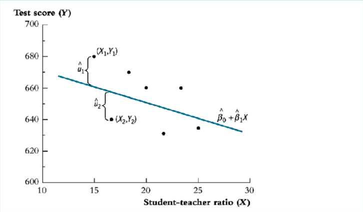
Estimation



- Intuitively, we would like to choose $\widehat{\beta}_0$ & $\widehat{\beta}_1$ to make all of these errors as small as possible. But how?

Estimation

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



- Intuitively, we would like to choose $\widehat{\beta}_0$ & $\widehat{\beta}_1$ to make all of these errors as small as possible. But how?
- Should we just minimize their sum? Should we set

$$\sum \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right) = 0?$$

- Let's see what happens if we do that.

- Suppose we have two data points:

$$X_1 = 1 \text{ and } Y_1 = 7 \qquad X_2 = 2 \text{ and } Y_2 = 9$$

- We can write $\widehat{Y}_i = 5 + 2X_i$
  - $\rightarrow \widehat{Y}_1 = 7$, $\widehat{Y}_2 = 9$

$$\begin{aligned} & \sum_{i=1}^{2} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right) \\ = \quad & \underbrace{(Y_1 - \widehat{Y}_1)}_{0} + \underbrace{(Y_2 - \widehat{Y}_2)}_{0} = 0 \end{aligned}$$

- Suppose we have two data points:

$$X_1 = 1 \text{ and } Y_1 = 7 \qquad X_2 = 2 \text{ and } Y_2 = 9$$

- We can write $\widehat{Y}_i = 5 + 2X_i$
  - $\rightarrow \widehat{Y}_1 = 7$, $\widehat{Y}_2 = 9$

$$\begin{aligned} & \sum_{i=1}^{2}\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right) \\ = \quad & \underbrace{(Y_1 - \widehat{Y}_1)}_{0} + \underbrace{(Y_2 - \widehat{Y}_2)}_{0} = 0 \end{aligned}$$

- We could also write $\widehat{Y}_i = 11 - 2X_i$
  - $\rightarrow \widehat{Y}_1 = 9$, $\widehat{Y}_2 = 7$

$$\begin{aligned} & \sum_{i=1}^{2}\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right) \\ = \quad & \underbrace{(Y_1 - \widehat{Y}_1)}_{-2} + \underbrace{(Y_2 - \widehat{Y}_2)}_{2} = 0 \end{aligned}$$

- Both solutions are equivalent under the criteria of

$$\sum \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right) = 0$$

- So $\sum \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right) = 0$ does not help to distinguish these two options (one of which is clearly wrong).

- Both solutions are equivalent under the criteria of

$$\sum\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right) = 0$$

- So $\sum\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right) = 0$ does not help to distinguish these two options (one of which is clearly wrong).
- Ideally, we would like a procedure that will set

$$\sum\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right) = 0$$

but identify the first case only.

Ordinary Least Squares (OLS)

- The OLS estimator chooses the regression coefficients by minimizing the sum of the **squared** prediction errors

$$\underset{\widehat{\beta}_0,\widehat{\beta}_1}{Min} \sum \left(Y_i - \widehat{Y}_i \right)^2 = \underset{\widehat{\beta}_0,\widehat{\beta}_1}{Min} \sum \left[Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i \right) \right]^2$$

[1] More reasons discussed shortly

Ordinary Least Squares (OLS)

- The OLS estimator chooses the regression coefficients by minimizing the sum of the **squared** prediction errors

$$\underset{\widehat{\beta}_0,\widehat{\beta}_1}{Min} \sum \left(Y_i - \widehat{Y}_i\right)^2 = \underset{\widehat{\beta}_0,\widehat{\beta}_1}{Min} \sum \left[Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)\right]^2$$

- Why not $\left|Y_i - \widehat{Y}_i\right|$? Because we'd like to use calculus.[1]

[1] More reasons discussed shortly

Ordinary Least Squares (OLS)

- The OLS estimator chooses the regression coefficients by minimizing the sum of the **squared** prediction errors

$$\underset{\widehat{\beta}_0,\widehat{\beta}_1}{Min}\sum\left(Y_i-\widehat{Y}_i\right)^2=\underset{\widehat{\beta}_0,\widehat{\beta}_1}{Min}\sum\left[Y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1X_i\right)\right]^2$$

- Why not $\left|Y_i-\widehat{Y}_i\right|$? Because we'd like to use calculus.[1]
- Taking partial derivatives yields

$$\begin{aligned}\frac{\partial}{\partial\widehat{\beta}_0}\sum\left[Y_i-\widehat{\beta}_0-\widehat{\beta}_1X_i\right]^2 &= -2\sum\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1X_i\right)\\ \frac{\partial}{\partial\widehat{\beta}_1}\sum\left[Y_i-\widehat{\beta}_0-\widehat{\beta}_1X_i\right]^2 &= -2\sum\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1X_i\right)X_i\end{aligned}$$

[1]More reasons discussed shortly

Ordinary Least Squares (OLS)

- Setting the partial derivatives equal to zero, collecting terms, dividing by n, and solving the resulting two equations in two unknowns for $\widehat{\beta}_0$ & $\widehat{\beta}_1$ yields:

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} = \frac{s_{XY}}{s_X^2} \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} \end{aligned}$$

Ordinary Least Squares (OLS)

- Setting the partial derivatives equal to zero, collecting terms, dividing by n, and solving the resulting two equations in two unknowns for $\widehat{\beta}_0$ & $\widehat{\beta}_1$ yields:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(X_i - \overline{X}\right)^2} = \frac{s_{XY}}{s_X^2} \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X}\end{aligned}$$

- So we can derive the estimating equations for $\widehat{\beta}_0$ & $\widehat{\beta}_1$ by minimizing the sum of squared prediction errors
    - Aside: In the same way, we can show $\overline{X}$ minimizes the sum of squared prediction errors and is the 'least squares' estimator of E(X).

Ordinary Least Squares (OLS)

- Setting the partial derivatives equal to zero, collecting terms, dividing by n, and solving the resulting two equations in two unknowns for $\widehat{\beta}_0$ & $\widehat{\beta}_1$ yields:

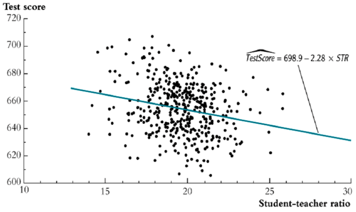
$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} = \frac{s_{XY}}{s_X^2} \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} \end{aligned}$$

- So we can derive the estimating equations for $\widehat{\beta}_0$ & $\widehat{\beta}_1$ by minimizing the sum of squared prediction errors
  - Aside: In the same way, we can show $\overline{X}$ minimizes the sum of squared prediction errors and is the 'least squares' estimator of E(X).
- Moreover, just like $\overline{X}$, $\widehat{\beta}_0$ & $\widehat{\beta}_1$ are themselves random variables. *(We'll derive their distributions shortly)*.

- Let's look at an example:

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 1 student, the estimated regression predicts that test scores will increase by 2.28 points.

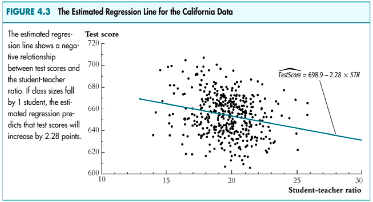


- The estimated regression line is

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

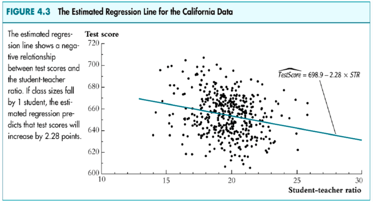
- So the expected impact on test scores of a one student increase in class size is -2.28 points.

Interpretation



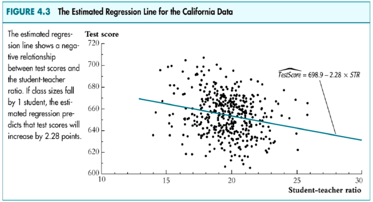
- So what then is the expected impact on test scores of a two student increase in class size?

Interpretation



- So what then is the expected impact on test scores of a two student increase in class size? $(2 \times -2.28 = -4.56$ points$)$

Interpretation



- So what then is the expected impact on test scores of a two student increase in class size? ($2 \times -2.28 = -4.56$ points)
- What is the expected test score in a district with 20 students per teacher?

Interpretation

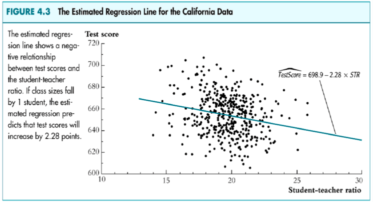
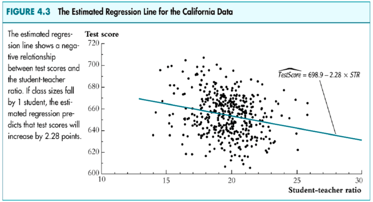


FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.

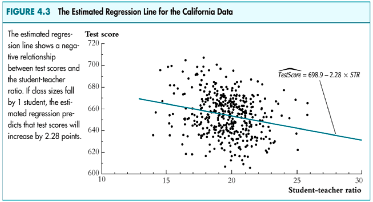
- So what then is the expected impact on test scores of a two student increase in class size? ($2 \times -2.28 = -4.56$ points)
- What is the expected test score in a district with 20 students per teacher? ($698.9 - 2.28 \cdot 20 = 653.3$ points)

Interpretation



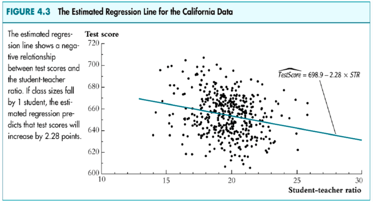
- So what then is the expected impact on test scores of a two student increase in class size? $(2 \times -2.28 = -4.56$ points$)$
- What is the expected test score in a district with 20 students per teacher? $(698.9 - 2.28 \cdot 20 = 653.3$ points$)$
- How about 30 students?

Interpretation



- So what then is the expected impact on test scores of a two student increase in class size? ($2 \times -2.28 = -4.56$ points)
- What is the expected test score in a district with 20 students per teacher? ($698.9 - 2.28 \cdot 20 = 653.3$ points)
- How about 30 students?
- How about 0 students?

Interpretation



- So what then is the expected impact on test scores of a two student increase in class size? $(2 \times -2.28 = -4.56$ points)
- What is the expected test score in a district with 20 students per teacher? $(698.9 - 2.28 \cdot 20 = 653.3$ points)
- How about 30 students?
- How about 0 students?
- You should be careful not to extrapolate beyond where you have data!

Estimation in Stata

So how do we run a regression in practice? We use a program with the OLS formulas built into it!

```
. reg 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
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(.52)}{2.28} \cdot STR$$

Note[2] that $SE(\widehat{\beta_0}) = 10.4$ & $SE(\widehat{\beta_1}) = .52$.

[2]We will discuss how they are calculated soon.

Estimation in Stata

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(.52)}{2.28} \cdot STR$$

Estimation in Stata

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(.52)}{2.28} \cdot STR$$

Of course, we can use these SEs to test hypotheses just as before. For example, suppose you want to test

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

Estimation in Stata

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(.52)}{2.28} \cdot STR$$

Of course, we can use these SEs to test hypotheses just as before. For example, suppose you want to test

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

t-stat $= \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)} = \frac{-2.28-0}{.52} = -4.39 \Rightarrow p$-value $= 2 \cdot \Phi(-4.39) \approx 0$

Estimation in Stata

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(.52)}{2.28} \cdot STR$$

Of course, we can use these SEs to test hypotheses just as before. For example, suppose you want to test

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

t-stat $= \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)} = \frac{-2.28-0}{.52} = -4.39 \Rightarrow p$-value $= 2 \cdot \Phi(-4.39) \approx 0$

So we reject the null (at any significance level).

Estimation in Stata

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(.52)}{2.28} \cdot STR$$

Of course, we can use these SEs to test hypotheses just as before. For example, suppose you want to test

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

t-stat $= \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)} = \frac{-2.28-0}{.52} = -4.39 \Rightarrow p$-value $= 2 \cdot \Phi(-4.39) \approx 0$

So we reject the null (at any significance level).

Alternatively, a 95% CI for β_1 is simply

$\widehat{\beta}_1 \pm 1.96 \cdot SE(\widehat{\beta}_1) = -2.28 \pm 1.02 = (-3.3, -1.26)$ (reject null)

The OLS Assumptions

- So why should we have faith in the OLS methodology?

The OLS Assumptions

- So why should we have faith in the OLS methodology?
- Do the OLS estimators have the same desirable properties that $\overline{X}$ had (unbiasedness, consistency, asymptotic normality, efficiency)?

The OLS Assumptions

- So why should we have faith in the OLS methodology?
- Do the OLS estimators have the same desirable properties that $\overline{X}$ had (unbiasedness, consistency, asymptotic normality, efficiency)?
- The answer is yes,

The OLS Assumptions

- So why should we have faith in the OLS methodology?
- Do the OLS estimators have the same desirable properties that $\overline{X}$ had (unbiasedness, consistency, asymptotic normality, efficiency)?
- The answer is yes, ... pending certain assumptions.
- The following three assumptions are enough to give us unbiasedness, consistency and asymptotic normality (which will let us build confidence intervals and conduct hypothesis tests).
- Efficiency will require an additional assumption that we'll discuss later.

The OLS Assumptions

The four assumptions of OLS are

OLS Assumption 1 Linearity (in parameters)

$y = \beta_0 + \beta_1 x + u$ is the data generating process

The OLS Assumptions

The four assumptions of OLS are

OLS Assumption 1 Linearity (in parameters)

$y = \beta_0 + \beta_1 x + u$ is the data generating process

OLS Assumption 2 Simple random sample

(X_i, Y_i) are *iid* draws from their joint distribution

The OLS Assumptions

The four assumptions of OLS are

OLS Assumption 1 Linearity (in parameters)

$y = \beta_0 + \beta_1 x + u$ is the data generating process

OLS Assumption 2 Simple random sample

(X_i, Y_i) are *iid* draws from their joint distribution

OLS Assumption 3 Some variation in X

$x_i, i = 1, 2, \ldots, n$ are not all identical values

The OLS Assumptions

The four assumptions of OLS are

OLS Assumption 1 Linearity (in parameters)

$y = \beta_0 + \beta_1 x + u$ is the data generating process

OLS Assumption 2 Simple random sample

(X_i, Y_i) are *iid* draws from their joint distribution

OLS Assumption 3 Some variation in X

$x_i, i = 1, 2, \ldots, n$ are not all identical values

OLS Assumption 4 Zero Conditional Mean

$\mathbb{E}(u|X) = 0$

The OLS Assumptions

The four assumptions of OLS are

OLS Assumption 1 Linearity (in parameters)

$y = \beta_0 + \beta_1 x + u$ is the data generating process

OLS Assumption 2 Simple random sample

(X_i, Y_i) are *iid* draws from their joint distribution

OLS Assumption 3 Some variation in X

$x_i, i = 1, 2, \ldots, n$ are not all identical values

OLS Assumption 4 Zero Conditional Mean

$\mathbb{E}(u|X) = 0$

OLS Assumption 5 Homoskedasticity

$Var(u|X) = \sigma^2$

OLS Assumption 1

We assume the 'true' relationship between Y and X is given by the population regression function:

$$y = \beta_0 + \beta_1 x + u$$

OLS Assumption 1

We assume the ‘true’ relationship between Y and X is given by the population regression function:

$$y = \beta_0 + \beta_1 x + u$$

This means we have identified the proper relationship between our economic variables and it is linear in parameters.

OLS Assumption 1

We assume the 'true' relationship between Y and X is given by the population regression function:

$$y = \beta_0 + \beta_1 x + u$$

This means we have identified the proper relationship between our economic variables and it is linear in parameters.

What about $y = \beta_0 * x^{\beta 1} u$? Does this fit our assumption?

OLS Assumption 1

We assume the ‘true’ relationship between Y and X is given by the population regression function:

$$y = \beta_0 + \beta_1 x + u$$

This means we have identified the proper relationship between our economic variables and it is linear in parameters.

What about $y = \beta_0 * x^{\beta 1} u$? Does this fit our assumption?

No ...

OLS Assumption 1

We assume the 'true' relationship between Y and X is given by the population regression function:

$$y = \beta_0 + \beta_1 x + u$$

This means we have identified the proper relationship between our economic variables and it is linear in parameters.

What about $y = \beta_0 * x^{\beta 1} u$? Does this fit our assumption?

No ... but we could work with it a bit to get something that could.

$$ln(y) = ln(\beta_0 * x^{\beta 1} u) = ln(\beta_0) + \beta_1 ln(x) + ln(u)$$

OLS Assumption 1

We assume the 'true' relationship between Y and X is given by the population regression function:

$$y = \beta_0 + \beta_1 x + u$$

This means we have identified the proper relationship between our economic variables and it is linear in parameters.

What about $y = \beta_0 * x^{\beta 1} u$? Does this fit our assumption?

No ... but we could work with it a bit to get something that could.

$$ln(y) = ln(\beta_0 * x^{\beta 1} u) = ln(\beta_0) + \beta_1 ln(x) + ln(u)$$

So, y and x may not have a linear relationship, but $ln(y)$ and $ln(x)$ might.

OLS Assumption 2

OLS Assumption 2 Simple random sample

$$(X_i, Y_i) \text{ are } \textit{iid} \text{ draws from their joint distribution}$$

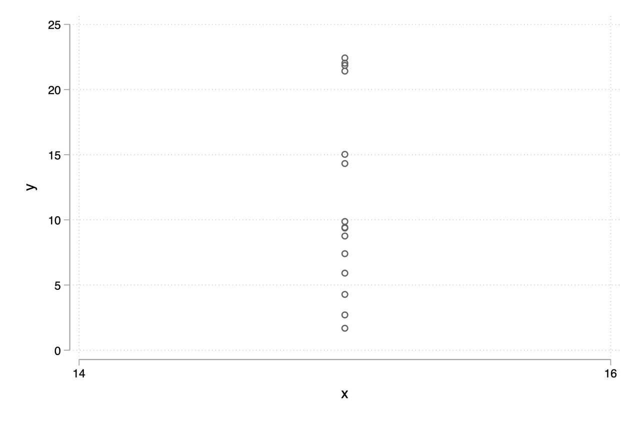
- Intuition: You have a random sample!
- OLS Assumption 2 is likely to hold in cross-sections, but is often violated in time series data.

OLS Assumption 3

OLS Assumption 3 Some variation in X

OLS Assumption 3

OLS Assumption 3 Some variation in X



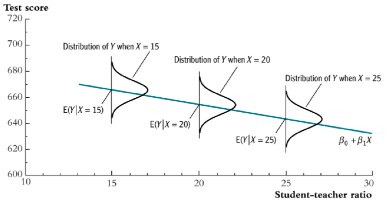
In this case, we cannot identify β_0 from β_1!

OLS Assumption 4

$$\mathbb{E}\left(u_i \mid X_i\right) = 0$$

- The conditional distribution of u_i given X_i has mean 0

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 $\beta_0 + \beta_1 X$. 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.

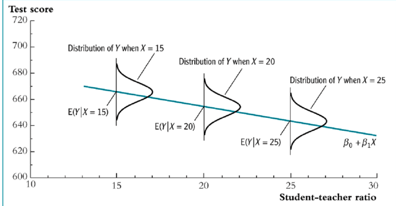
- Intuition: Given $X_i = 15$ (or any other value), the mean of the distribution of u_i is 0.
- In Figure 4.4 we can see that this means the conditional distribution is centered around the population regression line.

OLS Assumption 4

$$\mathbb{E}\left(u_i \mid X_i\right) = 0$$

- The conditional distribution of u_i given X_i has mean 0

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 $\beta_0 + \beta_1 X$. 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.

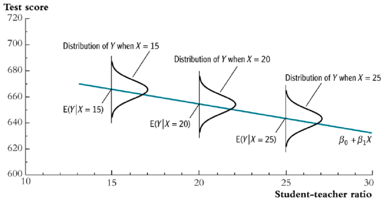
- Intuition: Given $X_i = 15$ (or any other value), the mean of the distribution of u_i is 0.
- In Figure 4.4 we can see that this means the conditional distribution is centered around the population regression line.

OLS Assumption 5

$$Var\left(u_i \mid X_i\right) = \sigma^2 \quad \forall \quad i$$

- The conditional distribution of u_i given X_i has variance that does not depend on the value of x

FIGURE 4.4 **The Conditional Probability Distributions and the Population Regression Line**



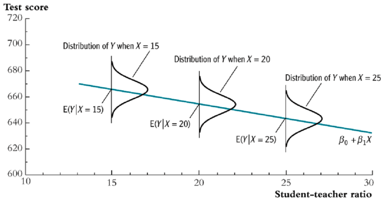
The figure shows the conditional probability of test scores for districts with class sizes of 15, 20, and 25 students. The mean of the conditional distribution of test scores, given the student-teacher ratio, $E(Y|X)$, is the population regression line $\beta_0 + \beta_1 X$. At a given value of X, Y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

OLS Assumption 5

$$Var\left(u_i \mid X_i\right) = \sigma^2 \quad \forall \quad i$$

- The conditional distribution of u_i given X_i has variance that does not depend on the value of x

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 $\beta_0 + \beta_1 X$. 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.

- Note that the variance does not change when $X = 15$ compared to when $X = 25$
- This assumption is not important for unbiasedness or consistency of the OLS estimator, but it will matter for efficiency (more on this shortly).

The OLS Assumptions

- In fact, a central purpose of these OLS assumptions is to allow us to derive these distributions (which turn out to be normal).
    - This will allow us to construct CIs and test hypotheses just like we did for μ.

The OLS Assumptions

- In fact, a central purpose of these OLS assumptions is to allow us to derive these distributions (which turn out to be normal).
    - This will allow us to construct CIs and test hypotheses just like we did for μ.
- They can also tell us that OLS is the BEST option when choosing among an array of different estimators.

The OLS Assumptions

- In fact, a central purpose of these OLS assumptions is to allow us to derive these distributions (which turn out to be normal).
  - This will allow us to construct CIs and test hypotheses just like we did for μ.
- They can also tell us that OLS is the BEST option when choosing among an array of different estimators.
- The flip side of the role of the OLS assumptions is to highlight situations in which OLS regressions might run into trouble.
  - Much of the second half of the course is focused on addressing these situations.

OLS Estimator Univariate, II

What's so great about OLS anyway???

- We have many options for how to pick $\hat{\beta}_0, \hat{\beta}_1$. Why do we care so much about OLS?

What's so great about OLS anyway???

- We have many options for how to pick $\hat{\beta}_0, \hat{\beta}_1$. Why do we care so much about OLS?
- OLS has a closed-form solution, so easy to calculate

What's so great about OLS anyway???

- We have many options for how to pick $\hat{\beta}_0, \hat{\beta}_1$. Why do we care so much about OLS?
- OLS has a closed-form solution, so easy to calculate ... less important with modern computing tools.

What's so great about OLS anyway???

- We have many options for how to pick $\hat{\beta}_0, \hat{\beta}_1$. Why do we care so much about OLS?
- OLS has a closed-form solution, so easy to calculate ... less important with modern computing tools.
- The Gauss-Markov Theorem!

What's so great about OLS anyway???

- We have many options for how to pick $\hat{\beta}_0, \hat{\beta}_1$. Why do we care so much about OLS?
- OLS has a closed-form solution, so easy to calculate ... less important with modern computing tools.
- The Gauss-Markov Theorem!
  - If our 5 assumptions hold...

What's so great about OLS anyway???

- We have many options for how to pick $\hat{\beta}_0, \hat{\beta}_1$. Why do we care so much about OLS?
- OLS has a closed-form solution, so easy to calculate ... less important with modern computing tools.
- The Gauss-Markov Theorem!
  - If our 5 assumptions hold...
  - then the least-squares estimator is the **B**est **L**inear **U**nbiased **E**estimator (BLUE)

What's so great about OLS anyway???

- We have many options for how to pick $\hat{\beta}_0, \hat{\beta}_1$. Why do we care so much about OLS?
- OLS has a closed-form solution, so easy to calculate ... less important with modern computing tools.
- The Gauss-Markov Theorem!
  - If our 5 assumptions hold...
  - then the least-squares estimator is the **B**est **L**inear **U**nbiased **E**estimator (BLUE)
- This means that (under the assumptions) no other unbiased estimator will have a lower variance for a given sample size than OLS - thus, OLS is efficient.

What's so great about OLS anyway???

- We have many options for how to pick $\hat{\beta}_0, \hat{\beta}_1$. Why do we care so much about OLS?
- OLS has a closed-form solution, so easy to calculate ... less important with modern computing tools.
- The Gauss-Markov Theorem!
  - If our 5 assumptions hold...
  - then the least-squares estimator is the **B**est **L**inear **U**nbiased **E**estimator (BLUE)
- This means that (under the assumptions) no other unbiased estimator will have a lower variance for a given sample size than OLS - thus, OLS is efficient.

- *Note that relaxing homoskedasticity and/or independence of observations does not make our estimator biased, but these situations mean our OLS estimator may not have the lowest possible variance anymore.*

Regression When X is a Binary Variable

- So far, we have only looked at examples where the regressor (X) is a "continuous" variable (e.g. dosage, class size).
- Regression Analysis can also be used when X is a binary or *dummy* variable (i.e. can only take on the values 0 and 1).
  - gender, drug treatment, democrat...
- Although the coefficients are calculated in exactly the same way when X is binary, the interpretation of β_1 differs.
  - Why? Because a regression with a binary regressor is equivalent to performing a difference of means analysis.
    - So β_1 isn't really a slope anymore...

Regression When X is a Binary Variable

- For example, suppose we look at the CPS data on earnings. Let's focus only on 1998.

TABLE 3.1 Hourly Earnings in the United States of Working College Graduates, Aged 25–34: Selected Statistics from the Current Population Survey, in 1998 Dollars

	Men			Women			Difference, Men vs. Women		
Year	$\overline{Y}_m$	s_m	n_m	$\overline{Y}_w$	s_w	n_w	$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	95% Confidence Interval for d
1992	17.57	7.50	1591	15.22	5.97	1371	2.35**	0.25	1.87–2.84
1994	16.93	7.39	1598	15.01	6.41	1358	1.92**	0.25	1.42–2.42
1996	16.88	7.29	1374	14.42	6.07	1235	2.46**	0.26	1.94–2.97
1998	17.94	7.86	1393	15.49	6.80	1210	2.45**	0.29	1.89–3.02

These estimates are computed using data on all full-time workers aged 25–34 from the CPS for the indicated years. The difference is significantly different from zero at the *5% or **1% significance level.

Regression When X is a Binary Variable

- Let Y_i be average hourly earnings in 1998 and D_i equal 1 if the worker is male and 0 if the worker is female.
- The population regression model with D_i as the regressor is

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

- Since D_i is not continuous, we can't really think of β_1 as a slope (because D_i only takes on 2 values, there's no "line").
- For this reason, we just call β_1 the coefficient on D_i, instead of the slope.
- So how do we interpret β_1 if it's not a slope? Let's look at what we have for each value of D_i.

Regression When X is a Binary Variable

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

Regression When X is a Binary Variable

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

- When $D_i = 0$ (the worker is female)

$$Y_i = \beta_0 + \beta_1 \cdot 0 + u_i = \beta_0 + u_i$$

Regression When X is a Binary Variable

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

- When $D_i = 0$ (the worker is female)

$$Y_i = \beta_0 + \beta_1 \cdot 0 + u_i = \beta_0 + u_i$$

- Since $E\left(Y_i \mid D_i = 0\right) = \beta_0$, β_0 is the population mean value of earnings for women.

Regression When X is a Binary Variable

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

- When $D_i = 0$ (the worker is female)

$$Y_i = \beta_0 + \beta_1 \cdot 0 + u_i = \beta_0 + u_i$$

- Since $E\left(Y_i \mid D_i = 0\right) = \beta_0$, β_0 is the population mean value of earnings for women.
- Whereas when $D_i = 1$ (the worker is male)

$$Y_i = \beta_0 + \beta_1 \cdot 1 + u_i = \beta_0 + \beta_1 + u_i$$

Regression When X is a Binary Variable

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

- When $D_i = 0$ (the worker is female)

$$Y_i = \beta_0 + \beta_1 \cdot 0 + u_i = \beta_0 + u_i$$

- Since $E\left(Y_i \mid D_i = 0\right) = \beta_0$, β_0 is the population mean value of earnings for women.
- Whereas when $D_i = 1$ (the worker is male)

$$Y_i = \beta_0 + \beta_1 \cdot 1 + u_i = \beta_0 + \beta_1 + u_i$$

- So $E\left(Y_i \mid D_i = 1\right) = \beta_0 + \beta_1$, the population mean value of earnings for men.
- β_1 is then the difference between the two population means.

Regression When X is a Binary Variable

$$Earnings_i = \beta_0 + \beta_1 Male_i + u_i$$

- Here's the result of the regression above using the 1998 data:

```
. reg earnings male if year == 1998, robust

Regression with robust standard errors                 Number of obs =    2603
                                                       F(  1,  2601) =   72.79
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0267
                                                       Root MSE      =  7.3876

------------------------------------------------------------------------------
             |               Robust
    earnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   2.451918    .287392     8.53   0.000     1.888378    3.015458
       _cons |   15.49195   .1954935    79.25   0.000     15.10861    15.87529
------------------------------------------------------------------------------
```

$$\widehat{Earnings} = \underset{(.20)}{15.49} + \underset{(0.29)}{2.45} \cdot Male$$

Year	$\overline{Y}_m$	s_m	n_m	$\overline{Y}_w$	s_w	n_w	$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	95% Confidence Interval for d
1998	17.94	7.86	1393	15.49	6.80	1210	2.45**	0.29	1.89–3.02

$$\widehat{Earnings} = \underset{(.20)}{15.49} + \underset{(0.29)}{2.45} \cdot Male$$

- $\widehat{\beta}_0 = 15.49$ is the average value of earnings for women.
- $\widehat{\beta}_0 + \widehat{\beta}_1 = 17.94$ is the average value of earnings for men.
- $\widehat{\beta}_1 = 2.45$ is the difference between the two sample averages.

Year	$\overline{Y}_m$	s_m	n_m	$\overline{Y}_w$	s_w	n_w	$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	95% Confidence Interval for d
1998	17.94	7.86	1393	15.49	6.80	1210	2.45**	0.29	1.89–3.02

$$\widehat{Earnings} = \underset{(.20)}{15.49} + \underset{(0.29)}{2.45} \cdot Male$$

- $\widehat{\beta}_0 = 15.49$ is the average value of earnings for women.
- $\widehat{\beta}_0 + \widehat{\beta}_1 = 17.94$ is the average value of earnings for men.
- $\widehat{\beta}_1 = 2.45$ is the difference between the two sample averages.
- Recall that, using Table 3.1, a 95% CI for the wage gap is

$$2.45 \pm 1.96 \cdot .29 = (1.89, 3.02)$$

Year	$\overline{Y}_m$	s_m	n_m	$\overline{Y}_w$	s_w	n_w	$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	95% Confidence Interval for d
1998	17.94	7.86	1393	15.49	6.80	1210	2.45**	0.29	1.89–3.02

$$\widehat{Earnings} = \underset{(.20)}{15.49} + \underset{(0.29)}{2.45} \cdot Male$$

- $\widehat{\beta}_0 = 15.49$ is the average value of earnings for women.
- $\widehat{\beta}_0 + \widehat{\beta}_1 = 17.94$ is the average value of earnings for men.
- $\widehat{\beta}_1 = 2.45$ is the difference between the two sample averages.
- Recall that, using Table 3.1, a 95% CI for the wage gap is

$$2.45 \pm 1.96 \cdot .29 = (1.89, 3.02)$$

- What is the 95% CI for β_1?

$$\widehat{\beta}_1 \pm 2.58 \cdot SE\left(\widehat{\beta}_1\right) = 2.45 \pm 1.96 \cdot .29 = (1.89, 3.02)$$

Regression When X is a Binary Variable

$$\widehat{Earnings} = \underset{(.20)}{15.49} + \underset{(0.29)}{2.45} \cdot Male$$

- We can test the hypothesis $H_0 : \beta_1 = 0 \;\; H_A : \beta_1 \neq 0$ by calculating the t-statistic

$$t^{act} = \frac{\widehat{\beta}_1 - 0}{SE\left(\widehat{\beta}_1\right)} = \frac{2.45}{0.29} = 8.45$$

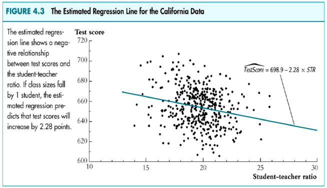
  and then calculating the *p-value*

$$p\text{-}value = 2\Phi\left(-\left|t^{act}\right|\right) \approx 0$$

- We can reject the null hypothesis at any positive level of significance (just as before).

Goodness of Fit

- So we've learned how to estimate β_0 & β_1 and how to test hypotheses and build CI's using these estimates.
- But how "good" is our regression?
- In other words, how close is the line to the actual data?



- Or, more precisely, how much of the variation in Y is our regression explaining?
- Can we measure how much better the regression does at estimating Y than just using $\overline{Y}$?
- We need to measure how close we are getting to the data...

Goodness of Fit

- It doesn't make sense to simply report the average of the errors since

$$\overline{\widehat{u}} = \sum \widehat{u}_i = 0$$

by construction: since

$$\widehat{u}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$$

we have

$$\sum \widehat{u}_i = n\overline{Y} - n\widehat{\beta}_0 - n\widehat{\beta}_1 \overline{X} = n\left(\overline{Y} - \widehat{\beta}_0 - \widehat{\beta}_1 \overline{X}\right) = 0$$

where we've used the OLS formula for $\widehat{\beta}_0 : \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$

- However, we can compute the average *squared* error.

Goodness of Fit

- The *Standard Error of the Regression* (*SER*) is an estimator of the standard deviation of u_i (note that $\left(\hat{u}_i - \overline{\hat{u}}\right)^2 = (\hat{u}_i)^2$)[3]

$$SER = s_{\hat{u}} \text{ where } s_{\hat{u}}^2 = \frac{1}{n-2}\sum \hat{u}_i^2 = \frac{1}{n-2}\sum \left(Y_i - \hat{Y}_i\right)^2 = \frac{SSR}{n-2}$$

- *SSR* stands for the *Sum of Squared Residuals*, which you should recognize as what the OLS procedure is minimizing
- But the *SER* depends on the scale of Y_i (\$, millions of \$).
- As always, we would like a *normalized* measure (like correlation).

[3] We proved that $\overline{\hat{u}} = 0$ on the previous slide.

Goodness of Fit

- We can normalize the *SER* using a measure of the *total* variation in Y, called the *Total Sum of Squares*:
$$TSS = \sum \left(Y_i - \overline{Y}\right)^2$$
- However, instead of focusing on what we **aren't** explaining, it makes more sense to focus on what we **are** explaining.
- The normalized measure we use is called the R^2.

Goodness of Fit: R-squared

- The R^2 is then just the *percentage of the total variation in Y "explained" by the estimated regression*:

$$R^2 = \frac{\sum\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} = \frac{ESS}{TSS} = \frac{\text{"explained variation"}}{\text{"total variation"}}$$

- *ESS* stands for the *Explained Sum of Squares*.
- Since $TSS = ESS + SSR$, we can also show that

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\text{"unexplained variation"}}{\text{"total variation"}}$$

where

$$\frac{\text{"unexplained variation"}}{\text{"total variation"}} = \frac{\sum\left(Y_i - \widehat{Y}_i\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2}$$

- Note that $0 \leq R^2 \leq 1$

Goodness of Fit: R-squared

- $R^2 = 1$ is a perfect fit (all the data points are on the regression line).
- $R^2 = 0$ means you are explaining none of the variation in Y (so your best guess for any Y_i is just the sample mean $\overline{Y}$).

Goodness of Fit: R-squared

- $R^2 = 1$ is a perfect fit (all the data points are on the regression line).
- $R^2 = 0$ means you are explaining none of the variation in Y (so your best guess for any Y_i is just the sample mean $\overline{Y}$).
- Looking at the Test Score example again, we find

```
. reg 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
------------------------------------------------------------------------------
```

Goodness of Fit: R-squared

- $R^2 = 1$ is a perfect fit (all the data points are on the regression line).
- $R^2 = 0$ means you are explaining none of the variation in Y (so your best guess for any Y_i is just the sample mean $\overline{Y}$).
- Looking at the Test Score example again, we find

```
. reg 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
------------------------------------------------------------------------------
```

- Here we see that we are only explaining about 5% of the variation in test scores with our regression.

R-squared & Correlation

```
. reg 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
------------------------------------------------------------------------------
```

- It turns out that there is a close link between R^2 in the univariate regression model and the sample correlation coefficient $r_{XY} = \frac{s_{XY}}{s_X s_Y}$
- R^2 is a measure of the fit of the linear model.
- The sample correlation (r_{XY}) is a measure of the linear relationship between two variables.

R-squared & Correlation

- In fact[4], $R^2 = r_{XY}^2$, which is where it got the name!
- In the Test Score data, for example, $r_{XY} = -.226$.

```
. pwcorr str testscr

             |      str  testscr
-------------+------------------
         str |   1.0000
     testscr |  -0.2264   1.0000
```

- We can see that $(-.226)^2 = .051$, which is the R^2 from the Test Score regression!
- This is useful to know since it gives us some idea of what a high or low R^2 should "look like".

[4] You can prove this using the definitions of R^2 and r_{XY}^2

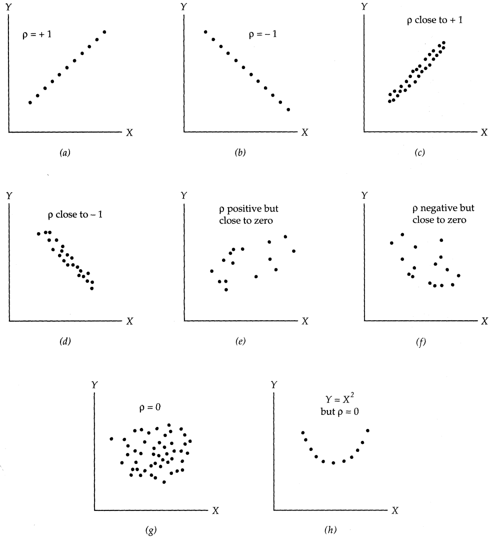
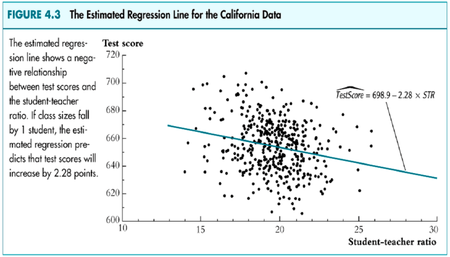


FIGURE 2-7
Some typical patterns of the correlation coefficient, ρ.

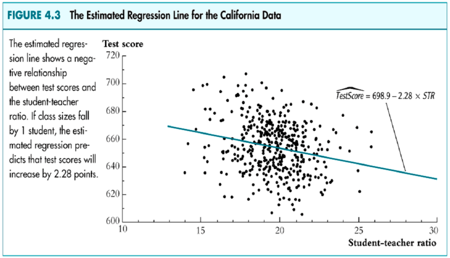
R-Squared Example



$$\widehat{TS} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \cdot STR, \quad R^2 = .05$$

- So this is what a "low" R^2 looks like.

R-Squared Example



$$\widehat{TS} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \cdot STR, \quad R^2 = .05$$

- So this is what a "low" R^2 looks like.
- Let's compare the Test Score example to a different application.

R-Squared Example

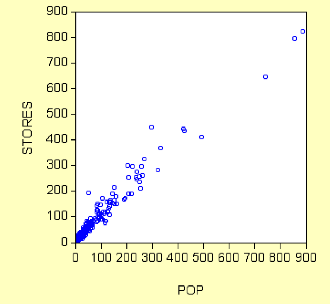
- We have data on the number of supermarkets (*Stores*) and population (*Pop*) for 320 Metropolitan Statistical Areas (MSAs).

R-Squared Example

- We have data on the number of supermarkets (*Stores*) and population (*Pop*) for 320 Metropolitan Statistical Areas (MSAs).
- *Stores* is in units of supermarket stores, *Pop* is in units of 10,000 people.

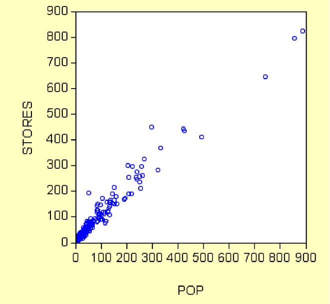
R-Squared Example

- We have data on the number of supermarkets (*Stores*) and population (*Pop*) for 320 Metropolitan Statistical Areas (MSAs).
- *Stores* is in units of supermarket stores, *Pop* is in units of 10,000 people.
- Here is a scatterplot of *Stores* versus *Pop*.



R-Squared Example

- We have data on the number of supermarkets (*Stores*) and population (*Pop*) for 320 Metropolitan Statistical Areas (MSAs).
- *Stores* is in units of supermarket stores, *Pop* is in units of 10,000 people.
- Here is a scatterplot of *Stores* versus *Pop*.



- It looks like a one to one relationship: about one store for every 10,000 people or so.

R-Squared Example

- How much of the variation in *Stores* do you think can be explained by variation in *Pop*?

```
. reg stores pop, robust

Regression with robust standard errors                 Number of obs =     320
                                                       F(  1,   318) = 1362.27
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.9591
                                                       Root MSE      =  20.599

------------------------------------------------------------------------------
             |               Robust
      stores |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pop |   .9596473   .0260004    36.91   0.000     .9084927    1.010802
       _cons |   10.10495   1.252446     8.07   0.000     7.640823    12.56908
------------------------------------------------------------------------------
```

An Hypothesis Test

$$\widehat{Stores} = \underset{(1.25)}{10.1} + \underset{(0.026)}{.96} \cdot Pop, \quad R^2 = .96$$

An Hypothesis Test

$$\widehat{Stores} = \underset{(1.25)}{10.1} + \underset{(0.026)}{.96} \cdot Pop, \quad R^2 = .96$$

- Can we test my claim about a one to one relationship?

$$\begin{aligned} H_0 &: \beta_1 = 1 \\ H_A &: \beta_A \neq 1 \end{aligned}$$

1. $SE(\beta_1) = .026$
2. $t = \frac{.96-1}{.026} = -1.55$
3. p-value $= 2\Phi(-1.55) = .12$

- So we can't reject the null.

Some Caveats

- A high R^2 means that a lot of the total variation is explained by the regression (data is tightly concentrated around the line).

Some Caveats

- A high R^2 means that a lot of the total variation is explained by the regression (data is tightly concentrated around the line).
- But R^2 does **not** tell you about the statistical significance of the coefficients (for this you need SEs).

Some Caveats

- A high R^2 means that a lot of the total variation is explained by the regression (data is tightly concentrated around the line).
- But R^2 does **not** tell you about the statistical significance of the coefficients (for this you need SEs).
- R^2 also does not prove whether our model is right or wrong: you can have a good model but a low R^2 because $Var(u_i)$ is large.

Some Caveats

- A high R^2 means that a lot of the total variation is explained by the regression (data is tightly concentrated around the line).
- But R^2 does **not** tell you about the statistical significance of the coefficients (for this you need SEs).
- R^2 also does not prove whether our model is right or wrong: you can have a good model but a low R^2 because $Var(u_i)$ is large.
- You can also have a bad model with $R^2 \approx 1$
  - Spurious corrlelation/regression: X & Y move together because of something else.
    - Regress the number of supermarkets on the number of cars (or video stores)...
    - Regress the GDP of Sweden on the GDP of Italy...