KU

The Simple Regression Model

Caio Vigo

The University of Kansas
Department of Economics

Spring 2020

These slides were based on *Introductory Econometrics* by Jeffrey M. Wooldridge (2015)

Topics

Definition of the Simple Regression Model

- **What type of analysis will we do?** Cross-sectional analysis
- **First step:** Clearly define what is your population (in what you are interested to study).
- **Second Step:** There are two variables, x and y, and we would like to "study how y varies with changes in x."
- **Third Step:** We assume we can collect a random sample from the population of interest.

Now we will learn to write our first econometric model, derive an estimator (**what's an estimator again?**) and use this estimator in our sample.

Introduction

We must confront three issues:

1. How do we allow factors other than x to affect y? There is never an exact relationship between two variables.
2. What is the functional relationship between y and x?
3. How can we be sure we are capturing a *ceteris paribus* relationship between y and x?

Introduction

Consider the following equation relating y to x:

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

which is assumed to hold in the population of interest.

- This equation defines the **simple linear regression model** (or *two-variable regression model, or bivariate linear regression model*).

Introduction

- y and x are not treated symmetrically. We want to explain y in terms of x.

x explains y

$$x \longrightarrow y$$

- **Example:**

size of the city x, **explains** number of crimes (y) **(not the other way around)**.

Terminology for Simple Regression

y	**x**
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Regressand	Regressor

The error term

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

This equation explicitly allows for other factors, contained in u, to affect y.

This equation also addresses the functional form issue (in a simple way).

Namely, y is assumed to be *linearly* related to x.

We call β_0 the **intercept parameter** and β_1 the **slope parameter**. These describe a population, and our ultimate goal is to estimate them.

The simple linear regression model equation

• The equation also addresses the *ceteris paribus* issue. In

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

all other factors that affect y are in u. We want to know how y changes when x changes, *holding* u *fixed*.

• Let Δ denote "change."Then holding u fixed means $\Delta u = 0$. So

$$\begin{aligned} \Delta y &= \beta_1 \Delta x + \Delta u \\ &= \beta_1 \Delta x \qquad \text{when } \Delta u = 0. \end{aligned}$$

• This equation effectively defines β_1 as a slope, with the only difference being the restriction $\Delta u = 0$.

The simple linear regression model equation

Example: Yield and Fertilizer

- A model to explain crop yield to fertilizer use is

$$yield = \beta_0 + \beta_1 fertilizer + u,$$

where u contains land quality, rainfall on a plot of land, and so on. The slope parameter, β_1, is of primary interest: it tells us how $yield$ changes when the amount of fertilizer changes, holding all else fixed.

The simple linear regression model equation

Example: Wage and Education

$$wage = \beta_0 + \beta_1 educ + u$$

where u contains somewhat nebulous factors ("ability") but also past workforce experience and tenure on the current job.

$$\Delta wage = \beta_1 \Delta educ \quad \text{when } \Delta u = 0$$

The simple linear regression model equation

We said we must confront three issues:

1. How do we allow factors other than x to affect y?
Answer: u

2. What is the functional relationship between y and x?
Answer: Linear (x has a linear effect on y)

3. How can we be sure we a capturing a ceteris paribus relationship between y and x?
Answer: Related with $\Delta u = 0$

- We have argued that the simple regression model

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

addresses each of them.

Relation between u and x

To estimate β_1 and β_0 from a random sample we also need to **restrict how u and x are related to each other.**

- Recall that x and u are properly viewed as having distributions in the population.

- What we must do is restrict the way in when u and x relate to each other in the **population**.

- First, we make a simplifying assumption that is without loss of generality: the average, or expected, value of u is zero in the population:

$$E(u) = 0$$

Relation between u and x

- Normalizing u should cause no impact in the most important parameter: β_1

- The presence of β_0 in

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

allows us to assume $E(u) = 0$.

- If the average of u is different from zero, we just adjust the intercept, leaving the slope the same. If $\alpha_0 = E(u)$ then we can write

$$y = (\beta_0 + \alpha_0) + \beta_1 x + (u - \alpha_0),$$

where the new error, $u - \alpha_0$, has a zero mean.

Relation between u and x

We need to restrict the dependence between u and x

- **Option 1:** Uncorrelated

We could assume u and x **uncorrelated** in the population:

$$Corr(x, u) = 0$$

It implies **only** that u and x are not **linearly** related. **Not good enough**.

Relation between u and x

- **Option 2:** Mean independence

The mean of the error (i.e., the mean of the unobservables) is the same across all slices of the population determined by values of x.

We represent it by:

$$E(u|x) = E(u), \text{ all values } x,$$

And we say that u is **mean independent** of x

Relation between u and x

- Suppose u is "ability" and x is years of education. We need, for example,

$$E(ability|x=8) = E(ability|x=12) = E(ability|x=16)$$

so that the average ability is the same in the different portions of the population with an 8^{th} grade education, a 12^{th} grade education, and a four-year college education.

Relation between u and x

- Combining $E(u|x) = E(u)$ (the substantive assumption) with $E(u) = 0$ (a normalization) gives

$$E(u|x) = 0, \text{ all values } x$$

- Called the **zero conditional mean assumption**.

The PRF

- First, recall the properties of conditional expectation. *(see slides with a review of Probability)*

- Now, take the conditional expectation of our *Simple Linear Regression Function*. Then, we get:

$$\begin{aligned} E(y|x) &= \beta_0 + \beta_1 x + E(u|x) \\ &= \beta_0 + \beta_1 x \end{aligned}$$

which shows the **population regression function** is a linear function of x.

The PRF

Figure: **Example:** The goal is to explain **weekly consumption expenditure** in terms of **weekly income**

$Y\downarrow$ \ $X\rightarrow$	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y, $	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of Y, $E(Y\|X)$	65	77	89	101	113	125	137	149	161	173

Source: Gujarati, Damodar (2002). Basic Econometrics.

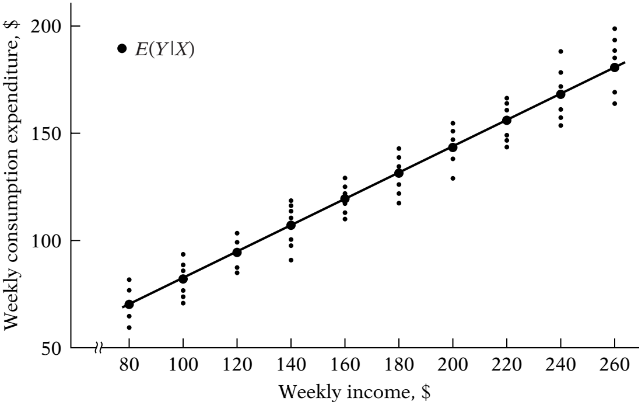
The PRF

Figure: Conditional Probabilities of the data

$p(Y \mid X_i) \downarrow$ \ $X \rightarrow$	80	100	120	140	160	180	200	220	240	260
Conditional probabilities $p(Y \mid X_i)$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	–	$\frac{1}{6}$	–	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	–	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	–	–	–	$\frac{1}{7}$	–	–	–	$\frac{1}{7}$	–	$\frac{1}{7}$
Conditional means of Y	65	77	89	101	113	125	137	149	161	173

Source: Gujarati, Damodar (2002). Basic Econometrics.

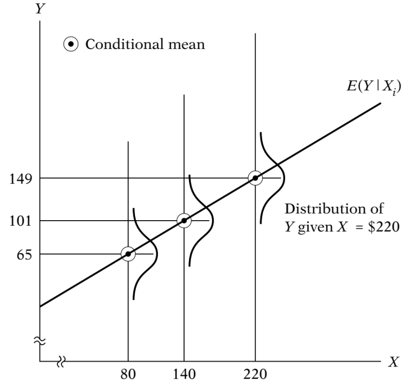
Figure: Conditional distribution of expenditure for various levels of income



Source: Gujarati, Damodar (2002). Basic Econometrics.

The PRF

Figure: The Population Regression Function (PRF)



Source: Gujarati, Damodar (2002). Basic Econometrics.

The PRF

- The straight line in the previous graph is the PRF, $E(y|x) = \beta_0 + \beta_1 x$. The conditional distribution of y at three different values of x are superimposed.

- For a given value of x, we see a range of y values: remember, $y = \beta_0 + \beta_1 x + u$, and u has a distribution in the population.

- In practice, we never know the **population intercept and slope.**

The PRF

- Assuming we know the PRF, consider this example:

Example

- Suppose for the population of students attending a university, we know the PRF:

$$E(\mathit{colGPA}|hsGPA) = 1.5 + 0.5\ hsGPA,$$

- For this example, what is y? what is x? What is the slope? What's the intercept?
- If $hsGPA = 3.6$ what's the expected college GPA? $1.5 + 0.5(3.6) = 3.3$

Topics

Deriving the OLS

- Given data on x and y, how can we estimate the population parameters, β_0 and β_1?

- Let $\{(x_i, y_i) : i = 1, 2, ..., n\}$ be a **random sample** of size n (the number of observations) from the population. Think of this as a random sample.

Deriving the OLS

Derivation: (On white board)

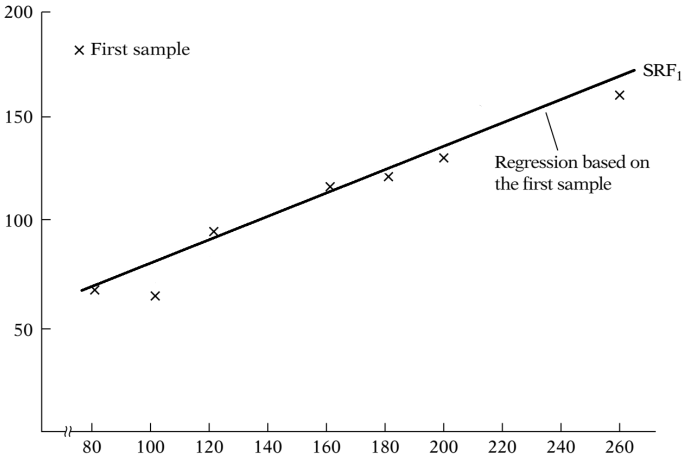
Estimator for β_0

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

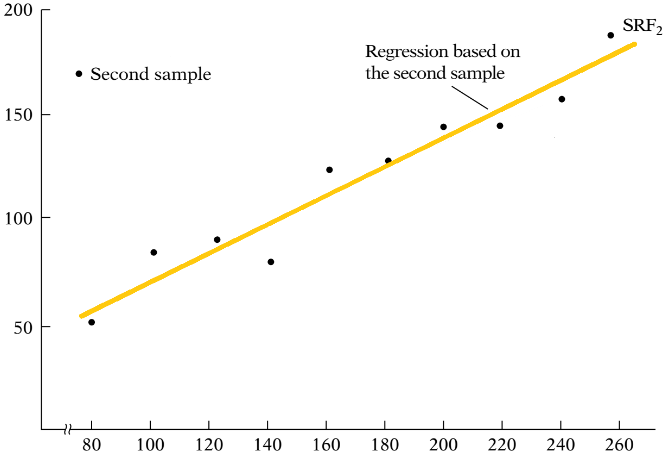
Estimator for β_1

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{Sample Covariance}(x, y)}{\text{Sample Variance}(x)} \\
&= \frac{S_{x,y}}{S_x^2} \\
&= \hat{\rho}_{x,y} \frac{\hat{\sigma}_y}{\hat{\sigma}_x}
\end{aligned}$$

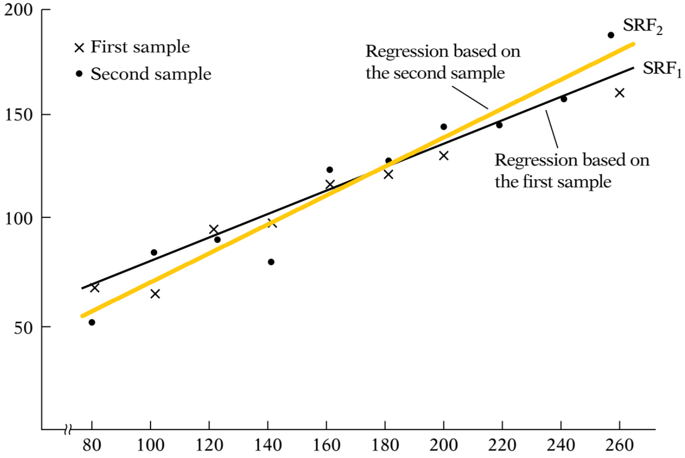
Deriving the OLS



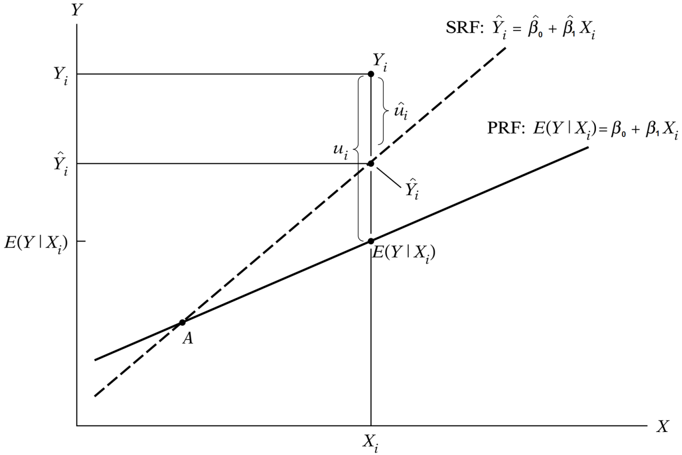
Deriving the OLS



Deriving the OLS



Deriving the OLS



Source: Gujarati, Damodar (2002). Basic Econometrics.

Interpreting the OLS Estimates

Example: Effects of Education on Hourly Wage

- Data: random sample from the US workforce population in 1976.

$wage$: dollars per hour,
$educ$: highest grade completed (years of education).

- The estimated equation is

$$\begin{aligned} \widehat{wage} &= -0.90 + 0.54\, educ \\ n &= 526 \end{aligned}$$

- Each additional year of schooling is estimated to be worth \$0.54.

Interpreting the OLS Estimates

The function

$$\widehat{wage} = -0.90 + 0.54\ educ$$

is the **OLS (or sample) regression line.**

Interpreting the OLS Estimates - R Output


```
> stargazer(regression_wage1, type='text', align=TRUE, digits=2)

===============================================
                        Dependent variable:
                    ---------------------------
                               wage
-----------------------------------------------
educ                         0.54***
                             (0.05)

Constant                      -0.90
                             (0.68)

-----------------------------------------------
Observations                   526
R2                            0.16
Adjusted R2                   0.16
Residual Std. Error      3.38 (df = 524)
F Statistic          103.36*** (df = 1; 524)
===============================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
```

Interpreting the OLS Estimates

- When we write the simple linear regression model,

$$wage = \beta_0 + \beta_1 educ + u,$$

it applies to the population, so we do not know β_0 and β_1.

- $\hat{\beta}_0 = -0.90$ and $\hat{\beta}_1 = 0.54$ are our *estimates* from this particular sample.

- These estimates may or may not be close to the population values. If we obtain another sample, the estimates would almost certainly change.

Interpreting the OLS Estimates

- If $educ = 0$,

the predicted wage is:

$$\widehat{wage} = -0.90 + 0.54(0) = -0.90$$

The predicted value does not fit in reality.

Mainly because when we extrapolate outside the majority range of our data can produce strange predictions.

Interpreting the OLS Estimates

- When $educ = 8$,

the predicted wage is:

$$\widehat{wage} = -0.90 + 0.54(8) = 3.42$$

which we can think of as our estimate of the average wage in the population when $educ = 8$.

Interpreting the OLS Estimates

Sample Regression Line (SRF)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

Also known as:

- OLS Regression Line
- Sample Regression Function
- OLS Regression Function
- Estimated Equation

Interpreting the OLS Estimates

Population Regression Function (PRF)

Since the simple linear regression model (or just econometric model) is:

$$y_i = \beta_0 + \beta_1 x_i + u$$

Then, the PRF is:

$$\Rightarrow E(y_i|x) = \beta_0 + \beta_1 x_i \qquad i = 1, 2, \ldots, n$$

Interpreting the OLS Estimates

Residuals

$$\hat{u}_i = y_i - \hat{y}_i \qquad i = 1, 2, \ldots, n$$

Error Term

$$\begin{aligned} u_i &= y_i - E(y|x) \\ &= y_i - \beta_0 - \beta_1 x_i \qquad i = 1, 2, \ldots, n \end{aligned}$$

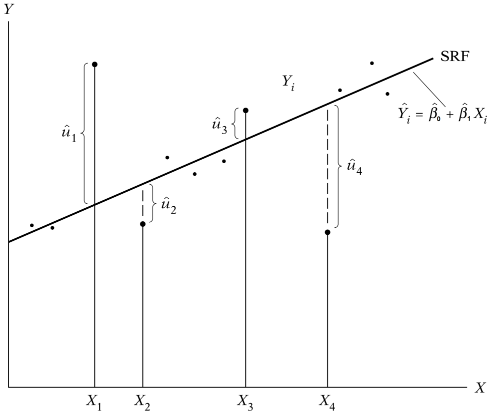
Topics

Properties of OLS on any Sample of Data

- Recall that the OLS residuals are

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \qquad , i = 1, 2, ..., n$$

Properties of OLS on any Sample of Data



Source: Gujarati, Damodar (2002). Basic Econometrics.

Properties of OLS on any Sample of Data

- Some residuals are positive, others are negative.
- If $\hat{u}_i$ is positive $\Rightarrow$ the line underpredicts y_i
- If $\hat{u}_i$ is negative $\Rightarrow$ the line overpredicts y_i

(1) The sum of the OLS residuals is 0

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

Algebraic Properties of OLS Statistics

(2) The sample covariance between the explanatory variables and the residuals is always zero

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

- Therefore the sample correlation between the x and $\hat{u}_i$ is also equal to zero.

- Because the $\hat{y}_i$ are linear functions of the x_i, the fitted values and residuals are uncorrelated, too:

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

Algebraic Properties of OLS Statistics

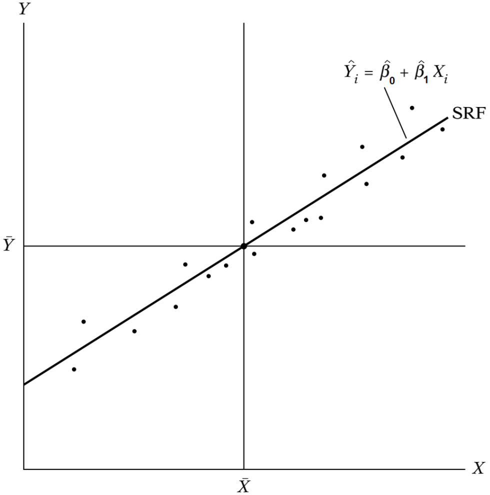
(3) The point $(\bar{x}, \bar{y})$ is always on the OLS regression line.

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

- That is, if we plug in the average for x, we predict the sample average for y.

KU

Goodness-of-Fit



Source: Gujarati, Damodar (2002). Basic Econometrics.

Goodness-of-Fit

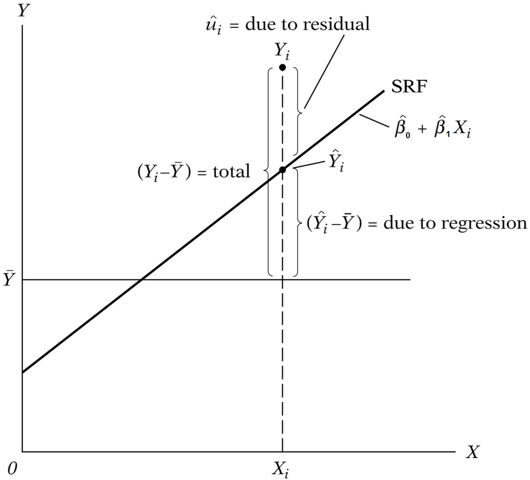
- For each observation, write

$$y_i = \hat{y}_i + \hat{u}_i$$

- Define:

$$\begin{aligned} \text{Total Sum of Squares} &= SST &=& \textstyle\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ \text{Explained Sum of Squares} &= SSE &=& \textstyle\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ \text{Residual Sum of Squares} &= SSR &=& \textstyle\sum_{i=1}^{n}\hat{u}_i^2 \end{aligned}$$

Goodness-of-Fit



Source: Gujarati, Damodar (2002). Basic Econometrics.

(Other names)

- SSR is also know as *Sum of Squared Residuals* or *Model Sum of Residuals*
- $SST = TSS$
- $SSE = ESS$
- $SSR = RSS$

Goodness-of-Fit

$$
\begin{aligned}
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
&= \sum_{i=1}^{n} [\hat{u}_i - (\hat{y}_i - \bar{y})]^2
\end{aligned}
$$

Using the fact that the fitted values and residuals are uncorrelated:

$$SST = SSE + SSR$$

The R-Squared

Goal: We want to evaluate how well the independet variable x explains the dependent variable y.

- We want to obtain the fraction of the sample variation in y that is explained by x.
- We will summarize it in one number: R^2 (or **coefficient of determination**.)

- Assuming $SST > 0$,

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

Goodness-of-Fit

- Since SSE cannot be greater than the SST, then:

$$0 \leq R^2 \leq 1$$

- $R^2 = 0 \Rightarrow$ **No linear relationship** *(between y_i and x_i)*.
- $R^2 = 1 \Rightarrow$ **Perfect linear relationship** (between y_i and x_i).
- As R^2 increases $\Rightarrow$ y_i gets closer and closer to the OLS regression line.

We should not focus only on R^2 to analyze our regression.

Goodness-of-Fit

Example (Wage)

$$\begin{aligned} \widehat{wage} &= -0.90 + 0.54\ educ \\ n &= 526, \qquad R^2 = .16 \end{aligned}$$

- Therefore, years of education explains only about 16% of the variation in hourly wage.

Goodness-of-Fit - R output


```
> summary(regression_wage1)

Call:
lm(formula = wage ~ educ, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.3396 -2.1501 -0.9674  1.1921 16.6085 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.90485    0.68497  -1.321    0.187    
educ         0.54136    0.05325  10.167   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.378 on 524 degrees of freedom
Multiple R-squared:  0.1648,	Adjusted R-squared:  0.1632 
F-statistic: 103.4 on 1 and 524 DF,  p-value: < 2.2e-16
```

Goodness-of-Fit - R output (using *stargazer*)


```
> stargazer(regression_wage1, type='text', align=TRUE, digits=2)

===============================================
                       Dependent variable:
                   ----------------------------
                               wage
-----------------------------------------------
educ                          0.54***
                              (0.05)

Constant                      -0.90
                              (0.68)

-----------------------------------------------
Observations                   526
R2                             0.16
Adjusted R2                    0.16
Residual Std. Error      3.38 (df = 524)
F Statistic          103.36*** (df = 1; 524)
===============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

Suggested Exercise

Suggested Exercise

Below you have a random sample with 10 data points. Your observations are (x_i, y_i). Find the $\hat{\beta}_0$, $\hat{\beta}_1$ and the R^2.

Obs. #	x_i	y_i	x_i	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	x_1	70	80	-41	-90	1681	8100	3690	65.18	4.82	2099.31	23.21
2	x_2	65	100	-46	-70	2116	4900	3220	75.36	-10.36	1269.95	107.40
3	x_3	90	120	-21	-50	441	2500	1050	85.55	4.45	647.93	19.84
4	x_4	95	140	-16	-30	256	900	480	95.73	-0.73	233.26	0.53
5	x_5	110	160	-1	-10	1	100	10	105.91	4.09	25.92	16.74
6	x_6	115	180	4	10	16	100	40	116.09	-1.09	25.92	1.19
7	x_7	120	200	9	30	81	900	270	126.27	-6.27	233.26	39.35
8	x_8	140	220	29	50	841	2500	1450	136.45	3.55	647.93	12.57
9	x_9	155	240	44	70	1936	4900	3080	146.64	8.36	1269.95	69.95
10	x_{10}	150	260	39	90	1521	8100	3510	156.82	-6.82	2099.31	46.49
	Sum	**1,110**	**1,700**	**0.00**	**0.00**	**8,890**	**33,000**	**16,800**	**1,110**	**0.00**	**8,553**	**337**

Topics

The effects of Changing Units of Measurement on OLS Statistics

Example

salary: Annual CEO's salary in thousands of dollars
roe: Average return on equity (measured in percentage)

$$\begin{aligned} \widehat{salary} &= 963.19 + 18.50\, roe \\ n &= 209,\ R^2 = .01 \end{aligned}$$

• A one unit increase in the independent variable (i.e. roe increases one percent) ⇒ increases the predicted salary by 18.501, or **$18,501.**

The effects of Changing Units of Measurement on OLS Statistics

- If we measure roe as a decimal (rather than a percent), what will happen to the intercept, slope, and R^2?

We want:

$$roedec = roe/100$$

- What if we measure salary in dollars (rather than thousands of dollars)? what will happen to the intercept, slope, and R^2?

We want:

$$salarydol = 1,000 \cdot salary$$

The effects of Changing Units of Measurement on OLS Statistics

Changing Units of Measurement

- If the dependent variable y is multiplied by a constant $c \Rightarrow c \cdot \hat{\beta}_0$ and $c \cdot \hat{\beta}_1$

- If the independent variable x is multiplied by a constant $c \Rightarrow \dfrac{1}{c} \cdot \hat{\beta}_1$

In general, changing the units of measurement of only the independent variable does not affect the intercept

The effects of Changing Units of Measurement on OLS Statistics

Example: CEO's salary - Original Regression

$$\begin{aligned}\widehat{salary} &= 963.19 + 18.50\ roe \\ n &= 209,\ R^2 = .01\end{aligned}$$

Example: CEO's salary - *roe* as a decimal

The new regression is:

$$\begin{aligned}\widehat{salary} &= 963.191 + 1,850.1\ roedec \\ n &= 209,\ R^2 = .01\end{aligned}$$

The effects of Changing Units of Measurement on OLS Statistics


```
> roedec<-ceosal1$roe*(1/100)
> regression_ceosal1c <- lm(salary ~ roedec, data = ceosal1)
> stargazer(regression_ceosal1c, type='text', align=TRUE, digits=2)

===============================================
                        Dependent variable:
                    ---------------------------
                              salary
-----------------------------------------------
roedec                       1,850.12*
                            (1,112.33)

Constant                     963.19***
                             (213.24)

-----------------------------------------------
Observations                    209
R2                             0.01
Adjusted R2                    0.01
Residual Std. Error     1,366.55 (df = 207)
F Statistic              2.77* (df = 1; 207)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

The effects of Changing Units of Measurement on OLS Statistics

Example: CEO's salary - Original Regression

$$\begin{aligned} \widehat{salary} &= 963.19 + 18.50\ roe \\ n &= 209,\ \ R^2 = .01 \end{aligned}$$

Example: CEO's salary - *salary* in dollars

The new regression is

$$\begin{aligned} \widehat{salarydol} &= 963,191 + 18,501\ roe \\ n &= 209,\ \ R^2 = .01 \end{aligned}$$

The effects of Changing Units of Measurement on OLS Statistics


```
> salarydol<-ceosal1$salary*1000
> regression_ceosal1b <- lm(salarydol ~ roe, data = ceosal1)
> stargazer(regression_ceosal1b, type='text', align=TRUE, digits=2)

===============================================
                       Dependent variable:
                   ----------------------------
                            salarydol
-----------------------------------------------
roe                         18,501.19*
                           (11,123.25)

Constant                   963,191.30***
                           (213,240.30)

-----------------------------------------------
Observations                   209
R2                             0.01
Adjusted R2                    0.01
Residual Std. Error   1,366,555.00 (df = 207)
F Statistic             2.77* (df = 1; 207)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Using the Natural Logarithm in Simple Regression

- Recall the **wage** example:

Example (Wage)

$$\begin{aligned}\widehat{wage} &= -0.90 + 0.54\ educ \\ n &= 526, \qquad R^2 = .16\end{aligned}$$

- Now, think about the econometric model and how this OLS Regression Function is interpreted.

- What the OLS Regression Line says may not fit how economically we see the problem.

Possible issue: the dollar value of another year of schooling is constant.

Using the Natural Logarithm in Simple Regression

- So the 16^{th} year of education is worth the same as the second.

- We expect additional years of schooling to be worth more, in dollar terms, than previous years.

- How can we incorporate an increasing effect? One way is to postulate a constant *percentage* effect.

- We can approximate percentage changes using the natural log.

Constant Percent Model

- Let the dependent variable be $\log(wage)$ and write a (new) simple linear regression model:

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

- Let's define $\log(wage)$ (write it as $lwage$) and run a new regression.

Using the Natural Logarithm in Simple Regression

	Dependent variable:
	lwage
educ	0.08***
	(0.01)
Constant	0.58***
	(0.10)
Observations	526
R2	0.19
Adjusted R2	0.18
Residual Std. Error	0.48 (df = 524)
F Statistic	119.58*** (df = 1; 524)
Note:	*p<0.1; **p<0.05; ***p<0.01

Using the Natural Logarithm in Simple Regression

$$
\begin{aligned}
\widehat{lwage} &= 0.58 + .08\ educ \\
n &= 526,\ R^2 = .19
\end{aligned}
$$

- The estimated return to each year of education is about 8%.

- **Attention:**

This R-squared is not directly comparable to the R-squared when $wage$ is the dependent variable. The total variation (SSTs) in $wage_i$ and $lwage_i$ that we must explain are completely different.

Using the Natural Logarithm in Simple Regression

Constant Elasticity Model

• We can use the log on both sides of the equation to get **constant elasticity models**. For example, if

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u$$

then

$$\beta_1 \approx \frac{\%\Delta salary}{\%\Delta sales}$$

• The elasticity is free of units of $salary$ and $sales$.

• A constant elasticity model for salary and sales makes more sense than a constant dollar effect.

Using the Natural Logarithm in Simple Regression

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-Level	y	x	$\Delta y = \beta_1 \Delta x$
Level-Log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-Level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-Log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

- Recall the **CEO salary** example, but now the independent variable is *sales*.

$$salary = \beta_o + \beta_1 sales + u$$

- Applying log on both variables (dependent and independent) we get:

Example (CEO salary)

$$\begin{aligned} log\widehat{(salary)} &= 4.82 + 0.26\ log(sales) \\ n &= 209, \qquad R^2 = .21 \end{aligned}$$

	Dependent variable:
	log(salary)
log(sales)	0.26***
	(0.03)
Constant	4.82***
	(0.29)
Observations	209
R2	0.21
Adjusted R2	0.21
Residual Std. Error	0.50 (df = 207)
F Statistic	55.30*** (df = 1; 207)
Note:	*p<0.1; **p<0.05; ***p<0.01

Using the Natural Logarithm in Simple Regression

- The estimated elasticity of CEO salary with respect to firms sales is about .26.
- A **10 percent** increase in sales is associated with a

$$.26(10) = 2.6$$

percent increase in salary.

Topics

1. Definition of the Simple Regression Model
2. Deriving the Ordinary Least Squares Estimates
3. Properties of OLS on any Sample of Data
4. Units of Measurement and Functional Form
   - Using the Natural Logarithm in Simple Regression
5. Expected Value of OLS

The 4 Assumptions for Unbiasedness

Goal: We want to study statistical properties of the OLS estimator

- In order to that, we will need to impose 4 assumptions.

The 4 Assumptions for Unbiasedness

Assumption SLR.1 (Linear in Parameters)

The population model can be written as

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

where β_0 and β_1 are the (unknown) population parameters.

- What linear in parameters mean?
- **Example of non linear in parameters on white board**

The 4 Assumptions for Unbiasedness

Assumption SLR.2 (Random Sampling)

We have a **random sample** of size n, $\{(x_i, y_i) : i = 1, ..., n\}$, following the population model.

The 4 Assumptions for Unbiasedness

Assumption SLR.3 (Sample Variation in the Explanatory Variable)

The sample outcomes on x_i are not all the same value.

- This is the same as saying the sample variance of $\{x_i : i = 1, ..., n\}$ is **not zero**.
- If in the population x does not change then we are not asking an interesting question.

The 4 Assumptions for Unbiasedness

Assumption SLR.4 (Zero Conditional Mean)

In the population, the error term has zero mean given any value of the explanatory variable:

$$E(u|x) = 0 \text{ for all } x.$$

- **Key assumption.**

- We can compute the OLS estimates whether or not this assumption holds.

The 4 Assumptions for Unbiasedness

Goal: We want to know if $\hat{\beta}_1$ is unbiased for β_1, and $\hat{\beta}_0$ is unbiased for β_0

- If,

$$
\begin{aligned}
E(\hat{\beta}_1) &= \beta_1 \\
E(\hat{\beta}_0) &= \beta_0
\end{aligned}
$$

Then, the **OLS estimator** is unbiased.

- **Demonstration:** On the white board.

Unbiasedness of OLS

Theorem: Unbiasedness of OLS

Under Assumptions SLR.1 through SLR.4

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1 \ ,$$

for any values of β_0 and β_1, i.e., $\hat{\beta}_0$ is unbiased for β_0, and $\hat{\beta}_1$ is unbiased for β_1

Unbiasedness of OLS

- Therefore, the four assumptions for the OLS estimator to be unbiased are:

SLR.1: (Linear in Parameters) $y = \beta_0 + \beta_1 x + u$
SLR.2: (Random Sampling)
SLR.3: (Sample Variation in x_i)
SLR.4: (Zero Conditional Mean) $E(u|x) = 0$

- If any of these assumptions fails, the OLS estimator will (generally) be biased.

- **To be discussed in the next chapter:** What are the omitted factors? Are they likely to be correlated with x? If so, SLR.4 fails and OLS will be biased.