Simple Linear Regression

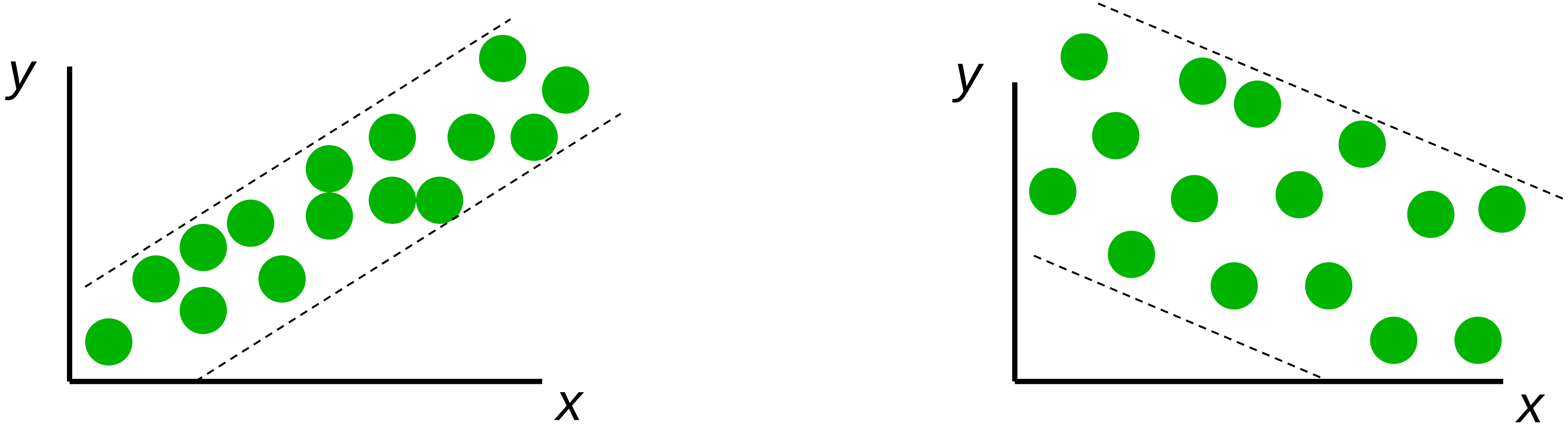
- Simple Regression
- Least Squares Method
- Coefficient of Determination
- Testing the Significance of the Slope of the Regression Equation

- Reading: Chapter 14
    - Section 14.2 – skip the test to determine the significance of the population correlation coefficient
    - Section 14.4 – skip the confidence interval and the prediction interval
    - Section 14.6 – skip for now, may discuss it later if time permits

Preliminaries: Correlation Analysis

Correlation coefficient measures both the **strength** and **direction** of a **linear** relationship between two variables

- Scatterplot is a good way to start observing the relationship
- In a scatterplot, we can see patterns and form expectations about form, direction and strength of the relationship



- Calculate the correlation coefficient:
  - The values of *r* range from -1.0, a strong negative relationship, to +1.0, a strong positive relationship

Some Examples…

Managerial decisions often are based on the relationship between two or more variables:

1. Relationship between advertising expenditures and sales
   - Predict sales for a given level of advertising expenditures
2. Relationship between daily high temperatures and the demand for electricity
   - Predict electricity usage on the basis of anticipated daily high temperatures

Relationship between life expectancy and income per person (cross country analysis)

- Is it positive? Is it strong? If yes, how big is the effect of income per person on life expectancy?

Our Goal

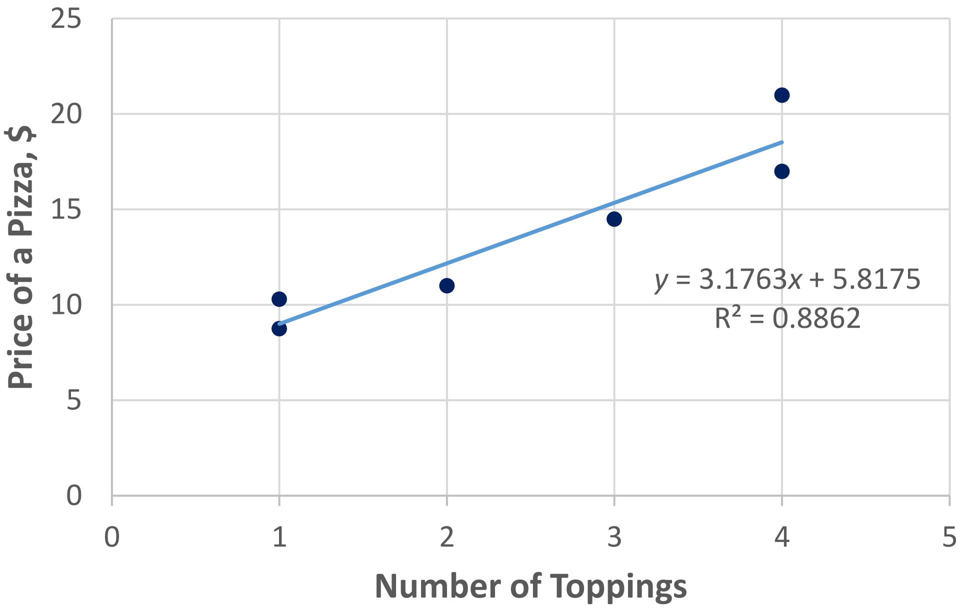
- We know something about relationship between two variables:
    - E.g., correlation coefficient
- Correlation only says "there seems to be a linear association between two variables", but does not tell us what that the association is
- We can say more about linear relationship between two variables with the **model**
    - Whether the change in one variable is associated with the big/small change in another variable

Our goal ⇒ develop such a model

Illustration Example

- Relationship between price of a pizza and the number of toppings

Order N	*Number of Toppings*	*Price, $*
1	1	8.75
2	2	11
3	1	10.3
4	3	14.5
5	4	17
6	4	21



Dependent and Independent Variables

An **independent variable, x,** explains the variation in another variable, which is called the **dependent variable, y**

- x = Number of toppings (horizontal axis)
- y = Price of a pizza (vertical axis)

For the regression analysis, the direction of the relationship matters

Variation in x explains variation in y, but not the reverse (direction is only one way)

Independent variable (x) → Dependent variable (y)

Simple Linear Regression

- Simple linear regression involves **one independent** variable and one dependent variable
  - Pizza example
- The relationship between the two variables is approximated by a straight line
- Regression analysis involving **two or more independent variables** is called multiple regression
  - You may want to include additional information about the order to help you explain the price of the pizza. For example, the gender of a restaurant visitor…

Simple Linear Regression Model

The equation that describes how y is related to x and an error term is called the regression model

The simple linear regression model (for the population) is:

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

where:

β_0 and β_1 are called parameters of the model;

ε is a random variable called the error term

Simple Linear Regression Equation

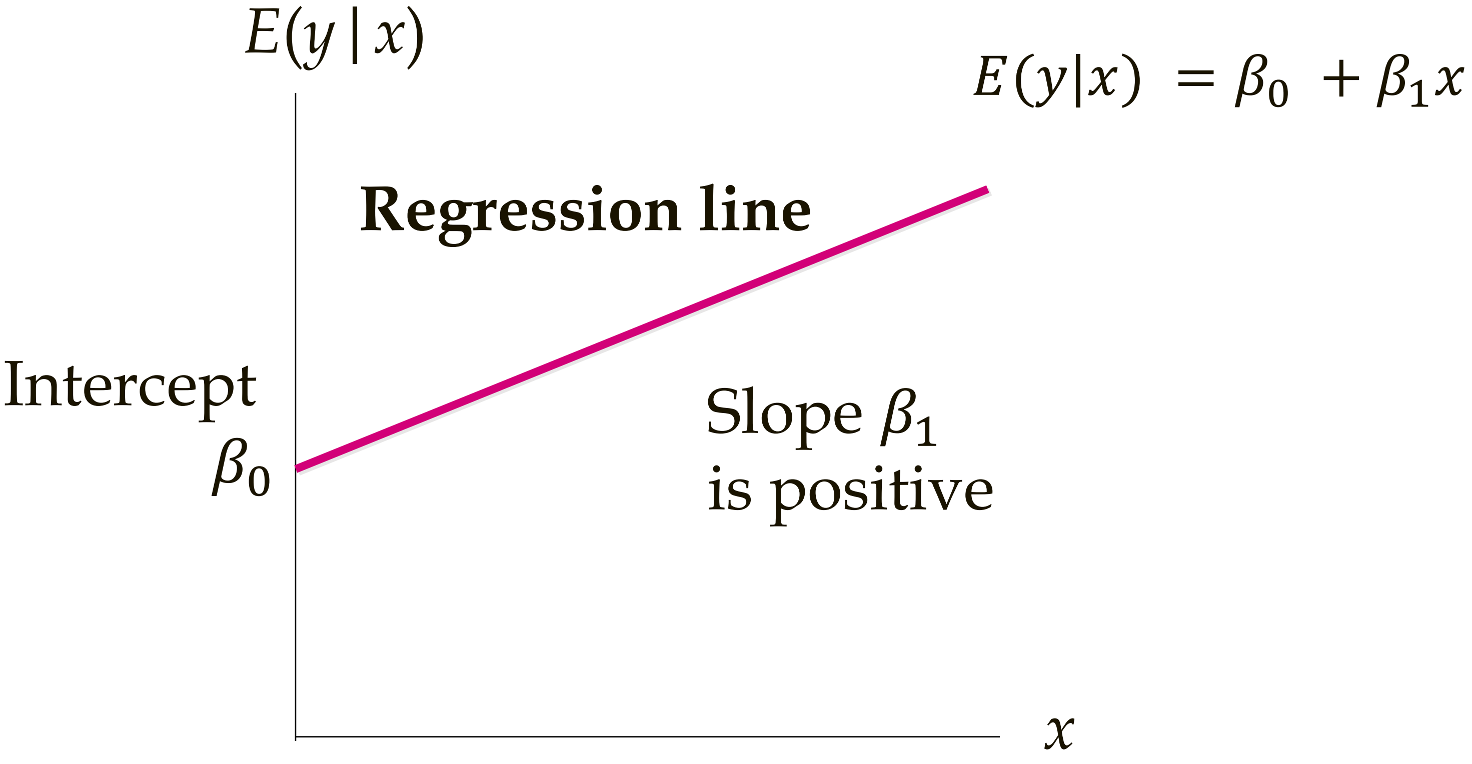
The simple linear regression equation is:

$$E(y|x) = \beta_0 + \beta_1 x$$

- Graph of the regression equation is a straight line
- β_0 is the y intercept of the regression line
- β_1 is the slope of the regression line
- $E(y|x)$ is the expected value of y for a given x value

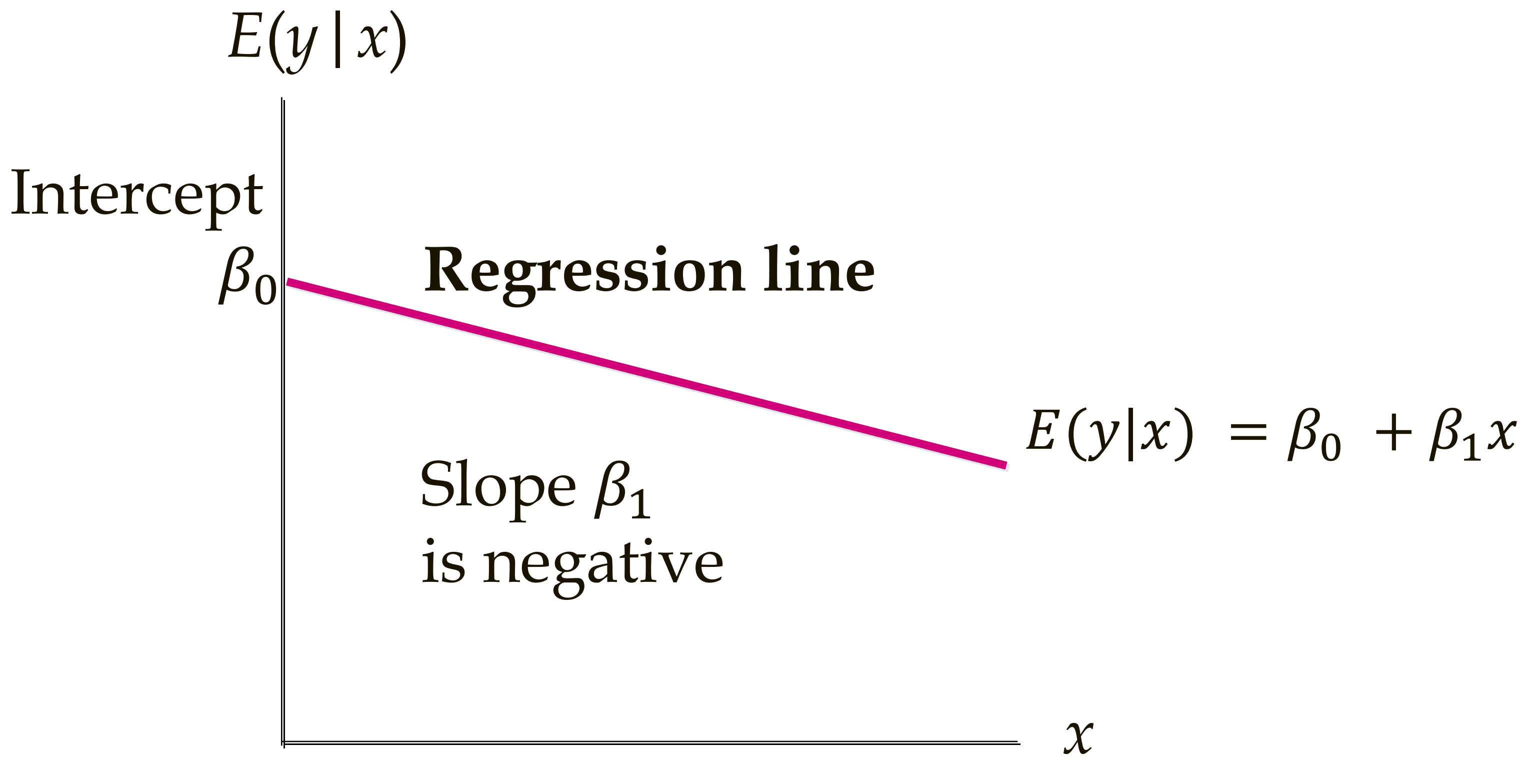
Simple Linear Regression Equation

Positive Linear Relationship



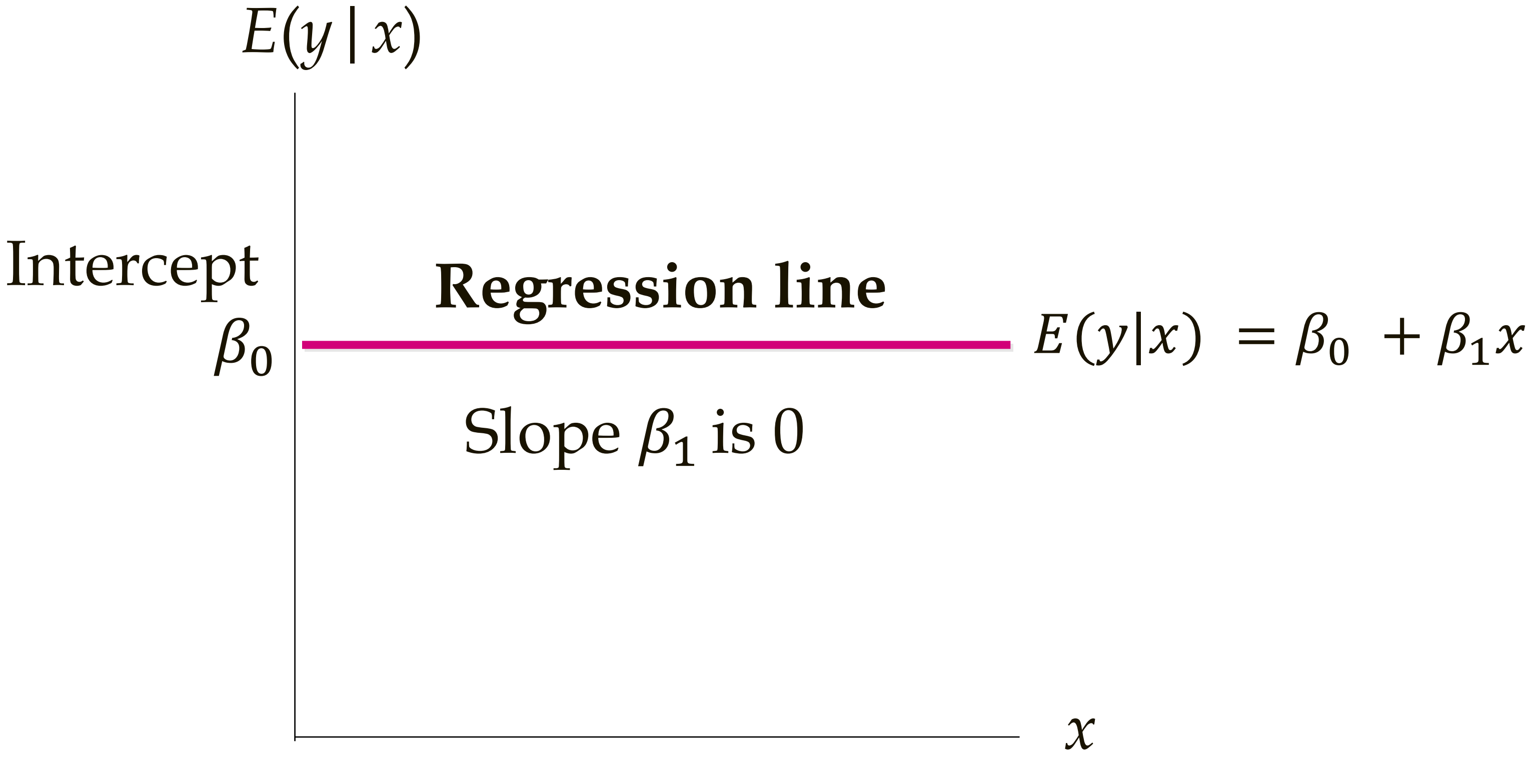
Simple Linear Regression Equation

Negative Linear Relationship



Simple Linear Regression Equation

No Relationship



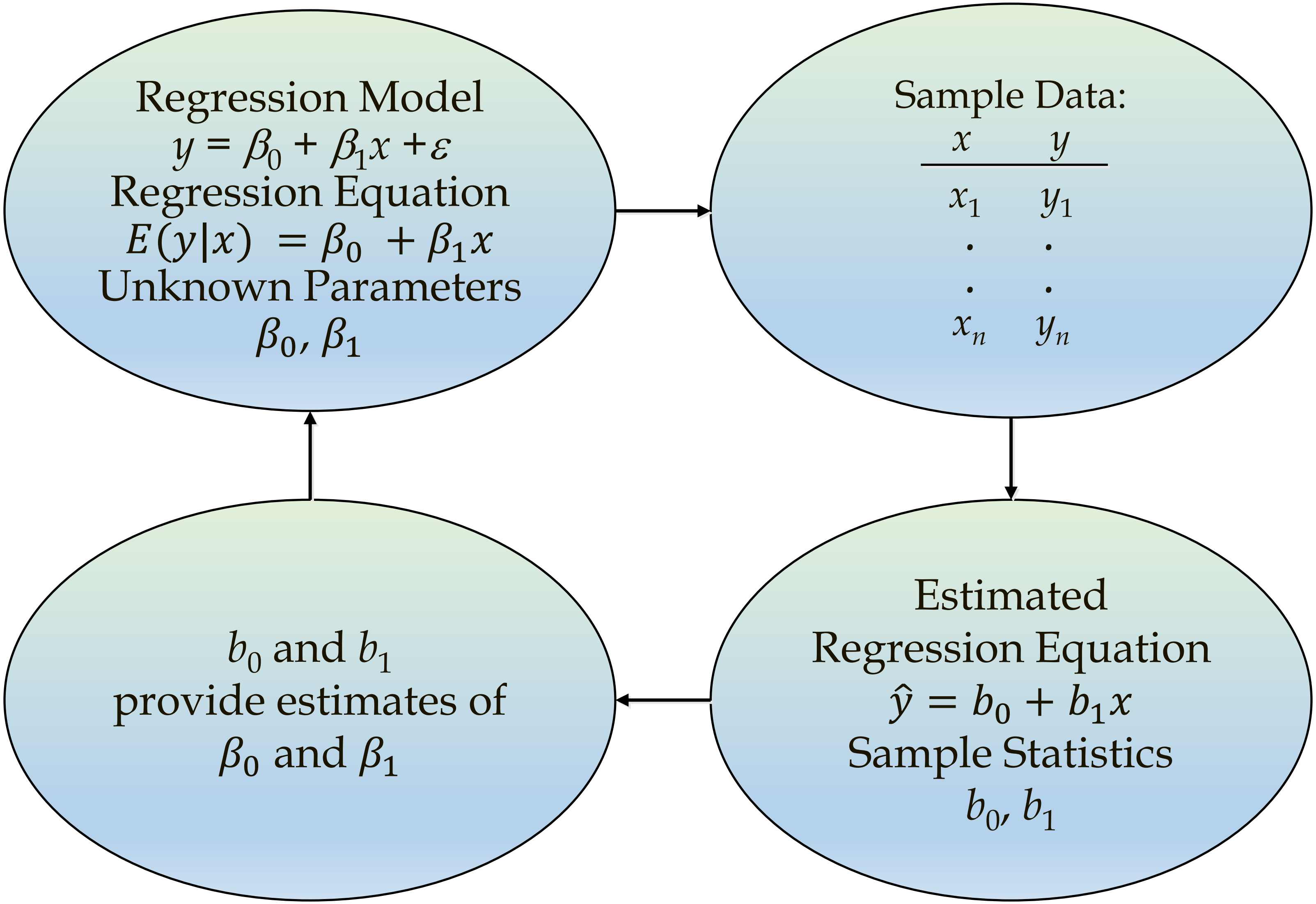
Estimated Simple Linear Regression Equation

The estimated simple linear regression equation

$$\hat{y} = b_0 + b_1 x$$

- b_0 is the y intercept of the estimated regression equation
- b_1 is the slope of the estimated regression equation
- $\hat{y}$ is the estimated (or predicted) value of y for a given x value
- The graph is called the *estimated regression line*

Estimation Process



Regression Model
$y = \beta_0 + \beta_1 x + \varepsilon$
Regression Equation
$E(y|x) = \beta_0 + \beta_1 x$
Unknown Parameters
β_0, β_1

Sample Data:

x	y
x_1	y_1
.	.
.	.
x_n	y_n

Estimated
Regression Equation
$\hat{y} = b_0 + b_1 x$
Sample Statistics
b_0, b_1

b_0 and b_1
provide estimates of
β_0 and β_1

Simple Regression Analysis

The difference between the actual data value and the predicted value is known as the **residual, e_i**:

$$e_i = y_i - \hat{y}_i$$

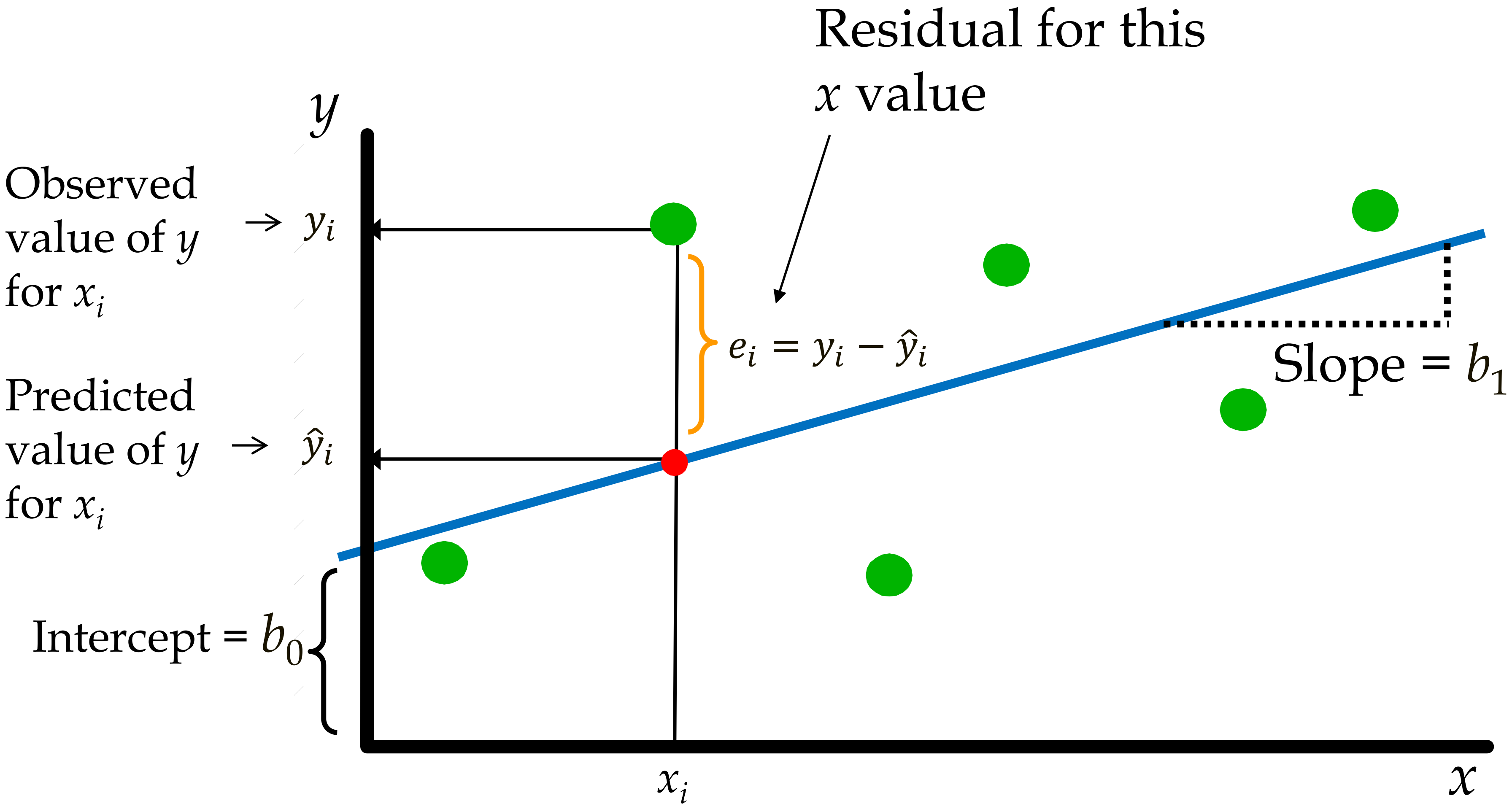
where:

e_i = The residual of the *i*th observation in the sample

y_i = The actual value of the dependent variable for the *i*th data point

$\hat{y}_i$ = The predicted value of the dependent variable for the *i*th data point

Simple Regression Analysis


Residual for this x value
y
Observed value of y for x_i → y_i
$e_i = y_i - \hat{y}_i$
Slope = b_1
Predicted value of y for x_i → $\hat{y}_i$
Intercept = b_0
x_i
x

On the scatter plot: Residual e_i for ith observation is a vertical distance between the data point and the estimated regression line.

Least Squares Method

The **least squares method** identifies the linear equation that ***best fits*** a set of ordered pairs (x_i, y_i)

- It is used to find the values for b_0 (the y-intercept) and b_1 (the slope of the line)

Goal ⇒ Find the line that minimizes the differences between actual values of y and predicted values $\hat{y}$

More formally, the Least Squares Method minimizes the sum of squares error (SSE):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where n = The number of observations

Least Squares Method

Least Squares Objective Function

$$\min \sum (y_i - \hat{y}_i)^2$$

If you are interested in math:

1. Plug in the estimated regression equation:
$$\hat{y} = b_0 + b_1 x$$
2. Minimize the objective function w.r.t. to b_0 and b_1
3. Unconditional optimization tools:
   - Take derivatives of the objective function w.r.t. $b_0,\ b_1$
   - Equate them to zero
   - Express b_0 and b_1 as a function of x, y and n

Least Squares Method

Regression Slope and *y*-intercept:

$$b_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

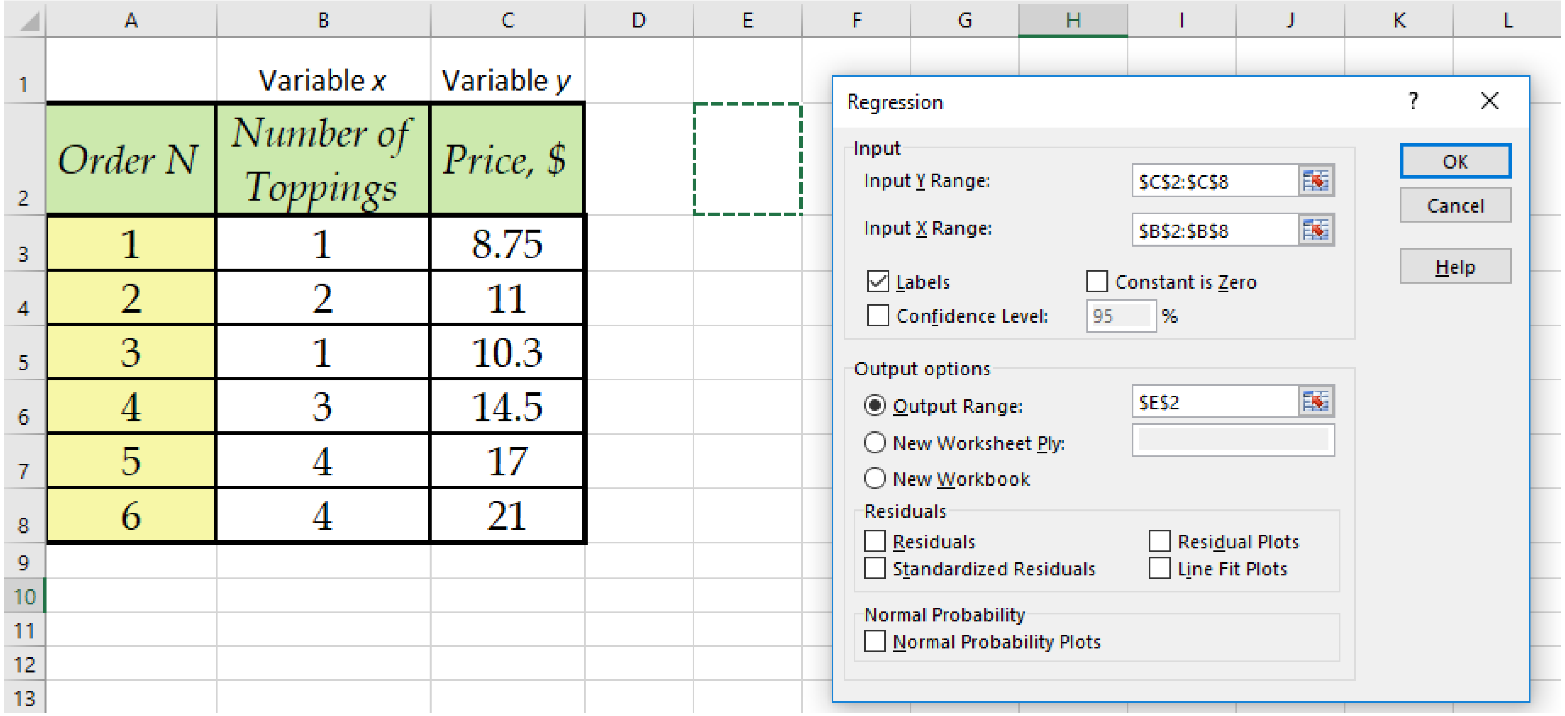
$$b_0 = \frac{\sum y}{n} - b_1\frac{\sum x}{n}$$

Note: Fortunately, virtually every statistical software package easily produces values for b_1 and b_0. So, we focus on interpreting the regression coefficients rather than performing the grueling calculations.

Excel: The Linear Regression Model

Pizza Example: Excel

- Choose **Data > Data Analysis > Regression**
- In the *Regression* dialog box, select corresponding *x* and *y* ranges and check *Labels*.



	Variable *x*	Variable *y*
Order N	*Number of Toppings*	*Price, $*
1	1	8.75
2	2	11
3	1	10.3
4	3	14.5
5	4	17
6	4	21

Regression

Input

Input Y Range: $C$2:$C$8

Input X Range: $B$2:$B$8

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $E$2

○ New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

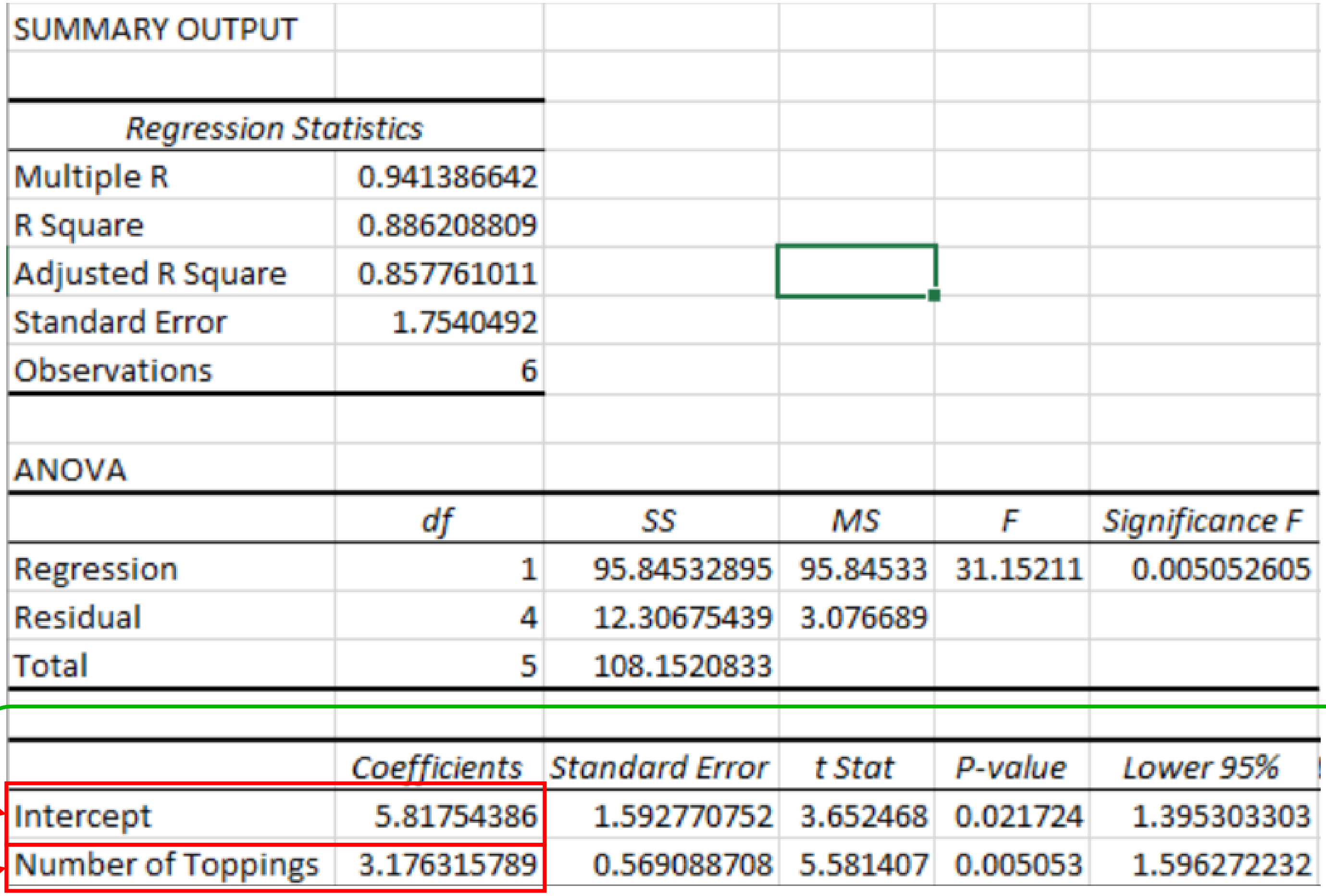
☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

Calculating the Slope and *y*-intercept Using Excel

The Excel-produced output from estimating the model for our pizza example:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.941386642
R Square	0.886208809
Adjusted R Square	0.857761011
Standard Error	1.7540492
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	95.84532895	95.84533	31.15211	0.005052605
Residual	4	12.30675439	3.076689		
Total	5	108.1520833			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	5.81754386	1.592770752	3.652468	0.021724	1.395303303
Number of Toppings	3.176315789	0.569088708	5.581407	0.005053	1.596272232

Information about coefficients

y-intercept

slope value

So the estimated regression equation: $\hat{y} = 5.8 + 3.18x$

Interpreting the Slope and *y*-intercept

The estimated slope b_1: *on average*, how much the dependent variable y *changes* when x *increases* by one unit

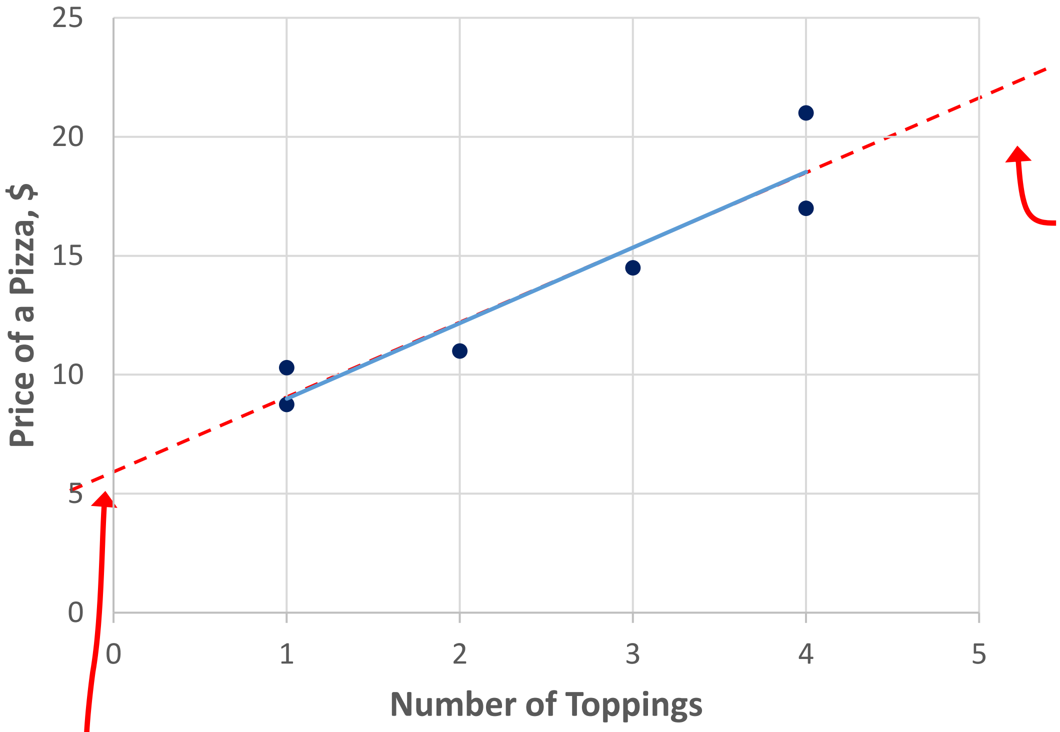
- Be specific about changes in y:
  - If $b_1 > 0$, then y *increases* by b_1 as x *increases* by 1
  - If $b_1 < 0$, then y *decreases* by b_1 as x *increases* by 1
- Make sure that you retain the unit of measurement for x and y

The estimated intercept b_0: it is not always possible to provide an economic interpretation of the intercept estimate b_0

- Mathematically, however, it represents the average value of y when x has a value of zero

Interpreting the Slope and *y*-intercept

Pizza example: the regression equation is: $\hat{y} = 5.8 + 3.18x$



Slope
= 3.18

Slope: On average, each additional topping *increases* the price of a pizza by $3.18

Intercept
= 5.8

Making a Prediction

Once the model is estimated:

- plug in the value of interest for independent variable x in the estimated regression to obtain the predicted value of $\hat{y}$

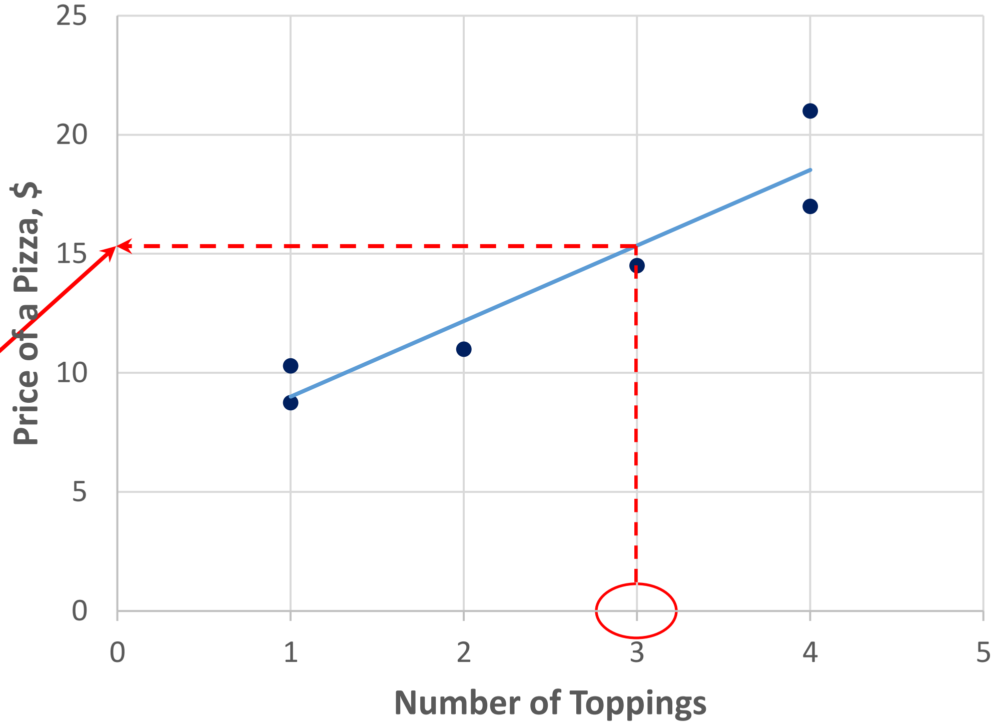
Pizza Example: the regression equation is: $\hat{y} = 5.8 + 3.18x$

What is the predicted price of a pizza for a person who ordered 3 toppings?

If we set $x = 3$ in the regression equation, we get:

$$\hat{y} = 5.8 + 3.18(3) = \$15.35$$

This a **point estimate** for the regression equation, given $x = 3$



Partitioning the Sum of Squares

The total sum of squares (*SST*) measures the total variation in the dependent variable

Total variation is made up of two parts:

Explained portion

$$SST = SSR + SSE$$

Unexplained portion

Total sum of Squares

$$SST = \sum (y_i - \bar{y})^2$$

Sum of Squares Regression

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

Sum of Squares Error

$$SSE = \sum (y_i - \hat{y}_i)^2$$

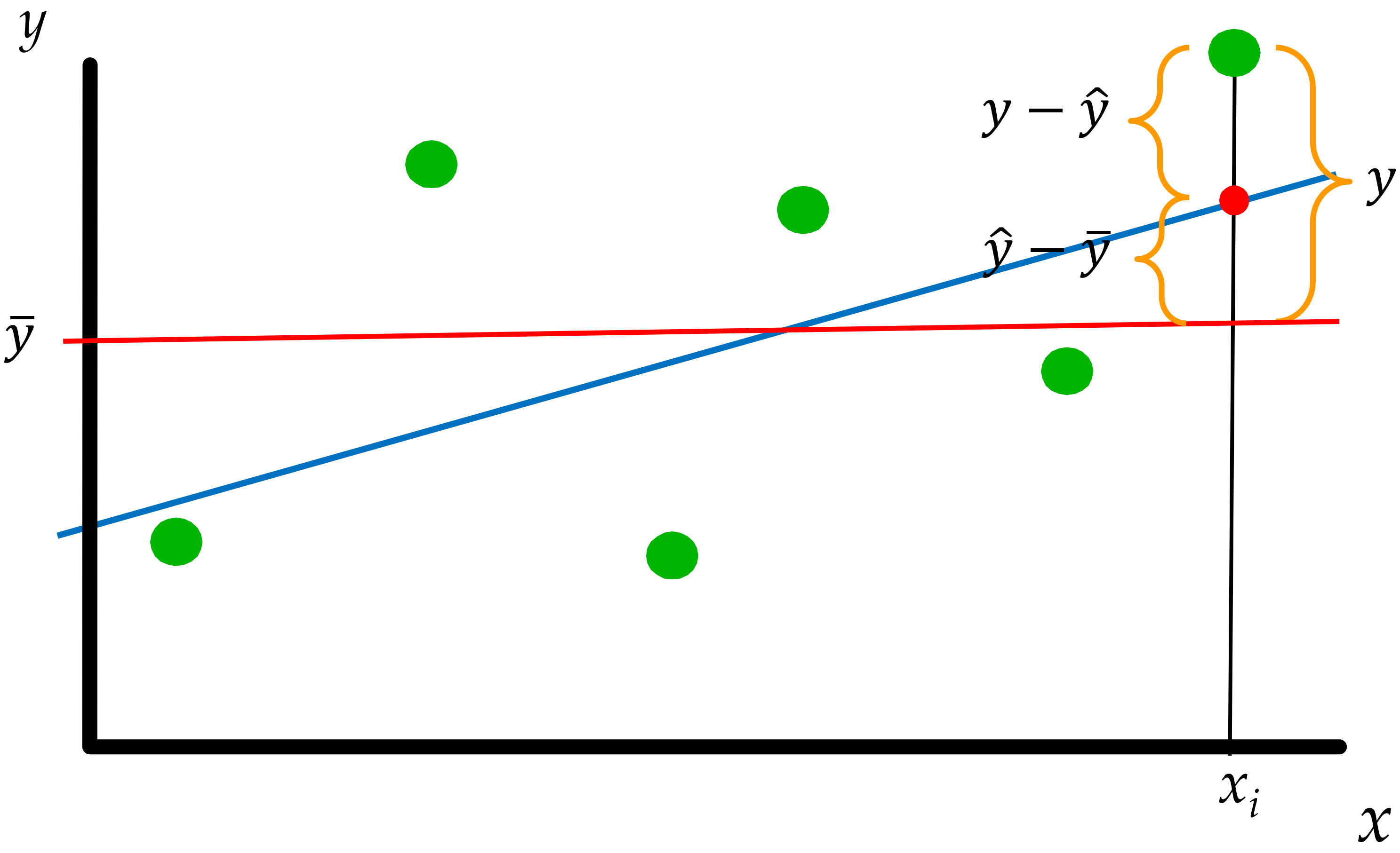
where:

y = A value of the dependent variable from the sample

$\bar{y}$ = The average value of the dependent variable from the sample

$\hat{y}$ = The estimated value of y for a given x value

Partitioning the Sum of Squares



Coefficient of Determination

The **coefficient of determination, R^2**, measures the percentage of the total variation of the dependent variable that is explained by the independent variable from a sample

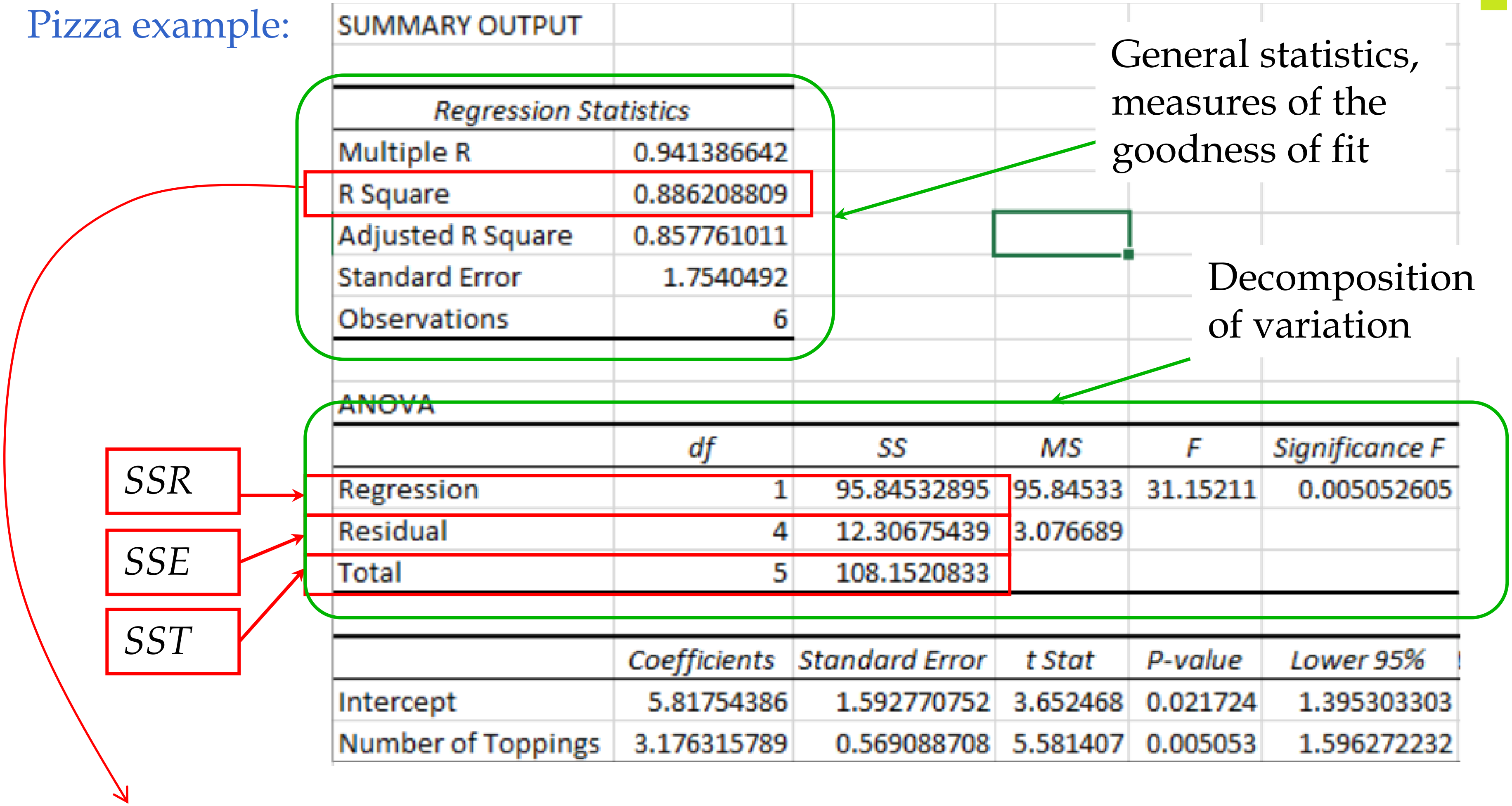
$$R^2 = \frac{SSR}{SST}$$

R^2 varies from 0% to 100% (0 to 1 if expressed as a fraction):

- Higher values of R^2 are more desirable because we would like to explain as much of the variation in the dependent variable as possible. In other words, higher values of R^2 indicate a higher predictive power of the regression

Partitioning the Sum of Squares

Pizza example:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.941386642
R Square	0.886208809
Adjusted R Square	0.857761011
Standard Error	1.7540492
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	95.84532895	95.84533	31.15211	0.005052605
Residual	4	12.30675439	3.076689		
Total	5	108.1520833			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	5.81754386	1.592770752	3.652468	0.021724	1.395303303
Number of Toppings	3.176315789	0.569088708	5.581407	0.005053	1.596272232

General statistics, measures of the goodness of fit

Decomposition of variation

$R^2 = 0.886 \Rightarrow$ 88.6% of the variation in the price of pizza (y) is explained by the number of toppings (x)

Hypothesis Test to Determine the Significance of the Coefficient of Determination

The **population coefficient of determination, ρ^2**, is unknown

The calculated value of R^2 represents the coefficient of determination for a random sample from the population

Use this hypothesis test to determine if the population coefficient of determination is significantly different from zero (based on the sample coefficient of determination):

$H_0 : \rho^2 \leq 0$ (none of the variation in y is explained by x)

$H_1 : \rho^2 > 0$ (x does explain a significant (non-zero) portion of the variation in y)

Hypothesis Test to Determine the Significance of the Coefficient of Determination

The *F*-test statistic is the appropriate test statistic for this hypothesis test

$$F = \frac{MSR}{MSE}$$

With degrees of freedom
$D_1 = k$
$D_2 = n - k - 1$

where:

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n - k - 1}$$

MSR = The mean square regression
MSE = The mean square error
SSR = The sum of squares regression
SSE = The sum of squares error
n = The number of observations in the sample
k = The number of independent variables

Hypothesis Test to Determine the Significance of the Coefficient of Determination

For the simple linear regression: $k = 1 \Rightarrow$

F test statistic simplifies:

$$F = \frac{SSR}{\left(\frac{SSE}{n-2}\right)}$$

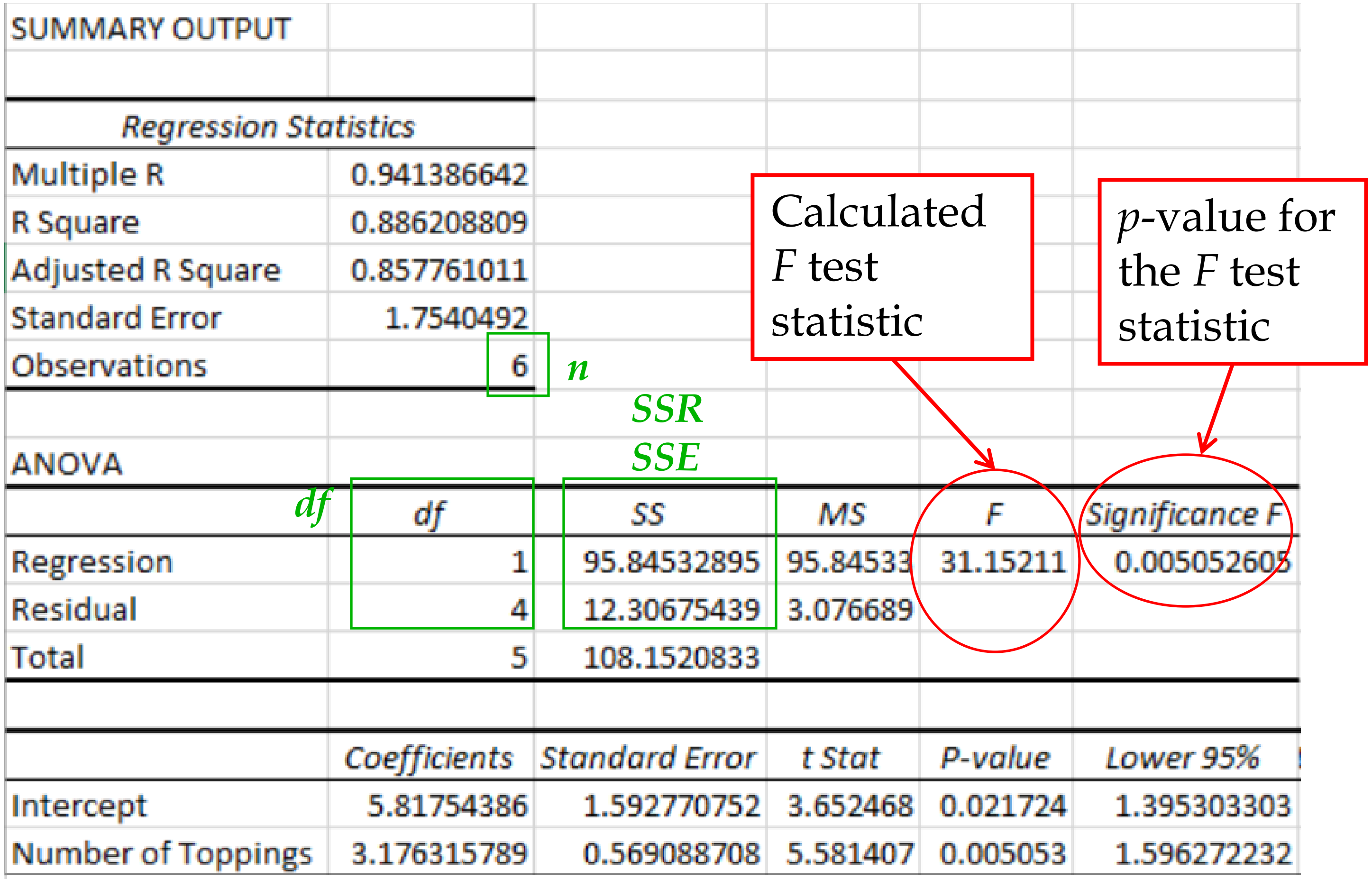
The critical value and the p-value are computed from F distribution with degrees of freedom:

$$D_1 = 1 \text{ and } D_2 = n - 2$$

Note: F statistic can be calculated manually or found in Excel output

Hypothesis Test to Determine the Significance of the Coefficient of Determination

Pizza example:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.941386642
R Square	0.886208809
Adjusted R Square	0.857761011
Standard Error	1.7540492
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	95.84532895	95.84533	31.15211	0.005052605
Residual	4	12.30675439	3.076689		
Total	5	108.1520833			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	5.81754386	1.592770752	3.652468	0.021724	1.395303303
Number of Toppings	3.176315789	0.569088708	5.581407	0.005053	1.596272232

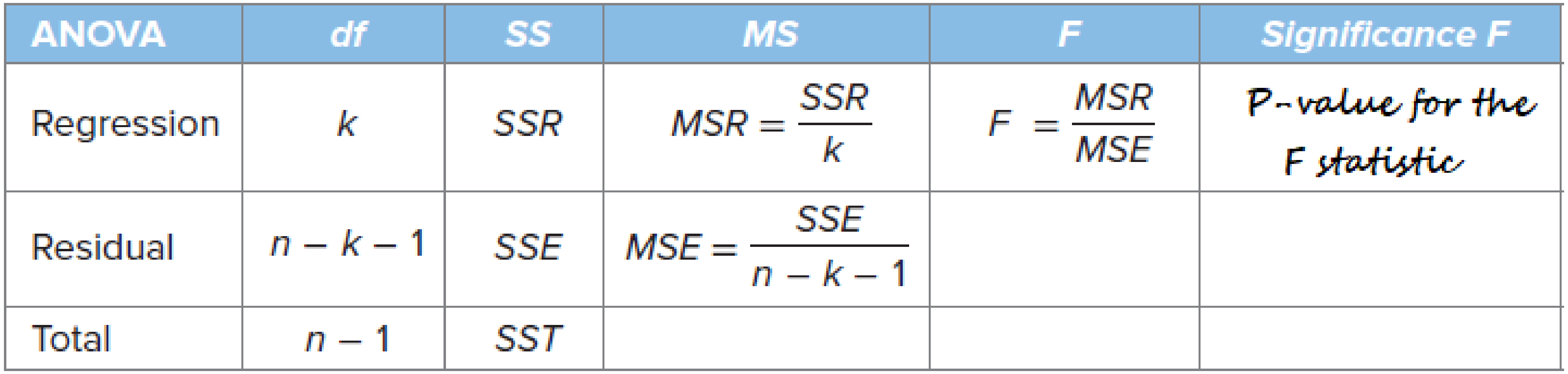
Calculated *F* test statistic

p-value for the *F* test statistic

Can substitute values in green to find *F* statistic:

$$F = \frac{SSR}{\left(\frac{SSE}{n-2}\right)}$$

Summary of the ANOVA Table



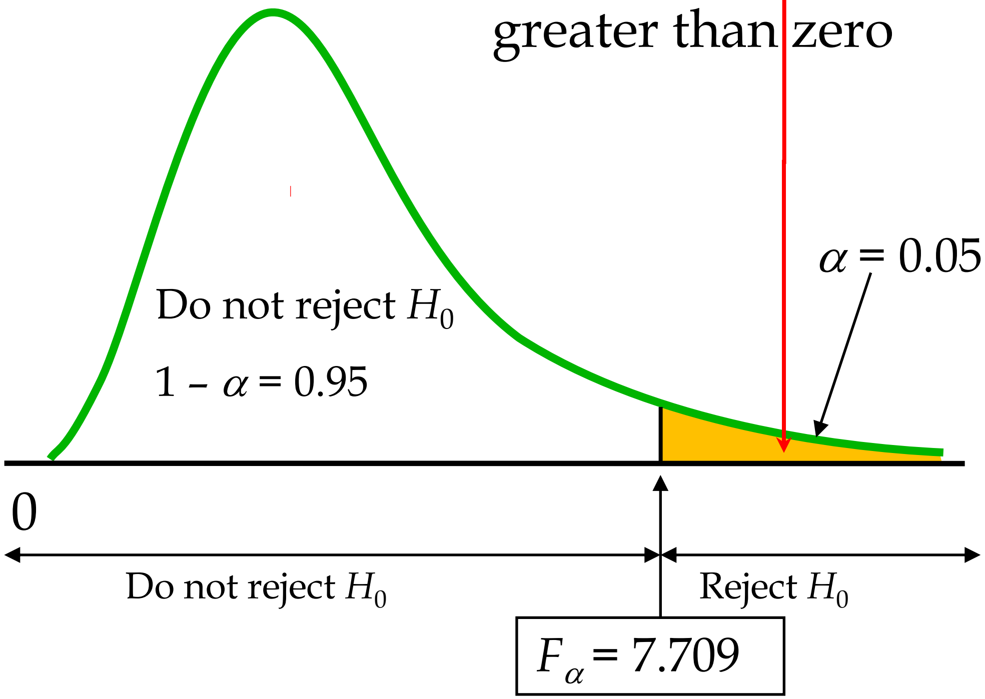
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$	P-value for the F statistic
Residual	$n - k - 1$	SSE	$MSE = \frac{SSE}{n - k - 1}$		
Total	$n - 1$	SST			

Hypothesis Test to Determine the Significance of the Coefficient of Determination

(**Skip in the class**) Critical value approach:

$$H_0 : \rho^2 \leq 0$$
$$H_1 : \rho^2 > 0$$

Since $F = 31.15 > F_\alpha = 7.709$ we **reject H_0** and conclude that the coefficient of determination is greater than zero



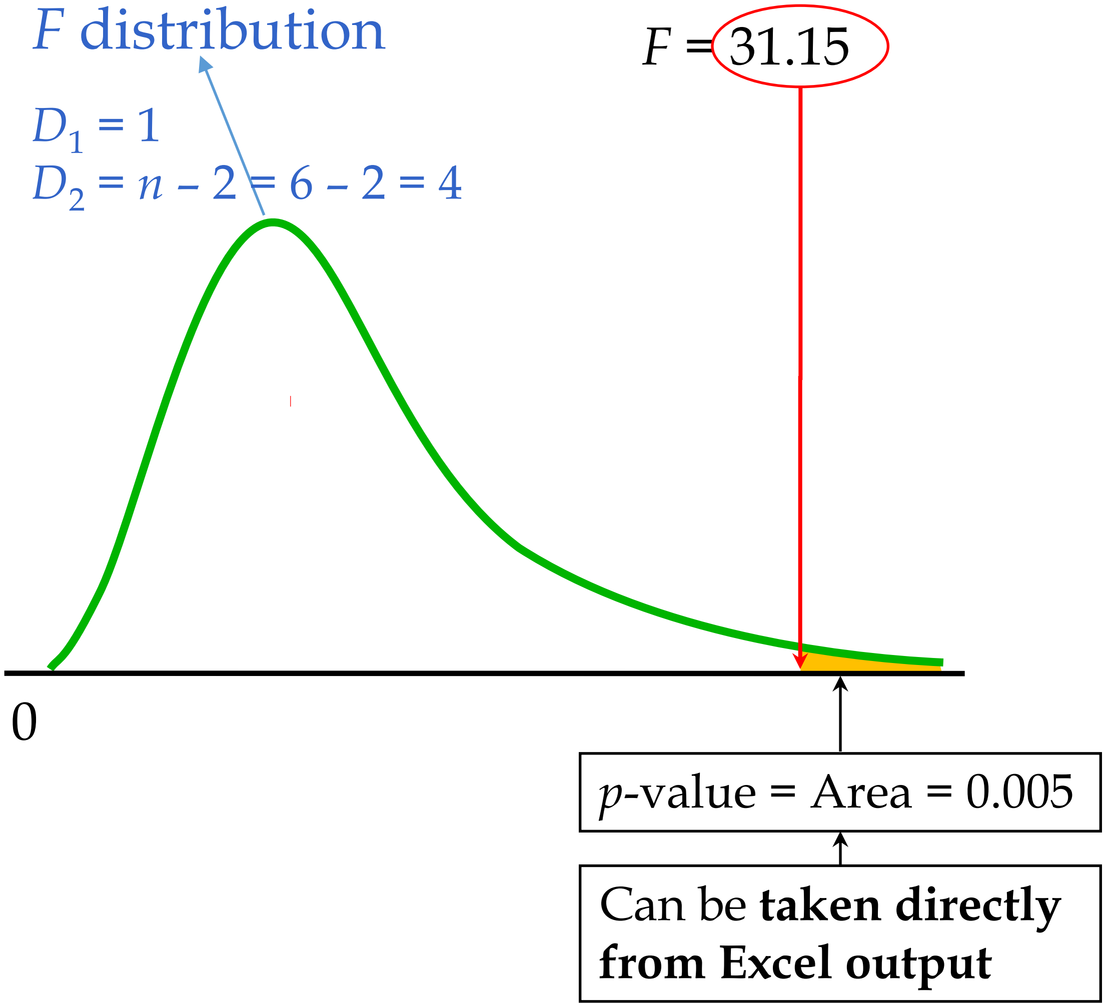
For this example:

$$D_1 = 1$$
$$D_2 = n - 2 = 6 - 2 = 4$$

The critical F-score for $\alpha = 0.05$ and degrees of freedom equal to 1 and 4 is $F_\alpha = 7.709$ (Excel is needed to perform these calculations)

Hypothesis Test to Determine the Significance of the Coefficient of Determination

Using pizza example output and a 5% significance level:



F distribution

$D_1 = 1$

$D_2 = n - 2 = 6 - 2 = 4$

$F = 31.15$

0

p-value = Area = 0.005

Can be **taken directly from Excel output**

The *p*-value is 0.005 < α = 0.05, so we *reject* the null hypothesis that there is no relationship between the price of a pizza and the number of toppings.

Therefore, we can conclude that the number of toppings explains non-zero portion of variation in the price of the pizza OR that there is a linear relationship between the price of a pizza and the number of toppings.

Standard Error of the Estimate

So far, we know one measure of the goodness-of-fit for the estimated regression: the coefficient of determination (R^2).

Now, we introduce one more measure: the standard error of the estimate (s_e).

- Generally, it measures the standard deviation of the residuals

$$s_e = \sqrt{\frac{SSE}{n-k-1}}$$

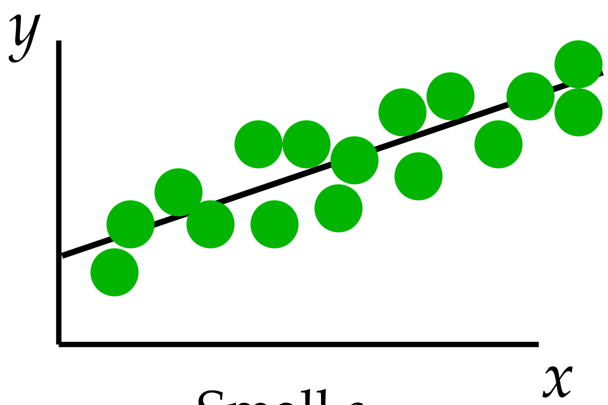
  where $SSE = \sum(y_i - \hat{y}_i)^2$ is the sum of squares error

- s_e can take any value between 0 and infinity (the closer it is to 0, the better the model fits the data)

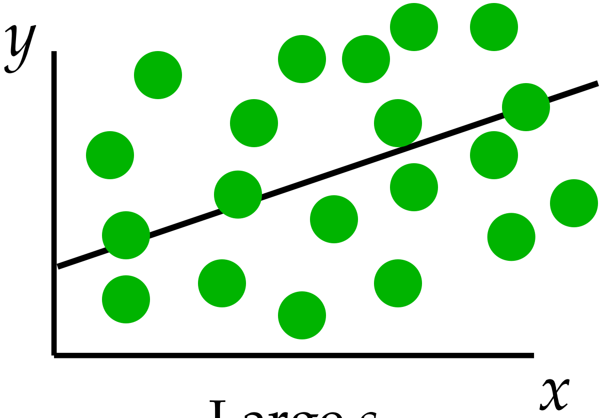
Standard Error of the Estimate

Standard error of the estimate, s_e, measures the variation of observed y values from the regression line

Graphically ⇒



Small s_e

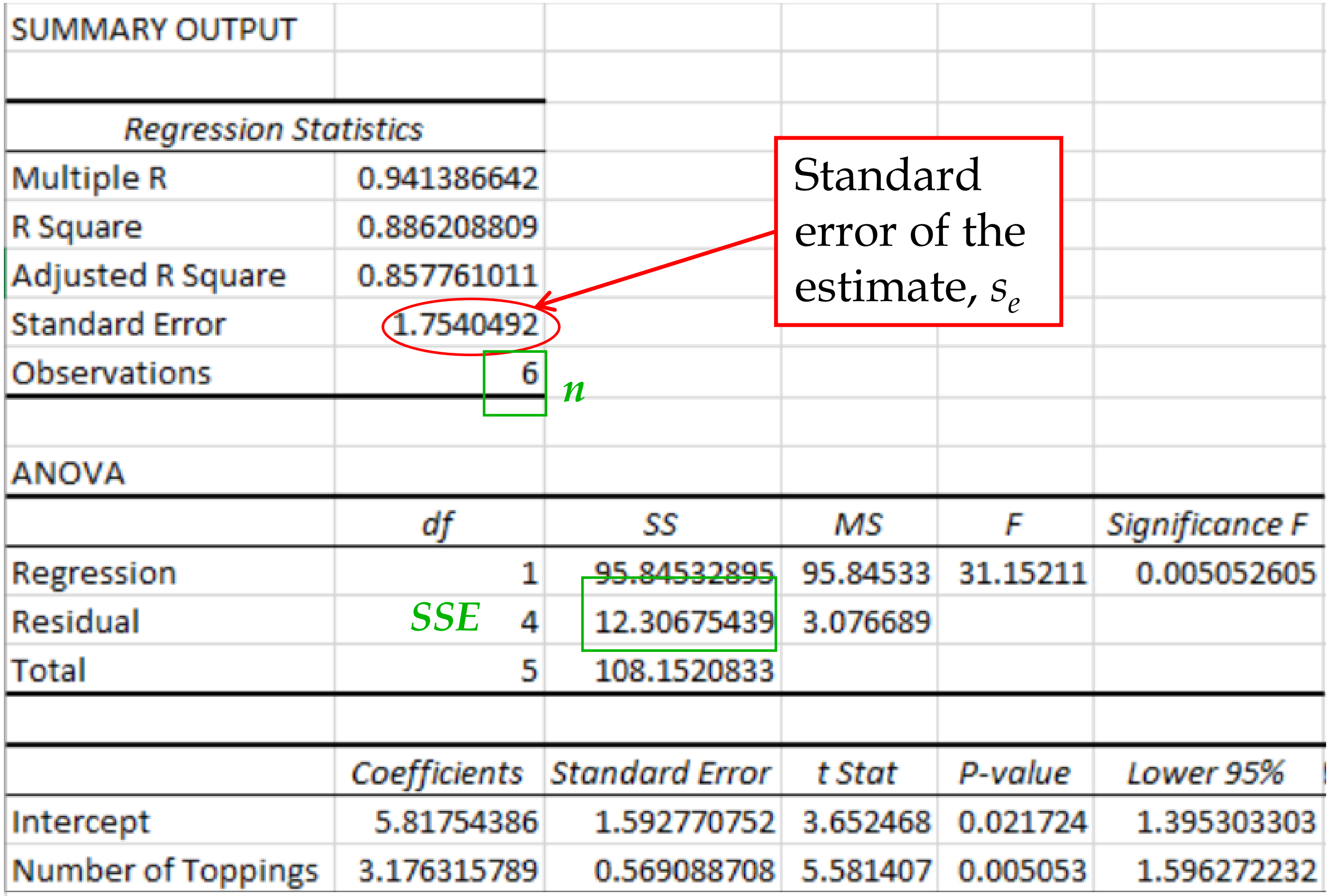


Large s_e

Standard Error of the Estimate

Excel reports s_e in the *Regression Statistics* portion of the regression output and refers to it as ***Standard Error***

Pizza example:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.941386642
R Square	0.886208809
Adjusted R Square	0.857761011
Standard Error	1.7540492
Observations	6

Standard error of the estimate, s_e

n

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	95.84532895	95.84533	31.15211	0.005052605
Residual	4	12.30675439	3.076689		
Total	5	108.1520833			

SSE

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	5.81754386	1.592770752	3.652468	0.021724	1.395303303
Number of Toppings	3.176315789	0.569088708	5.581407	0.005053	1.596272232

Can substitute values in green and $k = 1$ to find s_e:

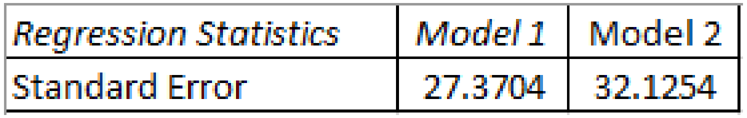
$$s_e = \sqrt{\frac{SSE}{n-k-1}}$$

Standard Error of the Estimate

The standard error of the estimate (s_e) is a useful measure to compare models:

- The lower the dispersion in the residuals $\Rightarrow$ the smaller s_e, which implies that the model provides a better fit for the sample data

For example, as measured by the standard error of the estimate, does a *Model 1* or *Model 2* provide a better fit for explaining the dependent variable y?



Regression Statistics	*Model 1*	Model 2
Standard Error	27.3704	32.1254

The standard error of the estimate for *Model 2* is greater than that for *Model 1* (32.1254 > 27.3704) $\Rightarrow$Therefore, Model 1 provides a better fit for the sample data

Testing the Significance of the Slope of the Regression Equation

If the population slope $\beta_1 = 0$, then x has no effect on y

⇒ can conclude that there is no relationship between the dependent and independent variables

Need a hypothesis test to determine if the population regression slope, β_1, is significantly different from zero, based on the sample regression slope, b_1

- One-tail and two tail tests are possible

Two-tail test is our focus:
easy with most software

$H_0 : \beta_1 = 0$ (There is no relationship between the independent and dependent variables)

$H_1 : \beta_1 \neq 0$ (There is a relationship between x and y)

Testing the Significance of the Slope of the Regression Equation

t test Statistic for the Regression Slope

$$t = \frac{b_1 - \beta_1}{s_b}$$

where:

b_1 = The sample regression slope
β_1 = The population regression slope from the H_0
s_b = The standard error of the slope

To find the critical value and p-value:

- Use a t distribution with $df = n - k - 1$
- $k = 1$ for the simple linear regression $\Rightarrow df = n - 2$

Sampling Distribution of b_1

The population regression slope, β_1, is unknown

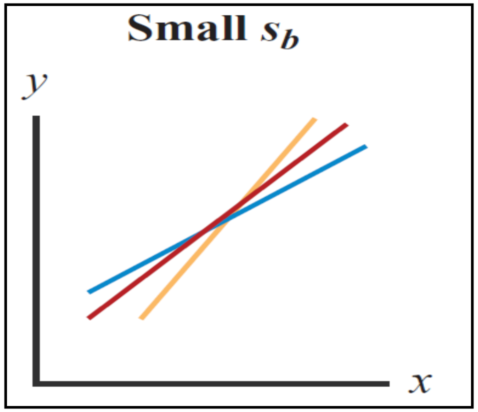
The calculated value of the slope, b_1, is from a random sample

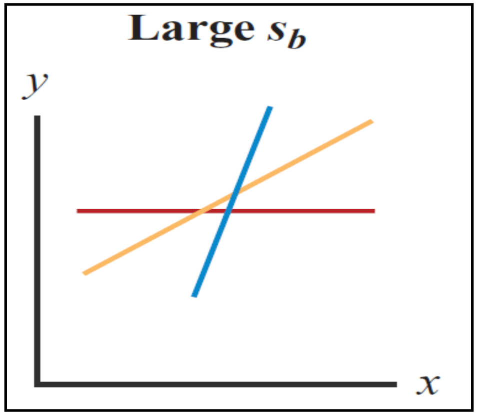
- Different samples $\Rightarrow$ Different values of b_1
- We can say that b_1 is a random variable which has its own sampling distribution!!!
- Generally, we would need to characterize the sampling distribution of b_1. But we won't do this our class. Still, we need to understand (intuitively) the standard error of the slope

The Standard Error of the Slope (s_b)

The **standard error of the slope, s_b,** measures the variation in the estimated slope of the regression equation, b_1 (as we estimate the regression for different samples)

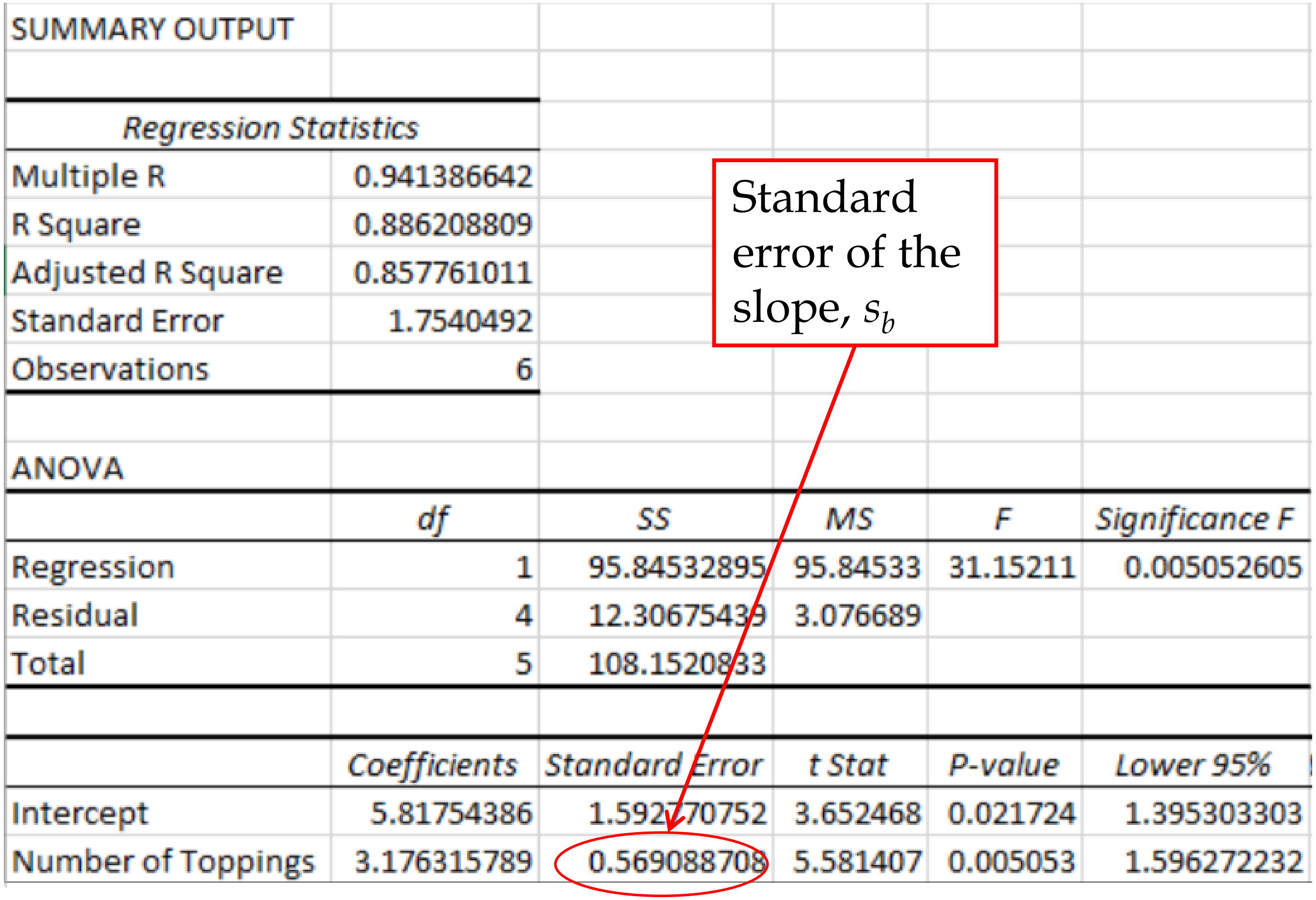
- The regression slope would vary if separate regressions were estimated with several sets of samples from the population


Small s_b


Large s_b

The Standard Error of the Slope (s_b)

Pizza example:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.941386642
R Square	0.886208809
Adjusted R Square	0.857761011
Standard Error	1.7540492
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	95.84532895	95.84533	31.15211	0.005052605
Residual	4	12.30675439	3.076689		
Total	5	108.1520833			

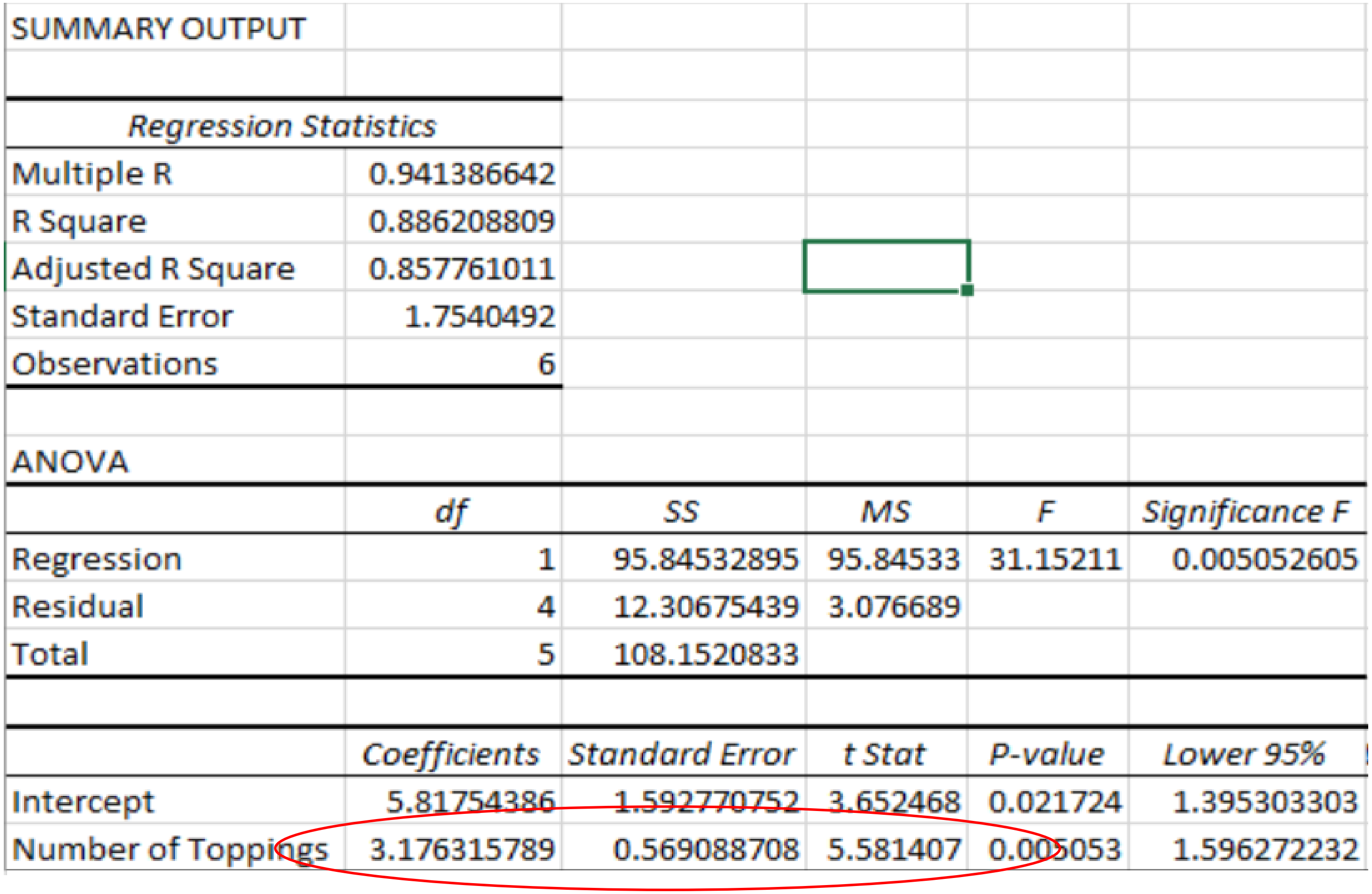
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	5.81754386	1.592770752	3.652468	0.021724	1.395303303
Number of Toppings	3.176315789	0.569088708	5.581407	0.005053	1.596272232

Standard error of the slope, s_b

Testing the Significance of the Slope of the Regression Equation

Knowing where to find s_b in the Excel output, we can easily calculate the test statistic. But Excel will simplify the analysis:

Pizza example:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.941386642
R Square	0.886208809
Adjusted R Square	0.857761011
Standard Error	1.7540492
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	95.84532895	95.84533	31.15211	0.005052605
Residual	4	12.30675439	3.076689		
Total	5	108.1520833			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	5.81754386	1.592770752	3.652468	0.021724	1.395303303
Number of Toppings	3.176315789	0.569088708	5.581407	0.005053	1.596272232

Testing the Significance of the Slope of the Regression Equation

Pizza example: **Use either the *p*-value from the regression output (easy!)** or the critical value (compute manually ☹)

$H_0 : \beta_1 = 0$ ⇒ Two tail test

⇒ Split α evenly between two tails

$H_1 : \beta_1 \neq 0$ ⇒ Find $t_{\alpha/2}$ with $df = n - 2 = 6 - 2 = 4$

If $\alpha = 0.05 \Rightarrow$ the critical *t*-value is $t_{\alpha/2} = 2.776$

Since $t = 5.58 > t_{\alpha/2} = 2.776$ (or, **from the output**, *p*-value = 0.005 < α), we **reject H_0** . Therefore, there is enough evidence to conclude that the population regression slope **is not equal to zero** OR that **there is A relationship** between the number of toppings and the price of a pizza

Testing Hypothesis about the Slope of the Regression Equation

A Word of Caution:

Generally, one-tail and two tail tests are possible with any hypothesized value of the population slope:

- For example, you might want to test

$$H_0 : \beta_1 \geq 5$$
$$H_1 : \beta_1 < 5$$

In this case, values from the regression output will not work:

- If hypothesized value is different from 0, recalculate the *t* test statistic using the standard error of the slope from the Excel output
- Recalculate the *p*-value using the *t* distribution with $df = n - k - 1$

A Short Guide to Writing an Empirical Paper

Things to pay attention to:

1. Structure of the paper:
   - How the paper is divided into sections
   - How each section serves a distinct purpose
2. Presentation of the descriptive statistics and empirical results of the paper

Structure of an Empirical Paper

Typically, we include sections such as:

1. Introduction
2. Data
3. Empirical Results
4. Conclusions

Introduction of the Paper

Conveys several things:

- What is the question that the paper asks?
- Why is the question important?
- How is the paper going to answer the question?
- How is the paper related to the existing work?

Note: The introduction is the most important part of any paper. No one will continue to read the paper if the introduction is confusing or poorly written.

Data Section

Tries to accomplish the following:

- State the sources of data
- Discuss the variables used and how they relate to the concepts that they are supposed to measure
- Present the data’s descriptive statistics:
  - Graphs or numerical measures, or both
  - Written discussion to make the data “come alive”

Descriptive Statistics

Graphs:

- Scatter plots: to show the relationship between variables
- Histograms: to illustrate distributions of variables
- Line charts: to talk about dynamics of a variable over time
- Other less standard graphs that help to make a point

Graphs, Example

"Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors" By Bertrand M., Goldin C., and Katz L.F.

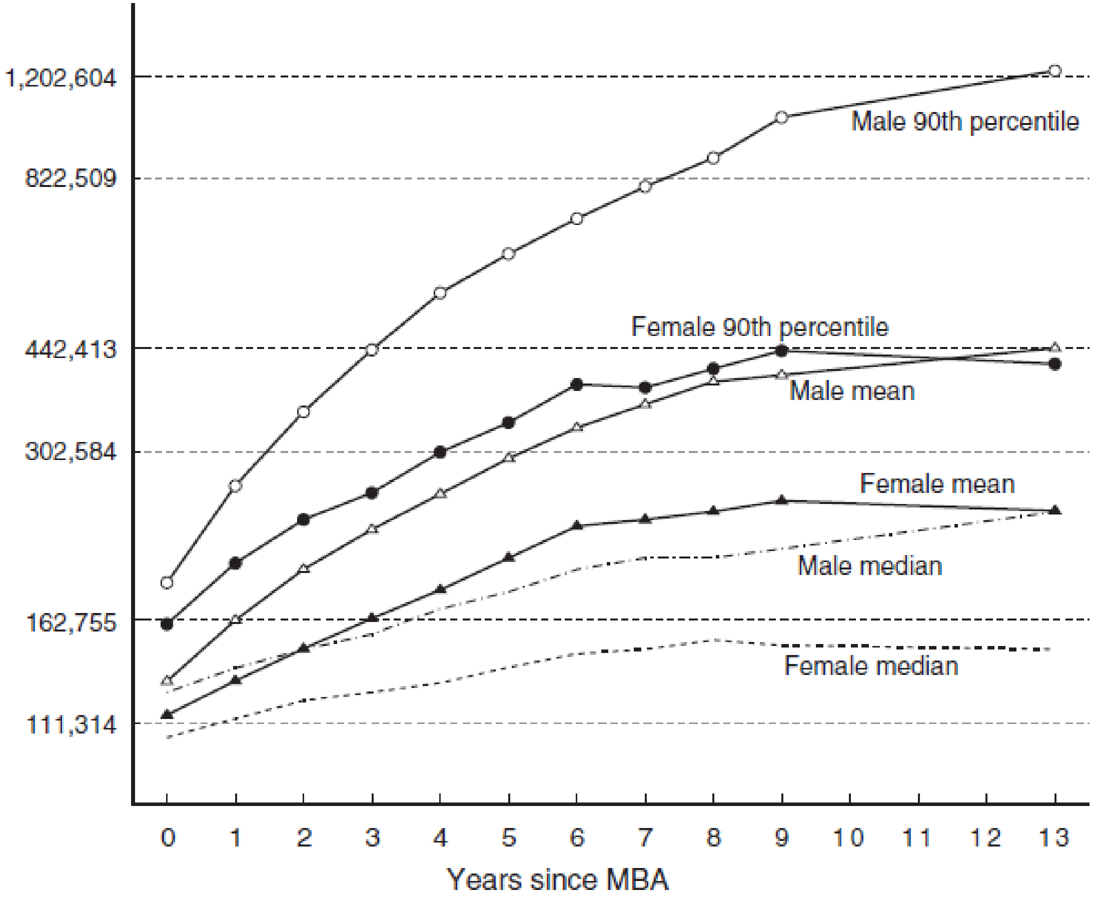


Figure 1. Male and Female Mean, Median, and Ninetieth Percentile Annual Salaries (2006 Dollars) by Years since MBA

Table, Example I

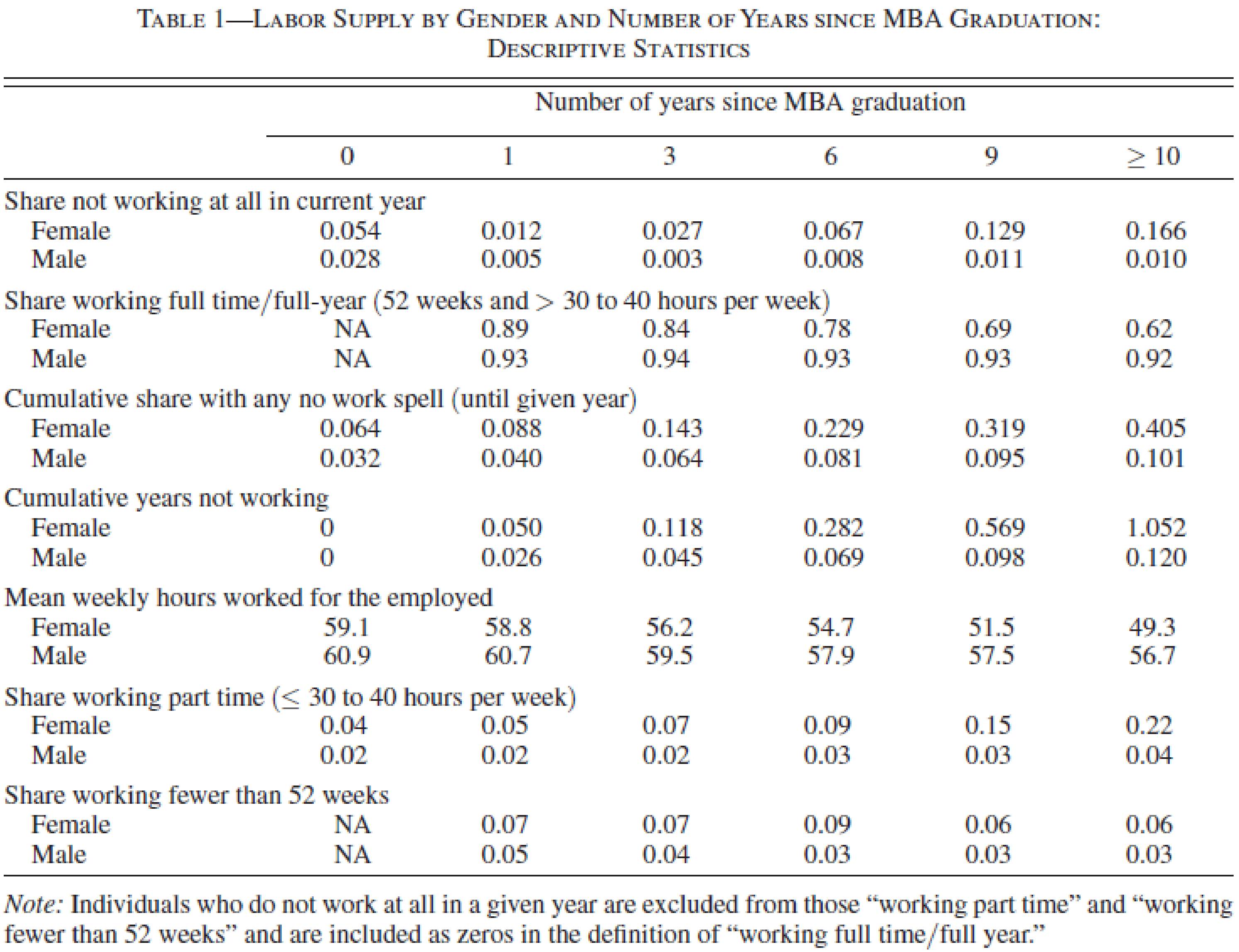


Table 1—Labor Supply by Gender and Number of Years since MBA Graduation: Descriptive Statistics

	Number of years since MBA graduation					
	0	1	3	6	9	≥ 10
Share not working at all in current year						
Female	0.054	0.012	0.027	0.067	0.129	0.166
Male	0.028	0.005	0.003	0.008	0.011	0.010
Share working full time/full-year (52 weeks and > 30 to 40 hours per week)						
Female	NA	0.89	0.84	0.78	0.69	0.62
Male	NA	0.93	0.94	0.93	0.93	0.92
Cumulative share with any no work spell (until given year)						
Female	0.064	0.088	0.143	0.229	0.319	0.405
Male	0.032	0.040	0.064	0.081	0.095	0.101
Cumulative years not working						
Female	0	0.050	0.118	0.282	0.569	1.052
Male	0	0.026	0.045	0.069	0.098	0.120
Mean weekly hours worked for the employed						
Female	59.1	58.8	56.2	54.7	51.5	49.3
Male	60.9	60.7	59.5	57.9	57.5	56.7
Share working part time (≤ 30 to 40 hours per week)						
Female	0.04	0.05	0.07	0.09	0.15	0.22
Male	0.02	0.02	0.02	0.03	0.03	0.04
Share working fewer than 52 weeks						
Female	NA	0.07	0.07	0.09	0.06	0.06
Male	NA	0.05	0.04	0.03	0.03	0.03

Note: Individuals who do not work at all in a given year are excluded from those “working part time” and “working fewer than 52 weeks” and are included as zeros in the definition of “working full time/full year.”

Table, Example II

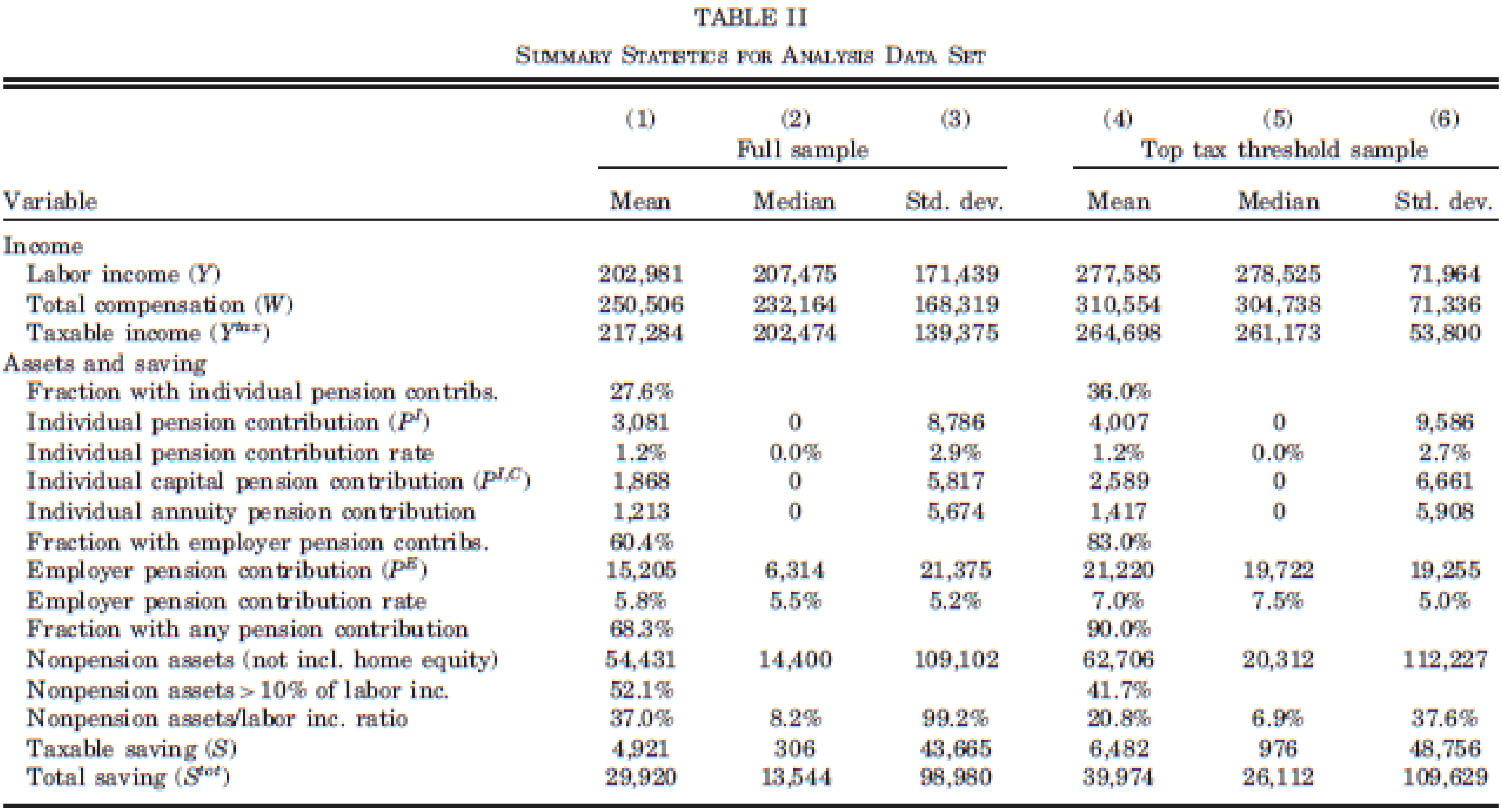


TABLE II

SUMMARY STATISTICS FOR ANALYSIS DATA SET

	(1)	(2) Full sample	(3)	(4)	(5) Top tax threshold sample	(6)
Variable	Mean	Median	Std. dev.	Mean	Median	Std. dev.
Income						
Labor income (Y)	202,981	207,475	171,439	277,585	278,525	71,964
Total compensation (W)	250,506	232,164	168,319	310,554	304,738	71,336
Taxable income (Y^{tax})	217,284	202,474	139,375	264,698	261,173	53,800
Assets and saving						
Fraction with individual pension contribs.	27.6%			36.0%		
Individual pension contribution (P^{I})	3,081	0	8,786	4,007	0	9,586
Individual pension contribution rate	1.2%	0.0%	2.9%	1.2%	0.0%	2.7%
Individual capital pension contribution ($P^{I,C}$)	1,868	0	5,817	2,589	0	6,661
Individual annuity pension contribution	1,213	0	5,674	1,417	0	5,908
Fraction with employer pension contribs.	60.4%			83.0%		
Employer pension contribution (P^{E})	15,205	6,314	21,375	21,220	19,722	19,255
Employer pension contribution rate	5.8%	5.5%	5.2%	7.0%	7.5%	5.0%
Fraction with any pension contribution	68.3%			90.0%		
Nonpension assets (not incl. home equity)	54,431	14,400	109,102	62,706	20,312	112,227
Nonpension assets > 10% of labor inc.	52.1%			41.7%		
Nonpension assets/labor inc. ratio	37.0%	8.2%	99.2%	20.8%	6.9%	37.6%
Taxable saving (S)	4,921	306	43,665	6,482	976	48,756
Total saving (S^{tot})	29,920	13,544	98,980	39,974	26,112	109,629

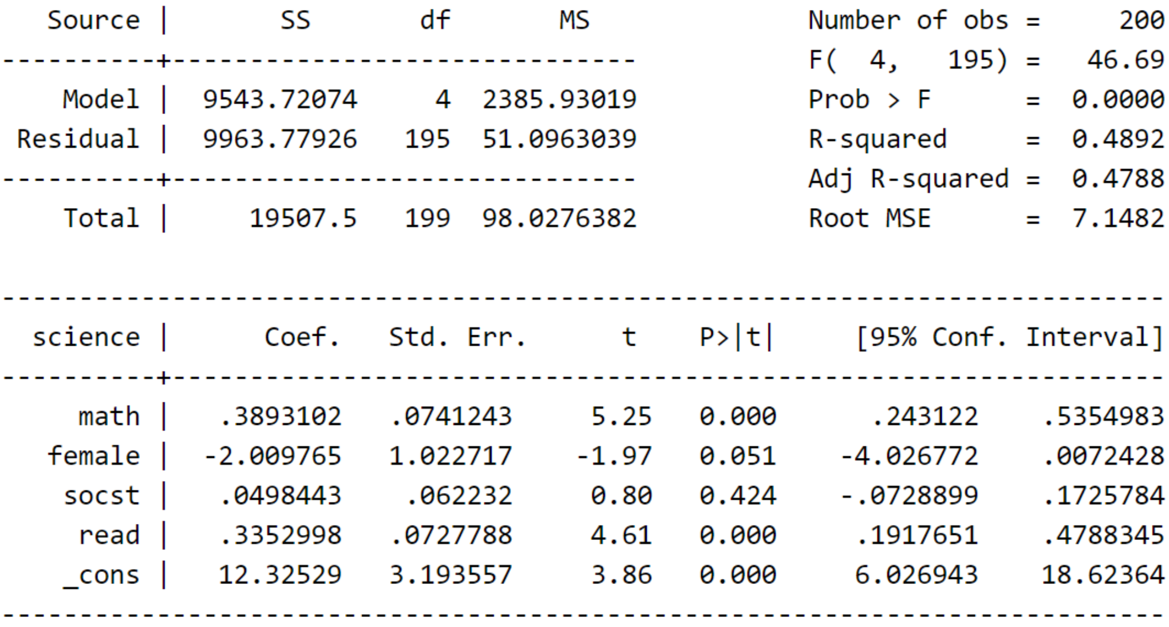
Additional Notes

1. Acknowledge the shortcomings of your data
2. Round the numbers in the tables and the text, and take into account:
   - Units of measurement
     - For example, if you round a faction to 0.5, this may hide a meaningful difference between 52% and 54% for two variables
   - Consider rounding that is easily perceived by a reader
     - For example, percentages may be rounded to a whole number if you are not trying to highlight presence of a small difference between two variables

Empirical Results Section

Different software produce regression outputs similar to Excel

Example from Stata:

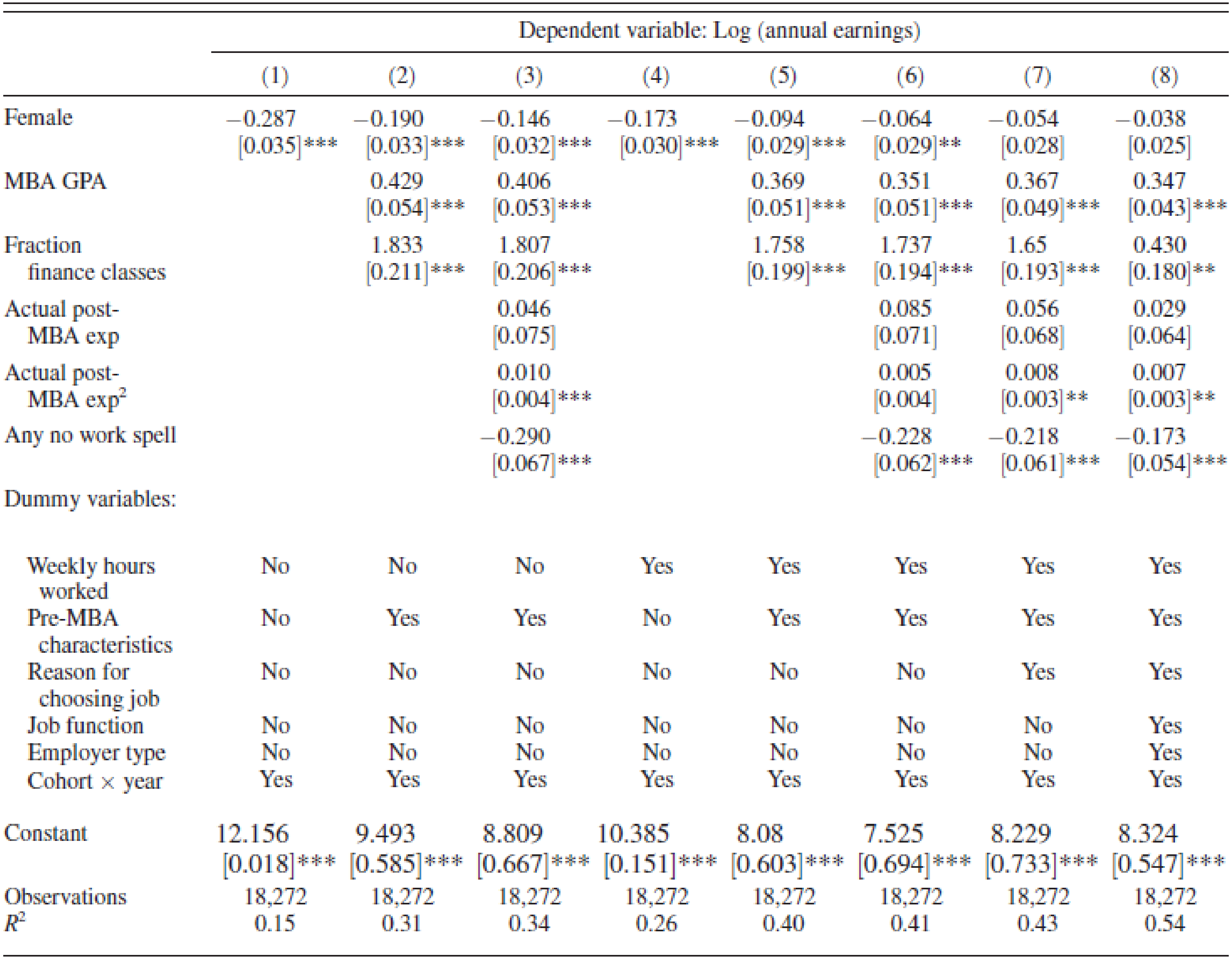


```
    Source |       SS       df       MS              Number of obs =     200
-----------+------------------------------           F(  4,   195) =   46.69
     Model |  9543.72074     4  2385.93019           Prob > F      =  0.0000
  Residual |  9963.77926   195  51.0963039           R-squared     =  0.4892
-----------+------------------------------           Adj R-squared =  0.4788
     Total |     19507.5   199  98.0276382           Root MSE      =  7.1482

------------------------------------------------------------------------------
   science |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
      math |   .3893102   .0741243     5.25   0.000      .243122    .5354983
    female |  -2.009765   1.022717    -1.97   0.051    -4.026772    .0072428
     socst |   .0498443    .062232     0.80   0.424    -.0728899    .1725784
      read |   .3352998   .0727788     4.61   0.000     .1917651    .4788345
     _cons |   12.32529   3.193557     3.86   0.000     6.026943    18.62364
------------------------------------------------------------------------------
```

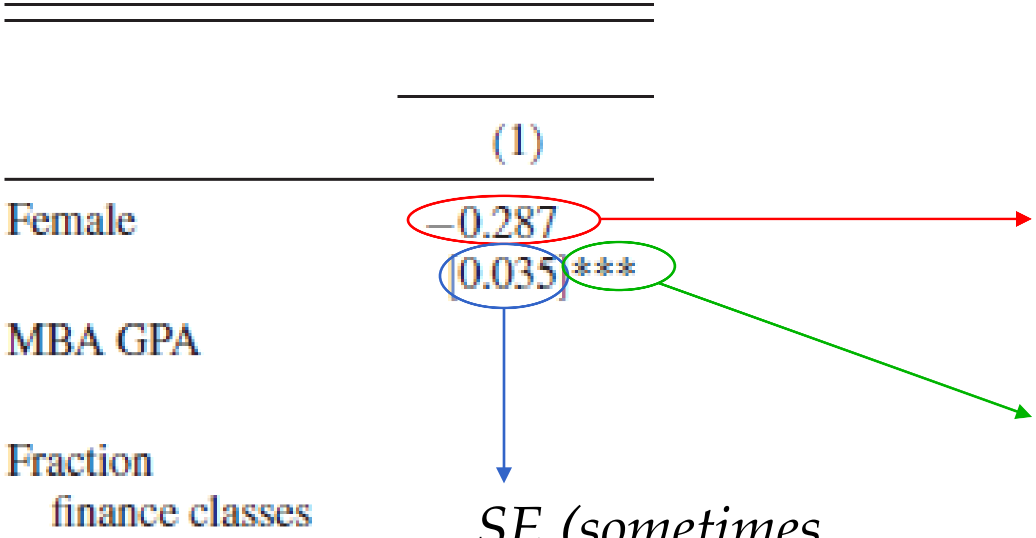
Do NOT include regression output generated by the software in the paper!

Good Presentation Style For the Regression



	Dependent variable: Log (annual earnings)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	−0.287	−0.190	−0.146	−0.173	−0.094	−0.064	−0.054	−0.038
	[0.035]***	[0.033]***	[0.032]***	[0.030]***	[0.029]***	[0.029]**	[0.028]	[0.025]
MBA GPA		0.429	0.406		0.369	0.351	0.367	0.347
		[0.054]***	[0.053]***		[0.051]***	[0.051]***	[0.049]***	[0.043]***
Fraction finance classes		1.833	1.807		1.758	1.737	1.65	0.430
		[0.211]***	[0.206]***		[0.199]***	[0.194]***	[0.193]***	[0.180]**
Actual post-MBA exp			0.046			0.085	0.056	0.029
			[0.075]			[0.071]	[0.068]	[0.064]
Actual post-MBA exp^2			0.010			0.005	0.008	0.007
			[0.004]***			[0.004]	[0.003]**	[0.003]**
Any no work spell			−0.290			−0.228	−0.218	−0.173
			[0.067]***			[0.062]***	[0.061]***	[0.054]***
Dummy variables:								
Weekly hours worked	No	No	No	Yes	Yes	Yes	Yes	Yes
Pre-MBA characteristics	No	Yes	Yes	No	Yes	Yes	Yes	Yes
Reason for choosing job	No	No	No	No	No	No	Yes	Yes
Job function	No	No	No	No	No	No	No	Yes
Employer type	No	No	No	No	No	No	No	Yes
Cohort × year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	12.156	9.493	8.809	10.385	8.08	7.525	8.229	8.324
	[0.018]***	[0.585]***	[0.667]***	[0.151]***	[0.603]***	[0.694]***	[0.733]***	[0.547]***
Observations	18,272	18,272	18,272	18,272	18,272	18,272	18,272	18,272
R^2	0.15	0.31	0.34	0.26	0.40	0.41	0.43	0.54

Regression Results Table: Closer Look



	(1)
Female	−0.287
	[0.035]***
MBA GPA	
Fraction finance classes	

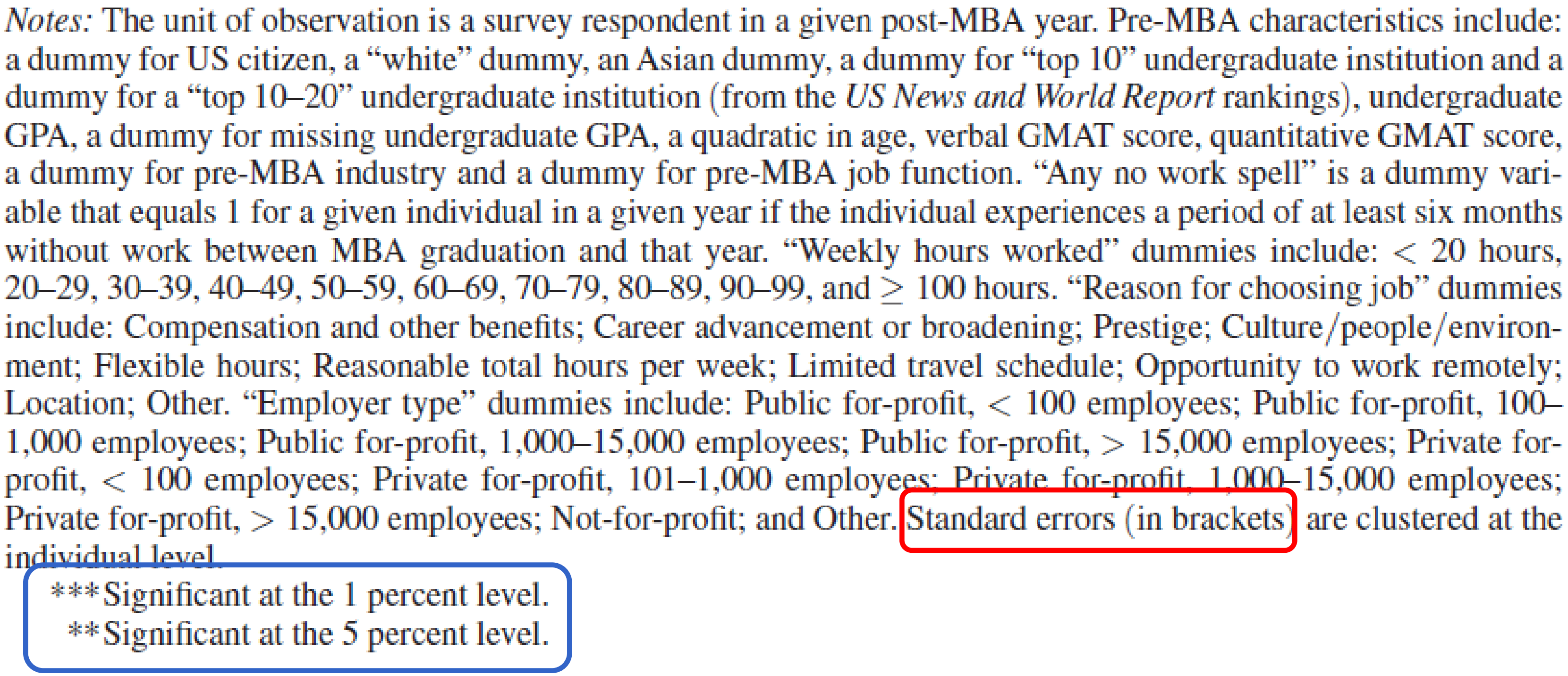
Estimated regression coefficient

SE (sometimes, t-statistic or the p-value. Read footnotes!)

*Asterisks are used to report significance of the coefficient. Here, *** means "Significant at 1%" (to be safe, always read footnotes to know what asterisks mean).*

This tells us that the p-value for this coefficient is less than 0.01 and we can reject H_0 at 1% significance level (even without seeing the exact p-value).

Importance of Footnotes



Notes: The unit of observation is a survey respondent in a given post-MBA year. Pre-MBA characteristics include: a dummy for US citizen, a "white" dummy, an Asian dummy, a dummy for "top 10" undergraduate institution and a dummy for a "top 10–20" undergraduate institution (from the *US News and World Report* rankings), undergraduate GPA, a dummy for missing undergraduate GPA, a quadratic in age, verbal GMAT score, quantitative GMAT score, a dummy for pre-MBA industry and a dummy for pre-MBA job function. "Any no work spell" is a dummy variable that equals 1 for a given individual in a given year if the individual experiences a period of at least six months without work between MBA graduation and that year. "Weekly hours worked" dummies include: $<$ 20 hours, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99, and $\geq$ 100 hours. "Reason for choosing job" dummies include: Compensation and other benefits; Career advancement or broadening; Prestige; Culture/people/environment; Flexible hours; Reasonable total hours per week; Limited travel schedule; Opportunity to work remotely; Location; Other. "Employer type" dummies include: Public for-profit, $<$ 100 employees; Public for-profit, 100–1,000 employees; Public for-profit, 1,000–15,000 employees; Public for-profit, $>$ 15,000 employees; Private for-profit, $<$ 100 employees; Private for-profit, 101–1,000 employees; Private for-profit, 1,000–15,000 employees; Private for-profit, $>$ 15,000 employees; Not-for-profit; and Other. Standard errors (in brackets) are clustered at the individual level.

***Significant at the 1 percent level.
**Significant at the 5 percent level.

Be careful reading footnotes for the tables because there are no strict rules for notation and papers may use different ways to present their results!

Conclusions and Take-Aways

Conclusions section of an empirical paper aims to:

1. Summarize the results
2. Explore the implications of the results
3. Point to future research

Main take-aways. When you write a paper:

- Describe the data / include descriptive statistics
- Do NOT include the entire regression output!
- Present regression results in a compact way (table/equation)
- Report SEs in parenthesis under the estimated coefficients
- Use additional notation (e.g. asterisks) to report significance at 1%, 5%, and 10% significance levels

Regression Analysis: Example

Superintendent: Hire more teachers?

- Trade-off:
  - (*Advantage*) Smaller classes and more individualized attention to the students
  - (*Disadvantage*) Need to spend more money
- Need more information to make an informed decision
  - **If she hires more teachers and reduces class sizes, what will the effect be on test scores?**

→ Need a quantitative assessment

Regression Model

Assuming *linear* relationship between test scores and class size:

Intercept

Slope

Error term
(other factors)

$$TestScores = \beta_0 + \beta_1 \times STR + \varepsilon$$

Dependent variable *y*
(average test scores
for the district)

Independent variable *x*
(student-teacher ratio:
average number of
students per teacher)

How is the Regression Model Helpful?

Main Question:

If the class size is reduced, what will the effect be on test scores?

$$TestScores = \beta_0 + \boxed{\beta_1} \times STR + \varepsilon$$

Can be used to answer the question!

β_1: Change in test scores for a unit change in STR

(for an increase in the average class size by 1 student)

Estimation: Descriptive Statistics, I

Next Step: Use the sample data to estimate the unknown slope β_1 and intercept β_0

Data:

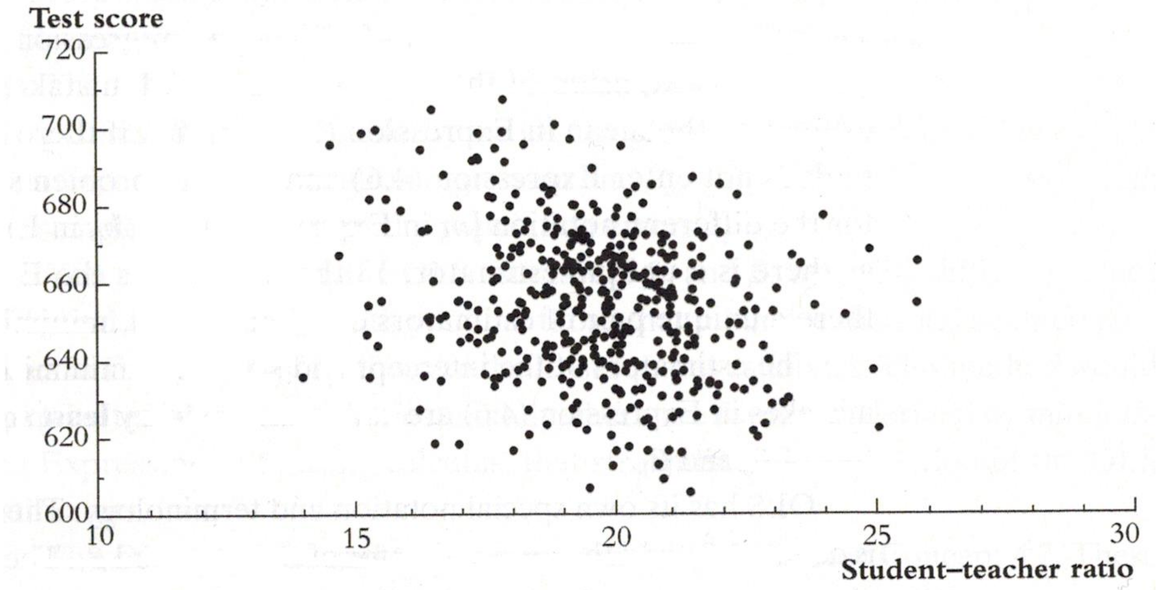
- 420 California school districts
- Test score and STR for each district

Descriptive statistics:

	Mean	**SD**	**Percentile**				
			10%	**40%**	**50%**	**60%**	**90%**
STR	19.6	1.9	17.3	19.3	19.7	20.1	21.9
Test Scores	654.2	19.1	630.4	649.1	654.5	659.4	679.1

Estimation: Descriptive Statistics, II

Scatter plot:



- Weak relationship between test scores and class size
- There are other determinants of test scores (besides class size)

Estimation

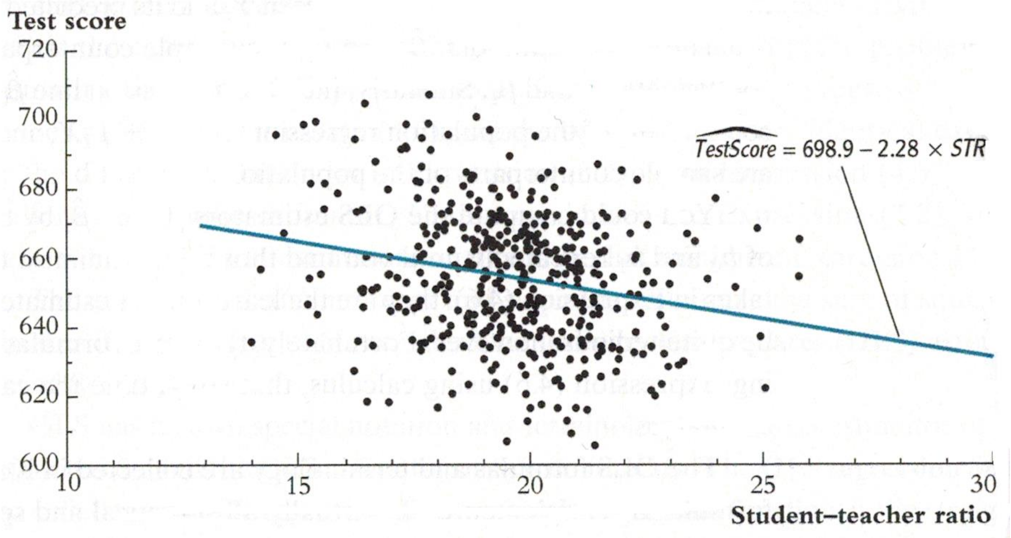
Estimating the regression model in Excel, we obtain:

- $b_0 = 698.9$
- $b_1 = -2.28$

Estimated regression equation:

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

Scatter plot and regression line:



Interpretation of the Results

Estimated regression equation:

$$\widehat{TestScores} = 698.9 \underbrace{- 2.28}_{} \times STR$$

$b_1 = -2.28$: *Increase* in STR by one student per class is, on average, associated with a *decline* in districtwide test scores by 2.28 points

Quantitative Assessment and Predictions

Estimated regression equation:

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

What happens if we are trying to **reduce** *STR* by 2?

$$(-2) \times (-2.28) = 4.56$$

Decrease in *STR*

Effect of a unit change in *STR*

Test scores are predicted to *increase* by 4.56

Predicted test score for a district with 20 students per teacher (*STR* = 20):

$$\widehat{TestScores} = 698.9 - 2.28 \times 20 = 653.3$$

Making Sense of Quantitative Results

Is the change in the test scores predicted by the estimated regression *large* or *small*?

Let's consider a district at the **median**:

	Mean	**SD**	**Percentile**				
			10%	**40%**	**50%**	**60%**	**90%**
STR	19.6	1.9	17.3	19.3	19.7	20.1	21.9
Test Scores	654.2	19.1	630.4	649.1	654.5	659.4	679.1

STR: decreases by 2

- It goes down from 19.7 to 17.7
- This moves the *STR* from 50th to almost 10th percentile

Test Scores: predicted to increase by 4.56

- It increases test scores from 654.5 to 659.1
- This moves test scores from 50th to almost 60th percentile

Other Regression Statistics

$R^2 = 0.051$

Independent variable *STR* explains 5.1% of the variation in the dependent variable *TestScores*

$s_e = 18.6$

The standard deviation of the residuals is 18.6 which means that there is a large spread of the scatterplot around the estimated regression line

Conclusion: other important factors influence the test scores that we left out in our regression

Additional Situation

Situation: A taxpayer claims that class size has
NO effect on test scores

Q: Can we reject the taxpayer's claim?

In regression language:

Q: Can we reject that $\beta_1 = 0$?

Estimated regression equation:

$$\widehat{TestScores} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR, \qquad R^2 = 0.051$$

Significance of the Slope

Estimated regression equation:

$$\widehat{TestScores} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR, \qquad R^2 = 0.051$$

Need to test the significance of the slope:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

$$t = \frac{b_1 - \beta_1}{s_1} = \frac{-2.28 - 0}{0.52} = \frac{-2.28}{0.52} = -4.38$$

p-value = 0.00001 (from Excel)

p-value < α (conventional values of α: 0.01, 0.05, 0.1)

Conclusion: reject H_0 and conclude that our sample provides sufficient evidence that class size affects test scores

z-scores to Keep in Mind

Confidence Level: $100(1-\alpha)\%$	Significance Level: $100(\alpha)\%$	Critical z-score: $z_{\alpha/2}$
80%	20%	$z_{0.10} = \pm 1.28$
90%	10%	$z_{0.05} = \pm 1.645$
95%	5%	$z_{0.025} = \pm 1.96$
98%	2%	$z_{0.01} = \pm 2.33$
99%	1%	$z_{0.005} = \pm 2.575$

Excel Time: Exercise 14.40 (Modified)

Suppose that the Internet retailer Buy.com would like to investigate the relationship between the amount of time in minutes a purchaser spends on its Web site and the amount of money he or she spend on an order. The data in the Excel file **Buy.com.xlsx** (*Excel Files* → *Ch 14*) show the relevant data for a random sample of 12 customers.

1. Define independent and dependent variables
2. Construct a scatter plot for these data and discuss your expectations about regression equation (e.g., the sign of the slope coefficient and the coefficient of determination)

Excel Time: Exercise 14.40 (Modified)

3. Calculate the slope and the intercept for the regression equation:
    - Use the Data Analysis tool
    - (*If time permits*) Perform calculations step-by-step to verify that you understand how the formulas work. Verify that you get the same results as in the Data Analysis output.
4. Write down the estimated regression equation
5. Interpret the value of the slope coefficient
6. Using the output generated by the Data Analysis, what are the values of SST, SSR and SSE. Can you explain what these values mean?

Excel Time: Exercise 14.40 (Modified)

7. Calculate the coefficient of determination using values of SST, SSR and SSE.
    - Verify that the calculated coefficient of determination is the same as the one provided in the Data Analysis output
    - How would you interpret the value of the coefficient of determination?
8. Using $\alpha = 0.1$, test the significance of the population coefficient of determination.
    - Calculate the value of the F statistic and verify that it is the same as reported in the Data Analysis output
    - Use the p-value approach (Data Analysis output can be used here)
    - (*If time permits*) Use the critical value approach (function F.INV.RT())

Excel Time: Exercise 14.40 (Modified)

9. Using $\alpha = 0.1$, test the significance of the slope.
   - Calculate the value of the *t* statistic and verify that it is the same as reported in the Data Analysis output
   - Use the *p*-value approach (Data Analysis output can be used here)
   - (*If time permits*) Use the critical value approach (function T.INV.2T())

10. (*More Advanced*) Using $\alpha = 0.1$, test the hypothesis that the slope is greater than 3 (let's say that you are interested to see if an additional minute on the web site increases the order by *more* than $3).
    - Calculate the value of the *t* statistic
    - Use the *p*-value approach or the critical value approach

Excel Time: Exercise 14.56-14.59 (Modified)

As a measure of productivity, Verizon Wireless records the number of customers each of its retail employees activates weekly. An activation is defined as either a new customer signing a cell phone contract or an existing customer renewing a contract. The data in the Excel file **activations.xlsx** (*Excel Files → Ch 14*) show the number of weekly activations for eight randomly selected employees along with their job-satisfaction levels rated on a scale of 1-10 (10 = Most satisfied). Verizon is interested in studying the relationship between satisfaction level of the employees and the number of activations.

1. Define independent and dependent variables
2. Construct a scatter plot for these data and discuss your expectations about regression equation (e.g., the sign of the slope coefficient and the coefficient of determination)

Excel Time: Exercise 14.56-14.59 (Continued)

3. Calculate the slope and the intercept for the regression equation:
   - Use Data Analysis add-in
4. Write down the estimated regression equation and interpret the value of the slope coefficient
5. Using the output generated by the Data Analysis, what are the values of SST, SSR and SSE. Can you explain what these values mean?

Excel Time: Exercise 14.56-14.59 (Continued)

6. Calculate the coefficient of determination using values of SST, SSR and SSE.
   - Verify that the calculated coefficient of determination is the same as the one provided in the Data Analysis output.
   - How would you interpret the value of the coefficient of determination?
7. Using $\alpha = 0.01$, test the significance of the population coefficient of determination.
   - Calculate the value of the F statistic and verify that it is the same as reported in the Data Analysis output
   - Use the p-value approach (Data Analysis output can be used here)

Excel Time: Exercise 14.56-14.59 (Continued)

8. Using $\alpha = 0.01$, test the significance of the regression slope.
   - Calculate the value of the *t* statistic and verify that it is the same as reported in the Data Analysis output
   - Use the *p*-value approach (Data Analysis output can be used here)

Excel Time: Exercise 14.60-14.63 (Extra Practice)

The American Board of Family Medicine would like to investigate the theory that the mother's shoe size can be used to predict an infant's birth weight in pounds. The data in **ABFM.xlsx** (*Excel Files → Ch 14*) records the shoe size of 10 mothers and the birth weight of their infants.

1. Define independent and dependent variables
2. Construct a scatter plot for these data and discuss your expectations about regression equation (e.g., the sign of the slope coefficient and the coefficient of determination)

Excel Time: Exercise 14.60-14.63 (Continued)

3. Calculate the slope and the intercept for the estimated regression equation
   - Use Data Analysis add-in
4. Interpret the value of the slope coefficient
5. Using the output from Data Analysis, what are the values of SST, SSR and SSE. Can you explain what these values mean?
6. Calculate the coefficient of determination using values of SST, SSR and SSE.
   - Verify that the calculated coefficient of determination is the same as the one in the Data Analysis output.
   - How would you interpret the coefficient of determination?

Excel Time: Exercise 14.60-14.63 (Continued)

7. Using $\alpha = 0.05$, test the significance of the population coefficient of determination.
   - Calculate the value of the *F* statistic and verify that it is the same as reported in the Data Analysis output
   - Use the *p*-value approach (Data Analysis output can be used here)
8. Using $\alpha = 0.05$, test the significance of the regression slope.
   - Calculate the value of the *t* statistic and verify that it is the same as reported in the Data Analysis output
   - Use the *p*-value approach (Data Analysis output can be used here)