Chapter 14, Part A
Simple Linear Regression

- Simple Linear Regression Model
- Least Squares Method
- Coefficient of Determination
- Model Assumptions
- Testing for Significance

Simple Linear Regression (1 of 6)

- Managerial decisions often are based on the relationship between two or more variables.
- Regression analysis can be used to develop an equation showing how the variables are related.
- The variable being predicted is called the dependent variable and is denoted by *y*.
- The variables being used to predict the value of the dependent variable are called the independent variables and are denoted by *x*.

Simple Linear Regression (2 of 6)

- Simple linear regression involves one independent variable and one dependent variable.
- The relationship between the two variables is approximated by a straight line.
- Regression analysis involving two or more independent variables is called multiple regression.

Simple Linear Regression Model (1 of 5)

- The equation that describes how y is related to x and an error term is called the regression model.
- The simple linear regression model 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 (2 of 5)

- The Simple Linear Regression Equation is:

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

Graph of the regression equation is a straight line.

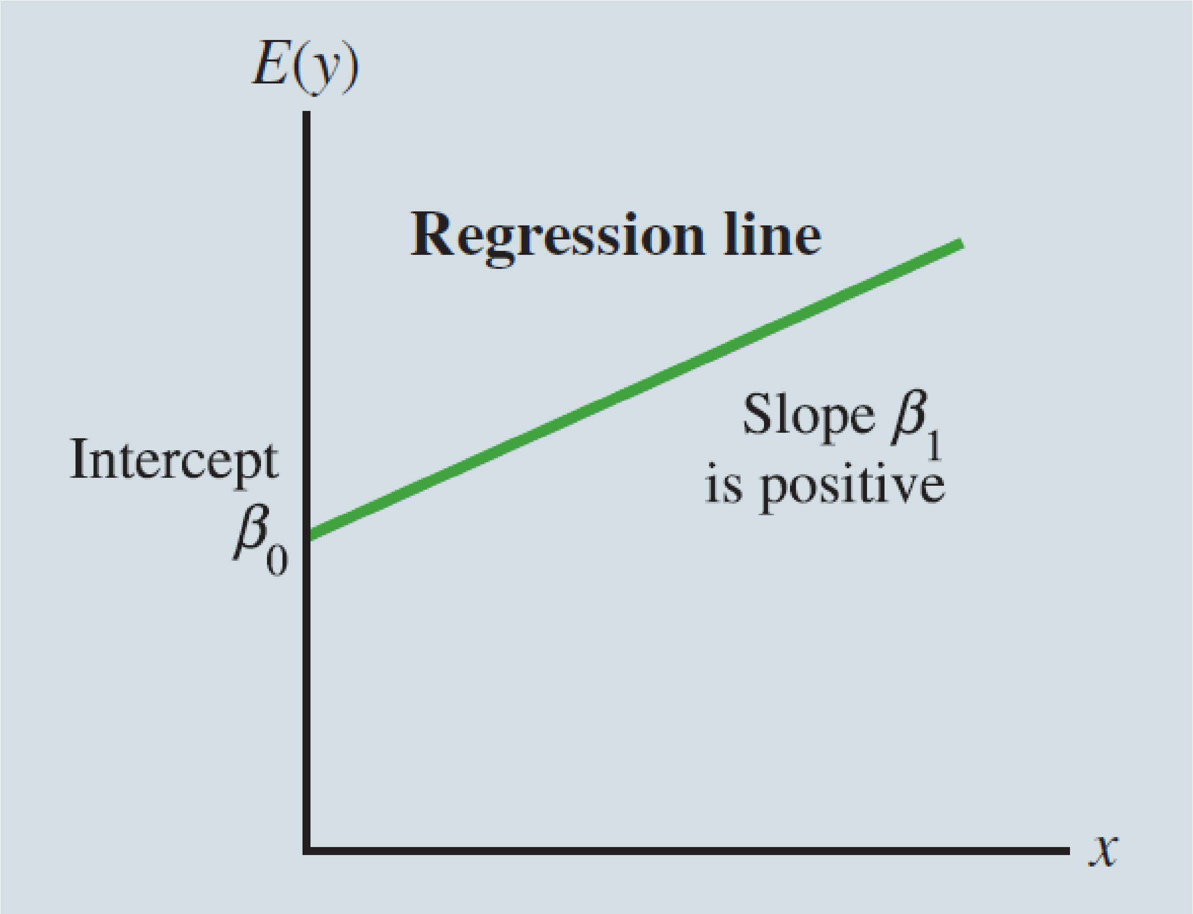
β_0 is the *y* intercept of the regression line.

β_1 is the slope of the regression line.

E(y) is the expected value of *y* for a given *x* value.

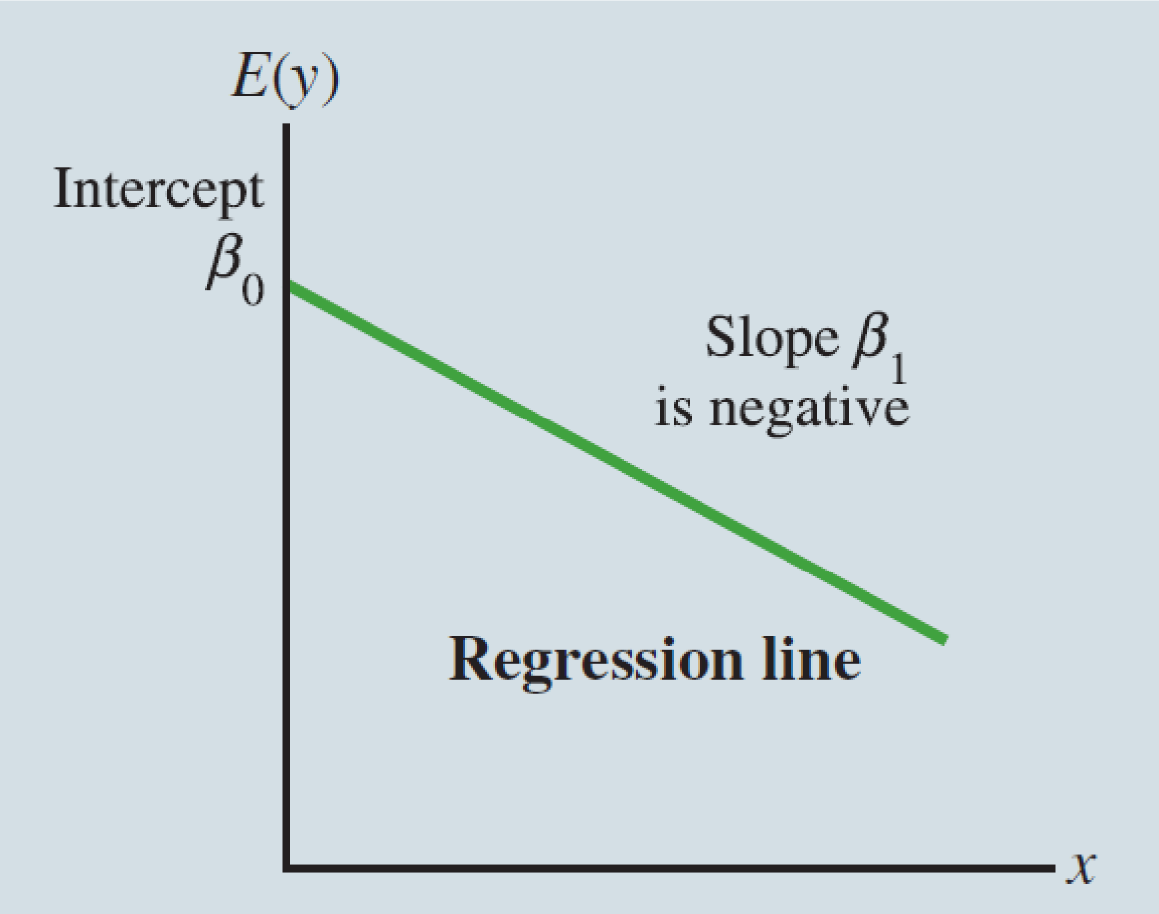
Simple Linear Regression Equation (3 of 5)

- Positive Linear Relationship



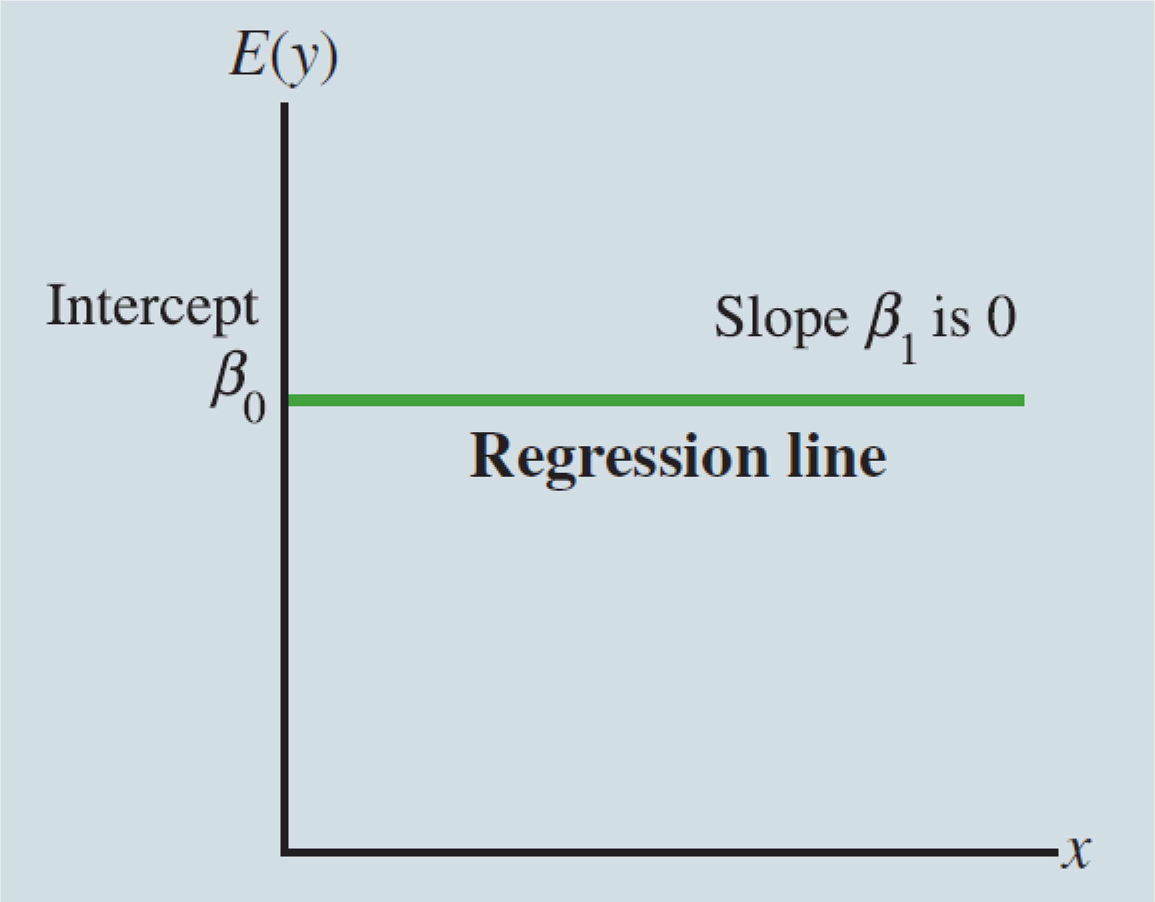
Simple Linear Regression Equation (4 of 5)

- Negative Linear Relationship



Simple Linear Regression Equation (5 of 5)

- No Relationship



Estimated Simple Linear Regression Equation

The estimated simple linear regression equation

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

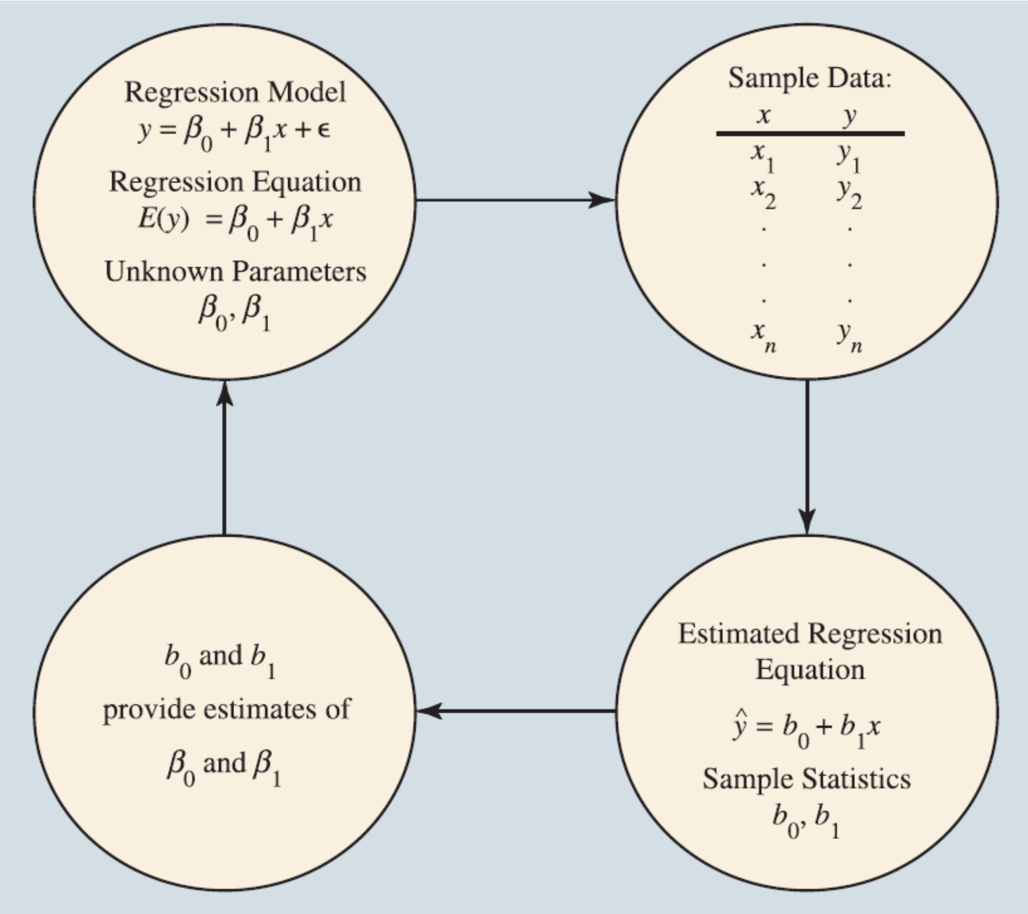
The graph is called the estimated regression line.

b_0 is the y intercept of the line.

b_1 is the slope of the line.

$\hat{y}$ is the estimated value of y for a given x value.

Estimation Process



Least Squares Method (1 of 3)

- Least Squares Criterion

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

where:

y_i = observed value of the dependent variable for the ith observation.

$\hat{y}$ = estimated value of the dependent variable for the ith observation.

Least Squares Method (2 of 3)

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

where:

x_i = value of independent variable for *i*th observation

y_i = value of dependent variable for *i*th observation

$\bar{x}$ = mean value for dependent variable

$\bar{y}$ = mean value for independent variable

Least Squares Method (3 of 3)

- *y*-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1\bar{x}$$

Simple Linear Regression (3 of 6)

Example: Armand's Pizza Parlor Restaurants

Data was collected from a sample of 10 Armand's Pizza Parlor Restaurants near college campuses. For the *i*th observation or restaurant in the sample, x_i is the size of the student population and y_i is the quarterly sales.

Simple Linear Regression (4 of 6)

Example: Armand's Pizza Parlor Restaurants

Restaurant	Student population (1000s)	Quarterly sales ($1000s)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Simple Linear Regression (5 of 6)

Example: Armand's Pizza Parlor Restaurants

i	X_i	Y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	**140**	**1300**			**2840**	**568**

CENGAGE

Simple Linear Regression (6 of 6)

Example: Armand's Pizza Parlor Restaurants

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

- *y*-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

- Estimated Regression Equation

$$\hat{y} = 60 + 5x$$

Estimated Regression Equation

Example: Armand's Pizza Parlor Restaurants

- Excel Worksheet

	A	B	C	D
1	Restaurant	Population	Sales	
2	1	2	58	
3	2	6	105	
4	3	8	88	
5	4	8	118	
6	5	12	117	
7	6	16	137	
8	7	20	157	
9	8	20	169	
10	9	22	149	
11	10	26	202	

Using Excel's Chart Tools for Scatter Diagram & Estimated Regression Equation (1 of 4)

Example: Armand's Pizza Parlor Restaurants

- Producing a Scatter Diagram
  - Step 1 Select cells B2:C11
  - Step 2 Click the **Insert** tab on the Ribbon
  - Step 3 In the Charts group, click **Insert Scatter (X,Y) or Bubble Chart**
  - Step 4 When the list of scatter diagram subtypes appears,

    Click **Scatter** (chart in upper left corner)

Using Excel's Chart Tools for Scatter Diagram & Estimated Regression Equation (2 of 4)

Example: Armand's Pizza Parlor Restaurants

- Editing a Scatter Diagram
  - Step 1 Click the **Chart Title** and replace it with *Armand's Pizza Parlors*
  - Step 2 Click the **Chart Elements** button
  - Step 3 When the list of chart elements appears:

    Click **Axis Titles** (creates placeholders for titles)

    Click **Gridlines** (to deselect gridlines option)

    Click **Trendline**

Using Excel's Chart Tools for Scatter Diagram & Estimated Regression Equation (3 of 4)

Example: Armand's Pizza Parlor Restaurants

- Editing a Scatter Diagram (continued)
  - Step 4 Click the horizontal **Axis Title** and replace it with *Student population (1000s)*
  - Step 5 Click the **Vertical (Value) Axis Title** and replace it with *Quarterly Sales ($1000s)*
  - Step 6 Select the **Format Trendline** option
  - Step 7 When the Format Trendline dialog box appears:

  Select **Display equation on chart**

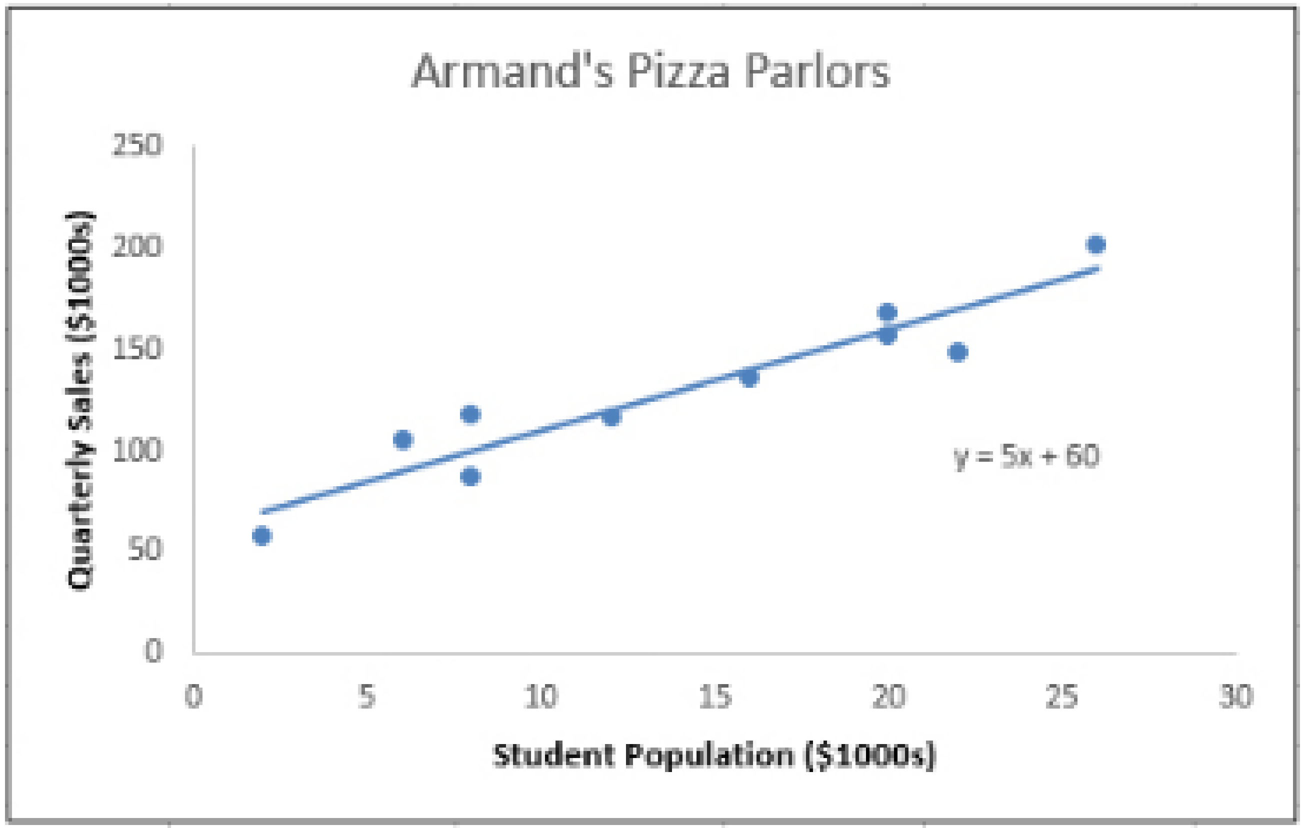
  Click the **Fill & Line** button

  In the **Dash type** box, select **Solid**

  Close the **Format Trendline** dialog box

Using Excel's Chart Tools for Scatter Diagram & Estimated Regression Equation (4 of 4)

Example: Armand's Pizza Parlor Restaurants



Coefficient of Determination (1 of 3)

- Relationship Among SST, SSR, SSE

$$SST = SSR + SSE$$

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

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Coefficient of Determination (2 of 3)

- The <u>coefficient of determination</u> is:

$$r^2 = \frac{\text{SSR}}{\text{SST}}$$

where:

SSR = sum of squares due to regression

SST = total sum of squares

Coefficient of Determination (3 of 3)

Example: Armand's Pizza Parlor Restaurants

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{14,200}{15,730} = .9027$$

The regression relationship is very strong; 90.27% of the variability in the sales can be explained by the linear relationship between the size of the student population and sales.

Using Excel to Compute the Coefficient of Determination (1 of 2)

- Adding r^2 Value to Scatter Diagram
  - Step 1 Right-click on the trendline and select the **Format Trendline** option
  - Step 2 When the Format Trendline dialog box appears:

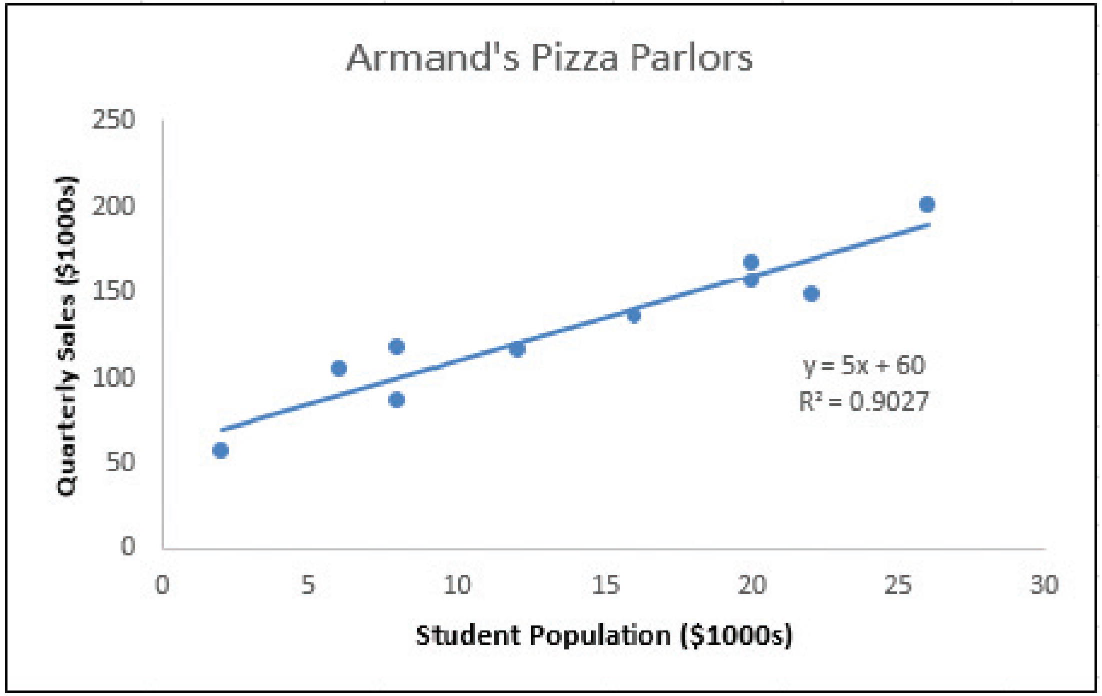
    Select **Display R-squared on chart**

    Close the **Format Trendline** dialog box

Using Excel to Compute the Coefficient of Determination (2 of 2)

<u>Example</u>: Armand’s Pizza Parlor Restaurants

- Adding r^2 Value to Scatter Diagram



Sample Correlation Coefficient (1 of 2)

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of Determination}}$$
$$= (\text{sign of } b_1)\sqrt{r^2}$$

where:

$$b_1 = \text{the slope of the estimated regression equation } \hat{y} = b_0 + b_1 x$$

Sample Correlation Coefficient (2 of 2)

<u>Example</u>: Armand's Pizza Parlor Restaurants

$$r_{xy} = (\text{sign of } b_1)\sqrt{r^2}$$

The sign of b_1 in the equation $\hat{y} = 10 + 5x$ is "+".

$$r_{xy} = +\sqrt{.9027}$$

$$r_{xy} = +.9501$$

CENGAGE

Assumptions About the Error Term ε

Model assumptions

1. The error term ε is a random variable with mean of zero.
2. The variance of ε, denoted by σ^2, is the same for all values of the independent variable.
3. The values of ε are independent.
4. The error term ε is a normally distributed random variable for all values of x.

Testing for Significance (1 of 3)

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.
- Two tests are commonly used:

t Test and *F* test

- Both the *t* test and *F* test require an estimate of σ^2, the variance of ε in the regression model.

Testing for Significance (2 of 3)

- An Estimate of σ^2

The mean square error (MSE) provides the estimate of σ^2, and the notation s^2 is also used.

$$s^2 = \text{MSE} = \frac{\text{SSE}}{(n-2)}$$

where:

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

Testing for Significance (3 of 3)

- An Estimate of σ
  - To estimate σ, we take the square root of s^2.
  - The resulting s is called the standard error of the estimate.

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}}$$

Testing for Significance: *t* Test (1 of 4)

- Hypotheses

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

- Test Statistic

$$t = \frac{b_1}{s_{b_1}} \quad \text{where } s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Testing for Significance: *t* Test (2 of 4)

- Rejection Rule

$$\text{Reject } H_0 \text{ if } p\text{-value} \leq \alpha$$

$$\text{or } t \leq -t_{\alpha/2} \text{ or } t \geq t_{\alpha/2}$$

where:

$t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom

Testing for Significance: *t* Test (3 of 4)

1. Determine the hypotheses.

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

2. Specify the level of significance.

$$\alpha = .01$$

3. Select the test statistic.

$$t = \frac{b_1}{s_{b_1}}$$

4. State the rejection rule.

Reject H_0 if p-value $\leq .01$ or $|t| > 3.355$ (with 2 degrees of freedom)

CENGAGE

Testing for Significance: *t* Test (4 of 4)

5. Compute the value of the test statistic.

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{.5803} = 8.62$$

6. Determine whether to reject H_0.

   $t = 3.355$ provides an area of .005 in the upper tail. Hence the *p*-value is less than .005. Also, $t = 8.62 > 3.355$. We can reject H_0.

Confidence Interval for β_1 (1 of 3)

We can use a 99% confidence interval for β_1 to test the hypotheses just used in the *t* test.

H_0 is rejected if the hypothesized value of β_1 is not included in the confidence interval for β_1.

Confidence Interval for β_1 (2 of 3)

- The form of a confidence interval for β_1 is:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

where:

b_1 is the point estimator,

$t_{\alpha/2} s_{b_1}$ is the margin of error, and

$t_{\alpha/2}$ is the *t* value providing an area of $\alpha/2$ in the upper tail of a *t* distribution with $n-2$ degrees of freedom

Confidence Interval for β_1 (3 of 3)

- Rejection Rule

  Reject H_0 if 0 is not included in the confidence interval for β_1.

- 99% Confidence Interval for β_1.

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3.355(.5803) = 5 \pm 1.95$$

or 3.05 to 6.95

- Conclusion

  0 is not included in the confidence interval. Reject H_0

Testing for Significance: *F* Test (1 of 4)

- Hypotheses

$$H_0: b_1 = 0$$
$$H_a: b_1 \neq 0$$

- Test Statistic

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

Testing for Significance: *F* Test (2 of 4)

- Rejection Rule

$$\text{Reject } H_0 \text{ if } p\text{-value } \leq \alpha \text{ or } F \geq F_\alpha$$

where:

F_α is based on an F distribution with 1 degree of freedom in the numerator and $n - 2$ degrees of freedom in the denominator.

Testing for Significance: *F* Test (3 of 4)

1. Determine the hypotheses.

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

2. Specify the level of significance.

$$\alpha = .01$$

3. Select the test statistic.

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

4. State the rejection rule.

Reject H_0 if *p*-value $\leq .01$ or $F \geq 11.26$ (with 1 d.f. in numerator and 8 d.f. in denominator).

Testing for Significance: *F* Test (4 of 4)

5. Compute the value of the test statistic.

$$F = \frac{MSR}{MSE} = \frac{14,200}{191.25} = 74.25$$

6. Determine whether to reject H_0.

$F = 11.26$ provides an area of .01 in the upper tail. Thus, the *p*-value corresponding to $F = 74.25$ is less than .01. Hence, we reject H_0.

The statistical evidence is sufficient to conclude that a significant relationship exists between the size of the student population and quarterly sales

Some Cautions about the Interpretation of Significance Tests

- Rejecting $H_0: \beta_1 = 0$ and concluding that the relationship between *x* and *y* is significant does not enable us to conclude that a cause-and-effect relationship is present between *x* and *y*.

- Because we are able to reject $H_0: \beta_1 = 0$ and demonstrate statistical significance does not enable us to conclude that there is a linear relationship between *x* and *y*.