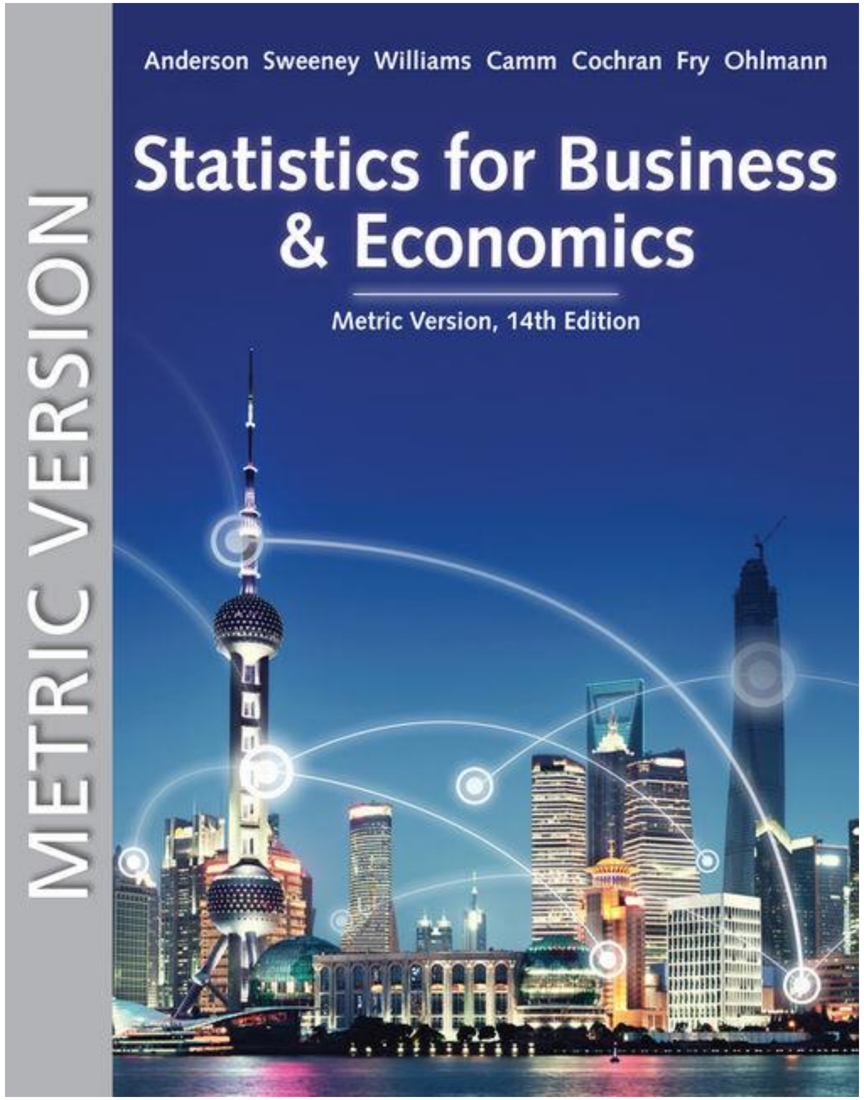
Statistics for Business and Economics (14e) Metric Version

Chapters 14~15 （迴歸分析）


Anderson Sweeney Williams Camm Cochran Fry Ohlmann
Statistics for Business & Economics
Metric Version, 14th Edition
METRIC VERSION

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Chapter 14 - Simple Linear Regression

14.1 - Simple Linear Regression Model

14.2 - Least Squares Method

14.3 - Coefficient of Determination

14.4 - Model Assumptions

14.5 - Testing for Significance

14.6 - Using the Estimated Regression Equation for Estimation and Prediction

14.7 - Computer Solution

14.8 - Residual Analysis: Validating Model Assumptions

14.9 - Residual Analysis: Outliers and Influential Observations

14.10 - Practical Advice: Big Data and Hypothesis Testing in Simple Linear Regression

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

量化模型

透過量化模型描述觀察結果：

觀察現象 = 模型 + 誤差

或是

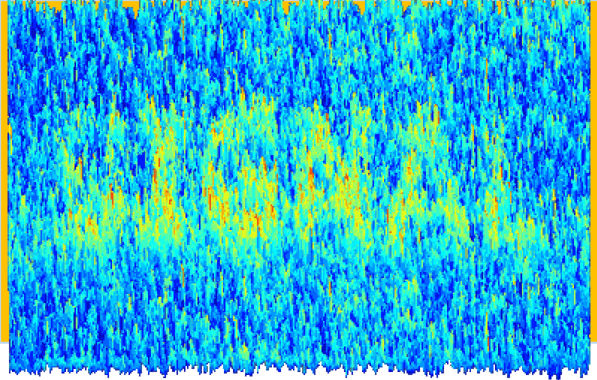
$y = f(x)$ + error；觀察值 = 訊號 + 雜訊

- 數量化模型的關鍵：

→量化目標值 y ：定義問題！

→選取關鍵變數： $x_1, x_2, \ldots, x_p$

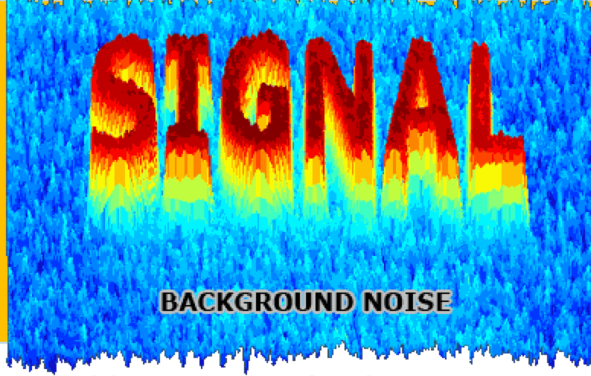
→建立量化模型：統計學習、機器學習。



http://www.ljagilamplighter.com/wp-content/uploads/noise.png

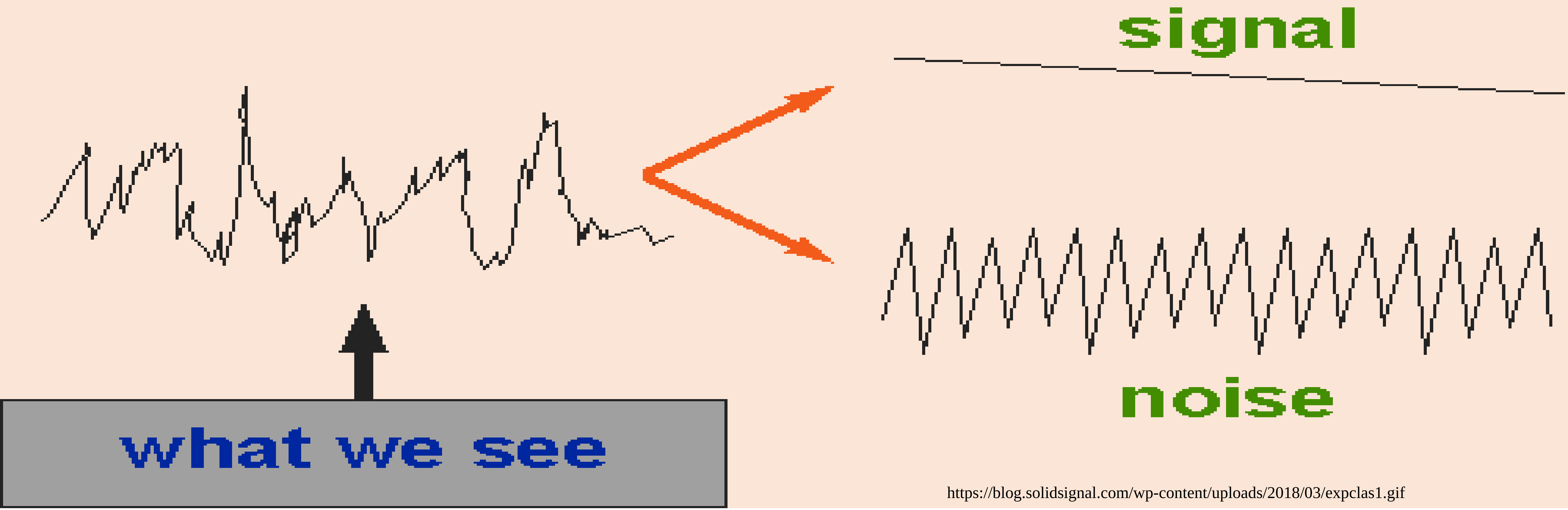
觀察現象 = 模型 + 誤差

Observation = Signal + Noise



http://www.ljagilamplighter.com/wp-content/uploads/noise_signal-mlab1_png_pagespeed_ce_b_GTiE6tAg.png

What we observe can be divided into:



https://blog.solidsignal.com/wp-content/uploads/2018/03/expclas1.gif

Simple Linear Regression (1 of 2)

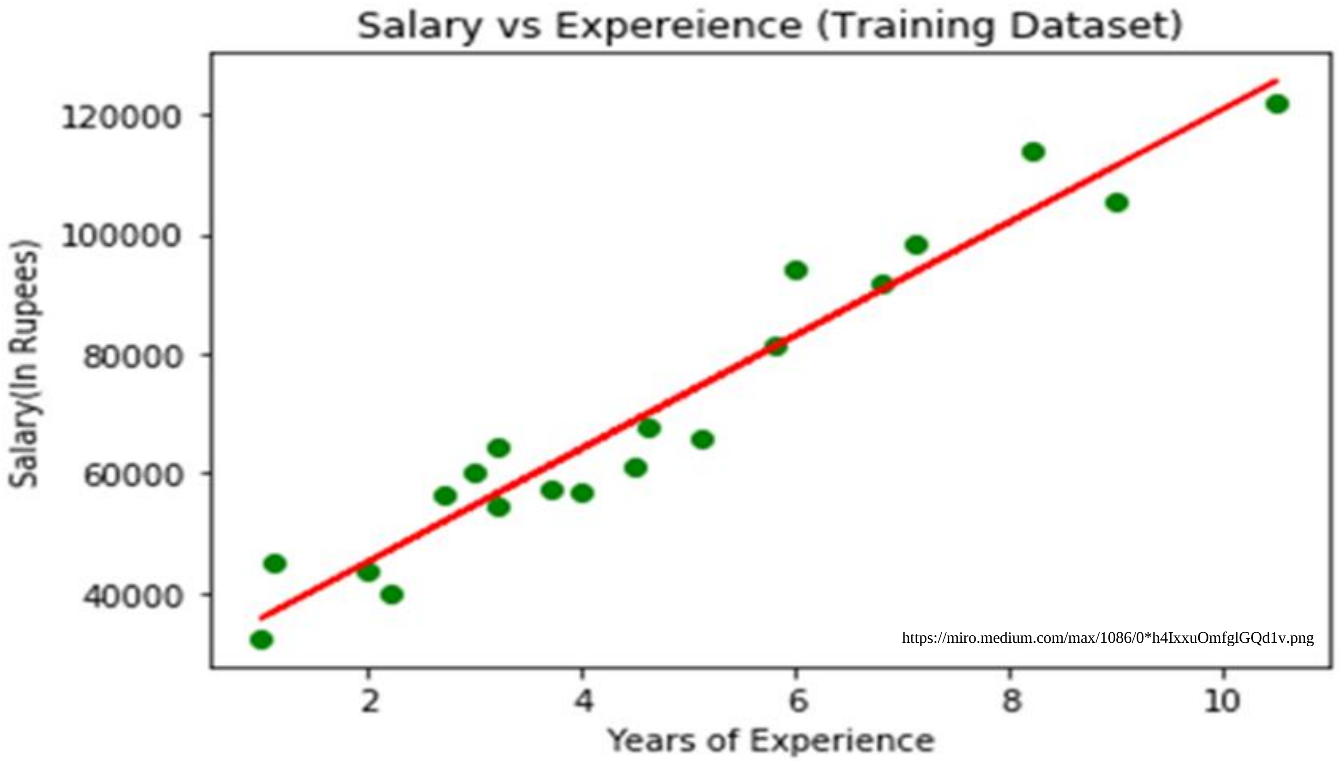
- Managerial decisions often are based on the relationship between two or more variables.
- Regression analysis can be used to develop an equation showing how the variables are related.
- The variable being predicted is called the dependent variable and is denoted by *y*.
- The variables being used to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

註：因變數（應變數、相依變數） vs. 自變數（獨立變數、解釋變數）

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Simple Linear Regression (2 of 2)

- 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.



https://miro.medium.com/max/1086/0*h4IxxuOmfglGQd1v.png

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or i
password-protected website or school-approved learning management system for classroom use.

什麼是迴歸分析？

Regression analysis

FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE *INDEPENDENT* OR *PREDICTOR* VARIABLE, AND y IS THE *DEPENDENT* OR *RESPONSE* VARIABLE. THE *REGRESSION* OR *PREDICTION* LINE HAS THE FORM

$$y = a+bx$$

https://analyticsbuddhu.files.wordpress.com/2018/03/cartoon_guide_regression.png

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 is

$$\mathrm{r} = \beta_1 \times \frac{\sigma_x}{\sigma_y}$$

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

where: β_0 and β_1 are the parameters of the model.

ε is a random variable called the error term.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Simple Linear Regression Equation (1 of 4)

The simple linear regression equation is

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

where: β_0 is the y intercept of the regression line.

β_1 are the slope of the regression line.

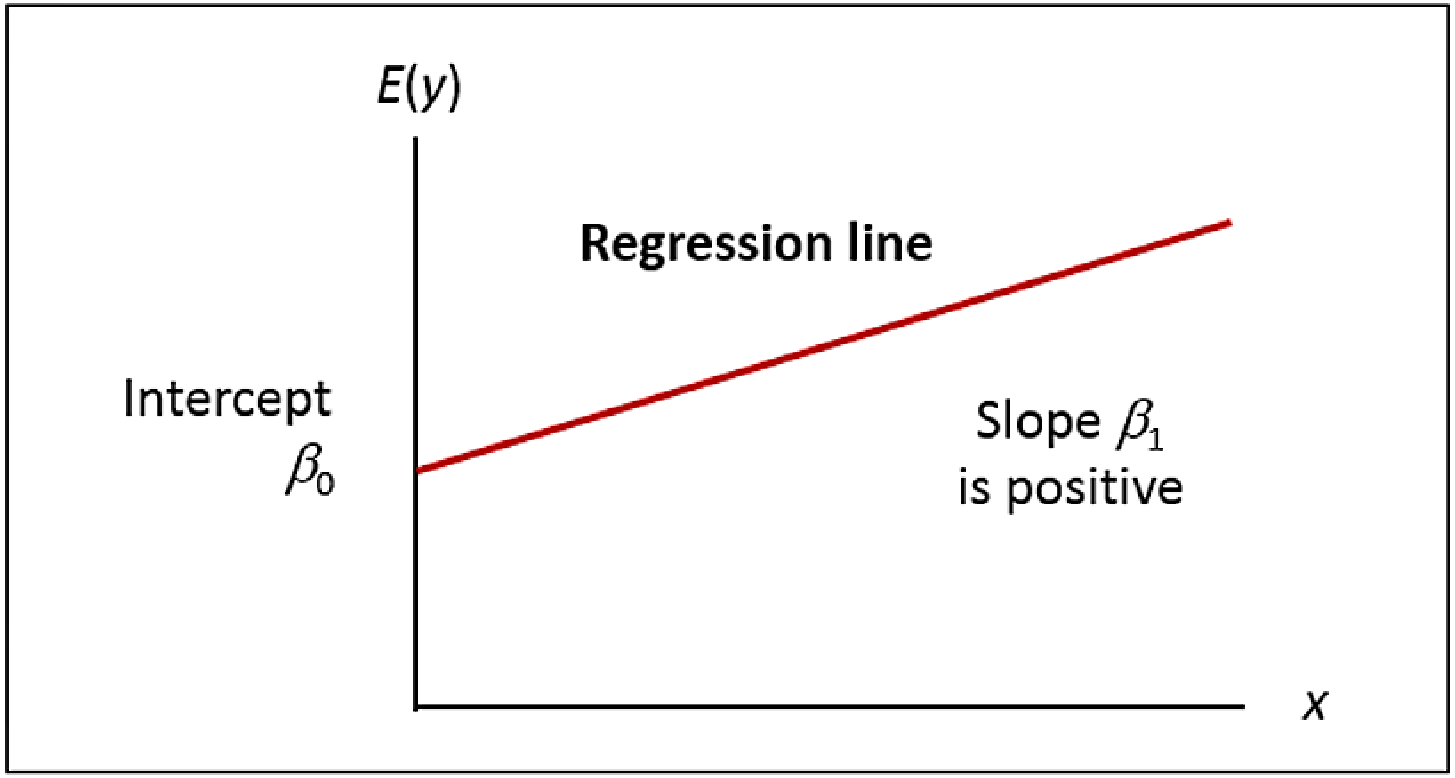
$E(y)$ is the expected value of y for a given x value.

The graph of the regression equation is a straight line.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Simple Linear Regression Equation (2 of 4)

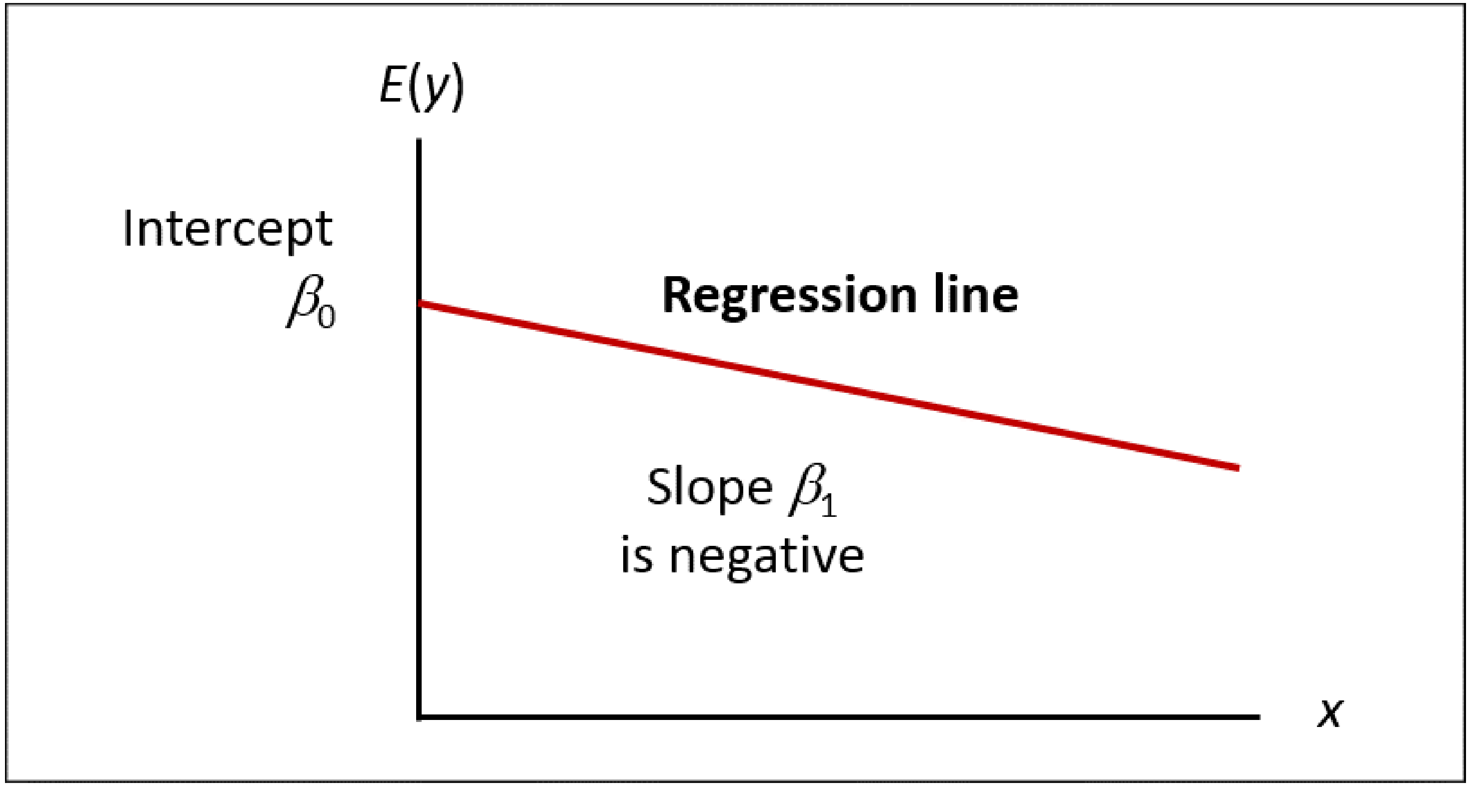
Positive Linear Relationship



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Simple Linear Regression Equation (3 of 4)

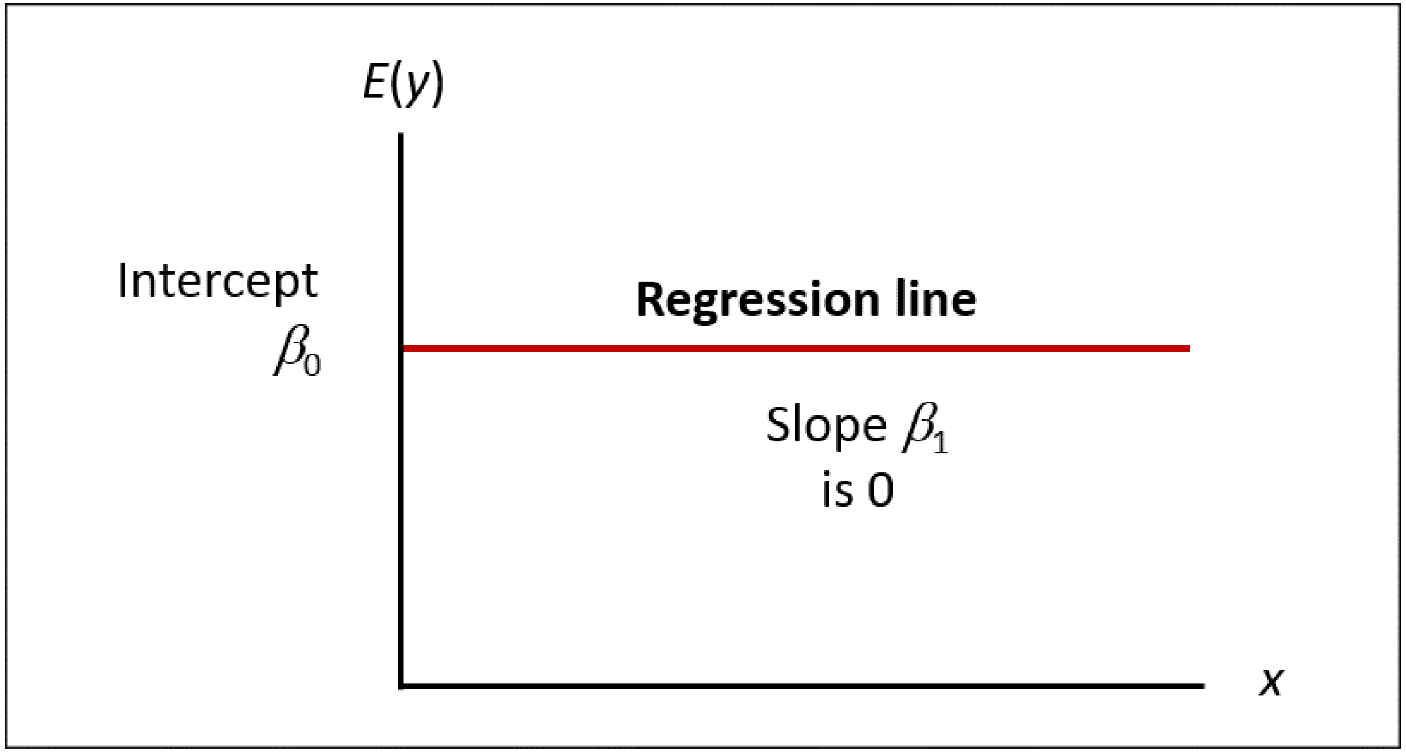
Negative Linear Relationship



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

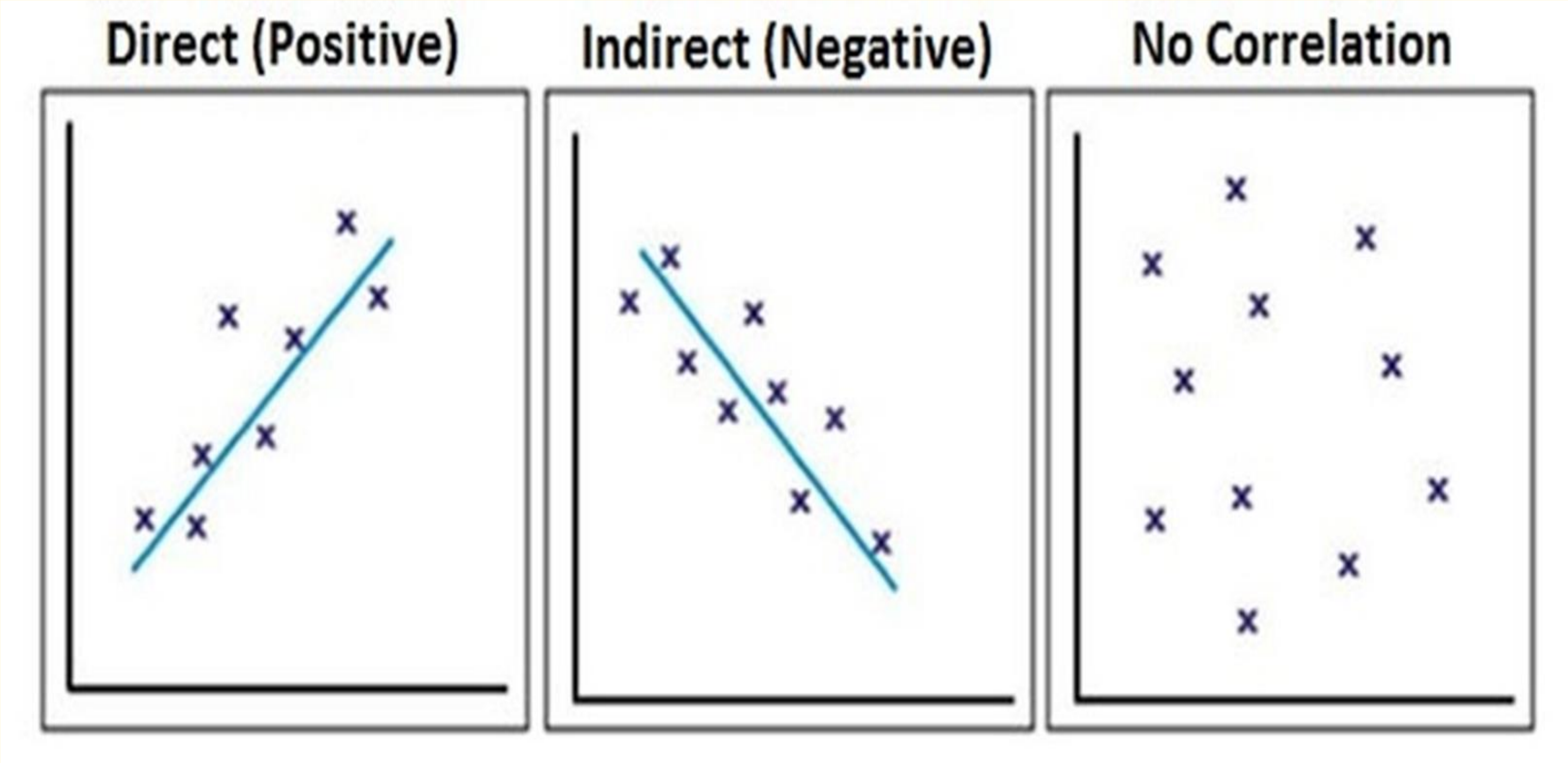
Simple Linear Regression Equation (4 of 4)

No Relationship



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

相關係數與迴歸係數



https://www.biologyforlife.com/uploads/2/2/3/9/22392738/correlation_1.jpg?688

Estimated Simple Linear Regression Equation

The estimated simple linear regression equation:

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

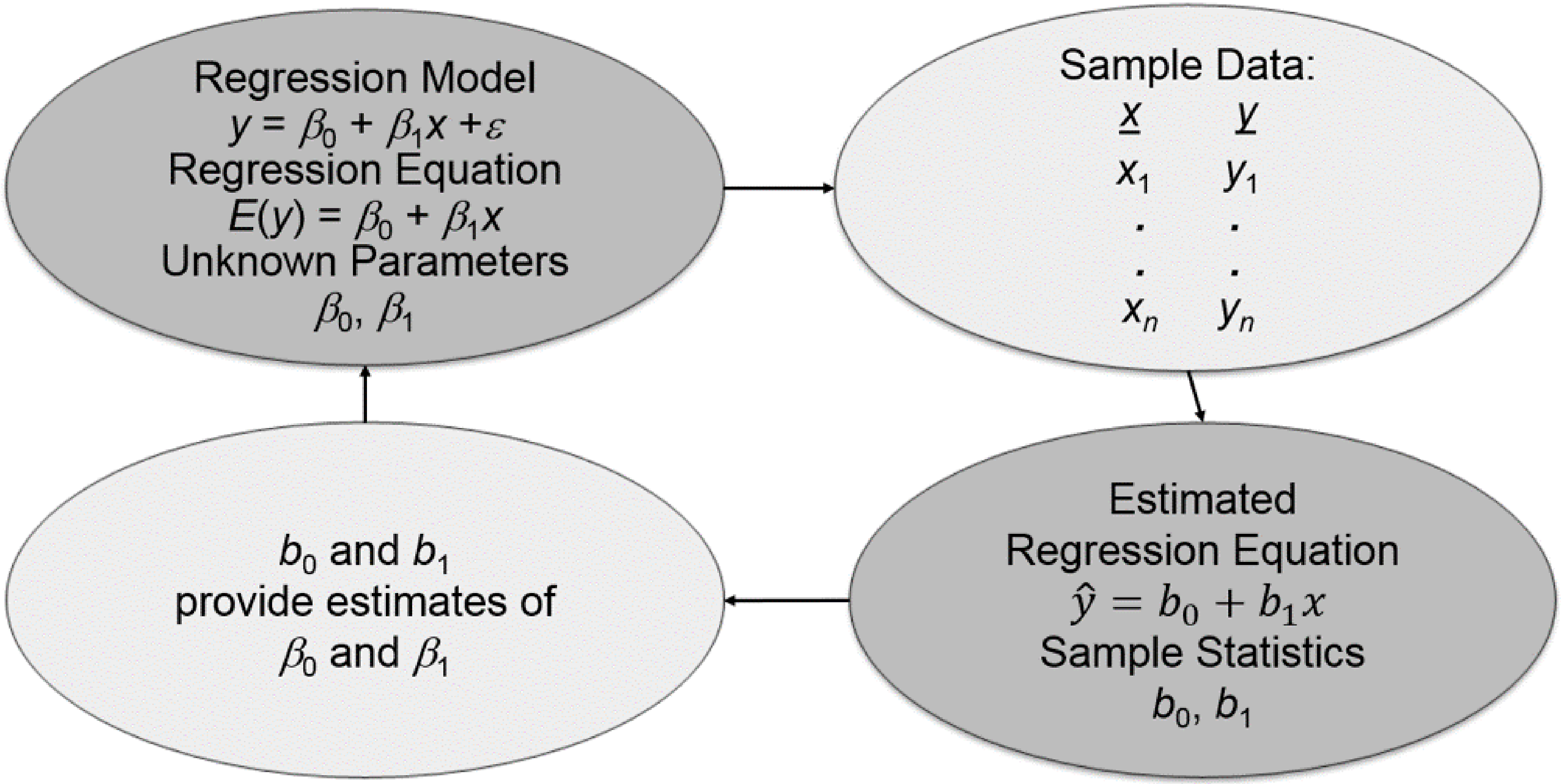
where: b_0 is the y intercept of the regression line.

b_1 are the slope of the regression line.

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

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Estimation Process



Regression Model

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

Regression Equation

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

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

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Least Squares Method (1 of 3)

Least Squares Criterion

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

where:

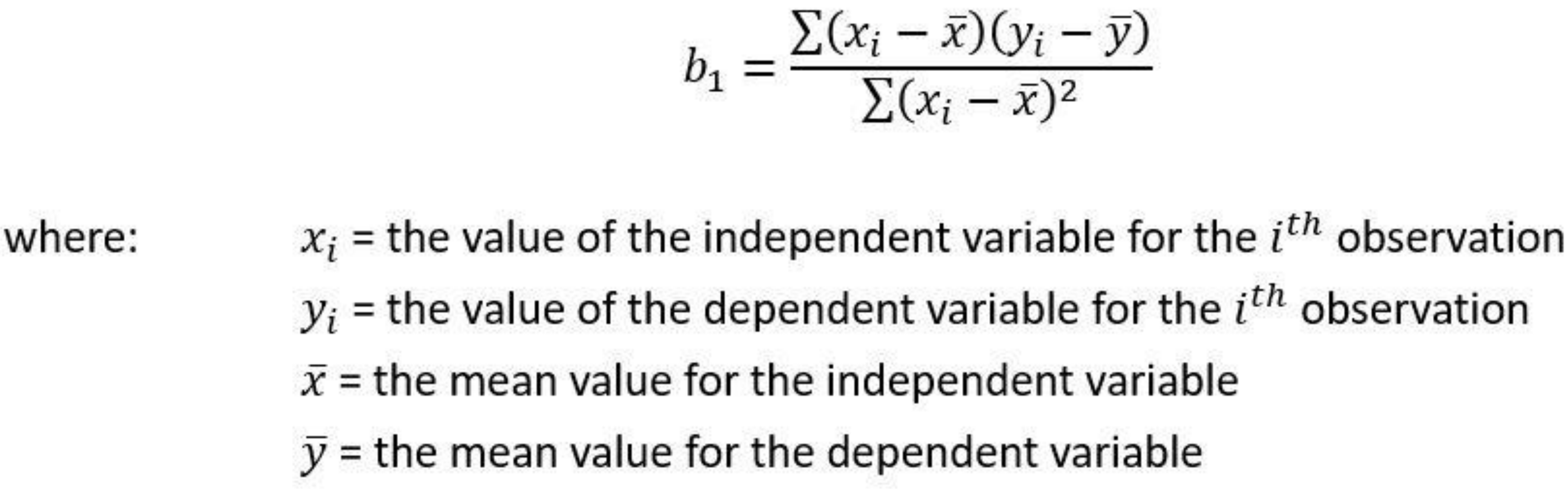
y_i = the observed value of the dependent variable for the i^{th} observation

$\hat{y}_i$ = the estimated value of the dependent variable for the i^{th} observation

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

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 = the value of the independent variable for the i^{th} observation

y_i = the value of the dependent variable for the i^{th} observation

$\bar{x}$ = the mean value for the independent variable

$\bar{y}$ = the mean value for the dependent variable

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Least Squares Method (3 of 3)

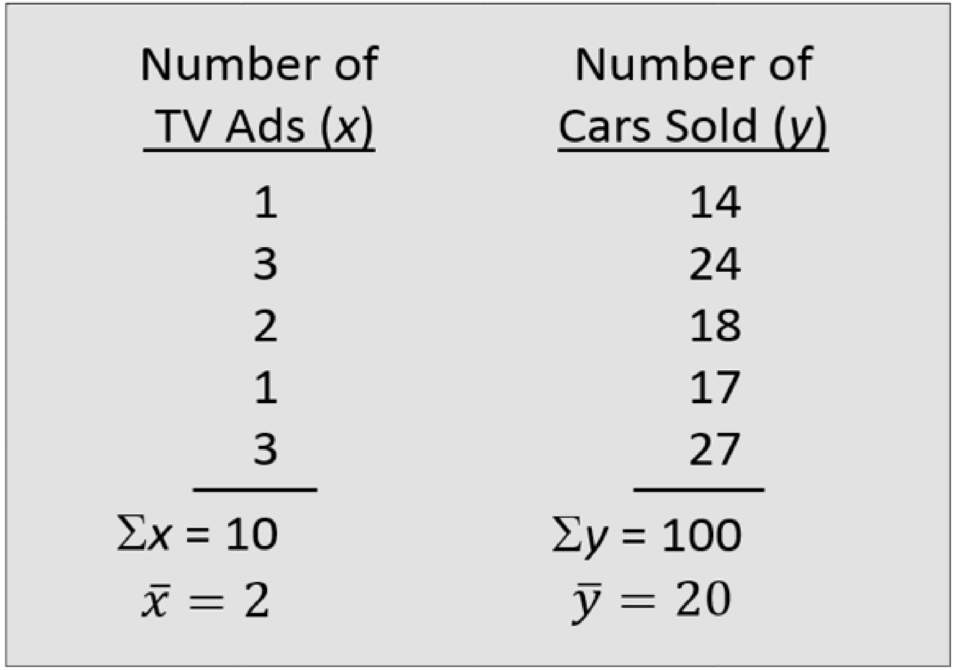
y-intercept for the estimated regression equation

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

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Simple Linear Regression

Reed Auto periodically has a special week-long sale. As part of the advertising campaign, Reed runs one or more television commercials during the weekend preceding the sale. Here are the data from a sample of 5 previous sales:



Number of TV Ads (x)	Number of Cars Sold (y)
1	14
3	24
2	18
1	17
3	27
$\Sigma x = 10$	$\Sigma y = 100$
$\bar{x} = 2$	$\bar{y} = 20$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Estimated Regression Equation

- Slope for the estimated regression equation

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{20}{4} = 5$$

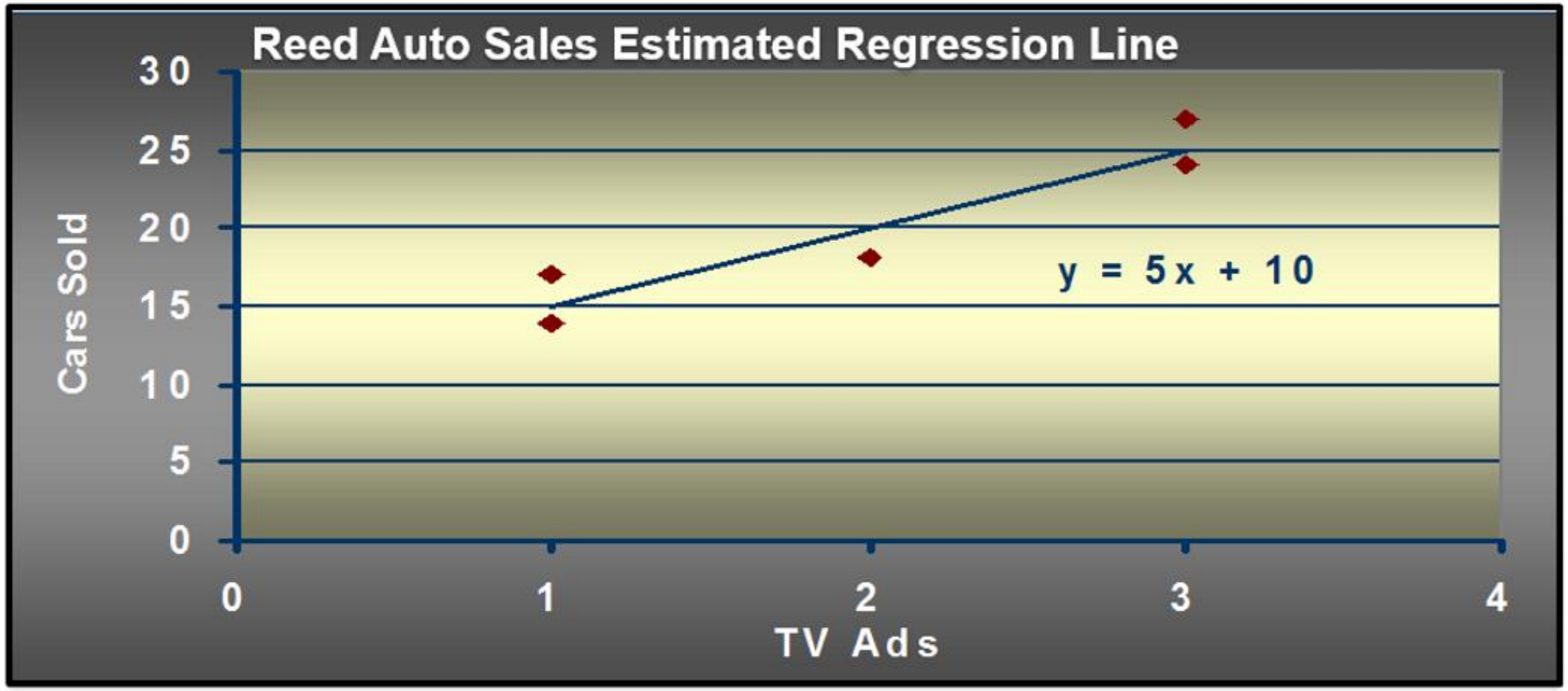
- y-intercept for the estimated regression equation

$$b_0 = \bar{y} - b_1\bar{x} = 20 - 5(2) = 10$$

- Estimated Regression Equation: $\hat{y} = 10 + 5x$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Using Excel’s Chart Tools for Scatter Diagram & Estimated Regression Equation



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Coefficient of Determination (1 of 3)

Relationship Among SST, SSR, SSE

$$\text{SST} = \text{SSR} + \text{SSE}$$

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

where:

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Coefficient of Determination (2 of 3)

The coefficient of determination is:

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

where:

SSR = sum of squares due to regression

SST = total sum of squares

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Coefficient of Determination (3 of 3)

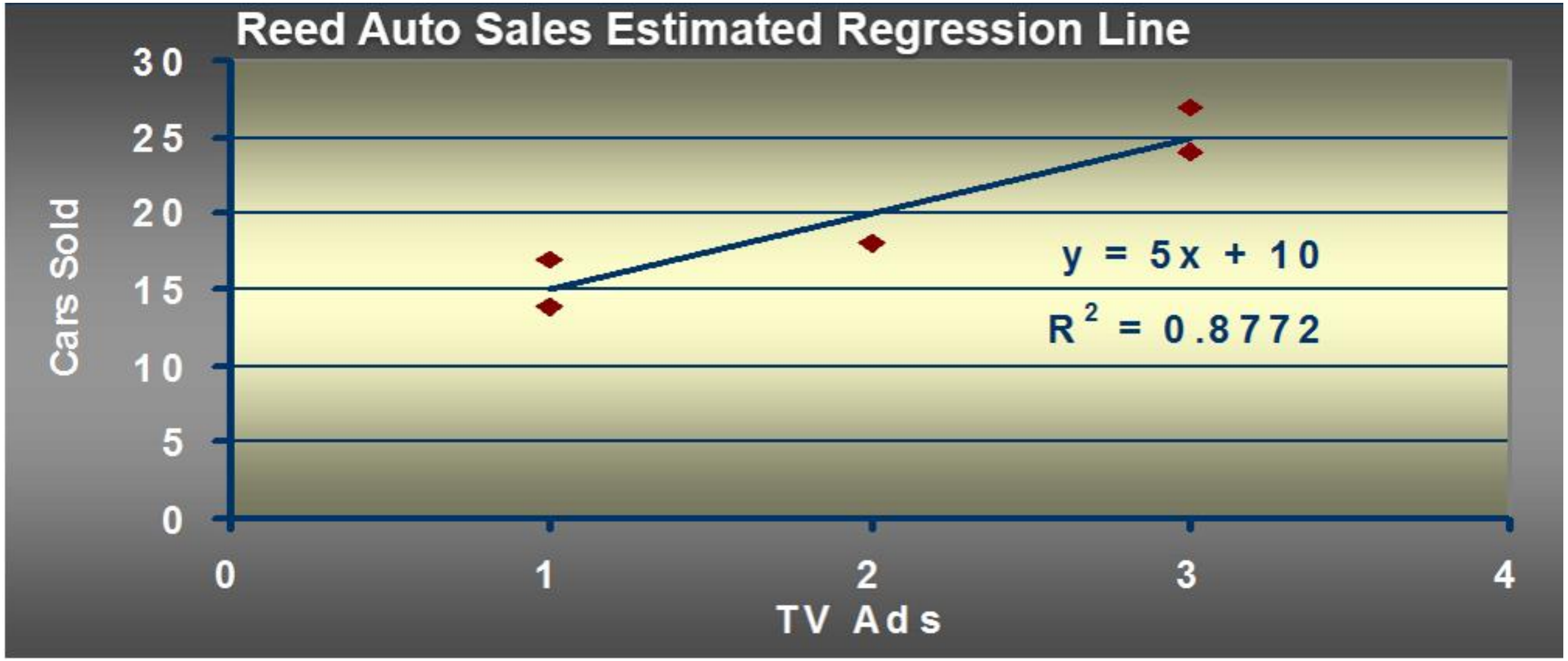
$$r^2 = \frac{SSR}{SST} = \frac{100}{114} = 0.8772$$

The regression relationship is very strong; 87.72% of the variability in the number of cars sold can be explained by the linear relationship between the number of TV ads and the number of cars sold.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Using Excel to Compute the Coefficient of Determination

Adding r^2 Value to Scatter Diagram



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Sample Correlation Coefficient (1 of 2)

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

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

where:

b_1= the slope of the estimated regression equation

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Sample Correlation Coefficient (2 of 2)

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

The sign of b_1 in the equation $\hat{y} = 10 + 5x$ is positive, so

$$r_{xy} = \sqrt{0.8772} = \mathbf{0.9366}$$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Assumptions About the Error Term ε

1. The error 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 ε is a normally distributed random variable.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

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 commonly used are the 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.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

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.

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

where:

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

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Testing for Significance (3 of 3)

To estimate *σ*, we take the square root of s^2.

The resulting *s* is called the standard error of the estimate.

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

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

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

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

Hypotheses:

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

Test Statistic:

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

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

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

Rejection Rule:

Reject H_0 if p –value $\leq \alpha$

Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where:

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

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

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

1. Determine the hypotheses. $H_0\colon\ \beta_1 = 0$

   $H_a\colon\ \beta_1 \neq 0$

2. Specify the level of significance. $\alpha = 0.05$

3. Select the test statistic. $t = \dfrac{b_1}{s_{b_1}}$

4. State the rejection rule. Reject H_0 if the *p*-value ≤ 0.05 or $|t| > 3.182$ with 3 degrees of freedom.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

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

5. Compute the value of the test statistic.

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{1.08} = 4.63$$

6. Determine whether to reject H_0.

$t = 4.4541$ provides an area of 0.01 in the upper tail. Hence the *p*-value < 0.02.

Also, $t = 4.63 > 3.182$, so we can reject H_0.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Confidence Interval for β_1 (1 of 3)

- We can use a 95% 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.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Confidence Interval for β_1 (2 of 3)

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

where

b_1 = the point estimator,

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

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

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

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.

- 95% Confidence Interval for β_1

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

$$5 \pm 3.182(1.08)$$

$$5 \pm 3.44$$

1.56 to 8.44

- Conclusion

  0 is not included in the confidence interval. Reject H_0.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Testing for Significance: F Test (1 of 3)

- Hypotheses:

$$H_0\colon \beta_1 = 0$$

$$H_a\colon \beta_1 \neq 0$$

- Test Statistic:

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

- Rejection Rule:

Reject H_0 if the *p*-value ≤ 0.05 or if $F \geq F_\alpha$

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.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Testing for Significance: F Test (2 of 3)

1. Determine the hypotheses. $H_0\colon \beta_1 = 0$

   $H_a\colon \beta_1 \neq 0$

2. Specify the level of significance. $\alpha = 0.05$

3. Select the test statistic. $F = \frac{MSR}{MSE}$

4. State the rejection rule. Reject H_0 if the *p*-value ≤ 0.05 or $F \geq 10.13$ with 1 *df* in the numerator and 3 *df* in the denominator.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Testing for Significance: F Test (3 of 3)

5. Compute the value of the test statistic.

$$F = \frac{MSR}{MSE} = \frac{100}{4.667} = 21.43$$

6. Determine whether to reject H_0.

   F = 17.44 provides an area of 0.025 in the upper tail Thus, the *p*-value corresponding to F = 21.43 is less than 0.025. Hence, we reject H_0.

   The statistical evidence is sufficient to conclude that we have a significant relationship between the number of TV ads aired and the number of cars sold.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Some Cautions about the Interpretation of Significance Tests

- Rejecting H_0: β_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.
- Just because we are able to reject H_0: β_1 = 0 and demonstrate statistical significance does not enable us to conclude that there is a linear relationship between x and y.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Using the Estimated Regression Equation for Estimation and Prediction (1 of 2)

- A confidence interval is an interval estimate of the *mean value of* y for a given value of x.
- A prediction interval is used whenever we want to *predict an individual value of* y for a new observation corresponding to a given value of x.
- The margin of error is larger for a prediction interval.

假設需要預測以下兩個問題：

1.對於那些積雪深度為7米的年份，預測灌溉面積

2.今年最大積雪深度為7米，預測今年灌溉面積

→第一個問題「**預測平均值**」，第2個問題「**預測個別值**」。

https://topic.alibabacloud.com/tc/a/what-is-the-difference-between-the-prediction-interval-the-confidence-interval-and-the-prediction-interval_8_8_31219070.html

變異數

- 由(2.29b)可以注意到，$\hat{Y}_h$抽樣分配的變異數受到X_h與$\overline{X}$距離多遠的影響，即受到$\boxed{(X_h-\overline{X})^2}$的影響。$X_h$離$\overline{X}$越遠，則$(X_h-\overline{X})^2$就越大，$\hat{Y}_h$的變異數也越大。圖 2.3 可以對這個現象做一個直觀的解釋。在同一組X值之下重複抽取兩個樣本所得到的兩條樣本迴歸線。

$$\hat{Y}_h = b_0 + b_1 X_h \qquad \sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n}+\frac{(X_h-\overline{X})^2}{\sum(X_i-\overline{X})^2}\right] \tag{2.29b}$$

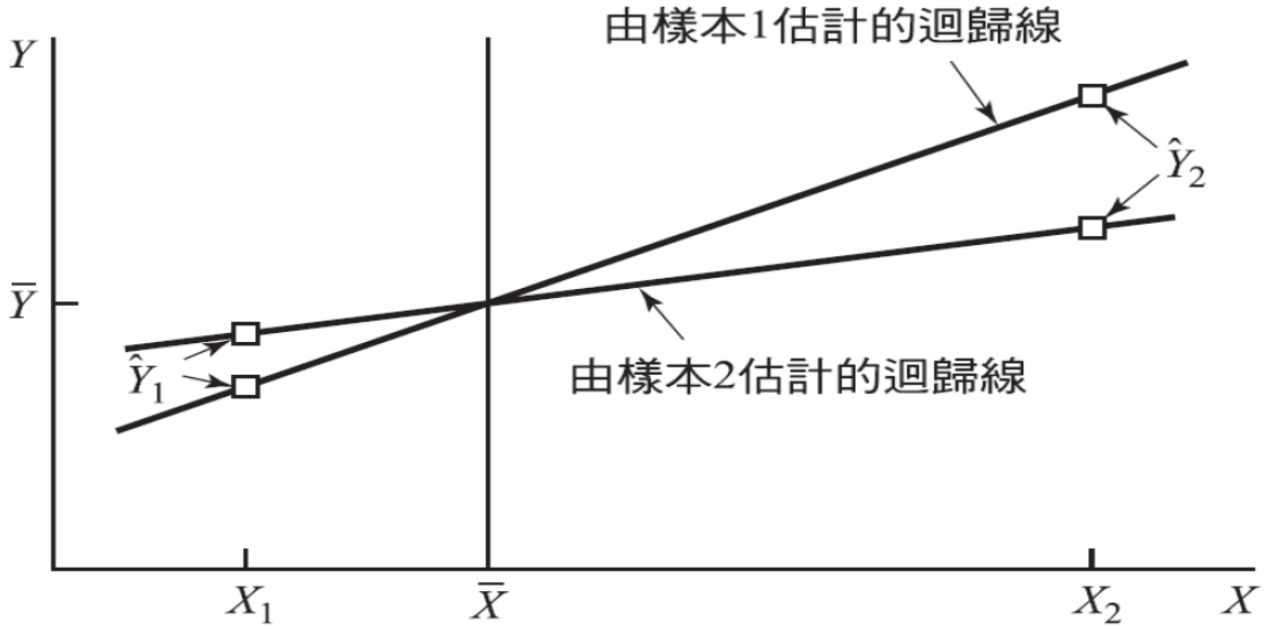


圖 2.3
具有相同平均數$\overline{Y}$及$\overline{X}$的兩樣本其不同樣本之b_1的變化對$\hat{Y}_h$的影響。

Using the Estimated Regression Equation for Estimation and Prediction (2 of 2)

- Confidence Interval Estimate of $E(y^*)$

$$\hat{y}^* \pm t_{\alpha/2}(s_{\hat{y}^*})$$

- Confidence Interval Estimate of $E(y^*)$

$$\hat{y}^* \pm t_{\alpha/2}(s_{\text{pred}})$$

where $1 - \alpha$ is the confidence coefficient and $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Point Estimation

If 3 TV ads are run prior to a sale, we expect the mean number of cars sold to be:

$$\hat{y} = 10 + 5(3) = 25 \text{ cars}$$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Confidence Interval for $E(y^*)$ (1 of 2)

Estimate of the Standard Deviation of $\hat{y}^*$:

$$s_{\hat{y}^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

$$s_{\hat{y}^*} = 2.16025\sqrt{\frac{1}{5} + \frac{(3-2)^2}{(1-2)^2 + (3-2)^2 + \cdots + (3-2)^2}}$$

$$s_{\hat{y}^*} = 2.16025\sqrt{\frac{1}{5} + \frac{1}{4}} = \mathbf{1.4491}$$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Confidence Interval for $E(y^*)$ (2 of 2)

The 95% confidence interval estimate of the mean number of cars sold when 3 TV ads are run is:

$$\hat{y}^* \pm t_{\alpha/2}(s_{\hat{y}^*})$$

$$25 \pm 3.1824(1.4491)$$

$$25 \pm 4.61$$

20.39 to 29.61 cars

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Prediction Interval for y^* (1 of 2)

Estimate of the Standard Deviation of an Individual Value of y^*

$$s_{\text{pred}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

$$s_{\text{pred}} = 2.16025\sqrt{1 + \frac{1}{5} + \frac{1}{4}}$$

$$s_{\text{pred}} = 2.16025(1.20416) = \mathbf{2.6013}$$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Prediction Interval for y^* (2 of 2)

The 95% prediction interval estimate of the number of cars sold in one particular week when 3 TV ads are run is:

$$\hat{y}^* \pm t_{\alpha/2}(s_{\text{pred}})$$

$$25 \pm 3.1824(2.6013)$$

$$25 \pm 8.28$$

16.72 to 33.28 cars

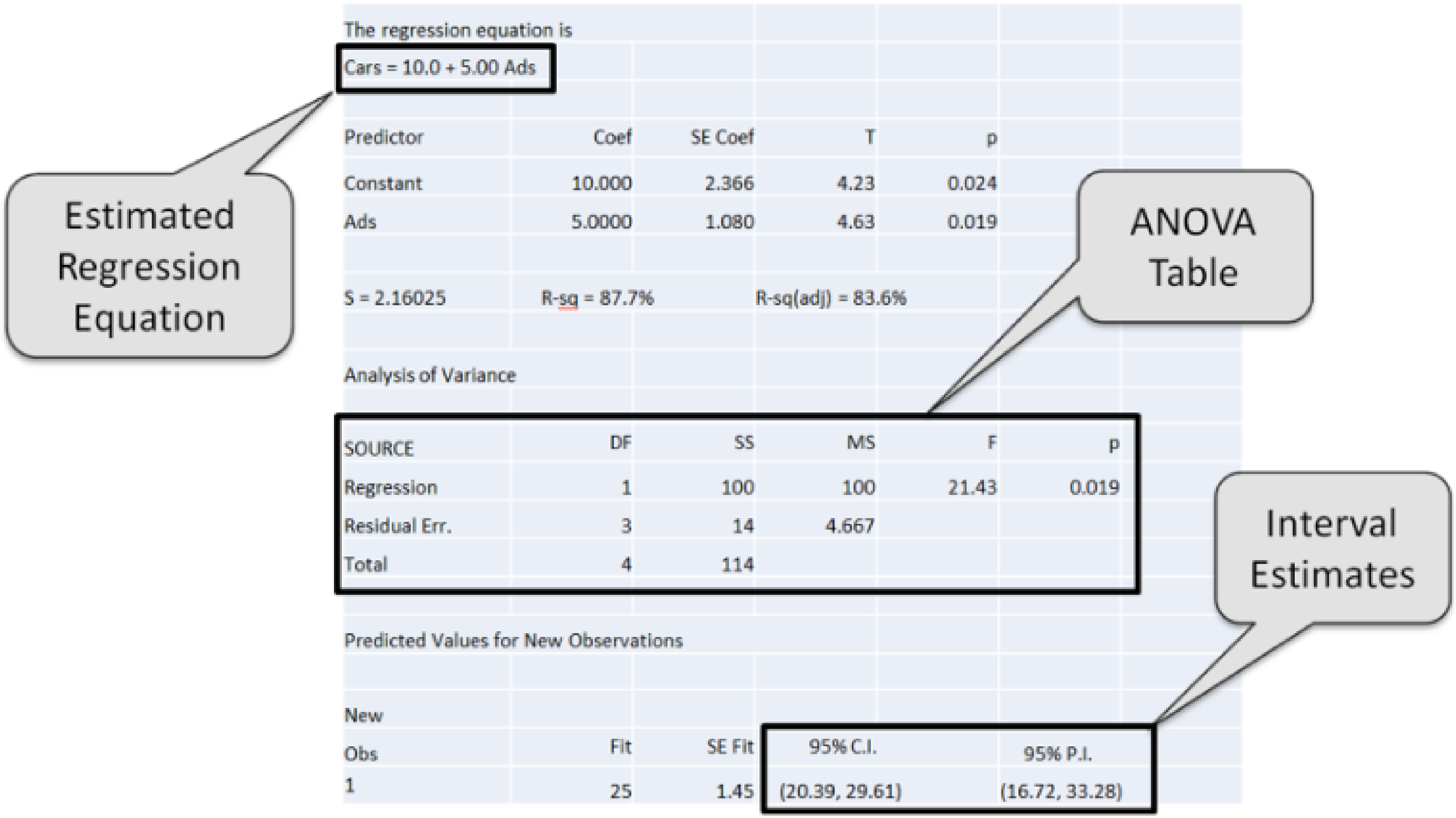
© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Computer Solution

- Up to this point, you have seen how Excel can be used for various parts of a regression analysis.
- Excel also has a comprehensive tool in its Data Analysis package called Regression.
- The Regression tool can be used to perform a complete regression analysis.
- Performing the regression analysis computations without the help of a computer can be quite time consuming.
- On the next slide we show Minitab output for the Reed Auto Sales example.
- Recall that the independent variable was named Ads and the dependent variable was named Cars in the example.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Minitab Output (1 of 2)



```
The regression equation is
Cars = 10.0 + 5.00 Ads

Predictor         Coef    SE Coef        T         p
Constant        10.000      2.366     4.23     0.024
Ads             5.0000      1.080     4.63     0.019

S = 2.16025     R-sq = 87.7%        R-sq(adj) = 83.6%

Analysis of Variance

SOURCE              DF         SS       MS         F         p
Regression           1        100      100     21.43     0.019
Residual Err.        3         14    4.667
Total                4        114

Predicted Values for New Observations

New
Obs                Fit     SE Fit    95% C.I.              95% P.I.
1                   25       1.45    (20.39, 29.61)        (16.72, 33.28)
```

Estimated Regression Equation

ANOVA Table

Interval Estimates

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Minitab Output (2 of 2)

- Minitab prints the estimated regression equation as Cars = 10 + 5 Ads.
- For each of the coefficients b_0 and b_1, the output shows its value, standard deviation, *t* value, and *p*-value.
- Minitab prints the standard error of the estimate, *s*, as well as information about the goodness of fit.
- The standard ANOVA table is printed.
- Also provided are the 95% confidence interval estimate of the expected number of cars sold and the 95% prediction interval estimate of the number of cars sold for an individual weekend with 3 ads.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Using Excel’s Regression Tool (1 of 4)

Excel Output (top portion)

	A	B	C
9			
10	*Regression Statistics*	*Regression Statistics*	
11	Multiple R	0.936585812	
12	R Square	0.877192982	
13	Adjusted R Square	0.83625731	
14	Standard Error	2.160246899	
15	Observations	5	
16			

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Using Excel’s Regression Tool (2 of 4)

Excel Output (middle portion)

	A	B	C	D	E	F
16						
17	ANOVA					
18		df	SS	MS	F	Significance F
19	Regression	1	100	100	21.4286	0.018986231
20	Residual	3	14	4.66667		
21	Total	4	114			
22						

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Using Excel's Regression Tool (3 of 4)

Excel Output (bottom-left portion)

	A	B	C	D	E
22					
23		*Coeffic.*	*Std. Err.*	*t Stat*	*P-values*
24	Intercept	10	2.36643	4.2258	0.02424
25	TV Ads	5	1.08012	4.6291	0.01899
26					

Note: Columns F – I are not shown

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Using Excel’s Regression Tool (4 of 4)

Excel Output (bottom-right portion)

	A	B	F	G	H	I
22						
23		*Coeffic.*	*Low. 95%*	*Up. 95%*	*Low. 95.0%*	*Up. 95.0%*
24	Intercept	10	2.46895	17.53105	2.46895044	17.5310496
25	TV Ads	5	1.562562	8.437438	1.56256189	8.43743811
	A	**B**	**F**	**G**	**H**	**I**

Note: Columns C – E are hidden

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Residual Analysis

- If the assumptions about the error term ε appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.
- The residuals provide the best information about ε.
- Residual for observation i:

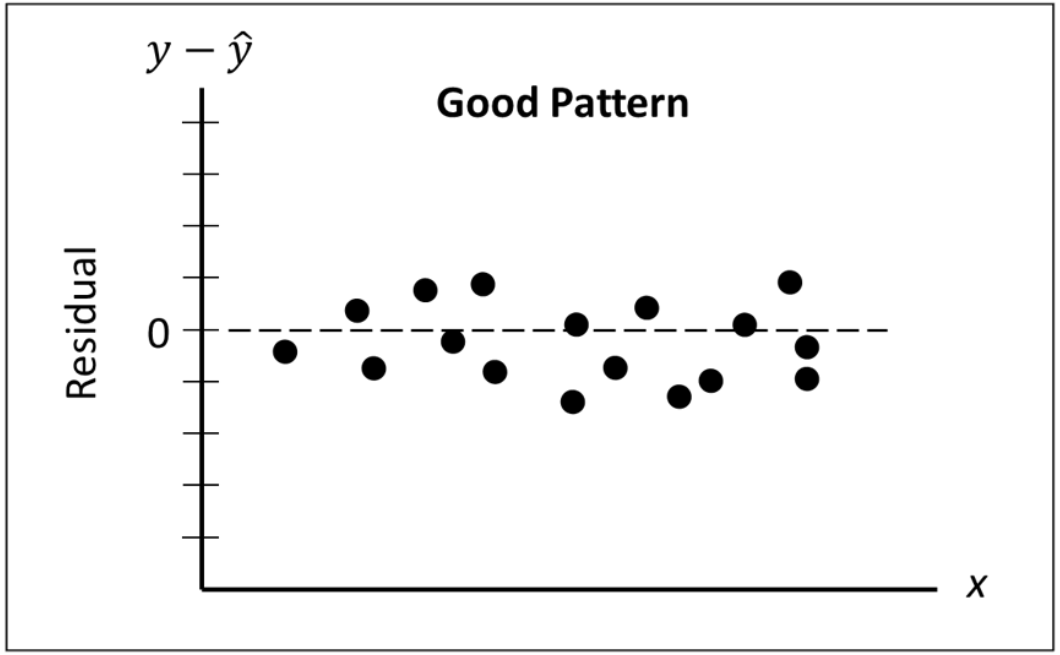
$$y_i - \hat{y}_i$$

- Much of the residual analysis is based on an examination of graphical plots.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

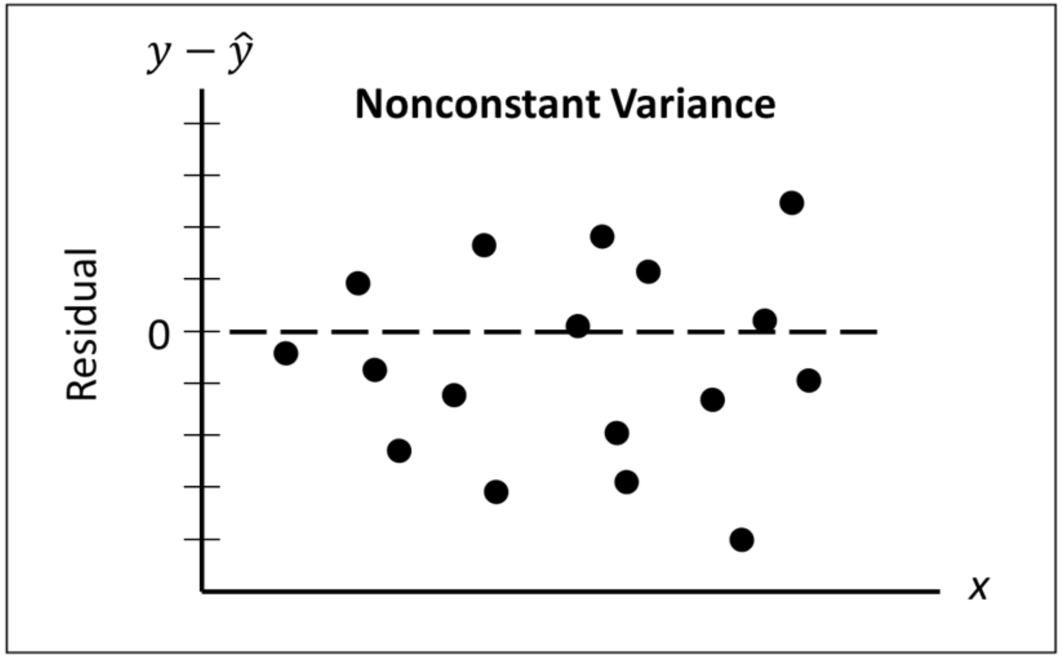
Residual Plot Against *x* (1 of 4)

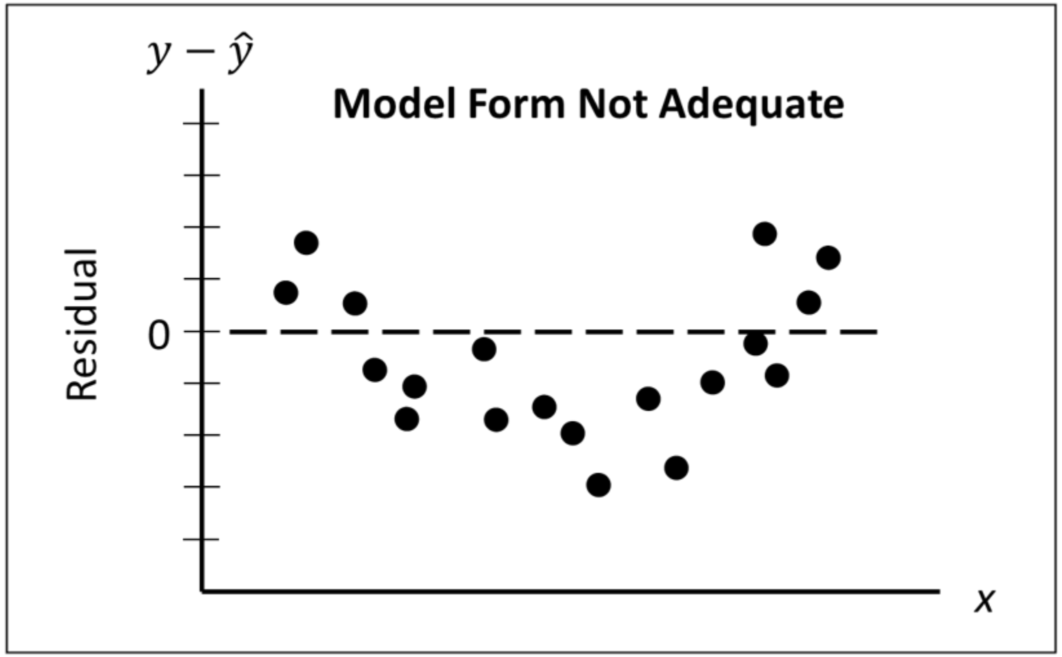
If the assumption that the variance of ε is the same for all values of *x* is valid, and the assumed regression model is an adequate representation of the relationship between the variables, then the residual plot should give an overall impression of a horizontal band of points.



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Residual Plot Against *x* (2 of 4)





© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

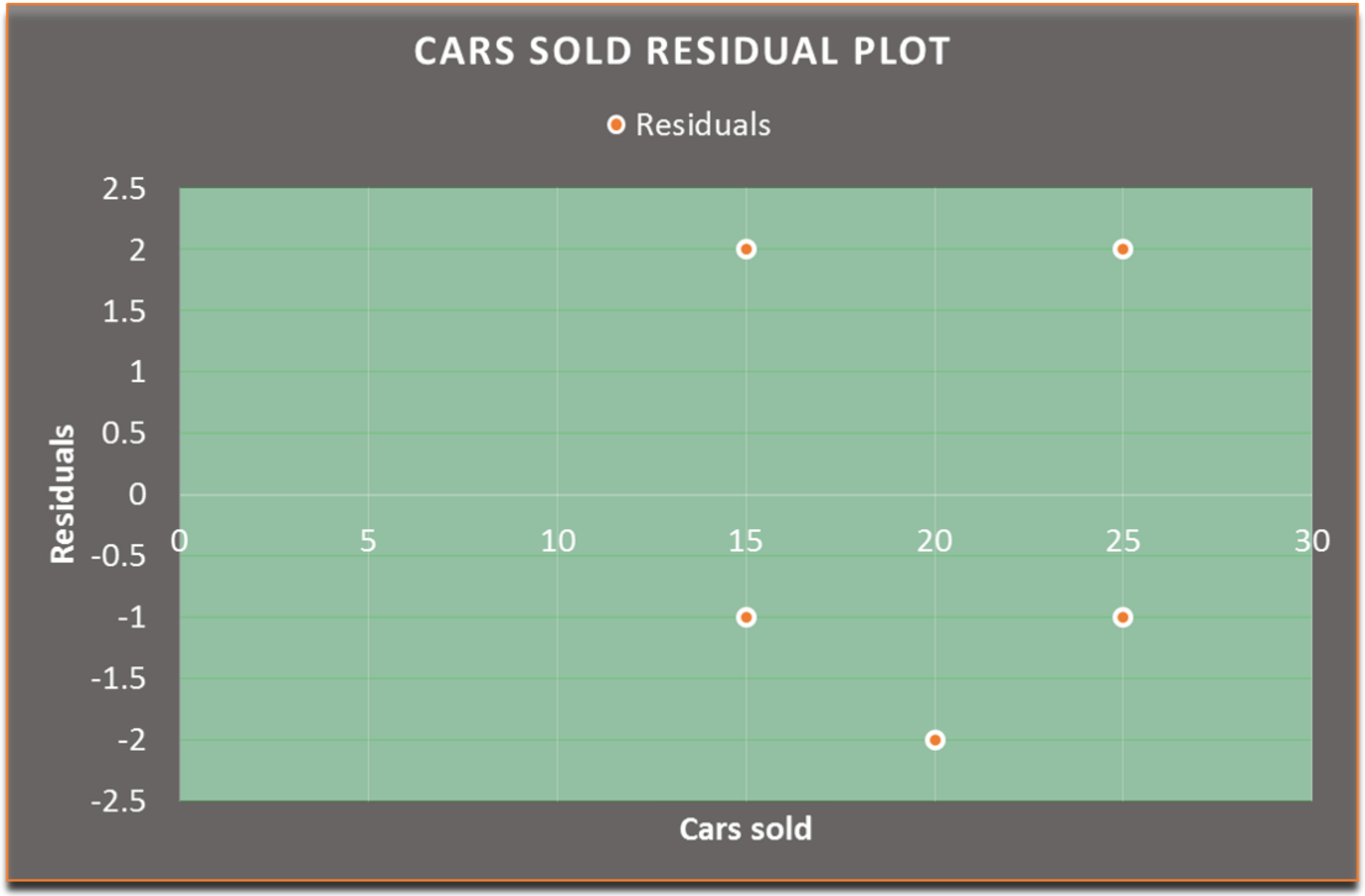
Residual Plot Against *x* (3 of 4)

Residuals

Observation	Predicted Cars Sold	Residuals
1	15	-1
2	25	-1
3	20	-2
4	15	2
5	25	2

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Residual Plot Against *x* (4 of 4)



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Standardized Residuals

Standardized Residual for Observation i

$$\frac{y_i - \hat{y}_i}{s_{y_i-\hat{y}_i}}$$

$$s_{y_i-\hat{y}_i} = s\sqrt{1 - h_i}$$

Where:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Standardized Residual Plot (1 of 3)

- The standardized residual plot can provide insight about the assumption that the error term ε has a normal distribution.
- If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a standard normal probability distribution.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Standardized Residual Plot (2 of 3)

Standardized Residuals

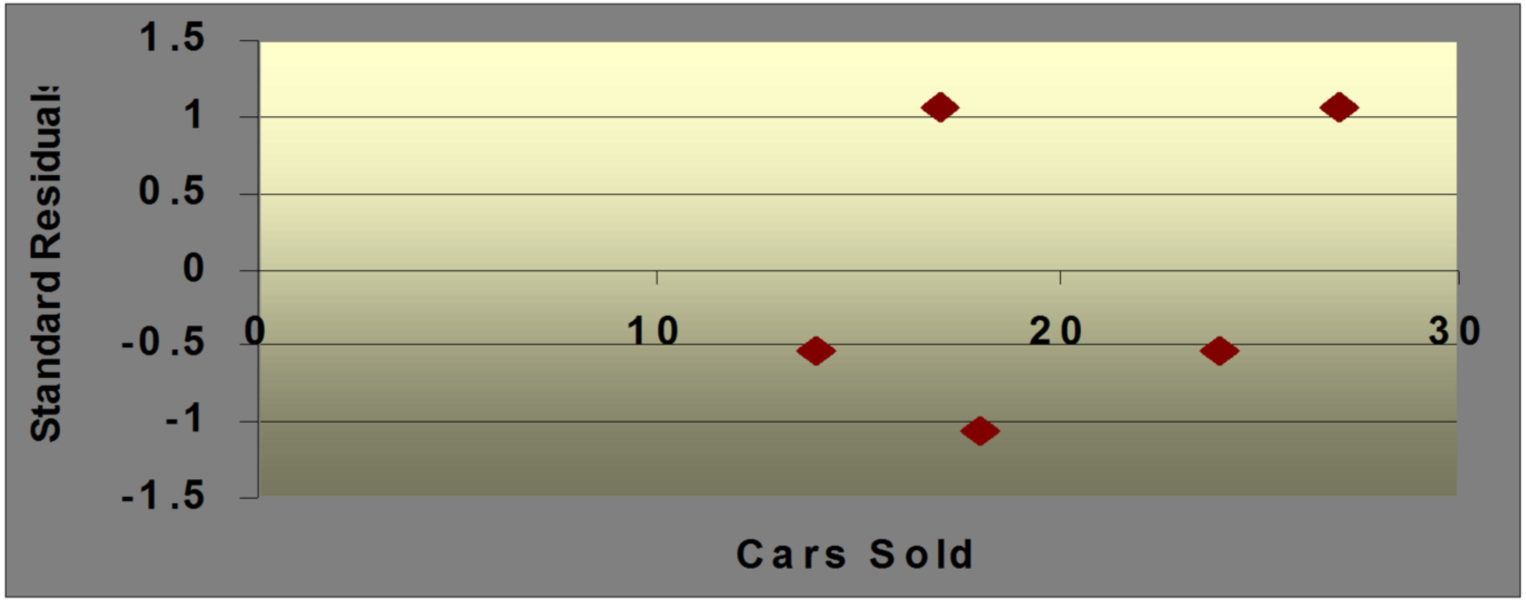
Observation	Predicted y	Residual	Standardized Residual
1	15	-1	-0.5345
2	25	-1	-0.5345
3	20	-2	-1.0690
4	15	2	1.0690
5	25	2	1.0690

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Standardized Residual Plot (3 of 3)

Standardized Residual Plot

All of the standardized residuals are between –1.5 and +1.5, indicating that there is no reason to question the assumption that ε has a normal distribution.



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Outliers and Influential Observations

Detecting Outliers

- An outlier is an observation that is unusual in comparison with the other data.
- Minitab classifies an observation as an outlier if its standardized residual value is

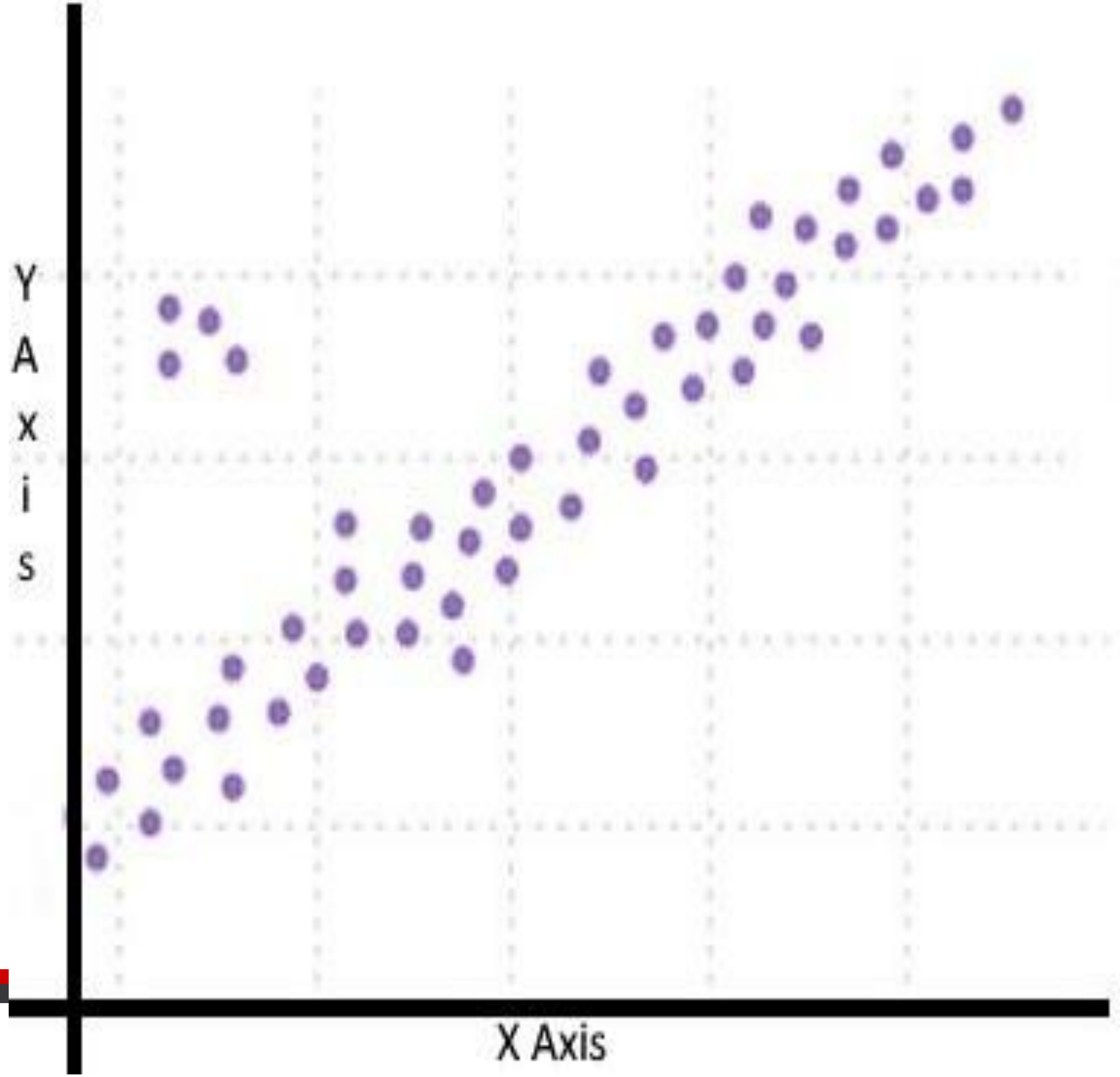
< -2 or $> +2$.

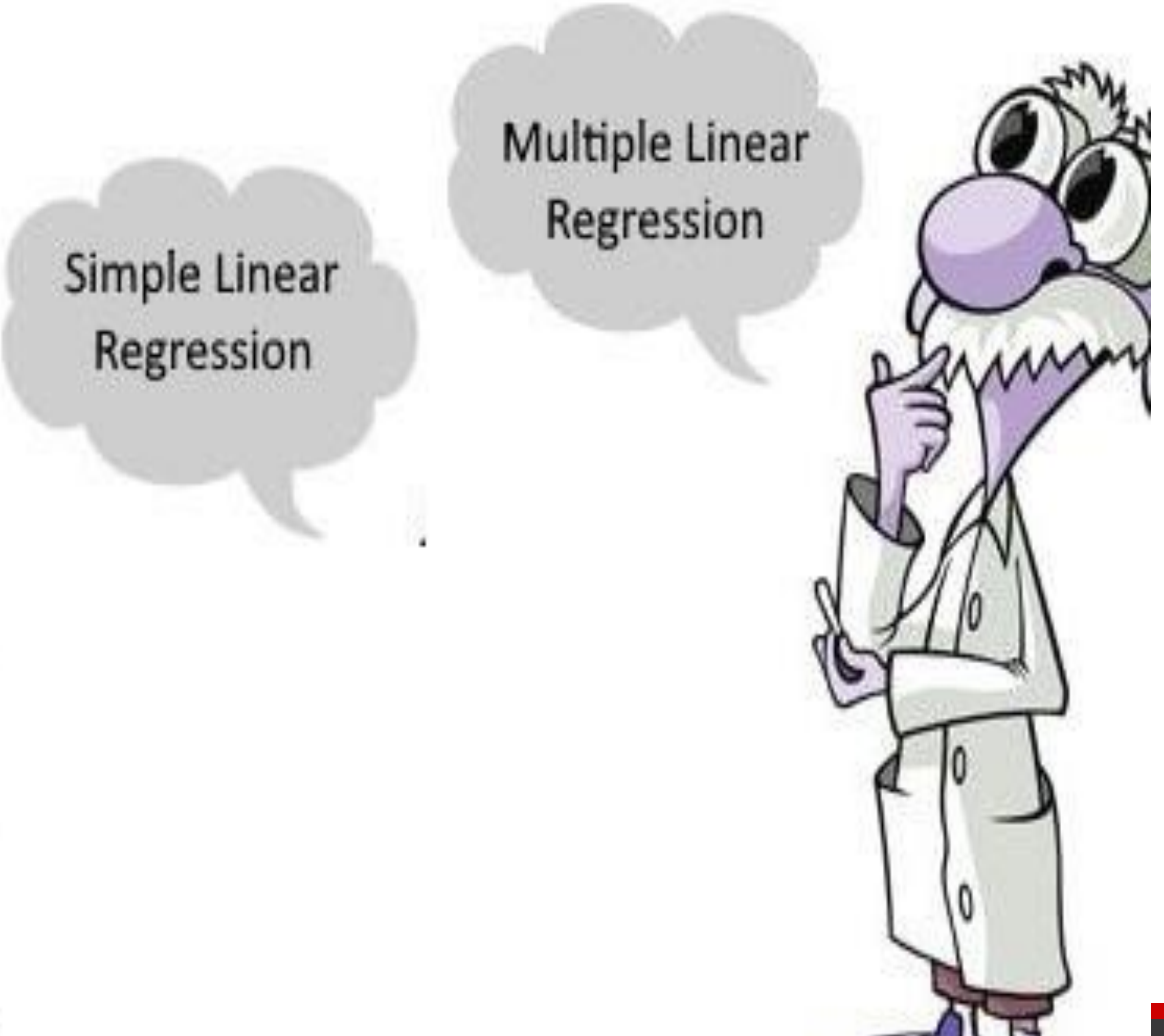
- This standardized residual rule sometimes fails to identify an unusually large observation as being an outlier.
- This rule's shortcoming can be circumvented by using studentized deleted residuals.

- The $|i^{th}$ studentized deleted residual$|$ will be larger than the $|i^{th}$ studentized residual$|$.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

為什麼要檢查殘差？



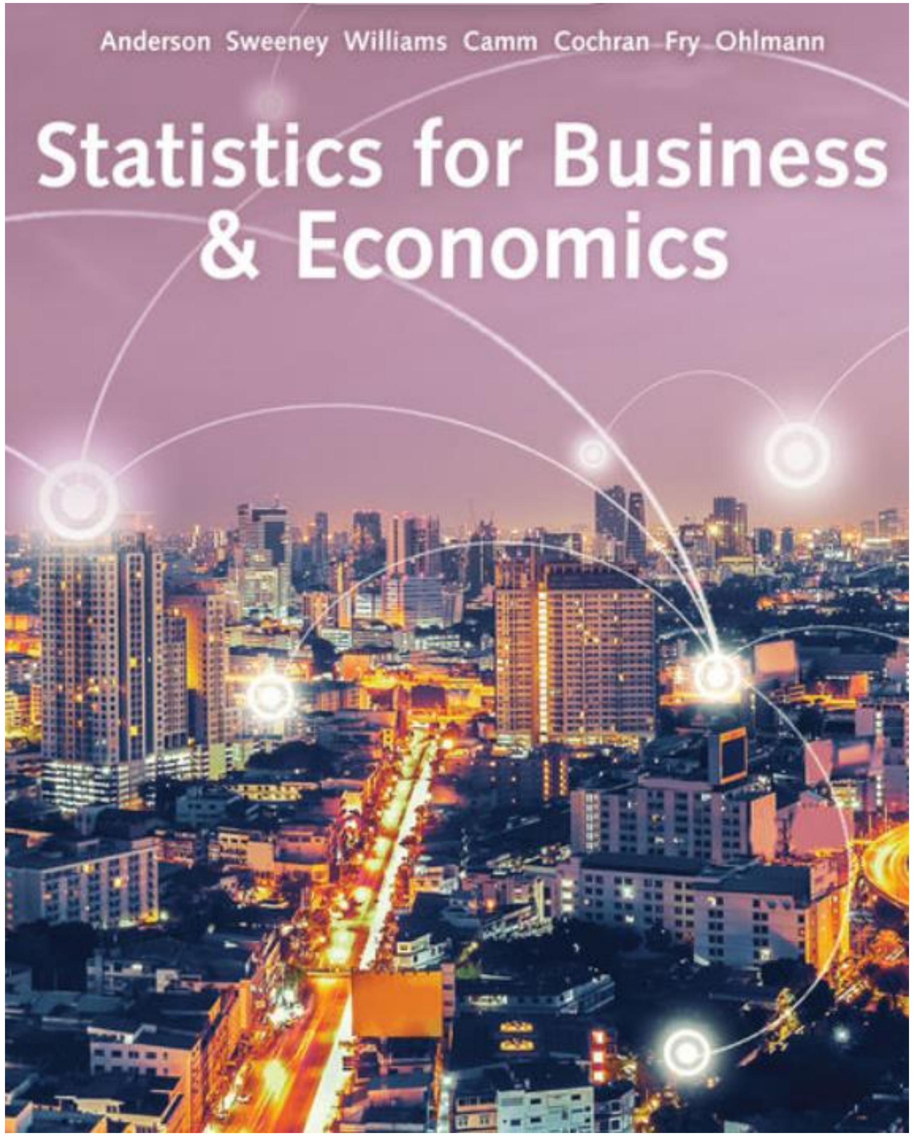


https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcR4zqNi6A64DUF7NycVkosxJNA0_F9vNTtU0L4bC9Fq2jssBJmARe9FMZq6LryC_X943pY&usqp=CAU

Statistics for Business and Economics (14e) Metric Version

Anderson, Sweeney, Williams, Camm, Cochran, Fry, Ohlmann

© 2020 Cengage Learning


Anderson Sweeney Williams Camm Cochran Fry Ohlmann
Statistics for Business & Economics

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Chapter 15 - Multiple Regression

15.1 - Multiple Regression Model

15.2 - Least Squares Method

15.3 - Multiple Coefficient of Determination

15.4 - Model Assumptions

15.5 - Testing for Significance

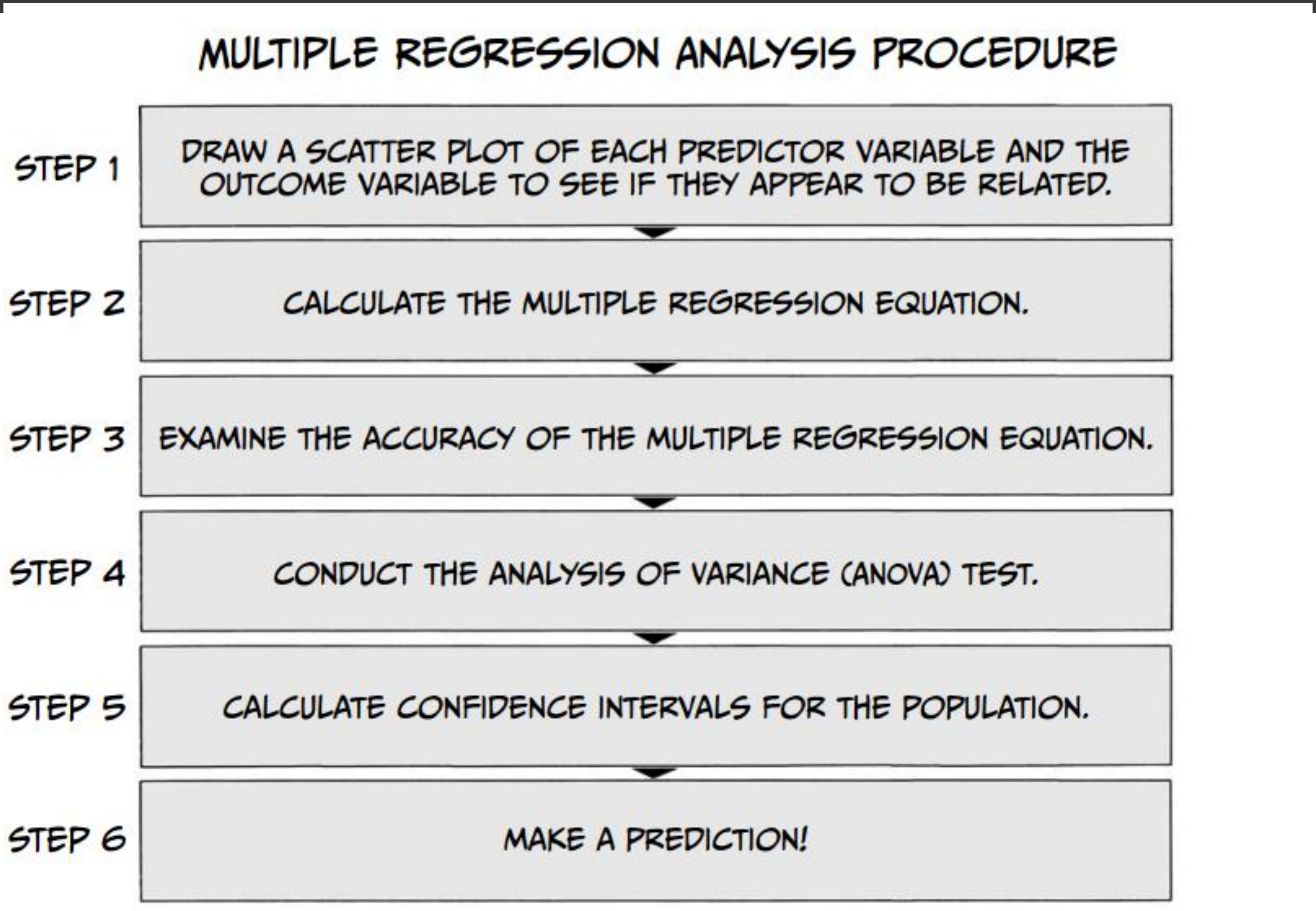
15.6 - Using the Estimated Regression Equation for Estimation and Prediction

15.7 - Categorical Independent Variables

15.8 - Residual Analysis

15.9 - Logistic Regression

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.



MULTIPLE REGRESSION ANALYSIS PROCEDURE

STEP 1 DRAW A SCATTER PLOT OF EACH PREDICTOR VARIABLE AND THE OUTCOME VARIABLE TO SEE IF THEY APPEAR TO BE RELATED.

STEP 2 CALCULATE THE MULTIPLE REGRESSION EQUATION.

STEP 3 EXAMINE THE ACCURACY OF THE MULTIPLE REGRESSION EQUATION.

STEP 4 CONDUCT THE ANALYSIS OF VARIANCE (ANOVA) TEST.

STEP 5 CALCULATE CONFIDENCE INTERVALS FOR THE POPULATION.

STEP 6 MAKE A PREDICTION!

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Multiple Regression（複迴歸）

- In this chapter we continue our study of regression analysis by considering situations involving two or more independent variables.
- This subject area, called multiple regression analysis, enables us to consider more factors and thus obtain better estimates than are possible with simple linear regression.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Multiple Regression Model

The equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, \dots, x_p$ and an error term is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

where β_0, β_1, β_2,..., β_p are the parameters, and ε is a random variable called the error term.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Multiple Regression Equation

The equation that describes how the mean value of y is related to $x_1, x_2, \ldots, x_p$ is:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

A simple random sample is used to compute the sample statistics b_0, b_1, b_2,…, b_p that are used as the point estimators of the parameters β_0, β_1, β_2,…, β_p.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Multiple Regression Model (1 of 3)

Example: Programmer Salary Survey

A software firm collected data for a sample of 20 computer programmers. A suggestion was made that regression analysis could be used to determine if salary was related to the years of experience and the score on the firm's programmer aptitude test.

The years of experience, score on the aptitude test, and corresponding annual salary ($1000s) for a sample of 20 programmers is shown on the next slide.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Multiple Regression Model (2 of 3)

Exper. (Yrs.)	Test score	Salary ($1000s)
4	78	24.0
7	100	43.0
1	86	23.7
5	82	34.3
10	84	38.0
0	75	22.2
1	80	23.1
6	83	30.0
6	91	33.0

Exper. (Yrs.)	Test score	Salary ($1000s)
9	88	38.0
2	73	26.6
10	75	36.2
6	74	29.0
8	87	34.0
4	79	30.1
6	94	33.9
3	70	28.2
3	89	30.0

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Multiple Regression Model (3 of 3)

Suppose we believe that salary (y) is related to the years of experience (x_1) and the score on the programmer aptitude test (x_2) by the following regression model:

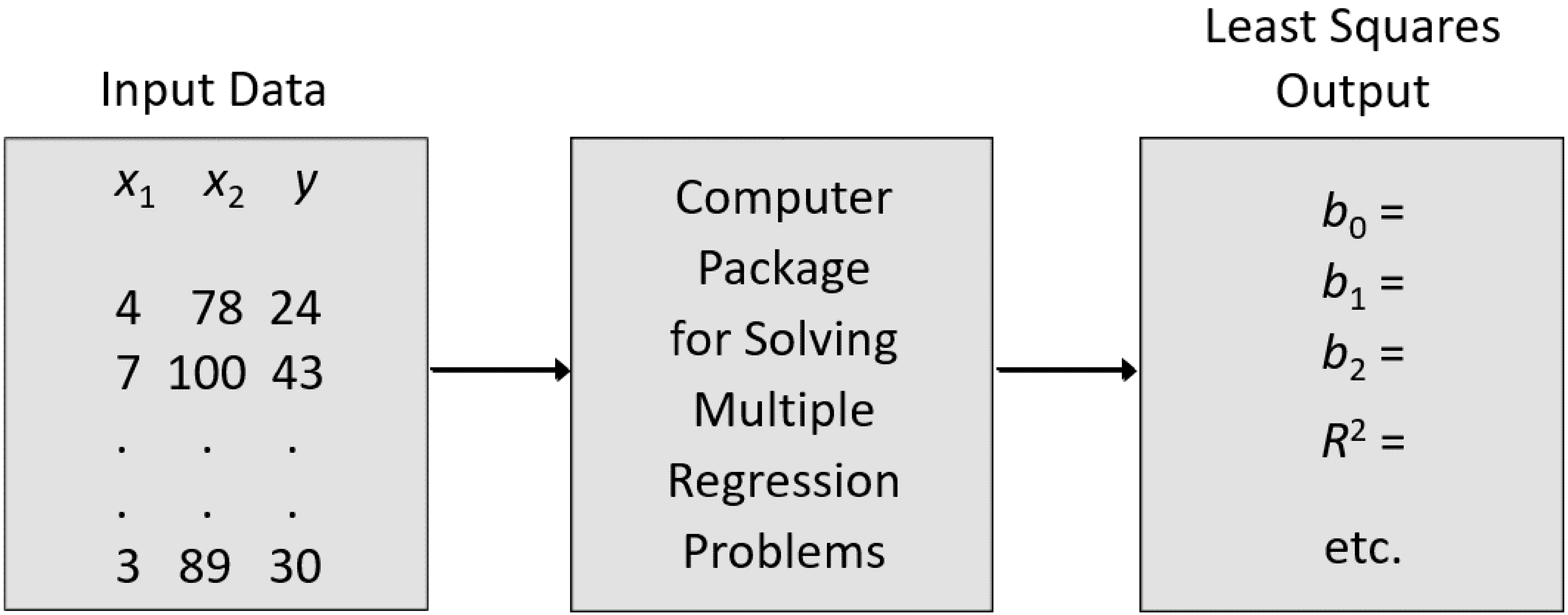
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

y = annual salary ($1000s)
x_1 = years of experience
x_2 = score on programmer aptitude test

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Solving for the Estimates of β_0, β_1, β_2 (1 of 2)



Input Data

x_1	x_2	y
4	78	24
7	100	43
.	.	.
.	.	.
3	89	30

Computer Package for Solving Multiple Regression Problems

Least Squares Output

$b_0 =$

$b_1 =$

$b_2 =$

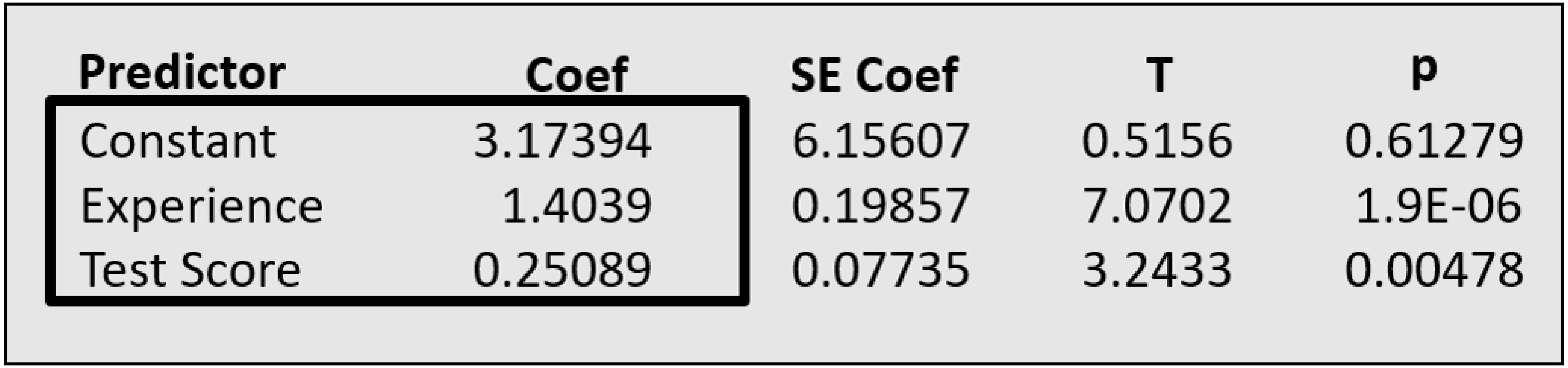
$R^2 =$

etc.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Solving for the Estimates of β_0, β_1, β_2 (2 of 2)

Regression Equation Output:



Predictor	Coef	SE Coef	T	p
Constant	3.17394	6.15607	0.5156	0.61279
Experience	1.4039	0.19857	7.0702	1.9E-06
Test Score	0.25089	0.07735	3.2433	0.00478

The estimated regression equation is:

SALARY = 3.174 + 1.404(EXPER) + 0.251(SCORE)

(Note: Predicted salary will be in thousands of dollars.)

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Testing for Significance

- In simple linear regression, the F and t tests provide the same conclusion.
- In multiple regression, the F and t tests have different purposes.
- The *F* test is referred to as the test for overall significance.

- If the F test shows an overall significance, the t test is used to determine whether each of the individual independent variables is significant.
- A separate t test is conducted for each of the independent variables in the model.
- We refer to each of these t tests as a test for individual significance.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Testing for Significance: F Test

- Hypotheses: $H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_p = 0$

  H_a: One or more of the parameters is not equal to zero.

- Test Statistic: $F = \frac{MSR}{MSE}$

- Rejection Rule: Reject H_0 if the *p*-value $\leq \alpha$ or if $F \geq F_\alpha$

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

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

F Test for Overall Significance (1 of 3)

- Hypotheses: H_0: $\beta_1 = \beta_2 = 0$

  H_a: One or more of the parameters is not equal to zero.

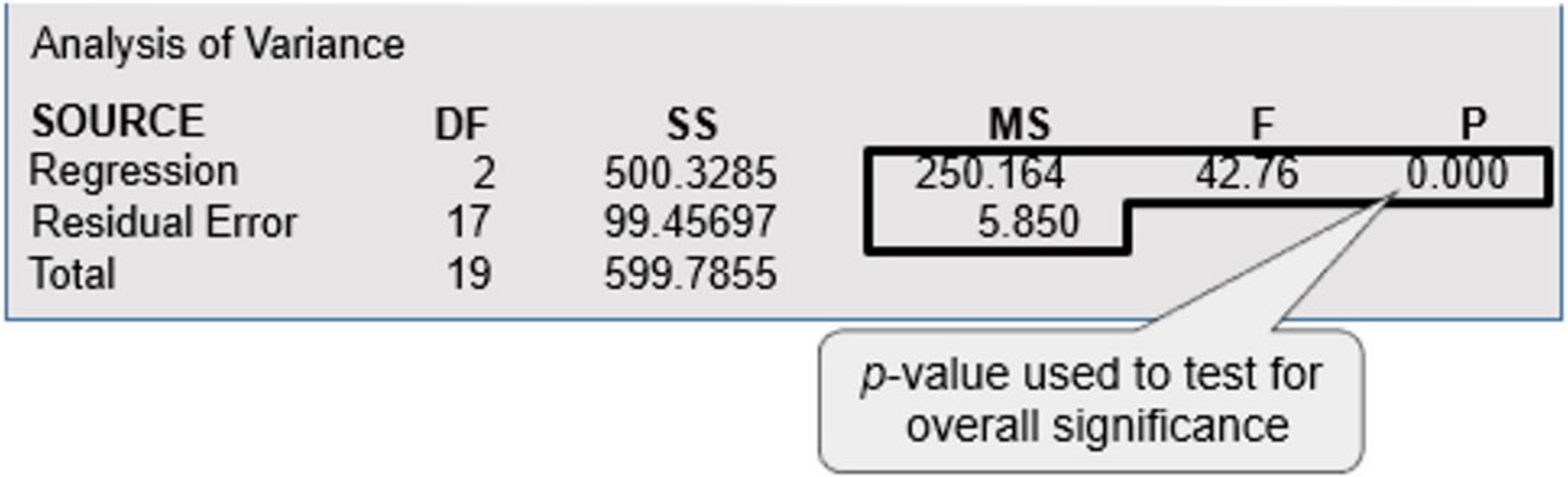
- Rejection Rule: For $\alpha = 0.05$ and *df* = 2, 17; $F_{0.05} = 3.59$

  Reject H_0 if the *p*-value ≤ 0.05 or $F \geq 3.59$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

F Test for Overall Significance (2 of 3)

ANOVA Output



Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	2	500.3285	250.164	42.76	0.000
Residual Error	17	99.45697	5.850		
Total	19	599.7855			

p-value used to test for overall significance

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

F Test for Overall Significance (3 of 3)

- Test Statistics: $F = \frac{MSR}{MSE} = \frac{250.16}{5.85} = 42.76$

- Conclusion: *p*-value ≤ 0.05 so we can reject H_0.

  Also, $F = 42.76 \geq 3.59$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Testing for Significance: *t* Test

- Hypotheses: $H_0\colon\ \beta_i = 0$

  $H_a\colon \beta_i \neq 0$

- Test Statistic: $t = \dfrac{b_i}{s_{b_i}}$

- Rejection Rule: Reject H_0 if the *p*-value $\leq \alpha$ or if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$ where $t_{\alpha/2}$ is based on a t distribution with $n - p - 1$ degrees of freedom.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

t Test for Significance of Individual Parameters (1 of 4)

- Hypotheses: $H_0\colon\ b_i = 0$

  $H_a\colon b_i \neq 0$

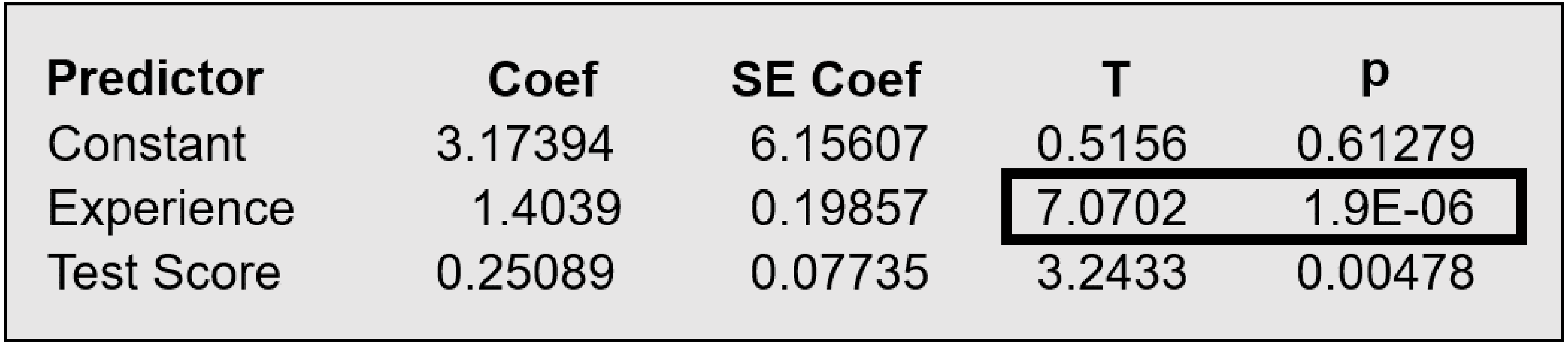
- Rejection Rule: For $\alpha = 0.05$ and *df* = 17; $t_{0.025} = 2.11$

  Reject H_0 if the *p*-value ≤ 0.05 or $t \leq -2.11$ or if $t \geq 2.11$

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

t Test for Significance of Individual Parameters (2 of 4)

Regression Equation Output

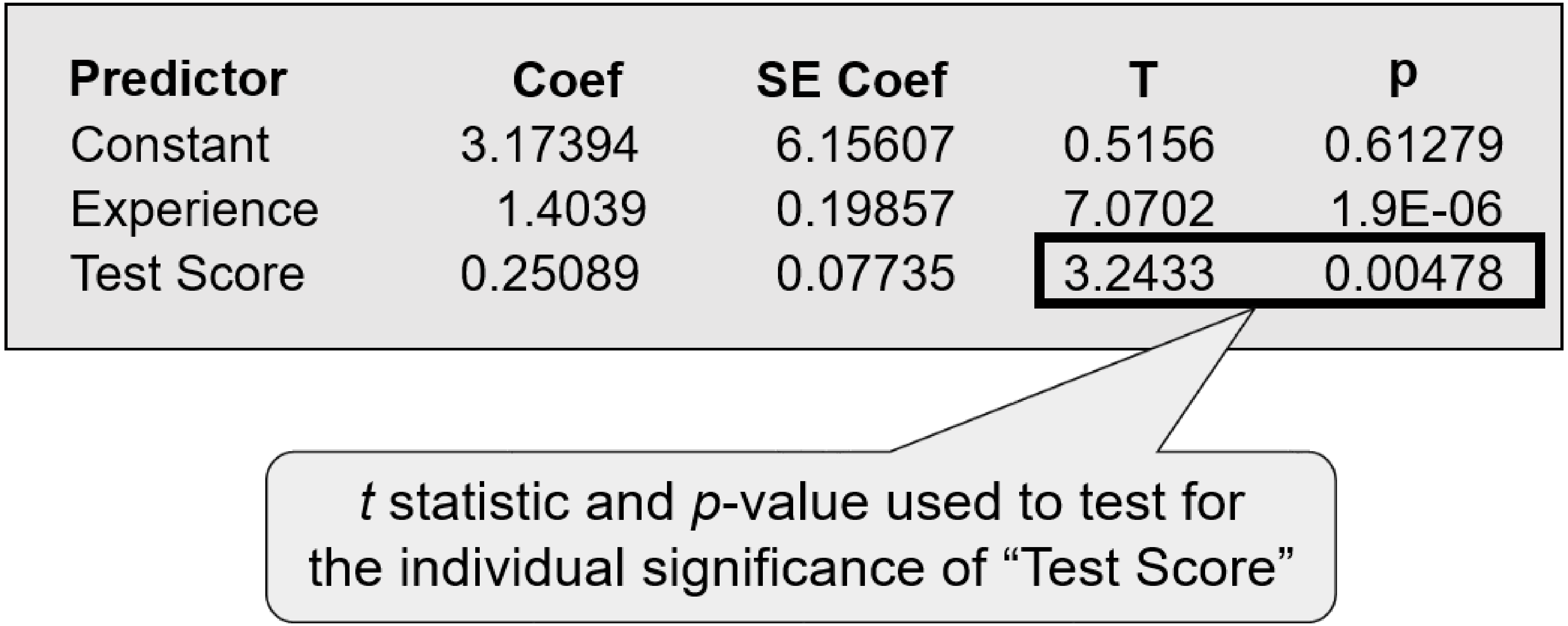


```
Predictor        Coef     SE Coef        T          p
Constant      3.17394     6.15607   0.5156    0.61279
Experience     1.4039     0.19857   7.0702    1.9E-06
Test Score    0.25089     0.07735   3.2433    0.00478
```

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

t Test for Significance of Individual Parameters (3 of 4)

Regression Equation Output



Predictor	Coef	SE Coef	T	p
Constant	3.17394	6.15607	0.5156	0.61279
Experience	1.4039	0.19857	7.0702	1.9E-06
Test Score	0.25089	0.07735	3.2433	0.00478

t statistic and *p*-value used to test for the individual significance of “Test Score”

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

t Test for Significance of Individual Parameters (4 of 4)

- Test Statistics:

$$t = \frac{b_1}{s_{b_1}} = \frac{1.4039}{0.1986} = 7.07$$

$$t = \frac{b_2}{s_{b_2}} = \frac{0.25089}{0.07735} = 3.24$$

- Conclusions

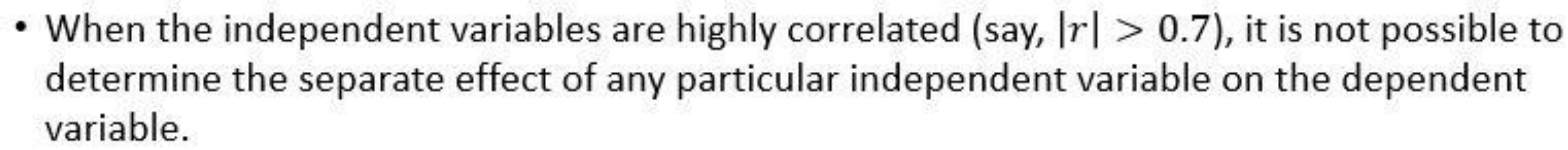
Reject both $H_0\text{: } \beta_1 = 0$ and $H_0\text{: } \beta_2 = 0$.

Both independent variables are significant.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Testing for Significance: Multicollinearity (共線性)

- The term multicollinearity refers to the correlation among the independent variables.



- When the independent variables are highly correlated (say, $|r| > 0.7$), it is not possible to determine the separate effect of any particular independent variable on the dependent variable.

- If the estimated regression equation is to be used only for predictive purposes, multicollinearity is usually not a serious problem.

- Every attempt should be made to avoid including independent variables that are highly correlated.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Chapter 16 - Regression Analysis: Model Building

- 16.1 – General Linear Model
- 16.2 – Determining When to Add or Delete Variables
- 16.3 – Analysis of a Larger Problem
- 16.4 – Variable Selection Procedures
- 16.5 – Multiple Regression Approach to Experimental Design
- 16.6 – Autocorrelation and the Durbin-Watson Test

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable Selection Procedures

1. Stepwise Regression
2. Forward Selection
3. Backward Elimination
4. Best-Subsets Regression

- The first three procedures are Iterative; one independent variable at a time is added or deleted based on the F statistic.
- The first 3 procedures are heuristics and therefore offer no guarantee that the best model will be found.
- In the fourth procedure different subsets of the independent variables are evaluated.

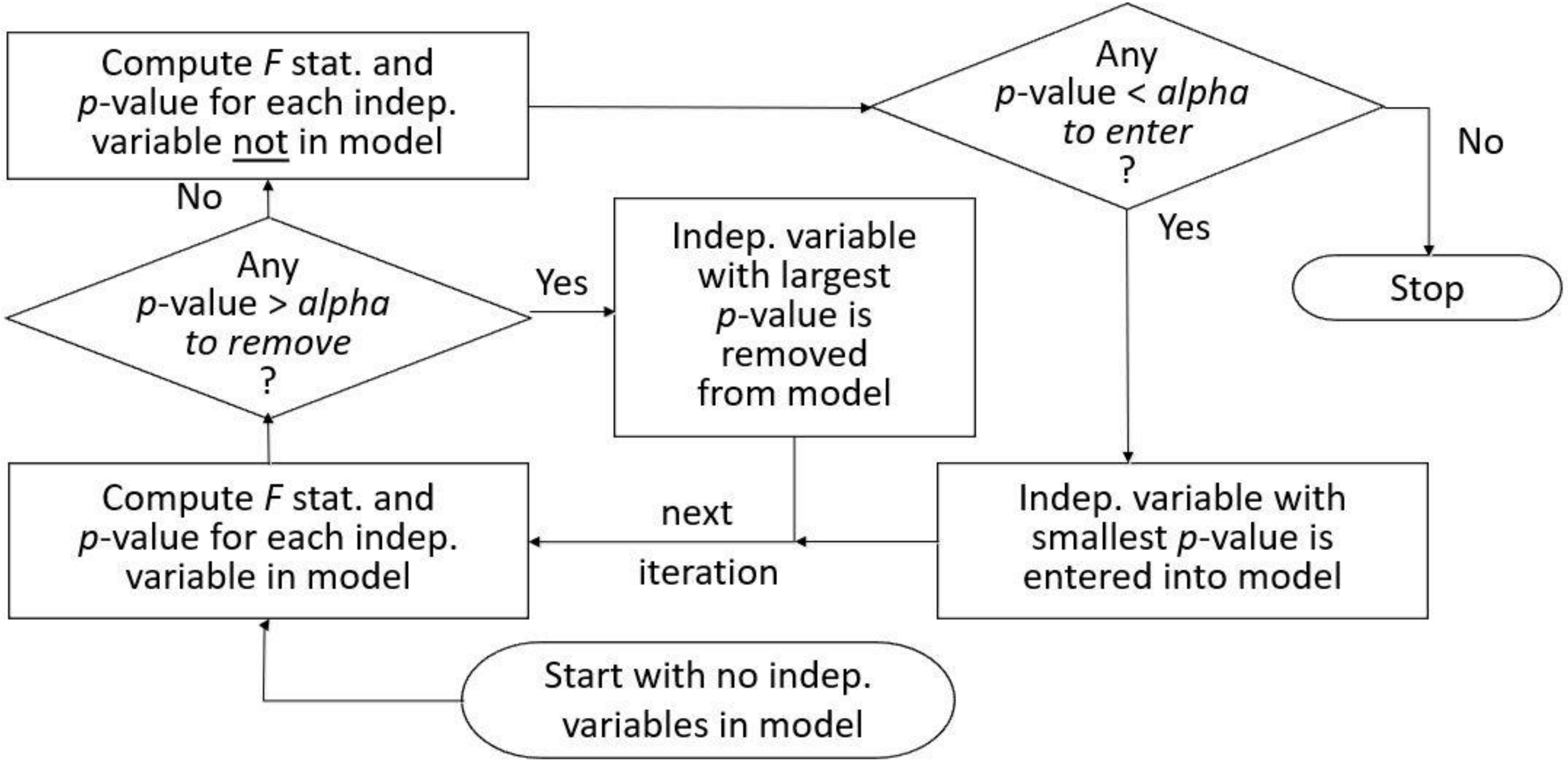
© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable Selection: Stepwise Regression (1 of 2)

- At each iteration, the first consideration is to see whether the least significant variable currently in the model can be removed because its F value is less than the user-specified or default *Alpha to remove*.
- If no variable can be removed, the procedure checks to see whether the most significant variable not in the model can be added because its F value is greater than the user-specified or default *Alpha to enter*.
- If no variable can be removed and no variable can be added, the procedure stops.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable Selection: Stepwise Regression (2 of 2)


Compute *F* stat. and *p*-value for each indep. variable not in model
Any *p*-value < *alpha to enter* ?
No
Stop
Yes
No
Any *p*-value > *alpha to remove* ?
Yes
Indep. variable with largest *p*-value is removed from model
Compute *F* stat. and *p*-value for each indep. variable in model
next iteration
Indep. variable with smallest *p*-value is entered into model
Start with no indep. variables in model

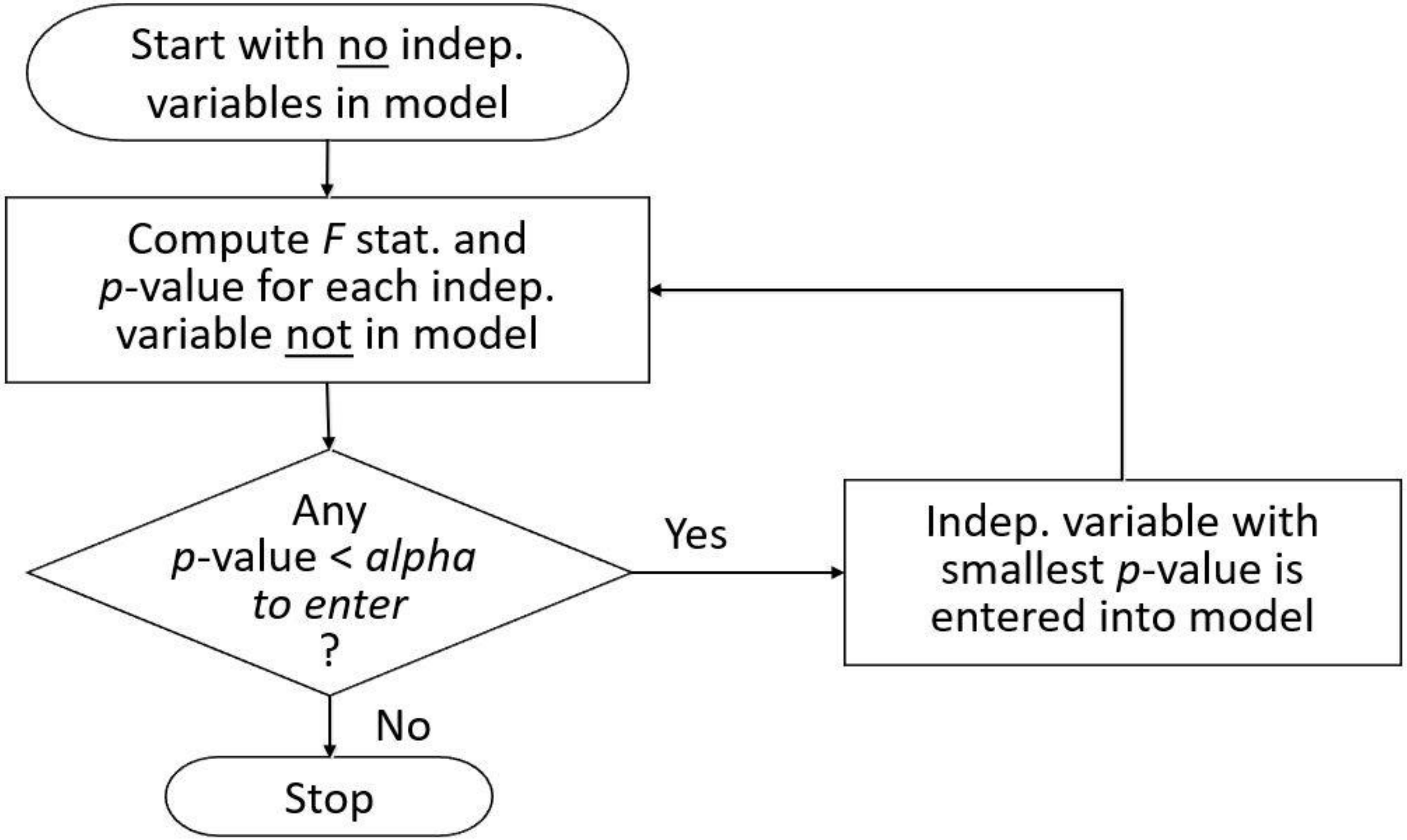
© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable Selection: Forward Selection (1 of 2)

- This procedure is similar to stepwise regression, but does not permit a variable to be deleted.
- It adds variables one at a time as long as a significant reduction in the error sum of squares (SSE) can be achieved.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable Selection: Forward Selection (2 of 2)



Start with no indep. variables in model
Compute *F* stat. and *p*-value for each indep. variable not in model
Any *p*-value < *alpha to enter* ?
Yes
Indep. variable with smallest *p*-value is entered into model
No
Stop

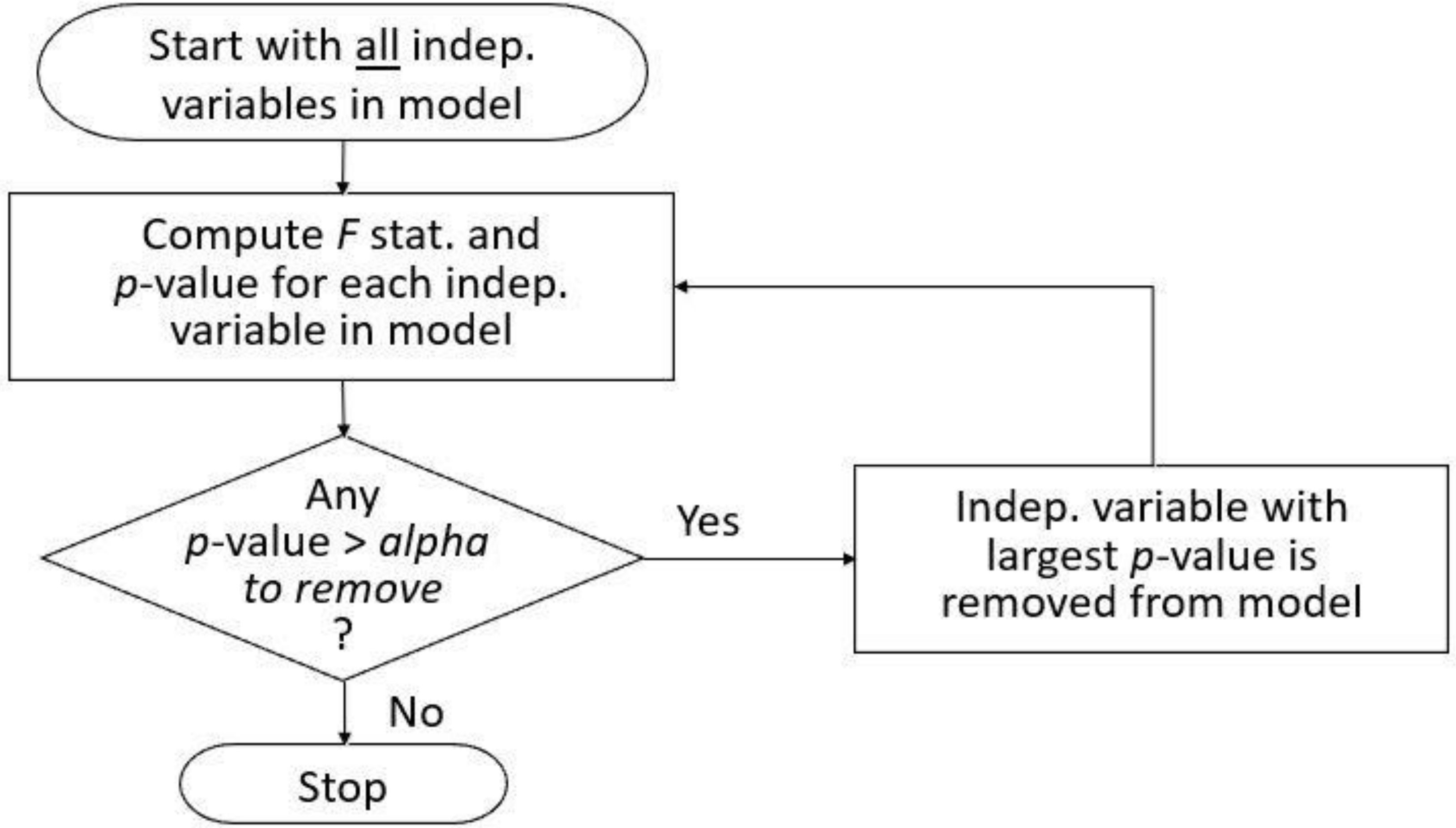
© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable Selection: Backward Elimination (1 of 2)

- This procedure begins with a model that includes all the independent variables the modeler wants considered.
- It then attempts to delete one variable at a time by determining whether the least significant variable currently in the model can be removed because its *p*-value is less than the user-specified or default value.
- Once a variable has been removed from the model, it cannot reenter at a subsequent step.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable Selection: Backward Elimination (2 of 2)


Start with all indep. variables in model
Compute *F* stat. and *p*-value for each indep. variable in model
Any *p*-value > *alpha to remove* ?
Yes
Indep. variable with largest *p*-value is removed from model
No
Stop

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable Selection: Best-Subsets Regression (1 of 2)

- The three preceding procedures are one-variable-at-a-time methods offering no guarantee that the best model for a given number of variables will be found.
- Minitab output identifies the two best one-variable estimated regression equations, the two best two-variable equations, and so on.
- Some software packages include best-subsets regression that enables the user to find, given a specified number of independent variables, the best regression model.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable Selection: Best-Subsets Regression (2 of 2)

Example: PGA Tour Data

The Professional Golfers Association keeps a variety of statistics regarding performance measures. Data include the average driving distance, percentage of drives that land in the fairway, percentage of greens hit in regulation, average number of putts, percentage of sand saves, and average score.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable-Selection Procedures (1 of 8)

Variable Names and Definitions

Drive: average length of a drive in yards

Fair: percentage of drives that land in the fairway

Green: percentage of greens hit in regulation (a par-3 green is “hit in regulation” if the player’s first shot lands on the green)

Putt: average number of putts for greens that have been hit in regulation

Sand: percentage of sand saves (landing in a sand trap and still scoring par or better)

Score: average score for an 18-hole round

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable-Selection Procedures (2 of 8)

Sample Data (Part 1)

Drive	Fair	Green	Putt	Sand	Score
277.6	.681	.667	1.768	.550	69.10
259.6	.691	.665	1.810	.536	71.09
269.1	.657	.649	1.747	.472	70.12
267.0	.689	.673	1.763	.672	69.88
267.3	.581	.637	1.781	.521	70.71
255.6	.778	.674	1.791	.455	69.76
272.9	.615	.667	1.780	.476	70.19
265.4	.718	.699	1.790	.551	69.73

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable-Selection Procedures (3 of 8)

Sample Data (Part 2)

Drive	Fair	Green	Putt	Sand	Score
272.6	.660	.672	1.803	.431	69.97
263.9	.668	.669	1.774	.493	70.33
267.0	.686	.687	1.809	.492	70.32
266.0	.681	.670	1.765	.599	70.09
258.1	.695	.641	1.784	.500	70.46
255.6	.792	.672	1.752	.603	69.49
261.3	.740	.702	1.813	.529	69.88
262.2	.721	.662	1.754	.576	70.27

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable-Selection Procedures (4 of 8)

Sample Data (Part 3)

Drive	Fair	Green	Putt	Sand	Score
260.5	.703	.623	1.782	.567	70.72
271.3	.671	.666	1.783	.492	70.30
263.3	.714	.687	1.796	.468	69.91
276.6	.634	.643	1.776	.541	70.69
252.1	.726	.639	1.788	.493	70.59
263.0	.639	.647	1.760	.374	70.81
253.5	.732	.693	1.797	.518	70.26
266.2	.681	.657	1.812	.472	70.96

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable-Selection Procedures (5 of 8)

Sample Correlation Coefficients

	Score	Drive	Fair	Green	Putt
Drive	-.154				
Fair	-.427	-.679			
Green	-.556	-.045	.421		
Putt	.258	-.139	.101	.354	
Sand	-.278	-.024	.265	.083	-.296

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable-Selection Procedures (6 of 8)

Best Subsets Regression of SCORE

Varsx	R-sq	R-sq(a)	C-p	s	D	F	G	P	S
1	30.9	27.9	26.9	.39685			X		
1	18.2	14.6	35.7	.43183		X			
2	54.7	50.5	12.4	.32872	X	X			
2	54.6	50.5	12.5	.32891			X	X	
3	60.7	55.1	10.2	.31318	X	X		X	
3	59.1	53.3	11.4	.31957	X	X	X		
4	72.2	66.8	4.2	.26913	X	X	X	X	
4	60.9	53.1	12.1	.32011	X	X		X	X
5	72.6	65.4	6.0	.27499	X	X	X	X	X

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable-Selection Procedures (7 of 8)

Minitab Output

The regression equation

Score = 74.678 - .0398(Drive) - 6.686(Fair) - 10.342(Green) + 9.858(Putt)

Predictor	**Coef**	**Stdev**	**t-ratio**	**p**
Constant	74.678	6.952	10.74	.000
Drive	-.0398	.01235	-3.22	.004
Fair	-6.686	1.939	-3.45	.003
Green	-10.342	3.561	-2.90	.009
Putt	9.858	3.180	3.10	.006

s = .2691 **R-sq** = 72.4% **R-sq(adj)** = 66.8%

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Variable-Selection Procedures (8 of 8)

Minitab Output

Analysis of Variance

SOURCE	**DF**	**SS**	**MS**	***F***	***P***
Regression	4	3.79469	.94867	13.10	.000
Error	20	1.44865	.07243		
Total	24	5.24334			

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Logistic regression

- Most important model for ***categorical response*** (y_i) data
- Categorical response with 2 levels (*binary*: 0 and 1)
- Categorical response with $\geq$ 3 levels (nominal or ordinal)
- Predictor variables (x_i) can take on *any* form: binary, categorical, and/or continuous

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Logistic Regression

- Models relationship between set of variables X_i
  - dichotomous (yes/no, smoker/nonsmoker,…)
  - categorical (social class, race, ...)
  - continuous (age, weight, gestational age, ...)

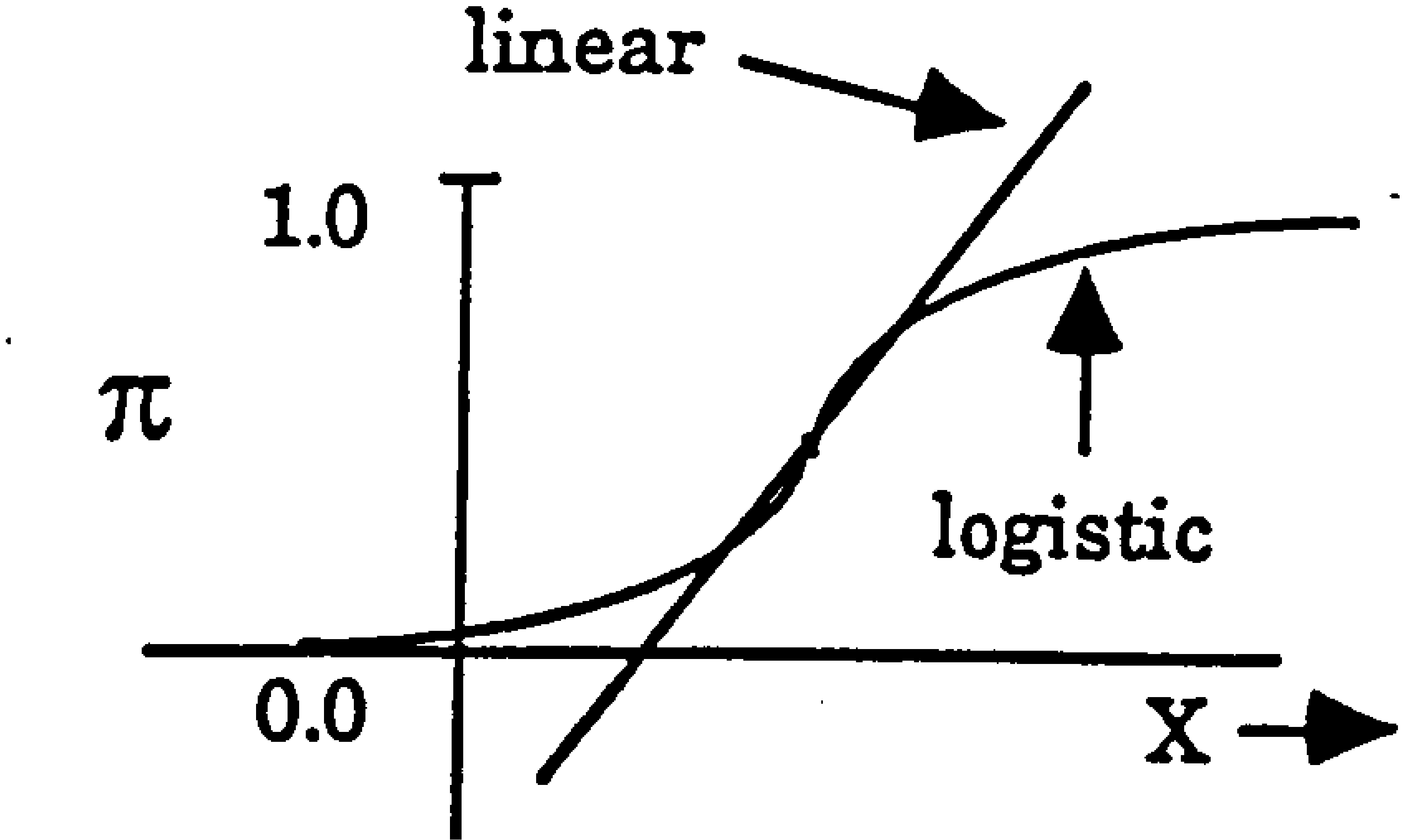
and

- dichotomous categorical response variable Y

e.g. Success/Failure, Remission/No Remission, Survived/Died, CHD/No CHD, Low Birth Weight/Normal Birth Weight…

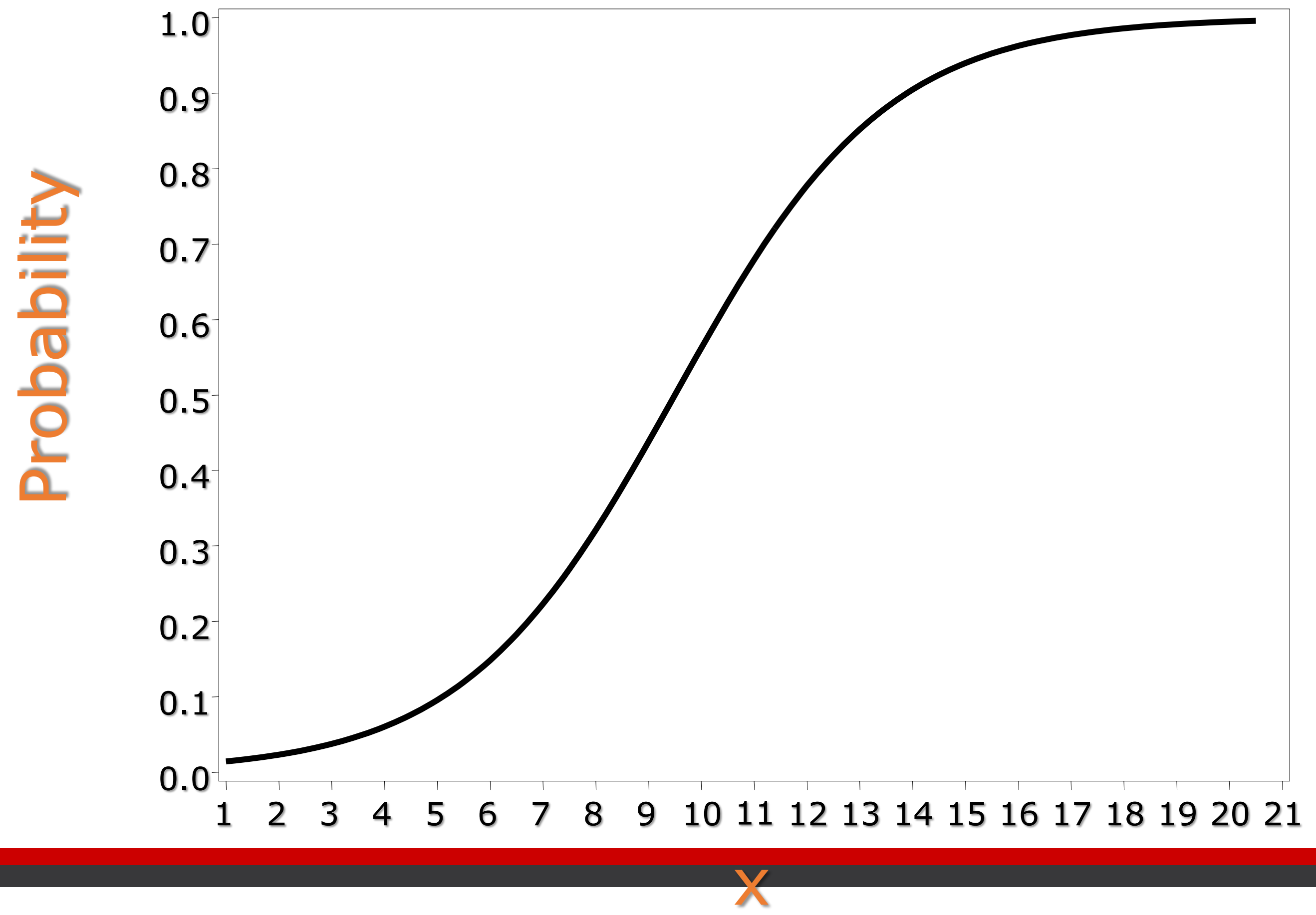
© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Sigmoid curve for logistic regression



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Logistic Regression Curve



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Logit Transformation

Logistic regression models transform probabilities called *logits*.

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

where

i indexes all cases (observations).

p_i is the probability the event (a sale, for example) occurs in the *i*th case.

log is the natural log (to the base *e*).

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Logistic regression model with a single continuous predictor

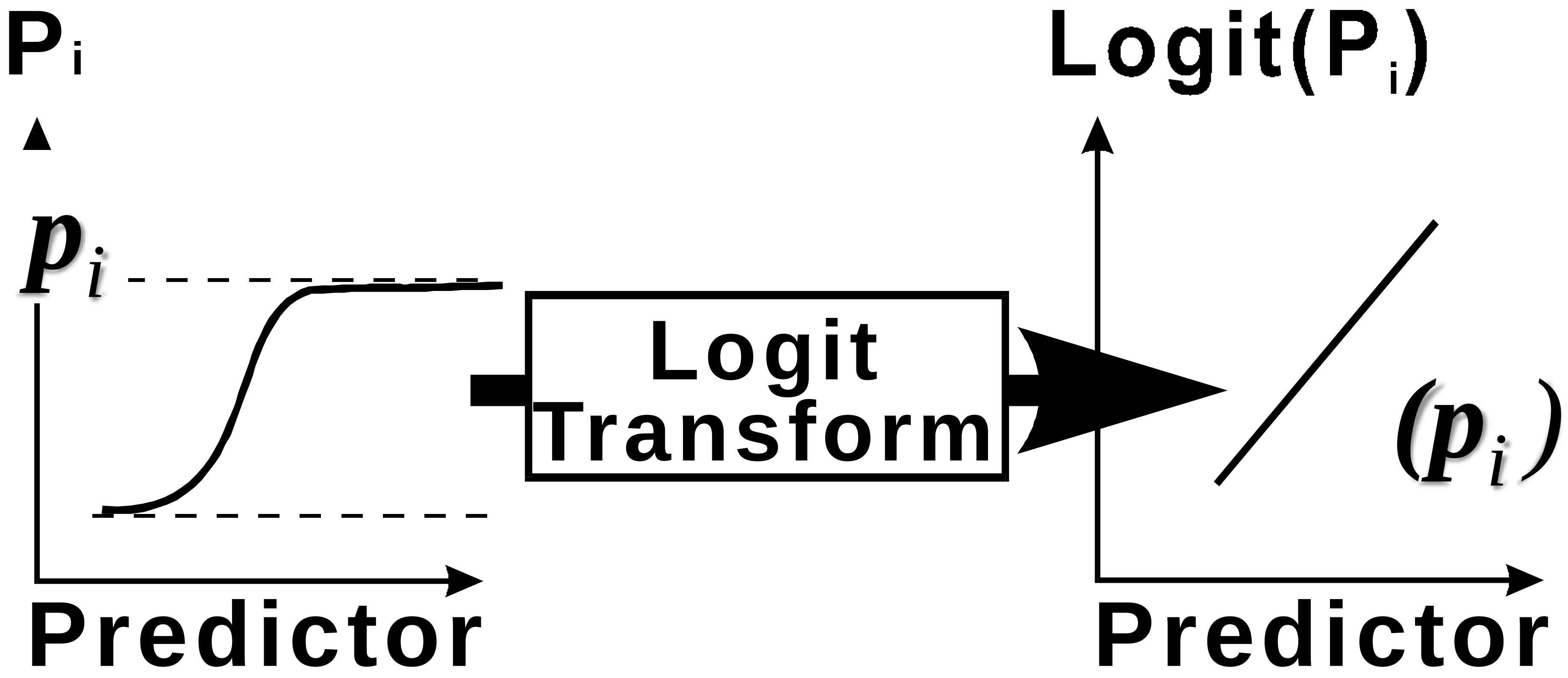
$$\text{logit}\ (p_i) = \log\ (\text{odds}) = \beta_0 + \beta_1 X_1$$

where

logit(p_i)	logit transformation of the probability of the event
β_0	intercept of the regression line
β_1	slope of the regression line

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Assumption



© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Interpretation of a single *continuous* parameter

- The sign (±) of β determines whether the **log odds** of y is increasing or decreasing *for every 1-unit increase* in x.
- If β > 0, there is an increase in the **log odds** of y for every 1-unit increase in x.
- If β < 0, there is a decrease in the **log odds** of y for every 1-unit increase in x.
- If β = 0 there is *no linear relationship* between the **log odds** and x.

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

Parameter Interpretation (ctd).

- Exponentiating both sides of the logit link function we get the following:

$$\left(\frac{p_i}{1-p_i}\right) = \text{odds} = \exp(\beta_0 + \beta_1 X_1) = e^{\beta 0}\ e^{\beta 1 X1}$$

- The odds increase **multiplicatively** by e^{β} for every 1-unit increase in *x*.
- Whether the increase is greater than 1 or less than one depends on whether $\beta > 0$ or $\beta < 0$.
- The odds at $X = x+1$ are e^{β} times the odds at $X = x$. Therefore, ***e*^β^ is an odds ratio!**

Logistic regression model with a single *categorical ($\geq$ 2 levels)* predictor

$$logit\ (p_i) = log\ (odds) = \beta_0 + \beta_k X_k$$

where

logit(p_i) logit transformation of the probability of the event

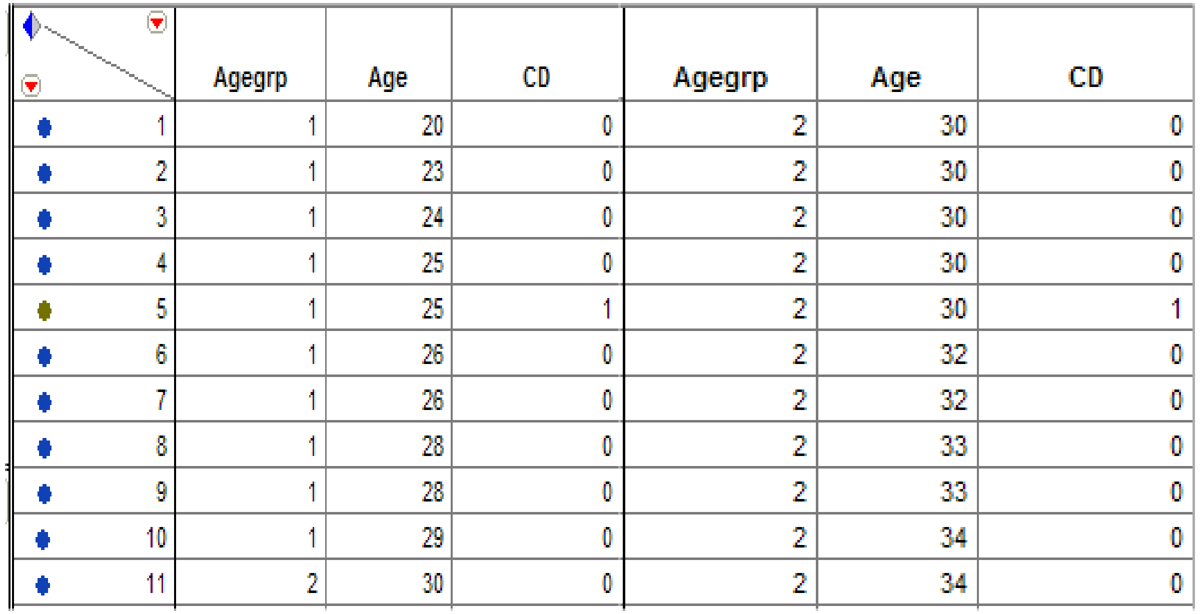
β_0 intercept of the regression line

β_k difference between the logits for category k vs. the reference category

© 2020 Cengage. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part, except for use as permitted in a license distributed with a certain product or service or otherwise on a password-protected website or school-approved learning management system for classroom use.

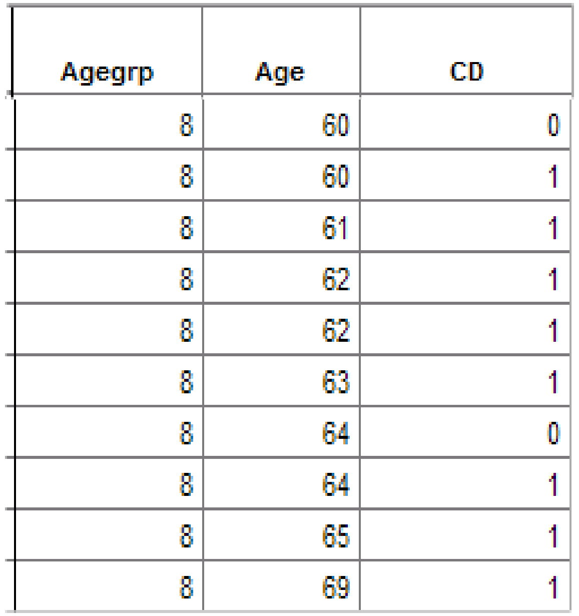
Logistic Regression

Example: Coronary Heart Disease (CD) and Age In this study sampled individuals were examined for signs of CD (present = 1/absent = 0) and its potential relationship with the age (yrs.) was considered.



	Agegrp	Age	CD	Agegrp	Age	CD
1	1	20	0	2	30	0
2	1	23	0	2	30	0
3	1	24	0	2	30	0
4	1	25	0	2	30	0
5	1	25	1	2	30	1
6	1	26	0	2	32	0
7	1	26	0	2	32	0
8	1	28	0	2	33	0
9	1	28	0	2	33	0
10	1	29	0	2	34	0
11	2	30	0	2	34	0

• • •

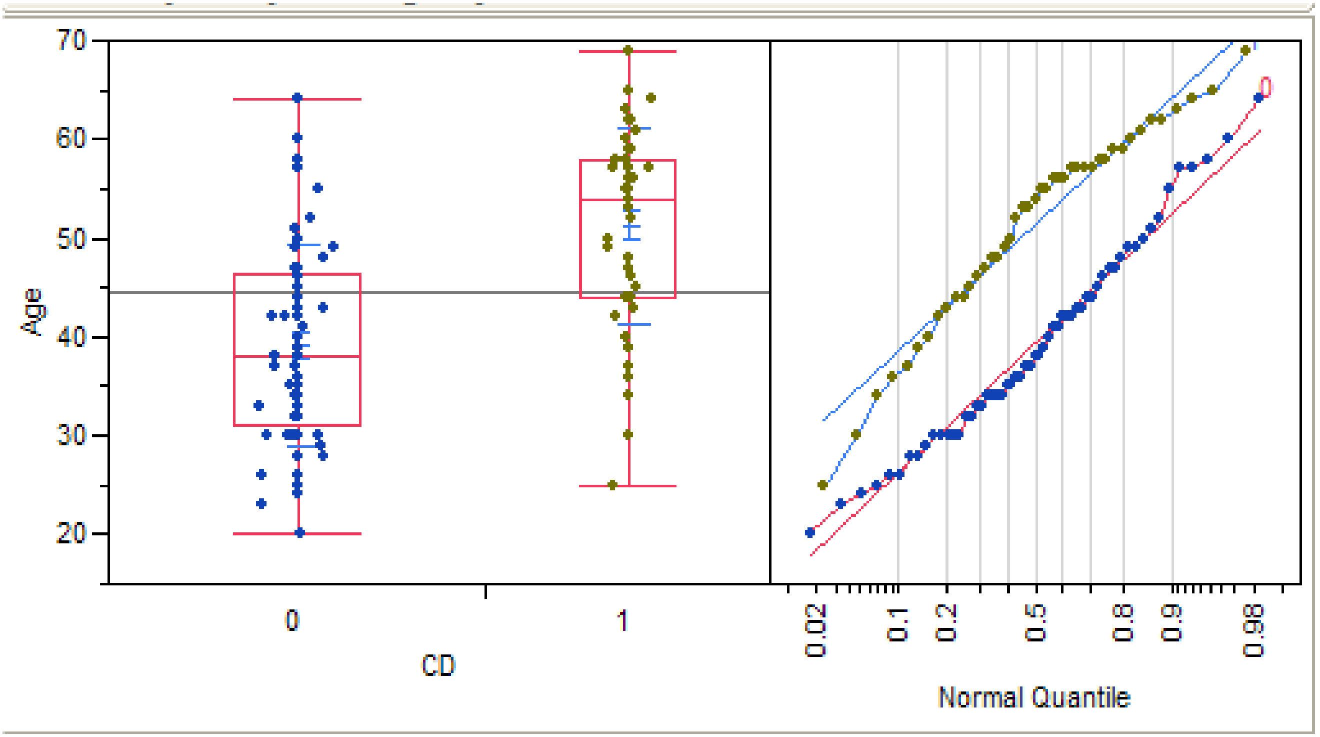


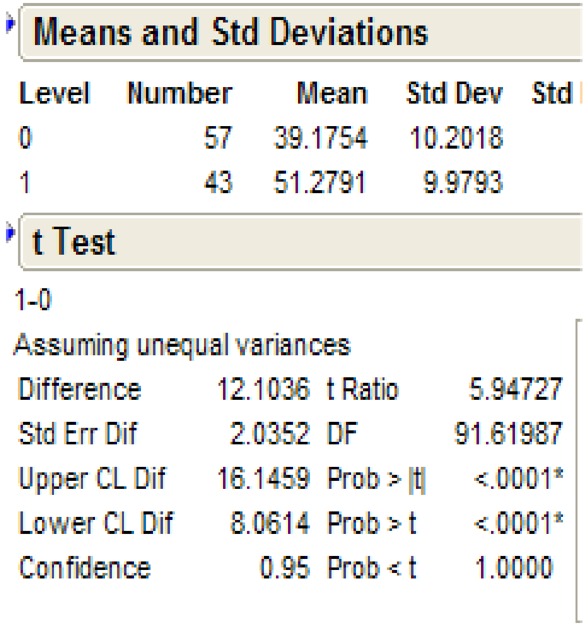
Agegrp	Age	CD
8	60	0
8	60	1
8	61	1
8	62	1
8	62	1
8	63	1
8	64	0
8	64	1
8	65	1
8	69	1

Note: There are 100 subjects participating in the study.

Logistic Regression

- How can we analyze these data?





Means and Std Deviations

Level	Number	Mean	Std Dev	Std
0	57	39.1754	10.2018	
1	43	51.2791	9.9793	

t Test

1-0

Assuming unequal variances

Difference	12.1036	t Ratio	5.94727
Std Err Dif	2.0352	DF	91.61987
Upper CL Dif	16.1459	Prob > |t|	<.0001*
Lower CL Dif	8.0614	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000

Non-pooled t-test

The mean age of the individuals with some signs of coronary heart disease is 51.28 years vs. 39.18 years for individuals without signs (t = 5.95, p < .0001).

© 2020 Cengage. M
password-protectec