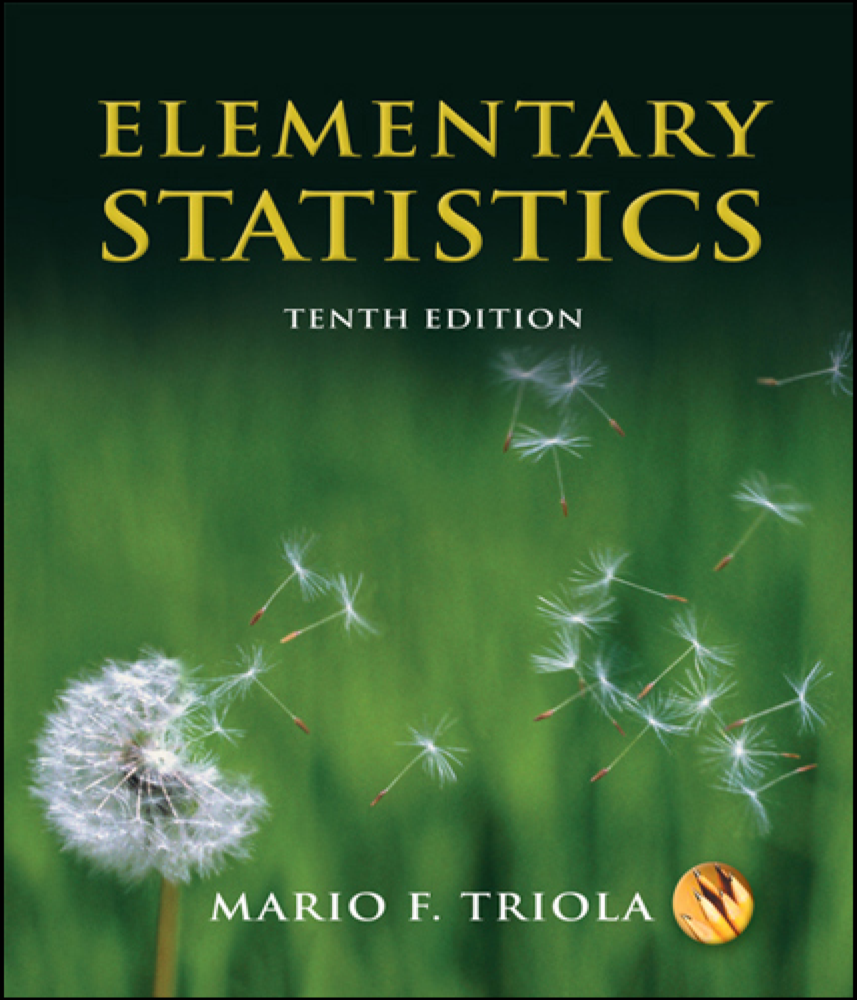
Lecture Slides



Elementary Statistics
Tenth Edition

and the Triola Statistics Series

by Mario F. Triola

Chapter 10
Correlation and Regression

10-1 Overview

10-2 Correlation

10-3 Regression

10-4 Variation and Prediction Intervals

10-5 Multiple Regression

10-6 Modeling

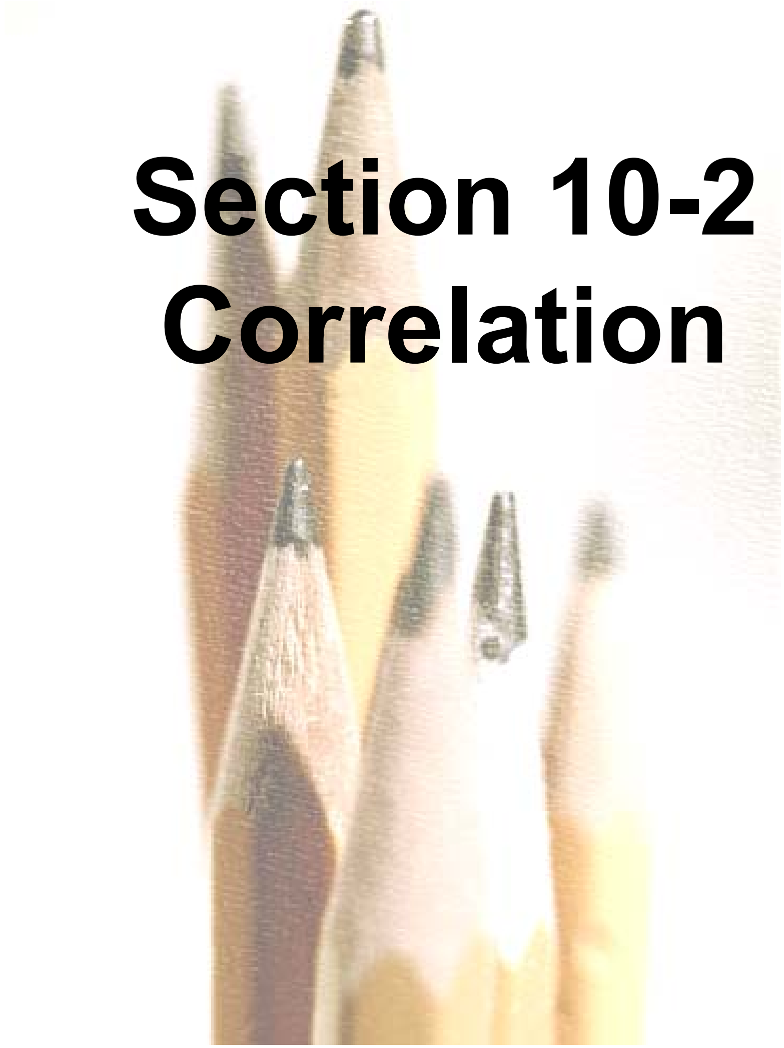
Section 10-1
Overview

Created by Erin Hodgess, Houston, Texas
Revised to accompany 10th Edition, Tom Wegleitner, Centreville, VA

Overview

This chapter introduces important methods for making inferences about a correlation (or relationship) between two variables, and describing such a relationship with an equation that can be used for predicting the value of one variable given the value of the other variable.

We consider sample data that come in pairs.



Section 10-2
Correlation

Created by Erin Hodgess, Houston, Texas
Revised to accompany 10th Edition, Tom Wegleitner, Centreville, VA

PEARSON
Addison Wesley

Key Concept

This section introduces the **linear correlation coefficient *r***, which is a numerical measure of the strength of the relationship between two variables representing quantitative data.

Because technology can be used to find the value of *r*, it is important to focus on the concepts in this section, without becoming overly involved with tedious arithmetic calculations.

Part 1: Basic Concepts of Correlation

Definition

A **correlation** exists between two variables when one of them is related to the other in some way.

Definition

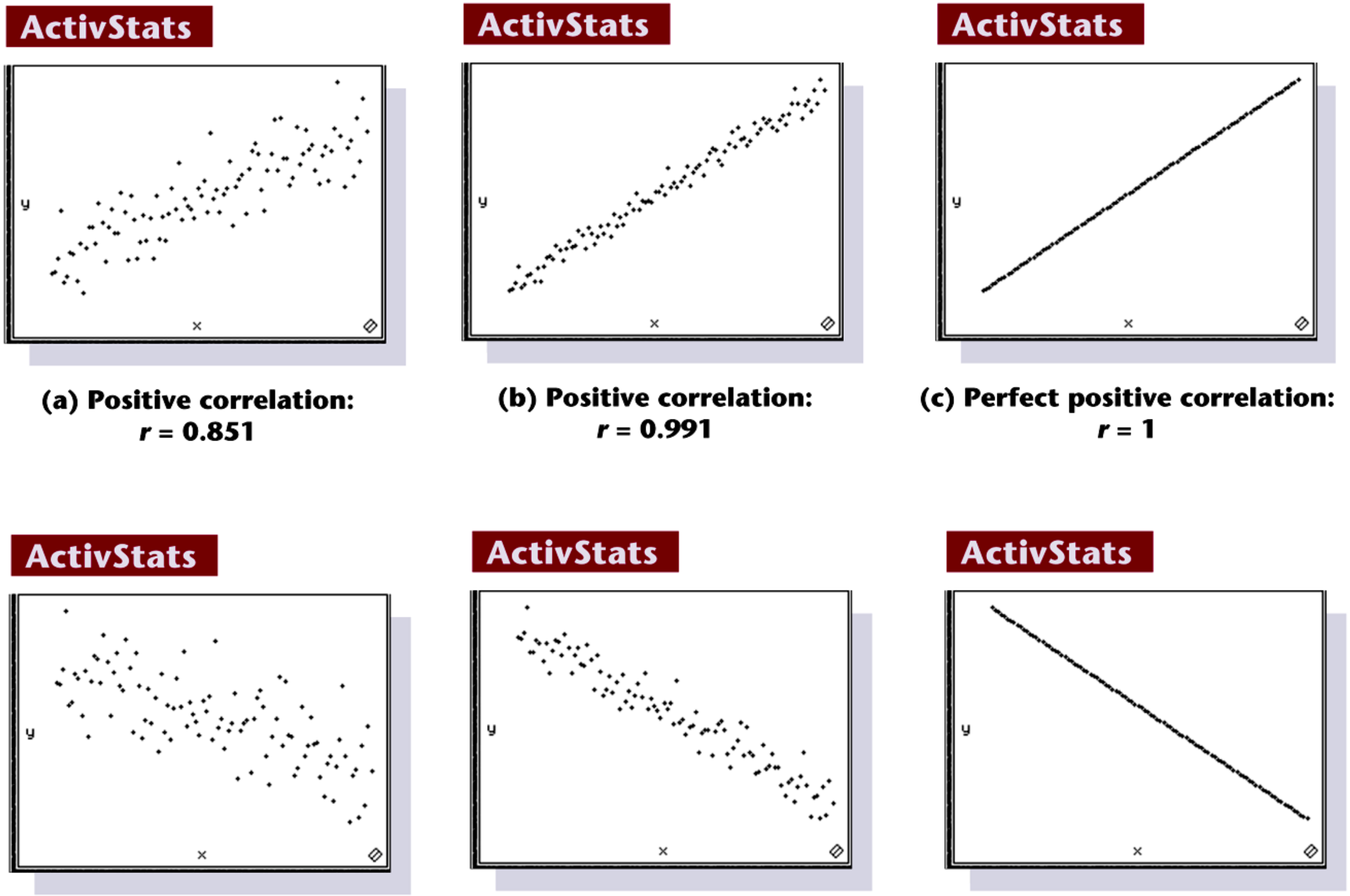
The **linear correlation coefficient** ***r*** measures the strength of the linear relationship between paired *x*- and *y*-quantitative values in a **sample**.

Exploring the Data

We can often see a relationship between two variables by constructing a scatterplot.

Figure 10-2 following shows scatterplots with different characteristics.

Scatterplots of Paired Data



(a) Positive correlation:
***r* = 0.851**

(b) Positive correlation:
***r* = 0.991**

(c) Perfect positive correlation:
***r* = 1**

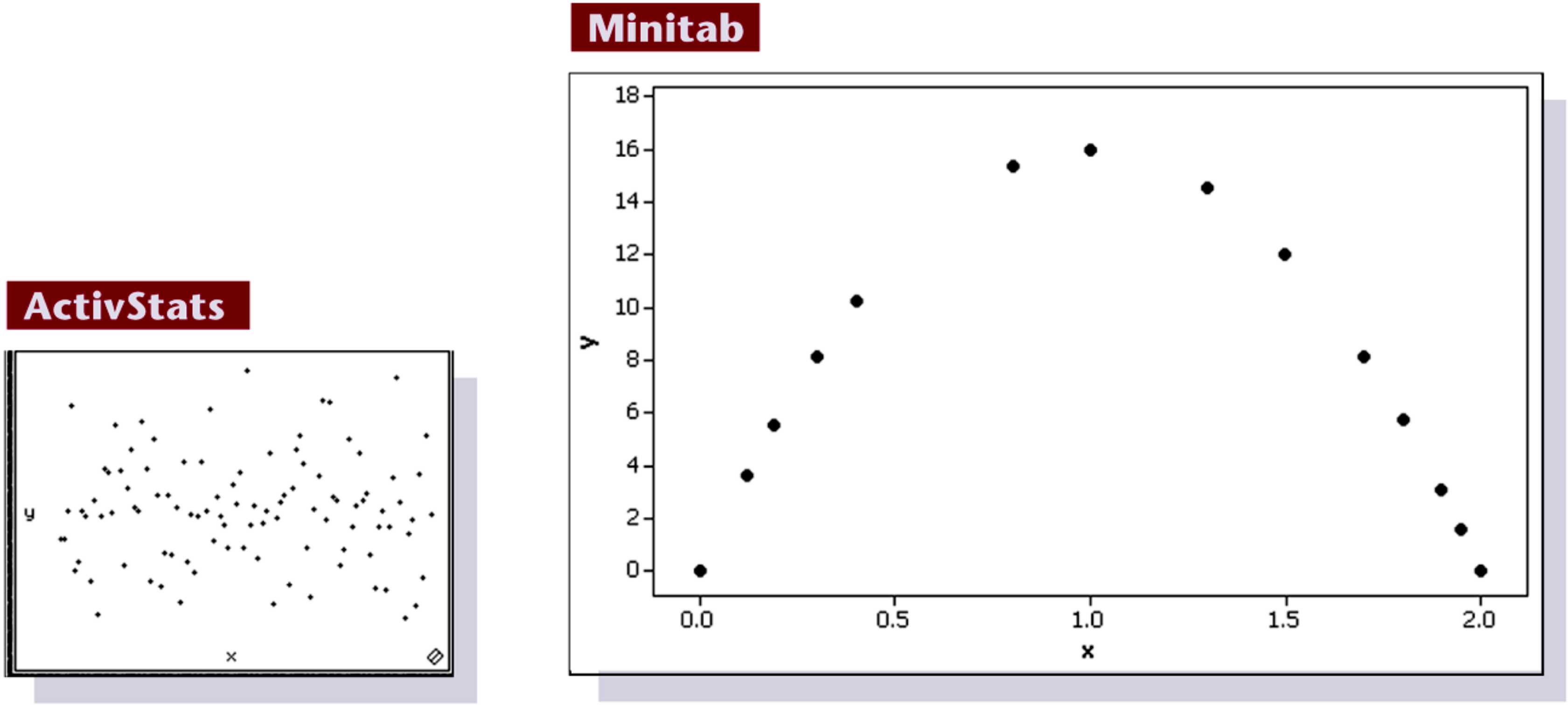
(d) Negative correlation:
***r* = −0.702**

(e) Negative correlation:
***r* = −0.965**

(f) Perfect negative correlation:
***r* = −1**

Figure 10-2

Scatterplots of Paired Data


ActivStats

Minitab

(g) No correlation: *r* = 0

(h) Nonlinear relationship: *r* = −0.087

Figure 10-2

Requirements

1. The sample of paired (x, y) data is a random sample of independent quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. The outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating *r* with and without the outliers included.

Notation for the Linear Correlation Coefficient

n represents the number of pairs of data present.

Σ denotes the addition of the items indicated.

Σx denotes the sum of all x-values.

Σx^2 indicates that each x-value should be squared and then those squares added.

$(\Sigma x)^2$ indicates that the x-values should be added and the total then squared.

Σxy indicates that each x-value should be first multiplied by its corresponding y-value. After obtaining all such products, find their sum.

r represents linear correlation coefficient for a sample.

ρ represents linear correlation coefficient for a population.

Formula

The **linear correlation coefficient** r measures the strength of a linear relationship between the paired values in a **sample**.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

Formula 10-1

Calculators can compute r

Interpreting *r*

Using Table A-6: If the absolute value of the computed value of *r* exceeds the value in Table A-6, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

Using Software: If the computed *P*-value is less than or equal to the significance level, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

Rounding the Linear Correlation Coefficient *r*

- Round to three decimal places so that it can be compared to critical values in Table A-6.
- Use calculator or computer if possible.

Example: Calculating *r*

Using the simple random sample of data below, find the value of *r*.

Data

x	3	1	3	5
y	5	8	6	4

Example: Calculating *r* - cont

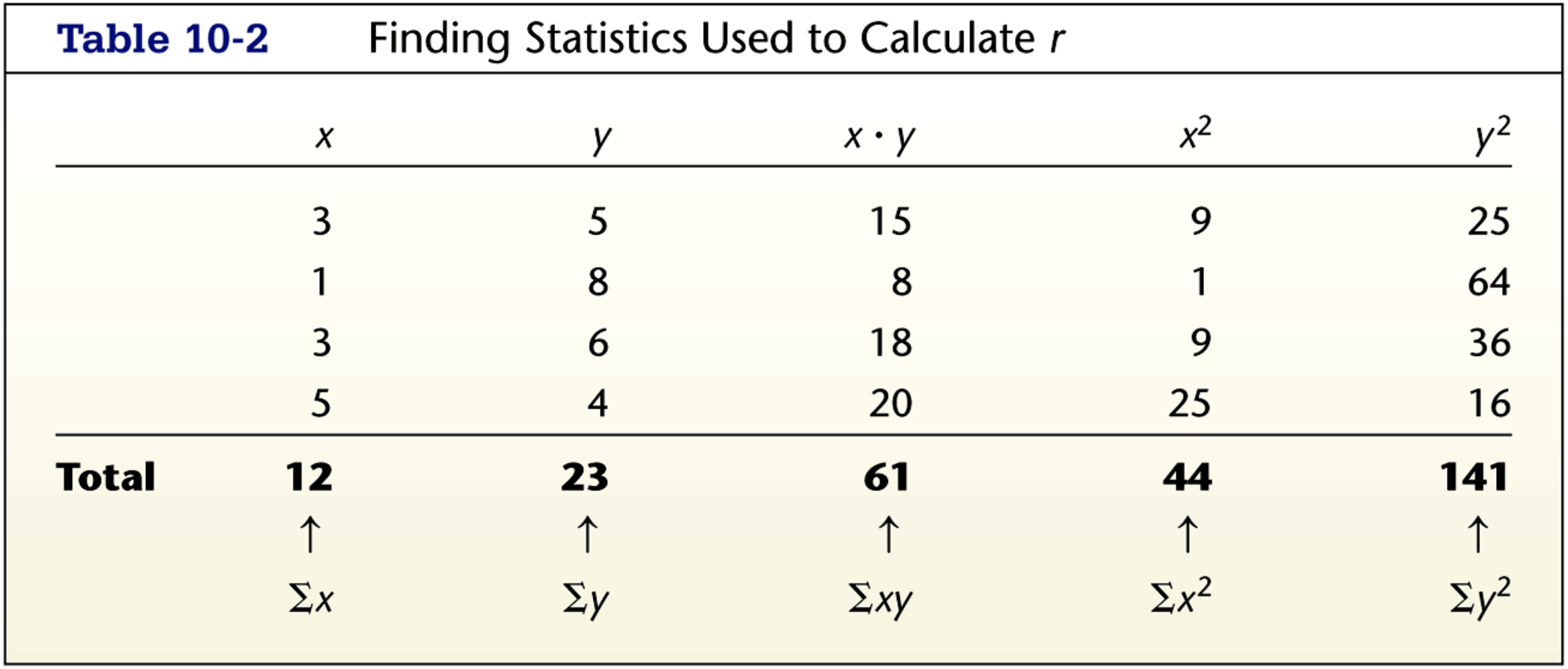


Table 10-2 Finding Statistics Used to Calculate *r*

	x	y	$x \cdot y$	x^2	y^2
	3	5	15	9	25
	1	8	8	1	64
	3	6	18	9	36
	5	4	20	25	16
Total	**12**	**23**	**61**	**44**	**141**
	↑	↑	↑	↑	↑
	Σx	Σy	Σxy	Σx^2	Σy^2

Example: Calculating *r* - cont

Data

x	3	1	3	5
y	5	8	6	4

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$$r = \frac{4(61) - (12)(23)}{\sqrt{4(44) - (12)^2}\ \sqrt{4(141) - (23)^2}}$$

$$r = \frac{-32}{33.466} = -0.956$$

Example: Calculating r - cont

Given r = - 0.956, if we use a 0.05 significance level we conclude that there is a linear correlation between x and y since the absolute value of r exceeds the critical value of 0.950. However, if we use a 0.01 significance level, we do not conclude that there is a linear correlation because the absolute value of r does not exceed the critical value of 0.999.

Example: Old Faithful

Given the sample data in Table 10-1, find the value of the linear correlation coefficient *r*, then refer to Table A-6 to determine whether there is a significant linear correlation between the duration and interval of the eruption times.

Using the same procedure previously illustrated, we find that *r* = 0.926.

Referring to Table A-6, we locate the row for which *n* = 8. Using the critical value for α = 0.05, we have 0.707. Because *r* = 0.926, its absolute value exceeds 0.707, so we conclude that there is a linear correlation between the duration and interval after the eruption times.

Properties of the Linear Correlation Coefficient *r*

1. $-1 \le r \le 1$

2. The value of *r* does not change if all values of either variable are converted to a different scale.

3. The value of *r* is not affected by the choice of *x* and *y*. Interchange all *x*- and *y*-values and the value of *r* will not change.

4. *r* measures strength of a linear relationship.

Interpreting r : Explained Variation

The value of r^2 is the proportion of the variation in y that is explained by the linear relationship between x and y.

Example: Old Faithful

Using the duration/interval data in Table 10-1, we have found that the value of the linear correlation coefficient r = 0.926. What proportion of the variation of the interval after eruption times can be explained by the variation in the duration times?

With $r = 0.926$, we get $r^2 = 0.857$.

We conclude that 0.857 (or about 86%) of the variation in interval after eruption times can be explained by the linear relationship between the duration times and the interval after eruption times. This implies that 14% of the variation in interval after eruption times cannot be explained by the duration times.

Common Errors Involving Correlation

1. **Causation**: It is wrong to conclude that correlation implies causality.

2. **Averages**: Averages suppress individual variation and may inflate the correlation coefficient.

3. **Linearity**: There may be some relationship between x and y even when there is no linear correlation.

Part 2: Formal Hypothesis Test

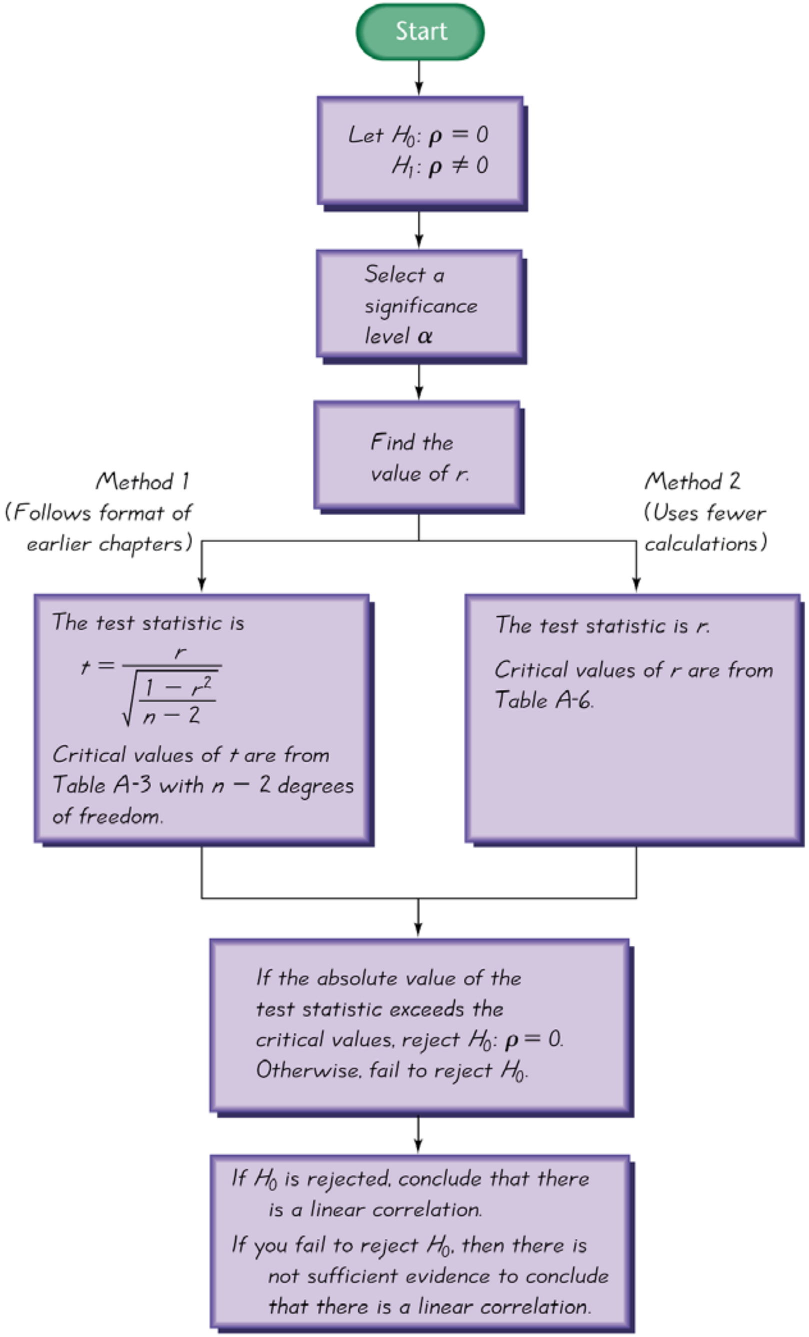
Formal Hypothesis Test

- We wish to determine whether there is a significant linear correlation between two variables.
- We present two methods.
- In both methods let

H_0: $\rho = 0$ (no significant linear correlation)

H_1: $\rho \neq 0$ (significant linear correlation)

Hypothesis Test for a Linear Correlation



Start

Let H_0: $\rho = 0$
H_1: $\rho \neq 0$

Select a significance level α

Find the value of r.

Method 1
(Follows format of earlier chapters)

Method 2
(Uses fewer calculations)

The test statistic is

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Critical values of t are from Table A-3 with $n - 2$ degrees of freedom.

The test statistic is r.

Critical values of r are from Table A-6.

If the absolute value of the test statistic exceeds the critical values, reject H_0: $\rho = 0$. Otherwise, fail to reject H_0.

If H_0 is rejected, conclude that there is a linear correlation.

If you fail to reject H_0, then there is not sufficient evidence to conclude that there is a linear correlation.

Figure 10-3

Method 1: Test Statistic is *t*

Test statistic:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Critical values:

Use Table A-3 with degrees of freedom = $n - 2$

Method 1 - cont

P-value:

Use Table A-3 with
degrees of freedom = $n - 2$

Conclusion:

If the absolute value of *t* is > critical value from Table A-3, reject H_0 and conclude that there is a linear correlation. If the absolute value of *t* ≤ critical value, fail to reject H_0; there is not sufficient evidence to conclude that there is a linear correlation.

Method 2:
Test Statistic is *r*

Test statistic: ***r***

Critical values: **Refer to Table A-6**

Conclusion:

If the absolute value of *r* is > critical value from Table A-6, reject H_0 and conclude that there is a linear correlation. If the absolute value of *r* ≤ critical value, fail to reject H_0; there is not sufficient evidence to conclude that there is a linear correlation.

Example: Old Faithful

Using the duration/interval data in Table 10-1, test the claim that there is a linear correlation between the duration of an eruption and the time interval after that eruption. Use Method 1.

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

$$t = \frac{0.926}{\sqrt{\frac{1 - 0.926^2}{8 - 2}}} = 6.008$$

Method 1: Test Statistic is *t*

(follows format of earlier chapters)

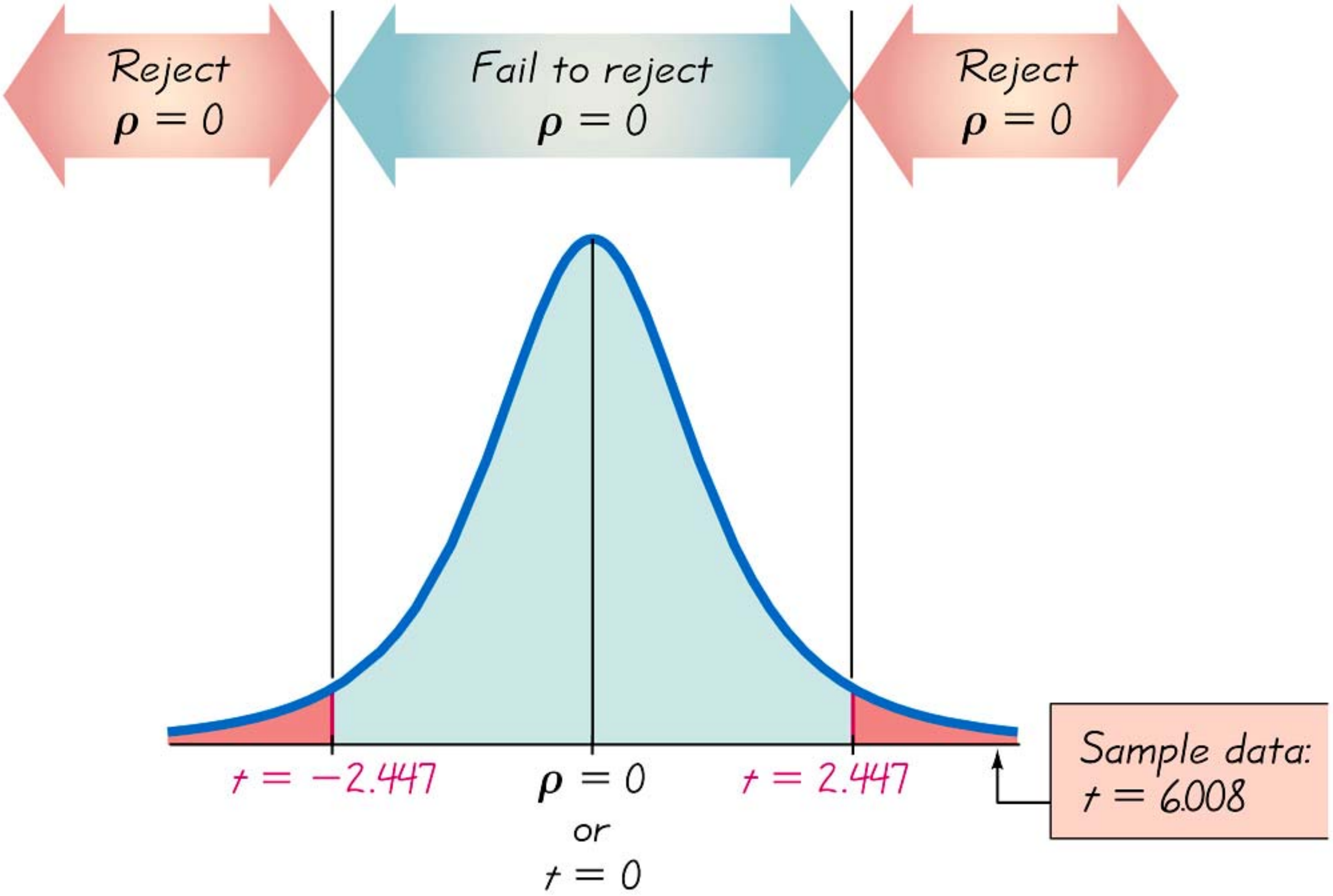


Figure 10-4

Example: Old Faithful

Using the duration/interval data in Table 10-1, test the claim that there is a linear correlation between the duration of an eruption and the time interval after that eruption. Use Method 2.

The test statistic is r = 0.926. The critical values of r = ±0.707 are found in Table A-6 with n = 8 and α = 0.05.

Method 2: Test Statistic is *r*

- **Test statistic: *r***
- **Critical values: Refer to Table A-6 (8 degrees of freedom)**

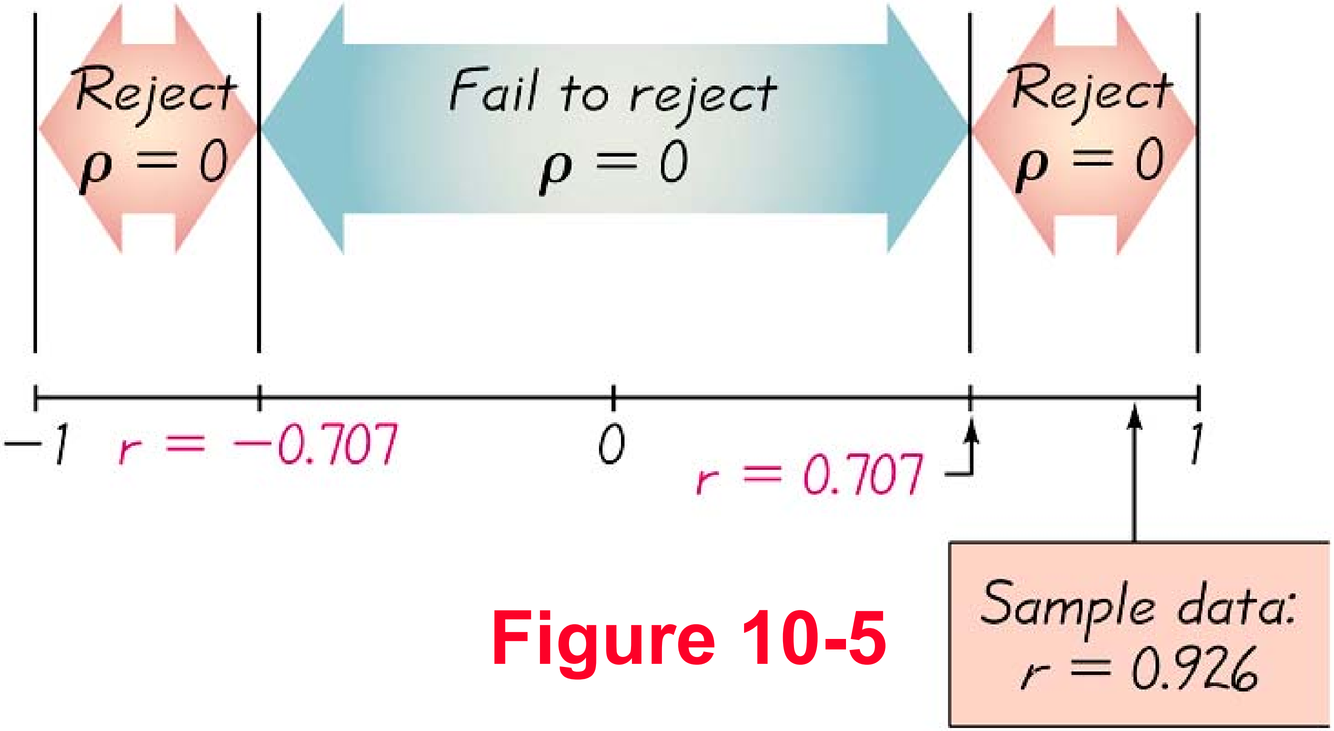


Figure 10-5

Example: Old Faithful

Using the duration/interval data in Table 10-1, test the claim that there is a linear correlation between the duration of an eruption and the time interval after that eruption.

**Using either of the two methods, we find that the absolute value of the test statistic does exceed the critical value (Method 1: 6.008 > 2.447.
Method 2: 0.926 > 0.707); that is, the test statistic falls in the critical region.**

We therefore reject the null hypothesis. There is sufficient evidence to support the claim of a linear correlation between the duration times of eruptions and the time intervals after the eruptions.

Justification for *r* Formula

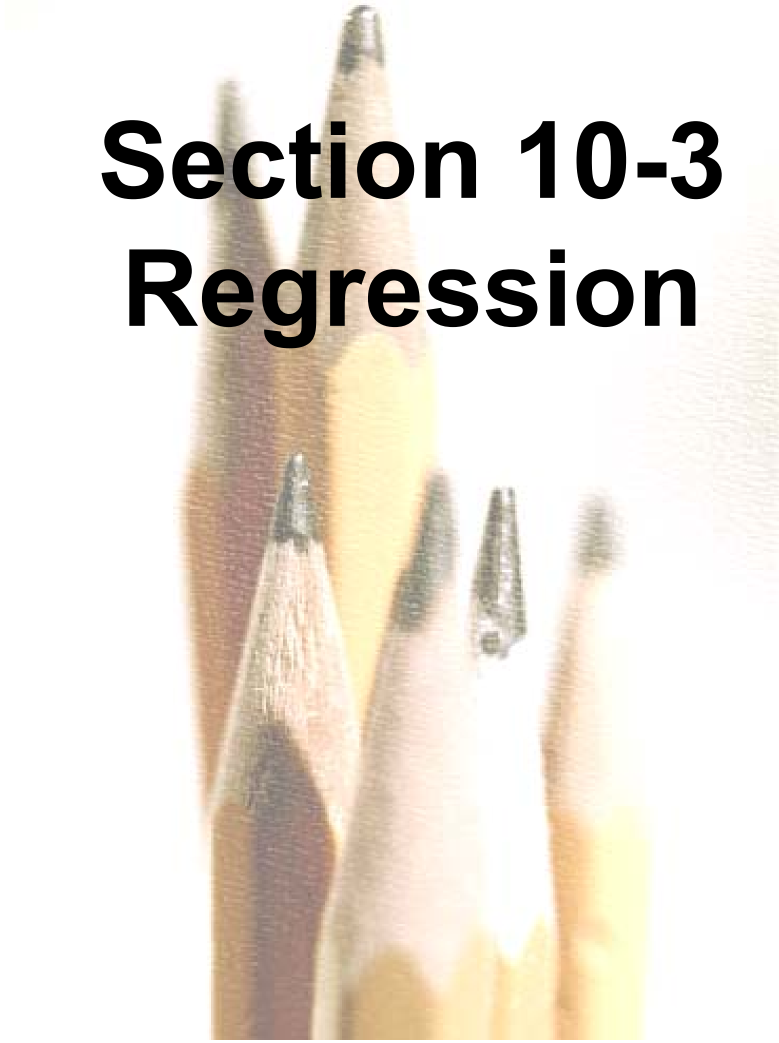
The text presents a detailed rationale for the use of Formula 10-1. The student should study it carefully.

Recap

In this section, we have discussed:

- **Correlation.**
- **The linear correlation coefficient *r*.**
- **Requirements, notation and formula for *r*.**
- **Interpreting *r*.**
- **Formal hypothesis testing (two methods).**
- **Rationale for calculating *r*.**

Section 10-3
Regression



Created by Erin Hodgess, Houston, Texas
Revised to accompany 10th Edition, Tom Wegleitner, Centreville, VA

Key Concept

The key concept of this section is to describe the relationship between two variables by finding the graph and the equation of the straight line that best represents the relationship.

The straight line is called a regression line and its equation is called the regression equation.

Part 1: Basic Concepts of Regression

Regression

The regression equation expresses a relationship between x (called the **independent variable**, **predictor variable** or **explanatory variable**), and $\hat{y}$ (called the **dependent variable** or **response variable**).

The typical equation of a straight line $y = mx + b$ is expressed in the form $\hat{y} = b_0 + b_1x$, where b_0 is the y-intercept and b_1 is the slope.

Requirements

1. The sample of paired (x, y) data is a random sample of quantitative data.

2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern.

3. Any outliers must be removed if they are known to be errors. Consider the effects of any outliers that are not known errors.

Definitions

- **Regression Equation**

Given a collection of paired data, the regression equation

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

algebraically describes the relationship between the two variables.

- **Regression Line**

The graph of the regression equation is called the regression line (or line of best fit, or least squares line).

Notation for Regression Equation

	Population Parameter	Sample Statistic
y-intercept of regression equation	β_0	b_0
Slope of regression equation	β_1	b_1
Equation of the regression line	$y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

Formulas for b_0 and b_1

Formula 10-2 $$b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$ (slope)

Formula 10-3 $$b_0 = \bar{y} - b_1 \bar{x}$$ (*y*-intercept)

calculators or computers can compute these values

Special Property

The regression line fits the sample points best.

Rounding the y-intercept b_0 and the Slope b_1

- Round to three significant digits.
- If you use the formulas 10-2 and 10-3, try not to round intermediate values.

Calculating the Regression Equation

Data

x	3	1	3	5
y	5	8	6	4

In Section 10-2, we used these values to find that the linear correlation coefficient of $r = –0.956$. Use this sample to find the regression equation.

Calculating the Regression Equation - cont

Data

x	3	1	3	5
y	5	8	6	4

$n = 4$

$\Sigma x = 12$

$\Sigma y = 23$

$\Sigma x^2 = 44$

$\Sigma y^2 = 141$

$\Sigma xy = 61$

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

$$b_1 = \frac{4(61) - (12)(23)}{4(44) - (12)^2}$$

$$b_1 = \frac{-32}{32} = -1$$

Calculating the Regression Equation - cont

Data

x	3	1	3	5
y	5	8	6	4

$n = 4$

$\Sigma x = 12$

$\Sigma y = 23$

$\Sigma x^2 = 44$

$\Sigma y^2 = 141$

$\Sigma xy = 61$

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

$$5.75 - (-1)(3) = 8.75$$

Calculating the Regression Equation - cont

Data

x	3	1	3	5
y	5	8	6	4

$n = 4$
$\Sigma x = 12$
$\Sigma y = 23$
$\Sigma x^2 = 44$
$\Sigma y^2 = 141$
$\Sigma xy = 61$

The estimated equation of the regression line is:

$$\hat{y} = 8.75 - 1x$$

Example: Old Faithful

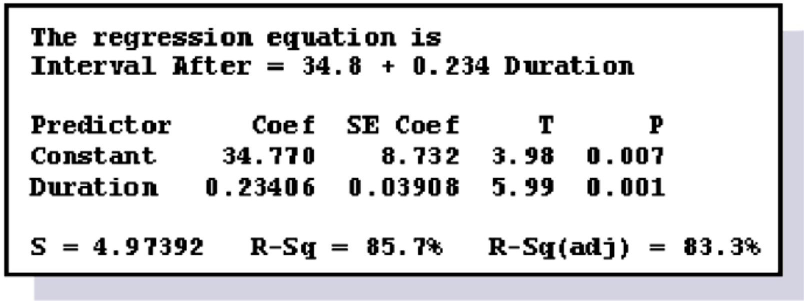
Given the sample data in Table 10-1, find the regression equation.

Using the same procedure as in the previous example, we find that b_1 = 0.234 and b_0 = 34.8. Hence, the estimated regression equation is:

$$\hat{y} = 34.8 + 0.234x$$

Example: Old Faithful - cont

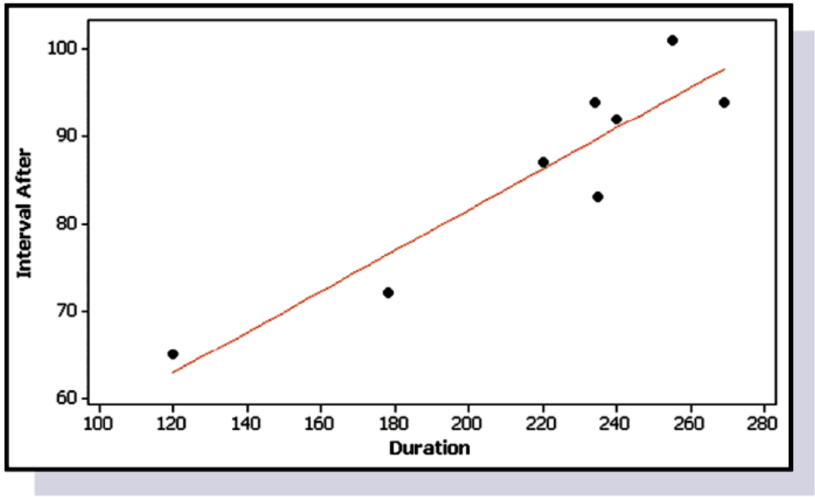
Given the sample data in Table 10-1, find the regression equation.



```
The regression equation is
Interval After = 34.8 + 0.234 Duration

Predictor      Coef   SE Coef      T       P
Constant     34.770     8.732   3.98   0.007
Duration    0.23406   0.03908   5.99   0.001

S = 4.97392   R-Sq = 85.7%    R-Sq(adj) = 83.3%
```



Part 2: Beyond the Basics of Regression

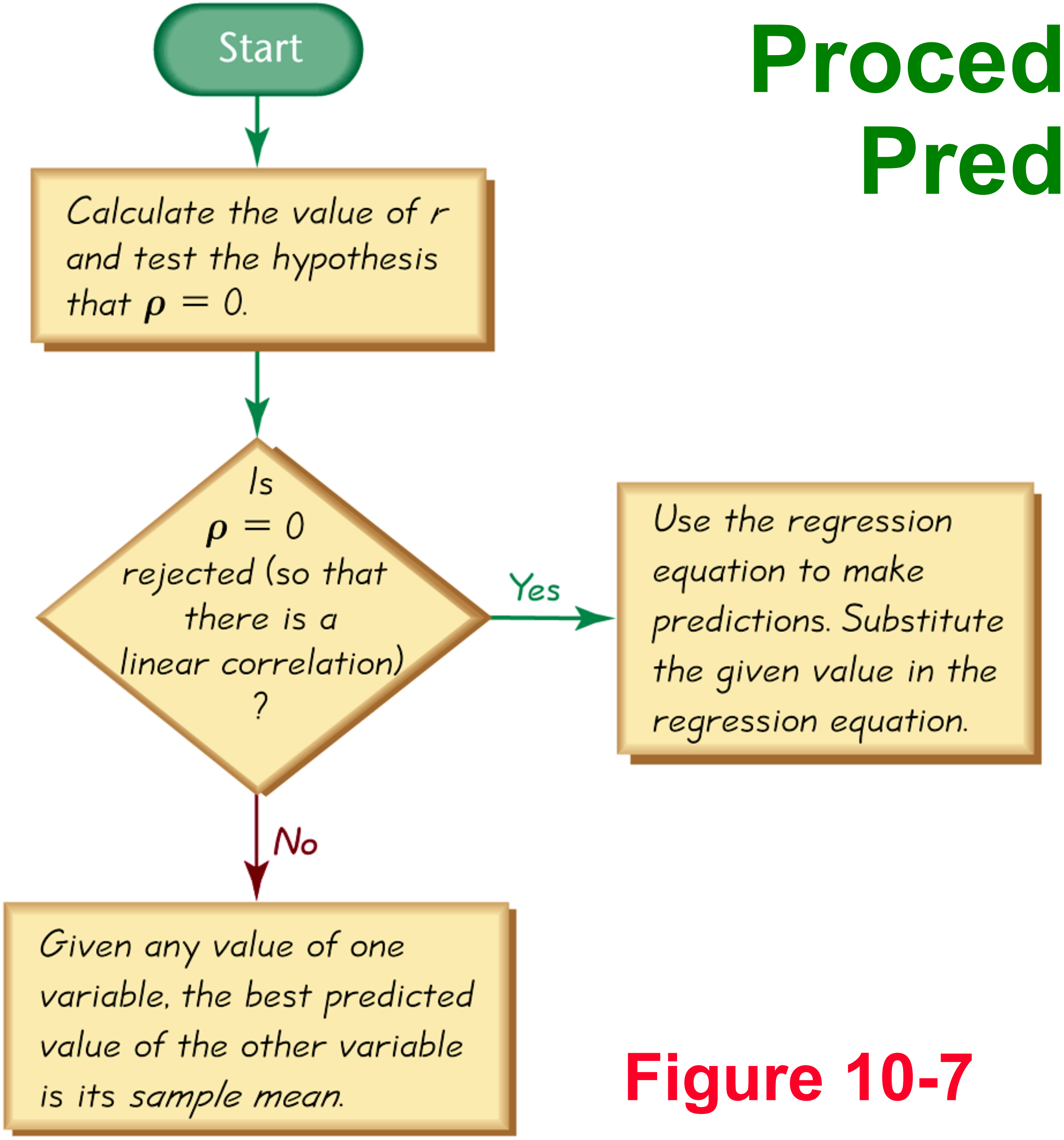
Predictions

In predicting a value of y based on some given value of x ...

1. If there is **not** a linear correlation, the best predicted y-value is $\bar{y}$.

2. If there is a linear correlation, the best predicted y-value is found by substituting the x-value into the regression equation.

Procedure for Predicting



Start

Calculate the value of r and test the hypothesis that $\rho = 0$.

Is $\rho = 0$ rejected (so that there is a linear correlation)?

Yes

Use the regression equation to make predictions. Substitute the given value in the regression equation.

No

Given any value of one variable, the best predicted value of the other variable is its sample mean.

Figure 10-7

Example: Old Faithful

Given the sample data in Table 10-1, we found that the regression equation is $\hat{y}$ = 34.8 + 0.234x. Assuming that the current eruption has a duration of x = 180 sec, find the best predicted value of y, the time interval after this eruption.

Example: Old Faithful

Given the sample data in Table 10-1, we found that the regression equation is $\hat{y}$ = 34.8 + 0.234x. Assuming that the current eruption has a duration of x = 180 sec, find the best predicted value of y, the time interval after this eruption.

We must consider whether there is a linear correlation that justifies the use of that equation. We do have a significant linear correlation (with r = 0.926).

Example: Old Faithful - cont

Given the sample data in Table 10-1, we found that the regression equation is $\hat{y} = 34.8 + 0.234x$. Assuming that the current eruption has a duration of $x = 180$ sec, find the best predicted value of y, the time interval after this eruption.

$$\hat{y} = 34.8 + 0.234x$$

$$34.8 + 0.234(180) = 76.9 \text{ min}$$

The predicted time interval is 76.9 min.

Guidelines for Using The Regression Equation

1. If there is no linear correlation, don’t use the regression equation to make predictions.
2. When using the regression equation for predictions, stay within the scope of the available sample data.
3. A regression equation based on old data is not necessarily valid now.
4. Don’t make predictions about a population that is different from the population from which the sample data were drawn.

Definitions

- ## Marginal Change

  The **marginal change** is the amount that a variable changes when the other variable changes by exactly one unit.

- ## Outlier

  An **outlier** is a point lying far away from the other data points.

- ## Influential Point

  An **influential point** strongly affects the graph of the regression line.

Definition

Residual

The residual for a sample of paired (x, y) data, is the difference (y - $\hat{y}$) between an observed sample y-value and the value of y, which is the value of y that is predicted by using the regression equation.

residual = observed y – predicted $y = y - \hat{y}$

Definitions

- **Least-Squares Property**

  A straight line has the **least-squares property** if the sum of the squares of the residuals is the smallest sum possible.

- **Residual Plot**

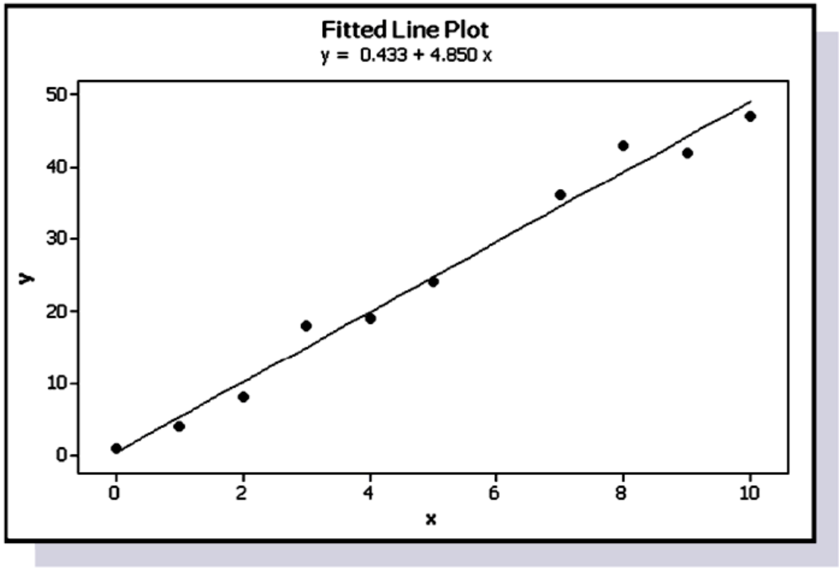
  A scatterplot of the (x, y) values after each of the y-coordinate values have been replaced by the residual value $y - \hat{y}$. That is, a **residual plot** is a graph of the points $(x, y - \hat{y})$

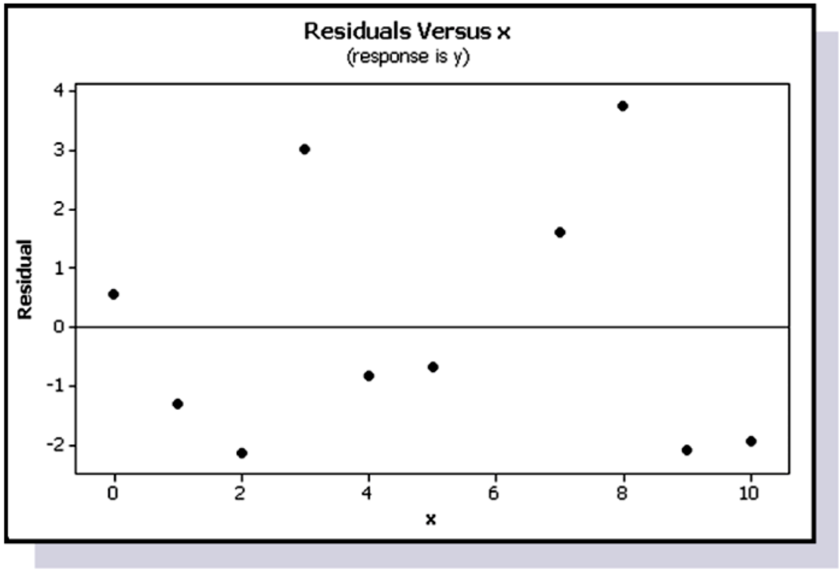
Residual Plot Analysis

If a residual plot does not reveal any pattern, the regression equation is a good representation of the association between the two variables.

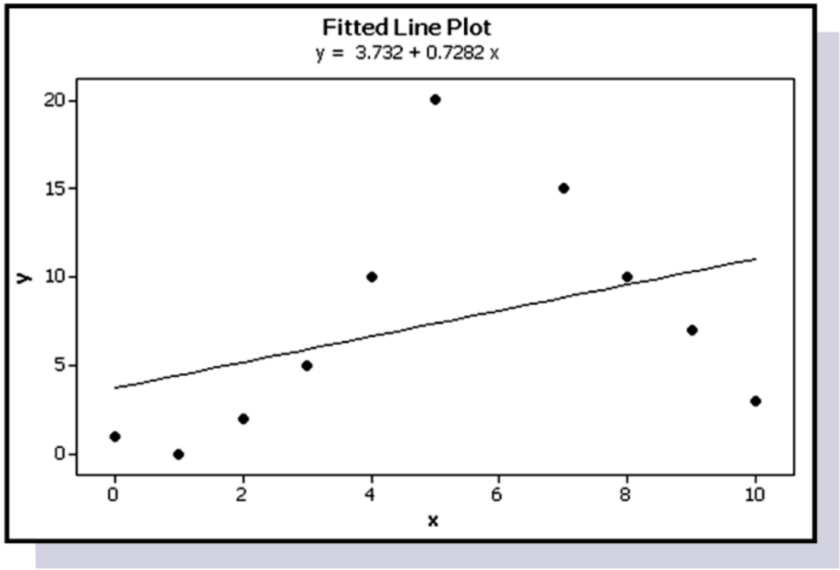
If a residual plot reveals some systematic pattern, the regression equation is not a good representation of the association between the two variables.

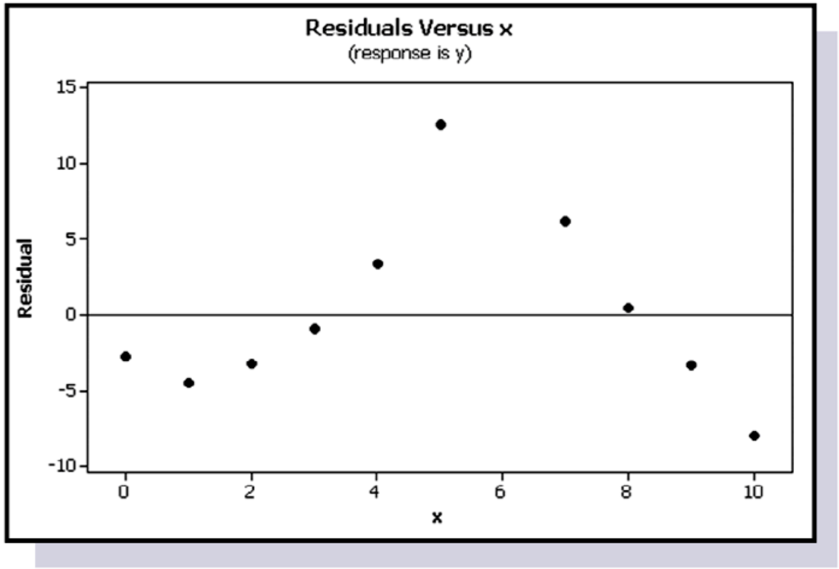
Residual Plots



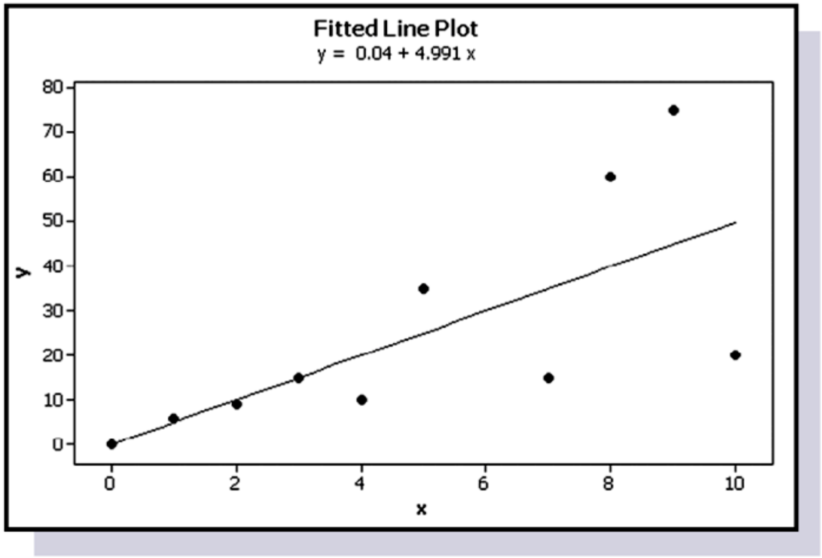


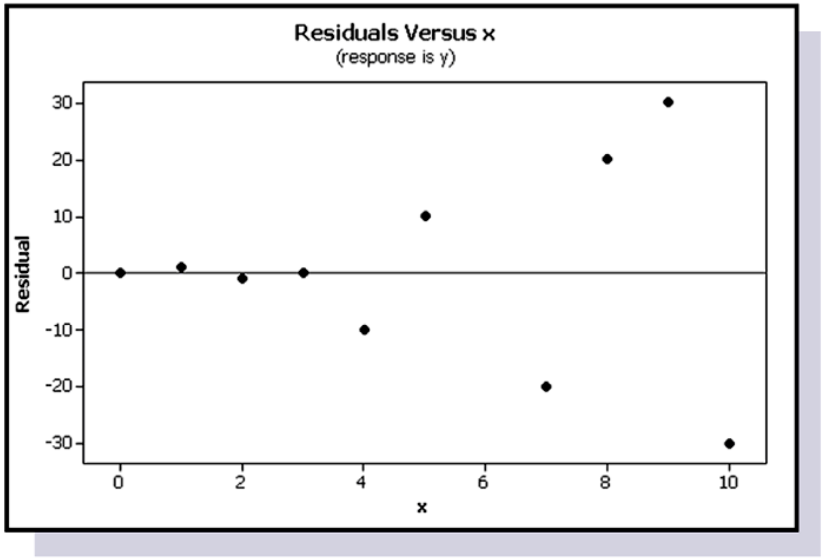
Residual Plots





Residual Plots

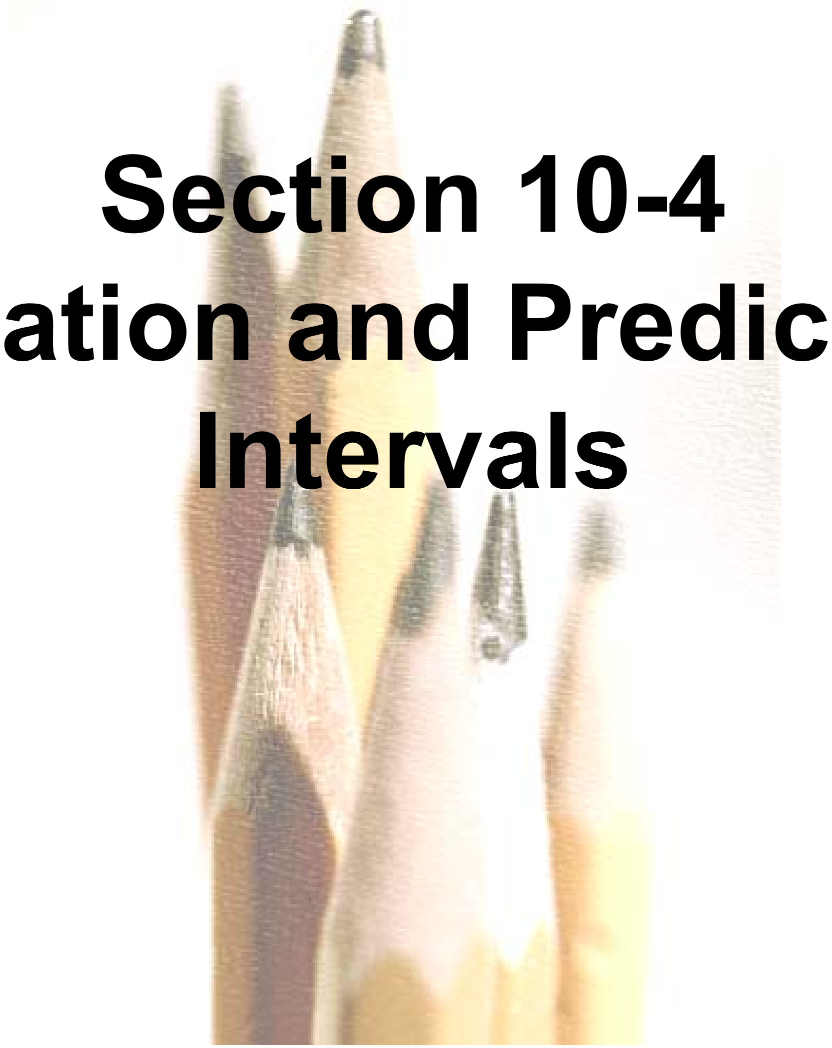




Recap

In this section we have discussed:

- The basic concepts of regression.
- Rounding rules.
- Using the regression equation for predictions.
- Interpreting he regression equation.
- Outliers
- Residuals and least-squares.
- Residual plots.



Section 10-4
Variation and Prediction Intervals

Created by Erin Hodgess, Houston, Texas
Revised to accompany 10th Edition, Tom Wegleitner, Centreville, VA

PEARSON
Addison Wesley

Key Concept

In this section we proceed to consider a method for constructing a prediction interval, which is an interval estimate of a predicted value of *y*.

Definition

Total Deviation

The total deviation of (x, y) is the vertical distance $y - \bar{y}$, which is the distance between the point (x, y) and the horizontal line passing through the sample mean $\bar{y}$.

Definition

Explained Deviation

The explained deviation is the vertical distance $\hat{y}$ - $\bar{y}$, which is the distance between the predicted y-value and the horizontal line passing through the sample mean $\bar{y}$.

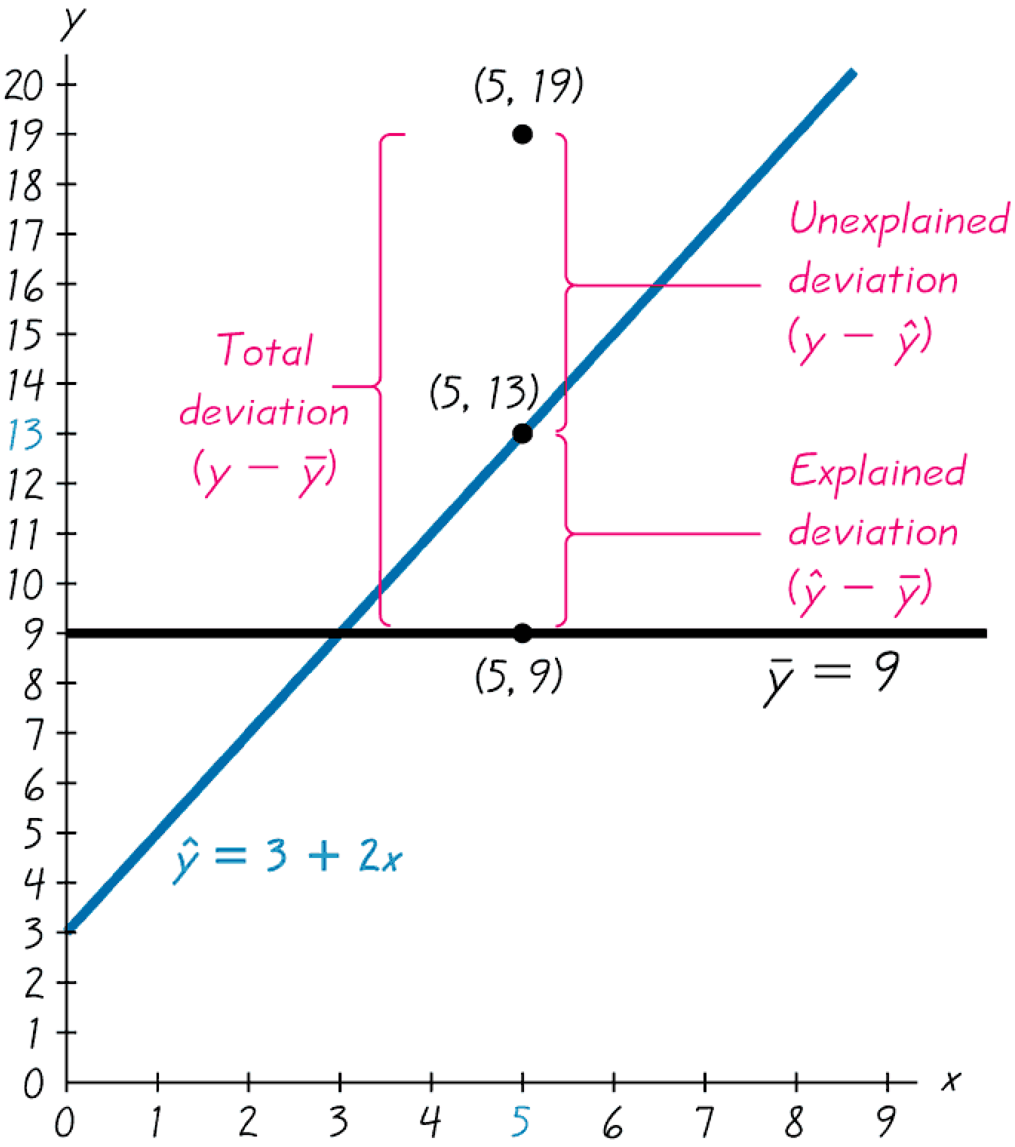
Definition

Unexplained Deviation

The unexplained deviation is the vertical distance $y - \hat{y}$, which is the vertical distance between the point (x, y) and the regression line. (The distance $y - \hat{y}$ is also called a residual, as defined in Section 10-3.)

Unexplained, Explained, and Total Deviation

Figure 10-9



Relationships

(total deviation) = (explained deviation) + (unexplained deviation)

$$(y - \bar{y}) \quad = \quad (\hat{y} - \bar{y}) \quad + \quad (y - \hat{y})$$

(total variation) = (explained variation) + (unexplained variation)

$$\Sigma\,(y - \bar{y})^2 = \Sigma\,(\hat{y} - \bar{y})^2 + \Sigma\,(y - \hat{y})^2$$

Formula 10-4

Definition

Coefficient of determination

is the amount of the variation in *y* that is explained by the regression line.

$$r^2 = \frac{\text{explained variation.}}{\text{total variation}}$$

The value of r^2 is the proportion of the variation in *y* that is explained by the linear relationship between *x* and *y*.

Definition

Standard Error of Estimate

The standard error of estimate, denoted by s_e is a measure of the differences (or distances) between the observed sample y-values and the predicted values $\hat{y}$ that are obtained using the regression equation.

Standard Error of Estimate

$$s_e = \sqrt{\frac{\sum (y-\hat{y})^2}{n-2}}$$

or

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

Formula 10-5

Example: Old Faithful

Given the sample data in Table 10-1, find the standard error of estimate s_e for the duration/interval data.

$n = 8$

$\Sigma y^2 = 60{,}204$

$\Sigma y = 688$

$\Sigma xy = 154{,}378$

$b_0 = 34.7698041$

$b_1 = 0.2340614319$

$$s_e = \sqrt{\frac{\Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy}{n - 2}}$$

$$s_e = \sqrt{\frac{60{,}204 - (34.7698041)(688) - (0.2340614319)(154{,}378)}{8 - 2}}$$

Example: Old Faithful - cont

Given the sample data in Table 10-1, find the standard error of estimate s_e for the duration/interval data.

$n = 8$

$\Sigma y^2 = 60{,}204$

$\Sigma y = 688$

$\Sigma xy = 154{,}378$

$b_0 = 34.7698041$

$b_1 = 0.2340614319$

$s_e = 4.973916052 = 4.97$ (rounded)

Prediction Interval for an Individual y

$$\hat{y} - E < y < \hat{y} + E$$

where

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$

x_0 represents the given value of x

$t_{\alpha/2}$ has $n - 2$ degrees of freedom

Example: Old Faithful

For the paired duration/interval after eruption times in Table 10-1, we have found that for a duration of 180 sec, the best predicted time interval after the eruption is 76.9 min. Construct a 95% prediction interval for the time interval after the eruption, given that the duration of the eruption is 180 sec (so that x = 180).

$$E = t_{\alpha/2}\, s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \overline{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$

$$E = (2.447)(4.973916052) \sqrt{1 + \frac{1}{8} + \frac{8(180 - 218.875)^2}{8(399{,}451) - (1751)^2}}$$

E = 13.4 **(rounded)**

Example: Old Faithful - cont

For the paired duration/interval after eruption times in Table 10-1, we have found that for a duration of 180 sec, the best predicted time interval after the eruption is 76.9 min. Construct a 95% prediction interval for the time interval after the eruption, given that the duration of the eruption is 180 sec (so that *x* = 180).

$$\hat{y} - E < y < \hat{y} + E$$

$$76.9 - 13.4 < y < 76.9 + 13.4$$

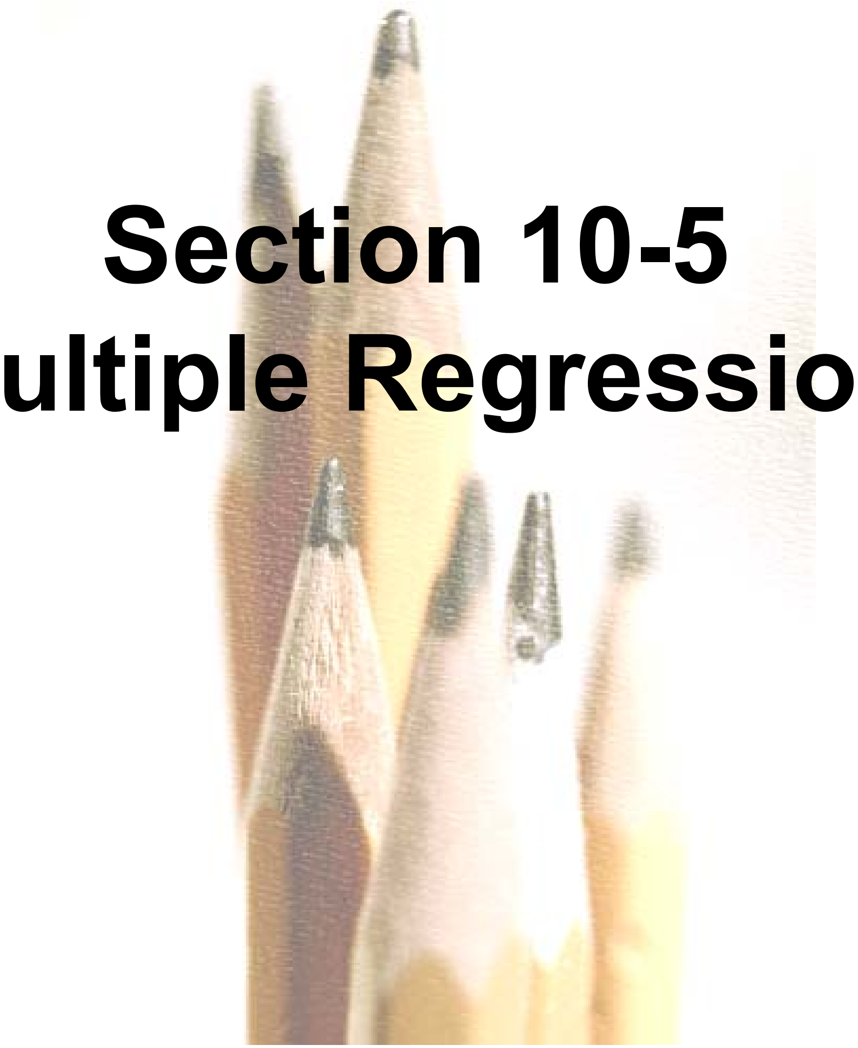
$$63.5 < y < 90.3$$

Recap

In this section we have discussed:

- Explained and unexplained variation.
- Coefficient of determination.
- Standard error estimate.
- Prediction intervals.

Section 10-5
Multiple Regression



Created by Erin Hodgess, Houston, Texas
Revised to accompany 10th Edition, Tom Wegleitner, Centreville, VA

Key Concept

This section presents a method for analyzing a linear relationship involving more than two variables.

We focus on three key elements:

1. The multiple regression equation.
2. The values of the adjusted R^2.
3. The P-value.

Definition

Multiple Regression Equation

A linear relationship between a response variable y and two or more predictor variables ($x_1, x_2, x_3 . . . , x_k$)

The general form of the multiple regression equation is

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + . . . + b_kx_k.$$

Notation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k$$

(General form of the estimated multiple regression equation)

n = sample size

k = number of predictor variables

$\hat{y}$ = predicted value of y

$x_1, x_2, x_3 \ldots, x_k$ are the predictor variables

Notation - cont

$ß_0$ = the y-intercept, or the value of y when all of the predictor variables are 0

b_0 = estimate of $ß_0$ based on the sample data

$ß_1, ß_2, ß_3 . . . , ß_k$ are the coefficients of the independent variables $x_1, x_2, x_3 . . . , x_k$

$b_1, b_2, b_3 . . . , b_k$ are the sample estimates of the coefficients $ß_1, ß_2, ß_3 . . . , ß_k$

Technology

Use a statistical software package such as

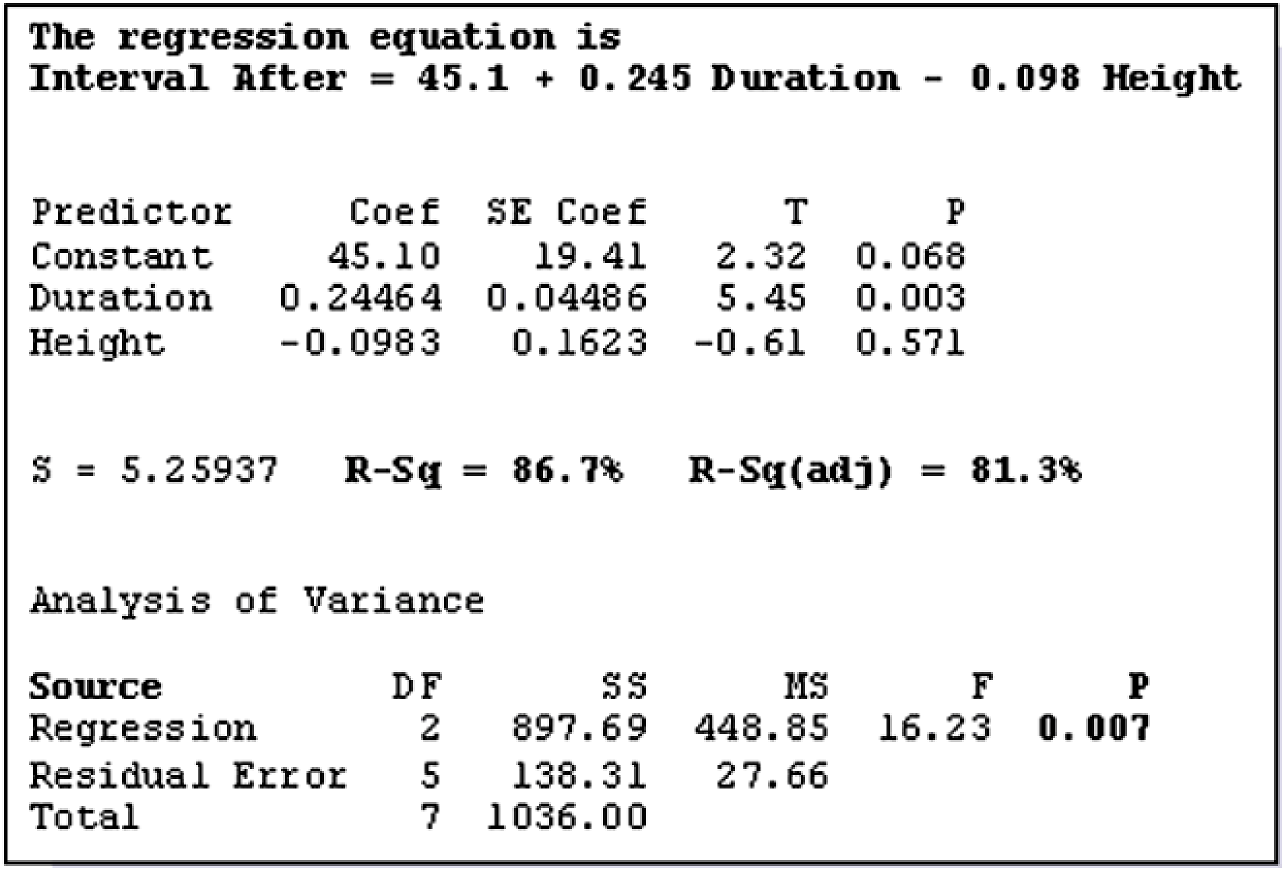
- **STATDISK**
- **Minitab**
- **Excel**

Example: Old Faithful

Using the sample data in Table 10-1 included with the Chapter Problem, find the multiple regression in which the response (y) variable is the time interval after an eruption and the predictor (x) variables are the duration time of the eruption and height of the eruption.

Example: Old Faithful - cont

The Minitab results are shown here:



```
The regression equation is
Interval After = 45.1 + 0.245 Duration - 0.098 Height


Predictor      Coef   SE Coef      T       P
Constant      45.10     19.41   2.32   0.068
Duration    0.24464   0.04486   5.45   0.003
Height      -0.0983    0.1623  -0.61   0.571


S = 5.25937   R-Sq = 86.7%   R-Sq(adj) = 81.3%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2   897.69  448.85  16.23  0.007
Residual Error   5   138.31   27.66
Total            7  1036.00
```

Example: Old Faithful - cont

The Minitab results give:

Interval After = 45.1 + 0.245 Duration – 0.098 Height

We can write this equation as:

$$y = 45.1 + 0.245\, x_1 - 0.098 x_2$$

Definition

- **Multiple coefficient of determination**

  The multiple coefficient of determination R^2 is a measure of how well the multiple regression equation fits the sample data.

- **Adjusted coefficient of determination**

  The adjusted coefficient of determination is the multiple coefficient of determination R^2 modified to account for the number of variables and the sample size.

Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

Formula 10-6

where n = sample size

k = number of independent (x) variables

Finding the Best Multiple Regression Equation

1. **Use common sense and practical considerations to include or exclude variables.**

2. **Consider the *P*-value.** **Select an equation having overall significance, as determined by the *P*-value found in the computer display.**

3. **Consider equations with high values of adjusted R^2 and try to include only a few variables.**

   - **If an additional predictor variable is included, the value of adjusted R^2 does not increase by a substantial amount.**
   - **For a given number of predictor (x) variables, select the equation with the largest value of adjusted R^2.**
   - **In weeding out predictor (x) variables that don't have much of an effect on the response (y) variable, it might be helpful to find the linear correlation coefficient r for each of the paired variables being considered.**

Dummy Variables

Many applications involve a dichotomous variable which has only two possible discrete values (such as male/female, dead/alive, etc.). A common procedure is to represent the two possible discrete values by 0 and 1, where 0 represents “failure” and 1 represents success.

A dichotomous variable with the two values 0 and 1 is called a dummy variable.

Logistic Regression

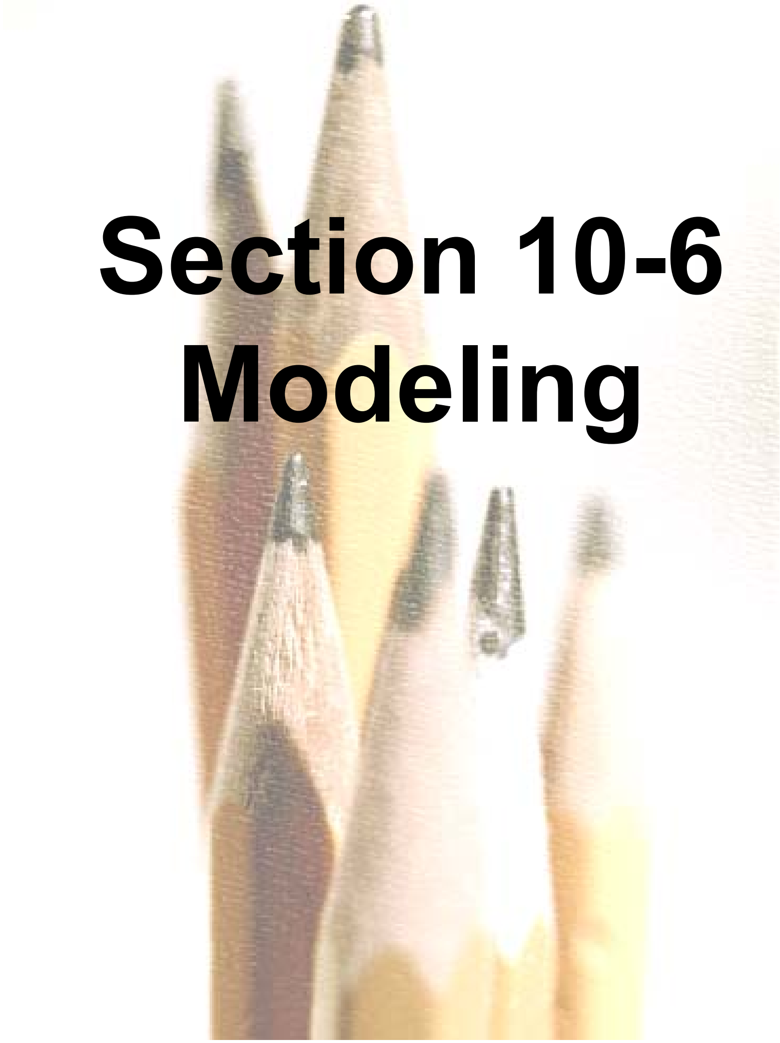
We can use the methods of this section if the dummy variable is the predictor variable. If the dummy variable is the response variable we need to use a method known as logistic regression.

A discussion of logistic regression is given in the text.

Recap

In this section we have discussed:

- The multiple regression equation.
- Adjusted R^2.
- Finding the best multiple regression equation.
- Dummy variables and logistic regression.



Section 10-6
Modeling

Created by Erin Hodgess, Houston, Texas
Revised to accompany 10th Edition, Tom Wegleitner, Centreville, VA

PEARSON
Addison Wesley

Key Concept

This section introduces some basic concepts of developing a **mathematical model**, which is a function that “fits” or describes real-world data.

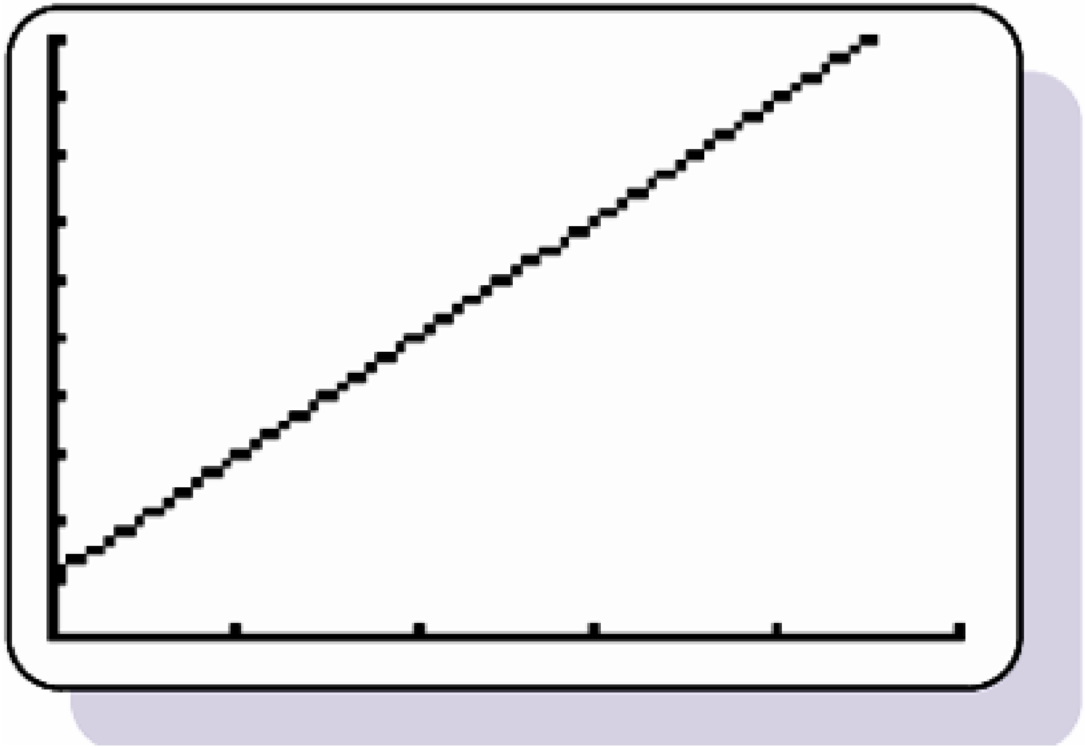
Unlike Section 10-3, we will not be restricted to a model that must be linear.

TI-83/84 Plus Generic Models

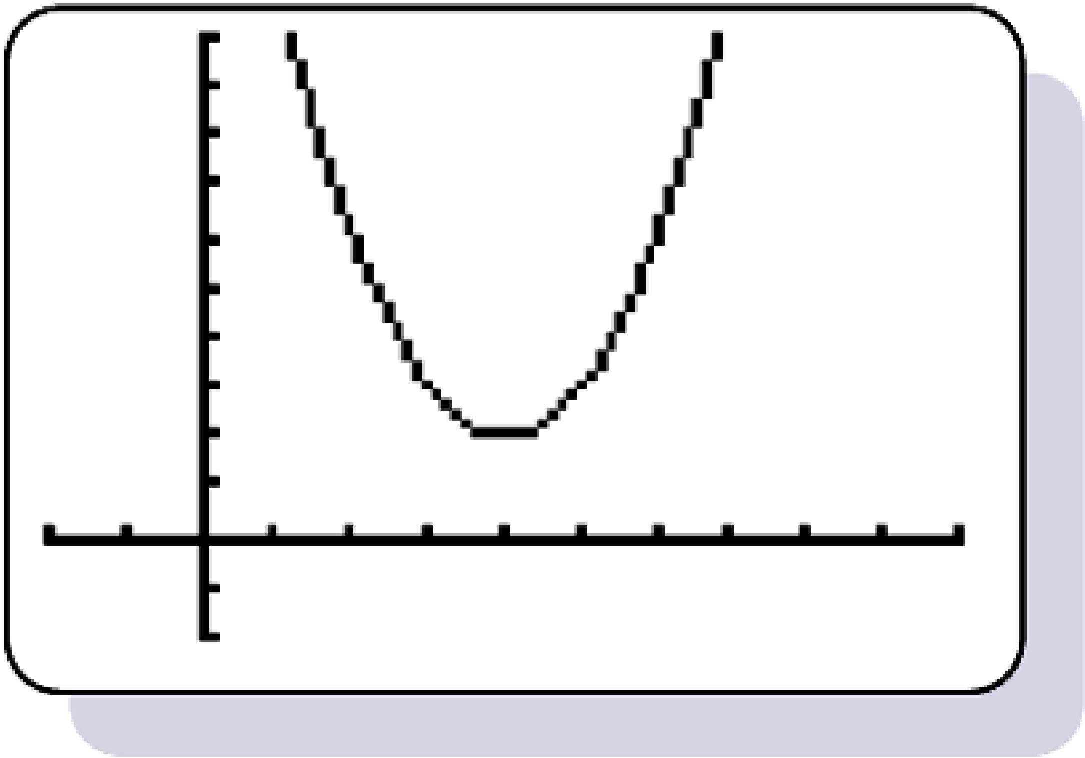
Linear:	$y = a + bx$
Quadratic:	$y = ax^2 + bx + c$
Logarithmic:	$y = a + b \ln x$
Exponential:	$y = ab^x$
Power:	$y = ax^b$

The slides that follow illustrate the graphs of some common models displayed on a TI-83/84 Plus Calculator

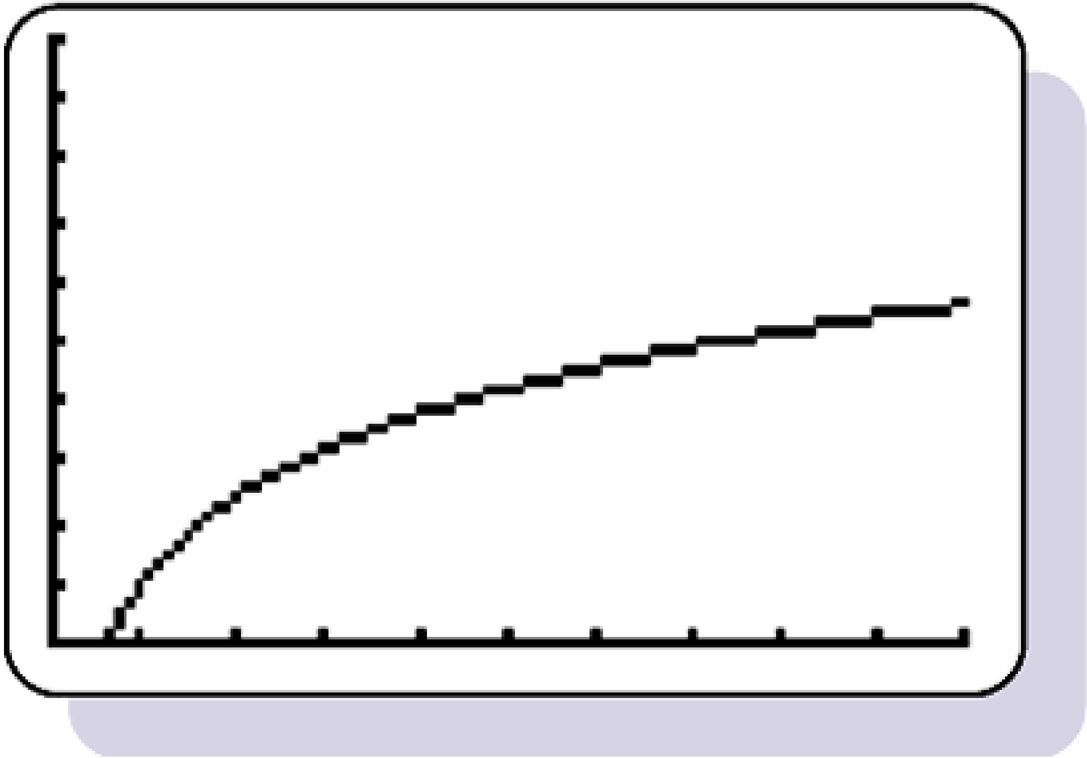
Linear: $y = 1 + 2x$



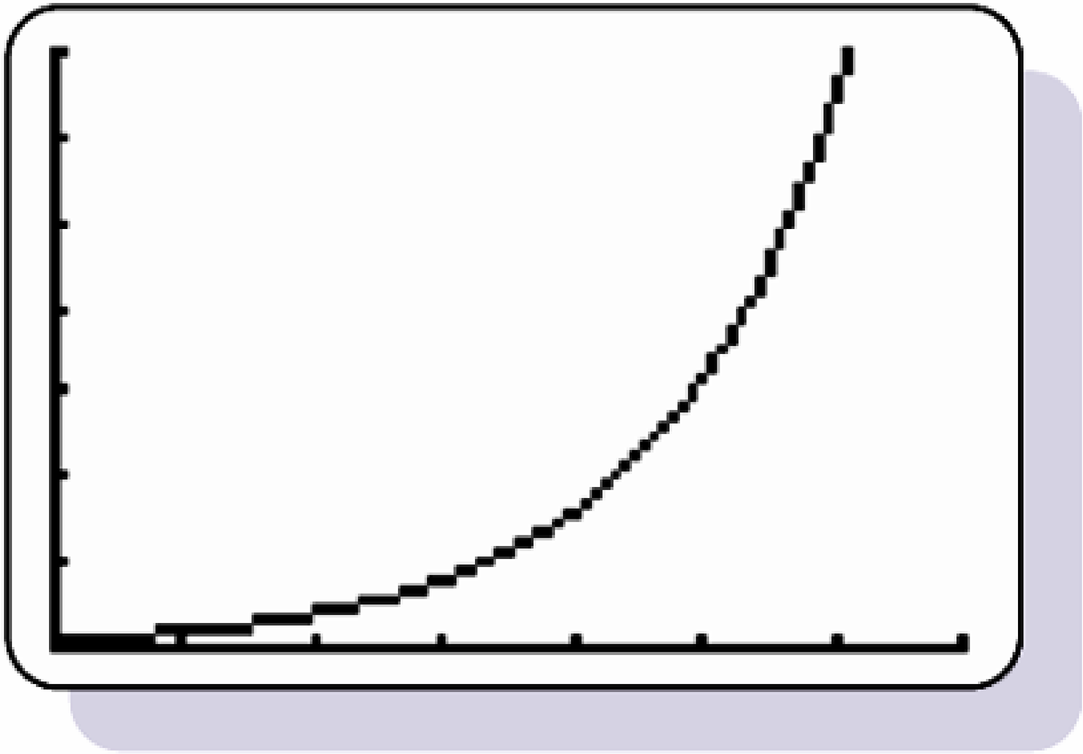
Quadratic: $y = 2x^2 - 8x + 9$



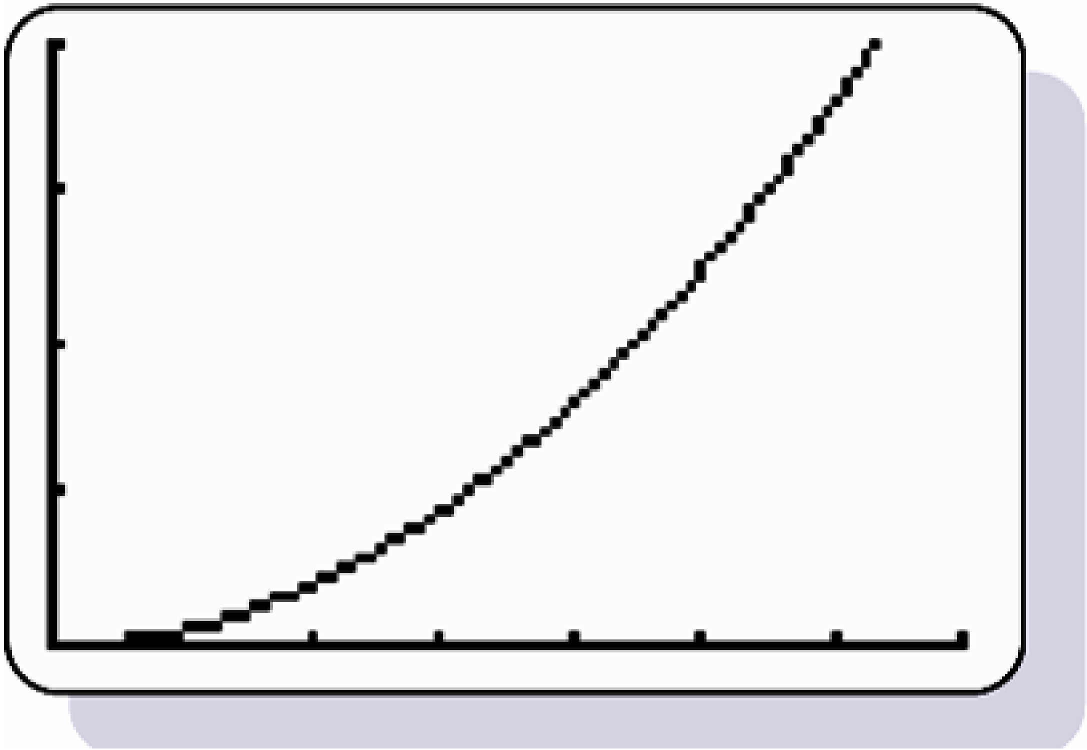
Logarithmic: $y = 1 + 2\ln x$



Exponential: $y = 2^x$



Power: $y = x^2$



Development of a Good Mathematical Model

- **Look for a Pattern in the Graph:** Examine the graph of the plotted points and compare the basic pattern to the known generic graphs of a linear function.
- **Find and Compare Values of R^2:** Select functions that result in larger values of R^2, because such larger values correspond to functions that better fit the observed points.
- **Think:** Use common sense. Don’t use a model that leads to predicted values known to be totally unrealistic.

Recap

In this section we have discussed:

- **The concept of mathematical modeling.**
- **Graphs from a TI-83/84 Plus calculator.**
- **Rules for developing a good mathematical model.**