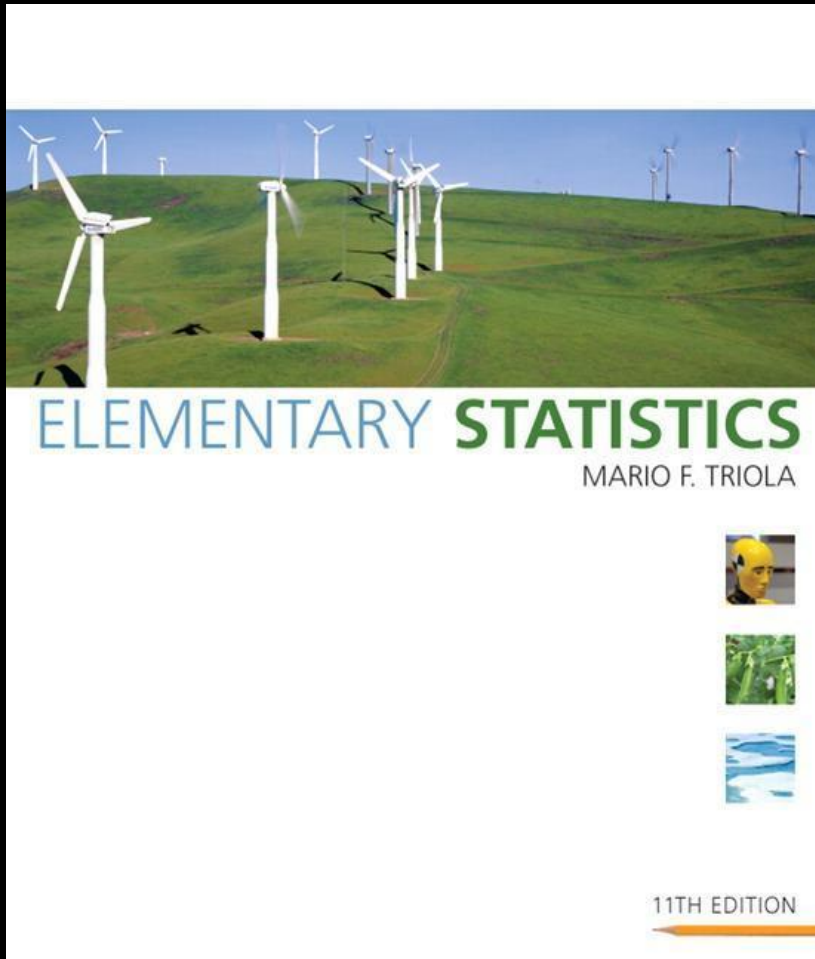


# Lecture Slides



## *Elementary Statistics* Eleventh Edition

and the Triola Statistics Series

by Mario F. Triola



# Chapter 10

## Correlation and Regression

**10-1 Review and Preview**

**10-2 Correlation**

**10-3 Regression**

**10-4 Variation and Prediction Intervals**

**10-5 Multiple Regression**

**10-6 Modeling**



**Section 10-1**  
**Review and Preview**

# Review

**In Chapter 9 we presented methods for making inferences from two samples. In Section 9-4 we considered two dependent samples, with each value of one sample somehow paired with a value from the other sample. In Section 9-4 we considered the differences between the paired values, and we illustrated the use of hypothesis tests for claims about the population of differences. We also illustrated the construction of confidence interval estimates of the mean of all such differences. In this chapter we again consider paired sample data, but the objective is fundamentally different from that of Section 9-4.**

# Preview

**In this chapter we introduce methods for determining whether a correlation, or association, between two variables exists and whether the correlation is linear. For linear correlations, we can identify an equation that best fits the data and we can use that equation to predict the value of one variable given the value of the other variable. In this chapter, we also present methods for analyzing differences between predicted values and actual values.**

# Preview

**In addition, we consider methods for identifying linear equations for correlations among three or more variables. We conclude the chapter with some basic methods for developing a mathematical model that can be used to describe nonlinear correlations between two variables.**



# **Section 10-2 Correlation**

# Key Concept

In part 1 of 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.

Using paired sample data (sometimes called bivariate data), we find the value of  $r$  (usually using technology), then we use that value to conclude that there is (or is not) a linear correlation between the two variables.



# Key Concept

**In this section we consider only linear relationships, which means that when graphed, the points approximate a straight-line pattern.**

**In Part 2, we discuss methods of hypothesis testing for correlation.**

# Part 1: Basic Concepts of Correlation

# Definition

**A **correlation** exists between two variables when the values of one are somehow associated with the values of the other in some way.**

# Definition

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

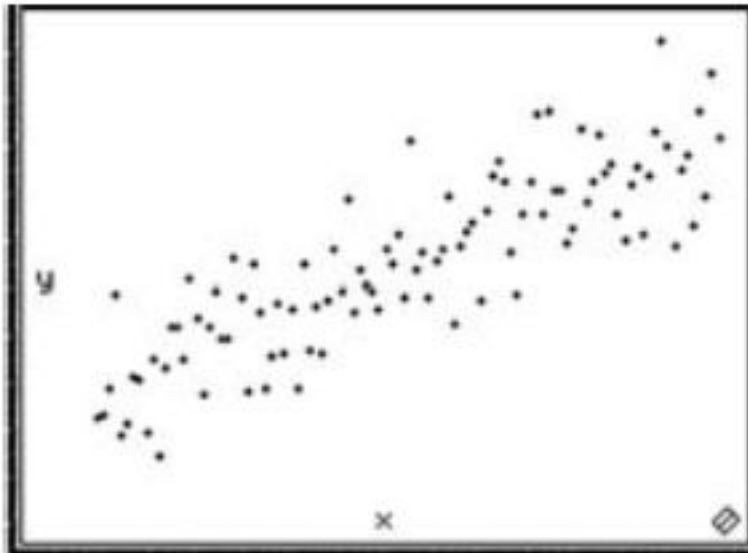
# Exploring the Data

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

**Figure 10-2 following shows scatterplots with different characteristics.**

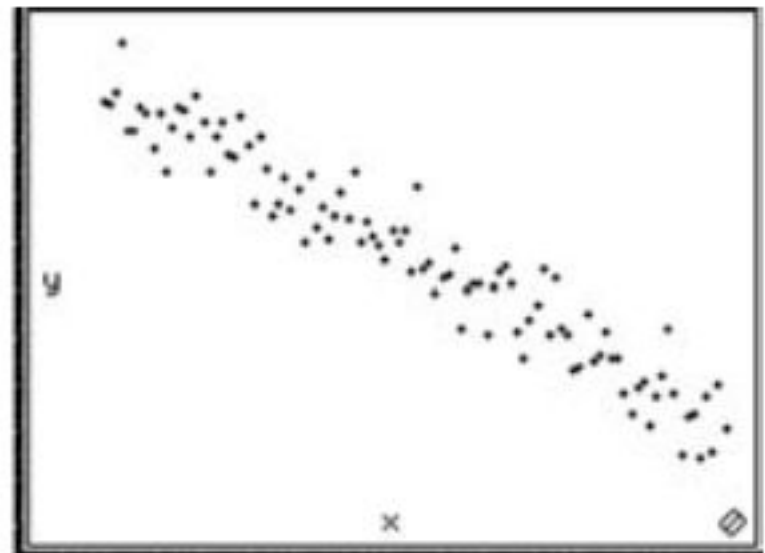
# Scatterplots of Paired Data

**ActivStats**



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

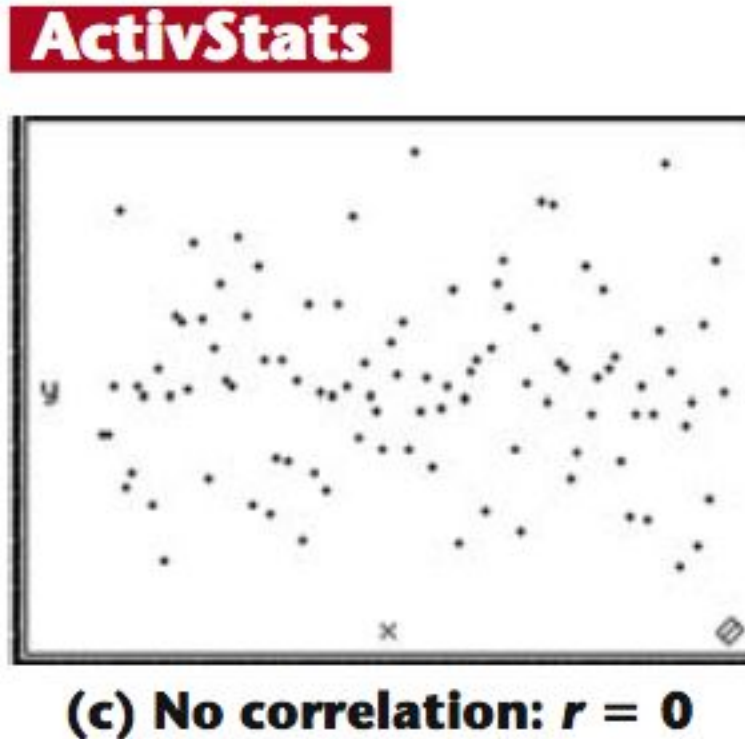
**ActivStats**



**(b) Negative correlation:**  
 $r = -0.965$

**Figure 10-2**

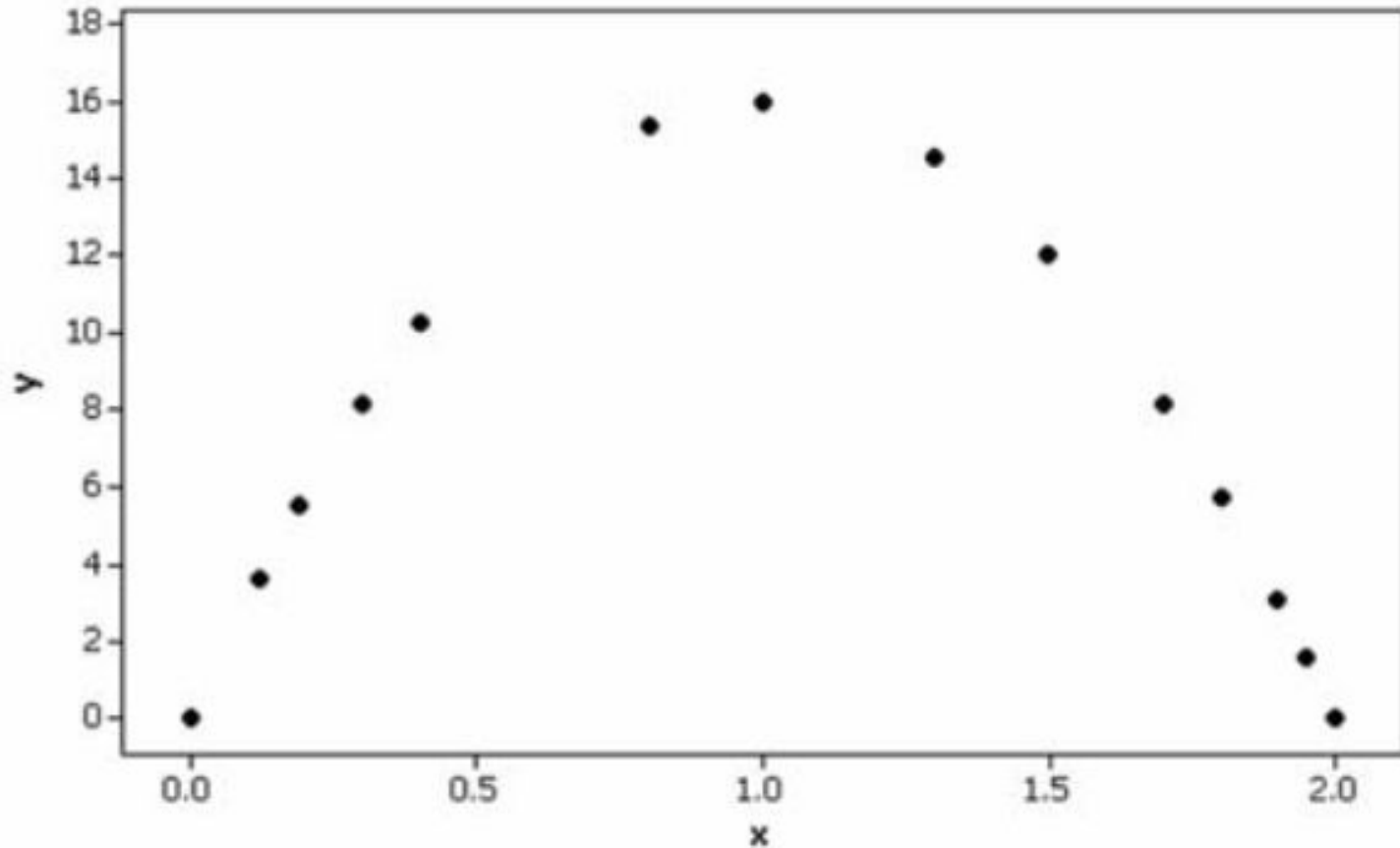
# Scatterplots of Paired Data



**Figure 10-2**

# Scatterplots of Paired Data

**Minitab**



**(d) Nonlinear relationship:  $r = -0.087$**

**Figure 10-2**



# Requirements

- 1. The sample of paired  $(x, y)$  data is a simple random sample of 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

$n$  = number of pairs of sample data

$\Sigma$  denotes the addition of the items indicated.

$\Sigma x$  denotes the sum of all  $x$ -values.

$\Sigma x^2$  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 then the total squared.

# Notation for the Linear Correlation Coefficient

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

$r$  = linear correlation coefficient for **sample** data.

$\rho$  = linear correlation coefficient for **population** data.

# Formula

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

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

**Formula 10-1**

**Computer software or calculators can compute  $r$**

# Interpreting $r$

**Using Table A-6:** If the absolute value of the computed value of  $r$ , denoted  $|r|$ , 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.

# Caution

**Know that the methods of this section apply to a *linear* correlation. If you conclude that there does not appear to be linear correlation, know that it is possible that there might be some other association that is not linear.**

# 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.

# Properties of the Linear Correlation Coefficient $r$

1.  $-1 \leq r \leq 1$
2. if all values of either variable are converted to a different scale, the value of  $r$  does not change.
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.
5.  $r$  is very sensitive to outliers, they can dramatically affect its value.

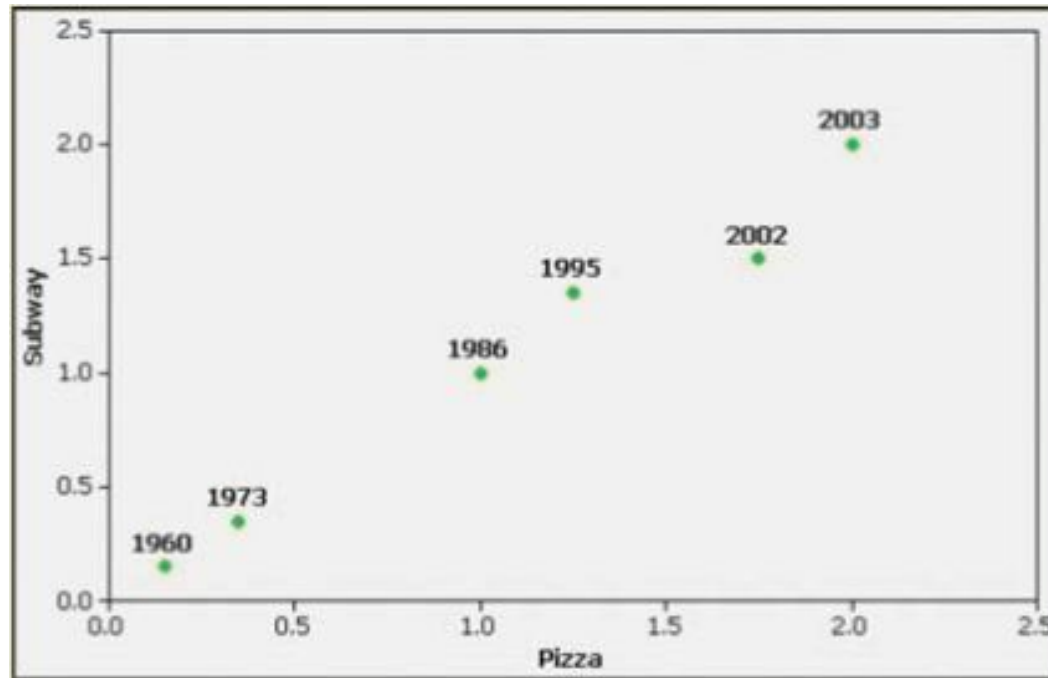


## Example:

**The paired pizza/subway fare costs from Table 10-1 are shown here in Table 10-2. Use computer software with these paired sample values to find the value of the linear correlation coefficient  $r$  for the paired sample data.**

**Requirements are satisfied: simple random sample of quantitative data; Minitab scatterplot approximates a straight line; scatterplot shows no outliers - see next slide**

# Example:



Using software or a calculator,  $r$  is automatically calculated:

## MINITAB

Correlations: Pizza, Subway

Pearson correlation of Pizza and Subway = 0.988  
P-Value = 0.000

# Interpreting the Linear Correlation Coefficient $r$

**We can base our interpretation and conclusion about correlation on a  $P$ -value obtained from computer software or a critical value from Table A-6.**

# Interpreting the Linear Correlation Coefficient $r$

**Using Computer Software to Interpret  $r$ :**

**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.**

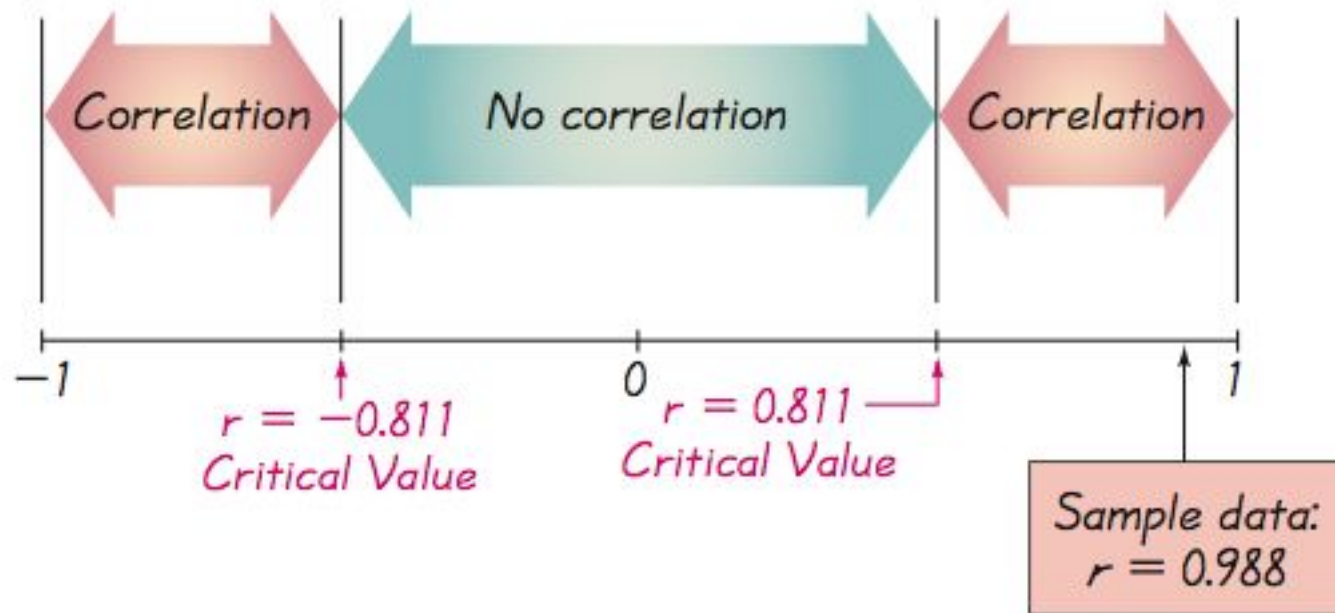
# Interpreting the Linear Correlation Coefficient $r$

**Using Table A-6 to Interpret  $r$ :**

**If  $|r|$  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.**

# Interpreting the Linear Correlation Coefficient $r$



**Critical Values from Table A-6 and the Computed Value of  $r$**

## Example:

**Using a 0.05 significance level, interpret the value of  $r = 0.117$  found using the 62 pairs of weights of discarded paper and glass listed in Data Set 22 in Appendix B. When the paired data are used with computer software, the  $P$ -value is found to be 0.364. Is there sufficient evidence to support a claim of a linear correlation between the weights of discarded paper and glass?**

## Example:

**Requirements are satisfied: simple random sample of quantitative data; scatterplot approximates a straight line; no outliers**

**Using Software to Interpret  $r$ :**

**The  $P$ -value obtained from software is 0.364. Because the  $P$ -value is not less than or equal to 0.05, we conclude that there is not sufficient evidence to support a claim of a linear correlation between weights of discarded paper and glass.**



## Example:

### Using Table A-6 to Interpret $r$ :

If we refer to Table A-6 with  $n = 62$  pairs of sample data, we obtain the critical value of 0.254 (approximately) for  $\alpha = 0.05$ . Because  $|0.117|$  does not exceed the value of 0.254 from Table A-6, we conclude that there is not sufficient evidence to support a claim of a linear correlation between weights of discarded paper and glass.

# 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:

Using the pizza subway fare costs in Table 10-2, we have found that the linear correlation coefficient is  $r = 0.988$ . What proportion of the variation in the subway fare can be explained by the variation in the costs of a slice of pizza?

**With  $r = 0.988$ , we get  $r^2 = 0.976$ .**

We conclude that 0.976 (or about 98%) of the variation in the cost of a subway fares can be explained by the linear relationship between the costs of pizza and subway fares. This implies that about 2% of the variation in costs of subway fares cannot be explained by the costs of pizza.

# 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.

# Caution

**Know that correlation does not imply causality.**

# Part 2: Formal Hypothesis Test

# Formal Hypothesis Test

**We wish to determine whether there is a significant linear correlation between two variables.**

# Hypothesis Test for Correlation Notation

$n$  = number of pairs of sample data

$r$  = linear correlation coefficient for a *sample* of paired data

$\rho$  = linear correlation coefficient for a *population* of paired data



# Hypothesis Test for Correlation Requirements

- 1. The sample of paired  $(x, y)$  data is a simple random sample of 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.**

# Hypothesis Test for Correlation Hypotheses

$H_0: \rho = 0$  (There is no linear correlation.)

$H_1: \rho \neq 0$  (There is a linear correlation.)

**Test Statistic:  $r$**

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

# Hypothesis Test for Correlation Conclusion

If  $|r| >$  critical value from Table A-6, reject  $H_0$  and conclude that there is sufficient evidence to support the claim of a linear correlation.

If  $|r| \leq$  critical value from Table A-6, fail to reject  $H_0$  and conclude that there is not sufficient evidence to support the claim of a linear correlation.

## Example:

Use the paired pizza subway fare data in Table 10-2 to test the claim that there is a linear correlation between the costs of a slice of pizza and the subway fares. Use a 0.05 significance level.

Requirements are satisfied as in the earlier example.

$H_0: \rho = 0$  (There is no linear correlation.)

$H_1: \rho \neq 0$  (There is a linear correlation.)

## Example:

The test statistic is  $r = 0.988$  (from an earlier Example). The critical value of  $r = 0.811$  is found in Table A-6 with  $n = 6$  and  $\alpha = 0.05$ . Because  $|0.988| > 0.811$ , we reject  $H_0: r = 0$ . (Rejecting “no linear correlation” indicates that there is a linear correlation.)

We conclude that there is sufficient evidence to support the claim of a linear correlation between costs of a slice of pizza and subway fares.

# Hypothesis Test for Correlation

## *P*-Value from a *t* Test

$H_0: \rho = 0$  (There is no linear correlation.)

$H_1: \rho \neq 0$  (There is a linear correlation.)

**Test Statistic: *t***

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

# Hypothesis Test for Correlation

## Conclusion

***P*-value:** Use computer software or use Table A-3 with  $n - 2$  degrees of freedom to find the *P*-value corresponding to the test statistic  $t$ .

If the *P*-value is less than or equal to the significance level, reject  $H_0$  and conclude that there is sufficient evidence to support the claim of a linear correlation.

If the *P*-value is greater than the significance level, fail to reject  $H_0$  and conclude that there is not sufficient evidence to support the claim of a linear correlation.

## Example:

Use the paired pizza subway fare data in Table 10-2 and use the  $P$ -value method to test the claim that there is a linear correlation between the costs of a slice of pizza and the subway fares. Use a 0.05 significance level.

Requirements are satisfied as in the earlier example.

$H_0: \rho = 0$  (There is no linear correlation.)

$H_1: \rho \neq 0$  (There is a linear correlation.)



## Example:

The linear correlation coefficient is  $r = 0.988$  (from an earlier Example) and  $n = 6$  (six pairs of data), so the test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.988}{\sqrt{\frac{1-0.988^2}{6-2}}} = 12.793$$

With  $df = 4$ , Table A-6 yields a  $P$ -value that is less than 0.01.

Computer software generates a test statistic of  $t = 12.692$  and  $P$ -value of 0.00022.

## Example:

Using either method, the  $P$ -value is less than the significance level of 0.05 so we reject  $H_0: \rho = 0$ .

We conclude that there is sufficient evidence to support the claim of a linear correlation between costs of a slice of pizza and subway fares.

# One-Tailed Tests

One-tailed tests can occur with a claim of a positive linear correlation or a claim of a negative linear correlation. In such cases, the hypotheses will be as shown here.

**Claim of *Negative* Correlation**

**Claim of *Positive* Correlation**

**(Left-tailed test)**

**(Right-tailed test)**

---

$$H_0: \rho = 0$$

$$H_0: \rho = 0$$

$$H_1: \rho < 0$$

$$H_1: \rho > 0$$

For these one-tailed tests, the *P*-value method can be used as in earlier chapters.

# Recap

**In this section, we have discussed:**

- ❖ **Correlation.**
- ❖ **The linear correlation coefficient  $r$ .**
- ❖ **Requirements, notation and formula for  $r$ .**
- ❖ **Interpreting  $r$ .**
- ❖ **Formal hypothesis testing.**



# **Section 10-3 Regression**

# Key Concept

In part 1 of this section we find the equation of the straight line that best fits the paired sample data. That equation algebraically describes the relationship between two variables.

The best-fitting straight line is called a **regression line** and its equation is called the **regression equation**.

In part 2, we discuss marginal change, influential points, and residual plots as tools for analyzing correlation and regression results.

# Part 1: Basic Concepts of Regression

# Regression

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



# Definitions

## ❖ Regression Equation

Given a collection of paired data, the regression equation

$$\hat{y} = b_0 + b_1x$$

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
<b>y-intercept of regression equation</b>	$\beta_0$	$b_0$
<b>Slope of regression equation</b>	$\beta_1$	$b_1$
<b>Equation of the regression line</b>	$y = \beta_0 + \beta_1 x$	$y_{\wedge} = b_0 + b_1 x$

# 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.**

# Formulas for $b_0$ and $b_1$

**Formula 10-3** 
$$b_1 = r \frac{s_y}{s_x}$$
 **(slope)**

**Formula 10-4** 
$$b_0 = \bar{y} - b_1 \bar{x}$$
 **(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 10-3 and 10-4, do not round intermediate values.

## Example:

Refer to the sample data given in Table 10-1 in the Chapter Problem. Use technology to find the equation of the regression line in which the explanatory variable (or  $x$  variable) is the cost of a slice of pizza and the response variable (or  $y$  variable) is the corresponding cost of a subway fare.

**Table 10-1** Cost of a Slice of Pizza, Subway Fare, and the CPI

Year	1960	1973	1986	1995	2002	2003
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00
CPI	30.2	48.3	112.3	162.2	191.9	197.8

# Example:

Requirements are satisfied: simple random sample; scatterplot approximates a straight line; no outliers

Here are results from four different technologies

## STATDISK

```
Regression Results:  
Y= b0 + b1x:  
Y Intercept, b0:      0.0345602  
Slope, b1:           0.9450214
```

## EXCEL

	<i>Coefficients</i>	<i>Standard Error</i>
Intercept	0.034560171	0.095012806
X Variable 1	0.945021381	0.074457849

## MINITAB

### Regression Analysis: Subway versus Pizza

```
The regression equation is  
Subway = 0.0346 + 0.945 Pizza
```

## TI-83/84 PLUS

```
LinRegTTest  
y=a+bx  
b≠0 and ρ≠0  
↑a=.034560171  
b=.9450213806  
s=.1229869984  
↓r2=.9757704494
```



## Example:

All of these technologies show that the regression equation can be expressed as  $\hat{y} = 0.0346 + 0.945x$ , where  $\hat{y}$  is the predicted cost of a subway fare and  $x$  is the cost of a slice of pizza.

We should know that the regression equation is an estimate of the true regression equation.

This estimate is based on one particular set of sample data, but another sample drawn from the same population would probably lead to a slightly different equation.

## Example:

**Graph the regression equation**

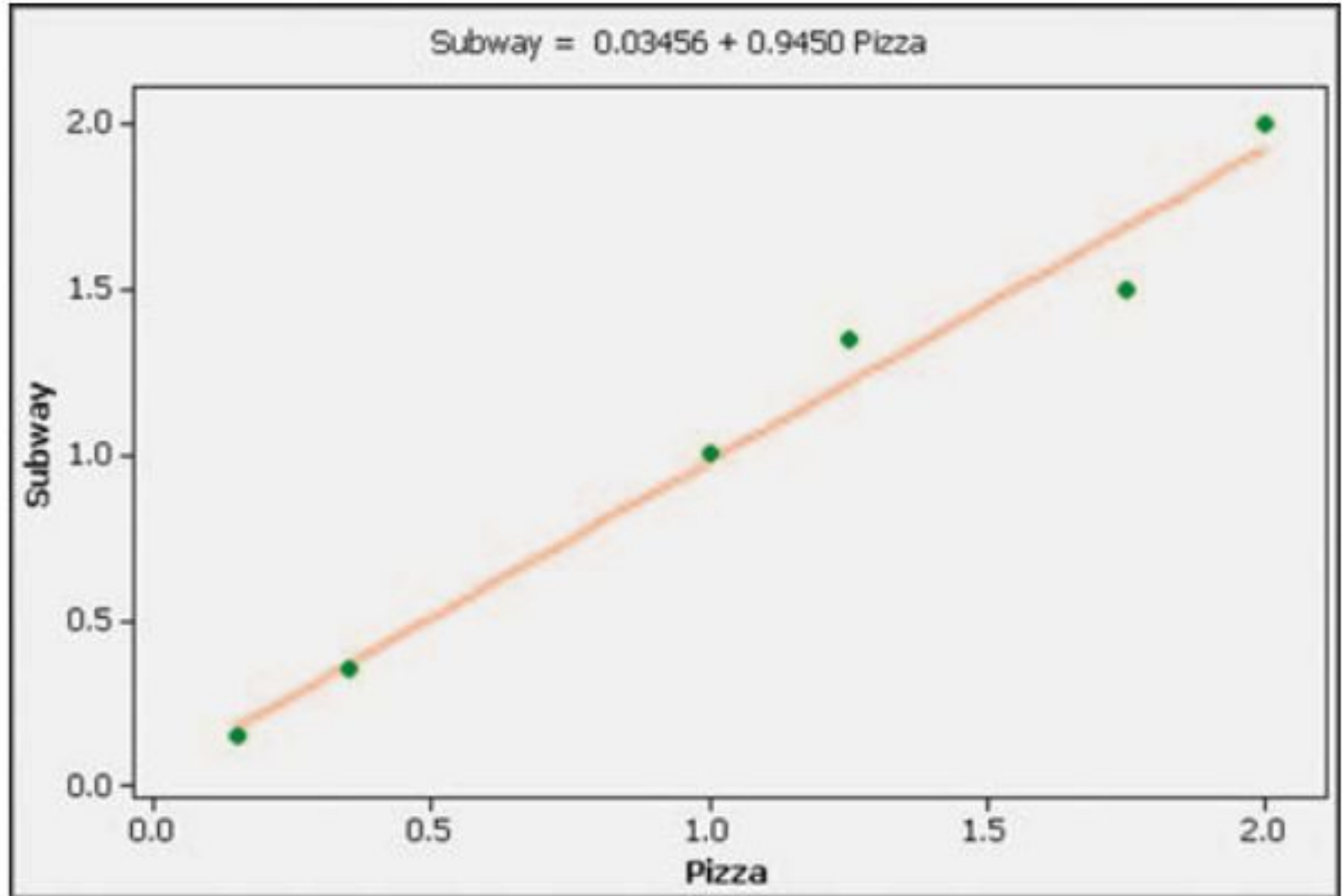
$$\hat{y} = 0.0346 + 0.945x$$

**(from the preceding Example) on the scatterplot of the pizza/subway fare data and examine the graph to subjectively determine how well the regression line fits the data.**

**On the next slide is the Minitab display of the scatterplot with the graph of the regression line included. We can see that the regression line fits the data quite well.**

# Example:

## MINITAB



# Using the Regression Equation for Predictions

- 1. Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.**
- 2. Use the regression equation for predictions only if the linear correlation coefficient  $r$  indicates that there is a linear correlation between the two variables (as described in Section 10-2).**

# Using the Regression Equation for Predictions

- 3. Use the regression line for predictions only if the data do not go much beyond the scope of the available sample data. (Predicting too far beyond the scope of the available sample data is called *extrapolation*, and it could result in bad predictions.)**
- 4. If the regression equation does not appear to be useful for making predictions, the best predicted value of a variable is its point estimate, which is its sample mean.**

# Strategy for Predicting Values of Y

## Strategy for Predicting Values of Y

Is the regression equation a good model?

- The regression line graphed in the scatterplot shows that the line fits the points well.
- $r$  indicates that there is a linear correlation.
- The prediction is not much beyond the scope of the available sample data.

Yes.  
The regression equation is a good model.

No.  
The regression equation is not a good model.

Substitute the given value of  $x$  into the regression equation  $\hat{y} = b_0 + b_1x$ .

Regardless of the value of  $x$ , the best predicted value of  $y$  is the value of  $\bar{y}$  (the mean of the  $y$  values).

# Using the Regression Equation for Predictions

If the regression equation is not a good model, the best predicted value of  $y$  is simply  $\hat{y}$ , the mean of the  $y$  values.

Remember, this strategy applies to linear patterns of points in a scatterplot.

If the scatterplot shows a pattern that is not a straight-line pattern, other methods apply, as described in Section 10-6.

# Part 2: Beyond the Basics of Regression



# Definitions

In working with two variables related by a regression equation, the **marginal change** in a variable is the amount that it changes when the other variable changes by exactly one unit. The slope  $b_1$  in the regression equation represents the marginal change in  $y$  that occurs when  $x$  changes by one unit.

# Definitions

In a scatterplot, an **outlier** is a point lying far away from the other data points.

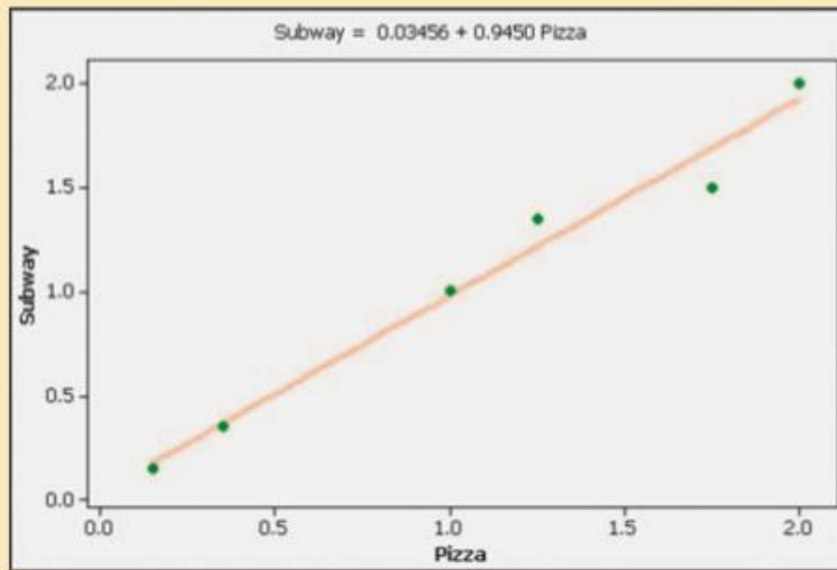
Paired sample data may include one or more **influential points**, which are points that strongly affect the graph of the regression line.

## Example:

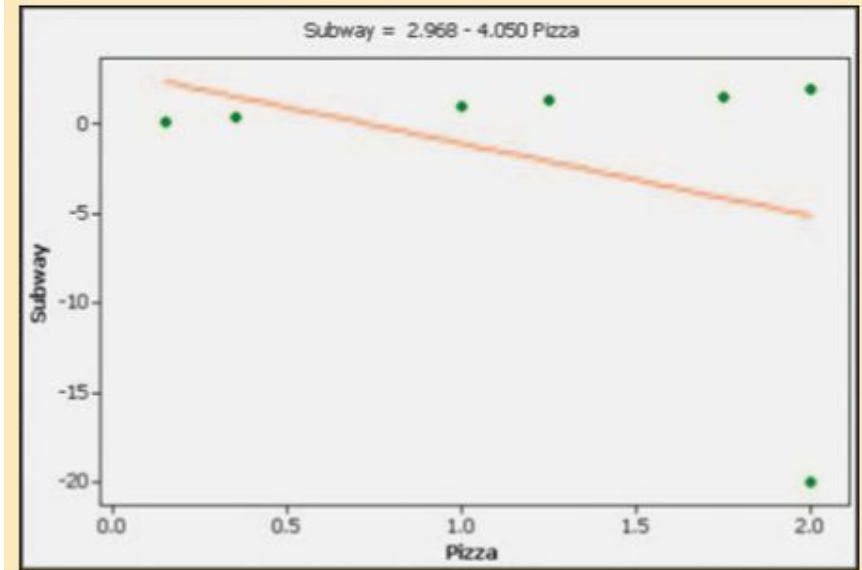
Consider the pizza subway fare data from the Chapter Problem. The scatterplot located to the left on the next slide shows the regression line. If we include this additional pair of data:  $x = 2.00, y = -20.00$  (pizza is still \$2.00 per slice, but the subway fare is \$-20.00 which means that people are paid \$20 to ride the subway), this additional point would be an influential point because the graph of the regression line would change considerably, as shown by the regression line located to the right.

# Example:

**PIZZA/SUBWAY DATA FROM THE CHAPTER PROBLEM**



**PIZZA/SUBWAY DATA WITH AN INFLUENTIAL POINT**



## Example:

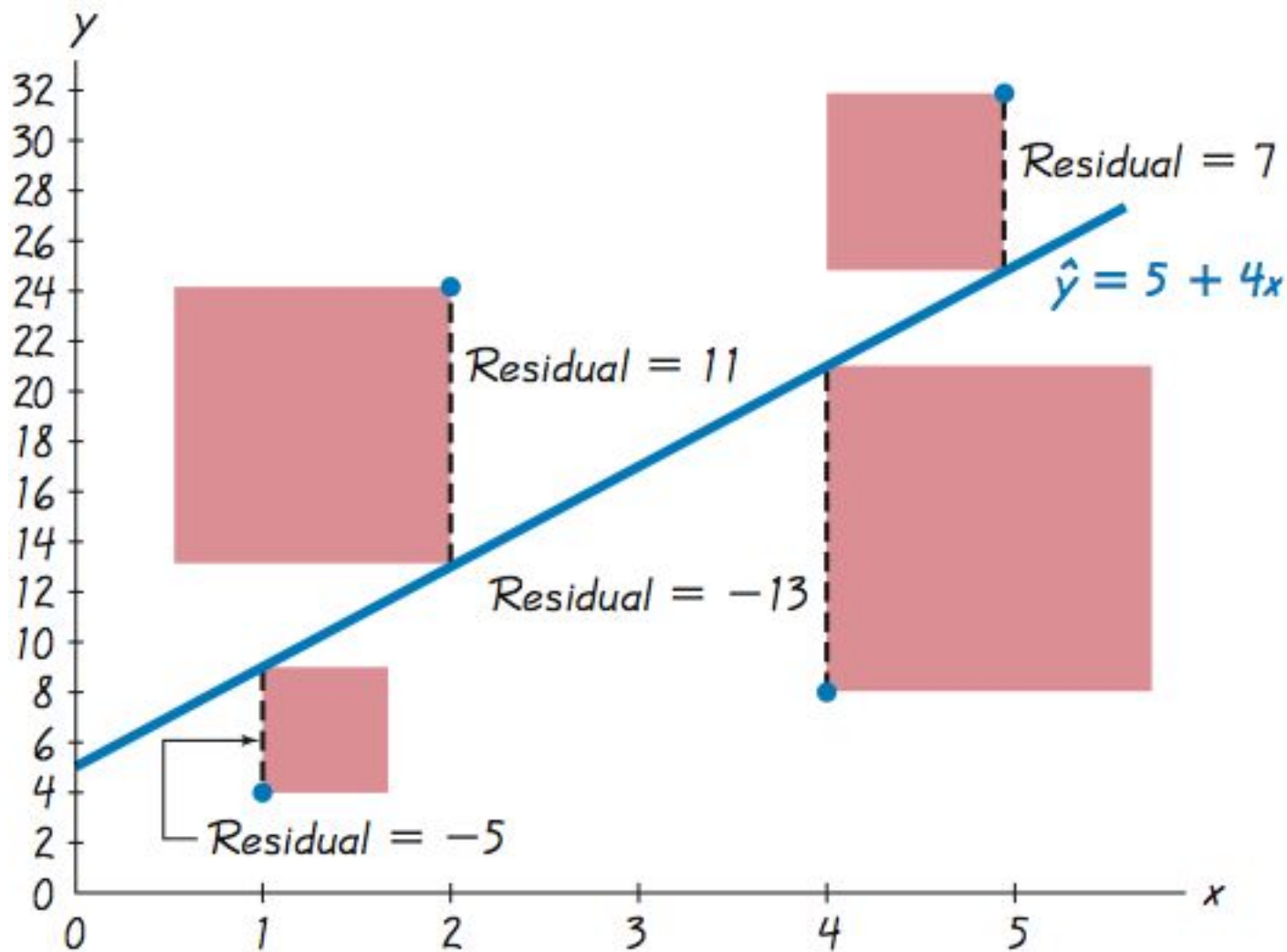
**Compare the two graphs and you will see clearly that the addition of that one pair of values has a very dramatic effect on the regression line, so that additional point is an influential point. The additional point is also an outlier because it is far from the other points.**

# Definition

For a pair of sample  $x$  and  $y$  values, the **residual** is the difference between the *observed* sample value of  $y$  and the  $y$ -value that is *predicted* by using the regression equation. That is,

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

# Residuals



# Definitions

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



# Definitions

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

# Residual Plot Analysis

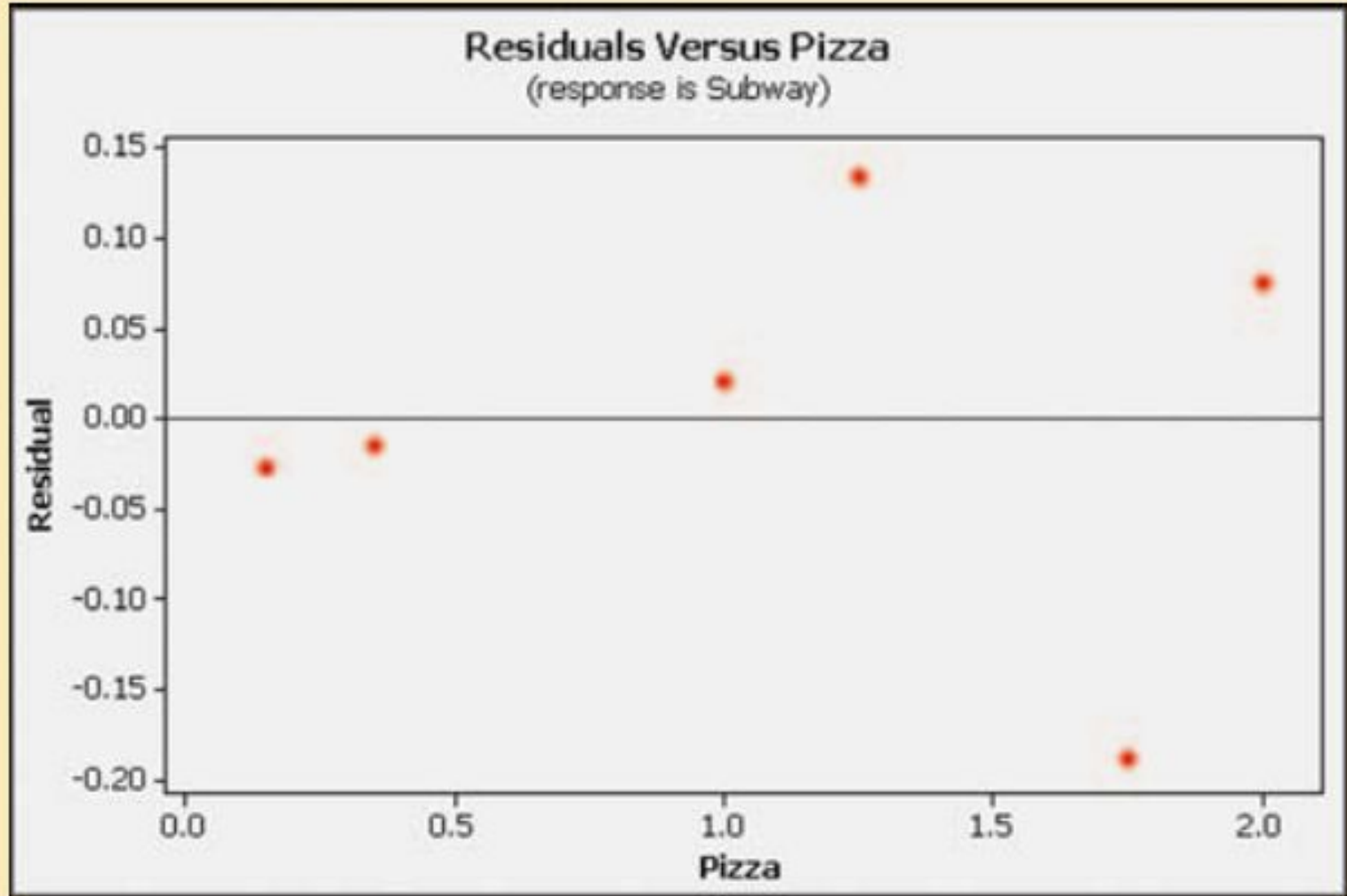
**When analyzing a residual plot, look for a pattern in the way the points are configured, and use these criteria:**

**The residual plot should not have an obvious pattern that is not a straight-line pattern.**

**The residual plot should not become thicker (or thinner) when viewed from left to right.**

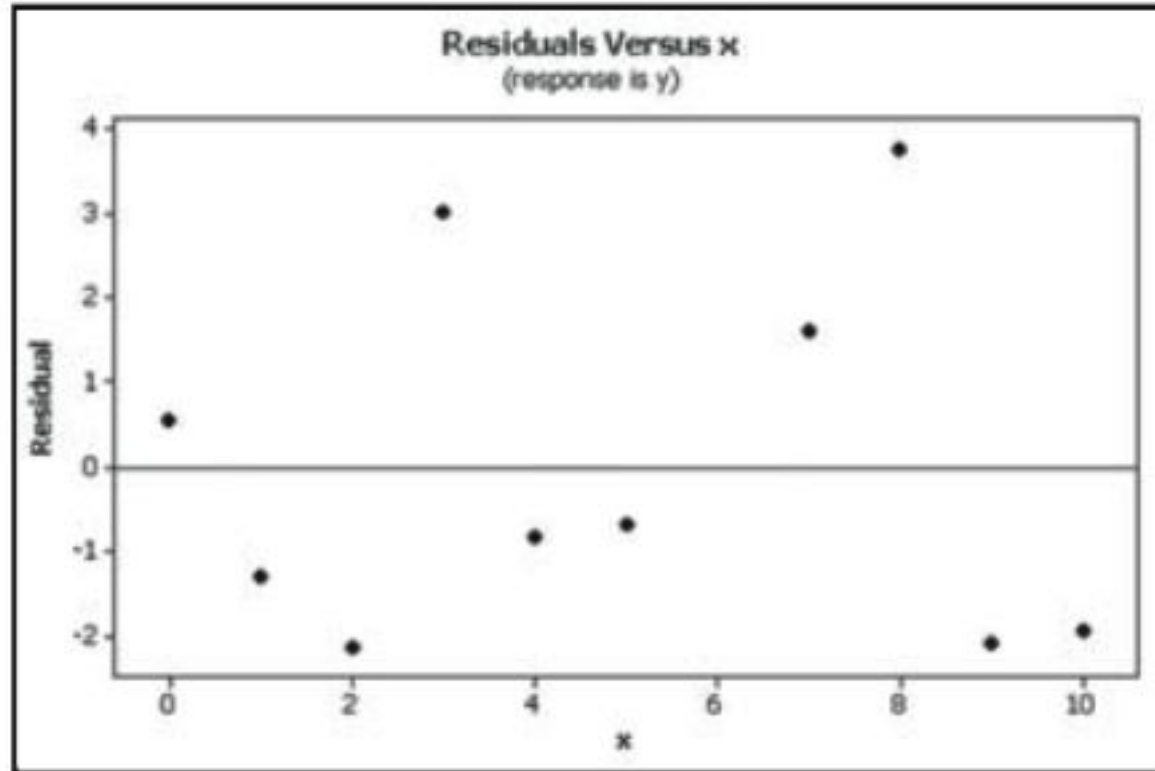
# Residuals Plot - Pizza/Subway

MINITAB



# Residual Plots

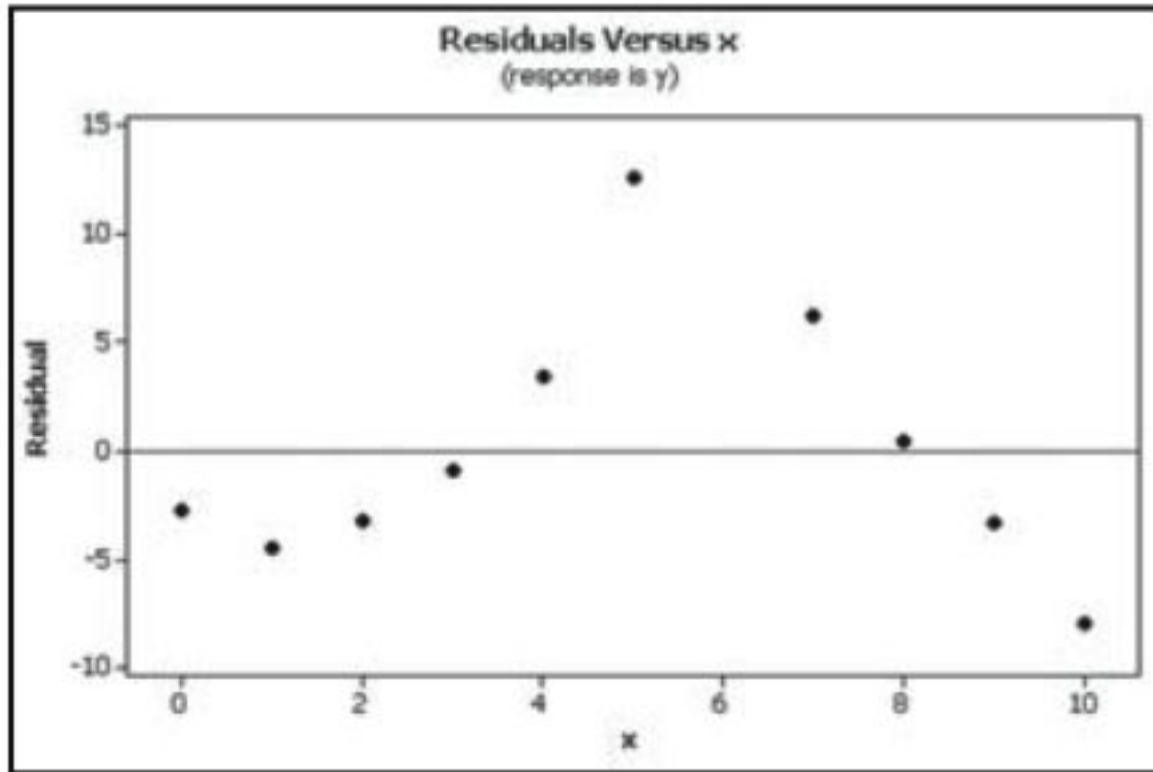
## MINITAB



**Residual Plot Suggesting that the Regression Equation is a Good Model**

# Residual Plots

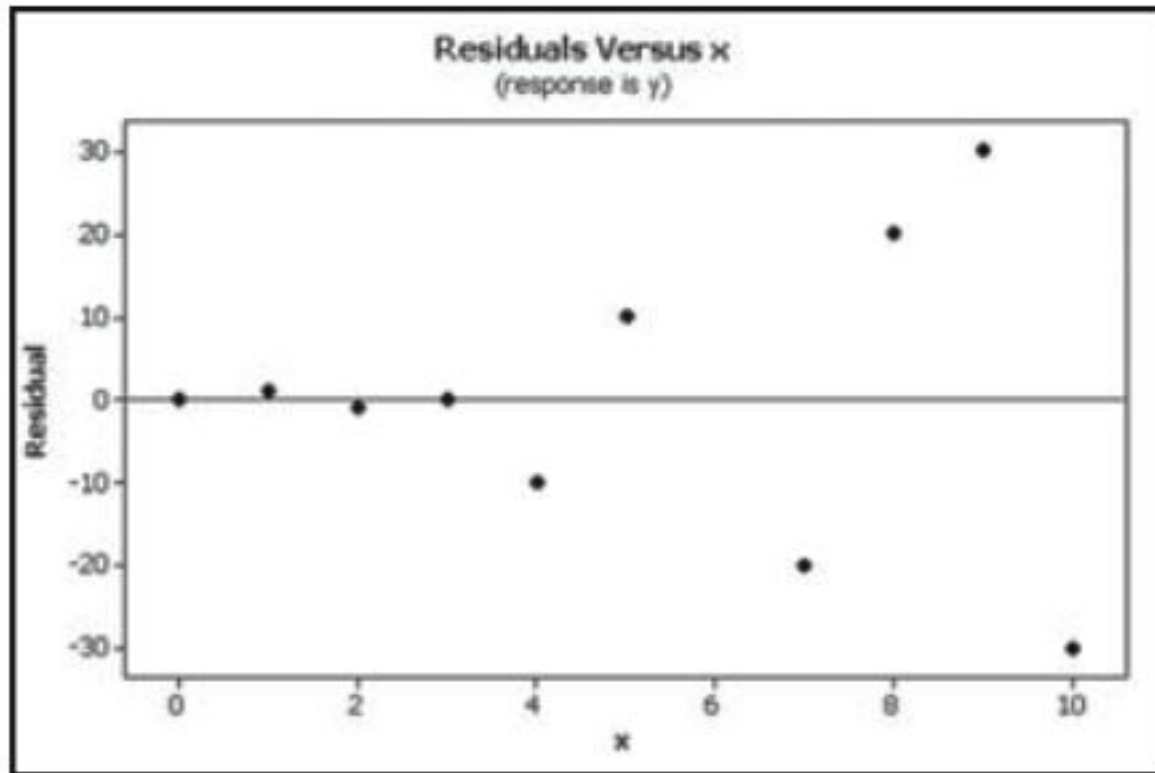
MINITAB



**Residual Plot with an Obvious Pattern,  
Suggesting that the Regression Equation  
Is Not a Good Model**

# Residual Plots

MINITAB



**Regression Plot that Becomes Thicker,  
Suggesting that the Regression Equation  
Is Not a Good Model**

# Complete Regression Analysis

- 1. Construct a scatterplot and verify that the pattern of the points is approximately a straight-line pattern without outliers. (If there are outliers, consider their effects by comparing results that include the outliers to results that exclude the outliers.)**
- 2. Construct a residual plot and verify that there is no pattern (other than a straight-line pattern) and also verify that the residual plot does not become thicker (or thinner).**

# Complete Regression Analysis

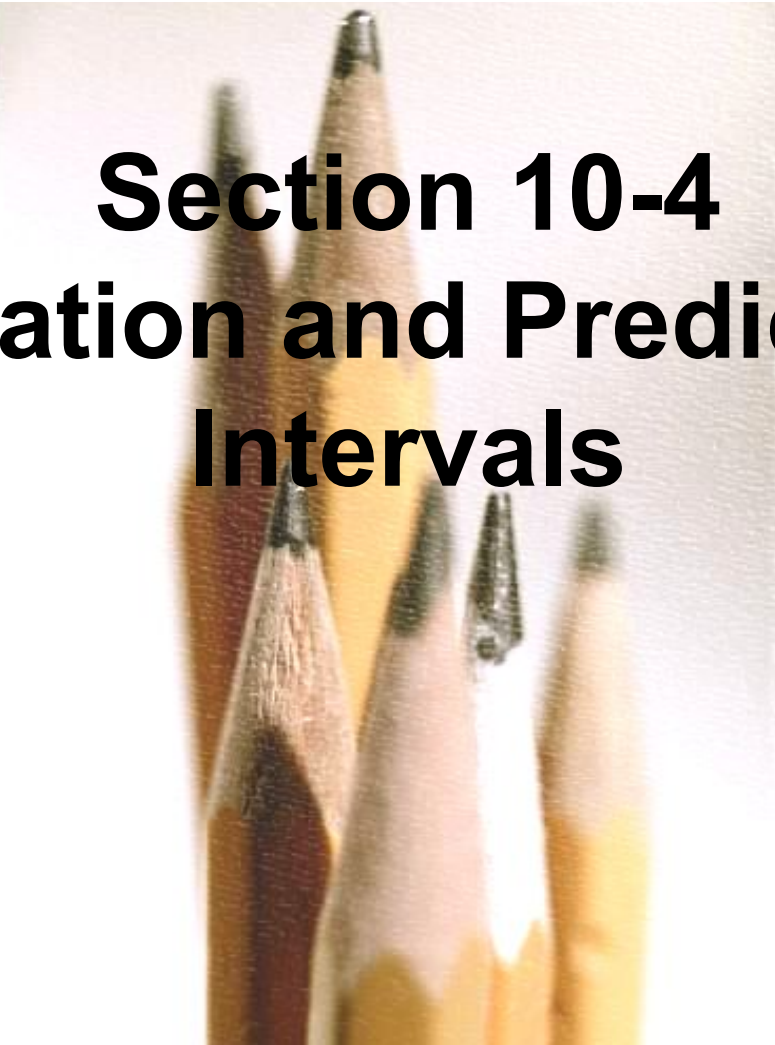
- 3. Use a histogram and/or normal quantile plot to confirm that the values of the residuals have a distribution that is approximately normal.**
- 4. Consider any effects of a pattern over time.**



# Recap

**In this section we have discussed:**

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



**Section 10-4**  
**Variation and Prediction**  
**Intervals**

# Key Concept

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

# Definition

**Assume that we have a collection of paired data containing the sample point  $(x, y)$ , that  $\hat{y}$  is the predicted value of  $y$  (obtained by using the regression equation), and that the mean of the sample  $y$ -values is  $\bar{y}$ .**

# Definition

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

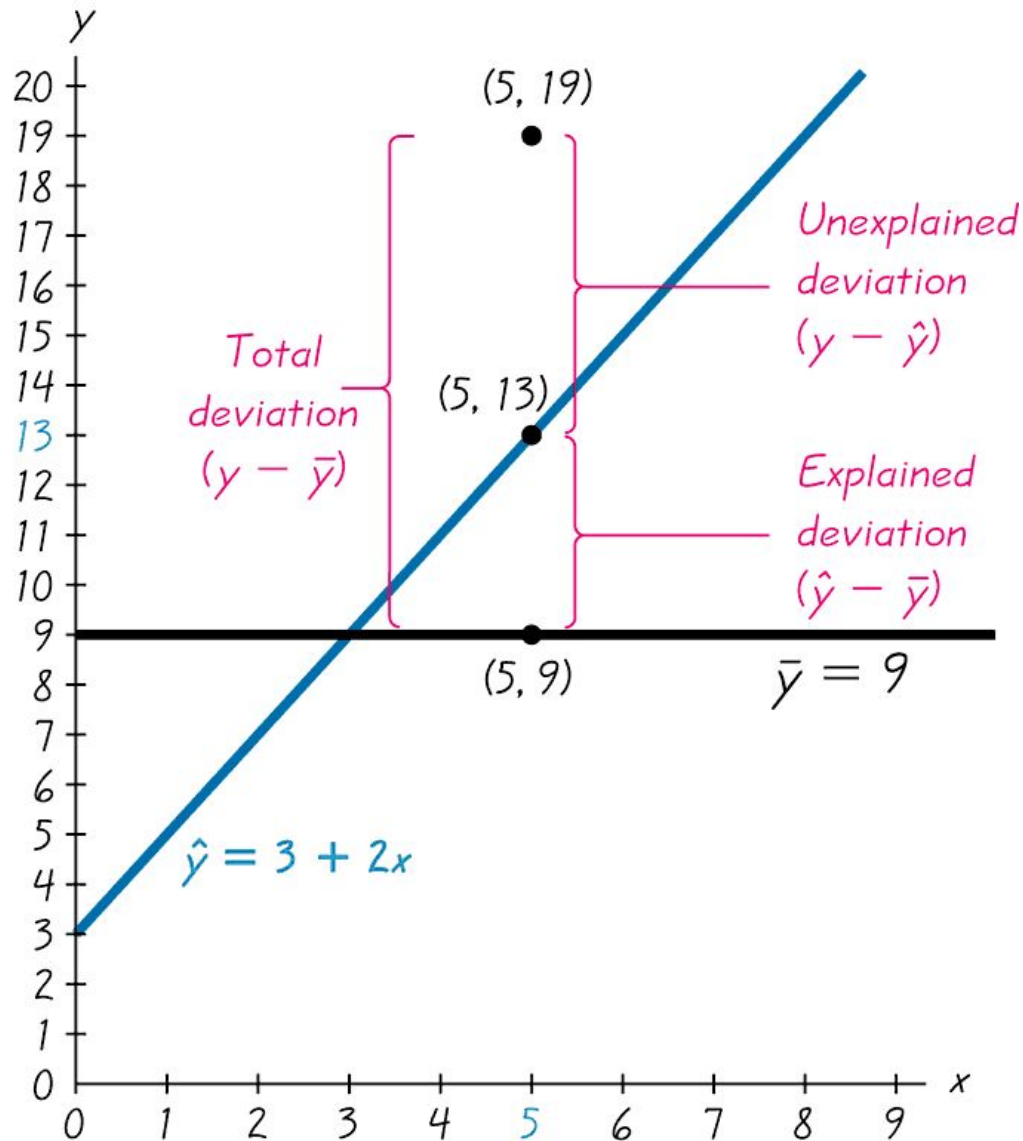
The **explained deviation** is the vertical distance  $\hat{y} - \bar{y}$ , which is the distance between the predicted  $y$ -value and the horizontal line passing through the sample mean  $\bar{y}$ .

# Definition

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 10-7





# Relationships

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

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

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

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

**Formula 10-5**

# 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

A **prediction interval**, is an interval estimate of a predicted value of  $y$ .

# Definition

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 10-6**

## Example:

Use Formula 10-6 to find the standard error of estimate  $s_e$  for the paired pizza/subway fare data listed in Table 10-1 in the Chapter Problem.

$$n = 6$$

$$\Sigma y^2 = 9.2175$$

$$\Sigma y = 6.35$$

$$\Sigma xy = 9.4575$$

$$b_0 = 0.034560171$$

$$b_1 = 0.94502138$$

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

$$s_e = \sqrt{\frac{9.2175 - (0.034560171)(6.35) - (0.94502138)(9.4575)}{6 - 2}}$$

$$s_e = 0.12298700 = 0.123$$

# Prediction Interval for an Individual $y$

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

where

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

$x_0$  represents the given value of  $x$   
 $t_{\alpha/2}$  has  $n - 2$  degrees of freedom

## Example:

For the paired pizza/subway fare costs from the Chapter Problem, we have found that for a pizza cost of \$2.25, the best predicted cost of a subway fare is \$2.16. Construct a 95% prediction interval for the cost of a subway fare, given that a slice of pizza costs \$2.25 (so that  $x = 2.25$ ).

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

$$E = (2.776)(0.12298700) \sqrt{1 + \frac{1}{6} + \frac{6(2.25 - 1.0833333)^2}{6(9.77) - (6.50)^2}}$$

$$E = (2.776)(0.12298700)(1.2905606) = 0.441$$



## Example:

Construct the confidence interval.

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

$$2.16 - 0.441 < y < 2.16 + 0.441$$

$$1.72 < y < 2.60$$

# Recap

**In this section we have discussed:**

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



**Section 10-5**  
**Multiple Regression**

# 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.

# Part 1: Basic Concepts of a Multiple Regression Equation

# Definition

A **multiple regression equation** expresses a linear relationship between a response variable  $y$  and two or more predictor variables  $(x_1, x_2, x_3 \dots, x_k)$

The general form of the multiple regression equation obtained from sample data is

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

# Notation

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

(General form of the multiple regression equation)

**$n$  = sample size**

**$k$  = number of predictor variables**

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

**$x_1, x_2, x_3 \dots, x_k$  are the predictor variables**

# Notation - cont

$\beta_0, \beta_1, \beta_2, \dots, \beta_k$  are the parameters for the multiple regression equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

$b_0, b_1, b_2, \dots, b_k$  are the *sample estimates* of the parameters  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$



# Technology

Use a statistical software package such as

❖ **STATDISK**

❖ **Minitab**

❖ **Excel**

❖ **TI-83/84**

## Example:

**Table 10-6 includes a random sample of heights of mothers, fathers, and their daughters (based on data from the National Health and Nutrition Examination). Find the multiple regression equation in which the response ( $y$ ) variable is the height of a daughter and the predictor ( $x$ ) variables are the height of the mother and height of the father.**

Height of Mother	Height of Father	Height of Daughter
63	64	58.6
67	65	64.7
64	67	65.3
60	72	61.0
65	72	65.4
67	72	67.4
59	67	60.9
60	71	63.1
58	66	60.0
72	75	71.1
63	69	62.2
67	70	67.2
62	69	63.4
69	62	68.4
63	66	62.2
64	76	64.7
63	69	59.6
64	68	61.0
60	66	64.0
65	68	65.4

# Example:

The Minitab results are shown here:

The regression equation is

**Height = 7.5 + 0.707 Mother + 0.164 Father**

Predictor	Coef	SE Coef	T	P
Constant	7.45	10.88	0.69	0.503
Mother	0.7072	0.1289	5.49	0.000
Father	0.1636	0.1266	1.29	0.213

**S = 1.93990      R-Sq = 67.5%      R-Sq(adj) = 63.7%**

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	132.997	66.499	17.67	<b>0.000</b>
Residual Error	17	63.975	3.763		
Total	19	196.972			

## Example:

From the display, we see that the multiple regression equation is

$$\text{Height} = 7.5 + 7.07\text{Mother} + 0.164\text{Father}$$

Using our notation presented earlier in this section, we could write this equation as

$$\hat{y} = 7.5 + 0.707x_1 + 0.164x_2$$

where  $\hat{y}$  is the predicted height of a daughter,  $x_1$  is the height of the mother, and  $x_2$  is the height of the father.

# Definition

- ❖ The **multiple coefficient of determination**  $R^2$  is a measure of how well the multiple regression equation fits the sample data.
- ❖ 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 Coefficient of Determination

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

where  $n$  = sample size

$k$  = number of predictor ( $x$ ) variables

**Formula 10-7**

# ***P*-Value**

**The *P*-value is a measure of the overall significance of the multiple regression equation. Like the adjusted  $R^2$ , this *P*-value is a good measure of how well the equation fits the sample data.**

# P-Value

The displayed Minitab *P*-value of 0.000 (rounded to three decimal places) is small, indicating that the multiple regression equation has good overall significance and is usable for predictions. That is, it makes sense to predict heights of daughters based on heights of mothers and fathers. The value of 0.000 results from a test of the null hypothesis that  $\beta_1 = \beta_2 = 0$ . Rejection of  $\beta_1 = \beta_2 = 0$  implies that at least one of  $\beta_1$  and  $\beta_2$  is not 0, indicating that this regression equation is effective in predicting heights of daughters.



# 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.

# Finding the Best Multiple Regression Equation

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.**

# Part 2: Dummy Variables and Logistic Equations

# Dummy Variable

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**.

As the name implies logistic regression involves the use of natural logarithms. This text book does not include detailed procedures for using logistic regression.

# 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**

# 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.

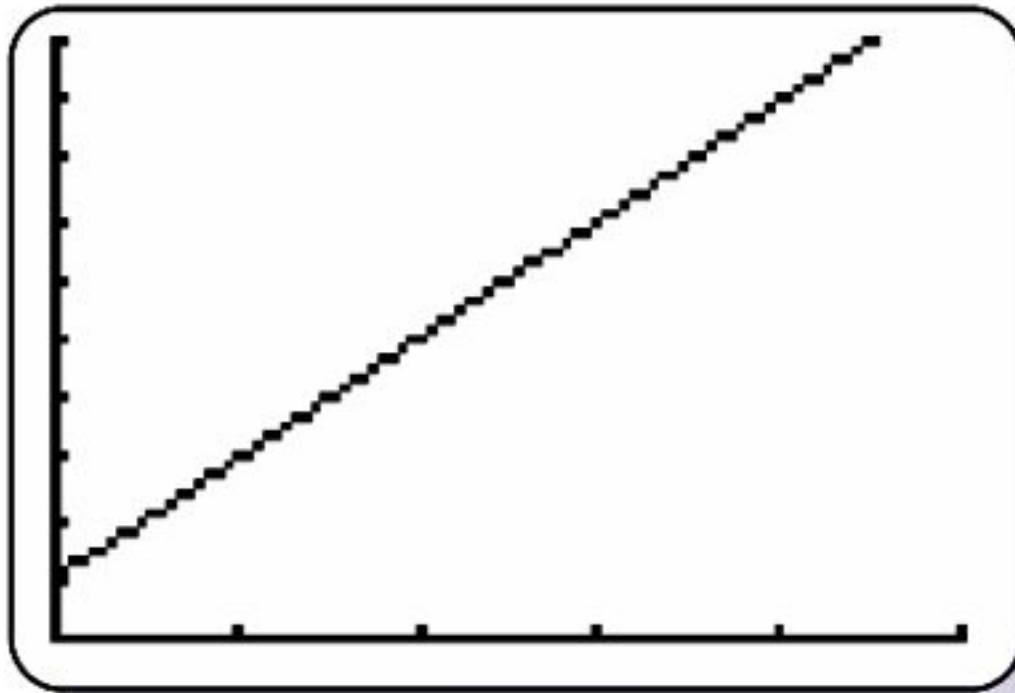


# TI-83/84 Plus Generic Models

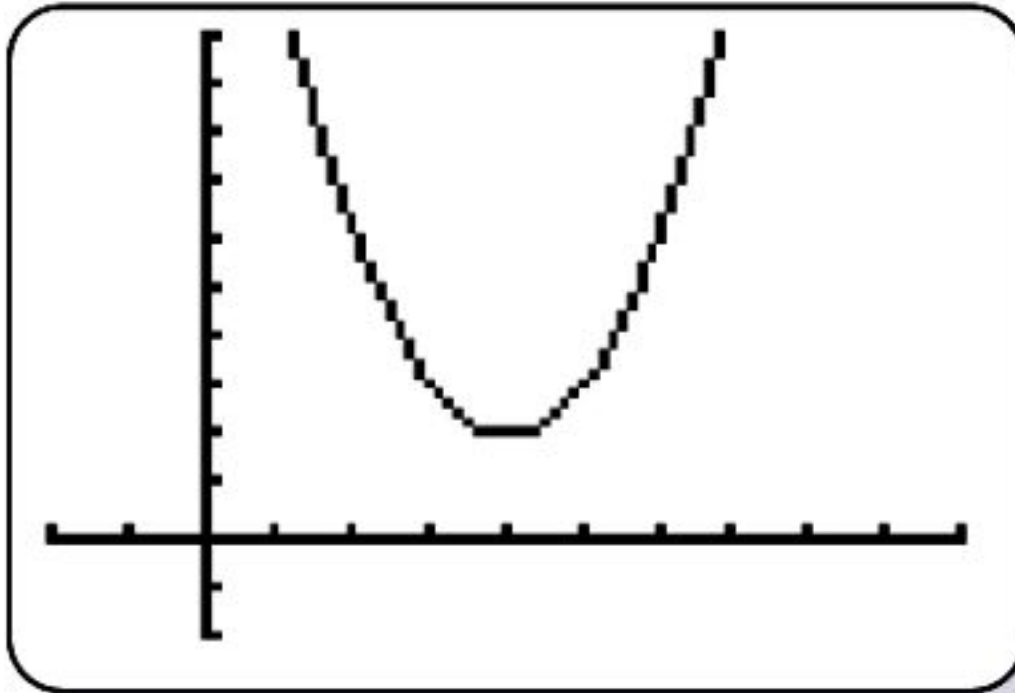
- ❖ **Linear:**  $y = a + bx$
- ❖ **Quadratic:**  $y = ax^2 + bx + c$
- ❖ **Logarithmic:**  $y = a + b \ln x$
- ❖ **Exponential:**  $y = ab^x$
- ❖ **Power:**  $y = ax^b$

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

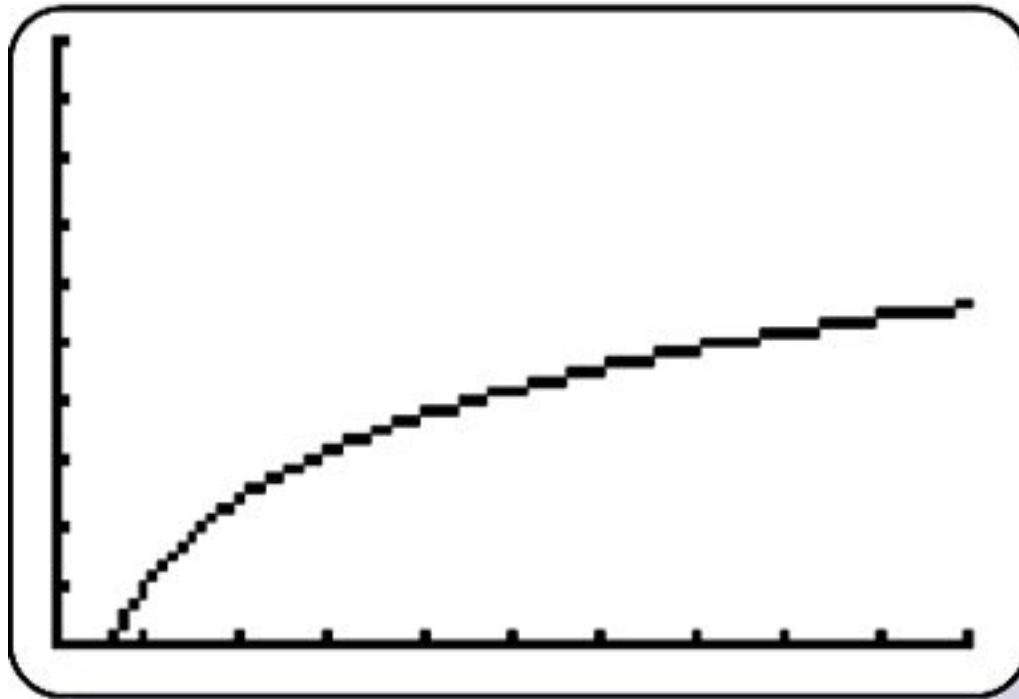
Linear:  $y = 1 + 2x$



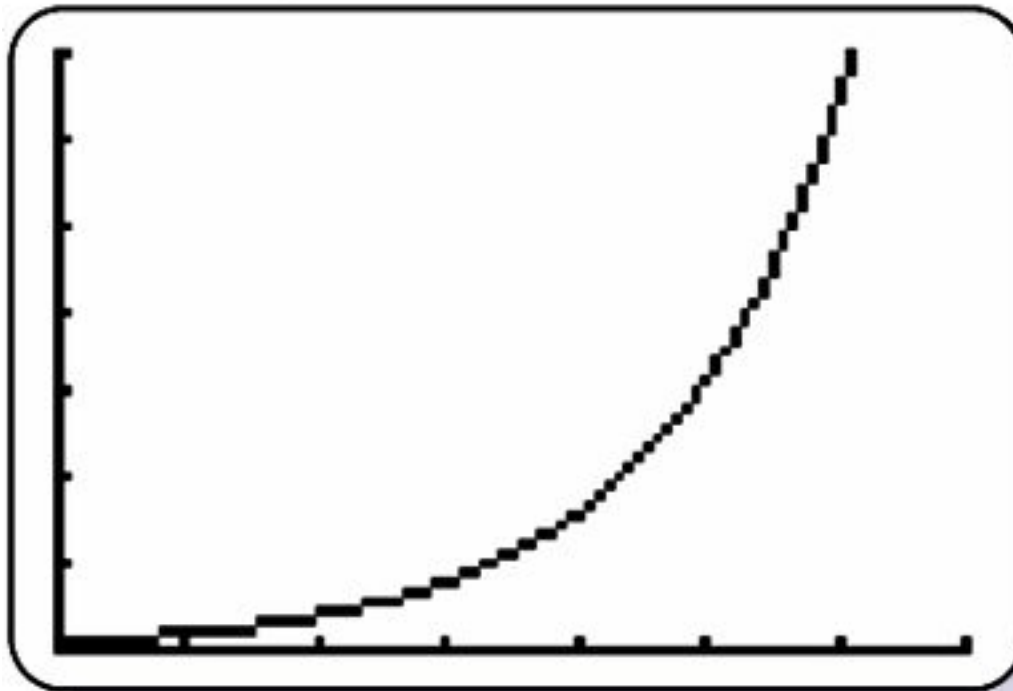
Quadratic:  $y = 2x^2 - 8x + 9$



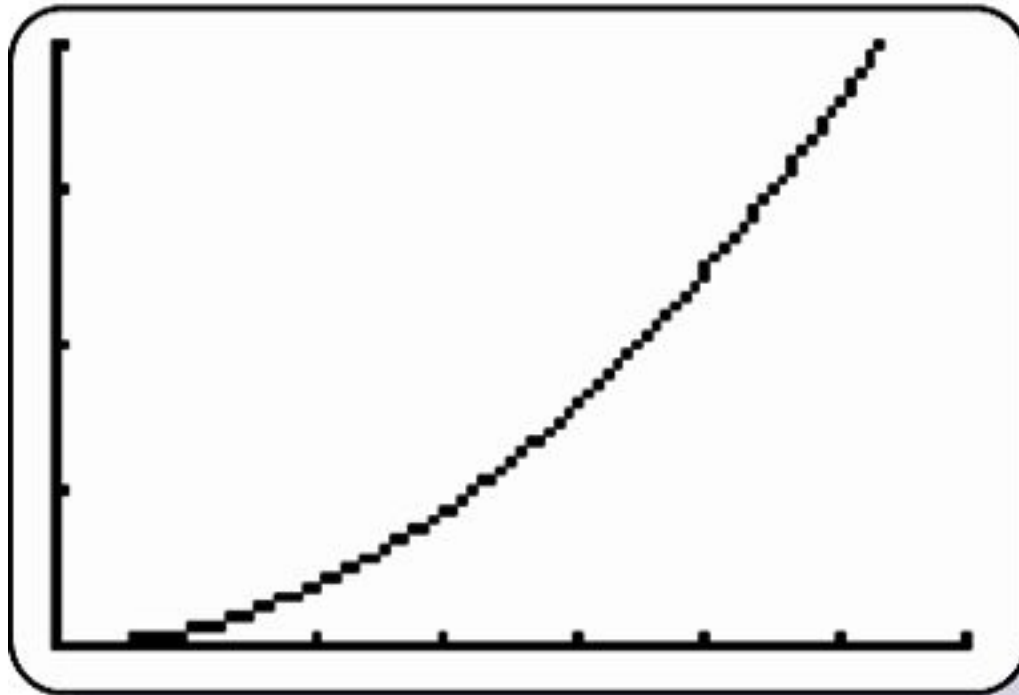
Logarithmic:  $y = 1 + 2\ln x$



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.



# Important Point

**“The best choice (of a model) depends on the set of data being analyzed and requires an exercise in judgment, not just computation.”**

# 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.**