Chapter 10
Correlation and Regression

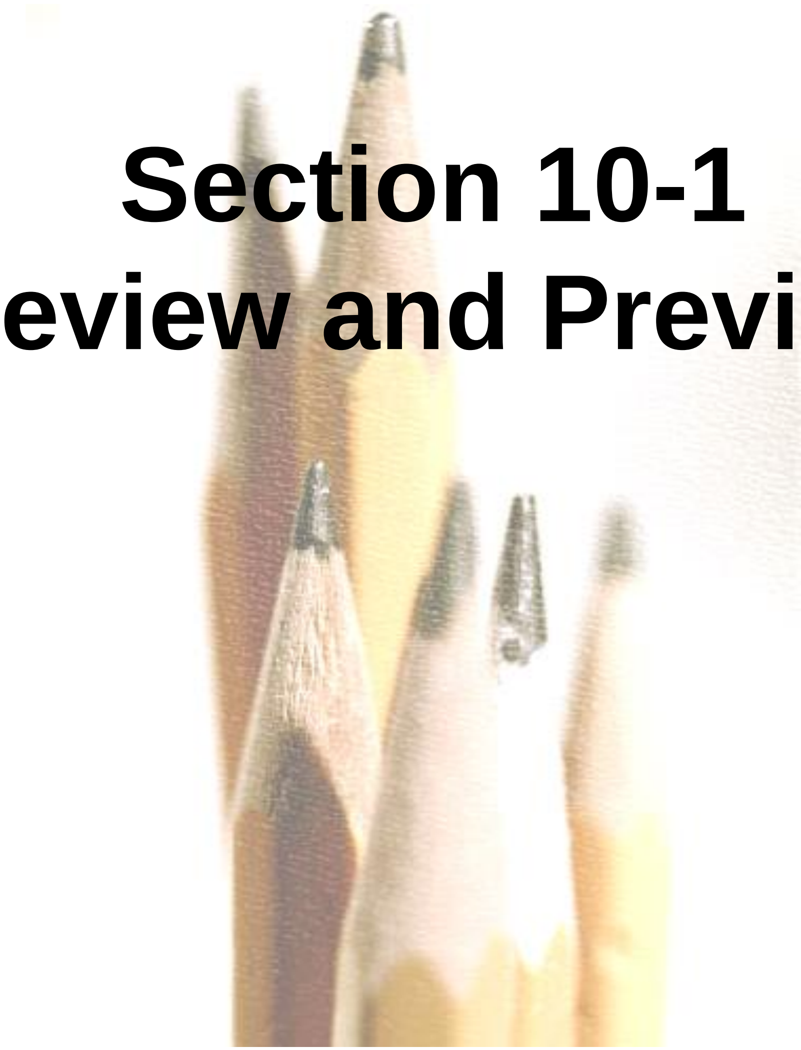
10-1 Review and Preview

10-2 Correlation

10-3 Regression

10-4 Variation and Prediction Intervals

Section 10-1
Review and Preview



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.

- Causal Relationship - One event triggers the occurrence of another event. A causal relationship is also referred to as cause and effect. It is mostly uni-directional. Example: Generate income to do expenditures
- Correlation - a mutual relationship or connection between two or more variables. Example: with the increase in the exercise level, the calories burned will also increase
- In some cases it may be possible that two variables possess correlation and they are causal i.e., exercise/ calorie burnt example
- In others, correlated variables may vary with variation in 3rd variable, like example in next slide

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

- If there is a correlation between two variables, how can it be described? Is there an *equation* that can be used to predict the cost of a subway fare given the cost of a slice of pizza?
- If we can predict the cost of a subway fare, how accurate is that prediction likely to be?
- Is there also a correlation between the CPI and the cost of a subway fare, and if so, is the CPI better for predicting the cost of a subway fare?

The Consumer Price Index (CPI) measures the monthly change in prices paid by consumers. It is a measure of inflation

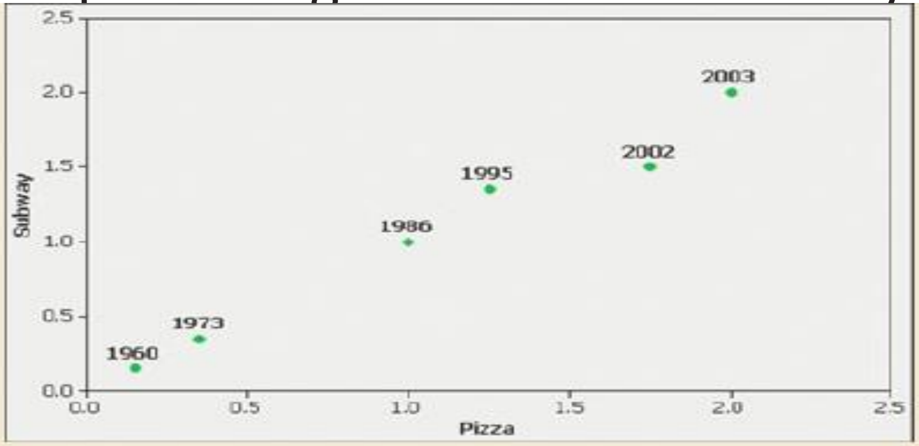
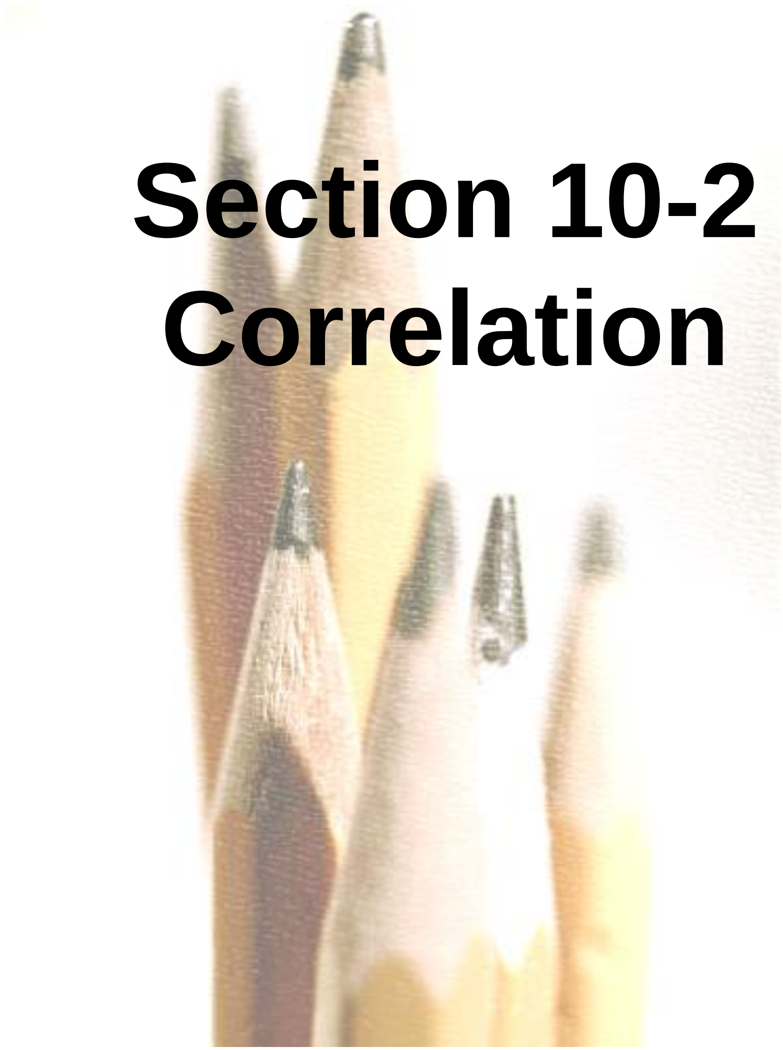


Figure 10-1 Scatterplot of Pizza Costs and Subway Costs

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, 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

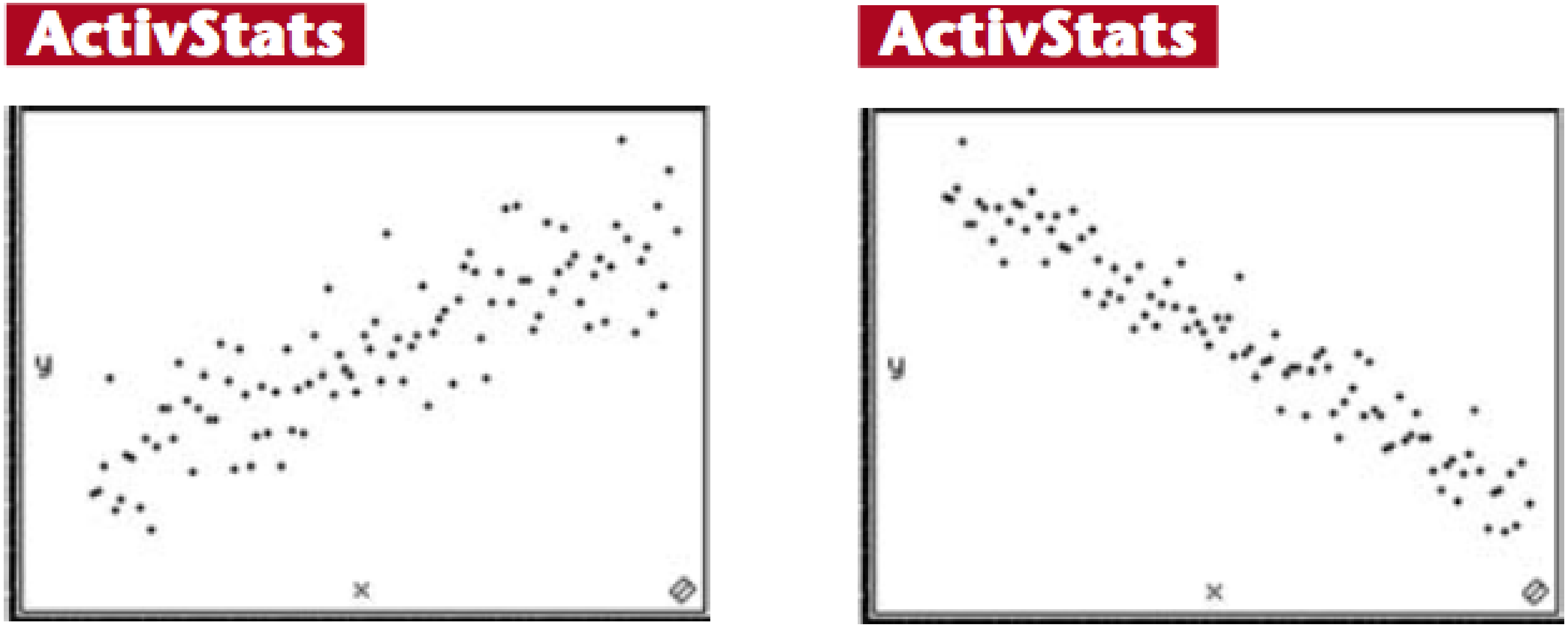
A correlation exists between two variables when the values of one are somehow associated with the values of the other in some way.

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



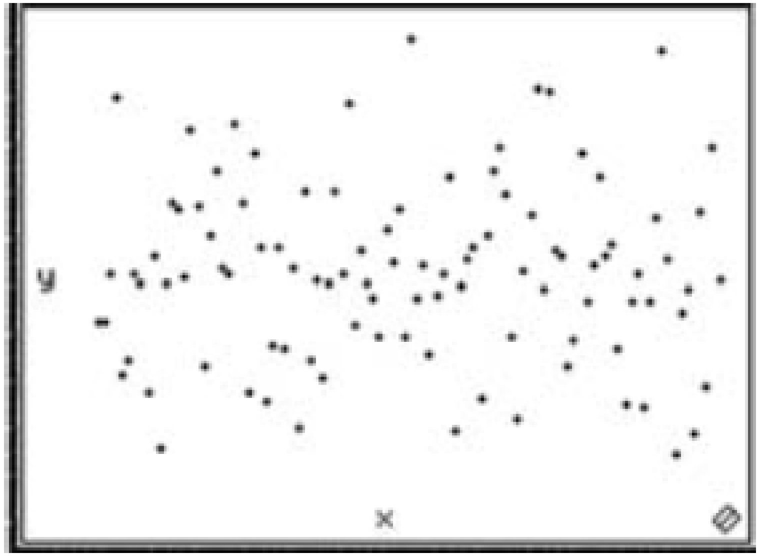
(a) Positive correlation:
***r* = 0.851**

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

Figure 10-2

Scatterplots of Paired Data

ActivStats

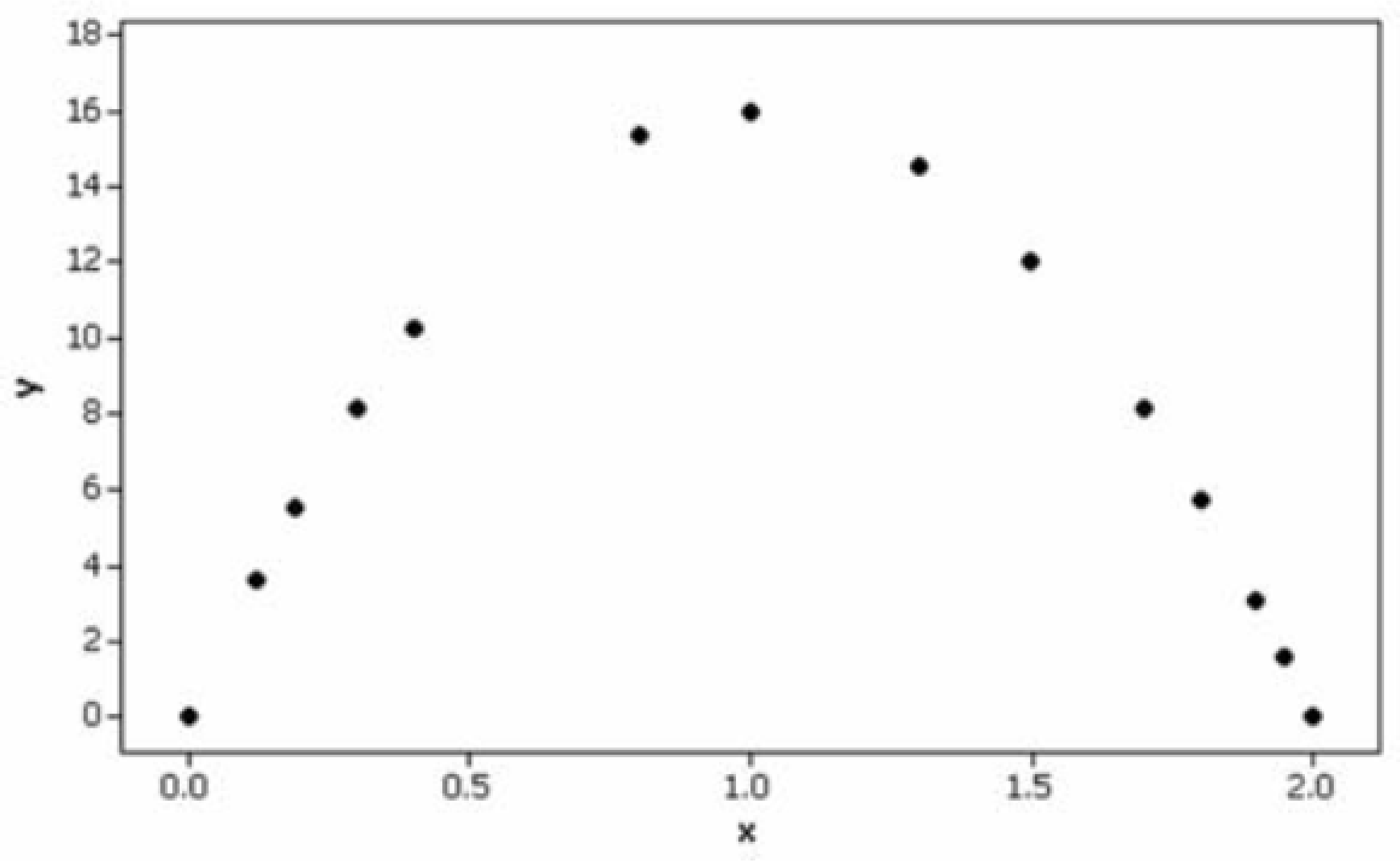


(c) No correlation: *r* = 0

Figure 10-2

Scatterplots of Paired Data

Minitab



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

Figure 10-2

Definition

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

Incase of population linear correlation coefficient is denoted by ρ

Calculating Linear Correlation Coefficient *r*

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.

Objective

Determine whether there is a linear correlation between two variables.

Calculating Linear Correlation Coefficient *r* - Terminologies

n = number of pairs of sample data

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 then the total squared.

Calculating Linear Correlation Coefficient *r* - Terminologies

Σxy 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.

ρ = linear correlation coefficient for population data.

Calculating Linear Correlation Coefficient *r* - 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

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.**

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. $-1 \le r \le 1$
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 formula with these paired sample values to find the value of the linear correlation coefficient *r* for the paired sample data.

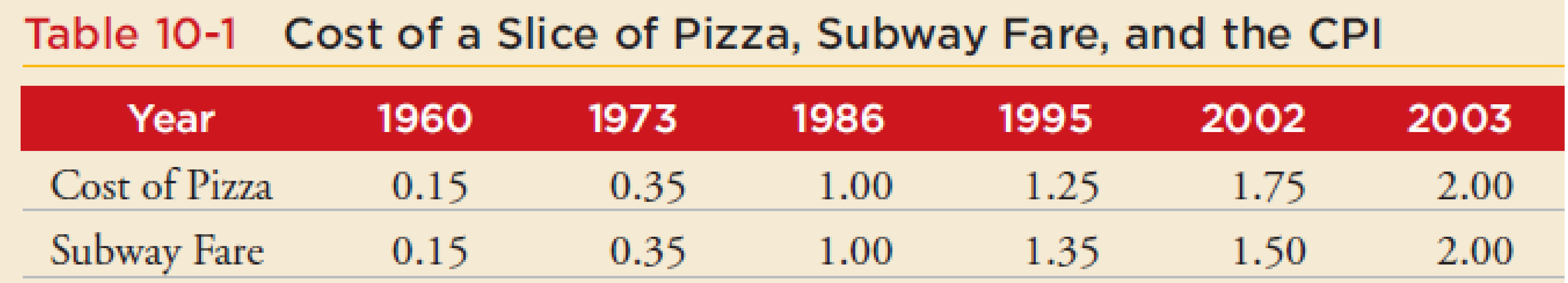
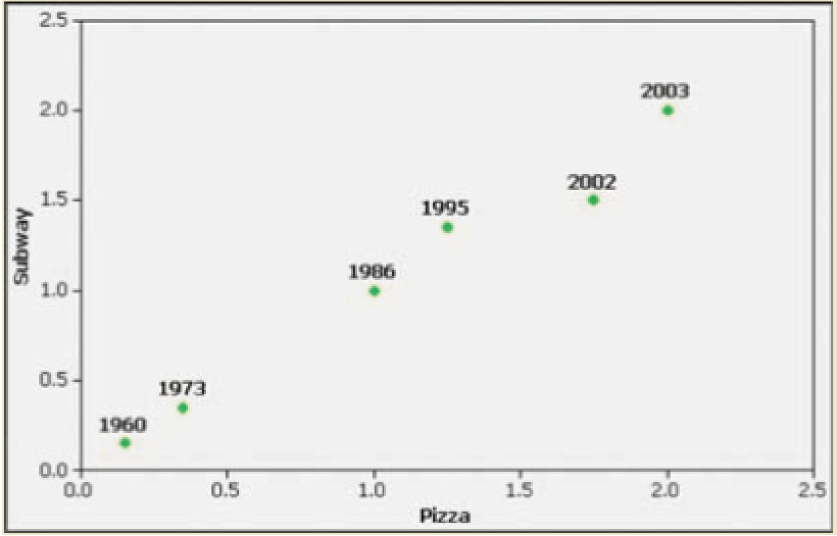


Table 10-1 Cost of a Slice of Pizza, Subway Fare, and the CPI

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

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

Example:



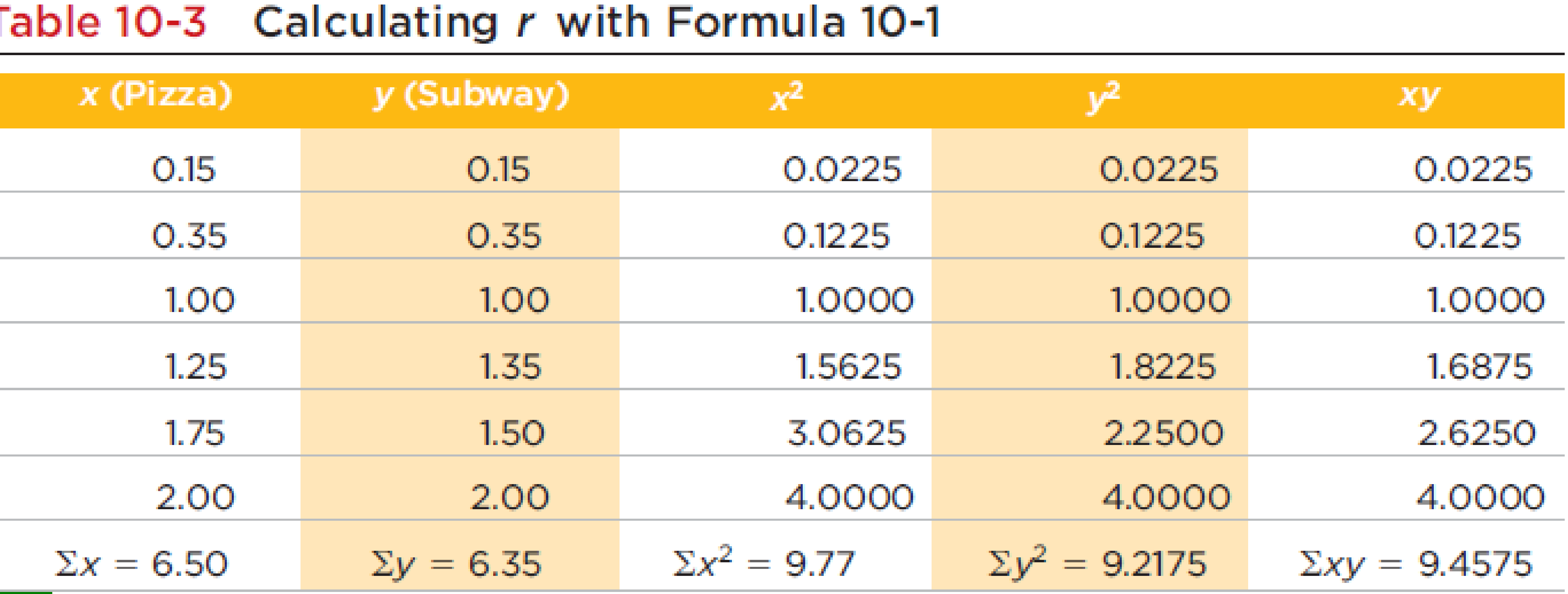
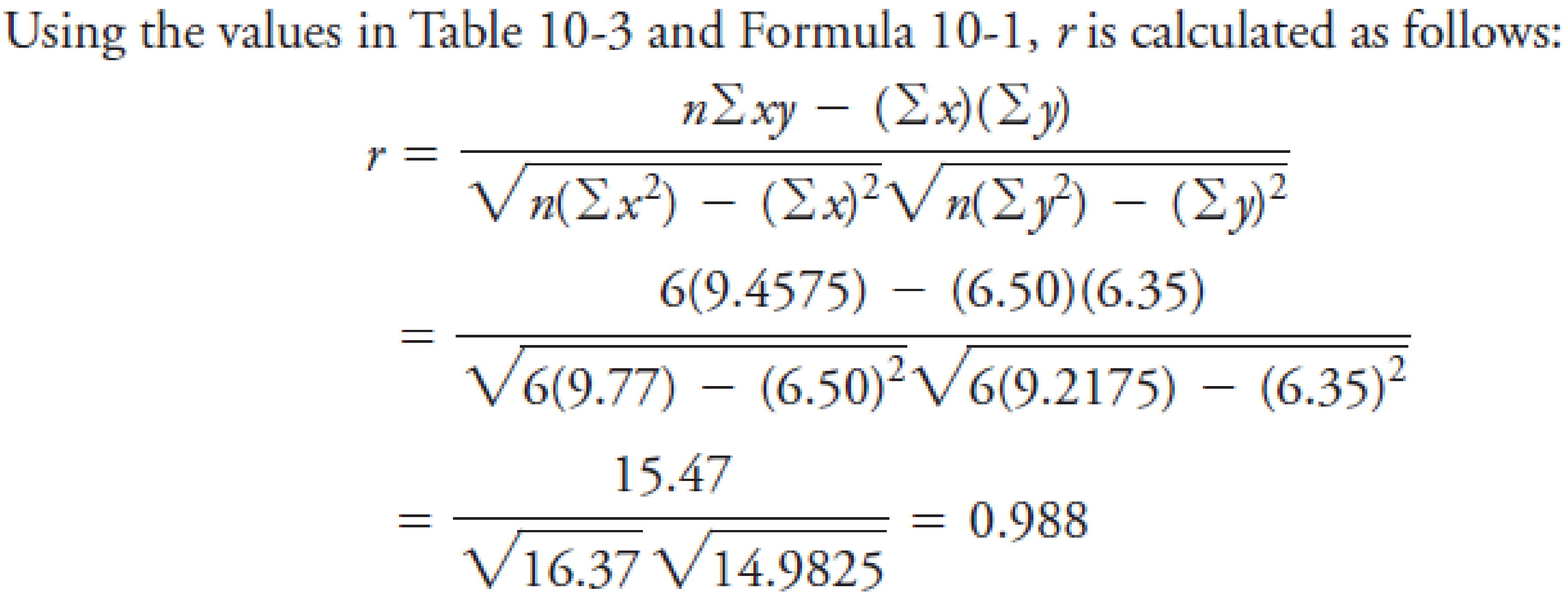


Table 10-3 Calculating *r* with Formula 10-1

x (Pizza)	*y* (Subway)	x^2	y^2	*xy*
0.15	0.15	0.0225	0.0225	0.0225
0.35	0.35	0.1225	0.1225	0.1225
1.00	1.00	1.0000	1.0000	1.0000
1.25	1.35	1.5625	1.8225	1.6875
1.75	1.50	3.0625	2.2500	2.6250
2.00	2.00	4.0000	4.0000	4.0000
$\Sigma x = 6.50$	$\Sigma y = 6.35$	$\Sigma x^2 = 9.77$	$\Sigma y^2 = 9.2175$	$\Sigma xy = 9.4575$



Using the values in Table 10-3 and Formula 10-1, r is calculated as follows:

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

$$= \frac{6(9.4575) - (6.50)(6.35)}{\sqrt{6(9.77) - (6.50)^2}\sqrt{6(9.2175) - (6.35)^2}}$$

$$= \frac{15.47}{\sqrt{16.37}\sqrt{14.9825}} = 0.988$$

Interpreting the Linear Correlation Coefficient *r*

We can base our interpretation and conclusion about correlation on a critical value from Table A-6.

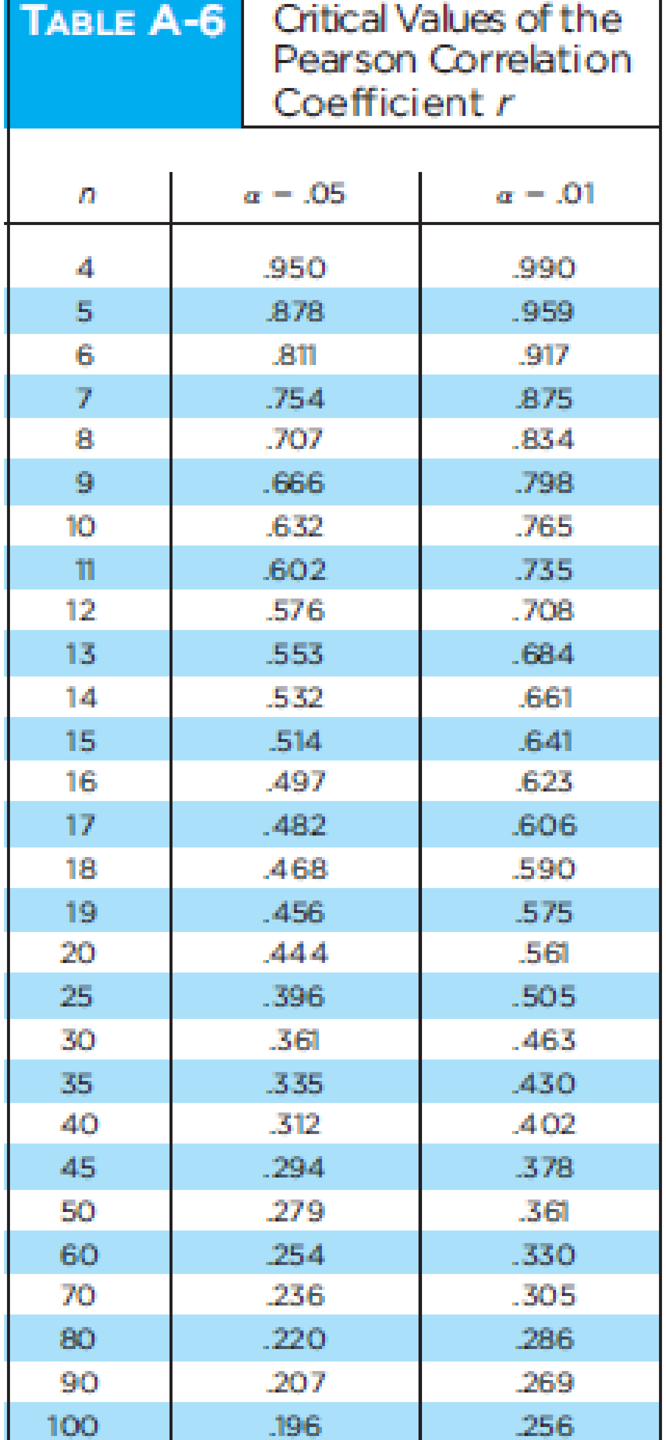


TABLE A-6 Critical Values of the Pearson Correlation Coefficient *r*

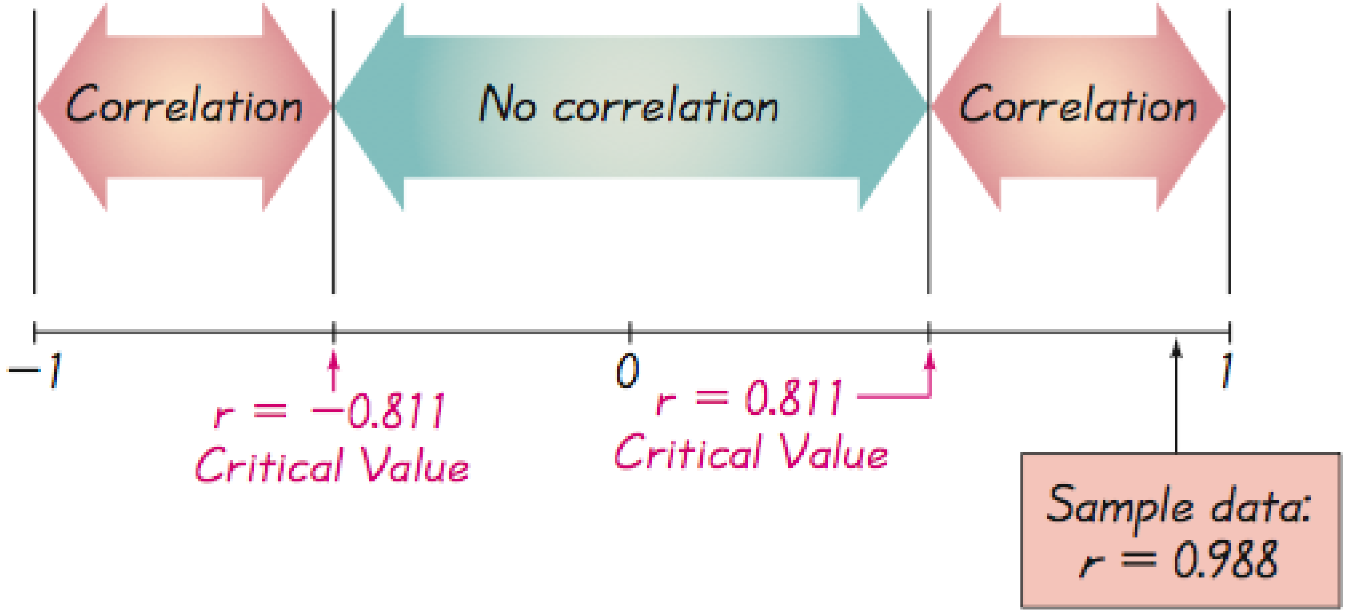
n	$\alpha = .05$	$\alpha = .01$
4	.950	.990
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286
90	.207	.269
100	.196	.256

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 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 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 α = 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**

ρ = 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.)

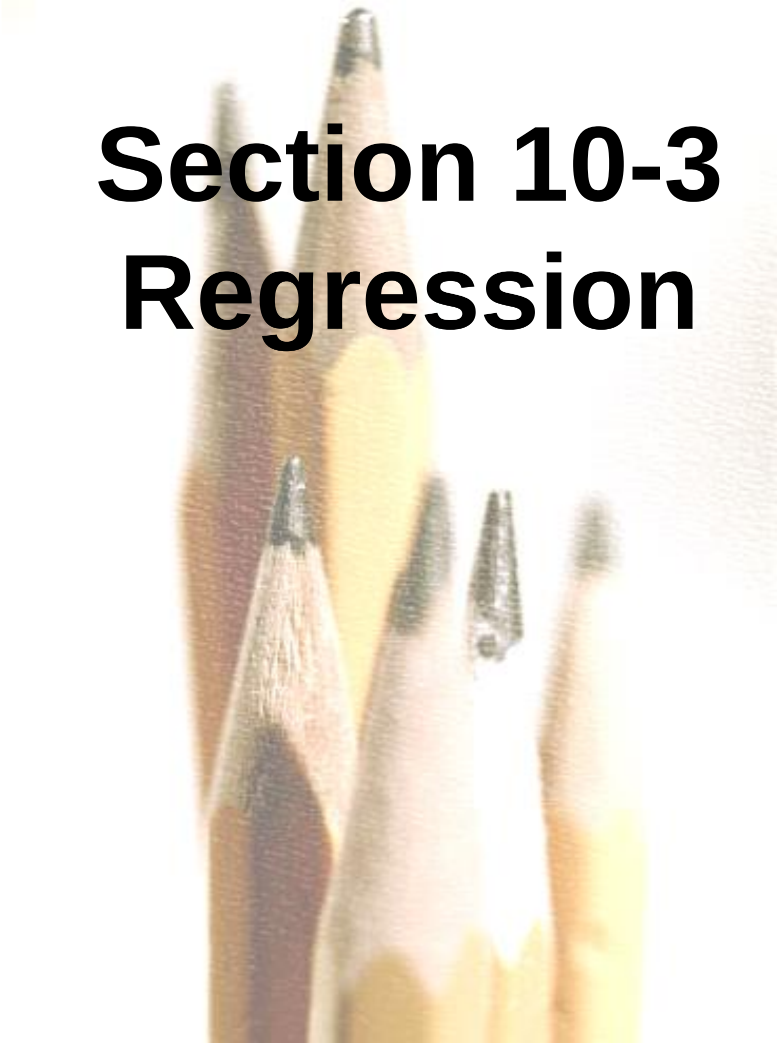
We conclude that there is sufficient evidence to support the claim of a linear correlation between costs of a slice of pizza and subway fares.

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

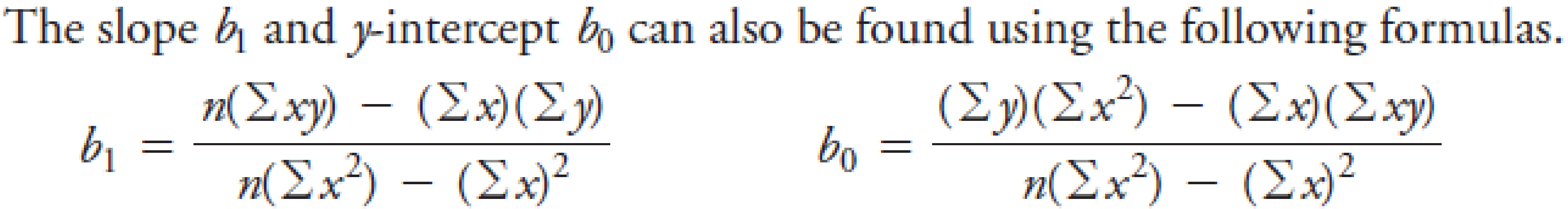
Given a collection of paired data, the regression equation

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

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). (The specific criterion used to determine which line fits "best" is the least-squares property, which will be described later.)



The slope b_1 and y-intercept b_0 can also be found using the following formulas.

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \qquad b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

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	$y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

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.

Another Formula 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)**

Where r is linear correlation coefficient, Sy is Standard deviation of y values, Sx is Standard deviation of x values. $\bar{x}$ and $\bar{y}$ are means of x values and y values, respectively

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:

Using Manual Calculations to Find the Regression Equation Refer to the sample data given in Table 10-1 in the Chapter Problem. Use Formulas 10-3 and 10-4 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.

$$b_1 = r\frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1\bar{x}$$

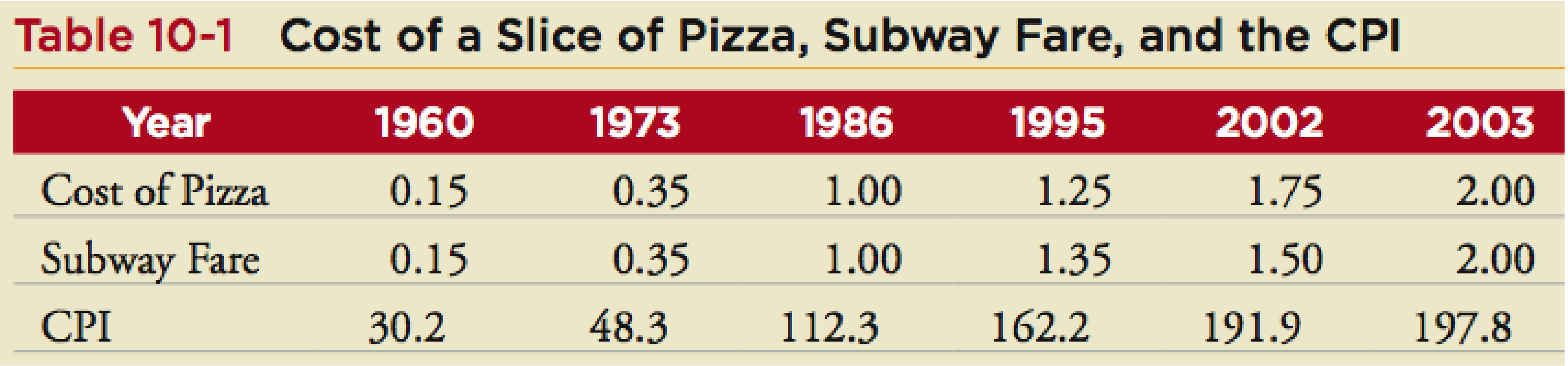
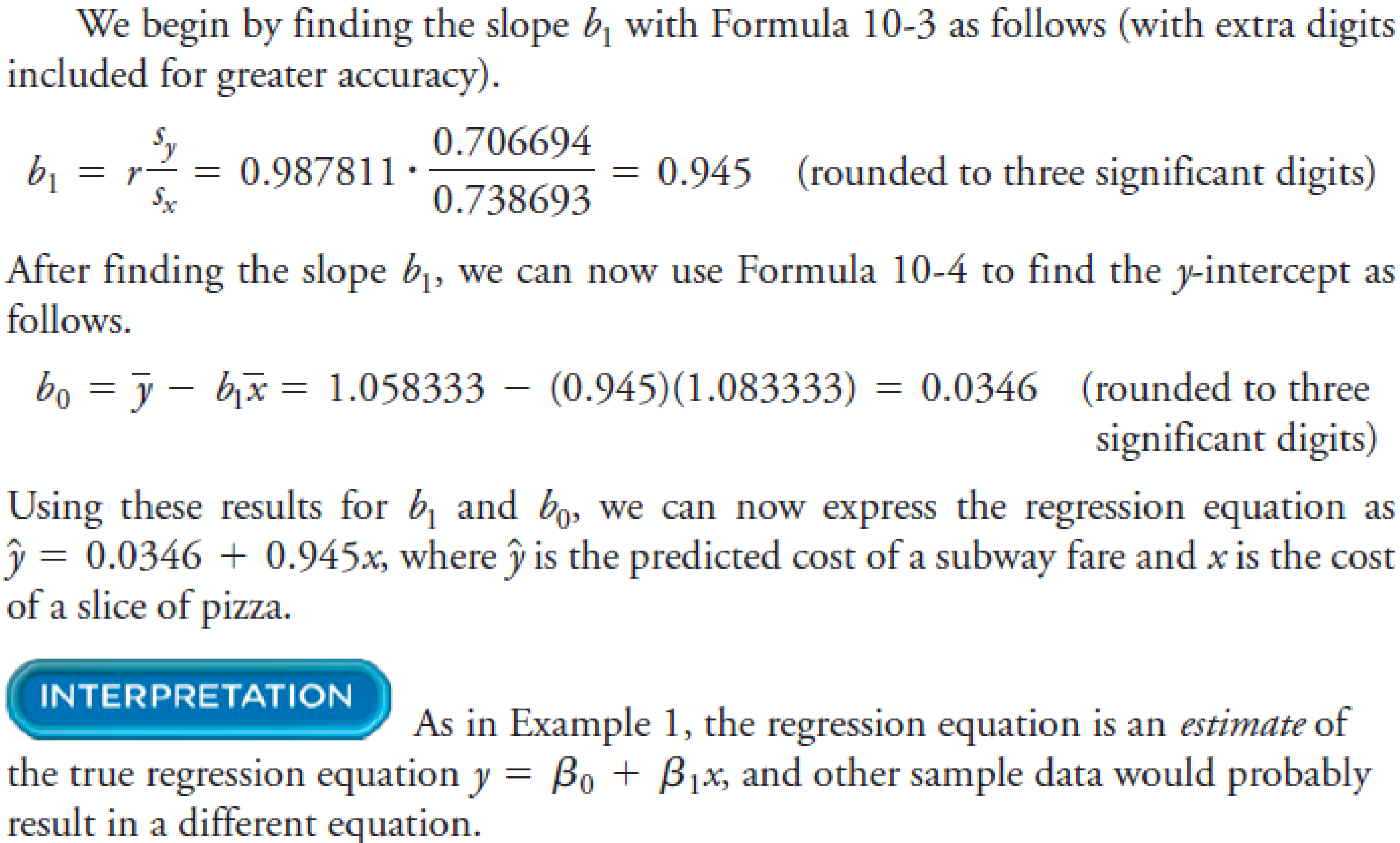


Table 10-1 Cost of a Slice of Pizza, Subway Fare, and the CPI

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:

(1) The data are assumed to be a simple random sample. (2) Figure 10-1 is a scatterplot showing a pattern of points that does appear to be a straight-line pattern. (3) There are no outliers. The requirements are satisfied.



We begin by finding the slope b_1 with Formula 10-3 as follows (with extra digits included for greater accuracy).

$$b_1 = r\frac{s_y}{s_x} = 0.987811 \cdot \frac{0.706694}{0.738693} = 0.945 \quad \text{(rounded to three significant digits)}$$

After finding the slope b_1, we can now use Formula 10-4 to find the y-intercept as follows.

$$b_0 = \bar{y} - b_1\bar{x} = 1.058333 - (0.945)(1.083333) = 0.0346 \quad \text{(rounded to three significant digits)}$$

Using these results for b_1 and b_0, we can now express the regression equation 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.

INTERPRETATION

As in Example 1, the regression equation is an *estimate* of the true regression equation $y = \beta_0 + \beta_1 x$, and other sample data would probably result in a different equation.

Example:

Hence, regression equation can be expressed as:
$\hat{y}$ = 0.0346 +0.945*x*, 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.

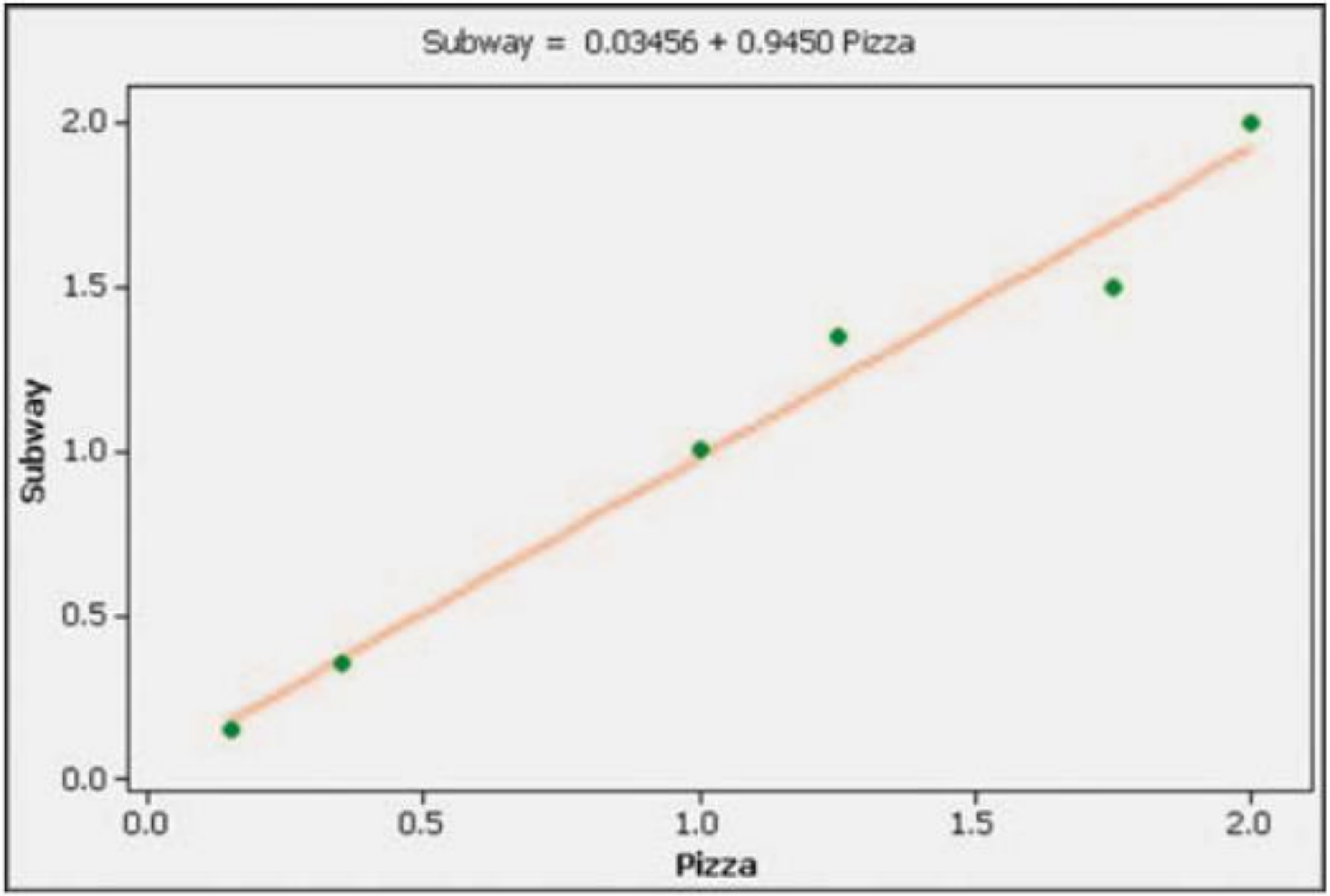
Example:

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.

Example:



We can see that the regression line fits the data quite well.

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.

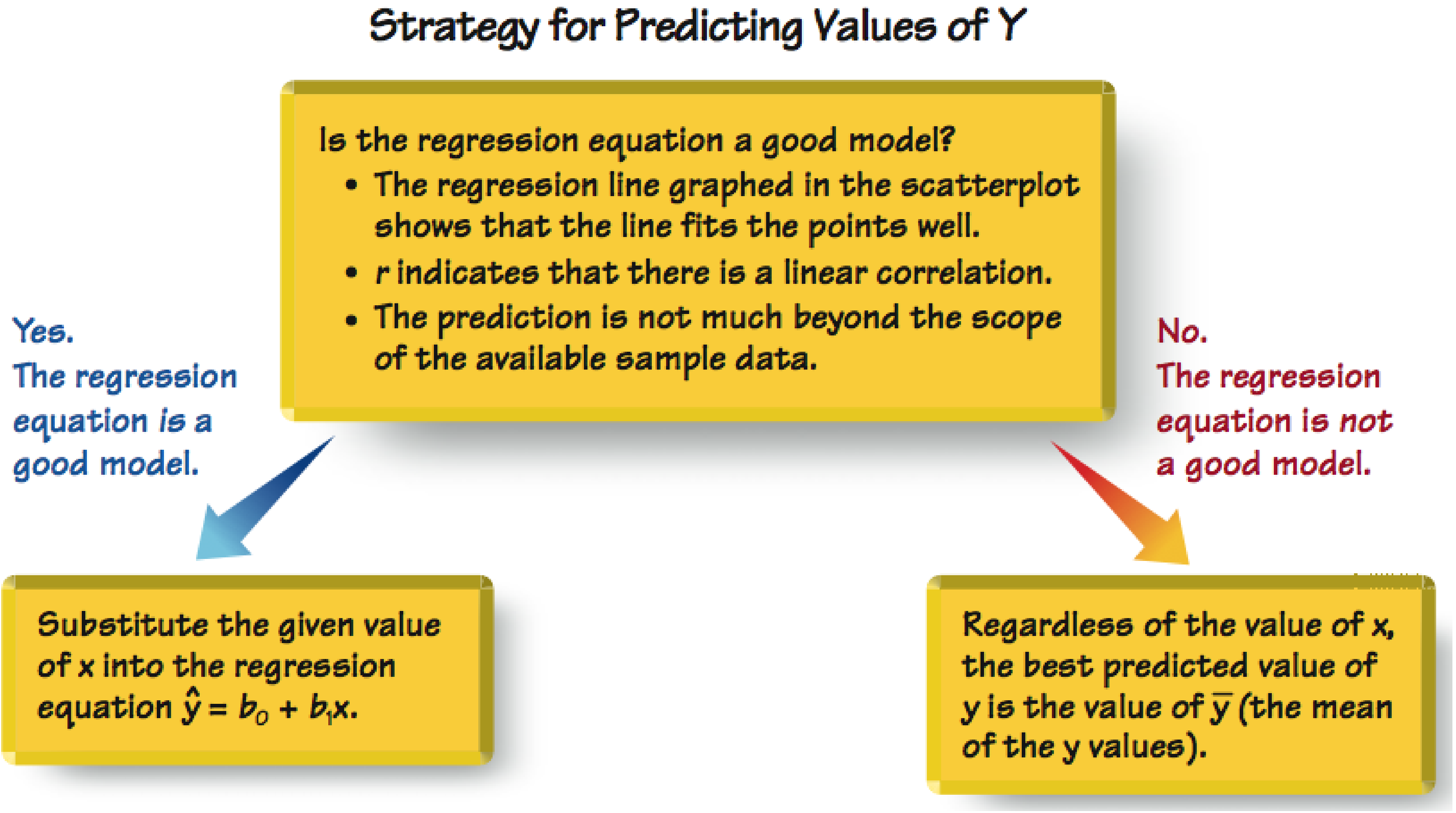
2. Use the regression equation for predictions only if the linear correlation coefficient *r* indicates that there is a linear correlation between the two variables (as described in Section 10-2).

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.

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

No.
The regression equation is *not* a good model.

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.

Example. Predicting Subway Fare Table 10-1 includes the pizza subway fare costs from the Chapter Problem. As of this writing, the cost of a slice of pizza in New York City was $2.25. Use the pizza subway fare data from Table 10-1 to predict the cost of a subway fare given that a slice of pizza costs $2.25.

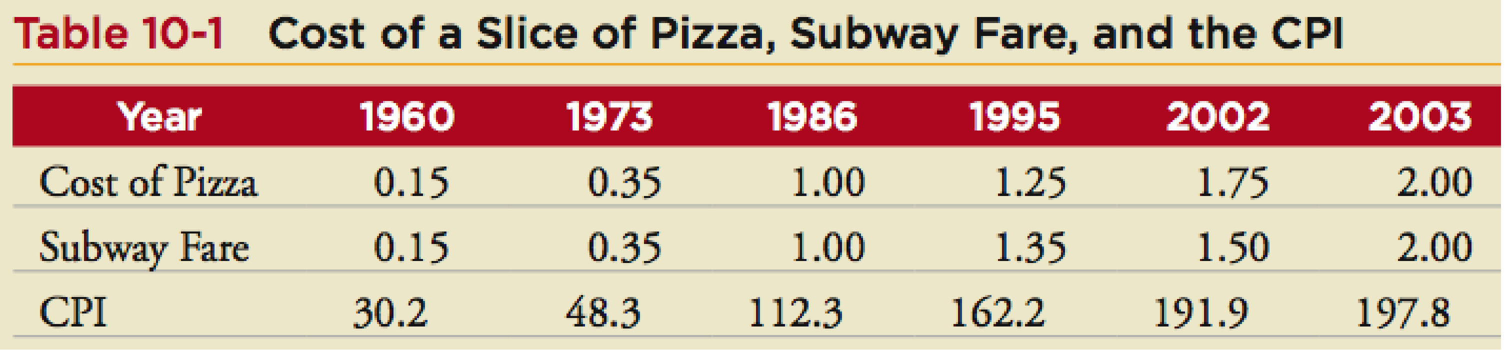


Table 10-1 Cost of a Slice of Pizza, Subway Fare, and the CPI

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

The regression equation (from the preceding Example)

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

substitute x=2.23 to get a predicted subway fare of $\grave{y}$ = $2.16

Note. Use the regression equation for predictions only if it is a good model. If the regression equation is not a good model, use the predicted value of $\bar{y}$.

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

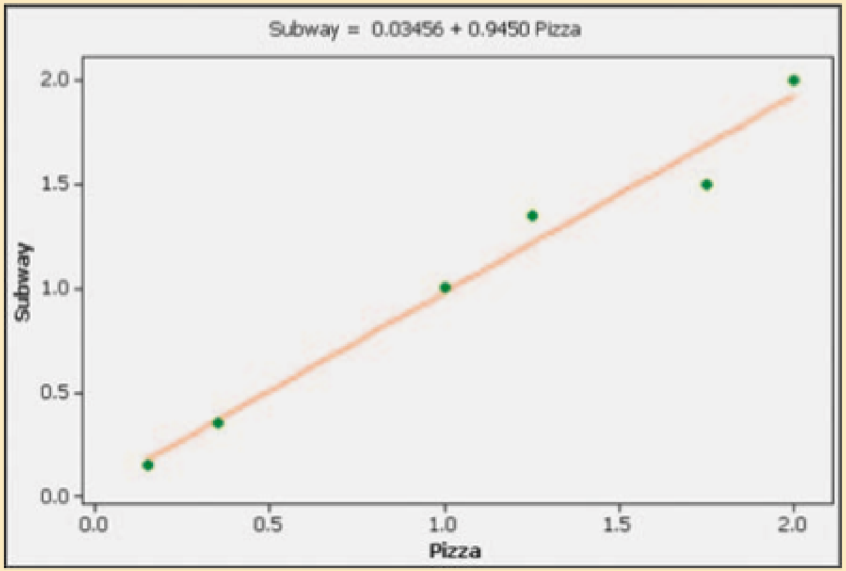
In a scatterplot, an outlier is a point lying far away from the other data points.

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

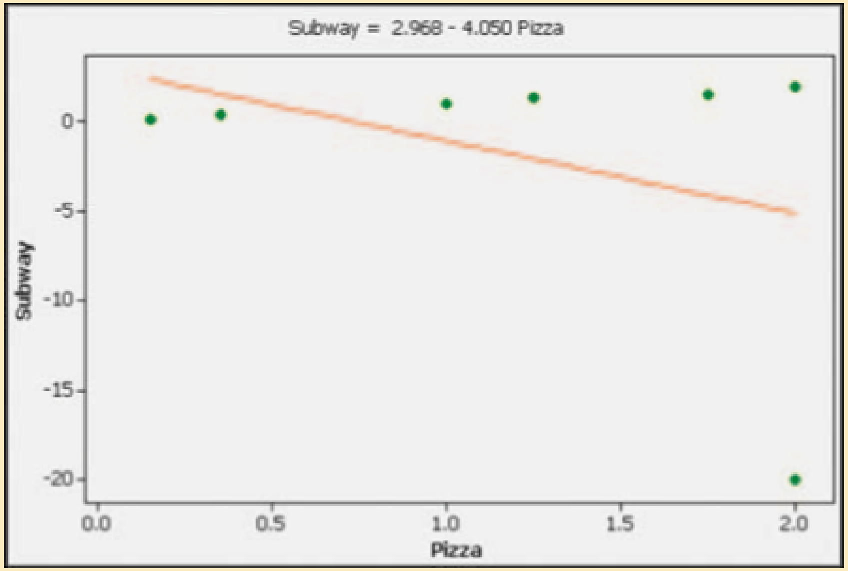
Example:

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:

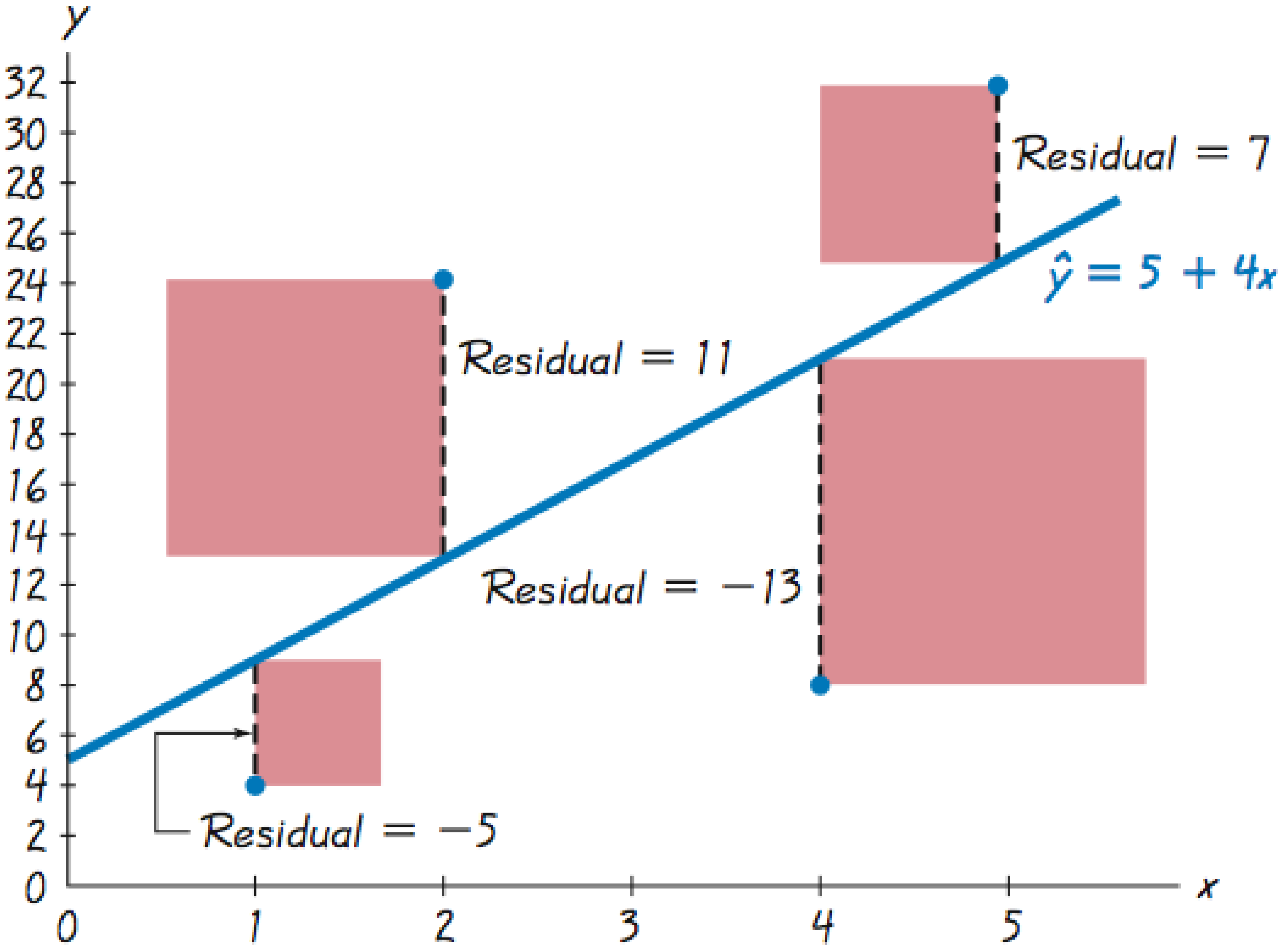
Compare the two graphs and you will see clearly that the addition of that one pair of values has a very dramatic effect on the regression line, so that additional point is an influential point. The additional point is also an outlier because it is far from the other points.

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,

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

Residuals



y
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
x
Residual = 7
$\hat{y} = 5 + 4x$
Residual = 11
Residual = −13
Residual = −5

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

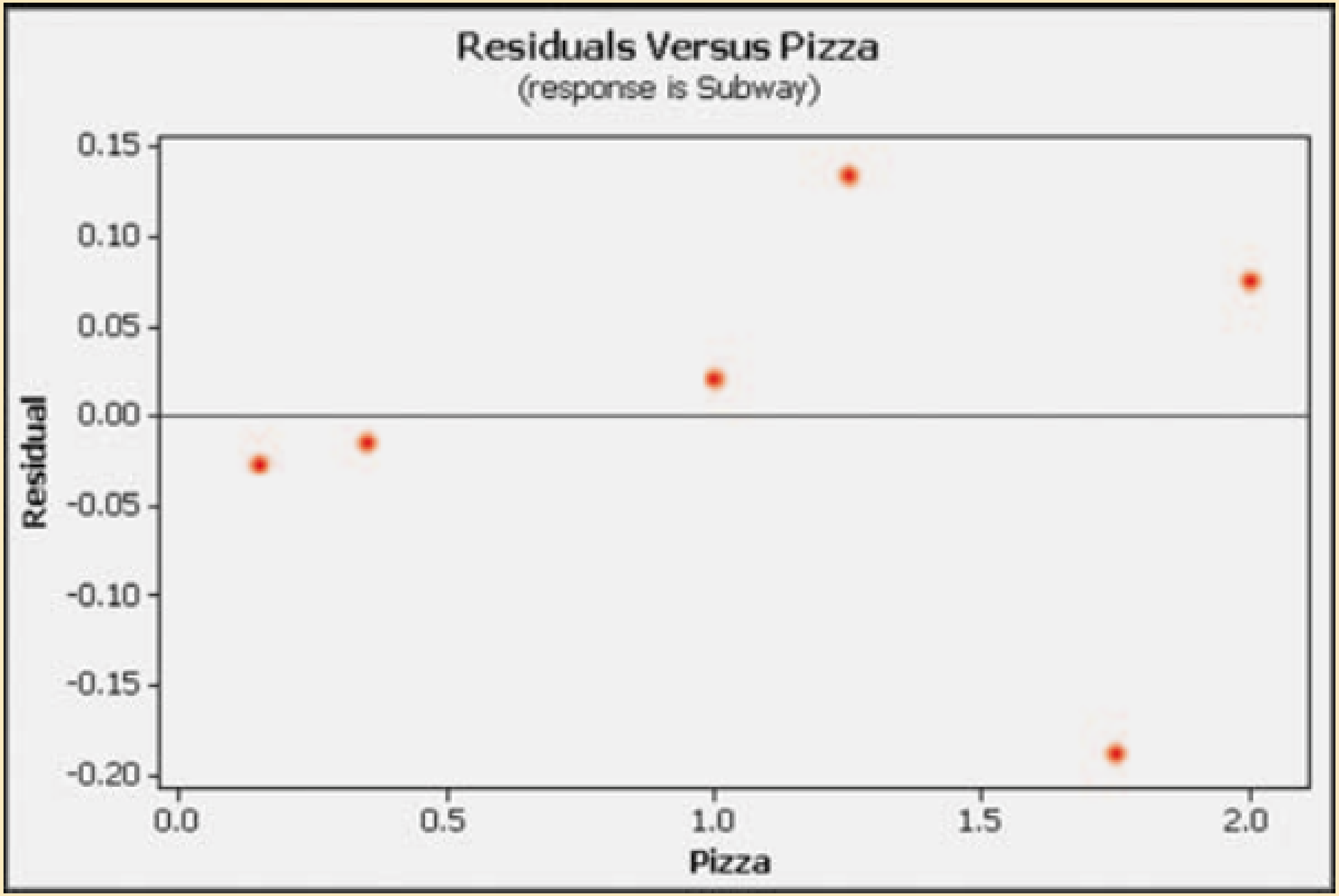
When analyzing a residual plot, look for a pattern in the way the points are configured, and use these criteria:

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.

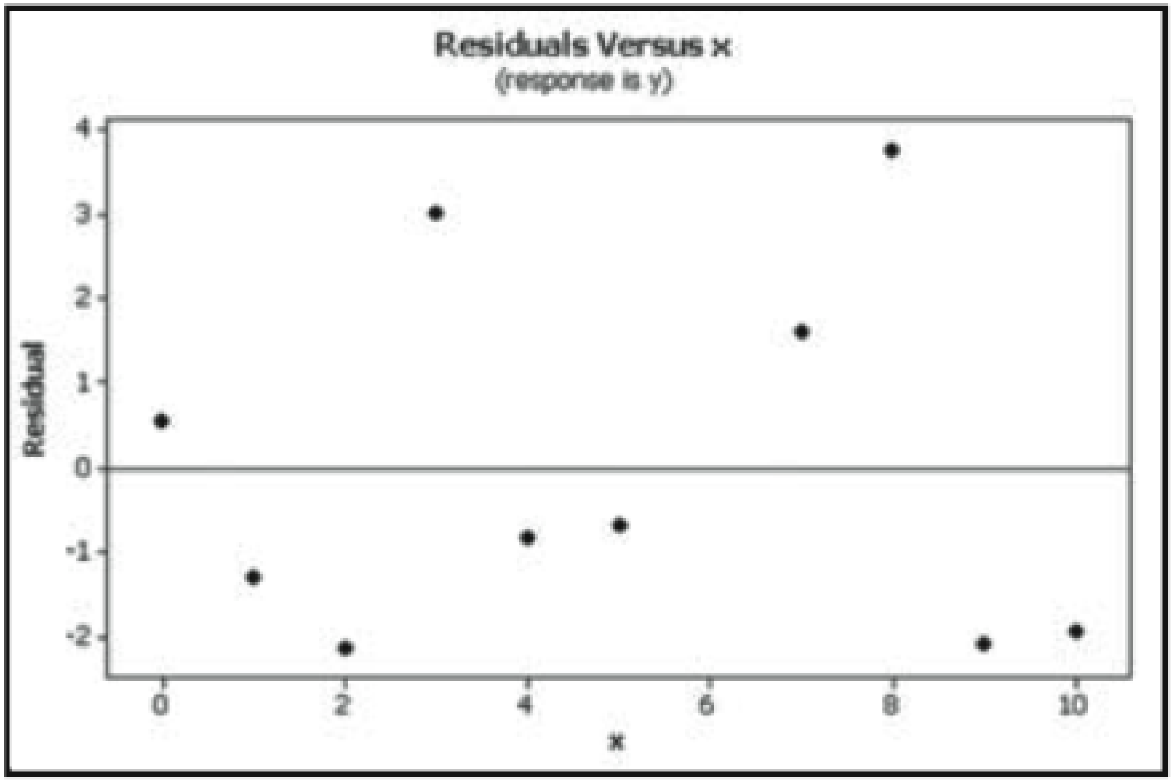
Residuals Plot - Pizza/Subway

MINITAB



Residual Plots

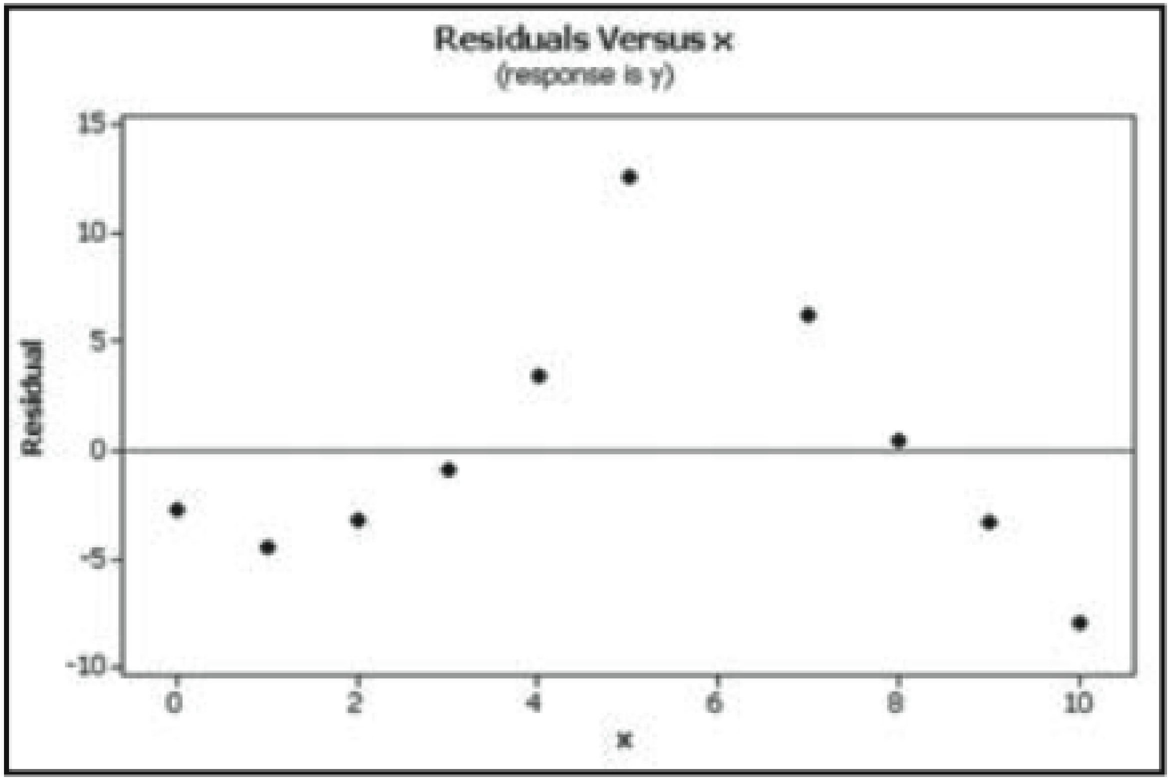
MINITAB



Residual Plot Suggesting that the Regression Equation is a Good Model

Residual Plots

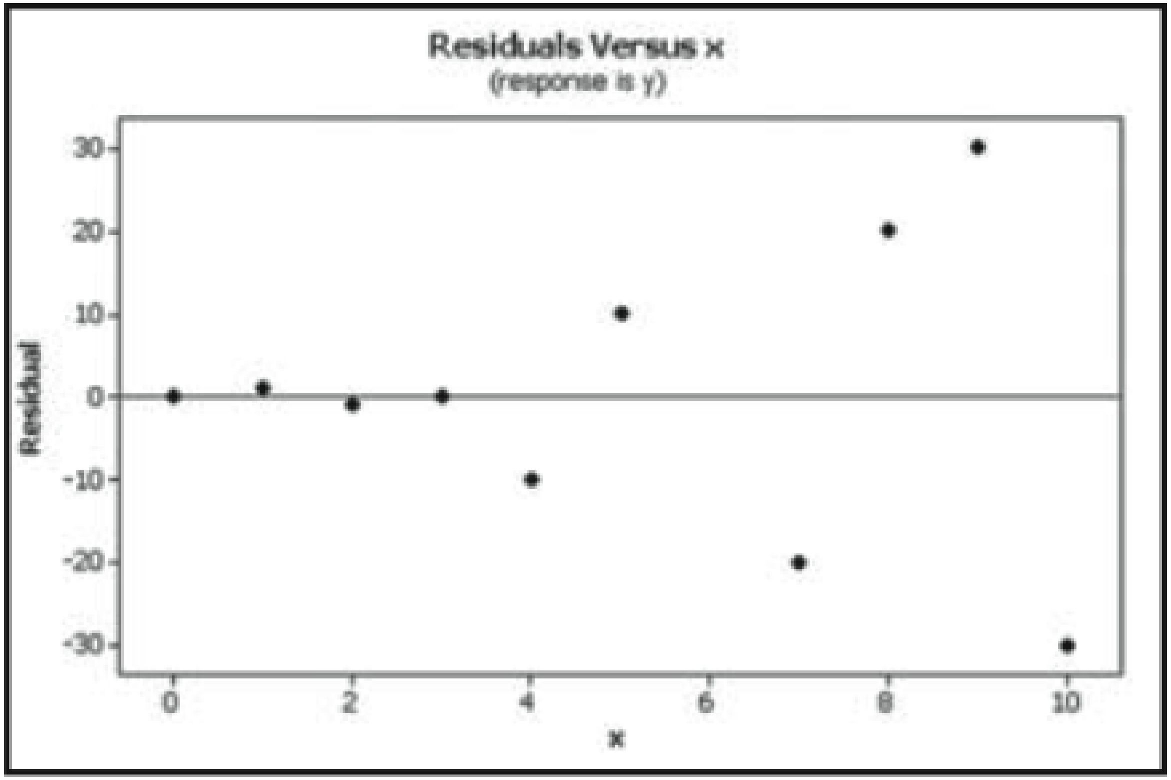
MINITAB



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.**