Chapter 9

Correlation and Regression

§ 9.1

Correlation

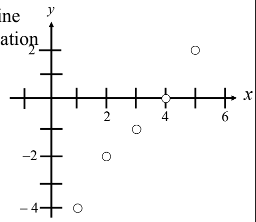
Correlation

A **correlation** is a relationship between two variables. The data can be represented by the ordered pairs (x, y) where x is the **independent** (or **explanatory**) **variable**, and y is the **dependent** (or **response**) **variable**.

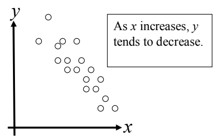
A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

Example:

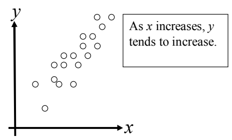
x	1	2	3	4	5
y	– 4	– 2	– 1	0	2



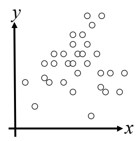
Linear Correlation



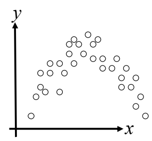
Negative Linear Correlation



Positive Linear Correlation



No Correlation



Nonlinear Correlation

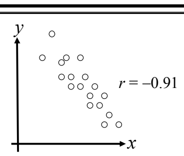
Correlation Coefficient

The **correlation coefficient** is a measure of the strength and the direction of a linear relationship between two variables. The symbol r represents the sample correlation coefficient. The formula for r is

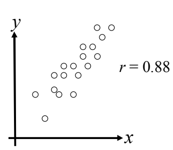
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$$

The range of the correlation coefficient is –1 to 1. If x and y have a strong positive linear correlation, r is close to 1. If x and y have a strong negative linear correlation, r is close to –1. If there is no linear correlation or a weak linear correlation, r is close to 0.

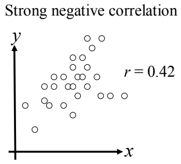
Linear Correlation



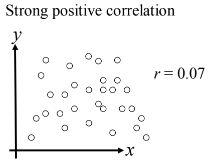
Strong negative correlation



Strong positive correlation



Weak positive correlation



Nonlinear Correlation

Calculating a Correlation Coefficient

Calculating a Correlation Coefficient

In Words	*In Symbols*
1. Find the sum of the x-values.	$\sum x$
2. Find the sum of the y-values.	$\sum y$
3. Multiply each x-value by its corresponding y-value and find the sum.	$\sum xy$
4. Square each x-value and find the sum.	$\sum x^2$
5. Square each y-value and find the sum.	$\sum y^2$
6. Use these five sums to calculate the correlation coefficient.	$r = \dfrac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$.

Continued.

Correlation Coefficient

Example:
Calculate the correlation coefficient r for the following data.

x	y	xy	x^2	y^2
1	− 3	− 3	1	9
2	− 1	− 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{5(9) - (15)(-1)}{\sqrt{5(55) - 15^2}\sqrt{5(15) - (-1)^2}}$$

$$= \frac{60}{\sqrt{50}\sqrt{74}} \approx 0.986$$

There is a strong positive linear correlation between x and y.

Correlation Coefficient

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.
a.) Display the scatter plot.
b.) Calculate the correlation coefficient r.

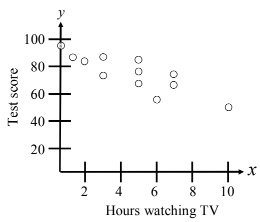
Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Continued.

Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50



Continued.

Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$\Sigma x = 54$ $\Sigma y = 908$ $\Sigma xy = 3724$ $\Sigma x^2 = 332$ $\Sigma y^2 = 70836$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{12(3724) - (54)(908)}{\sqrt{12(332) - 54^2}\sqrt{12(70836) - (908)^2}} \approx -0.831$$

There is a strong negative linear correlation.

As the number of hours spent watching TV increases, the test scores tend to decrease.

Testing a Population Correlation Coefficient

Once the sample correlation coefficient r has been calculated, we need to determine whether there is enough evidence to decide that the population correlation coefficient ρ is significant at a specified level of significance.

One way to determine this is to use Table 11 in Appendix B.

If $|r|$ is greater than the critical value, there is enough evidence to decide that the correlation coefficient ρ is significant.

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

For a sample of size $n = 6$, ρ is significant at the 5% significance level, if $|r| > 0.811$.

Testing a Population Correlation Coefficient

Finding the Correlation Coefficient ρ

In Words	*In Symbols*
1. Determine the number of pairs of data in the sample.	Determine n.
2. Specify the level of significance.	Identify α.
3. Find the critical value.	Use Table 11 in Appendix B.
4. Decide if the correlation is significant.	If $\|r\|$ > critical value, the correlation is significant. Otherwise, there is not enough evidence to support that the correlation is significant.
5. Interpret the decision in the context of the original claim.	

Testing a Population Correlation Coefficient

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

The correlation coefficient $r \approx -0.831$.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Is the correlation coefficient significant at $\alpha = 0.01$?

Continued.

Testing a Population Correlation Coefficient

Example continued:

$r \approx -0.831$

$n = 12$

$\alpha = 0.01$

Appendix B: Table 11

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684

$|r| > 0.708$

Because, the population correlation is significant, there is enough evidence at the 1% level of significance to conclude that there is a significant linear correlation between the number of hours of television watched during the weekend and the scores of each student who took a test the following Monday.

Hypothesis Testing for ρ

A hypothesis test can also be used to determine whether the sample correlation coefficient r provides enough evidence to conclude that the population correlation coefficient ρ is significant at a specified level of significance.

A hypothesis test can be one tailed or two tailed.

$\begin{cases} H_0: \rho \geq 0 \text{ (no significant negative correlation)} \\ H_a: \rho < 0 \text{ (significant negative correlation)} \end{cases}$ Left-tailed test

$\begin{cases} H_0: \rho \leq 0 \text{ (no significant positive correlation)} \\ H_a: \rho > 0 \text{ (significant positive correlation)} \end{cases}$ Right-tailed test

$\begin{cases} H_0: \rho = 0 \text{ (no significant correlation)} \\ H_a: \rho \neq 0 \text{ (significant correlation)} \end{cases}$ Two-tailed test

Hypothesis Testing for ρ

The *t*-Test for the Correlation Coefficient

A ***t*-test** can be used to test whether the correlation between two variables is significant. The **test statistic** is r and the **standardized test statistic**

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

follows a t-distribution with $n - 2$ degrees of freedom.

In this text, only two-tailed hypothesis tests for ρ are considered.

Hypothesis Testing for ρ

Using the *t*-Test for the Correlation Coefficient ρ

In Words	*In Symbols*
1. State the null and alternative hypothesis.	State H_0 and H_a.
2. Specify the level of significance.	Identify α.
3. Identify the degrees of freedom.	d.f. = $n - 2$
4. Determine the critical value(s) and rejection region(s).	Use Table 5 in Appendix B.

Hypothesis Testing for *ρ*

Using the *t*-Test for the Correlation Coefficient *ρ*

In Words	*In Symbols*
5. Find the standardized test statistic.	$t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$
6. Make a decision to reject or fail to reject the null hypothesis.	If t is in the rejection region, reject H_0. Otherwise fail to reject H_0.
7. Interpret the decision in the context of the original claim.	

Hypothesis Testing for *ρ*

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

The correlation coefficient $r \approx -0.831$.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Test the significance of this correlation coefficient significant at $\alpha = 0.01$?

Continued.

Hypothesis Testing for *ρ*

Example continued:

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

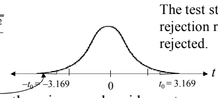
The level of significance is $\alpha = 0.01$.

Degrees of freedom are d.f. = 12 – 2 = 10.

The critical values are $-t_0 = -3.169$ and $t_0 = 3.169$.

The standardized test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.831}{\sqrt{\frac{1-(-0.831)^2}{12-2}}} \approx -4.72.$$



The test statistic falls in the rejection region, so H_0 is rejected.

At the 1% level of significance, there is enough evidence to conclude that there is a significant linear correlation between the number of hours of TV watched over the weekend and the test scores on Monday morning.

Correlation and Causation

The fact that two variables are strongly correlated does not in itself imply a cause-and-effect relationship between the variables.

If there is a significant correlation between two variables, you should consider the following possibilities.

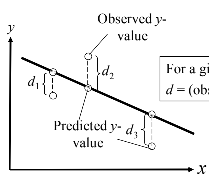
1. **Is there a direct cause-and-effect relationship between the variables?**
   Does x cause y?
2. **Is there a reverse cause-and-effect relationship between the variables?**
   Does y cause x?
3. **Is it possible that the relationship between the variables can be caused by a third variable or by a combination of several other variables?**
4. **Is it possible that the relationship between two variables may be a coincidence?**

§ 9.2

Linear Regression

Residuals

After verifying that the linear correlation between two variables is significant, next we determine the equation of the line that can be used to predict the value of y for a given value of x.



For a given x-value,
d = (observed y-value) – (predicted y-value)

Each data point d_i represents the difference between the observed y-value and the predicted y-value for a given x-value on the line. These differences are called **residuals**.

Regression Line

A **regression line**, also called a **line of best fit**, is the line for which the sum of the squares of the residuals is a minimum.

The Equation of a Regression Line

The equation of a regression line for an independent variable x and a dependent variable y is

$$\hat{y} = mx + b$$

where $\hat{y}$ is the predicted y-value for a given x-value. The slope m and y-intercept b are given by

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \text{ and } b = \bar{y} - m\bar{x} = \frac{\sum y}{n} - m\frac{\sum x}{n}$$

where $\bar{y}$ is the mean of the y-values and $\bar{x}$ is the mean of the x-values. The regression line always passes through $(\bar{x}, \bar{y})$.

Regression Line

Example:

Find the equation of the regression line.

x	y	xy	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{5(9) - (15)(-1)}{5(55) - (15)^2} = \frac{60}{50} = 1.2$$

Continued.

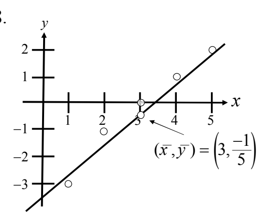
Regression Line

Example continued:

$$b = \bar{y} - m\bar{x} = \frac{-1}{5} - (1.2)\frac{15}{5} = -3.8$$

The equation of the regression line is

$\hat{y} = 1.2x - 3.8.$



Regression Line

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Find the equation of the regression line.

b.) Use the equation to find the expected test score for a student who watches 9 hours of TV.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$\Sigma x = 54$ $\quad \Sigma y = 908$ $\quad \Sigma xy = 3724$ $\quad \Sigma x^2 = 332$ $\quad \Sigma y^2 = 70836$

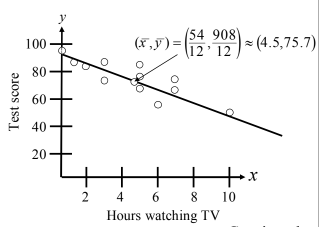
Regression Line

Example continued:

$$m = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{12(3724) - (54)(908)}{12(332) - (54)^2} \approx -4.067$$

$$b = \bar{y} - m\bar{x}$$
$$= \frac{908}{12} - (-4.067)\frac{54}{12}$$
$$\approx 93.97$$

$\hat{y} = -4.07x + 93.97$



Continued.

Regression Line

Example continued:

Using the equation $\hat{y} = -4.07x + 93.97$, we can predict the test score for a student who watches 9 hours of TV.

$\hat{y} = -4.07x + 93.97$
$= -4.07(9) + 93.97$
$= 57.34$

A student who watches 9 hours of TV over the weekend can expect to receive about a 57.34 on Monday's test.

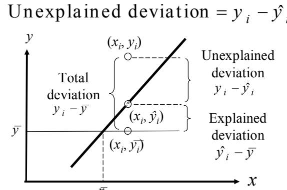
§ 9.3

Measures of Regression and Prediction Intervals

Variation About a Regression Line

To find the total variation, you must first calculate the **total deviation**, the **explained deviation**, and the **unexplained deviation**.

$$\text{Total deviation} = y_i - \bar{y}$$
$$\text{Explained deviation} = \hat{y}_i - \bar{y}$$
$$\text{Unexplained deviation} = y_i - \hat{y}_i$$



Variation About a Regression Line

The **total variation** about a regression line is the sum of the squares of the differences between the *y*-value of each ordered pair and the mean of *y*.

$$\text{Total variation} = \Sigma(y_i - \bar{y})^2$$

The **explained variation** is the sum of the squares of the differences between each predicted *y*-value and the mean of *y*.

$$\text{Explained variation} = \Sigma(\hat{y}_i - \bar{y})^2$$

The **unexplained variation** is the sum of the squares of the differences between the *y*-value of each ordered pair and each corresponding predicted *y*-value.

$$\text{Unexplained variation} = \Sigma(y_i - \hat{y}_i)^2$$

$$\text{Total variation} = \text{Explained variation} + \text{Unexplained variation}$$

Coefficient of Determination

The **coefficient of determination r^2** is the ratio of the explained variation to the total variation. That is,

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Example:
The correlation coefficient for the data that represents the number of hours students watched television and the test scores of each student is $r \approx -0.831$. Find the coefficient of determination.

$$r^2 \approx (-0.831)^2$$
$$\approx 0.691$$

About 69.1% of the variation in the test scores can be explained by the variation in the hours of TV watched. About 30.9% of the variation is unexplained.

The Standard Error of Estimate

When a $\hat{y}$-value is predicted from an x-value, the prediction is a point estimate.

An interval can also be constructed.

The **standard error of estimate s_e** is the standard deviation of the observed y_i-values about the predicted $\hat{y}$-value for a given x_i-value. It is given by

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

where n is the number of ordered pairs in the data set.

The closer the observed y-values are to the predicted y-values, the smaller the standard error of estimate will be.

The Standard Error of Estimate

Finding the Standard Error of Estimate

In Words	*In Symbols*
1. Make a table that includes the column heading shown.	x_i, y_i, $\hat{y}_i$, $(y_i - \hat{y}_i)$, $(y_i - \hat{y}_i)^2$
2. Use the regression equation to calculate the predicted y-values.	$\hat{y} = mx_i + b$
3. Calculate the sum of the squares of the differences between each observed y-value and the corresponding predicted y-value.	$\Sigma(y_i - \hat{y}_i)^2$
4. Find the standard error of estimate.	$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}$

The Standard Error of Estimate

Example:
The regression equation for the following data is
$\hat{y} = 1.2x - 3.8.$
Find the standard error of estimate.

x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	– 3	– 2.6	0.16
2	– 1	– 1.4	0.16
3	0	– 0.2	0.04
4	1	1	0
5	2	2.2	0.04
			$\Sigma = 0.4$

Unexplained variation

$$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{0.4}{5-2}} \approx 0.365$$

The standard deviation of the predicted y value for a given x value is about 0.365.

The Standard Error of Estimate

Example:
The regression equation for the data that represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday is
$\hat{y} = -4.07x + 93.97.$
Find the standard error of estimate.

Hours, x_i	0	1	2	3	3	5
Test score, y_i	96	85	82	74	95	68
$\hat{y}_i$	93.97	89.9	85.83	81.76	81.76	73.62
$(y_i - \hat{y}_i)^2$	4.12	24.01	14.67	60.22	175.3	31.58

Hours, x_i	5	5	6	7	7	10
Test score, y_i	76	84	58	65	75	50
$\hat{y}_i$	73.62	73.62	69.55	65.48	65.48	53.27
$(y_i - \hat{y}_i)^2$	5.66	107.74	133.4	0.23	90.63	10.69

Continued.

The Standard Error of Estimate

Example continued:

$$\Sigma(y_i - \hat{y}_i)^2 = 658.25$$

Unexplained variation

$$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{658.25}{12-2}} \approx 8.11$$

The standard deviation of the student test scores for a specific number of hours of TV watched is about 8.11.

Prediction Intervals

Two variables have a **bivariate normal distribution** if for any fixed value of x, the corresponding values of y are normally distributed and for any fixed values of y, the corresponding x-values are normally distributed.

A prediction interval can be constructed for the true value of y.

Given a linear regression equation $\hat{y} = mx + b$ and x_0, a specific value of x, a ***c*-prediction interval** for y is

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

where

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

The point estimate is $\hat{y}$ and the margin of error is E. The probability that the prediction interval contains y is c.

Prediction Intervals

Construct a Prediction Interval for *y* for a Specific Value of *x*

In Words	*In Symbols*
1. Identify the number of ordered pairs in the data set n and the degrees of freedom.	$\text{d.f.} = n - 2$
2. Use the regression equation and the given x-value to find the point estimate $\hat{y}$.	$\hat{y} = mx_i + b$
3. Find the critical value t_c that corresponds to the given level of confidence c.	Use Table 5 in Appendix B.

Continued.

Prediction Intervals

Construct a Prediction Interval for *y* for a Specific Value of *x*

In Words	*In Symbols*
4. Find the standard error of estimate s_e.	$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$
5. Find the margin of error E.	$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - (\sum x)^2}}$
6. Find the left and right endpoints and form the prediction interval.	Left endpoint: $\hat{y} - E$ Right endpoint: $\hat{y} + E$ Interval: $\hat{y} - E < y < \hat{y} + E$

Prediction Intervals

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

$\hat{y} = -4.07x + 93.97$ $s_e \approx 8.11$

Construct a 95% prediction interval for the test scores when 4 hours of TV are watched.

Continued.

Prediction Intervals

Example continued:
Construct a 95% prediction interval for the test scores when the number of hours of TV watched is 4.

There are $n - 2 = 12 - 2 = 10$ degrees of freedom.

The point estimate is

$\hat{y} = -4.07x + 93.97 \quad = -4.07(4) + 93.97 \quad = 77.69.$

The critical value $t_c = 2.228$, and $s_e = 8.11$.

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

$77.69 - 8.11 = 69.58$ $77.69 + 8.11 = 85.8$

You can be 95% confident that when a student watches 4 hours of TV over the weekend, the student's test grade will be between 69.58 and 85.8.

§ 9.4

Multiple Regression

Multiple Regression Equation

In many instances, a better prediction can be found for a dependent (response) variable by using more than one independent (explanatory) variable.

For example, a more accurate prediction of Monday's test grade from the previous section might be made by considering the number of other classes a student is taking as well as the student's previous knowledge of the test material.

A **multiple regression equation** has the form

$$\hat{y} = b + m_1x_1 + m_2x_2 + m_3x_3 + \ldots + m_kx_k$$

where $x_1, x_2, x_3, \ldots, x_k$ are independent variables, b is the y-intercept, and y is the dependent variable.

* Because the mathematics associated with this concept is complicated, technology is generally used to calculate the multiple regression equation.

Predicting *y*-Values

After finding the equation of the multiple regression line, you can use the equation to predict y-values over the range of the data.

Example:
The following multiple regression equation can be used to predict the annual U.S. rice yield (in pounds).

$$\hat{y} = 859 + 5.76x_1 + 3.82x_2$$

where x_1 is the number of acres planted (in thousands), and x_2 is the number of acres harvested (in thousands). *(Source: U.S. National Agricultural Statistics Service)*

a.) Predict the annual rice yield when $x_1 = 2758$, and $x_2 = 2714$.

b.) Predict the annual rice yield when $x_1 = 3581$, and $x_2 = 3021$.

Continued.

Predicting *y*-Values

Example continued:

a.) $\hat{y} = 859 + 5.76x_1 + 3.82x_2$

$= 859 + 5.76(2758) + 3.82(2714)$

$= 27{,}112.56$

The predicted annual rice yield is 27,1125.56 pounds.

b.) $\hat{y} = 859 + 5.76x_1 + 3.82x_2$

$= 859 + 5.76(3581) + 3.82(3021)$

$= 33{,}025.78$

The predicted annual rice yield is 33,025.78 pounds.