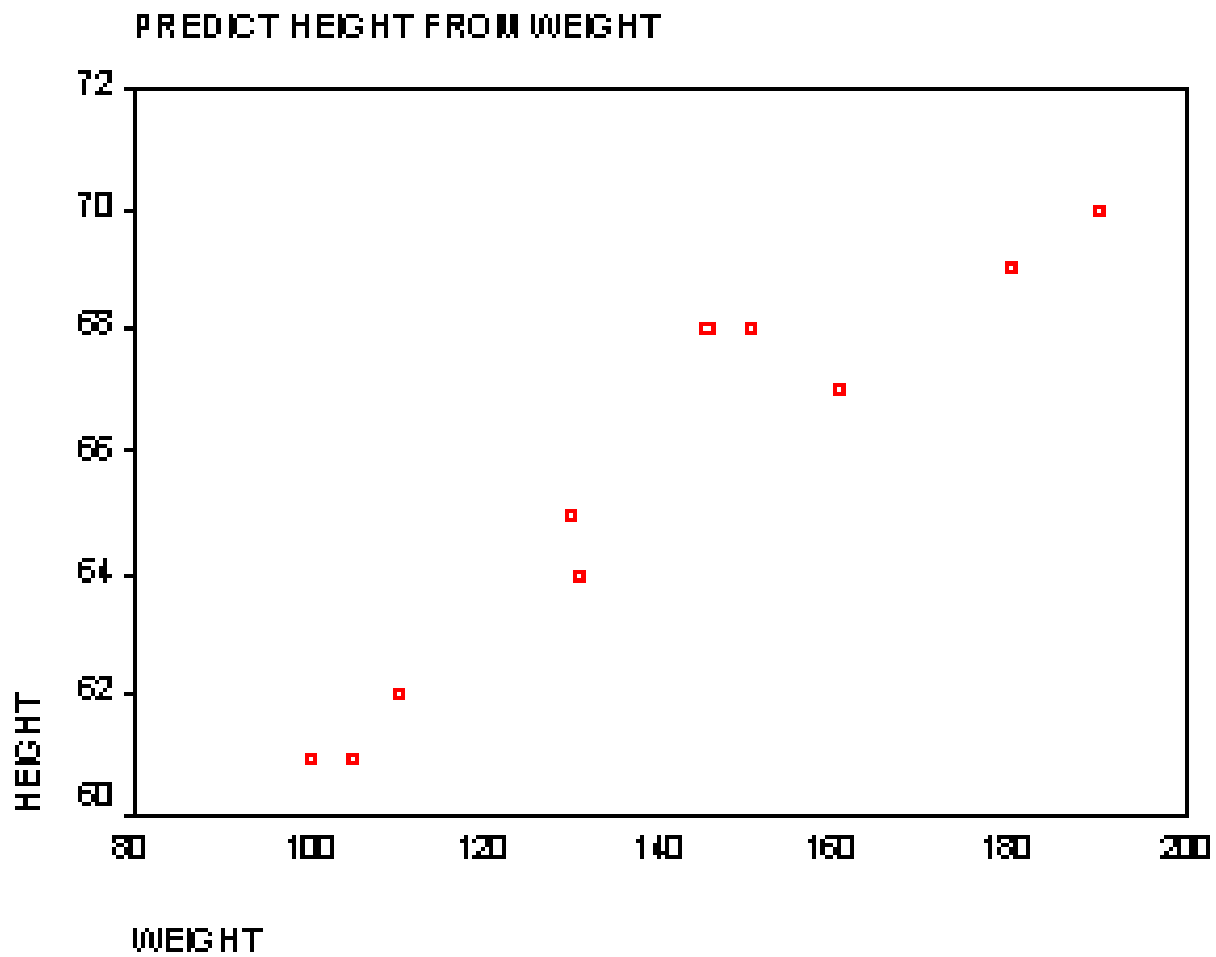



# LINEAR REGRESSION

# SLR: Simple Linear Regression :

- ▶ **Simple Linear Regression**
- ▶ Simple linear regression is when you want to predict values of one variable, given values of another variable. For example, you might want to predict a person's height (in inches) from his weight (in pounds).
- ▶ Imagine a sample of ten people for whom you know their height and weight. You could plot the values on a graph, with weight on the x axis and height on the y axis.

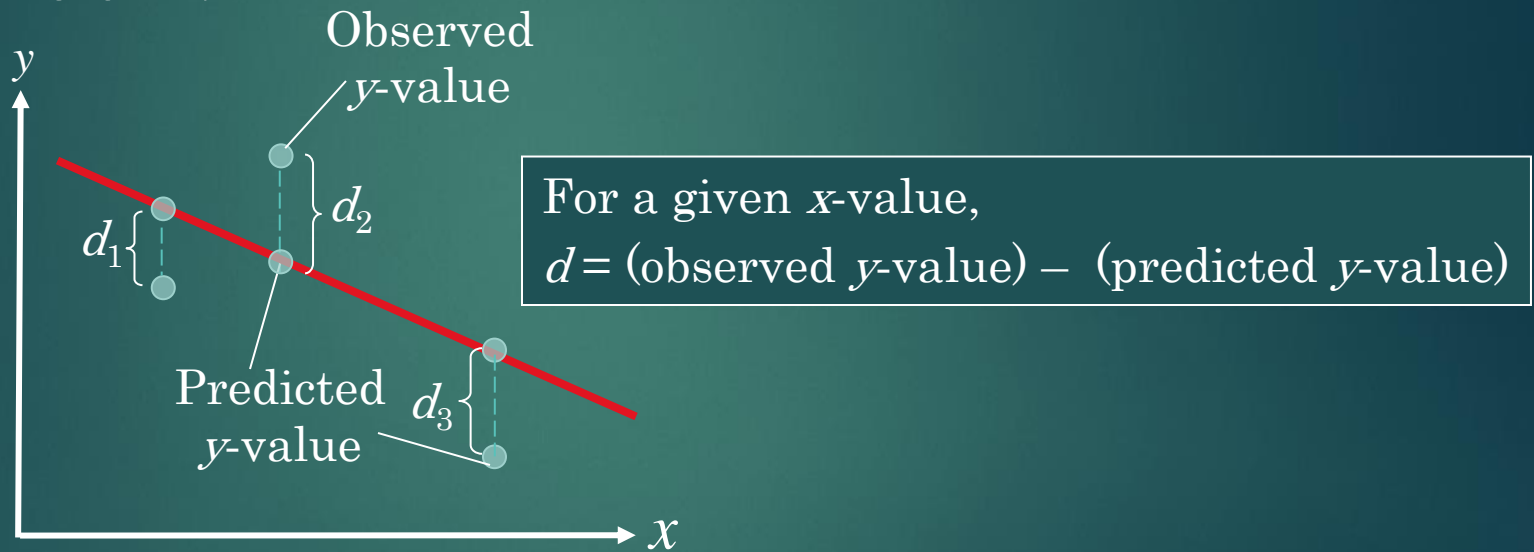
- If there were a perfect linear relationship between height and weight, then all 10 points on the graph would fit on a straight line. But, this is never the case (unless your data are rigged). If there is a (nonperfect) linear relationship between height and weight (presumably a positive one), then you would get a cluster of points on the graph which slopes upward. In other words, people who weigh a lot should be taller than those people who are of less weight. (See graph.)



- 
- ▶ The purpose of regression analysis is to come up with an equation of a line that fits through that cluster of points with the minimal amount of deviations from the line. The deviation of the points from the line is called "error." Once you have this regression equation, if you knew a person's weight, you could then predict their height. Simple linear regression is actually the same as a bivariate correlation between the independent and dependent variable.

# 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  $x$ .



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

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

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
$\Sigma x = 15$	$\Sigma y = -1$	$\Sigma xy = 9$	$\Sigma x^2 = 55$	$\Sigma y^2 = 15$

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

Continued.



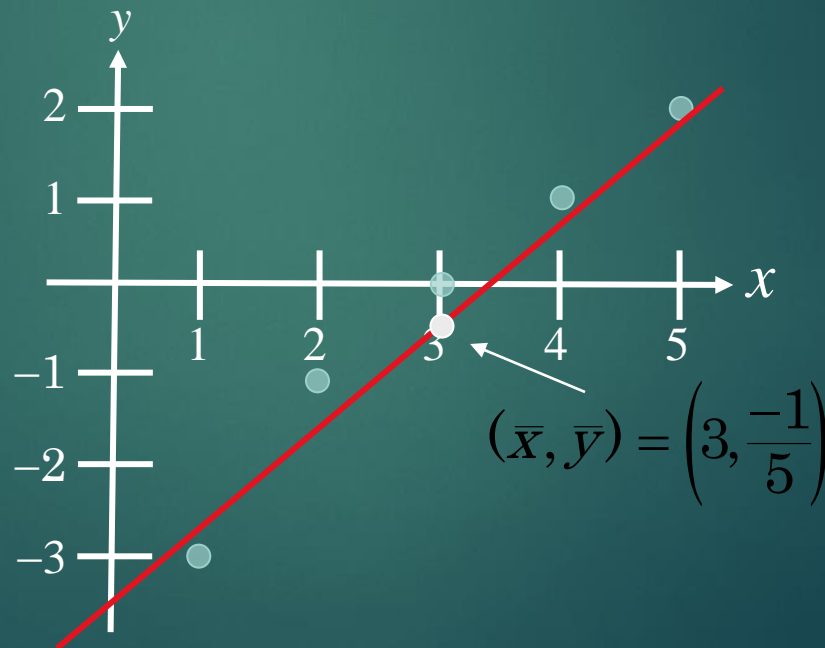
# Regression Line

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.

- Find the equation of the regression line.
- 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

$$\Sigma x = 54$$

$$\Sigma y = 908$$

$$\Sigma xy = 3724$$

$$\Sigma x^2 = 332$$

$$\Sigma y^2 = 70836$$

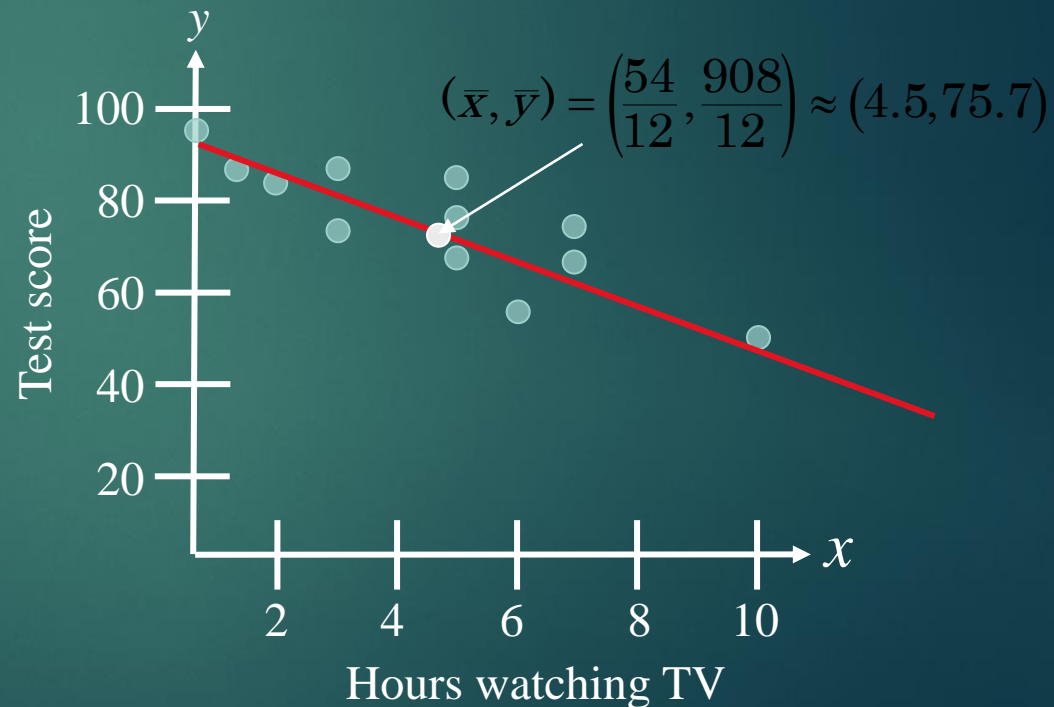
# Regression Line

Example continued:

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

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

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



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.

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

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

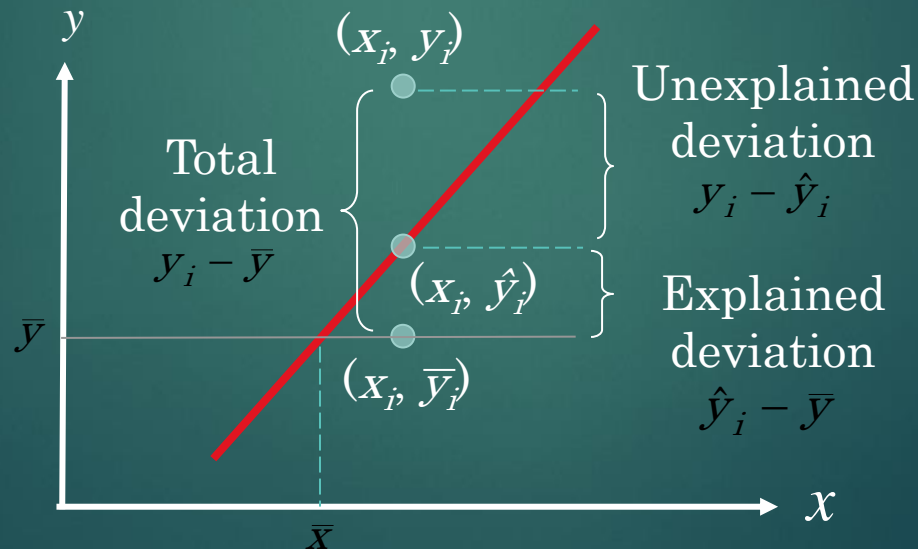
# Variation About a

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  $y$ .

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

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

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

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.

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

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

# 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.

$$\begin{aligned} r^2 &\approx (-0.831)^2 \\ &\approx 0.691 \end{aligned}$$

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  $y$ -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

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

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*

1. Make a table that includes the column heading shown.
2. Use the regression equation to calculate the predicted  $y$ -values.
3. Calculate the sum of the squares of the differences between each observed  $y$ -value and the corresponding predicted  $y$ -value.
4. Find the standard error of estimate.

### *In Symbols*

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

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

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

# 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:**

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

└─ Unexplained  
variation

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

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*

1. Identify the number of ordered pairs in the data set  $n$  and the degrees of freedom.
2. Use the regression equation and the given  $x$ -value to find the point estimate  $\hat{y}$ .
3. Find the critical value  $t_c$  that corresponds to the given level of confidence  $c$ .

### *In Symbols*

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

$$\text{d.f.} = n - 2$$

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

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

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

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

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.

# 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

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

where  $x_1, x_2, x_3, \dots, x_k$  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:

$$\begin{aligned}\text{a.) } \hat{y} &= 859 + 5.76x_1 + 3.82x_2 \\ &= 859 + 5.76(2758) + 3.82(2714) \\ &= 27,112.56\end{aligned}$$

The predicted annual rice yield is 27,1125.56 pounds.

$$\begin{aligned}\text{b.) } \hat{y} &= 859 + 5.76x_1 + 3.82x_2 \\ &= 859 + 5.76(3581) + 3.82(3021) \\ &= 33,025.78\end{aligned}$$

The predicted annual rice yield is 33,025.78 pounds.