Linear Regression

Objectives

1. Determine and interpret the linear correlation coefficient
2. Determine the linear regression equation
3. Determine and Interpret the Coefficient of Determination

Linear Correlation Coefficient

In the previous section, we examined correlation types (positive, negative, or none) with the help of the means of the explanatory (x) and response variables (y).

Linear Correlation Coefficient

In the previous section, we examined correlation types (positive, negative, or none) with the help of the means of the explanatory (x) and response variables (y).

In this section, we will examine the correlation type the way it is done in the real world: calculating the linear correlation coefficient (r).

Linear Correlation Coefficient

Correlation Coefficient

The **correlation coefficient**, r, is a numerical value with $-1 \leq r \leq 1$ that measures the type of linear correlation of a bivariate dataset.

Linear Correlation Coefficient

Correlation Coefficient

The **correlation coefficient**, r, is a numerical value with $-1 \leq r \leq 1$ that measures the type of linear correlation of a bivariate dataset.

- $r > 0$: positive linear correlation

Linear Correlation Coefficient

Correlation Coefficient

The **correlation coefficient**, r, is a numerical value with $-1 \leq r \leq 1$ that measures the type of linear correlation of a bivariate dataset.

- $r > 0$: positive linear correlation
- $r = 0$: no linear correlation

Linear Correlation Coefficient

Correlation Coefficient

The **correlation coefficient**, r, is a numerical value with $-1 \leq r \leq 1$ that measures the type of linear correlation of a bivariate dataset.

- $r > 0$: positive linear correlation
- $r = 0$: no linear correlation
- $r < 0$: negative linear correlation

Linear Correlation Coefficient

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

Linear Correlation Coefficient

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

We will use technology to calculate r

Linear Correlation Coefficient

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

We will use technology to calculate r

The closer r is to 1 (or -1), the more the data points "fall in line"

Linear Correlation Coefficient

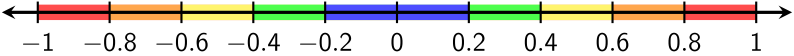
$$r = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sqrt{\sum(x - \overline{x})^2 \cdot \sum(y - \overline{y})^2}}$$

We will use technology to calculate r

The closer r is to 1 (or -1), the more the data points “fall in line”

The closer r is to 0, the more the data points resemble a “cloud”

Interpreting r



None

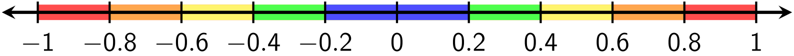
Weak

Moderate

Strong

Very Strong

Interpreting r



None

Weak

Moderate

Strong

Very Strong

Note: These interpretations are not universal.

Example 1

Find and interpret the linear correlation coefficient, r, for each.

(a)

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

Example 1

Find and interpret the linear correlation coefficient, r, for each.

(a)

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

$r \approx 0.9588$

Example 1

Find and interpret the linear correlation coefficient, r, for each.

(a)

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

$r \approx 0.9588$

Very strong positive linear correlation

Example 1

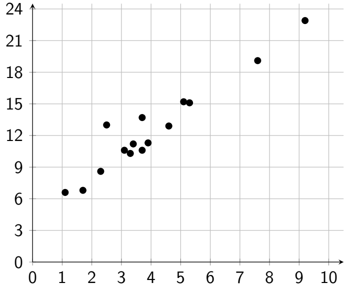
Find and interpret the linear correlation coefficient, r, for each.

(a)

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

$r \approx 0.9588$

Very strong positive linear correlation



Example 1

(b)

x	y
7.6	11.0
9.2	3.6
6.3	8.9
1.1	14.9
6.7	8.1
3.9	12.0
4.6	9.4
2.3	10.3
5.1	11.4
5.3	12.4
2.5	9.0
3.4	8.9
3.1	14.2
1.7	10.9
3.7	13.3

Example 1

(b)

x	y
7.6	11.0
9.2	3.6
6.3	8.9
1.1	14.9
6.7	8.1
3.9	12.0
4.6	9.4
2.3	10.3
5.1	11.4
5.3	12.4
2.5	9.0
3.4	8.9
3.1	14.2
1.7	10.9
3.7	13.3

$r \approx -0.6273$

Example 1

(b)

x	y
7.6	11.0
9.2	3.6
6.3	8.9
1.1	14.9
6.7	8.1
3.9	12.0
4.6	9.4
2.3	10.3
5.1	11.4
5.3	12.4
2.5	9.0
3.4	8.9
3.1	14.2
1.7	10.9
3.7	13.3

$r \approx -0.6273$

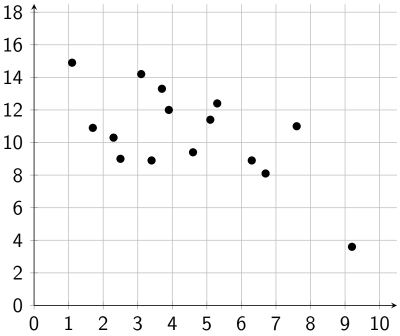
Strong negative linear correlation

(b)

x	y
7.6	11.0
9.2	3.6
6.3	8.9
1.1	14.9
6.7	8.1
3.9	12.0
4.6	9.4
2.3	10.3
5.1	11.4
5.3	12.4
2.5	9.0
3.4	8.9
3.1	14.2
1.7	10.9
3.7	13.3

$r \approx -0.6273$

Strong negative linear correlation



Example 1

(c)

x	y
6.9	3.4
7.7	4.5
0.9	9.8
3.4	1.5
8.9	3.3
5.7	8.9
3.1	8.4
2.2	8.1
4.5	6.8
4.1	0.5
5.0	0.4
7.8	8.4
2.5	3.1
6.1	9.0
1.1	8.5

Example 1

(c)

x	y
6.9	3.4
7.7	4.5
0.9	9.8
3.4	1.5
8.9	3.3
5.7	8.9
3.1	8.4
2.2	8.1
4.5	6.8
4.1	0.5
5.0	0.4
7.8	8.4
2.5	3.1
6.1	9.0
1.1	8.5

$r \approx -0.2218$

Example 1

(c)

x	y
6.9	3.4
7.7	4.5
0.9	9.8
3.4	1.5
8.9	3.3
5.7	8.9
3.1	8.4
2.2	8.1
4.5	6.8
4.1	0.5
5.0	0.4
7.8	8.4
2.5	3.1
6.1	9.0
1.1	8.5

$r \approx -0.2218$

Weak negative correlation

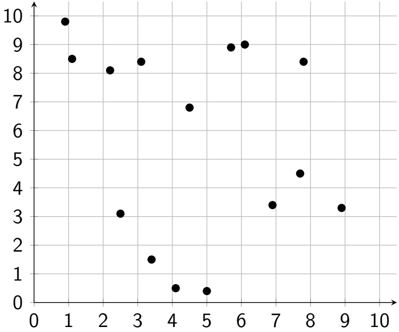
Example 1

(c)

x	y
6.9	3.4
7.7	4.5
0.9	9.8
3.4	1.5
8.9	3.3
5.7	8.9
3.1	8.4
2.2	8.1
4.5	6.8
4.1	0.5
5.0	0.4
7.8	8.4
2.5	3.1
6.1	9.0
1.1	8.5

$r \approx -0.2218$

Weak negative correlation



Objectives

1. Determine and interpret the linear correlation coefficient
2. Determine the linear regression equation
3. Determine and Interpret the Coefficient of Determination

Linear Regression Equation

While determining the linear correlation coefficient is valuable, it is also helpful to be able to predict data values not contained in the data set.

Linear Regression Equation

While determining the linear correlation coefficient is valuable, it is also helpful to be able to predict data values not contained in the data set.

To do this, we can create the **least squares regression equation**, (also called the *line of best fit*) which will **minimize** the total squared distance each data point is from the line:

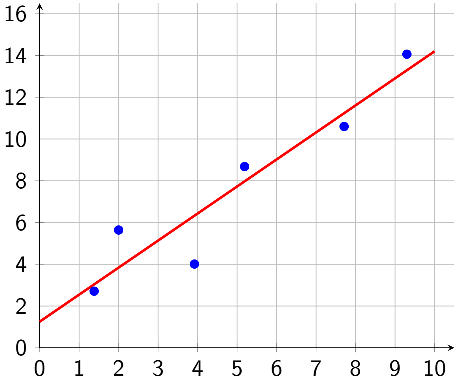
Linear Regression Equation

While determining the linear correlation coefficient is valuable, it is also helpful to be able to predict data values not contained in the data set.

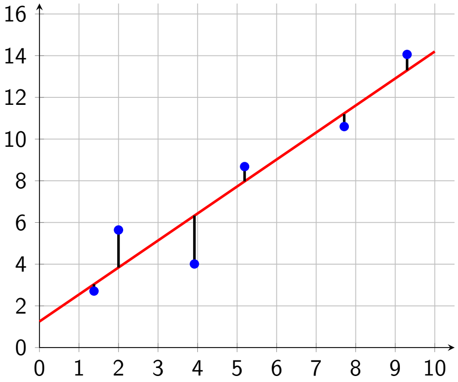
To do this, we can create the **least squares regression equation**, (also called the *line of best fit*) which will **minimize** the total squared distance each data point is from the line:

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

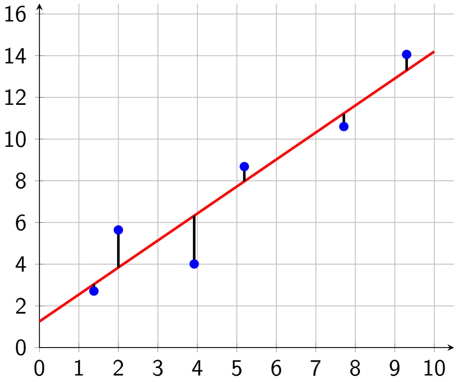
Line of Best Fit



Least Squares Regression Equation

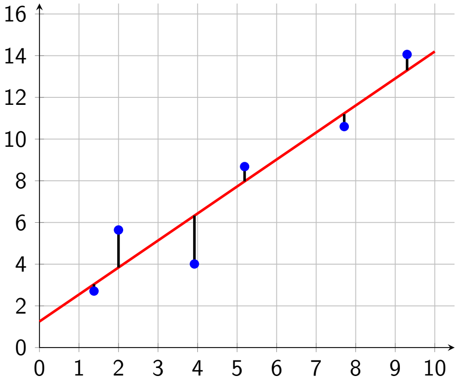


Least Squares Regression Equation



The black lines are **residuals**.

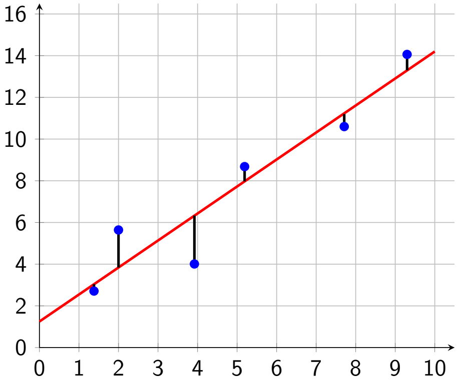
Least Squares Regression Equation



The black lines are **residuals**.

Like deviations from the mean, the sum of the residuals is 0.

Least Squares Regression Equation

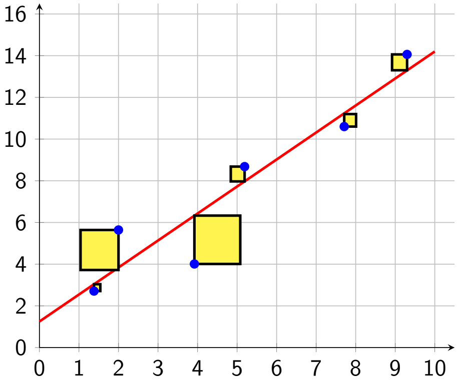


The black lines are **residuals**.

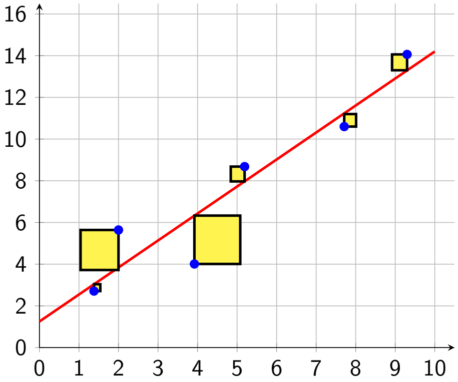
Like deviations from the mean, the sum of the residuals is 0.

So we need to square the deviations so the negatives don't cancel the positives.

Least Squares Regression Equation



Least Squares Regression Equation



The line of best fit minimizes the sum of the areas of the squares.

Slope and y-intercept

We will be using technology to find the equation of the line of best fit.

Slope and y-intercept

We will be using technology to find the equation of the line of best fit.

Below are the formulas for calculating the slope, m, and y-intercept, b:

Slope and y-intercept

We will be using technology to find the equation of the line of best fit.

Below are the formulas for calculating the slope, m, and y-intercept, b:

$$m = r\left(\frac{\sigma_y}{\sigma_x}\right)$$

and

$$b = \overline{y} - m\left(\overline{x}\right)$$

Example 2

Find the least squares regression equation for the following dataset.

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

Example 2

Find the least squares regression equation for the following dataset.

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

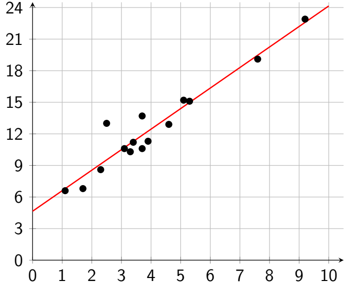
$$\hat{y} = 1.95x + 4.65$$

Example 2

Find the least squares regression equation for the following dataset.

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

$$\hat{y} = 1.95x + 4.65$$



Example 3

Given the regression equation $\hat{y} = 1.95x + 4.65$, predict the values of the following response variables for each explanatory variable.

(a) $x = 6$

Example 3

Given the regression equation $\hat{y} = 1.95x + 4.65$, predict the values of the following response variables for each explanatory variable.

(a) $x = 6$

Since 6 between the minimum and maximum values of x in our dataset, finding its y-coordinate is called **interpolation**.

.

Example 3

Given the regression equation $\hat{y} = 1.95x + 4.65$, predict the values of the following response variables for each explanatory variable.

(a) $x = 6$

Since 6 between the minimum and maximum values of x in our dataset, finding its y-coordinate is called **interpolation**.

$$\hat{y} = 1.95x + 4.65$$

.

Example 3

Given the regression equation $\hat{y} = 1.95x + 4.65$, predict the values of the following response variables for each explanatory variable.

(a) $x = 6$

Since 6 between the minimum and maximum values of x in our dataset, finding its y-coordinate is called **interpolation**.

$$\begin{aligned} \hat{y} &= 1.95x + 4.65 \\ &= 1.95(6) + 4.65. \end{aligned}$$

Example 3

Given the regression equation $\hat{y} = 1.95x + 4.65$, predict the values of the following response variables for each explanatory variable.

(a) $x = 6$

Since 6 between the minimum and maximum values of x in our dataset, finding its y-coordinate is called **interpolation**.

$$\begin{aligned}\hat{y} &= 1.95x + 4.65 \\ &= 1.95(6) + 4.65. \\ &= 16.35\end{aligned}$$

Example 3

Given the regression equation $\hat{y} = 1.95x + 4.65$, predict the values of the following response variables for each explanatory variable.

(a) $x = 6$

Since 6 between the minimum and maximum values of x in our dataset, finding its y-coordinate is called **interpolation**.

$$\begin{aligned} \hat{y} &= 1.95x + 4.65 \\ &= 1.95(6) + 4.65. \\ &= 16.35 \end{aligned}$$

The predicted value when $x = 6$ is $y = 16.35$

Residuals

Suppose we actually obtain a datapoint and realize that the actual value of y when $x = 6$ is 16, not the predicted 16.35.

Residuals

Suppose we actually obtain a datapoint and realize that the actual value of y when $x = 6$ is 16, not the predicted 16.35.

The residual, denoted ϵ, would be

$$\epsilon = \text{predicted value} - \text{observed value}$$

$$\epsilon = 16 - 16.35$$

$$\epsilon = -0.35$$

Residuals

Suppose we actually obtain a datapoint and realize that the actual value of y when $x = 6$ is 16, not the predicted 16.35.

The residual, denoted ϵ, would be

$$\epsilon = \text{predicted value} - \text{observed value}$$

$$\epsilon = 16 - 16.35$$

$$\epsilon = -0.35$$

We could then add that observation to our dataset and use it to create a better linear regression equation.

Example 3

(b) $x = 11$

Example 3

(b) $x = 11$

Since 11 is outside of the x values in our dataset, finding its y-coordinate is called **extrapolation**.

Example 3

(b) $\quad x = 11$

Since 11 is outside of the x values in our dataset, finding its y-coordinate is called **extrapolation**.

$$\hat{y} = 1.95x + 4.65$$

Example 3

(b) $\quad x = 11$

Since 11 is outside of the x values in our dataset, finding its y-coordinate is called **extrapolation**.

$$\begin{aligned}\hat{y} &= 1.95x + 4.65 \\ &= 1.95(11) + 4.65\end{aligned}$$

Example 3

(b) $x = 11$

Since 11 is outside of the x values in our dataset, finding its y-coordinate is called **extrapolation**.

$$\begin{aligned}\hat{y} &= 1.95x + 4.65 \\ &= 1.95(11) + 4.65 \\ &= 26.1\end{aligned}$$

Example 3

(b) $\quad x = 11$

Since 11 is outside of the x values in our dataset, finding its y-coordinate is called **extrapolation**.

$$\begin{aligned}\hat{y} &= 1.95x + 4.65\\ &= 1.95(11) + 4.65\\ &= 26.1\end{aligned}$$

The predicted value when $x = 11$ is $y = 26.1$.

Objectives

1. Determine and interpret the linear correlation coefficient
2. Determine the linear regression equation
3. Determine and Interpret the Coefficient of Determination

Coefficient of Determination

We looked at how to calculate the value of the linear correlation coefficient r.

Coefficient of Determination

We looked at how to calculate the value of the linear correlation coefficient r.

The value r^2 is called the **coefficient of determination** and it tells us what percentage of the variability in the response (y) variable is due to the linear relationship of x and y.

Coefficient of Determination

We looked at how to calculate the value of the linear correlation coefficient r.

The value r^2 is called the **coefficient of determination** and it tells us what percentage of the variability in the response (y) variable is due to the linear relationship of x and y.

The rest may be due to such things as lurking variables.

Coefficient of Determination

We looked at how to calculate the value of the linear correlation coefficient r.

The value r^2 is called the **coefficient of determination** and it tells us what percentage of the variability in the response (y) variable is due to the linear relationship of x and y.

The rest may be due to such things as lurking variables.

Since $-1 \leq r \leq 1$, $0 \leq r^2 \leq 1$

But What Exactly Does r^2 Do?

Without our regression equation, the best predictor for response variables (y values) would be $\overline{y}$.

But What Exactly Does r^2 Do?

Without our regression equation, the best predictor for response variables (y values) would be $\overline{y}$.

However, the value of r^2 represents how much of a decrease in prediction error we get from using our regression equation rather than $\overline{y}$.

But What Exactly Does r^2 Do?

Without our regression equation, the best predictor for response variables (y values) would be $\overline{y}$.

However, the value of r^2 represents how much of a decrease in prediction error we get from using our regression equation rather than $\overline{y}$.

Note: If r (and hence r^2) is close to 0, then the linear regression equation is not going to be a good predictor. In this case, use $\hat{y} = \bar{y}$ as the linear predictor equation.

Example 4

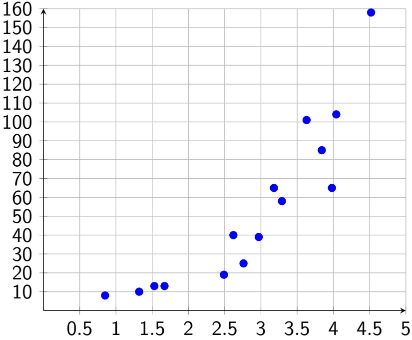
Determine and interpret the value of r^2 given the paw width (in inches) and dog's weight (in pounds) below.

Paw	Weight
1.32	10
1.67	13
2.76	25
3.98	65
2.97	39
2.49	19
3.84	85
4.04	104
4.52	158
3.18	65
3.29	58
3.63	101
0.85	8
4.62	157
1.53	13

Example 4

Determine and interpret the value of r^2 given the paw width (in inches) and dog's weight (in pounds) below.

Paw	Weight
1.32	10
1.67	13
2.76	25
3.98	65
2.97	39
2.49	19
3.84	85
4.04	104
4.52	158
3.18	65
3.29	58
3.63	101
0.85	8
4.62	157
1.53	13



Example 4

Paw	Weight
1.32	10
1.67	13
2.76	25
3.98	65
2.97	39
2.49	19
3.84	85
4.04	104
4.52	158
3.18	65
3.29	58
3.63	101
0.85	8
4.62	157
1.53	13

$$r^2 \approx 0.8049$$

Example 4

Paw	Weight
1.32	10
1.67	13
2.76	25
3.98	65
2.97	39
2.49	19
3.84	85
4.04	104
4.52	158
3.18	65
3.29	58
3.63	101
0.85	8
4.62	157
1.53	13

$$r^2 \approx 0.8049$$

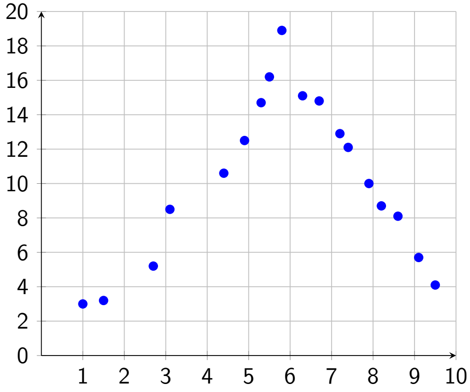
The prediction error used to predict y is about 80.5% smaller than using $\overline{y}$ as a predictor.

Other Types of Regression

Since we only get values of r for *linear* regression, we will not calculate a linear correlation coefficient for a data set that does not appear linear.

Other Types of Regression

Since we only get values of r for *linear* regression, we will not calculate a linear correlation coefficient for a data set that does not appear linear.



Other Types of Regression

We can still calculate r^2 using the more-traditional formula than was presented in this section; although you would likely want to let technology calculate it for you.