Chapter 4 Describing the Relation between Two Variables

Section 4.1 Scatter Diagrams and Correlation

Objectives

1. Draw and interpret scatter diagrams
2. Describe the properties of the linear correlation coefficient
3. Compute and interpret the linear correlation coefficient
4. Determine whether a linear relation exists between two variables
5. Explain the difference between correlation and causation

The **response** (or **dependent**) **variable** is the variable whose value can be explained by the value of the **explanatory** or **predictor** or **independent variable**.

1 Draw and Interpret Scatter Diagrams

A **scatter diagram** is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The explanatory variable is plotted on the horizontal axis, and the response variable is plotted on the vertical axis.

Example 1 Drawing a Scatter Diagram

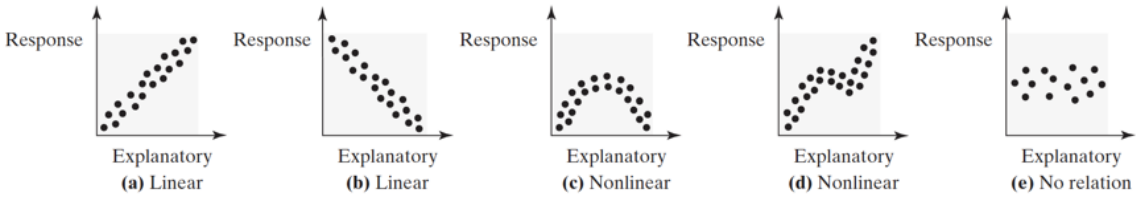
The following data are based on a study for drilling rock. The researchers wanted to determine whether the time it takes to dry drill a distance of 5 feet in rock increases with the depth at which the drilling begins. So, depth at which drilling begins is the explanatory variable, x, and time (in minutes) to drill five feet is the response variable, y. Draw a scatter diagram of the data. The data are in StatCrunch.

Depth at Which Drilling Begins (in feet), x	Time to Drill Five Feet (in minutes), y
35	5.88
50	5.99
75	6.74
95	6.10
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.90

Data from Penner, R., and Watts, D.G. "Mining Information." *The American Statistician*, Vol. 45, No. 1, Feb. 1991, p. 6.

Alternative Example 1

Go to www.zillow.com and select a city in which you would like to build a model for predicting the selling price of a home. Under "Listing Type", select Recently Sold. Enter some parameters for the home (number of bedrooms, home type). Randomly select about 15 recently sold homes and record the Zestimate (Zillow's estimate of the home's value) and the Sale Price. Draw a scatter diagram of the data treating Zestimate as the explanatory variable.



Two variables that are linearly related are **positively associated** when above-average values of one variable are associated with above-average values of the other variable and below-average values of one variable are associated with below-average values of the other variable. That is, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable also increases.

Two variables that are linearly related are **negatively associated** when above-average values of one variable are associated with below-average values of the other variable. That is, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.

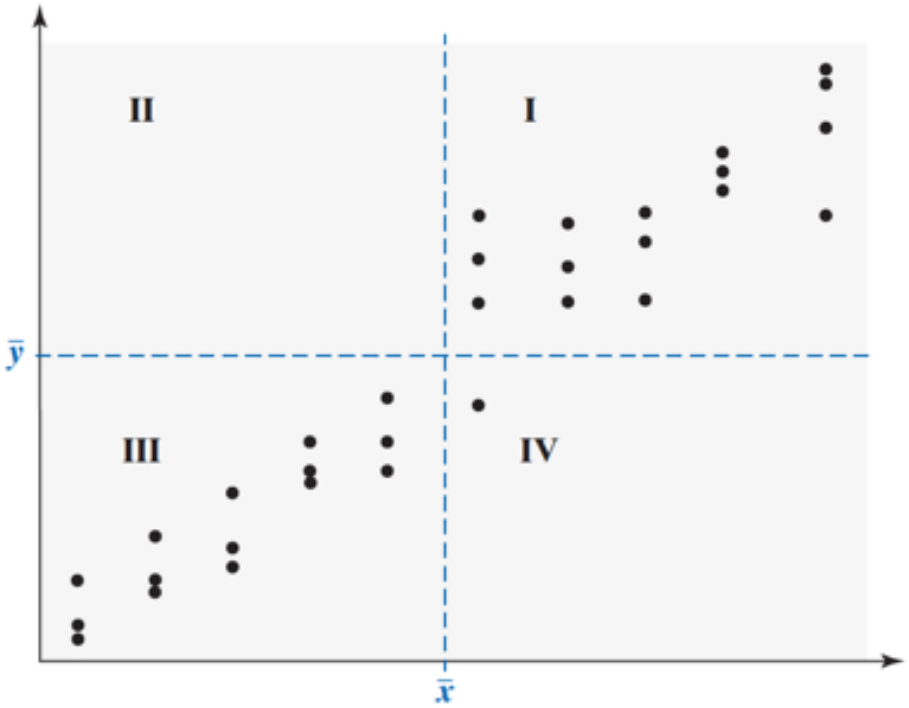
2 Describe the Properties of the Linear Correlation Coefficient

The **linear correlation coefficient** or **Pearson product moment correlation coefficient** is a measure of the strength and direction of the linear relation between two quantitative variables. The Greek letter ρ (rho) represents the population correlation coefficient, and r represents the sample correlation coefficient. We present only the formula for the sample correlation coefficient.

Sample Linear Correlation Coefficient*

$$r = \frac{\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n - 1} \quad (1)$$

where x_i is the ith observation of the explanatory variable
$\bar{x}$ is the sample mean of the explanatory variable
s_x is the sample standard deviation of the explanatory variable
y_i is the ith observation of the response variable
$\bar{y}$ is the sample mean of the response variable
s_y is the sample standard deviation of the response variable
n is the number of individuals in the sample



Applet Activity – Exploring the Properties of the Linear Correlation Coefficient

Load the Correlation by Eye Applet that is located at www.pearsonhighered.com/sullivanstats . Or, from StatCrunch, select Applets > Correlation by eye . Select the "Randomly generated" radio button and click Compute!.

1. Click "Reset" at the top of the applet to clear the data from the scatter diagram.
a. Create a scatter diagram of 12 to 15 observations with positive association. Click "Show" at the bottom of the applet to show the correlation coefficient of the data in the scatter diagram. Draw the scatter diagram (or copy it from the applet) and record the correlation coefficient below.

b. Move some of the observations from the scatter diagram and note how the correlation coefficient changes as the positive association strengthens and weakens.
c. Align the points in the scatter diagram in a straight line with positive slope. What is the value of the linear correlation coefficient?

2. Click “Reset” at the top of the applet to clear the data from the scatter diagram.
a. Create a scatter diagram of 12 to 15 observations with negative association. Click “Show” at the bottom of the applet to show the correlation coefficient of the data in the scatter diagram. Draw the scatter diagram and record the correlation coefficient below.

b. Move some of the observations from the scatter diagram and note how the correlation coefficient changes as the negative association strengthens and weakens.
c. Align the points in the scatter diagram in a straight line with negative slope. What is the value of the linear correlation coefficient?

3. a. Click “Reset” at the top of the applet to clear the data from the scatter diagram. Create a scatter diagram with no association. What is the value of the correlation coefficient?

b. Click “Reset” at the top of the applet to clear the data from the scatter diagram. Create a scatter diagram in an upside-down U-shaped pattern. Draw the scatter diagram (or copy it from the applet) and record the correlation coefficient below.

c. What does a correlation coefficient of 0 suggest?

4. a. Click “Reset” at the top of the applet to clear the data from the scatter diagram. In the lower-left corner of the applet, draw a scatter diagram of 8 to 10 observations with a correlation coefficient around 0.8.

b. Add another point in the upper-right corner of the applet that roughly lines up with the other points in the scatter diagram. What is the new value of the correlation coefficient?

c. Move the additional point around the scatter diagram and note how the correlation coefficient changes. Is the correlation coefficient a resistant measure? Why or why not?

5. a. Click “Reset” at the top of the applet to clear the data from the scatter diagram. Draw a scatter diagram with six points arranged vertically in a straight line. What is the value of the correlation coefficient? Why?

b. Add a seventh point to the right side of the scatter diagram and move the point around until the correlation coefficient is approximately 0.75.

c. Click “Reset” at the top of the applet to clear the data from the scatter diagram. Draw a scatter diagram with approximately seven points in a U-shaped pattern near the lower-left corner of the applet. Add an eighth point to the scatter diagram and move it around until the correlation coefficient is approximately 0.75. Draw the scatter diagram (or copy it from the applet).

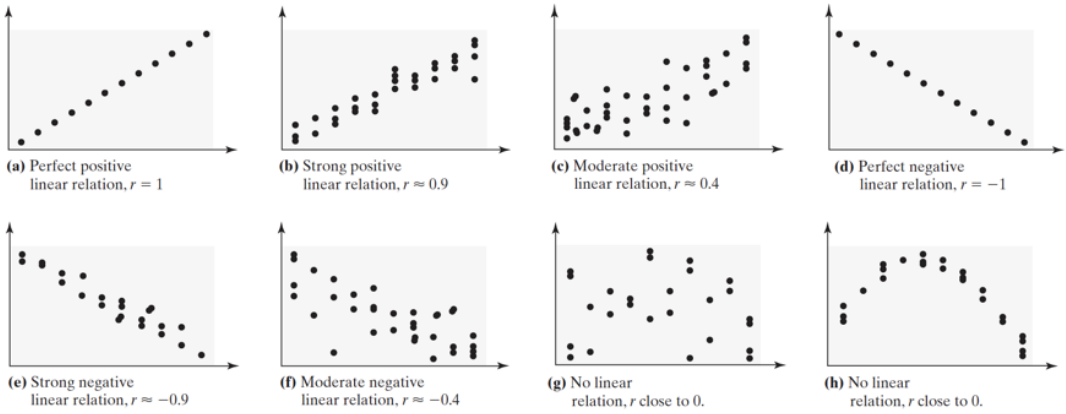
d. Explain why the correlation coefficient should not be used exclusively to judge linear association without also using a scatter diagram.

Properties of the Linear Correlation Coefficient

1. The linear correlation coefficient is always between -1 and 1, inclusive. That is, $-1 \le r \le 1$.
2. If $r = +1$, then a perfect positive linear relation exists between the two variables. See Figure 4(a).
3. If $r = -1$, then a perfect negative linear relation exists between the two variables. See Figure 4(d).
4. The closer r is to $+1$, the stronger is the evidence of positive association between the two variables. See Figures 4(b) and 4(c).
5. The closer r is to -1, the stronger is the evidence of negative association between the two variables. See Figures 4(e) and 4(f).
6. If r is close to 0, then little or no evidence exists of a *linear* relation between the two variables. So ***r* close to 0 does not imply no relation, just no *linear* relation**. See Figures 4(g) and 4(h).
7. The linear correlation coefficient is a unitless measure of association. So the unit of measure for x and y plays no role in the interpretation of r.
8. The correlation coefficient is not resistant. Therefore, an observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.

CAUTION!
A linear correlation coefficient close to 0 does not imply that there is no relation, just no linear relation. For example, although the scatter diagram drawn in Figure 4(h) indicates that the two variables are related, the linear correlation coefficient is close to 0.

Figure 4



(a) Perfect positive linear relation, $r = 1$
(b) Strong positive linear relation, $r \approx 0.9$
(c) Moderate positive linear relation, $r \approx 0.4$
(d) Perfect negative linear relation, $r = -1$
(e) Strong negative linear relation, $r \approx -0.9$
(f) Moderate negative linear relation, $r \approx -0.4$
(g) No linear relation, r close to 0.
(h) No linear relation, r close to 0.

3 Compute and Interpret the Linear Correlation Coefficient

Example 2 Computing the Linear Correlation Coefficient by Hand

Determine the linear correlation coefficient of the drilling data.

Depth, x	Time, y	$\frac{x_i - \bar{x}}{s_x}$	$\frac{y_i - \bar{y}}{s_y}$	$\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
35	5.88	-1.74712	-1.41633	2.474501
50	5.99	-1.45992	-1.27544	1.862051
75	6.74	-0.98126	-0.31486	0.308958
95	6.1	-0.59833	-1.13456	0.678839
120	7.47	-0.11967	0.620111	-0.07421
130	6.93	0.0718	-0.07151	-0.00513
145	6.42	0.358998	-0.72471	-0.26017
155	7.97	0.550463	1.260501	0.693859
160	7.92	0.646196	1.196462	0.773149
175	7.62	0.933394	0.812228	0.758129
185	6.89	1.12486	-0.12274	-0.13807
190	7.9	1.220592	1.170846	1.429126
$\bar{x} = 126.25$	$\bar{y} = 6.985833$	$\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$		= 8.501037
$S_x = 52.22874$	$S_y = 0.780774$			

$$r = \frac{\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$
$$= \frac{8.501037}{12-1}$$
$$= 0.773$$

Example 3 Using StatCrunch to Determine the Linear Correlation Coefficient

Use StatCrunch to find the linear correlation coefficient of the drilling data or Zillow data.

4 Determine Whether a Linear Relation Exists between Two Variables

Testing for a Linear Relation

Step 1 Determine the absolute value of the correlation coefficient.

Step 2 Find the critical value in Table II from Appendix A for the given sample size.

Step 3 If the absolute value of the correlation coefficient is greater than the critical value, we say a linear relation exists between the two variables. Otherwise, no linear relation exists.

Example 4 Does a Linear Relation Exist?

Determine whether a linear relation exsits between depth at which drilling begins and time to drill five feet (or the Zestimate and Sale Price).

5 Explain the Difference between Correlation and Causation

According to data obtained from the Statistical Abstract of the United States, the correlation between the percentage of the female population with a bachelor's degree and the percentage of births to unmarried mothers since 1990 is 0.940.
Does this mean that a higher percentage of females with bachelor's degrees causes a higher percentage of births to unmarried mothers?

Certainly not! The correlation exists only because both percentages have been increasing since 1990. It is this relation that causes the high correlation. In general, time series data (data collected over time) may have high correlations because each variable is moving in a specific direction over time (both going up or down over time; one increasing, while the other is decreasing over time).
When data are observational, we cannot claim a causal relation exists between two variables. We can only claim causality when the data are collected through a designed experiment.

Another way that two variables can be related even though there is not a causal relation is through a *lurking variable*.
A **lurking variable** is related to both the explanatory and response variable.

For example, ice cream sales and crime rates have a very high correlation. Does this mean that local governments should shut down all ice cream shops? No! The lurking variable is temperature. As air temperatures rise, both ice cream sales and crime rates rise.

Example – Beware the Lurking Variable

Open the data set "SAT versus Teacher Salaries" in StatCrunch (in the SullyStats group).

(a) Draw a scatter diagram treating "Avg Teacher Salary (000s)" as the explanatory variable and "Overall SAT" as the response variable. What type of relation appears to exist between teacher's salary and SAT score?

(b) Find the linear correlation between teacher salary and overall SAT score. Does a linear relation appear to exist between these two variables? If so, what type of relation? Is this consistent with what you would expect?

(c) The variable "Percent Taking" is a qualitative variable where "low" means less than or equal to 22% of eligible students took the SAT; "med" means between 23 and 49% of eligible students took the SAT; "high" means at least 50% of eligible students took the SAT. Draw a scatter diagram using a different plotting symbol for each classification of Percent Taking.

(d) Compute the linear correlation coefficient for each classification of Percent Taking. Does a linear relation appear to exist between these two variables for each classification?

(e) Summarize the dangers of drawing conclusions about relationships between variables without considering lurking variables.

4.2 Least-Squares Regression

Objectives

1. Find the least-squares regression line and use the line to make predictions
2. Interpret the slope and the *y*-intercept of the least-squares regression line
3. Compute the sum of squared residuals

Example 1 Finding an Equation that Describes Linearly Related Data

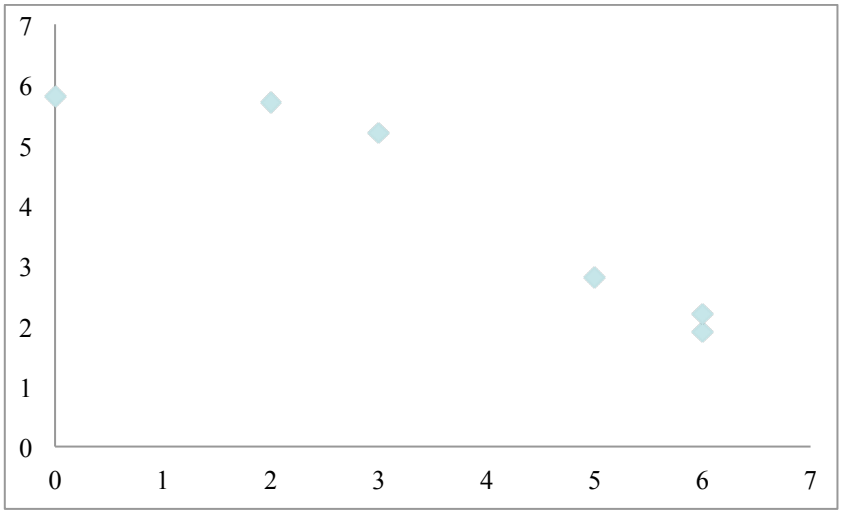
Use the following sample data:

x	0	2	3	5	6	6
y	5.8	5.7	5.2	2.8	1.9	2.2

(a) Find a linear equation that relates x (the explanatory variable) and y (the response variable) by selecting two points and finding the equation of the line containing the points.

Choose (2, 5.7) and (6, 1.9).

(b) Graph the equation on the scatter diagram.



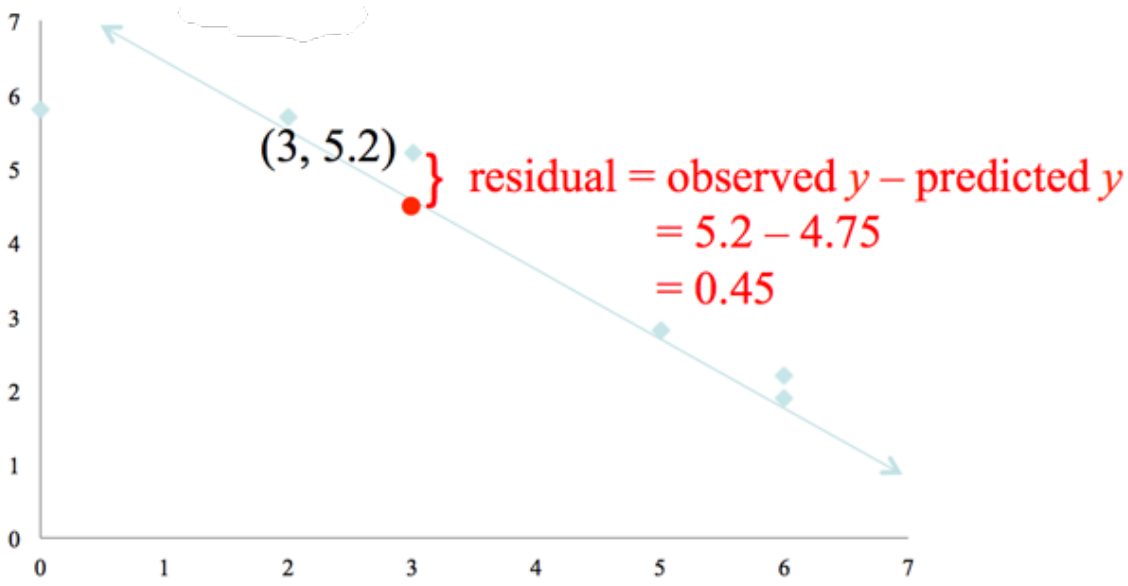
(c) Use the equation to predict y if $x = 3$.

1 Find the Least-Squares Regression Line and Use the Line to Make Predictions

The difference between the observed value of y and the predicted value of y is the error, or **residual**.

Using the line from the last example, and the predicted value at $x = 3$:

residual = observed y – predicted y
= 5.2 – 4.75
= 0.45



Least-Squares Regression Criterion

The **least-squares regression line** is the line that minimizes the sum of the squared errors (or residuals). This line minimizes the sum of the squared vertical distance between the observed values of y and those predicted by the line, $\hat{y}$ (read "y-hat"). We represent this as "minimize Σ residuals^2".

Activity What is "Least-Squares?"

Go to StatCrunch. Select Applets > Regression > by eye. Click Compute! Click Reset to remove all the point from the scatter diagram. Create a scatter diagram with positive association and 8 points.

Using the endpoints of the green line, attempt to fit a linear model to the data that minimizes the sum of squared residuals (or errors), SSE. Each point has a square whose side represents the residual of the point. Therefore, the area of the square is the value of the squared residual (because the area of a square equals side^2). So your goal should be to make the sum of the areas of the squares as small as possible. The area of the squares is given under the heading SSE. Pay attention to the "fit" of the line versus the value of the sum of squared residuals (or errors), SSE. Check the "Regression" box. Compare the SSE for the green line to the least-squares regression line. Notice that it is not possible to find a line whose SSE is less than the regression line.

The Least-Squares Regression Line

The equation of the least-squares regression line is given by

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

where

$$b_1 = r \cdot \frac{s_y}{s_x} \text{ is the } \textbf{slope} \text{ of the least-squares regression line*} \qquad \textbf{(2)}$$

and

$$b_0 = \bar{y} - b_1 \bar{x} \text{ is the } \textbf{y-intercept} \text{ of the least-squares regression line} \qquad \textbf{(3)}$$

Note: $\bar{x}$ is the sample mean and s_x is the sample standard deviation of the explanatory variable x; $\bar{y}$ is the sample mean and s_y is the sample standard deviation of the response variable y.

Example 2 Finding the Least-Squares Regression Line

Using the drilling or Zillow data.

(a) Find the least-squares regression line.
(b) Predict the drilling time if drilling starts at 130 feet. Choose a Zestimate from to make a prediction if you are using Zillow data].
(c) Is the observed drilling time at 130 feet above, or below, average.
(d) Draw the least-squares regression line on the scatter diagram of the data.

2 Interpret the Slope and the *y*-Intercept of the Least-Squares Regression Line

Interpretation of Slope:
The slope of the regression line is 0.0116. For each additional foot of depth we start drilling, the time to drill five feet increases by 0.0116 minutes, on average.

Interpretation of the *y*-Intercept: The *y*-intercept of the regression line is 5.5273. To interpret the *y*-intercept, we must first ask two questions:
1. Is 0 a reasonable value for the explanatory variable?
2. Do any observations near $x = 0$ exist in the data set?
A value of 0 is reasonable for the drilling data (this indicates that drilling begins at the surface of Earth. The smallest observation in the data set is $x = 35$ feet, which is reasonably close to 0. So, interpretation of the *y*-intercept is reasonable.
The time to drill five feet when we begin drilling at the surface of Earth is 5.5273 minutes.

Caution If the least-squares regression line is used to make predictions based on values of the explanatory variable that are much larger or much smaller than the observed values, we say the researcher is working **outside the scope of the model**. Never use a least-squares regression line to make predictions outside the scope of the model because we can't be sure the linear relation continues to exist.

4.3 Diagnostics on the Least-Squares Regression Line

Objectives

1. Compute and interpret the coefficient of determination
2. Perform residual analysis on a regression model
3. Identify influential observations

1 Compute and Interpret the Coefficient of Determination

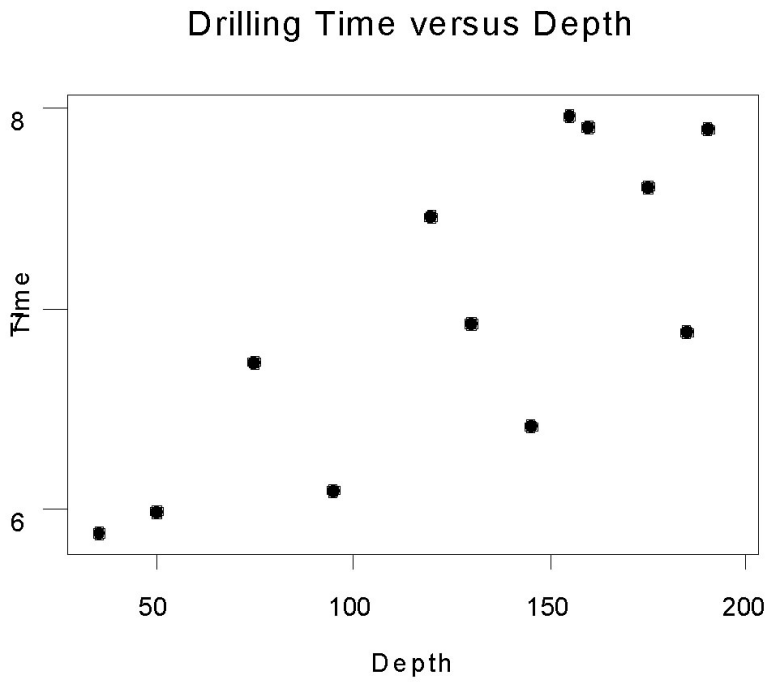
The **coefficient of determination, R^2,** measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

The coefficient of determination is a number between 0 and 1, inclusive. That is, $0 \le R^2 \le 1$.

If $R^2 = 0$ the line has no explanatory value

If $R^2 = 1$ means the line explains 100% of the variation in the response variable.

Let's consider the drilling data.



Sample Statistics

	Mean	Standard Deviation
Depth	126.2	52.2
Time	6.99	0.781

Correlation Between Depth and Time: 0.773

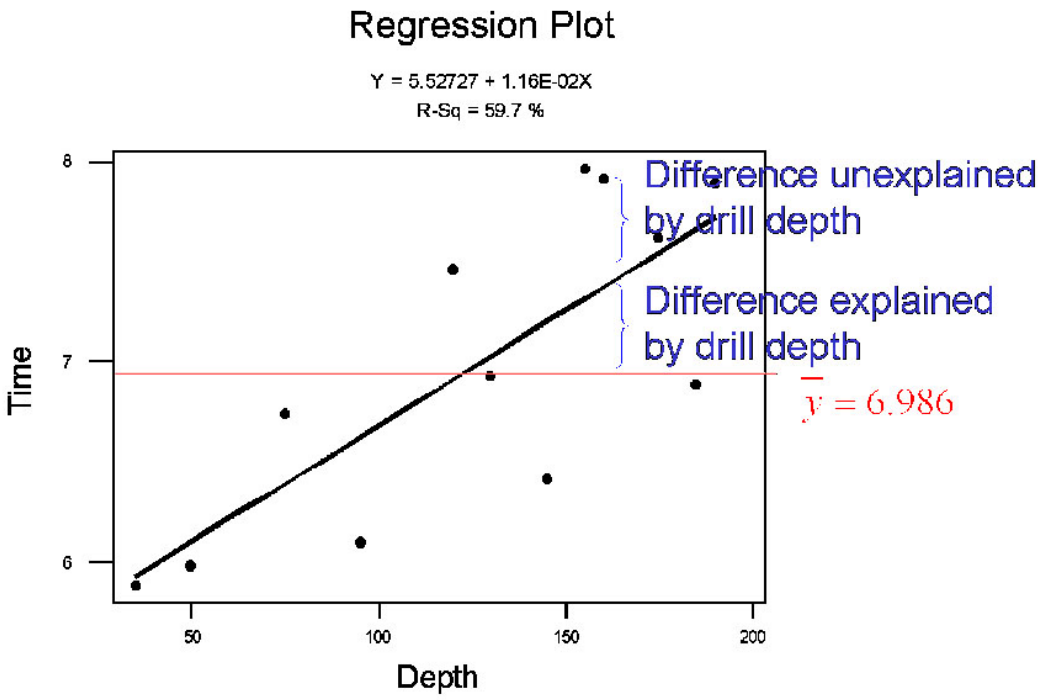
Regression Analysis

The regression equation is

Time = 5.53 + 0.0116 Depth

Suppose we were asked to predict the time to drill an additional 5 feet, but we did not know the current depth of the drill. What would be our best "guess"?

Now suppose that we are asked to predict the time to drill an additional 5 feet if the current depth of the drill is 160 feet?



The difference between the observed value of the response variable and the mean value of the response variable is called the **total deviation** and is equal to

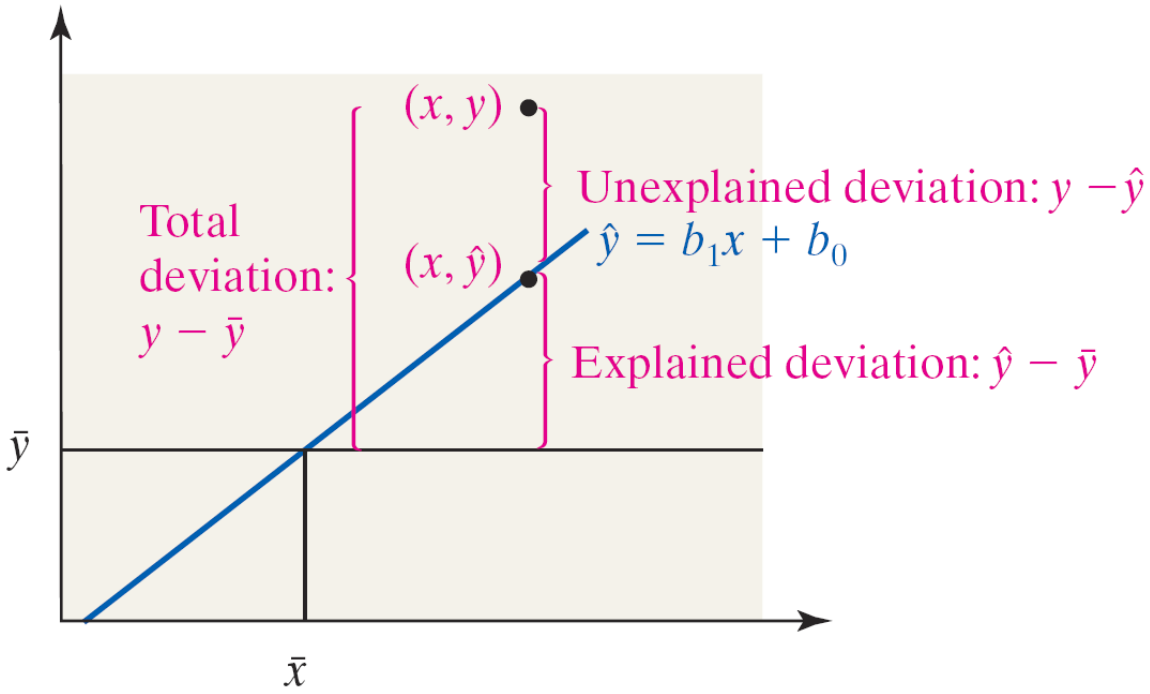
$$y - \bar{y}$$

The difference between the predicted value of the response variable and the mean value of the response variable is called the **explained deviation** and is equal to

$$\hat{y} - \bar{y}$$

The difference between the observed value of the response variable and the predicted value of the response variable is called the **unexplained deviation** and is equal to

$$y - \hat{y}$$



$$\underset{\text{Deviation}}{\text{Total}} = \underset{\text{Deviation}}{\text{Unexplained}} + \underset{\text{Deviation}}{\text{Explained}}$$

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

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

Total Variation = Unexplained Variation + Explained Variation

$$1 = \frac{\text{Unexplained Variation}}{\text{Total Variation}} + \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

To determine R^2 for the linear regression model simply square the value of the linear correlation coefficient.

Squaring the linear correlation coefficient to obtain the coefficient of determination works only for the least-squares linear regression model

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

The method does not work in general.

Example 1 Determining the Coefficient of Determination

Find and interpret the coefficient of determination for the drilling (or Zillow) data.

Activity Understanding the Coefficient of Determination

Go to our class group in StatCrunch. Open "4_3_Activity1.txt".

(a) All three data sets have the same *y*-values. What is the standard deviation of *y*? What is the variance of *y*?

(b) Draw a scatter diagram of each data set. Which data set has the explanatory variable *x* that best explains the variability in *y*? Why?

(c) Find the least-squares regression model between *x* and *y* for each data set. Draw the least-squares regression line on each scatter diagram.

(d) Find the linear correlation coefficient between *x* and *y* for each data set. Use the linear correlation coefficient to find the coefficient of determination between *x* and *y* for each data set.

(e) Fill in the following table.

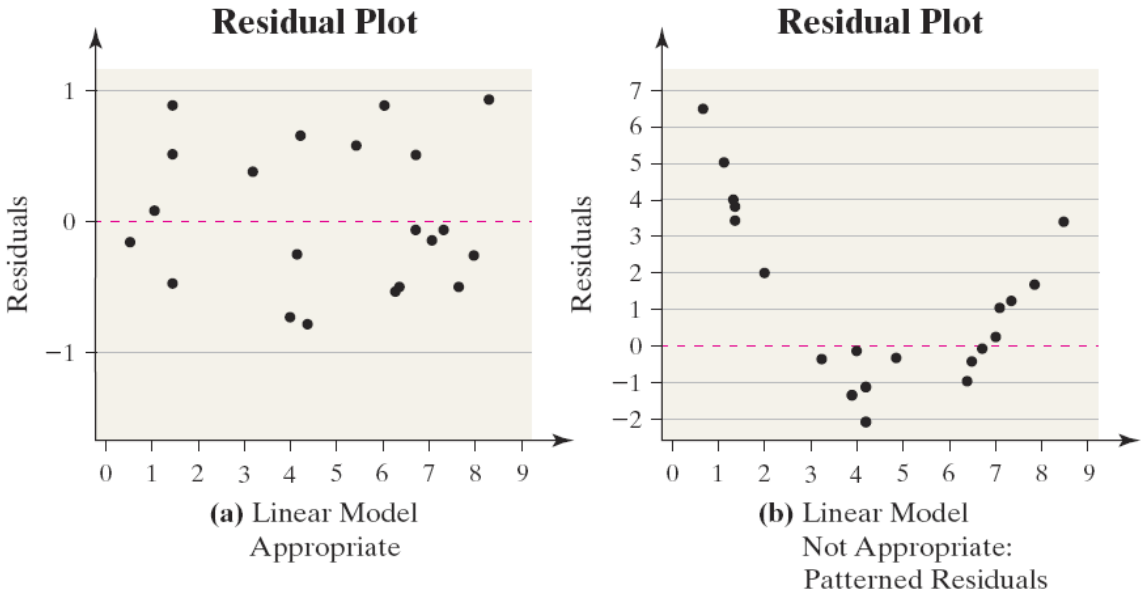
Data Set	Coefficient of Determination, R^2	Interpretation
A	____	____of the variability in y is explained by the least-squares regression line.
B	____	____of the variability in y is explained by the least-squares regression line.
C	____	____of the variability in y is explained by the least-squares regression line.

2 Perform Residual Analysis on a Regression Model

Residuals play an important role in determining the adequacy of the linear model. In fact, residuals can be used for the following purposes:

- To determine whether a linear model is appropriate to describe the relation between the predictor and response variables.
- To determine whether the variance of the residuals is constant.
- To check for outliers.

If a plot of the residuals against the predictor variable shows a discernable pattern, such as a curve, then the response and predictor variable may not be linearly related.



(a) Linear Model Appropriate

(b) Linear Model Not Appropriate: Patterned Residuals

Example 2 Is a Linear Model Appropriate?

A chemist has a 1000-gram sample of a radioactive material. She records the amount of radioactive material remaining in the sample every day for a week and obtains the following data.

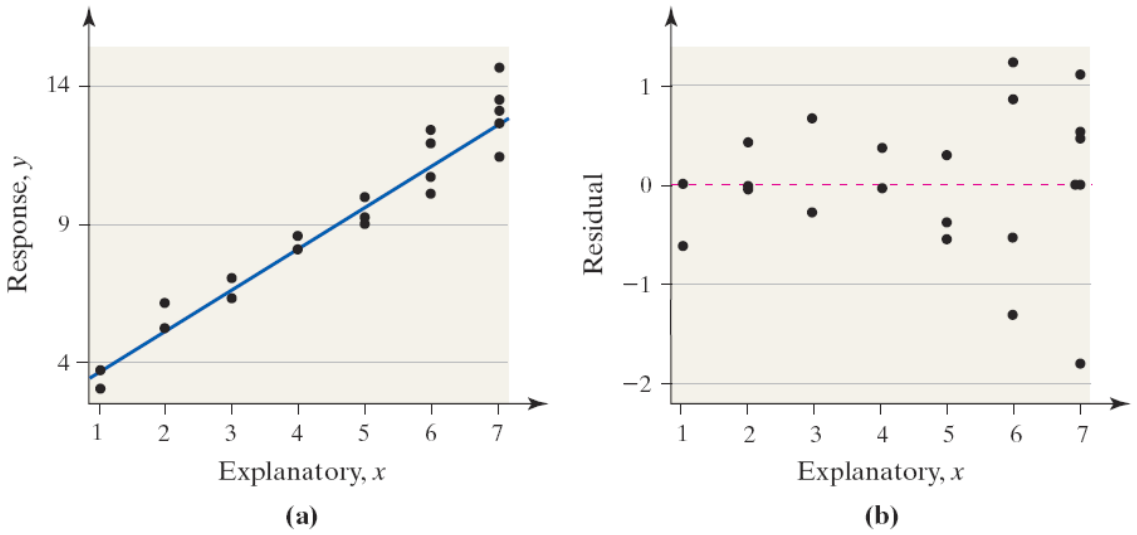
Day	Weight (in grams)
0	1000.0
1	897.1
2	802.5
3	719.8
4	651.1
5	583.4
6	521.7
7	468.3

(a) Draw a scatter diagram of the data and find the correlation between day and weight. Is there a linear relation between day and weight based on these results?

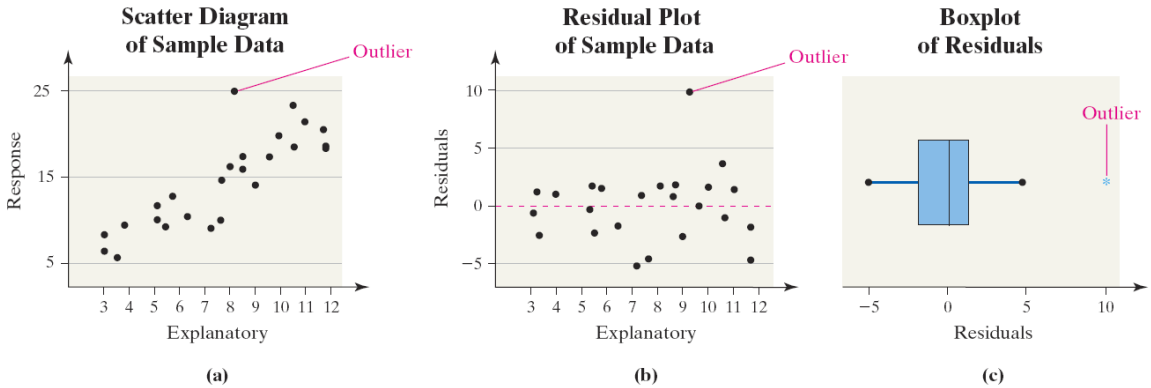
(b) Find the least-squares regression line and draw a residual plot. Based on the residual plot, what do you conclude?

If a plot of the residuals against the explanatory variable shows the spread of the residuals increasing or decreasing as the explanatory variable increases, then a strict requirement of the linear model is violated.

This requirement is called **constant error variance**. The statistical term for constant error variance is **homoscedasticity.**



A plot of residuals against the explanatory variable may also reveal outliers. These values will be easy to identify because the residual will lie far from the rest of the plot.



Example 3 Residual Analysis

Draw a residual plot of the drilling time (or Zillow) data. Comment on the appropriateness of the linear least-squares regression model.

3 Identify Influential Observations

An **influential observation** is an observation that significantly affects the least-squares regression line's slope and/or y-intercept, or the value of the correlation coefficient.

Activity Influential Observations

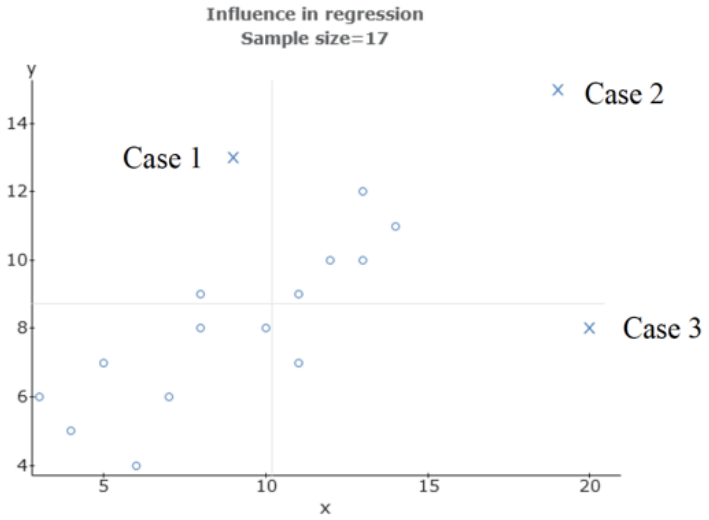
Load the Regression Influence Applet located at www.pearsonhighered.com/sullivanstats

1. a. Click "Add point." Add a point at (9, 13) by typing 9, 13 in the dialogue box. Click OK. Using your mouse, click and draw a box around the point (9, 13). You should see the point change to an x.
b. Check the box "Not selected" to draw the green regression line with the new point excluded. Look at the values of the intercept and slope in the table below the graph. Does it appear that the slope and/or y-intercept changed significantly?

2. a. Click “Reset” to reset the applet.
b. Add a point at (19, 15). Using your mouse, click and draw a box around the point (19, 15). You should see the point change to an x.
c. Check the box “Not selected” to draw the green regression line with the new point excluded. Look at the values of the intercept and slope in the table below the graph. Does it appear that the slope and/or y-intercept changed significantly?

3. a. Click “Reset” to reset the applet.
b. Add a point at (20, 8). Using your mouse, click and draw a box around the point (20, 8). You should see the point change to an x.
c. Check the box “Not selected” to draw the green regression line with the new point excluded. Look at the values of the intercept and slope in the table below the graph. Does it appear that the slope and/or y-intercept changed significantly?

The cases from the previous exploration are shown in the scatter diagram below. Case 1 is an outlier because the y -value is large relative to the x -value. In Case 2, both the x -value and y -value are large, but the point follows the overall pattern of the data. Finally, in Case 3, the x -value is large, and the y -value is not consistent with the pattern of the data.



Influence is affected by two factors:

(1) the relative vertical position of the observation (residuals) and
(2) the relative horizontal position of the observation (leverage).

Leverage is a measure that depends on how much the observation's value of the explanatory variable differs from the mean value of the explanatory variable.

4. Using these terms,
a. Case 1 has ____ (low/high) leverage and a ____ (small/large) residual.
b. Case 2 has_____ (low/high) leverage and a____ (small/large) residual.
c. Case 3 has _____ (low/high) leverage and a _____ (small/large) residual.

From the previous exploration, you should conclude that observations such as Case 3 tend to be influential.

Example 4 Influential Observations

Suppose an additional data point is added to the drilling data. At a depth of 300 feet, it took 12.49 minutes to drill 5 feet. Is this point influential? [Or add a house that may be influential to your Zillow data]

Section 4.4 Contingency Tables and Association

Objectives

1. Compute the marginal distribution of a variable
2. Use the conditional distribution to identify association among categorical data
3. Explain Simpson's Paradox

A professor at a community college conducted a study to assess the effectiveness of delivering Intermediate Algebra via traditional lecture-based method, online delivery (no classroom instruction), and through an emporium-style method of instruction in which students learn the material through a course management system with mastery learning required for each learning module. The grades students received in each of the courses were tallied.

Grade	Traditional	Online	Emporium
A	19	8	26
B	36	19	56
C	48	29	17
D	60	51	38
F	20	30	12
W	57	47	22

The table is referred to as a **contingency table,** or **two-way table**, because it relates two categories of data. The **row variable** is grade, because each row in the table describes the grade received for each group. The **column variable** is delivery method. Each box inside the table is referred to as a **cell**.

A **marginal distribution** of a variable is a frequency or relative frequency distribution of either the row or column variable in the contingency table.

Example 1 Determining Frequency and Relative Frequency Marginal Distributions

Determine the frequency and relative frequency marginal distributions for course grade and delivery method.

2 Use the Conditional Distribution to Identify Association among Categorical Data

A **conditional distribution** lists the relative frequency of each category of the response variable, given a specific value of the explanatory variable in the contingency table.

Example 2 Determining a Conditional Distribution

Construct a conditional distribution of course grade by method of delivery. Comment on any type of association that may exist between course grade and delivery method.

Example 3 Drawing a Bar Graph of a Conditional Distribution

Use the results of Example 3 to draw a conditional bar graph of grade distribution by method of delivery.

3 Explain Simpson's Paradox

Example 4 Illustrating Simpson's Paradox

Insulin dependent (or Type 1) diabetes is a disease that results in the permanent destruction of insulin-producing beta cells of the pancreas. Type 1 diabetes is lethal unless treatment with insulin injections replaces the missing hormone. Individuals with insulin independent (or Type 2) diabetes can produce insulin internally. The data shown in the table below represent the survival status of 902 patients with diabetes by type over a 5-year period.

	Type 1	Type 2	Total
Survived	253	326	579
Died	105	218	323
	358	544	902

From the table, the proportion of patients with Type 1 diabetes who died was 105/358 = 0.29; the proportion of patients with Type 2 diabetes who died was 218/544 = 0.40. Based on this, we might conclude that Type 2 diabetes is more lethal than Type 1 diabetes.

However, Type 2 diabetes is usually contracted after the age of 40. If we account for the variable age and divide our patients into two groups (those 40 or younger and those over 40), we obtain the data in the table below.

	Type 1		Type 2		Total
	≤ 40	> 40	≤ 40	> 40	
Survived	129	124	15	311	579
Died	1	104	0	218	323
	130	228	15	529	902

Of the diabetics 40 years of age or younger, the proportion of those with Type 1 diabetes who died is 1/130 = 0.008; the proportion of those with Type 2 diabetes who died is 0/15 = 0.

Of the diabetics over 40 years of age, the proportion of those with Type 1 diabetes who died is 104/228 = 0.456; the proportion of those with Type 2 diabetes who died is 218/529 = 0.412.

The lurking variable age led us to believe that Type 2 diabetes is the more dangerous type of diabetes.

Simpson's Paradox describes a situation in which an association between two variables inverts or goes away when a third variable is introduced to the analysis.