

CHAPTER 2

LOOKING AT DATA - RELATIONSHIPS

SECTION 2.1

OVERVIEW

The first chapter provides the tools to explore several types of variables one by one, but in most instances the data of interest are a collection of variables that may exhibit some kind of relationships among themselves. Typically, these relationships are more interesting than the behavior of the variables individually. If we think that one of the variables, x , may explain or even cause changes in another variable, y , we call x an **explanatory variable** and y a **response variable**.

The first tool we consider for examining the relationship between variables is the **scatterplot**. Scatterplots show us two quantitative variables at a time, such as the weight of a car and its MPG (miles per gallon). Using colors or different symbols, we can add information to the plot about a third variable which is categorical in nature. For example, if in our plot we wanted to distinguish between cars with manual or automatic transmissions, we might use a circle to plot the cars with manual transmissions and a cross to plot the cars with automatic transmissions.

When drawing a scatterplot, we need to pick one variable to be on the horizontal axis and the other to be on the vertical axis. When there is a

2 Chapter 2

response variable and an explanatory variable, the explanatory variable is always placed on the horizontal axis. In cases where there is no explanatory-response variable distinction, either variable can go on the horizontal axis. After drawing the scatterplot by hand or using a computer, the scatterplot should be examined for

an **overall pattern** which may tell us about any relationship between the variables and for **deviations** from it. You should be looking for the **direction**, **form**, and **strength** of the overall pattern. In terms of direction, **positive association** occurs when the variables both take on high values together, while **negative association** occurs if one variable takes high values when the other takes on low values. In many cases, when an association is present, the variables appear to have a **linear relationship**. The plotted values seem to cluster around a line. If the line slopes up to the right, the association is positive; and if the line slopes down to the right, the association is negative. As always, look for **outliers**. The outlier may be far away in terms of the horizontal variable, the vertical variable, or far away from the overall pattern of the relationship.

GUIDED SOLUTIONS

Exercise 2.1

KEY CONCEPTS - explanatory and response variables

a) When examining the relationship between two variables, if you hope to show that one variable can be used to explain variation in the other, remember that the response variable measures the outcome of the study, while the explanatory variable explains changes in the response variable. When you just want to explore the relationship between two variables like score on the math and verbal SAT, then the explanatory-response variable distinction is not important.

In this case, it seems reasonable to view the time spent studying as explaining the grade on the exam. Thus, the grade on the exam is the response and the time spent studying is the explanatory variable. Now try the other parts on your own.

b)

c)

d)

e)

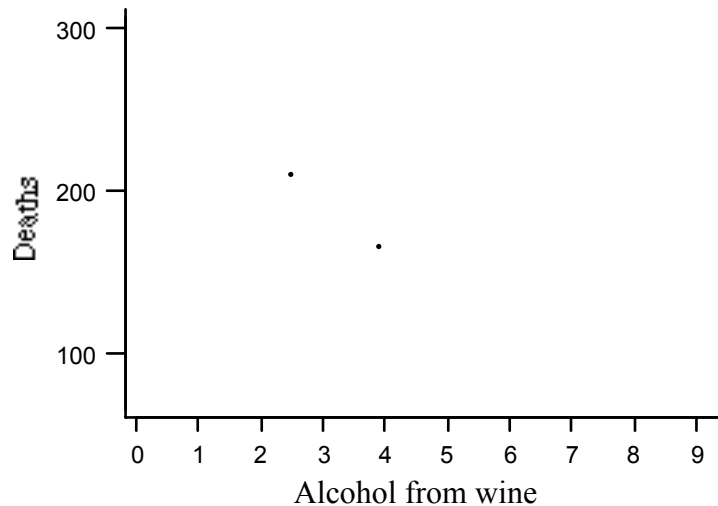
Exercise 2.7

4 Chapter 2

KEY CONCEPTS - drawing and interpreting a scatterplot

a) When drawing a scatterplot, we first need to pick one variable (the explanatory variable) to be on the horizontal axis and the other (the response) to

be on the vertical axis. In this data set we are interested in the "effect" of drinking moderate amounts of wine on yearly deaths from heart disease. So wine consumption is the explanatory variable and deaths from heart disease is the response. We have drawn the points corresponding to Australia and Austria in the plot below. Although you will generally draw scatterplots on the computer, drawing a small one like this by hand makes sure that you understand what the points represent.



b) We are looking for the form and strength of the relationship. Can the relationship be described with a straight line? Section 2.3 discusses formal methods for drawing a straight line through a set of data, but for now just try to draw a straight line to follow the overall pattern in the scatterplot in part (a) above. Do the points seem to follow the line that you have drawn or are there significant deviations from the pattern? How tight is the scatter about that line?

c) Is the association positive or negative? Do countries with higher wine consumption tend to have higher or lower death rates? You need to be careful with the language you use to describe the relationship. In this example the countries may differ on many other factors besides wine consumption, which may explain the lower death rates due to heart attacks. So avoid the use of expressions such as drinking more wine lowers the risk of heart disease, or wine

6 Chapter 2

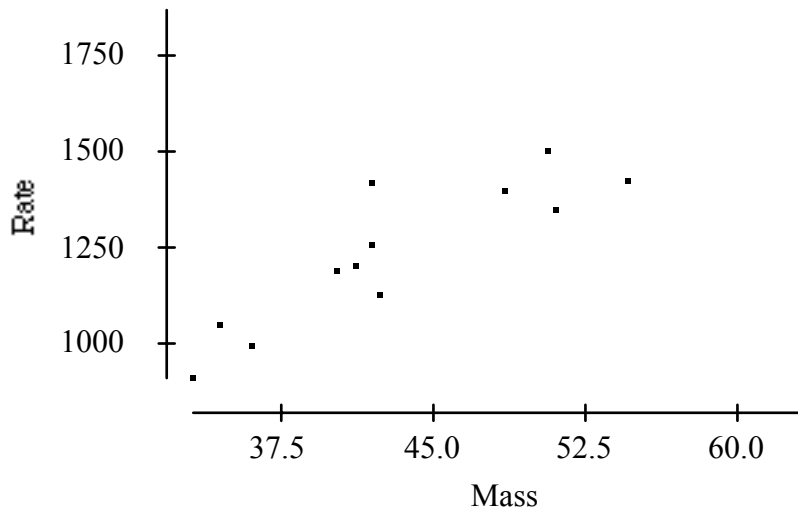
produces a lower risk of heart disease, as these expressions imply causation. Try and explain in simple language what the data has to say.

Exercise 2.11

KEY CONCEPTS - drawing and interpreting a scatterplot, adding a categorical variable to a scatterplot

a) When drawing a scatterplot, we first need to pick one variable (the explanatory variable) to be on the horizontal axis and the other (the response) to be on the vertical axis. In this data set we are interested in the "effect" of lean body mass on metabolic rate. So lean body mass is the explanatory variable and metabolic rate is the response variable in the plot in the following figure. Although you will generally draw scatterplots on the computer, drawing a small one like this by hand makes sure that you understand what the points represent.

The scatterplot with the points for the females is given below. Add the data for the males to this plot using a different color or plotting symbol.



b) Here are some guidelines for examining scatterplots: Do the data show any association? **Positive association** is when the variables both take on high values together. **Negative association** is when one variable takes high values and the other takes on low values. If the plotted values seem to form a line, the variables may have a **linear relationship**. If the line slopes up to the right, the association is positive. If the line seems to slope down to the right, the association is negative. Are there any **clusters** of data? Clusters are distinct groups of observations. As always, look for outliers. An **outlier** may be far away in terms of the horizontal variable or the vertical variable, or far away from the overall pattern of the relationship.

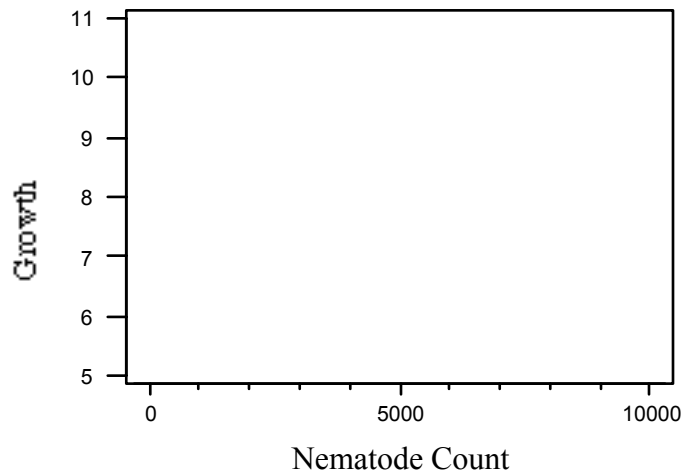
For our plot, is the overall association positive or negative? What is the overall form of the relationship? How strong is the overall relationship?

Is the pattern of the relationship for the men similar to that for the female subjects? If not, how do the male subjects as a group differ from the female subjects as a group?

Exercise 2.16

KEY CONCEPTS - categorical variables in scatter plots

a) In the scatterplot, there should be four observations above each of the levels of nematode count. After adding these points to the graph below, compute the mean of the four observations at each level of nematode count, and put each mean on the graph. Then connect the four means.



b) What sorts of changes do you see in the means as the nematode count increases?

COMPLETE SOLUTIONS

Exercise 2.1

a) A complete solution was provided in the Guided Solutions.

10 Chapter 2

b) We would probably simply want to explore the relationship between weight and height.

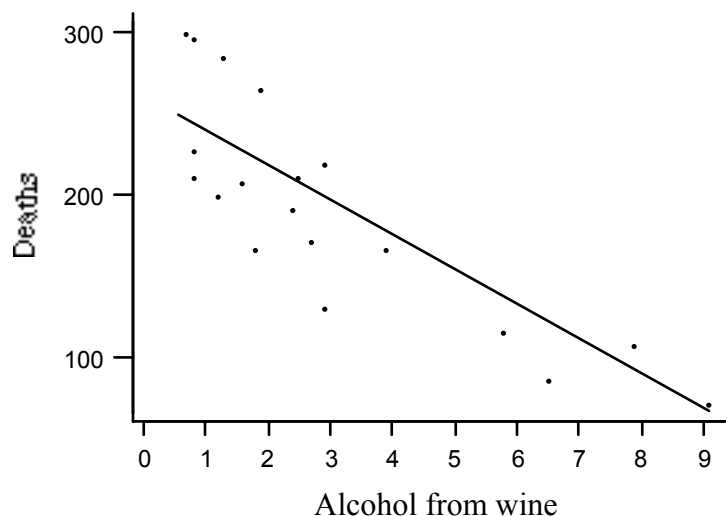
c) We would probably view inches of rain as explaining the yield of corn. Thus, the response is the yield of corn in bushels and the explanatory variable is inches of rain in the growing season.

d) We would probably simply want to explore the relationship between a student's scores on the SAT math exam and scores on the SAT verbal exam.

e) We would probably view a family's income as explaining the years of education their eldest child receives. Thus, the response is the years of education that the eldest child receives and the explanatory variable is the family's income.

Exercise 2.7

a), b)

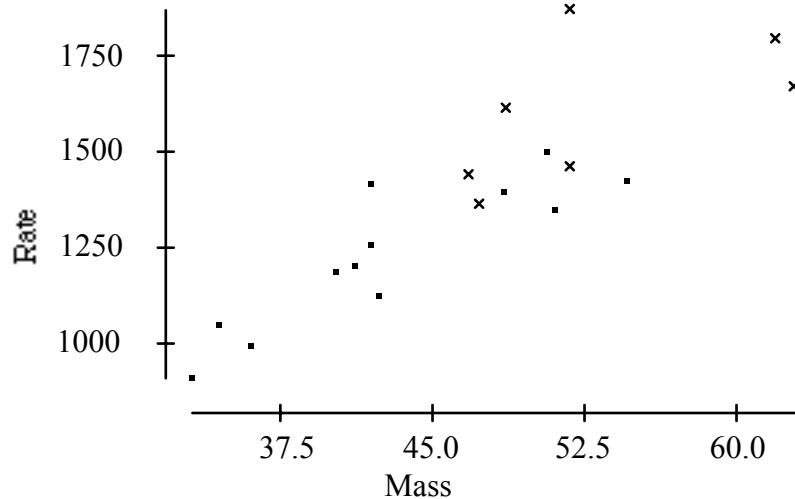


A line does not do a bad job of describing the general pattern. The relationship is moderately strong. In Section 2.2 we will give a numerical measure which describes the strength of the linear relationship.

c) There is a negative association between wine consumption and deaths from heart disease. Those countries in which wine consumption is higher, tend to have a lower rate of deaths from heart disease. (Note: there is nothing in this language which implies causation.)

Exercise 2.11

a) We add the men to the plot. Men are indicated by the x's in the plot on the next page.

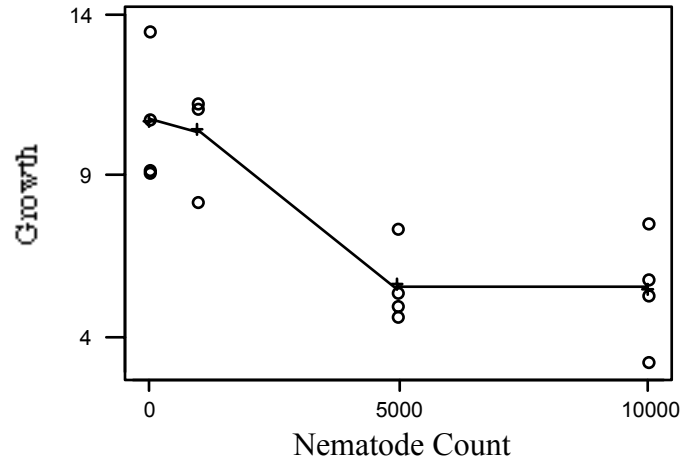


b) As lean body mass increases, or as you move from left to right across the horizontal axis in the scatterplot, the points in the plot tend to rise. This indicates that the association between the variables is positive. The form of the relationship appears to be linear since a straight line seems to be a reasonable approximation to the overall trend in the plot. The relationship is not perfect, but it appears to be moderately strong.

The pattern of the relationship is roughly the same for men and women. The strength of the relationship for females appears to be slightly stronger than for males. The most striking difference between the points corresponding to male and female subjects is that the men are clustered in the upper right of the plot. This is not surprising, since men tend to be larger than women.

Exercise 2.16

a) In the graph below, the circles correspond to the observations and the pluses to the means at each level of nematode count.



b) The level of growth may decrease slightly when going from 0 to a 1000 count, but then it drops off considerably by 5000 and seems to stay at that level until 10,000. In making these statements we are making some assumptions about the average growth between the levels of nematode counts in our experiments (interpolating). We claimed the growth level wasn't changing between 5000 and 10,000 but we have no data between those values. More problematic is when the drop in level occurs to get from the level at 1000 to the level at 5000. It would be helpful to have had an observation at 3000, as the data suggests there was a fairly sharp decrease in growth at some nematode count between 1000 and 5000, but we can't say more.

SECTION 2.2

OVERVIEW

Scatterplots provide a visual tool for looking at the relationship between two variables. Unfortunately our eyes are not good tools for judging the strength of the relationship. Changes in the scale or the amount of white space in the graph can easily affect our judgment as to the strength of the relationship. **Correlation** is a numerical measure we will use to show the strength of **linear association**.

The correlation can be calculated using the formula

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

where \bar{x} and \bar{y} are the respective means for the two variables X and Y , and s_x and s_y are their respective standard deviations. In practice, you will probably compute the value of r using computer software or a calculator that finds r from entering the values of the x 's and y 's. When computing a correlation coefficient there is no need to distinguish between the explanatory and response variables, even in cases where this distinction exists. The value of r will not change if we switch x and y .

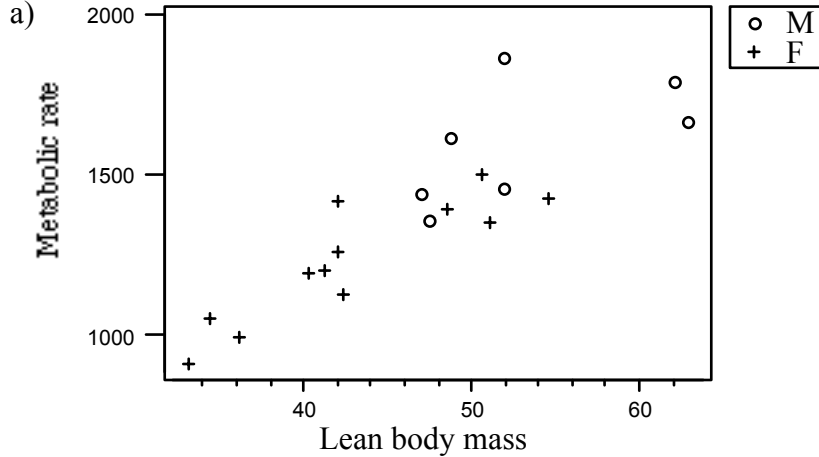
When r is positive it means that there is a positive linear association between the variables and when it is negative there is a negative linear association. The value of r is always between 1 and -1. Values close to 1 or -1 show a strong association while values near 0 show a weak association. As with means and standard deviations, the value of r is strongly affected by outliers. Their presence can make the correlation much different than what it might be with the outlier removed. Finally, remember that the correlation is a

measure of straight line association. There are many other types of association between two variables, but these patterns will not be captured by the correlation coefficient.

GUIDED SOLUTIONS

Exercise 2.20

KEY CONCEPTS - interpreting and computing the correlation coefficient



Should the sign of the correlation coefficient be the same for men and women? Is either relationship "stronger?" Are there outliers in either group that might raise or lower the value of the correlation coefficient?

b) Try and use a computer package or a calculator to compute the value of the correlation coefficient. If you do not have access to a calculator or computer package, the required "hand" computations are illustrated below for the men.

$$\begin{aligned} \bar{x} &= 53.10 & s_x &= 6.69 \\ \bar{y} &= 1600.00 & s_y &= 189.2 \end{aligned}$$

We summarize the calculations for the correlation r in the following table

x	$\frac{x - \bar{x}}{s_x}$	y	$\left(\frac{y - \bar{y}}{s_y}\right)$	$\left(\frac{x - \bar{x}}{s_x}\right) \left(\frac{y - \bar{y}}{s_y}\right)$
62.0	1.33034	1792	1.01480	1.35003
62.9	1.46487	1666	0.34884	0.51100
47.4	-0.85202	1362	-1.25793	1.07178
48.7	-0.65770	1614	0.07400	-0.04867

51.9	-0.17937	1460	-0.73996	0.13273
51.9	-0.17937	1867	1.41121	-0.25313
46.9	-0.92676	1439	-0.85095	0.78862

18 Chapter 2

The sum of the values in the last column above is 3.5524. Thus the correlation is

$$r = 3.5524/6 = 0.592 \text{ for the men.}$$

Now you need to either repeat the above "hand" calculation for the 12 women, or learn how to do the calculation on a computer package or calculator.

Women's correlation coefficient =

c) Mean body mass for men =

Mean body mass for women =

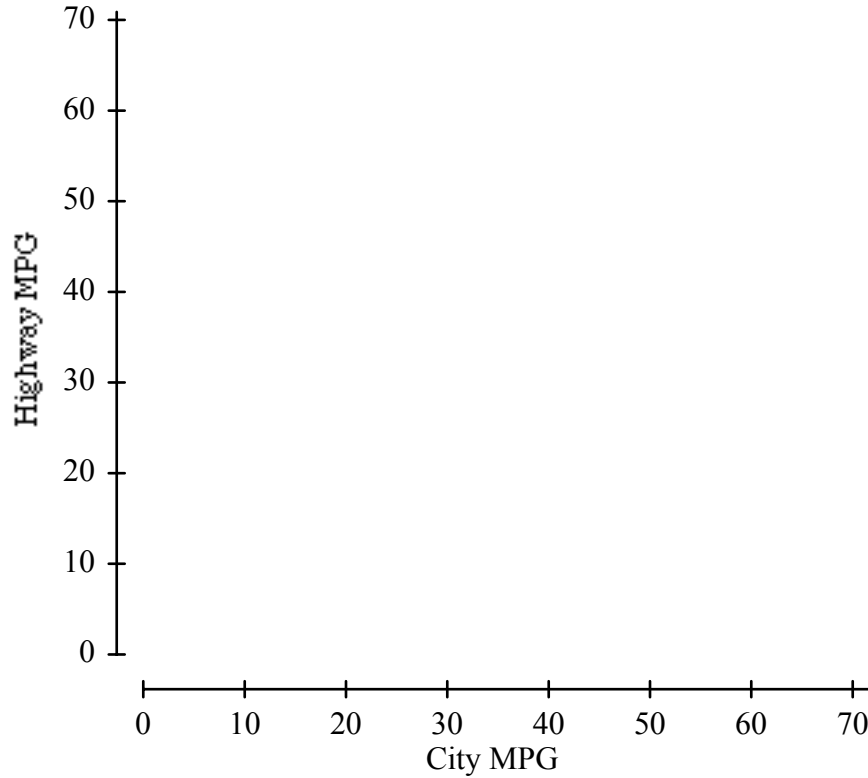
To determine whether the fact that men are heavier on average than women influences the correlation, ask yourself the following. Is the relationship between lean body mass and metabolic rate the same for all weights between 40 and 65 kilograms (the range of all the data)? If the relationship is different for heavier people than lighter people, then the fact that men are heavier could effect the value of the correlation since then men and women might have a different relationship between body mass and metabolic rate (with possibly differing strengths as measured by the correlation coefficient).

d) Is this a linear transformation? What is the effect of a linear transformation on the correlation coefficient?

Exercise 2.25

KEY CONCEPTS - computing the correlation coefficient, the effect of outliers

a) Make your scatterplot in the axes provided.



Does the Insight extend the linear pattern of the other cars, or is it far from the line they form?

b) Compute the correlations and enter the results in the space provided. Use statistical software if available. If you are computing the correlations by hand, you may find it useful to organize your calculations as we did in Exercise 2.20.

Correlation with all observations =

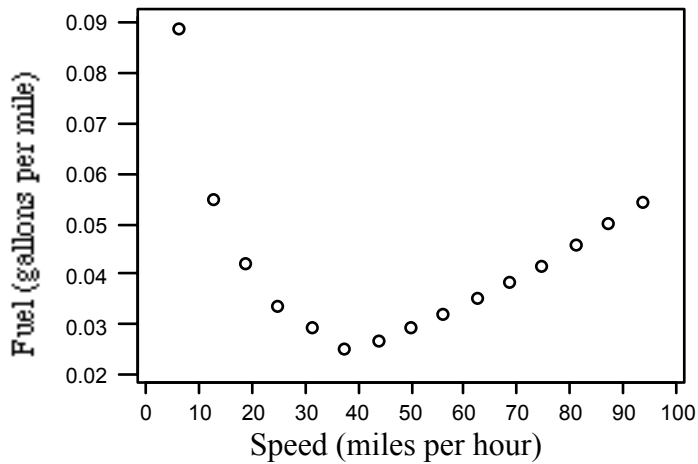
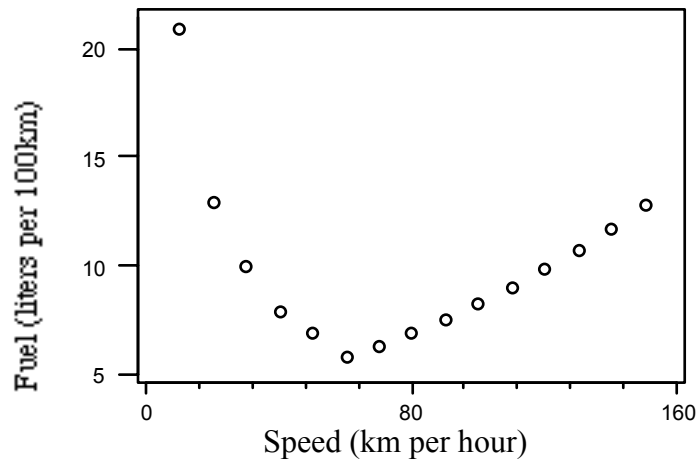
Correlation without the Insight =

Explain the difference in the two values based on your answer to (a).

Exercise 2.29

KEY CONCEPTS - effect of transformations on the correlation coefficient

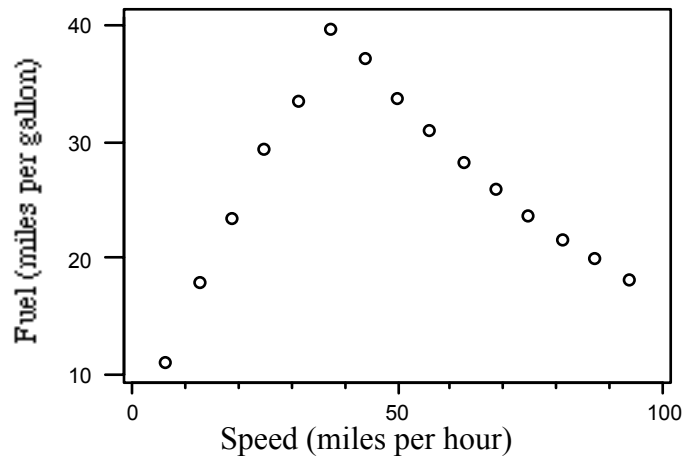
a) The first plot is speed (km / h) vs. fuel (liters / 100 km) and the second plot is speed (miles / hour) vs. fuel (gallons / mile). Both are virtually identical except for the labeling of the axes which corresponds to the units that the variables are measured in.



The transformation to change kilometers to miles is $\text{miles} = \text{kilometers} / 1.609$ and is a linear transformation. What's the transformation to change fuel used in liters/100km to fuel used in gallons/mile? Is it a linear transformation? What is the effect of these two transformations on the numerical value of the correlation coefficient? (**Note:** Since the relationship between speed and fuel consumption

is not linear, do you think the correlation coefficient is a good summary of the "strength" of the relationship?)

b)



What is the transformation to change fuel used in gallons/mile to miles per gallon? What is the transformation to change liters /100 km to miles per gallon? Is this transformation linear? What is the effect on the correlation coefficient?

Note: The relationship is still not linear so the correlation coefficient is again not a very good summary measure.

Exercise 2.33

KEY CONCEPTS - interpreting the correlation coefficient

The problem is that a correlation close to zero and the quote "good researchers tend to be poor teachers, and vice versa" are not the same. What does a

correlation close to zero mean? What would be true about the correlation if "good researchers tend to be poor teachers, and vice versa"?

COMPLETE SOLUTIONS

Exercise 2.20

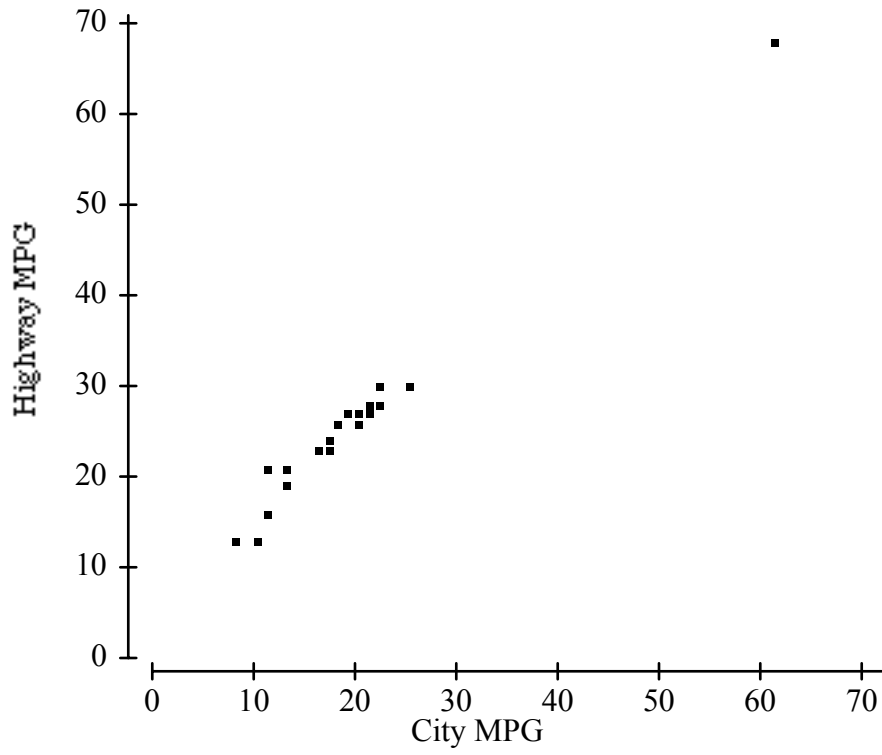
- a) The relationship for the women seems to be tighter around a line. The men's observation of mass = 51.9 and rate = 1867 lowers the value of their correlation.
- b) Men's correlation coefficient = 0.592
Women's correlation coefficient = 0.876.
- c) Mean body mass for men = 53.10.
Mean body mass for women = 43.03

The relationship between lean body mass and metabolic rate is roughly a straight line over the range of the data. Thus the fact that men are heavier than women on average would not, of itself, influence the correlation. However, we do notice that the range of values of lean body mass is larger for women than for men. Because the values are more spread out horizontally for women, this may be partly responsible for the larger correlation.

- d) Kilograms = $2.2 \times$ Pounds is a linear transformation with $a = 0$ and $b = 2.2$. The value of the correlation is unchanged under linear transformations of the variables.

Exercise 2.25

- a)



The Insight (outlier in the upper right corner) appears to extend the linear pattern of the other cars.

b) Using statistical software, we obtained the following.

$$\begin{array}{ll} \text{Correlation with all observations} & = 0.991 \\ \text{Correlation without the Insight} & = 0.956 \end{array}$$

The correlation is closer to 1 when we include the Insight. As we noted in (a), the Insight appears to extend the linear pattern of the other cars. In so doing, it actually strengthens the visual impression of the linear pattern because it is far from the other points in the plot.

Exercise 2.29

a) The transformation here is $\text{fuel}_{(\text{gallons per mile})} = (1.609/378.5) \times \text{fuel}_{(\text{liters per 100 km})}$ and is linear. Since the transformation of x and y are both linear, the correlation coefficient is the same for the original and transformed variables. The value of the correlation coefficient is -0.172 . Since the relationship is clearly nonlinear, and the correlation only measures the strength of a linear relationship, it would be a mistake to interpret this as a negative relationship between fuel consumption and speed.

b) To go from gallons per mile to miles per gallon we need to take the reciprocal $\text{miles per gallon} = 1/\text{gallons per mile}$, or

$$\text{mpg} = 378.5 / (1.609 \times \text{fuel}_{(\text{liters per 100 km})}).$$

This is not a linear transformation, so the numerical value of the correlation coefficient will change. The value of the correlation coefficient is now -0.043 . A value near zero would suggest a weak relationship but that refers to a weak **linear** relationship. There is clearly a very strong relationship between speed and miles per gallon, albeit nonlinear.

Exercise 2.33

If the correlation were close to zero, there would be no particular linear relationship. Good researchers would be just as likely as bad researchers to be good or bad teachers. The statement that "good researchers tend to be poor teachers, and vice versa" implies that the correlation is negative, not zero.

SECTION 2.3

OVERVIEW

If a scatterplot shows a linear relationship which is moderately strong as measured by the correlation, we would like to draw a line on the scatterplot to summarize the relationship. In the case where there is a response and an explanatory variable, the **least-squares regression** line often provides a good

summary of this relationship. A straight line relating y to x has the form $y = a + bx$ where b is the **slope** of the line and a is the **intercept**. The least squares regression line is the straight line $\hat{y} = a + bx$ which minimizes the sum of the squares of the vertical distances between the line and the observed values y . The formula for the slope of the least squares line is

$$b = r \frac{s_y}{s_x}$$

and for the intercept is $a = \bar{y} - b\bar{x}$, where \bar{x} and \bar{y} are the means of the x and y variables, s_x and s_y are their respective standard deviations and r is the value of the correlation coefficient. Typically, the equation of the least squares regression line is obtained by computer software or a calculator with a regression function.

Regression can be used to predict the value of y for any value of x . Just substitute the value of x into the equation of the least squares regression line to get the predicted value for y . Predicting values of y for x values in the range of those x 's we observed is called interpolation and is fine to do. However, be careful about **extrapolation** (using the line for prediction beyond the range of x values covered by the data). Extrapolation may lead to misleading results if the pattern found in the range of the data does not continue outside the range.

Correlation and regression are clearly related as can be seen from the equation for the slope, b . However, the more important connection is how r^2 , the square of the correlation coefficient, measures the strength of the regression. r^2 tells us the fraction of the variation in y that is explained by the regression of y on x . The closer r^2 is to 1 the better the regression describes the connection between x and y .

GUIDED SOLUTIONS

Exercise 2.37

KEY CONCEPTS - review of straight lines

a) In order to give the equation of a straight line, $y = a + bx$, the first thing is to figure out which variable will play the role of x and which will be y . In this problem

$x =$

$y =$

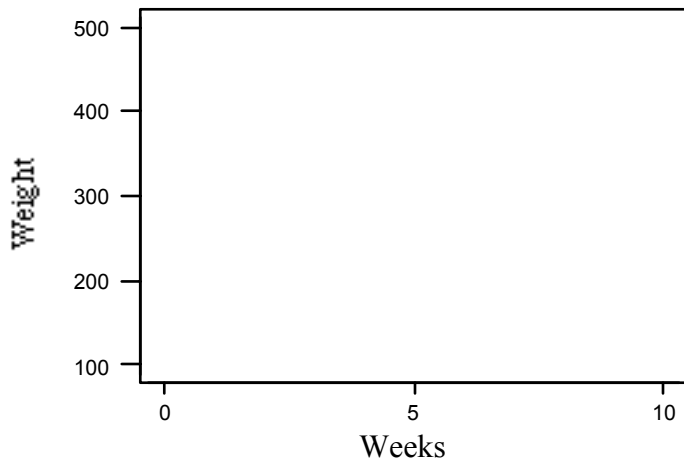
The only thing remaining is to determine which piece of information represents the intercept or value of y when $x = 0$, and which represents the slope, how much y increases with each unit increase in x . In this problem

$a =$

$b =$

The equation of the line is then

b) Draw the graph on the axes below. Remember, drawing a straight line only requires that you find two points on the line and connect them. The value at zero is easy, so you just need to pick a second value.

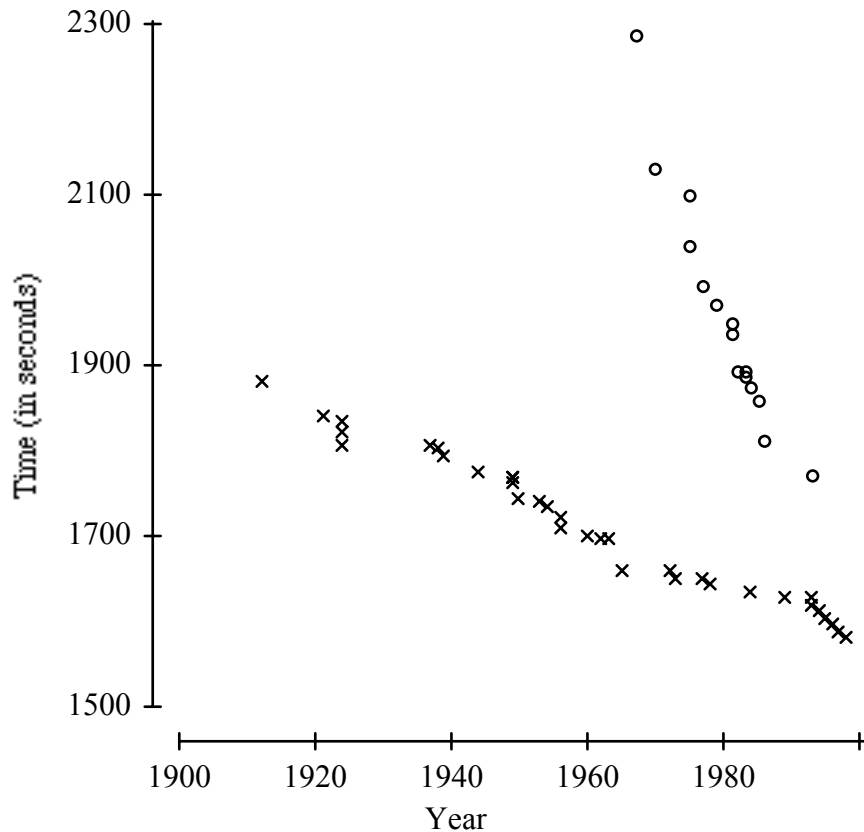


c) When predicting the weight at 2 years, remember the units that age is measured in before doing the calculation. Do you think that a rat's growth pattern will remain the same outside of the range of data? Why or why not?

Exercise 2.45

KEY CONCEPTS - scatterplots, least square regression, interpreting the slope, extrapolation

For purposes of reference, a scatterplot of the world record times (y) against year (x) is given on the next page. The symbol o is used to denote points corresponding to women and x to denote points corresponding to men.



a) Statistical software produced the following information for regressing record time on year.

MEN

	Coefficients	Standard Error	t Stat
Intercept	8167.02	189.40	43.1
Year	-3.29278	0.0966	-34.1

WOMEN

	Coefficients	Standard Error	t Stat
Intercept	41373.2	2717	15.2
Year	-19.9046	1.372	-14.5

Based on this information, give the equation of the least-squares regression lines for predicting record time from year for men and women separately.

MEN

WOMEN

b) What do the slopes tell us about the progress of men and women in the 10,000 meter run? The scatterplot given at the beginning of the problem may help you visualize these slopes.

c) Based on the scatterplot when would you estimate that the women's world record will be the same as the men's?

Exercise 2.51

KEY CONCEPTS - units of measurement for descriptive measures, effect of transformations on summary measures

a) The mean and standard deviation are discussed in Section 2 of Chapter 1 of the text. The correlation is discussed in Section 2 of this chapter. Refer to these sections for help if you have forgotten how to compute the mean, standard deviation, or correlation, or if you have forgotten the units of measurement for these quantities. Compute \bar{x} , s_x , \bar{y} , s_y and r and then enter the values below. These computations can be done easily using software.

$$\begin{array}{ll} \bar{x} = & s_x = \\ \bar{y} = & s_y = \\ r = & \end{array}$$

What are the units of measurement for each of these descriptive measures?

b) You need to figure which of these descriptive measures would have new units of measurement, a new value and what the effect of the linear transformation inches = $(1/2.54) \times$ centimeters has on each quantity.

c) Recall that the slope is $b = r \frac{s_y}{s_x}$. Compute this from the quantities you calculated in parts (a) and (b).

$b =$

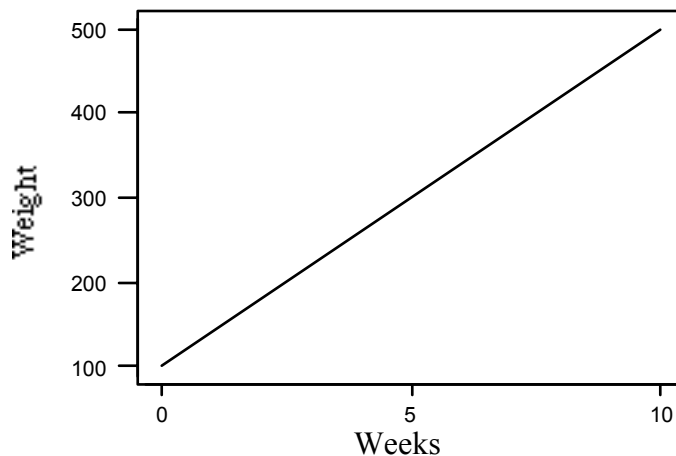
Exercise 2.53**KEY CONCEPTS** - r versus r^2

Recall that r^2 tells us the fraction of the variation in y that is explained by the regression of y on x . Is the number 16% the value of r or r^2 ? If it is r^2 , what is the sign of r ?

COMPLETE SOLUTIONS**Exercise 2.37**

a) The weight is the y variable and age is the x variable. The intercept is the weight at age = 0 or birth and so $a = 100$ grams. The slope is how much the weight goes up with each week and $b = 40$ grams per week. The equation of the line is then $y = 100 + 40x$. Don't forget to include the units in the slope and intercept.

b) At zero the weight is equal to 100 grams, and at 10 weeks the value of weight is $100 + 40(10) = 500$ grams. Just plot these two points on the graph and connect them to get the line.



c) This is an example of extrapolation. Rats do not continue to grow at the same rate for two years, just as people wouldn't necessarily grow at the same rate for 20 years. At 2 years or 104 weeks (remember that x is measured in

weeks), the predicted weight is $100 + 40(104) = 4260$ grams which would be quite a rat!

Exercise 2.45

a) The equations of the least-squares regression line can be determined from the information provided in the guided solution, or you can determine this from your own statistical software. The entries in the coefficients column give the values of the intercept and slope, respectively. We find

MEN

$$\text{Record Time} = 8167.02 - 3.29278 \times \text{Year}$$

WOMEN

$$\text{Record Time} = 41,373.2 - 19.9046 \times \text{Year}$$

b) Both slopes are negative, indicating that the world record times are going down over time. For men, the decrease in the world record time is 3.29278 seconds per year and for women the decrease is 19.9046 seconds per year *over the range of the data*. The world record times for women have been decreasing at a faster rate than for men over the range of the data.

c) The scatterplot suggests that by approximately the year 2000 men and women will have the same world record times. A more careful mathematical calculation using the equations of the least-squares regression lines shows that the times will be the same in 1999. Of course, the year 2000 has passed and the world record time for women is still greater than for men. So much for extrapolation!

Exercise 2.51

a) Using statistical software we found

$$\bar{x} = 95 \qquad s_x = 53.3854$$

$$\bar{y} = 12.6611 \qquad s_y = 8.4967$$

$$r = 0.996$$

The units of measurement are \bar{x} = minutes, \bar{y} = centimeters, s_x = minutes, s_y = centimeters, and r = unitless.

b) The linear transformation inches = $(1/2.54) \times$ centimeters changes the mean from \bar{y} in centimeters to $(1/2.54) \times \bar{y}$ in inches. Likewise, it changes the standard deviation from s_y in centimeters to $(1/2.54) \times s_y$ in inches. The correlation r is unchanged because it is unitless. Thus we get

$$\text{new } \bar{y} \text{ in inches} = 12.6611/2.54 = 4.9847$$

$$\text{new } s_y \text{ in inches} = 8.4967/2.54 = 3.3452$$

$$\text{new } r = 0.996.$$

c) We compute

$$b = r \frac{s_y}{s_x} = 0.996 \frac{3.3452}{53.3854} = 0.0624$$

Exercise 2.53

16% is the value of r^2 . Hence $r^2 = 0.16$ and $r = \sqrt{0.16} = 0.4$. We take the positive square root because the problem states that, in general, students who attended a higher percentage of their classes earned a higher grade which corresponds to a positive association.

SECTION 2.4

OVERVIEW

Plots of the **residuals**, which are the differences between the observed and predicted values of the response variable, are very useful for examining the fit of a regression line. Features to look out for in a residual plot are unusually large values of the residuals (outliers), nonlinear patterns, and uneven variation about the horizontal line through zero (corresponding to uneven variation about the regression line).

The effects of **lurking variables**, variables other than the explanatory variable which may also affect the response, can often be seen by plotting the residuals versus such variables. Linear or nonlinear trends in such a plot are evidence of a lurking variable. If the time order of the observations is known, it is good practice to plot the residuals versus time order to see if time can be considered a lurking variable.

Influential observations are individual points whose removal would cause a substantial change in the regression line. Influential observations are often outliers in the horizontal direction but they need not have large residuals.

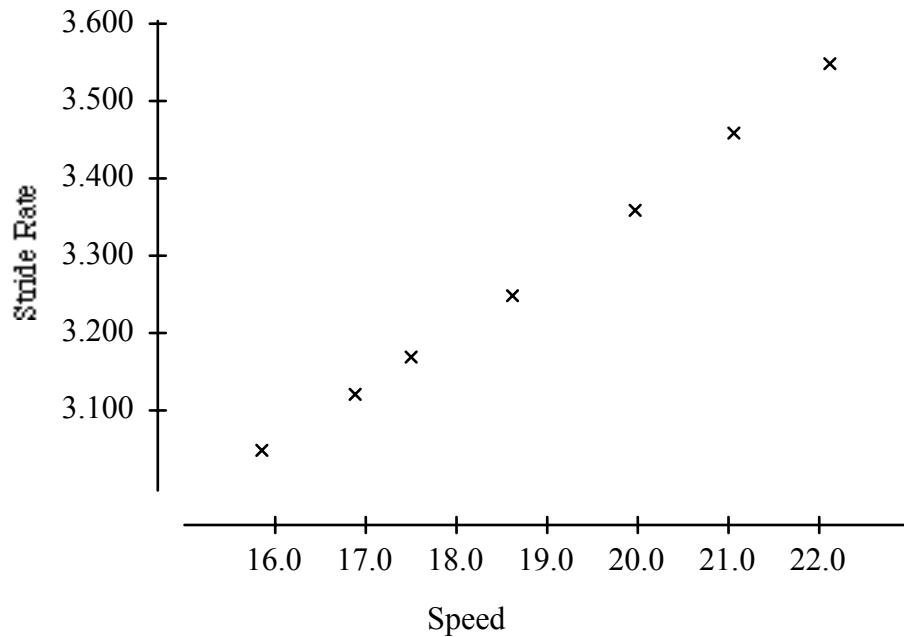
Correlation and regression must be interpreted with caution. Plots of the data, including residual plots, help make sure the relationship is roughly linear and help to detect outliers and influential observations. The presence of lurking variables can make a correlation or regression misleading. *Always remember*

that association, even strong association, does not imply a cause-and-effect relationship between two variables.

A correlation based on averages is usually higher than if we had data for individuals. A correlation based on data with a restricted range is often lower than would be the case if we had observed the full range of the variables.

GUIDED SOLUTIONS**Exercise 2.65****KEY CONCEPTS** - residuals

a) If you are using statistical software, you should enter the data and use the software to create a scatterplot. Although we are giving you many of the plots, it is a good idea to make sure you understand in each plot why one variable was designated the response and the other the explanatory variable.



What is the general trend in your scatterplot? Does it appear to be adequately described by a straight line or is curvature present?

b) If you do not have access to a computer or a calculator that will compute the least-squares regression line, you will have to do computations by hand. If you are using statistical software (or a calculator that will compute the equation of the least-squares regression line), you should enter the data and use the software (calculator) to calculate the equation of the least-squares regression line. Write the equation in the space provided below.

c) Recall that the residual for a given speed x is

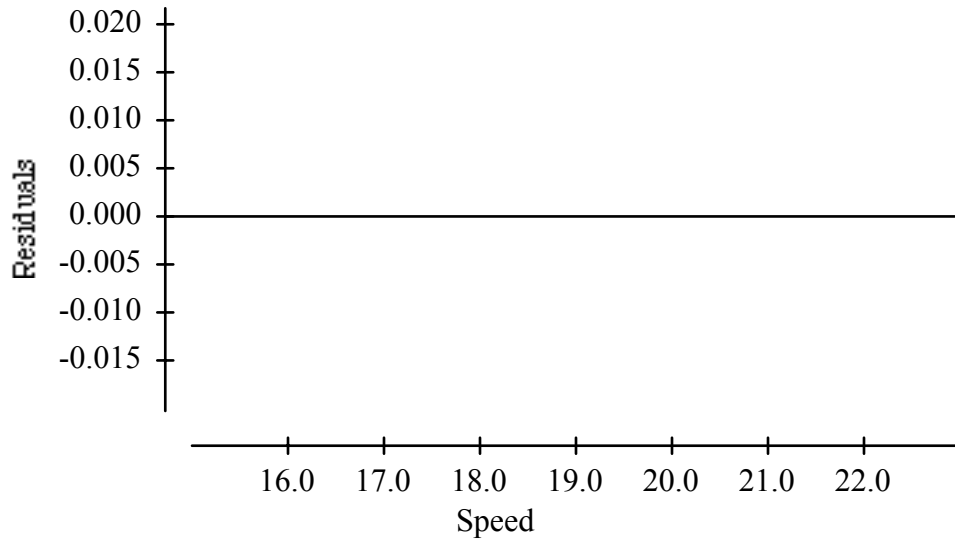
$$\begin{aligned} \text{residual} &= \text{observed stride rate} - \text{predicted stride rate} \\ \text{predicted stride rate} &= a + bx \end{aligned}$$

and $a + bx$ is the equation of the least-squares regression line from (b). Statistical software can be used to calculate the residuals directly. If you use statistical software, fill in the values in the residual column only in the table below. If you are calculating the residuals by hand, complete the table below to aid you in systematically calculating the residuals.

speed	observed stride rate	predicted stride rate	residual = observed - predicted
15.86	3.05		
16.88	3.12		
17.50	3.17		
18.62	3.25		
19.97	3.36		
21.06	3.46		
22.11	3.55		

Add the entries in the residual column to verify that the residuals sum to 0 (except for rounding error).

d) If you are using statistical software, the software should allow you to create a plot of the residuals directly. Plot the seven residuals on the axes below.



48 Chapter 2

Does it appear that the residuals have a random scatter or is there a pattern present? Do you think that a linear fit is appropriate for these data?

Do any of the points in your plot appear to be influential? Ask yourself if their removal would cause a substantial change in the least-squares line (or if they appear to be outliers in the horizontal direction).

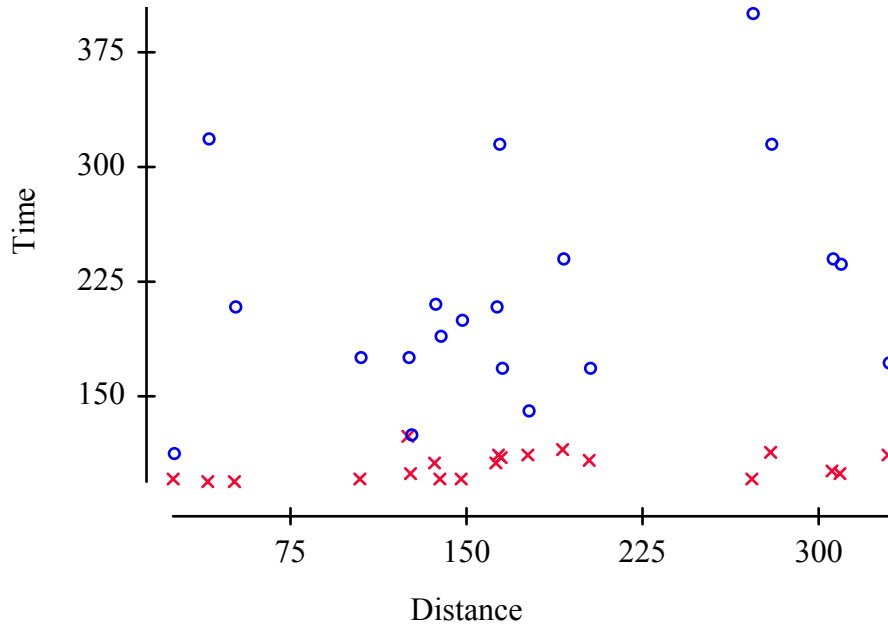
Note: For classes that discussed the topic of DFFITS, you can also calculate the DFFITS to determine if any observations are influential.

Are you provided with any information that would allow you to plot the observations against the time they were made? You might ask yourself, in what form would this information have to be in order to make such a plot (think about what the actual data values represent).

Exercise 2.74

KEY CONCEPTS - regression including a categorical variable, lurking variable

a) A plot of the data is given below. The x corresponds to the right hand times and the circles to the left hand times.



b) In describing the pattern, look for the most striking details. Are there any clear trends for the left-hand observations? The right-hand observations? Are there any differences between the left-hand and right-hand observations?

c) The calculations of the least-squares regression lines of time on distance for each hand are best done using statistical software. Be sure to analyze the left-hand and right-hand observations separately. If you must do them by hand, approach the calculations systematically. Write the two equations of the regression lines in the space provided below.

regression line for left-hand data:

regression line for right-hand data:

Now draw these lines carefully on the scatterplot in (a).

Which line appears to do the better job of predicting time from distance? You should compute the correlation r (or better yet, r^2) associated with each regression line as a possible numerical summary describing the success of the two lines. Can you reconcile the values of your numerical summaries with a visual inspection of the plot?

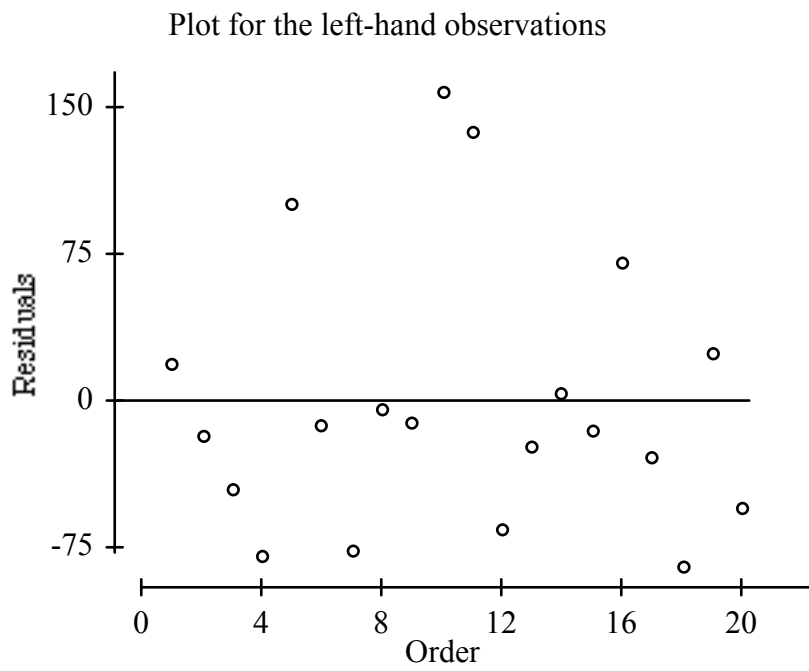
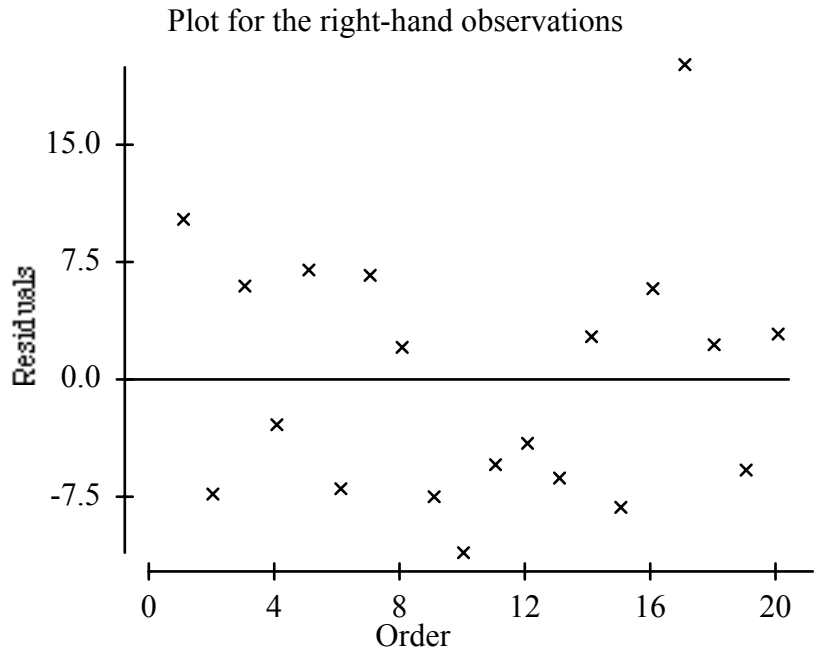
d) If you are plotting manually, you first calculate the residuals using

$$\text{residual} = \text{time} - [99.364 + 0.028(\text{distance})]$$

for the right-hand observations and

$$\text{residual} = \text{time} - [171.548 + 0.262(\text{distance})]$$

for the left-hand observations. Then you plot your results, remembering that the vertical axis represents the values of the residuals and the horizontal axis time. If you are using statistical software, your plots of the residuals from each regression against the time order of the trials should look similar to those given on the next page. Note the difference in the scales for the vertical axes in the two plots.



Are any trends or patterns visible in these plots?

Exercise 2.76**KEY CONCEPTS** - correlations based on averaged data

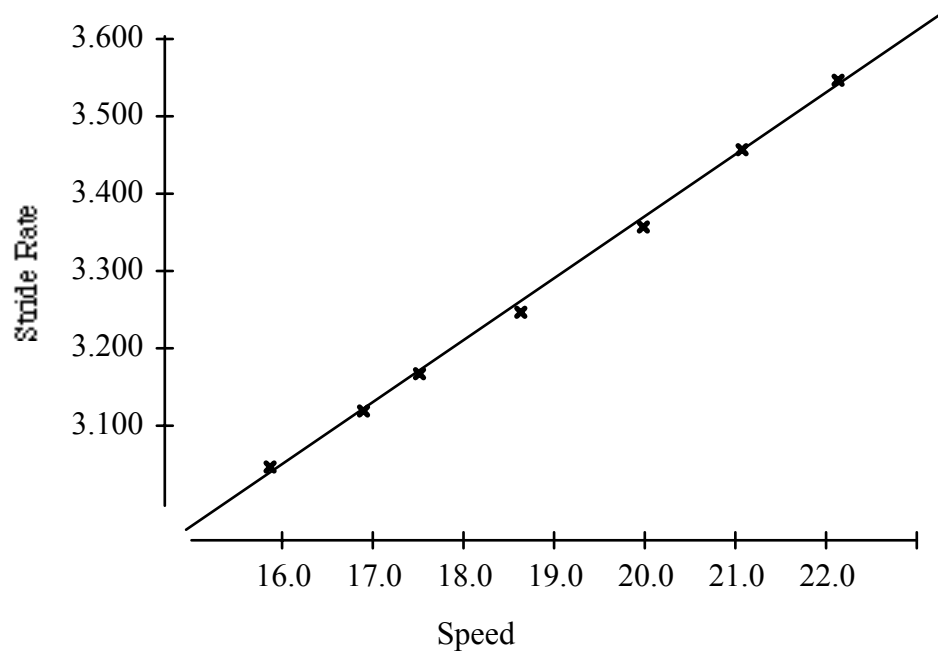
Computing the correlation is most easily done using statistical software or a calculator that computes correlation. Regarding whether the correlation would increase or decrease if we had data on the individual stride rates of all 21 runners, note that a correlation based on averages over many individuals is usually higher than the correlation between the same variables based on data for individuals.

COMPLETE SOLUTIONS**Exercise 2.65**

- a) A plot of the data is given in the Guided Solutions. A straight line "appears" to describe the data quite well.
- b) The least-squares regression line

$$\text{Stride Rate} = 1.77 + 0.08(\text{Speed})$$

A scatterplot with this line drawn in is given below.



c) For each of the speeds given, we substitute the value into the equation of the least-squares regression line to compute the predicted value of stride rate. The residual is then computed as

$$\text{residual} = \text{observed stride rate} - \text{predicted stride rate}$$

For example, for a speed of $x = 15.86$ we compute

$$\text{predicted stride rate} = 1.77 + 0.08(15.86) = 1.77 + 1.27 = 3.04$$

and hence

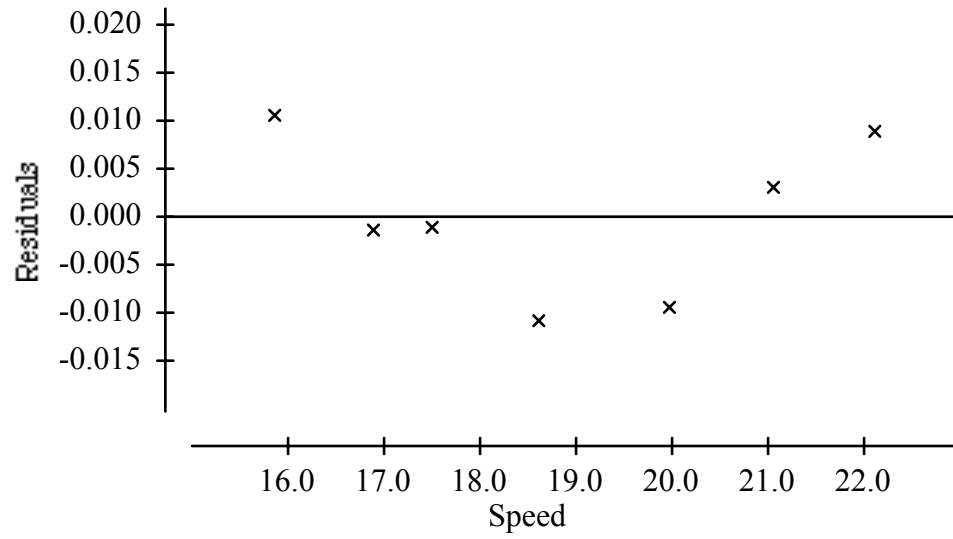
$$\text{residual} = \text{observed stride rate} - \text{predicted stride rate} = 3.05 - 3.04 = 0.01.$$

We summarize the results in the table below (to two decimal places)

speed	observed stride rate	predicted stride rate	residual
15.86	3.05	$1.77 + 0.08(15.86) = 3.04$	0.01
16.88	3.12	$1.77 + 0.08(16.88) = 3.12$	0.00
17.50	3.17	$1.77 + 0.08(17.50) = 3.17$	0.00
18.62	3.25	$1.77 + 0.08(18.62) = 3.26$	-0.01
19.97	3.36	$1.77 + 0.08(19.97) = 3.37$	-0.01
21.06	3.46	$1.77 + 0.08(21.06) = 3.46$	0.00
22.11	3.55	$1.77 + 0.08(22.11) = 3.54$	0.01

While these calculations can be done by hand, they are much more easily obtained using statistical software and should agree (to two decimal places) with the above. If you sum the entries in the residual column, you can easily see they sum to 0.

d) A plot of the residuals against speed is given below



The pattern is curved (U-shaped) and indicates that the linear fit is not completely adequate. Recall that the scatterplot in (a) suggested that the linear fit was quite good. This demonstrates that the residual plot can be more informative than the scatterplot concerning the adequacy of the linear fit.

There do not appear to be any influential observations in the plot. Since we are not told the time at which observations were made we cannot plot the residuals against the time observations were made. Note that since observations are actually the averages for 21 runners, we would have to know that all observations on the 21 runners at a given speed were made at the same time in order to even be able to make such a plot.

Note: For classes that discussed DFFITS, the values of DFFITS are given below. These are most easily computed using statistical software.

Speed	Stride Rate	DFFITS
15.86	3.05	1.6635023
16.88	3.12	-0.08839246
17.50	3.17	-0.05792908
18.62	3.25	-0.59243516
19.97	3.36	-0.55913409
21.06	3.46	0.24486537
22.11	3.55	1.4562468

The first and last entries listed are the largest values of DFFITS and hence the most influential observations. Neither point appears to be particularly influential, however.

Exercise 2.74

- The plot is given in the Guided Solutions.
- There does not appear to be any clear relation between distance and time. The most striking feature of the plot is that nearly all the left-hand observations (the o's) lie above the right-hand observations (the x's). This means that left-hand trials had longer times for a given distance than right-hand trials. The subject performed consistently better with the right-hand, suggesting the subject is right-handed.
- Least-squares regression for the right-hand observations.

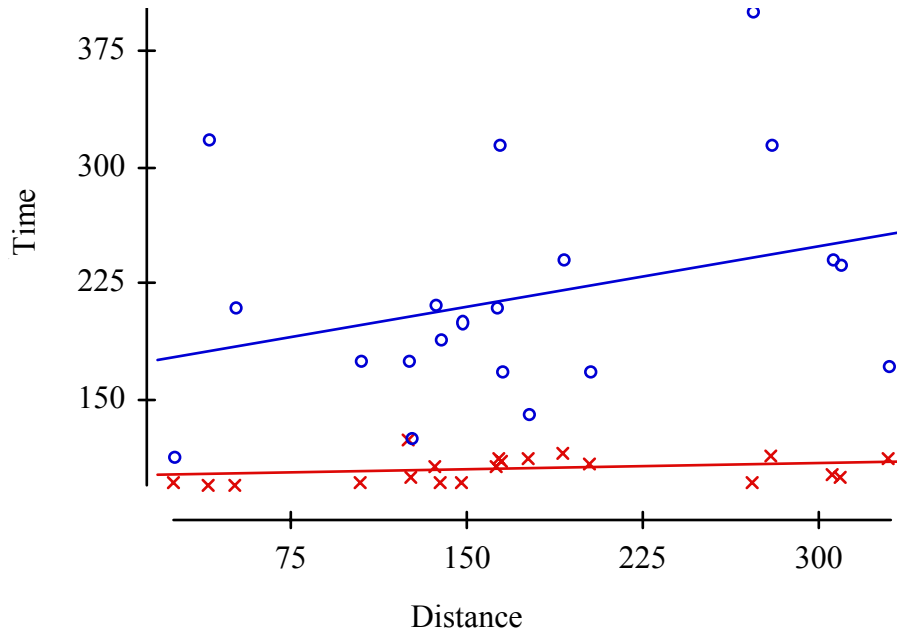
$$\text{time} = 99.364 + 0.028(\text{distance})$$

58 Chapter 2

Least-squares regression for the left-hand observations.

$$\text{time} = 171.548 + 0.262(\text{distance})$$

These lines have been drawn on our scatterplot, reproduced on the next page.



Visually, the regression line for the right-hand data appears to do a better job of predicting time from distance, since the x's appear to be more tightly clustered about this line than the o's are about the regression line for the left-hand data. This would suggest that the value of r^2 would be much higher for the right-hand data. The regression for the right-hand observations has $r^2 = 9.3\%$, i.e., the least-squares regression of time on distance explains 9.3% of the variation in the time values. The regression for the left-hand observations has $r^2 = 10.1\%$, i.e., the least-squares regression of time on distance explains 10.1% of the variation in the time values. Thus the regression for the left-hand observations actually explains a higher percentage of the variation in time than for the right-hand observations.

For the right-hand data there is little variation in times to explain. The times are all quite close to the mean value of 104.25 and this mean is a pretty good predictor of times *without* using distance as an explanatory variable (this is why the least squares regression line is almost horizontal and why r^2 is so low. Distance adds little information to help with predictions). For the left-hand the r^2 value agrees with our visual impression of the plot. There is a lot of variability and distance explains little, resulting in a low r^2 .

d) From the equations of the two least-squares regression lines, we can calculate the residuals using the general expression

residual = observed value of time - predicted value of time

For the right-hand observations, this becomes

$$\text{residual} = \text{time} - [99.364 + 0.028(\text{distance})]$$

For the left-hand observations, this becomes

$$\text{residual} = \text{time} - [171.548 + 0.262(\text{distance})]$$

The following tables summarize the results of the above calculations.

Order	residuals(right-hand)	residuals(left-hand)
1	10.2369110	18.503049
2	-7.2858363	-17.829839
3	5.9622312	-44.786498
4	-2.9367652	-79.600335
5	7.0157385	100.708530
6	-7.0176886	-11.607554
7	6.6488317	-76.685980
8	2.0273459	-4.184077
9	-7.5505414	-10.278813
10	-11.049462	158.350190
11	-5.5037426	137.909280
12	-4.0652097	-65.033543
13	-6.3311991	-22.997814
14	2.7628582	3.620668
15	-8.1009221	-14.377606
16	5.7144264	71.165760
17	20.0824860	-28.422228
18	2.2988909	-84.930362
19	-5.8267679	24.920744
20	2.9184153	-54.443579

These residuals can then be plotted against the order in which the observations were taken. The plots are given in the Guided Solutions.

There is no clear pattern in either plot that would suggest the subject got better in later trials due to learning or got worse due to fatigue. Time order does not appear to be a lurking variable.

Exercise 2.76

The value of $r = 0.999$.

This correlation is based on averaged data, namely the average stride rates of 21 runners at each of the seven values of speed (note that such data might have been collected by having the runners run on a treadmill where speed can be controlled). If we had the data on the individual stride rates of all 21 runners, we would expect the correlation to decrease (be less than 0.999).

SECTION 2.5

OVERVIEW

An observed association between two variables can be due to several things. It can be due to a **cause-and-effect** relationship between the variables. It can also be due to the effects of **lurking variables**, i.e., variables not directly studied that may effect the response and possibly the explanatory variable. Lurking variables may operate through **common response**, in which case changes in both the explanatory and response variables are caused by changes in the lurking variable. Lurking variables may also cause **confounding**, in which case both the explanatory variable and the lurking variables cause changes in the response, but we cannot distinguish their individual effects.

The best way to determine if an association is due to a cause-and-effect relationship between the explanatory variable and the response is through an **experiment** in which we control the influences of other variables. In the absence of good experimental evidence, be cautious in accepting claims of causation. Good evidence requires an association that appears consistently in many studies, a clear explanation for the alleged cause-and-effect relationship, and careful examination of possible lurking variables.

GUIDED SOLUTIONS

Exercise 2.82

KEY CONCEPTS - explaining causation, lurking variables

Ask yourself the following questions.

- Was the study an **experiment** in which the influences of other variables were controlled?
- If the study was not an experiment,

Is there information that the observed association appears consistently in many studies?

Is there a clear explanation for the alleged cause-and-effect relationship?

Is there evidence that possible lurking variables have been ruled out as possible causes?

Exercise 2.87

KEY CONCEPTS - lurking variables

Ask yourself, what sorts of students are likely to study foreign language for at least two years. How are such students likely to do in other subjects?

Exercise 2.89

KEY CONCEPTS - evidence for causation in studies that are not experiments

In order to determine what kinds of information you would seek in records, ask yourself

Would the records contain any information that the observed association appears consistently in many studies or settings?

Would the records provide a clear explanation for the alleged cause-and-effect relationship?

Would information in the records allow you to identify or rule out possible lurking variables as causes?

COMPLETE SOLUTIONS

Exercise 2.82

The data are intriguing since they are based on such a large number of operations, but by themselves they do not prove that anesthetic C is causing more deaths than the others. The main weakness of the study is that it was not an experiment in which the influences of other factors were controlled. Some case might still be made for causation, but we would need to know that similar results have been observed in other studies, would need a clear explanation as to why anesthetic C might cause more deaths than the others, and would need evidence that the effects of lurking variables have been ruled out. Unfortunately, no such information is given.

A possible lurking variable that needs to be ruled out is the type of operations for which the anesthetics are used. Is anesthetic C used more often than the others in operations involving long, difficult, and risky surgery? If so, this might explain why the death rate is higher for anesthetic C.

Exercise 2.87

The explanatory variable in this study is whether or not a student has studied a foreign language for at least two years. The response variable is student's score on an English achievement test.

Probably the most important lurking variable is “quality of student” (how hardworking the student is, innate intelligence, innate language ability). Students that are willing to take at least two years of foreign language are also likely to be more serious or talented students that are likely to work hard or do well in other subjects.

Exercise 2.89

The records are not likely to provide us with a clear explanation as to why anesthetic C would have a higher death rate. This probably requires a medical, biological, or chemical explanation.

The records would also allow us to see if the anesthetics are used for different purposes. Perhaps certain anesthetics are used in simple, routine types of surgery, while others are used in complicated, more risky types of surgery. In general, we should look for other patterns of association to identify lurking variables.

The records would also allow us to see if the pattern is in a variety of settings which would allow us to rule out possible lurking variables. We might look to see if the pattern repeats itself if we restrict to cases from specific geographic regions, specific types of hospitals, specific types of operations, or specific types of patients.

SECTION 2.6**OVERVIEW**

Many relationships between two quantitative variables are nonlinear rather than linear. Sometimes, nonlinear relationships can be changed into linear relationships by **transforming** one or both variables. **Power transformations** consist of transforming a variable t to the variable t^p and are the most commonly used transformations. It is sometimes convenient to consider the **ladder of power transformations** that corresponds to transforming t to

$$\frac{t^p - 1}{p}$$

The logarithm $\log t$ fits into the ladder of power transformations and corresponds to $p = 0$.

A function $f(t)$ is **monotonic** if it changes in one direction (only increases or only decreases) as t increases. The power transformation is monotonic when the variable we are transforming takes only positive values. In this case there is an inverse transformation that returns the transformed data back to their original

values. The effect of power transformations on data becomes stronger as we move farther away from a linear transformation, i.e., as p moves farther away from the value 1.

When we have reason to believe that data are governed by some mathematical model, power transformations are very useful. The **exponential growth model** $y = ab^x$ becomes linear when we plot $\log y$ against x . The **power law model** $y = ax^p$ becomes linear when we plot $\log y$ against $\log x$.

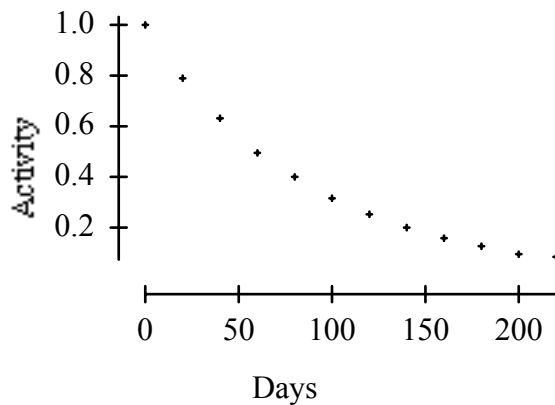
We can fit exponential growth and power models to data by first transforming the data so that the relationship of the transformed variables looks linear, then fitting the least-squares regression line to the transformed data, and finally doing the inverse transformation.

GUIDED SOLUTIONS

Exercise 2.93

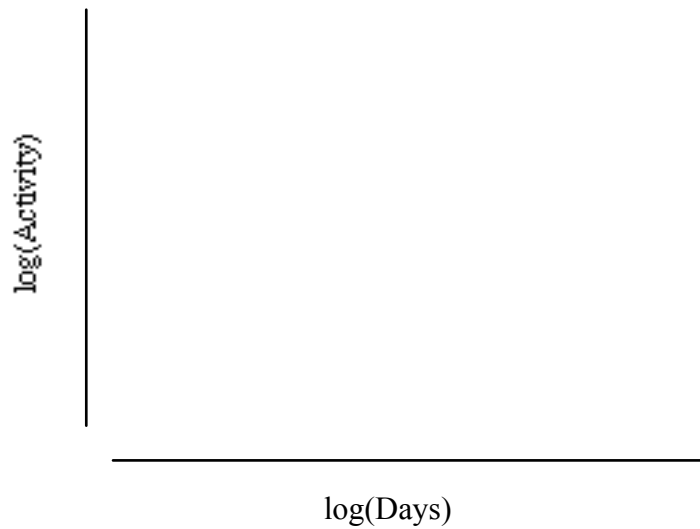
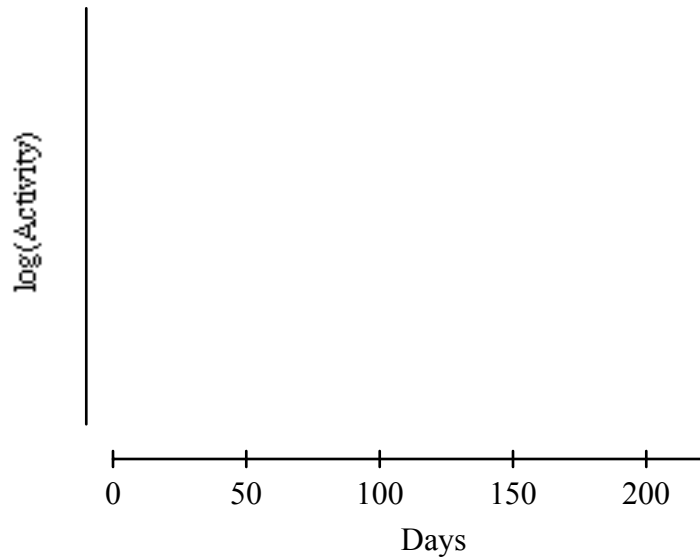
KEY CONCEPTS - exponential growth and power models

a) In this example we would consider Activity to be the response variable (y) and Days to be the explanatory variable (x). A scatterplot of these variables is given below.



This has a distinct curvature and is nonlinear. To determine if an exponential growth model is more appropriate, we can plot $\log(\text{Activity})$ against Days. If the plot is linear with a positive slope, then the exponential growth model provides a good description of the data. If the plot is linear with a negative slope, then the exponential decay model provides a good description of the data. To determine if a power law model is more appropriate, we can plot

$\log(\text{Activity})$ against $\log(\text{Days})$. If the plot is linear, then the power law model provides a good description of the data. Make both plots and determine which model best describes the data. You can either make the plots by hand in the space provided on the next page, or you can use statistical software.



b) Fit a least-squares regression line to the transformed data. What is the equation of the line in terms of the transformed data?

Equation:

Now use the appropriate inverse transformations to express the equation in terms of the original data. Recall that the inverse transformation of $\log(x)$ is $10^{\log(x)}$.

Equation:

c) Is there any feature of the scatterplots you made in (a) that would indicate whether the data might be from actual laboratory experiments or calculated from a theoretical model?

Exercise 2.105

KEY CONCEPTS - transformations, power law

To fit a power law, first compute $\log(\text{Weight})$ and $\log(\text{Lifespan})$. Then fit a least-squares regression model to the logarithms. Write the equation of your regression model below.

Equation: $\log(\text{Lifespan}) = \underline{\hspace{2cm}} + \underline{\hspace{2cm}} \times \log(\text{Weight})$

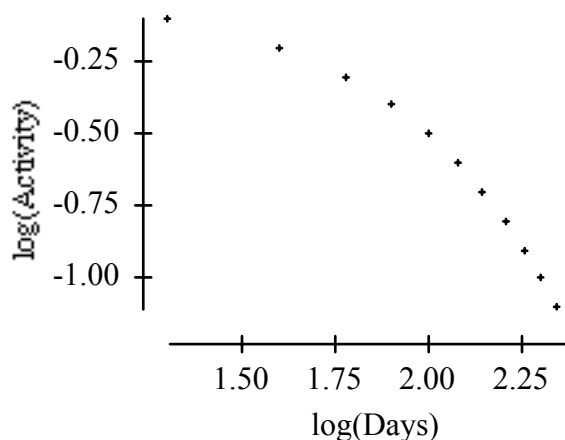
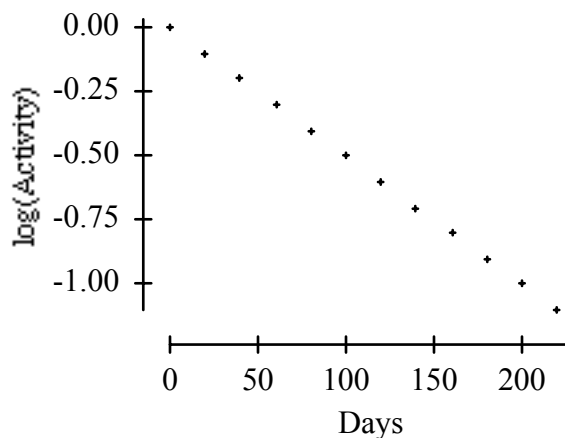
Recall that the slope of this model corresponds to the power p of the power law model $\text{Lifespan} = a \times (\text{Weight})^p$. Is the slope approximately 0.2? You may also want to examine a plot of $\log(\text{Lifespan})$ against $\log(\text{Weight})$ to assess whether a power law model seems reasonable. What do you conclude?

Substitute a weight of 143 into the least-squares regression model to predict $\log(\text{Lifespan})$, then raise 10 to this power to get the predicted Lifespan for humans. What do you find?

COMPLETE SOLUTIONS

Exercise 2.93

a) The plots of $\log(\text{Activity})$ against Days and $\log(\text{Activity})$ against $\log(\text{Days})$ are given on the next page.



The plot of $\log(\text{Activity})$ against Days is a straight line with a negative slope. This suggests that an exponential decay model is a good description of the data.

b) Using statistical software, we found the equation of the least-squares line for $\log(\text{Activity})$ as a function of Days to be

$$\text{Equation: } \log(\text{Activity}) = -0.00002907 - 0.00501799 \times \text{Days}$$

We apply the inverse transformation on each side of this equation by raising 10 to power equal to the quantity given on each side. This yields

$$\text{Equation: } \text{Activity} = 10^{-0.00002907 - 0.00501799 \times \text{Days}}$$

c) The plot of $\log(\text{Activity})$ against Days is nearly a perfect straight line. The fit is so good that it is more consistent with data calculated from a theoretical model than data from an actual laboratory experiment. One would expect data

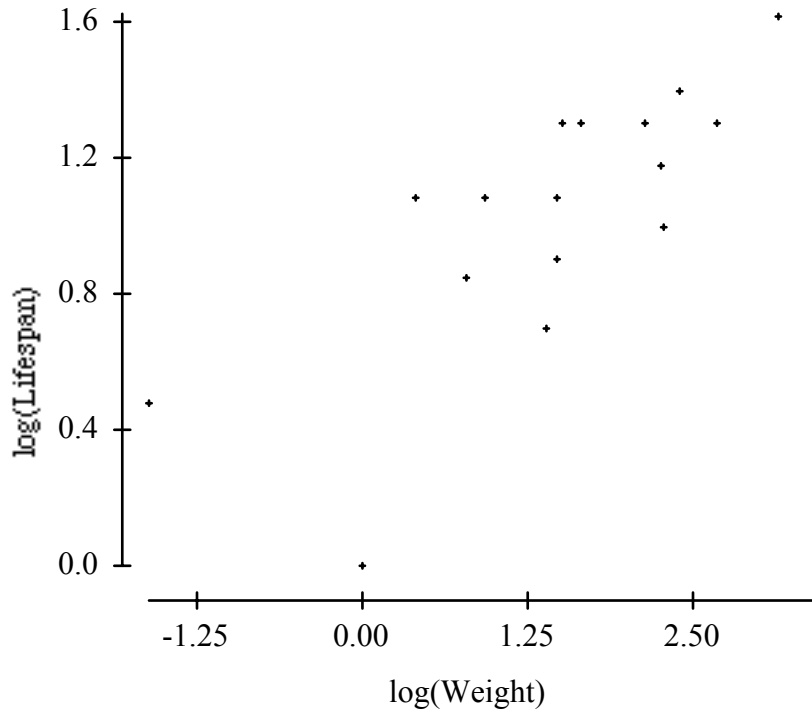
from an actual experiment to contain some measurement error and hence not to have such a perfect fit.

Exercise 2.105

Using statistical software we obtain the following equation for the least-squares regression of $\log(\text{Lifespan})$ on $\log(\text{Weight})$.

$$\text{Equation: } \log(\text{Lifespan}) = 0.667 + 0.257 \times \log(\text{Weight})$$

The slope is 0.257, so this yields a power law model with $p = 0.257$. This is reasonably close to 0.2. A plot of $\log(\text{Lifespan})$ against $\log(\text{Weight})$ is given below.



The general pattern is linear, and since the fitted model has a power near 0.2, a power law model with power 0.2 may not be an unreasonable model.

If we substitute a Weight of 143 into our least-squares regression equation, we would predict

$$\log(\text{Lifespan}) = 0.667 + 0.257 \times \log(143) = 1.22.$$

Thus we would predict

$$\text{Lifespan} = 10^{1.22} = 16.6 \text{ years.}$$

Obviously humans are an exception to the rule.