

Chapter 8

Introduction to linear regression

8.1 Fitting a line, residuals, and correlation

8.2 Least squares regression

8.3 Types of outliers in linear regression

8.4 Inference for linear regression

Linear regression is a very powerful statistical technique. Many people have some familiarity with regression just from reading the news, where straight lines are overlaid on scatterplots. Linear models can be used for prediction or to evaluate whether there is a linear relationship between two numerical variables.



For videos, slides, and other resources, please visit
www.openintro.org/os

8.1 Fitting a line, residuals, and correlation

It's helpful to think deeply about the line fitting process. In this section, we define the form of a linear model, explore criteria for what makes a good fit, and introduce a new statistic called *correlation*.

8.1.1 Fitting a line to data

Figure 8.1 shows two variables whose relationship can be modeled perfectly with a straight line. The equation for the line is

$$y = 5 + 64.96x$$

Consider what a perfect linear relationship means: we know the exact value of y just by knowing the value of x . This is unrealistic in almost any natural process. For example, if we took family income (x), this value would provide some useful information about how much financial support a college may offer a prospective student (y). However, the prediction would be far from perfect, since other factors play a role in financial support beyond a family's finances.

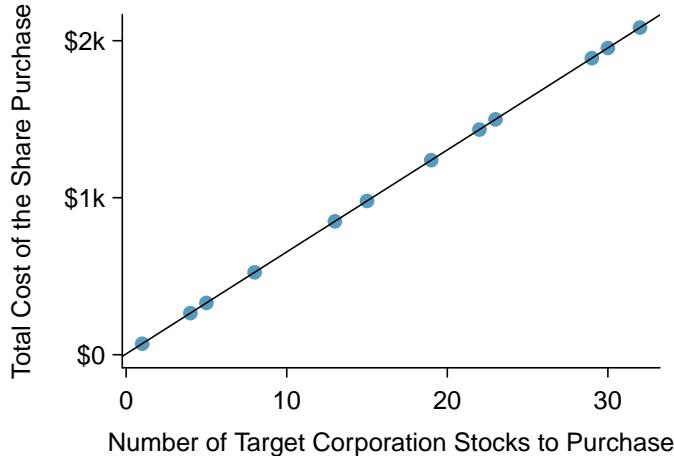


Figure 8.1: Requests from twelve separate buyers were simultaneously placed with a trading company to purchase Target Corporation stock (ticker TGT, December 28th, 2018), and the total cost of the shares were reported. Because the cost is computed using a linear formula, the linear fit is perfect.

Linear regression is the statistical method for fitting a line to data where the relationship between two variables, x and y , can be modeled by a straight line with some error:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The values β_0 and β_1 represent the model's parameters (β is the Greek letter *beta*), and the error is represented by ε (the Greek letter *epsilon*). The parameters are estimated using data, and we write their point estimates as b_0 and b_1 . When we use x to predict y , we usually call x the explanatory or **predictor** variable, and we call y the response; we also often drop the ε term when writing down the model since our main focus is often on the prediction of the average outcome.

It is rare for all of the data to fall perfectly on a straight line. Instead, it's more common for data to appear as a *cloud of points*, such as those examples shown in Figure 8.2. In each case, the data fall around a straight line, even if none of the observations fall exactly on the line. The first plot shows a relatively strong downward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y . The second plot shows an upward trend that, while evident, is not as strong as the first. The last plot shows a

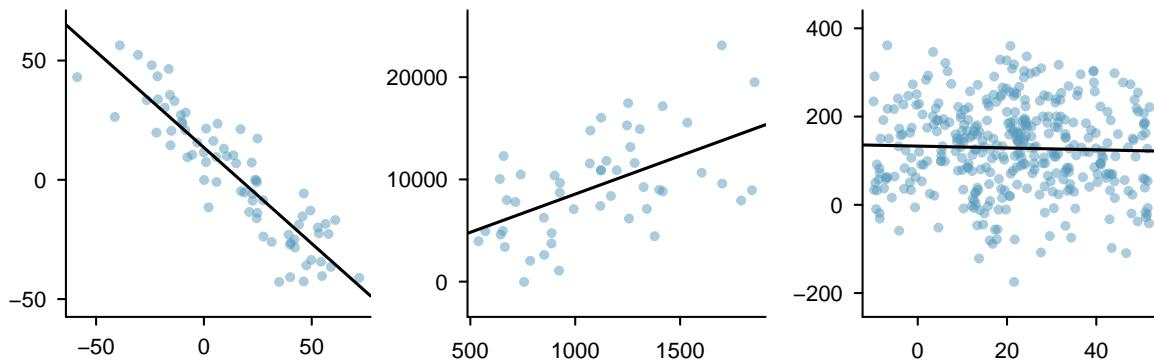


Figure 8.2: Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

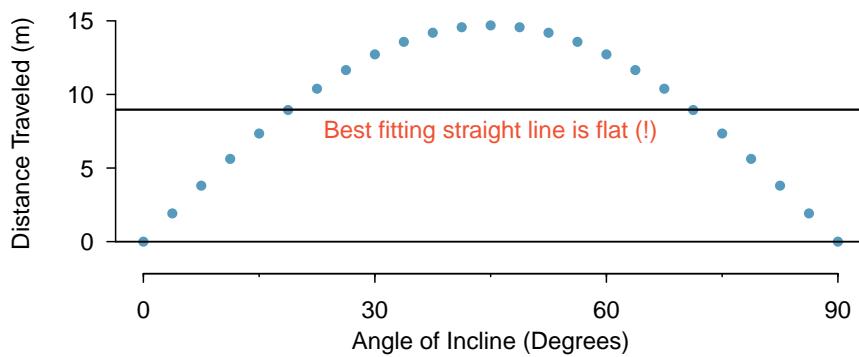


Figure 8.3: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

very weak downward trend in the data, so slight we can hardly notice it. In each of these examples, we will have some uncertainty regarding our estimates of the model parameters, β_0 and β_1 . For instance, we might wonder, should we move the line up or down a little, or should we tilt it more or less? As we move forward in this chapter, we will learn about criteria for line-fitting, and we will also learn about the uncertainty associated with estimates of model parameters.

There are also cases where fitting a straight line to the data, even if there is a clear relationship between the variables, is not helpful. One such case is shown in Figure 8.3 where there is a very clear relationship between the variables even though the trend is not linear. We discuss nonlinear trends in this chapter and the next, but details of fitting nonlinear models are saved for a later course.

8.1.2 Using linear regression to predict possum head lengths

Brushtail possums are a marsupial that lives in Australia, and a photo of one is shown in Figure 8.4. Researchers captured 104 of these animals and took body measurements before releasing the animals back into the wild. We consider two of these measurements: the total length of each possum, from head to tail, and the length of each possum's head.

Figure 8.5 shows a scatterplot for the head length and total length of the possums. Each point represents a single possum from the data. The head and total length variables are associated: possums with an above average total length also tend to have above average head lengths. While the relationship is not perfectly linear, it could be helpful to partially explain the connection between these variables with a straight line.



Figure 8.4: The common brushtail possum of Australia.

Photo by Greg Schechter (<https://flic.kr/p/9BAFbR>). CC BY 2.0 license.

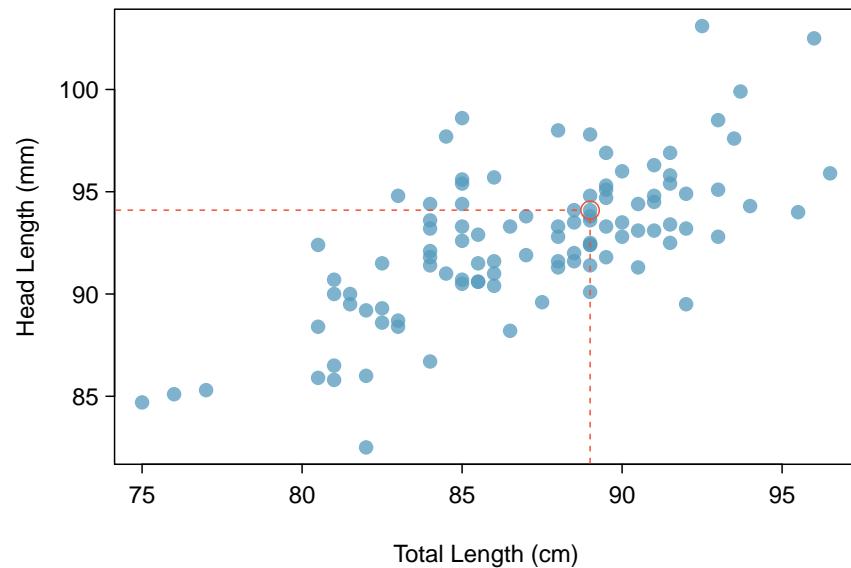


Figure 8.5: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89cm is highlighted.

We want to describe the relationship between the head length and total length variables in the possum data set using a line. In this example, we will use the total length as the predictor variable, x , to predict a possum's head length, y . We could fit the linear relationship by eye, as in Figure 8.6. The equation for this line is

$$\hat{y} = 41 + 0.59x$$

A “hat” on y is used to signify that this is an estimate. We can use this line to discuss properties of possums. For instance, the equation predicts a possum with a total length of 80 cm will have a head length of

$$\begin{aligned}\hat{y} &= 41 + 0.59 \times 80 \\ &= 88.2\end{aligned}$$

The estimate may be viewed as an average: the equation predicts that possums with a total length of 80 cm will have an average head length of 88.2 mm. Absent further information about an 80 cm possum, the prediction for head length that uses the average is a reasonable estimate.

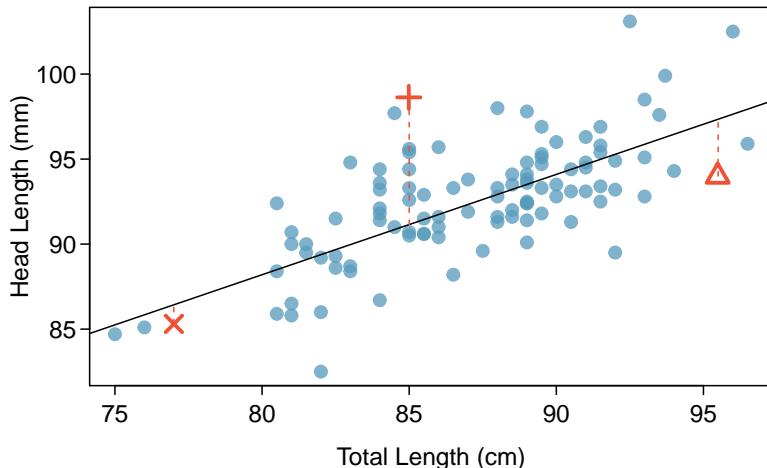


Figure 8.6: A reasonable linear model was fit to represent the relationship between head length and total length.

EXAMPLE 8.1

What other variables might help us predict the head length of a possum besides its length?

(E)

Perhaps the relationship would be a little different for male possums than female possums, or perhaps it would differ for possums from one region of Australia versus another region. In Chapter 9, we'll learn about how we can include more than one predictor. Before we get there, we first need to better understand how to best build a simple linear model with one predictor.

8.1.3 Residuals

Residuals are the leftover variation in the data after accounting for the model fit:

$$\text{Data} = \text{Fit} + \text{Residual}$$

Each observation will have a residual, and three of the residuals for the linear model we fit for the possum data is shown in Figure 8.6. If an observation is above the regression line, then its residual, the vertical distance from the observation to the line, is positive. Observations below the line have negative residuals. One goal in picking the right linear model is for these residuals to be as small as possible.

Let's look closer at the three residuals featured in Figure 8.6. The observation marked by an “ \times ” has a small, negative residual of about -1; the observation marked by “+” has a large residual of about +7; and the observation marked by “ \triangle ” has a moderate residual of about -4. The size of a residual is usually discussed in terms of its absolute value. For example, the residual for “ \triangle ” is larger than that of “ \times ” because $| -4 |$ is larger than $| -1 |$.

RESIDUAL: DIFFERENCE BETWEEN OBSERVED AND EXPECTED

The residual of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit (\hat{y}_i) :

$$e_i = y_i - \hat{y}_i$$

We typically identify \hat{y}_i by plugging x_i into the model.

EXAMPLE 8.2

The linear fit shown in Figure 8.6 is given as $\hat{y} = 41 + 0.59x$. Based on this line, formally compute the residual of the observation $(77.0, 85.3)$. This observation is denoted by “ \times ” in Figure 8.6. Check it against the earlier visual estimate, -1.

We first compute the predicted value of point “ \times ” based on the model:

$$\hat{y}_\times = 41 + 0.59x_\times = 41 + 0.59 \times 77.0 = 86.4$$

Next we compute the difference of the actual head length and the predicted head length:

$$e_\times = y_\times - \hat{y}_\times = 85.3 - 86.4 = -1.1$$

The model’s error is $e_\times = -1.1$ mm, which is very close to the visual estimate of -1mm. The negative residual indicates that the linear model overpredicted head length for this particular possum.

GUIDED PRACTICE 8.3

If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?¹

GUIDED PRACTICE 8.4

Compute the residuals for the “+” observation $(85.0, 98.6)$ and the “ \triangle ” observation $(95.5, 94.0)$ in the figure using the linear relationship $\hat{y} = 41 + 0.59x$.²

Residuals are helpful in evaluating how well a linear model fits a data set. We often display them in a **residual plot** such as the one shown in Figure 8.7 for the regression line in Figure 8.6. The residuals are plotted at their original horizontal locations but with the vertical coordinate as the residual. For instance, the point $(85.0, 98.6)_+$ had a residual of 7.45, so in the residual plot it is placed at $(85.0, 7.45)$. Creating a residual plot is sort of like tipping the scatterplot over so the regression line is horizontal.

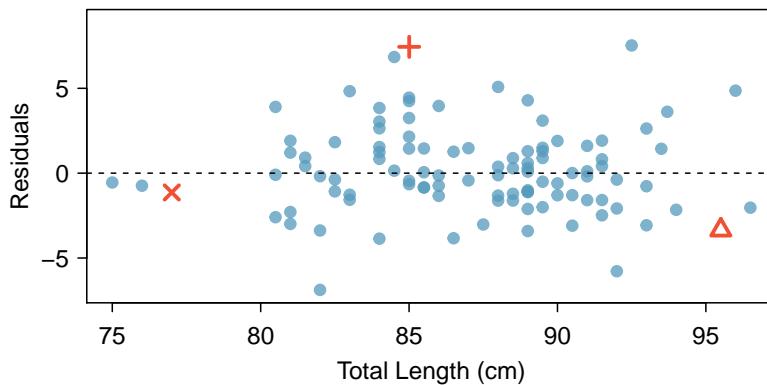


Figure 8.7: Residual plot for the model in Figure 8.6.

¹If a model underestimates an observation, then the model estimate is below the actual. The residual, which is the actual observation value minus the model estimate, must then be positive. The opposite is true when the model overestimates the observation: the residual is negative.

²(+) First compute the predicted value based on the model:

$$\hat{y}_+ = 41 + 0.59x_+ = 41 + 0.59 \times 85.0 = 91.15$$

Then the residual is given by

$$e_+ = y_+ - \hat{y}_+ = 98.6 - 91.15 = 7.45$$

This was close to the earlier estimate of 7.

(\triangle) $\hat{y}_\Delta = 41 + 0.59x_\Delta = 97.3$. $e_\Delta = y_\Delta - \hat{y}_\Delta = -3.3$, close to the estimate of -4.

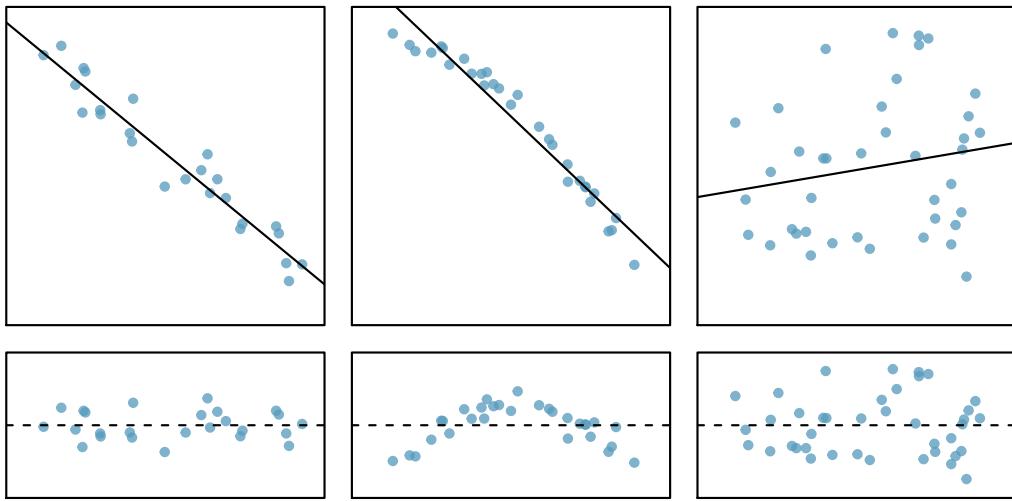


Figure 8.8: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

EXAMPLE 8.5

One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. Figure 8.8 shows three scatterplots with linear models in the first row and residual plots in the second row. Can you identify any patterns remaining in the residuals?

In the first data set (first column), the residuals show no obvious patterns. The residuals appear to be scattered randomly around the dashed line that represents 0.

The second data set shows a pattern in the residuals. There is some curvature in the scatterplot, which is more obvious in the residual plot. We should not use a straight line to model these data. Instead, a more advanced technique should be used.

The last plot shows very little upwards trend, and the residuals also show no obvious patterns. It is reasonable to try to fit a linear model to the data. However, it is unclear whether there is statistically significant evidence that the slope parameter is different from zero. The point estimate of the slope parameter, labeled b_1 , is not zero, but we might wonder if this could just be due to chance. We will address this sort of scenario in Section 8.4.

8.1.4 Describing linear relationships with correlation

We've seen plots with strong linear relationships and others with very weak linear relationships. It would be useful if we could quantify the strength of these linear relationships with a statistic.

CORRELATION: STRENGTH OF A LINEAR RELATIONSHIP

Correlation, which always takes values between -1 and 1, describes the strength of the linear relationship between two variables. We denote the correlation by R .

We can compute the correlation using a formula, just as we did with the sample mean and standard deviation. This formula is rather complex,³ and like with other statistics, we generally perform

³Formally, we can compute the correlation for observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ using the formula

$$R = \frac{1}{n-1} \sum_{i=1}^n \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

where \bar{x} , \bar{y} , s_x , and s_y are the sample means and standard deviations for each variable.

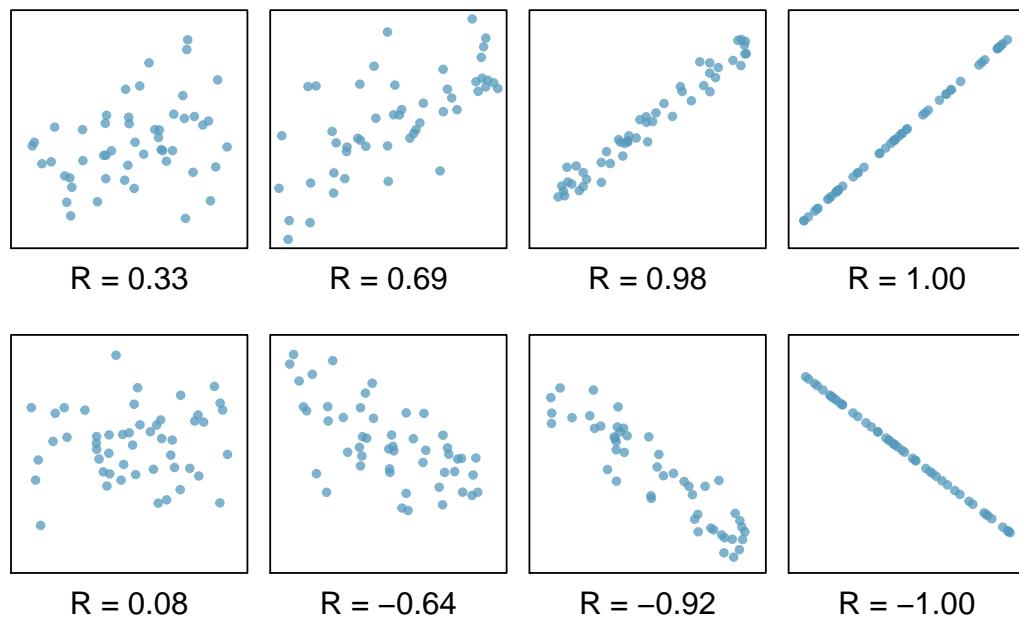


Figure 8.9: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

the calculations on a computer or calculator. Figure 8.9 shows eight plots and their corresponding correlations. Only when the relationship is perfectly linear is the correlation either -1 or 1. If the relationship is strong and positive, the correlation will be near +1. If it is strong and negative, it will be near -1. If there is no apparent linear relationship between the variables, then the correlation will be near zero.

The correlation is intended to quantify the strength of a linear trend. Nonlinear trends, even when strong, sometimes produce correlations that do not reflect the strength of the relationship; see three such examples in Figure 8.10.

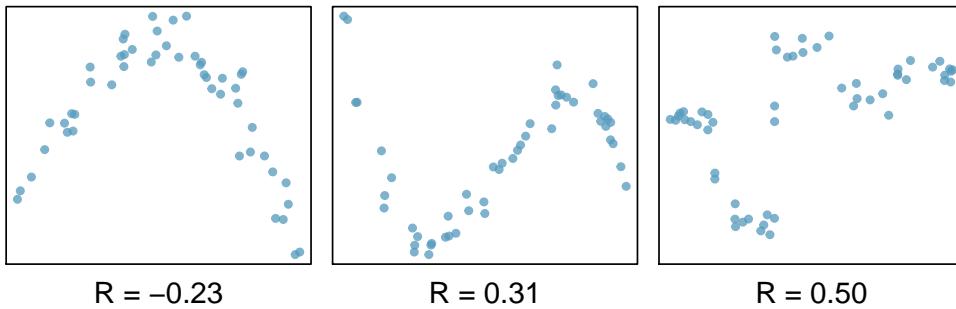


Figure 8.10: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, because the relationship is nonlinear, the correlation is relatively weak.

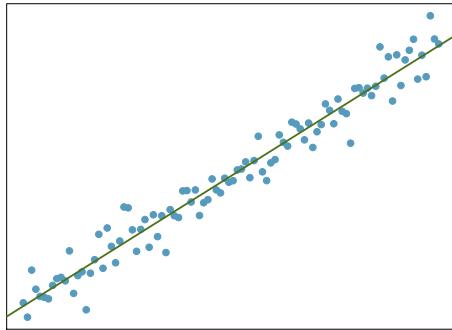
GUIDED PRACTICE 8.6

No straight line is a good fit for the data sets represented in Figure 8.10. Try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in your fit.⁴

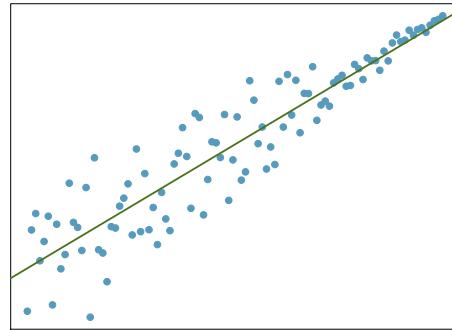
⁴We'll leave it to you to draw the lines. In general, the lines you draw should be close to most points and reflect overall trends in the data.

Exercises

8.1 Visualize the residuals. The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.

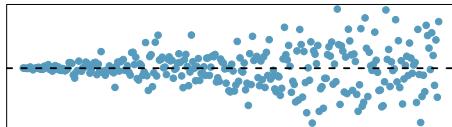


(a)

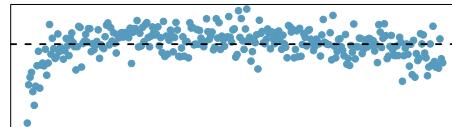


(b)

8.2 Trends in the residuals. Shown below are two plots of residuals remaining after fitting a linear model to two different sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.

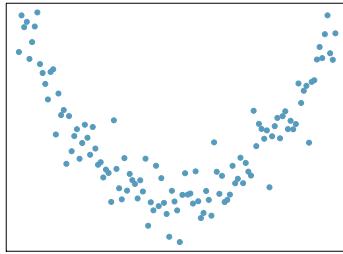


(a)

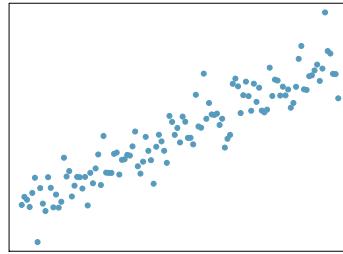


(b)

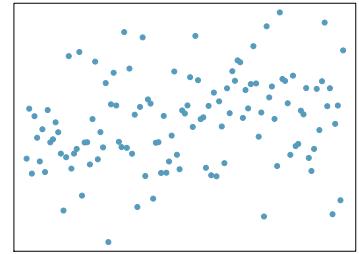
8.3 Identify relationships, Part I. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.



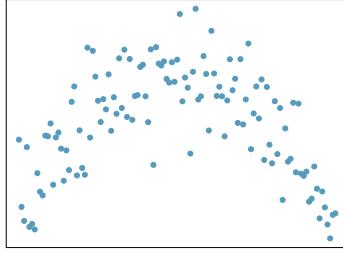
(a)



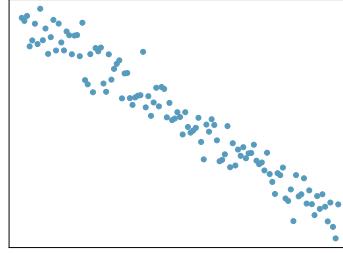
(b)



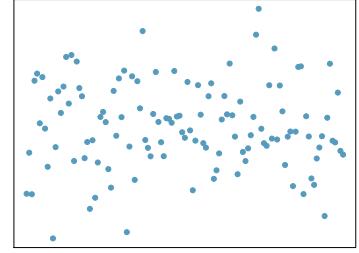
(c)



(d)

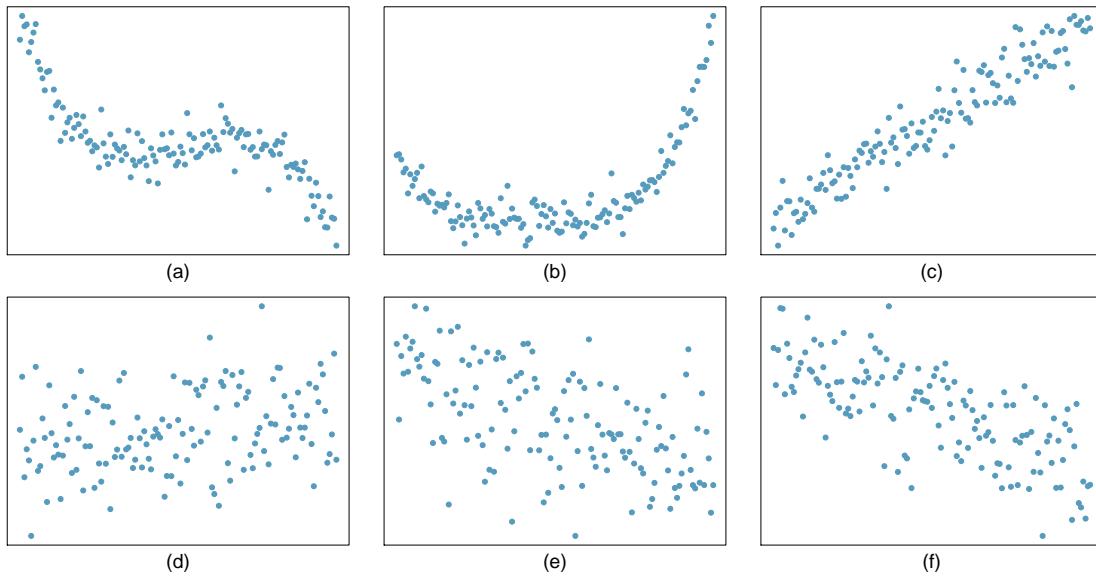


(e)



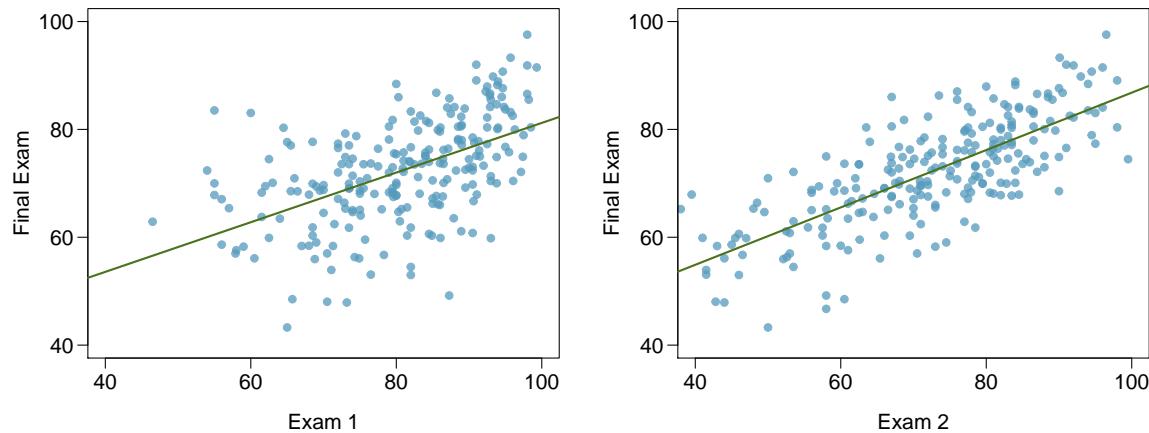
(f)

8.4 Identify relationships, Part II. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.

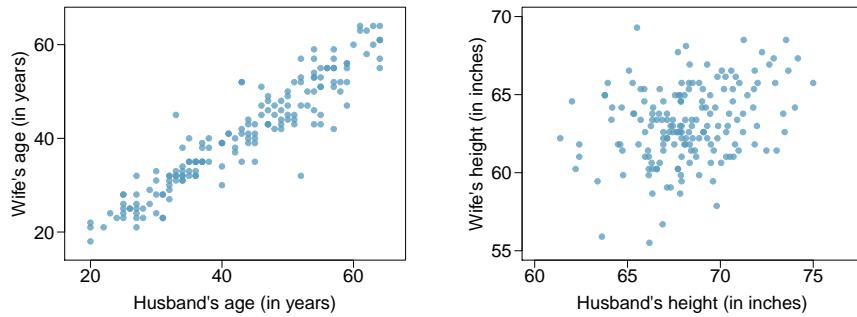


8.5 Exams and grades. The two scatterplots below show the relationship between final and mid-semester exam grades recorded during several years for a Statistics course at a university.

- Based on these graphs, which of the two exams has the strongest correlation with the final exam grade? Explain.
- Can you think of a reason why the correlation between the exam you chose in part (a) and the final exam is higher?



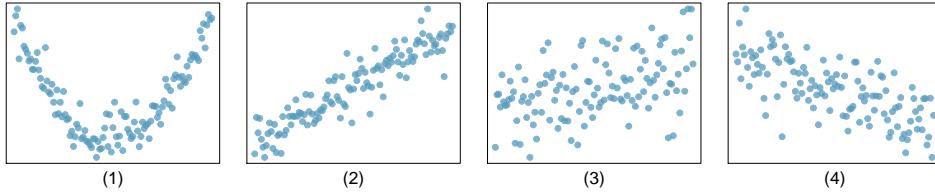
8.6 Husbands and wives, Part I. The Great Britain Office of Population Census and Surveys once collected data on a random sample of 170 married couples in Britain, recording the age (in years) and heights (converted here to inches) of the husbands and wives.⁵ The scatterplot on the left shows the wife's age plotted against her husband's age, and the plot on the right shows wife's height plotted against husband's height.



- (a) Describe the relationship between husbands' and wives' ages.
- (b) Describe the relationship between husbands' and wives' heights.
- (c) Which plot shows a stronger correlation? Explain your reasoning.
- (d) Data on heights were originally collected in centimeters, and then converted to inches. Does this conversion affect the correlation between husbands' and wives' heights?

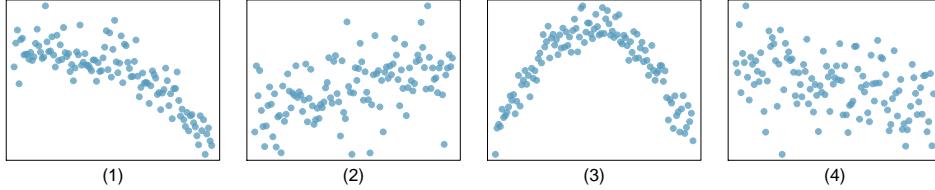
8.7 Match the correlation, Part I. Match each correlation to the corresponding scatterplot.

- (a) $R = -0.7$
- (b) $R = 0.45$
- (c) $R = 0.06$
- (d) $R = 0.92$

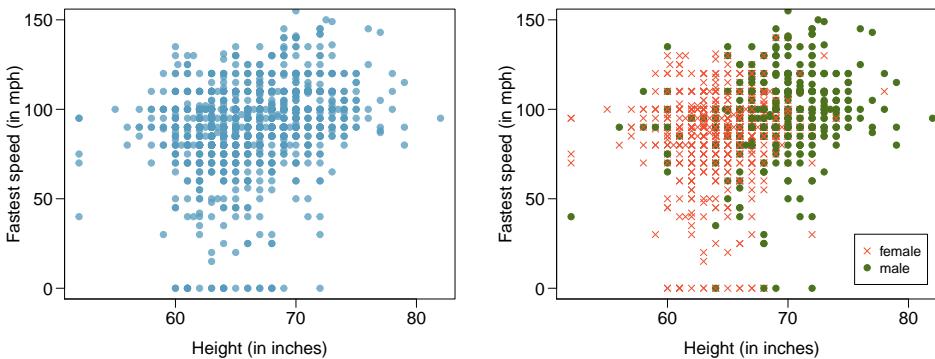


8.8 Match the correlation, Part II. Match each correlation to the corresponding scatterplot.

- (a) $R = 0.49$
- (b) $R = -0.48$
- (c) $R = -0.03$
- (d) $R = -0.85$



8.9 Speed and height. 1,302 UCLA students were asked to fill out a survey where they were asked about their height, fastest speed they have ever driven, and gender. The scatterplot on the left displays the relationship between height and fastest speed, and the scatterplot on the right displays the breakdown by gender in this relationship.



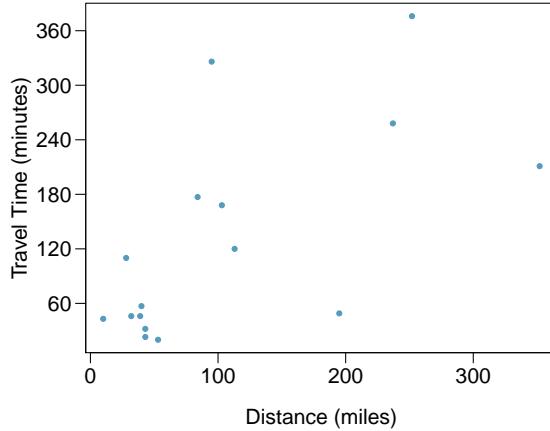
- (a) Describe the relationship between height and fastest speed.
- (b) Why do you think these variables are positively associated?
- (c) What role does gender play in the relationship between height and fastest driving speed?

⁵D.J. Hand. *A handbook of small data sets*. Chapman & Hall/CRC, 1994.

8.10 Guess the correlation. Eduardo and Rosie are both collecting data on number of rainy days in a year and the total rainfall for the year. Eduardo records rainfall in inches and Rosie in centimeters. How will their correlation coefficients compare?

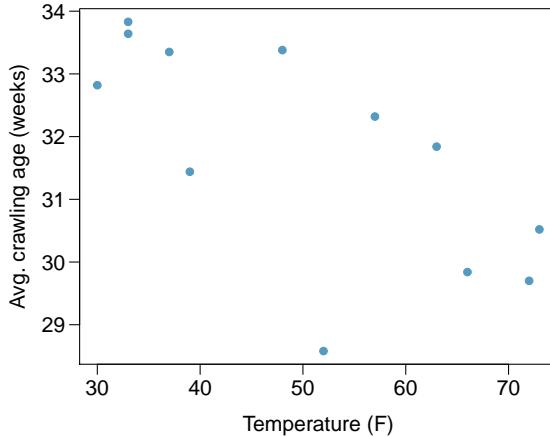
8.11 The Coast Starlight, Part I. The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).

- (a) Describe the relationship between distance and travel time.
- (b) How would the relationship change if travel time was instead measured in hours, and distance was instead measured in kilometers?
- (c) Correlation between travel time (in miles) and distance (in minutes) is $r = 0.636$. What is the correlation between travel time (in kilometers) and distance (in hours)?



8.12 Crawling babies, Part I. A study conducted at the University of Denver investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months.⁶ Infants born during the study year were split into twelve groups, one for each birth month. We consider the average crawling age of babies in each group against the average temperature when the babies are six months old (that's when babies often begin trying to crawl). Temperature is measured in degrees Fahrenheit ($^{\circ}\text{F}$) and age is measured in weeks.

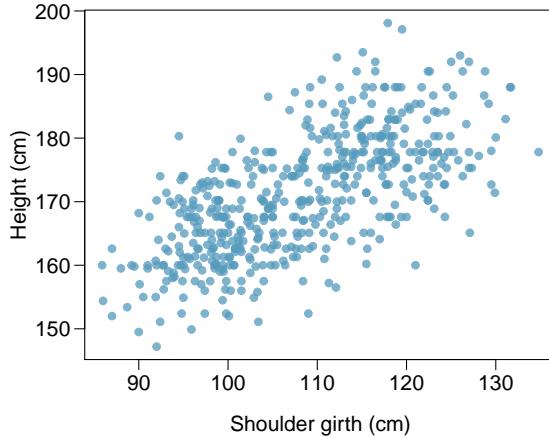
- (a) Describe the relationship between temperature and crawling age.
- (b) How would the relationship change if temperature was measured in degrees Celsius ($^{\circ}\text{C}$) and age was measured in months?
- (c) The correlation between temperature in $^{\circ}\text{F}$ and age in weeks was $r = -0.70$. If we converted the temperature to $^{\circ}\text{C}$ and age to months, what would the correlation be?



⁶J.B. Benson. “Season of birth and onset of locomotion: Theoretical and methodological implications”. In: *Infant behavior and development* 16.1 (1993), pp. 69–81. ISSN: 0163-6383.

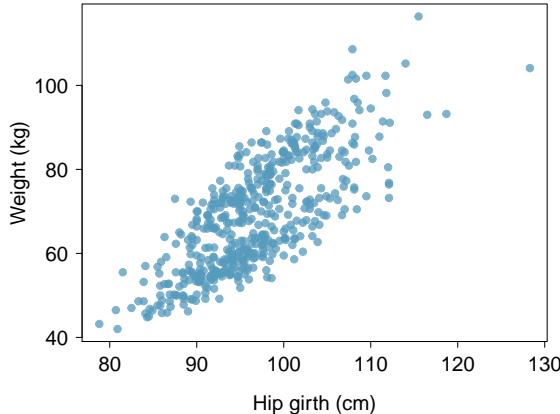
8.13 Body measurements, Part I. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals.⁷ The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.

- (a) Describe the relationship between shoulder girth and height.
- (b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?



8.14 Body measurements, Part II. The scatterplot below shows the relationship between weight measured in kilograms and hip girth measured in centimeters from the data described in Exercise 8.13.

- (a) Describe the relationship between hip girth and weight.
- (b) How would the relationship change if weight was measured in pounds while the units for hip girth remained in centimeters?



8.15 Correlation, Part I. What would be the correlation between the ages of husbands and wives if men always married women who were

- (a) 3 years younger than themselves?
- (b) 2 years older than themselves?
- (c) half as old as themselves?

8.16 Correlation, Part II. What would be the correlation between the annual salaries of males and females at a company if for a certain type of position men always made

- (a) \$5,000 more than women?
- (b) 25% more than women?
- (c) 15% less than women?

⁷G. Heinz et al. "Exploring relationships in body dimensions". In: *Journal of Statistics Education* 11.2 (2003).

8.2 Least squares regression

Fitting linear models by eye is open to criticism since it is based on an individual's preference. In this section, we use *least squares regression* as a more rigorous approach.

8.2.1 Gift aid for freshman at Elmhurst College

This section considers family income and gift aid data from a random sample of fifty students in the freshman class of Elmhurst College in Illinois. Gift aid is financial aid that does not need to be paid back, as opposed to a loan. A scatterplot of the data is shown in Figure 8.11 along with two linear fits. The lines follow a negative trend in the data; students who have higher family incomes tended to have lower gift aid from the university.

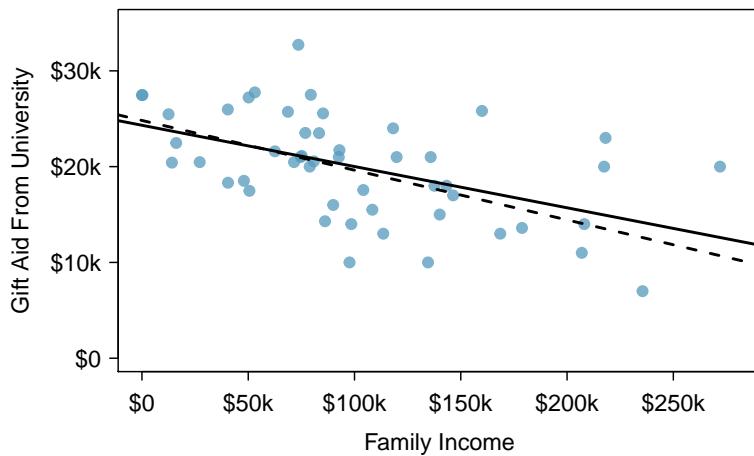


Figure 8.11: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are fit to the data, the solid line being the *least squares line*.



GUIDED PRACTICE 8.7

Is the correlation positive or negative in Figure 8.11?⁸

8.2.2 An objective measure for finding the best line

We begin by thinking about what we mean by “best”. Mathematically, we want a line that has small residuals. The first option that may come to mind is to minimize the sum of the residual magnitudes:

$$|e_1| + |e_2| + \cdots + |e_n|$$

which we could accomplish with a computer program. The resulting dashed line shown in Figure 8.11 demonstrates this fit can be quite reasonable. However, a more common practice is to choose the line that minimizes the sum of the squared residuals:

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

⁸Larger family incomes are associated with lower amounts of aid, so the correlation will be negative. Using a computer, the correlation can be computed: -0.499.

The line that minimizes this **least squares criterion** is represented as the solid line in Figure 8.11. This is commonly called the **least squares line**. The following are three possible reasons to choose this option instead of trying to minimize the sum of residual magnitudes without any squaring:

1. It is the most commonly used method.
2. Computing the least squares line is widely supported in statistical software.
3. In many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by 2. Squaring the residuals accounts for this discrepancy.

The first two reasons are largely for tradition and convenience; the last reason explains why the least squares criterion is typically most helpful.⁹

8.2.3 Conditions for the least squares line

When fitting a least squares line, we generally require

Linearity. The data should show a linear trend. If there is a nonlinear trend (e.g. left panel of Figure 8.12), an advanced regression method from another book or later course should be applied.

Nearly normal residuals. Generally, the residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about influential points, which we'll talk about more in Sections 8.3. An example of a residual that would be a potentially concern is shown in Figure 8.12, where one observation is clearly much further from the regression line than the others.

Constant variability. The variability of points around the least squares line remains roughly constant. An example of non-constant variability is shown in the third panel of Figure 8.12, which represents the most common pattern observed when this condition fails: the variability of y is larger when x is larger.

Independent observations. Be cautious about applying regression to **time series** data, which are sequential observations in time such as a stock price each day. Such data may have an underlying structure that should be considered in a model and analysis. An example of a data set where successive observations are not independent is shown in the fourth panel of Figure 8.12. There are also other instances where correlations within the data are important, which is further discussed in Chapter 9.

GUIDED PRACTICE 8.8

Should we have concerns about applying least squares regression to the Elmhurst data in Figure 8.11?¹⁰

⁹There are applications where the sum of residual magnitudes may be more useful, and there are plenty of other criteria we might consider. However, this book only applies the least squares criterion.

¹⁰The trend appears to be linear, the data fall around the line with no obvious outliers, the variance is roughly constant. These are also not time series observations. Least squares regression can be applied to these data.

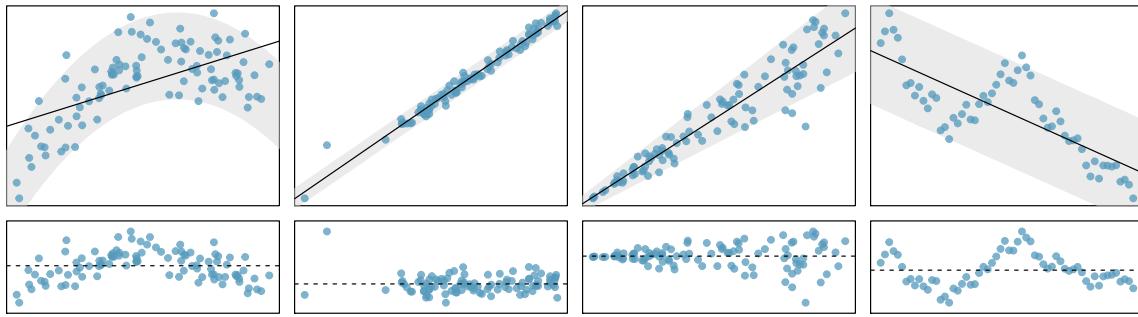


Figure 8.12: Four examples showing when the methods in this chapter are insufficient to apply to the data. First panel: linearity fails. Second panel: there are outliers, most especially one point that is very far away from the line. Third panel: the variability of the errors is related to the value of x . Fourth panel: a time series data set is shown, where successive observations are highly correlated.

8.2.4 Finding the least squares line

For the Elmhurst data, we could write the equation of the least squares regression line as

$$\widehat{aid} = \beta_0 + \beta_1 \times \text{family_income}$$

Here the equation is set up to predict gift aid based on a student's family income, which would be useful to students considering Elmhurst. These two values, β_0 and β_1 , are the parameters of the regression line.

As in Chapters 5, 6, and 7, the parameters are estimated using observed data. In practice, this estimation is done using a computer in the same way that other estimates, like a sample mean, can be estimated using a computer or calculator. However, we can also find the parameter estimates by applying two properties of the least squares line:

- The slope of the least squares line can be estimated by

$$b_1 = \frac{s_y}{s_x} R$$

where R is the correlation between the two variables, and s_x and s_y are the sample standard deviations of the explanatory variable and response, respectively.

- If \bar{x} is the sample mean of the explanatory variable and \bar{y} is the sample mean of the vertical variable, then the point (\bar{x}, \bar{y}) is on the least squares line.

Figure 8.13 shows the sample means for the family income and gift aid as \$101,780 and \$19,940, respectively. We could plot the point $(101.8, 19.94)$ on Figure 8.11 on page 317 to verify it falls on the least squares line (the solid line).

Next, we formally find the point estimates b_0 and b_1 of the parameters β_0 and β_1 .

	Family Income (x)	Gift Aid (y)
mean	$\bar{x} = \$101,780$	$\bar{y} = \$19,940$
sd	$s_x = \$63,200$	$s_y = \$5,460$
		$R = -0.499$

Figure 8.13: Summary statistics for family income and gift aid.

GUIDED PRACTICE 8.9

(G) Using the summary statistics in Figure 8.13, compute the slope for the regression line of gift aid against family income.¹¹

You might recall the **point-slope** form of a line from math class, which we can use to find the model fit, including the estimate of b_0 . Given the slope of a line and a point on the line, (x_0, y_0) , the equation for the line can be written as

$$y - y_0 = \text{slope} \times (x - x_0)$$

IDENTIFYING THE LEAST SQUARES LINE FROM SUMMARY STATISTICS

To identify the least squares line from summary statistics:

- Estimate the slope parameter, $b_1 = (s_y/s_x)R$.
- Noting that the point (\bar{x}, \bar{y}) is on the least squares line, use $x_0 = \bar{x}$ and $y_0 = \bar{y}$ with the point-slope equation: $y - \bar{y} = b_1(x - \bar{x})$.
- Simplify the equation, which would reveal that $b_0 = \bar{y} - b_1\bar{x}$.

EXAMPLE 8.10

Using the point $(101780, 19940)$ from the sample means and the slope estimate $b_1 = -0.0431$ from Guided Practice 8.9, find the least-squares line for predicting aid based on family income.

Apply the point-slope equation using $(101.78, 19.94)$ and the slope $b_1 = -0.0431$:

$$\begin{aligned} y - y_0 &= b_1(x - x_0) \\ y - 19,940 &= -0.0431(x - 101,780) \end{aligned}$$

Expanding the right side and then adding 19,940 to each side, the equation simplifies:

$$\widehat{\text{aid}} = 24,327 - 0.0431 \times \text{family_income}$$

Here we have replaced y with $\widehat{\text{aid}}$ and x with family_income to put the equation in context. The final equation should always include a “hat” on the variable being predicted, whether it is a generic “ y ” or a named variable like “ aid ”.

A computer is usually used to compute the least squares line, and a summary table generated using software for the Elmhurst regression line is shown in Figure 8.14. The first column of numbers provides estimates for b_0 and b_1 , respectively. These results match those from Example 8.10 (with some minor rounding error).

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	24319.3	1291.5	18.83	<0.0001
family_income	-0.0431	0.0108	-3.98	0.0002

Figure 8.14: Summary of least squares fit for the Elmhurst data. Compare the parameter estimates in the first column to the results of Example 8.10.

¹¹Compute the slope using the summary statistics from Figure 8.13:

$$b_1 = \frac{s_y}{s_x} R = \frac{5,460}{63,200} (-0.499) = -0.0431$$

EXAMPLE 8.11

Examine the second, third, and fourth columns in Figure 8.14. Can you guess what they represent? (If you have not reviewed any inference chapter yet, skip this example.)

We'll describe the meaning of the columns using the second row, which corresponds to β_1 . The first column provides the point estimate for β_1 , as we calculated in an earlier example: $b_1 = -0.0431$. The second column is a standard error for this point estimate: $SE_{b_1} = 0.0108$. The third column is a t -test statistic for the null hypothesis that $\beta_1 = 0$: $T = -3.98$. The last column is the p-value for the t -test statistic for the null hypothesis $\beta_1 = 0$ and a two-sided alternative hypothesis: 0.0002. We will get into more of these details in Section 8.4.

EXAMPLE 8.12

Suppose a high school senior is considering Elmhurst College. Can she simply use the linear equation that we have estimated to calculate her financial aid from the university?

She may use it as an estimate, though some qualifiers on this approach are important. First, the data all come from one freshman class, and the way aid is determined by the university may change from year to year. Second, the equation will provide an imperfect estimate. While the linear equation is good at capturing the trend in the data, no individual student's aid will be perfectly predicted.

8.2.5 Interpreting regression model parameter estimates

Interpreting parameters in a regression model is often one of the most important steps in the analysis.

EXAMPLE 8.13

The intercept and slope estimates for the Elmhurst data are $b_0 = 24,319$ and $b_1 = -0.0431$. What do these numbers really mean?

Interpreting the slope parameter is helpful in almost any application. For each additional \$1,000 of family income, we would expect a student to receive a net difference of $\$1,000 \times (-0.0431) = -\43.10 in aid on average, i.e. $\$43.10$ less. Note that a higher family income corresponds to less aid because the coefficient of family income is negative in the model. We must be cautious in this interpretation: while there is a real association, we cannot interpret a causal connection between the variables because these data are observational. That is, increasing a student's family income may not cause the student's aid to drop. (It would be reasonable to contact the college and ask if the relationship is causal, i.e. if Elmhurst College's aid decisions are partially based on students' family income.)

The estimated intercept $b_0 = 24,319$ describes the average aid if a student's family had no income. The meaning of the intercept is relevant to this application since the family income for some students at Elmhurst is \$0. In other applications, the intercept may have little or no practical value if there are no observations where x is near zero.

INTERPRETING PARAMETERS ESTIMATED BY LEAST SQUARES

The slope describes the estimated difference in the y variable if the explanatory variable x for a case happened to be one unit larger. The intercept describes the average outcome of y if $x = 0$ and the linear model is valid all the way to $x = 0$, which in many applications is not the case.

8.2.6 Extrapolation is treacherous

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6th it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on.

Stephen Colbert
April 6th, 2010¹²

Linear models can be used to approximate the relationship between two variables. However, these models have real limitations. Linear regression is simply a modeling framework. The truth is almost always much more complex than our simple line. For example, we do not know how the data outside of our limited window will behave.

EXAMPLE 8.14

Use the model $\widehat{\text{aid}} = 24,319 - 0.0431 \times \text{family_income}$ to estimate the aid of another freshman student whose family had income of \$1 million.

We want to calculate the aid for $\text{family_income} = 1,000,000$:

$$24,319 - 0.0431 \times \text{family_income} = 24,319 - 0.0431 \times 1,000,000 = -18,781$$

The model predicts this student will have -\$18,781 in aid (!). However, Elmhurst College does not offer *negative aid* where they select some students to pay extra on top of tuition to attend.

Applying a model estimate to values outside of the realm of the original data is called **extrapolation**. Generally, a linear model is only an approximation of the real relationship between two variables. If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analyzed.

8.2.7 Using R^2 to describe the strength of a fit

We evaluated the strength of the linear relationship between two variables earlier using the correlation, R . However, it is more common to explain the strength of a linear fit using R^2 , called **R-squared**. If provided with a linear model, we might like to describe how closely the data cluster around the linear fit.

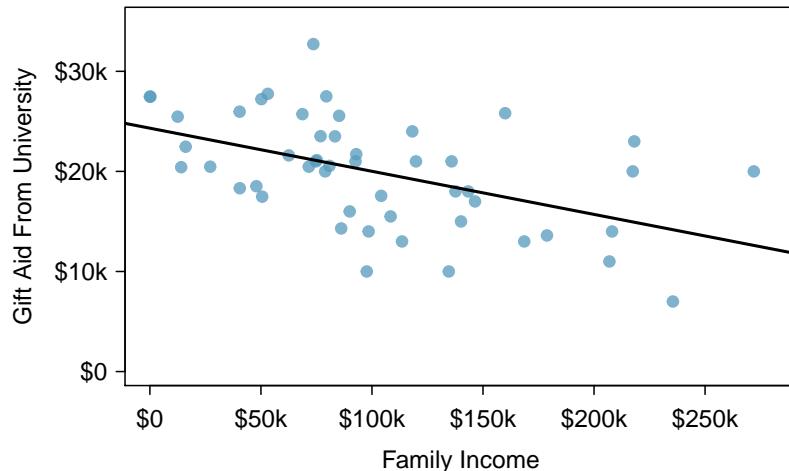


Figure 8.15: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College, shown with the least squares regression line.

¹²www.cc.com/video-clips/l4nk0q

The R^2 of a linear model describes the amount of variation in the response that is explained by the least squares line. For example, consider the Elmhurst data, shown in Figure 8.15. The variance of the response variable, aid received, is about $s_{\text{aid}}^2 \approx 29.8$ million. However, if we apply our least squares line, then this model reduces our uncertainty in predicting aid using a student's family income. The variability in the residuals describes how much variation remains after using the model: $s_{\text{RES}}^2 \approx 22.4$ million. In short, there was a reduction of

$$\frac{s_{\text{aid}}^2 - s_{\text{RES}}^2}{s_{\text{aid}}^2} = \frac{29,800,000 - 22,400,000}{29,800,000} = \frac{7,500,000}{29,800,000} = 0.25$$

or about 25% in the data's variation by using information about family income for predicting aid using a linear model. This corresponds exactly to the R-squared value:

$$R = -0.499$$

$$R^2 = 0.25$$

GUIDED PRACTICE 8.15

If a linear model has a very strong negative relationship with a correlation of -0.97, how much of the variation in the response is explained by the explanatory variable?¹³

8.2.8 Categorical predictors with two levels

Categorical variables are also useful in predicting outcomes. Here we consider a categorical predictor with two levels (recall that a *level* is the same as a *category*). We'll consider Ebay auctions for a video game, *Mario Kart* for the Nintendo Wii, where both the total price of the auction and the condition of the game were recorded. Here we want to predict total price based on game condition, which takes values `used` and `new`. A plot of the auction data is shown in Figure 8.16.

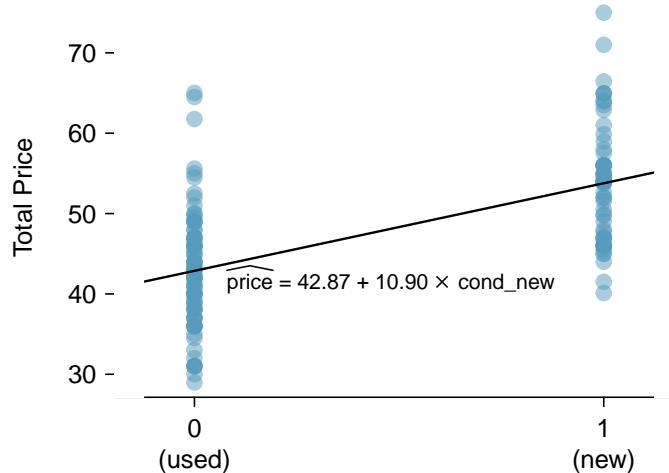


Figure 8.16: Total auction prices for the video game *Mario Kart*, divided into used ($x = 0$) and new ($x = 1$) condition games. The least squares regression line is also shown.

To incorporate the game condition variable into a regression equation, we must convert the categories into a numerical form. We will do so using an **indicator variable** called `cond_new`, which takes value 1 when the game is new and 0 when the game is used. Using this indicator variable, the linear model may be written as

$$\widehat{\text{price}} = \beta_0 + \beta_1 \times \text{cond_new}$$

¹³About $R^2 = (-0.97)^2 = 0.94$ or 94% of the variation is explained by the linear model.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	42.87	0.81	52.67	<0.0001
cond_new	10.90	1.26	8.66	<0.0001

Figure 8.17: Least squares regression summary for the final auction price against the condition of the game.

The parameter estimates are given in Figure 8.17, and the model equation can be summarized as

$$\widehat{\text{price}} = 42.87 + 10.90 \times \text{cond_new}$$

For categorical predictors with just two levels, the linearity assumption will always be satisfied. However, we must evaluate whether the residuals in each group are approximately normal and have approximately equal variance. As can be seen in Figure 8.16, both of these conditions are reasonably satisfied by the auction data.

EXAMPLE 8.16

Interpret the two parameters estimated in the model for the price of *Mario Kart* in eBay auctions.

The intercept is the estimated price when `cond_new` takes value 0, i.e. when the game is in used condition. That is, the average selling price of a used version of the game is \$42.87.

The slope indicates that, on average, new games sell for about \$10.90 more than used games.

INTERPRETING MODEL ESTIMATES FOR CATEGORICAL PREDICTORS

The estimated intercept is the value of the response variable for the first category (i.e. the category corresponding to an indicator value of 0). The estimated slope is the average change in the response variable between the two categories.

We'll elaborate further on this topic in Chapter 9, where we examine the influence of many predictor variables simultaneously using multiple regression.

Exercises

8.17 Units of regression. Consider a regression predicting weight (kg) from height (cm) for a sample of adult males. What are the units of the correlation coefficient, the intercept, and the slope?

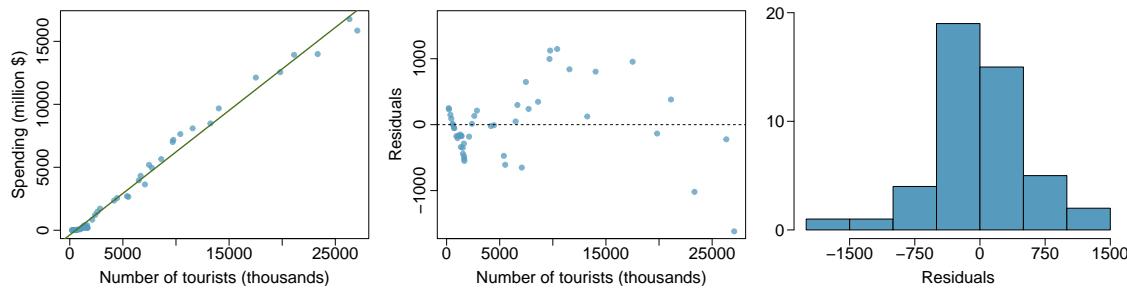
8.18 Which is higher? Determine if I or II is higher or if they are equal. Explain your reasoning. For a regression line, the uncertainty associated with the slope estimate, b_1 , is higher when

- there is a lot of scatter around the regression line or
- there is very little scatter around the regression line

8.19 Over-under, Part I. Suppose we fit a regression line to predict the shelf life of an apple based on its weight. For a particular apple, we predict the shelf life to be 4.6 days. The apple's residual is -0.6 days. Did we over or under estimate the shelf-life of the apple? Explain your reasoning.

8.20 Over-under, Part II. Suppose we fit a regression line to predict the number of incidents of skin cancer per 1,000 people from the number of sunny days in a year. For a particular year, we predict the incidence of skin cancer to be 1.5 per 1,000 people, and the residual for this year is 0.5. Did we over or under estimate the incidence of skin cancer? Explain your reasoning.

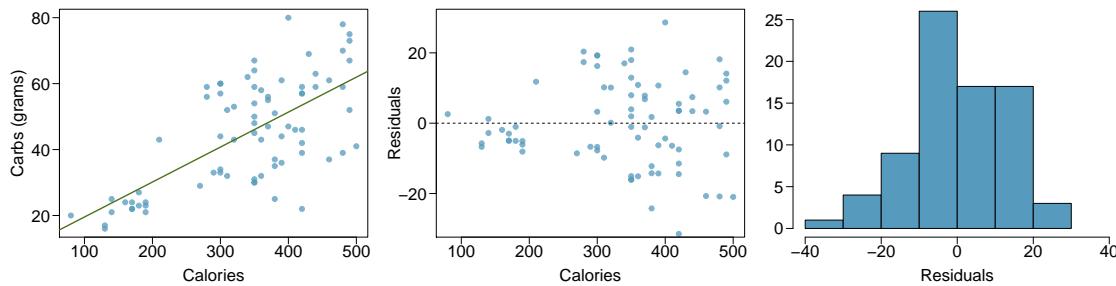
8.21 Tourism spending. The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year.¹⁴ Three plots are provided: scatterplot showing the relationship between these two variables along with the least squares fit, residuals plot, and histogram of residuals.



- Describe the relationship between number of tourists and spending.
- What are the explanatory and response variables?
- Why might we want to fit a regression line to these data?
- Do the data meet the conditions required for fitting a least squares line? In addition to the scatterplot, use the residual plot and histogram to answer this question.

¹⁴Association of Turkish Travel Agencies, Foreign Visitors Figure & Tourist Spendings By Years.

8.22 Nutrition at Starbucks, Part I. The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.¹⁵ Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content.



- Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.
- In this scenario, what are the explanatory and response variables?
- Why might we want to fit a regression line to these data?
- Do these data meet the conditions required for fitting a least squares line?

8.23 The Coast Starlight, Part II. Exercise 8.11 introduces data on the Coast Starlight Amtrak train that runs from Seattle to Los Angeles. The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

- Write the equation of the regression line for predicting travel time.
- Interpret the slope and the intercept in this context.
- Calculate R^2 of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret R^2 in the context of the application.
- The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.
- It actually takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.
- Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

8.24 Body measurements, Part III. Exercise 8.13 introduces data on shoulder girth and height of a group of individuals. The mean shoulder girth is 107.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67.

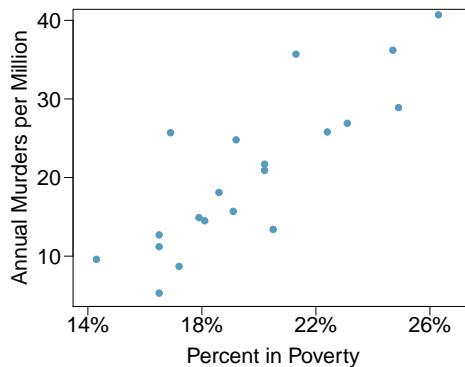
- Write the equation of the regression line for predicting height.
- Interpret the slope and the intercept in this context.
- Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.
- A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.
- The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.
- A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

¹⁵Source: Starbucks.com, collected on March 10, 2011, www.starbucks.com/menu/nutrition.

8.25 Murders and poverty, Part I. The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000
$s = 5.512$	$R^2 = 70.52\%$		$R^2_{adj} = 68.89\%$	

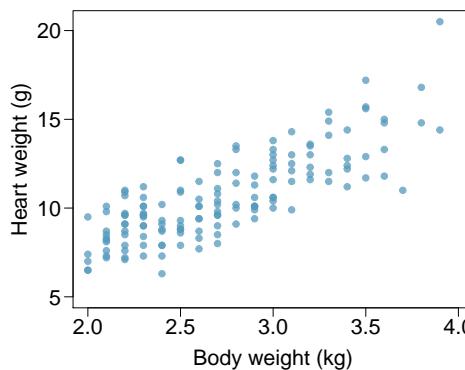
- (a) Write out the linear model.
- (b) Interpret the intercept.
- (c) Interpret the slope.
- (d) Interpret R^2 .
- (e) Calculate the correlation coefficient.



8.26 Cats, Part I. The following regression output is for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000
$s = 1.452$	$R^2 = 64.66\%$		$R^2_{adj} = 64.41\%$	

- (a) Write out the linear model.
- (b) Interpret the intercept.
- (c) Interpret the slope.
- (d) Interpret R^2 .
- (e) Calculate the correlation coefficient.



8.3 Types of outliers in linear regression

In this section, we identify criteria for determining which outliers are important and influential. Outliers in regression are observations that fall far from the cloud of points. These points are especially important because they can have a strong influence on the least squares line.

EXAMPLE 8.17

There are six plots shown in Figure 8.18 along with the least squares line and residual plots. For each scatterplot and residual plot pair, identify the outliers and note how they influence the least squares line. Recall that an outlier is any point that doesn't appear to belong with the vast majority of the other points.

- (1) There is one outlier far from the other points, though it only appears to slightly influence the line.
- (2) There is one outlier on the right, though it is quite close to the least squares line, which suggests it wasn't very influential.
- (3) There is one point far away from the cloud, and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud doesn't appear to fit very well.
- (4) There is a primary cloud and then a small secondary cloud of four outliers. The secondary cloud appears to be influencing the line somewhat strongly, making the least square line fit poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated.
- (5) There is no obvious trend in the main cloud of points and the outlier on the right appears to largely control the slope of the least squares line.
- (6) There is one outlier far from the cloud. However, it falls quite close to the least squares line and does not appear to be very influential.

Examine the residual plots in Figure 8.18. You will probably find that there is some trend in the main clouds of (3) and (4). In these cases, the outliers influenced the slope of the least squares lines. In (5), data with no clear trend were assigned a line with a large trend simply due to one outlier (!).

LEVERAGE

Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with **high leverage**.

Points that fall horizontally far from the line are points of high leverage; these points can strongly influence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line – as in cases (3), (4), and (5) of Example 8.17 – then we call it an **influential point**. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line.

It is tempting to remove outliers. Don't do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly. For instance, if a financial firm ignored the largest market swings – the “outliers” – they would soon go bankrupt by making poorly thought-out investments.

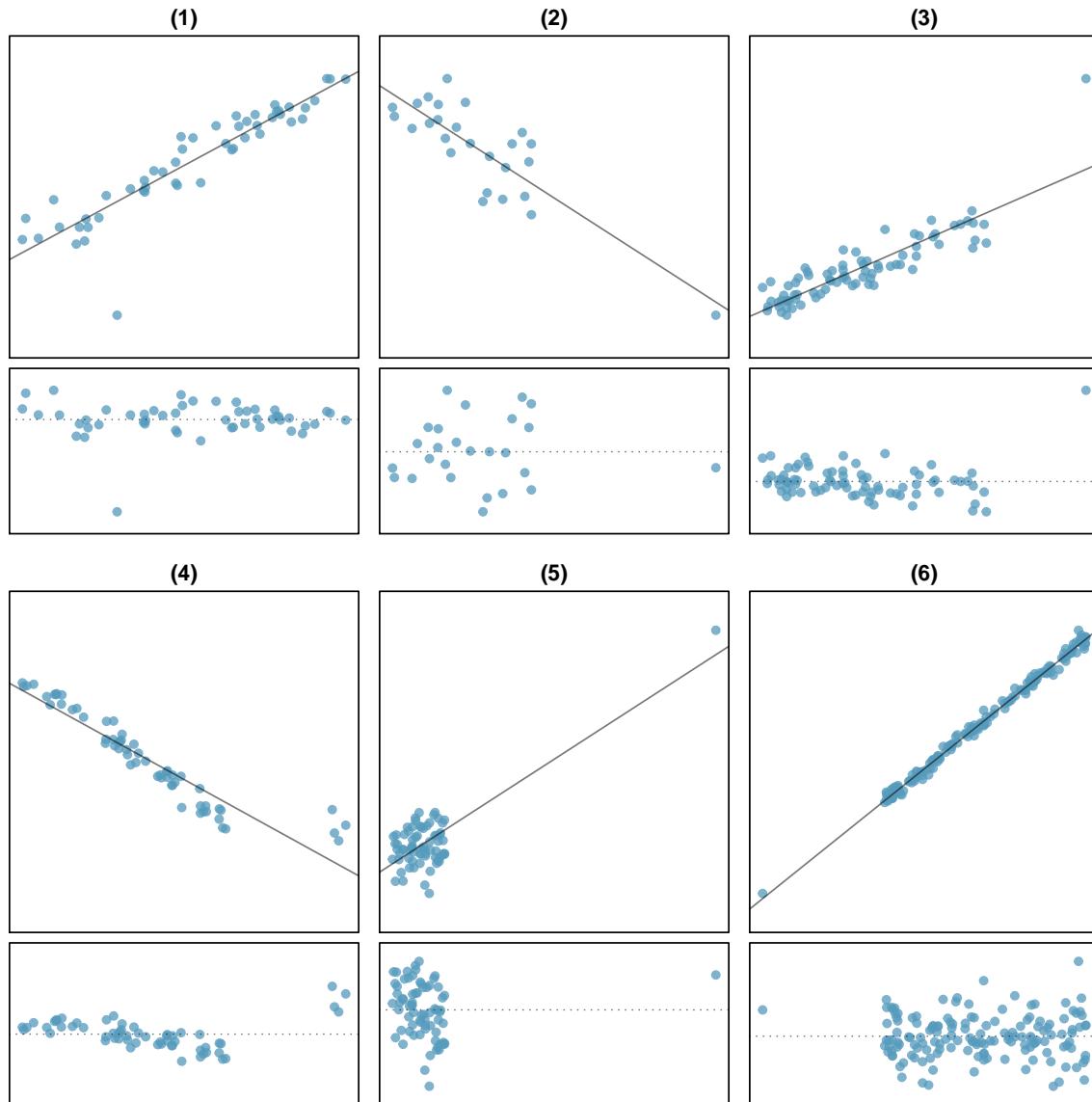
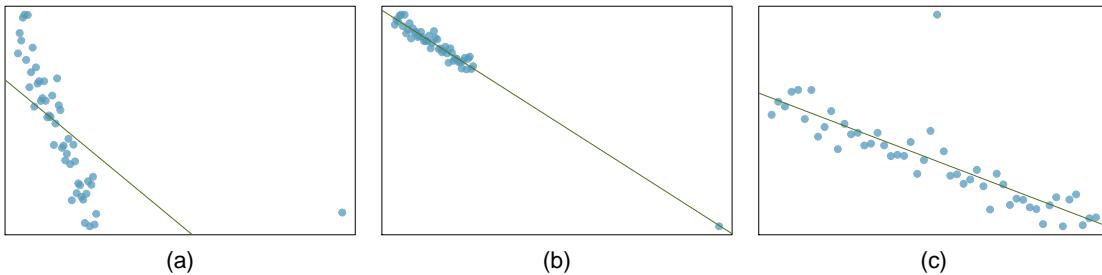


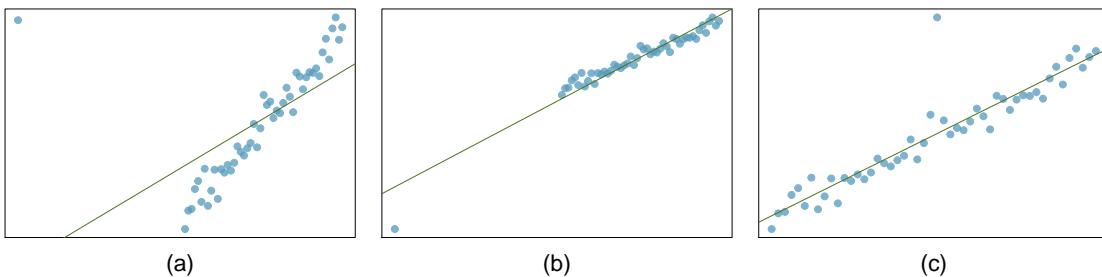
Figure 8.18: Six plots, each with a least squares line and residual plot. All data sets have at least one outlier.

Exercises

8.27 Outliers, Part I. Identify the outliers in the scatterplots shown below, and determine what type of outliers they are. Explain your reasoning.

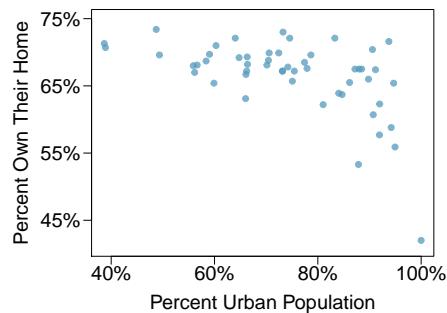


8.28 Outliers, Part II. Identify the outliers in the scatterplots shown below and determine what type of outliers they are. Explain your reasoning.



8.29 Urban homeowners, Part I. The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas.¹⁶ There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.

- (a) Describe the relationship between the percent of families who own their home and the percent of the population living in urban areas.
- (b) The outlier at the bottom right corner is District of Columbia, where 100% of the population is considered urban. What type of an outlier is this observation?



8.30 Crawling babies, Part II. Exercise 8.12 introduces data on the average monthly temperature during the month babies first try to crawl (about 6 months after birth) and the average first crawling age for babies born in a given month. A scatterplot of these two variables reveals a potential outlying month when the average temperature is about 53°F and average crawling age is about 28.5 weeks. Does this point have high leverage? Is it an influential point?

¹⁶United States Census Bureau, 2010 Census Urban and Rural Classification and Urban Area Criteria and Housing Characteristics: 2010.

8.4 Inference for linear regression

In this section, we discuss uncertainty in the estimates of the slope and y -intercept for a regression line. Just as we identified standard errors for point estimates in previous chapters, we first discuss standard errors for these new estimates.

8.4.1 Midterm elections and unemployment

Elections for members of the United States House of Representatives occur every two years, coinciding every four years with the U.S. Presidential election. The set of House elections occurring during the middle of a Presidential term are called midterm elections. In America's two-party system, one political theory suggests the higher the unemployment rate, the worse the President's party will do in the midterm elections.

To assess the validity of this claim, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2018, with the exception of those elections during the Great Depression. Figure 8.19 shows these data and the least-squares regression line:

$$\begin{aligned} \text{\% change in House seats for President's party} \\ = -7.36 - 0.89 \times (\text{unemployment rate}) \end{aligned}$$

We consider the percent change in the number of seats of the President's party (e.g. percent change in the number of seats for Republicans in 2018) against the unemployment rate.

Examining the data, there are no clear deviations from linearity, the constant variance condition, or substantial outliers. While the data are collected sequentially, a separate analysis was used to check for any apparent correlation between successive observations; no such correlation was found.

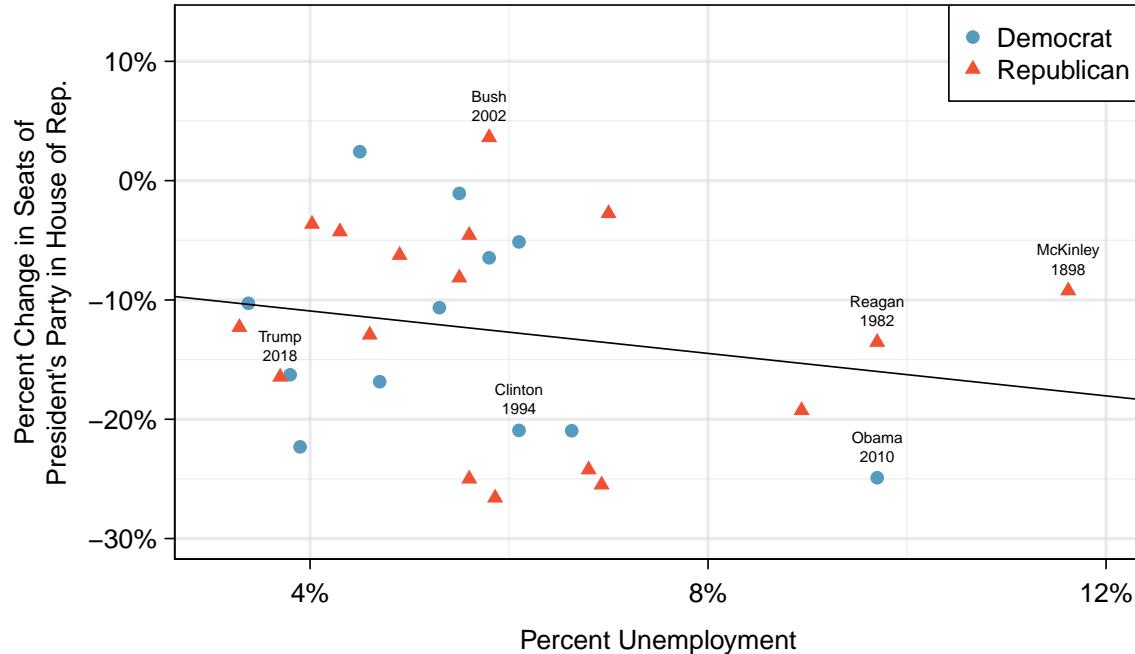


Figure 8.19: The percent change in House seats for the President's party in each election from 1898 to 2010 plotted against the unemployment rate. The two points for the Great Depression have been removed, and a least squares regression line has been fit to the data.

GUIDED PRACTICE 8.18

(G) The data for the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively. Do you agree that they should be removed for this investigation? Why or why not?¹⁷

There is a negative slope in the line shown in Figure 8.19. However, this slope (and the y-intercept) are only estimates of the parameter values. We might wonder, is this convincing evidence that the “true” linear model has a negative slope? That is, do the data provide strong evidence that the political theory is accurate, where the unemployment rate is a useful predictor of the midterm election? We can frame this investigation into a statistical hypothesis test:

$H_0: \beta_1 = 0$. The true linear model has slope zero.

$H_A: \beta_1 \neq 0$. The true linear model has a slope different than zero. The unemployment is predictive of whether the President’s party wins or loses seats in the House of Representatives.

We would reject H_0 in favor of H_A if the data provide strong evidence that the true slope parameter is different than zero. To assess the hypotheses, we identify a standard error for the estimate, compute an appropriate test statistic, and identify the p-value.

8.4.2 Understanding regression output from software

Just like other point estimates we have seen before, we can compute a standard error and test statistic for b_1 . We will generally label the test statistic using a T , since it follows the t -distribution.

We will rely on statistical software to compute the standard error and leave the explanation of how this standard error is determined to a second or third statistics course. Figure 8.20 shows software output for the least squares regression line in Figure 8.19. The row labeled *unemp* includes the point estimate and other hypothesis test information for the slope, which is the coefficient of the unemployment variable.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-7.3644	5.1553	-1.43	0.1646
unemp	-0.8897	0.8350	-1.07	0.2961
<i>df</i> = 27				

Figure 8.20: Output from statistical software for the regression line modeling the midterm election losses for the President’s party as a response to unemployment.

EXAMPLE 8.19

What do the first and second columns of Figure 8.20 represent?

The entries in the first column represent the least squares estimates, b_0 and b_1 , and the values in the second column correspond to the standard errors of each estimate. Using the estimates, we could write the equation for the least square regression line as

$$\hat{y} = -7.3644 - 0.8897x$$

where \hat{y} in this case represents the predicted change in the number of seats for the president’s party, and x represents the unemployment rate.

¹⁷We will provide two considerations. Each of these points would have very high leverage on any least-squares regression line, and years with such high unemployment may not help us understand what would happen in other years where the unemployment is only modestly high. On the other hand, these are exceptional cases, and we would be discarding important information if we exclude them from a final analysis.

We previously used a t -test statistic for hypothesis testing in the context of numerical data. Regression is very similar. In the hypotheses we consider, the null value for the slope is 0, so we can compute the test statistic using the T (or Z) score formula:

$$T = \frac{\text{estimate} - \text{null value}}{\text{SE}} = \frac{-0.8897 - 0}{0.8350} = -1.07$$

This corresponds to the third column of Figure 8.20.

EXAMPLE 8.20

Use the table in Figure 8.20 to determine the p-value for the hypothesis test.

(E) The last column of the table gives the p-value for the two-sided hypothesis test for the coefficient of the unemployment rate: 0.2961. That is, the data do not provide convincing evidence that a higher unemployment rate has any correspondence with smaller or larger losses for the President's party in the House of Representatives in midterm elections.

INFERENCE FOR REGRESSION

We usually rely on statistical software to identify point estimates, standard errors, test statistics, and p-values in practice. However, be aware that software will not generally check whether the method is appropriate, meaning we must still verify conditions are met.

EXAMPLE 8.21

Examine Figure 8.15 on page 322, which relates the Elmhurst College aid and student family income. How sure are you that the slope is statistically significantly different from zero? That is, do you think a formal hypothesis test would reject the claim that the true slope of the line should be zero?

While the relationship between the variables is not perfect, there is an evident decreasing trend in the data. This suggests the hypothesis test will reject the null claim that the slope is zero.

GUIDED PRACTICE 8.22

Figure 8.21 shows statistical software output from fitting the least squares regression line shown in Figure 8.15. Use this output to formally evaluate the following hypotheses.¹⁸

H_0 : The true coefficient for family income is zero.

H_A : The true coefficient for family income is not zero.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	24319.3	1291.5	18.83	<0.0001
family_income	-0.0431	0.0108	-3.98	0.0002
<i>df = 48</i>				

Figure 8.21: Summary of least squares fit for the Elmhurst College data, where we are predicting the gift aid by the university based on the family income of students.

¹⁸We look in the second row corresponding to the family income variable. We see the point estimate of the slope of the line is -0.0431, the standard error of this estimate is 0.0108, and the t -test statistic is $T = -3.98$. The p-value corresponds exactly to the two-sided test we are interested in: 0.0002. The p-value is so small that we reject the null hypothesis and conclude that family income and financial aid at Elmhurst College for freshman entering in the year 2011 are negatively correlated and the true slope parameter is indeed less than 0, just as we believed in Example 8.21.

8.4.3 Confidence interval for a coefficient

Similar to how we can conduct a hypothesis test for a model coefficient using regression output, we can also construct a confidence interval for that coefficient.

EXAMPLE 8.23

Compute the 95% confidence interval for the `family_income` coefficient using the regression output from Table 8.21.

The point estimate is -0.0431 and the standard error is $SE = 0.0108$. When constructing a confidence interval for a model coefficient, we generally use a t -distribution. The degrees of freedom for the distribution are noted in the regression output, $df = 48$, allowing us to identify $t_{48}^* = 2.01$ for use in the confidence interval.

We can now construct the confidence interval in the usual way:

$$\text{point estimate} \pm t_{48}^* \times SE \quad \rightarrow \quad -0.0431 \pm 2.01 \times 0.0108 \quad \rightarrow \quad (-0.0648, -0.0214)$$

We are 95% confident that with each dollar increase in `family_income`, the university's gift aid is predicted to decrease on average by \$0.0214 to \$0.0648.

CONFIDENCE INTERVALS FOR COEFFICIENTS

Confidence intervals for model coefficients can be computed using the t -distribution:

$$b_i \pm t_{df}^* \times SE_{b_i}$$

where t_{df}^* is the appropriate t -value corresponding to the confidence level with the model's degrees of freedom.

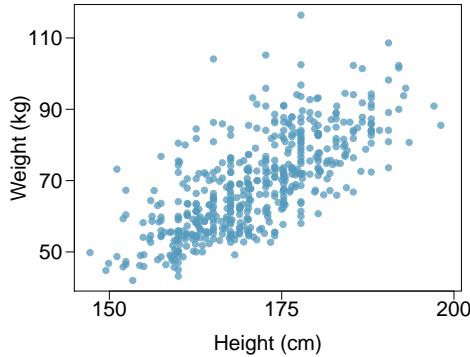
On the topic of intervals in this book, we've focused exclusively on confidence intervals for model parameters. However, there are other types of intervals that may be of interest, including prediction intervals for a response value and also confidence intervals for a mean response value in the context of regression. These two interval types are introduced in an online extra that you may download at

www.openintro.org/d?file=stat_extra_linear_regression_supp

Exercises

In the following exercises, visually check the conditions for fitting a least squares regression line. However, you do not need to report these conditions in your solutions.

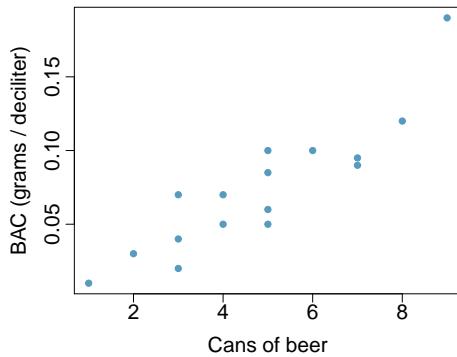
8.31 Body measurements, Part IV. The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.



	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-105.0113	7.5394	-13.93	0.0000
height	1.0176	0.0440	23.13	0.0000

- (a) Describe the relationship between height and weight.
- (b) Write the equation of the regression line. Interpret the slope and intercept in context.
- (c) Do the data provide strong evidence that an increase in height is associated with an increase in weight? State the null and alternative hypotheses, report the p-value, and state your conclusion.
- (d) The correlation coefficient for height and weight is 0.72. Calculate R^2 and interpret it in context.

8.32 Beer and blood alcohol content. Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood.¹⁹ The scatterplot and regression table summarize the findings.

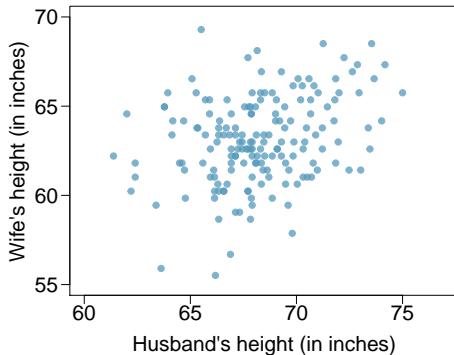


	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.0127	0.0126	-1.00	0.3320
beers	0.0180	0.0024	7.48	0.0000

- (a) Describe the relationship between the number of cans of beer and BAC.
- (b) Write the equation of the regression line. Interpret the slope and intercept in context.
- (c) Do the data provide strong evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion.
- (d) The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R^2 and interpret it in context.
- (e) Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study?

¹⁹J. Malkevitch and L.M. Lesser. *For All Practical Purposes: Mathematical Literacy in Today's World*. WH Freeman & Co, 2008.

8.33 Husbands and wives, Part II. The scatterplot below summarizes husbands' and wives' heights in a random sample of 170 married couples in Britain, where both partners' ages are below 65 years. Summary output of the least squares fit for predicting wife's height from husband's height is also provided in the table.

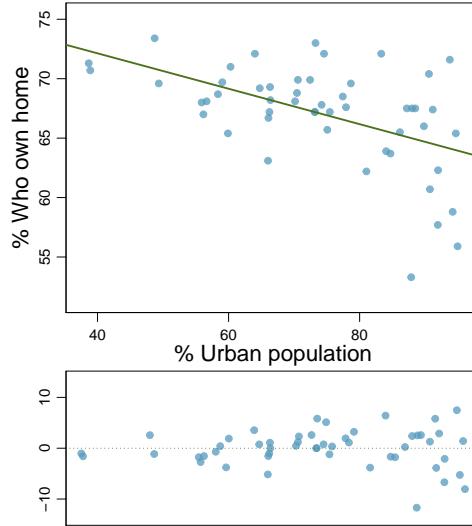


	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	43.5755	4.6842	9.30	0.0000
height_husband	0.2863	0.0686	4.17	0.0000

- Is there strong evidence that taller men marry taller women? State the hypotheses and include any information used to conduct the test.
- Write the equation of the regression line for predicting wife's height from husband's height.
- Interpret the slope and intercept in the context of the application.
- Given that $R^2 = 0.09$, what is the correlation of heights in this data set?
- You meet a married man from Britain who is 5'9" (69 inches). What would you predict his wife's height to be? How reliable is this prediction?
- You meet another married man from Britain who is 6'7" (79 inches). Would it be wise to use the same linear model to predict his wife's height? Why or why not?

8.34 Urban homeowners, Part II. Exercise 8.29 gives a scatterplot displaying the relationship between the percent of families that own their home and the percent of the population living in urban areas. Below is a similar scatterplot, excluding District of Columbia, as well as the residuals plot. There were 51 cases.

- For these data, $R^2 = 0.28$. What is the correlation? How can you tell if it is positive or negative?
- Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for these data?



8.35 Murders and poverty, Part II. Exercise 8.25 presents regression output from a model for predicting annual murders per million from percentage living in poverty based on a random sample of 20 metropolitan areas. The model output is also provided below.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$$s = 5.512 \quad R^2 = 70.52\% \quad R_{adj}^2 = 68.89\%$$

- (a) What are the hypotheses for evaluating whether poverty percentage is a significant predictor of murder rate?
- (b) State the conclusion of the hypothesis test from part (a) in context of the data.
- (c) Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context of the data.
- (d) Do your results from the hypothesis test and the confidence interval agree? Explain.

8.36 Babies. Is the gestational age (time between conception and birth) of a low birth-weight baby useful in predicting head circumference at birth? Twenty-five low birth-weight babies were studied at a Harvard teaching hospital; the investigators calculated the regression of head circumference (measured in centimeters) against gestational age (measured in weeks). The estimated regression line is

$$\widehat{\text{head circumference}} = 3.91 + 0.78 \times \text{gestational age}$$

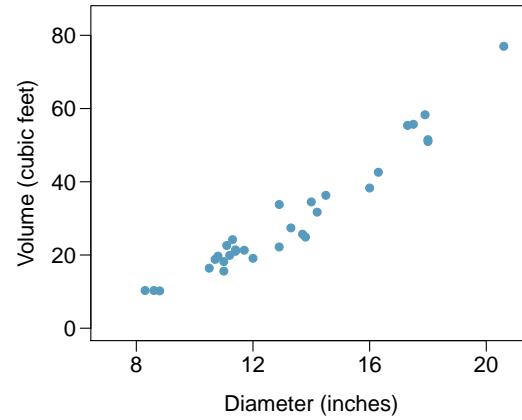
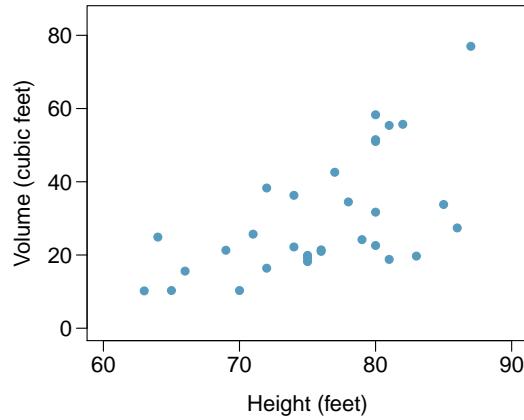
- (a) What is the predicted head circumference for a baby whose gestational age is 28 weeks?
- (b) The standard error for the coefficient of gestational age is 0.35, which is associated with $df = 23$. Does the model provide strong evidence that gestational age is significantly associated with head circumference?

Chapter exercises

8.37 True / False. Determine if the following statements are true or false. If false, explain why.

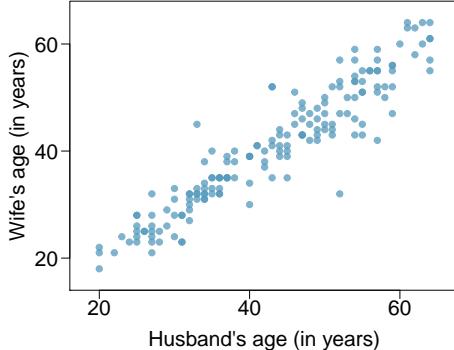
- (a) A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation of 0.5.
- (b) Correlation is a measure of the association between any two variables.

8.38 Trees. The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground.²⁰



- (a) Describe the relationship between volume and height of these trees.
- (b) Describe the relationship between volume and diameter of these trees.
- (c) Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

8.39 Husbands and wives, Part III. Exercise 8.33 presents a scatterplot displaying the relationship between husbands' and wives' ages in a random sample of 170 married couples in Britain, where both partners' ages are below 65 years. Given below is summary output of the least squares fit for predicting wife's age from husband's age.



	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.5740	1.1501	1.37	0.1730
age_husband	0.9112	0.0259	35.25	0.0000

df = 168

- (a) We might wonder, is the age difference between husbands and wives consistent across ages? If this were the case, then the slope parameter would be $\beta_1 = 1$. Use the information above to evaluate if there is strong evidence that the difference in husband and wife ages differs for different ages.
- (b) Write the equation of the regression line for predicting wife's age from husband's age.
- (c) Interpret the slope and intercept in context.
- (d) Given that $R^2 = 0.88$, what is the correlation of ages in this data set?
- (e) You meet a married man from Britain who is 55 years old. What would you predict his wife's age to be? How reliable is this prediction?
- (f) You meet another married man from Britain who is 85 years old. Would it be wise to use the same linear model to predict his wife's age? Explain.

²⁰Source: R Dataset, stat.ethz.ch/R-manual/R-patched/library/datasets/html/trees.html.

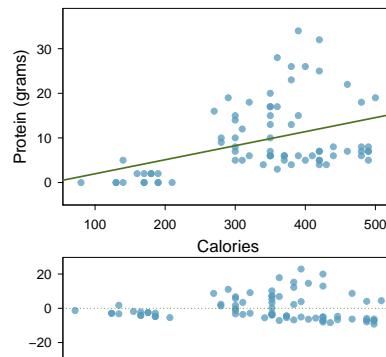
8.40 Cats, Part II. Exercise 8.26 presents regression output from a model for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cat. The model output is also provided below.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000

$$s = 1.452 \quad R^2 = 64.66\% \quad R_{adj}^2 = 64.41\%$$

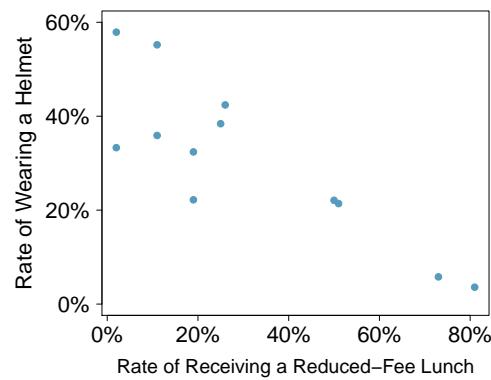
- (a) We see that the point estimate for the slope is positive. What are the hypotheses for evaluating whether body weight is positively associated with heart weight in cats?
- (b) State the conclusion of the hypothesis test from part (a) in context of the data.
- (c) Calculate a 95% confidence interval for the slope of body weight, and interpret it in context of the data.
- (d) Do your results from the hypothesis test and the confidence interval agree? Explain.

8.41 Nutrition at Starbucks, Part II. Exercise 8.22 introduced a data set on nutrition information on Starbucks food menu items. Based on the scatterplot and the residual plot provided, describe the relationship between the protein content and calories of these menu items, and determine if a simple linear model is appropriate to predict amount of protein from the number of calories.



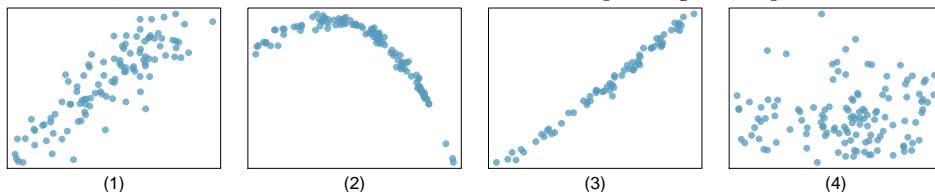
8.42 Helmets and lunches. The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (`lunch`) and the percentage of bike riders in the neighborhood wearing helmets (`helmet`). The average percentage of children receiving reduced-fee lunches is 30.8% with a standard deviation of 26.7% and the average percentage of bike riders wearing helmets is 38.8% with a standard deviation of 16.9%.

- (a) If the R^2 for the least-squares regression line for these data is 72%, what is the correlation between `lunch` and `helmet`?
- (b) Calculate the slope and intercept for the least-squares regression line for these data.
- (c) Interpret the intercept of the least-squares regression line in the context of the application.
- (d) Interpret the slope of the least-squares regression line in the context of the application.
- (e) What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual in the context of the application.



8.43 Match the correlation, Part III. Match each correlation to the corresponding scatterplot.

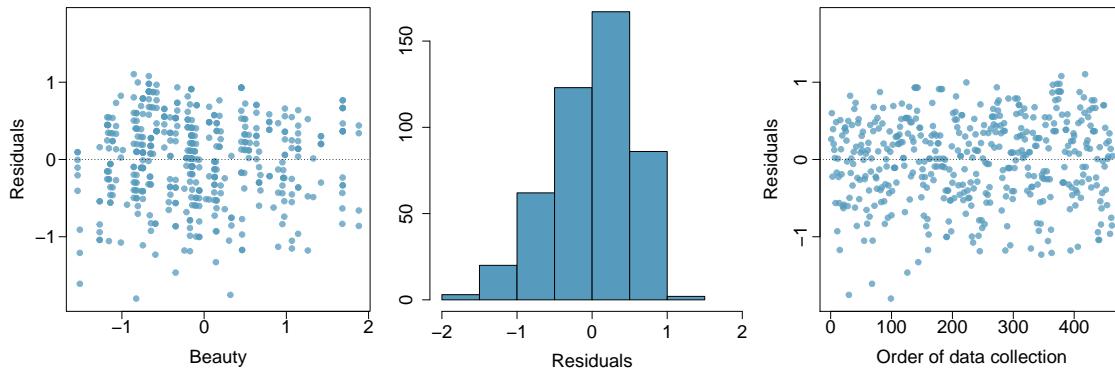
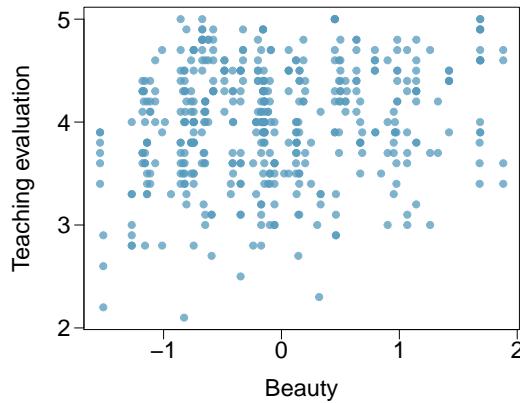
- (a) $r = -0.72$
- (b) $r = 0.07$
- (c) $r = 0.86$
- (d) $r = 0.99$



8.44 Rate my professor. Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors.²¹ The scatterplot below shows the relationship between these variables, and regression output is provided for predicting teaching evaluation score from beauty score.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty	[]	0.0322	4.13	0.0000

- (a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.
- (b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.
- (c) List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.



²¹Daniel S Hamermesh and Amy Parker. “Beauty in the classroom: Instructors’ pulchritude and putative pedagogical productivity”. In: *Economics of Education Review* 24.4 (2005), pp. 369–376.