Chapter 8

Introduction to linear regression

Linear regression is a very powerful statistical technique. Many people have some familiarity with regression just from reading the news, where graphs with 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.

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

$$y = 5 + 57.49x$$

Imagine what a perfect linear relationship would mean: you would 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 y a college may offer a prospective student. However, there would still be variability in financial support, even when comparing students whose families have similar financial backgrounds.

Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line:

β_0, β_1
Linear model parameters

$$y = \beta_0 + \beta_1 x \tag{8.1}$$

where β_0 and β_1 represent two model parameters (β is the Greek letter *beta*). (This use of β has nothing to do with the β we used to describe the probability of a Type II error.) These 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.

It is rare for all of the data to fall on a straight line, as seen in the three scatterplots 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 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

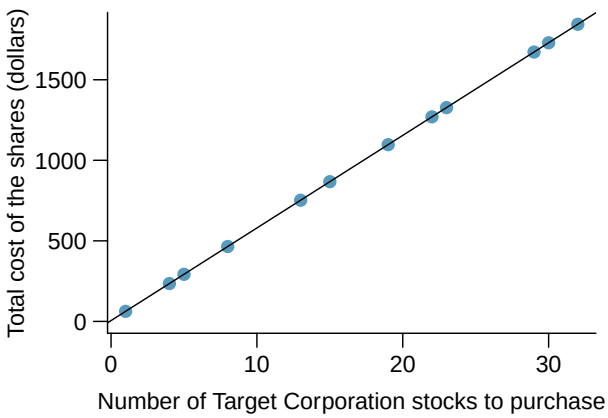

Figure 8.1: Requests from twelve separate buyers were simultaneously placed with a trading company to purchase Target Corporation stock (ticker `TGT`, April 26th, 2012), and the total cost of the shares were reported. Because the cost is computed using a linear formula, the linear fit is perfect.

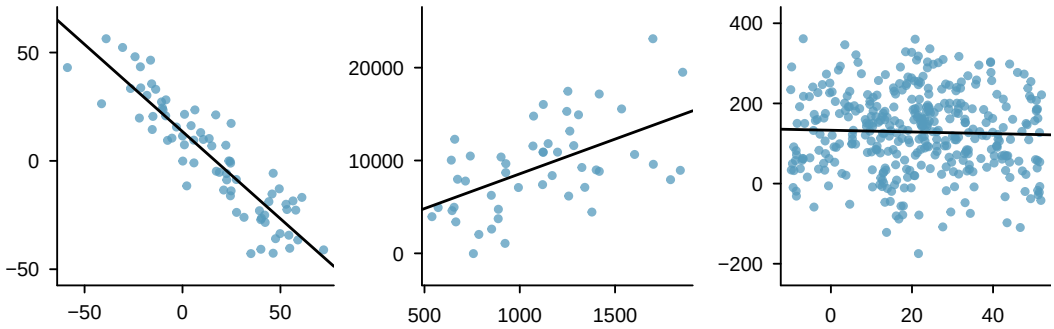

Figure 8.2: Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

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 different criteria for line-fitting, and we will also learn about the uncertainty associated with estimates of model parameters.

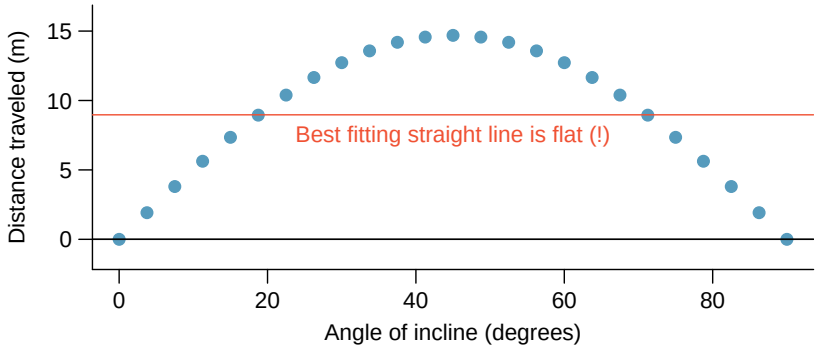

Figure 8.3: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

We will also see examples in this chapter 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 strong relationship between the variables even though the trend is not linear. We will discuss nonlinear trends in this chapter and the next, but the details of fitting nonlinear models are saved for a later course.

8.1 Line fitting, residuals, and correlation

It is helpful to think deeply about the line fitting process. In this section, we examine criteria for identifying a linear model and introduce a new statistic, *correlation*.

8.1.1 Beginning with straight lines

Scatterplots were introduced in Chapter 1 as a graphical technique to present two numerical variables simultaneously. Such plots permit the relationship between the variables to be examined with ease. Figure 8.4 shows a scatterplot for the head length and total length of 104 brushtail possums from Australia. 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.

Straight lines should only be used when the data appear to have a linear relationship, such as the case shown in the left panel of Figure 8.6. The right panel of Figure 8.6 shows a case where a curved line would be more useful in understanding the relationship between the two variables.

Caution: Watch out for curved trends
We only consider models based on straight lines in this chapter. If data show a nonlinear trend, like that in the right panel of Figure 8.6, more advanced techniques should be used.

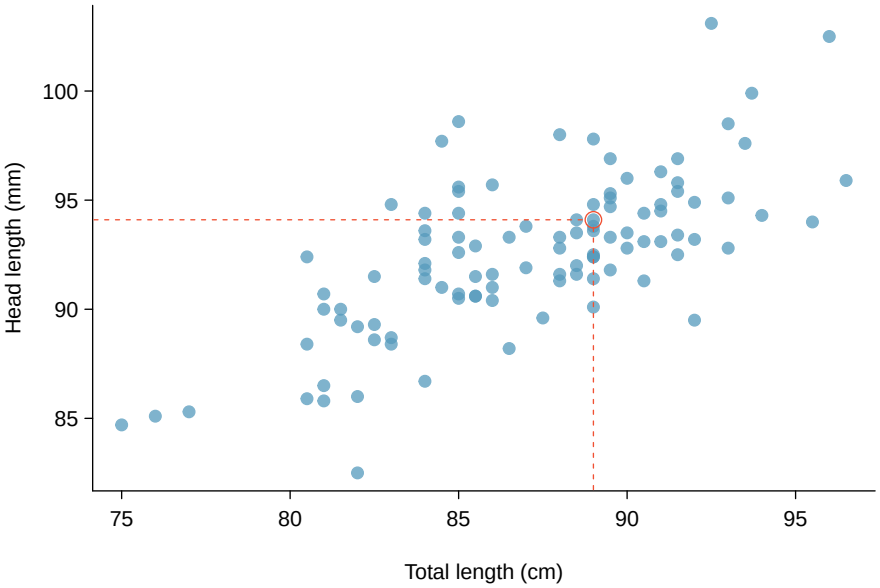

Figure 8.4: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89cm is highlighted.

Figure 8.5: The common brushtail possum of Australia.

Photo by wollombi on Flickr: www.flickr.com/photos/wollombi/58499575

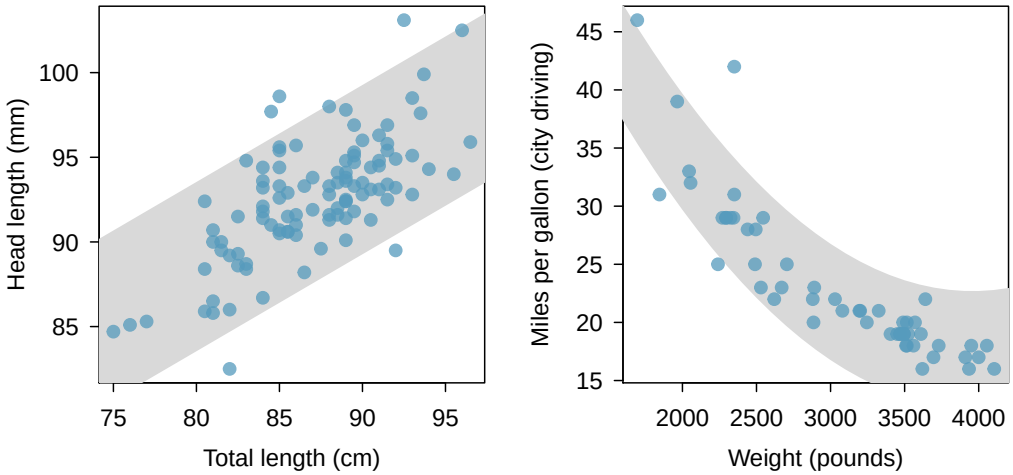

Figure 8.6: The figure on the left shows head length versus total length, and reveals that many of the points could be captured by a straight band. On the right, we see that a curved band is more appropriate in the scatterplot for `weight` and `mpgCity` from the `cars` data set.

8.1.2 Fitting a line by eye

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 8.7. The equation for this line is

$$\hat{y} = 41 + 0.59x \tag{8.2}$$

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

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

A "hat" on y is used to signify that this is an estimate. This 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.

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. 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.

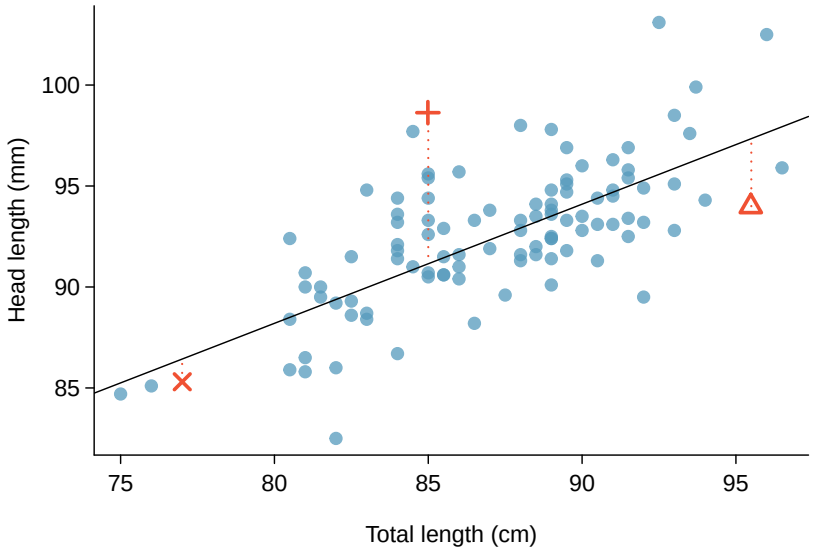

Figure 8.7: A reasonable linear model was fit to represent the relationship between head length and total length.

Three observations are noted specially in Figure 8.7. The observation marked by an "×" has a small, negative residual of about -1; the observation marked by "+" has a large residual of about +7; and the observation marked by "△" 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 "△" is larger than that of "×" 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 $(\hat{y}_i)$:

$$\text{residual}_i = y_i - \hat{y}_i$$

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

Example 8.3 The linear fit shown in Figure 8.7 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 "×" on the plot. Check it against the earlier visual estimate, -1.

We first compute the predicted value of point "×" 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:

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

This is very close to the visual estimate of -1.

⊙ **Guided Practice 8.4** If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?[1]

⊙ **Guided Practice 8.5** Compute the residuals for the observations $(85.0, 98.6)$ ("+" in the figure) and $(95.5, 94.0)$ ("△") 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 8.8 for the regression line in Figure 8.7. 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.

From the residual plot, we can better estimate the **standard deviation of the residuals**, often denoted by the letter s. The standard deviation of the residuals tells us the average size of the residuals. As such, it is a measure of the average deviation between the y values and the regression line. In other words, it tells us the average prediction error using the linear model.

● **Example 8.6** Estimate the standard deviation of the residuals for predicting head length from total length using the regression line. Also, interpret the quantity in context.

To estimate this graphically, we use the residual plot. The approximate 68, 95 rule for standard deviations applies. Approximately 2/3 of the points are within ± 2.5 and approximately 95% of the points are within ± 5, so 2.5 is a good estimate for the standard deviation of the residuals. On average, the prediction of head length is off by about 2.5 cm.

Standard deviation of the residuals
The standard deviation of the residuals, often denoted by the letter s, tells us the average error in the predictions using the regression model. It can be estimated from a residual plot.

[1] If 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

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

This was close to the earlier estimate of 7.

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

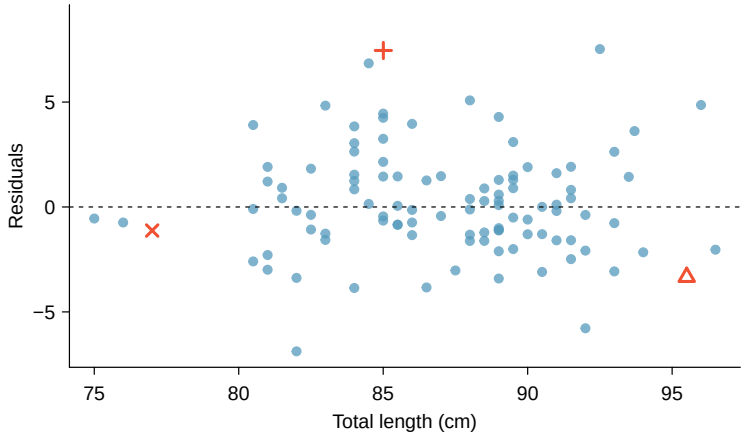

Figure 8.8: Residual plot for the model in Figure 8.7.

Example 8.7 One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. Figure 8.9 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.

8.2 Fitting a line by least squares regression

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

This section considers family income and gift aid data from a random sample of fifty students in the 2011 freshman class of Elmhurst College in Illinois.[5] 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 8.12 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.

⊙ **Guided Practice 8.10** Is the correlation positive or negative in Figure 8.12?[6]

8.2.1 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. Perhaps our criterion could minimize the sum of the residual magnitudes:

$$|y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + \cdots + |y_n - \hat{y}_n| \tag{8.11}$$

which we could accomplish with a computer program. The resulting dashed line shown in Figure 8.12 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:

$$(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 \tag{8.12}$$

The line that minimizes this **least squares criterion** is represented as the solid line in Figure 8.12. This is commonly called the **least squares line**. The following are three possible reasons to choose Criterion (8.12) over Criterion (8.11):

1. It is the most commonly used method.

2. Computing the line based on Criterion (8.12) is much easier by hand and in most 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 Criterion (8.12) is typically most helpful.[7]

[5] These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled *What Students Really Pay to Go to College* published online by *The Chronicle of Higher Education*: chronicle.com/article/What-Students-Really-Pay-to-Go/131435

[6] Larger 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.

8.2.2 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 8.13), 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 will discuss in greater depth in Section 8.3. An example of non-normal residuals is shown in the second panel of Figure 8.13.

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 8.13.

These conditions are best checked using a residual plot. If a residual plot has no pattern, such as a U-shape or the presence of outliers or non-constant variability in the residuals, then the conditions above may be considered to be satisfied.

TIP: Use a residual plot to determine if a linear model is appropriate
When a residual plot appears as a random cloud of points, a linear model is generally appropriate. If a residual plot has any type of pattern, a linear model is not appropriate.

Be cautious about applying regression to data collected sequentially in what is called a **time series**. Such data may have an underlying structure that should be considered in a model and analysis.

[7] There are applications where Criterion (8.11) may be more useful, and there are plenty of other criteria we might consider. However, this book only applies the least squares criterion.

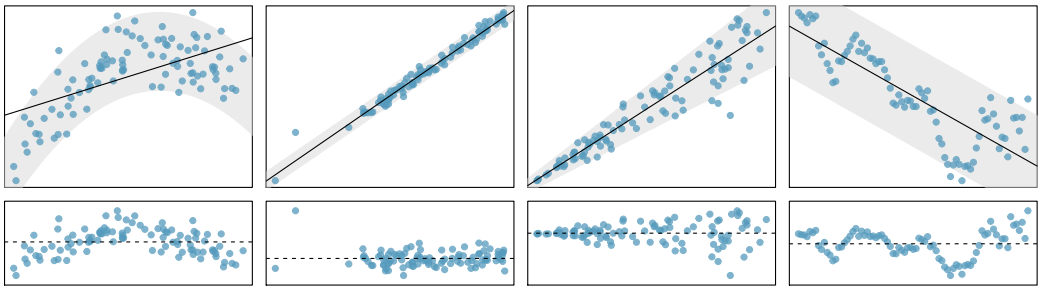

Figure 8.13: Four examples showing when the methods in this chapter are insufficient to apply to the data. In the left panel, a straight line does not fit the data. In the second panel, there are outliers; two points on the left are relatively distant from the rest of the data, and one of these points is very far away from the line. In the third panel, the variability of the data around the line increases with larger values of x. In the last panel, a time series data set is shown, where successive observations are highly correlated.

⊙ **Guided Practice 8.13** Should we have concerns about applying least squares regression to the Elmhurst data in Figure 8.12?[8]

8.2.3 Finding the least squares line

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

$$\widehat{aid} = \beta_0 + \beta_1 \times family\_income$$

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 4-6, 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

$$b_1 = r\frac{s_y}{s_x} \tag{8.14}$$

  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 mean of the horizontal variable (from the data) and $\bar{y}$ is the mean of the vertical variable, then the point $(\bar{x}, \bar{y})$ is on the least squares line. Plugging this point in for x and y in the least squares equation and solving for b_0 gives

$$\bar{y} = b_0 + b_1\bar{x} \qquad\qquad b_0 = \bar{y} - b_1\bar{x} \tag{8.15}$$

[8]The 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.

When solving for the y-intercept, first find the slope, b_1, and plug the slope and the point $(\bar{x}, \bar{y})$ into the least squares equation.

We use b_0 and b_1 to represent the point estimates of the parameters β_0 and β_1.

b_0, b_1 Sample estimates of β_0, β_1

⊙ **Guided Practice 8.16** Table 8.14 shows the sample means for the family income and gift aid as \$101,800 and \$19,940, respectively. Plot the point $(101.8, 19.94)$ on Figure 8.12 on page 334 to verify it falls on the least squares line (the solid line).[9]

	family income, in \$1000s ("x")	gift aid, in \$1000s ("y")
mean	$\bar{x} = 101.8$	$\bar{y} = 19.94$
sd	$s_x = 63.2$	$s_y = 5.46$
		$r = -0.499$

Table 8.14: Summary statistics for family income and gift aid.

⊙ **Guided Practice 8.17** Using the summary statistics in Table 8.14, compute the slope and y-intercept for the regression line of gift aid against family income. Write the equation of the regression line.[10]

We mentioned earlier that a computer is usually used to compute the least squares line. A summary table based on computer output is shown in Table 8.15 for the Elmhurst data. The first column of numbers provides estimates for b_0 and b_1, respectively. Compare these to the result from Guided Practice 8.17.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.3193	1.2915	18.83	0.0000
family_income	-0.0431	0.0108	-3.98	0.0002

Table 8.15: Summary of least squares fit for the Elmhurst data. Compare the parameter estimates in the first column to the results of Guided Practice 8.17.

[9] If you need help finding this location, draw a straight line up from the x-value of 100 (or thereabout). Then draw a horizontal line at 20 (or thereabout). These lines should intersect on the least squares line.

[10] Apply Equations (8.14) and (8.15) with the summary statistics from Table 8.14 to compute the slope and y-intercept:

$$b_1 = r\frac{s_y}{s_x} = (-0.499)\frac{5.46}{63.2} = -0.0431$$
$$b_0 = \bar{y} - b_1\bar{x} = 19.94 - (-0.0431)(101.8) = 24.3$$
$$\hat{y} = 24.3 - 0.0431x \qquad \text{or} \qquad \widehat{aid} = 24.3 - 0.0431\text{family\_income}$$

Example 8.18 Examine the second, third, and fourth columns in Table 8.15. Can you guess what they represent?

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: -0.0431. The second column is a standard error for this point estimate: 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 8.19 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.

8.2.4 Interpreting regression line parameter estimates

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

Example 8.20 The slope and intercept estimates for the Elmhurst data are -0.0431 and 24.3. 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 $b_0 = 24.3$ (in \$1000s) 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 in a linear model

- The slope, b_1, describes the estimated difference in the y variable if the explanatory variable x for a case happened to be one unit larger.
- The y-intercept, b_0, describes the average or predicted outcome of y if $x = 0$. The linear model must be valid all the way to $x = 0$ for this to make sense, which in many applications is not the case.

8.2.5 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 6^{th} 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, 2010 [11]

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 8.21 Use the model $\widehat{aid} = 24.3 - 0.0431 \times family\_income$ to estimate the aid of another freshman student whose family had income of $1 million.

Recall that the units of family income are in $1000s, so we want to calculate the aid for $family\_income = 1000$:

$$\widehat{aid} = 24.3 - 0.0431 \times family\_income$$
$$\widehat{aid} = 24.3 - 0.431(1000) = -18.8$$

The model predicts this student will have -$18,800 in aid (!). Elmhurst College cannot (or at least does not) require any 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.

8.2.6 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 coefficient, r. However, it is more common to explain the strength of a linear fit using R^2, called **R-squared** or the **explained variance**. If provided with a linear model, we might like to describe how closely the data cluster around the linear fit.

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

[11] http://www.colbertnation.com/the-colbert-report-videos/269929/

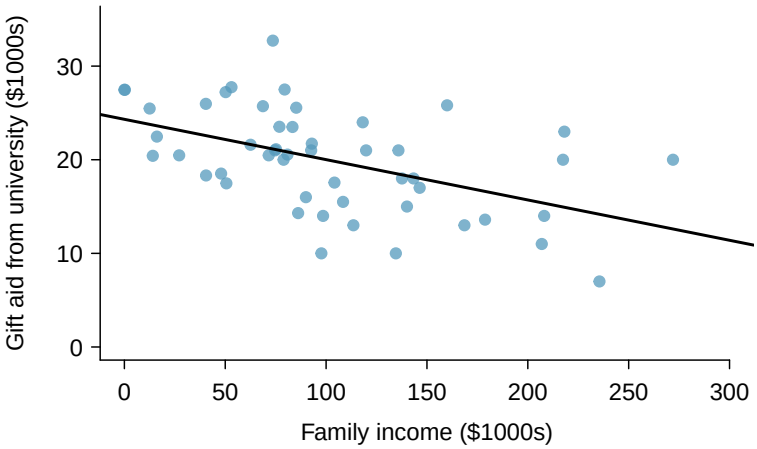

Figure 8.16: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College, shown with the least squares regression line.

Figure 8.16. The variance of the response variable, aid received, is $s^2_{aid} = 29.8$. 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: $s^2_{RES} = 22.4$. In short, there was a reduction of

$$\frac{s^2_{aid} - s^2_{RES}}{s^2_{aid}} = \frac{29.8 - 22.4}{29.8} = \frac{7.5}{29.8} = 0.25$$

This is how we compute the R^2 value.[12] It also corresponds to the square of the correlation coefficient, r, that is, $R^2 = r^2$.

$$R^2 = 0.25 \qquad\qquad r = -0.499$$

R^2 is the explained variance

R^2 is always between 0 and 1, inclusive. It tells us the proportion of variation in the y values that is explained by a regression model. The higher the value of R^2, the better the model "explains" the reponse variable.

⊙ **Guided Practice 8.22** 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]

⊙ **Guided Practice 8.23** If a linear model has an R^2 or explained variance of 0.94, what is the correlation coefficient?[14]

[12] $R^2 = 1 - \frac{s^2_{RES}}{s^2_y}$

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

[14] We take the square root of R^2 and get 0.97, but we must be careful, because r could be 0.97 *or* -0.97. Without knowing the slope or seeing the scatterplot, we have no way of knowing if r is positive or negative.

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 8.26 There are six plots shown in Figure 8.19 along with the least squares line and residual plots. For each scatterplot and residual plot pair, identify any obvious 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 8.19. 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 8.26 – 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

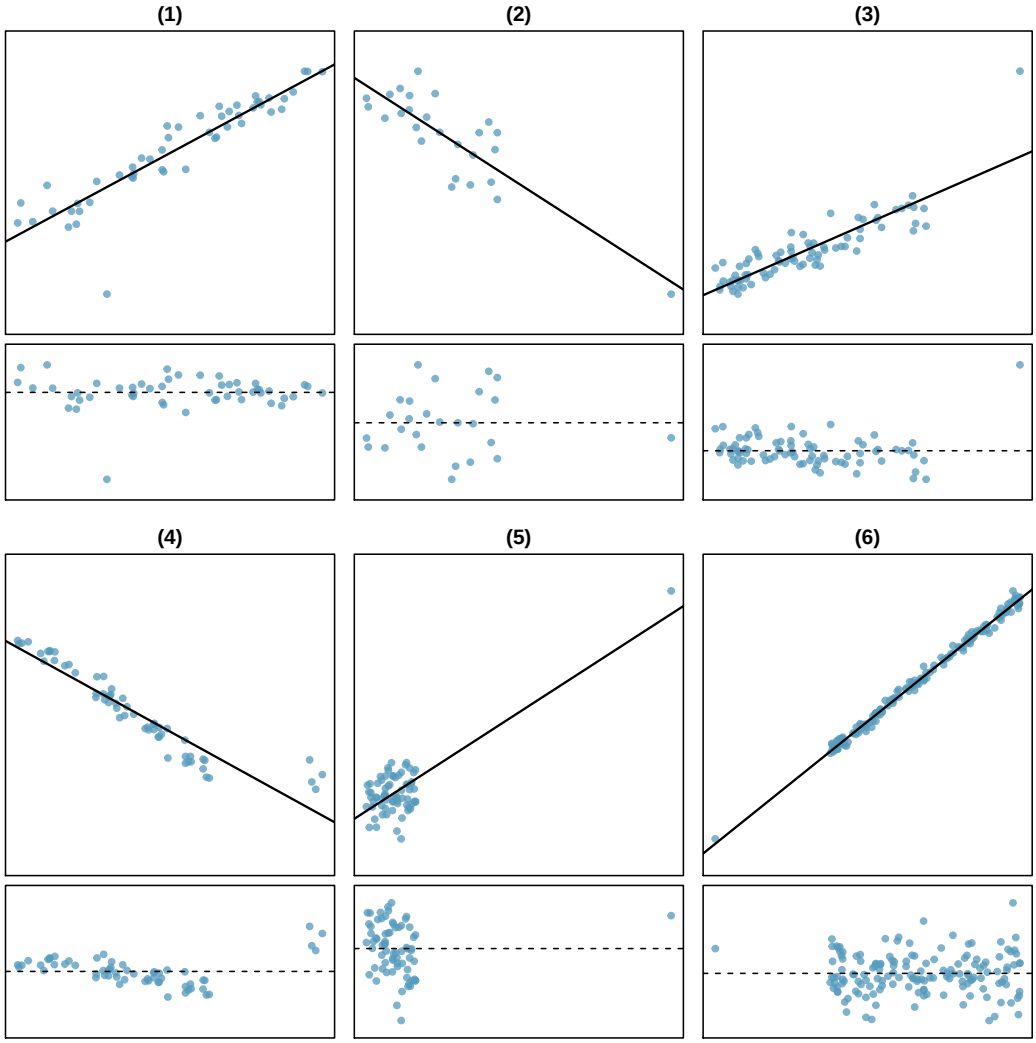

Figure 8.19: Six plots, each with a least squares line and residual plot. All data sets have at least one outlier.

financial firm ignored the largest market swings – the "outliers" – they would soon go bankrupt by making poorly thought-out investments.

Caution: Don't ignore outliers when fitting a final model
If there are outliers in the data, they should not be removed or ignored without a good reason. Whatever final model is fit to the data would not be very helpful if it ignores the most exceptional cases.

Caution: Outliers for a categorical predictor with two levels
Be cautious about using a categorical predictor when one of the levels has very few observations. When this happens, those few observations become influential points.

Appendix A

End of chapter exercise solutions

8 Introduction to linear regression

8.1 (a) The residual plot will show randomly distributed residuals around 0. The variance is also approximately constant. (b) The residuals will show a fan shape, with higher variability for smaller x. There will also be many points on the right above the line. There is trouble with the model being fit here.

8.3 (a) Strong relationship, but a straight line would not fit the data. (b) Strong relationship, and a linear fit would be reasonable. (c) Weak relationship, and trying a linear fit would be reasonable. (d) Moderate relationship, but a straight line would not fit the data. (e) Strong relationship, and a linear fit would be reasonable. (f) Weak relationship, and trying a linear fit would be reasonable.

8.5 (a) Exam 2 since there is less of a scatter in the plot of final exam grade versus exam 2. Notice that the relationship between Exam 1 and the Final Exam appears to be slightly nonlinear. (b) Exam 2 and the final are relatively close to each other chronologically, or Exam 2 may be cumulative so has greater similarities in material to the final exam. Answers may vary for part (b).

8.7 (a) $R = -0.7 \rightarrow (4)$. (b) $R = 0.45 \rightarrow (3)$. (c) $R = 0.06 \rightarrow (1)$. (d) $R = 0.92 \rightarrow (2)$.

8.9 (a) The relationship is positive, weak, and possibly linear. However, there do appear to be some anomalous observations along the left where several students have the same height that is notably far from the cloud of the other points. Additionally, there are many students who appear not to have driven a car, and they are represented by a set of points along the bottom of the scatterplot. (b) There is no obvious explanation why simply being tall should lead a person to drive faster. However, one confounding factor is gender. Males tend to be taller than females on average, and personal experiences (anecdotal) may suggest they drive faster. If we were to follow-up on this suspicion, we would find that sociological studies confirm this suspicion. (c) Males are taller on average and they drive faster. The gender variable is indeed an important confounding variable.

8.11 (a) There is a somewhat weak, positive, possibly linear relationship between the distance traveled and travel time. There is clustering near the lower left corner that we should take special note of. (b) Changing the units will not change the form, direction or strength of the relationship between the two variables. If longer distances measured in miles are associated with longer travel time measured in minutes, longer distances measured in kilometers will be associated with longer travel time measured in hours. (c) Changing units doesn't affect correlation: $R = 0.636$.

8.13 (a) There is a moderate, positive, and linear relationship between shoulder girth and height. (b) Changing the units, even if just for one of the variables, will not change the form, direction or strength of the relationship between the two variables.

8.15 In each part, we may write the husband ages as a linear function of the wife ages: (a) $age_H = age_W + 3$; (b) $age_H = age_W - 2$; and (c) $age_H = 2 \times age_W$. Therefore, the correlation will be exactly 1 in all three parts. An alternative way to gain insight into this solution is to create a mock data set, such as a data set of 5 women with ages 26, 27, 28, 29, and 30 (or some other set of ages). Then, based on the description, say for part (a), we can compute their husbands' ages as 29, 30, 31, 32, and 33. We can plot these points to see they fall on a straight line, and they always will. The same approach can be applied to the other parts as well.

8.17 (a) There is a positive, very strong, linear association between the number of tourists and spending. (b) Explanatory: number of tourists (in thousands). Response: spending (in millions of US dollars). (c) We can predict spending for a given number of tourists using a regression line. This may be useful information for determining how much the country may want to spend in advertising abroad, or to forecast expected

revenues from tourism. (d) Even though the relationship appears linear in the scatterplot, the residual plot actually shows a nonlinear relationship. This is not a contradiction: residual plots can show divergences from linearity that can be difficult to see in a scatterplot. A simple linear model is inadequate for modeling these data. It is also important to consider that these data are observed sequentially, which means there may be a hidden structure that it is not evident in the current data but that is important to consider.

8.19 (a) First calculate the slope: $b_1 = R \times s_y/s_x = 0.636 \times 113/99 = 0.726$. Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$: $\bar{y} = b_0 + b_1 \times \bar{x}$. Plug in $\bar{x}$, $\bar{y}$, and b_1, and solve for b_0: 51. Solution: $\widehat{travel\ time} = 51 + 0.726 \times distance$. (b) b_1: For each additional mile in distance, the model predicts an additional 0.726 minutes in travel time. b_0: When the distance traveled is 0 miles, the travel time is expected to be 51 minutes. It does not make sense to have a travel distance of 0 miles in this context. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself. (c) $R^2 = 0.636^2 = 0.40$. About 40% of the variability in travel time is accounted for by the model, i.e. explained by the distance traveled. (d) $\widehat{travel\ time} = 51 + 0.726 \times distance = 51 + 0.726 \times 103 \approx 126$ minutes. (Note: we should be cautious in our predictions with this model since we have not yet evaluated whether it is a well-fit model.) (e) $e_i = y_i - \hat{y}_i = 168 - 126 = 42$ minutes. A positive residual means that the model underestimates the travel time. (f) No, this calculation would require extrapolation.

8.21 (a) $\sqrt{R^2} = 0.849$. Since the trend is negative, R is also negative: $R = -0.849$. (b) $b_0 = 55.34$. $b_1 = -0.537$. (c) For a neighborhood with 0% reduced-fee lunch, we would expect 55.34% of the bike riders to wear helmets. (d) For every additional percentage point of reduced fee lunches in a neighborhood, we would expect 0.537% fewer kids to be wearing helmets. (e) $\hat{y} = 40 \times (-0.537) + 55.34 = 33.86$, $e = 40 - \hat{y} = 6.14$. There are 6.14% more bike riders wearing helmets than predicted by the regression model in this neighborhood.

8.23 (a) The outlier is in the upper-left corner. Since it is horizontally far from the center of the data, it is a point with high leverage. Since the slope of the regression line would be very different if fit without this point, it is also an influential point. (b) The outlier is located in the lower-left corner. It is horizontally far from the rest of the data, so it is a high-leverage point. The line again would look notably different if the fit excluded this point, meaning it the outlier is influential. (c) The outlier is in the upper-middle of the plot. Since it is near the horizontal center of the data, it is not a high-leverage point. This means it also will have little or no influence on the slope of the regression line.

8.25 (a) There is a negative, moderate-to-strong, somewhat linear relationship between percent of families who own their home and the percent of the population living in urban areas in 2010. There is one outlier: a state where 100% of the population is urban. The variability in the percent of homeownership also increases as we move from left to right in the plot. (b) The outlier is located in the bottom right corner, horizontally far from the center of the other points, so it is a point with high leverage. It is an influential point since excluding this point from the analysis would greatly affect the slope of the regression line.