CHAPTER 9

Regression Wisdom

A S ***Activity:*** **Construct a Plot with a Given Slope.** How's your feel for regression lines? Can you make a scatterplot that has a specified slope?

Regression may be the most widely used Statistics method. It is used every day throughout the world to predict customer loyalty, numbers of admissions at hospitals, sales of automobiles, and many other things. Because regression is so widely used, it's also widely abused and misinterpreted. This chapter presents examples of regressions in which things are not quite as simple as they may have seemed at first, and shows how you can still use regression to discover what the data have to say.

Getting the "Bends": When the Residuals Aren't Straight

We can't *know* whether the **Linearity Assumption** is true, but we can see if it's *plausible* by checking the **Straight Enough Condition.**

No regression analysis is complete without a display of the residuals to check that the linear model is reasonable. Because the residuals are what is "left over" after the model describes the relationship, they often reveal subtleties that were not clear from a plot of the original data. Sometimes these are additional details that help confirm or refine our understanding. Sometimes they reveal violations of the regression conditions that require our attention.

The fundamental assumption in working with a linear model is that the relationship you are modeling is, in fact, linear. That sounds obvious, but when you fit a regression, you can't take it for granted. Often it's hard to tell from the scatterplot you looked at before you fit the regression model. Sometimes you can't see a bend in the relationship until you plot the residuals.

Jessica Meir and Paul Ponganis study emperor penguins at the Scripps Institution of Oceanography's Center for Marine Biotechnology and Biomedicine at the University of California at San Diego. Says Jessica:

> *Emperor penguins are the most accomplished divers among birds, making routine dives of 5–12 minutes, with the longest recorded dive over 27 minutes. These birds can also dive to depths of over 500 meters! Since air-breathing animals like penguins must hold their breath while submerged, the duration of any given dive depends on how much oxygen is in the bird's body at the beginning of the dive, how quickly that oxygen gets used,*

and the lowest level of oxygen the bird can tolerate. The rate of oxygen depletion is primarily determined by the penguin's heart rate. Consequently, studies of heart rates during dives can help us understand how these animals regulate their oxygen consumption in order to make such impressive dives.

The researchers equip emperor penguins with devices that record their heart rates during dives. Here's a scatterplot of the *Dive Heart Rate* (beats per minute) and the *Duration* (minutes) of dives by these high-tech penguins.

The scatterplot looks fairly linear with a moderately strong negative association ($R^2 = 71.5\%$). The linear regression equation

$$\widehat{DiveHeartRate} = 96.9 - 5.47\ Duration$$

says that for longer dives, the average *Dive Heart Rate* is lower by about 5.47 beats per dive minute, starting from a value of 96.9 beats per minute.

The scatterplot of the residuals against *Duration* holds a surprise. The Linearity Assumption says we should not see a pattern, but instead there's a bend, starting high on the left, dropping down in the middle of the plot, and rising again at the right. Graphs of residuals often reveal patterns such as this that were easy to miss in the original scatterplot.

Now looking back at the original scatterplot, you may see that the scatter of points isn't really straight. There's a slight bend to that plot, but the bend is much easier to see in the residuals. Even though it means rechecking the Straight Enough Condition *after* you find the regression, it's always a good idea to check your scatterplot of the residuals for bends that you might have overlooked in the original scatterplot.

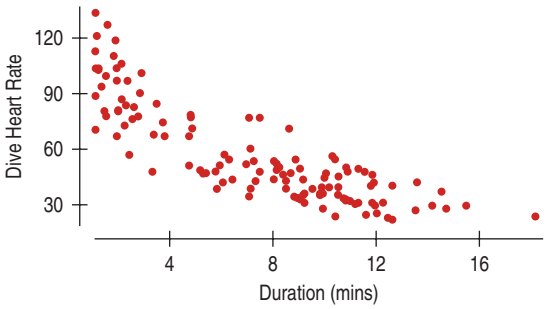


FIGURE 9.1

The scatterplot of Dive Heart Rate *in beats per minute (bpm) vs.* Duration *(minutes) shows a strong, roughly linear, negative association.*

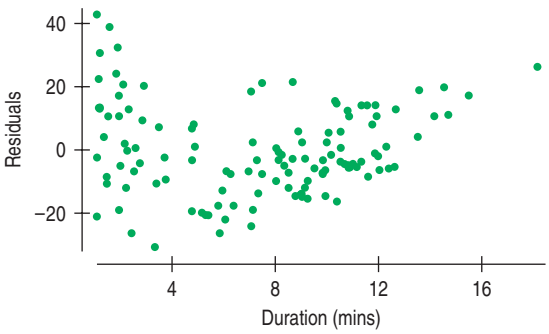


FIGURE 9.2

Plotting the residuals against Duration *reveals a bend. It was also in the original scatterplot, but here it's easier to see.*

Sifting Residuals for Groups

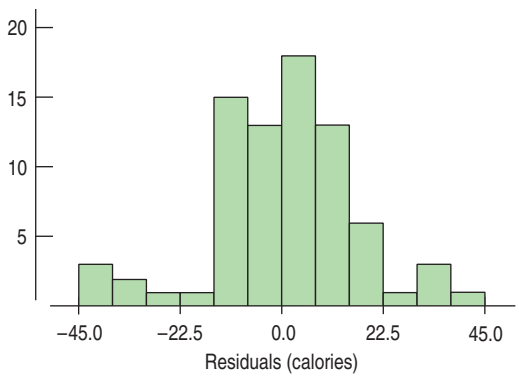


FIGURE 9.3

A histogram of the regression residuals shows small modes both above and below the central large mode. These may be worth a second look.

In the Step-By-Step analysis in Chapter 8 to predict *Calories* from *Sugar* content in breakfast cereals, we examined a scatterplot of the residuals. Our first impression was that it had no particular structure—a conclusion that supported using the regression model. But let's look again.

Here's a histogram of the residuals. How would you describe its shape? It looks like there might be small modes on both sides of the central body of the data. One group of cereals seems to stand out as having large negative residuals, with fewer calories than we might have predicted, and another stands out with large positive residuals. The calories in these cereals were underestimated by the model. Whenever we suspect multiple modes, we ask whether they are somehow different.

On the next page is the residual plot, with the points in those modes marked. Now we can see that those two groups stand away from the central pattern in the scatterplot. The high-residual cereals are Just Right Fruit & Nut; Muesli Raisins, Dates & Almonds; Peaches & Pecans; Mueslix Crispy Blend; and Nutri-Grain Almond Raisin. Do these cereals seem to have something in common? They all present themselves as "healthy." This might be surprising, but in fact, "healthy" cereals

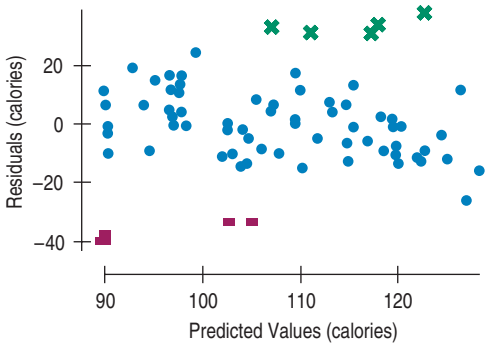


FIGURE 9.4

A scatterplot of the residuals vs. predicted values for the cereal regression. The green "x" points are cereals whose calorie content is higher than the linear model predicts. The red "–" points show cereals with fewer calories than the model predicts. Is there something special about these cereals?

often contain more fat, and therefore more calories, than we might expect from looking at their sugar content alone.

The low-residual cereals are Puffed Rice, Puffed Wheat, three bran cereals, and Golden Crisps. You might not have grouped these cereals together before. What they have in common is a low calorie count *relative to their sugar content*—even though their sugar contents are quite different.

These observations may not lead us to question the overall linear model, but they do help us understand that other factors may be part of the story. An examination of residuals often leads us to discover groups of observations that are different from the rest.

When we discover that there is more than one group in a regression, we may decide to analyze the groups separately, using a different model for each group. Or we can stick with the original model and simply note that there are groups that are a little different. Either way, the model will be wrong, but useful, so it will improve our understanding of the data.

Subsets

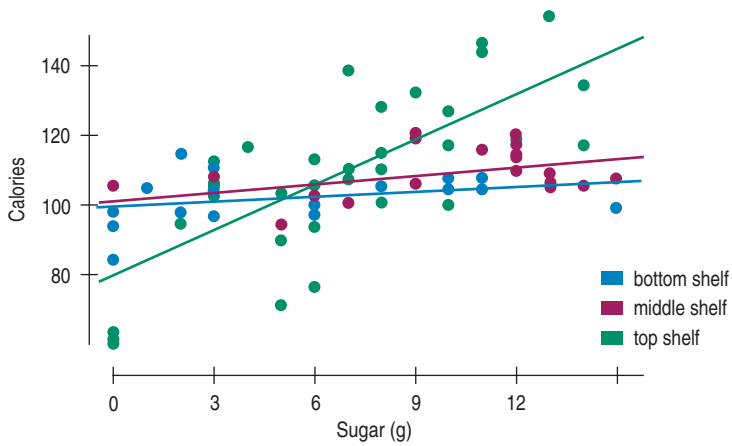
Here's an important unstated condition for fitting models: **All the data must come from the same population.**

Cereal manufacturers aim cereals at different segments of the market. Supermarkets and cereal manufacturers try to attract different customers by placing different types of cereals on certain shelves. Cereals for kids tend to be on the "kid's shelf," at their eye level. Toddlers wouldn't be likely to grab a box from this shelf and beg, "Mom, can we please get this All-Bran with Extra Fiber?"

Should we take this extra information into account in our analysis? Figure 9.5 shows a scatterplot of *Calories* and *Sugar*, colored according to the shelf on which the cereals were found and with a separate regression line fit for each. The top shelf is clearly different. We might want to report two regressions, one for the top shelf and one for the bottom two shelves.[1]

FIGURE 9.5

Calories *and* Sugar *colored according to the shelf on which the cereal was found in a supermarket, with regression lines fit for each shelf individually. Do these data appear homogeneous? That is, do all the cereals seem to be from the same population of cereals? Or are there different kinds of cereals that we might want to consider separately?*



Extrapolation: Reaching Beyond the Data

Linear models give a predicted value for each case in the data. Put a new x-value into the equation, and it gives a predicted value, $\hat{y}$, to go with it. But when the new x-value lies far from the data we used to build the regression, how trustworthy is the prediction?

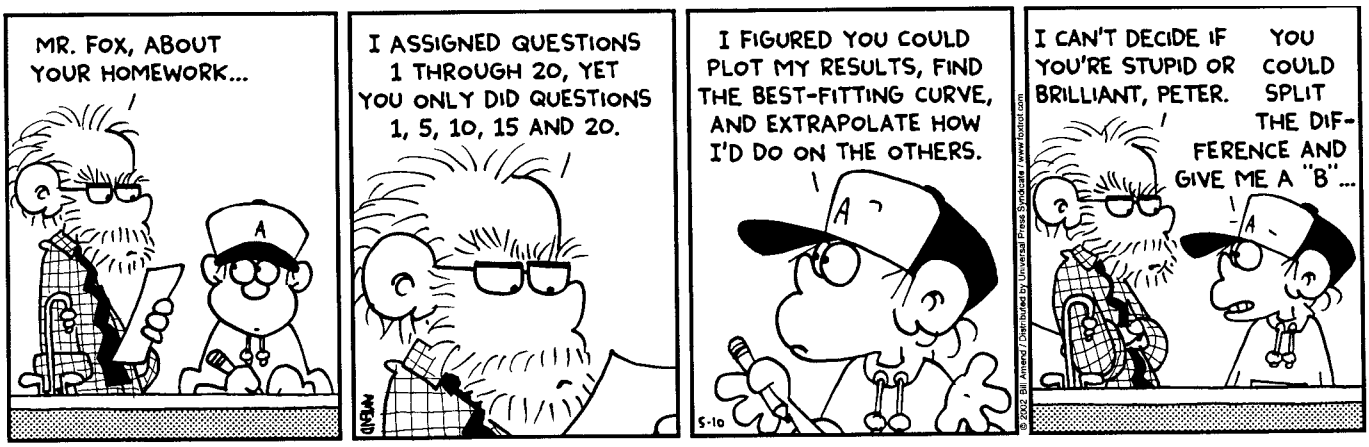
[1] More complex models can take into account both sugar content and shelf information. This kind of *multiple regression* model is a natural extension of the model we're using here. You can learn about such models in Chapter 29 on the DVD.

A S ***Case Study:*** **Predicting Manatee Kills.** Can we use regression to predict the number of manatees that will be killed by power boats this year?

"Prediction is difficult, especially about the future."
—Niels Bohr, Danish physicist

The simple answer is that the farther the new x-value is from $\bar{x}$, the less trust we should place in the predicted value. Once we venture into new x territory, such a prediction is called an **extrapolation.** Extrapolations are dubious because they require the very questionable assumption that nothing about the relationship between x and y changes even at extreme values of x and beyond.

Extrapolations can get us into deep trouble. When the x-variable is *Time,* extrapolation becomes an attempt to peer into the future. People have always wanted to see into the future, and it doesn't take a crystal ball to foresee that they always will. In the past, seers, oracles, and wizards were called on to predict the future. Today mediums, fortune-tellers, and Tarot card readers still find many customers.



Those with a more scientific outlook may use a linear model as their digital crystal ball. Linear models are based on the x-values of the data at hand and cannot be trusted beyond that span. Some physical phenomena do exhibit a kind of "inertia" that allows us to guess that current systematic behavior will continue, but regularity can't be counted on in phenomena such as stock prices, sales figures, hurricane tracks, or public opinion.

Extrapolating from current trends is so tempting that even professional forecasters make this mistake, and sometimes the errors are striking. In the mid-1970s, oil prices surged and long lines at gas stations were common. In 1970, oil cost about $17 a barrel (in 2005 dollars)—about what it had cost for 20 years or so. But then, within just a few years, the price surged to over $40. In 1975, a survey of 15 top econometric forecasting models (built by groups that included Nobel prize–winning economists) found predictions for 1985 oil prices that ranged from $300 to over $700 a barrel (in 2005 dollars). How close were these forecasts?

Here's a scatterplot of oil prices from 1972 to 1981 (in 2005 dollars).

When the Data Are Years. . .
. . . we usually don't enter them as four-digit numbers. Here we used 0 for 1970, 10 for 1980, and so on. Or we may simply enter two digits, using 82 for 1982, for instance. Rescaling years like this often makes calculations easier and equations simpler. We recommend you do it, too. But be careful: If 1982 is 82, then 2004 is 104 (not 4), right?

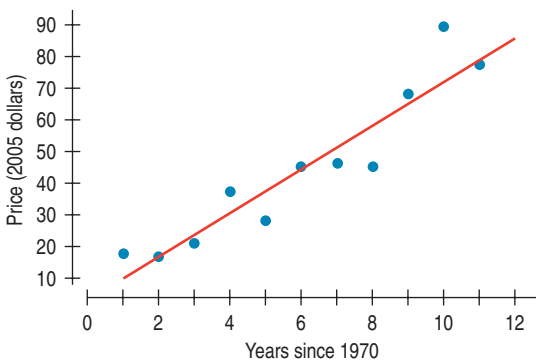


FIGURE 9.6
The scatterplot shows an average increase in the price of a barrel of oil of over $7 per year from 1971 to 1982.

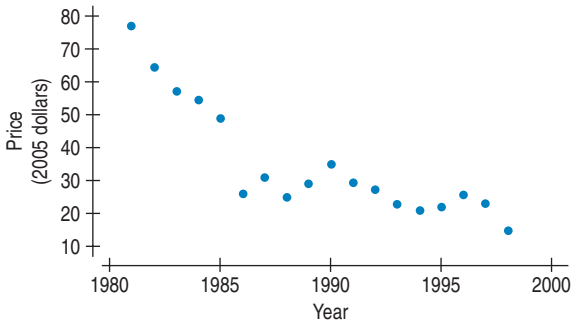


FIGURE 9.7

This scatterplot of oil prices from 1981 to 1998 shows a fairly constant decrease of about $3 per barrel per year.

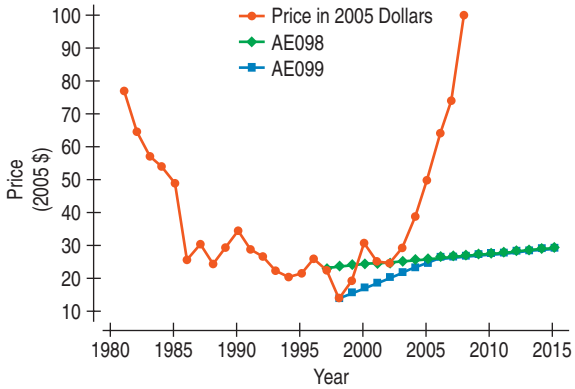


FIGURE 9.8

Here are the EIA forecasts with the actual prices from 1981 to 2008. Neither forecast predicted the sharp run-up in the past few years.

The regression model

$$\widehat{Price} = -0.85 + 7.39\ Years\ since\ 1970$$

says that prices had been going up 7.39 dollars per year, or nearly $74 in 10 years. If you assume that they would *keep going up*, it's not hard to imagine almost any price you want.

So, how did the forecasters do? Well, in the period from 1982 to 1998 oil prices didn't exactly continue that steady increase. In fact, they went down so much that by 1998, prices (adjusted for inflation) were the lowest they'd been since before World War II.

Not one of the experts' models predicted that.

Of course, these decreases clearly couldn't continue, or oil would be free by now. The Energy Information Administration offered two *different* 20-year forecasts for oil prices after 1998, and both called for relatively modest increases in oil prices. So, how accurate have *these* forecasts been? Here's a timeplot of the EIA's predictions and the actual prices (in 2005 dollars).

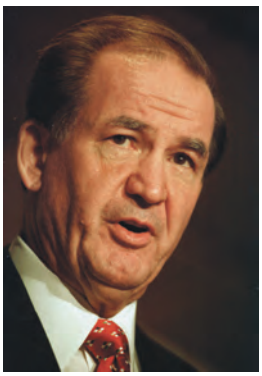
Oops! They seemed to have missed the sharp run-up in oil prices in the past few years.

Where do you think oil prices will go in the next decade? *Your* guess may be as good as anyone's!

Of course, knowing that extrapolation is dangerous doesn't stop people. The temptation to see into the future is hard to resist. So our more realistic advice is this:

> *If you must extrapolate into the future, at least don't believe that the prediction will come true.*

Outliers, Leverage, and Influence



The outcome of the 2000 U.S. presidential election was determined in Florida amid much controversy. The main race was between George W. Bush and Al Gore, but two minor candidates played a significant role. To the political right of the main party candidates was Pat Buchanan, while to the political left was Ralph Nader. Generally, Nader earned more votes than Buchanan throughout the state. We would expect counties with larger vote totals to give more votes to each candidate. Here's a regression relating *Buchanan's* vote totals by county in the state of Florida to *Nader's:*

Dependent variable is: Buchanan
R-squared = 42.8%

Variable	Coefficient
Intercept	50.3
Nader	0.14

The regression model,

$$\widehat{Buchanan} = 50.3 + 0.14\ Nader,$$

says that, in each county, Buchanan received about 0.14 times (or 14% of) the vote Nader received, starting from a base of 50.3 votes.

This seems like a reasonable regression, with an R^2 of almost 43%. But we've violated all three Rules of Data Analysis by going straight to the regression table without making a picture.

Here's a scatterplot that shows the vote for Buchanan in each county of Florida plotted against the vote for Nader. The striking **outlier** is Palm Beach County.

"Nature is nowhere accustomed more openly to display her secret mysteries than in cases where she shows traces of her workings apart from the beaten path."

—William Harvey (1657)

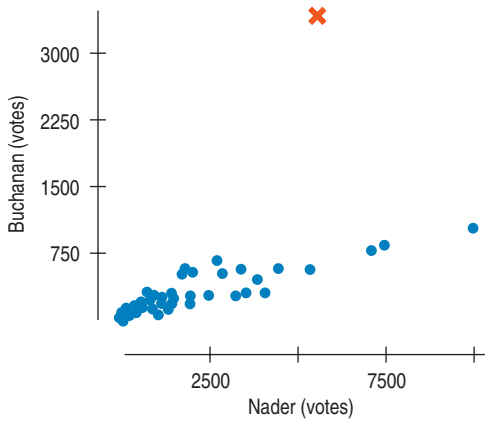


FIGURE 9.9

Votes received by Buchanan against votes for Nader in all Florida counties in the presidential election of 2000. The red "x" point is Palm Beach County, home of the "butterfly ballot."

The so-called "butterfly ballot," used only in Palm Beach County, was a source of controversy. It has been claimed that the format of this ballot confused voters so that some who intended to vote for the Democrat, Al Gore, punched the wrong hole next to his name and, as a result, voted for Buchanan.

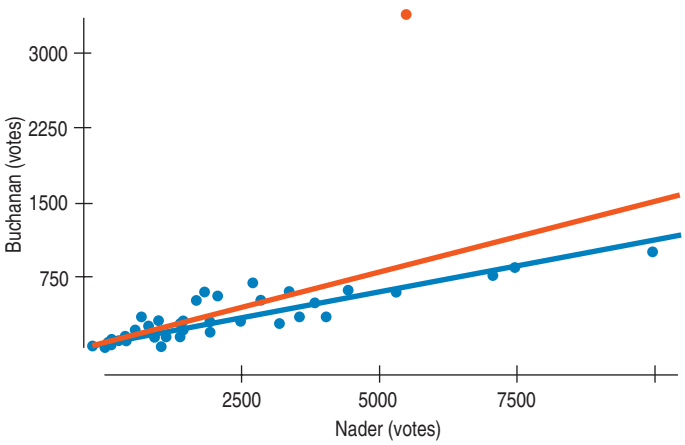
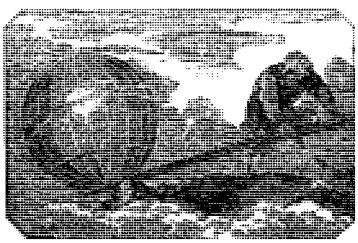


FIGURE 9.10

The red line shows the effect that one unusual point can have on a regression.

The scatterplot shows a strong, positive, linear association, and one striking point. With Palm Beach removed from the regression, the R^2 jumps from 42.8% to 82.1% and the slope of the line changes to 0.1, suggesting that Buchanan received only about 10% of the vote that Nader received. With more than 82% of the variability of the Buchanan vote accounted for, the model when Palm Beach is omitted certainly fits better. Palm Beach County now stands out, not as a Buchanan stronghold, but rather as a clear violation of the model that begs for explanation.

One of the great values of models is that, by establishing an idealized behavior, they help us to see when and how data values are unusual. In regression, a point can stand out in two different ways. First, a data value can have a large residual, as Palm Beach County does in this example. Because they seem to be different from the other cases, points whose residuals are large always deserve special attention.



"Give me a place to stand and I will move the Earth."

—Archimedes (287–211 BCE)

A data point can also be unusual if its x-value is far from the mean of the x-values. Such a point is said to have high **leverage**. The physical image of a lever is exactly right. We know the line must pass through $(\bar{x}, \bar{y})$, so you can picture that point as the fulcrum of the lever. Just as sitting farther from the hinge on a see-saw gives you more leverage to pull it your way, points with values far from $\bar{x}$ pull more strongly on the regression line.

A point with high leverage has the potential to change the regression line. But it doesn't always use that potential. If the point lines up with the pattern of the other points, then including it doesn't change our estimate of the line. By sitting so far from $\bar{x}$, though, it may strengthen the relationship, inflating the correlation and R^2. How can you tell if a high-leverage point actually changes the model? Just fit the linear model twice, both with and without the point in question. We say that a point is **influential** if omitting it from the analysis gives a very different model.[2]

Influence depends on both leverage and residual; a case with high leverage whose y-value sits right on the line fit to the rest of the data is not influential.

AS ***Activity:* Leverage.** You may be surprised to see how sensitive to a single influential point a regression line is.

[2] Some textbooks use the term *influential point* for any observation that influences the slope, intercept, or R^2. We'll reserve the term for points that influence the slope.

TI-*nspire*

Influential points. Try to make the regression line's slope change dramatically by dragging a point around in the scatterplot.

"For whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways."

—Francis Bacon (1561–1626)

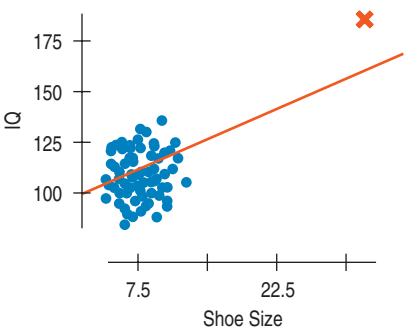


FIGURE 9.11

Bozo's extraordinarily large shoes give his data point high leverage in the regression. Wherever Bozo's IQ falls, the regression line will follow.

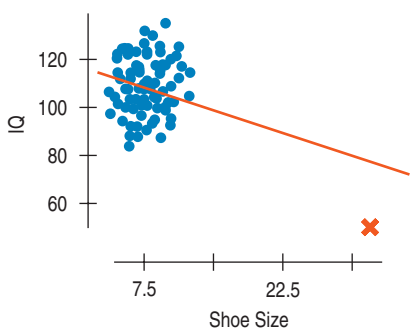


FIGURE 9.12

If Bozo's IQ were low, the regression slope would change from positive to negative. A single influential point can change a regression model drastically.

Removing that case won't change the slope, even if it does affect R^2. A case with modest leverage but a very large residual (such as Palm Beach County) can be influential. Of course, if a point has enough leverage, it can pull the line right to it. Then it's highly influential, but its residual is small. The only way to be sure is to fit both regressions.

Unusual points in a regression often tell us more about the data and the model than any other points. We face a challenge: The best way to identify unusual points is against the background of a model, but good models are free of the influence of unusual points. (That insight's at least 400 years old. See the sidebar.) Don't give in to the temptation to simply delete points that don't fit the line. You can take points out and discuss what the model looks like with and without them, but arbitrarily deleting points can give a false sense of how well the model fits the data. Your goal should be understanding the data, not making R^2 as big as you can.

In 2000, George W. Bush won Florida (and thus the presidency) by only a few hundred votes, so Palm Beach County's residual is big enough to be meaningful. It's the rare unusual point that determines a presidency, but all are worth examining and trying to understand.

A point with so much influence that it pulls the regression line close to it can make its residual deceptively small. Influential points like that can have a shocking effect on the regression. Here's a plot of *IQ* against *Shoe Size*, again from the fanciful study of intelligence and foot size in comedians we saw in Chapter 7. The linear regression output shows

Dependent variable is: IQ
R-squared = 24.8%

Variable	Coefficient
Intercept	93.3265
Shoe size	2.08318

Although this is a silly example, it illustrates an important and common potential problem: Almost all of the variance accounted for ($R^2 = 24.8\%$) is due to *one* point, namely, Bozo. Without Bozo, there is little correlation between *Shoe Size* and *IQ*. Look what happens to the regression when we take him out:

Dependent variable is: IQ
R-squared = 0.7%

Variable	Coefficient
Intercept	105.458
Shoe size	−0.460194

The R^2 value is now 0.7%—a very weak linear relationship (as one might expect!). One single point exhibits a great influence on the regression analysis.

What would have happened if Bozo hadn't shown his comic genius on IQ tests? Suppose his measured *IQ* had been only 50. The slope of the line would then drop from 0.96 IQ points/shoe size to −0.69 IQ points/shoe size. No matter where Bozo's *IQ* is, the line tends to follow it because his *Shoe Size*, being so far from the mean *Shoe Size*, makes this a high-leverage point.

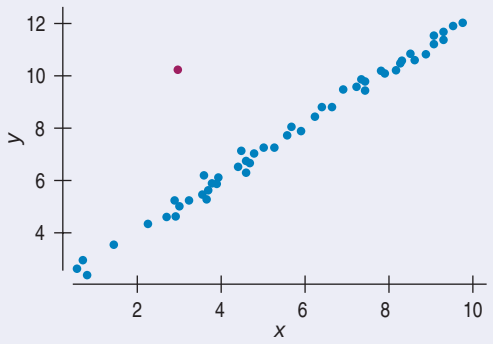
Even though this example is far fetched, similar situations occur all the time in real life. For example, a regression of sales against floor space for hardware stores that looked primarily at small-town businesses could be dominated in a similar way if The Home Depot were included.

Warning: Influential points can hide in plots of residuals. Points with high leverage pull the line close to them, so they often have small residuals. You'll see influential points more easily in scatterplots of the original data or by finding a regression model with and without the points.

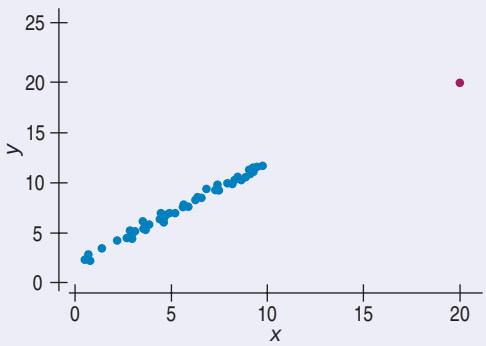
JUST CHECKING

Each of these scatterplots shows an unusual point. For each, tell whether the point is a high-leverage point, would have a large residual, or is influential.

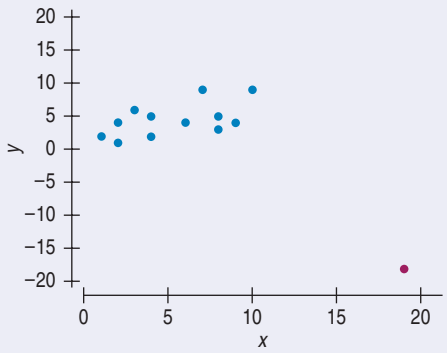
1.



2.



3.



Lurking Variables and Causation

One common way to interpret a regression slope is to say that "a change of 1 unit in x results in a change of b_1 units in y." This way of saying things encourages causal thinking. Beware.

In Chapter 7, we tried to make it clear that no matter how strong the correlation is between two variables, there's no simple way to show that one variable causes the other. Putting a regression line through a cloud of points just increases the temptation to think and to say that the x-variable *causes* the y-variable. Just to make sure, let's repeat the point again: No matter how strong the association, no matter how large the R^2 value, no matter how straight the line, there is no way to conclude from a regression alone that one variable *causes* the other. There's always the possibility that some third variable is driving both of the variables you have observed. With observational data, as opposed to data from a designed experiment, there is no way to be sure that a **lurking variable** is not the cause of any apparent association.

Here's an example: The scatterplot shows the *Life Expectancy* (average of men and women, in years) for each of 41 countries of the world, plotted against the square root of the number of *Doctors* per person in the country. (The square root is here to make the relationship satisfy the Straight Enough Condition, as we saw back in Chapter 7.)

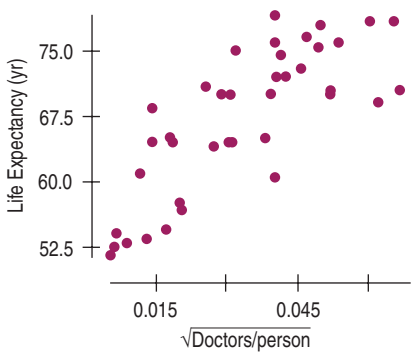


FIGURE 9.13

The relationship between Life Expectancy *(years) and availability of* Doctors *(measured as* $\sqrt{\text{doctors per person}}$*) for countries of the world is strong, positive, and linear.*

The strong positive association ($R^2 = 62.4\%$) seems to confirm our expectation that more *Doctors* per person improves healthcare, leading to longer lifetimes and a greater *Life Expectancy*. The strength of the association would *seem* to argue that we should send more doctors to developing countries to increase life expectancy.

That conclusion is about the consequences of a change. Would sending more doctors increase life expectancy? Specifically, do doctors *cause* greater life expectancy? Perhaps, but these are observed data, so there may be another explanation for the association.

On the next page, the similar-looking scatterplot's x-variable is the square root of the number of *Televisions* per person in each country. The positive association in this scatterplot is even *stronger* than the association in the previous plot

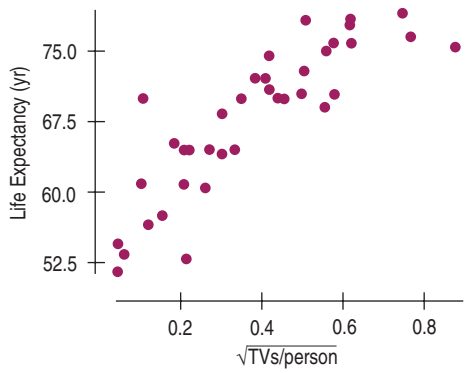


FIGURE 9.14

To increase life expectancy, don't send doctors, send TVs; they're cheaper and more fun. Or maybe that's not the right interpretation of this scatterplot of life expectancy *against availability of TVs (as* $\sqrt{TVs\ per\ person}$*).*

($R^2 = 72.3\%$). We can fit the linear model, and quite possibly use the number of TVs as a way to predict life expectancy. Should we conclude that increasing the number of TVs actually extends lifetimes? If so, we should send TVs instead of doctors to developing countries. Not only is the correlation with life expectancy higher, but TVs are much cheaper than doctors.

What's wrong with this reasoning? Maybe we were a bit hasty earlier when we concluded that doctors *cause* longer lives. Maybe there's a lurking variable here. Countries with higher standards of living have both longer life expectancies *and* more doctors (and more TVs). Could higher living standards cause changes in the other variables? If so, then improving living standards might be expected to prolong lives, increase the number of doctors, and increase the number of TVs.

From this example, you can see how easy it is to fall into the trap of mistakenly inferring causality from a regression. For all we know, doctors (or TVs!) *do* increase life expectancy. But we can't tell that from data like these, no matter how much we'd like to. Resist the temptation to conclude that x causes y from a regression, no matter how obvious that conclusion seems to you.

Working with Summary Values

Scatterplots of statistics summarized over groups tend to show less variability than we would see if we measured the same variable on individuals. This is because the summary statistics themselves vary less than the data on the individuals do—a fact we will make more specific in coming chapters.

In Chapter 7 we looked at the heights and weights of individual students. There we saw a correlation of 0.644, so R^2 is 41.5%.

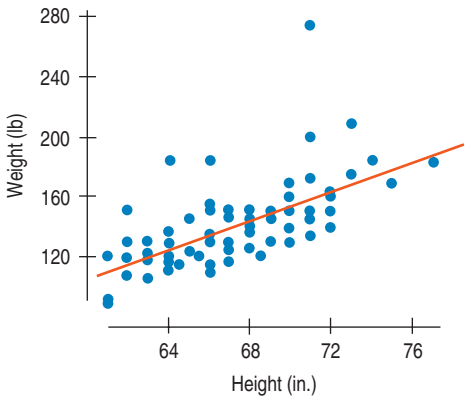


FIGURE 9.15

Weight *(lb) against* Height *(in.) for a sample of men. There's a strong, positive, linear association.*

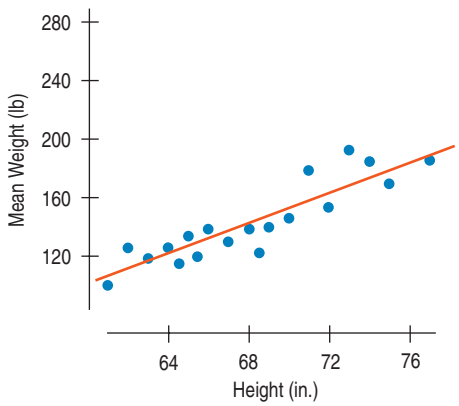


FIGURE 9.16

Mean Weight *(lb) shows a stronger linear association with* Height *than do the weights of individuals. Means vary less than individual values.*

Suppose, instead of data on individuals, we knew only the mean weight for each height value. The scatterplot of mean weight by height would show less scatter. And the R^2 would increase to 80.1%.

Scatterplots of summary statistics show less scatter than the baseline data on individuals and can give a false impression of how well a line summarizes the data. There's no simple correction for this phenomenon. Once we're given summary data, there's no simple way to get the original values back.

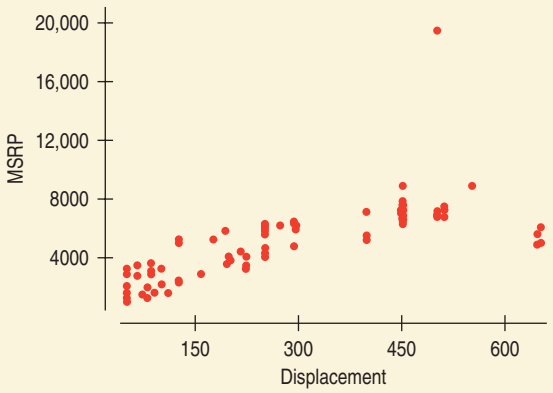
In the life expectancy and TVs example, we have no good measure of exposure to doctors or to TV on an individual basis. But if we did, we should expect the scatterplot to show more variability and the corresponding R^2 to be smaller. The bottom line is that you should be a bit suspicious of conclusions based on regressions of summary data. They may look better than they really are.

FOR EXAMPLE Using several of these methods together

Motorcycles designed to run off-road, often known as dirt bikes, are specialized vehicles.

We have data on 104 dirt bikes available for sale in 2005. Some cost as little as $3000, while others are substantially more expensive. Let's investigate how the size and type of engine contribute to the cost of a dirt bike. As always, we start with a scatterplot.

Here's a scatterplot of the manufacturer's suggested retail price (*MSRP*) in dollars against the engine *Displacement*, along with a regression analysis:



Dependent variable is: MSRP
R-squared = 49.9% s = 1737

Variable	Coefficient
Intercept	2273.67
Displacement	10.0297

Question: What do you see in the scatterplot?

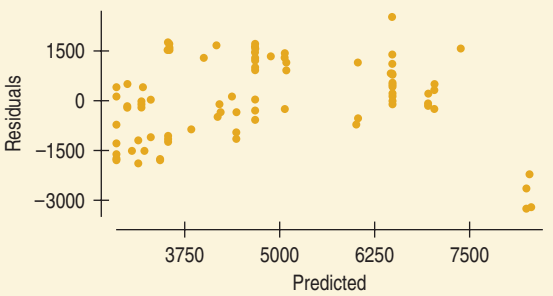
There is a strong positive association between the engine displacement of dirt bikes and the manufacturer's suggested retail price. One of the dirt bikes is an outlier; its price is more than double that of any other bike.

The outlier is the Husqvarna TE 510 Centennial. Most of its components are handmade exclusively for this model, including extensive use of carbon fiber throughout. That may explain its $19,500 price tag! Clearly, the TE 510 is not like the other bikes. We'll set it aside for now and look at the data for the remaining dirt bikes.

Question: What effect will removing this outlier have on the regression? Describe how the slope, R^2, and s_e will change.

The TE 510 was an influential point, tilting the regression line upward. With that point removed, the regression slope will get smaller. With that dirt bike omitted, the pattern becomes more consistent, so the value of R^2 should get larger and the standard deviation of the residuals, s_e, should get smaller.

With the outlier omitted, here's the new regression and a scatterplot of the residuals:



Dependent variable is: MSRP
R-squared = 61.3% s = 1237

Variable	Coefficient
Intercept	2411.02
Displacement	9.05450

Question: What do you see in the residuals plot?

The points at the far right don't fit well with the other dirt bikes. Overall, there appears to be a bend in the relationship, so a linear model may not be appropriate.

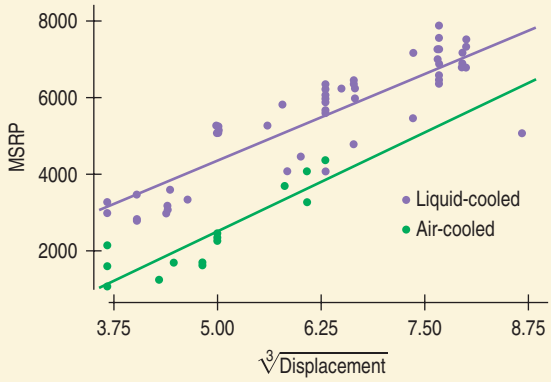
Let's try a re-expression. Here's a scatterplot showing *MSRP* against the cube root of *Displacement* to make the relationship closer to straight. (Since displacement is measured in cubic centimeters, its cube root has the simple units of centimeters.) In addition, we've colored the plot according to the cooling

method used in the bike's engine: liquid or air. Each group is shown with its own regression line, as we did for the cereals on different shelves.

Question: What does this plot say about dirt bikes?

There appears to be a positive, linear relationship between MSRP and the cube root of Displacement. In general, the larger the engine a bike has, the higher the suggested price. Liquid-cooled dirt bikes, however, typically cost more than air-cooled bikes with comparable displacement. A few liquid-cooled bikes appear to be much less expensive than we might expect, given their engine displacements.

[Jiang Lu, Joseph B. Kadane, and Peter Boatwright, "The Dirt on Bikes: An Illustration of CART Models for Brand Differentiation," provides data on 2005-model bikes.]



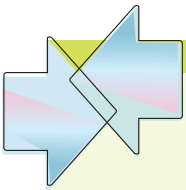
WHAT CAN GO WRONG?

This entire chapter has held warnings about things that can go wrong in a regression analysis. So let's just recap. When you make a linear model:

- **Make sure the relationship is straight.** Check the Straight Enough Condition. Always examine the residuals for evidence that the Linearity Assumption has failed. It's often easier to see deviations from a straight line in the residuals plot than in the scatterplot of the original data. Pay special attention to the most extreme residuals because they may have something to add to the story told by the linear model.
- **Be on guard for different groups in your regression.** Check for evidence that the data consist of separate subsets. If you find subsets that behave differently, consider fitting a different linear model to each subset.
- **Beware of extrapolating.** Beware of extrapolation beyond the x-values that were used to fit the model. Although it's common to use linear models to extrapolate, the practice is dangerous.
- **Beware especially of extrapolating into the future!** Be especially cautious about extrapolating into the future with linear models. To predict the future, you must assume that future changes will continue at the same rate you've observed in the past. Predicting the future is particularly tempting and particularly dangerous.
- **Look for unusual points.** Unusual points always deserve attention and may well reveal more about your data than the rest of the points combined. Always look for them and try to understand why they stand apart. A scatterplot of the data is a good way to see high-leverage and influential points. A scatterplot of the residuals against the predicted values is a good tool for finding points with large residuals.
- **Beware of high-leverage points and especially of those that are influential.** Influential points can alter the regression model a great deal. The resulting model may say more about one or two points than about the overall relationship.
- **Consider comparing two regressions.** To see the impact of outliers on a regression, it's often wise to run two regressions, one with and one without the extraordinary points, and then to discuss the differences.
- **Treat unusual points honestly.** If you remove enough carefully selected points, you can always get a regression with a high R^2 eventually. But it won't give you much understanding. Some variables are not related in a way that's simple enough for a linear model to fit very well. When that happens, report the failure and stop.

(*continued*)

- **Beware of lurking variables.** Think about lurking variables before interpreting a linear model. It's particularly tempting to explain a strong regression by thinking that the x-variable *causes* the y-variable. A linear model alone can never demonstrate such causation, in part because it cannot eliminate the chance that a lurking variable has caused the variation in both x and y.
- **Watch out when dealing with data that are summaries.** Be cautious in working with data values that are themselves summaries, such as means or medians. Such statistics are less variable than the data on which they are based, so they tend to inflate the impression of the strength of a relationship.



CONNECTIONS

We are always alert to things that can go wrong if we use statistics without thinking carefully. Regression opens new vistas of potential problems. But each one relates to issues we've thought about before.

It is always important that our data be from a single homogeneous group and not made up of disparate groups. We looked for multiple modes in single variables. Now we check scatterplots for evidence of subgroups in our data. As with modes, it's often best to split the data and analyze the groups separately.

Our concern with unusual points and their potential influence also harks back to our earlier concern with outliers in histograms and boxplots—and for many of the same reasons. As we've seen here, regression offers such points new scope for mischief.

The risks of interpreting linear models as causal or predictive arose in Chapters 7 and 8. And they're important enough to mention again in later chapters.

WHAT HAVE WE LEARNED?

We've learned that there are many ways in which a data set may be unsuitable for a regression analysis.

- Watch out for more than one group hiding in your regression analysis. If you find subsets of the data that behave differently, consider fitting a different regression model to each subset.
- The **Straight Enough Condition** says that the relationship should be reasonably straight to fit a regression. Somewhat paradoxically, sometimes it's easier to see that the relationship is not straight *after* fitting the regression by examining the residuals. The same is true of outliers.
- The **Outlier Condition** actually means two things: Points with large residuals or high leverage (especially both) can influence the regression model significantly. It's a good idea to perform the regression analysis with and without such points to see their impact.

And we've learned that even a good regression doesn't mean we should believe that the model says more than it really does.

- Extrapolation far from $\bar{x}$ can lead to silly and useless predictions.
- Even an R^2 near 100% doesn't indicate that x causes y (or the other way around). Watch out for lurking variables that may affect both x and y.
- Be careful when you interpret regressions based on *summaries* of the data sets. These regressions tend to look stronger than the regression based on all the individual data.

Terms

Extrapolation 203. Although linear models provide an easy way to predict values of y for a given value of x, it is unsafe to predict for values of x far from the ones used to find the linear model equation. Such extrapolation may pretend to see into the future, but the predictions should not be trusted.

Outlier 205. Any data point that stands away from the others can be called an outlier. In regression, outliers can be extraordinary in two ways: by having a large residual or by having high leverage.

Leverage 206. Data points whose x-values are far from the mean of x are said to exert leverage on a linear model. High-leverage points pull the line close to them, and so they can have a large effect on the line, sometimes completely determining the slope and intercept. With high enough leverage, their residuals can be deceptively small.

Influential point 206. If omitting a point from the data results in a very different regression model, then that point is called an influential point.

Lurking variable 208. A variable that is not explicitly part of a model but affects the way the variables in the model appear to be related is called a lurking variable. Because we can never be certain that observational data are not hiding a lurking variable that influences both x and y, it is never safe to conclude that a linear model demonstrates a causal relationship, no matter how strong the linear association.

Skills

THINK

- Understand that we cannot fit linear models or use linear regression if the underlying relationship between the variables is not itself linear.
- Understand that data used to find a model must be homogeneous. Look for subgroups in data before you find a regression, and analyze each separately.
- Know the danger of extrapolating beyond the range of the x-values used to find the linear model, especially when the extrapolation tries to predict into the future.
- Understand that points can be unusual by having a large residual or by having high leverage.
- Understand that an influential point can change the slope and intercept of the regression line.
- Look for lurking variables whenever you consider the association between two variables. Understand that a strong association does not mean that the variables are causally related.
- Know how to display residuals from a linear model by making a scatterplot of residuals against predicted values or against the x-variable, and know what patterns to look for in the picture.

SHOW

- Know how to look for high-leverage and influential points by examining a scatterplot of the data and how to look for points with large residuals by examining a scatterplot of the residuals against the predicted values or against the x-variable. Understand how fitting a regression line with and without influential points can add to your understanding of the regression model.
- Know how to look for high-leverage points by examining the distribution of the x-values or by recognizing them in a scatterplot of the data, and understand how they can affect a linear model.

TELL

- Include diagnostic information such as plots of residuals and leverages as part of your report of a regression.
- Report any high-leverage points.
- Report any outliers. Consider reporting analyses with and without outliers, to assess their influence on the regression.
- Include appropriate cautions about extrapolation when reporting predictions from a linear model.
- Discuss possible lurking variables.

REGRESSION DIAGNOSIS ON THE COMPUTER

Most statistics technology offers simple ways to check whether your data satisfy the conditions for regression. We have already seen that these programs can make a simple scatterplot. They can also help us check the conditions by plotting residuals.