

chapter 9 Re-expressing Data: Get It Straight!



A S **Activity: Re-expressing Data.**
Should you re-express data? Actually, you already do.

How fast can you go on a bicycle? If you measure your speed, you probably do it in miles per hour or kilometers per hour. In a 12-mile-long time trial in the 2005 Tour de France, Dave Zabriskie *averaged* nearly 35 mph (54.7 kph), beating Lance Armstrong by 2 seconds. You probably realize that's a tough act to follow. It's fast. You can tell that at a glance because you have no trouble thinking in terms of distance covered per time.

OK, then, if you averaged 12.5 mph (20.1 kph) for a mile *run*, would *that* be fast? Would it be fast for a 100-m dash? Even if you run the mile often, you probably have to stop and calculate. Running a mile in under 5 minutes (12 mph) is fast. A mile at 16 mph would be a world record (that's a 3-minute, 45-second mile). There's no single *natural* way to measure speed. Sometimes we use time over distance; other times we use the *reciprocal*, distance over time. Neither one is *correct*. We're just used to thinking that way in each case.

So, how does this insight help us understand data? All quantitative data come to us measured in some way, with units specified. But maybe those units aren't the best choice. It's not whether meters are better (or worse) than fathoms or leagues. **What we're talking about is a different type of re-expression: applying a function, such as a square root, log, or reciprocal to the data.** You already use some of them, even though you may not know it. For example, the Richter scale of earthquake strength (logs), the decibel scale for sound intensity (logs), the f/stop scale for camera aperture openings (squares), and the gauges of shotguns (square roots) all include simple functions of this sort.

Why bother? As with speeds, some expressions of the data may be easier to think about. And some may be much easier to analyze with statistical methods. We've seen that symmetric distributions are easier to summarize and straight scatterplots are easier to model with regressions. We often look to re-express our data if doing so makes them more suitable for our methods.

Beyond $b_0 + b_1x$

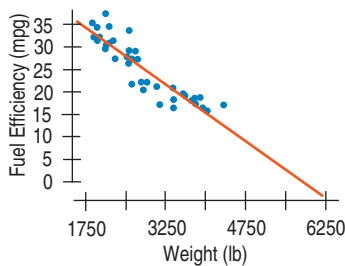
Scan through any Physics book. Most equations have powers, reciprocals, or logs.

Straight to the Point

We know from common sense and from physics that heavier cars need more fuel, but exactly how does a car's weight affect its fuel efficiency? Here are the scatterplot

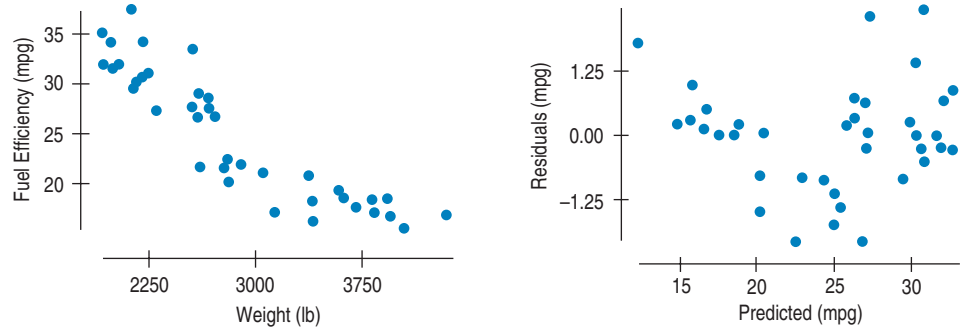
Figure 9.1

Fuel Efficiency (mpg) vs. Weight for 38 cars as reported by Consumer Reports. The scatterplot shows a negative direction, roughly linear shape, and strong relationship. However, the residuals from a regression of *Fuel Efficiency* on *Weight* reveal a bent shape when plotted against the predicted values. Looking back at the original scatterplot, you may be able to see the bend.

**Figure 9.2**

Extrapolating the regression line gives an absurd answer for vehicles that weigh as little as 6000 pounds.

of *Weight* (in pounds) and *Fuel Efficiency* (in miles per gallon) for 38 cars, and the residuals plot:



Hmm . . . Even though R^2 is 81.6%, the residuals don't show the random scatter we were hoping for. The shape is clearly bent. Looking back at the first scatterplot, you can probably see the slight bending. Think about the regression line through the points. How heavy would a car have to be to have a predicted gas mileage of 0? It looks like the *Fuel Efficiency* would go negative at about 6000 pounds. A Hummer H2 weighs about 6400 pounds. The H2 is hardly known for fuel efficiency, but it does get more than the *minus 5 mpg* this regression predicts. Extrapolation is always dangerous, but it's more dangerous the more the model is wrong, because wrong models tend to do even worse the farther you get from the middle of the data.

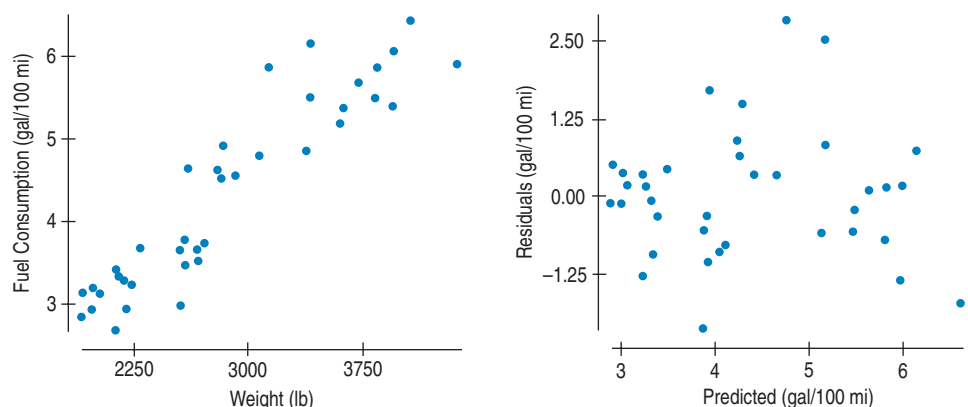
The bend in the relationship between *Fuel Efficiency* and *Weight* is the kind of failure to satisfy the conditions for an analysis that we can repair by re-expressing the data. Instead of looking at miles per gallon, we could take the reciprocal and work with gallons per hundred miles.¹

“Gallons Per Hundred Miles—What an Absurd Way to Measure Fuel Efficiency! Who Would Ever Do It That Way?”

Not all re-expressions are easy to understand, but in this case the answer is “Everyone except U.S. drivers.” Most of the world measures fuel efficiency in liters per 100 kilometers (L/100 km). This is the same reciprocal form (fuel amount per distance driven) and differs from gallons per 100 miles only by a constant multiple of about 2.38. It has been suggested that most of the world says, “I’ve got to go 100 km; how much gas do I need?” But Americans say, “I’ve got 10 gallons in the tank. How far can I drive?” In much the same way, re-expressions “think” about the data differently but don’t change what they mean.

Figure 9.3

The reciprocal ($1/y$) is measured in gallons per mile. Gallons per 100 miles gives more meaningful numbers. The reciprocal is more nearly linear against *Weight* than the original variable, but the re-expression changes the direction of the relationship. The residuals from the regression of *Fuel Consumption* (gal/100 mi) on *Weight* show less of a pattern than before.



¹Multiplying by 100 to get gallons per 100 miles simply makes the numbers easier to think about: You might have a good idea of how many gallons your car needs to drive 100 miles, but probably a much poorer sense of how much gas you need to go just 1 mile.



The direction of the association is positive now, since we’re measuring gas consumption and heavier cars consume more gas per mile. The relationship is much straighter, as we can see from a scatterplot of the regression residuals.

This is more the kind of boring residuals plot (no direction, no particular shape, no outliers, no bends) that we hope to see, so we have reason to think that the Straight Enough Condition is now satisfied. And here’s the payoff: What does the reciprocal model say about the Hummer? The regression line fit to *Fuel Consumption* vs. *Weight* predicts somewhere near 9.7 gallons for a car weighing 6400 pounds. What does this mean? It means the car is predicted to use 9.7 gallons for every 100 miles, or in other words,

$$\frac{100 \text{ miles}}{9.7 \text{ gallons}} = 10.3 \text{ mpg.}$$

That’s a much more reasonable prediction and very close to the reported value of 11.0 miles per gallon (of course, *your* mileage may vary . . .).

Goals of Re-expression

We re-express data for several reasons. Each of these goals helps make the data more suitable for analysis by our methods.

Goal 1

Make the distribution of a variable (as seen in its histogram, for example) more symmetric. It’s easier to summarize the center of a symmetric distribution, and for nearly symmetric distributions, we can use the mean and standard deviation. If the distribution is unimodal, then the resulting distribution may be closer to the Normal model, allowing us to use the 68–95–99.7 Rule.

Here are a histogram, quite skewed, showing the *Assets* of 77 companies selected from the Forbes 500 list (in \$100,000) and the more symmetric histogram after taking logs.

| | |
|-------|--|
| Who | 77 large companies |
| What | Assets, sales, and market sector |
| Units | \$100,000 |
| How | Public records |
| When | 1986 |
| Why | By <i>Forbes</i> magazine in reporting on the Forbes 500 for that year |

AS **Simulation: Re-expression in Action.** Slide the re-expression power and watch the histogram change.

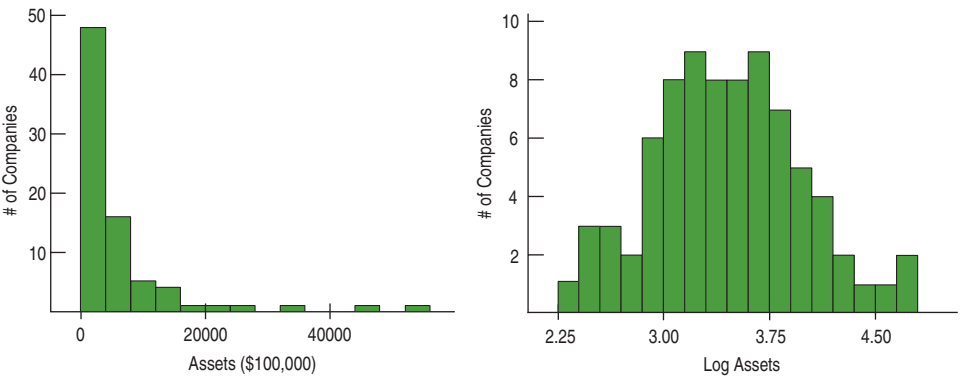


Figure 9.4 The distribution of the *Assets* of large companies is skewed to the right. Data on wealth often look like this. Taking logs makes the distribution more nearly symmetric.

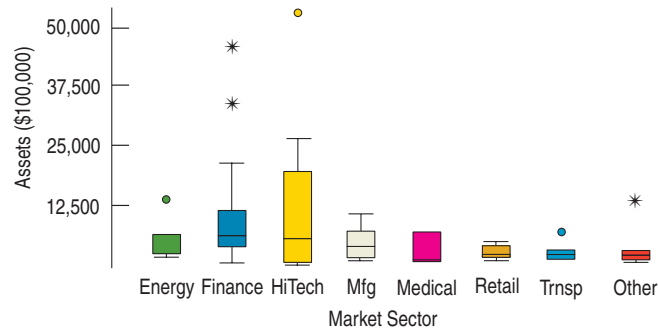
Goal 2

Make the spread of several groups (as seen in side-by-side boxplots) more alike, even if their centers differ. Groups that share a common spread are easier to compare. We’ll see methods later in the book that can be applied only to groups with a common standard deviation. We saw an example of re-expression for comparing groups with boxplots in Chapter 4.

Here are the *Assets* of these companies by *Market Sector*:

Figure 9.5

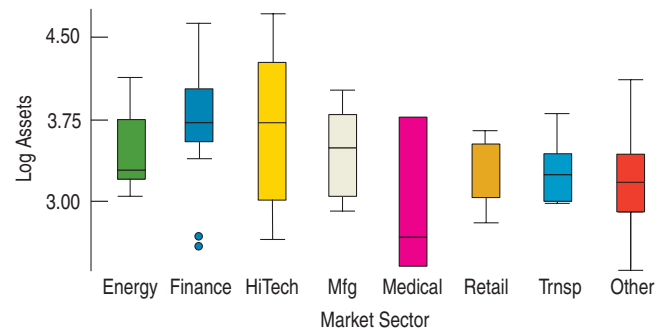
Assets of large companies by Market Sector. It's hard to compare centers or spreads, and there seem to be a number of high outliers.



Taking logs makes the individual boxplots more symmetric and gives them spreads that are more nearly equal.

Figure 9.6

After re-expressing by logs, it's much easier to compare across market sectors. The boxplots are more nearly symmetric, most have similar spreads, and the companies that seemed to be outliers before are no longer extraordinary. Two new outliers have appeared in the finance sector. They are the only companies in that sector that are not banks. Perhaps they don't belong there.



Doing this makes it easier to compare assets across market sectors. It can also reveal problems in the data. Some companies that looked like outliers on the high end turned out to be more typical. But two companies in the finance sector now stick out. Unlike the rest of the companies in that sector, they are not banks. They may have been placed in the wrong sector, but we couldn't see that in the original data.

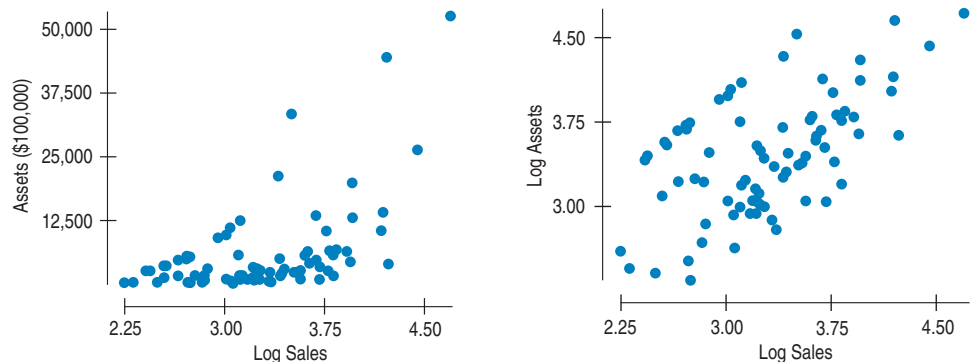
Goal 3

Make the form of a scatterplot more nearly linear. Linear scatterplots are easier to model. We saw an example of scatterplot straightening in Chapter 6. The greater value of re-expression to straighten a relationship is that we can fit a linear model once the relationship is straight.

Here are *Assets* of the companies plotted against the logarithm of *Sales*, clearly bent. Taking logs makes things much more linear.

Figure 9.7

Assets vs. *Log Sales* shows a positive association (bigger sales go with bigger assets) but a bent shape. Note also that the points go from tightly bunched at the left to widely scattered at the right; the plot "thickens." In the second plot, *Log Assets* vs. *Log Sales* shows a clean, positive, linear association. And the variability at each value of *x* is about the same.



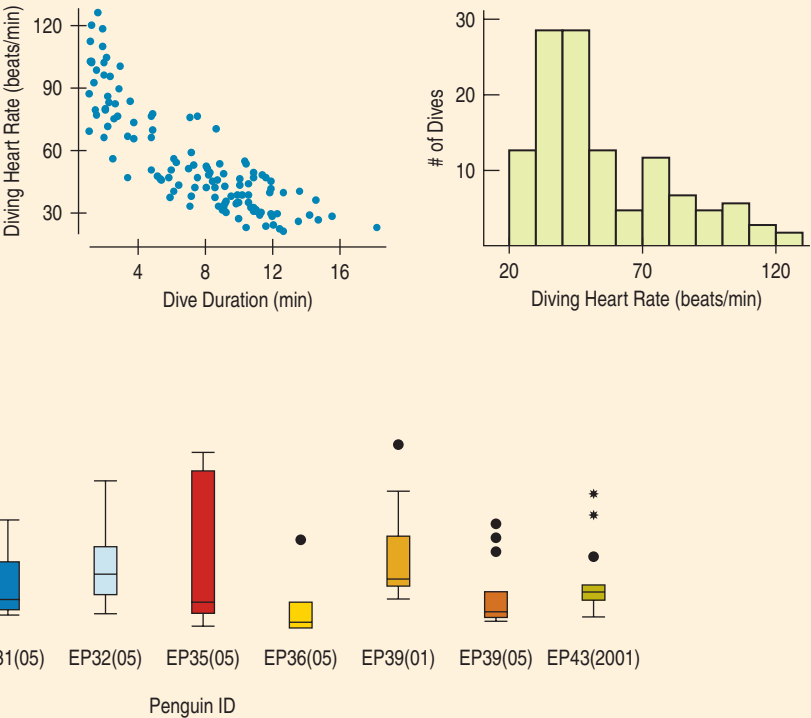
Goal 4

Make the scatter in a scatterplot spread out evenly rather than thickening at one end. Having consistent scatter is a condition of many methods of Statistics, as we'll see in later chapters. This goal is closely related to Goal 2, but it often comes along with Goal 3. Indeed, a glance back at the scatterplot (Figure 9.7) shows that the plot for *Assets* is much more spread out on the right than on the left, while the plot for *log Assets* has roughly the same variation in *log Assets* for any *x*-value.

For Example RECOGNIZING WHEN A RE-EXPRESSION CAN HELP

In Chapter 8, we saw the awesome ability of emperor penguins to slow their heart rates while diving. Here are three displays relating to the diving heart rates:
(The boxplots show the diving heart rates for each of the 9 penguins whose dives were tracked. The names are those given by the researchers; EP = emperor penguin.)

QUESTION: What features of each of these displays suggest that a re-expression might be helpful?



ANSWER: The scatterplot shows a curved relationship, concave upward, between the duration of the dives and penguins' heart rates. Re-expressing either variable may help to straighten the pattern.

The histogram of heart rates is skewed to the high end. Re-expression often helps to make skewed distributions more nearly symmetric.

The boxplots each show skewness to the high end as well. The medians are low in the boxes, and several show high outliers.

The Ladder of Powers

Activity: Re-expression in Action. Here's the animated version of the Ladder of Powers. Slide the power and watch the change.

How can we pick a re-expression to use? The secret is to choose a re-expression from a simple family of functions that includes powers and the logarithm.² We raise each data value to the same power: $\frac{1}{2}$, for example, by taking square roots. Or -1 , by finding reciprocals. The good news is that the family of re-expressions line up in order, so that the farther you move away from the original data (the "1" position), the greater the effect on any curvature. This fact lets you search systematically for a re-expression that works, stepping a bit farther from "1" or taking a step back toward "1" as you see the results.

²Don't be scared. You may have learned lots of properties of logarithms or done some messy calculations. Relax! You won't need that stuff here.

Where to start? It turns out that certain kinds of data are more likely to be helped by particular re-expressions. Knowing that gives you a good place to start your search, and from there you can look around a bit for a useful re-expression. We call this collection of re-expressions the **Ladder of Powers**.

| Power | Name | Comment |
|-------|--|---|
| 2 | The square of the data values, y^2 . | Try this for unimodal distributions that are skewed to the left. |
| 1 | The raw data—no change at all. This is “home base.” The farther you step from here up or down the ladder, the greater the effect. | Data that can take on both positive and negative values with no bounds are less likely to benefit from re-expression. |
| 1/2 | The square root of the data values, \sqrt{y} . | Counts often benefit from a square root re-expression. For counted data, start here. |
| “0” | Although mathematicians define the “0-th” power differently, ³ for us the place is held by the logarithm. You may feel uneasy about logarithms. Don’t worry; the computer or calculator does the work. ⁴ | Measurements that cannot be negative, and especially values that grow by percentage increases such as salaries or populations, often benefit from a log re-expression. When in doubt, start here. If your data have zeros, try adding a small constant to all values before finding the logs. |
| -1/2 | The (negative) reciprocal square root, $-1/\sqrt{y}$. | An uncommon re-expression, but sometimes useful. Changing the sign to take the <i>negative</i> of the reciprocal square root preserves the direction of relationships, making things a bit simpler. |
| -1 | The (negative) reciprocal, $-1/y$. | Ratios of two quantities (miles per hour, for example) often benefit from a reciprocal. (You have about a 50–50 chance that the original ratio was taken in the “wrong” order for simple statistical analysis and would benefit from re-expression.) Often, the reciprocal will have simple units (hours per mile). Change the sign if you want to preserve the direction of relationships. If your data have zeros, try adding a small constant to all values before finding the reciprocal. |

Ti-*inspire*

Re-expression. See a curved relationship become straighter with each step on the Ladder of Powers.

The Ladder of Powers orders the effects that the re-expressions have on data. If you try, say, taking the square roots of all the values in a variable and it helps, but not enough, then move farther down the ladder to the logarithm or reciprocal root. Those re-expressions will have a similar, but even stronger, effect on your data. If you go too far, you can always back up. But don’t forget—when you take a negative power, the *direction* of the relationship will change. That’s OK. You can always change the sign of the response variable if you want to keep the same direction. With modern technology, finding a suitable re-expression is no harder than the push of a button.



Just Checking

1. You want to model the relationship between the number of birds counted at a nesting site and the temperature (in degrees Celsius). The scatterplot of counts vs. temperature shows an upwardly curving pattern, with more birds spotted at higher temperatures. What transformation (if any) of the bird counts might you start with?
2. You want to model the relationship between prices for various items in Paris and in Hong Kong. The

scatterplot of Hong Kong prices vs. Parisian prices shows a generally straight pattern with a small amount of scatter. What transformation (if any) of the Hong Kong prices might you start with?

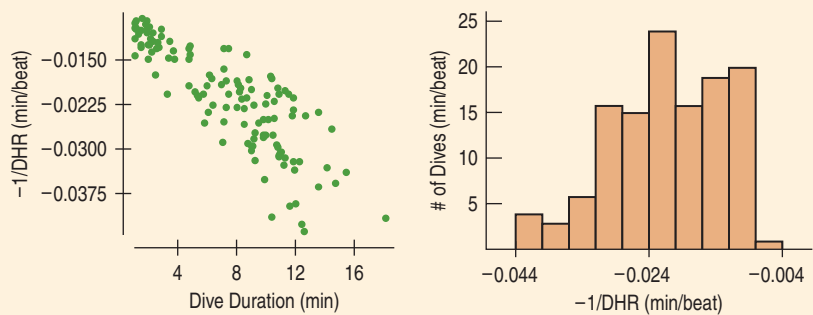
3. You want to model the population growth of the United States over the past 200 years. The scatterplot shows a strongly upwardly curved pattern. What transformation (if any) of the population might you start with?

³You may remember that for any nonzero number y , $y^0 = 1$. This is not a very exciting transformation for data; every data value would be the same. We use the logarithm in its place.

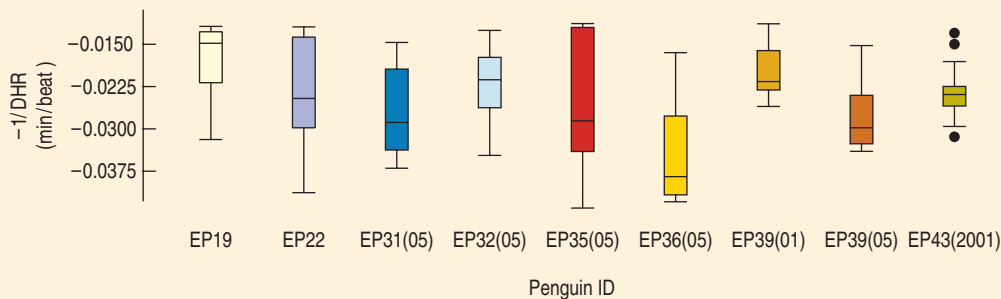
⁴Your calculator or software package probably gives you a choice between “base 10” logarithms and “natural (base e)” logarithms. Don’t worry about that. It doesn’t matter at all which you use; they have exactly the same effect on the data. If you want to choose, base 10 logarithms can be a bit easier to interpret.

For Example TRYING A RE-EXPRESSION

RECAP: We’ve seen curvature in the relationship between emperor penguins’ diving heart rates and the duration of the dive. Let’s start the process of finding a good re-expression. Heart rate is in beats per minute; maybe heart “speed” in minutes per beat would be a better choice. Here are the corresponding displays for this reciprocal re-expression (as we often do, we’ve changed the sign to preserve the order of the data values):



QUESTION: Were the re-expressions successful?



ANSWER: The scatterplot bends less than before, but now may be slightly concave downward. The histogram is now slightly skewed to the low end. Most of the boxplots have no outliers. These boxplots seem better than the ones for the raw heart rates.

Overall, it looks like I may have moved a bit “too far” on the ladder of powers. Halfway between “1” (the original data) and “−1” (the reciprocal) is “0,” which represents the logarithm. I’d try that for comparison.

Step-by-Step Example RE-EXPRESSION TO STRAIGHTEN A SCATTERPLOT



Standard (monofilament) fishing line comes in a range of strengths, usually expressed as “test pounds.” Five-pound test line, for example, can be expected to withstand a pull of up to five pounds without breaking. The convention in selling fishing line is that the price of a spool doesn’t vary with strength. Instead, the length of line on the spool varies. Higher test pound line is thicker, though, so spools of fishing line hold about the same amount of material. Some spools hold line that is thinner and longer, some fatter and shorter. Let’s look at the *Length* and *Strength* of spools of monofilament line manufactured by the same company and sold for the same price at one store.

Question: How are the *Length* on the spool and the *Strength* related? And what re-expression will straighten the relationship?

THINK ➡ Plan State the problem.

I want to fit a linear model for the length and strength of monofilament fishing line.

(continued)

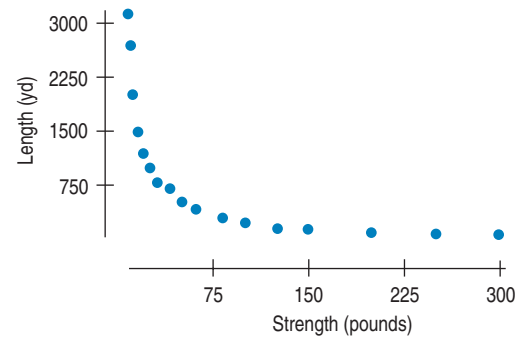
Variables Identify the variables and report the W's.

Plot Check that even if there is a curve, the overall pattern does not reach a minimum or maximum and then turn around and go back. An up-and-down curve can't be fixed by re-expression.

I have the *length* and "pound test" *strength* of monofilament fishing line sold by a single vendor at a particular store. Each case is a different strength of line, but all spools of line sell for the same price.

Let *Length* = length (in yards) of fishing line on the spool

Strength = the test strength (in pounds).



The plot shows a negative direction and an association that has little scatter but is not straight.

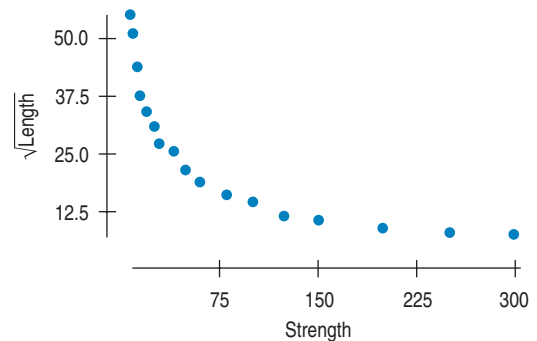
SHOW ➡ Mechanics Try a re-expression.

The lesson of the Ladder of Powers is that if we're moving in the right direction but have not had sufficient effect, we should go farther along the ladder. This example shows improvement, but is still not straight.

(Because *Length* is an amount of something and cannot be negative, we probably should have started with logs. This plot is here in part to illustrate how the Ladder of Powers works.)

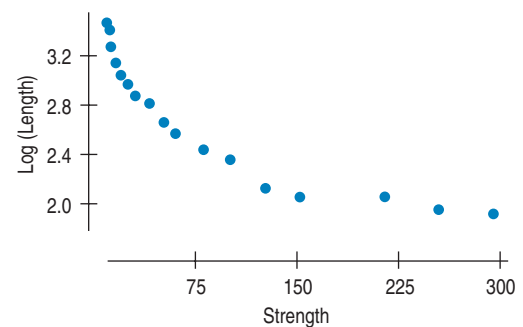
Stepping from the $1/2$ power to the "0" power, we try the logarithm of *Length* against *Strength*.

Here's a plot of the square root of *Length* against *Strength*:



The plot is less bent, but still not straight.

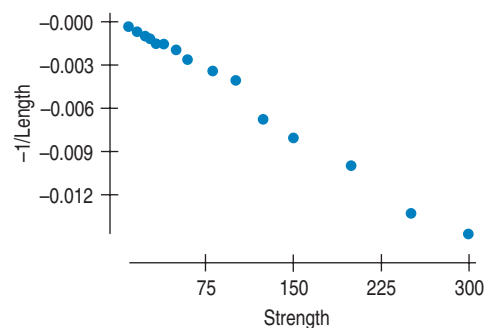
The scatterplot of the logarithm of *Length* against *Strength* is even less bent:



The straightness is improving, so we know we're moving in the right direction. But since the plot of the logarithms is not yet straight, we know we haven't gone far enough. To keep the direction consistent, change the sign and re-express to $-1/\text{Length}$.

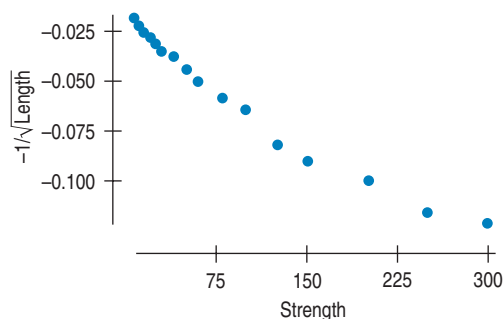
We may have to choose between two adjacent re-expressions. For most data analyses, it really doesn't matter which we choose.

This is much better, but still not straight, so I'll take another step to the -1 power, or reciprocal.



Maybe now I moved too far along the ladder.

A half-step back is the $-1/2$ power: the reciprocal square root.



TELL ➡ **Conclusion** Specify your choice of re-expression. If there's some natural interpretation (as for gallons per 100 miles), give that.

It's hard to choose between the last two alternatives. Either of the last two choices is good enough. I'll choose the $-1/2$ power.

Now that the re-expressed data satisfy the Straight Enough Condition, we can fit a linear model by least squares. We find that

$$\frac{-1}{\sqrt{\text{Length}}} = -0.023 - 0.000373 \text{ Strength}.$$

We can use this model to predict the length of a spool of, say, 35-pound test line:

$$\frac{-1}{\sqrt{\text{Length}}} = -0.023 - 0.000373 \times 35 = -0.036$$

We could leave the result in these units ($-1/\sqrt{\text{yards}}$). Sometimes the new units may be as meaningful as the original, but here we want to transform the predicted value back into yards. Fortunately, each of the re-expressions in the Ladder of Powers can be reversed.

To reverse the process, we first take the reciprocal: $\sqrt{\text{Length}} = -1/(-0.036) = 27.778$. Then squaring gets us back to the original units:

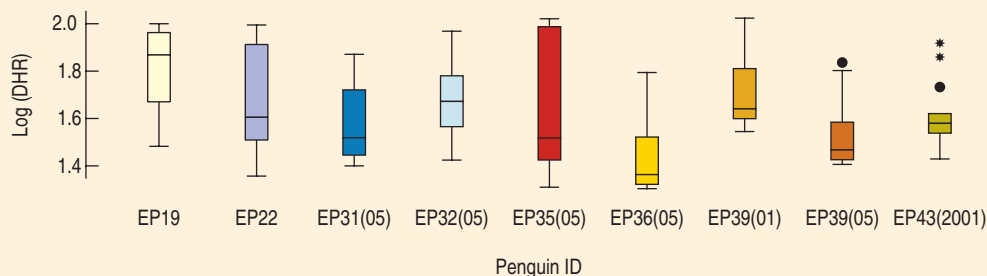
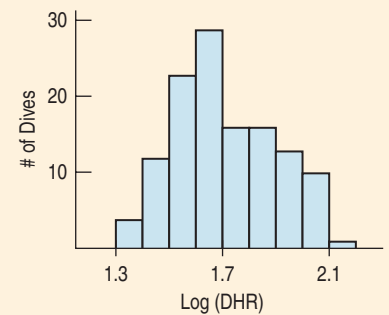
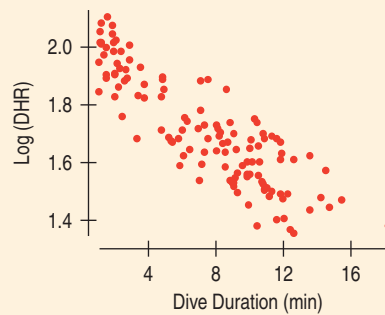
$$\widehat{\text{Length}} = 27.778^2 = 771.6 \text{ yards}.$$

This may be the most painful part of the re-expression. Getting back to the original units can sometimes be a little work. Nevertheless, it's worth the effort to always consider re-expression. Re-expressions extend the reach of all of your Statistics tools by helping more data to satisfy the conditions they require. Just think how much more useful this course just became!

For Example COMPARING RE-EXPRESSIONS

RECAP: We've concluded that in trying to straighten the relationship between *Diving Heart Rate* and *Dive Duration* for emperor penguins, using the reciprocal re-expression goes a bit "too far" on the ladder of powers. Now we try the logarithm. Here are the resulting displays:

QUESTION: Comment on these displays. Now that we've looked at the original data (rung 1 on the Ladder), the reciprocal (rung -1), and the logarithm (rung 0), which re-expression of *Diving Heart Rate* would you choose?



ANSWER: The scatterplot is now more linear and the histogram is symmetric. The boxplots are still a bit skewed to the high end, but less so than for the original *Diving Heart Rate* values. We don't expect real data to cooperate perfectly, and the logarithm seems like the best compromise re-expression, improving several different aspects of the data.

TI Tips RE-EXPRESSIONS DATA TO ACHIEVE LINEARITY



```
log(LTUIT)→L1
{3.815976001 3...
```



Let's revisit the Arizona State tuition data. Recall that back in Chapter 7 when we tried to fit a linear model to the yearly tuition costs, the residuals plot showed a distinct curve. Residuals are high (positive) at the left, low in the middle of the decade, and high again at the right.

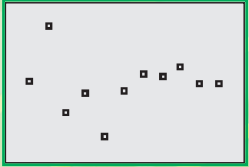
This curved pattern indicates that data re-expression may be in order. If you have no clue what re-expression to try, the Ladder of Powers may help. We just used that approach in the fishing line example. Here, though, we can play a hunch. It is reasonable to suspect that tuition increases at a relatively consistent percentage year by year. This suggests that using the logarithm of tuition may help.

- Tell the calculator to find the logs of the tuitions, and store them as a new list. Remember that you must import the name TUIT from the LIST NAMES menu. The command is `log(LTUIT) STO L1`.
- Check the scatterplot for the re-expressed data by changing your STATPLOT specifications to `Xlist:YR` and `Ylist:L1`. (Don't forget to use 9: ZoomStat to resize the window properly.)

The new scatterplot looks quite linear, but it's really the residuals plot that will tell the story. Remember that the TI automatically finds and stores the residuals whenever you ask it to calculate a regression.

(continued)

```
LinReg
y=a+bx
a=3.815541881
b=.0175535352
r=.9908736906
r=.9954263863
```



```
Y1(11)
4.008630769
ln^(Ans)
10200.71864
```

- Perform the regression for the *logarithm of tuition* vs. *year* with the command `LinReg(a + bx)`, setting `Xlist:YR`, `Ylist:L1`, and `RegEQ:Y1` (or on an older calculator, `LinReg(a+bx) 1YR, L1, Y1`). That both creates the residuals and reports details about the model (storing the equation for later use).
- Now that the residuals are stored in `RESID`, set up a new scatterplot, this time specifying `Xlist:YR` and `Ylist:RESID`.

While the residuals for the second and fifth years are comparatively large, the curvature we saw above is gone. The pattern in these residuals seem essentially horizontal and random. This re-expressed model is probably more useful than the original linear model.

Do you know what the model's equation is? Remember, it involves a log re-expression. The calculator does not indicate that; be sure to *Think* when you write your model!

$$\log \widehat{tu\hat{it}} = 3.816 + 0.018 \text{ yr}$$

And you have to *Think* some more when you make an estimate using the calculator's equation. Notice that this model does not actually predict tuition; rather, it predicts the *logarithm* of the tuition.

For example, to estimate the 2001 tuition we must first remember that in entering our data we designated 1990 as year 0. That means we'll use 11 for the year 2001 and evaluate `Y1(11)`.

No, we're not predicting the tuition to be \$4! That's the log of the estimated tuition. Since logarithms are exponents, $\log(\widehat{tu\hat{it}}) = 4$ means $\widehat{tu\hat{it}} = 10^4$, or about \$10,000. When you are working with models that involve re-expressions, you'll often need to "backsolve" like this to find the correct predictions.

Plan B: Attack of the Logarithms

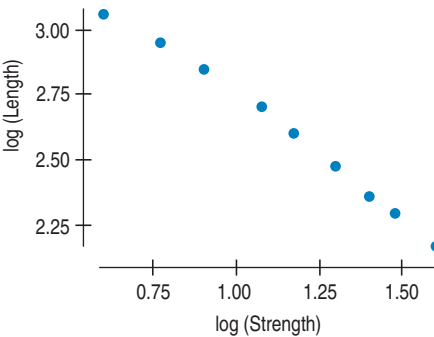


Figure 9.8
Plotting $\log(\text{Length})$ against $\log(\text{Strength})$ gives a straighter shape.

The Ladder of Powers is often successful at finding an effective re-expression. Sometimes, though, the curvature is more stubborn, and we're not satisfied with the residual plots. What then?

When none of the data values is zero or negative, logarithms can be a helpful ally in the search for a useful model. Try taking the logs of both the x - and y -variables. Then re-express the data using some combination of x or $\log(x)$ vs. y or $\log(y)$. You may find that one of these works pretty well.

| Model Name | x -axis | y -axis | Comment |
|-------------|-----------|-----------|---|
| Exponential | x | $\log(y)$ | This model is the "0" power in the ladder approach, useful for values that grow by percentage increases. |
| Logarithmic | $\log(x)$ | y | A wide range of x -values, or a scatterplot descending rapidly at the left but leveling off toward the right, may benefit from trying this model. |
| Power | $\log(x)$ | $\log(y)$ | The Goldilocks model: When one of the ladder's powers is too big and the next is too small, this one may be just right. |

When we tried to model the relationship between the length of fishing line and its strength, we were torn between the -1 power and the $-1/2$ power. The first showed slight upward curvature, and the second downward. Maybe there's a better power between those values.

The scatterplot shows what happens when we graph the logarithm of *Length* against the logarithm of *Strength*. Technology reveals that the equation of our log-log model is

$$\widehat{\log(\text{Length})} = 4.49 - 1.08 \log(\text{Strength}).$$

It's interesting that the slope of this line (-1.08) is a power⁵ we didn't try. After all, the ladder can't have every imaginable rung.

A warning, though! Don't expect to be able to straighten every curved scatterplot you find. It may be that there just isn't a very effective re-expression to be had. You'll certainly encounter situations when nothing seems to work the way you wish it would. Don't set your sights too high—you won't find a perfect model. Keep in mind: We seek a *useful* model, not perfection (or even "the best").

TI Tips USING LOGARITHMIC RE-EXPRESSIONS



```
log(L1)→L3
(-3 -2.69897000...
log(L2)→L4
(.4471580313 .6...
```



```
LinReg
y=a+bx
a=1.93880413
b=.4969548956
r²=.9993420212
r=.9996709565
```

In Chapter 6 we looked at data showing the relationship between the f/stop of a camera's lens and its shutter speed. Let's use the attack of the logarithms to model this situation.

| | | | | | | | | |
|------------------------------------|--------|-------|-------|-------|------|------|------|-----|
| Shutter speed: | 1/1000 | 1/500 | 1/250 | 1/125 | 1/60 | 1/30 | 1/15 | 1/8 |
| f/stop: | 2.8 | 4 | 5.6 | 8 | 11 | 16 | 22 | 32 |

- Enter these data into your calculator, shutter *speed* in L1 and f/stop in L2.
- Create the scatterplot with Xlist:L1 and Ylist:L2. See the curve?
- Find the logarithms of each variable's values. Keep track of where you store everything so you don't get confused! We put $\log(\text{speed})$ in L3 and $\log(f/\text{stop})$ in L4.
- Make three scatterplots:
 - f/stop vs. $\log(\text{speed})$ using Xlist:L3 and Ylist:L2
 - $\log(f/\text{stop})$ vs. speed using Xlist:L1 and Ylist:L4
 - $\log(f/\text{stop})$ vs. $\log(\text{speed})$ using Xlist:L3 and Ylist:L4
- Pick your favorite. We liked $\log(f/\text{stop})$ vs. $\log(\text{speed})$ a lot! It appears to be very straight. (Don't be misled—this is a situation governed by the laws of Physics. Real data are not so cooperative. Don't expect to achieve this level of perfection often!)
- Remember that before you check the residuals plot, you first have to calculate the regression. In this situation all the errors in the residuals are just round-off errors in the original f/stops .
- Use your regression to write the equation of the model. Remember: The calculator does not know there were logarithms involved. You have to Think about that to be sure you write your model correctly.⁶

$$\log(\widehat{f/\text{stop}}) = 1.94 + 0.497\log(\text{speed})$$

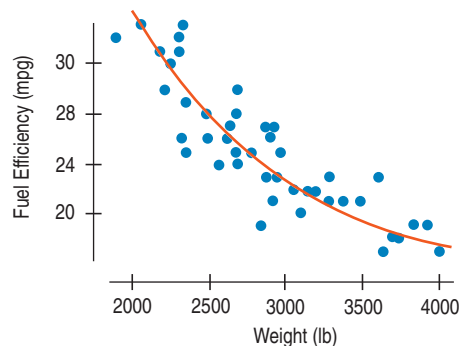
Why Not Just Use a Curve?

When a clearly curved pattern shows up in the scatterplot, why not just fit a curve to the data? We saw earlier that the association between the *Weight* of a car and its *Fuel Efficiency* was not a straight line. Instead of trying to find a way to straighten the plot, why not find a curve that seems to describe the pattern well?

We can find "curves of best fit" using essentially the same approach that led us to linear models. You won't be surprised, though, to learn that the mathematics and the calculations are considerably more difficult for curved models. Many calculators and

⁵For logarithms, $-1.08 \log(\text{Strength}) = \log(\text{Strength}^{-1.08})$.

⁶See the slope, 0.497? Just about 0.5. That's because the actual relationship involves the square root of shutter speeds. Technically the f/stop listed as 2.8 should be $2\sqrt{2} \approx 2.8284$. Rounding off to 2.8 makes sense for photographers, but it's what led to the minor errors you saw in the residuals plot.



computer packages do have the ability to fit curves to data, but this approach has many drawbacks.

Straight lines are easy to understand. We know how to think about the slope and the y-intercept, for example. We often want some of the other benefits mentioned earlier, such as making the spread around the model more nearly the same everywhere. In later chapters you will learn more advanced statistical methods for analyzing linear associations.

We give all of that up when we fit a model that is not linear. For many reasons, then, it is usually better to re-express the data to straighten the plot.

TI Tips SOME SHORTCUTS TO AVOID

Your calculator offers many regression options in the STAT CALC menu. There are three that automate fitting simple re-expressions of y or x :

- 9:LnReg—fits a logarithmic model ($\hat{y} = a + b \ln x$)
- 0:ExpReg—fits an exponential model ($\hat{y} = ab^x$)
- A:PwrReg—fits a power model ($\hat{y} = ax^b$)

In addition, the calculator offers two other functions:

- 5:QuadReg—fits a quadratic model ($\hat{y} = ax^2 + bx + c$)
- 6:CubicReg—fits a cubic model ($\hat{y} = ax^3 + bx^2 + cx + d$)

These two models have a form we haven't seen, with several x -terms. Because x , x^2 , and x^3 are likely to be highly correlated with each other, the quadratic and cubic models are almost sure to be unreliable to fit, difficult to understand, and dangerous to use for predictions that are even slight extrapolations. We recommend that you be very wary of models of this type.

Let's try out one of the calculator shortcuts; we'll use the Arizona State tuition data. (For the last time, we promise!) This time, instead of re-expressing *tuition* to straighten the scatterplot, we'll have the calculator do more of the work.

Which model should you use? You could always just play hit-and-miss, but knowing something about the data can save a lot of time. If tuition increases by a consistent percentage each year, then the growth is exponential. The Exp Reg results all look very good: R^2 is high, the curve appears to fit the points quite well, and the residuals plot is acceptably random.

The equation of the model is $\widehat{tuitt} = 6539.46(1.041^{year})$.

Notice, though that this is the same residuals plot we saw when we re-expressed the data and fit a line to the logarithm of *tuition*. That's the calculator just did the very same thing. This new equation may look different, but it is equivalent to our earlier model $\log \widehat{tuitt} = 3.816 + 0.018 \text{ year}$.

Not easy to see that, is it? Here's how it works:

Initially we used a logarithmic re-expression to create a linear model:

$$\log \hat{y} = a + bx$$

Rewrite that equation in exponential form:

$$\hat{y} = 10^{a+bx}$$

Simplify, using the laws of exponents:

$$\hat{y} = 10^a(10^b)^x$$

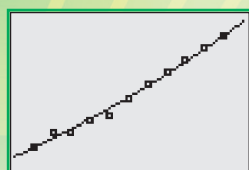
Let $10^a = a$ and $10^b = b$ (different a and b !)

$$\hat{y} = ab^x$$

See? Your linear model created by logarithmic re-expression is the same as the calculator model created by ExpReg. In fact, three of the special TI functions correspond to a simple regression model involving re-expression.

```
ExpReg LVR, LTUIT
Y1
```

```
ExpReg
y=a*b^x
a=6539.459906
b=1.041246454
r^2=.9908736906
r=.9954263863
```



| Type of Model | Re-expression Equation | Calculator's Curve | |
|---------------|-------------------------------|--------------------|-------------------------|
| | | Command | Equation |
| Logarithmic | $\hat{y} = a + b \log x$ | LnReg | $\hat{y} = a + b \ln x$ |
| Exponential | $\log \hat{y} = a + bx$ | ExpReg | $\hat{y} = ab^x$ |
| Power | $\log \hat{y} = a + b \log x$ | PwrReg | $\hat{y} = ax^b$ |

Be careful. It may look like the calculator fit these equations to the data by minimizing the sum of squared residuals, but it really didn't do that. It handles the residuals differently, and the difference matters. If you use a statistics program to fit an "exponential model," it will probably fit the exponential form of the equation and give you a different answer.

You've seen two ways to handle bent relationships:

- straighten the data, then fit a line, or
- use the calculator shortcut to create a curve.

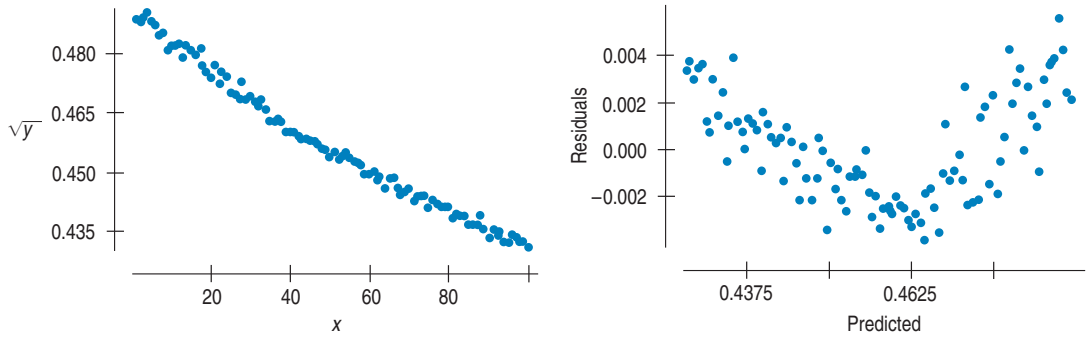
Note that the calculator does not have a shortcut for every model you might want to use—models involving square roots or reciprocals, for instance. And remember: The calculator may be quick, but there are real advantages to finding *linear* models by actually re-expressing the data. That's the approach you should always use.

WHAT CAN GO WRONG?

Occam's Razor If you think that simpler explanations and simpler models are more likely to give a true picture of the way things work, then you should look for opportunities to re-express your data and simplify your analyses.

The general principle that simpler explanations are likely to be the better ones is known as Occam's Razor, after the English philosopher and theologian William of Occam (1284–1347).

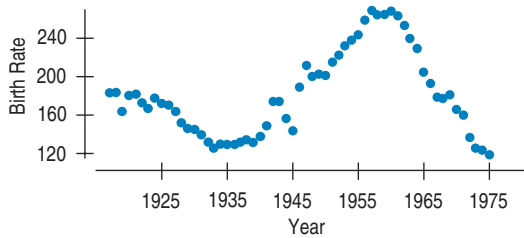
- **Don't get seduced by ExpReg and its calculator cousins.** Those so-called "curved" regression options look enticing, but don't go there. This course is about *linear* regression. If you see a curve, re-express the data to achieve linearity and then fit a line. Equations of lines are easier to interpret, and will be far easier to work with later on when we do more advanced statistical analyses.
- **Don't expect your model to be perfect.** In Chapter 5 we quoted statistician George Box: "All models are wrong, but some are useful." Be aware that the real world is a messy place and data can be uncooperative. Don't expect to find one elusive re-expression that magically irons out every kink in your scatterplot and produces perfect residuals. You aren't looking for the Right Model, because that mythical creature doesn't exist. Find a useful model and use it wisely.
- **Don't stray too far from the ladder.** It's wise not to stray too far from the powers that we suggest in the Ladder of Powers. Stick to powers between 2 and -2 . Even in that interval, you should prefer the simpler powers in the ladder to those in the cracks. A square root is easier to understand than the 0.413 power. That simplicity may compensate for a slightly less straight relationship.
- **Don't choose a model based on R^2 alone.** You've tried re-expressing your data to straighten a curved relationship and found a model with a high R^2 . Beware: That doesn't mean the pattern is straight now. On the next page is a plot of a relationship with an R^2 of 98.3%. The R^2 is about as high as we could ask for, but if you look closely, you'll see that there's a consistent bend. Plotting the residuals from the least squares line makes the bend much easier to see.



Remember the basic rule of data analysis: *Make a picture*. Before you fit a line, always look at the pattern in the scatterplot. After you fit the line, check for linearity again by plotting the residuals.

- **Beware of multiple modes.** Re-expression can often make a skewed unimodal histogram more nearly symmetric, but it cannot pull separate modes together. A suitable re-expression may, however, make the separation of the modes clearer, simplifying their interpretation and making it easier to separate them to analyze individually.
- **Watch out for scatterplots that turn around.** Re-expression can straighten many bent relationships but not those that go up and then down or down and then up. You should refuse to analyze such data with methods that require a linear form.

Figure 9.9
The shape of the scatterplot of *Birth Rates* (births per 100,000 women) in the United States shows an oscillation that cannot be straightened by re-expressing the data.



- **Watch out for zero or negative data values.** It’s impossible to re-express negative values by any power that is not a whole number on the Ladder of Powers or to re-express values that are zero for negative powers. One possible cure for zeros and small negative values is to add a constant ($\frac{1}{2}$ and $\frac{1}{6}$ are often used) to bring all the data values above zero.



What Have We Learned?

We’ve learned that when the conditions for regression are not met, a simple re-expression of the data may help. There are several reasons to consider a re-expression:

- To make the distribution of a variable more symmetric (as we saw in Chapter 4)
- To make the spread across different groups more similar
- To make the form of a scatterplot straighter
- To make the scatter around the line in a scatterplot more consistent

We’ve learned that when seeking a useful re-expression, taking logs is often a good, simple starting point. To search further, the Ladder of Powers or the log–log approach can help us find a good re-expression.

We’ve come to understand that our models won’t be perfect, but that re-expression can lead us to a useful model.

Terms

Re-expression

We re-express data by taking the logarithm, the square root, the reciprocal, or some other mathematical operation of all values of a variable. (p. 232)

Ladder of Powers

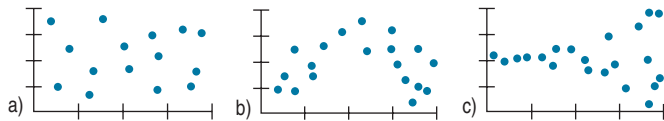
The Ladder of Powers places in order the effects that many re-expressions have on the data. (p. 237)

On the Computer RE-EXPRESSION

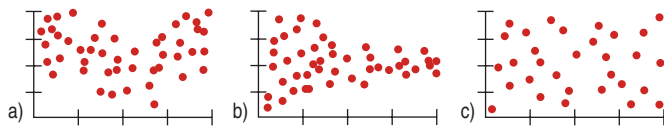
Computers and calculators make it easy to re-express data. Most statistics packages offer a way to re-express and compute with variables. Some packages permit you to specify the power of a re-expression with a slider or other moveable control, possibly while watching the consequences of the re-expression on a plot or analysis. This, of course, is a very effective way to find a good re-expression.

Exercises

1. **Residuals** Suppose you have fit a linear model to some data and now take a look at the residuals. For each of the following possible residuals plots, tell whether you would try a re-expression and, if so, why.

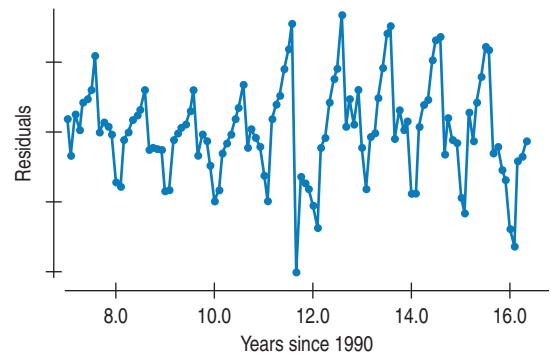


2. **Residuals** Suppose you have fit a linear model to some data and now take a look at the residuals. For each of the following possible residuals plots, tell whether you would try a re-expression and, if so, why.

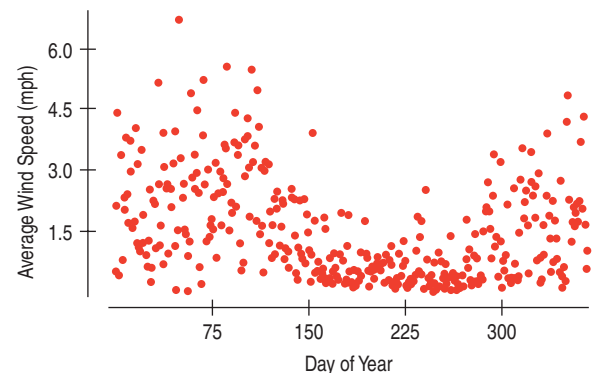


3. **Oakland passengers revisited** In Chapter 8, Exercise 15, we created a linear model describing the trend in the number of passengers departing from the Oakland (CA) airport each month since the start of 1997. Here's the residual plot, but with lines added to show the order of the values in time:

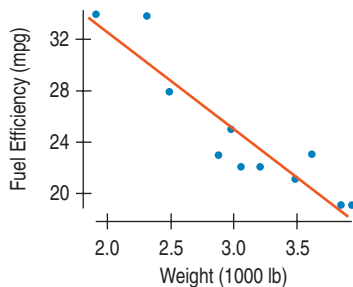
- a) Can you account for the pattern shown here?
b) Would a re-expression help us deal with this pattern? Explain.



4. **Hopkins winds, revisited** In Chapter 4, we examined the wind speeds in the Hopkins forest over the course of a year. Here's the scatterplot we saw then:



- a) Describe the pattern you see here.
b) Should we try re-expressing either variable to make this plot straighter? Explain.
5. **Models** For each of the models listed below, predict y when $x = 2$.
- a) $\ln \hat{y} = 1.2 + 0.8x$ d) $\hat{y} = 1.2 + 0.8 \ln x$
b) $\sqrt{\hat{y}} = 1.2 + 0.8x$ e) $\log \hat{y} = 1.2 + 0.8 \log x$
c) $\frac{1}{\hat{y}} = 1.2 + 0.8x$
6. **More models** For each of the models listed below, predict y when $x = 2$.
- a) $\hat{y} = 1.2 + 0.8 \log x$ d) $\hat{y}^2 = 1.2 + 0.8x$
b) $\log \hat{y} = 1.2 + 0.8x$ e) $\frac{1}{\sqrt{\hat{y}}} = 1.2 + 0.8x$
c) $\ln \hat{y} = 1.2 + 0.8 \ln x$
7. **Gas mileage** As the example in the chapter indicates, one of the important factors determining a car's *Fuel Efficiency* is its *Weight*. Let's examine this relationship again, for 11 cars.
- a) Describe the association between these variables shown in the scatterplot.

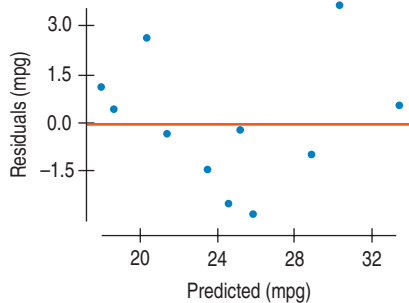


- b) Here is the regression analysis for the linear model. What does the slope of the line say about this relationship?

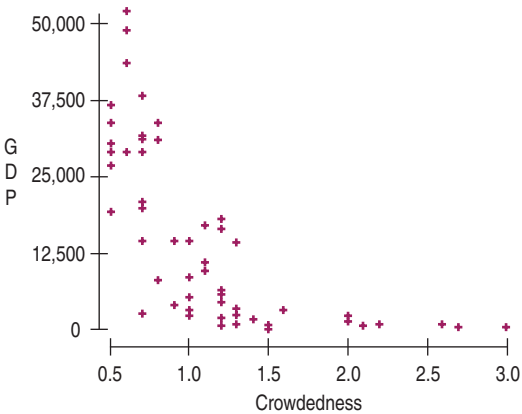
Dependent variable is: Fuel Efficiency
R-squared = 85.9%

| Variable | Coefficient |
|-----------|-------------|
| Intercept | 47.9636 |
| Weight | -7.65184 |

- c) Do you think this linear model is appropriate? Use the residuals plot to explain your decision.



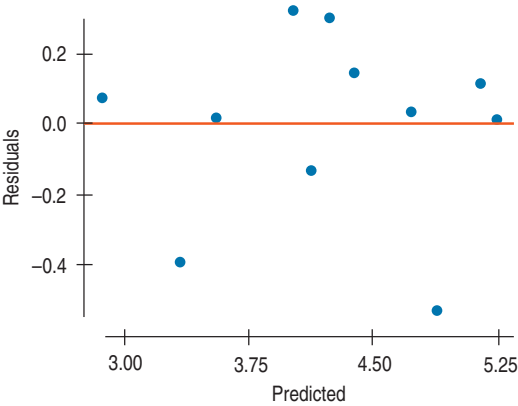
8. **Crowdedness** In a *Chance* magazine article (Summer 2005), Danielle Vasilescu and Howard Wainer used data from the United Nations Center for Human Settlements to investigate aspects of living conditions for several countries. Among the variables they looked at were the country's per capita gross domestic product (*GDP*, in \$) and *Crowdedness*, defined as the average number of persons per room living in homes there. This scatterplot displays these data for 56 countries:



- a) Explain why you should re-express these data before trying to fit a model.
b) What re-expression of *GDP* would you try as a starting point?
9. **Gas mileage revisited** Let's try the re-expressed variable *Fuel Consumption* (gal/100 mi) to examine the fuel efficiency of the 11 cars in Exercise 7. Here are the revised regression analysis and residuals plot:

Dependent variable is: Fuel Consumption
R-squared = 89.2%

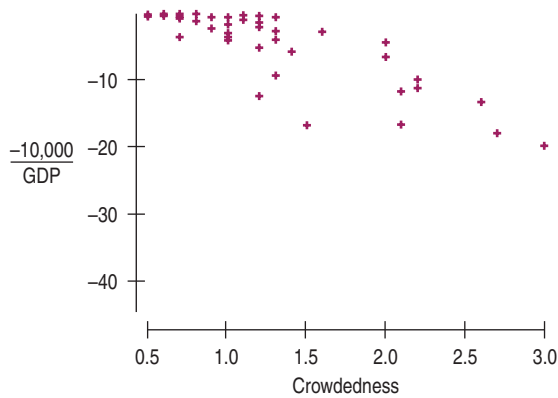
| Variable | Coefficient |
|-----------|-------------|
| Intercept | 0.624932 |
| Weight | 1.17791 |



- a) Explain why this model appears to be better than the linear model.

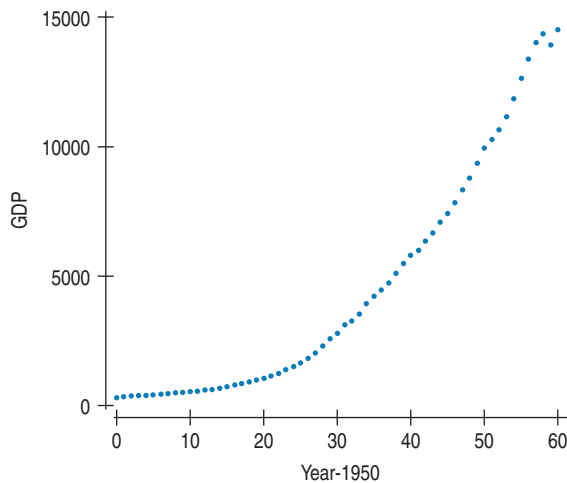
- b) Using the regression analysis above, write an equation of this model.
- c) Interpret the slope of this line.
- d) Based on this model, how many miles per gallon would you expect a 3500-pound car to get?

10. Crowdedness again In Exercise 8 we looked at United Nations data about a country's *GDP* and the average number of people per room (*Crowdedness*) in housing there. For a re-expression, a student tried the reciprocal $-10000/\text{GDP}$, representing the number of people per \$10,000 of gross domestic product. Here are the results, plotted against *Crowdedness*:



- a) Is this a useful re-expression? Explain.
- b) What re-expression would you suggest this student try next?

11. GDP The scatterplot shows the *gross domestic product* (*GDP*) of the United States in billions of dollars plotted against years since 1950.

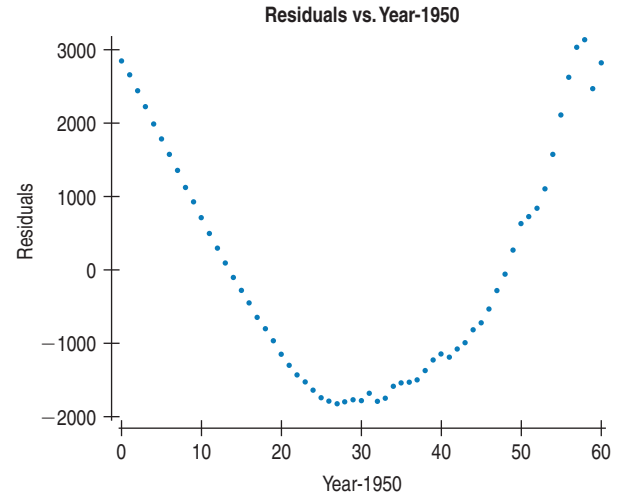


A linear model fit to the relationship looks like this:

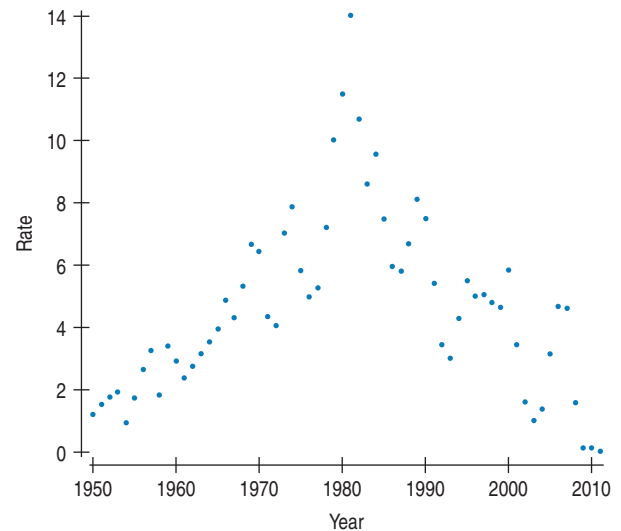
Dependent variable is: GDP
R-squared = 87.6% $s = 1597.7456$

| Variable | Coefficient |
|-----------|-------------|
| Intercept | -2561.3552 |
| Year-1950 | 237.74577 |

- a) Does the value 87.6% suggest that this is a good model? Explain.
- b) Here's a scatterplot of the residuals. Now do you think this is a good model for these data? Explain?



12. Treasury Bills The 3-month Treasury bill interest rate is watched by investors and economists. Here's a scatterplot of the 3-month Treasury bill rate since 1950:

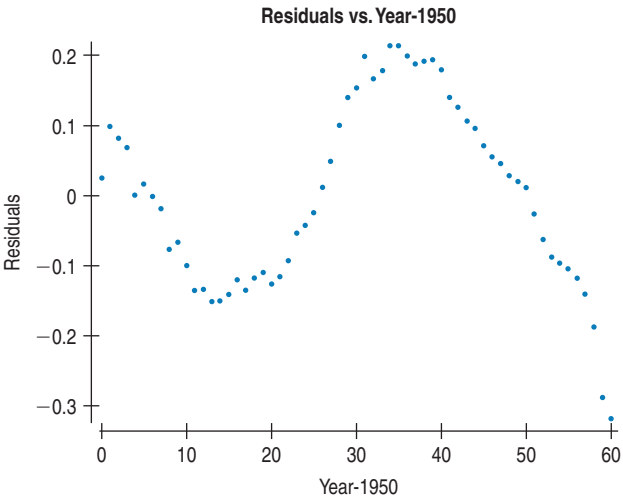


Clearly, the relationship is not linear. Can it be made nearly linear with a re-expression? If so, which one would you suggest? If not, why not?

13. Better GDP model? Consider again the post-1950 trend in U.S. GDP we examined in Exercise 11. Here are a regression and (on the next page) a residual plot when we use the log of GDP in the model. Is this a better model for GDP? Explain.

Dependent variable is: logGDP
R-squared = 98.9% $s = 0.13185$

| Variable | Coefficient |
|-----------|-------------|
| Intercept | 5.6579766 |
| Year-1950 | 0.070734456 |



- T 14. Pressure** Scientist Robert Boyle examined the relationship between the volume in which a gas is contained and the pressure in its container. He used a cylindrical container with a moveable top that could be raised or lowered to change the volume. He measured the *Height* in inches by counting equally spaced marks on the cylinder, and measured the *Pressure* in inches of mercury (as in a barometer). Some of his data are listed in the table. Create an appropriate model.

| | | | | | | |
|-----------------|------|------|------|------|-------|-------|
| Height | 48 | 44 | 40 | 36 | 32 | 28 |
| Pressure | 29.1 | 31.9 | 35.3 | 39.3 | 44.2 | 50.3 |
| Height | 24 | 20 | 18 | 16 | 14 | 12 |
| Pressure | 58.8 | 70.7 | 77.9 | 87.9 | 100.4 | 117.6 |

- T 15. Brakes** The table below shows stopping distances in feet for a car tested 3 times at each of 5 speeds. We hope to create a model that predicts *Stopping Distance* from the *Speed* of the car.

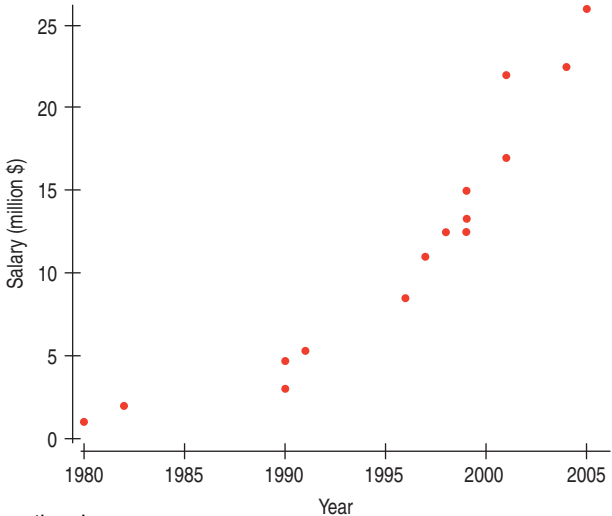
| Speed (mph) | Stopping Distances (ft) |
|--------------------|--------------------------------|
| 20 | 64, 62, 59 |
| 30 | 114, 118, 105 |
| 40 | 153, 171, 165 |
| 50 | 231, 203, 238 |
| 60 | 317, 321, 276 |

- a) Explain why a linear model is not appropriate.
b) Re-express the data to straighten the scatterplot.
c) Create an appropriate model.
d) Estimate the stopping distance for a car traveling 55 mph.
e) Estimate the stopping distance for a car traveling 70 mph.
f) How much confidence do you place in these predictions? Why?
- T 16. Pendulum** A student experimenting with a pendulum counted the number of full swings the pendulum made in 20 seconds for various lengths of string. Here are her data.

| | | | | | | | | | | |
|-------------------------|-----|----|------|------|----|----|----|----|----|------|
| Length (in.) | 6.5 | 9 | 11.5 | 14.5 | 18 | 21 | 24 | 27 | 30 | 37.5 |
| Number of Swings | 22 | 20 | 17 | 16 | 14 | 13 | 13 | 12 | 11 | 10 |

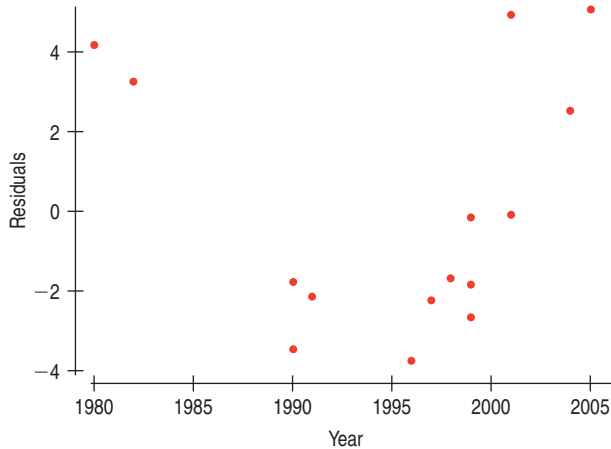
- a) Explain why a linear model is not appropriate for using the *Length* of a pendulum to predict the *Number of Swings* in 20 seconds.
b) Re-express the data to straighten the scatterplot.
c) Create an appropriate model.
d) Estimate the number of swings for a pendulum with a 4-inch string.
e) Estimate the number of swings for a pendulum with a 48-inch string.
f) How much confidence do you place in these predictions? Why?
- T 17. Baseball salaries 2012** Ballplayers have been signing ever larger contracts. The highest salaries (in millions of dollars per season) for some notable players are given in the table and plotted below by year.

| Player | Year | Salary (million \$) |
|------------------|-------------|----------------------------|
| Nolan Ryan | 1980 | 1.0 |
| George Foster | 1982 | 2.0 |
| Kirby Puckett | 1990 | 3.0 |
| Jose Canseco | 1990 | 4.7 |
| Roger Clemens | 1991 | 5.3 |
| Ken Griffey, Jr. | 1996 | 8.5 |
| Albert Belle | 1997 | 11.0 |
| Pedro Martinez | 1998 | 12.5 |
| Mike Piazza | 1999 | 12.5 |
| Mo Vaughn | 1999 | 13.3 |
| Kevin Brown | 1999 | 15.0 |
| Carlos Delgado | 2001 | 17.0 |
| Alex Rodriguez | 2001 | 22.0 |
| Manny Ramirez | 2004 | 22.5 |
| Alex Rodriguez | 2005 | 26.0 |



- a) Examine the scatterplot above. Does it look straight?
Given what you know about money and inflation, would you expect it to be straight?

Here is the residual plot for regression of *Year* vs. *Salary*.

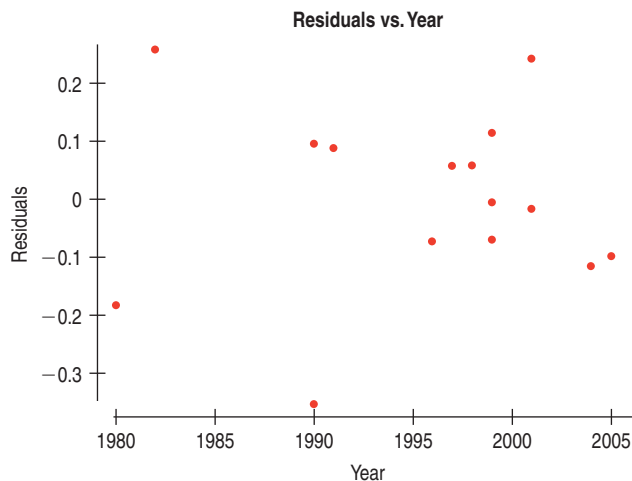


- b) What does this residual plot tell you about using the regression model for *Year* vs. *Salary*?

The log of salary was computed in an attempt to straighten the data. Regression was run on *Year* vs. \ln *Salary*. Here are the results and the residual plot.

Dependent variable is: \ln Salary
R-squared = 97.2% $s = 0.16478$

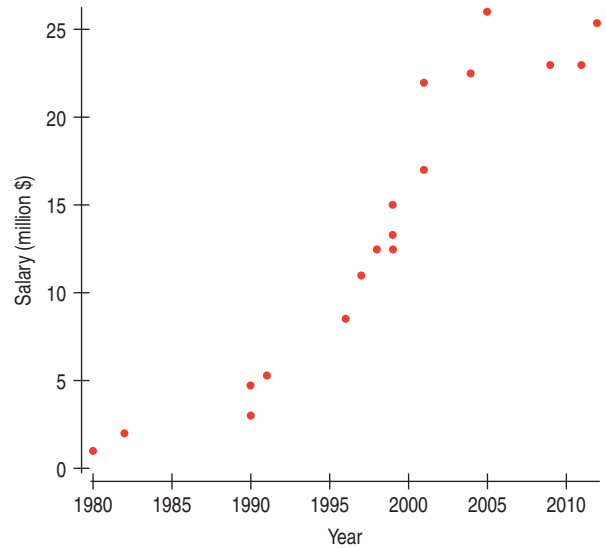
| Variable | Coefficient |
|-----------|-------------|
| Intercept | -251.28738 |
| Year | 0.1270045 |



- c) Was this transformation successful? Use this model to predict the top salary in 2015.
d) After A-Rod's record setting salary, the next three biggest contracts were:

| | | |
|------|---------------|--------|
| 2009 | CC Sabathia | \$23 |
| 2011 | Joe Mauer | \$23 |
| 2012 | Albert Pujols | \$25.4 |

Adding the new data to the scatterplot, we see:



Given this recent trend, how do you feel about your 2015 prediction? Describe how the new data and the graph change your perspective on this model.

- T 18. Planet distances and years 2012** At a meeting of the International Astronomical Union (IAU) in Prague in 2006, Pluto was determined not to be a planet, but rather the largest member of the Kuiper belt of icy objects. Let's examine some facts. Here is a table of the 9 sun-orbiting objects formerly known as planets:

| Planet | Position Number | Distance from Sun (million miles) | Length of Year (Earth years) |
|---------|-----------------|-----------------------------------|------------------------------|
| Mercury | 1 | 36 | 0.24 |
| Venus | 2 | 67 | 0.61 |
| Earth | 3 | 93 | 1.00 |
| Mars | 4 | 142 | 1.88 |
| Jupiter | 5 | 484 | 11.86 |
| Saturn | 6 | 887 | 29.46 |
| Uranus | 7 | 1784 | 84.07 |
| Neptune | 8 | 2796 | 164.82 |
| Pluto | 9 | 3707 | 247.68 |

- a) Plot the *Length* of the year against the *Distance* from the sun. Describe the shape of your plot.
b) Re-express one or both variables to straighten the plot. Use the re-expressed data to create a model describing the length of a planet's year based on its distance from the sun.
c) Comment on how well your model fits the data.

- T 19. Planet distances and order 2012** Let's look again at the pattern in the locations of the planets in our solar system seen in the table in Exercise 18.

- a) Re-express the distances to create a model for the *Distance* from the sun based on the planet's *Position*.
- b) Based on this model, would you agree with the International Astronomical Union that Pluto is not a planet? Explain.

T 20. Planets 2012, part 3 The asteroid belt between Mars and Jupiter may be the remnants of a failed planet. If so, then Jupiter is really in position 6, Saturn is in 7, and so on. Repeat Exercise 19, using this revised method of numbering the positions. Which method seems to work better?

T 21. Eris: Planets 2012, part 4 In July 2005, astronomers Mike Brown, Chad Trujillo, and David Rabinowitz announced the discovery of a sun-orbiting object, since named Eris,⁷ that is 5% larger than Pluto. Eris orbits the sun once every 560 earth years at an average distance of about 6300 million miles from the sun. Based on its *Position*, how does Eris's *Distance* from the sun (re-expressed to logs) compare with the prediction made by your model of Exercise 19?

T 22. Models and laws: Planets 2012, part 5 The model you found in Exercise 18 is a relationship noted in the 17th century by Kepler as his Third Law of Planetary Motion. It was subsequently explained as a consequence of Newton's Law of Gravitation. The models for Exercises 19–21 relate to what is sometimes called the Titius-Bode “law,” a pattern noticed in the 18th century but lacking any scientific explanation.

Compare how well the re-expressed data are described by their respective linear models. What aspect of the model of Exercise 18 suggests that we have found a physical law? In the future, we may learn enough about a planetary system around another star to tell whether the Titius-Bode pattern applies there. If you discovered that another planetary system followed the same pattern, how would it change your opinion about whether this is a real natural “law”? What would you think if the next system we find does not follow this pattern?

23. Logs (not logarithms) The value of a log is based on the number of board feet of lumber the log may contain. (A board foot is the equivalent of a piece of wood 1 inch thick, 12 inches wide, and 1 foot long. For example, a 2" × 4" piece that is 12 feet long contains 8 board feet.) To estimate the amount of lumber in a log, buyers measure the diameter inside the bark at the smaller end. Then they look in a table based on the Doyle Log Scale. The table below shows the estimates for logs 16 feet long.

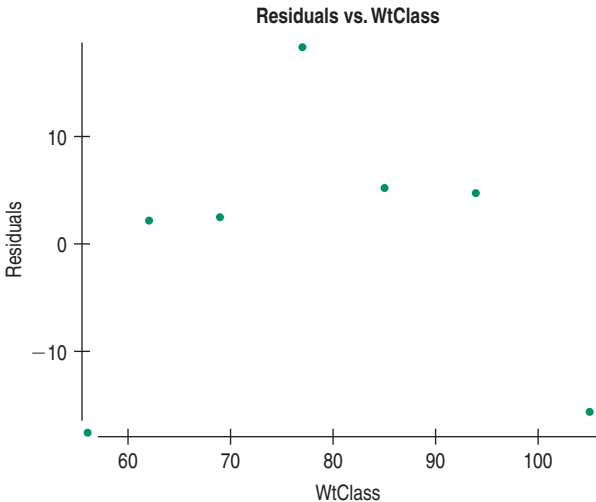
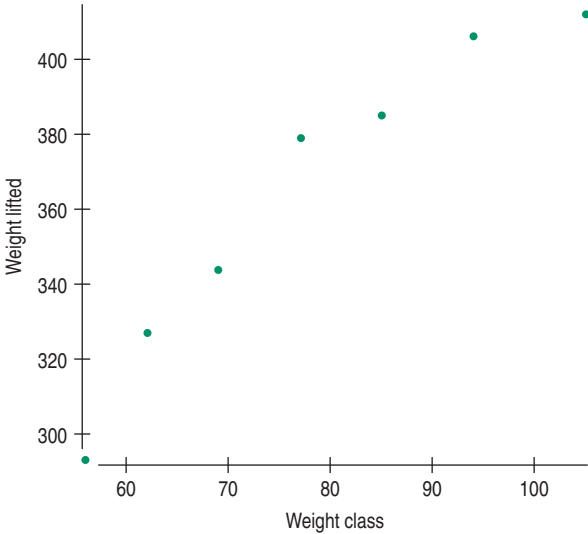
| Diameter of Log | 8" | 12" | 16" | 20" | 24" | 28" |
|-----------------|----|-----|-----|-----|-----|-----|
| Board Feet | 16 | 64 | 144 | 256 | 400 | 576 |

⁷Eris is the Greek goddess of warfare and strife who caused a quarrel among the other goddesses that led to the Trojan war. In the astronomical world, Eris stirred up trouble when the question of its proper designation led to the raucous meeting of the IAU in Prague where IAU members voted to demote Pluto and Eris to dwarf-planet status—<http://www.gps.caltech.edu/~mbrown/planetlila/#paper>.

- a) What model does this scale use?
- b) How much lumber would you estimate that a log 10 inches in diameter contains?
- c) What does this model suggest about logs 36 inches in diameter?

T 24. Weightlifting 2012 Listed below are the gold medal-winning men's weight-lifting performances at the 2012 Olympics, followed by some analysis.

| Weight Class (kg) | Name (Country) | Weight Lifted (kg) |
|-------------------|-----------------------------|--------------------|
| 56 | Yun Om (Korea) | 293 |
| 62 | Un Kim (Korea) | 327 |
| 69 | Qingfeng Lin (China) | 344 |
| 77 | Xiaojun Lu (China) | 379 |
| 85 | Adrian Zielinski (Poland) | 385 |
| 94 | Ilya Ilyin (Kazakhstan) | 406 |
| 105 | Oleksiy Torokhtiy (Ukraine) | 412 |

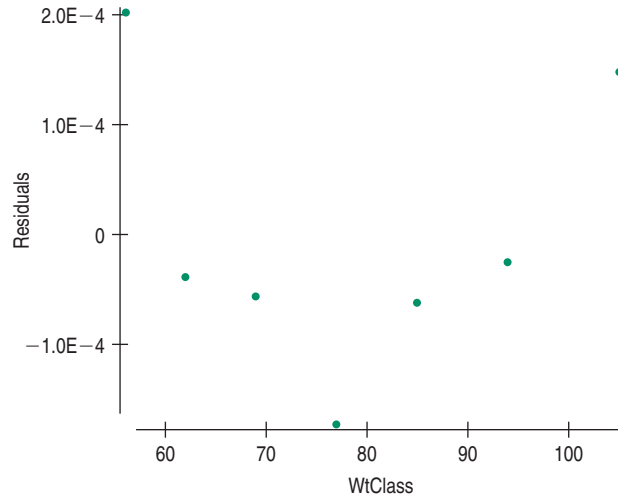


Dependent variable is: WtLifted
R-squared = 91.8% s = 13.7446

| Variable | Coefficient |
|-----------|-------------|
| Intercept | 176.607 |
| WtClass | 2.39 |

a) What does the residual plot tell you about the need to re-express?

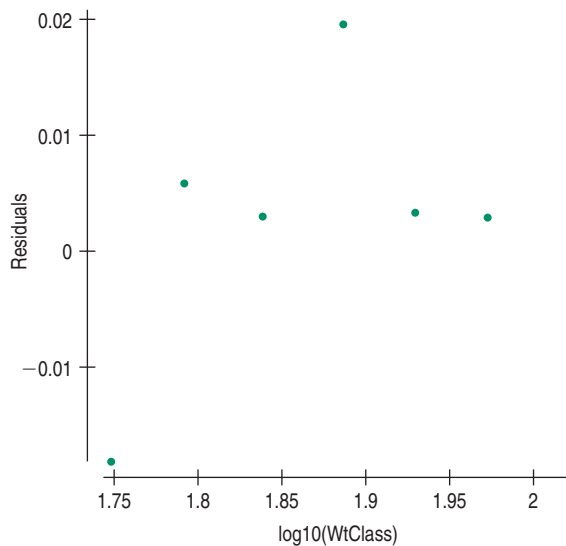
We reexpressed these data two ways, first using the reciprocal and then taking the logs of both variables. Here are the results.



Dependent variable is: 1 / WtLifted
R-squared = 86.8% s = 0.000143

| Variable | Coefficient |
|-----------|--------------|
| Intercept | 0.00427 |
| WtClass | -0.000019006 |

Residuals vs. log10(WtClass)



Dependent variable is: log(WtLifted)
R-squared = 94.1% s = 0.0144

| Variable | Coefficient |
|--------------|-------------|
| Intercept | 1.5479 |
| log(WtClass) | 0.5360 |

b) What does the residual plot tell you about the success of the re-expressions?

- T 25. Life expectancy** The data in the table below list the *Life Expectancy* for white males in the United States every decade during the last century (1 = 1900 to 1910, 2 = 1911 to 1920, etc.). Create a model to predict future increases in life expectancy. (National Vital Statistics Report)

| Decade | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|-----------|------|------|------|------|------|------|------|------|------|------|------|
| Life exp. | 48.6 | 54.4 | 59.7 | 62.1 | 66.5 | 67.4 | 68.0 | 70.7 | 72.7 | 74.9 | 76.5 |

- T 26. Lifting record weight 2012** In 2012, Xiaojun Lu from China set a world record in the 77kg weight class with a lift of 379 kg.
- Use the reciprocal re-expression in Exercise 24 to calculate his residual. Interpret this residual in context.
 - Does the sign of the residual surprise you, given that Xiaojun set a world record?
- T 27. Slower is cheaper?** Researchers studying how a car's *Fuel Efficiency* varies with its *Speed* drove a compact car 200 miles at various speeds on a test track. Their data are shown in the table.

| Speed (mph) | 35 | 40 | 45 | 50 | 55 | 60 | 65 | 70 | 75 |
|-----------------|------|------|------|------|------|------|------|------|------|
| Fuel Eff. (mpg) | 25.9 | 27.7 | 28.5 | 29.5 | 29.2 | 27.4 | 26.4 | 24.2 | 22.8 |

Create a linear model for this relationship and report any concerns you may have about the model.

- T 28. Orange production** The table below shows that as the number of oranges on a tree increases, the fruit tends to get smaller. Create a model for this relationship, and express any concerns you may have.

| Number of Oranges/Tree | Average Weight/Fruit (lb) |
|------------------------|---------------------------|
| 50 | 0.60 |
| 100 | 0.58 |
| 150 | 0.56 |
| 200 | 0.55 |
| 250 | 0.53 |
| 300 | 0.52 |
| 350 | 0.50 |
| 400 | 0.49 |
| 450 | 0.48 |
| 500 | 0.46 |
| 600 | 0.44 |
| 700 | 0.42 |
| 800 | 0.40 |
| 900 | 0.38 |

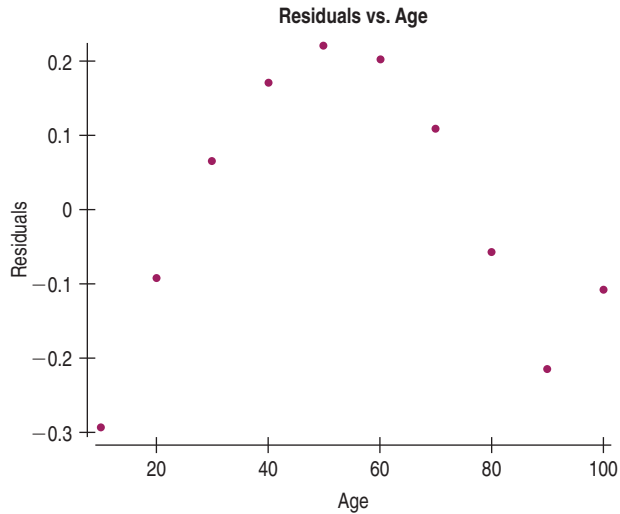
- T 29. Years to live 2008** Insurance companies and other organizations use actuarial tables to estimate the remaining life spans of their customers. The table below shows the predicted additional years of life for Hispanic females of various ages in the United States, according to a 2008 National Vital Statistics Report. (www.cdc.gov/nchs/deaths.htm)

| Age | Years to Live |
|-----|---------------|
| 10 | 73.9 |
| 20 | 64 |
| 30 | 54.2 |
| 40 | 44.5 |
| 50 | 35.1 |
| 60 | 26.1 |
| 70 | 17.8 |
| 80 | 10.6 |
| 90 | 5.3 |
| 100 | 2.6 |

Here are the results of a re-expression.

Dependent variable is: sqrt(YrsToLive)
R-squared = 99.4% s = 0.19068

| Variable | Coefficient |
|-----------|-------------|
| Intercept | 9.6867 |
| Age | −0.0797 |



- a) Evaluate the success of the regression.
- b) Predict the lifespan of an 18-year-old Hispanic woman.
- c) Are you satisfied that your model could predict the life expectancy of a friend of yours?

- T 30. Tree growth** A 1996 study examined the growth of grapefruit trees in Texas, determining the average trunk *Diameter* (in inches) for trees of varying *Ages*:

| Age (yr) | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|------|------|
| Diameter (in.) | 2.1 | 3.9 | 5.2 | 6.2 | 6.9 | 7.6 | 8.3 | 9.1 | 10.0 | 11.4 |

- a) Fit a linear model to these data. What concerns do you have about the model?
- b) If data had been given for individual trees instead of averages, would you expect the fit to be stronger, less strong, or about the same? Explain.

Just Checking ANSWERS

- Counts are often best transformed by using the square root.
- None. The relationship is already straight.
- Even though, technically, the population values are counts, you should probably try a stronger transformation like log(population) because populations grow in proportion to their size.

Review of part

Exploring Relationships Between Variables

Quick Review

You have now survived your second major unit of Statistics. Here's a brief summary of the key concepts and skills:

- We treat data two ways: as categorical and as quantitative.
- To explore relationships in categorical data, check out Chapter 2.
- To explore relationships in quantitative data:
 - Make a picture. Use a scatterplot. Put the explanatory variable on the x -axis and the response variable on the y -axis.
 - Describe the association between two quantitative variables in terms of direction, form, and strength.
 - The amount of scatter determines the strength of the association.
 - If, as one variable increases so does the other, the association is positive. If one increases as the other decreases, it's negative.
 - If the form of the association is linear, calculate a correlation to measure its strength numerically, and do a regression analysis to model it.
 - Correlations closer to -1 or $+1$ indicate stronger linear associations. Correlations near 0 indicate weak linear relationships, but other forms of association may still be present.
 - The line of best fit is also called the least squares regression line because it minimizes the sum of the squared residuals.
- The regression line predicts values of the response variable from values of the explanatory variable.
- A residual is the difference between the true value of the response variable and the value predicted by the regression model.
- The slope of the line is a rate of change, best described in “ y -units” per “ x -unit.”
- R^2 gives the fraction of the variation in the response variable that is accounted for by the model.
- The standard deviation of the residuals measures the amount of scatter around the line.
- Outliers and influential points can distort any of our models.
- If you see a pattern (a curve) in the residuals plot, your chosen model is not appropriate; use a different model. You may, for example, straighten the relationship by re-expressing one of the variables.
- To straighten bent relationships, re-express the data using logarithms or a power (squares, square roots, reciprocals, etc.).
- Always remember that an association is not necessarily an indication that one of the variables causes the other.

Need more help with some of this? Try rereading some sections of Chapters 6 through 9. Starting right here on this very page are more opportunities to review these concepts and skills.

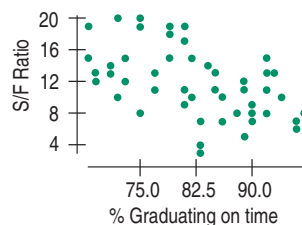
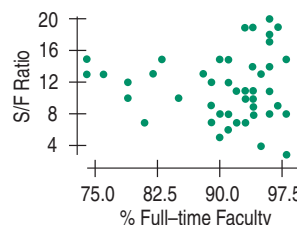
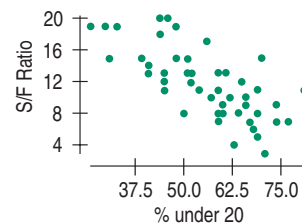
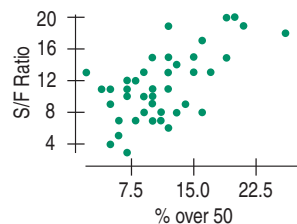
“One must learn by doing the thing; though you think you know it, you have no certainty until you try.”

—Sophocles (495–406 BCE)

Review Exercises

1. **College** Every year *US News and World Report* publishes a special issue on many U.S. colleges and universities. The scatterplots below have *Student/Faculty Ratio* (number of students per faculty member) for the colleges and universities on the y -axes plotted against 4 other variables. The correct correlations for these scatterplots appear in this list. Match them.

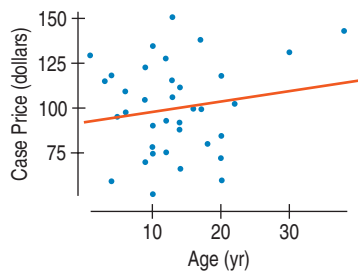
–0.98 –0.71 –0.51 0.09 0.23 0.69



2. Togetherness Are good grades in high school associated with family togetherness? A random sample of 142 high school students was asked how many meals per week their families ate together. Their responses produced a mean of 3.78 meals per week, with a standard deviation of 2.2. Researchers then matched these responses against the students' grade point averages (GPAs). The scatterplot appeared to be reasonably linear, so they created a line of regression. No apparent pattern emerged in the residuals plot. The equation of the line was $\widehat{GPA} = 2.73 + 0.11 \text{ Meals}$.

- a) Interpret the y-intercept in this context.
- b) Interpret the slope in this context.
- c) What was the mean GPA for these students?
- d) If a student in this study had a negative residual, what did that mean?
- e) Upon hearing of this study, a counselor recommended that parents who want to improve the grades their children get should get the family to eat together more often. Do you agree with this interpretation? Explain.

3. Vineyards Here are the scatterplot and regression analysis for *Case Prices* of 36 wines from vineyards in the Finger Lakes region of New York State and the *Ages* of the vineyards.

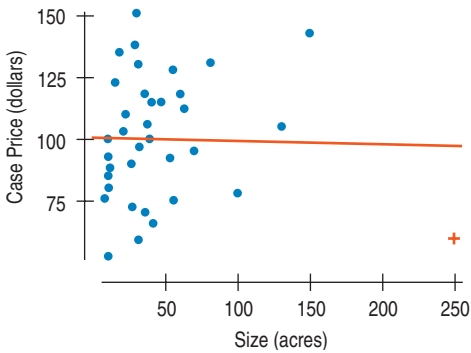


Dependent variable is: Case Price
R-squared = 2.7%

| Variable | Coefficient |
|----------|-------------|
| Constant | 92.7650 |
| Age | 0.567284 |

- a) Does it appear that vineyards in business longer get higher prices for their wines? Explain.
- b) What does this analysis tell us about vineyards in the rest of the world?
- c) Write the regression equation.
- d) Explain why that equation is essentially useless.

4. Vineyards again Instead of *Age*, perhaps the *Size* of the vineyard (in acres) is associated with the price of the wines. Look at the scatterplot:



- a) Do you see any evidence of an association?
- b) What concern do you have about this scatterplot?
- c) If the red "+" data point is removed, would the correlation become stronger or weaker? Explain.
- d) If the red "+" data point is removed, would the slope of the line increase or decrease? Explain.

5. More twins 2009? As the table shows, the number of twins born in the United States has been increasing. (www.cdc.gov/nchs/births.htm)

| Year | Twin Births | Year | Twin Births |
|------|-------------|------|-------------|
| 1980 | 68,339 | 1995 | 96,736 |
| 1981 | 70,049 | 1996 | 100,750 |
| 1982 | 71,631 | 1997 | 104,137 |
| 1983 | 72,287 | 1998 | 110,670 |
| 1984 | 72,949 | 1999 | 114,307 |
| 1985 | 77,102 | 2000 | 118,916 |
| 1986 | 79,485 | 2001 | 121,246 |
| 1987 | 81,778 | 2002 | 125,134 |
| 1988 | 85,315 | 2003 | 128,665 |
| 1989 | 90,118 | 2004 | 132,219 |
| 1990 | 93,865 | 2005 | 133,122 |
| 1991 | 94,779 | 2006 | 137,085 |
| 1992 | 95,372 | 2007 | 138,961 |
| 1993 | 96,445 | 2008 | 138,660 |
| 1994 | 97,064 | 2009 | 137,217 |

- a) Find the equation of the regression line for predicting the number of twin births.
- b) Explain in this context what the slope means.
- c) Predict the number of twin births in the United States for the year 2014. Comment on your faith in that prediction.
- d) Comment on the residuals plot.

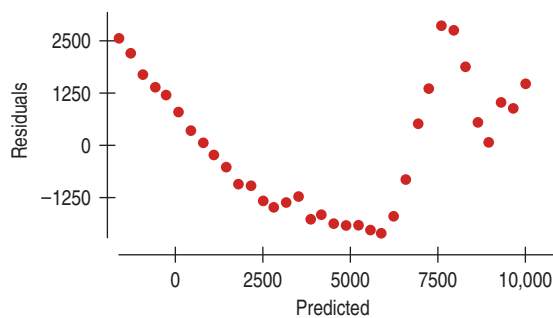
6. Dow Jones 2012 The Dow Jones stock index measures the performance of the stocks of America's largest

companies (finance.yahoo.com). A regression of the Dow prices on years 1972–2012 looks like this:

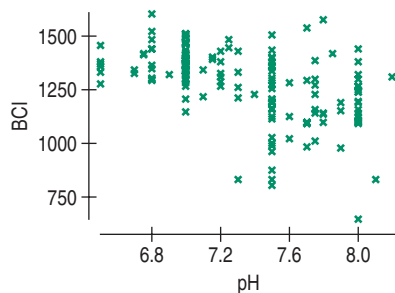
Dependent variable is Dow Index
R-squared = 83.9% $s = 1659$

| Variable | Coefficient |
|-----------|-------------|
| Intercept | -667396 |
| Year | 337.605 |

- What is the correlation between *Dow Index* and *Year*?
- Write the regression equation.
- Explain in this context what the equation says.
- Here's a scatterplot of the residuals. Which assumption(s) of the regression analysis appear to be violated?



- 7. Acid rain** Biologists studying the effects of acid rain on wildlife collected data from 163 streams in the Adirondack Mountains. They recorded the *pH* (acidity) of the water and the *BCI*, a measure of biological diversity, and they calculated $R^2 = 27\%$. Here's a scatterplot of *BCI* against *pH*:



- What is the correlation between *pH* and *BCI*?
- Describe the association between these two variables.
- If a stream has average *pH*, what would you predict about the *BCI*?
- In a stream where the *pH* is 3 standard deviations above average, what would you predict about the *BCI*?

- 8. Manatees 2010** Marine biologists warn that the growing number of powerboats registered in Florida threatens the existence of manatees. The data in the table come from the Florida Fish and Wildlife Conservation Commission (myfwc.com/research/manatee/) and the National Marine Manufacturers Association (www.nmma.org/).

| Year | Manatees Killed | Power Registrations (in 1000s) | Year | Manatees Killed | Power Registrations (in 1000s) |
|------|-----------------|--------------------------------|------|-----------------|--------------------------------|
| 1982 | 13 | 447 | 1997 | 53 | 716 |
| 1983 | 21 | 460 | 1998 | 38 | 716 |
| 1984 | 24 | 481 | 1999 | 35 | 716 |
| 1985 | 16 | 498 | 2000 | 49 | 735 |
| 1986 | 24 | 513 | 2001 | 81 | 860 |
| 1987 | 20 | 512 | 2002 | 95 | 923 |
| 1988 | 15 | 527 | 2003 | 73 | 940 |
| 1989 | 34 | 559 | 2004 | 69 | 946 |
| 1990 | 33 | 585 | 2005 | 79 | 974 |
| 1992 | 33 | 614 | 2006 | 92 | 988 |
| 1993 | 39 | 646 | 2007 | 73 | 992 |
| 1994 | 43 | 675 | 2008 | 90 | 932 |
| 1995 | 50 | 711 | 2009 | 97 | 949 |
| 1996 | 47 | 719 | 2010 | 83 | 914 |

- In this context, which is the explanatory variable?
- Make a scatterplot of these data and describe the association you see.
- Find the correlation between *Boat Registrations* and *Manatee Deaths*.
- Interpret the value of R^2 .
- Does your analysis prove that powerboats are killing manatees?

- 9. A manatee model 2010** Continue your analysis of the manatee situation from the previous exercise.

- Create a linear model of the association between *Manatee Deaths* and *Powerboat Registrations*.
- Interpret the slope of your model.
- Interpret the y-intercept of your model.
- How accurately did your model predict the high number of manatee deaths in 2010?
- Which is better for the manatees, positive residuals or negative residuals? Explain.
- What does your model suggest about the future for the manatee?

- 10. Grades** A Statistics instructor created a linear regression equation to predict students' final exam scores from their midterm exam scores. The regression equation was $\widehat{Fin} = 10 + 0.9 \widehat{Mid}$.

- If Susan scored a 70 on the midterm, what did the instructor predict for her score on the final?
- Susan got an 80 on the final. How big is her residual?
- If the standard deviation of the final was 12 points and the standard deviation of the midterm was 10 points, what is the correlation between the two tests?
- How many points would someone need to score on the midterm to have a predicted final score of 100?
- Suppose someone scored 100 on the final. Explain why you can't estimate this student's midterm score from the information given.

- f) One of the students in the class scored 100 on the midterm but got overconfident, slacked off, and scored only 15 on the final exam. What is the residual for this student?
- g) No other student in the class “achieved” such a dramatic turnaround. If the instructor decides not to include this student’s scores when constructing a new regression model, will the R^2 value of the regression increase, decrease, or remain the same? Explain.
- h) Will the slope of the new line increase or decrease?

11. Traffic Highway planners investigated the relationship between traffic *Density* (number of automobiles per mile) and the average *Speed* of the traffic on a moderately large city thoroughfare. The data were collected at the same location at 10 different times over a span of 3 months. They found a mean traffic *Density* of 68.6 cars per mile (cpm) with standard deviation of 27.07 cpm. Overall, the cars’ average *Speed* was 26.38 mph, with standard deviation of 9.68 mph. These researchers found the regression line for these data to be $\widehat{Speed} = 50.55 - 0.352 \text{ Density}$.

- a) What is the value of the correlation coefficient between *Speed* and *Density*?
- b) What percent of the variation in average *Speed* is explained by traffic *Density*?
- c) Predict the average *Speed* of traffic on the thoroughfare when the traffic *Density* is 50 cpm.
- d) What is the value of the residual for a traffic *Density* of 56 cpm with an observed *Speed* of 32.5 mph?
- e) The data set initially included the point *Density* = 125 cpm, *Speed* = 55 mph. This point was considered an outlier and was not included in the analysis. Will the slope increase, decrease, or remain the same if we redo the analysis and include this point?
- f) Will the correlation become stronger, weaker, or remain the same if we redo the analysis and include this point (125,55)?
- g) A European member of the research team measured the *Speed* of the cars in kilometers per hour (1 km \approx 0.62 miles) and the traffic *Density* in cars per kilometer. Find the value of his calculated correlation between speed and density.

T 12. Cramming One Thursday, researchers gave students enrolled in a section of basic Spanish a set of 50 new vocabulary words to memorize. On Friday the students took a vocabulary test. When they returned to class the following Monday, they were retested—without advance warning. Here are the test scores for the 25 students.

| Fri. | Mon. | Fri. | Mon. | Fri. | Mon. |
|------|------|------|------|------|------|
| 42 | 36 | 48 | 37 | 39 | 41 |
| 44 | 44 | 43 | 41 | 46 | 32 |
| 45 | 46 | 45 | 32 | 37 | 36 |
| 48 | 38 | 47 | 44 | 40 | 31 |
| 44 | 40 | 50 | 47 | 41 | 32 |
| 43 | 38 | 34 | 34 | 48 | 39 |
| 41 | 37 | 38 | 31 | 37 | 31 |
| 35 | 31 | 43 | 40 | 36 | 41 |
| 43 | 32 | | | | |

- a) What is the correlation between *Friday* and *Monday* scores?
- b) What does a scatterplot show about the association between the scores?
- c) What does it mean for a student to have a positive residual?
- d) What would you predict about a student whose *Friday* score was one standard deviation below average?
- e) Write the equation of the regression line.
- f) Predict the *Monday* score of a student who earned a 40 on Friday.

13. Correlations What factor most explains differences in *Fuel Efficiency* among cars? Below is a correlation matrix exploring that relationship for the car’s *Weight*, *Horsepower*, engine size (*Displacement*), and number of *Cylinders*.

| | MPG | Weight | Horsepower | Displacement | Cylinders |
|--------------|--------|--------|------------|--------------|-----------|
| MPG | 1.000 | | | | |
| Weight | −0.903 | 1.000 | | | |
| Horsepower | −0.871 | 0.917 | 1.000 | | |
| Displacement | −0.786 | 0.951 | 0.872 | 1.000 | |
| Cylinders | −0.806 | 0.917 | 0.864 | 0.940 | 1.000 |

- a) Which factor seems most strongly associated with *Fuel Efficiency*?
- b) What does the negative correlation indicate?
- c) Explain the meaning of R^2 for that relationship.

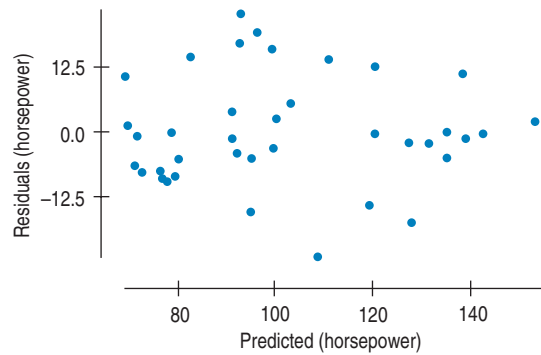
T 14. Autos revisited Look again at the correlation table for cars in the previous exercise.

- a) Which two variables in the table exhibit the strongest association?
- b) Is that strong association necessarily cause-and-effect? Offer at least two explanations why that association might be so strong.
- c) Engine displacements for U.S.-made cars are often measured in cubic inches. For many foreign cars, the units are either cubic centimeters or liters. How would changing from cubic inches to liters affect the calculated correlations involving *Displacement*?
- d) What would you predict about the *Fuel Efficiency* of a car whose engine *Displacement* is one standard deviation above the mean?

T 15. Cars, one more time! Can we predict the *Horsepower* of the engine that manufacturers will put in a car by knowing the *Weight* of the car? Here are the regression analysis and residuals plot:

Dependent variable is: Horsepower
R-squared = 84.1%

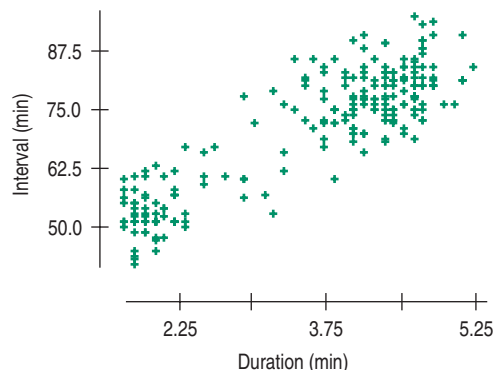
| Variable | Coefficient |
|-----------|-------------|
| Intercept | 3.49834 |
| Weight | 34.3144 |



- Write the equation of the regression line.
- Do you think the car's *Weight* is measured in pounds or thousands of pounds? Explain.
- Do you think this linear model is appropriate? Explain.
- The highest point in the residuals plot, representing a residual of 22.5 horsepower, is for a Chevy weighing 2595 pounds. How much horsepower does this car have?

16. Colorblind Although some women are colorblind, this condition is found primarily in men. Why is it wrong to say there's a strong correlation between *Sex* and *Colorblindness*?

- T 17. Old Faithful** There is evidence that eruptions of Old Faithful can best be predicted by knowing the duration of the previous eruption.
- Describe what you see in the scatterplot of *Intervals* between eruptions vs. *Duration* of the previous eruption.



- Write the equation of the line of best fit. Here's the regression analysis:

Dependent variable is: Interval
R-squared = 77.0%

| Variable | Coefficient |
|-----------|-------------|
| Intercept | 33.9668 |
| Duration | 10.3582 |

- Carefully explain what the slope of the line means in this context.
- How accurate do you expect predictions based on this model to be? Cite statistical evidence.

- If you just witnessed an eruption that lasted 4 minutes, how long do you predict you'll have to wait to see the next eruption?
- So you waited, and the next eruption came in 79 minutes. Use this as an example to define a residual.

- T 18. Which croc?** The ranges inhabited by the Indian gharial crocodile and the Australian saltwater crocodile overlap in Bangladesh. Suppose a very large crocodile skeleton is found there, and we wish to determine the species of the animal. Wildlife scientists have measured the lengths of the heads and the complete bodies of several crocs (in centimeters) of each species, creating the regression analyses below:

Indian Crocodile

Dependent variable is: IBody
R-squared = 97.2%

| Variable | Coefficient |
|-----------|-------------|
| Intercept | -69.3693 |
| IHead | 7.40004 |

Australian Crocodile

Dependent variable is: ABody
R-squared = 98.0%

| Variable | Coefficient |
|-----------|-------------|
| Intercept | -20.2245 |
| AHead | 7.71726 |

- Do the associations between the sizes of the heads and bodies of the two species appear to be strong? Explain.
- In what ways are the two relationships similar? Explain.
- What is different about the two models? What does that mean?
- The crocodile skeleton found had a head length of 62 cm and a body length of 380 cm. Which species do you think it was? Explain why.

- T 19. How old is that tree?** One can determine how old a tree is by counting its rings, but that requires cutting the tree down. Can we estimate the tree's age simply from its diameter? A forester measured 27 trees of the same species that had been cut down, and counted the rings to determine the ages of the trees.

| Diameter (in.) | Age (yr) | Diameter (in.) | Age (yr) |
|----------------|----------|----------------|----------|
| 1.8 | 4 | 10.3 | 23 |
| 1.8 | 5 | 14.3 | 25 |
| 2.2 | 8 | 13.2 | 28 |
| 4.4 | 8 | 9.9 | 29 |
| 6.6 | 8 | 13.2 | 30 |
| 4.4 | 10 | 15.4 | 30 |
| 7.7 | 10 | 17.6 | 33 |
| 10.8 | 12 | 14.3 | 34 |
| 7.7 | 13 | 15.4 | 35 |
| 5.5 | 14 | 11.0 | 38 |
| 9.9 | 16 | 15.4 | 38 |
| 10.1 | 18 | 16.5 | 40 |
| 12.1 | 20 | 16.5 | 42 |
| 12.8 | 22 | | |

- Find the correlation between *Diameter* and *Age*. Does this suggest that a linear model may be appropriate? Explain.

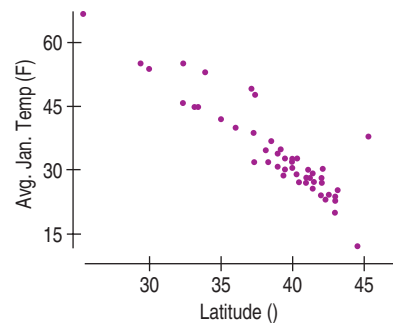
- b) Create a scatterplot and describe the association.
- c) Create the linear model.
- d) Check the residuals. Explain why a linear model is probably not appropriate.
- e) If you used this model, would it generally overestimate or underestimate the ages of very large trees? Explain.

- T 20. Improving trees** In the last exercise you saw that the linear model had some deficiencies. Let's create a better model.
- a) Perhaps the cross-sectional area of a tree would be a better predictor of its age. Since area is measured in square units, try re-expressing the data by squaring the diameters. Does the scatterplot look better?
 - b) Create a model that predicts *Age* from the square of the *Diameter*.
 - c) Check the residuals plot for this new model. Is this model more appropriate? Why?
 - d) Estimate the age of a tree 18 inches in diameter.
- 21. New homes** A real estate agent collects data to develop a model that will use the *Size* of a new home (in square feet) to predict its *Sale Price* (in thousands of dollars). Which of these is most likely to be the slope of the regression line: 0.008, 0.08, 0.8, or 8? Explain.
- T 22. Smoking and pregnancy 2006** The Child Trends Data Bank monitors issues related to children. The table shows a 50-state average of the percent of expectant mothers who smoked cigarettes during their pregnancies.

| Year | % Smoking While Pregnant | Year | % Smoking While Pregnant |
|------|--------------------------|------|--------------------------|
| 1990 | 19.2 | 1999 | 14.1 |
| 1991 | 18.7 | 2000 | 14.0 |
| 1992 | 17.9 | 2001 | 13.8 |
| 1993 | 16.8 | 2002 | 13.3 |
| 1994 | 16.0 | 2003 | 12.7 |
| 1995 | 15.4 | 2004 | 10.9 |
| 1996 | 15.3 | 2005 | 10.1 |
| 1997 | 14.9 | 2006 | 10.0 |
| 1998 | 14.8 | | |

- a) Create a scatterplot and describe the trend you see.
 - b) Find the correlation.
 - c) How is the value of the correlation affected by the fact that the data are averages rather than percentages for each of the 50 states?
 - d) Write a linear model and interpret the slope in context.
- T 23. No smoking?** The downward trend in smoking you saw in the last exercise is good news for the health of babies, but will it ever stop?
- a) Explain why you can't use the linear model you created in Exercise 22 to see when smoking during pregnancy will cease altogether.
 - b) Create a model that could estimate the year in which the level of smoking would be 0%.
 - c) Comment on the reliability of such a prediction.

- 24. Tips** It's commonly believed that people use tips to reward good service. A researcher for the hospitality industry examined tips and ratings of service quality from 2645 dining parties at 21 different restaurants. The correlation between ratings of service and tip percentages was 0.11. (M. Lynn and M. McCall, "Gratitude and Gratitude." *Journal of Socio-Economics* 29: 203–214)
- a) Describe the relationship between *Quality of Service* and *Tip Size*.
 - b) Find and interpret the value of R^2 in this context.
- 25. US cities** Data from 50 large U.S. cities show the mean *January Temperature* and the *Latitude*. Describe what you see in the scatterplot.



- 26. Correlations** The study of U.S. cities in Exercise 25 found the mean *January Temperature* (degrees Fahrenheit), *Altitude* (feet above sea level), and *Latitude* (degrees north of the equator) for 55 cities. Here's the correlation matrix:
- | | Jan. Temp | Latitude | Altitude |
|-----------|-----------|----------|----------|
| Jan. Temp | 1.000 | | |
| Latitude | -0.848 | 1.000 | |
| Altitude | -0.369 | 0.184 | 1.000 |
- a) Which seems to be more useful in predicting *January Temperature*—*Altitude* or *Latitude*? Explain.
 - b) If the *Temperature* were measured in degrees Celsius, what would be the correlation between *Temperature* and *Latitude*?
 - c) If the *Temperature* were measured in degrees Celsius and the *Altitude* in meters, what would be the correlation? Explain.
 - d) What would you predict about the *January Temperatures* in a city whose *Altitude* is two standard deviations higher than the average *Altitude*?
- 27. Winter in the city** Summary statistics for the data relating the latitude and average January temperature for 55 large U.S. cities are given below.

| Variable | Mean | StdDev |
|-------------|----------|--------|
| Latitude | 39.02 | 5.42 |
| JanTemp | 26.44 | 13.49 |
| Correlation | = -0.848 | |

- What percent of the variation in January *Temperatures* can be explained by variation in *Latitude*?
- What is indicated by the fact that the correlation is negative?
- Write the equation of the line of regression for predicting January *Temperature* from *Latitude*.
- Explain what the slope of the line means.
- Do you think the y -intercept is meaningful? Explain.
- The latitude of Denver is 40° N. Predict the mean January temperature there.
- What does it mean if the residual for a city is positive?

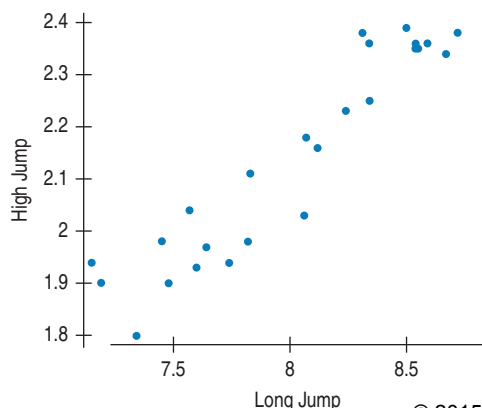
28. Depression The September 1998 issue of the *American Psychologist* published an article by Kraut et al. that reported on an experiment examining “the social and psychological impact of the Internet on 169 people in 73 households during their first 1 to 2 years online.” In the experiment, 73 households were offered free Internet access for 1 or 2 years in return for allowing their time and activity online to be tracked. The members of the households who participated in the study were also given a battery of tests at the beginning and again at the end of the study. The conclusion of the study made news headlines: Those who spent more time online tended to be more depressed at the end of the experiment. Although the paper reports a more complex model, the basic result can be summarized in the following regression of *Depression* (at the end of the study, in “depression scale units”) vs. *Internet Use* (in mean hours per week):

Dependent variable is: Depression
 R-squared = 4.6%
 $s = 0.4563$

| Variable | Coefficient |
|--------------|-------------|
| Intercept | 0.5655 |
| Internet use | 0.0199 |

The news reports about this study clearly concluded that using the Internet causes depression. Discuss whether such a conclusion can be drawn from this regression. If so, discuss the supporting evidence. If not, say why not.

- T 29. Jumps 2012** How are Olympic performances in various events related? The plot shows winning long-jump and high-jump distances, in meters, for the Summer Olympics from 1912 through 2012.



- Describe the association.
- Do long-jump performances somehow influence the high-jumpers? How do you account for the relationship you see?
- The correlation for the plotted data is 0.913. If we converted the jump lengths to centimeters by multiplying by 100, would that make the actual correlation higher or lower?
- What would you predict about the long jump in a year when the high-jumper jumped one standard deviation better than the average high jump?

- T 30. Modeling jumps 2012** Here are the summary statistics for the Olympic long jumps and high jumps displayed in the previous exercise.

| Event | Mean | StdDev |
|-----------|-------|--------|
| High Jump | 2.148 | 0.1939 |
| Long Jump | 8.05 | 0.5136 |

Correlation = 0.9125

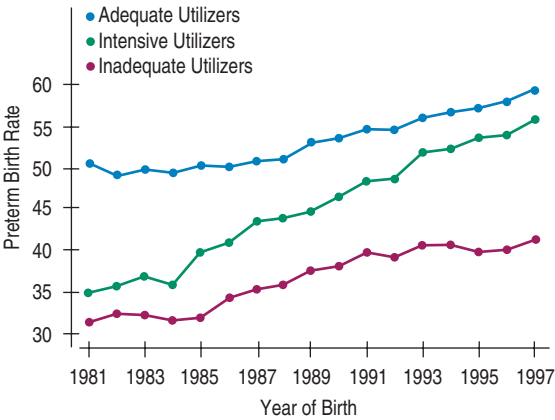
- Write the equation of the line of regression for estimating *High Jump* from *Long Jump*.
- Interpret the slope of the line.
- In a year when the long jump is 8.9 m, what high jump would you predict?
- Why can't you use this line to estimate the long jump for a year when you know the high jump was 2.25 m?
- Write the equation of the line you need to make that prediction.

- 31. French** Consider the association between a student's score on a French vocabulary test and the weight of the student. What direction and strength of correlation would you expect in each of the following situations? Explain.

- The students are all in third grade.
- The students are in third through twelfth grades in the same school district.
- The students are in tenth grade in France.
- The students are in third through twelfth grades in France.

- 32. Twins** Twins are often born after a pregnancy that lasts less than 9 months. On the next page is a graph from the *Journal of the American Medical Association (JAMA)* showing the rate of preterm twin births in the United States over the past 20 years. In this study, *JAMA* categorized mothers by the level of prenatal medical care they received: inadequate, adequate, or intensive.

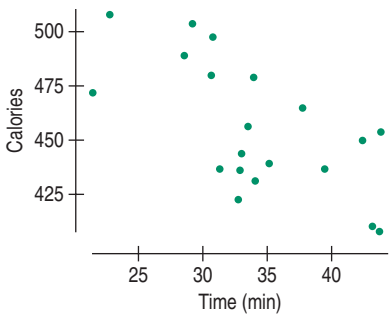
- Describe the overall trend in preterm twin births.
- Describe any differences you see in this trend, depending on the level of prenatal medical care the mother received.
- Should expectant mothers be advised to cut back on the level of medical care they seek in the hope of avoiding preterm births? Explain.



Preterm Birth Rate per 100 live twin births among U.S. twins by intensive, adequate, and less than adequate prenatal care utilization, 1981–1997. (JAMA 284[2000]: 335–341)

T 33. Lunchtime Create and interpret a model for the toddlers’ lunchtime data presented below. The table and graph show the number of minutes the kids stayed at the table and the number of calories they consumed.

| Calories | Time | Calories | Time |
|----------|------|----------|------|
| 472 | 21.4 | 450 | 42.4 |
| 498 | 30.8 | 410 | 43.1 |
| 465 | 37.7 | 504 | 29.2 |
| 456 | 33.5 | 437 | 31.3 |
| 423 | 32.8 | 489 | 28.6 |
| 437 | 39.5 | 436 | 32.9 |
| 508 | 22.8 | 480 | 30.6 |
| 431 | 34.1 | 439 | 35.1 |
| 479 | 33.9 | 444 | 33.0 |
| 454 | 43.8 | 408 | 43.7 |



34. Gasoline Since clean-air regulations have dictated the use of unleaded gasoline, the supply of leaded gas in New York state has diminished. The table below was given on the August 2001 New York State Math B exam, a statewide achievement test for high school students.

| Year | 1984 | 1988 | 1992 | 1996 | 2000 |
|------------------|------|------|------|------|------|
| Gallons (1000's) | 150 | 124 | 104 | 76 | 50 |

- a) Create a linear model and predict the number of gallons that will be available in 2005.
- b) The exam then asked students to estimate the year when leaded gasoline will first become unavailable, expecting them to use the model from part a to answer the question. Explain why that method is incorrect.
- c) Create a model that *would* be appropriate for that task, and make the estimate.
- d) The “wrong” answer from the other model is fairly accurate in this case. *Why?*

T 35. Tobacco and alcohol Are people who use tobacco products more likely to consume alcohol? Here are data on household spending (in pounds) taken by the British Government on 11 regions in Great Britain. Do tobacco and alcohol spending appear to be related? What questions do you have about these data? What conclusions can you draw?

| Region | Alcohol | Tobacco |
|------------------|---------|---------|
| North | 6.47 | 4.03 |
| Yorkshire | 6.13 | 3.76 |
| Northeast | 6.19 | 3.77 |
| East Midlands | 4.89 | 3.34 |
| West Midlands | 5.63 | 3.47 |
| East Anglia | 4.52 | 2.92 |
| Southeast | 5.89 | 3.20 |
| Southwest | 4.79 | 2.71 |
| Wales | 5.27 | 3.53 |
| Scotland | 6.08 | 4.51 |
| Northern Ireland | 4.02 | 4.56 |

T 36. Football weights The Sears Cup was established in 1993 to honor institutions that maintain a broad-based athletic program, achieving success in many sports, both men’s and women’s. Since its Division III inception in 1995, the cup has been won by Williams College in every year except one. Their football team has a 85.3% winning record under their current coach. Why does the football team win so much? Is it because they’re heavier than their opponents? The table shows the average team weights for selected years from 1973 to 1993.

| Year | Weight (lb) | Year | Weight (lb) |
|------|-------------|------|-------------|
| 1973 | 185.5 | 1983 | 192.0 |
| 1975 | 182.4 | 1987 | 196.9 |
| 1977 | 182.1 | 1989 | 202.9 |
| 1979 | 191.1 | 1991 | 206.0 |
| 1981 | 189.4 | 1993 | 198.7 |

- a) Fit a straight line to the relationship between *Weight* and *Year*.
- b) Does a straight line seem reasonable?

- c) Predict the average weight of the team for the year 2003. Does this seem reasonable?
 d) What about the prediction for the year 2103? Explain.
 e) What about the prediction for the year 3003? Explain.

37. Models Find the predicted value of y , using each model for $x = 10$.

a) $\hat{y} = 2 + 0.8 \ln x$ b) $\log \hat{y} = 5 - 0.23x$

c) $\frac{1}{\sqrt{\hat{y}}} = 17.1 - 1.66x$

- T 38. Williams vs Texas** Here are the average weights of the football team for the University of Texas for various years in the 20th century.

| Year | 1905 | 1919 | 1932 | 1945 | 1955 | 1965 |
|-------------|------|------|------|------|------|------|
| Weight (lb) | 164 | 163 | 181 | 192 | 195 | 199 |

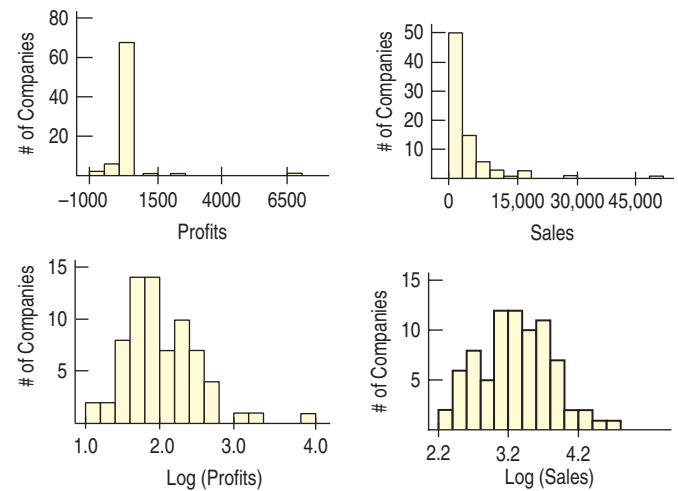
- a) Fit a straight line to the relationship of *Weight* by *Year* for Texas football players.
 b) According to these models, in what year will the predicted weight of the Williams College team from Exercise 36 first be more than the weight of the University of Texas team?
 c) Do you believe this? Explain.

39. Vehicle weights The Minnesota Department of Transportation hoped that they could measure the weights of big trucks without actually stopping the vehicles by using a newly developed “weigh-in-motion” scale. After installation of the scale, a study was conducted to find out whether the scale’s readings correspond to the true weights of the trucks being monitored. In Exercise 46 of Chapter 6, you examined the scatterplot for the data they collected, finding the association to be approximately linear with $R^2 = 93\%$. Their regression equation is $\widehat{Wt} = 10.85 + 0.64 \text{ Scale}$, where both the scale reading and the predicted weight of the truck are measured in thousands of pounds.

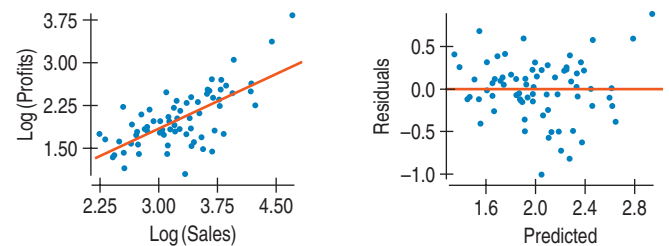
- a) Estimate the weight of a truck if this scale read 31,200 pounds.
 b) If that truck actually weighed 32,120 pounds, what was the residual?
 c) If the scale reads 35,590 pounds, and the truck has a residual of -2440 pounds, how much does it actually weigh?
 d) In general, do you expect estimates made using this equation to be reasonably accurate? Explain.
 e) If the police plan to use this scale to issue tickets to trucks that appear to be overloaded, will negative or positive residuals be a greater problem? Explain.

40. Profit How are a company’s profits related to its sales? Let’s examine data from 71 large U.S. corporations. All amounts are in millions of dollars.

- a) Histograms of *Profits* and *Sales* and histograms of the logarithms of *Profits* and *Sales* appear below. Why are the re-expressed data better for regression?



- b) Here are the scatterplot and residuals plot for the regression of logarithm of *Profits* vs. $\text{Log}(\text{Sales})$. Do you think this model is appropriate? Explain.



- c) Here’s the regression analysis. Write the equation.

Dependent variable is: Log Profit
 R-squared = 48.1%

| Variable | Coefficient |
|-----------|-------------|
| Intercept | -0.106259 |
| LogSales | 0.647798 |

- d) Use your equation to estimate profits earned by a company with sales of 2.5 billion dollars. (That’s 2500 million.)

- T 41. Down the drain** Most water tanks have a drain plug so that the tank may be emptied when it’s to be moved or repaired. How long it takes a certain size of tank to drain depends on the size of the plug, as shown in the table. Create a model.

| Plug Dia (in.) | $\frac{3}{8}$ | $\frac{1}{2}$ | $\frac{3}{4}$ | 1 | $1\frac{1}{4}$ | $1\frac{1}{2}$ | 2 |
|-------------------|---------------|---------------|---------------|----|----------------|----------------|---|
| Drain Time (min.) | 140 | 80 | 35 | 20 | 13 | 10 | 5 |

- 42. Chips** A start-up company has developed an improved electronic chip for use in laboratory equipment. The company needs to project the manufacturing cost, so it develops a spreadsheet model that takes into account the purchase of production equipment, overhead, raw materials, depreciation, maintenance, and other business costs. The spreadsheet estimates the cost of producing 10,000 to 200,000 chips per year, as seen in the table. Develop a regression model to predict *Costs* based on the *Level* of production.

| Chips Produced (1000s) | Cost per Chip (\$) | Chips Produced (1000s) | Cost per Chip (\$) |
|---------------------------|-----------------------|---------------------------|-----------------------|
| 10 | 146.10 | 90 | 47.22 |
| 20 | 105.80 | 100 | 44.31 |
| 30 | 85.75 | 120 | 42.88 |
| 40 | 77.02 | 140 | 39.05 |
| 50 | 66.10 | 160 | 37.47 |
| 60 | 63.92 | 180 | 35.09 |
| 70 | 58.80 | 200 | 34.04 |
| 80 | 50.91 | | |

Practice Exam

I. Multiple Choice

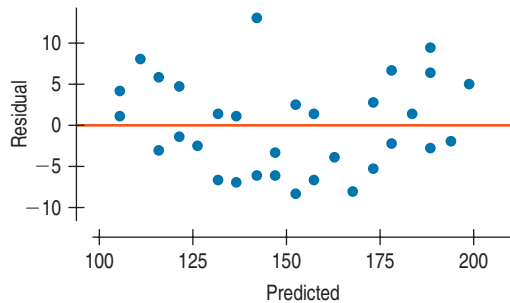
(Questions 1–3) Based on data collected over several sessions, a statistically minded trainer of office typists modeled the linear relationship between the number of hours of training a typist receives and the typist's speed (in words per minute) with the equation $\widehat{speed} = 10.6 + 5.4 \text{ hour}$.

- Which of these statements best interprets this equation?
 - Typists increase their speed by 10.6 wpm for every 5.4 hours of training.
 - Typists increase their speed by 5.4 wpm for every 10.6 hours of training.
 - A typist who trains for an additional hour will benefit with a speed increase of 5.4 wpm.
 - On average, typists tend to increase their speed by roughly 5.4 wpm for every hour of training.
 - For every 5.4 hours of training, typists can increase their speed from 10.6 wpm to faster.
- Which is the best interpretation of the y-intercept for this model?
 - People who can't type need about 10.6 hours of training.
 - Before undergoing this training, typists' average speed was about 10.6 words per minute.
 - The y-intercept is meaningless here because no one types at 0 wpm.
 - The y-intercept is meaningless here because none of the typists had 0 hours of training.
 - In regression models, the slope has meaning, but not the y-intercept.
- After some training, one of the typists was told that the speed he attained had a residual of 4.3 words per minute. How should he interpret this?
 - He types slower than the model predicted, given the amount of time he spent training.
 - He types faster than the model predicted, given the amount of time he spent training.
 - He can't interpret his residual without also knowing the correlation.

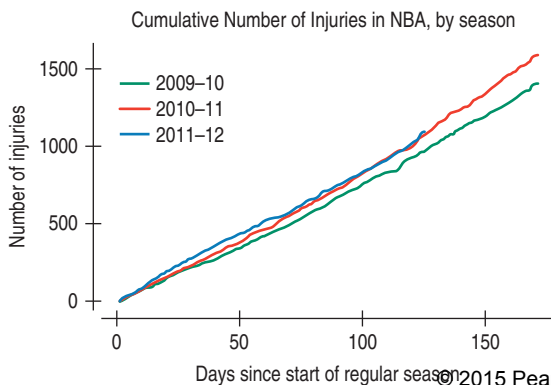
- He can't interpret his residual without also knowing the size of other people's residuals.
- He can't interpret his residual without also knowing the standard deviation of the residuals.

- The Bureau of Labor Statistics looked at the association between students' GPAs in high school ($\widehat{gpa_HS}$) and their freshmen GPAs at a University of California school ($\widehat{gpa_U}$). The resulting least-squares regression equation is $\widehat{gpa_U} = 0.22 + 0.72\widehat{gpa_HS}$. Calculate the residual for a student with a 3.8 in high school who achieved a freshman GPA of 3.5.
 - 0.844
 - 0.544
 - 2.956
 - 0.544
 - 0.844
- In April of 2012, the Centers for Disease Control and Prevention announced that birth rates for U.S. teenagers reached historic lows. From 2009 to 2010 the rate declined 9%, to a level of 34.3 births per 1000 women aged 15–19. Which of these conclusions is an example of extrapolation in this context?
 - There was a decreasing trend in teenage birth rates at the time of this study.
 - Time is an explanatory variable in the change of teenage birth rates.
 - By 2014, teenage birth rates will be 36% lower and set new records.
 - There is a linear relationship between year and teenage birth rates.
 - None of these is an example of extrapolation.
- An engineer studying the performance of a certain type of bolt predicts the failure rate (bolts per 1000) from the load (in pounds) using the model $\log(\widehat{fail}) = 1.04 + 0.0013\text{load}$. If these bolts are subjected to a load of 600 pounds, what failure rate should we expect?
 - 0.26
 - 0.60
 - 1.82
 - 6.17
 - 66.07

7. A researcher analyzing some data created a linear model with $R^2 = 94\%$ and having the residuals plot seen here. What should she conclude?



- A) The linear model is appropriate, because about the half the residuals are positive and half negative.
 B) The linear model is appropriate, because the value of R^2 is quite high.
 C) The linear model is not appropriate, because the value of R^2 is not high enough.
 D) The linear model is not appropriate, because the residuals plot shows curvature.
 E) The linear model is not appropriate, because the residuals plot identifies an outlier.
8. Researchers at UC San Francisco discovered that high plasma levels of vitamins B, C, D, and E are associated with better cognitive performance. "Each standard deviation higher plasma level for these vitamins predicted a global cognitive score 0.28 standard deviations better," the researchers reported. Which value are the researchers interpreting in this statement?
- A) the correlation coefficient between plasma level and cognitive score
 B) the y-intercept of the regression model predicting cognitive score from plasma level
 C) the slope of the regression model predicting cognitive score from plasma level
 D) the standard deviation of the regression model's residuals
 E) R^2 for the regression model
9. This graph shows the relationship the number of days since the NBA season began and the number of injuries, over the course of three different seasons. In 2011–12, the season was shortened by a labor strike.

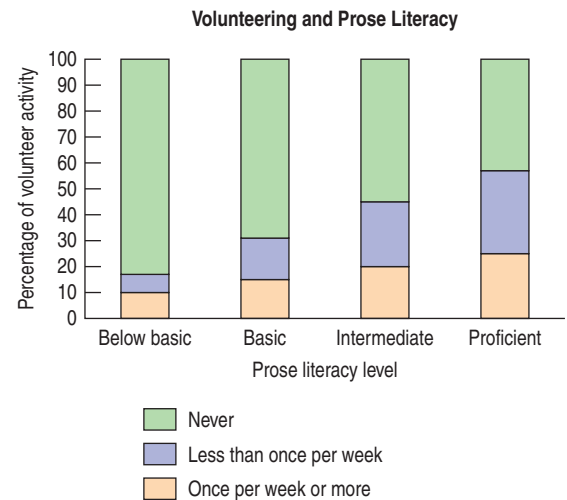


Of statements A–D, which of the following is NOT a correct conclusion that can be drawn from this graph?

- A) There is a fairly strong linear relationship between days since the start of the season and the number of injuries.
 B) At first the rate of injuries was higher during the strike-shortened season.
 C) As the strike-shortened season continued the number of injuries became similar to the other two seasons.
 D) While the strike-shortened season had more injuries initially, we cannot know for certain if the strike caused the difference or if it was attributable to other variables.
 E) All of A–D are correct.
10. Which of statements A–D is true?

- A) An influential point always has a large residual.
 B) An influential point changes the slope of the regression equation.
 C) An influential point decreases the value of R^2 .
 D) An influential point does not affect the y-intercept.
 E) Statements A–D are all false.

(Questions 11–12) The segmented bar charts below depict the data from the NAAL (National Assessment of Adult Literacy) conducted in 2003.



11. Which of the following is greatest?
- A) The number of people who volunteer once per week or more and test Below Basic on Prose Literacy.
 B) The number of people who volunteer less than once per week and test Basic on Prose Literacy.
 C) The number of people who never volunteer and test Proficient on Prose Literacy.
 D) The number of people who volunteer less than once per week and test Intermediate in Prose Literacy.
 E) It is impossible to determine which is greatest without knowing the actual number of people at each literacy level.
12. Based on the segmented bar graphs, does there appear to be an association between volunteerism and literacy level?
- A) Yes, all three bars have the same number of segments.
 B) Yes, because all three bars have the same height.

- C) Yes, because the corresponding segments of the three bars have different heights.
 D) No, because the corresponding segments of the three bars have different heights.
 E) No, because the sums of the 3 proportions in each bar are identical.
13. A TV weatherman's end-of-year analysis of the local weather showed that among all the years for which records had been kept, the past year's average temperature had a z -score of 2.8. What does that mean?
- A) The past year's average temperature was 2.8° higher than the historical mean.
 B) The past year's average temperature was 2.8 standard deviations above the historical mean.
 C) The past year's average temperature was 2.8% higher than the historical mean.
 D) The past year's temperatures had a standard deviation of 2.8° .
 E) The past year had 2.8 times as many days with above average temperatures as is typical for that area.
14. In Statsville there's a city-wide speed limit of 30 mph. If you are caught speeding the fine is \$100 plus \$10 for every mile per hour you were over the speed limit. For example, if you're ticketed for going 45 mph, your fine is $100 + 10(45 - 30) = \$250$. Last month all the drivers who were fined for speeding averaged 42 mph with a standard deviation of 7 mph. What were the mean and standard deviation of the fines?
- A) \$120 and \$70 B) \$220 and \$7
 C) \$220 and \$70 D) \$220 and \$170
 E) \$420 and \$70
15. Among those Statsville drivers fined for speeding, the fastest 10% were caught exceeding how many miles per hour?
- A) 37.0 B) 48.3 C) 51.0 D) 58.3
 E) It cannot be determined from the information given.
- a) Write the equation of the least square regression line.
 b) Interpret R^2 in this context.
 c) Interpret the equation in this context.
 d) This student's girlfriend tried out his model on a pencil she had used for 5 hours, and found a residual of -0.88 cm. How long was her pencil at that time?
 e) Should she have expected this model to describe the rate for her pencils? Why or why not?
2. Energy drinks come in different-sized packages: pouches, small bottles, large bottles, twin-packs, 6-packs, and so on. How is the price related to the amount of beverage? Data collected on a variety of packages revealed a mean size of 140.17 ounces with a standard deviation of 78.23 ounces. The packages had an average price of \$3.66 with a standard deviation of \$1.50, and the correlation between size and price was $r = 0.91$. A scatterplot of these data suggested that the assumptions needed for regression were reasonable.
- a) Interpret the value of r in context.
 b) Compute the slope of the least-squares regression line for predicting the price of an energy drink. Include the proper units in your answer.
 c) Write the equation for the least-squares regression line for these data.
 d) For this model the standard deviation of the residuals was $s = 0.26$. Explain what that means in context.
3. The Pew Research Center conducted two surveys, one in December 2011 and another in November 2012, asking people about their reading habits. Pew reported the percentage of people who read at least one e-book in the past year, given that they had read at least one book. Those percentages, broken down by age group, are shown in the table below.

| Date of Poll | Age Group | | | | |
|---------------|-----------|-------|-------|-------|-----|
| | 16–17 | 18–29 | 30–49 | 50–64 | 65+ |
| December 2011 | 13 | 25 | 25 | 19 | 12 |
| November 2012 | 28 | 31 | 41 | 23 | 20 |

II. Free Response

1. A diligent statistics student recorded the length of his faithful #2 pencil as he worked away on his homework. He discovered a strong linear relationship between the number of hours that he worked and the length of his pencil. Here is the regression analysis for these data.

Dependent variable: length (cm)

$$R^2 = 92.3\% \quad R^2(\text{adj}) = 89.5\%$$

| | coeff | se | t ratio | p value |
|-----------|--------|-------|---------|---------|
| constant | 17.047 | 0.128 | 23.58 | <0.0001 |
| time (hr) | -1.914 | 0.047 | 35.28 | <0.0001 |

- a) Create an appropriate graphical display that allows a comparison of responses between the two years and also among the different age groups.
 b) Write a few sentences comparing e-book readership in the two time periods.
 c) Is there an association between age and the growth in e-book readership? Use evidence from the table or your graph to justify your answer.