CHAPTER 10

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 that meters are better (or worse) than fathoms or leagues. What we're talking about is re-expressing the data another way by applying a function, such as a square root, log, or reciprocal. 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.

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

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.

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

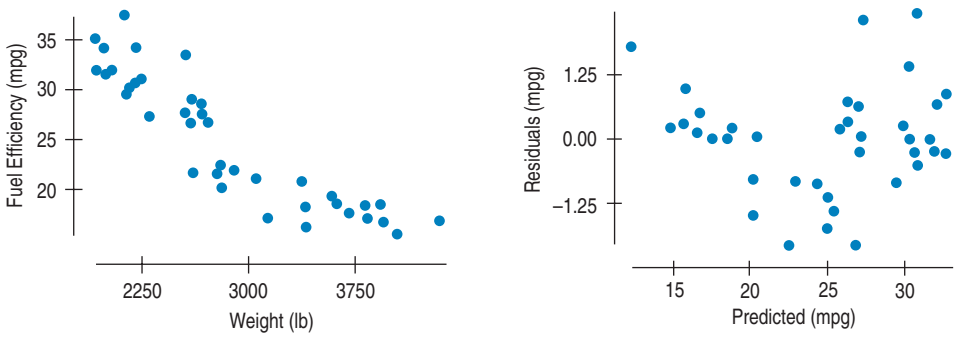


FIGURE 10.1

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.*

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]

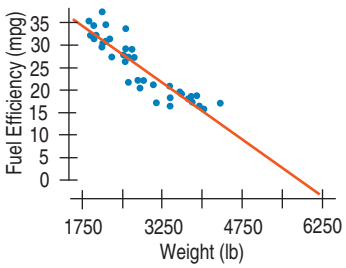


FIGURE 10.2

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

"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.

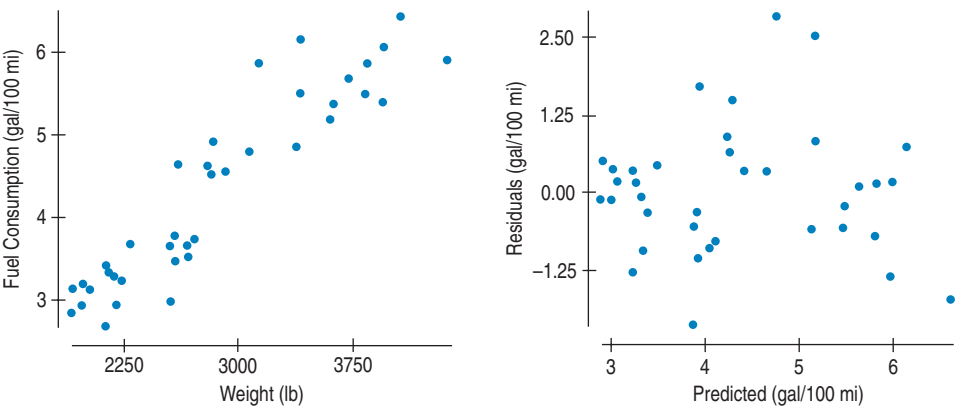


FIGURE 10.3

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.*

[1] Multiplying 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. Now 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 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,

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

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 *Forbes* magazine in reporting on the Forbes 500 for that year

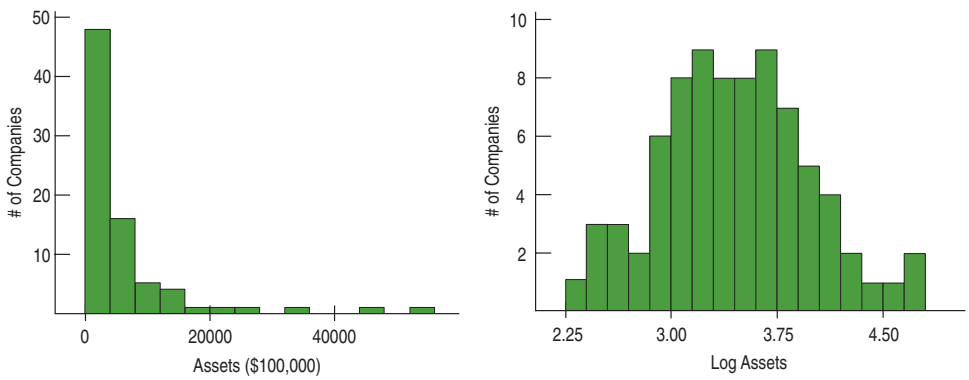


Figure 10.4

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.*

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

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 5.

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

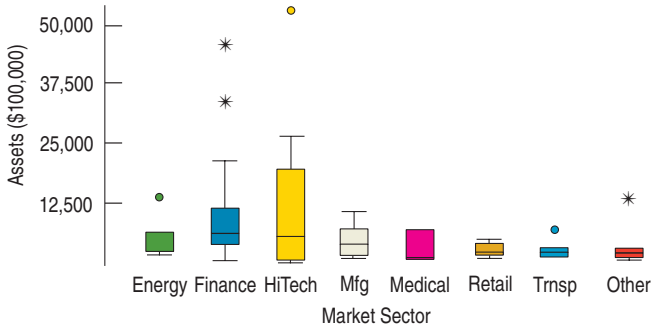


FIGURE 10.5
Assets *of large companies by* Market Sector. *It's hard to compare centers or spreads, and there seem to be a number of high outliers.*

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

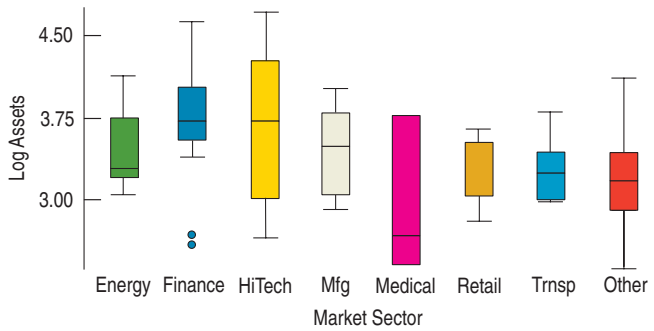


FIGURE 10.6
After re-expressing by logs, it's much easier to compare across market sectors. The boxplots are more nearly symmetric, most have similar spreads, and the companies that seemed to be outliers before are no longer extraordinary. Two new outliers have appeared in the finance sector. They are 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 7. 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.

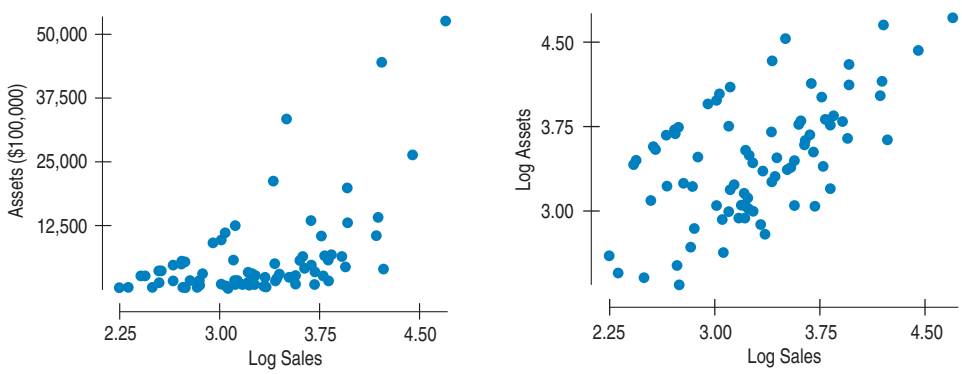


FIGURE 10.7
Assets *vs. log* Sales *shows a positive association (bigger sales go with bigger assets) but a bent shape. Note also that the points go from tightly bunched at the left to widely scattered at the right; the plot "thickens." In the second plot, log* Assets *vs. log* Sales *shows a clean, positive, linear association. And the variability at each value of x is about the same.*

GOAL 4

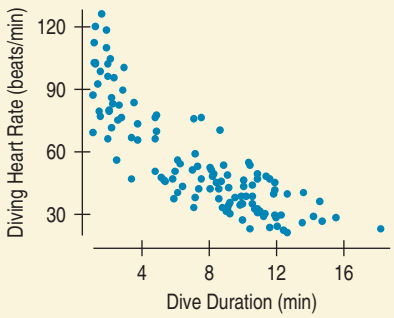
Make the scatter in a scatterplot spread out evenly rather than thickening at one end. Having an even scatter is a condition of many methods of Statistics, as we'll see in later chapters. This goal is closely related to Goal 2, but it often comes along with Goal 3. Indeed, a glance back at the scatterplot (Figure 10.7) 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.

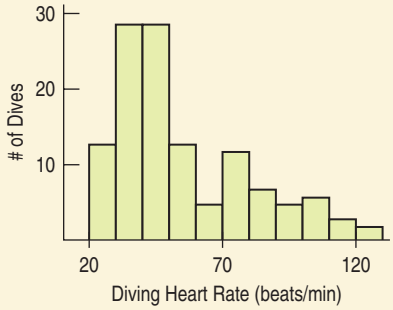
FOR EXAMPLE Recognizing when a re-expression can help

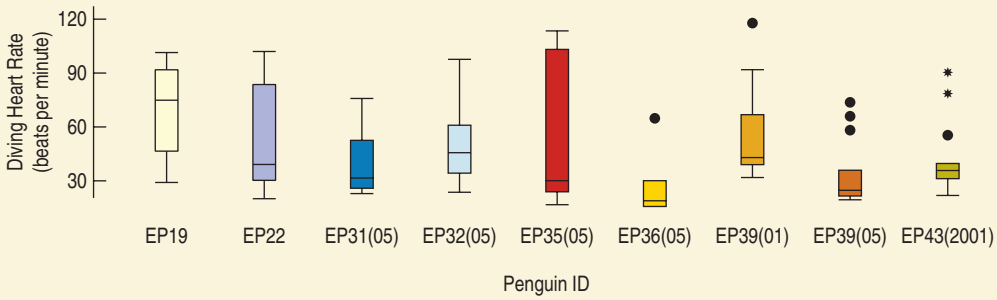
In Chapter 9, we saw the awesome ability of emperor penguins to slow their heart rates while diving. Here are three displays relating to the diving heart rates:

(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?







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

A S ***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? Some kinds of data favor certain re-expressions. But even starting from a suggested one, it's always a good idea to look around a bit. Fortunately, the re-expressions line up in order, so it's easy to slide up and down to find the best one. The trick is to choose our re-expressions from a simple family that includes the most common ways to re-express data. More important, the members of the family line up in order, so that the farther you move away from the original data (the "1" position), the greater is the effect on the data. This fact lets you search systematically for a re-expression that

TI-nspire

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

works, stepping a bit farther from "1" or taking a step back toward "1" as you see the results.

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 for a 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,[2] 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.[3]	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 *negative* 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.

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?

[2] You 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.

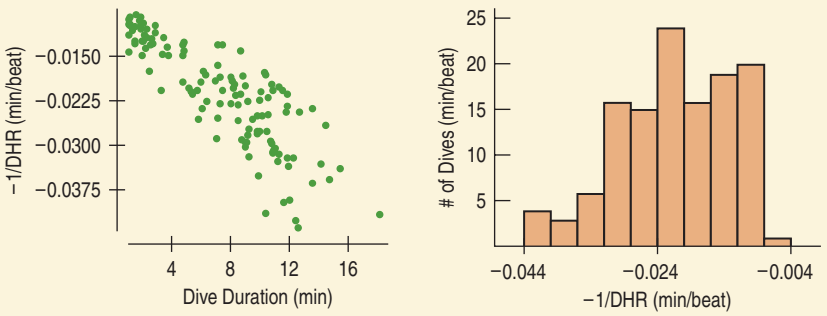
[3] Your calculator or software package probably gives you a choice between "base 10" logarithms and "natural (base e)" logarithms. Don't worry about that. It doesn't matter at all which you use; they have exactly the same effect on the data. If you want to choose, base 10 logarithms can be a bit easier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

Scientific laws often include simple re-expressions. For example, in Psychology, Fechner's Law states that sensation increases as the logarithm of stimulus intensity ($S = k \log R$).

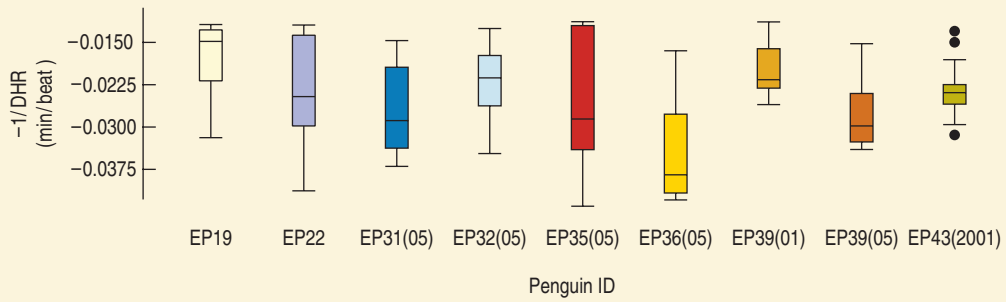
The Ladder of Powers orders the effects that the re-expressions have on data. If you try, say, taking the square roots of all the values in a variable and it helps, but not enough, then move farther down the ladder to the logarithm or reciprocal root. Those re-expressions will have a similar, but even 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.

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?



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-expressing 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.

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

Plan State the problem.

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

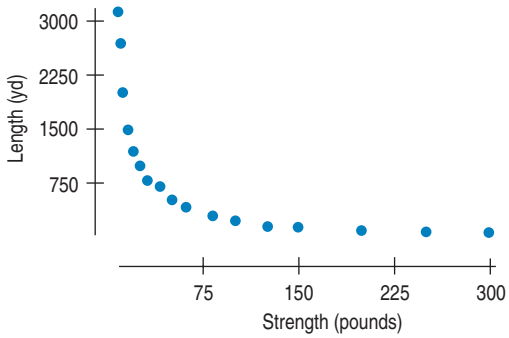
Variables Identify the variables and report the W's.

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).

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.



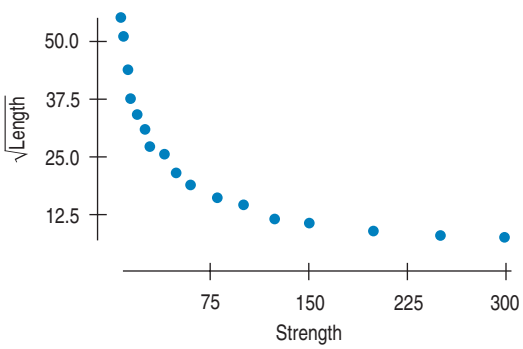
The plot shows a negative direction and an association that has little scatter but is not straight.

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.)

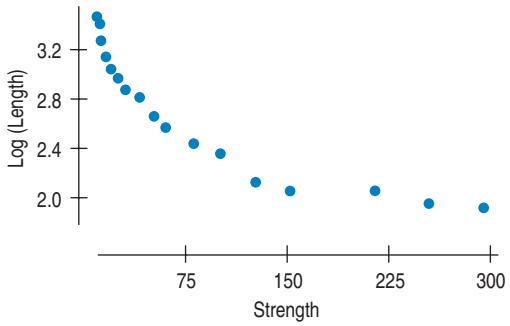
Here's a plot of the square root of *Length* against *Strength*:



The plot is less bent, but still not straight.

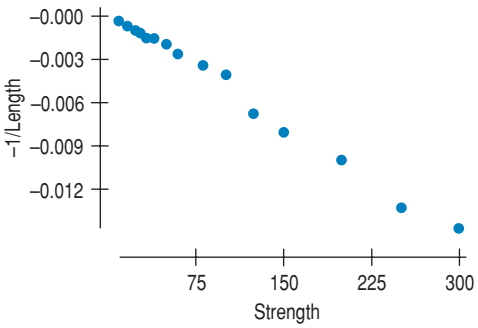
Stepping from the 1/2 power to the "0" power, we try the logarithm of *Length* against *Strength*.

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 $-1/Length$.

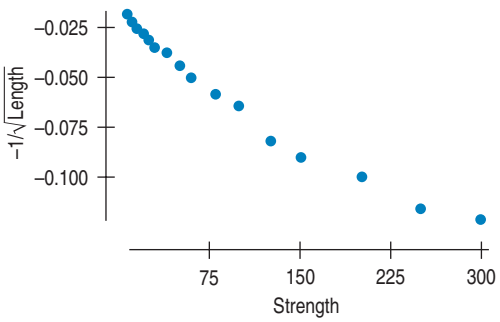
This is much better, but still not straight, so I'll take another step to the "−1" power, or reciprocal.



Maybe now I moved too far along the ladder.

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

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



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

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

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

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

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.

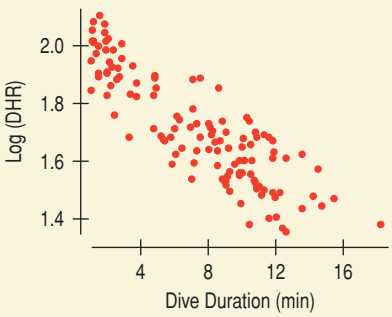
To reverse the process, we first take the reciprocal: $\sqrt{\widehat{Length}} = -1/(-0.036) = 27.778$. Then squaring gets us back to the original units:

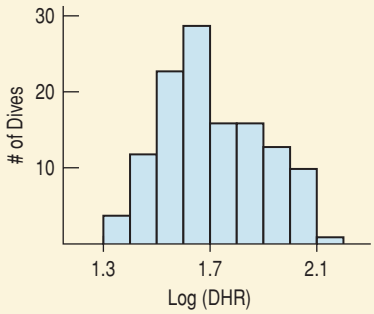
$$\widehat{Length} = 27.778^2 = 771.6\ yards.$$

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!

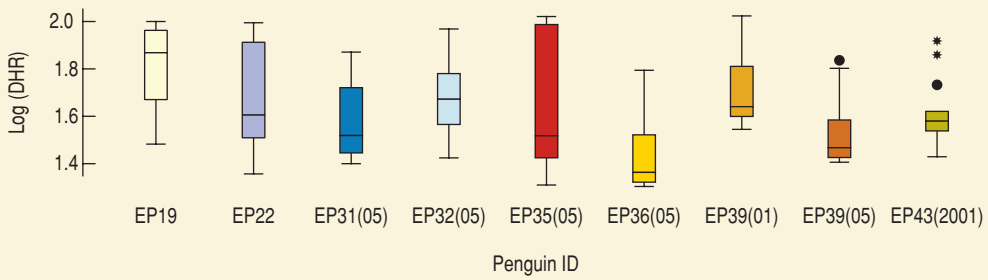
FOR EXAMPLE Comparing re-expressions

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:





Questions: 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?



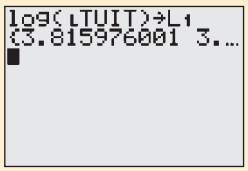
The scatterplot is now more linear and the histogram is symmetric. The boxplots are still a bit skewed to the high end, but less so than for the original Diving Heart Rate values. We don't expect real data to cooperate perfectly, and the logarithm seems like the best compromise re-expression, improving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

TI Tips

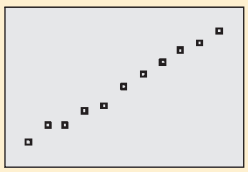
Re-expressing data to achieve linearity

Let's revisit the Arizona State tuition data. Recall that back in Chapter 8 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.

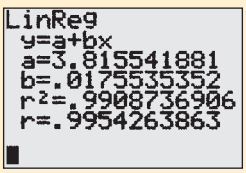
This curved pattern indicates that data re-expression may be in order. If you have no clue what re-expression to try, the Ladder of Powers may help. We just used that approach in the fishing line example. Here, though, we can play a hunch. It is reasonable to suspect that tuition increases at a relatively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



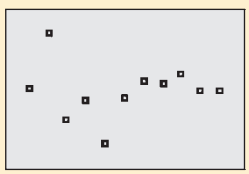
- Tell the calculator to find the logs of the tuitions, and store them as a new list. Remember that you must import the name `TUIT` from the `LIST NAMES` menu. The command is `log(ʟTUIT) STO L1`.
- Check the scatterplot for the re-expressed data by changing your `STATPLOT` specifications to `Xlist:YR` 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.



- Perform the regression for the logarithm of *tuition* vs. *year* with the command `LinReg(a+bx)ʟYR,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`.



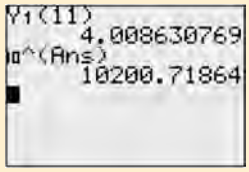
While the residuals for the second and fifth years are comparatively large, the curvature we saw above is gone. The pattern in these residuals seem essentially horizontal and random. This re-expressed model is probably more useful than the original linear model.

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!

$$\log \widehat{tuit} = 3.816 + 0.018\, yr$$

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, $\log(\widehat{tuit}) = 4$ means $\widehat{tuit} = 10^4$, 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

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.

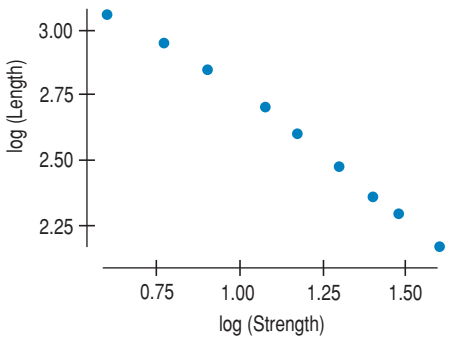


FIGURE 10.8
Plotting log (Length) *against log* (Strength) *gives a straighter shape.*

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

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

It's interesting that the slope of this line (−1.08) is a power[4] 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

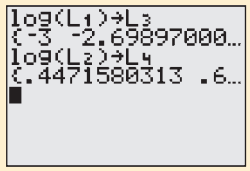
Using logarithmic re-expressions

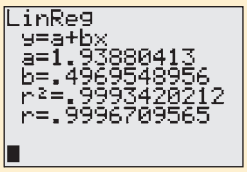
In Chapter 7 we looked at data showing the relationship between the f/stop of a camera's lens and its shutter speed. Let's use the attack of the logarithms to model this situation.

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?

[4] For logarithms, $-1.08 \log(Strength) = \log(Strength^{-1.08})$.



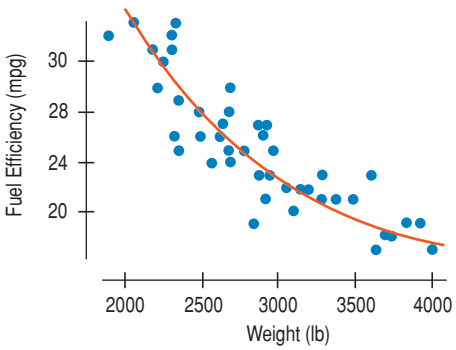


- Find the logarithms of each variable's values. Keep track of where you store everything so you don't get confused! We put log(*speed*) in L3 and log(*f*/*stop*) in L4.
- Make three scatterplots:
  - *f*/*stop* vs. log(*speed*) using Xlist:L3 and Ylist:L2
  - log(*f*/*stop*) vs. *speed* using Xlist:L1 and Ylist:L4
  - log(*f*/*stop*) vs. log(*speed*) using Xlist:L3 and Ylist:L4
- Pick your favorite. We liked log(*f*/*stop*) vs. log(*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.[5]

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

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 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 ($\hat{y} = a + b\ln x$)

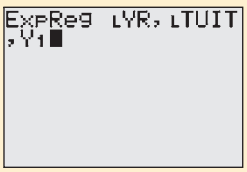
[5] See the slope, 0.497? Just about 0.5. That's because the actual relationship involves the square root of shutter speeds. Technically the *f*/stop listed as 2.8 should be $2\sqrt{2} \approx 2.8284$. Rounding off to 2.8 makes sense for photographers, but it's what led to the minor errors you saw in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

- `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:

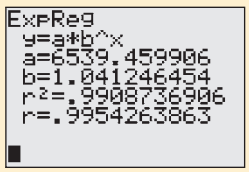
- `5:QuadReg`—fits a quadratic model ($\hat{y} = ax^2 + bx + c$)
- `6:CubicReg`—fits a cubic model ($\hat{y} = ax^3 + bx^2 + cx + d$)

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 even slightly outside the range of the data. We recommend that you be very wary of models of this type.

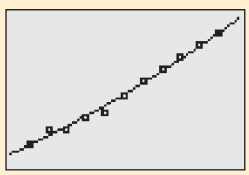


Let's try out one of the calculator shortcuts; we'll use the Arizona State tuition data. (For the last time, we promise!) This time, instead of re-expressing *tuition* to straighten the scatterplot, we'll have the calculator do more of the work.

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.

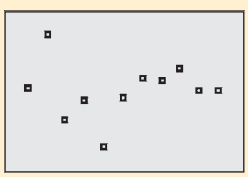


- Choose the exponential model, and specify your variables by importing `YR` and `TUIT` from the list names menu. And, because you'll want to graph the curve later, save its equation by adding `Y1` (from `VARS`, `Y-VARS`, `Function`) to the command `ExpReg LYR, LTUIT, Y1`.
- Set up the scatterplot. `ZoomStat` should show you the curve too.
- Graph the residuals plot.



This all looks 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 $\widehat{tuit} = 6539.46(1.041^{year})$.



Notice 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 because what the calculator just did is mathematically the very same thing. This new equation may look different, but it is equivalent to our earlier model $\log \widehat{tuit} = 3.816 + 0.018\, year$.

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`.

Three of the special TI functions correspond to a simple regression model involving re-expression. The calculator presents the results in an equation of a different form, but it doesn't actually fit that equation. Instead it is just doing the re-expression for you automatically.

Here are the equivalent models for the two approaches.

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 is fitting these equations to the data by minimizing the sum of squared residuals, but it isn't really doing 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. So think of these TI functions as just shortcuts for fitting linear regressions to re-expressed versions of your data.

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 we strongly recommend you use.

Occam's Razor

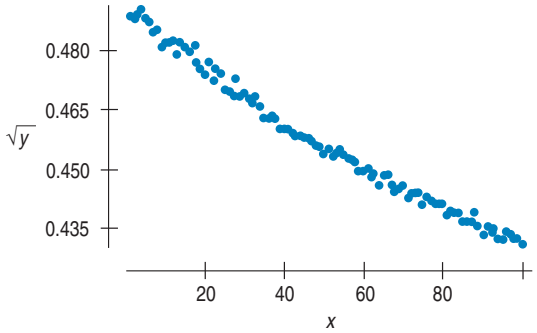
If you think that simpler explanations and simpler models are more likely to give a true picture of the way things work, then you should look for opportunities to re-express your data and simplify your analyses.

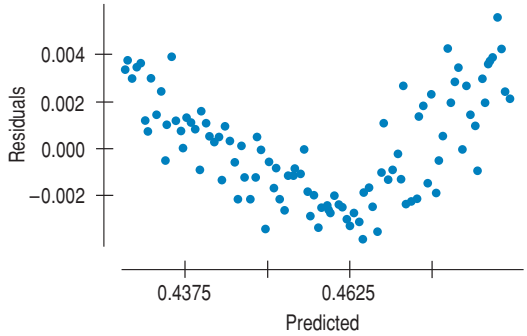
The general principle that simpler explanations are likely to be the better ones is known as Occam's Razor, after the English philosopher and theologian William of Occam (1284–1347).

WHAT CAN GO WRONG?

- **Don't expect your model to be perfect.** In Chapter 6 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. Taking the y-values to an extremely high power may artificially inflate R^2, but it won't give a useful or meaningful model, so it doesn't really simplify anything. It's better to stick to powers between 2 and -2. Even in that range, 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.

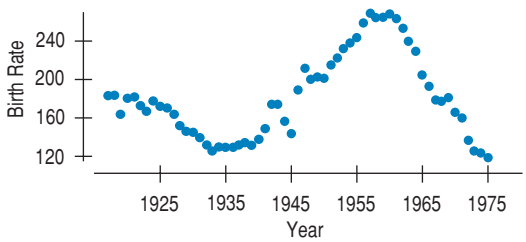


FIGURE 10.9
The shape of the scatterplot of Birth Rates *(births per 100,000 women) in the United States shows an oscillation that cannot be straightened by re-expressing the data.*

- **Watch out for 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. Most statistics programs will just mark the result of trying to re-express such values "missing" if they can't be re-expressed. But that might mean that when you try a re-expression, you inadvertently lose a bunch of data values. The effect of that loss may be surprising and may substantially change your analysis. Because you are likely to be working with a computer package or calculator, take special care that you do not lose otherwise good data values when you choose a re-expression.

  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.
- **Watch for data far from 1.** Data values that are all very far from 1 may not be much affected by re-expression unless the range is very large. Re-expressing numbers between 1 and 100 will have a much greater effect than re-expressing numbers between 100,001 and 100,100. When all your data values are large (for example, working with years), consider subtracting a constant to bring them back near 1. (For example, consider "years since 1950" as an alternative variable for re-expression. Unless your data start at 1950, then avoid creating a zero by using "years since 1949.")

CONNECTIONS

We have seen several ways to model or summarize data. Each requires that the data have a particular simple structure. We seek symmetry for summaries of center and spread and to use a Normal model. We seek equal variation across groups when we compare groups with boxplots or want to compare their centers. We seek linear shape in a scatterplot so that we can use correlation to summarize the scatter and regression to fit a linear model.

Data do often satisfy the requirements to use Statistics methods. But often they do not. Our choice is to stop with just displays, to use much more complex methods, or to re-express the data so that we can use the simpler methods we have developed.

In this fundamental sense, this chapter connects to everything we have done thus far and to all of the methods we will introduce throughout the rest of the book. Re-expression greatly extends the reach and applicability of all of these methods.

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 5)
- 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 224. We re-express data by taking the logarithm, the square root, the reciprocal, or some other mathematical operation on all values of a variable.

Ladder of Powers 226. The Ladder of Powers places in order the effects that many re-expressions have on the data.

Skills

THINK

- Recognize when a well-chosen re-expression may help you improve and simplify your analysis.
- Understand the value of re-expressing data to improve symmetry, to make the scatter around a line more constant, or to make a scatterplot more linear.
- Recognize when the pattern of the data indicates that no re-expression can improve the structure of the data.

SHOW

- Know how to re-express data with powers and how to find an effective re-expression for your data using your statistics software or calculator.
- Be able to reverse any of the common re-expressions to put a predicted value or residual back into the original units.

TELL

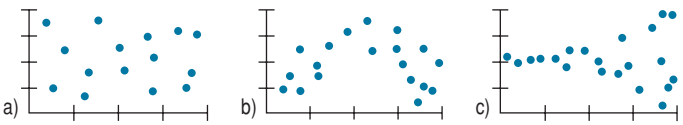
- Be able to describe a summary or display of a re-expressed variable, making clear how it was re-expressed and giving its re-expressed units.
- Be able to describe a regression model fit to re-expressed data in terms of the re-expressed variables.

RE-EXPRESSION ON THE COMPUTER

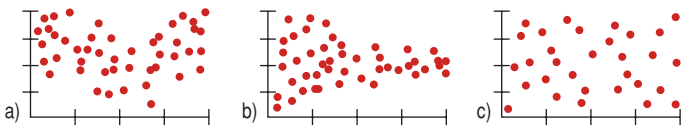
Computers and calculators make it easy to re-express data. Most statistics packages offer a way to re-express and compute with variables. Some packages permit you to specify the power of a re-expression with a slider or other moveable control, possibly while watching the consequences of the re-expression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

EXERCISES

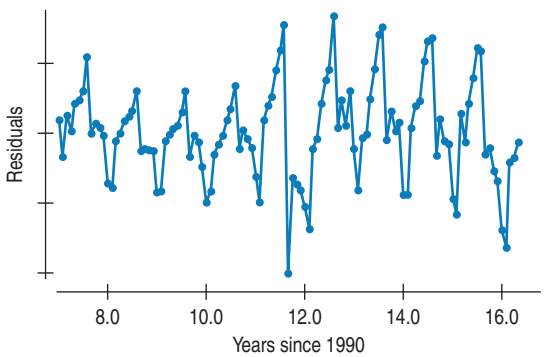
1. Residuals. Suppose you have fit a linear model to some data and now take a look at the residuals. For each of the following possible residuals plots, tell whether you would try a re-expression and, if so, why.



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.

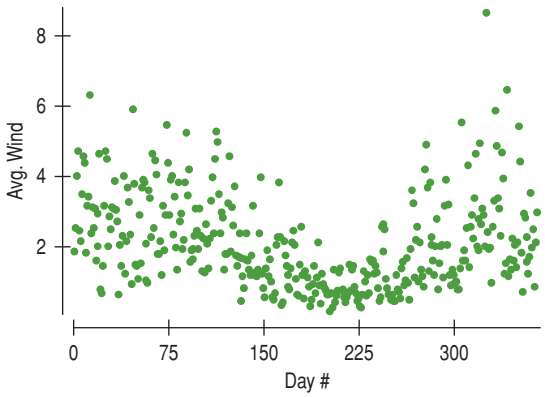


T **3. Airline passengers revisited.** In Chapter 9, Exercise 9, we created a linear model describing the trend in the number of passengers departing from the Oakland (CA) airport each month since the start of 1997. Here's the residual plot, but with lines added to show the order of the values in time:



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

T **4. Hopkins winds, revisited.** In Chapter 5, 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) $\ln \hat{y} = 1.2 + 0.8x$
b) $\sqrt{\hat{y}} = 1.2 + 0.8x$
c) $\frac{1}{\hat{y}} = 1.2 + 0.8x$
d) $\hat{y} = 1.2 + 0.8 \ln x$
e) $\log \hat{y} = 1.2 + 0.8 \log x$

6. More models. For each of the models listed below, predict y when $x = 2$.

a) $\hat{y} = 1.2 + 0.8 \log x$
b) $\log \hat{y} = 1.2 + 0.8x$
c) $\ln \hat{y} = 1.2 + 0.8 \ln x$
d) $\hat{y}^2 = 1.2 + 0.8x$
e) $\frac{1}{\sqrt{\hat{y}}} = 1.2 + 0.8x$

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 on the next page.