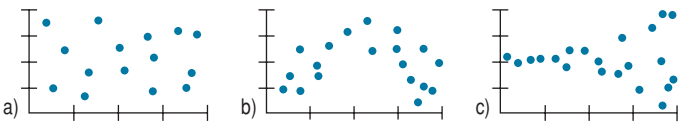
RE-EXPRESSION ON THE COMPUTER

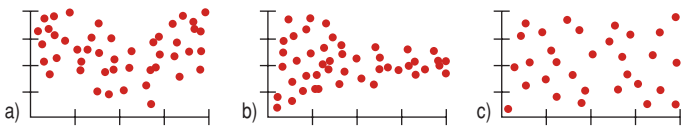
Computers and calculators make it easy to re-express data. Most statistics packages offer a way to re-express and compute with variables. Some packages permit you to specify the power of a re-expression with a slider or other moveable control, possibly while watching the consequences of the re-expression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

EXERCISES

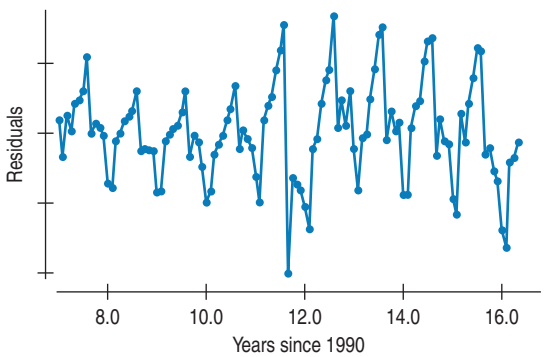
1. Residuals. Suppose you have fit a linear model to some data and now take a look at the residuals. For each of the following possible residuals plots, tell whether you would try a re-expression and, if so, why.



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.

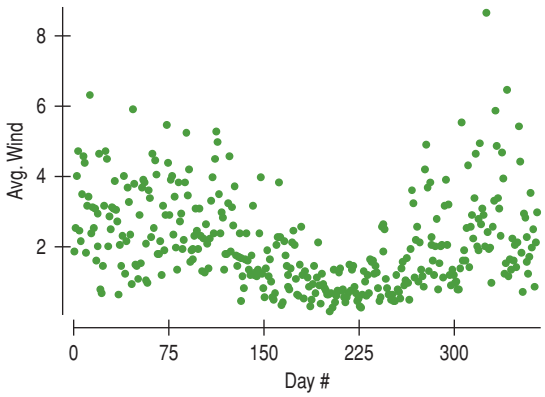


T **3. Airline passengers revisited.** In Chapter 9, Exercise 9, we created a linear model describing the trend in the number of passengers departing from the Oakland (CA) airport each month since the start of 1997. Here's the residual plot, but with lines added to show the order of the values in time:



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 $x = 2$.

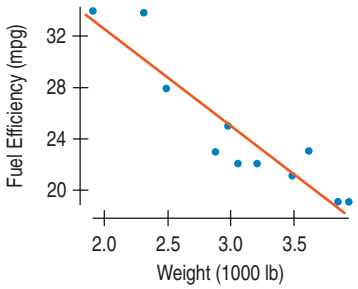
a) $\ln \hat{y} = 1.2 + 0.8x$
b) $\sqrt{\hat{y}} = 1.2 + 0.8x$
c) $\frac{1}{\hat{y}} = 1.2 + 0.8x$
d) $\hat{y} = 1.2 + 0.8 \ln x$
e) $\log \hat{y} = 1.2 + 0.8 \log x$

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.

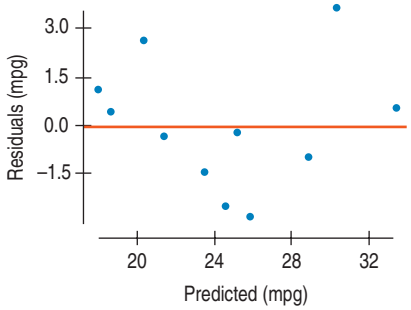


b) Here is the regression analysis for the linear model. What does the slope of the line say about this relationship?

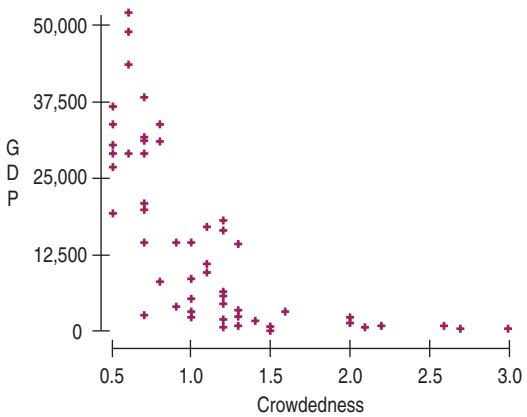
Dependent variable is: Fuel Ef ficiency
R-squared = 85.9%

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.



T **8. Crowdedness.** In a *Chance* magnitude 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:



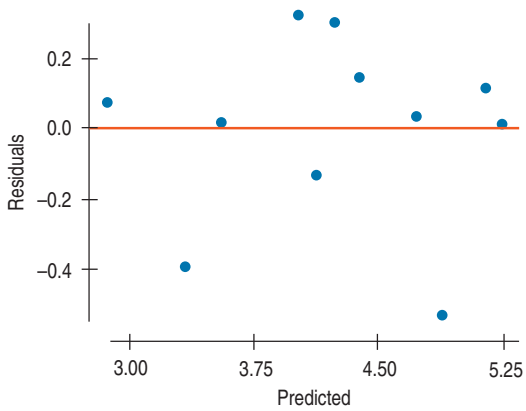
a) Explain why you should re-express these data before trying to fit a model.

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:

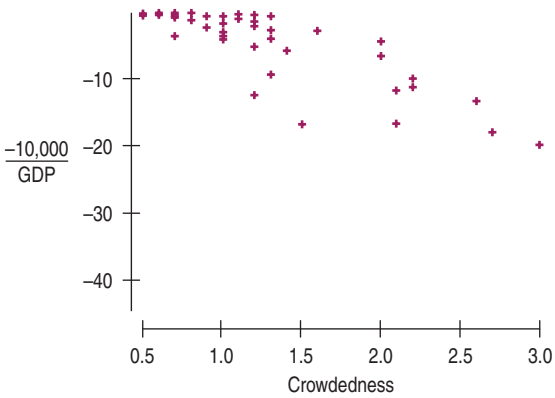
Dependent variable is: Fuel Consumption
R-squared = 89.2%

Variable	Coefficient
Intercept	0.624932
Weight	1.17791



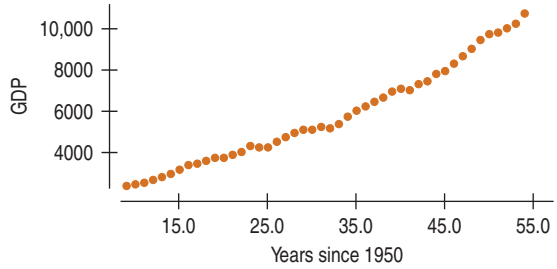
a) Explain why this model appears to be better than the linear model.
b) Using the regression analysis above, write an equation of this model.
c) Interpret the slope of this line.
d) Based on this model, how many miles per gallon would you expect a 3500-pound car to get?

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/*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.

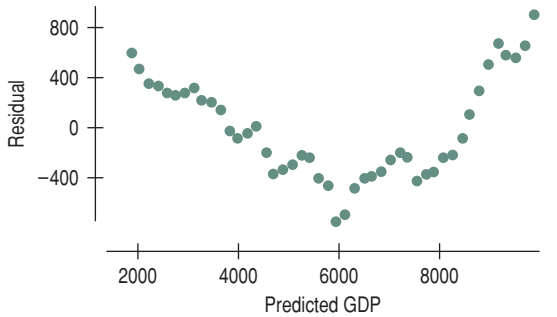


A linear model fit to the relationship looks like this:

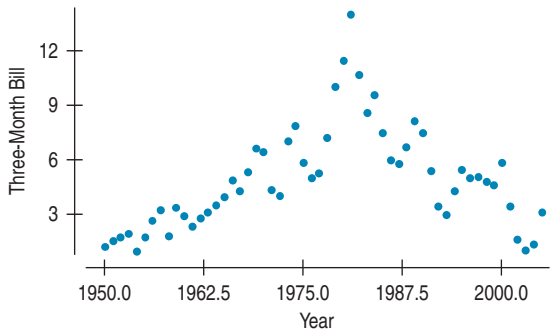
Dependent variable is: GDP
R-squared = 97.2% s = 406.6

Variable	Coefficient
Intercept	240.171
Year-1950	177.689

a) Does the value 97.2% 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:

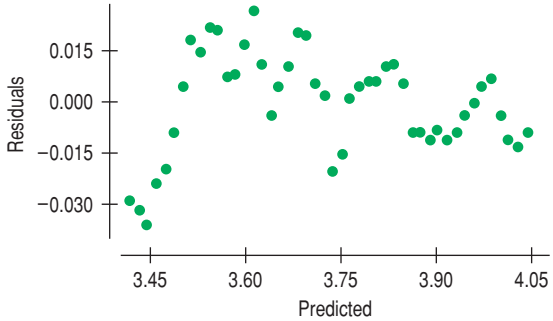


Clearly, the relationship is not linear. Can it be made nearly linear with a re-expression? If so, which one would you suggest? If not, why not?

13. Better GDP model? Consider again the post-1950 trend in U.S. GDP we examined in Exercise 11. Here are a regression and 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 = 99.4% s = 0.0150

Variable	Coefficient
Intercept	3.29092
Year-1950	0.013881



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

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?

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. Her data are shown on the next page.

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 2005.** Ballplayers have been signing ever larger contracts. The highest salaries (in millions of dollars per season) for some notable players are given in the following table.

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 a scatterplot of the data. Does it look straight?
b) Find the regression of *Salary* vs. *Year* and plot the residuals. Do they look straight?
c) Re-express the data, if necessary, to straighten the relationship.
d) What model would you report for the trend in salaries?

T **18. Planet distances and years 2006.** 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 2006.** 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 2006, 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 2006, part 4.** In July 2005, astronomers Mike Brown, Chad Trujillo, and David Rabinowitz announced the discovery of a sun-orbiting object, since named Eris,[6] 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 2006 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

[6] Eris 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.

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

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 2004. Listed below are the gold medal-winning men's weight-lifting performances at the 2004 Olympics.

Weight Class (kg)	Winner (country)	Weight Lifted (kg)
56	Halil Mutlu (Turkey)	295.0
62	Zhiyong Shi (China)	325.0
69	Guozheng Zhang (China)	347.5
77	Taner Sagir (Turkey)	375.0
85	George Asanidze (Georgia)	382.5
94	Milen Dobrev (Bulgaria)	407.5
105	Dmitry Berestov (Russia)	425.0

a) Create a linear model for the *Weight Lifted* in each *Weight Class*.
b) Check the residuals plot. Is your linear model appropriate?
c) Create a better model.
d) Explain why you think your new model is better.
e) Based on your model, which of the medalists turned in the most surprising performance? Explain.

T 25. Life expectancy. The data in the next column 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
Life exp.	48.6	54.4	59.7	62.1	66.5	67.4	68.0	70.7	72.7	74.9

T 26. Lifting more weight 2004. In Exercise 24 you examined the winning weight-lifting performances for the 2004 Olympics. One of the competitors turned in a performance that appears not to fit the model you created.
a) Consider that competitor to be an outlier. Eliminate that data point and re-create your model.
b) Using this revised model, how much would you have expected the outlier competitor to lift?
c) Explain the meaning of the residual from your new model for that competitor.

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 2003. Insurance companies and other organizations use actuarial tables to estimate the remaining lifespans of their customers. On the next page are the estimated additional years of life for black males in the United States, according to a 2003 National Vital Statistics Report. (www.cdc.gov/nchs/deaths.htm)

Age	10	20	30	40	50	60	70	80	90	100
Years Left	60.3	50.7	41.8	32.9	24.8	17.9	12.1	7.9	5.0	3.0

a) Find a re-expression to create an appropriate model.
b) Predict the lifespan of an 18-year-old black man.
c) Are you satisfied that your model has accounted for the relationship between *Years Left* and *Age*? Explain.

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

1. Counts are often best transformed by using the square root.
2. None. The relationship is already straight.
3. 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.

PART II

REVIEW OF PART II

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 3.
- 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 7 through 10. And go on to the next page for 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)