Transforming to Achieve Linearity

Date:

Essential Question: Why do we transform data to make it appear linear?

Questions:

Notes:

In a previous chapter, we learned how to analyze relationships between two quantitative variables that showed a linear pattern. When two-variable data show a curved relationship, we must develop new techniques for finding an appropriate model.

This section describes several simple transformations of data that can straighten a nonlinear pattern. Once the data have been transformed to achieve linearity, we can use least-squares regression to generate a useful model for making predictions.

If the conditions for regression inference are met, we can estimate or test a claim about the slope of the population (true) regression line using the transformed data.

Transforming with Powers and Roots

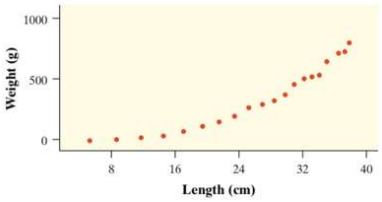
When you visit a pizza parlor, you order a pizza by its diameter, 10, 12 or 14 inches. But the amount of pizza you get to eat depends on the area of the pizza. The area of a circle is πr^2 (π times the square of its radius). So the are of a round pizza with diameter x is $area = \pi\left(\frac{x}{2}\right)^2 = \frac{\pi}{4}x^2$. This is a power model of the form $y = ax^p$.

Although a power model of the form $y = ax^p$ describes the relationship between x and y in this setting, there is a linear relationship between x^p and y. If we transform the values of the explanatory variable x by raising them to the p power, and graph the points (x^p, y), the scatterplot should have a linear form.

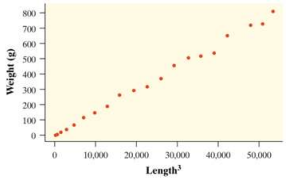
example: Fishing contest

Imagine that you have been put in charge of organizing a fishing tournament in which prizes will be given for the heaviest Atlantic Ocean rockfish caught. You know that many of the fish caught during the tournament will be measured and released. You are also aware that using delicate scales to try to weigh a fish that is flopping around in a moving boat will probably not yield very accurate results. It would be much easier to measure the length of the fish while on the boat.

Length:	5.2	8.5	11.5	14.3	16.8	19.2	21.3	23.3	25.0	26.7
Weight:	2	8	21	38	69	117	148	190	264	293
Length:	28.2	29.6	30.8	32.0	33.0	34.0	34.9	36.4	37.1	37.7
Weight:	318	371	455	504	518	537	651	719	726	810



Because length is one-dimensional and weight (like volume) is three-dimensional, a power model of the form $weight = a(length)^3$ should describe the relationship.



This transformation of the explanatory variable helps us produce a graph that is quite linear.

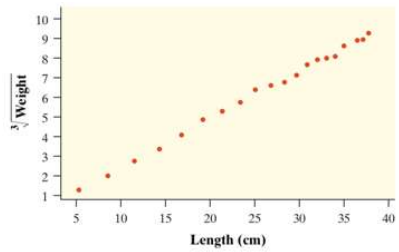
Summary:

Date:

Questions:

Notes:

Another way to transform the data to achieve linearity is to take the cube root of the weight values and graph the cube root of weight versus length. Note that the resulting scatterplot also has a linear form.



Once we straighten out the curved pattern in the original scatterplot, we fit a least-squares line to the transformed data.

When experience or theory suggests that the relationship between two variables is described by a power model of the form $y = ax^p$, you now have two strategies for transforming the data to achieve linearity.

1. Raise the values of the explanatory variable *x* to the *p* power and plot the points (x^p, y).
2. Take the p^{th} root of the values of the response variable *y* and plot the points $\left(x, \sqrt[p]{y}\right)$.

Transforming with Logarithms

To achieve linearity from a power model, we apply the logarithm transformation to both variables. Here are the details:

1. A power model has the form $y = ax^p$, where *a* and *p* are constants.
2. Take the logarithm of both sides of this equations. Using properties of logarithms, we get
$$\log y = \log(ax^p) = \log a + p \log x$$
The equation $\log y = \log a + p \log x$ shows that taking the logarithm of both variables results in a linear relationship between log *x* and log *y*.
3. Look carefully: the power *p* in the power model becomes the slope of the straight line links log *y* to log *x*.

If a power model describes the relationship between two variables, a scatterplot of the logarithms of both variables should produce a linear pattern. Then we can fit an LSRL to the transformed data and use the linear model to make predictions.

Summary:

Date:

Questions:

Notes:

example: New Planet

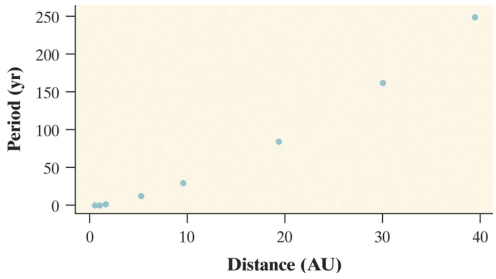
On July 31, 2005, a team of astronomers announced that they had discovered what appeared to be a new planet in our solar system. They had first observed this object almost two years earlier using a telescope at Caltech's Palomar Observatory in California.

Originally named UB313, the potential planet is bigger than Pluto and has an average distance of about 9.5 billion miles from the sun. (For reference, Earth is about 93 million miles from the sun.)

Could this new astronomical body, now called Eris, be a new planet? At the time of the discovery, there were nine known planets in our solar system.

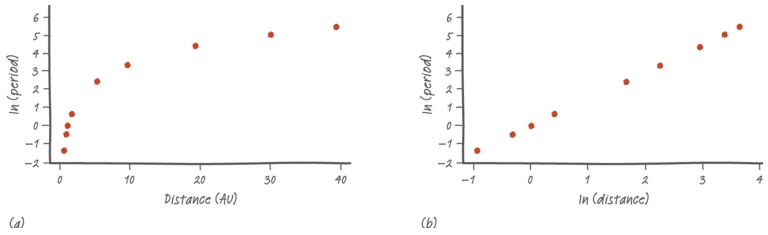
Here are data on the distance from the sun and period of revolution of those planets. Note that distance is measured in astronomical units (AU), the number of earth distances the object is from the sun.

Planet	Distance from sun (astronomical units)	Period of revolution (Earth years)
Mercury	0.387	0.241
Venus	0.723	0.615
Earth	1.000	1.000
Mars	1.524	1.881
Jupiter	5.203	11.862
Saturn	9.539	29.456
Uranus	19.191	84.070
Neptune	30.061	164.810
Pluto	39.529	248.530



There appears to be a strong curved relationship between distance from the sun and period of revolution.

The graphs below show the results of two different transformations of the data. Explain why a power model would provide a more appropriate description of the relationship between period of revolution and distance from the sun than an exponential model.



The scatterplot of ln(period) versus distance is clearly curved, so an exponential model would not be appropriate. However, the graph of ln(period) versus ln(distance) has a strong linear pattern, indicating that a power model would be more appropriate.

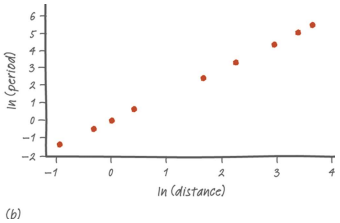
Summary:

Date:

Questions:

Notes:

Minitab output from a linear regression analysis on the transformed data is shown below.



Predictor	Coef	SE Coef	T	P
Constant	0.0002544	0.0001759	1.45	0.191
ln(distance)	1.49986	0.00008	18598.27	0.000

S = 0.000393364 R-Sq = 100.0% R-Sq(adj) = 100.0%

Give the equation of the LSRL. Be sure to define any variables you use.

$$\widehat{\ln(\text{period})} = 0.0002544 + 1.49986 \ln(\text{distance})$$

Let's use our model to predict the period of revolution for Eris, which is 9,500,000,000/93,000,000 = 102.1505 AU from the sun.

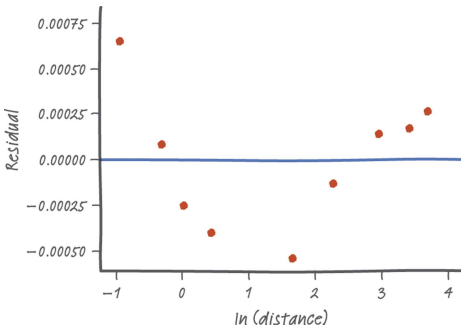
$$\widehat{\ln(\text{period})} = 0.0002544 + 1.49986 \ln(\text{distance})$$

$$\widehat{\ln(\text{period})} = 0.0002544 + 1.49986 \ln(102.15) = 6.939$$

$$\widehat{\text{period}} = e^{6.939} \approx 1032 \text{ years}$$

Question: Do you expect your prediction to be too high, too low, or just right?

A residual plot for the linear transformation is shown below.



Eris' value for ln(*distance*) is 6.939, which would fall at the right of the residual plot, where all of the residuals are positive.

Because residual = actual y - predicted y seems likely to be positive, we would expect our prediction to be too low.

Summary:

Date:

Questions:

Notes:

Sometimes the relationship between y and x is based on repeated multiplication by a constant factor. That is, each time x increases by 1 unit, the value of y is multiplied by b. An exponential model of the form $y = ab^x$ describes such multiplicative growth.

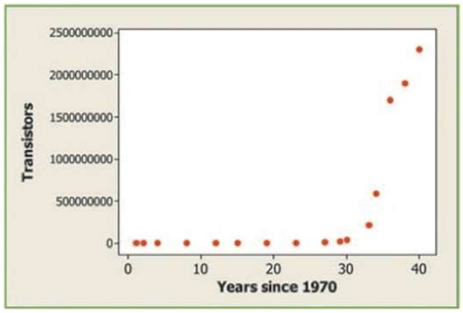
We can transform this model using logarithms.

$y = ab^x$
$\ln y = \ln(ab^x)$ take the logarithm of both sides
$\ln y = \ln a + \ln b^x$ use properties of logs: $\log(a \cdot b) = \log a + \log b$
$\ln y = \ln a + x \cdot \ln b$ use properties of log: $\log b^x = x \cdot \log b$

example: Transistors for Microprocessors

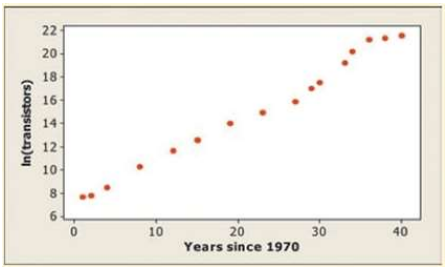
Gordon Moore, one of the founders of Intel Corporation, predicted in 1965 that the number of transistors on an integrated circuit chip would double every 18 months. This is Moore's law, one way to measure the revolution in computing. Below are data on the dates and number of transistors for Intel microprocessors.

Processor	Date	Transistors
4004	1971	2,250
8008	1972	2,500
8080	1974	5,000
8086	1978	29,000
286	1982	120,000
386	1985	275,000
486 DX	1989	1,180,000
Pentium	1993	3,100,000
Pentium II	1997	7,500,000
Pentium III	1999	24,000,000
Pentium 4	2000	42,000,000
Itanium 2	2003	220,000,000
Itanium 2 w/9MB cache	2004	592,000,000
Dual-core Itanium 2	2006	1,700,000,000
Six-core Xeon 7400	2008	1,900,000,000
8-core Xeon Nehalem-EX	2010	2,300,000,000



The graph sows the growth in the number of transistors on a computer chip from 1971 to 2010. Notice that we used "years since 1970" as the explanatory variable. If Moore's law is correct, then an exponential model should describe the relationship between the variables.

A scatterplot of the natural logarithm of the number of transistors on a computer hip versus years since 1970 is shown below.



If an exponential model describes the relationship between two variables x and y, then we expect a scatterplot of (x, ln y) to be roughly linear. As you can see, the scatterplot of ln(transistors) versus years since 1970 has a fairly linear pattern.

Summary:

Date:

Questions:

Notes:

Minitab output from a linear regression analysis on the transformed data is shown below.

Predictor	Coef	SE Coef	T	P
Constant	7.0647	0.2672	26.44	0.000
Years since 1970	0.36583	0.01048	34.91	0.000

S = 0.544467 R-Sq = 98.9% R-Sq(adj) = 98.8%

The least-squares regression line would be:

$$\widehat{\ln(\text{transistors})} = 7.0647 + 0.36583(\text{years since } 1970)$$

Use your model to predict the number of transistors on an Intel computer chip in 2020.

$$\widehat{\ln(\text{transistors})} = 7.0647 + 0.36583(\text{years since } 1970)$$

Because 2020 is 50 years since 1970, we have

$$\widehat{\ln(\text{transistors})} = 7.0647 + 0.36583(50) = 25.3562$$

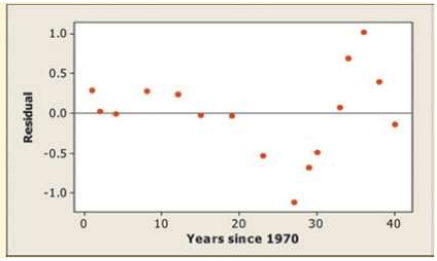
To find the predicted number of transistors, we use the definition of a logarithm as an exponent:

$$\widehat{\ln(\text{transistors})} = 25.3562 \Rightarrow \widehat{\log_e(\text{transistors})} = 25.3562$$

$$\widehat{\text{transistors}} = e^{25.362} \approx 1.028 \cdot 10^{11}$$

This model predicts that an Intel chip made in 2020 will have about 100 billion transistors.

A residual plot for the linear regression is shown below.



The residual plot shows a distinct pattern, with the residuals going from positive to negative to positive as we move from left to right. But the residuals are small in size relative to the transformed *y*-values.

Also, the scatterplot of the transformed data is much more linear than the original scatterplot. We feel reasonably comfortable using this model to make predictions about the number of transistors on a computer chip.

Summary: