(S2016)

Section 12.2 - Transforming to Achieve Linearity Part 1 (pp. 765-771)

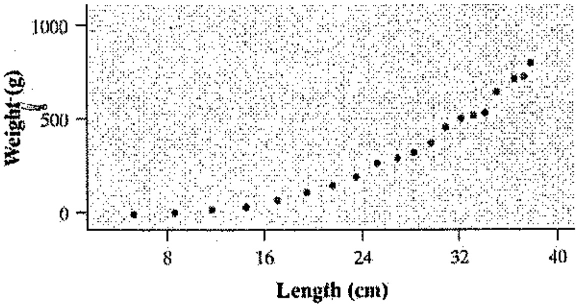
In Chapter 3, we learned how to analyze relationships between two quantitative variables that showed a linear pattern. When two-variable data show a curved relationship, we must develop new techniques for finding an appropriate model. This section describes several simple transformations of data that 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. And 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.

Applying a function such as the logarithm or square root to a quantitative variable is called transforming the data. We will see in this section that understanding how simple functions work helps us choose and use transformations to straighten nonlinear patterns.

Example - 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. What you need is a way to convert the length of the fish to its weight.

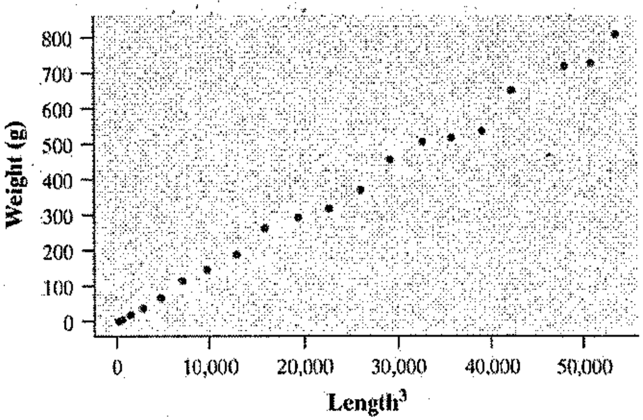
Length:	5.2	8.5	11.5	14.3	16.8	19.2	21.3	23.3	25.0	26.7
Weight:	2	8	21	38	69	117	148	190	264	293
Length:	28.2	29.6	30.8	32.0	33.0	34.0	34.9	36.4	37.1	37.7
Weight:	318	371	455	504	518	537	651	719	726	810



Scatterplot of rockfish weight vs. length

NOTE CLEARLY CURVED SHAPE.

LENGTH IS 1-DIMENSIONAL WHEREAS WT IS 3-D.



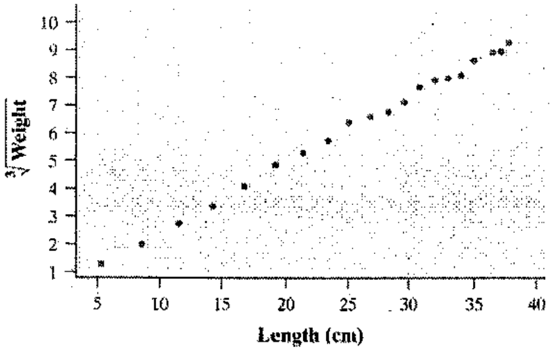
Scatterplot of weight vs. length^3

CUBE LENGTHS (VOLUME)

$\Rightarrow$ LINEARITY

$\Downarrow$

LIN. REG. TECHNIQUES CAN NOW BE USED.



Scatterplot of $\sqrt[3]{weight}$ vs. length

ANOTHER OPTION IS TO TAKE $\sqrt[3]{WT}$ SINCE LENGTH IS SORT OF = $\sqrt[3]{VOLUME}$

⇓

LINEAR PLOT

⇓

LINEAR REGRESSION TECHNIQUES.

Here is Minitab output from separate regression analyses of the two sets of transformed Atlantic Ocean rockfish data.

$\widehat{WT} = 4.066 + 0.015(WT^3)$

Transformation 1: (length^3, weight)

```
Predictor      Coef       SE Coef       T       P
Constant      4.066      6.902         0.59    0.563
Length^3      0.0146774  0.0002404    61.07    0.000
S = 18.8412  R-Sq = 99.5%  R-Sq(adj) = 99.5%
```

Transformation 2: (length, $\sqrt[3]{\text{weight}}$)

```
Predictor      Coef       SE Coef       T       P
Constant     -0.02204    0.07762     -0.28    0.780
Length        0.246616   0.002868    86.00    0.000
S = 0.124161  R-Sq = 99.8%  R-Sq(adj) = 99.7%
```

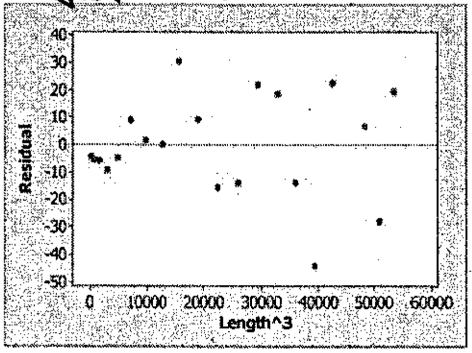
$\sqrt[3]{WT} = -0.022 + 0.247(WT)$

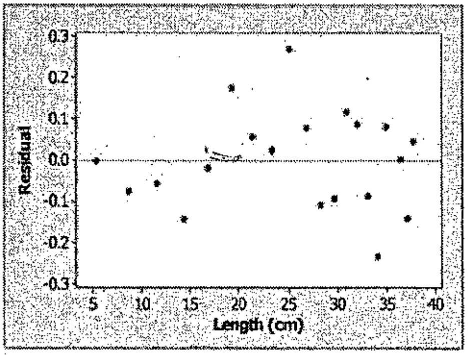
* COMPARE r^2 VALUES: 99.5% VS 99.8%

* COMPARE S VALUES: 18.8412 VS 0.124161

↑ DISCUSS ↑

WT $\sqrt[3]{WT}$





① GIVE LSR LINES:

$\widehat{WT} = 4.066 + 0.0146774(\text{LENGTH}^3)$

$\sqrt[3]{\widehat{WT}} = -0.02204 + 0.246616(\text{LENGTH})$

② FISH = 36 cm ⇒ FIND PREDICTED WT.

$\widehat{WT} = 4.066 + 0.00146774(36)^3 = \boxed{688.9g}$

$\sqrt[3]{\widehat{WT}} = -0.02204 + 0.246616(36) = 8.856$

$\widehat{WT} = (8.856)^3 = \boxed{694.6g}$

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 pth root of the values of the response variable y and plot the points $(x, \sqrt[p]{y})$

What if you have no idea what power to choose? You could guess and test until you find a transformation that works. Some technology comes with built-in sliders that allow you to dynamically adjust the power and watch the scatterplot change shape as you do.

(37) (A) (1) $\sqrt{\text{LENGTH}}$, PERIOD

$\widehat{\text{PERIOD}} = -0.08594 + 0.209999\sqrt{L}$

(2) L , PERIOD^2

$\widehat{\text{PERIOD}^2} = -0.15465 + 0.042836(\text{LENGTH}^2)$

(B) $\widehat{\text{PERIOD}} = -0.08594 + 0.209999\sqrt{80} = \boxed{1.792\text{ SEC}}$

$\widehat{\text{PERIOD}^2} = -0.15465 + 0.042836(80^2) = 3.269 \Rightarrow$

$\widehat{\text{PERIOD}} = \sqrt{3.269} = \boxed{1.808\text{ SEC}}$

HW: p. 786 problems 33, 35.

Section 12.2 - Transforming to Achieve Linearity Part 2 (pp. 771-785)

Transforming with logarithms - Not all curved relationships are described by power models. Some relationships can be described by a logarithmic model of the form

$y = a + b \log x$.

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.

If an exponential model of the form $y = ab^x$ describes the relationship between x and y, we can use logarithms to transform the data to produce a linear relationship.

$$\log y = \log ab^x \Rightarrow \log y = \log a + \log b^x \Rightarrow \text{Log } y = \log a + x \log b$$

We can rearrange the final equation as $\log y = \log a + (\log b)x$. Notice that log a and log b are constants because a and b are constants.

So the equation gives a linear model relating the explanatory variable x to the transformed variable log y.

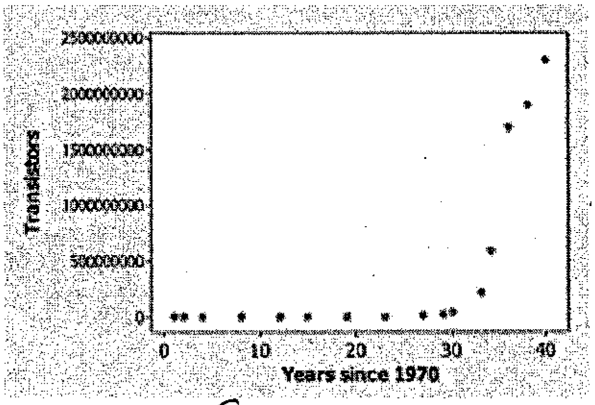
Thus, if the relationship between two variables follows an exponential model, and we plot the logarithm (base 10 or base e) of y against x, we should observe a straight-line pattern in the transformed data.

If we fit a least-squares regression line to the transformed data, we can find the predicted value of the logarithm of y for any value of the explanatory variable x by substituting our x-value into the equation of the line.

To obtain the corresponding prediction for the response variable y, we have to "undo" the logarithm transformation to return to the original units of measurement. One way of doing this is to use the definition of a logarithm as an exponent: $\log_b a = x \Rightarrow b^x = a$

Example - Moore's Law and Computer Chips (p. 773)

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. Here are data on the dates and number of transistors for Intel microprocessors:



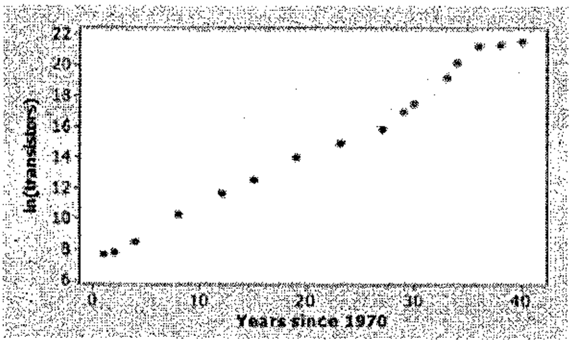
(EXPONENTIAL)

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

(a) A scatterplot of the natural logarithm (log base e or ln) of the number of transistors on a computer chip versus years since 1970 is shown. Based on this graph, explain why it would be reasonable to use an exponential model to describe the relationship between number of transistors and years since 1970.

SINCE ln(TRANSISTORS) IS LINEAR, WE EXPECT THE ORIGINAL RELATIONSHIP BETWEEN # OF TRANSISTORS + YRS SINCE 1970 TO BE EXPONENTIAL.

(b) Minitab output from a linear regression analysis on the transformed data is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.



```
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%
```

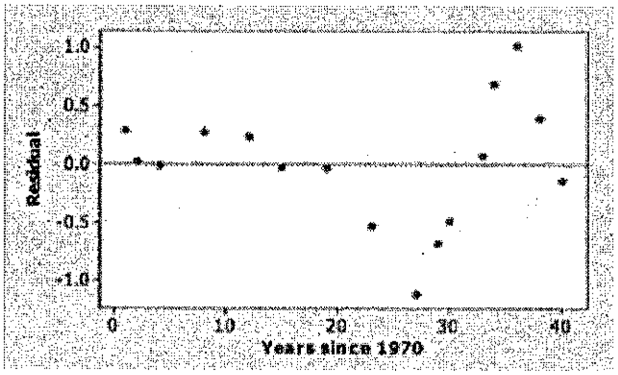
$$\widehat{\ln(\text{TRANS})} = 7.0647 + 0.36583\left(\begin{smallmatrix}\text{YRS}\\ \text{SINCE}\\ 1970\end{smallmatrix}\right)$$

(c) USE MODEL TO PREDICT # OF TRANSISTORS IN 2020.

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

$$e^{\widehat{\ln(\text{TRANS})}} = e^{25.3562} = 1.028 \times 10^{11} \text{ TRANSISTORS.}$$

(d) A residual plot for the linear regression in part (b) is shown below. Discuss what this graph tells you about the appropriateness of the model.



- DISTINCT PATTERN
- RESIDS ARE REALLY SMALL COMPARED TO TRANSFORMED VALUES
- SCATTERPLOT OF TRANSFORMED DATA IS MUCH MORE LINEAR THAN ORIGINAL DATA.

* REASONABLY COMFORTABLE WITH MODEL.

Power Models Again

When we apply the logarithm transformation to the response variable y in an exponential model, we produce a linear relationship. To achieve linearity from a power model, we apply the logarithm transformation to both variables. Here are the details:

A power model has the form $y = ax^p$, where a and p are constants.

Take the logarithm of both sides of this equation. Using properties of logarithms,

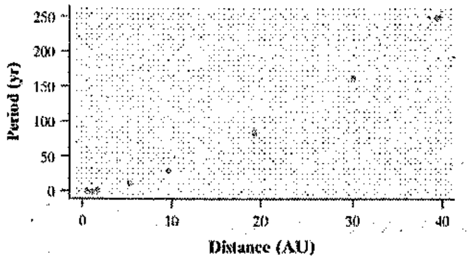
$$\log y = \log(ax^p) = \log a + \log(x^p) = \log a + p \log x$$

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 that 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 a least-squares regression line to the transformed data and use the linear model to make predictions.

Example (p. 778) - On July 31, 2005, a team of astronomers announced that they had discovered what appeared to be a new planet in our solar system. Originally named UB313, the potential planet is bigger than Pluto and has an average distance of about 9.5 billion 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 (in astronomical units, AU) and period of revolution of those planets.

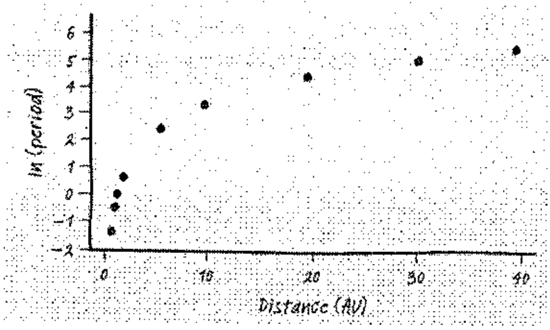


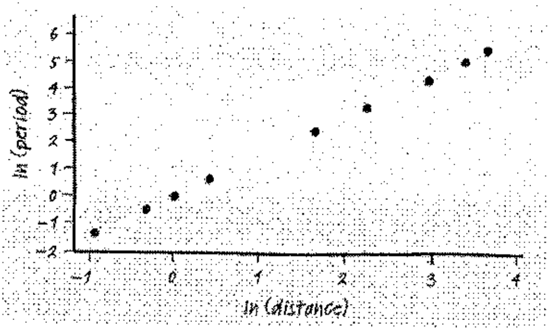
Planet	Distance from sun (astronomical units)	Period of revolution (Earth years)
Mercury	0.387	0.241
Venus	0.723	0.615
Earth	1.000	1.000
Mars	1.524	1.881
Jupiter	5.203	11.862
Saturn	9.539	29.456
Uranus	19.191	84.070
Neptune	30.061	164.810
Pluto	39.529	248.530

Describe the relationship between distance from the sun and period of revolution.

STRONG CURVILINEAR POSITIVE RELATIONSHIP BETWEEN DISTANCE FROM SUN (AU) AND PERIOD OF REVOLUTION (YRS).

(a) Based on the scatterplots below, 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.





ln FOR EXP MODEL (CURVED)

ln - ln ⇒ POWER MODEL (LINEAR) BETTER!

(b) Minitab output from a linear regression analysis on the transformed data (ln(distance), ln(period)) is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.

```
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%
```

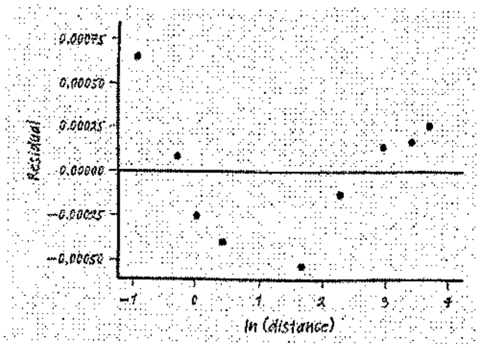
$$\widehat{\ln(\text{PERIOD})} = 0.0002544 + 1.49986 \ln(\text{DIST})$$

(c) Use your model from part (b) to predict the period of revolution for Eris, which is 9,500,000,000/93,000,000 = 102.15 AU from the sun. Show your work.

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

$$\widehat{\text{PERIOD}} = e^{6.939} \approx \boxed{1032 \text{ YRS}}$$

(d) A residual plot for the linear regression in part (b) is shown below. Do you expect your prediction in part (c) to be too high, too low, or just right? Justify your answer.



$\ln(102.15) = 4.626$

4.626 WOULD FALL ON RIGHT SIDE OF PLOT. SINCE RESID = ACTUAL - PREDICTED, WE EXPECT OUR PREDICTION TO BE LOW.

HW: p. 788 problems 35, 37, 39, 41, 45-50.