Chapter 10

Re-expressing Data: ***Get it Straight!***

It's easier than you think!

- **Don't assume that some re-expression will always work**
- **We don't need a perfect model, but a *useful* one !**

Keep In Mind…

- When we re-express for one reason, we often end up helping other aspects
- Logarithms straighten out the exponential trend and pull in the long right tail in the histogram
- Helps deal with potential "infinite" quantities
- Leads to simpler models

Benefits

- Mathematics and calculations are more difficult
- Straight lines are easy to understand
  - We know how to think about the slope and y-intercept

Why not just use a curve?

Reading Ch. 10 Quiz
5 min (10 points)

1. Name two situations/reasons we would want to consider re-expressing a data set.
2. What is the Ladder of Powers?
3. What is one often reliable method to re-express data and make it more linear?
4. Why isn't it better to simply use a curve to model data?
5. Name one of the benefits of re-expressing data.

Re-expressing Data

- Re-expression is another name for changing the scale of (transforming) the data.
- It's not cheating!
- We do this on a daily basis
- Ex: bike speed vs. running speed
  - ***Bike speed***: 15 mph $\frac{\text{distance}}{\text{time}}$
  - ***Running speed***: 6 min. in one mile $\frac{\text{time}}{\text{distance}}$

Re-expressing Data

- Re-expressed variables are common in scientific and social laws and models. Logs, reciprocals, roots, and inverse squares show up in physics, chemistry, psychology, and economics.
- Note the difference between creating a model and the wisdom of using it. Here, we have to create the model and then check to see if we should have. We need the residuals to decide whether the model is appropriate, but we need the model to fit the residuals.

Re-expressing Data

- Make sure that a re-expression can be meaningful.
  - Once we re-express, decide if the model is appropriate
  - Create a model
  - Plot the residuals. If there's a curve, build another model
  - Once we find a model that has random, unstructured residuals, interpret and use it.
- When the appropriate model is found, then
  - Ask how strong is the model
  - Look at the pattern
  - R^2--when interpreting keep in mind that it is still variability, but it is variability in the re-expressed variables and NOT the original

Re-expressing Data

  - Correlation is strength of a linear association so discuss "r" only if the reexpression makes the relationship linear.
- Residual plots are a ***signal-and-noise*** issue. A scatterplot shows the mixture of
  - ***signal*** (the underlying association between the variables)
  - ***noise*** (the random variation unaccounted for by the association)

Re-expressing Data

- The residual plot shows us the variation that remains undescribed by the model. If the plot appears to be random—just noise– we know we have captured the whole signal. If, however, there remains a curve in the residual plot, then we know we missed some of the signal. The model does not tell the whole story, so you have to look for a better model.

Example:

- *In a scatterplot of height and weight, we know that taller people generally weigh more—that's the signal. But not all people who are 6 feet tall are the same weight. The variation is the noise. We assume this variation is random. We seek regression model that describes the signal—the underlying relationship between height and weight.*

Re-expressing Data

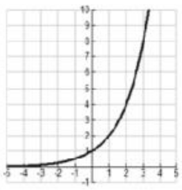
Once we have found a model that is appropriate, ask how ***strong*** it is.

- Look at the ***size of the residuals***.
  - Can be misleading when using re-expressions—difficult to interpret the actual size of the residuals—we care more about the pattern.
- Be careful about interpreting R^2
  - Note that it describes the model's effectiveness in accounting for the variability in ***re-expressed*** variables, not the original.
  - Correlation measures the strength of a linear association—can only talk about ***r*** if we find a re-expression that makes the relationship linear.

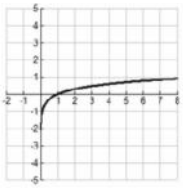
Re-expressing Data WATCH OUT!!!

- To write the correct equation for your model
  - Pay careful attention to the re-expression use.
  - Just knowing that the coefficients of the linear model are 1.2 and 0.55 is NOT enough. If you use logarithmic re-expression, the correct model is not just $\hat{y} = 1.2 + 0.55x$, its $\log(\hat{y}) = 1.2 + 0.55x$
  - Need to know that model represents *exponential* growth.
  - Must be able to make predictions from the equation. Here, start with a value of ***x*** = 2, find $\log(\hat{y}) = 1.2 + 0.55(2) = 2.3$
  - Now "backsolve" to get $\hat{y} = 10^{2.3} = 199.526 \approx 200$

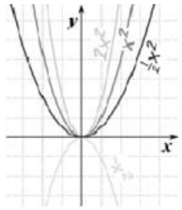
Recall:



Exponential function



Logarithmic function



Power function

Equivalent Models

Type of Model	Re-expression Equation	***Calculator's* Command**	***Curve* 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	$\hat{y} = ax^b$

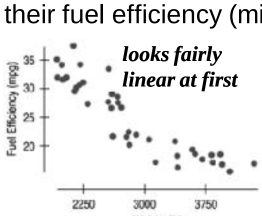
Equivalent Models

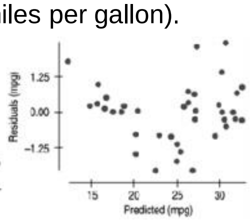
Type of Model	Model Equation	Transformation	Re-expression
Logarithmic	$\hat{y} = \boldsymbol{a} + \boldsymbol{b}\ln x$	$(\log \boldsymbol{x}, \boldsymbol{y})$	$\hat{y} = \boldsymbol{a} + \boldsymbol{b}\log \boldsymbol{x}$
Exponential	$\hat{y} = \boldsymbol{ab}^x$	$(\boldsymbol{x}, \log \boldsymbol{y})$	$\log \hat{y} = \boldsymbol{a} + \boldsymbol{bx}$
Power	$\hat{y} = \boldsymbol{ax}^b$	$(\log \boldsymbol{x}, \log \boldsymbol{y})$	$\log \hat{y} = \boldsymbol{a} + \boldsymbol{b}\log \boldsymbol{x}$

Straight to the Point

- We cannot use a linear model unless the relationship between the two variables is **linear**. Often re-expression can save the day, straightening bent relationships so that we can fit and use a simple linear model.
- If the relationship is nonlinear (which we can verify by examining the **residual plot**) we can try **re-expressing** the data.
- To re-express the data, we perform some mathematical operation on the data values such as taking the **reciprocal**, taking the **logarithm** , or taking the **square root**. Two simple ways to re-express data are with **logarithms** and **reciprocals**.
- Re-expressions (change of units, change of scale) can be seen in everyday life—everybody does it.

For example, consider the relationship between the weight of cars (in pounds) and their fuel efficiency (miles per gallon).

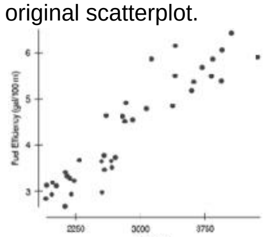


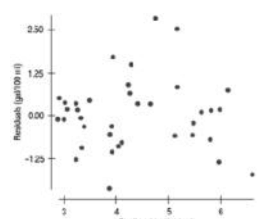


What do the scatterplot and residual plots reveal?

A look at the residuals plot shows a problem – a curved pattern – therefore, linear model is not appropriate.

If we take the **reciprocal** of the *y*-values (as gallons per hundred miles), we get the following scatterplot and residual plot and eliminate the bend in the original scatterplot.





- What do these plots reveal?
- ***That the relationship between weight and gal/100 mi (reciprocal of mpg) is linear.***

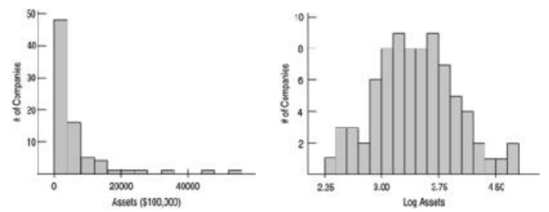
Goals of Re-expression

There are several reasons we may want to re-express our data:

1) To make the distribution of a variable more **symmetric**.
2) To make the **spreads** of several groups more alike.
3) To make the form of a scatterplot more **linear**.
4) To make the scatter in a scatterplot more **evenly spread** .

Goals of Re-expression

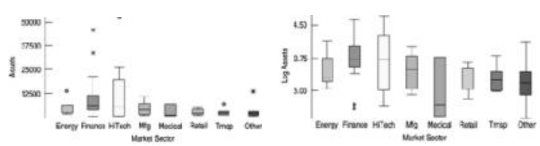
- **Goal 1:** Make the distribution of a variable (as seen in its histogram, for example) more **symmetric**.



- The skewed distribution is made much more nearly symmetric by taking logs.

Goals of Re-expression

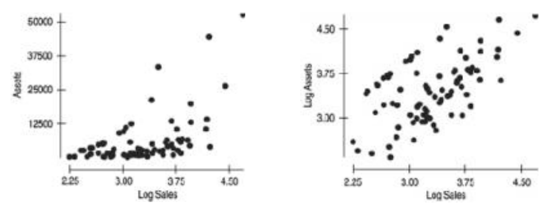
- **Goal 2:** Make the **spread** of several groups (as seen in side-by-side boxplots) **more alike** (not following like a fan shape), even if their centers differ.



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

Goals of Re-expression

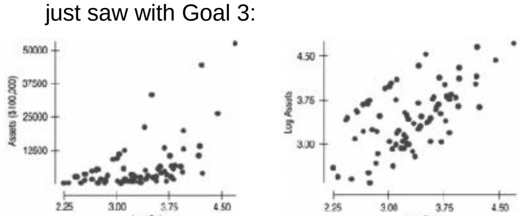
- **Goal 3:** Make the form of a scatterplot more **nearly linear**.



- The greater value of re-expression to straighten a relationship is that we can fit a linear model once the relationship is straight. This allows us to describe the relationship easier—allows us to use a linear model and all that goes with it.

Goals of Re-expression

- **Goal 4:** Make the scatter in a **scatterplot spread out evenly** rather than thickening at one end.
  - This can be seen in the two scatterplots we just saw with Goal 3:



- Groups that share a common spread are easier to compare.

Goals of Re-expression

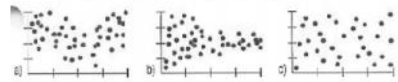
- **REMEMBER:** ***The model won't be perfect, but the re-expression can lead us to a useful model.***
- **You should recognize when the pattern of the data indicates that no re-expression can improve the structure of the data.**
- **You have to show how to re-express data with powers and how to find an effective re-expression for your data using the calculator.**
- **You should be able to reverse any of the common re-expressions to put a predicted value or residual back into original units.**

Goals of Re-expression

- **REMEMBER:** ***The model won't be perfect, but the re-expression can lead us to a useful model.***
- **You should be able to describe a summary or display of a re-expressed variable and clearly indicate how it was re-expressed and give its re-expressed units.**
- **You should be able to describe a regression model fit to re-expressed data in terms of the re-expressed variables.**

PRACTICE

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.



a) The residuals plot shows a curved pattern. Re-express to straighten the relationship.

b) The residuals plot shows a fan shape. Re-express to equalize spread.

c) The residuals plot shows no pattern. No re-expression is needed.

PRACTICE

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) $\log \hat{y} = 1.2 + 0.8 \log$

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) $\frac{1}{\sqrt{\hat{y}}} = 1.2 + 0.8x$

e)
$\log \hat{y} = 1.2 + 0.8 \log x$
$\log \hat{y} = 1.2 + 0.8 \log(2)$
$\log \hat{y} = 1.440823997...$
$\hat{y} = 10^{1.4408...}$
$\hat{y} = 27.59$

a)
$\ln \hat{y} = 1.2 + 0.8x$
$\ln \hat{y} = 1.2 + 0.8(2)$
$\ln \hat{y} = 2.8$
$\hat{y} = e^{2.8} = 16.44$

b)
$\sqrt{\hat{y}} = 1.2 + 0.8x$
$\sqrt{\hat{y}} = 1.2 + 0.8(2)$
$\sqrt{\hat{y}} = 2.8$
$\hat{y} = 2.8^2 = 7.84$

c)
$\frac{1}{\hat{y}} = 1.2 + 0.8x$
$\frac{1}{\hat{y}} = 1.2 + 0.8(2)$
$\frac{1}{\hat{y}} = 2.8$
$\hat{y} = \frac{1}{2.8} = 0.36$

d)
$\hat{y} = 1.2 + 0.8 \ln x$
$\hat{y} = 1.2 + 0.8 \ln(2)$
$\hat{y} = 1.75$

The Ladder of Powers

- There is a family of simple re-expressions that move data toward our goals in a consistent way. This collection of re-expressions is called the **Ladder of Powers**.
- The Ladder of Powers orders the ***effects*** that the re-expressions have on data.
- Members of the family line up in order.
  - The farther you move away from the original data (the "1" position), the greater the effect on the data.
- This fact allows you to search systematically for a re-expression that works, either stepping back from "1" or taking a step towards "1" as you see the results.

The Ladder of Powers

Power: 2

- Re-expression: y^2
- Comment: Use on left-skewed data

Power: 1

- Re-expression: y
- Comment: This is the raw data. No re-expression. Do not re-express the data if they are already well-behaved.

Power: 1/2

- Re-expression: $\sqrt{y}$
- Comment: Use on count data or when scatter in a scatterplot tends to increase as the explanatory variable increases.

The Ladder of Powers

Power: "0"

- Re-expression: $\log(y)$
- Comment: Not really the "0" power. Use on right-skewed data. Measurements cannot be negative or zero; values that grow by %; when in doubt, start here!

Power: -1/2, -1

- Re-expression: $\frac{1}{\sqrt{y}}, \quad -\frac{1}{y}$
- Comment: Use on right-skewed data. Measurements cannot be negative or zero. Use on ratios.

NOTE:
*The text lists very specific situations for which each of these might be an appropriate transformation, but we are not bound by these guidelines, as the ultimate goal is to **find a transformation that works**!*

The Ladder of Powers

Power	Name	Comment
2	Square of data values	Try with unimodal distributions that are skewed to the left.
1	Raw data	Data with positive and negative values and no bounds are less likely to benefit from re-expression.
½	Square root of data values	Counts often benefit from a square root re-expression.
"0"	We'll use logarithms here	Measurements that cannot be negative (salaries, population) often benefit from a log re-expression.
–½	Reciprocal square root	An uncommon re-expression, but sometimes useful.
–1	The reciprocal of the data	Ratios of two quantities (e.g., mph) often benefit from a reciprocal.

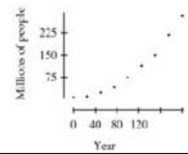
Example 1) (x, y)

consider the population growth in the US.

- We scale the years as we enter the data. We could use 1, 2, 3, 4, … or 0, 25, 50, … (Caution! Be careful using 0 or negative numbers as data values. Taking logs of 0 or negative values can make some points "go missing" and just disappear quietly from the analysis.)

Year	Population (millions)
1800	5
1825	11
1850	23
1875	44
1900	76
1925	114
1950	151
1975	215
2000	285

- We begin with the scatterplot and see a clear curve, concave upward.
- The association between year (measured in years since 1800) and U.S. population (in millions) is strong positive and curved. We cannot use a regression line to model this relationship without re-expressing the data first.



Example 1)

consider the population growth in the US.

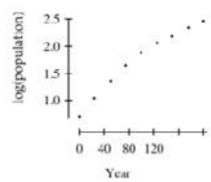
- Now we use our Ladder of Powers. First we'll try the zero power, the logarithm of the population. We start there because we suspect that population might increase by a roughly equal percentage each year (and hence that growth is exponential), or simply because it's a good place to start if we're not sure what to do.

Example 1) $(x, \log y)$

consider the population growth in the US.

- The change in the new scatterplot is dramatic. The scatterplot of log(population) and year still has a curve and it bends in the wrong way. We have gone too far on the Ladder of Powers.

- This is a clear indication that we have gone too far on the ladder and should retreat toward the original data (the "1" rung). That suggests the 1/2 power, so we find the square roots of the populations and plot them against the years.



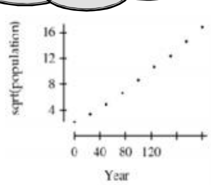
Example 1) $(x, \sqrt{y})$

consider the population growth in the US.

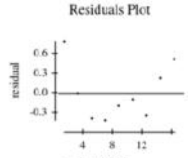
- The scatterplot of sqrt(population) and year is still a bit curved, but straight enough to fit a line. The model

$$\sqrt{\hat{Pop}} = 1.46275 + 0.07457(Year)$$

- Has a high value of R^2, and, although the residuals plot shows some pattern, the residuals are all very small. This is a good model.



Residuals Plot



Dependent variable is: $\sqrt{Pop}$
No Selector
R squared = 99.3% R squared (adjusted) = 99.2%
s = 0.4500 with 9 - 2 = 7 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	208.526	1	208.526	1030
Residual	1.41756	7	0.202508	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	1.46275	0.2766	5.29	0.0011
Year	0.074570	0.0023	32.1	≤ 0.0001

Example 2)

- During a science lab, students heated water, allowed it to cool, and recorded the temperature over time. They computed the difference between the water temperature and the room temperature. The results are in the table.

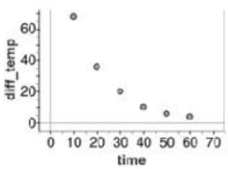
Time (in minutes)	10	20	30	40	50	60
Difference in temp. (degrees F)	68	36	20	10	6	4

1) Sketch a scatterplot.
2) Newton's Law of Cooling suggests an exponential function is appropriate. Reexpress the data using logarithms and sketch a new scatterplot.
3) Write the equation of the least-squares regression line for the transformed data. Draw the regression line on the scatterplot in question 2.
4) Use the equation to predict the difference in temperature after 45 minutes.
5) Use the equation to predict the difference in temperature at time 0 minutes. What does this value represent?

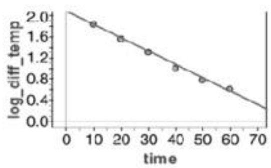
Example 2)

Time (in minutes)	10	20	30	40	50	60
Difference in temp. (degrees F)	68	36	20	10	6	4

1) Sketch a scatterplot.



2) Newton's Law of Cooling suggests an exponential function is appropriate. Reexpress the data using logarithms and sketch a new scatterplot.



Example 2)

Time (in minutes)	10	20	30	40	50	60
Difference in temp. (degrees F)	68	36	20	10	6	4

3) Write the equation of the least-squares regression line for the transformed data. Draw the regression line on the scatterplot in question 2.

$(x, \log y)$

$\log(\widehat{difftemp}) = 2.057 - 0.025time$

4) Use the equation to predict the difference in temperature after 45 minutes.

$\log(\widehat{difftemp}) = 2.057 - 0.025(45) = 0.932$

$\widehat{difftemp} = 10^{0.932} = 8.551$

Recall:

$\log y = x \Leftrightarrow 10^x = y$

5) Use the equation to predict the difference in temperature at time 0 minutes. What does this value represent?

$\log(\widehat{difftemp}) = 2.057$

$\widehat{difftemp} = 10^{2.057} = 114.025$

- This represents the model's prediction of the difference in the temperature at the beginning of the experiment.

Goals of Re-expression

1. World Population (United Nations database)
   a) Model (Scale the years!)

$\sqrt{\widehat{Pop}} = 21.35 + 0.566(Years\ since\ 1900)$

   b) Prediction for 2005? ________

6525 million people.

Year	Population (millions)
1950	2519
1955	2755
1960	3020
1965	3334
1970	3691
1975	4066
1980	4430
1985	4825
1990	5255
1995	5662
2000	6057

Goals of Re-expression

2. Mortgages (Republic National Bank, founded 1970)
   a) 1970 – 1988 Model (Scale the years somehow…)

$\log(\widehat{Mortgage}) = 0.1462 + 0.07404(Years\ since\ 1970)$

   b) Compare post-1990 mortgages to the previous trend.

- The model predict the mortgage amounts in 1990, 1995, and 2000 to be $42.4 million, $99.3 million, and $233.0 million, respectively. These predictions are all much higher than the actual amounts.
- The model is not valid for these years. .

Year	Million $
1970	1.2
1972	2.5
1974	2.9
1976	3.1
1978	5.8
1980	8.3
1982	10.8
1984	14.7
1986	21.8
1988	29.7
1990	32.4
1995	39.5
2000	49.7

Goals of Re-expression

3. Light Intensity
   a) Model

$\frac{1}{\sqrt{\widehat{Intensity}}} = 0.00006 + 0.022(Distance)$

   b) Intensity at
   1'? 2,136 cp
   12'? 14.8 cp
   30'? 2.36 cp

Distance	Candlepower
2 feet	531.2
5	84.3
8	33.6
10	21.1
15	9.5
20	5.3
25	3.4

Re-expressing Data Using Logarithms

- An equation of the form ***y = a + bx*** is used to model **linear** data.
- The process of transforming nonlinear data into linear data is called **linearization**.
- In order to linearize certain types of data we use properties of **logarithms**.

- **PROBLEM:** We cannot use least-squares regression for the nonlinear data because least-squares regression depends upon correlation, which only measures the strength of **linear** relationships.

- **SOLUTION:** We transform the *nonlinear data* into *linear data*, and then use least-squares regression to determine the best fitting **line** for the transformed data.
- Finally, do a **reverse** transformation to turn the linear equation back into a nonlinear equation which will model our original *nonlinear data*.

Linearizing Exponential Functions:

(We want to write an exponential function of the form $y = a \cdot b^x$ as a function of the form $y = a + bx$).

$y = a \cdot b^x$ (***x*** , ***y*** are *variables* and ***a*** , ***b*** are constants)

- This is in the general form ***y = a + bx***, which is linear.
- So, the graph of (var1, var2) is linear. This means the graph of (x, log y) is linear.

CONCLUSIONS:

- If the graph of **log y vs. x** is linear, then the graph of **y vs. x** is exponential.
- If the graph of **y vs. x** is exponential, then the graph of **log y vs. x** is linear.
- Once we have linearized our data, we can use least-squares regression on the transformed data to find the best fitting linear model.

PROPERTIES OF LOGARITHMS:

- 1) $\log(AB) =$
- 2) $\log\left(\dfrac{A}{B}\right) =$
- 3) $\log x^p =$

Plan B: Attack of the Logarithms

- 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*-variable.
- Then re-express the data using some combination of *x* or log(*x*) vs. *y* or log(*y*).

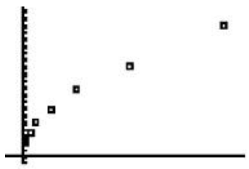
Plan B: Attack of the Logarithms

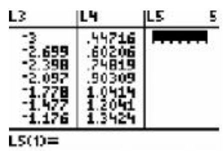
Model Name	x-axis	y-axis	Comment
Exponential	x	$\log(y)$	This model is the "0" power in the ladder approach, useful for values that grow by percentage increases.
Logarithmic	$\log(x)$	y	A wide range of *x*-values, or a scatterplot descending rapidly at the left but leveling off toward the right, may 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.

Let's Try It! (p. 233)

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

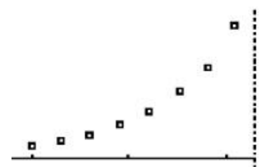
- Shutter speed and *f*/stop of the lens
  - L1: shutter speed
  - L2: *f*/stop
- Curved stat plot
  - Try logarithms
  - Take log of L1→L3
  - Take log of L2→L4

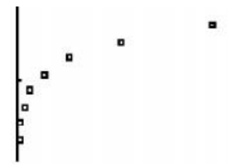




Let's Try It! (p. 233)

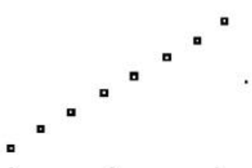
- Scatterplot #1: Xlist→L3, Ylist→L2
- Scatterplot #2: Xlist→L1, Ylist→L4

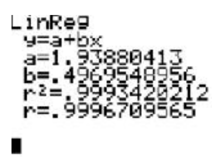




Let's Try It! (p. 233)

- Use Scatterplot #3: LinReg L3, L4
- LinReg L3, L4





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

PRACTICE:

- Linearize the Case 1 data and find the least-squares regression line for the transformed data.

x (mos.)	0	48	96	144	192	240
y ($)	100	161.22	259.93	419.06	675.62	1089.30

PRACTICE:

- Then, do a reverse transformation to turn the linear equation back into an exponential equation.
- Compare this to the equation the calculator gives when performing exponential regression on the Case 1 data

Linearizing Power Functions:

(We want to write a power function of the form as a function of the form $y = a + bx$).

$y = ax^b$ (***x*** , ***y*** are variables and ***a*** , ***b*** are constants)

- This is in the general form **y = a + bx**, which is linear.
- So, the graph of $(\log x, \log y)$ (var1, var2) is linear. This means the graph of is linear.

Case 2: Consider the following set of Nonlinear Data representing the average length and weight at different ages for Atlantic Ocean rockfish:

x: age (years)	0	4	8	12	16	20
y: weight (grams)	0	48	192	432	768	1200

PRACTICE:

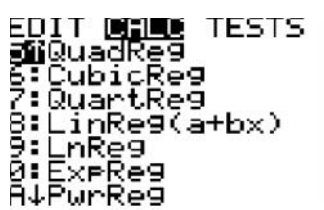
- Linearize the data for Case 2 and find the least-squares regression line for the transformed data.
- Linearize the data for Case 2 and find the least-squares regression line for the transformed data.
- Then, do a reverse transformation to turn the linear equation back into a power equation.
- Compare this to the equation the calculator gives when performing power regression on the Case 2 data.

Multiple Benefits

- We often choose a re-expression for one reason and then discover that it has helped other aspects of an analysis.
- For example, a re-expression that makes a histogram more symmetric might also straighten a scatterplot or stabilize variance.
- A single re-expression may improve each of our goals at the same time.
- Re-expression certainly simplifies efforts to analyze and understand relationships.
- Simpler explanations and simpler models tend to give a true picture of the relationship.

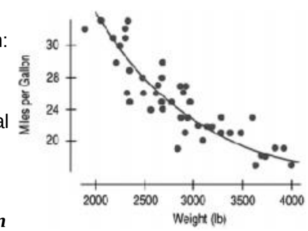
TI Tips

- Regressions that automatically and appropriately re-express the data:



Why Not Just Use a Curve?

- If there's a curve in the scatterplot, why not just fit a curve to the data?
- Benefits to linear approach:
  - Contextual meaning of slope and *y*-intercept
  - More advanced statistical methods for analyzing linear associations
- ***It is usually better to re-express the data to straighten the plot.***

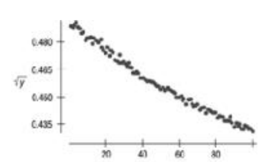


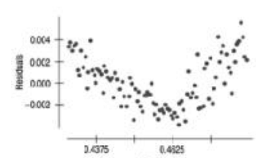
Why Not Just Use a Curve?

- The mathematics and calculations for "curves of best fit" are considerably more difficult than "lines of best fit."
- Besides, straight lines are easy to understand.
  - We know how to think about the slope and the *y*-intercept.

What Can Go Wrong?

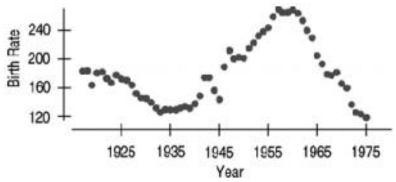
- Don't expect your model to be perfect.
- Don't stray too far from the ladder.
- Don't choose a model based on R^2 alone:





What Can Go Wrong?

- Beware of multiple modes.
  - Re-expression cannot pull separate modes together.
- Watch out for scatterplots that turn around.
  - Re-expression can straighten many bent relationships, but not those that go up then down, or down then up.



What Can Go Wrong?

- 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.
- 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. If all the data values are large (e.g., years), consider subtracting a constant to bring them back near 1.
  - Re-expressing data with a range from 1 to 1000 is far more effective than re-expressing data with a range of 100,000 to 100,100.

What have we learned?

- When the conditions for regression are not met, a simple re-expression of the data may help.
- A re-expression may make the:
  - Distribution of a variable more symmetric.
  - Spread across different groups more similar.
  - Form of a scatterplot straighter.
  - Scatter around the line in a scatterplot more consistent.

What have we learned? (cont.)

- 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.
- Our models won't be perfect, but re-expression can lead us to a useful model.