AP Statistics – Lesson Notes - Chapter 10: Re-expressing Data

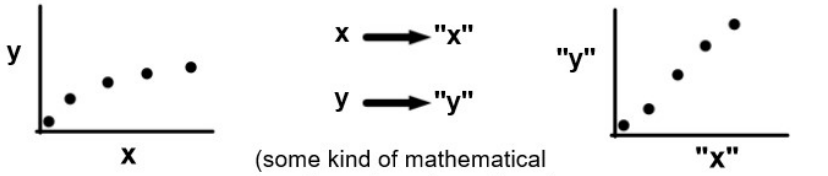
Re-expressing data

Sometimes the data collected is not in a form that allows us to use our analysis tools. A histogram may be too skewed to use Normal distribution percentage calculations, or a scatterplot may not be linear enough to use linear regression techniques.

In these cases, sometimes we can apply a mathematical function to the data to 're-express' or 'transform' the data to a set of values which allow us to use our analysis tools.

Textbook vs simpler method

The textbook uses a variety of function types to straighten data, but nearly all cases can be handled by just two functions: exponential and power functions.



Original data

(some kind of mathematical transformation of x and/or y)

Transformed data
Re-expressed
Straightened

Our method...just two transformation models...

Exponential Model

$$y = k^x$$

$$\log(y) = \log(k^x)$$

$$\log(y) = x\log(k)$$

Take Log of Y only to straighten the data.

Power Model

$$y = x^k$$

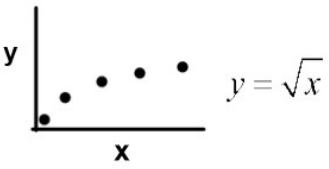
$$\log(y) = \log(x^k)$$

$$\log(y) = k\log(x)$$

Take Log of X and Y to straighten the data.

Here is how it works

Let's say you have some data which is curving in a way that resembles a square root function...



$y = \sqrt{x}$

This can be represented by a power model...

$$y = x^{1/2}$$

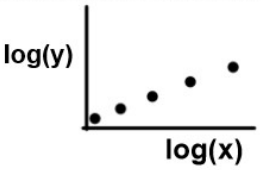
...so if we took the log of both sides...

$$\log(y) = \frac{1}{2}\log(x)$$

...we could define new variables for x and y which are the log of the original variables...

$$"y" = \log(y)$$
$$"x" = \log(x)$$

$$"y" = \frac{1}{2}"x"$$



...which 'straightens' the data.

We don't need to know which model will work best ahead of time

We just try both the exponential and the power model and judge which one does a better job of straightening the data by fitting an LSRL to each straightened model and examining the residual.

The model which results in the least 'pattern' in the residuals straightens the data best and is the one we choose to use.

Exponential Model	**Power Model**
Take Log of Y only to straighten the data.	Take Log of X and Y to straighten the data.

Example 1: Find an LSRL to model the data.

x	y
1/1000	2.8
1/500	4
1/250	5.6
1/125	8
1/60	11
1/30	16
1/15	22
1/8	32

1) Enter the data into L1, L2 and display a scatterplot.

2) Try a Linear Regression (LinReg).
What is the LSRL equation? What is r^2?

3) Plot the residuals. Should we use a linear model for this data?

4) Use list L3 to store log(x) and list L4 to :store log(y)
- Edit L3 to be: L3=LOG(L1)
- Edit L4 to be: L4=LOG(L2)

We'll go ahead and do both because we aren't sure yet whether the exponential model or the power model will straighten the data better.

L1 x	L2 y	L3 log(x)	L4 log(y)
1/1000	2.8	-3	.44716
1/500	4	-2.699	.60206
1/250	5.6	-2.398	.74819
1/125	8	-2.097	.90309
1/60	11	-1.778	1.0414
1/30	16	-1.477	1.2041
1/15	22	-1.176	1.2324
1/8	32	-.9031	1.5051

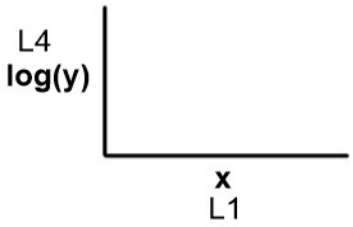
5) Try using an exponential model: $\hat{y} = k^x$

- take log of both sides: $\log\left(\hat{y}\right) = \log\left(k^x\right)$
- move exponent down: $\log\left(\hat{y}\right) = x\log\left(k\right)$

For this model, we use the original x, but the logarithm of y.

L1 and L4

6) Look at the scatterplot of L4 vs. L1:



Does it look straightened?

7) Now try a linear regression (LinReg) on L1,L4. What is the LSRL equation? What is r^2?

8) Plot the residuals. Is this a good linear model?

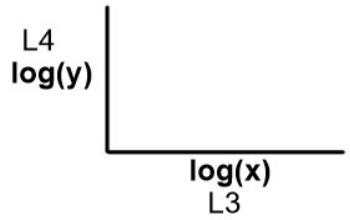
9) Now try using a power model: $\hat{y} = x^k$

- take log of both sides: $\log(\hat{y}) = \log(x^k)$
- move exponent down: $\boxed{\log(\hat{y})} = k\boxed{\log(x)}$

For this model, we use the logarithm of x, and the logarithm of y.

L3 and L4

10) Look at the scatterplot of L4 vs. L3:



Does it look straightened?

11) Now try a linear regression (LinReg) on L3,L4. What is the LSRL equation? What is r^2?

12) Plot the residuals. Is this a good linear model?

Based on the residuals, which is better power or exponential?

13) Using the best model: If x = 1/40 what is the predicted y?

Example 2: Find an LSRL to model the data.

x	y
1	27
2	36
3	45
4	181
5	306
6	1093
7	2459

1) Enter the data into L1, L2 and display a scatterplot.

2) Try a Linear Regression (LinReg).
What is the LSRL equation? What is r^2?

3) Plot the residuals. Should we use a linear model for this data?

L1	L2	L3	L4
x	y	log(x)	log(y)
1	27	0	1.4314
2	36	.30103	1.5563
3	45	.47712	1.6532
4	181	.60206	2.2577
5	306	.69897	2.4857
6	1093	.77815	3.0386
7	2459	.8451	3.3908

4) Use list L3 to store log(x) and list L4 to :store log(y)

- Edit L3 to be: L3=LOG(L1)
- Edit L4 to be: L4=LOG(L2)

We'll go ahead and do both because we aren't sure yet whether the exponential model or the power model will straighten the data better.

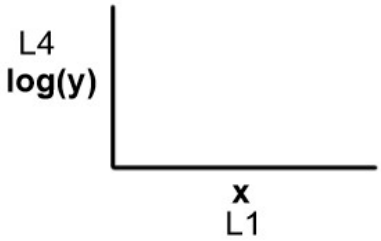
5) Try using an exponential model: $\hat{y} = k^x$

- take log of both sides: $\log(\hat{y}) = \log(k^x)$
- move exponent down: $\log(\hat{y}) = x\log(k)$

For this model, we use the original x, but the logarithm of y.

L1 and L4

6) Look at the scatterplot of L4 vs. L1:



Does it look straightened?

7) Now try a linear regression (LinReg) on L1,L4. What is the LSRL equation? What is r^2?

8) Plot the residuals. Is this a good linear model?

9) Now try using a power model: $\hat{y} = x^k$

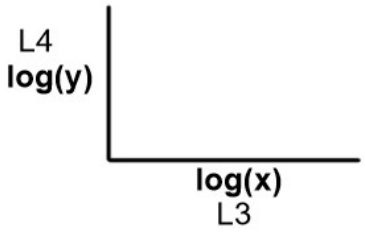
- take log of both sides: $\log(\hat{y}) = \log(x^k)$

- move exponent down: $\boxed{\log(\hat{y})} = k\boxed{\log(x)}$

For this model, we use the logarithm of x, and the logarithm of y.

L3 and L4

10) Look at the scatterplot of L4 vs. L3:



Does it look straightened?

11) Now try a linear regression (LinReg) on L3,L4. What is the LSRL equation? What is r^2?

12) Plot the residuals. Is this a good linear model?

Based on the residuals, which is better power or exponential?

13) Using the best model: If x = 8 what is the predicted y?

Summarizing...

When a scatterplot shows data is not-linear we can try straightening the data by re-expressing either or both variables as the logarithms of the data values.

Power Model: Try taking logs of both x and y.

Exponential Model: Try taking log of only y and using original x.

(You can also try taking log of only x.)

Use whichever method results in:
- A scatterplot which looks straight.
- Produces a LinReg with a high coefficient of determination r^2.
- Has a residuals plot with no pattern.

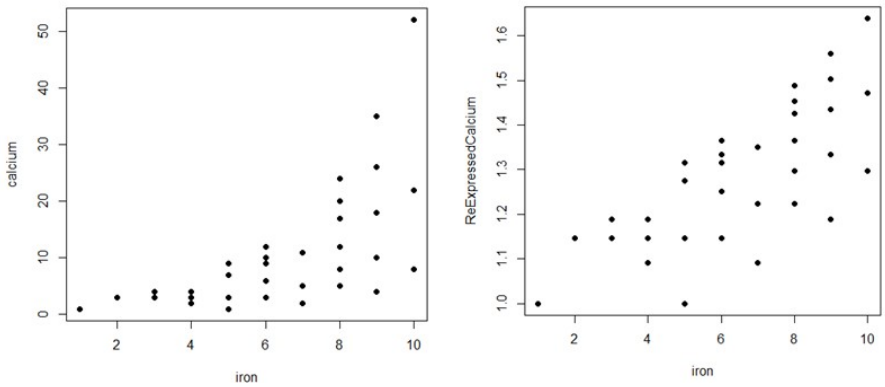
When do we try re-expressing data?

There are 4 general 'goals' in re-expressing data:

a) To make a scatterplot more linear.
b) To make the scatter in a scatterplot spread out evenly rather than in a fan shape.
c) To make the distribution of a variable more symmetric (e.g. a histogram).
d) To make the spread of several groups more alike (in side-by-side boxplots).

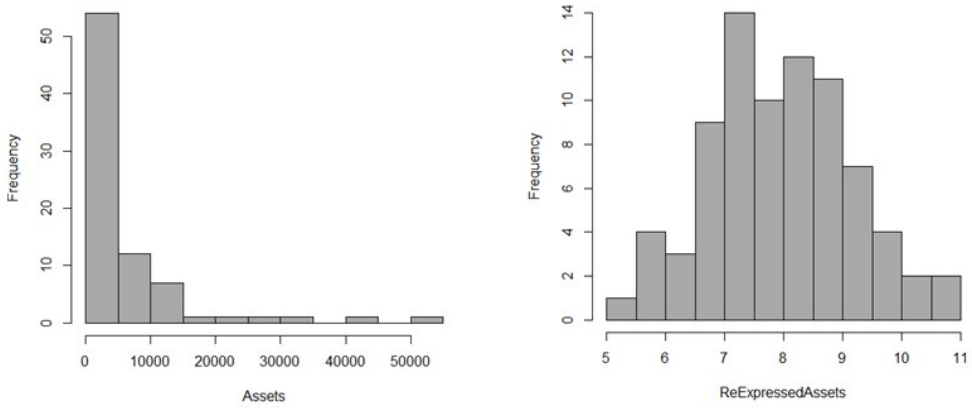
Example: a) Making a scatterplot more linear
b) Make a fan-shaped scatterplot spread more evenly

The following scatterplot compares bloodstream calcium and iron levels:



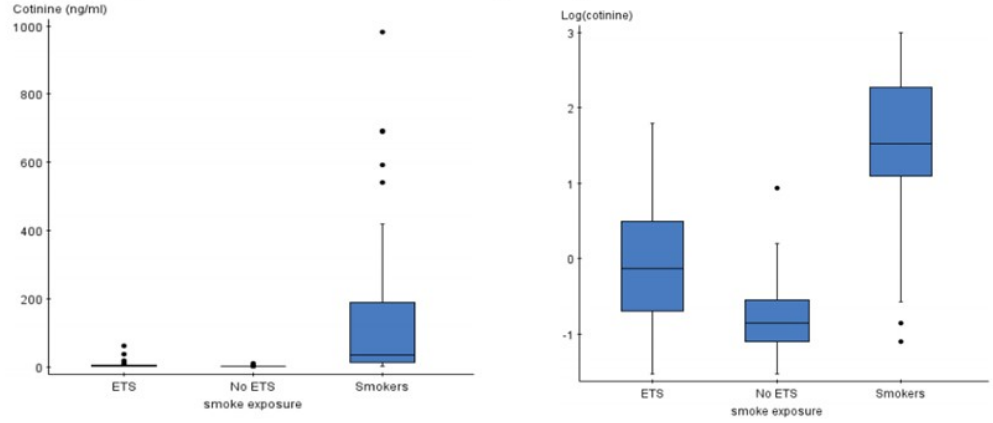
Example: c) Making a distribution of a variable more symmetric

The following is a histogram of data showing the number of companies with different amounts of assets:



Example: d) Making the spread of several groups more alike

The following boxplots show bloodstream concentrations of Cotinine (a nicotine metabolite) for 3 groups (smokers, non-smokers with no 2nd hand smoke exposure, non-smokers with 2nd hand exposure)



Caution: Remember that to avoid misrepresentation, you need to re-transform results back to original units.

Why not just fit to a curve?

Our calculator can fit data to a curve. Why not just use the curve model instead of straightening and them modeling with a line? There are a few reasons:

- With lines we know how to interpret slope which we lose if we don't have a line.
- Re-expressing data allows for benefits other than fitting an equation to a dataset (for example, making histograms symmetrical and spreads more even).
- In later chapters we will learn additional techniques, some of which only work with linear models.