AP Stats – Chap 9
Re-expressing Data: Get It Straight!

- Re-expressions "think" about the data differently but **DO NOT** change what they mean.

- **4 Goals of Re-expression**
  - Make the distribution of a variable (histogram) more symmetric
  - Make the spread of several groups (boxplots) more alike
  - Make a scatterplot more nearly linear
  - Make a scatterplot spread out evenly rather than following a fan shape

- Occam's Razor – Simpler explanations are likely to be the better ones

- Don't expect **ANY** model to be perfect! You aren't looking for the "right" model. You're looking for a "useful" one!

- Don't choose a model based on the R-squared alone.

- Watch out for scatterplots that turn around.

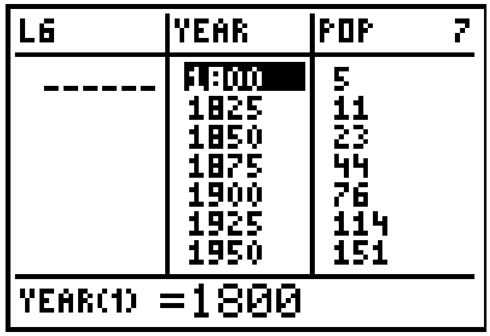
- Watch out for negative or zero data values when re-expressing.

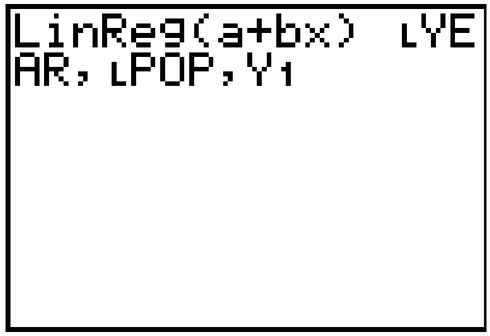
Population Growth in the US:

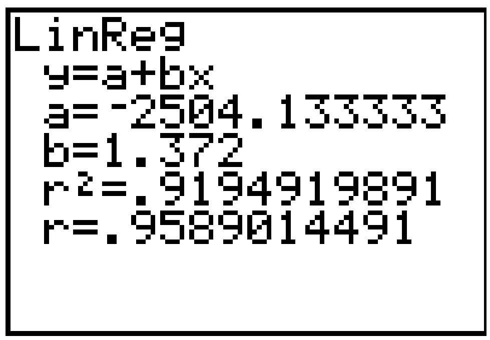
Year	Population (millions)
1800	5
1825	11
1850	23
1875	44
1900	76
1925	114
1950	151
1975	215
2000	285

Population Growth in the US:

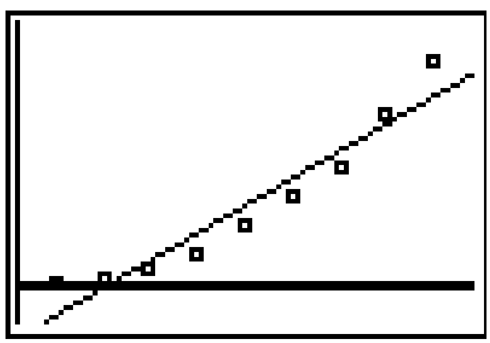
Year	Population (millions)
1800	5
1825	11
1850	23
1875	44
1900	76
1925	114
1950	151
1975	215
2000	285



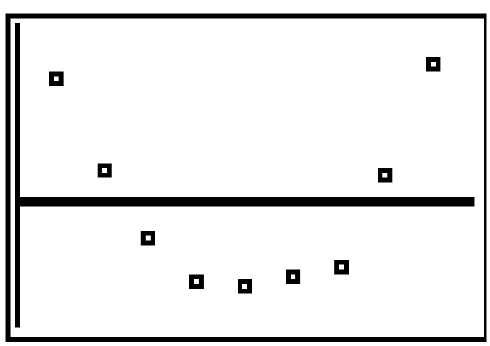




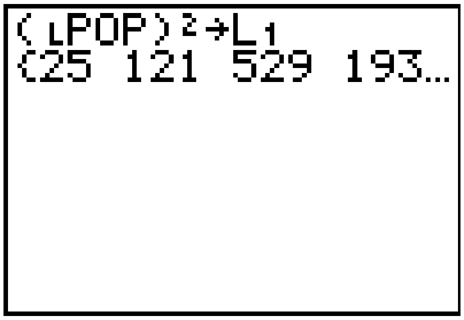
original scatterplot:

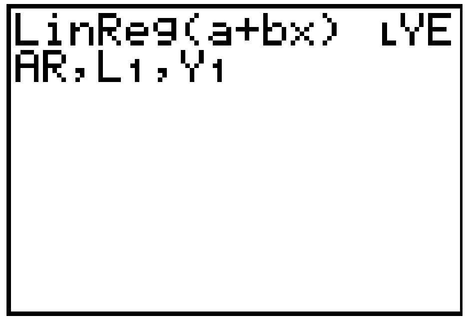


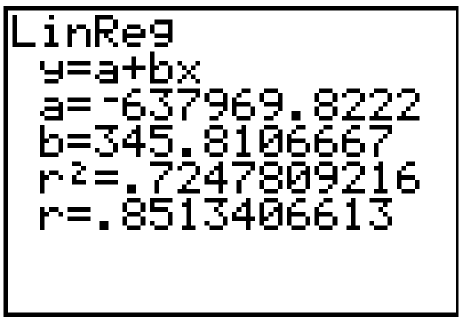
original residual plot:

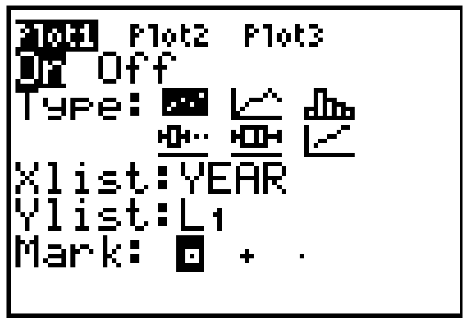


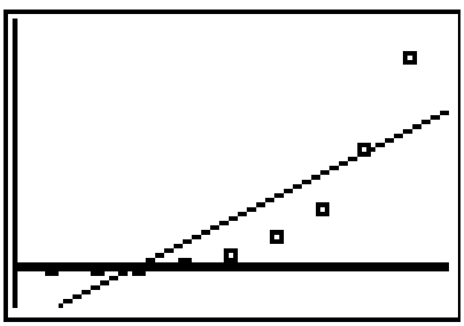
let’s try moving up the ladder to the square power.

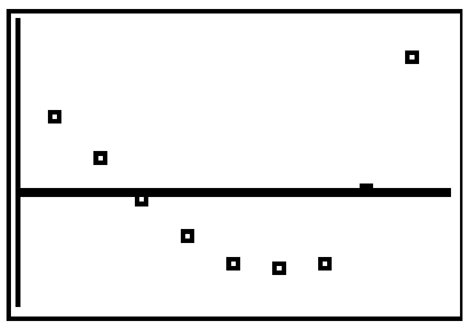






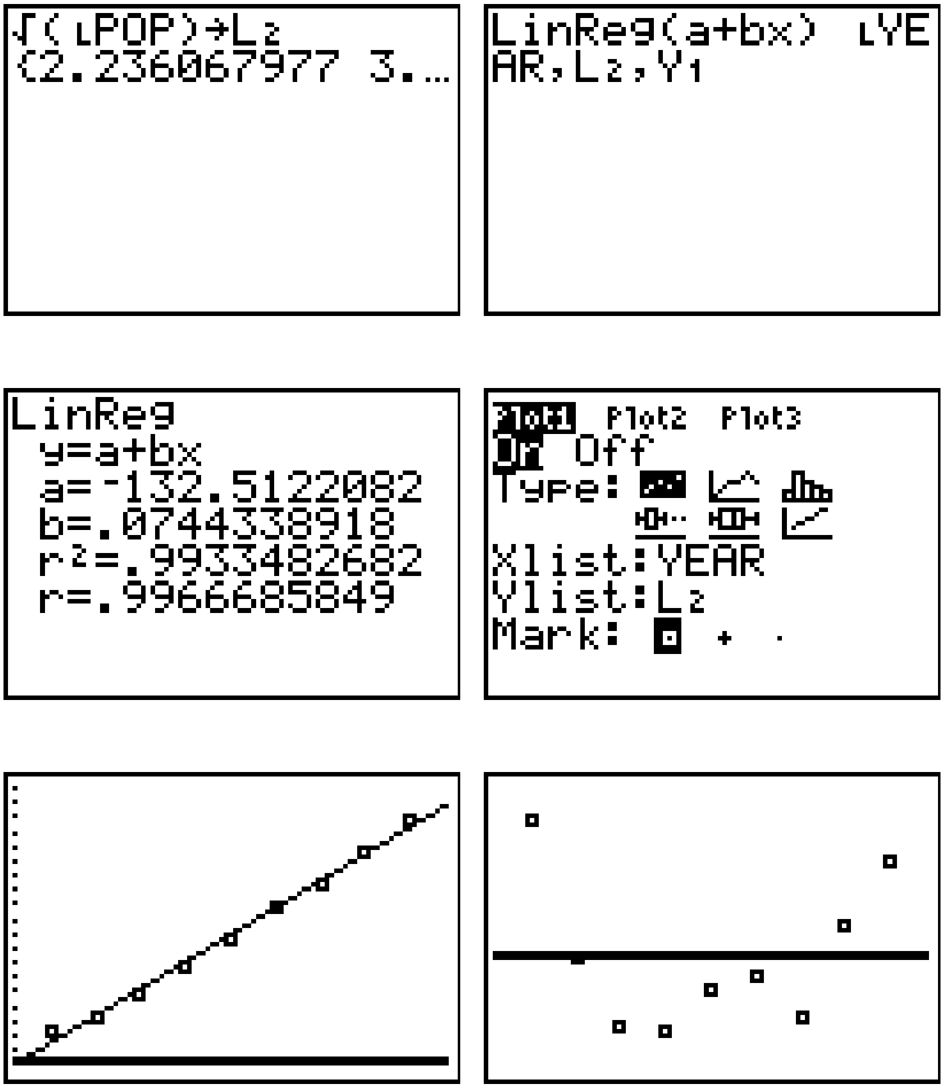






not any better. ☹

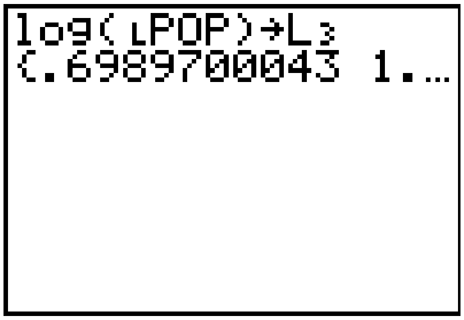
let’s try moving in the other direction on the ladder...the ½ (square root) power.

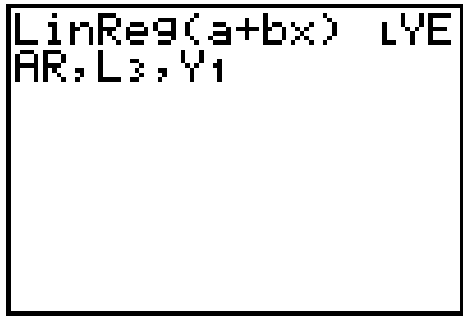


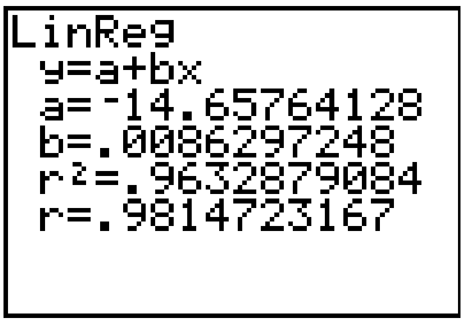
better scatterplot, but still a pattern in the residuals.

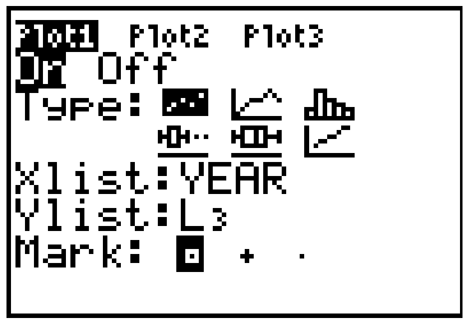
can we do better if we keep moving down the ladder?

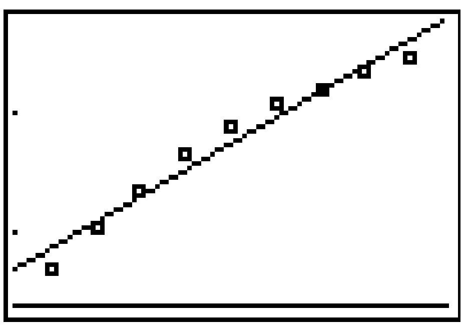
now...let’s try the “0” (log) rung.











STOP! scatterplot starts to bend back the other way! this is a sign that we went too far.

so...square-root re-expression is the best we can do.

Model equation: $\sqrt{\widehat{\text{pop}}} = -132.51 + .0744(\text{year})$

To estimate a year, you need to remove the re-expression as the **LAST** step!

$$\sqrt{\widehat{\text{pop}}} = -132.51 + .0744(2005)$$

$$\sqrt{\widehat{\text{pop}}} = 16.662$$

$$\widehat{\text{pop}} = 277.62$$

The model would estimate approximately 277.62 million people in 2005.

The Ladder of Powers

Power	Name	Comment
2	The square of the data values, y^2.	Try this for unimodal distributions that are skewed to the left.
1	The raw data – no changes at all. This is "home base." The farther you step from here up or down the ladder, the greater the effect.	Data that can take on both positive and negative values with no boundaries are less likely to benefit from re-expression.
½	The square root of the data values, $\sqrt{y}$.	Counts often benefit from a square root re-expression. For counted data, start here.
"0"	Although mathematicians define anything to the zero power as equal to one, for us this place is held for logarithms.	Measurements that cannot be negative, and especially values that grow by percentage increases (salaries or populations) often benefit from a log re-expression. When in doubt, start here. If your data have zeros, try adding a small constant to all values before finding the logs.
- ½	The negative reciprocal square root, $\frac{-1}{\sqrt{y}}$.	An uncommon re-expression, but sometimes useful. Changing the sign to take the negative of the reciprocal square root preserves the direction of the relationship, which can be a bit simpler.
-1	The negative reciprocal, $\frac{-1}{y}$.	Ratios of two quantities (miles per hour, for example), often benefit from a reciprocal. You have about a 50-50 chance that the original ratio was taken in the "wrong" order and would benefit from this re-expression. Change the sign if you want to preserve the direction of the relationships. If your data have zeros, try adding a small constant to all values before finding the reciprocal.

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.