Review:

Suppose we wanted to model a teacher's salary (y) based on the number of years he has been teaching (x). Using a random sample of 13 salaries (in 1000's), the following model was calculated:

$\hat{y} = 45.59 + 0.798x \qquad r^2 = .587 \qquad s_e = 5.31$

Interpret the slope and constant term of this model:

Interpret the values of r^2 and s_e.

Interpret the value of $\hat{y}$ when x = 3.

When people have colds, common wisdom says to avoid dairy products since they produce extra mucus. However, this claim was challenged by researchers in Australia in 1990. They infected volunteers with a cold virus and assigned them at random to drink different amounts of milk. Meanwhile, all used tissues were weighed to measure nasal mucus secretions (based on Nutrition Action 12-04).

Glasses of Milk	Ounces of Mucus
0	0.7
0	0.3
0	0.6
1	0.5
1	0.8
1	1.3
2	0.2
2	0.9
2	1.1
3	0.5
3	0.6
3	1.2
4	0.5
4	0.8
4	1.3

a. How many factors were used in this experiment?

b. How many levels?

c. What are the treatments?

d. What are the experimental units?

e. How is randomization used?

f. How is replication used?

g. Sketch the scatterplot and least squares regression line.

h. Interpret the slope of the LSRL in context.

i. Interpret the value of r^2 in context.

j. Interpret the value $s_e = 0.35$ in context.

13.2 Inference for the Slope of a Least Squares Line

In most studies, when we use a data set to calculate regression lines, the data we are using comes from a random sample or a randomized experiment. Thus, the values we find are really just estimates of the *true* values.

Notation:

Estimated LSRL:	$\hat{y} = a + bx$	or $\hat{y} = b_0 + b_1 x$
True LSRL:	$\hat{y} = \alpha + \beta x$	or $\hat{y} = \beta_0 + \beta_1 x$
Estimated correlation coefficient:	r	
True correlation coefficient:	ρ ("rho")	
Estimated standard deviation:	s_e	
True standard deviation:	σ_e	

5 Steps:

1. I will conduct a linear regression t-test for β (α =.05).

2. Ho: $\beta = 0$
   Ha: $\beta > 0$

   Note: Many questions will ask if there is a "useful linear relationship" between two variables. This is simply asking if there is a linear association that is useful for descriptions, predictions, etc. In this case, the alternate hypothesis would be $\beta \neq 0$.

3. Conditions:
   a. Treatments assigned at random? Given.
      Note: for observational studies, the condition is: "random sample from population of interest?"

   b. The linear model is appropriate?

   c. The variability of the residuals stays constant for all values of x?

d. The residuals are approximately normally distributed

4. Test statistic: $t = \frac{b - \beta}{s_b}$ with n - 2 degrees of freedom

- s_b is the standard deviation of the slope. It describes how variable the sample slope is for the given sample size, etc. This is the variability we saw in our dotplot earlier.
- This is different than s_e which is the standard deviation of the residuals, although they are related:

$$s_b = \frac{s_e}{s_x\sqrt{n-1}} = \frac{s_e}{\sqrt{\sum(x-\bar{x})^2}} = \frac{\sqrt{\frac{\sum(y-\hat{y})}{n-2}}}{\sqrt{\sum(x-\bar{x})^2}}$$

- TI-83: LinRegTTest: enter the lists, freq = 1, Ha, and leave RegEQ blank.
- Note: The TI-83 and most statistics software only test for $\beta = 0$, so if you are testing for something else, you need to calculate t by hand.
- Note: on the TI-83, s is s_e, not s_b!
- Note: on the TI-83, after a regression line is calculated, the residuals are stored in a list called Resid. Use this list to check the conditions.

13.2 Confidence Intervals for the True Slope

Suppose that a researcher was studying the relationship between the outdoor temperature in ˚C (x) and the depth of a woodpeckers nest in cm (y). Based on a sample of 12 nests, the slope was -.345 with a standard deviation of .06. Does this give evidence that woodpeckers dig deeper when it's colder outside? You may assume the conditions for inference have been met.

Suppose that previous research suggested that for every increase of 1˚C, the depth of a nest decreased by one-third of a centimeter. Does this new study disagree with previous research?

Confidence Intervals for the Slope of a Least Squares Line

1. We will calculate a 95% t-interval for β.

2. Conditions: same as for the hypothesis test

3. CI $= b \pm t^*(s_b)$ where t* has df = n – 2

Interpret the slope in the context of the problem.

Could we use this interval to answer both questions above?

Ex: The following is weather information for a random sample of SoCal cities:

Average High Temp (in ˚F):	70	71	73	74	76	76	77	72	72
Annual Precipitation (in inches):	15	13	9	12	8	10	7	11	12

Estimate the average change in precipitation associated with an increase of 1˚F in average high temperature for Southern California cities.

- Note: The output on the TI-83 does not give s_b (it gives $s = s_e$). But, since $t = \dfrac{b-0}{s_b}, s_b = \dfrac{b}{t}$

5.5 Modeling Non-Linear Data

When statisticians look at a scatterplot, they often use the phrases "signal" and "noise." The signal is the underlying form of the data and the noise is the random variation from that form.

For example, when we studied the relationship between fat and calories in pizza, we saw that pizza with more fat has more calories (that's the signal) but even pizzas with the same amount of fat can have different amounts of calories (that's the noise).

If the residual plot looks randomly scattered, then our model has captured the whole signal and only the noise remains. If the residual plot shows a pattern, however, then our model has missed some of the signal and we should seek a better model.

Therefore, when deciding which model to use, the most important thing to do is consider the residual plot for each model.

Here are some data showing the highest individual baseball salary (in millions of dollars) for various years between 1980 and 2001.

Years Since 1980	Salary ($ millions)
0	1
2	2.04
9	3
10	4.7
11	5.3
16	8.5
17	11
18	12.5
19	15
21	25.2

Note: We often shift the data on the x-axis so it is closer to the origin. This makes the computations and interpretations easier in many cases. In this case, x = 17 means year = 1997.

Sketch a scatterplot of this data. Make a residual plot and discuss if the linear model is appropriate.

In general, there are 2 approaches for finding models for non-linear data.

1. Fit a curve to the data.
2. Make the data straight and fit a line.

For several reasons, the second method is preferred in AP Statistics, although in the real world, option 1 is usually preferred.

To straighten the data, we can employ all sorts of transformations: squaring, square rooting, finding the reciprocal, etc. In AP Stat land, our transformation of choice will be the logarithm! We can use either the base 10 logarithm (log) or the natural logarithm (ln) and get similar models. In general, we will use ln since this is what most statisticians prefer, but you should be able to use both.

Since the original data seems to look like exponential growth, lets take the natural log of the salaries and make a scatterplot of year vs. ln(salary).

Since the transformed data looks fairly linear, find the LSRL for these data and make a residual plot.

Note: Our new model is in the form: $\ln \hat{y} = 0.1300 + .1360x$. Models in this form are called _____ since they can be re-expressed as $\hat{y} = ab^x$.

The residual plot still shows a possible U shaped curve, but it is much more scattered than the linear model. Thus, the exponential model will be more useful than the linear model. Remember, there is no perfect model, but some are more useful than others.

Use this model to predict the highest salary for the years 1993 and 2005. Do these seem reasonable?

Note: This method only works when the y-values are positive (you cannot find a log of 0 or a negative number) and start close to 0. If you need to shift the y-values, remember to shift back when making predictions!

5.5 More Non-Linear Data

Here are some data about our solar system:

Planet	Distance from Sun (million miles)	Length of Year (Earth years)
Mercury	36	0.24
Venus	67	0.61
Earth	93	1
Mars	142	1.88
Jupiter	484	11.86
Saturn	887	29.46
Uranus	1784	84.07
Neptune	2796	164.82
Pluto	3666	247.68

Sketch a scatterplot and make a residual plot. Does the linear model seem appropriate?

Transform the data by taking ln (y). Sketch a scatterplot and make a residual plot. Does the exponential model seem appropriate?

Transform the data by taking ln (x) and ln (y). Sketch a scatterplot and make a residual plot. Does this model seem appropriate?

Note: Models in the form $\ln \hat{y} = a + b(\ln x)$ are called _____ since they can be re-expressed as $\hat{y} = ax^b$.

In 2002, researchers discovered another object orbiting the sun 4 *billion* miles away from the sun. About how long should it take to orbit the sun?

A student experimenting with a pendulum counted the number of full swings of the pendulum made in 20 seconds with various amounts of string.

Length (in)	Number of Swings
6.5	22
9	20
11.5	17
14.5	16
18	14
21	13
24	13
27	12
30	11
37.5	10

Use the methods we have learned to find the most appropriate model for predicting the number of full swings. Give justification for your choice. Use the model you choose to predict the number of swings for a pendulum of length 10".

5.5 Conclusion

There are many more transformations we can try when logarithms don't work. Sometimes taking a square root will help straighten out a data set. Sometimes squaring a data set or using the reciprocals of a data set will work.

Besides transforming data to make it linear, you can also use many types of mathematical functions in an attempt to fit the data. Some of these can be found in the stat:calc menu on the TI-83.

The final decision about which model to use should be based on the residual plots. The model whose residual plot has the most random scatter does the best job of capturing the true form (signal) of the relationship.