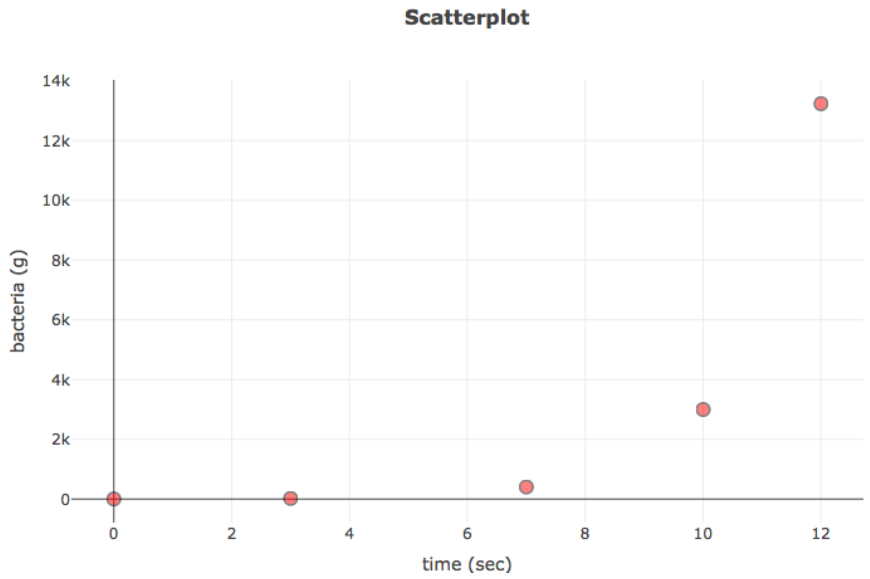
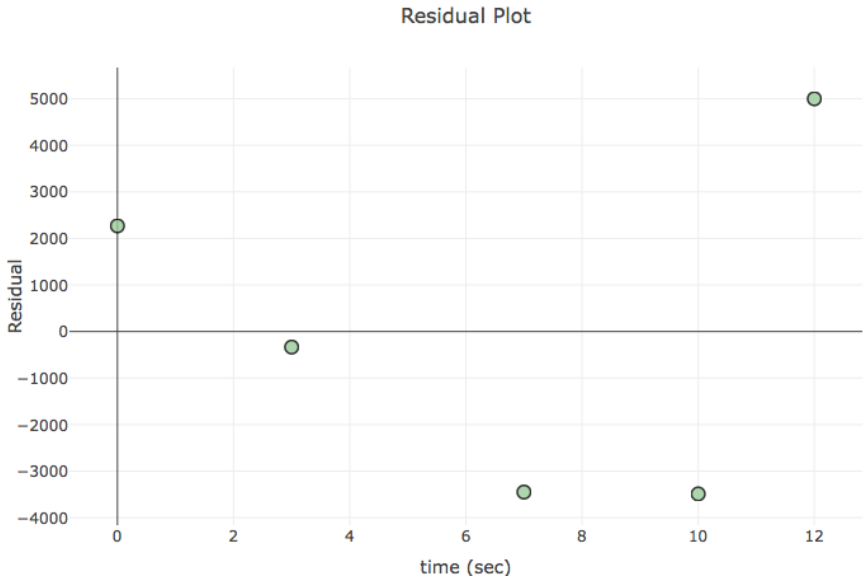
Transforming Non-Linear Data

Crash Course

A researcher has a bacteria growing in a laboratory culture. She measures the weight at various time increments. Her data, a scatterplot, and a residual plot for linear regression are below.

time(sec)	0	3	7	10	12
bacteria(g)	2.94	22.5	404.1	2997.35	13234.7

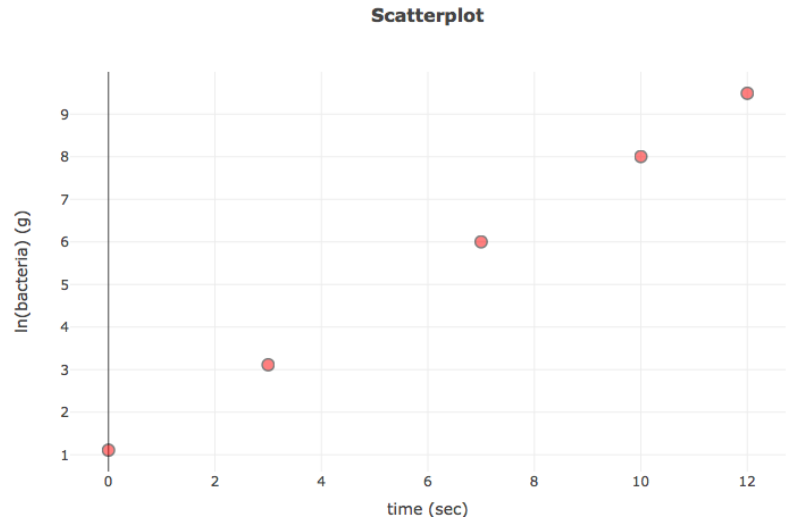


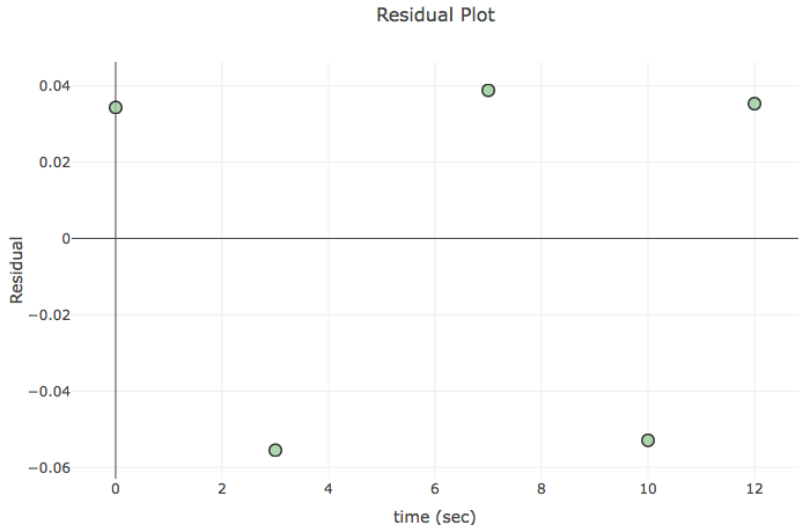


1. What does the residual plot tell you about the appropriateness of a linear model?

Transforming Non-Linear Data
Crash Course

What now? Answer: use some math skills + technology to transform the data and make it linear! As the researcher suspects her culture grows exponentially, she uses natural log of the weight to attempt to transform the data and make it linear. Here are the new plots.





2. What can you conclude about the appropriateness of a linear model for these transformed data?
3. The transformed regression equation is $\ln(\widehat{bacteria}) = 1.0736 + 0.6985(time)$. Use this equation to predict the weight of the bacteria at:
    a. 5 seconds
    b. 11 seconds
    c. 15 seconds
    d. Describe a concern you might have about your answer to question (c).

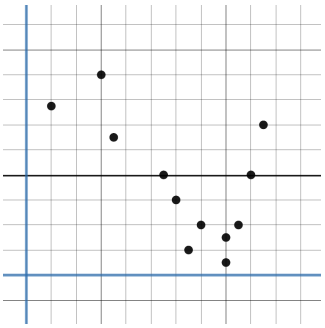
Transforming Non-Linear Data
Crash Course

You do not need to know how to choose the correct transformation. However, you should be able to evaluate the success of a transformation (by looking at the residual plot) and also be able to use a transformed equation to make a prediction.

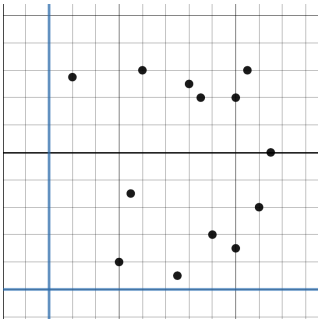
Here are some regression models that were used to model the relationship between time of day(between 6 a.m. and 12 noon, measured using hours i.e., 7.5= 7:30 a.m.) and the temperature (°F). Use each model to predict the temperature at 12 noon. The expected high was 85° F.

4. Model A: log(TEMP) = 1.56 + .02512 (hour)
5. Model B: sqrt(TEMP) = 5.89+ 0.25 (hour)
6. Model C: ln(TEMP) = 2.598 + 0.75 ln(hour)
7. Model D: TEMP = 41.568 + 3.256 (hour)
8. Model E: cubrt(TEMP)= 3.656 + 0.055 (hour)
9. Model F: ln(TEMP) = 3.65 + 0.065 (hour)
10. Model G: log(TEMP) = 0.73 + 1.112 * log (HOUR)
11. Model H: sq(TEMP) = 5234 + 167 (hour)
12. Here are the residual plots for some of the models. Comment on the success of each.

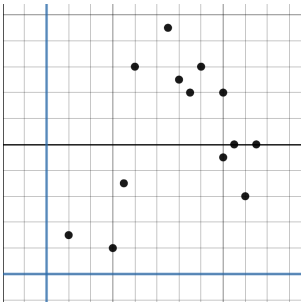
Model D



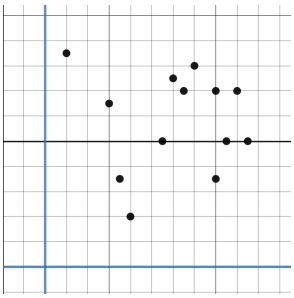
Model E



Model H

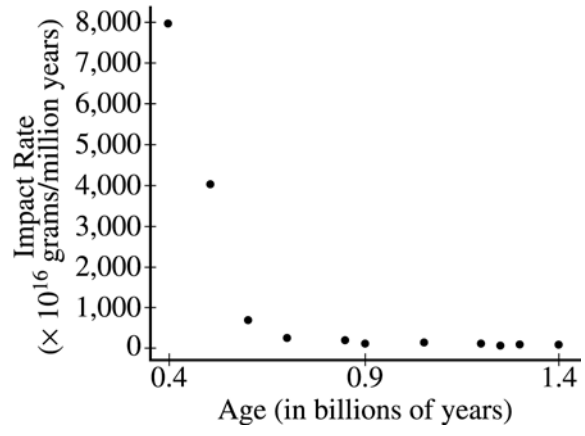


Model B



Check for Understanding
AP Exam 2004B

1. The Earth's Moon has many impact craters that were created when the inner solar system was subjected to heavy bombardment of small celestial bodies. Scientists studied 11 impact craters on the Moon to determine whether there was any relationship between the age of the craters (based on radioactive dating of lunar rocks) and the impact rate (as deduced from the density of the craters). The data are displayed in the scatterplot below.



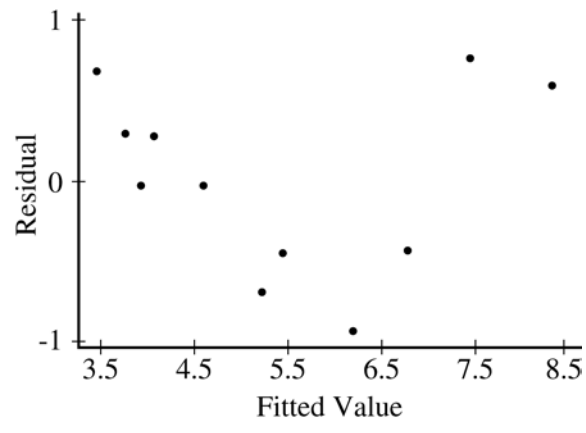
(a) Describe the nature of the relationship between impact rate and age.

Prior to fitting a linear regression model, the researchers transformed both impact rate and age by using logarithms. The following computer output and residual plot were produced.

Regression Equation: ln(rate) = 4.82 – 3.92 ln(age)

Predictor	Coef	SE Coef	T	P
Constant	4.8247	0.1931	24.98	0.000
ln(age)	–3.9232	0.4514	–8.69	0.000

S = 0.5977 R-Sq = 89.4% R-Sq (adj) = 88.2%



(b) Interpret the value of r^2.

(c) Comment on the appropriateness of this linear regression for modeling the relationship between the transformed variables.

Transforming Non-Linear Data
Crash Course

Answer Key

1. The residual plot has a clear curve. This indicates that the data does not follow a linear pattern. The relationship between time and bacteria is clearly curved.
2. The residual plot is now random. The data has been transformed to a linear pattern. Time and ln(bacteria) can be modeled using a linear model.
3. Plug in the number of seconds, then take e^x of both sides.
    a. 96.17 g
    b. 6355.65 g
    c. 103,891 g
    d. This may be extrapolation. We cannot be sure that the bacteria continues at the same growth rate.

4. Model A: 72.68°
5. B: 79.21°
6. C: 86.63°
7. D: 80.64°
8. E: 80.40°
9. F: 83.93°
10. G: 85.12°
11. H: 85.08°

12. Models D and H were both unsuccessful in linearizing the data. Both of these residual plots show a curved pattern. Models B and E both have random residual plots showing that the transformation successfully linearized the data.

Check for Understanding
2004B #1
1a. There is a strong, curved, negative relationship between Age and Impact Rate.
1b. 89.4% of the variation in ln(rate) has been successfully described by using ln(age).
1c. The residual plot shows a curved pattern. Thus the transformation, ln(rate) and ln(age), did not successfully linearize the data and the linear model is not appropriate.