Assignment 4.1 AP Statistics Name:

In addition to this assignment, please do questions from textbook on pp265 #4.1, 4.2, 4.3, pp 276 #4.5, 4.6, 4.7, pp289 4.15, 4.16

1. Definitions: Define each of the following using your own words:

Transforming/re expressing data: Applying a function (ie. logarithm or square root) to a quantitative variable.

Linear growth: Linear growth increases by a fixed amount in each equal time period.

Exponential growth: Exponential growth increases by a fixed percentage of the previous total in each equal time period.

2. *Algebraic properties of logarithms:*

$$log_b x = y \text{ if and only if } b^y = x$$

Rules:

- $log_b(MN) = log_b M + log_b N$
- $log_b(M/N) = log_b M - log_b N$
- $log_b x^p = p\, log_b x$

Note: Where are other models:

Power law model $\hat{y} = ax^b$ $log\, \hat{y} = log\, a + b\, log\, x$

Power b becomes the slope of the straight line that links $log\, y$ to $log\, x$.

Exponential model: $\hat{y} = ab^x$ $log\, \hat{y} = log\, a + x\, log\, b$

Logarithmic model: $\hat{y} = a + b\, log\, x$

3. Given each set of data, indicate which regression is best for modelling it. Perform the regression on your Ti-83 and write the equation:

i)

x	5	10	15	20	25	30	35	40	42
y	1.1	3.1	8.7	26.5	84.1	252.77	1350.3	2368.22	3000

ii)

x	4	6	9	16	18	21	23	25	26
y	33.5	112.5	252.2	1003.2	1403	1950.2	2395	2796	3300

iii)

x	1	2	3	4	5	6	9	12	14
y	24	31.6	36	39	41.3	43.5	48.5	51.5	55.3

4. Which regression would be best when comparing each pair of variables: Quadratic, Exponential, Logarithmic, Power, Logistic, or Linear. Explain your

i) The surface area of a salmon vs the amount of lice infection

A power regression would be best. The units of surface area is in the form of units^2, so the exponent for the variable will be 2.

ii) The size of a lake vs the amount of aquatic life

The size of a lake is measured in volume, which is units cubed. So a power regression would be best.

iii) Height of a child vs the age of the child

As the child grows from infancy, his/her height will grow rapidly in the beginning. As the child gets older, the height will not grow as fast. Instead it will eventually plateau, leveling off. A logistic regression would be best.

5. A scatterplot of a company's revenue versus time indicates a possible exponential relationship. A linear regression of $y = \log(revenue\ in\ \$10{,}000)$ against $x = years(\text{since } 2005)$ gives $\hat{y} = 0.875 + 0.59x$ with $r = 0.72$.

a) What does the slope represent?

It's given that the correlation is an exponential relationship,

$$\log(rev.) = 0.875 + 0.59x$$

$$\mathrm{Re}v(in\$10{,}000) = 10^{0.875+0.59x}$$

$$\mathrm{Re}v(in\$10{,}000) = 10^{0.875} \times 10^{0.59x}$$

$$\mathrm{Re}v = 7.4989 \times (3.89045^x)$$

The slope of 0.59 indicates an estimated percentage growth of $10^{0.59} \times 100\%$ each year since 2005. So the predicted company revenue increases by 389.045% each year since 2005.

b) What does the Y-intercept represent?

The Y-intercept of 0.875 represents the revenue (in $10,000) at the 2005. The predicted revenue at 2005 will be $10^{0.875}$, which is 7.49894 (in $10,000). In 2005, the predicted company revenue will be $74,989.

c) What is the equation in exponential form?

$$\mathrm{Re}venue = 10^{0.875} \times 10^{0.59x}$$

d) What is the predicted revenue for the year 2009?

$$\hat{y} = 0.875 + 0.59x$$

$$\hat{y} = 0.875 + 0.59(2009 - 2005)$$

$$\hat{y} = 3.235$$

$$\mathrm{Re}v = 10^{3.235}$$

$$\mathrm{Re}v = 1717.908(in\$10{,}000)$$

$$\mathrm{Re}v = \$17{,}179{,}083.87$$

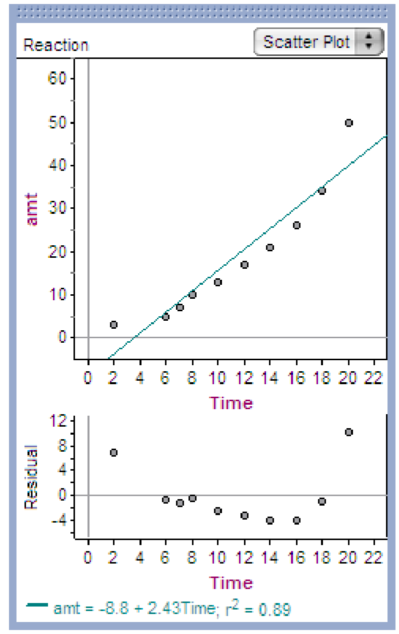
e) What percent in the variation in revenue can be explained by the variation in time with the LSRL?

$r = 0.72, \quad r^2 \approx 0.5184$. 51.84% of the variation in revenue can be explained by the variation in time with the LSRL

f) With this equation, can we use the revenue to predict the year? Explain

No, because regression both explanatory and response variables to be explicitly stated. Switching both variables will yield a different LSRL because residuals will be calculated differently.

6. Some data were collected during a chemical reaction in which reactants A and B were reacting to form products C and D. As the products were formed, scientists measured their masses (in grams) at several times (in minutes) during the experiment. The Fathom screen shot displays a scatterplot with the least-squares regression line superimposed, and a residual plot.

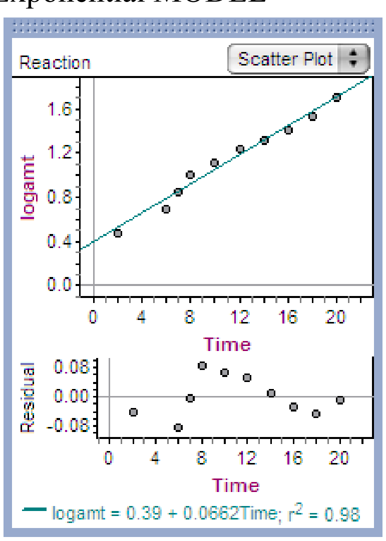


a. How well does the linear model fit these data? Justify your answer.

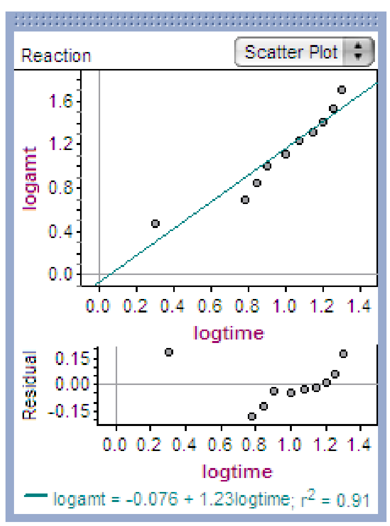
The linear model does not fit well because the scatterplot shows a definite curved pattern even though r=0.9434. The residual plot definitely shows a curved shape also. Residuals with lowest and highest time are positive while all the other times are negative in residual.

b. We used Fathom to transform the data and to produce the two screen shots below. Would an exponential model or a power model provide a better description of the relationship between reaction time and amount of product? Justify your answer.

Exponential MODEL



POWER MODEL



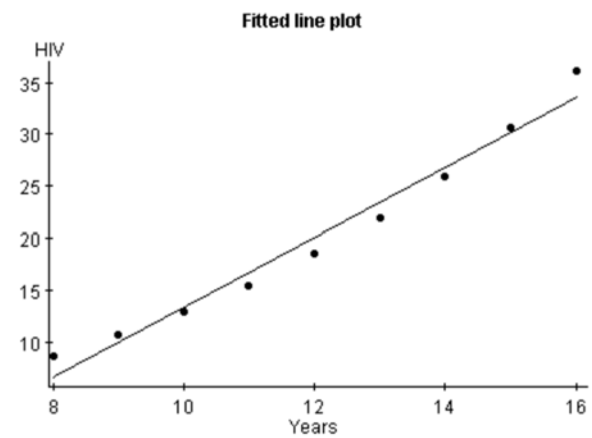
The exponential model provides a better description because data points tighter and closer to the LSRL. Residuals graph of the exponential also shows a more random graph, half of the residuals are positive/half negative. In addition the correlation is also stronger at r=0.99.

c. Use the model you chose in Question 6b to predict the amount of product that had been produced after 5 minutes. Show your method.

$$\log(amt) = 0.39 + 0.0662(time)$$
$$= 0.39 + 0.0662(5)$$
$$= 0.721$$
$$amt = 10^{0.721} = 5.2602\,grams$$

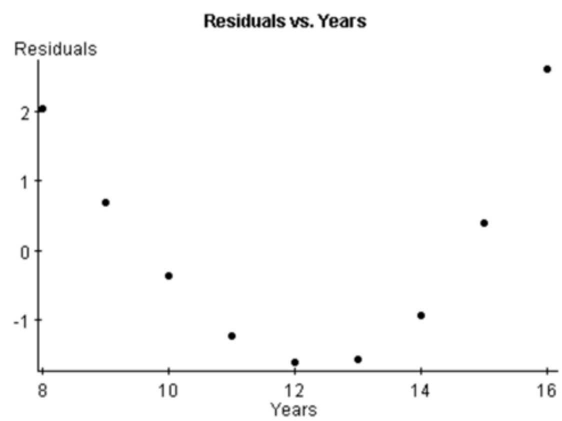
7. This dataset contains the global estimate of cumulative HIV cases worldwide. A scatterplot with the least-squares regression line superimposed and a residual plot from CrunchIt! are displayed.

Years since 1980	HIV infection (millions)
8	8.7
9	10.7
10	13.0
11	15.5
12	18.5
13	21.9
14	25.9
15	30.6
16	36.2



a. How well does the linear model fit these data? Justify your answer.

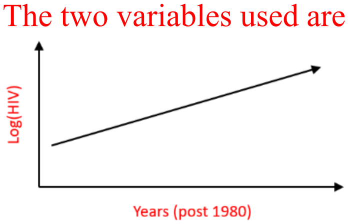
The linear model does not fit well because the scatter plot shows a slightly upward curve. Also the residual graph shows a definite curved pattern Based on this, the LSRL will not be a good fit.



b. Some experts warned that the number of cases of HIV infection was growing exponentially. Use your calculator to perform an appropriate transformation of the data to achieve linearity if the underlying relationship between the variables is exponential. Perform least-squares regression on the transformed data. Record your regression equation below. Define any variables you use.

$$\log(HIV) = 0.3409 + 0.0766(Years)$$

Y-intercept: 0.3409
Slope: 0.0766
The two variables used are "Years since 1980" and Number of HIV infections (in millions)



c. Use your model from Question 7c to predict the number of HIV infections in the year 2000.

$$\log(HIV) = 0.3409 + 0.0766(20)$$

$$year \quad 2000 : n = 20 \qquad \log(HIV) = 1.8729$$

$$HIV = 10^{1.8729} \approx 74.6277$$

Therefore, in the year 2000, the predicted amount of HIV infections in the year 2000 will be approximately 74.63 million people.