Chapter 10 – Re-expressing Data: Get It Straight!

1. **Residuals.**

   a) The residuals plot shows no pattern. No re-expression is needed.

   b) The residuals plot shows a curved pattern. Re-express to straighten the relationship.

   c) The residuals plot shows a fan shape. Re-express to equalize spread.

2. **Residuals.**

   a) The residuals plot shows a curved pattern. Re-express to straighten the relationship.

   b) The residuals plot shows a fan shape. Re-express to equalize spread.

   c) The residuals plot shows no pattern. No re-expression is needed.

3. **Airline passengers revisited.**

   a) The residuals cycle up and down because there is a yearly pattern in the number of passengers departing Oakland, California.

   b) A re-expression should not be tried. A cyclic pattern such as this one cannot be helped by re-expression.

4. **Hopkins winds, revisited.**

   a) The plot shows a wavy pattern, indicating a pattern that continues year to year as part of an annual cycle.

   b) A re-expression should not be tried. A cyclic pattern such as this one cannot be helped by re-expression.

5. **Models.**

   a) $\ln \hat{y} = 1.2 + 0.8x$
   $\ln \hat{y} = 1.2 + 0.8(2)$
   $\ln \hat{y} = 2.8$
   $\hat{y} = e^{2.8} = 16.44$

   b) $\sqrt{\hat{y}} = 1.2 + 0.8x$
   $\sqrt{\hat{y}} = 1.2 + 0.8(2)$
   $\sqrt{\hat{y}} = 2.8$
   $\hat{y} = 2.8^2 = 7.84$

   c) $\dfrac{1}{\hat{y}} = 1.2 + 0.8x$
   $\dfrac{1}{\hat{y}} = 1.2 + 0.8(2)$
   $\dfrac{1}{\hat{y}} = 2.8$
   $\hat{y} = \dfrac{1}{2.8} = 0.36$

   d) $\hat{y} = 1.2 + 0.8\ln x$
   $\hat{y} = 1.2 + 0.8\ln(2)$
   $\hat{y} = 1.75$

   e) $\log \hat{y} = 1.2 + 0.8\log x$
   $\log \hat{y} = 1.2 + 0.8\log(2)$
   $\log \hat{y} = 1.440823997...$
   $\hat{y} = 10^{1.4408...}$
   $\hat{y} = 27.59$

6. More models.

a)
$$\hat{y} = 1.2 + 0.8\log x$$
$$\hat{y} = 1.2 + 0.8\log(2)$$
$$\hat{y} = 1.44$$

b)
$$\log \hat{y} = 1.2 + 0.8x$$
$$\log \hat{y} = 1.2 + 0.8(2)$$
$$\log \hat{y} = 2.8$$
$$\hat{y} = 10^{2.8} = 630.96$$

c)
$$\ln \hat{y} = 1.2 + 0.8\ln x$$
$$\ln \hat{y} = 1.2 + 0.8\ln(2)$$
$$\ln \hat{y} = 1.7545...$$
$$\hat{y} = e^{1.7545..} = 5.78$$

d)
$$\hat{y}^2 = 1.2 + 0.8x$$
$$\hat{y}^2 = 1.2 + 0.8(2)$$
$$\hat{y}^2 = 2.8$$
$$\hat{y} = \sqrt{2.8} = 1.67$$

e)
$$\frac{1}{\sqrt{\hat{y}}} = 1.2 + 0.8x$$
$$\frac{1}{\sqrt{\hat{y}}} = 1.2 + 0.8(2)$$
$$\frac{1}{\sqrt{\hat{y}}} = 2.8$$
$$\hat{y} = \frac{1}{2.8^2} = 0.128$$

7. Gas mileage.

a) The association between weight and gas mileage of cars is fairly linear, strong, and negative. Heavier cars tend to have lower gas mileage.

b) For each additional thousand pounds of weight, the linear model predicts a decrease of 7.652 miles per gallon in gas mileage.

c) The linear model is not appropriate. There is a curved pattern in the residuals plot. The model tends to underestimate gas mileage for cars with relatively low and high gas mileages, and overestimates the gas mileage of cars with average gas mileage.

8. Crowdedness.

a) The scatterplot shows that the relationship between Crowdedness and GDP is strong, negative, and curved. Re-expression may yield an association that is more linear.

b) Start with logs, since GDP is non-negative. A plot of the log of GDP against Crowdedness score may be straighter.

9. Gas mileage revisited.

a) The residuals plot for the re-expressed relationship is much more scattered. This is an indication of an appropriate model.

b) The linear model that predicts the number of gallons per 100 miles in gas mileage from the weight of a car is: $\widehat{Gal}/100 = 0.625 + 1.178(Weight)$.

c) For each additional 1000 pounds of weight, the model predicts that the car will require an additional 1.178 gallons to drive 100 miles.

d)

$$\hat{Gal}/100 = 0.625 + 1.178(Weight)$$
$$\hat{Gal}/100 = 0.625 + 1.178(3.5)$$
$$\hat{Gal}/100 = 4.748$$

According to the model, a car that weighs 3500 pounds (3.5 thousand pounds) is expected to require approximately 4.748 gallons to drive 100 miles, or 0.04748 gallons per mile.

This is $\dfrac{1}{0.04748} \approx 21.06$ miles per gallon.

10. Crowdedness again.

a) This re-expression is not useful. The student has gone too far down the ladder of powers. We now see marked downward curvature and increasing scatter.

b) The next step would be to try a "weaker" re-expression, like reciprocal square root or log of GDP. Having gone too far, the student should move back "up" the ladder of powers.

11. GDP.

a) Although more than 97% of the variation in GDP can be accounted for by the model, the residuals plot should be examined to determine whether or not the model is appropriate.

b) This is not a good model for these data. The residuals plot shows curvature.

12. Treasury bills.

Re-expression should not be tried. An erratic trend that is positive then negative cannot be straightened by re-expression.

13. Better GDP model?

This is a much better model. Although there is still some pattern in the residuals, it is a much weaker pattern than the previous model.

14. Pressure.

The scatterplot at the right shows a strong, curved, negative association between the height of the cylinder and the pressure inside. Because of the curved nature of the association, a linear model is not appropriate.

Boyle's Pressure and Volume
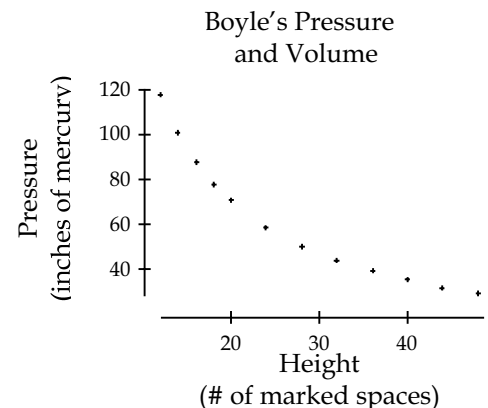
Re-expressing the pressure as the reciprocal of the pressure produces a scatterplot that is much straighter. Computer regression output for the height versus the reciprocal of pressure is below.

Dependent variable is: **recip pressure**
No Selector
R squared = 100.0%   R squared (adjusted) = 100.0%
s = 0.0001 with 12 - 2 = 10 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.000841	1	0.000841	75241
Residual	0.000000	10	0.000000	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-7.66970e-5	0.0001	-0.982	0.3494
Height	7.13072e-4	0.0000	274	≤ 0.0001

Height vs.
Reciprocal of Pressure
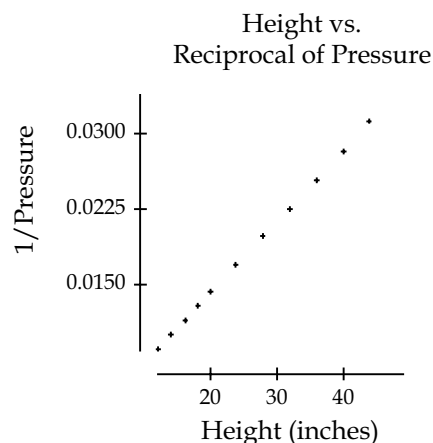
The reciprocal re-expression is very straight (perfectly straight, as far as the statistical software is concerned!). $R^2 = 100\%$, meaning that 100% of the variability in the reciprocal of pressure is explained by the model. The equation of the model is:

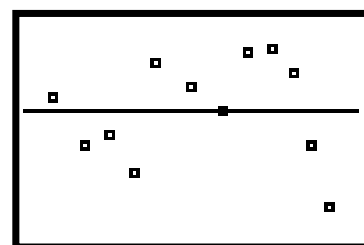$$\frac{1}{\widehat{Pressure}} = -0.000077 + 0.000713(Height).$$

Re-expressing each variable using logarithms is also a good model. TI-83 regression output for the log-log re-expression is given below.
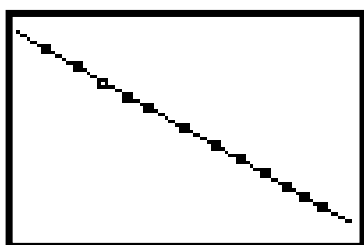
| Scatterplot of Log(pressure) vs. Log(height) | Regression Output | Residuals Plot |
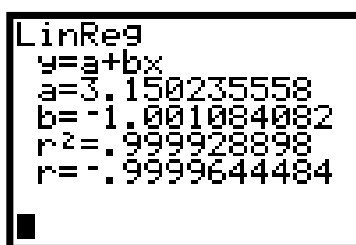
There is a strong, negative, linear relationship between log(*Pressure*) and log(*Height*).

$\log(\widehat{Pressure}) = 3.150 - 1.001(\log(Height))$ models the situation well, explaining nearly 100% of the variability in the logarithm of pressure. The residuals plot is fairly scattered, indicating an appropriate model.

15. Brakes.

a) The association between speed and stopping distance is strong, positive, and appears straight. Higher speeds are generally associated with greater stopping distances. The linear regression model, with equation $\widehat{Distance} = -65.9 + 5.98(Speed)$, has $R^2 = 96.9\%$, meaning that the model explains 96.9% of the variability in stopping distance. However, the residuals plot has a curved pattern. The linear model is not appropriate. A model using re-expressed variables should be used.

Stopping Distance and Speed

Residuals Plot
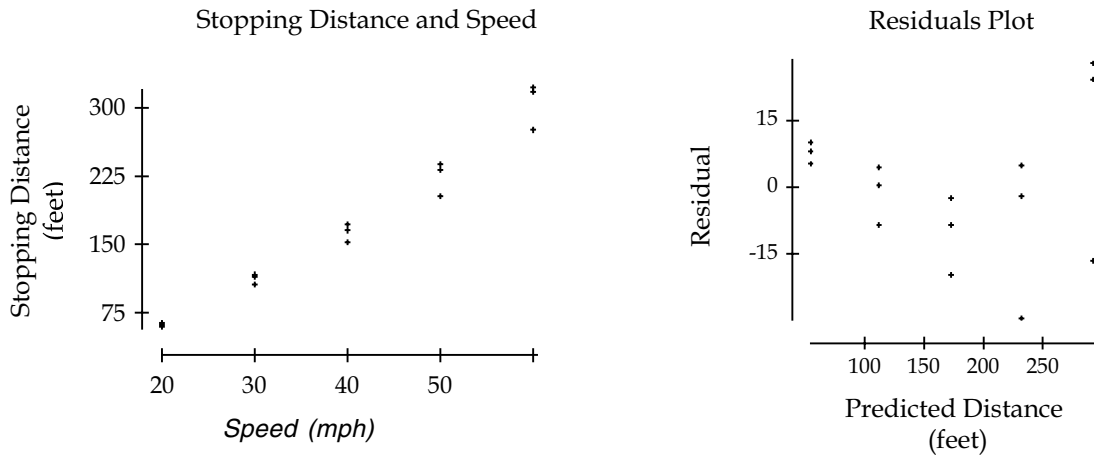
b) Stopping distances appear to be relatively higher for higher speeds. This increase in the rate of change might be able to be straightened by taking the square root of the response variable, stopping distance. The scatterplot of Speed versus $\sqrt{Distance}$ seems like it might be a bit straighter.

Square Root
Re-expression

Residuals Plot
Square Root Re-expression
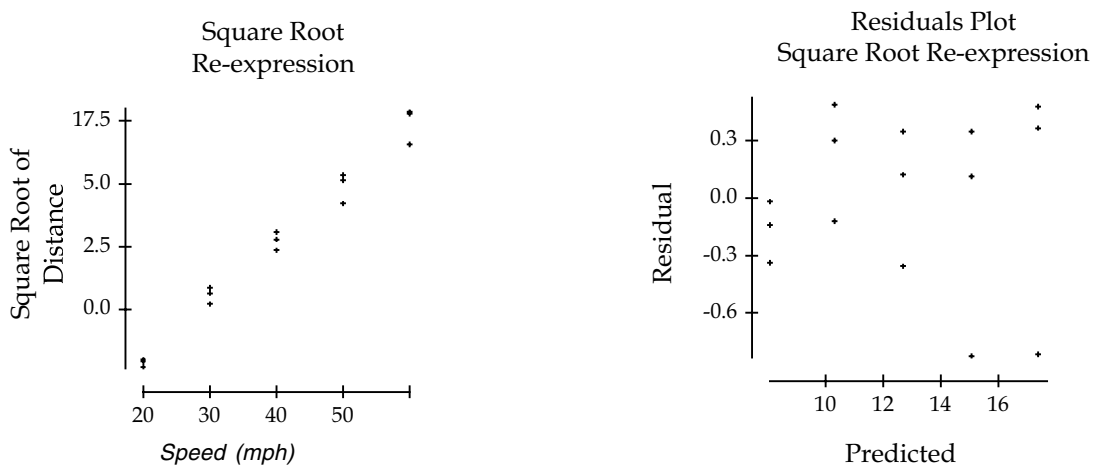
c) The model for the re-expressed data is $\sqrt{\widehat{Distance}} = 3.303 + 0.235(Speed)$. The residuals plot shows no pattern, and $R^2 = 98.4\%$, so 98.4% of the variability in the square root of the stopping distance can be explained by the model.

d)

$$\sqrt{Dis\hat{t}ance} = 3.303 + 0.235(Speed)$$

$$\sqrt{Dis\hat{t}ance} = 3.303 + 0.235(55)$$

$$\sqrt{Dis\hat{t}ance} = 16.228$$

$$Dis\hat{t}ance = 16.228^2 \approx 263.4$$

According to the model, a car traveling 55 mph is expected to require approximately 263.4 feet to come to a stop.

e)

$$\sqrt{Dis\hat{t}ance} = 3.303 + 0.235(Speed)$$

$$\sqrt{Dis\hat{t}ance} = 3.303 + 0.235(70)$$

$$\sqrt{Dis\hat{t}ance} = 19.753$$

$$Dis\hat{t}ance = 19.753^2 \approx 390.2$$

According to the model, a car traveling 70 mph is expected to require approximately 390.2 feet to come to a stop.

f) The level of confidence in the predictions should be quite high. R^2 is high, and the residuals plot is scattered. The prediction for 70 mph is a bit of an extrapolation, but should still be reasonably close.

16. Pendulum.

a) The scatterplot shows the association between the length of string and the number of swings a pendulum took every 20 seconds to be strong, negative, and curved. A pendulum with a longer string tended to take fewer swings in 20 seconds. The linear model is not appropriate, because the association is curved.

Pendulum

b) Curvature in a negative relationship sometimes is an indication of a reciprocal relationship. Try re-expressing the response variable with the reciprocal.
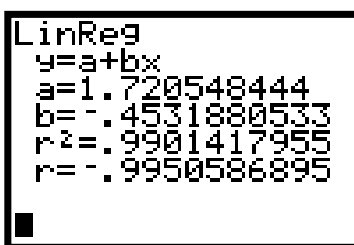

Reciprocal Re-expression

Residuals Plot
Reciprocal Re-expression
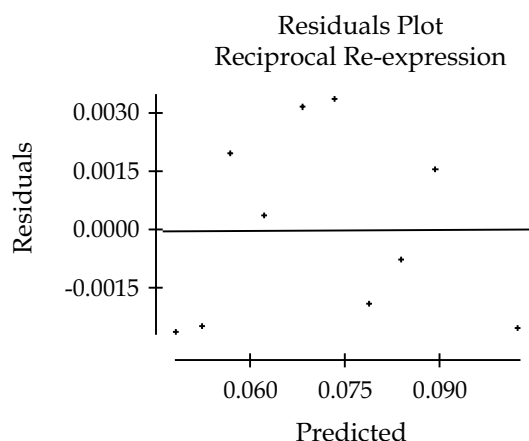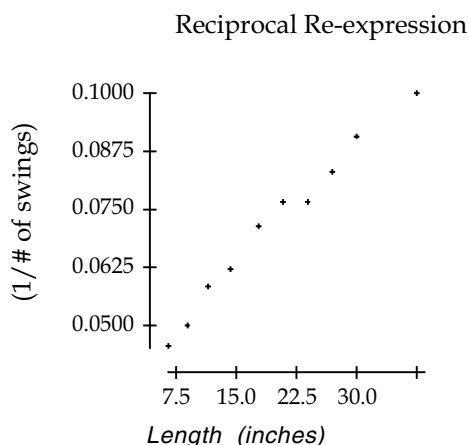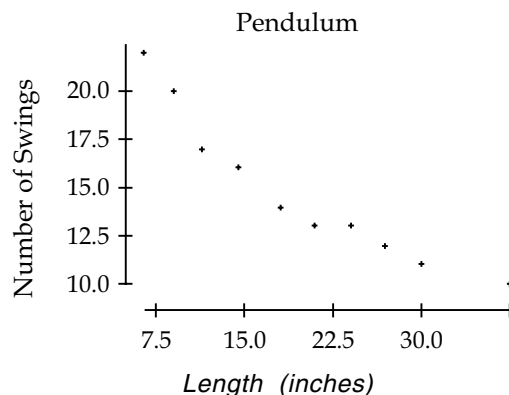
c) The reciprocal re-expression yields the model $\dfrac{1}{Swings} = 0.0367 + 0.00176(Length)$. The

residuals plot is scattered, and R^2 = 98.1%, indicating that the model explains 98.1% of the variability in the reciprocal of the number of swings. The model is both appropriate and accurate.
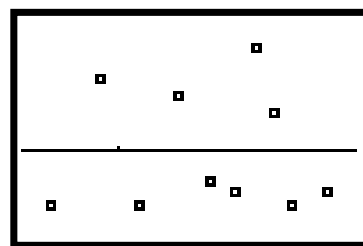
Re-expressing each variable with logarithms also results in a good model, and has the added benefit of maintaining the direction of the relationship between number of swings and the length of the string. TI-83 regression output for the log- log re-expression is given below.
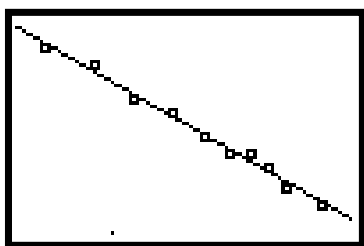
Scatterplot of Log(swings) vs. Log(length)	Regression Output	Residuals Plot

There is a strong, negative, linear relationship between log(swings) and log(length).

$\log(\widehat{Swings}) = 1.721 - 0.453(\log(Length))$ explains 99% of the variability in log(swings), and the residuals plot is scattered. The log-log re-expression is useful, as well.

d) Using the reciprocal model:

$$\frac{1}{\widehat{Swings}} = 0.0367 + 0.00176(Length) = 0.0367 + 0.00176(4) = 0.04374$$

$$\widehat{Swings} = \frac{1}{0.04374} \approx 22.9$$

According to the reciprocal model, a pendulum with a 4″ string is expected to swing approximately 22.9 times in 20 seconds.

Using the log-log model:

To make estimates involving logs, it is a good idea to use as much accuracy as possible. If the equation of the model is stored in the calculator, the model predicts that a pendulum with a 4″ string will swing approximately 28 times in 20 seconds.

e) Using the reciprocal model:

$$\frac{1}{\widehat{Swings}} = 0.0367 + 0.00176(Length) = 0.0367 + 0.00176(48) = 0.12118$$

$$\widehat{Swings} = \frac{1}{0.12118} \approx 8.3. \text{ The model predicts 8.3 swings in 20 seconds for a 48″ string.}$$
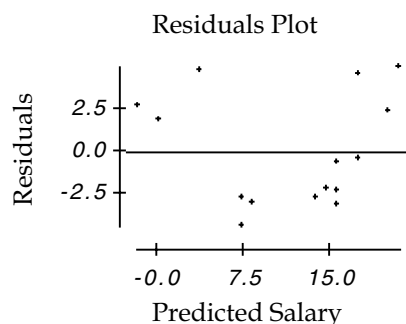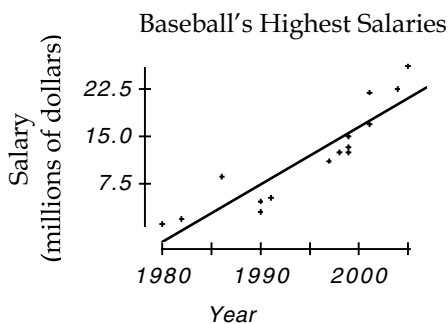
Using the log-log model:

To make estimates involving logs, it is a good idea to use as much accuracy as possible. If the equation of the model is stored in the calculator, the model predicts that a pendulum with a 48″ string will swing approximately 9.1 times in 20 seconds.

f) Confidence in the predictions is fairly high. The model is appropriate, as indicated by the scattered residuals plot, and accurate, indicated by the high value of R^2. The only concern is the fact that these predictions are slight extrapolations. The lengths of the strings aren't too far outside the range of the data, so the predictions should be reasonably accurate.

17. Baseball salaries 2005.

a) The scatterplot of year versus highest salary is moderately strong, positive and possibly curved, yet straight enough to try fitting a linear model. The highest salary has generally increased over the years.

b) The linear model, $\hat{Salary} = -1802.94 + 0.9097(Year)$, is not an appropriate model. The residuals plot shows a strong bend.

**c)** Re-expression using the logarithm of the salaries straightens the plot significantly.
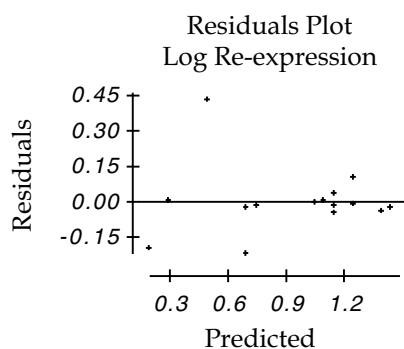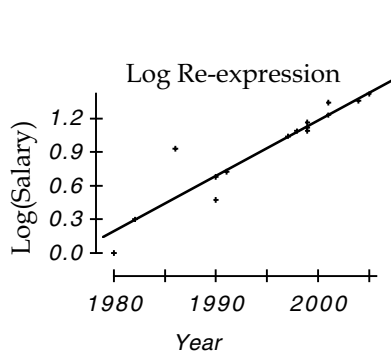

d) $\log(\hat{Salary}) = -98.509 + 0.04985(Year)$ is a good model for predicting the highest salary from the year. The residuals plot is scattered and $R^2 = 87.9\%$, indicating that 87.9% of the variability in highest salary can be explained by the model.

18. Planet distances and years 2006.

a) The association between distance from the sun and planet year is strong, positive, and curved concave upward. Generally, planets farther from the sun have longer years than closer planets.

Planetary Distance and Length of Year
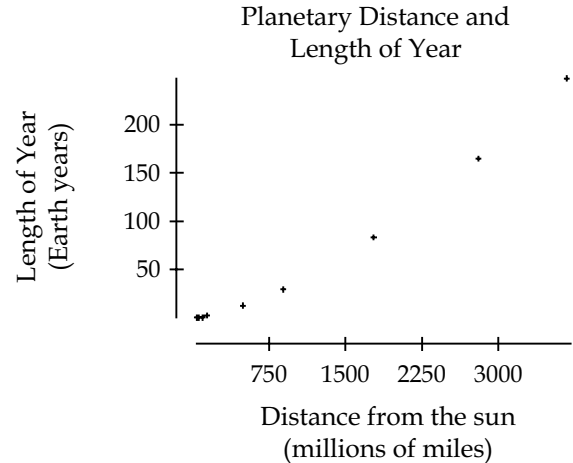
b) The rate of change in length of year per unit distance appears to be increasing, but not exponentially. Re-expressing with the logarithm of each variable may straighten a plot such as this. The scatterplot and residuals plot for the linear model relating log(Distance) and log(Length of Year) appear below.
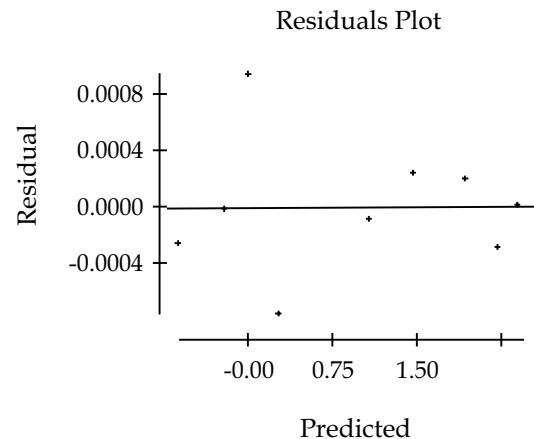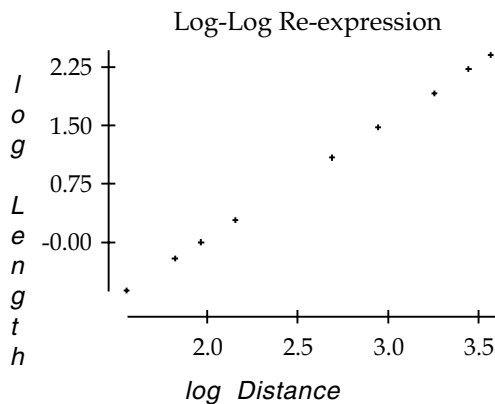
The regression model for the log-log re-expression is :
$$\log(Len\hat{g}th) = -2.95 + 1.5(\log(Distance)).$$

c) $R^2 = 100\%$, so the model explains 100% of the variability in the log of the length of the planetary year, at least according to the accuracy of the statistical software. The residuals plot is scattered, and the residuals are all extremely small. This is a very accurate model.

19. Planet distances and order 2006.

a) The association between planetary position and distance from the sun is strong, positive, and curved (below left). A good re-expression of the data is position versus Log(Distance). The scatterplot with regression line (below center) shows the straightened association. The equation of the model is $\log(\widehat{Distance}) = 1.245 + 0.271(Position)$. The residuals plot (below right) may have some pattern, but after trying several re-expressions, this is the best that can be done. $R^2 = 98.2\%$, so the model explains 98.2% of the variability in the log of the planets distance from the sun.
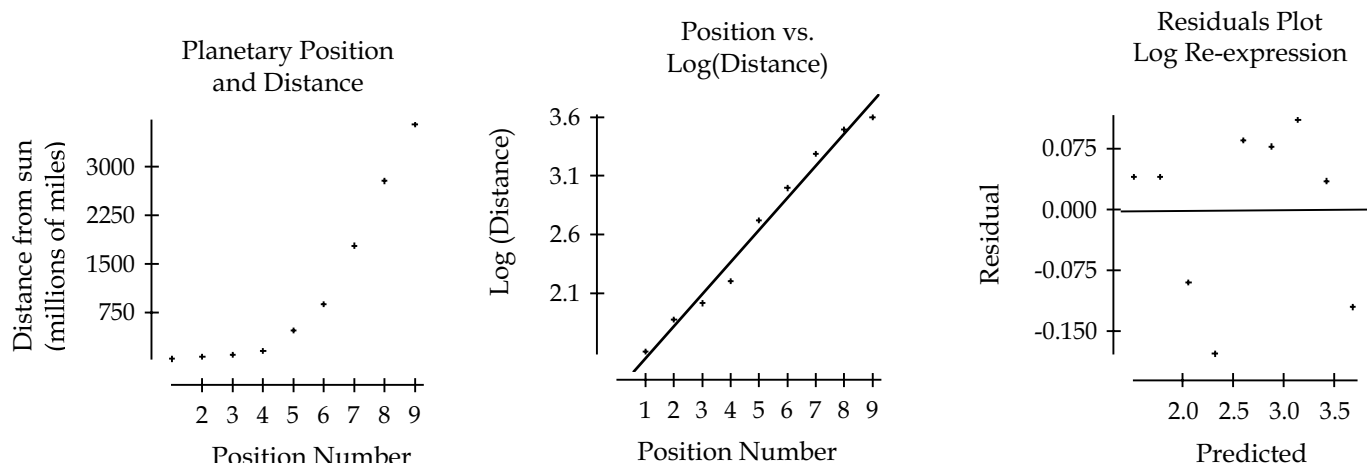
b) At first glance, this model appears to provide little evidence to support the contention of the astronomers. Pluto appears to fit the pattern, although Pluto's distance from the sun is a bit less than expected. A model generated without Pluto does not have a dramatically improved residuals plot, does not have a significantly higher R^2, nor a different slope. Pluto does not appear to be influential.

But don't forget that a logarithmic scale is being used for the vertical axis. The higher up the vertical axis you go, the greater the effect of a small change.

$$\log(\widehat{Distance}) = 1.24418 + 0.271229(Position)$$
$$\log(\widehat{Distance}) = 1.24418 + 0.271229(9)$$
$$\log(\widehat{Distance}) = 3.685241$$
$$\widehat{Distance} = 10^{3.685241} \approx 4844$$

According to the model, the 9th planet in the solar system is predicted to be approximately 4844 million miles away from the sun. Pluto is actually 3707 million miles away.

Pluto doesn't fit the pattern for position and distance in the solar system. In fact, the model made with Pluto included isn't a good one, because Pluto influences those predictions. The model without Pluto, $\log(\widehat{Distance}) = 1.20267 + 0.283680(Position)$, works much better. It has a high R^2, and scattered residuals plot. This new model predicts that the 9th planet should be a whopping 5699 million miles away from the sun! There is evidence that the IAU is correct. Pluto doesn't behave like planet in its relation to position and distance.

20. Planets 2006, part 3.

Using the revised planetary numbering system, and straightening the scatterplot using the same methods as in Exercise 9, the new model, $\log(\hat{Distance}) = 1.32 + 0.23(Position)$, is a slightly better fit. The residuals plot is more scattered, and R^2 is slightly higher, with the improved model explaining 99.5% of the variability in the log of distance from the sun.
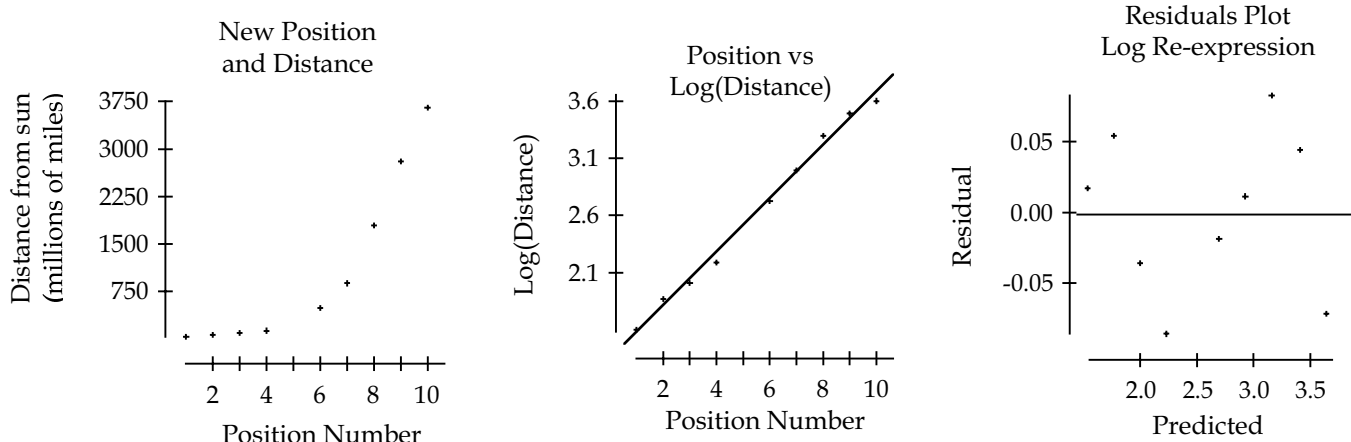
Pluto still doesn't fit very well. The new model predicts that Pluto, as 10th planet, should be about 4169 million miles away. That's about 462 million miles farther away. A better model yet is $\log(\hat{Distance}) = 1.28514 + 0.238826(Position)$, a model made with the new numbering system and with Pluto omitted.

21. Eris: Planets 2006, part 4.

A planet ninth from the sun was predicted, in a previous exercise, to be about 4844 million miles away from the sun. This distance is much shorter than the actual distance of Eris, about 6300 miles.

22. Models and laws: Planets 2006 part 5

The re-expressed data relating distance and year length are better described by their model than the re-expressed data relating position and distance. The model relating distance and year length has $R^2 = 100\%$, and a very scattered residuals plot (with miniscule residuals), possibly a natural "law". If planets in another solar system followed the Titius-Bode pattern, this belief would be reinforced. Similarly, if data were acquired from planets in another solar system that did not follow this pattern, we would be unlikely to think that this relationship was a universal law.

23. Logs (not logarithms).

a) The association between the diameter of a log and the number of board feet of lumber is strong, positive, and curved. As the diameter of the log increases, so does the number of board feet of lumber contained in the log.

The model used to generate the table used by the log buyers is based upon a square root re-expression. The values in the table correspond exactly to the model

$$\sqrt{Bo\hat{a}rdFeet} = -4 + Diameter.$$

b)

$$\sqrt{Bo\hat{a}rdFeet} = -4 + Diameter$$

$$\sqrt{Bo\hat{a}rdFeet} = -4 + (10)$$

$$\sqrt{Bo\hat{a}rdFeet} = 6$$

$$Bo\hat{a}rdFeet = 36$$

According to the model, a log 10″ in diameter is expected to contain 36 board feet of lumber.

**c)**

$$\sqrt{Bo\hat{a}rdFeet} = -4 + Diameter$$

$$\sqrt{Bo\hat{a}rdFeet} = -4 + (36)$$

$$\sqrt{Bo\hat{a}rdFeet} = 32$$

$$Bo\hat{a}rdFeet = 1024$$

According to the model, a log 36″ in diameter is expected to contain 1024 board feet of lumber.

Normally, we would be cautious of this prediction, because it is an extrapolation beyond the given data, but since this is a prediction made from an exact model based on the volume of the log, the prediction will be accurate.
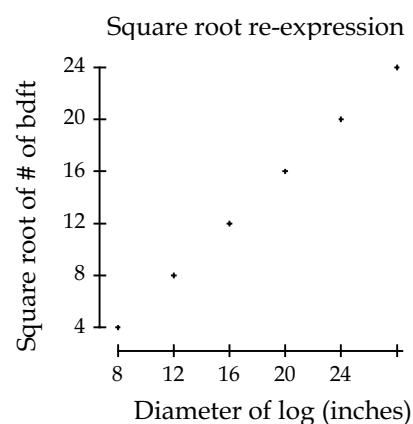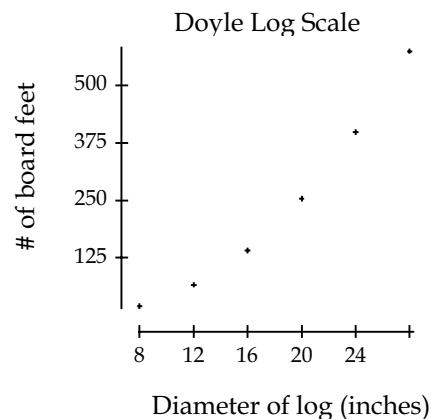
24. Weightlifting 2004.

a) The association between weight class and weight lifted for gold medal winners in weightlifting at the 2004 Olympics is strong, positive, and curved. The linear model that best fits the data is $L\hat{i}ft = 164.97 + 2.56(WeightClass)$.
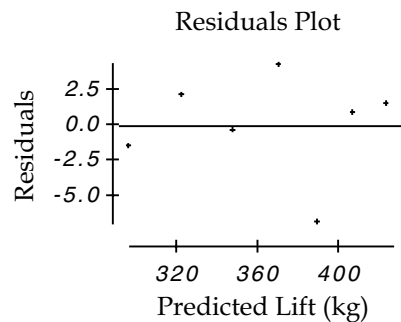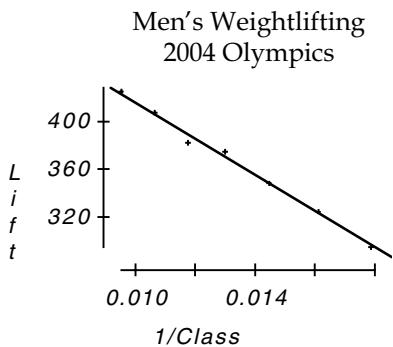
Although this model accounts for 96.3% of the variability in weight lifted, it does not fit the data well.
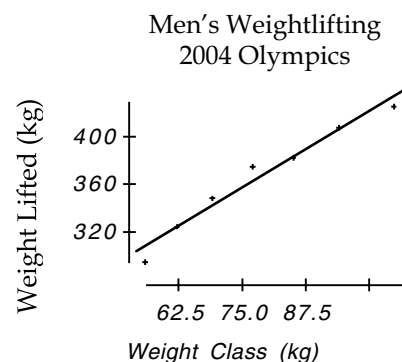
b) The residuals plot for the linear model shows a curved pattern, indicating that the linear model has failed to model the association well. A re-expressed model or a curved model might fit the association between weight class and weight lifted better than the linear model.
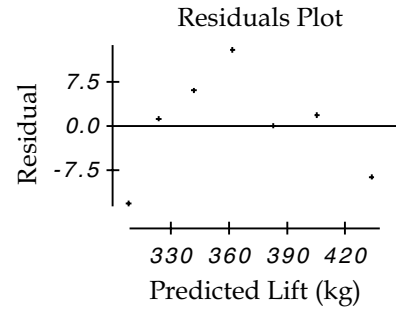
Residuals Plot

c) Re-expressing Weight Class using the reciprocal produces a scatterplot that is much straighter. $\hat{Lift} = 568.727 - 15243\left(\dfrac{1}{Class}\right)$.

Men's Weightlifting
2004 Olympics

Residuals Plot

d) The reciprocal model is a better model, since the residuals plot shows little pattern. Additionally, the model accounts for 99.4% of the variability in weight lifted.

e) George Asanidze's lift had a large, negative residual. This means that he lifted less than expected.

25. Life expectancy.

The association between year and life expectancy is strong, curved and positive. As the years passed, life expectancy for white males has increased. The bend (in mathematical language, the change in curvature) in the scatterplot makes it impossible to straighten the association by re-expression.
If all of the data points are to be used, the linear model, $\hat{LifeExp} = -454.938 + 0.265697(Year)$ is probably the best model we can make. It doesn't fit the data well, but accounts for 93.6% of the variability in life expectancy.

Life Expectancy
White Males
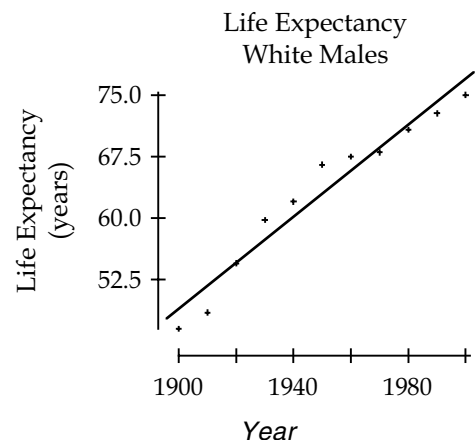
In order to make better future predictions, we might use only a subset of the data points. The scatterplot is very linear from 1970 to 2000, and these recent years are likely to be more indicative of the current trend in life expectancy. Using only these four data points, we might be able to make more accurate future predictions.

The model $Lif\hat{e}Exp = -379.020 + 0.227(Year)$ explains 99.6% of the variability in life expectancy, and has a scattered residuals plot. Great care should be taken in using this model for predictions, since it is based upon only 4 points, and, as always, prediction into the future is risky. This is especially true since the response variable is life expectancy, which cannot be expected to increase at the same rate indefinitely.

Life Expectancy
White Males
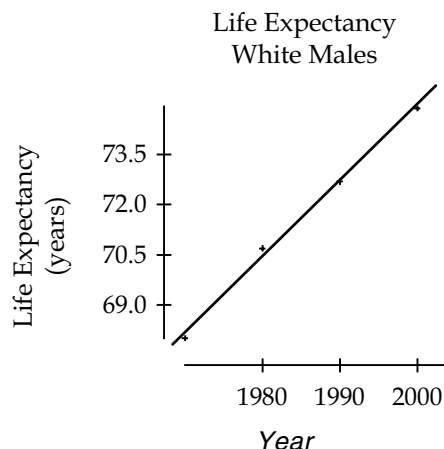
26. Lifting more weight 2004.

a) Without Asanidze, the reciprocal model is: $L\hat{i}ft = 573.257 - 15491.5\left(\dfrac{1}{Class}\right)$.

b) According to the reciprocal model, the winner of the 85 kg weight class was predicted to lift 391.0 kg.

c) Since Asanidze actually lifted 382.5 kg, his residual was 382.5 – 391 = – 8.5. He lifted 8.5 kg less than predicted.

27. Slower is cheaper?

The scatterplot shows the relationship between speed and mileage of the compact car. The association is extremely strong and curved, with mileage generally increasing as speed increases, until around 50 miles per hour, then mileage tends to decrease as speed increases. The linear model is a very poor fit, but the change in direction means that re-expression cannot be used to straighten this association.

Speed and Mileage
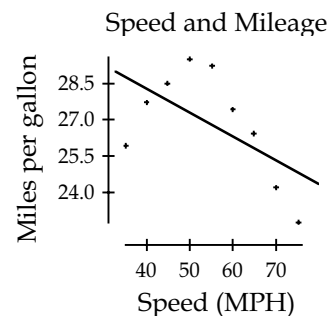
28. Orange production.

The association between the number of oranges per tree and the average weight is strong, negative, and appears linear at first look. Generally, trees that contain larger numbers of oranges have lower average weight per orange. The residuals plot shows a strong curved pattern. The data should be re-expressed.

Orange Production
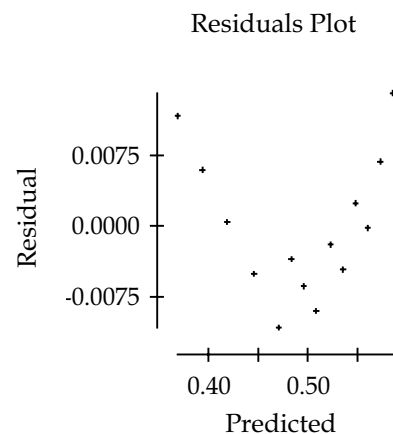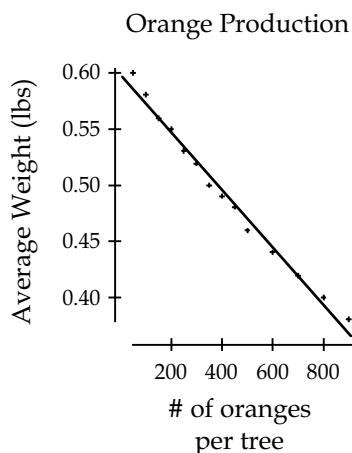
Residuals Plot

Plotting the number of oranges per tree and the reciprocal of the average weight per orange straightens the relationship considerably. The residuals plot shows little pattern and the value of R^2 indicates that the model explains 99.8% of the variability in the reciprocal of the average weight per orange. The more appropriate model is:

$$\frac{1}{Ave.\hat{wt}} = 1.603 + 0.00112(\#Oranges/Tree)$$

.

Reciprocal Re-expression

Residuals Plot
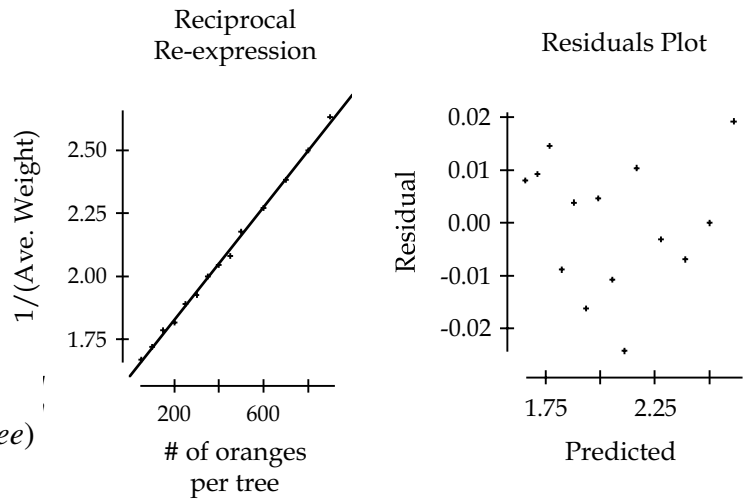
29. Years to live 2003.

a) The association between the age and estimated additional years of life for black males is strong, curved, and negative. Older men generally have fewer estimated years of life remaining. The square root re-expression of the data, $\sqrt{Year\hat{s}Left} = 8.465 - 0.06926(Age)$, straightens the data considerably, but has an extremely patterned residuals plot. The model is not a mathematically appropriate model, but fits so closely that it should be fine for predictions within the range of data. The model explains 99.8% of the variability in the estimated number of additional years.

Number of Years Left Black Males

Square Root Re-expression
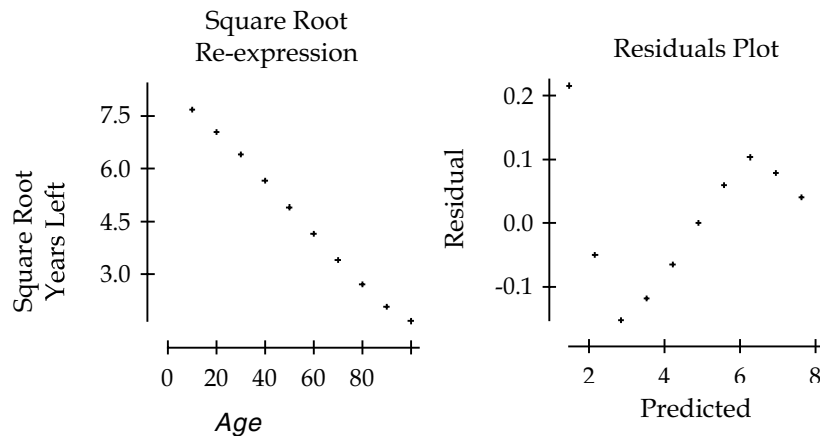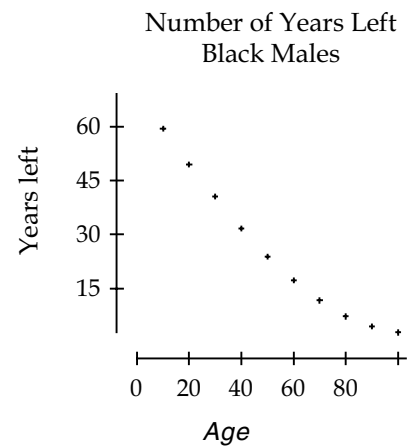
Residuals Plot

b)

$$\sqrt{Year\hat{s}Left} = 8.465 - 0.06926(Age)$$
$$\sqrt{Year\hat{s}Left} = 8.465 - 0.06926(18)$$
$$\sqrt{Year\hat{s}Left} = 7.21832$$
$$Year\hat{s}Left = 7.21832^2 \approx 52.1$$

According to the model, an 18-year-old black male is expected to live and additional 52.1 years, for a total age of 70.1 years.

c) The residuals plot is extremely patterned, so the model is not appropriate. However, the residuals are very small, making for a tight fit. Since 18 years is within the range of the data, the prediction should be at least reasonable.

30. Tree growth.

a) The association between age and average diameter of grapefruit trees is strong, curved, and positive. Generally, older trees have larger average diameters.

Grapefruit Trees
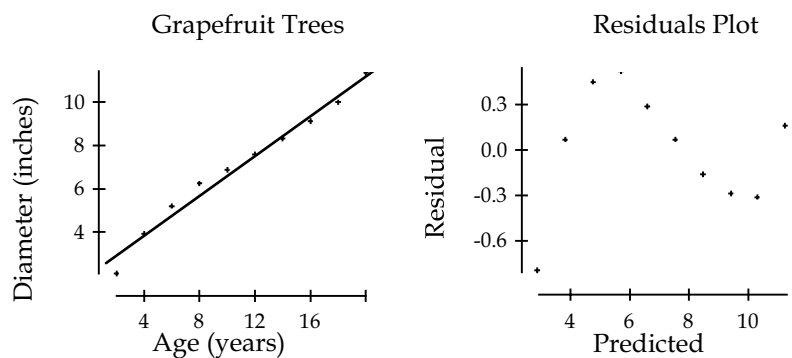
Residuals Plot

The linear model for this association, $Average\hat{D}iameter = 1.973 + 0.463(Age)$ is not appropriate. The residuals plot shows a clear pattern.

Because of the change in curvature in the association, these data cannot be straightened by re-expression.

b) If diameters from individual trees were given, instead of averages, the association would have been weaker. Individual observations are more variable than averages.