Chapter 8 – Regression Wisdom

Section 8.1

1. Credit card spending.

The different segments are not scattered at random throughout the residual plot. Each segment may have a different relationship, which would affect the accuracy of any predictions made with the model that ignores the differences between segments.

2. Revenue and talent cost.

a) There is a positive (revenue increases with talent cost), linear, moderately strong relationship between talent cost and total revenue. There is a possible outlier, an act with both high talent costs and high revenue stands apart from the others.

b) Both venues show an increase of revenue with talent cost.

c) The larger venue has greater variability. Revenue for that venue is more difficult to predict. Additionally, the larger venue typically has both higher total revenue and higher talent cost.

3. Market segments.

Yes, it is clear that the relationship between January and December spending is not the same for all five segments. Using one overall model to predict January spending would be very misleading.

4. Revenue and ticket sales.

a) There is a positive (revenue increases with ticket sales), linear, strong association between ticket sales and total revenue.

b) Both show a strong, positive association. Each may show some curvature.

c) They differ primarily in magnitude of the values. The larger venue typically has both higher total revenue and higher ticket sales. The larger venue has greater variability. Revenue for that venue is more difficult to predict.

Section 8.2

5. Cell phone costs.

Your friend is extrapolating. It is impossible to know if a trend like this will continue so far into the future.

6. Stopping times.

Since the model used only data from compact cars, you cannot be certain that this model extends to a large SUV that is much heavier.

7. Revenue and large venues.

a) $\widehat{Revenue} = -14,228 + 36.87\,TicketSales = -14,228 + 36.87(10,000) = 354,472;$ According to the model, a venue that seats 10,000 would be expected to generate $354,472 in revenue, if it were to sell out.

b) An extrapolation this far from the data is unreliable. We only have data up to about 3000 seats. 10,000 seats is well above that.

8. Revenue and advanced sales.

The point has high leverage and is an outlier. It will reduce the slope of a fitted regression line and make the line a poor model for the data.

Section 8.3

9. Abalone.

This observation was influential. After it was removed, the correlation and the slope of the regression line both changed by a large amount.

10. Abalone again.

No. Some data points will have higher residuals than others. While large residuals should be looked at carefully, it is not proper to simply remove all those data points. Furthermore, high leverage points often have small residuals, since the can dominate a regression, shifting the regression line toward themselves. Outliers should primarily be identified by looking at the scatterplot, not the residuals.

Section 8.4

11. Skinned knees.

No. There is a lurking variable, seasonal temperature. In warm weather, more children will go outside and play, and if there are more children playing, there will be more skinned knees.

12. Cell phones and life expectancy.

No. There is a lurking variable, wealth. Wealthier countries typically have more cell phones and better healthcare.

Section 8.5

13. Grading.

Individual student scores will vary greatly. The class averages will have much less variability and may disguise important patterns.

14. Average GPA.

The individual GPAs for each team are going to vary widely. Also, the rest of the team may hide a few individuals with low GPAs. These summaries are a risky method for predicting the students' graduation rates.

Section 8.6

15. 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.

16. More 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.

17. BK protein.

The goal of this re-expression is to improve homoscedasticity. We desire more equal spread between groups.

18. TVs and life expectancy.

The goal of this re-expression is to straighten the plot.

Section 8.7

19. BK protein again.

The log re-expression is still preferable. The square root doesn't make the spreads as nearly equal. The reciprocal clearly goes too far on the Ladder of Powers.

20. TVs and life expectancy.

The bend in the plot now goes other way, so we have moved too far on the Ladder of Powers.

Chapter Exercises.

21. Marriage age 2015.

a) The trend in age at first marriage for American women is very strong over the entire time period recorded on the graph, but the direction and form are different for different time periods. The trend appears to be somewhat linear, and consistent at around 22 years, up until about 1940, when the age seemed to drop dramatically, to under 21. From 1940 to about 1970, the trend appears non-linear and slightly positive. From 1975 to the present, the trend again appears linear and positive. The marriage age rose rapidly during this time period.

b) The association between age at first marriage for American women and year is strong over the entire time period recorded on the graph, but some time periods have stronger trends than others.

c) The correlation, or the measure of the degree of linear association is not high for this trend. The graph, as a whole, is non-linear. However, certain time periods, like 1975 to present, have a high correlation.

d) Overall, the linear model is not appropriate. The scatterplot is not Straight Enough to satisfy the condition. You could fit a linear model to the time period from 1975 to 2003, but this seems unnecessary. The ages for each year are reported, and, given the fluctuations in the past, extrapolation seems risky.

22. Smoking 2014.

a) The percent of men 18–24 who are smokers decreased dramatically between 1965 and 1990, but the trend has not been consistent since then, though it may be decreasing in a linear fashion since about 1998.

b) The association between percent of men 18–24 who smoke and year is very strong from 1965 to 1990, but is erratic after 1990.

c) A linear model is not an appropriate model for the trend in the percent of males 18–24 who are smokers. The relationship is not straight.

23. Human Development Index 2015.

a) Fitting a linear model to the association between HDI and GDPPC would be misleading, since the relationship is not straight.

b) If you fit a linear model to these data, the residuals plot will be curved downward.

24. HDI 2015 revisited.

a) Fitting a linear model to the association between the number of cell phones and HDI would be misleading, since the relationship is not straight.

b) The residuals plot will be curved downward.

25. Good model?

a) The student's reasoning is not correct. A scattered residuals plot, not high R^2, is the indicator of an appropriate model. Once the model is deemed appropriate, R^2 is used as a measure of the strength of the model.

b) The model may not allow the student to make accurate predictions. The data may be curved, in which case the linear model would not fit well.

26. Bad model?

a) The student's model may, in fact, be appropriate. Low R^2 simply means that the model is not accurate. The model explains only 13% of the variability in the response variable. If the residuals plot shows no pattern, this model may be appropriate.

26. (continued)

b) The predictions are not likely to be very accurate, but they may be the best that the student can get. $R^2 = 13\%$ indicates a great deal of scatter around the regression line, but if the residuals plot is not patterned, there probably isn't a better model. The two variables that are being studied by the student have a weak association.

27. Movie dramas.

a) The units for the slopes of these lines are millions of dollars per minutes of running time.

b) The slopes of the regression lines are about the same. Dramas and movies from other genres have costs for longer movies that increase at the same rate.

c) The regression line for dramas has a lower *y*-intercept. Regardless of running time, dramas cost about 20 million dollars less than other genres of movies of the same running time.

28. Smoking 2014, women and men.

a) Smoking rates for both men and women in the United States have decreased significantly, but not linearly, over the time period from 1965 to 2014.

b) Smoking rates are generally lower for women than for men. The exception is 1985, where the smoking rate for women was slightly higher than the smoking rate for men.

c) The trend in the smoking rates for women seems a bit straighter than the trend for men. The apparent curvature in the scatterplot for the men could possibly be due to just a few points, and not indicate a serious violation of the linearity condition.

29. Oakland passengers 2016.

a) There are several features to comment on in this plot. There is a strong monthly pattern around the general trend. From 1997 to 2008, passengers increased fairly steadily with a notable exception of Sept. 2001, probably due to the attack on the twin towers. Then sometime in late 2008, departures dropped dramatically, possibly due to the economic crisis. Recently, they have been recovering, but not at the same rate as their previous increase.

b) The trend was fairly linear until late 2008, then passengers dropped suddenly.

c) The trend since 2009 has been linear (overall, ignoring monthly oscillations) If the increase continues to be linear, the predictions should be reasonable for the short term.

30. Tracking hurricanes 2016.

a) According to the linear model, tracking errors averaged about 133 nautical miles in 1970, and have decreased an average of 2.06 nautical miles per year since then.

b) Residuals based on this model have a standard deviation of 15.44 nautical miles.

c) The linear model for the trend in predicting error is $\widehat{Error} = 133.130 - 2.067(Year - 1970)$.

$\widehat{Error} = 133.130 - 2.067(Year - 1970) = 133.130 - 2.067(50) \approx 29.78$; The model predicts an error of 29.78 nautical miles in 2020. This is consistent with the goal of achieving average tracking errors below 45 nautical miles. Of course, this is dependent on a continuation of the pattern.

d) A tracking error of 25 nautical miles is likely to be achieved by about 2023, if the trend fit by the regression model continues, but this is an extrapolation beyond the data. Given their current rate of improvement, 25 nautical miles seems overly optimistic. (This is determined by testing different years in the model. This model is NOT meant to be used "backwards", substituting in desired errors.)

e) We should be cautious in assuming that the improvements in prediction will continue at the same rate. They may improve faster, and perhaps the goal will be met. We can't say with any certainty.

31. Unusual points.

a) **1)** The point has high leverage and a small residual.
2) The point is not influential. It has the *potential* to be influential, because its position far from the mean of the explanatory variable gives it high leverage. However, the point is not *exerting* much influence, because it reinforces the association.
3) If the point were removed, the correlation would become weaker. The point heavily reinforces the positive association. Removing it would weaken the association.
4) The slope would remain roughly the same, since the point is not influential.

b) **1)** The point has high leverage and probably has a small residual.
2) The point is influential. The point alone gives the scatterplot the appearance of an overall negative direction, when the points are actually fairly scattered.
3) If the point were removed, the correlation would become weaker. Without the point, there would be very little evidence of linear association.
4) The slope would increase, from a negative slope to a slope near 0. Without the point, the slope of the regression line would be nearly flat.

c) **1)** The point has moderate leverage and a large residual.
2) The point is somewhat influential. It is well away from the mean of the explanatory variable, and has enough leverage to change the slope of the regression line, but only slightly.
3) If the point were removed, the correlation would become stronger. Without the point, the positive association would be reinforced.
4) The slope would increase slightly, becoming steeper after the removal of the point. The regression line would follow the general cloud of points more closely.

d) **1)** The point has little leverage and a large residual.
2) The point is not influential. It is very close to the mean of the explanatory variable, and the regression line is anchored at the point $(\bar{x}, \bar{y})$, and would only pivot if it were possible to minimize the sum of the squared residuals. No amount of pivoting will reduce the residual for the stray point, so the slope would not change.
3) If the point were removed, the correlation would become slightly stronger, decreasing to become more negative. The point detracts from the overall pattern, and its removal would reinforce the association.
4) The slope would remain roughly the same. Since the point is not influential, its removal would not affect the slope.

32. More unusual points.

a) **1)** The point has high leverage and a large residual.
2) The point is influential. It is well away from the mean of the explanatory variable, and has enough leverage to change the slope of the regression line.
3) If the point were removed, the correlation would become stronger. Without the point, the positive association would be reinforced.
4) The slope would increase, becoming steeper after the removal of the point. The regression line would follow the general cloud of points more closely.

b) **1)** The point has high leverage and a small residual.
2) The point is influential. The point alone gives the scatterplot the appearance of an overall positive direction, when the points are actually fairly scattered.
3) If the point were removed, the correlation would become weaker. Without the point, there would be very little evidence of linear association.
4) The slope would decrease, from a positive slope to a slope near 0. Without the point, the slope of the regression line would be nearly flat.

32. (continued)

c) **1)** The point has little leverage and a large residual.

2) The point is not influential. It is very close to the mean of the explanatory variable, and the regression line is anchored at the point $(\bar{x}, \bar{y})$, and would only pivot if it were possible to minimize the sum of the squared residuals. No amount of pivoting will reduce the residual for the stray point, so the slope would not change.

3) If the point were removed, the correlation would become slightly stronger. The point detracts from the overall pattern, and its removal would reinforce the association.

4) The slope would remain roughly the same. Since the point is not influential, its removal would not affect the slope.

d) **1)** The point has high leverage and a small residual.

2) The point is not influential. It has the *potential* to be influential, because its position far from the mean of the explanatory variable gives it high leverage. However, the point is not *exerting* much influence, because it reinforces the association.

3) If the point were removed, the correlation would become weaker. The point heavily reinforces the association. Removing it would weaken the association.

4) The slope would remain roughly the same, since the point is not influential.

33. The extra point.

1) Point e is very influential. Its addition will give the appearance of a strong, negative correlation like $r = -0.90$.

2) Point d is influential (but not as influential as point e). Its addition will give the appearance of a weaker, negative correlation like $r = -0.40$.

3) Point c is directly below the middle of the group of points. Its position is directly below the mean of the explanatory variable. It has no influence. Its addition will leave the correlation the same, $r = 0.00$.

4) Point b is almost in the center of the group of points, but not quite. Its addition will give the appearance of a very slight positive correlation like $r = 0.05$.

5) Point a is very influential. Its addition will give the appearance of a strong, positive correlation like $r = 0.75$.

34. The extra point revisited.

1) Point d is influential. Its addition will pull the slope of the regression line toward point d, resulting in the steepest negative slope, a slope of –0.45.

2) Point e is very influential, but since it is far away from the group of points, its addition will only pull the slope down slightly. The slope is –0.30.

3) Point c is directly below the middle of the group of points. Its position is directly below the mean of the explanatory variable. It has no influence. Its addition will leave the slope the same, 0.

4) Point b is almost in the center of the group of points, but not quite. It has very little influence, but what influence it has is positive. The slope will increase very slightly with its addition, to 0.05.

5) Point a is very influential. Its addition will pull the regression line up to its steepest positive slope, 0.85.

35. What's the cause?

1) High blood pressure may cause high body fat.

2) High body fat may cause high blood pressure.

3) Both high blood pressure and high body fat may be caused by a lurking variable, such as a genetic or lifestyle trait.

36. What's the effect?

1) Playing computer games may make kids more violent.

2) Violent kids may like to play computer games.

3) Playing computer games and violence may both be caused by a lurking variable such as the child's home life or a genetic predisposition to aggressiveness.

37. Reading.

a) The principal's description of a strong, positive trend is misleading. First of all, "trend" implies a change over time. These data were gathered during one year, at different grade levels. To observe a trend, one class's reading scores would have to be followed through several years. Second, the strong, positive relationship only indicates the yearly improvement that would be expected, as children get older. For example, the 4th graders are reading at approximately a 4th grade level, on average. This means that the school's students are progressing adequately in their reading, not extraordinarily. Finally, the use of average reading scores instead of individual scores increases the strength of the association.

b) The plot appears very straight. The correlation between grade and reading level is very high, probably between 0.9 and 1.0.

c) If the principal had made a scatterplot of all students' scores, the correlation would have likely been lower. Averaging reduced the scatter, since each grade level has only one point instead of many, which inflates the correlation.

d) If a student is reading at grade level, then that student's reading score should equal his or her grade level. The slope of that relationship is 1. That would be "acceptable", according to the measurement scale of reading level. Any slope greater than 1 would indicate above grade level reading scores, which would certainly be acceptable as well. A slope less than 1 would indicate below grade level average scores, which would be unacceptable.

38. Grades.

Perhaps the best way to start is to discuss the type of graph that would have been useful. The admissions officer should have made a scatterplot with a coordinate for each freshman, matching each individual's SAT score with his or her respective GPA. Then, if the cloud of points was straight enough, the officer could have attempted to fit a linear model, and assessed its appropriateness and strength.

As is, the graph of combined SAT score versus mean Freshman GPA indicates, very generally, that higher SAT achievement is associated with higher mean Freshman GPA, but that's about it.

The first concern is the SAT scores. They have been grouped into categories. We cannot perform any type of regression analysis, because this variable is not quantitative. We don't even know how many students are in each category. There may be one student with an SAT score in the 1500s, and 300 students in the 1200s. On this graph, these possibilities are given equal weight!

Even if the SAT scores were at all useful to us, the GPAs given are averages, which would make the association appear stronger than it actually is.

Finally, a connected line graph isn't a useful model. It doesn't simplify the situation at all, and may, in fact, give the false impression that we could interpolate between the data points.

39. Heating.

a) The model predicts a decrease in $2.13 in heating cost for an increase in temperature of 1° Fahrenheit. Generally, warmer months are associated with lower heating costs.

b) When the temperature is 0° Fahrenheit, the model predicts a monthly heating cost of $133.

c) When the temperature is around 32° Fahrenheit, the predictions are generally too high. The residuals are negative, indicating that the actual values are lower than the predicted values.

39. (continued)

d) $\widehat{Cost} = 133 - 2.13(Temp) = 133 - 2.13(10) = \$111.70$; According to the model, the heating cost in a month with average daily temperature 10° Fahrenheit is expected to be $111.70.

e) The residual for a 10° day is approximately –$6, meaning that the actual cost was $6 less than predicted, or $111.70 – $6 = $105.70.

f) The model is not appropriate. The residuals plot shows a definite curved pattern. The association between monthly heating cost and average daily temperature is not linear.

g) A change of scale from Fahrenheit to Celsius would not affect the relationship. Associations between quantitative variables are the same, no matter what the units.

40. Speed.

a) The model predicts that as speed increases by 1 mile per hour, the fuel economy is expected to decrease by 0.1 miles per gallon.

b) For this model, the *y*-intercept is the predicted mileage at a speed of 0 miles per hour. It's not possible to get 32 miles per gallon if you aren't moving.

c) The residuals are negative for the higher gas mileages. This means that the model is predicting higher than the actual mileage.

d) $\widehat{mpg} = 32 - 0.1mph = 32 - 0.1(50) = 27$; When a car is driven at 50 miles per hour, the model predicts mileage of 27 miles per gallon.

e) $\widehat{mpg} = 32 - 0.1mph = 32 - 0.1(45) = 27.5$; When a car is driven at 45 miles per hour, the model predicts mileage of 27.5 miles per gallon. From the graph, the residual at 27.5 mpg is +1. The actual gas mileage is 27.5 + 1 = 28.5 mpg.

f) The association between fuel economy and speed is probably quite strong, but not linear.

g) The linear model is not the appropriate model for the association between fuel economy and speed. The residuals plot has a clear pattern. If the linear model were appropriate, we would expect scatter in the residuals plot.

41. TBill rates 2016.

a) $r = \sqrt{R^2} = \sqrt{0.776} = 0.881$; The correlation between rate and year is +0.881, since the scatterplot shows a positive association.

b) According to the model, treasury bill rates during this period increased at about 0.25% per year, starting from an interest rate of about 0.61%in 1950.

c) The linear regression equation predicting interest rate from year is $\widehat{Rate} = 0.61149 + 0.24788(Year - 1950)$ $= 0.61149 + 0.24788(70) = 17.96$.

According to the model, the interest rate is predicted to be about 18% in the year 2020.

d) This prediction is not likely to have been a good one. Extrapolating 40 years beyond the final year in the data would be risky, and unlikely to be accurate.

42. Marriage age, 2015.

- **a)** The correlation between age difference and year is $r = \sqrt{R^2} = \sqrt{0.775} \approx -0.880$. The negative value is used since the scatterplot shows that the association is negative, strong, and linear.
- **b)** The linear regression model that predicts age difference from year is: $\widehat{(Men - Women)} = 33.604 - 0.01582\,Year$. This model predicts that each passing year is associated with a decrease of approximately 0.01646 years in the difference between male and female marriage age. A more meaningful comparison might be to say that the model predicts a decrease of approximately 0.1646 years in the age difference for every 10 years that pass.
- **c)** $\widehat{(Men - Women)} = 33.604 - 0.01582\,Year = 33.604 - 0.01582(2020) = 1.6476;$ According to the model, the age difference between men and women at first marriage is expected to be approximately 1.65 years. (This figure is very sensitive to the number of decimal places used in the model.)
- **d)** The latest data point is for the year 2015. Extrapolating for 2020 is risky because it depends on the assumption that the trend in age at first marriage will continue in the same manner.

43. TBill rates 2016 revisited.

- **a)** Treasure bill rates peaked around 1980 and decreased afterward. This regression model has a negative slope and a high intercept.
- **b)** The model that predicts the interest rate on 3-month Treasury bills from the number of years since 1950 is $\widehat{Rate} = 18.5922 - 0.29646(Year - 1950)$. This model predicts the interest rate to be –2.16%%, a negative rate, which doesn't make sense. This is much lower than the prediction made with the other model.
- **c)** Even though we separated the data, there is no way of knowing if this trend will continue. And the rate cannot become negative, so we have clearly extrapolated far beyond what the data can support.
- **d)** It is clear from the scatterplot that we can't count on TBill rates to change in a linear way over many years, since we have witnessed one change in direction already. It would not be wise to use any regression model to predict rates.

44. Marriage age 2015 again.

- **a)** The linear model is appropriate, since the scatterplot of the relationship between difference in age at first marriage and the year is reasonably straight, and the residuals plot is scattered. However, this is the kind of pattern that we can't expect to continue in a linear way, so the model may not be very useful. In the most recent years, the difference seems to be growing, which may suggest a change in behavior.
- **b)** For every 10 years that pass, the model predicts a decrease of approximately 0.23 years in average age difference at first marriage.
- **c)** The y-intercept is the prediction of the model in year 0, over 2000 years ago. An extrapolation that far into the past is not meaningful. The earliest year for which we have data is 1980.

45. Gestation.

- **a)** The association would be stronger if humans were removed. The point on the scatterplot representing human gestation and life expectancy is an outlier from the overall pattern and detracts from the association. Humans also represent an influential point. Removing the humans would cause the slope of the linear regression model to increase, following the pattern of the non-human animals much more closely.
- **b)** The study could be restricted to non-human animals. This appears justifiable, since one could point to a number of environmental factors that could influence human life expectancy and gestation period, making them incomparable to those of animals.
- **c)** The correlation is moderately strong. The model explains 72.2% of the variability in gestation period of non-human animals.

45. (continued)

d) For every year increase in life expectancy, the model predicts an increase of approximately 15.5 days in gestation period.

e) $\widehat{Gest} = -39.5172 + 15.4980\,LifEx = -39.5172 + 15.4980(20) \approx 270.4428;$ According to the linear model, monkeys with a life expectancy of 20 years are expected to have gestation periods of about 270.5 days. Care should be taken when assessing the accuracy of this prediction. First of all, the residuals plot has not been examined, so the appropriateness of the model is questionable. Second, it is unknown whether or not monkeys were included in the original 17 non-human species studied. Since monkeys and humans are both primates, the monkeys may depart from the overall pattern as well.

46. Swim the lake 2016.

a) Only 1.1% of the variability in lake swim times is accounted for by the linear model.

b) The slope of the regression, 2.544, means that the model predicts that lake swim times are increasing by about 2.544 minutes per year. This means that lake swimmers are generally getting slower. However, this model has very weak predicting power, and an outlier, so we shouldn't put too much faith in our prediction.

c) Removing this outlier is probably a good idea, since it doesn't belong with the other data points, but its removal probably wouldn't change the regression much. The fact that the point has a large residual indicates that it didn't have much leverage. If it had leverage, it would have dominated the regression, and had a small residual. It would be nice to have a scatterplot to look at, in addition to the residuals plot. There could be other outliers that don't show up in the residuals plot.

47. Elephants and hippos.

a) Hippos are more of a departure from the pattern. Removing that point would make the association appear to be stronger.

b) The slope of the regression line would increase, pivoting away from the hippos point.

c) Anytime data points are removed, there must be a justifiable reason for doing so, and saying, "I removed the point because the correlation was higher without it" is not a justifiable reason.

d) Elephants are an influential point. With the elephants included, the slope of the linear model is 15.4980 days gestation per year of life expectancy. When they are removed, the slope is 11.6 days per year. The decrease is significant.

48. Another swim.

a) The smaller value of s_e means that errors in prediction are smaller for this model than the original model.

b) The regression accounts for only 2.8% of the variation in lake swim times, but it appears that Lake Ontario swimmers are getting slower, at a rate of about 3.024 seconds per year. But again, with such a low value of R^2, we shouldn't put much faith in this prediction.

49. Marriage age 2015 predictions.

a) The linear model used to predict average female marriage age from year is $\widehat{Age} = -112.543 + 0.068479\,Year.$ The residuals plot shows a clear pattern. The model predicts that each year that passes is associated with an increase of 0.068479 years in the average female age at first marriage. The model predicts that the average female marriage age in 2025 will be approximately 26.13 years.

b) Don't place too much faith in this prediction. The residuals plot shows a clear pattern, indicating that this model is not appropriate. Additionally, this prediction is for a year that is 10 years higher than the highest year for which we have an average female marriage age. Extrapolation is risky when making predictions.

c) An extrapolation of more than 50 years into the future would be absurd. There is no reason to believe the trend would continue.

49. (continued)

d) The linear model used to predict average female marriage age from year is: $\widehat{Age} = -274.742 + 0.149983Year.$ The residuals plot shows a clear pattern. The model predicts that each year that passes is associated with an increase of 0.149983 years in the average female age at first marriage. The model predicts that the average female marriage age in 2025 will be approximately 28.97 years.

e) Don't place too much faith in this prediction. Though the residuals are quite small, the residuals plot shows a clear pattern, indicating that this model is not appropriate. Additionally, this prediction is for a year that is 10 years higher than the highest year for which we have an average female marriage age. Extrapolation is risky when making predictions. This model is better than the previous one, but still not a good model.

f) An extrapolation of more than 50 years into the future would be absurd. There is no reason to believe the trend would continue. Again, this would be an unreasonable extrapolation.

50. Bridges covered.

a) The linear model is $\widehat{Condition} = -33.4115 + 0.0196Year$, so a bridge built in 1853 is expected to have a condition of 2.9073. The residual is 4.57 – 2.91 = 1.66.

b) A point to the left of the overall group, and higher than expected would have high leverage. This point would pull the regression line toward it, lowering the regression slope.

c) The covered bridge does not fit the pattern we see in the scatterplot, so including it would lower R^2.

d) According to the linear model, a bridge built in 1972 is expected to have condition equal to $-33.4115 + 0.0196(1972) = 5.24$, which is closer to the actual condition of 4.57. When you consider the restoration, the bridge isn't remarkable.

51. Fertility and life expectancy 2014.

a) The association between fertility rate and life expectancy is moderate, linear, and negative. The residual plot is reasonably scattered with no evidence of nonlinearity, so we can fit the regression model. But there seems to be an outlier, which could be affecting the regression model.

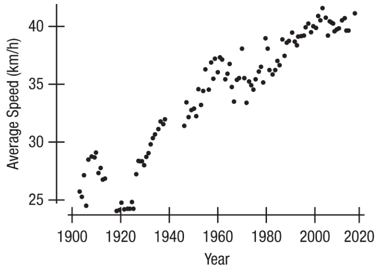
b) There is one outlier, Niger, with a higher life expectancy than typical for its large family size.

c) R^2 = 63.9%, so 63.9% of the variability in life expectancy is explained by variability in the fertility rate.

d) The government leaders should not suggest that women have fewer children in order to raise the life expectancy. Although there is evidence of an association between the fertility rate and life expectancy, this does not mean that one causes the other. There may be lurking variables involved, such as economic conditions, social factors, or level of health care.

52. Tour de France 2017.

a) The association between average speed and year is positive, moderate, but not quite linear. Generally, average speed of the winner has been increasing over time. There are several periods where the relationship is curved, but since 1950, the relationship has been much more linear. There are no races between 1915 and 1918 or between 1940 and 1946, presumably because of the two World Wars in Europe at the times.



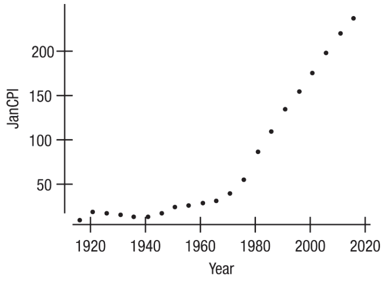
52. (continued)

b) $\widehat{Avgspeed} = -254.33 + 0.146\,Year$

c) The conditions for regression are not met. Although the variables are quantitative, and there are no outliers, the relationship is not straight enough in the early part of the 20th century to fit a regression line.

53. Inflation 2016.

a) The trend in Consumer Price Index is strong, non-linear, and positive. Generally, CPI has increased over the years, but the rate of increase has become much greater since approximately 1971. Other characteristics include fluctuations in CPI in the years prior to 1950.



b) Answers may vary. In order to effectively predict the CPI over the next decade, use only the most recent trend. The trend since 1971 is straight enough to apply the linear model. Prior to 1971, the trend is radically different from that of recent years, and is of no use in predicting CPI for the next decade.

The linear model that predicts CPI from year is $\widehat{CPI} = -8797.47 + 4.48388\,Year$. $R^2 = 99.7\%$, meaning that the model predicts 99.7% of the variability in CPI. The residuals plot shows some pattern, but the residuals are small, so the linear model is appropriate.

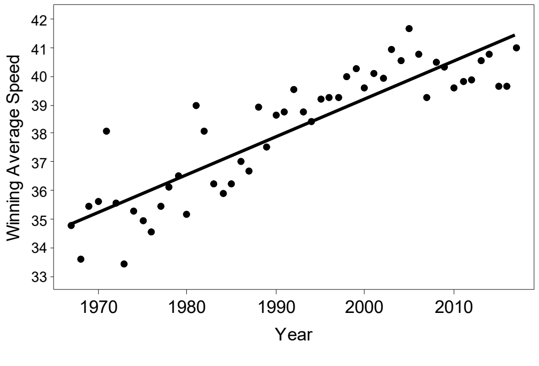
According to the model, the CPI is expected to increase by $4.48 each year, for 1971–2016.

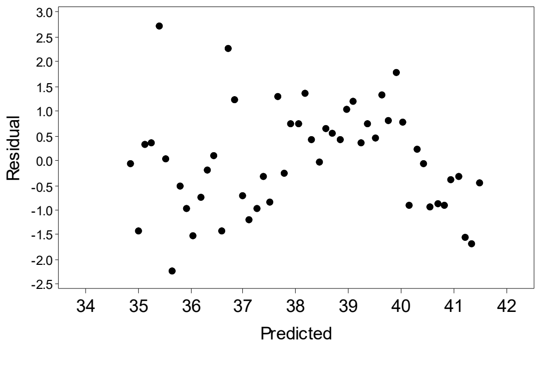
$\widehat{CPI} = -8797.47 + 4.48388\,Year = -8797.47 + 4.48388(2025) \approx 282$; As with any model, care should be taken when extrapolating. If the pattern continues, the model predicts that the CPI in 2025 will be approximately $282..

54. Second stage 2016.

a) The relationship between average winning speed and the year is much straighter with this subset of years.

The conditions for inference are met. Winning average speed and year are quantitative variables, and the scatterplot is Straight Enough to try regression. The residuals plot shows little pattern.





54. (continued)

```
Model Summary
S          R-sq       R-sq(adj)     R-sq(pred)
1.05326    78.15%     77.68%        76.02%

Coefficients
Term       Coef       SE Coef       T-Value      P-Value     VIF
Constant   –236.6     21.2          –11.17       0.000
Year       0.1379     0.0106        12.96        0.000       1.00
```

The new linear model is $\widehat{Avgspeed} = -236.568 + 0.137937\,Year$.

b) The 1999–2005 speeds are too high (due to doping), raising the line. Subsequent speeds are slower.

c) No, extrapolation to next year's average winning speed would not be advisable. As mentioned, the model was affected by the speeds that were later disqualified due to doping. Recent speeds seem to have leveled off at around 40 kph, so extrapolating a continued increase in speed won't work well.

55. Oakland passengers 2016 revisited.

a) The residuals cycle up and down because there is a yearly pattern in the number of passengers departing Oakland, California. There is also a sudden decrease in passenger traffic after 2008.

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

56. 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.

57. 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.

58. 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.

59. 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.

59. (continued)

d) $\widehat{Gal/100} = 0.625 + 1.178(Weight) = 0.625 + 1.178(3.5) = 4.748$; 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.

60. 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) Possibly. These outliers are exerting influence on the regression. A next step would be to try a "weaker" re-expression, like the reciprocal square root or log of GDP.

c) We could not see the outliers in the original scatterplot because the bend hid the outliers.

d) The GDP per capita of The Gambia in 2005 was about $433. The article was written in 2005, but the data may have been from slightly earlier. Correcting this GDP doesn't really make that much of a difference in the scatterplot. The Gambia is not as far away from the other data points now, but is still an outlier.

61. USGDP 2016.

a) Although nearly 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.

62. TBill rates 2016, once again.

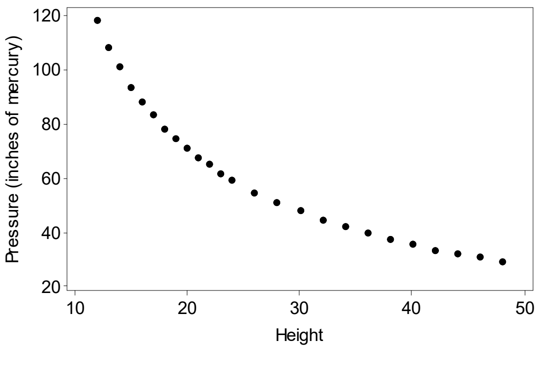
Re-expression should not be tried. An erratic trend that is positive then negative cannot be straightened by re-expression.

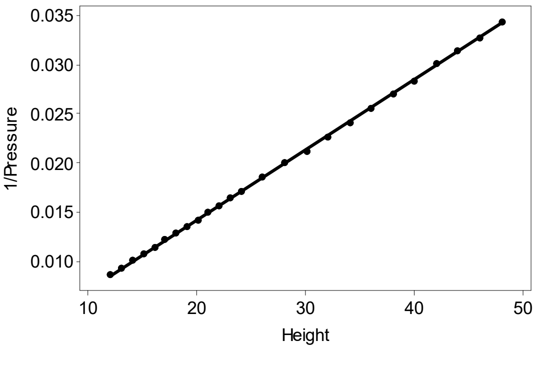
63. Better GDP model?

There is still a pattern in the residuals. This much pattern still indicates an inappropriate model. Since re-expressing with logarithms went too far, changing the curvature in the other direction, we should move up the ladder of powers. A square root re-expression might straighten the plot.

64. Boyle.

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.





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 on the next page.

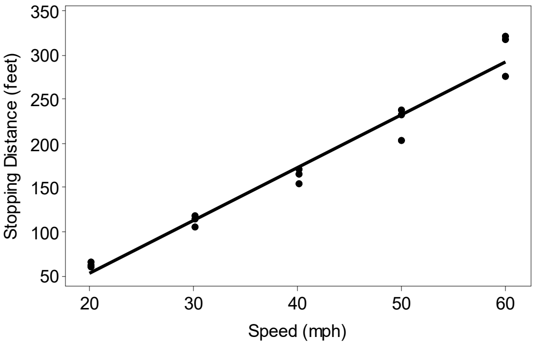
64. (continued)

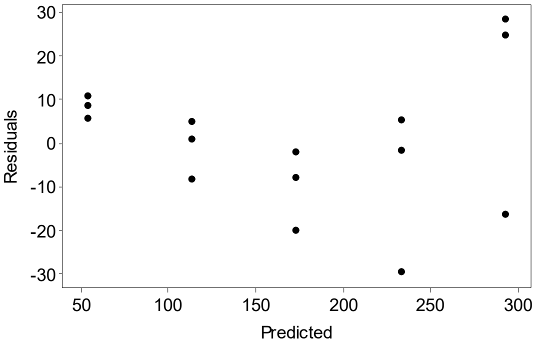
```
Dependent variable is:        recip pressure
No Selector
R squared = 100.0%    R squared (adjusted) = 100.0%
s =  0.0001  with  12 - 2 = 10  degrees of freedom

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
```

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

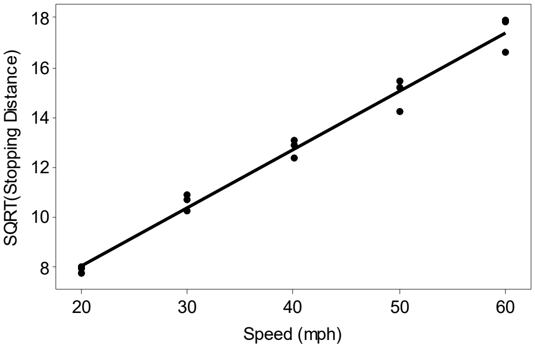
65. Brakes.

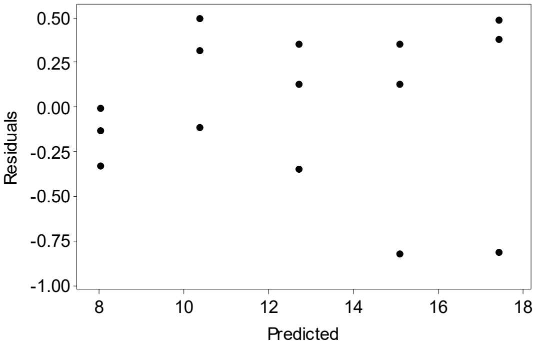
a) The association between speed and stopping distance is strong, positive, and appears straight. Higher speeds are generally associated with greater stopping distances. The linear regression model, with equation $\widehat{Distance} = -65.9 + 5.98(Speed)$, has R^2 = 96.9%, meaning that the model explains 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.





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.





65. (continued)

c) The model for the re-expressed data is $\widehat{\sqrt{Distance}} = 3.303 + 0.235(Speed)$. 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.

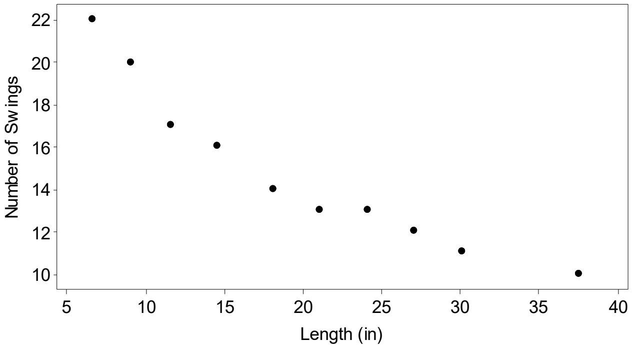
d) $\widehat{\sqrt{Distance}} = 3.303 + 0.235(Speed) = 3.303 + 0.235(55) = 16.228$, so $\widehat{Distance} = 16.228^2 \approx 263.4$; According to the model, a car traveling 55 mph is expected to require approximately 263.4 feet to come to a stop.

e) $\widehat{\sqrt{Distance}} = 3.303 + 0.235(Speed) = 3.303 + 0.235(70) = 19.753$, so $\widehat{Distance} = 19.753^2 \approx 390.2$; According to the model, a car traveling 70 mph is expected to require approximately 390.2 feet to come to a stop.

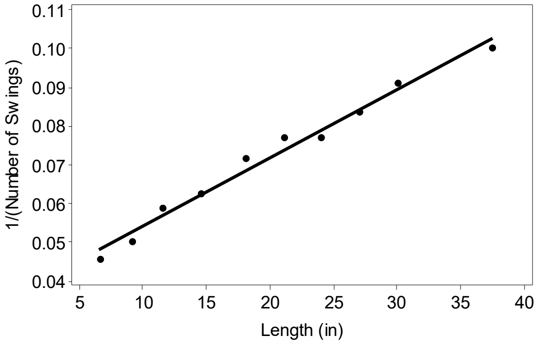
f) The level of confidence in the predictions should be quite high. R^2 is high, and the residuals plot is scattered. The prediction for 70 mph is a bit of an extrapolation, but should still be reasonably close.

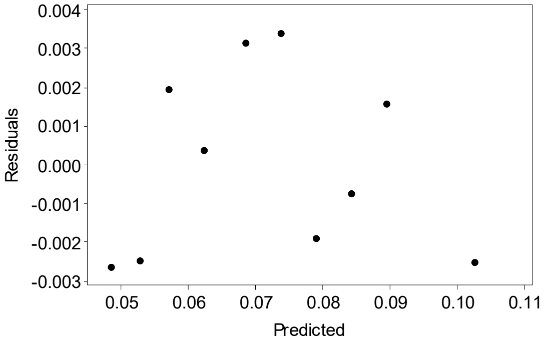
66. Pendulum.

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.



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





66. (continued)

c) The reciprocal re-expression yields the model $\widehat{\frac{1}{Swings}} = 0.0367 + 0.00176(Length)$. 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.

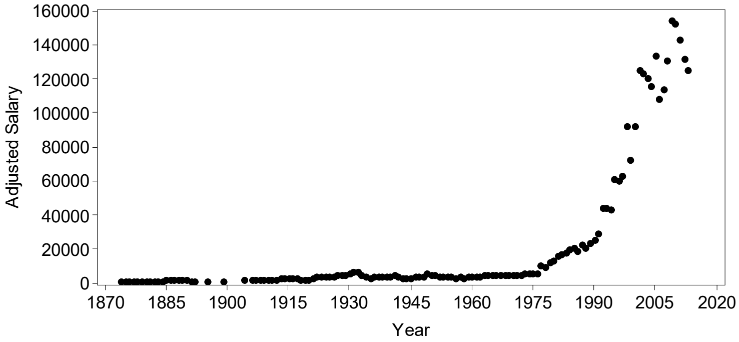
d) $\widehat{\frac{1}{Swings}} = 0.0367 + 0.00176(Length) = 0.0367 + 0.00176(4) = 0.04374$, so $\widehat{Swings} = \frac{1}{0.04374} \approx 22.9$; According to the reciprocal model, a pendulum with a 4" string is expected to swing approximately 22.9 times in 20 seconds.

e) $\widehat{\frac{1}{Swings}} = 0.0367 + 0.00176(Length) = 0.0367 + 0.00176(48) = 0.12118$, so $\widehat{Swings} = \frac{1}{0.12118} \approx 8.3$; The model predicts 8.3 swings in 20 seconds for a 48" string.

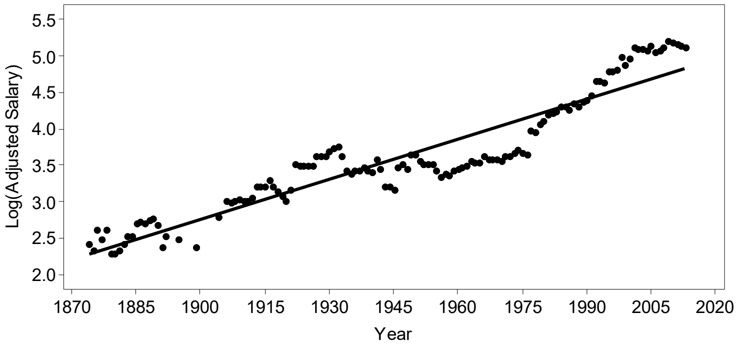
f) Confidence in the predictions is fairly high. The model is appropriate, as indicated by the scattered residuals plot, and accurate, indicated by the high value of R^2. The only concern is the fact that these predictions are slight extrapolations. The lengths of the strings aren't too far outside the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

67. Baseball salaries 2015.

a) The association between year and highest salary is curved, strong, and positive. Salaries were flat for many years, and began to increase in the late 1970s, then increased more rapidly as in recent years. The trend is not linear and the plot is not straight enough for a regression model.



b) Re-expression using the logarithm of the adjusted salaries straightens the plot significantly.



67. (continued)

c) Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.278834	87.68%	87.58%	87.32%

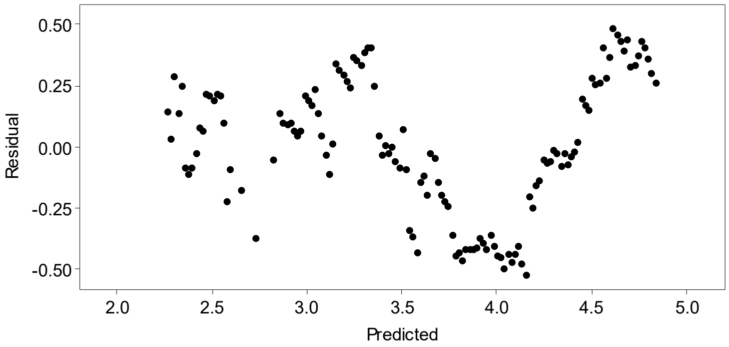
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	–32.3897	1.19	–27.15	0.000	
Year	0.018492	0.000613	30.18	0.000	1.00

The equation of the model is $\widehat{\log(AdjSalary)} = -32.3897 + 0.018492(Year)$. Salaries have been increasing on a log scale by about 0.018 logmillion \$/year. Transforming back to dollars, that's a bit more than a million dollars/year.

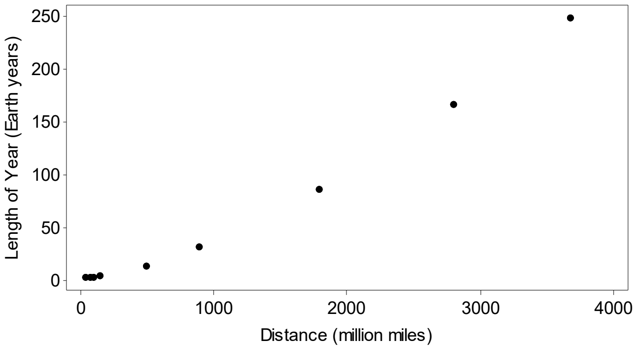
d) The plot of the residuals for the logarithmic model is at the right.

e) The residuals show a cyclic pattern. Salaries were lower than the model predicts at the beginning of the 20th century and again in the 1960s. They may have recently hit a high point and started turning lower, but it is difficult to tell. The model based on the logarithmic re-expression may be the best model we can find, but it doesn't explain the pattern in highest baseball salary over time.



68. Planets.

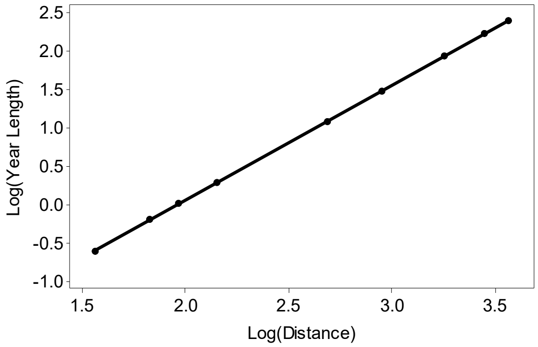
a) The association between distance from the sun and planet year is strong, positive, and curved concave upward. Generally, planets farther from the sun have longer years than closer planets.

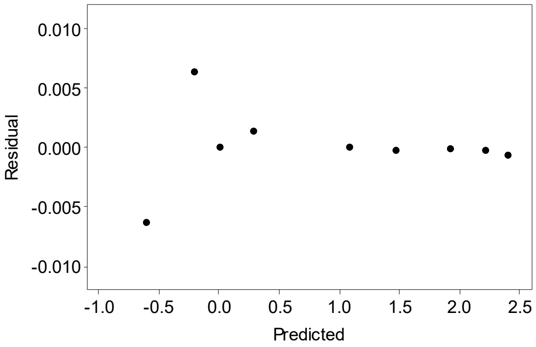


68. (continued)

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 for the linear model relating log(Distance) and log(Length of Year) appears below to the left. The residuals plot of the logarithmic re-expression, shown below to the right, shows little pattern. The log-log model appears to be a good one.

The regression model for the log-log re-expression is $\widehat{\log(Length)} = -2.95 + 1.5(\log(Distance))$.

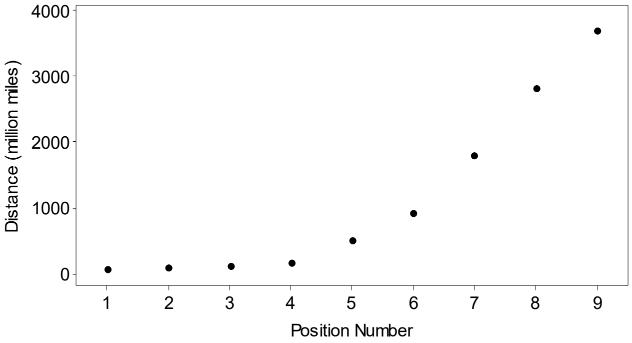




c) R^2 = 100%, so the model explains 100% of the variability in the log of the length of the planetary year, at least according to the accuracy of the statistical software. The residuals plot is scattered, and the residuals are all extremely small. This is a very accurate model.

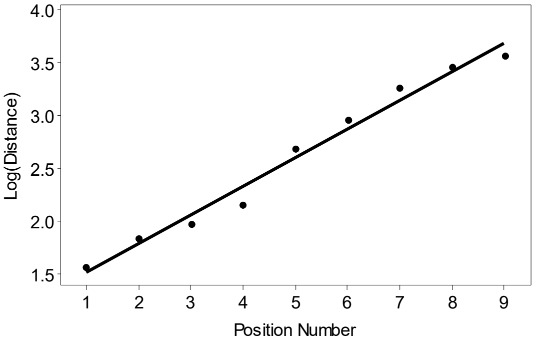
69. Is Pluto a planet?

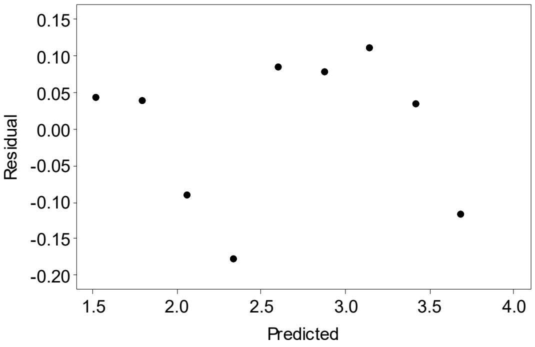
a) The association between planetary position and distance from the sun is strong, positive, and curved.



69. (continued)

A good re-expression of the data is Log(Distance) vs. Position. The scatterplot with regression line shows the straightened association. The equation of the model is $\widehat{\log(Distance)} = 1.246 + 0.271(Position)$. The residuals plot (below right) may have some pattern, but after trying several re-expressions, this is the best that can be done. $R^2 = 98.1\%$, so the model explains 98.1% of the variability in the log of the planet's distance from the sun.





b) At first glance, this model appears to provide little evidence to support the contention of the International Astronomical Union. 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.

$\widehat{\log(Distance)} = 1.245807 + 0.270792(Position) = 1.245807 + 0.270792(9) = 3.682935,$

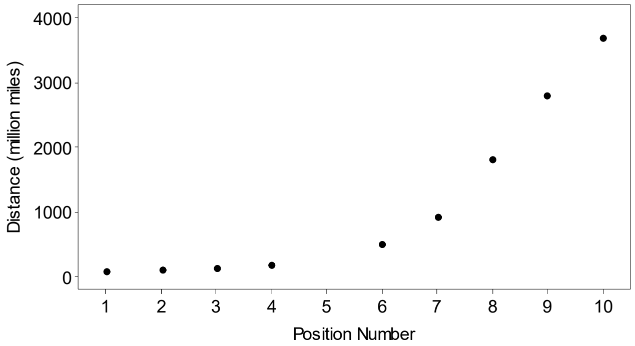
so $\widehat{Distance} = 10^{3.682935} \approx 4819;$ According to the model, the 9th planet in the solar system is predicted to be approximately 4819 million miles away from the sun. Pluto is actually 3672 million miles away.

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.

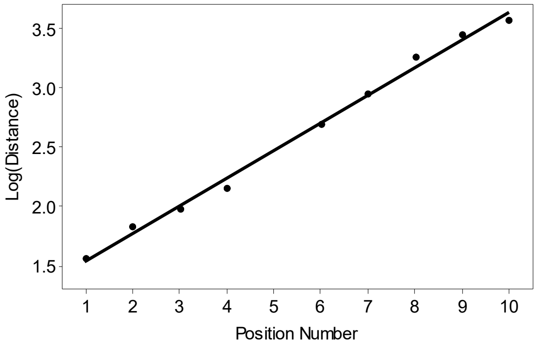
The model without Pluto, $\widehat{\log(Distance)} = 1.203650 + 0.283439(Position)$, works much better. It has a high R^2, and scattered residuals plot. This new model predicts that the 9th planet should be a whopping 5683 million miles away from the sun! There is evidence that the IAU is correct. Pluto doesn't behave like a planet in its relation to position and distance.

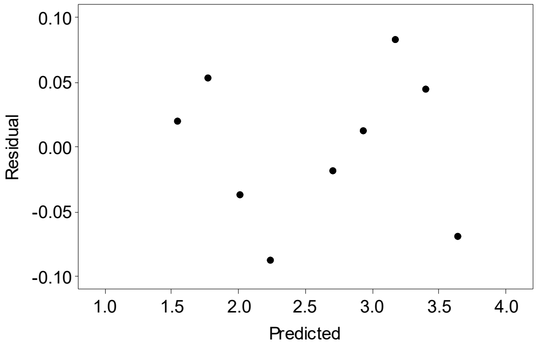
70. Planets, and asteroids.

Using the revised planetary numbering system, and straightening the scatterplot using the same methods as in Exercises 68 and 69, the new model, $\widehat{\log(Distance)} = 1.305655 + 0.232961(Position)$, is a slightly better fit. The residuals plot is more scattered, and R^2 is slightly higher, with the improved model explaining 99.4% 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 the 10th planet, should be about 4169 million miles away. That's about 497 million miles farther away than it is. A better model yet is $\widehat{\log(Distance)} = 1.28514 + 0.238826(Position)$, a model made with the new numbering system and with Pluto omitted.





71. Planets, and Eris.

$\widehat{\log(Distance)} = 1.28514 + 0.238826(Position) = 1.28514 + 0.238826(10) \approx 3.6734$, so

$\widehat{Distance} = 10^{3.6734} \approx 4714$; A planet tenth from the sun (with the asteroid belt as a failed planet and Pluto not included as a planet) is predicted to be about 4714 million miles away from the sun.

This distance is much shorter than the actual distance of Eris, about 6300 miles.

$\widehat{\log(Distance)} = 1.306310 + 0.232822(Position) = 1.306310 + 0.232822(11) = 3.867352$, so

$\widehat{Distance} = 10^{3.867352} \approx 7368$; Similarly, with the asteroid belt as a failed planet, and Pluto as the tenth planet, Eris, as the eleventh planet, is predicted to be 7368 million miles away, much farther away than it actually is. Like Pluto, Eris doesn't behave like a planet in its relation to position and distance.

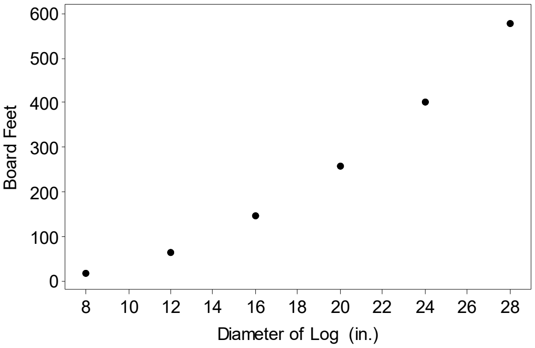
72. Planets, models, and laws.

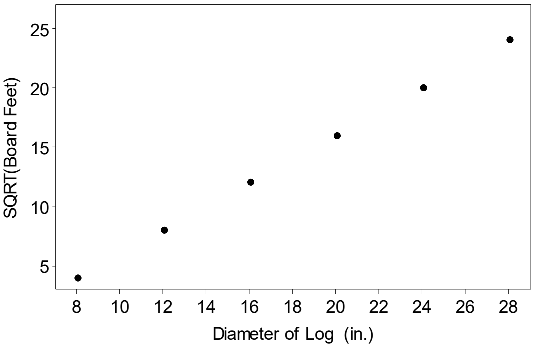
The re-expressed data relating distance and year length are better described by their model than the re-expressed data relating position and distance. The model relating distance and year length has R^2 = 100%, and a very scattered residuals plot (with miniscule residuals), possibly a natural "law". If planets in another planetary 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.

73. 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 $\widehat{\sqrt{BoardFeet}} = -4 + Diameter$.



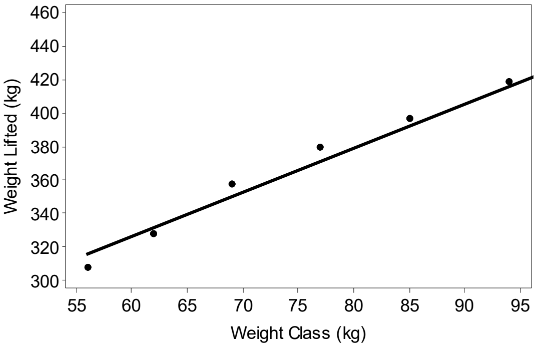


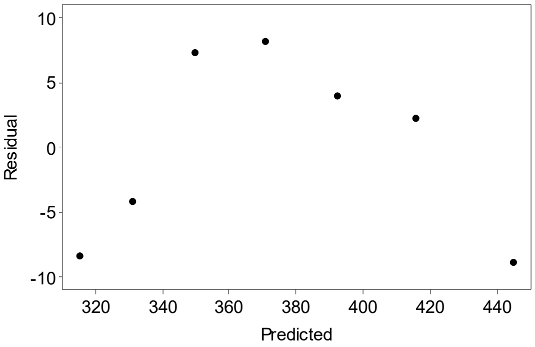
b) $\widehat{\sqrt{BoardFeet}} = -4 + Diameter = -4 + (10) = 6$, so $\widehat{BoardFeet} = 6^2 = 36$; According to the model, a log 10" in diameter is expected to contain 36 board feet of lumber.

c) $\widehat{\sqrt{BoardFeet}} = -4 + Diameter = -4 + (36) = 32$, so $\widehat{BoardFeet} = 32^2 = 1024$; 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.

74. Weightlifting 2016.

a) The association between weight class and weight lifted for world record-holders in weightlifting is strong, positive, and curved. The linear model that best fits the data is $\widehat{Lift} = 167.31 + 2.644(WeightClass)$. Although this model accounts for 97.7% of the variability in weight lifted, it does not fit the data well.



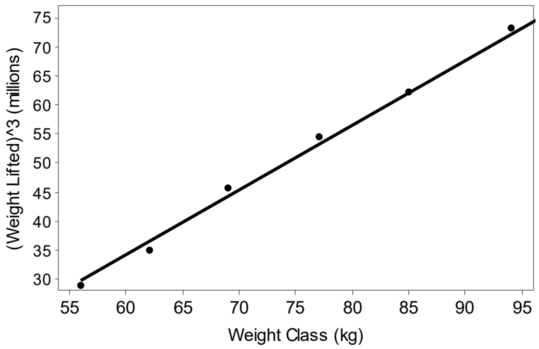


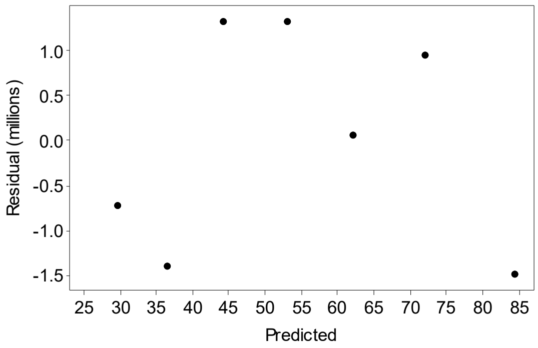
74. (continued)

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 might fit the association between weight class and weight lifted better than the linear model.

c) Answers may vary. The plot is curved downward, so move up the ladder of powers. Going further than a cube is overly complex. The model is $\widehat{Lift^3} = -32,873,285 + 1,116,730(Class)$.

We could also move the variable *Class* down the ladder of powers.





d) The cubic model is a better model, since the residuals plot shows little pattern. Additionally, the model accounts for 99.6% of the variability in weight lifted.

e) $\widehat{Lift} = 167.31 + 2.644(WeightClass)$

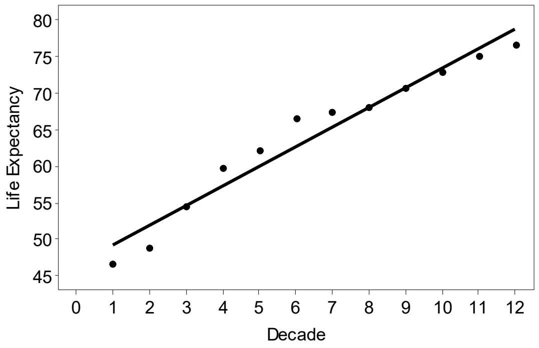
$\widehat{Lift} = 167.31 + 2.644(WeightClass) = 167.31 + 2.644(157) = 582.42$; According to the linear model, the world record of a 157 kg lifter is approximately 582.42 kg.

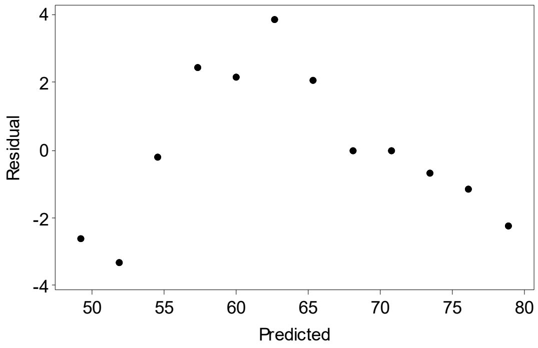
$\widehat{Lift^3} = -32,873,285 + 1,116,730(Class) = -32,873,285 + 1,116,730(157) = 142,453,325$, so $\widehat{Lift} = \sqrt[3]{142,453,325} \approx 522.26$; According to the cubic model, the world record of a 157 kg lifter is approximately 522.26 kg.

75. Life expectancy history.

The association between year and life expectancy is strong, curved and positive. As the years passed, life expectancy for white males has increased.

The linear model, $\widehat{LifeExp} = 46.57 + 2.68(Dec)$, explains 95.0% of the variability in the life expectancy of white males, but has a residuals plot that reveals a strong pattern.

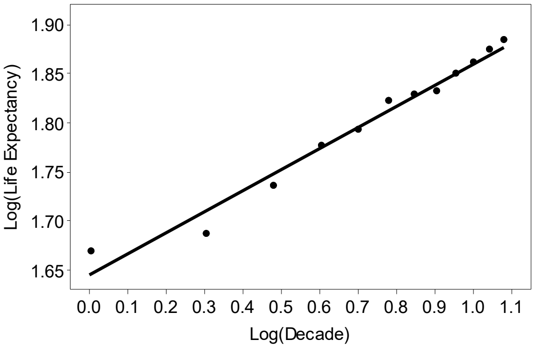


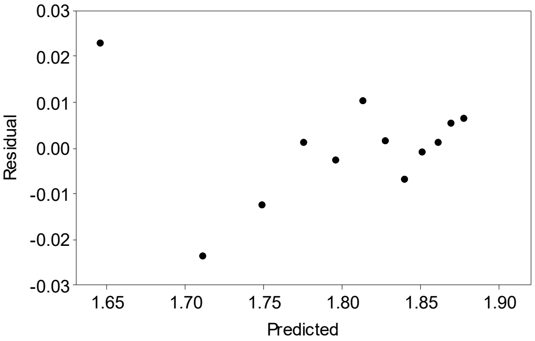


75. (continued)

The association between Log(*Year*) and Log(*Life Expectancy*) is strong, positive, and reasonably straight. The model is $\widehat{\log_{10}(LifeExp)} = 1.646 + 0.215(\log_{10}(Year))$.

This model explains 97.4% of the variability in the logarithm of the Life Expectancy. The residuals plot shows some pattern, but seems more scattered than the residuals plot for the linear model.

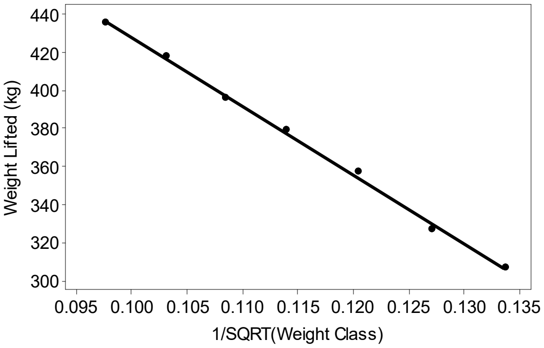


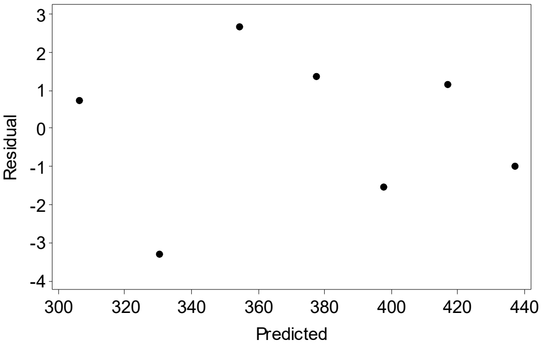


76. Lifting more weight.

a) Answers may vary. The reciprocal square root re-expression seems to straighten the scatterplot significantly.

$$\widehat{Lift} = 790.90 - \left(\frac{3626.46}{\sqrt{WeightClass}}\right)$$





b) The residual plot shows more scatter, and R^2 is higher, at 99.81%. This model appears to be the better one.

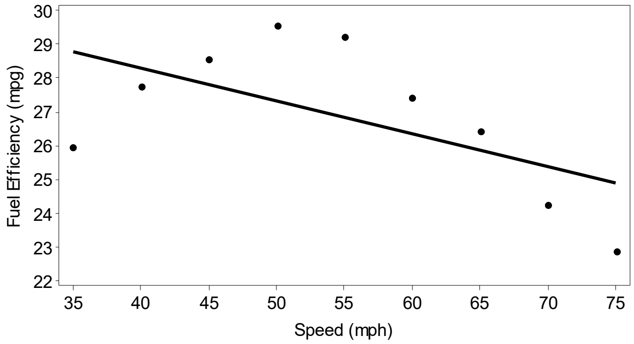
c) $\widehat{Lift} = 790.90 - \left(\frac{3626.46}{\sqrt{WeightClass}}\right) = 790.90 - \left(\frac{3626.46}{\sqrt{157}}\right) = 501.48$; The new model predicts that the 157 kg world record holder will lift approximately 501.48 kg.

d) This prediction is probably better, since the residuals are more scattered and the scatterplot is very straight.

e) All of the models over-predicted, but this one came the closest.

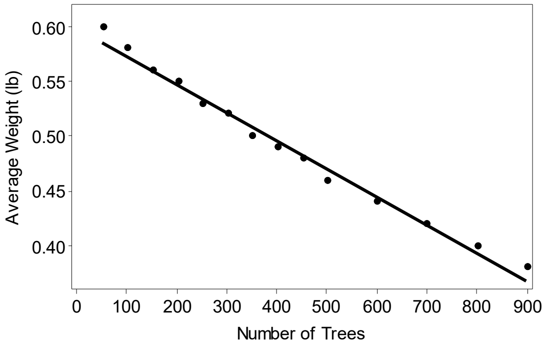
77. Slower is cheaper?

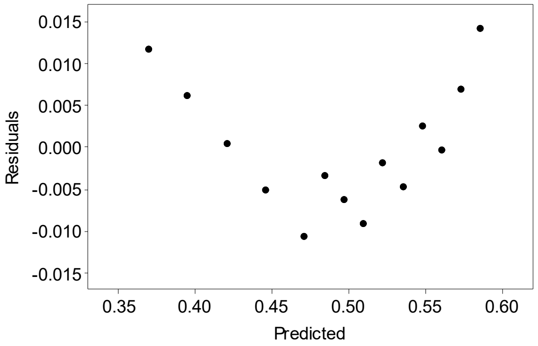
The scatterplot shows the relationship between speed and mileage of the compact car. The association is extremely strong and curved, with mileage generally increasing as speed increases, until around 50 miles per hour, then mileage tends to decrease as speed increases. The linear model is a very poor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



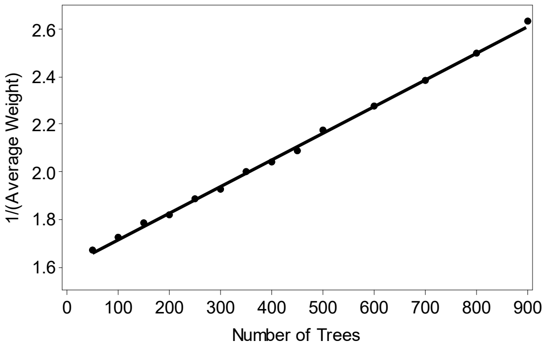
78. Oranges.

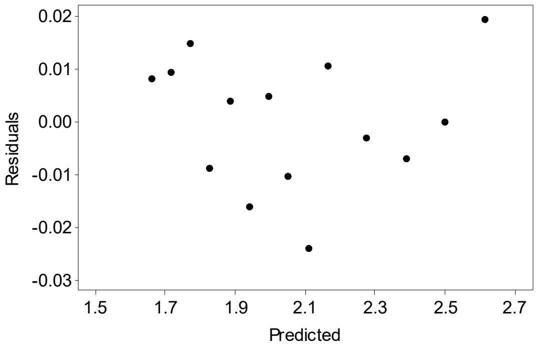
The association between the number of oranges per tree and the average weight is strong, negative, and appears linear at first look. Generally, trees that contain larger numbers of oranges have lower average weight per orange. The residuals plot shows a strong curved pattern. The data should be re-expressed.





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



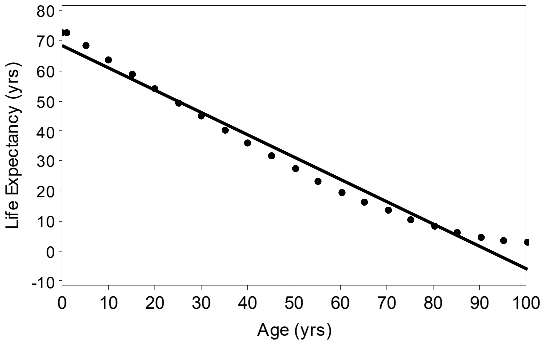


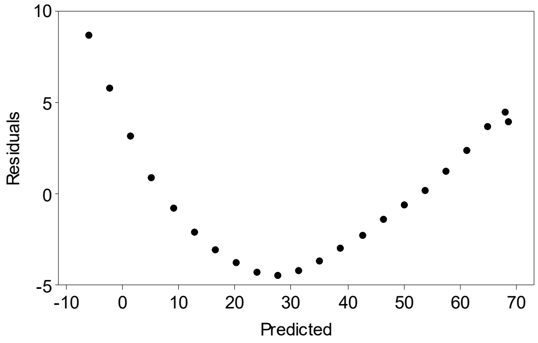
79. Years to live, 2016.

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 equation of the linear model that predicts life expectancy from age of black males is $\widehat{LifeExp} = 68.63 - 0.747\,Age.$

The model is not a good fit. The residuals plot shows a curved pattern.

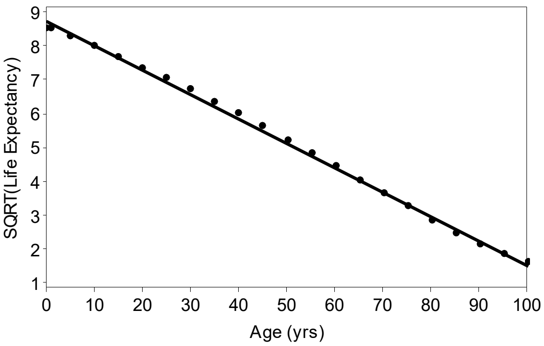


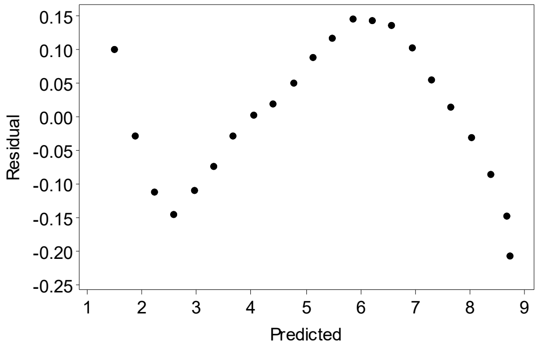


b) Answers may vary. The square root re-expression straightens the data considerably, but has an extremely patterned residuals plot. The model is not a mathematically appropriate model, but fits so closely that it should be fine for predictions within the range of data. The model explains 99.8% of the variability in the estimated number of additional years.

The equation of the square root model is $\widehat{\sqrt{LifeExp}} = 8.722 - 0.0721\,Age.$

$\widehat{\sqrt{LifeExp}} = 8.722 - 0.0721\,Age = 8.722 - 0.0721(18) = 7.4242,$ So $\widehat{LifeExp} = 7.4242^2 \approx 55.12;$ According to the model, an 18-year-old black male is expected to have 55.12 years of life remaining, and live to be approximately 73.12 years of age.

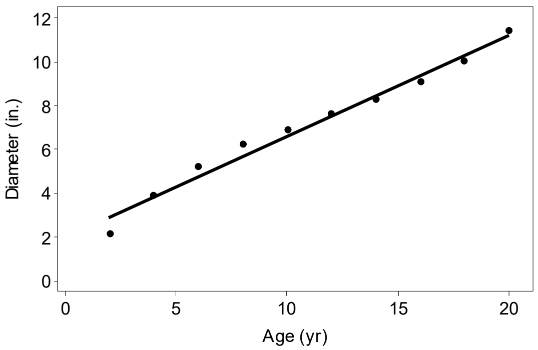


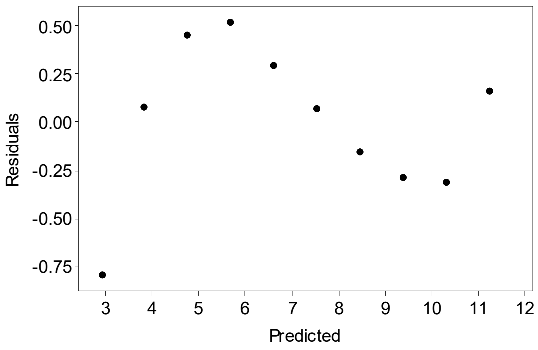


c) The residuals plot is extremely patterned, so the model is not appropriate. However, the residuals are very small, making for a tight fit. Since 18 years is within the range of the data, the prediction should be at least reasonable.

80. 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.





The linear model for this association, $\widehat{AverageDiameter} = 1.973 + 0.463(Age)$ 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.