Review of Part II – Exploring Relationships Between Variables

R2.1. College.

% over 50: $r = 0.69$	The only moderate, positive correlation in the list.
% under 20: $r = –0.71$	Moderate, negative correlation (–0.98 is too strong)
% Full-time Fac.: $r = 0.09$	No correlation.
% Gr. on time: $r = –0.51$	Moderate, negative correlation (not as strong as %under 20)

R2.2. Togetherness.

a) If no meals are eaten together, the model predicts a GPA of 2.73.

b) For an increase of one meal per week eaten together, the model predicts an increase of 0.11 in GPA.

c) The model will predict the mean GPA for the mean number of meals, 3.78.
$\widehat{GPA} = 2.73 + 0.11 Meals = 2.73 + 0.11(3.78) = 3.15$; The mean GPA is 3.15.

d) A negative residual means that the student’s actual GPA was lower than the GPA predicted by the model. The model over-predicted the student’s GPA.

e) Although there is evidence of an association between GPA and number of meals eaten together per week, this is not necessarily a cause-and-effect relationship. There may be other variables that are related to GPA and meals, such as parental involvement and family income.

R2.3. Vineyards.

a) There does not appear to be an association between ages of vineyards and the price of products.
$r = \sqrt{R^2} = \sqrt{0.027} = 0.164$, indicating a very weak association, at best. The model only explains 2.7% of the variability in case price. Furthermore, the regression equation appears to be influenced by two outliers, products from vineyards over 30 years old, with relatively high case prices.

b) This analysis tells us nothing about vineyards worldwide. There is no reason to believe that the results for the Finger Lakes region are representative of the vineyards of the world.

c) The linear equation used to predict case price from age of the vineyard is
$\widehat{CasePrice} = 92.765 + 0.567284 Years.$

d) This model is not useful because only 2.7% of the variability in case price is accounted for by the ages of the vineyards. Furthermore, the slope of the regression line seems influenced by the presence of two outliers, products from vineyards over 30 years old, with relatively high case prices.

R2.4. Vineyards again.

a) There is no evidence of an association between vineyard size and case price.

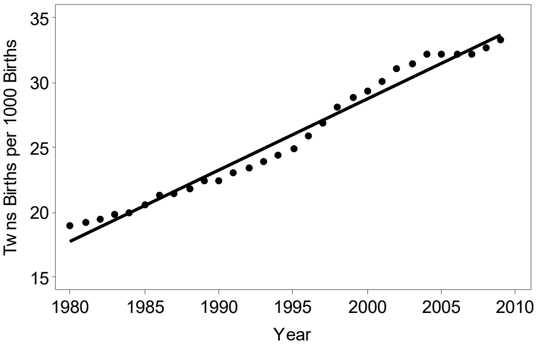
b) One vineyard is approximately 250 acres, with a relatively low case price. This point has high leverage.

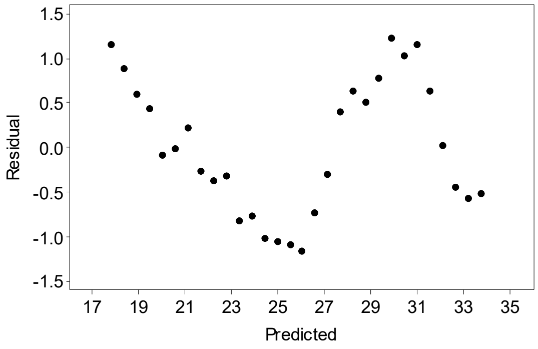
c) If the point were removed, the correlation would be expected to increase, from a slightly negative correlation, to a correlation that is slightly positive. The point is an outlier in the *x*-direction and low in the *y*-direction. It is giving the association the artificial appearance of a slightly negative relationship.

d) If the point were removed, the slope would be expected to increase, from slightly negative to slightly positive. The point is “pulling” the regression line down.

R2.5. Twins by year 2014.

a) The association between year and the twin birth rate is strong, positive, and appears non-linear. Generally, the rate of twin births has increased over the years. The linear model that predicts the rate of twin births from the year is $\widehat{Twins} = 17.77 + 0.551(\mathit{Years\ Since\ 1980})$.

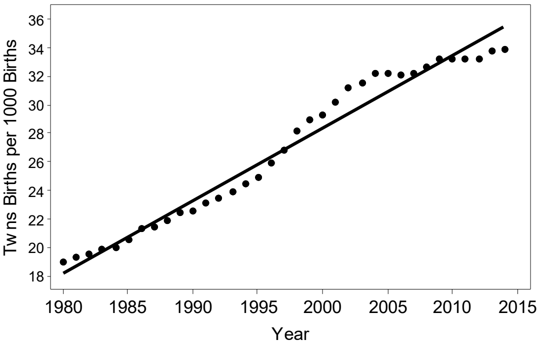


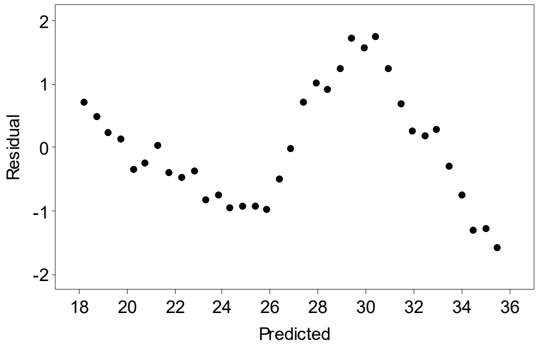


b) For each year that passes, the model predicts that the twin birth rate will increase by an average of approximately 0.55 twin births per 1000 live births.

c) $\widehat{Twins} = 17.77 + 0.551(\mathit{Years\ Since\ 1980}) = 17.77 + 0.551(34) = 36.504$; According to the model, the twin birth rate is expected to be 36.50 twin births per 1000 live births in the US in 2014. This is reasonably close to the actual rate of 33.22 twin births per 1000 women, even though this was an extrapolation based on a model with a highly patterned residuals plot. However, being close in hindsight is not justification for using this model to make this prediction.

d) The linear model that predicts the rate of twin births from the year is $\widehat{Twins} = 18.21 + 0.508(\mathit{Years\ Since\ 1980})$.





This model fits very well, with $R^2 = 97\%$. However, the residuals plot still shows the fluctuating pattern that troubled us before. There may be more to understand about twin births than is available from these data.

R2.6. Dow Jones 2015.

a) $r = \sqrt{R^2} = \sqrt{0.884} = 0.940$; Since the slope of the regression equation is positive, we know that the correlation is also positive.

b) The linear model that predicts Dow from year is $\widehat{DJIA} = -2891.43 + 379.917(\mathit{Year} - 1970)$.

R2.6. (continued)

c) This model predicts that the Dow was expected to be –2891.43 points in the year 1970, which doesn't have contextual meaning. Furthermore, the model predicts that the Dow is expected to increase by approximately 380 points each year, on average.

d) The residuals plot shows a definite pattern. A single linear model is not appropriate. Before attempting to fit a linear model, look at the scatterplot. If it is not straight enough, the linear model cannot be used.

R2.7. Streams.

a) $r = \sqrt{R^2} = \sqrt{0.27} = -0.5196$; The association between pH and BCI appears negative in the scatterplot, so use the negative value of the square root.

b) The association between pH and BCI is negative, moderate, and linear. Generally, higher pH is associated with lower BCI. Additionally, BCI appears more variable for higher values of pH.

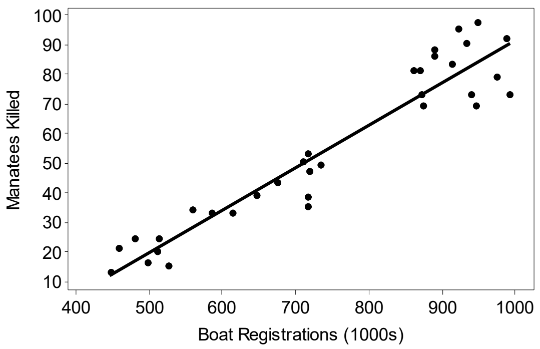
c) In a stream with average pH, the BCI would be expected to be average, as well.

d) $r \cdot (3) = -0.5196(3) = -1.56$; In a stream where the pH is 3 standard deviations above average, the BCI is expected to be 1.56 standard deviations below the mean level of BCI.

R2.8. Manatees 2015.

a) The explanatory variable is the number of powerboat registrations. This is the relationship about which the biologists are concerned. They believe that the high number of manatees killed is related to the increase in powerboat registrations.

b) The association between the number of powerboat registrations and the number of manatees killed in Florida is fairly strong, linear, and positive. Higher numbers of powerboat registrations are associated with higher numbers of manatees killed.



c) The correlation between the number of powerboat registrations and the number of manatees killed in Florida is $r = 0.951$.

d) $R^2 = 90.4\%$; Variability in the number of powerboat registrations accounts for 90.4% of the variability in the number of manatees killed.

e) There is an association between the number of powerboat registrations and the number of manatees killed, but that is no reason to assume a cause-and-effect relationship. There may be lurking variables that affect one or the other of the variables.

R2.9. Streams II.

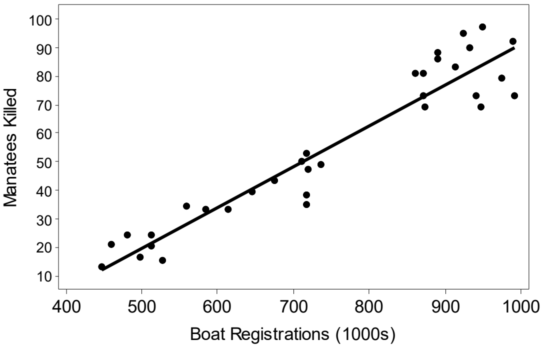
a) The association between pH and BCI is negative, moderate, and linear. Generally, higher pH is associated with lower BCI. The scatterplot (shown in Exercise 7) is straight enough to find a linear regression model. The linear regression model is $\widehat{BCI} = 2733.37 - 197.694pH$.

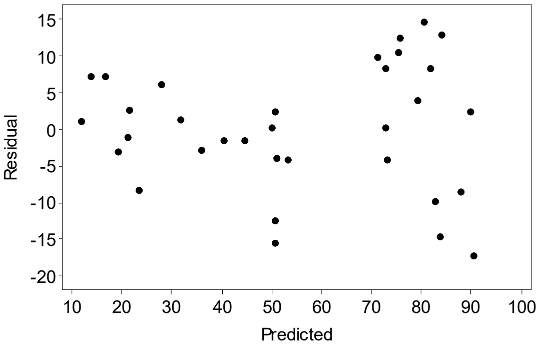
b) The model predicts that *BCI* decreases by an average of 197.69 point per point of *pH.*

c) $\widehat{BCI} = 2733.37 - 197.694pH = 2733.37 - 197.694(8.2) = 1112.2792$; The model predicts that a stream with *pH* of 8.2 will have a BCI of approximately 1112.3.

R2.10. A manatee model 2015.

a) The association between the number of powerboat registrations and the number of manatees killed is straight enough to try a linear model. The residuals plot is scattered, so the linear model is appropriate. $\widehat{ManateeDeaths} = -52.42 + 0.1439\,PowerboatRegistrations$ is the best fitting model.





b) For every additional 10,000 powerboats registered, the model predicts that an additional 1.439 manatees will be killed on average.

c) The model predicts that if no powerboats were registered, the number of manatee deaths would be approximately –52.41. This is an extrapolation beyond the scope of the data, and doesn't have much contextual meaning.

d) Negative residuals are better for the manatees. A negative residual suggests that the actual number of kills was below the number of kills predicted by the model.

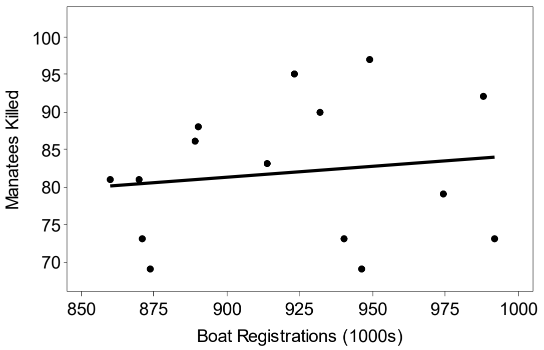
e) $\widehat{ManateeDeaths} = 56.07 + 0.028\,PowerboatRegistrations$ is the linear model for predicting manatee deaths from powerboat registrations, for the years 2001 to 2015.

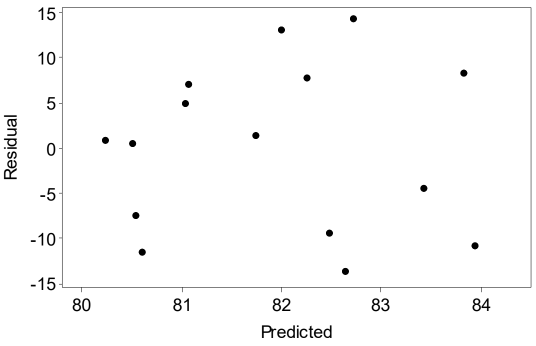
The residuals plot shows no pattern, indicating an appropriate model.

For every additional 10,000 powerboats registered, the model predicts that an additional 0.28 manatees will be killed on average.

This slope predicts almost no change in the number of manatee deaths with increases in the number or powerboats registered.

R2.10. (continued)





f) As the number of powerboat registrations has increased, the number of manatee deaths has increased. However, in the last 15 years, the slope has decreased dramatically. We don't know whether powerboat registrations will continue to increase. Even if they do increase, we don't know whether the trend will resume. It appears that efforts by the state of Florida to protect manatees may be having some effect.

R2.11. Streams III.

a) The regression model that predicts *BCI* from *pH* and water hardness is $\widehat{BCI} = 2342.95 - 137.833pH - 0.337210Hard$.

b) This model is slightly better than the model in the prior exercise. R^2 is slightly higher and *s* is slightly lower.

c) $\widehat{BCI} = 2342.95 - 137.833pH - 0.337210Hard = 2342.95 - 137.833(8.2) - 0.337210(205) = 1143.59135$; According to the multiple linear regression model, a stream with a *pH* of 8.2 and a hardness value of 205 is predicted to have *BCI* of approximately 1144.

d) $1309 - 1144 = 165$; This is the residual.

e) After allowing for the effects of *Hardness* of the water, *BCI* decreases by about 137.8 per unit of *pH.*

R2.12. Final exam.

a) $\widehat{Fin} = 10 + 0.9Mid = 10 + 0.9(70) = 73$; According to the model, Susan is predicted to earn a score of 73 on the final exam.

b) Susan's residual is 80 – 73 = 7 points. She scored 7 points higher than predicted.

c) $b_1 = r\frac{s_y}{s_x} \Rightarrow 0.9 = r\frac{12}{10} \Rightarrow r = 0.75$; The correlation between midterm exam score and final exam score is 0.75.

d) $\widehat{Fin} = 10 + 0.9Mid \Rightarrow 100 = 10 + 0.9Mid \Rightarrow Mid = 100$; In order to have a predicted final exam score of 100, a student would need to have a midterm exam score of 100, as well.

e) This linear model is designed to predict final exam scores based upon midterm exam scores. It does not predict midterm scores from final exam scores. In order to predict in this direction, a linear model would have to be generated with final exam score as the explanatory variable and midterm exam score as the response variable. (Notice that part d is NOT predicting midterm from final, but rather asking what *actual* midterm score is required to result in a *prediction* of 100 for the final exam score.)

R2.12. (continued)

f) From part (d), a student with a midterm score of 100 is predicted have a final exam score of 100. The student's residual is 15 – 100 = –85.

g) The R^2 value of the regression will increase. This student's large negative residual would detract from the overall pattern of the data, allowing the model to explain less of the variability in final exam score. Removing it would increase the strength of the association.

h) The slope of the linear model would increase. This student's large negative residual would "pull" the regression line down, perhaps even making the association appear negative. The removal of this point would allow the line to snap back up to the true positive association.

R2.13. **Traffic.**

a) $b_1 = r\frac{s_y}{s_x} \Rightarrow -0.352 = r\frac{9.68}{27.07} \Rightarrow r = -0.984$; The correlation between traffic density and speed is $r = -$ 0.984

b) $R^2 = (-0.984)^2 = 0.969$; The variation in the traffic density accounts for 96.9% of the variation in speed.

c) $\widehat{Speed} = 50.55 - 0.352\,Density = 50.55 - 0.352(50) = 32.95$; According to the linear model, when traffic density is 50 cars per mile, the average speed of traffic on a moderately large city thoroughfare is expected to be 32.95 miles per hour.

d) $\widehat{Speed} = 50.55 - 0.352\,Density = 50.55 - 0.352(56) = 30.84$; According to the linear model, when traffic density is 56 cars per mile, the average speed of traffic on a moderately large city thoroughfare is expected to be 30.84 miles per hour. If traffic is actually moving at 32.5 mph, the residual is 32.5—30.84 = 1.66 miles per hour.

e) $\widehat{Speed} = 50.55 - 0.352\,Density = 50.55 - 0.352(125) = 6.55$; According to the linear model, when traffic density is 125 cars per mile, the average speed of traffic on a moderately large city thoroughfare is expected to be 6.55 miles per hour. The point with traffic density 125 cars per minute and average speed 55 miles per hour is considerably higher than the model would predict. If this point were included in the analysis, the slope would increase.

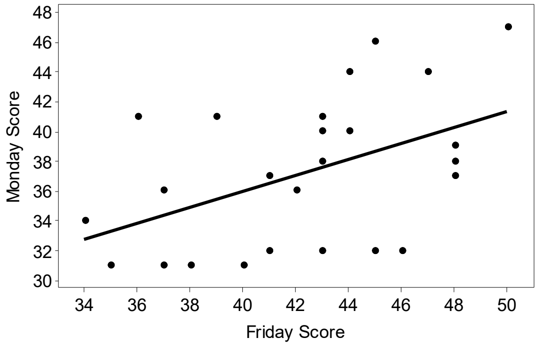
f) The correlation between traffic density and average speed would become weaker. The influential point (125, 55) is a departure from the pattern established by the other data points.

g) The correlation would not change if kilometers were used instead of miles in the calculations. Correlation is a "unitless" measure of the degree of linear association based on *z*-scores, and is not affected by changes in scale. The correlation would remain the same, $r = -0.984$.

R2.14. Cramming.

a) The correlation between the Friday scores and the Monday scores on the Spanish Test is $r = 0.473$.

b) The scatterplot shows a weak, linear, positive association between Friday score and Monday score. Generally, students who scored high on Friday also tended to score high on Monday.



c) A student with a positive residual scored higher on Monday's test than the model predicted.

d) A student with a Friday score that is one standard deviation below average is expected to have a Monday score that is 0.473 standard deviations below Monday's average score. The distribution of scores for Monday had mean 37.24 points and standard deviation 5.02 points, so the student's score is predicted to be approximately 37.24 – (0.473)(5.02) = 34.87.

e) The regression equation for the linear model that predicts Monday score from Friday score is: $\widehat{Monday} = 14.59 + 0.536\,Friday$.

f) $\widehat{Monday} = 14.5921 + 0.535666\,Friday = 14.5921 + 0.535666(40) \approx 36.0$; According to the model, a student with a Friday score of 40 is expected to have a Monday score of about 36.0.

R2.15. Cars, correlations.

a) Weight, with a correlation of –0.903, seems to be most strongly associated with fuel economy, since the correlation has the largest magnitude (distance from zero). However, without looking at a scatterplot, we can't be sure that the relationship is linear. Correlation might not be an appropriate measure of the strength of the association if the association is non-linear.

b) The negative correlation between weight and fuel economy indicates that, generally, cars with higher weights tend to have lower mileages than cars with lower weights. Once again, this is only correct if the association between weight and fuel economy is linear.

c) $R^2 = (-0.903)^2 = 0.815$; The variation in weight accounts for 81.5% of the variation in mileage. Once again, this is only correct if the association between weight and fuel economy is linear.

R2.16. Cars, associations.

a) *Displacement* and *weight* show the strongest association, with a correlation of 0.951. Generally, cars with larger engines are heavier than cars with smaller engines. However, without looking at a scatterplot, we can't be sure that the relationship is linear. Correlation might not be an appropriate measure of the strength of the association if the association is non-linear.

b) The strong correlation between displacement and weight is not necessarily a sign of a cause-and-effect relationship. Price of the car might be confounded with weight and displacement. More expensive luxury cars may have extra features that result in higher weights. One of these features might be a larger engine. Another difficulty with assigning a cause and an effect is that we can't be sure of the direction of the relationship. Certainly, the larger engine adds to the weight of the car, but maybe the larger engine is needed to power heavier cars.

R2.16. (continued)

c) The correlation would not change if cubic centimeters or liters were used instead of cubic inches in the calculations. Correlation is a "unitless" measure of the degree of linear association based on *z*-scores, and is not affected by changes in scale.

d) As long as the association between fuel economy and engine displacement was linear, a car whose engine displacement is one standard deviation above the mean would be predicted to have a fuel economy that is 0.786 standard deviations below the mean. (The correlation between the variables is –0.786, so a change in direction is indicated.) If the relationship were non-linear, the relative fuel economy could not be determined.

R2.17. Cars, horsepower.

a) The linear model that predicts the horsepower of an engine from the weight of the car is $\widehat{Horsepower} = 3.49834 + 34.3144\,Weight$.

b) The weight is measured in thousands of pounds. The slope of the model predicts an increase of about 34.3 horsepower for each additional unit of weight. 34.3 horsepower for each additional thousand pounds makes more sense than 34.3 horsepower for each additional pound.

c) Since the residuals plot shows no pattern, the linear model is appropriate for predicting horsepower from weight.

d) $\widehat{Horsepower} = 3.49843 + 34.3144\,Weight = 3.49843 + 34.3144(2.595) \approx 92.544$; According to the model, a car weighing 2595 pounds is expected to have 92.543 horsepower. The actual horsepower of the car is: $92.544 + 22.5 \approx 115.0$ horsepower.

R2.18. Cars, fuel efficiency.

a) After accounting for the linear effect of *Drive Ratio*, the model predicts a decrease in *MPG* of an average of 10.8 miles per gallon for each additional 1000 pounds of vehicle weight.

b) Since $R^2 = 89.5\%$, the model should be successful at predicting the fuel efficiency of a car. However, we would want to see the residuals plot to make sure that the model is appropriate.

c) The relationship is determined by the physics of car and engine design, so it should be roughly the same for modern cars. Exceptions for modern cars would include electric and hybrid vehicles.

R2.19. Cars, more efficient?

a) The residual plot is curved, suggesting that the conditions for inference are not satisfied. Specifically, the Straight Enough condition is not satisfied.

b) The reciprocal model is preferred to the model in Exercise 18. The new residual plot has no pattern, and appears to satisfy the conditions for inference. There is, however, one outlier to consider.

R2.20. Colorblind.

Sex and *Colorblindness* are both categorical variables. Correlation is a measure of the strength of a linear relationship between quantitative variables. The proper terminology is to say gender is associated with colorblindness.

R2.21. Old Faithful again.

a) The association between the duration of eruption and the interval between eruptions of Old Faithful is fairly strong, linear, and positive. Long eruptions are generally associated with long intervals between eruptions. There are also two distinct clusters of data, one with many short eruptions followed by short intervals, the other with many long eruptions followed by long intervals, with only a few medium eruptions and intervals in between.

b) The linear model used to predict the interval between eruptions is: $\widehat{Interval} = 33.9668 + 10.3582\,Duration$.

R2.21. (continued)

c) As the duration of the previous eruption increases by one minute, the model predicts an increase of about 10.4 minutes in the interval between eruptions.

d) $R^2 = 77.0\%$, so the model accounts for 77% of the variability in the interval between eruptions. The predictions should be fairly accurate, but not precise. Also, the association appears linear, but we should look at the residuals plot to be sure that the model is appropriate before placing too much faith in any prediction.

e) $\widehat{Interval} = 33.9668 + 10.3582\,Duration = 33.9668 + 10.3582(4) \approx 75.4$; According to the model, if an eruption lasts 4 minutes, the next eruption is expected to occur in approximately 75.4 minutes.

f) The actual eruption at 79 minutes is 3.6 minutes later than predicted by the model. The residual is 79 – 75.4 = 3.6 minutes. In other words, the model under-predicted the interval.

R2.22. Crocodile lengths.

a) The associations between the head sizes and body sizes for the crocodiles appear to be strong. 97.2% of the variability in Indian Crocodile length and 98.1% of the variability in Australian Crocodile length is accounted for by the variability in head size. (This assertion is only valid if the association between head and body length is linear for each crocodile.)

b) The slopes of the two models are similar. Indian Crocodiles are predicted to increase in length 7.4 centimeters for each centimeter increase in head length, and Australian Crocodiles are predicted to increase in length 7.72 centimeters for each centimeter increase in head length. (These predictions are only valid if the association between head and body length is linear for each crocodile.) The values of R^2 are also similar.

c) The two models have different values for the *y*-intercept. According to the models, the Indian Crocodile is smaller.

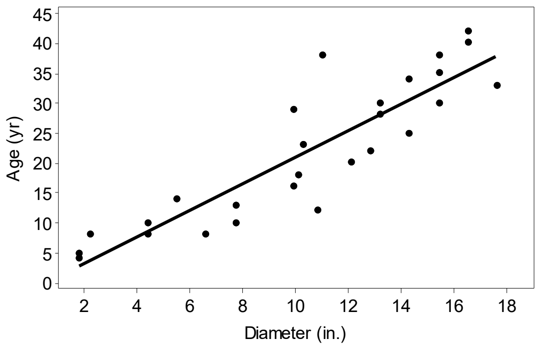
d) Indian Crocodile Model: $\widehat{IBody} = -69.3693 + 7.40004\,IHead = -69.3693 + 7.40004(62) = 389.43318$

Australian Crocodile Model: $\widehat{ABody} = -21.3429 + 7.82761\,AHead = -21.3429 + 7.82761(62) = 463.96892$

The appropriate models predict body lengths of 389.4 centimeters and 464.0 centimeters for Indian Crocodiles and Australian Crocodiles, respectively. The actual length of 380 centimeters indicates that this is probably an Indian Crocodile. The prediction is closer to the actual length for that model.

R2.23. How old is that tree?

a) The correlation between tree diameter and tree age is $r = 0.888$. Although the correlation is moderately high, this does not suggest that the linear model is appropriate. We must look at a scatterplot in order to verify that the relationship is straight enough to try the linear model. After finding the linear model, the residuals plot must be checked. If the residuals plot shows no pattern, the linear model can be deemed appropriate.

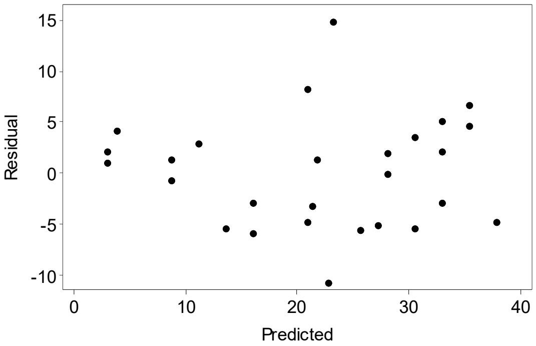


R2.23. (continued)

b) The association between diameter and age of these trees is fairly strong, somewhat linear, and positive. Trees with larger diameters are generally older.

c) The linear model that predicts age from diameter of trees is: $\widehat{Age} = -0.974424 + 2.20552\,Diameter$. This model explains 78.9% of the variability in age of the trees.

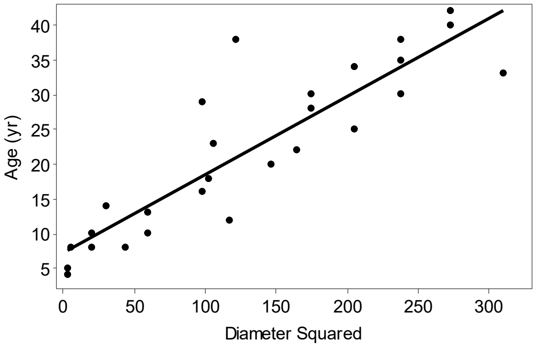
d) The residuals plot shows a curved pattern, so the linear model is not appropriate. Additionally, there are several trees with large residuals.



e) The largest trees are generally above the regression line, indicating a positive residual. The model is likely to underestimate these values.

R2.24. Improving trees.

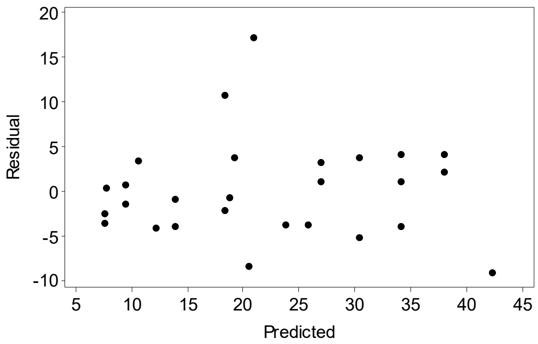
a) The re-expressed data, diameter squared versus age, is straighter than the original. This model appears to fit much better.



b) The linear model that predicts age from diameter squared is: $\widehat{Age} = 7.23961 + 0.113011\,Diameter^2$. This model explains 78.7% of the variability in the age of the trees.

R2.24. (continued)

c) The residuals plot shows random scatter. This model appears to be appropriate, although there are still some trees with large residuals.



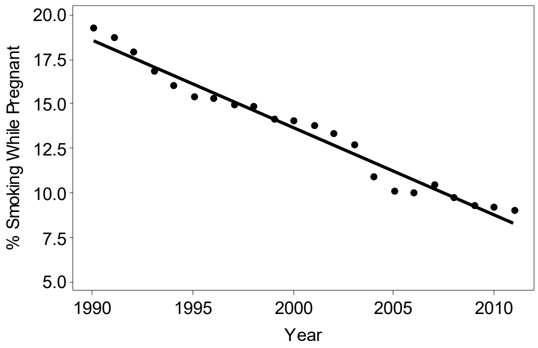
d) $\widehat{Age} = 7.23961 + 0.113011\,Diameter^2 = 7.23961 + 0.113011\left(18^2\right) \approx 43.855;$ According to the model, a tree with a diameter of 18" is expected to be approximately 43.9 years old.

R2.25. Big screen.

TV screen sizes might vary from 19 to 70 inches. A TV with a screen that was 10 inches larger would be predicted to cost 10(0.03) = +0.3, 10(0.3) = +3, 10(3) = +30, or 10(30) = +300. Notice that the TV costs are measure in hundreds of dollars, so the potential price changes for getting a TV 10 inches larger are \$30, \$300, \$3000, and \$30,000. Only \$300 is reasonable, so the slope must be 0.3.

R2.26. Smoking and pregnancy 2011.

a) The association between year and the percent of expectant mothers who smoked cigarettes during their pregnancies is strong, somewhat cyclical, but overall roughly linear, and negative. The percentage has decreased steadily since 1990.

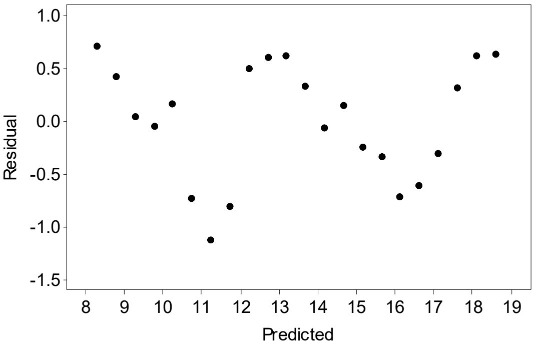


b) The correlation between year and percent of expectant mothers who smoked cigarettes during their pregnancies is $r = -0.985$. This may not be an appropriate measure of strength, since the scatterplot shows a slight bend.

c) The use of averages instead of individual percentages for each of the 50 states results in a correlation that is artificially strong. The correlation of the averaged data is "more negative" than the correlation of the individual percentages would have been.

R2.26. (continued)

d) The linear model that predicts the percent of expectant mothers who smoked during their pregnancies from the year is: $\widehat{Percentage} = 993.35 - 0.48984\,Year$. This model accounts for 97.1% of the variability in the percent. According to this model, for each year that passes, the average percent of women who smoked while pregnant decreases by 0.4898%. This model does not appear to be appropriate, since the residuals plot shows a pattern. However, it is unlikely that we can provide a better model, since the scatterplot shows a bend that cannot be straightened by re-expression.



R2.27. No smoking?

a) The model from Exercise 26 is for predicting the percent of expectant mothers who smoked during their pregnancies from the year, not the year from the percent.

b) The model that predicts the year from the percent of expectant mothers who smoked during pregnancy is: $\widehat{Year} = 2027.13 - 1.98233(\%)$. This model predicts that 0% of mothers will smoke during pregnancy in $2027.13 - 1.98233(0) \approx 2027$.

c) The lowest data point corresponds to 9.0% of expectant mothers smoking during pregnancy in 2011. The prediction for 0% is an extrapolation outside the scope of the data. There is no reason to believe that the model will be accurate at that point.

R2.28. Tips.

a) Without the data we can’t tell whether the relationship between Quality of service and Tip size is straight enough for the correlation to be appropriate.

b) R^2 = 1.21%. Only about 1% of the variability in tip size is explained by variability in the rating of service.

R2.29. U.S. Cities.

There is a strong, roughly linear, negative association between mean January temperature and latitude. U.S. cities with higher latitudes generally have lower mean January temperatures. There are two outliers, cities with higher mean January temperatures than the pattern would suggest.

R2.30. Correlations.

a) *Latitude* is the better predictor of average January temperature when the relationship between latitude and temperature is linear. The correlation, –0.848, is stronger than the correlation between altitude and temperature, –0.369.

b) The correlation would be the same, –0.848. Correlation is a measure of the degree of linear association between two quantitative variables and is unaffected by changes in units.

R2.30. (continued)

c) The correlation would be the same, –0.369. Correlation is a measure of the degree of linear association between two quantitative variables and is unaffected by changes in units.

d) (–0.369)(2) = –0.738. If a city has an altitude 2 standard deviations above the mean, its average January temperature is expected to be 0.738 standard deviations below the mean average January temperature.

R2.31. Winter in the city.

a) $R^2 = (-0.848)^2 \approx 0.719$. The variation in latitude explains 71.9% of the variability in average January temperature.

b) The negative correlation indicates that the as latitude increases, the average January temperature generally decreases.

c)

$$b_1 = r\frac{s_y}{s_x}$$
$$b_1 = (-0.848)\frac{13.49}{5.42}$$
$$b_1 = -2.1106125$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1\overline{x}$$
$$26.44 = b_0 - 2.1106125(39.02)$$
$$b_0 = 108.79610$$

The equation of the linear model for predicting January temperature from latitude is $\widehat{JanTemp} = 108.796 - 2.111\,Latitude$

d) For each additional degree of latitude, the model predicts a decrease of approximately 2.1°F in average January temperature.

e) The model predicts that the mean January temperature will be approximately 108.8°F when the latitude is 0°. This is an extrapolation, and may not be meaningful.

f) $\widehat{JanTemp} = 108.796 - 2.111\,Latitude = 108.796 - 2.111(40) \approx 24.4$; According to the model, the mean January temperature in Denver is expected to be 24.4°F.

g) In this context, a positive residual means that the actual average temperature in the city was higher than the temperature predicted by the model. In other words, the model underestimated the average January temperature.

R2.32. Depression and the Internet.

First of all, no association between variables can imply a cause-and-effect relationship. There may be lurking variables that explain the increase in both Internet use and depression. Additionally, provided the association is linear, only 4.6% of the variability in depression level can be explained by variability in Internet use. This is a very weak linear association at best.

R2.33. Olympic Jumps 2016.

a) The association between Olympic long jump distances and high jump heights is strong, linear, and positive. Years with longer long jumps tended to have higher high jumps. There is one departure from the pattern. The year in which the Olympic gold medal long jump was the longest had a shorter gold medal high jump than we might have predicted.

b) There is an association between long jump and high jump performance, but it is likely that training and technique have improved over time and affected both jump performances.

R2.33. (continued)

c) The correlation would be the same, 0.910. Correlation is a measure of the degree of linear association between two quantitative variables and is unaffected by changes in units.

d) In a year when the high jumper jumped one standard deviation better than the average jump, the long jumper would be predicted to jump $r = 0.910$ standard deviations above the average long jump.

R2.34. Modeling jumps 2016.

a)

$$b_1 = r\frac{s_y}{s_x} \qquad b_1 = (0.910)\frac{0.195271}{0.507606} \qquad b_1 = 0.3500679858$$

$$\hat{y} = b_0 + b_1 x \qquad \overline{y} = b_0 + b_1\overline{x} \qquad 2.15667 = b_0 + 0.3500679858(8.06222) \qquad b_0 = -0.6656551165$$

The linear model that predicts high jump heights from long jump distances is $\widehat{High} = -0.666 + 0.350\,Long$.

(Using technology, instead of calculating by hand, we get a model of $\widehat{High} = -0.665 + 0.350\,Long$.)

b) For each additional meter jumped in the long jump, the model predicts an increase of approximately 0.350 meters in the high jump.

c) $\widehat{High} = -0.665 + 0.350\,Long = -0.665 + 0.350(8.9) \approx 2.45$; According to the model, the high jump height is expected to be approximately 2.45 meters in a year when the long jump distance is 8.9 meters.

d) This equation cannot be used to predict long jump distance from high jump height, because it was specifically designed to predict high jump height from long jump distance.

e)

$$b_1 = r\frac{s_y}{s_x} \qquad b_1 = (0.910)\frac{0.507606}{0.195271} \qquad b_1 = 2.365540505$$

$$\hat{y} = b_0 + b_1 x \qquad \overline{y} = b_0 + b_1\overline{x} \qquad 8.06222 = b_0 + 2.365540505(2.15667) \qquad b_0 = 2.960529759$$

The linear model that predicts long jump distances from high jump distances is $\widehat{Long} = 2.961 + 2.366\,High.$

(Using technology, instead of calculating by hand, we get a model of $\widehat{Long} = 2.961 + 2.365\,High$.)

R2.35. French.

a) Most of the students would have similar weights. Regardless of their individual French vocabularies, the correlation would be near 0.

b) There are two possibilities. If the school offers French at all grade levels, then the correlation would be positive and strong. Older students, who typically weigh more, would have higher scores on the test, since they would have learned more French vocabulary. If French is not offered, the correlation between weight and test score would be near 0. Regardless of weight, most students would have horrible scores.

c) The correlation would be near 0. Most of the students would have similar weights and vocabulary test scores. Weight would not be a predictor of score.

d) The correlation would be positive and strong. Older students, who typically weigh more, would have higher test scores, since they would have learned more French vocabulary.

R2.36. Twins.

a) There is a strong, fairly linear, positive trend in pre-term twin birth rates. As the year of birth increases, the pre-term twin birth rate increases.

b) The highest pre-term twin birth rate is for mothers receiving "adequate" prenatal care, and the lowest pre-term twin birth rate is for mothers receiving "inadequate" prenatal care. The slope is about the same for these relations. Mothers receiving "intensive" prenatal care had a pre-term twin birth rate that was higher than mothers receiving "inadequate" care and lower than mothers receiving "adequate" prenatal care. However, the rate of increase in pre-term twin birth rate is greater for mothers receiving "intensive" prenatal care than the rate of increase for the other groups.

c) Avoiding medical care would not be a good idea. There are likely lurking variables explaining the differences in pre-term twin birth rate. For instance, the level of pre-natal care may actually be determined by complications early in the pregnancy that may result in a pre-term birth.

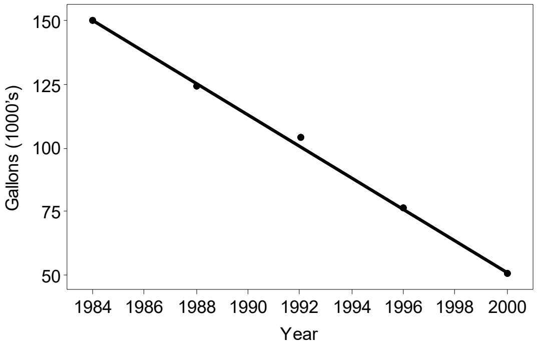
R2.37. Lunchtime.

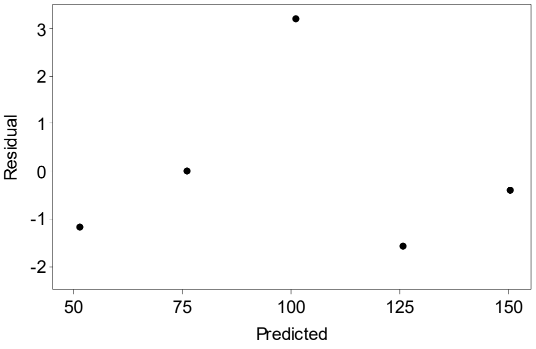
The association between time spent at the table and number of calories consumed by toddlers is moderate, roughly linear, and negative. Generally, toddlers who spent a longer time at the table consumed fewer calories than toddlers who left the table quickly. The scatterplot between time at the table and calories consumed is straight enough to justify the use of the linear model. The linear model that predicts the time number of calories consumed by a toddler from the time spent at the table is $\widehat{Calories} = 560.7 - 3.08\,Time$. For each additional minute spent at the table, the model predicts that the number of calories consumed will be approximately 3.08 fewer. Only 42.1% of the variability in the number of calories consumed can be accounted for by the variability in time spent at the table. The residuals plot shows no pattern, so the linear model is appropriate, if not terribly useful for prediction.

R2.38. Gasoline.

a) The association between the year and the number of gallons of leaded gasoline available is linear, very strong, and negative. As the years have passed, the number of gallons of leaded gasoline has decreased steadily. The linear model that predicts the number of gallons available based upon the year is $\widehat{ThousandGallons} = 12{,}451.2 - 6.2\,Year$.

The residuals plot shows no pattern, so the linear model is appropriate. R^2 = 99.8%, so the variability in year accounts for 99.8% of the variability in the number of gallons available. According to the model, there will be approximately –41,800 gallons available in 2015. This is an extrapolation, and isn't meaningful.





b) The model is designed to predict the number of gallons available based on the year. The question asks the students to predict the year based on the number of gallons available, and models only predict in one direction.

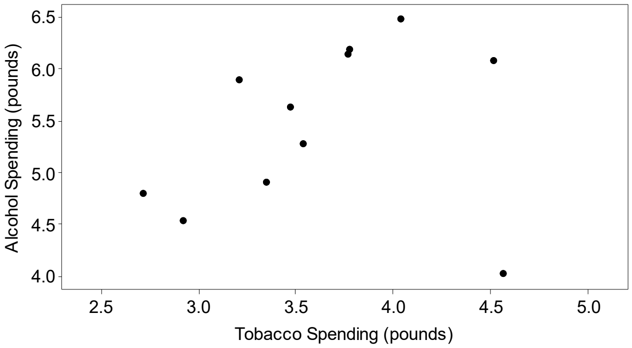
R2.38. (continued)

c) The linear regression model that predicts the year from the number of gallons available is: $\widehat{Year} = 2008.22 - 0.161\,ThousandGallons$ The model predicts that 0 gallons of leaded gasoline will be available in about 2008.

d) The association between year and the number of gallons of leaded gasoline available is very strong. In fact, it is so strong that the models actually do a decent job of predicting in the wrong direction! The model designed to minimize the sum of squared residuals in the response direction is actually pretty good at minimizing the sum of the squared residuals in the explanatory direction.

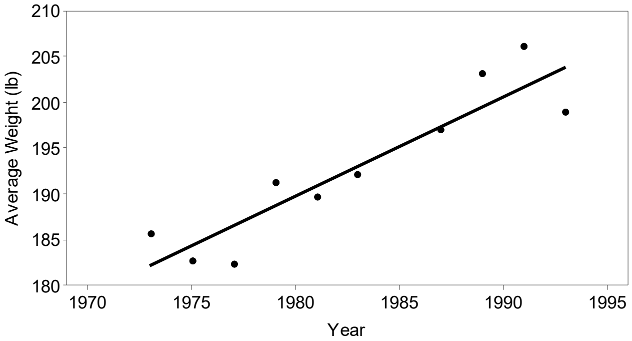
R2.39. Tobacco and alcohol.

The first concern about these data is that they consist of averages for regions in Great Britain, not individual households. Any conclusions reached can only be about the regions, not the individual households living there. The second concern is the data point for Northern Ireland. This point has high leverage, since it has the highest household tobacco spending and the lowest household alcohol spending. With this point included, there appears to be only a weak positive association between tobacco and alcohol spending. Without the point, the association is much stronger. In Great Britain, with the exception of Northern Ireland, higher levels of household spending on tobacco are associated with higher levels of household spending on tobacco. It is not necessary to make the linear model, since we have the household averages for the regions in Great Britain, and the model wouldn't be useful for predicting in other countries or for individual households in Great Britain.



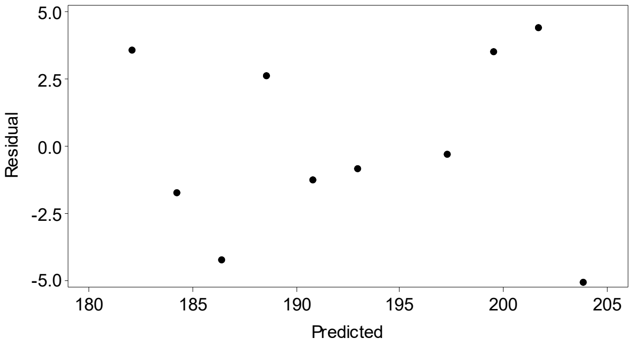
R2.40. Williams football.

a) The association between year and average weight of the members of the Williams College football team is strong, linear, and positive. Over the years, there has been a general increase in the average weight of the team. The linear model that predicts average weight from year is $\widehat{AvgWeight} = -1971.26 + 1.09137\,Year$. According to the model, average weight has increased by approximately one pound per year since 1973. The model accounts for 83.8% of the variability in average weight.



R2.40. (continued)

b) The residuals plot shows no pattern, so the linear model is appropriate. However, there are a couple of things to remember. First, the model is for predicting the average weight of the players on the team. Individual weights would be much more variable. Second, since we are dealing with weights, it is not reasonable to use the model for extrapolations. There is no reason to believe that average weights before 1973 or after 1993 would follow the model.



c) $\widehat{AvgWeight} = -1971.26 + 1.09137\,Year = -1971.26 + 1.09137(2015) \approx 227.85$; The model predicts that the average weight of the Williams College football team will be approximately 227.85 pounds in 2015. This prediction might be pretty close, but we shouldn't place to much faith in it. 2015 is 22 years later than the last year for which we have data.

d) The model predicts that the average weight of the Williams College football team will be approximately 323.88 pounds in 2103. This is not reasonable. The prediction is based upon an extrapolation of 110 years.

e) The model predicts that the average weight of the Williams College football team will be approximately 1306.11 pounds in 3003. This is absurd. The prediction is based upon an extrapolation of 1010 years.

R2.41. Models.

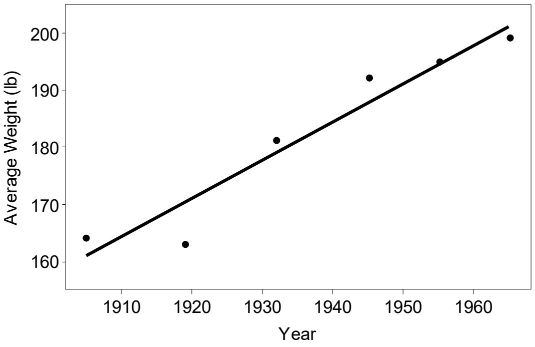
a) $\hat{y} = 2 + 0.8\ln x = 2 + 0.8\ln(10) \approx 3.842$

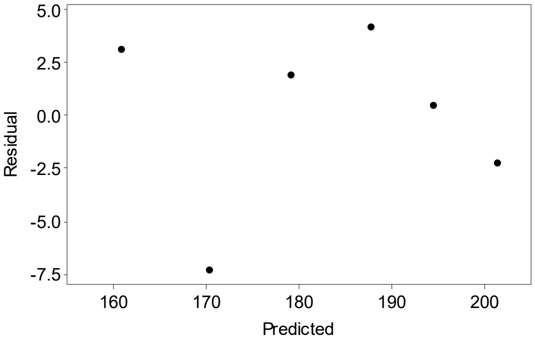
b) $\log \hat{y} = 5 - 0.23x = 5 - 0.23(10) = 2.7$, so $\hat{y} = 10^{2.7} \approx 501.187$

c) $\dfrac{1}{\sqrt{\hat{y}}} = 17.1 - 1.66x = 17.1 - 1.66(10) = 0.5$, so $\hat{y} = \dfrac{1}{0.5^2} = 4$

R2.42. Williams vs. Texas.

a) The association between year and average weight for the University of Texas football team is strong, roughly linear, and positive. The average weight has generally gone up over time. The linear model is: $\widehat{Ave.Weight} = -1121.66 + 0.67326\,Year$. This model explains 92.8% of the variability in average weight, but the residuals plot shows a possible pattern. The linear model may not be appropriate.





b) $-1971.26 + 1.091\,Year = -1121.66 + 0.673(Year) \Rightarrow 0.4181\,Year = 849.60 \Rightarrow Year \approx 2032.5$; According to these models, the predicted weights will be the same sometime during the year 2033. The average weight of the Williams College team is predicted to be more than the average weight of the University of Texas team any time after 2033.

c) This information is not likely to be accurate. The year 2033 is an extrapolation for both of the models, each of which has been shown to be of little use for even small extrapolations.

R2.43. Vehicle weights.

a) $\widehat{Wt} = 10.85 + 0.64\,scale = 10.85 + 0.64(31.2) = 30.818$; According to the model, a truck with a scale weight of 31,200 pounds is expected to weigh 30,818 pounds.

b) If the actual weight of the truck is 32,120 pounds, the residual is 32,120 – 30,818 = 1302 pounds. The model underestimated the weight.

c) $\widehat{Wt} = 10.85 + 0.64\,scale = 10.85 + 0.64(35.590) = 33.6276$; The predicted weight of the truck is 33,627.6 pounds. If the residual is –2440 pounds, the actual weight of the truck is 33,627.6 – 2440 = 31,187.6 pounds.

d) $R^2 = 93\%$, so the model explains 93% of the variability in weight, but some of the residuals are 1000 pounds or more. If we need to be more accurate than that, then this model will not work well.

e) Negative residuals will be more of a problem. Police would be issuing tickets to trucks whose weights had been overestimated by the model. The U.S. justice system is based upon the principle of innocence until guilt is proven. These truckers would be unfairly ticketed, and that is worse than allowing overweight trucks to pass. Also, trucking companies might be inclined to take the ticket to court.

R2.44. Companies.

a) The re-expressed data are more symmetric, with no outliers. That's good for regression because there is less of a chance for influential points. (Additionally, symmetric distributions of the explanatory and response variables will help ensure that the residuals around a line through the data are more unimodal and symmetric. We will learn more about why this is important in a later chapter.)

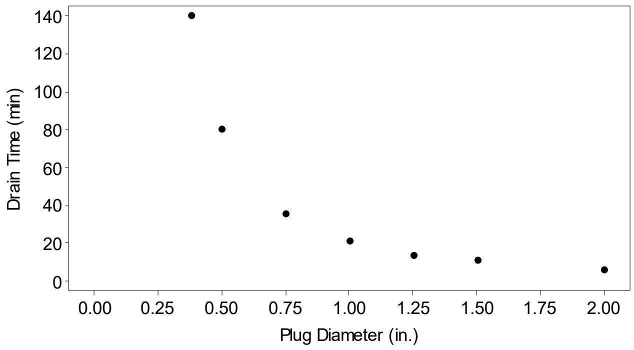
b) The association between log(sales) and log(profit) is linear, positive, and strong. The residuals plot shows no pattern. This model appears to be appropriate, and is surely better than the model generated with the original data.

c) The linear model is: $\widehat{\log(Profit)} = -0.106259 + 0.647798\ \log\left(Sales\right)$

d) $\widehat{\log(Profit)} = -0.106259 + 0.647798\ \log\left(Sales\right) = -0.106259 + 0.647798\ \log\left(2500\right) = 2.0949197,$ so $10^{2.0949197} \approx 124.43;$ According to the model, a company with sales of 2.5 billion dollars is expected to have profits of about 124.43 million dollars.

R2.45. Down the drain.

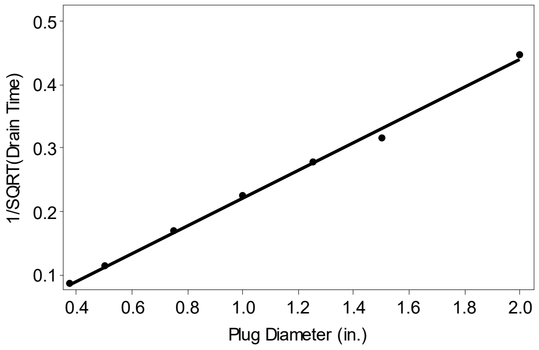
The association between diameter of the drain plug and drain time of this water tank is strong, curved, and negative. Tanks with larger drain plugs have lower drain times. The linear model is not appropriate for the curved association, so several re-expressions of the data were tried.

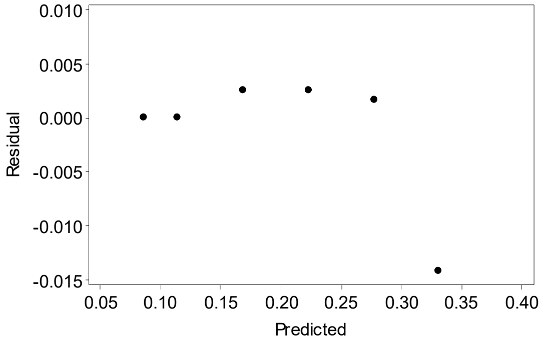


The best one was the reciprocal square root re-expression, resulting in the equation

$$\widehat{\frac{1}{\sqrt{DrainTime}}} = 0.00243 + 0.219\,Diameter\ .$$

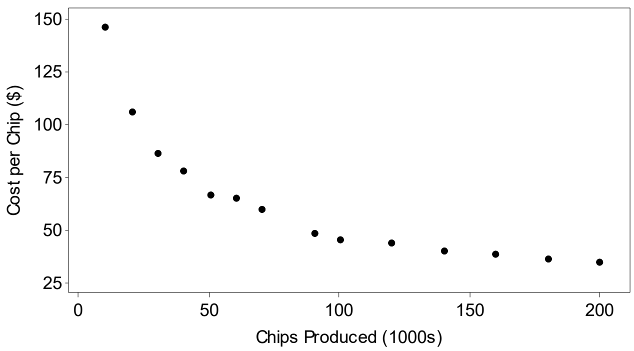
The re-expressed data is nearly linear, and although the residuals plot might still indicate some pattern and has one large residual, this is the best of the models examined. The model explains 99.7% of the variability in drain time.

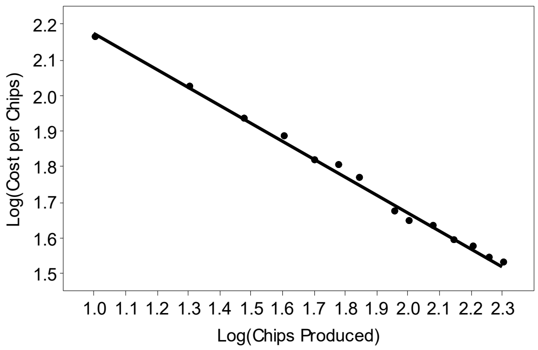


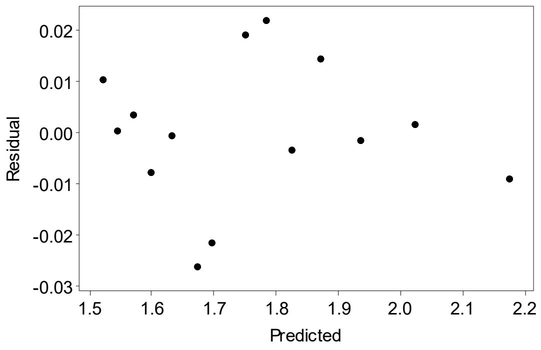


R2.46. Chips.

The association between the number of chips produced and the cost per chip is strong, negative, and curved. As the number of chips produced increases, the cost per chip decreases. The linear model is not appropriate for this curved association, so several re-expressions of the data were tried. Re-expressing each variable using logarithms results in a scatterplot that is nearly linear.







The model, $\widehat{\log\left(CostperChip\right)} = 2.67492 - 0.501621\ \log\left(ChipsProduced\right)$, has a residuals plot that shows no pattern and $R^2 = 99.5\%$. The model accounts for 99.5% of the variability in the cost per chip.

R2.47. Companies assets and sales.

a) According to the model, *LogAssets* increase on average by 0.868 Log$ per Log$ of sales for both banks and other companies.

b) Bank assets are, on average, 0.946 Log$ higher than other companies, after allowing for the linear effect of *LogSales.*

c) Yes, the use of the variable *Banks* is appropriate for these data. On the plot, the regression lines are roughly parallel.

R2.48. Real estate.

a) According to the model, houses in this market cost about $110/square foot of living area whether they are in a suburb or not.

b) Suburban houses cost, on average, $104,610 more even after allowing for their living area.

c) The slopes of the regression lines for suburban and other homes are nearly parallel, so an indicator variable is appropriate. However, R^2 is small, so we should take care in interpreting the coefficients.

R2.49. Real estate, bathrooms.

a) After accounting for the linear effect of *Living Area*, houses with another *Bathroom* on average, have a *Price* that is $75,020 higher.

b) The slope of the partial regression line is the *Bathrooms* coefficient, 75,020.3.

R2.50. Real estate, bedrooms.

a) Among houses with the same *Living Area*, each *Bedroom* lowers the price of a home by $43,346, on average. (Perhaps that living area would be better devoted to other uses.)

b) *Living Area* and *Bedrooms* are collinear. The regression coefficient is not about how the number of bedrooms affects the price in general, but only among houses of a given living area.

c) The slope of the partial regression line is the *Bedrooms* coefficient, -43,346.8.

d) That doesn't look like a good idea, since for a given *Living Area*, homes with another bedroom tend to have lower values. But keep in mind that the regression model doesn't predict what would happen if a change were made.

R2.51. Penguins again.

a) After accounting for the linear effect of *Depth*, the *log(Heart rate)* decreases by 0.045 bpm per minute *Duration* of the dive.

b) Yes, the coefficient of *Duration* is a good estimate of the relationship of *Duration* to (log) *Heart rate*. There is a strong, negative, linear relationship. The longer the *Duration*, the slower the *Heart rate*.

R2.52. Doctors and life expectancy and TVs.

a) $\sqrt{TV / person}$ and $\sqrt{Doctors / person}$ are highly correlated (because both are related to the general state of a country's economy). The regression coefficient describes the effect of $\sqrt{Doctors / person}$ among countries with roughly the same number of $\sqrt{TV / person}$.

b) The slope of the partial regression line is the $\sqrt{TV / person}$ coefficient, 22.9032.

c) Without North Korea, the slope of the partial regression line would be greater, so North Korea is making the $\sqrt{TV / person}$ coefficient smaller.

d) Without North Korea, the slope of the partial regression line would be greater, so its omission would make the $\sqrt{TV / person}$ coefficient larger.