Chapter 2 Looking at Data: Relationships

2.1. (a) The cases are students. (b)Number of friends and time (average amount spent on Facebook per week). **(c)** Both variables are quantitative because both are numeric and arithmetic operations are possible.

2.2. The variable as defined would be categorical because each student is categorized by the group into which he/she falls.

One advantage is that it might simplify the analysis, or at least it might simplify the process of describing the results. A subtler issue is that it is not clear whether finding an average is appropriate for the original numbers; technica lly, averages are not appropriate for a quantitative measurement unless the variable is measured on an "interval" scale for which the variable resources to cope is certainly not.

2.3. Cases: cups of Mocha Frappuccino. Variables: size and price (both quantitative).

2.4. Answers will vary. For example, people with high stress levels may have necessarily developed coping mechanism s, so we would expect the amount of stress to explain or cause changes in the resources to cope variable.

2.5. (a)Tweets. **(b)** Click count and length of tweet are quantitative. Day of week and sex are categorical. Time of day could be quantitative (in the hours/minutes format ashh:mm)or categorical (if morning, afternoon, etc.). **(c)** Click counts is the response because the other variables can possibly explain the number of click counts. The others could all be potentially explanatory because they can be controlled by the researcher or are already set.

2.6. (a) Neither. There is no reason to believe that a person who likes to sing also likes to dance or vice versa. **(b)** Number of pages is the explanatory variable, which should explain or cause changes to the cost of a new copy of the text book, which is the response. (**c)** Number of alcoholic drinks is the explanatory variable, which should explain or cause changes in the blood alcohol content, which is the response. (**d)** The dose of vitamin D is the explanatory variable, which should explain or cause changes to the change in total bone mineral content, which is the response.

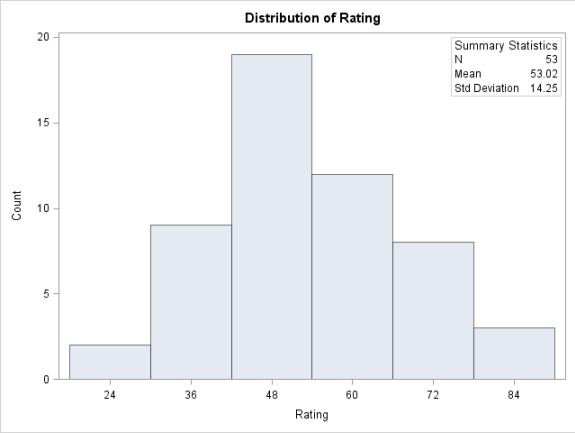
2.7. Answers will vary. Some possible variables are condition of the book (with values poor, good, or excellent), number of pages, and binding type (with values hardback or paperbound), in addition to purchase price and buy-back price. Cases would be individual textbooks. Here, we would likely be interested in the relationship between the buy-back price (response variable) based on the other explanatory variables such as the number of pages. We could also use the categorical variable, condition of the book, to group the data.

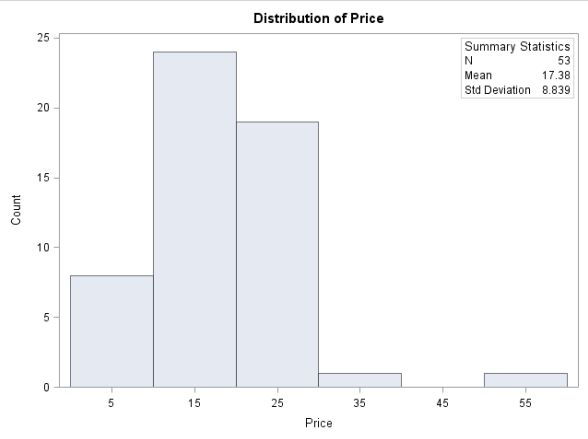
2.8. Answers will vary. Some possible variables are sex, age, etc., in addition to protein intake and fat intake. Cases would be first-year college students. Here, we would likely be interested in the relationship between protein and fat intake.

2.9. Answers will vary. Some possible variables are university, size, etc., in addition to the average number of tickets sold and the percentage of games won. Cases would be individual teams. Here, we would likely be interested if there is a relationship between the average number of tickets sold and the percentage of games won.

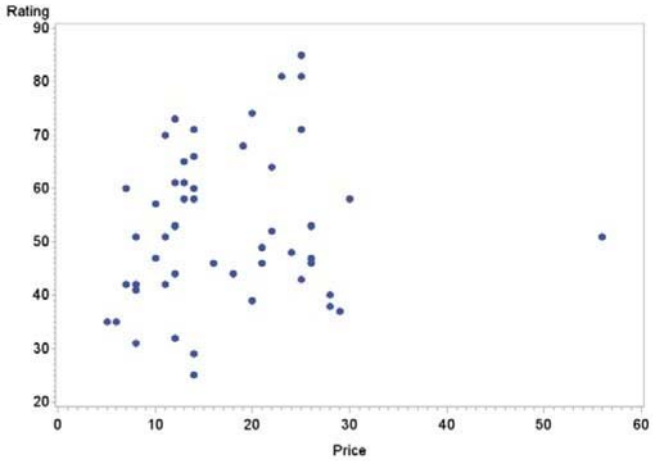
2.10. (a) The data set has 53 cases. (b)There are three variables: Price (with values 5-30), Rating (with values 31-81, and Type (with values liquid or powder). **(c)** Price and Rating are quantitative; Type is categorical. (d) Answers will vary.

2.11. Rating looks Normally distributed. Price also looks somewhat normal but has a high outlier.

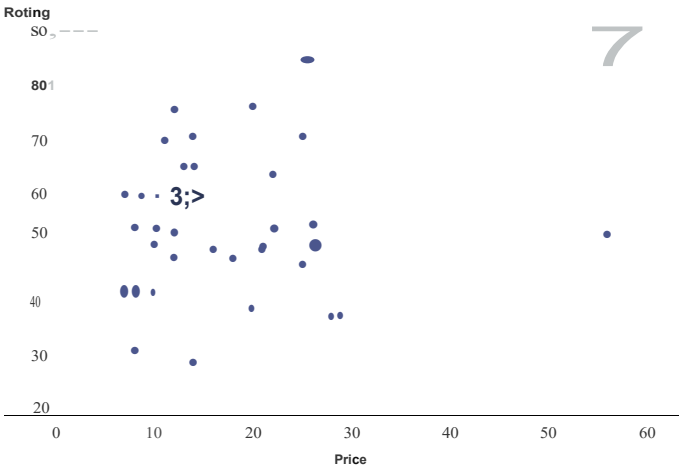




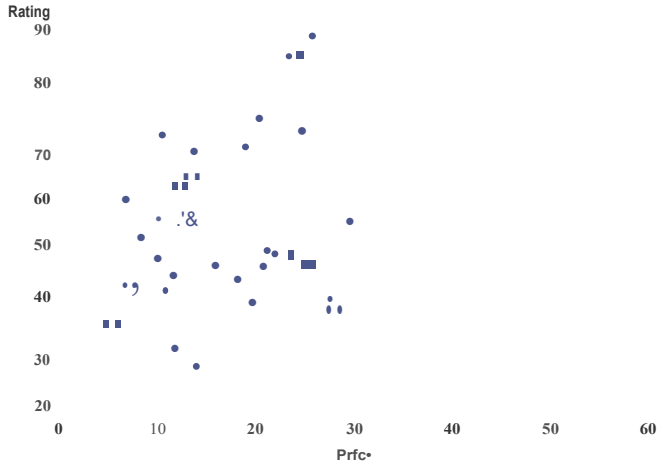
2.12.(a)



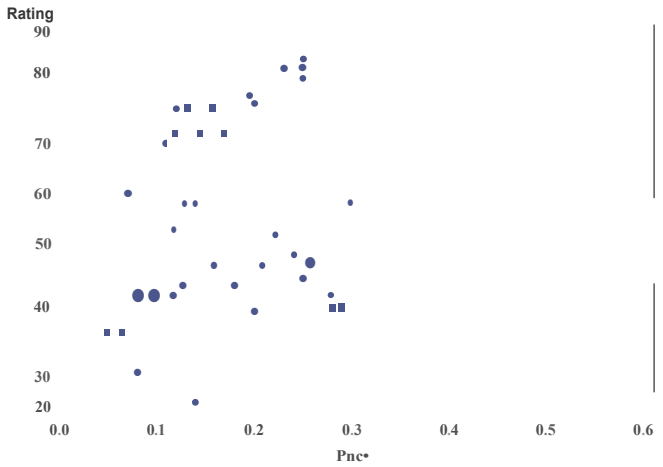
(b) The duplicate points are circled in the plot.



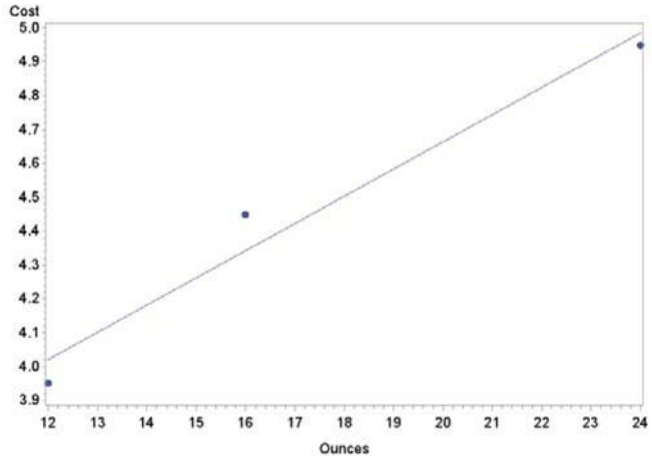
(c) Answers will vary. Interpretations of the plot could be misleading depending on how many duplicate points are hidden. (**d)**



2.13. (b) Shown below. **(c)** There is no difference in the plot except for the scale on the *x* axis.

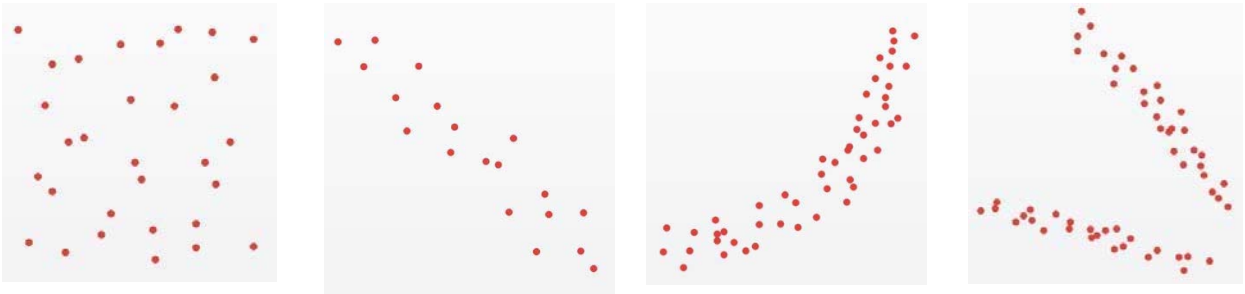


2.14. The size in ounces is the explanatory variable, which should explain or cause changes in the cost. The scatterplot shows that there is a relationship, as ounces increases so does the cost, but the cost increase between 12 and 16 ozis greater than it is between 16 and 24 oz.



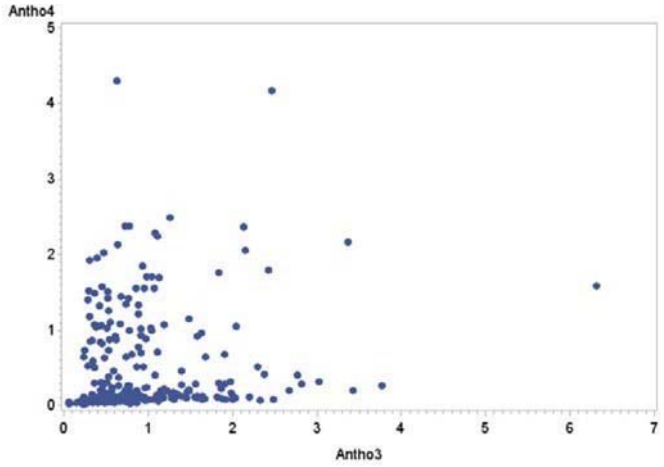
2.15. Answers may vary. Most will likely prefer the plot with the log transformation because it spreads out the data points and makes the plot easier to read and interpret.

2.16. Shown in order below.(a) No apparent relationship. **(b)** A strong negative linear relationship. **(c)** A strong positive relationship that is not linear. **(d)** A more complicated relationship. Answers will vary. Below is an example of two distinct populations with separate relationships plotted together.

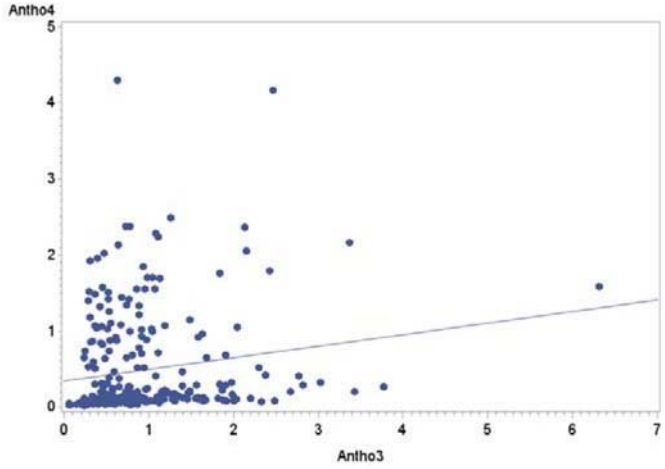


2.17.(a) A negative association means that low values of one variable are associated with *high* values of the other variable. **(b)** A stemplot is used for only a single quantitative variable. **(c)** We put the response variable on *they* axis and the explanatory variable on the *x* axis.

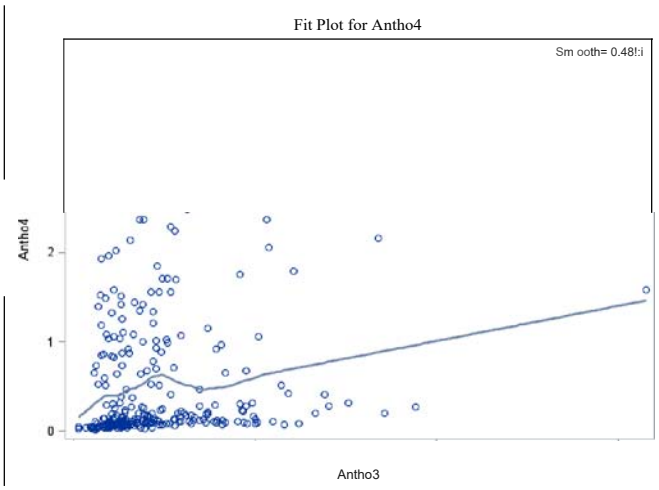
2.18. (a)



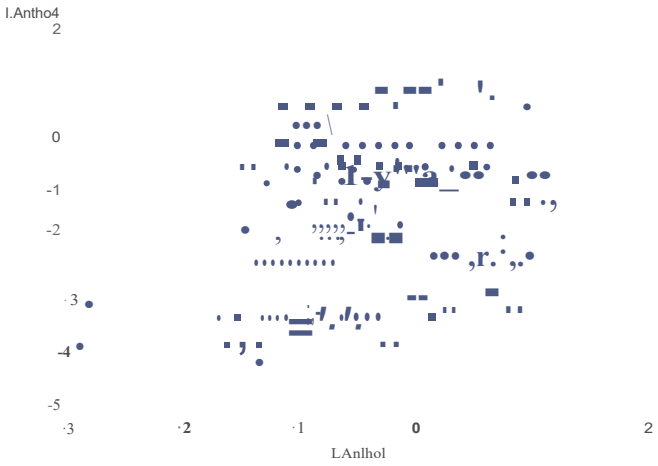
(b)The form is quite scattered; the direction is positive; the strength is super weak. (c) There are several outliers: one for Antho3 and two for Antho4. (d) Adding a line would not be useful because the relationship is not linear.



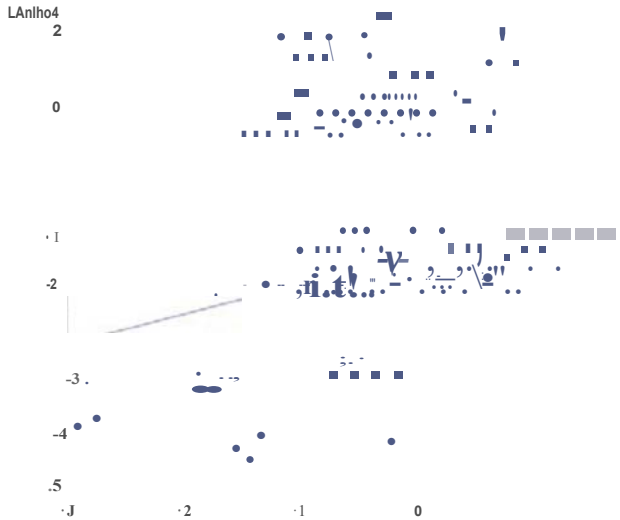
(e) Answers may vary depending on software used; there does not seem to be much of arelationship, either linear or smooth, in this scatterplot.



2.19. (a)

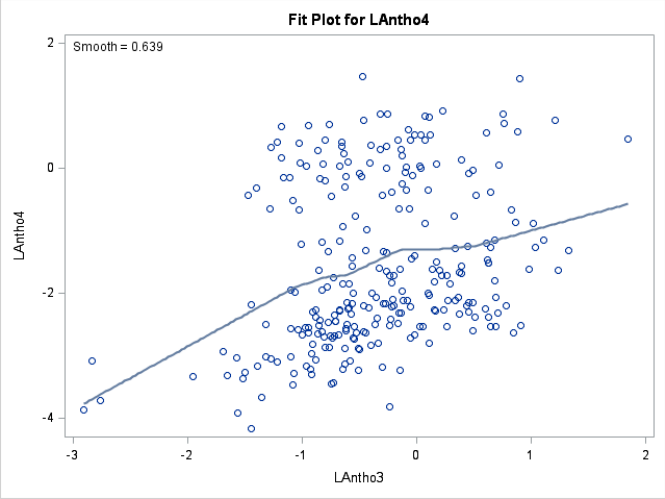


(b) The form is somewhat linear; the direction is positive; the strength is still weak. **(c)** There are a couple possible low outliers for Antho3. (**d)** Adding a line could be useful because the relationship is somewhat linear (shown below).



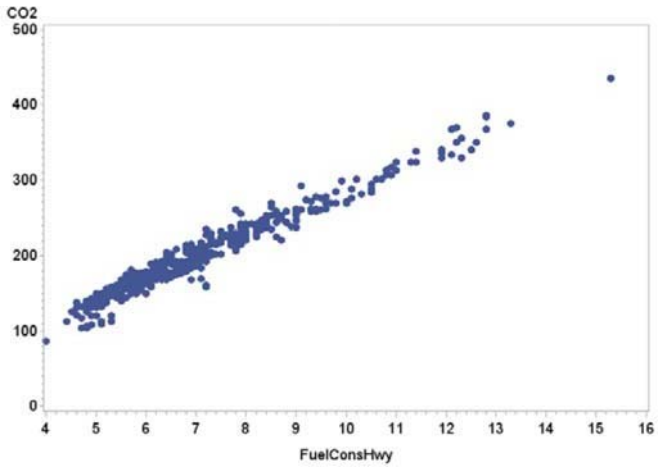
LAnlhoJ

(e) Answers will vary depending on software used; the smoothing does not contribute much in describing the relationship.

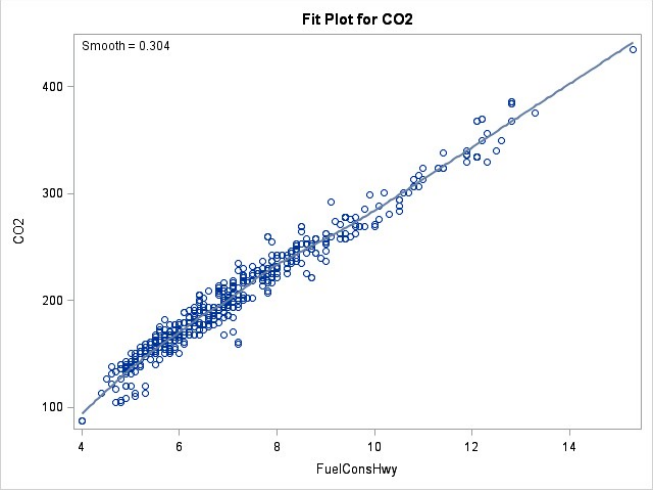


2.20. (a) The analysis done using the log transformed data gave a much better understanding of the relationship between Antho4 and Antho3 than the original raw data did. We could accurately see a relationship between the variables once the log transformation was done.(b) Answers may vary, but most should prefer the log transformed data.

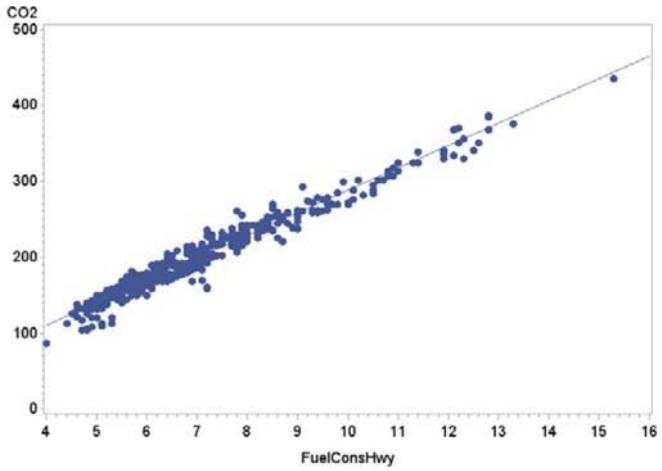
2.21. (a)



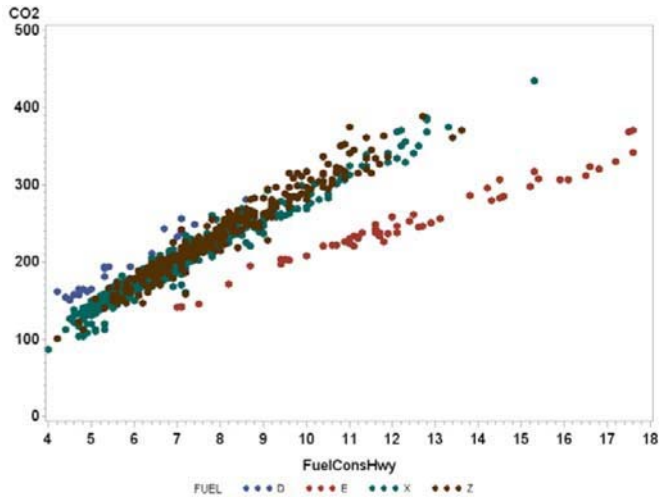
(b) The form is linear; the direction is positive; the strength is very strong. **(c)** There is one outlier with an unusually high value for both variables. **(d)** Yes, the line shows the direction and strength. (**e**)Answers will vary depending on software used; the smoothing does not help at all because the relationship is quite linear.



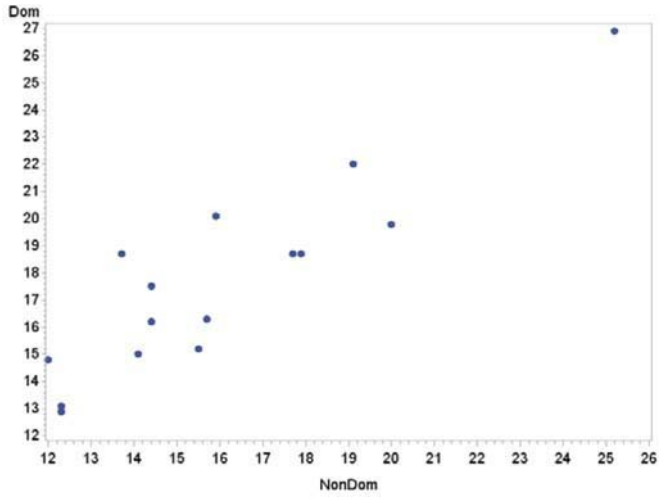
2.22. (a) Plot shown below. The line explains the relationship well. **(b)** Yes, the analysis supports using a straight-line relationship.



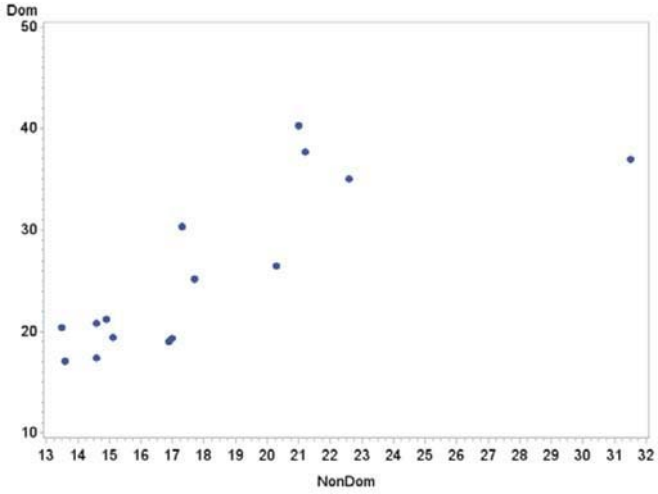
2.23. (a) Plot shown below. For all fuel types, as highway fuel consumption increases, so does carbon dioxide emissions.(b) Vehicles with fuel type D have the largest emissions, while vehicles with fuel type E have the smallest emissions. The other two types, X and Z, have fairly similar emissions.



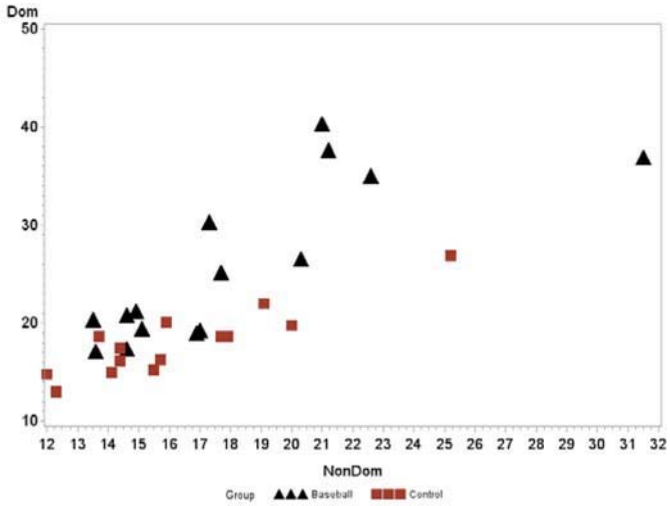
2.24. (a) Shown below. **(b)** As nondominant arm strength increases, so does dominant arm strength. There is one outlier with an extremely high nondominant arm strength. **(c)** The form is linear. The direction is positive. The strength is strong. (**d)** There is one outlier with an extremely high nondominant arm strength. (**e)** Yes, the relationship is linear.



2.25. (a)Shown below. **(b)** As nondominant arm strength increases, so does dominant arm strength. There is one outlier with an extremely high nondominant arm strength. **(c)** The form is linear. The direction is positive. The strength is strong. (**d)** There is one outlier with an extremely high nondominant arm strength. (**e)** Yes, the relationship is linear except for the outlier.

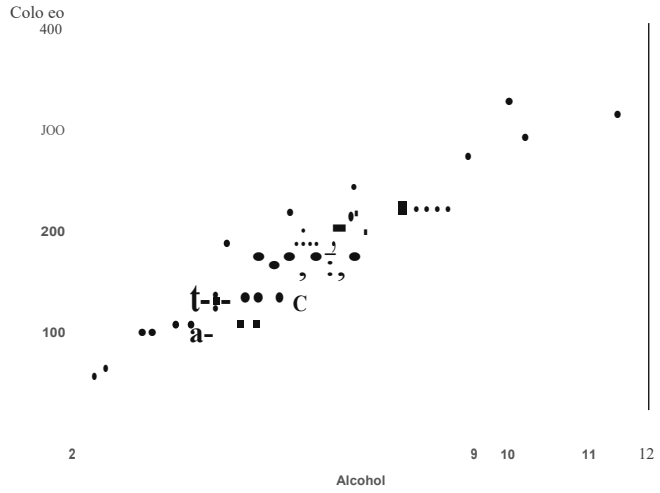


2.26. (a) Shown below. **(b)** Both groups have increasing linear relationships; the relationship for the baseball players seems weaker mostly due to the outlier on the far right. Baseball players also are generally stronger in both arms.

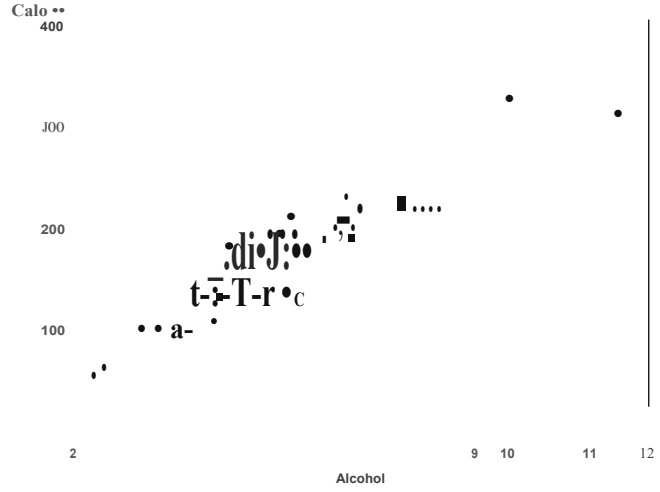


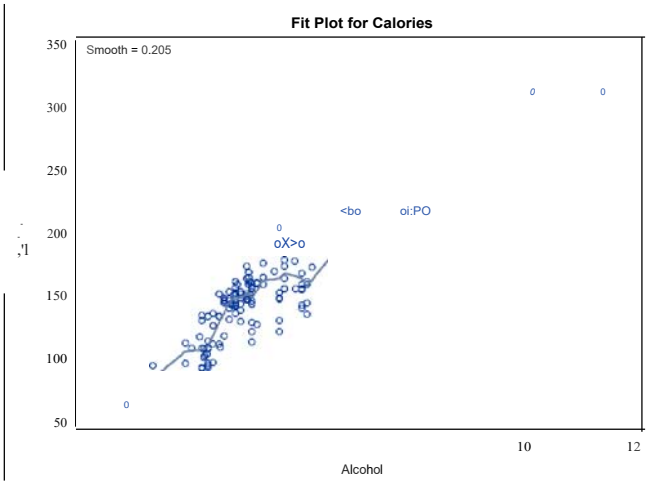
2.27. Parents' income is explanatory, and college debt is the response. Both variables are quantitative. We would expect a negative association: for parents with lower incomes, student college debt would be high and vice versa.

2.28. (a) Overall the relationship is linear and positive. **(b)** O'Doul's is the outlier; it has almost no alcohol. **(c)** Shown below.(d) The relationship is linear, positive, and quite strong.

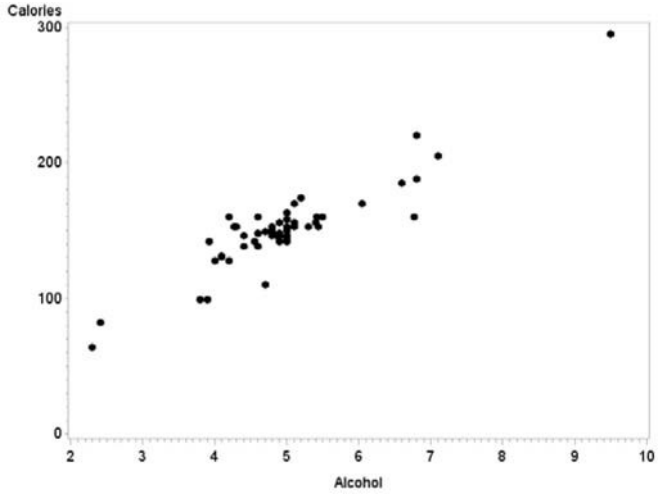


2.29. (a) Shown below. **(b)** Answers will vary depending on software used; the smoothing does not help, as the relationship is quite linear.

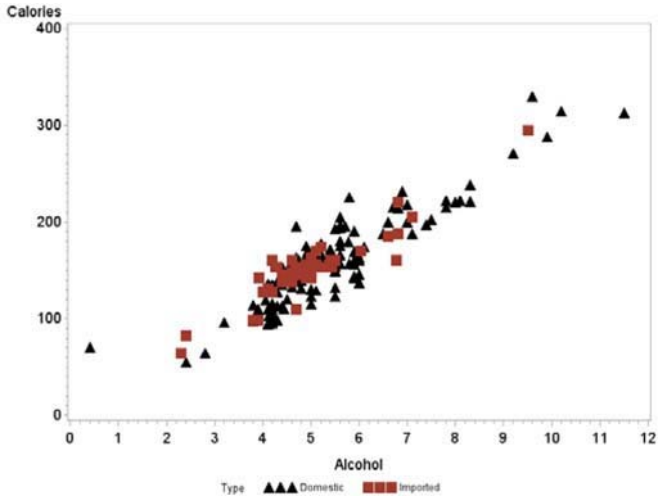




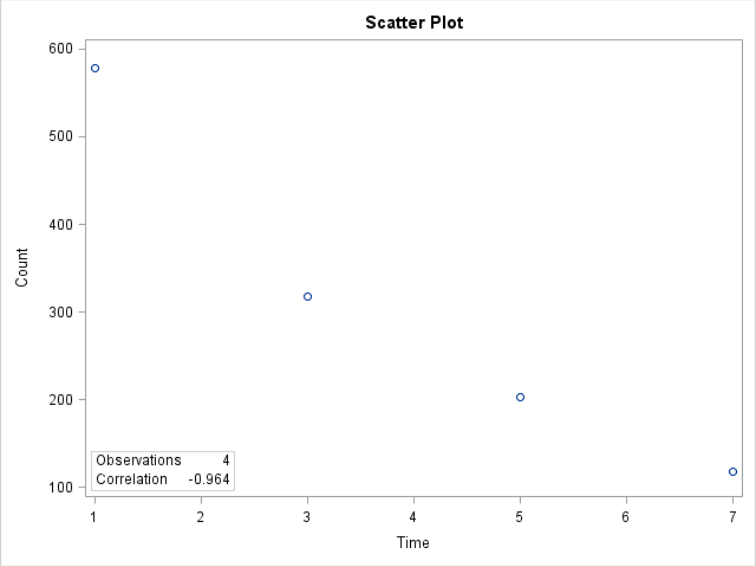
2.30. The scatterplot shows the relationship is linear, positive, and strong. McEwan's Scotch Ale is an outlier with an unusually large alcohol amount, but it does fit the pattern.



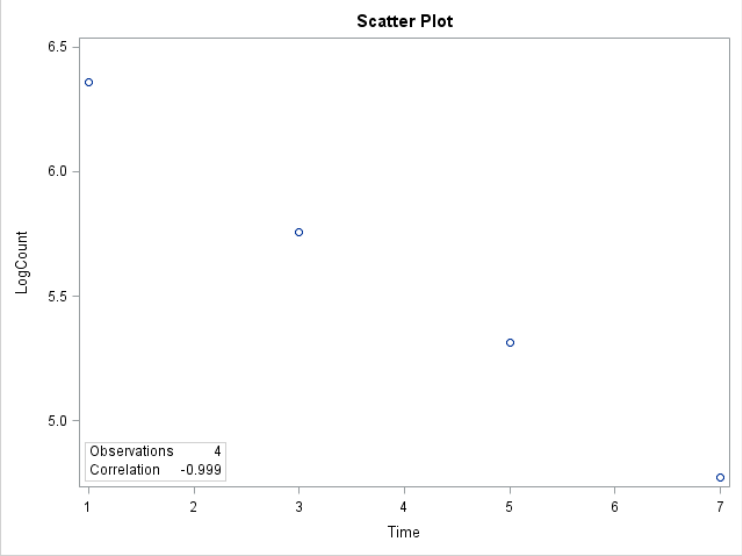
2.31. The relationships between calories and alcohol content are quite similar for domestic and imported beers. Also, the outlier for the imported beers no longer is an outlier, as there are several other domestic beers that have a similar alcohol content.



2.32. (a) Plot shown below, we would expect time to explain the count, so time should be on the *x* axis.(b) As time increases, the count goes down. **(c)** The form is curved; the direction is negative; the strength is very strong. (**d)** The first data point at time 1 is somewhat of an outlier because it does not line up as well as the other times do. (e)No, a curve might fit the data better than a simple linear trend.

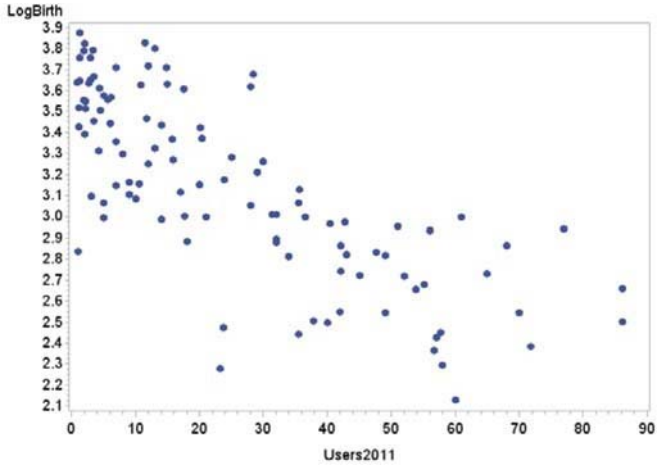


2.33. (a) Shown below. **(b)** As time increases, log count goes down. (c) The form is linear; the direction is negative; the strength is extremely strong. (d) There are no outliers. (e) Yes, the relationship is very linear, almost a perfect line.

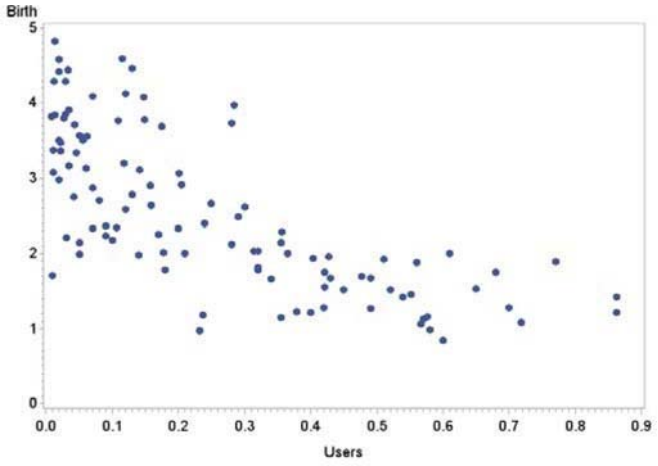


2.34. (a) The relationship is decreasing (negative) and seems curved. It appears that births decrease until about 40 Internet users per 100 people and then do not change. There is also a fair amount of scatter (especially when there are few Internet users per 100 people), so the relationship could be said to be moderate.(b) Association is not causation. Countries with many Internet users are also more developed. They may practice more birth control, for example.

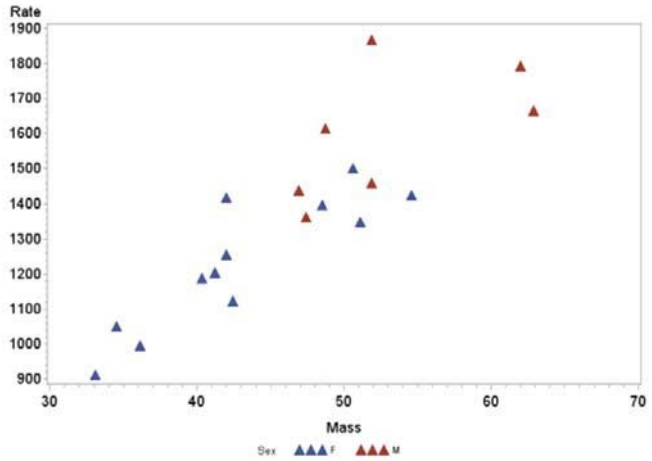
2.35.(a) Shown below. **(b)** The plot is more linear than the original scatterplot. **(c)** Answers may vary but the log transformed data should be preferred because it straightens out the relationship.



2.36. (b) The new variables are constant multiples of the original variables. For example, Birth=0 + 0.1BirthRate2011. **(c)** Shown below. **(d)** There is no change in relationship with the transformed variables. **(e)** Answers may vary. The original values are all scaled to be between 0 and 100, which is generally quite easy to interpret.



2.37. (a) Shown below. **(b)** The association is positive, linear, and strong. (c)Overall, the relationship is strong, but it is stronger for women than for men. Male subjects generally have both larger lead body mass and higher metabolic rates than women.



2.38r = 0.671.

2.39. (a) This is a linear transformation. Dollars= 0 + 0.01xCents. (b)r = 0.671. **(c)** They are the same. **(d)** Changing the units does not change the correlation.

2.40. If the relationship is not linear, computing *r* makes no sense; it only describes the strength and direction for linear relationships.

2.41. (a) Strong and positive. **(b)** Strong and negative. **(c)** Weak and negative. **(d)** No linear relationship.

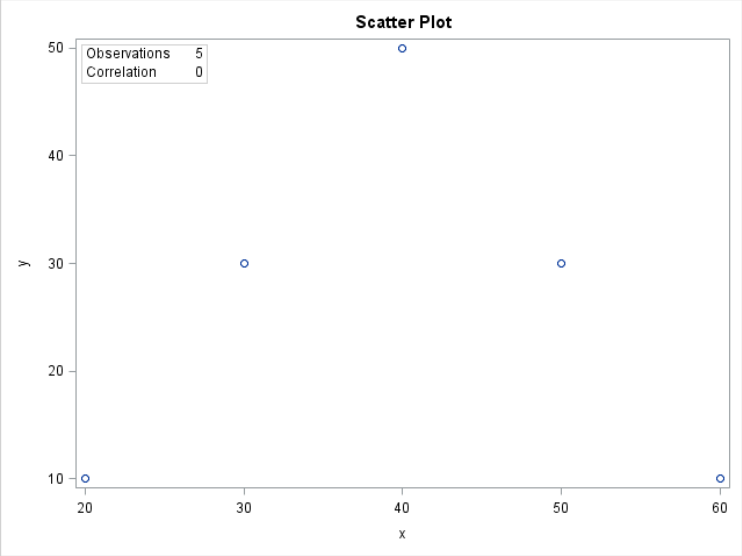
2.42.(a)r = 0.163. **(b)** No, the plot is not linear. **(c)** No, if they were approximately equal, the correlation would be close to 1.

2.43. (a)r = 0.298. **(b)** Probably. The plot is somewhat linear. **(c)** No,if they were approximately equal, the correlation would be closer to 1.

2.44.r = 0.981. Because of the high correlation, we know there is a strong linear relationship between CO2 emissions and highway fuel consumption. CO2 emissions could be well explained by the highway fuel consumption.

2.45. For fuel type D: $r = 0.977$. For fuel type E: $r = 0.983$. For fuel type X: $r = 0.981$. For fuel type Z: $r = 0.976$. The correlations for all four fuel types of vehicles are similar, around 0.98. The relationship between CO2 emissions and highway fuel consumption is linear and strong for all four fuel types.

2.46. (a)Shown below. **(b)** The relationship betweenx andy is strong, but it is not linear; it has a curved relationship or parabola. **(c)** $r = 0$. **(d)** The correlation is only good for measuring the strength of a linear relationship.



2.47. (a)r = 0.905. (b)Yes, the correlation is appropriate because the pattern is linear. There is one outlier, but it fits the overall pattern.

2.48. (a)r = 0.794. **(b)** The correlation is appropriate because the pattern is somewhat linear; however, there is an outlier that does not match the overall pattern that could be lowering the value of the correlation.

2.49. The person who wrote the article interpreted a correlation close to O as if it were a correlation close to -1 (implying a negative association between teaching ability and research productivity). Professor McDaniel's findings mean there is little linear association between research and teaching; for example, knowing that a professor is a good researcher gives little information about whether she is a good or bad teacher.

2.50. (a)r = -0.964. **(b)** Because there is an obvious curve in the data, the correlation is not a good numerical summary for these variables.

2.51. (a)r = -0.999. **(b)** Correlation is a good numerical summary here because the scatterplot is strongly linear. (c)Both have a correlation value that is close to 1, which would indicate a strong relationship. However, the correlation only makes sense if there is a linear relationship. The relationship in the previous exercise is not linear.

2.52. (a)r would be 1. (The relationship would be exactly StoreBrand = 0 + 0.8*BrandName.) (b) *r* would be 1. (The relationship would be exactly StoreBrand = -2 + BrandName.)

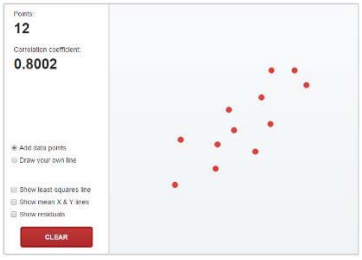
2.53. (a)r = 0.905. **(b)** The correlation does a good job of describing the relationship because it is quite linear; however, there is one outlier in this dataset, O'Doul's, with an extremely low alcohol percent.

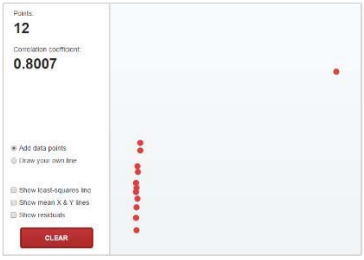
2.54. (a) Removing O'Doul's, $r = 0.908$. **(b)** Removing outliers that do not fit the overall pattern generally strengthens the correlation; here, the correlation did go up a bit but was already quite high to begin with.

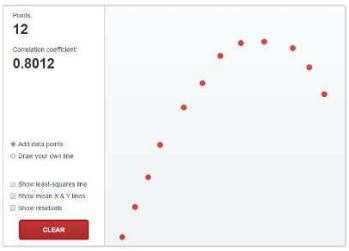
2.55.r = 0.912. Both correlations for the imported and domestic beers are quite similar, especially when the outlier O'Doul's is removed. The relationships between calories and percent alcohol for both types of beers are linear and strong and quite similar in pattern.

2.56. Answers may vary. **(a)** The correlation goes up and should be close to -1. **(b)** If you drag the point down far enough, you can make the correlation zero or even positive.

2.57. Answers may vary. (a)With only two points, the correlation will be **1** or-1 because they form a perfect straight line. **(b)- (d)** Shown below.

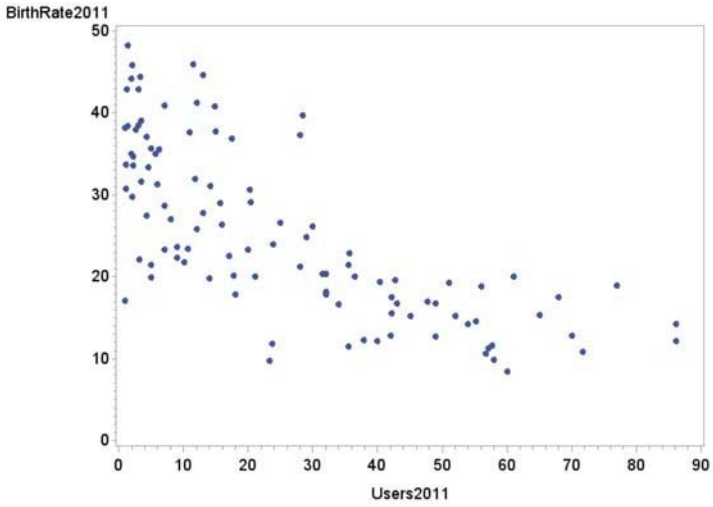






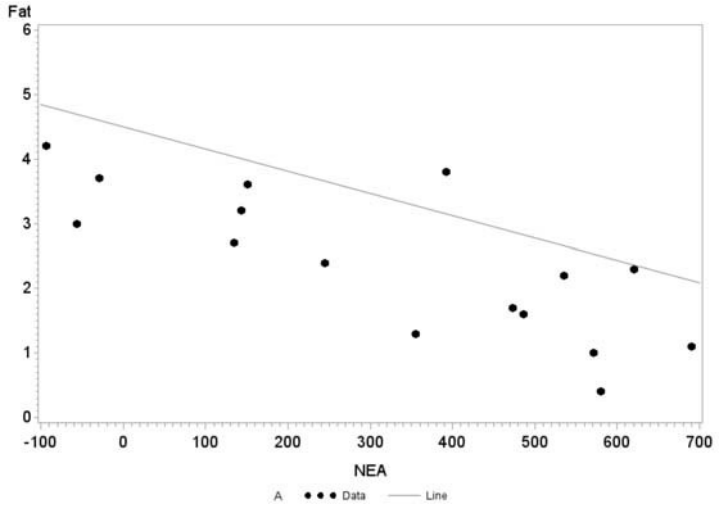
2.58. The correlation is $r = 0.48107$. The correlation is drastically lowered by the one outlier.

2.59.(a)r = -0.72971. **(b)** The correlation is not a good numerical summary for this relationship because there is a curvature in the plot.



2.60.(a)Because occupation has a categorical (nominal) scale, we cannot compute the correlation between occupation and anything. (There may be a strong *association* between these variables; some writers and speakers use "correlation" as a synonym for "association." It is much better to retain the more specific meaning.) (b)A correlation $r = 1.19$ is impossible because -1 *::Sr* ::S +1 always. (c)Neither variable (sex and color) is quantitative.

2.61. (a)Almost all the data points are below the line.



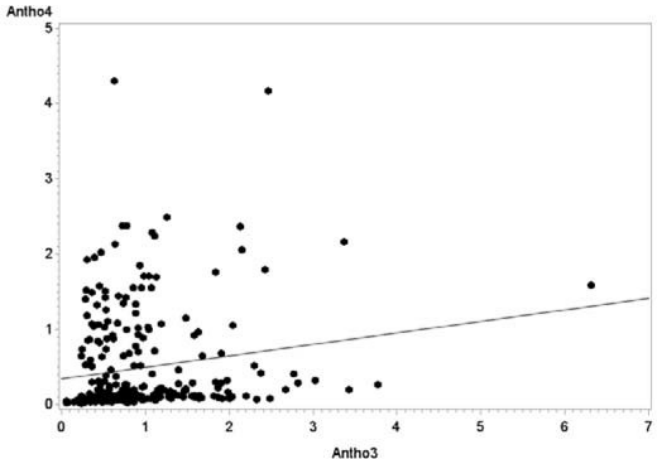
2.62. The predicted fat gain = 3.505 - (0.00344 x 250) = 2.645.

2.63. Predictions for 300 and 600 would be trustworthy because they are within the range of the data for NEA. We would not trust predictions from - 300 and 800 because they are outside the range of the data. This would be extrapolation.

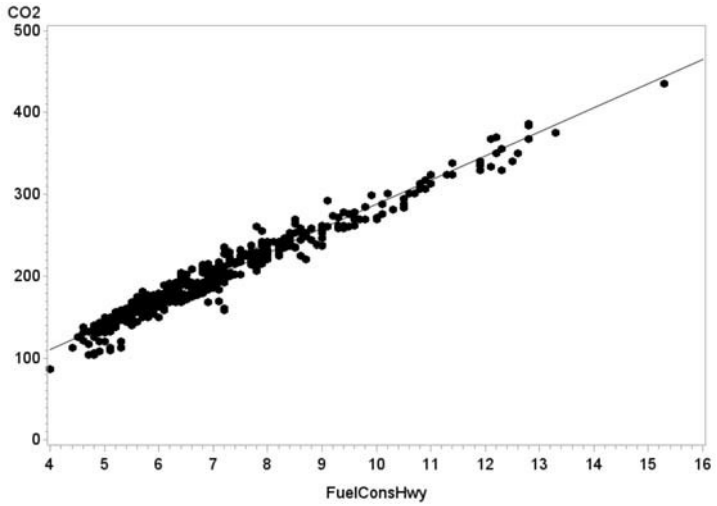
2.64. The values for $r2$ are given the table below. As the correlation *r* moves away from 0, the value of $r2$ mcreases.

r	-0.9	-0.5	-0.2	0	0.2	0.5	0.9
r^2	0.81	0.25	0.04	0	0.04	0.25	0.81

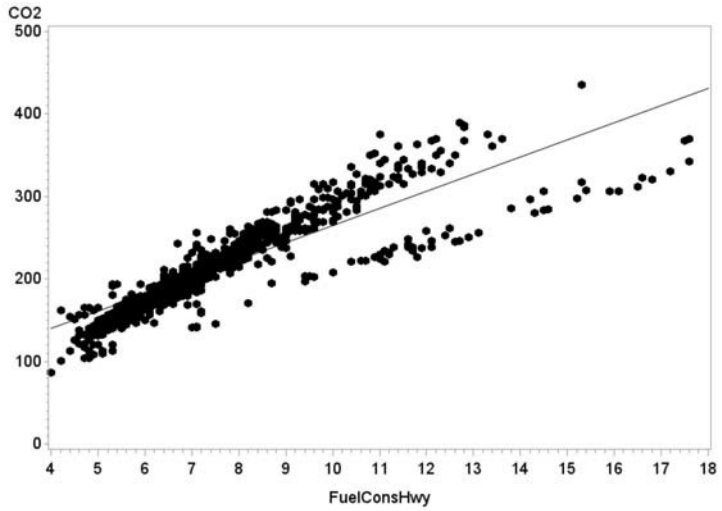
2.65. (a) y = 0.34475 + 0.15185Antho3. **(b)** The scatterplot is shown below. **(c)** The line does not fit the data well. Several of the data points are very far from the line.(d) y = 0.34475 + 0.15185(1.5) = 0.572525.



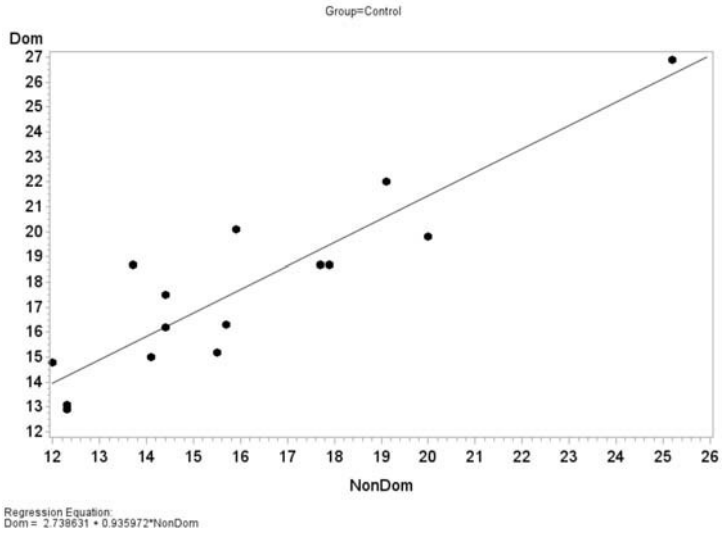
2.66. (a) $\hat{y}$ = -6.69394 + 29.44378FuelConsHwy. **(b)** The scatterplot is shown below. **(c)** We would expect the line to represent the data quite well because all the data points, even including the outlier, fall close to the line. **(d)** $\hat{y}$ = -6.69394 + 29.44378(8.0) = 228.8563.



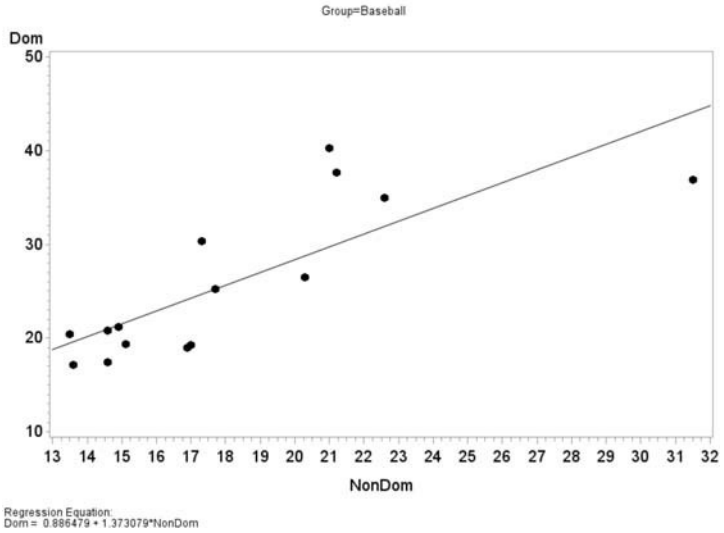
2.67. (a) $\hat{y}$ = 56.75012 + 20.78205FuelConsHwy. **(b)** The scatterplot is shown below. **(c)** A single regression line would not be a good fit for the four types of vehicles, even though the correlations were all very close. From the plot, we can see there are several different lines that need to be accounted for with separate regression lines.



2.68. (a) - (b) Plot shown below. **(c)** For the control group, there is a strong linear relationship between bone strength of the dominant arm and the nondominant arm. For each unit increase in nondominant, the dominant arm bone strength increases by 0.936.



2.69. (a) - (b) Plot shown below. **(c)** For the baseball group, there is a strong linear relationship between bone strength of the dominant arm and the nondominant arm. For each unit increase in the nondominant, the dominant arm bone strength increases by 1.373. The increase is greater than in the control group.



2.70. Predicted bone strength is 2.74 + 0.936*16 = 17.716 $cm^4/1000$.

2.71. Predicted bone strength is 0.886 + 1.373*16 = 22.854 $cm^4/1000$.

2.72. (a) Baseball players will have 22.854- 17.716 = 5.138 $cm^4/1000$ more bone strength than a non-baseball player when they both have nondominant bone strength of 16 $cm^4/1000$.(b) Answers will vary. One possible explanation is that baseball players still use the nondominant arm in batting and lift weights (which would strengthen both arms). **(c)** We note that, as the nondominant arm gets stronger, the difference becomes larger (baseball players always having more bone strength than the controls).

	12	16	20
Baseball players	17.36	22.85	28.35
Non-Baseball players	13.97	17.72	21.46
Difference	3.39	5.138	6.886

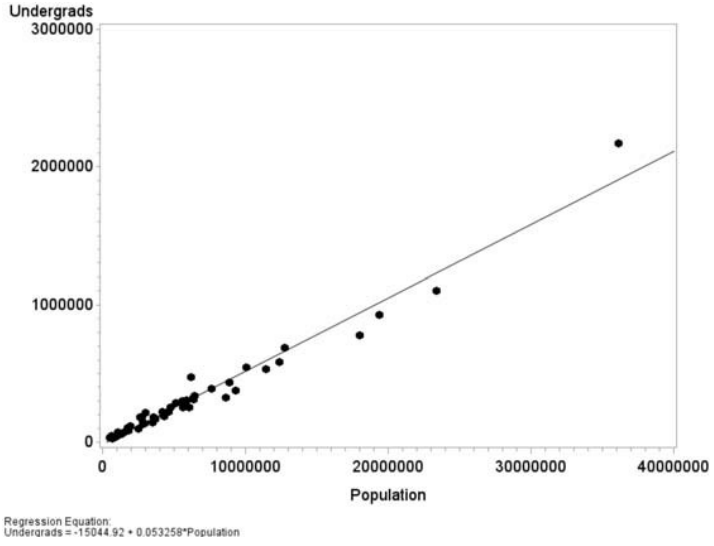
2.73. (a)-(d) below. (e) In terms ofleast squares, the first line is a better description of the relationship between time and radioactive counts. The sum of its squared differences (the least squares that is minimized in regression) is 8440.2; the sum for the second line is 187,706.

		Count = 602.8 – (74.7 × time)			(d) Count = 500 – (100 × time)		
Time	Count	Predicted (a)	Difference (b)	Squared Difference (c)	Predicted	Difference	Squared Difference
1	578	528.1	49.9	2490.01	400	178	31,684
3	317	378.7	−61.7	3806.89	200	117	13,689
5	203	229.3	−26.3	691.69	0	203	41,209
7	118	79.9	38.1	1451.61	−200	318	101,124

2.74. (a)-(d) below. **(e)** Once again, the first line is a better predictor; the sum of its squared differences is 0.0039 (to four decimal places), whereas the sum for the second line is more than 1.76.

		Log Count = 6.593 – (0.2606 × time)			(d) Log Count = 7 – (0.2 × time)		
Time	Log Count	Predicted (a)	Difference (b)	Squared Difference (c)	Predicted	Difference	Squared Difference
1	6.35957	6.3324	0.0271	0.0007	6.8	−0.4404	0.1940
3	5.75890	5.8112	−0.0523	0.0027	6.4	−0.6411	0.4110
5	5.31321	5.290	0.02321	0.0005	6.0	−0.6868	0.4717
7	4.77068	4.7688	0.00188	0.0000	5.6	−0.8293	0.6878

2.75. (a) and **(d)** Plot shown below. **(b)** The relationship is linear, positive, and strong. There are several outliers with large population values. **(c)** The slope is $b_1 = \frac{s}{n}_{xs} = 0.98367\sqrt{\frac{358,460}{6,620,733}} = 0.05326$. The intercept is $b_0 = y - b/x$ = 302,136 - 0.05326(5, 955, 551) = -15057 (-15045 from software). The regression line is j) = -15057 +0.05326x.



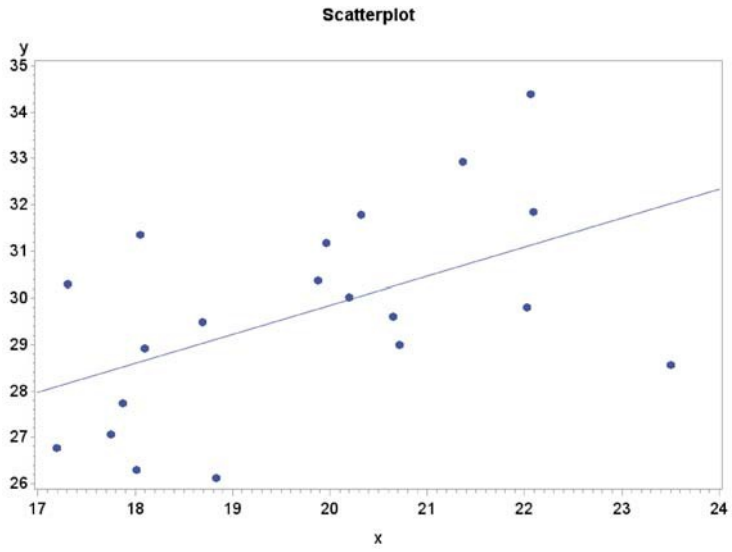
2.76. The slope is $b_1 = r\frac{s_y}{s_x} = 0.97081\left(\frac{165{,}270}{3{,}310{,}957}\right) = 0.04846$. The intercept is $b_0 = \bar{y} - b_1\bar{x} = 220{,}134 - 0.04846(4{,}367{,}448) = 8487$ (8491.90705 from software). The regression line is $\hat{y} = 8487 + 0.04846x$.

2.77. (a) $\hat{y} = -15047 + 0.05326(4{,}000{,}000) = 197{,}993$ (197,987 from software). **(b)** $\hat{y} = 8487 + 0.04846(4000000) = 202{,}327$ (202,328 from software).(c)In this situation, the outliers did not change the prediction for the median-sized state. Although they have populations much larger than the rest of the states, the outliers follow the pattern of the regression line, so they are not influencing our regression equation drastically.

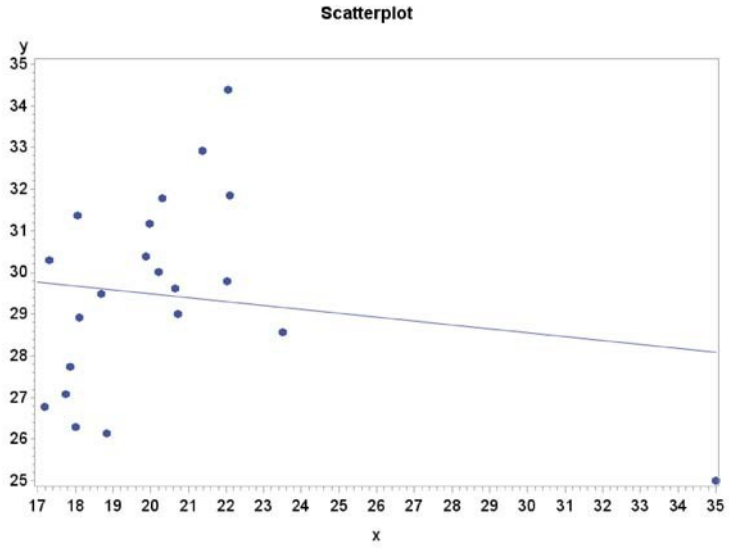
2.78.(a) $\hat{y} = -15044.917 + 0.053x$.(b)$r^2 = 0.968$. **(c)** 96.8% of the variation in the number of undergraduates is accounted for by the population size. (**d)** The software does not report the nature of the relationship; it is assuming a linear relationship in the calculations shown.

2.79.(a) $\hat{y} = 8491.907 + 0.048x$.(b)$r^2 = 0.942$. **(c)** 94.2% of the variation in the number of undergraduates is accounted for by the population size. (**d)** The software does not report the nature of the relationship; it is assuming a linear relationship in the calculations shown.

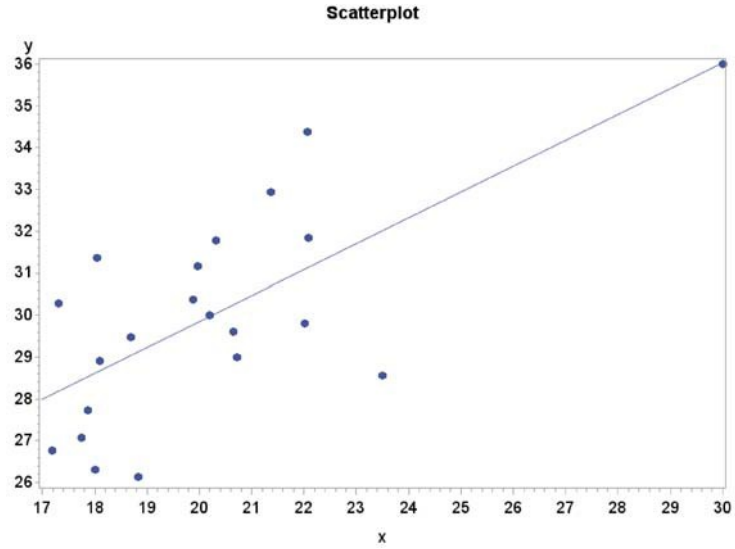
2.80. (a) Plot shown below. There seems to be a weak positive linear relationship betweeny and *x*. **(b)** $\hat{y} = 17.38036 + 0.62332x$. (c)$r^2 = 0.2737$ or 27.37%. **(d)** The *x* variable only accounts for 27.37% of the variation in *y*, so the relationship is fairly weak.



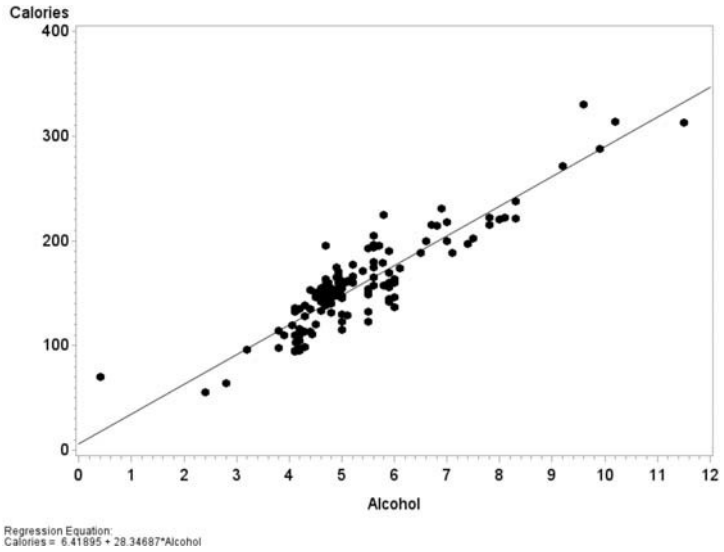
2.81. (a) Plot shown below. There seems to be a weak negative linear relationship between *y* and *x*, but with one extreme outlier with a very high *x* value. **(b)** $\hat{y} = 31.38274 - 0.09425x$. **(c)** $r^2 = 0.0224$ or 2.24%. (**d)** Here, the outlier completely eliminated all evidence of a regression line. The *x* variable accounts for only 2.24% of the variation *iny*, meaning there is no linear relationship at all. However, we know this is wrong because it is mostly due to the outlier unnaturally twisting the regression line.

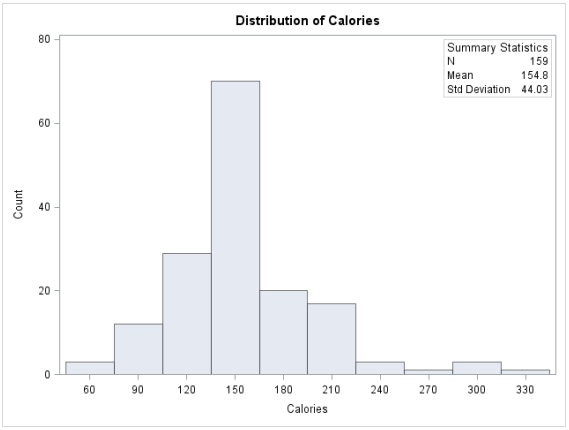


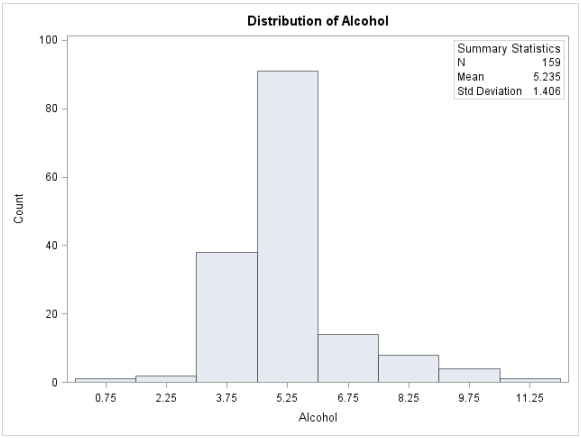
2.82. **(a)** Plot shown below. There seems to be a strong positive linear relationship betweeny andx but with one extreme outlier with high values for both *x* and *y,* but this outlier does fit the overall pattern. $\hat{y}$ = 17.47170+0.61861x. $r2=$ 0.4843 or 48.43%. There is a jump in the R-square value from 27% up to 48% just from this one observation, indicating that this observation is unnaturally influential in the analysis. **(b)**In this exercise, 2.81, the outlier drastically decreased the relationship between *y* and *x*, changing the $r2$ from 27% to 2%. In exercise 2.82, the outlier drastically increased the relationship between *y* and *x*, increasing the $r2$ from 27% to 48%. This demonstrates that a single outlier can be influential and can mislead our interpretation of the relationship between *y* and *x* if not careful.



2.83. (a) $\hat{y}$ = 6.41895 + 28.34687x. **(b)** $r2$ = 0.8193; 81.93% of the variation in calories is explained by the relationship with percent alcohol. **(c)** The relationship between calories and percent alcohol is linear, positive, and strong; however, there does seem to be one low outlier, O'Doul's, with a very low alcohol content.

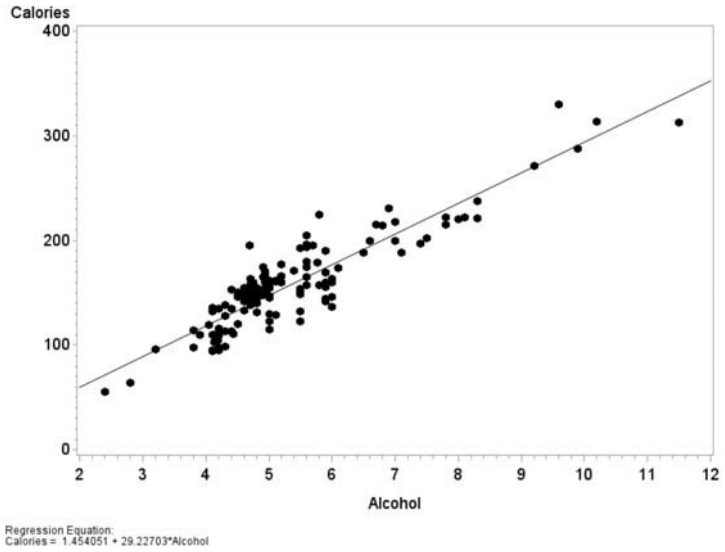


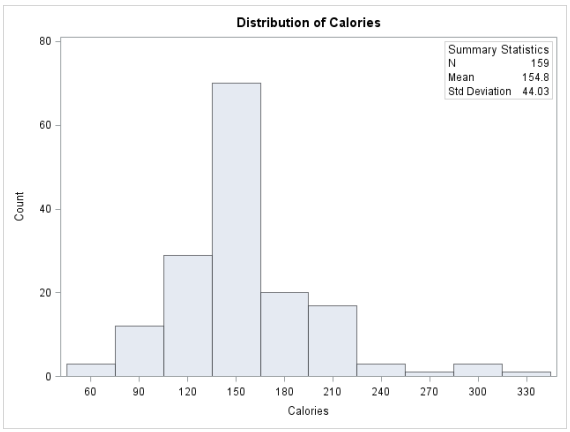


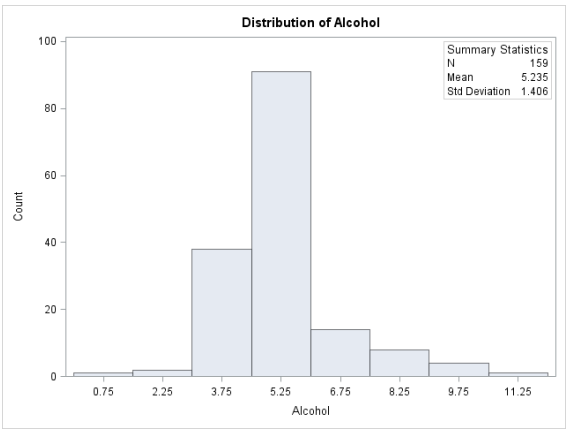


2.84. (a) y = 1.45405 + 29.22703x.(b)r2 = 0.8248; 82.48% of the variation in calories is explained by the relationship with percent alcohol. **(c)** The relationship between calories and percent alcohol is linear, positive, and strong. **(d)** In this situation, removing the outlier did not change the relationship much. The regression line is slightly different, and the $r2$ value went up slightly; overall, we are getting similar results.

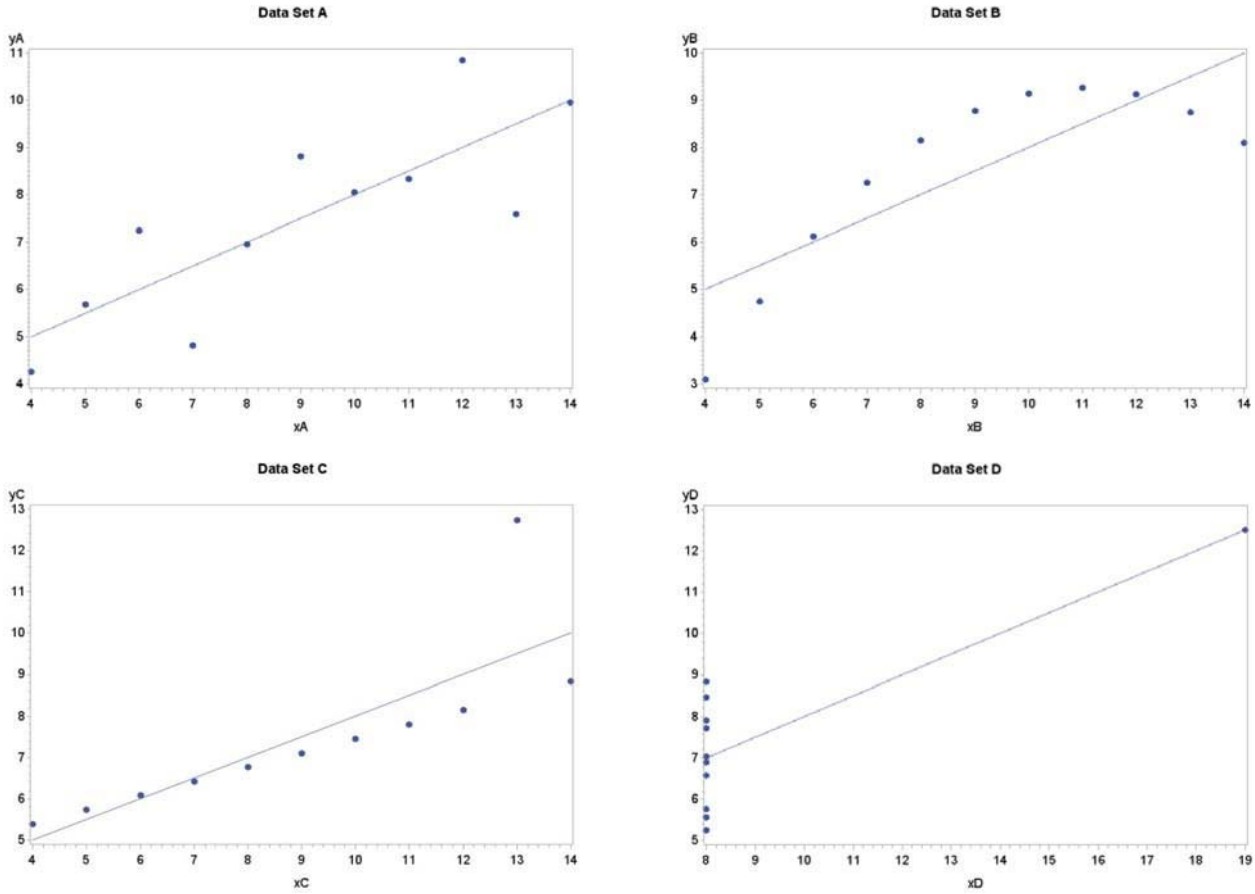
Note: *This is primarily because the outlier was quite close to the original regression line so that it is not influential on the regression analysis.*







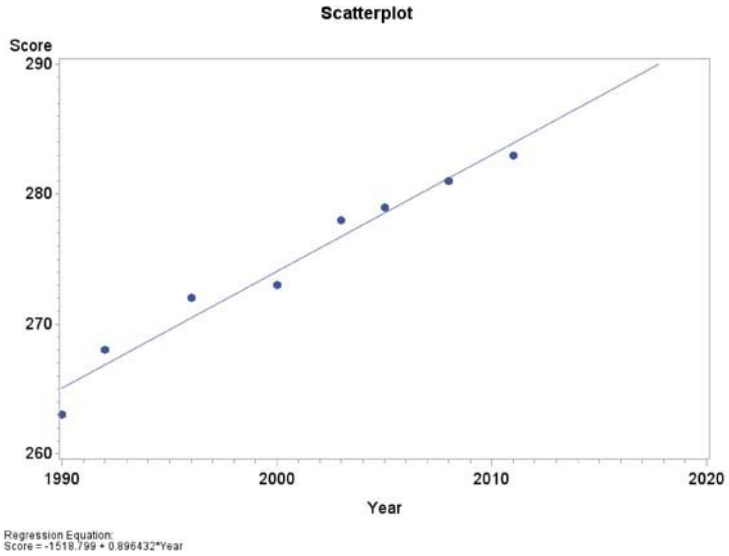
2.85. (a) The correlations and regression lines for all four datasets are essentially the same: $r = 0.82$ and $\hat{y} = 3 + 0.5x$. For $x = 10$, $\hat{y} = 3 + 0.5(10) = 8$. **(b)** Plots shown below.(c) For Data A, the regression line is a reasonable fit for the data. For Data B, there is obviously a curve in the scatterplot, and a transformation is needed before a regression should be run. For Data C, this is a perfect relationship but with one outlier, which is influential; a regression would not be valid for this data. For Data D, there is no relationship between y and x at all; only the extreme x outlier even makes the regression line possible, but it is definitely inappropriate to use regression on this dataset. Only for Data Set A should regression be used.



2.86. (a) Plot shown below. (b)The means and standard deviations are shown in the table below. The correlation is 0.98202. Using these, we get: b = 0.98202 ($\frac{6.864765}{7.520211}$) = 0.89643.

b_0 = 274.625 -0.89643(2000.63) = -1518.79918. The regression line is then j> = -1518.79918 + 0.89643Year. The percent of the year-to-year variation in scores explained by the linear trend is l = 0.9644. **(c)** Software gives the same equation.

	Mean	Std Dev
Year	2000.63	
	7.520211	



2.87. (a) $y = 15 - 2(4) = 7$. **(b)** For each 1unit increase in x, y decreases by 2. **(c)** The intercept is 15.

2.88. The means and standard deviations are shown in the table below. The correlation is 0.865.

For regressing metabolic rate on body mass, we get: $b_1 = 0.865\left(\frac{257.504}{8.2844}\right) = 26.8868;$

$h_0 = 1369.53 - 26.8868(46.7421) = 112.7845.$ The regression line is then j) = 112.7845 + 26.8868x.

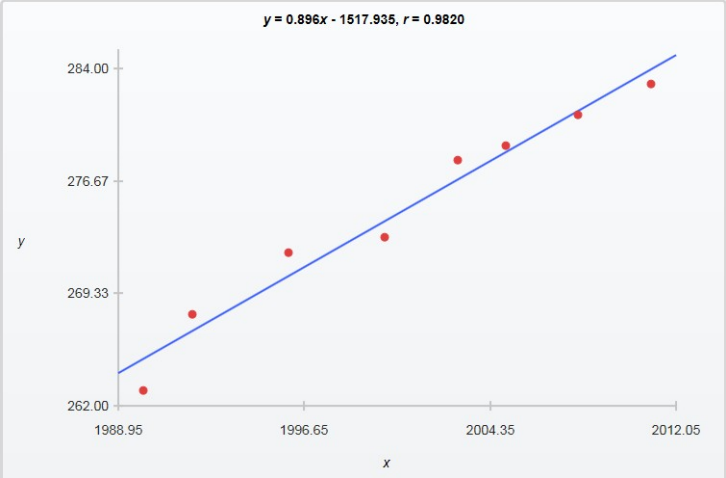
For regressing body mass on metabolic rate, we get: $b_1 = 0.865\left(\frac{8.2844}{257.504}\right) = 0.0278;$

$h_0 = 46.7421 - 0.0278(1369.53) = 8.6692.$ The regression line is then j) = 8.6692 + 0.0278x.

The units for the first equation slope are calories/kg; units for the second are kg/calorie.

	Mean	Std Dev
Mass	46.7421	8.2844
Rate	1369.53	257.504

2.89. The results from the applet are shown as follows. The regression equation is $y = 0.896x - 1517.935$. This is a strong, positive relationship because $r = 0.982$. We can conclude that NAEP scores are steadily increasing about 0.896 points per year.



2.90. The slope is $b_1 = r \frac{s_Y}{s_X}$, and the intercept is $b_0 = \bar{Y} - b_1\bar{X}$. The equation of the line is $\hat{Y} = b_0 + b_1 X$.

When $x = \bar{x}$, $y = (\bar{y} - b_1\bar{x}) + b_1\bar{x} = \bar{y}$.

2.91. r = --/0l6 = 0.4. r has a positive sign because students who attended a higher percent of classes earned higher grades.

2.92. We have $\hat{y}$ = 3.505-0.00344(250) = *2.645k g.* The residual is e = 2.4- 2.645 = - 0.245.

2.93. The residuals sum to 0.01.

2.94. The residuals are calculated as Dominant - (2.74 + 0.936 xNondominant).

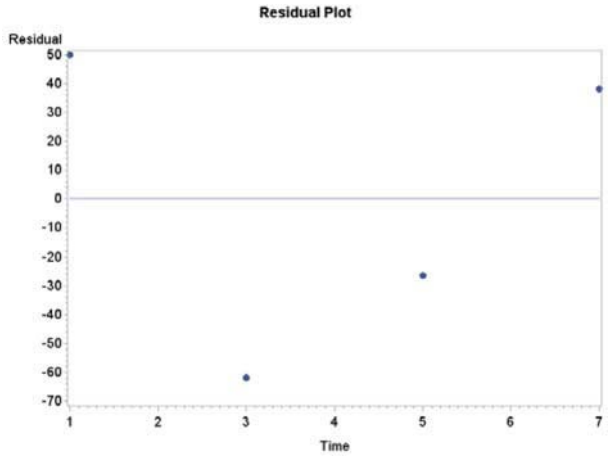
Nondominant	Dominant	Residual	ID	Nondominant	Dominant	Residual
12.0	14.8	0.83	7.0	12.3	13.1	-1.15
20.0	19.8	-1.66	8.0	14.4	17.5	1.28

2.95. The residuals are calculated as Dominant - (0.886 + 1.373 xNondominant).

Nondominant	Dominant	Residual	ID	Nondominant	Dominant	Residual
21.0	40.3	10.58	7.0	31.5	36.9	- 7.24
14.6	20.8	-0.13	8.0	14.9	21.2	- 0.14

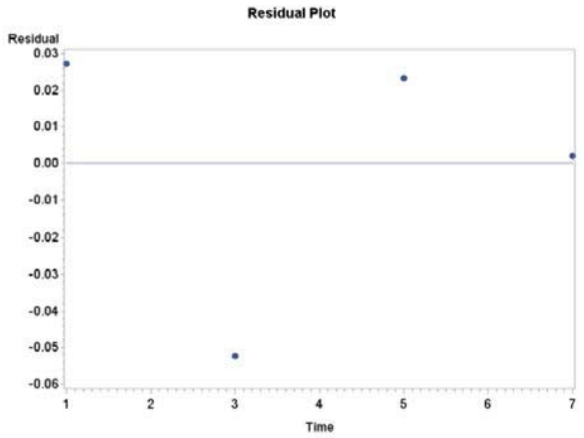
2.96. (a) - (b) The table and plot are shown below. **(c)** There is a clear curve in the residual plot; this is not a good model for these data.

Time	Count	Predicted	Residual
1	578	528.1	49.9
3	317	378.7	-61.7
5	203	229.3	-26.3
7	118	79.9	38.1

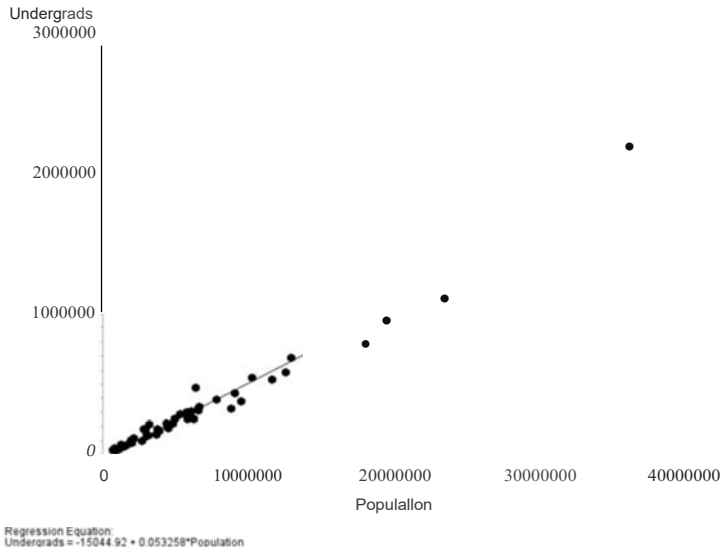


2.97. (a) - (b) The table and plot are shown below. **(c)** The residual plot looks random; the model using logs is much better.

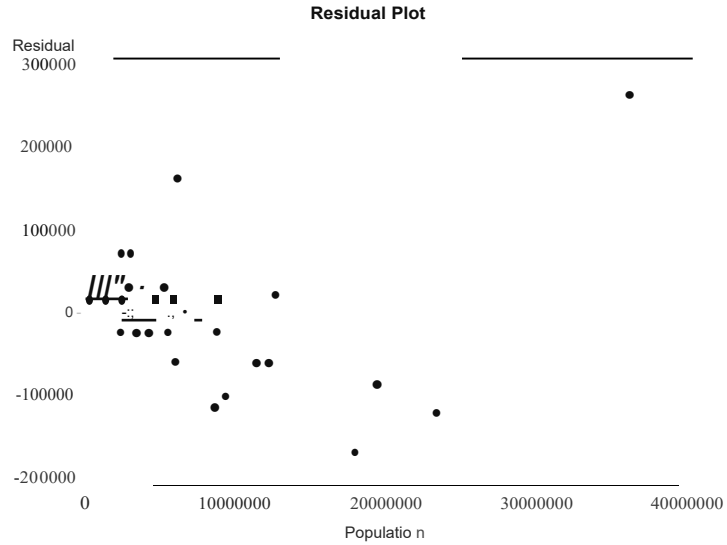
Time	LogCount	Predicted	Residual
1	6.35957	6.33244	0.02713
3	5.75890	5.81121	-0.05231
5	5.31321	5.28997	0.02323
7	4.77068	4.76874	0.00195



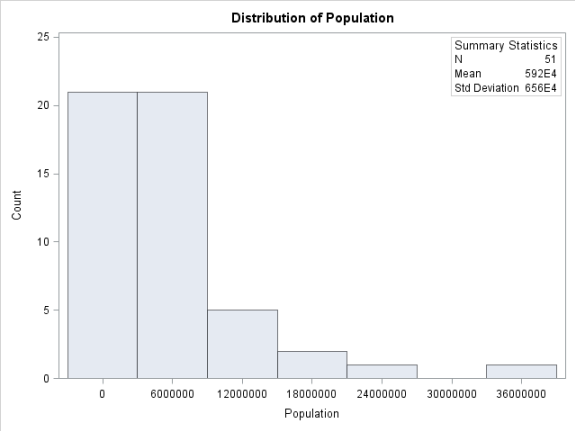
2.98. (a) Plot shown below.



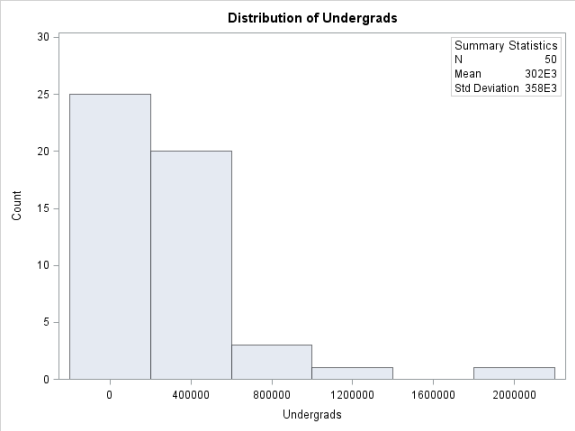
(b) Residual plot below.



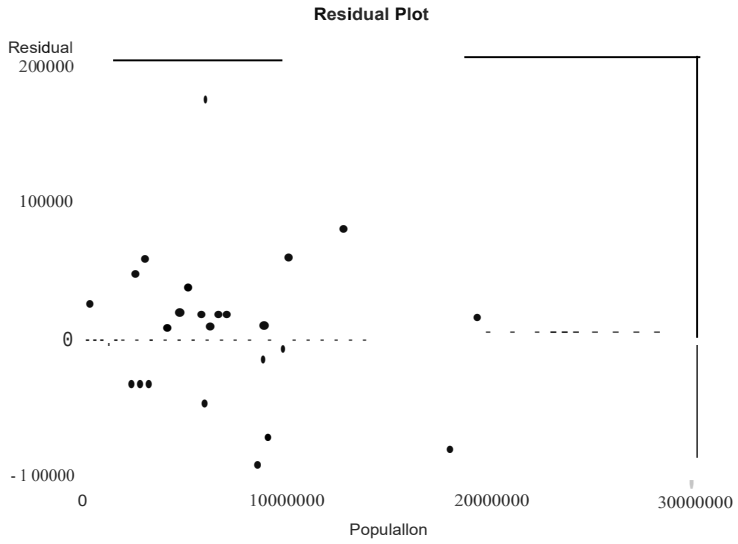
(c). The histogram below shows that California does appear as a slight outlier for the Population.



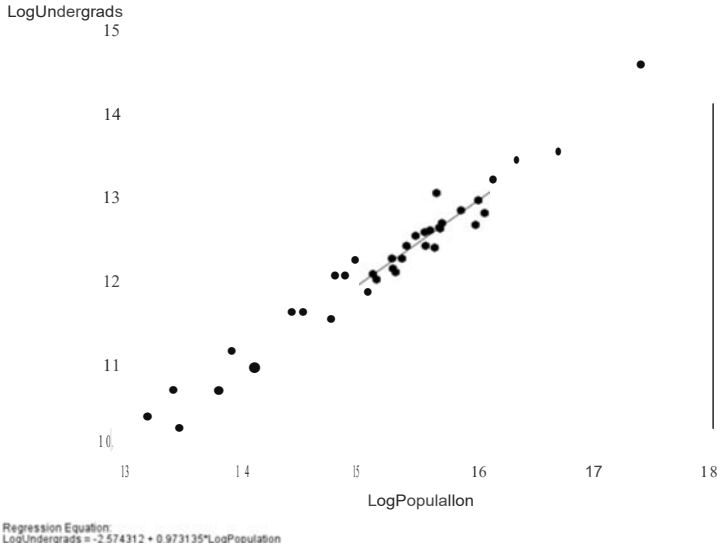
(d).The histogram below shows that California does appear as a slight outlier for Undergrads.



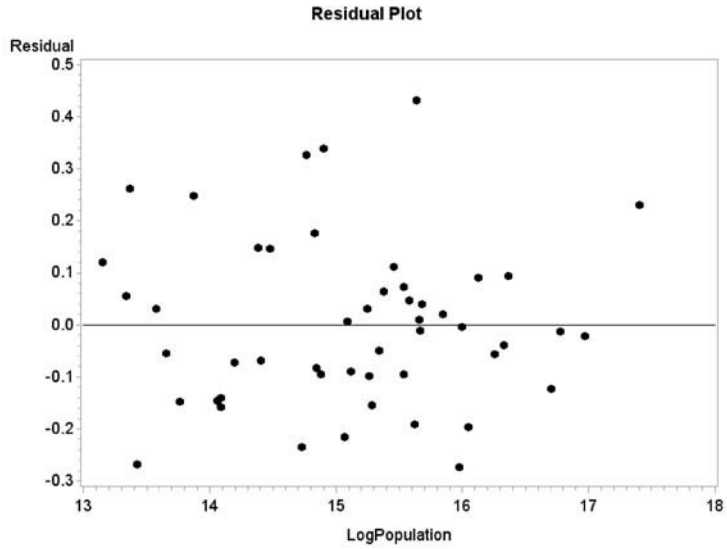
(e) At first, looking at the scatterplot in part **(a),** California follows the pattern with the other states and does not appear as an outlier in terms of the relationship. However, once we see the residual plot in part **(b**), it is clear that California is indeed an outlier for this relationship and does not match the pattern for the other states. It is clear where the regression line "should" be and that California pulling the regression line unnaturally. **(f)** It is likely that California is influencing the regression line. Below is the residual plot with California removed; overall, the plot looks fairly random, and we no longer have the "pull" by California that twisted the relationship.



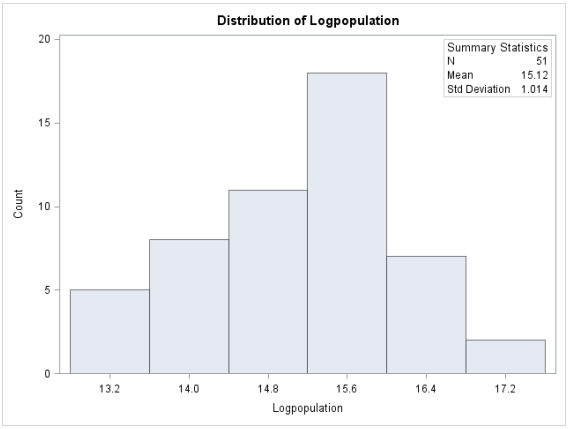
2.99. (a) Plot shown below.



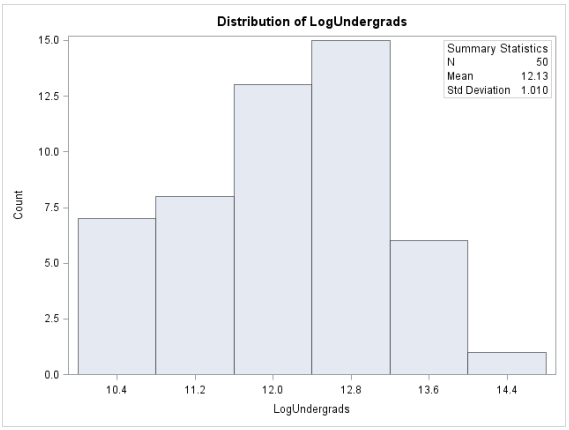
(b). Residual plot below.



(c). The histogram below shows that California does not appear as an outlier for LogPopulation.

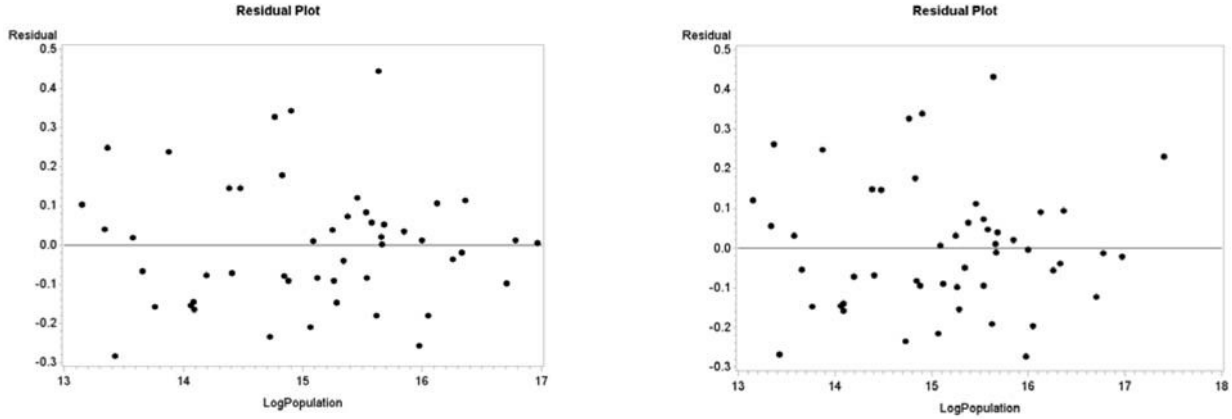


(d). The histogram below shows that California does not appear as an outlier for LogUndergrads.

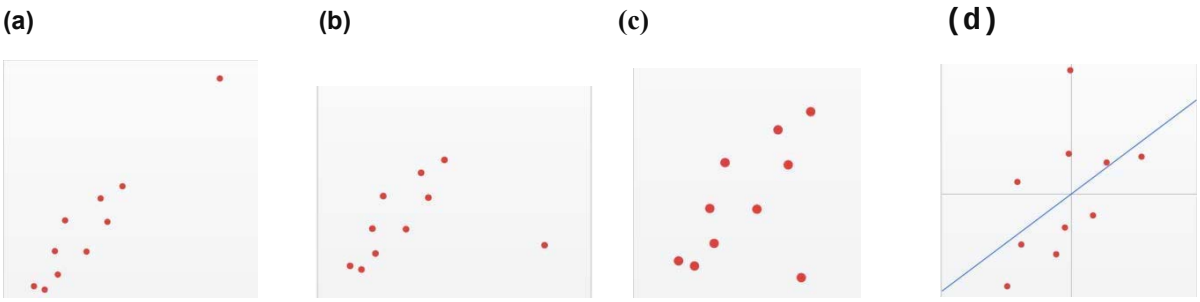


(e). Looking at the scatterplot in part **(a),** California follows the pattern with the other states using the log transformations; it does not appear as an outlier in terms of the relationship. The residual plot in part **(b)** confirms this. **(f)** California is not influencing the regression line using the log transformations. Below are

residual plots with and without California removed; overall, the plot looks nearly identical. Using logs seems to fix the influential nature of California for this dataset.



2.100. Examples will vary.(a) Generally, any extreme *x* outlier that follows the pattern of the rest of the data will not be influential to the regression. **(b)** Generally, any extreme *x* outlier that does not follow the pattern of the rest will be influential to the regression. **(c)** A point close to the outside of the *x* range is not an outlier; but such a point can still be influential. (**d)** An outlier with *x* coordinate exactly equal to $\bar{X}$ will change the intercept but not the slope.



2.101. (a) If the line is pulled toward the influential point, the observation will not necessarily have a large residual. **(b)** High correlation is always present if there is causation. **(c)** Extrapolation is using a regression to predict for x-values outside the range of the data (here, using 20, for example).

2.102. (a) If the residuals are all negative, one does not have a least-squares regression line. The line goes through the middle of the data; some residuals will be positive and some negative. Residuals must sum to 0. **(b)** A strong negative relationship DOES imply an association between the variables. **(c)** Lurking variables cannot always be measured. It would be difficult, for example, to measure "parental involvement" when examining a relationship between grades and television watching.

2.103. The Internet use does not cause people to have fewer babies. Possible lurking variables are economic status of the country, levels of education, etc.

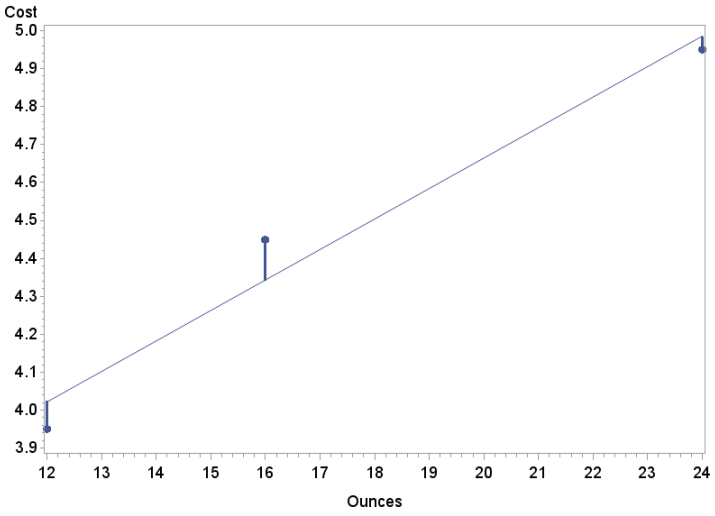
2.104. For example, a student who in the past might have received a grade of B (and a lower SAT score) now receives an A (but has a lower SAT score than an A student in the past). While this is a bit of an oversimplification, this means that today's A students are yesterday's A and B students, today's **B** students are yesterday's C students, and so on. Because of the grade inflation, we are not comparing students with equal abilities in the past and today.

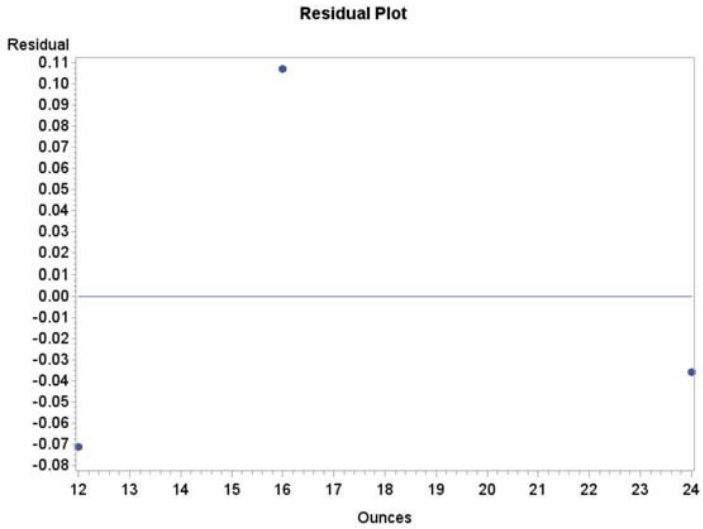
2.105. Answers will vary. For example, a reasonable explanation is that the cause-and-effect relationship goes in the other direction: Doing well makes students or workers feel good about themselves, rather than vice versa.

2.106. Patients suffering from more serious illnesses are more likely to go to larger hospitals (which may have more or better facilities) for treatment. They are also likely to require more time to recuperate afterward.

2.107. The explanatory and response variables were "consumption of herbal tea" and "cheerfulness /health." The most important lurking variable is social interaction; many nursing home residents may have been lonely before students started visiting.

2.108.(a) We put size on the *x* axis because we expect it to explain cost. As size in ounces increases, so does price. **(b)** j> = 3.05714 + 0.08036x.(c) The lines are on the plot.





(d) The residuals are-0.0714, 0.1071and-0.0357; they do sum to 0.