Unit 2 Homework

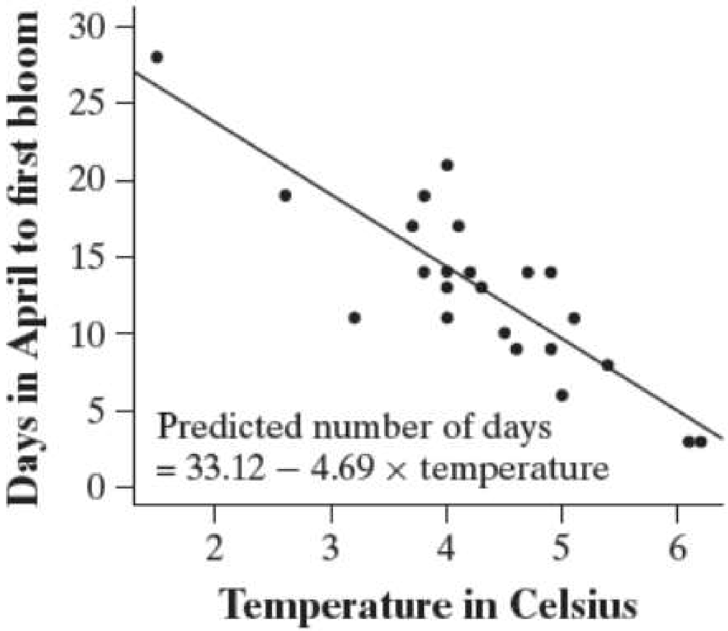
Assignment 5

R3.1 (a) There is a moderate, positive linear association between gestation and life span. Without the outliers at the top and in the upper-right, the association appears moderately strong, positive and curved.

(b) The hippopotamus makes the correlation closer to 0 because it decreases the strength of what would otherwise be a moderately strong positive association. Because this point's x coordinate is very close to $\overline{x}$, it won't influence the slope very much. However, it makes the y intercept higher because its y coordinate is so large compared to the rest of the values. Because it has such a large residual, it increases the standard deviation of the residuals.

(c) Because the Asian elephant is in the positive, linear pattern formed by most of the data values, it will make the correlation closer to 1. Also, because the point is likely to be above the least-squares regression line, it will "pull up" the line on the right side, making the slope larger and the intercept smaller. Because this point is likely to have a small residual, it decreases the standard deviation of the residuals.

R3.4 (a) The scatterplot is shown below. Average March temperature was chosen as the explanatory variable because changes in March temperature probably have an effect on the date of first bloom. Also, we are told that we want to predict the date of first bloom from the temperature.

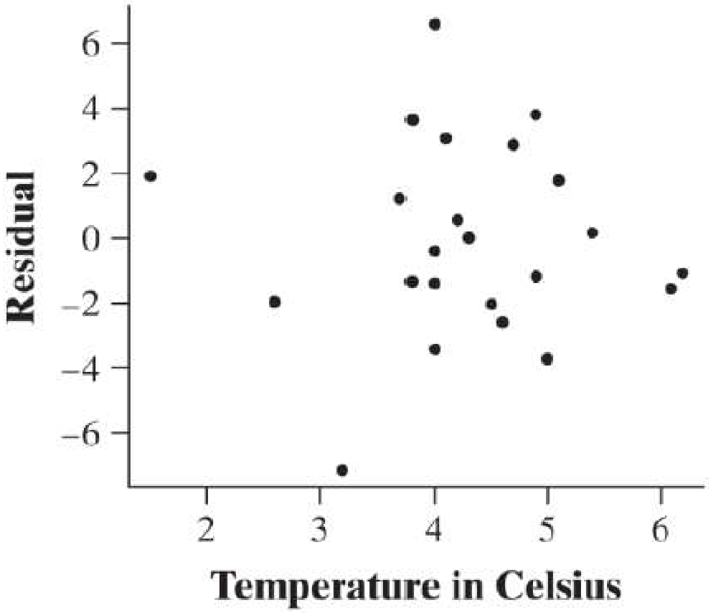


(b) The correlation is $r = -0.85$. The least-squares regression equation is $\hat{y} = 33.12 - 4.69x$ where y represents the number of days and x represents the temperature. The correlation tells us that there is a strong, negative linear association between the average March temperature and the days in April until first bloom. The slope tells us that for every 1 degree increase in average March temperature, the predicted number of days in April until first bloom decreases by 4.69. The y intercept tells us that if the average March temperature was 0 degrees Celsius, the predicted number of days in April to first bloom is 33.12 (May 3). However, $x = 0$ is outside of the range of data, so this prediction is an extrapolation.

(c) No, $x = 8.2$ is well beyond the values of x we have in the data set. This prediction would be an extrapolation.

(d) The predicted number of days until 1st bloom is $\hat{y} = 33.12 - 4.69(4.5) = 12.015$. The residual is $y - \hat{y} = 10 - 12.015 = -2.015$. In this year, the actual date of first bloom occurred about 2 days earlier than predicted based on the average March temperature.

(e) The residual plot is given below. There is no leftover pattern in the residuals, indicating that a linear model is appropriate.



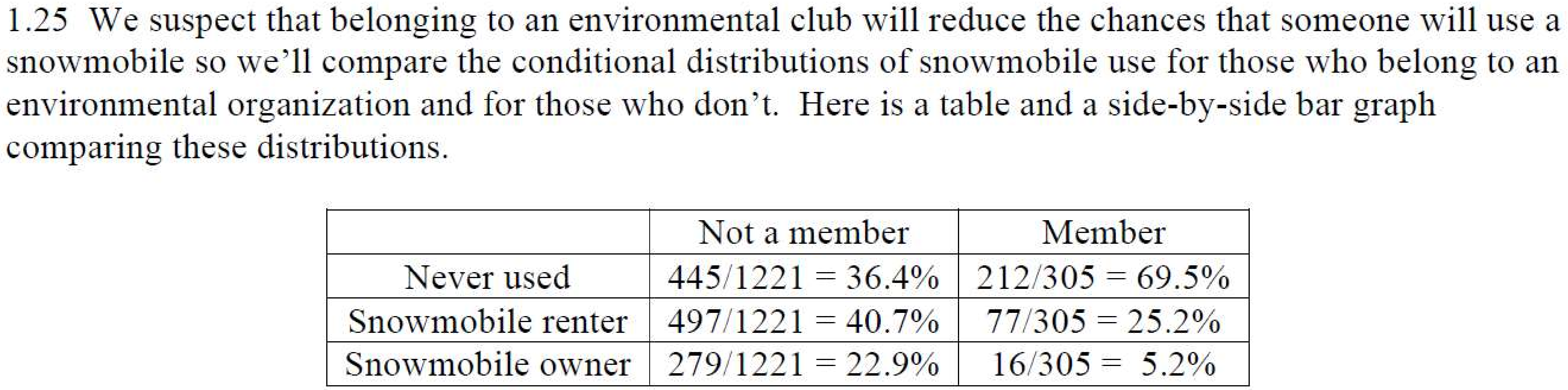
R3.5 (a) The slope of the regression line for predicting final-exam score from pre-exam totals is $b = 0.6\left(\frac{8}{30}\right) = 0.16$. The y intercept of the regression line is $a = 75 - 0.16(280) = 30.2$. Thus, the equation of the least-squares regression line is $\hat{y} = 30.2 + 0.16x$, where y = final exam score and x = total score before the final examination.

(b) Julie's predicted final exam score is $\hat{y} = 30.2 + 0.16(300) = 78.2$.

(c) Of all the lines that the professor could use to summarize the relationship between final exam score and total points before the final exam, the least-squares regression line is the one that has the smallest sum of squared residuals.

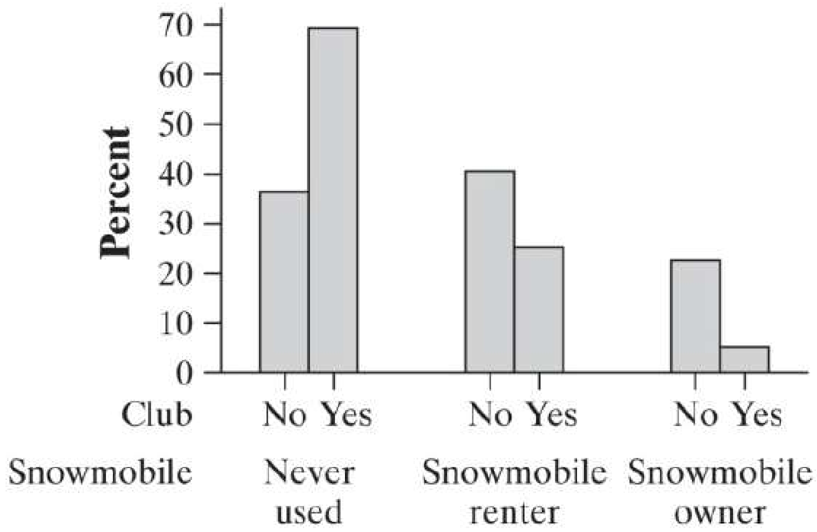
(d) Because $r^2 = 0.36$, only 36% of the variability in the final exam scores is accounted for by the linear model relating final exam scores to total score before the final exam. More than half (64%) of the variation in final exam scores is *not* accounted for by the least squares regression line, so Julie has a good reason to think this is not a good estimate.

Assignment 6



1.25 We suspect that belonging to an environmental club will reduce the chances that someone will use a snowmobile so we'll compare the conditional distributions of snowmobile use for those who belong to an environmental organization and for those who don't. Here is a table and a side-by-side bar graph comparing these distributions.

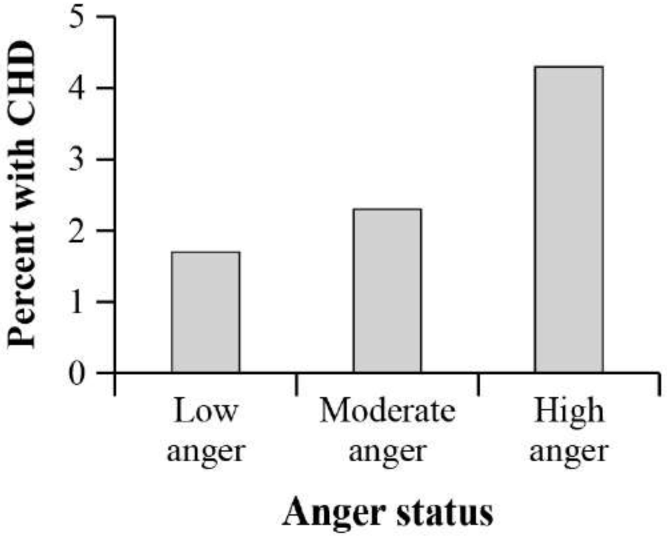
	Not a member	Member
Never used	445/1221 = 36.4%	212/305 = 69.5%
Snowmobile renter	497/1221 = 40.7%	77/305 = 25.2%
Snowmobile owner	279/1221 = 22.9%	16/305 = 5.2%



Based on these data, there is an association between environmental club membership and snowmobile use. The visitors who are members of an environmental club are much more likely to have never used a snowmobile than visitors who are not members of an environmental club. Those in an environmental club are less likely to have rented or owned a snowmobile than visitors who are not in an environmental club.

1.26 We suspect that people with different anger levels will have different rates of CHD, so we'll compare the conditional distributions of CHD for each anger level. Here is a table and a side-by-side bar graph comparing these distributions.

	Low anger	Moderate Anger	High Anger
CHD	53/3110 = 1.7%	110/4731 = 2.3%	27/633 = 4.3%
No CHD	3057/3110 = 98.3%	4621/4731 = 97.7%	606/633 = 95.7%



The data do support the study's conclusion about the relationship between anger and heart disease. The percent of the people in the study with CHD increases as the anger level increases.

1.27 d

1.28 b

1.29 d

1.30 d

1.33 d