

STATISTICS 101 - Homework 3 Answers

Due Wednesday, June 8, 2005

This assignment is worth a total of 100 points.

1. (20pts) An educational foundation would like to give scholarships to high school seniors who will be successful in college. The foundation wishes to see if there is a relationship between the score on a verbal aptitude test as a predictor of success in college and thus help them decide who should get the scholarships. The verbal aptitude test is on a scale of 200 to 800 and GPA is on a scale from 0 to 4. The plot that is appended to the back of this assignment is a plot of GPA versus the verbal aptitude test score for 50 students randomly selected from all students at a large public university.

- (a) (2pts) From the plot, what is the lowest GPA? What verbal aptitude score is associated with the lowest GPA?

The lowest GPA is approximately 1.4. The verbal score is approximately 460.

- (b) (2pts) From the plot, what is the highest GPA? What verbal aptitude score is associated with the highest GPA?

The highest GPA score is approximately 3.9. The verbal score is approximately 720.

- (c) (4pts) Describe the general pattern of the relationship between verbal aptitude score and GPA.

The general relationship between verbal aptitude score and GPA is positive and linear. As verbal aptitude increases, GPA increases as well. The relationship is not exact however, since there is quite a bit of scatter of the points.

- (d) (3pts) The value of the correlation coefficient for these 50 pairs of verbal aptitude score and GPA is 0.516. However, there appears to be an unusual pair or outlier. What are the verbal aptitude score and GPA for that apparent outlier? If this apparent outlier were removed, would the correlation coefficient calculated using the remaining 49 students be smaller than, about the same as, or larger than the 0.516? Explain briefly.

The outlier is located at verbal aptitude score 360 and GPA 2.4. If you draw approximate \bar{y} and \bar{x} axes on the graph, you will see that the outlier is contributing a negative value to the correlation coefficient. So removing the point would cause r to increase.

- (e) (9pts) Below are summary data for the 49 observations after eliminating the one outlier. Calculate the value of the correlation coefficient, r . Does this agree with your assessment in (d)?

First, we need to calculate the standard deviation for both X and Y.

$$s_x = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{246,074.88}{48}} = 71.6$$

$$s_y = \sqrt{\frac{\sum(Y - \bar{Y})^2}{n - 1}} = \sqrt{\frac{16.71}{48}} = 0.59$$

Now, we can calculate r .

$$r = \frac{1}{n - 1} \frac{\sum(X - \bar{X})(Y - \bar{Y})}{s_x s_y} = \frac{1}{48} \frac{1083.0}{71.6(0.59)} = 0.534$$

2. (36 pts) Can the length of a person's forearm be used to predict the length of a person's foot? We will collect data on the two variables during Lab 4. The data below were taken from a sample of 25 women during a previous semester.

Forearm (cm)	Foot (cm)	Forearm (cm)	Foot (cm)	Forearm (cm)	Foot (cm)
24	24	23	25.5	24.5	25.5
24	27.5	26	27	26	28
25	30	24.5	25	26	30
25.5	26.5	27.5	28	26	28
23.5	27	24.5	27.5	25	24
27	28.5	26	25.5	24	27
29	31	24	26	27	26.5
27	29	25	25	28	27
28	28				

10 points for JMP Output.

- (a) (3 pts) What is the explanatory variable? What is the response variable? Briefly explain your choice.
 The explanatory variable is the length of the person's forearm and the response variable is the length of the person's foot. We are using the forearm length to predict the foot length, so the foot length is the response variable.
- (b) (4 pts) Describe the general relationship between the two variables.
 The two variables have a positive linear relationship. The relationship appears to be moderately weak with no outliers.
- (c) (3 pts) Give the value for the slope of the least squares regression line. Give an interpretation of this value within the context of the problem.
 The slope of the regression line is 0.63. This means that for every 1 cm increase in the length of a person's forearm, the length of the person's foot will increase by an average of 0.63 cm.
- (d) (5 pts) Give the value for the intercept of the least squares regression line. Give an interpretation of this value within the context of the problem. Does this interpretation make sense? Explain your answer.
 The value of the intercept is 10.92. This means that when the length of a person's forearm is 0cm, the predicted foot length is 10.92cm. This interpretation does not make sense given the data since a person would not have a forearm length of 0cm.
- (e) (3 pts) Give the equation of the least squares regression line for this problem. Use this equation to predict the length of a person's foot given the length of their forearm is 29 cm.
 The equation of the least squares regression line is $\hat{y} = 10.92 + 0.63x$ where y stands for the foot length and x stands for the forearm length.
 If the forearm length is 29 cm, we would predict the foot length to be $\hat{y} = 10.92 + 0.63(29) = 29.19$ cm.
- (f) (3 pts) One woman had a forearm length of 29 cm. What is the residual for this woman?
 The women with forearm length of 29 cm had foot length of 31 cm. The residual is the observed value y minus the predicted value \hat{y} . $y = 31$ and $\hat{y} = 29.19$ so the residual is $y - \hat{y} = 31 - 29.19 = 1.81$ cm.

- (g) (2 pts) Would you use the least squares regression line to predict the length of a person's foot if their forearm length was 32 cm?

No, the maximum forearm length from the data is 29 cm. Predicting a foot length using a forearm length of 32 cm would be extrapolation.

- (h) (3 pts) Give the value of R^2 for this regression. Give an interpretation of this value in the context of the problem.

The value of R^2 is 30.05%. This means that 30.05% of the variation in foot length can be explained by the linear regression with forearm length.

3. (44 pts) We often hear reports about the relationship between diet and health. The data below give fat intake (grams) *per capita* per day (X) and the death rate (deaths per 100,000 people) from colon cancer (Y) for thirty nations in the year 1975.

Nation	Fat Intake per day (X)	Death rate (Y)	Nation	Fat Intake per day (X)	Death rate (Y)
Phillipines	28	4.5	Czechoslovakia	95	15.0
Japan	39	3.0	Hungary	100	14.0
Taiwan	46	3.5	Spain	101	8.5
Colombia	47	5.1	Finland	115	14.0
Chile	52	9.0	Austria	118	17.2
Panama	55	7.5	Australia	128	19.0
Mexico	57	3.9	Norway	129	17.0
Bulgaria	68	9.0	Sweden	129	18.5
Portugal	70	13.0	Ireland	134	21.5
Yugoslavia	70	7.0	Switzerland	138	22.0
Hong Kong	71	10.0	Belgium	140	21.0
Puerto Rico	77	5.5	United Kingdom	142	24.5
Italy	87	16.0	Canada	143	23.0
Poland	90	10.5	United States	148	21.0
Greece	95	7.5	New Zealand	152	23.0

10 points for JMP output.

- (a) (20 pts) Describe the distributions of fat intake and death rate. Make sure to include in your description the five number summary, the mean and standard deviation, and the shape of the histogram. Are there any outliers?

The distribution of fat intake is bimodal with a large mode from 125-150 and a smaller mode from 50-75. There is not a definite skew present in either the histogram or the stem and leaf plot. The mean is 95.47 and the median is 95, indicating a similar value for the center of the distribution. The five number summary is 28, 65.25, 95, 130.25, 152 and the standard deviation is 37.39.

The distribution of death rate is bimodal with a large mode from 5-10 and a smaller mode from 20-25. There is not a definite skew present in either the histogram or the stem and leaf plot. The mean is 13.14 and the median is 13.5, indicating a similar value for the center of the distribution. The five number summary is 3, 7.375, 13.5, 19.5, 24.5 and the standard deviation is 6.84.

- (b) (14 pts) Describe the scatterplot of fat intake vs. death rate. Give the regression equation for predicting death rate from fat intake, give an interpretation of the slope of

the regression equation, and give an interpretation of the R^2 value for the regression. Finally describe the residual plot, and make note of any potential problems with the regression.

Fat Intake and Death Rate have a strong positive linear relationship. There does not appear to be any outliers. The regression equation for predicting Death Rate from Fat Intake is $\hat{y} = -3.04 + 0.17x$, where \hat{y} is predicted death rate and x is fat intake. The slope of the equation is 0.17, indicating that for every 1 gram increase in the per capita fat intake, the predicted death rate from colon cancer will increase an average of 0.17 (deaths per 100,000 people). The R^2 value for the regression is 0.8592 indicating that 85.92% of the variation in death rate can be explained by the linear regression with fat intake. The residual plot is a general cloud of points. There does not appear to be any pattern to the residuals.

College GPA vs. Verbal Aptitude Score

