Statistics 110
PRACTICE MIDTERM EXAM

Open notes. Calculator required.

1. The R output below shows a regression analysis of data from 84 medium-sized counties in the US. For each county, X = percentage of adults in the county having at least a high-school diploma, and Y = crime rate (crimes reported per 100,000 residents) in a given year.

```
>lm(formula = Crime ~ Diploma, data = Midterm)

            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
Diploma      -170.58      41.57  -4.103 9.57e-05 ***

Residual standard error: 2356 on 82 degrees of freedom
Multiple R-squared: 0.1703,   Adjusted R-squared: 0.1602
F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05

>anova(Midterm)
Analysis of Variance Table

Response: Crime
          Df    Sum Sq  Mean Sq F value    Pr(>F)
Diploma    1  93462942 93462942  16.834 9.571e-05 ***
Residuals 82 455273165  5552112
```

a. Write the *population* version of the regression model.

b. Write the *estimated* (sample) regression function.

c. Interpret the slope in the context of this situation.

d. According to the last US Census, 82.7% of Orange County adults have a high school diploma. Round this number to 83%, and obtain a point estimate for the crime rate in Orange County.

Problem 1, continued…

e. A 95% confidence interval for μ_Y when $x^* = 70$ is 7702 to 9453. Interpret this interval in words, in the context of this situation.

f. The results indicate that counties with higher percentages of high-school graduates tend to have lower crime rates. Can we conclude from this study that having a high school diploma causes people to be less likely to commit crimes, in other words, that higher high-school graduation rates cause crime to be lower? Explain your answer.

2. Suppose that a plot of the residuals (Y axis) versus the predicted values (X axis) from a simple linear regression shows a "fan" or "megaphone" shape, with the residuals increasingly spread out as the predicted values increase.

a. Which regression assumption would appear to be violated?

b. Would it be more appropriate to do a transformation on the X values, or on the Y values? Explain.

3. What assumption is being examined by looking at a normal probability plot? Be specific.

4. A regression equation is to be fit for predicting Y = resting pulse rate using the predictor variables X_1 = number of minutes of exercise per week and X_2 = gender, with 1 = male and 0 = female. Explain in words what the coefficient attached to X_2 represents.

5. A company offers a training course for the Math SAT. They give their students a test at the end of the course, graded from 0 to 100. They would like to use that test in the future to predict how well students will score on the Math SAT. They have scores on their test and the Math SAT for a sample of students. Thus, X = score on the company's test and Y = score on the Math SAT, which ranges from 200 to 800. They plan to use the usual simple linear regression model.

a. Would the intercept have a useful meaning in this example? Explain your answer.

b. One of the company analysts states that the intercept should be fixed at 200, because that's the lowest the SAT Math score can be. Suppose the intercept is set to 200 for this situation. Write the population model.

c. Write the sum that is to be minimized to get the least squares regression line, if the model you wrote in Part b is used.

MULTIPLE CHOICE

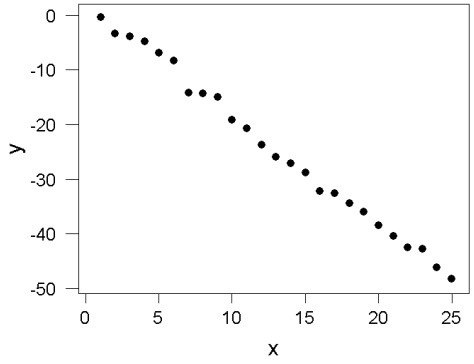
1. In a linear regression analysis with the usual assumptions, which one of the following quantities is the same for all individual units in the analysis?
   A. μ_y
   B. $\sigma\varepsilon$
   C. e_i
   D. $\hat{Y}_i$

2. A regression line is used for all of the following *except* one. Which one is *not* a valid use of a regression line in general?
   A. to estimate the average value of Y at a specified value of X.
   B. to predict the value of Y for an individual, given that individual's X-value.
   C. to estimate the change in Y for a one-unit change in X.
   D. to determine if a change in X causes a change in Y.

3. Which choice is *not* an appropriate description of $\hat{Y}$ in a regression equation?
   A. Estimated response
   B. Predicted response
   C. Estimated average response
   D. Observed response

4. Which of the following is the *best* way to determine whether or not there is a statistically significant linear relationship between two quantitative variables?
   A. Compute a regression line from a sample and see if the sample slope is 0.
   B. Compute the correlation coefficient and see if it is greater than 0.05.
   C. Conduct a test of the null hypothesis that the population slope is 0.
   D. Conduct a test of the null hypothesis that the population intercept is 0.

5. Shown below is a scatterplot of Y versus X.



   Which choice is most likely to be the approximate value of R^2?
   A. –99.5%
   B. 2.0%
   C. 50.0%
   D. 99.5%