Statistics 110

PRACTICE MIDTERM EXAM KEY

Note that pages have been condensed on this key to fit on 3 pages, to save paper if you print it.

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.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \textit{or} \quad \mu_Y = \beta_0 + \beta_1 X_i$$

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

$$\hat{Y}_i = 20517.6 - 170.58\ X_i \quad \textit{or} \quad Y_i = 20517.6 - 170.58\ X_i + e_i$$

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

If two counties differ by 1% in adults with a high-school diploma, on average the crime rate would be 170.58 lower (per 100,000 residents) for the county with the higher percentage of high school diplomas.

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.

$$\hat{Y} = 20517.6 - 170.58\ X = 20517.6 - 170.58\ (83) = 6359.46.$$

This is the predicted number of crimes reported per 100,000 residents.

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.

For the population of all medium-sized counties in which 70% of adults have a high-school diploma, we are 95% certain that the population mean crime rate that year was between 7702 and 9453 crimes reported per 100,000 residents.

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.

No. The data clearly come from an observational study, because you can't randomly assign different counties to have specific high-school graduation rates. There are many possible confounding variables that could influence crime rates, such as income levels, county services provided, etc. You can't make cause and effect conclusions in observational studies.

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?

The assumption of equal variance, i.e. that the variance (or standard deviation) of the errors (or the population of Y values at each X) is equal across the range of X values.

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

A transformation on the Y values would be more appropriate. A transformation on the X values does not change the variance of the Y values, and in this situation, we need to change the variance of the Y values to make them equal across the range of X's.

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

The assumption that the error terms ε in the population are normally distributed.

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.

The coefficient would be present in the model for males, and absent in the model for females, so it represents the average difference in pulse rates for males and females with number of minutes of exercise per week held constant. (Notice that the model forces us to assume that the difference between males and females is constant across the range of exercise amounts.)

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.

It might. It depends on whether or not some students score at or near 0 on the company's test. If so, then the intercept would be the predicted SAT score for students who score 0 on the company test. If all students score much higher, then the intercept would not have a useful meaning.

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.

$$Y_i = 200 + \beta_1 X_i + \varepsilon_i \quad \textit{or} \quad \mu_Y = 200 + \beta_1 X_i$$

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.

$$\Sigma(\varepsilon_i)^2 = \Sigma(Y_i - 200 - \beta_1 X_i)^2$$

MULTIPLE CHOICE (Correct answers in bold)

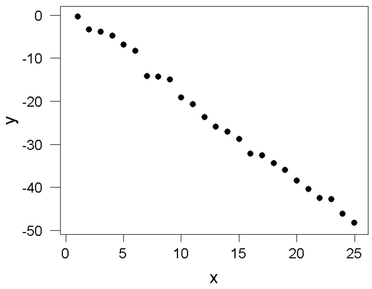
1. In a linear regression analysis with the usual assumptions, which one of the following quantities is the same for all individual units in the analysis?
   A. μ_y
   B. $\sigma\varepsilon$
   C. e_i
   D. $\hat{Y}_i$

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%**