Analytics Primer

Quiz 4

Honor Pledge:

I certify that I have not received or given unauthorized aid in taking this exam.

Signed: _____

Printed Name: KEY

Date: _____

There are 15 questions. The first 10 are worth 2 points each. The last 6 are worth different point values. There will be some True/False, Multiple Choice, and Short Answer questions.

1. Although an over-specified multiple regression model provides coefficient estimates that have higher variance, the estimates are still unbiased.
   a. True.
   b. False.

2. When we compute a two sample hypothesis test for means where we assume variances of the two populations are equal, why do we use the pooled standard deviation?
   a. The pooled estimate is an estimate derived from information prior to the sample.
   b. Both of the sample standard deviations are estimating the same population standard deviation so both should be used in an "average" of the two.
   c. We cannot use the t-distribution if we don't have a pooled standard deviation.
   d. Both b and c, but not a.

3. Which of the following is *not* a characteristic of the F distribution?
   a. Symmetric.
   b. Skewed Right.
   c. Only takes on positive values.
   d. Has two sets of degrees of freedom.

4. You are trying to test if the average GPA of males in your statistics class is larger than the average GPA of females in your statistics class. You sample 20 males and 20 females that are paired by degree major and take the difference in their GPA's. What distribution must be approximately Normal for inference from this test to be possible?
   a. Either the male or the female distribution of GPA.
   b. The distribution of both male and female GPA.
   c. The distribution of the difference in GPA for males and females.
   d. None of these because the sample size is 40 people and therefore large enough to need no distributional assumption.

5. Which of the following are signs and problems that are caused by multicollinearity?
   a. The signs of the coefficients change after a new variable is added to the model.
   b. Coefficients have extreme shifts in value when a new variable is added to the model.
   c. The standard deviation of the error in the model increases when a new variable is added to the model.
   d. All of the above.

6. An analyst is building a multiple regression model to model income. In the model, one of the analyst's variables is a categorical variable for age group that has four categories – teen, young adult, middle-aged, elderly. He is interested in testing if there is a significant difference between the teenage group and the other groups. How should he code his dummy variable?
   a. Use response coding with 4 dummy variables (one for each category).
   b. Use response coding with 3 dummy variables for young adult, middle-aged, and elderly.
   c. Use effects coding with 3 dummy variables for young adult, middle-aged, and elderly.
   d. Multiple regression cannot determine relationships between categorical variables and continuous variables.

7. You and a partner are afraid that you have underspecified a multiple regression model by only estimating a simple linear regression model. Your partner tells you that at least you don't have to worry about bias estimates of the coefficients. Is your partner correct?
   a. No, the coefficients in the simple linear regression model would be biased if there is an underspecified model.
   b. Yes, underspecified models do not have biased estimates of coefficients in the model.
   c. There is always bias in simple linear regression model coefficients.
   d. Both a and c, but not b.

8. List two of the assumptions of a One-way ANOVA.

   1. Populations are Normal
   2. Populations have equal variance

   Independence

9. The manager of a local fast food chain wants to calculate a hypothesis test to determine if the proportion of males that order french fries with a burger is higher than the proportion of females that order french fries with a burger. The manager sampled 87 males where 85 of them ordered fries with their burger and 57 females where 42 of them ordered fries with their burger. Is this a valid sample to conduct the test?
   a. Yes, the total sample size of 129 is considered large.
   b. Yes, both samples had more than 5 successes in the sample.
   c. Yes, the two sample sizes are both larger than 30.
   d. None of the above.

10. Write the null and alternative hypotheses for a One-way ANOVA for a factor with 4 levels.

H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: At least two means are unequal

An analyst calculating the needed number of reams of paper to order for each department in a college for a given year comes up with the following multiple regression model. Use the following information to answer questions 11 and 12.

$$x_1 = Number\ of\ professors\ in\ department$$
$$x_2 = Number\ of\ students\ in\ department$$
$$x_3 = Number\ of\ printers\ in\ department$$
$$x_4 = \begin{cases} 1 & if\ dept.\ in\ humanities \\ 0 & if\ dept.\ not\ in\ humanities \end{cases}$$

$$\hat{y}_i = 14.98 + 2.31x_{1,i} + 1.96x_{2,i} + 0.47x_{3,i} + 52.18x_{4,i}$$

11. (5 points) Interpret the meaning of the coefficient on variable x_4 in terms of the problem.

All else held constant, if a department is a humanities department, it needs 52.18 more reams of paper on average.

12. (5 points) Based solely on the names of the variables, is there any potential for multicollinearity in this model? Explain.

Yes, an argument can be made all variables are related to each other, especially the first three variables.

A group of consultants are based in two large cities. They are trying to determine if the clientele in one is different in age than the clientele in the other city. They sampled 150 clients from each city to determine their age. The first city sampled had an average age of 45.7 years with a standard deviation of 8.1. The second city sampled had an average age of 39.8 years with a standard deviation of 9.2. Use this information to answer questions 13 – 15.

13. (5 points) Paired sampling was not used in this study. Briefly explain the difference between a paired difference test and a two-sample test of means.

A paired difference test matches observations, from both samples, on similarities to eliminate possible biases in the sampling. It therefore focuses on the difference of the populations and not the populations themselves.

14. (5 points) Run a hypothesis test to determine if the variances of the two populations are equal with a significance of 0.05. State all of the steps to the hypothesis test.

$H_0: \sigma_1^2 = \sigma_2^2$

$H_a: \sigma_1^2 \neq \sigma_2^2$

$F = \frac{s_2^2}{s_1^2} = \frac{9.2^2}{8.1^2} = 1.29$

$F^*_{149,149} \Rightarrow F^*_{100,100} = 1.483 > 1.29 \Rightarrow$ P-value $> \alpha$

DO NOT REJECT H_0

Not enough evidence to say the two variances are different.

15. (10 points) Based on your answer to question 14, conduct the appropriate two sample hypothesis test. State all the steps to the hypothesis test.

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{1,0} - \mu_{2,0})}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$= \frac{45.7 - 39.8 - 0}{8.67 \sqrt{\frac{1}{150} + \frac{1}{150}}} = 5.89$

$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$

$= \sqrt{\frac{149 \cdot 8.1^2 + 149 \cdot 9.2^2}{150+150-2}}$

$= 8.67$

P-value $< 0.001 < \alpha$

REJECT H_0

Cities have different ages of clientele.