902189898

Test 3

STAT 021

Swarthmore College

Do not flip this page until instructed to do so.

Test organization: There are 12 questions in total on this test and they are organized into three subsections: the first 4 questions are matching or True/False with explanation questions, the next 5 questions are free response short answer and should not require more than a sentence or two to answer. The last section contains 3 long answer free response questions that require more than a couple of sentences to answer fully. There are a total of 60 points possible on this test. The last section explains an extra credit opportunity. If you need additional scratch paper you may come to the front of the class and pick some up.

Instructions: Answer each question to the best of your ability and raise your hand if you are confused by any of the wording in the questions or suspect a typo. For the short and long answer questions show all your work and provide enough justification and/or explanation in order to get full credit or to be considered for partial credit. You do not need a calculator to evaluate any expressions. For any calculation problems, simply writing out the formula to find the answer will suffice.

First and Last Name: Satchel Tsai

Swarthmore Username: Stsai1

Take a deep breath.

You have prepared for this test and with a clear and well-rested mind, you are ready to show me what you have learned this semester. The purpose of this test is to measure your understanding of the material we have covered this semester. This is nothing more than a metric for me to evaluate your preparedness to think statistically at this particular moment in time and in this particular setting. This is not a perfect measure of your knowledge and does not predict your future statistical skills.

Section 1: Matching and True/False problems

1. (5 points)

Suppose we are modeling the weight of birds (in kg) as a linear function of a categorical predictor variable for bird type (with levels pigeon, sparrow, and finch) and a numeric predictor for bird age. Given a "full" model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon,$$

where $x_1 = \begin{cases} 1, \text{ if sparrow} \\ 0, \text{ otherwise} \end{cases}$, $x_2 = \begin{cases} 1, \text{ if finch} \\ 0, \text{ otherwise} \end{cases}$ and x_3 is the age of the bird (in months), match the questions below to their corresponding null hypotheses.

a) For newly hatched birds (of age zero months), is there a statistically discernible difference in the weights of these three different bird types?

b) Does the effect of age on a bird's weight depend on what type of bird it is?

c) Given we are only comparing birds of the same age, is there a statistically significant difference in the mean weight of sparrows and pigeons?

d) Given we are only comparing pigeons, is the effect of age on a bird's weight statistically significant?

e) Is there statistically discernible evidence of a linear relationship between bird age and type and bird weight?

1. c $H_0 : \beta_1 = 0$
2. a $H_0 : \beta_1 = \beta_2 = 0$
3. d $H_0 : \beta_3 = 0$
4. b $H_0 : \beta_4 = \beta_5 = 0$
5. e $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

2. (5 points)

Determine which of the following statements about MLR models are true and false. For each statement that is false, provide a brief explanation as to why it is false.

(a) If predictors are collinear, then removing one variable will have no influence on the point estimate of another variable's coefficient.

false unless predictors are perfectly collinear, removing one variable will influence the point estimate of another variable's coefficient.

(b) If a regression model's first variable has a coefficient of $\hat{\beta}_1 = 5.7$, then if we are able to influence the data so that an observation will have a value of x_1 be one unit larger than it was before, the value of y_1 for this observation would increase by 5.7 units.

false. $\hat{\beta}_1$ represents correlation, not causation.

(c) As the total sample size increases, the degrees of freedom for the residuals increases as well.

true.

3. (5 points)

Determine which of the following statements about ANOVA models are true and false. For each statement that is false, provide a brief explanation as to why it is false.

If the null hypothesis that the means of four groups are all the same is rejected from an ANOVA model and overall F-test at a 5% significance level, then ...

(a) We can then conclude that all the means are different from one another.

false. we can conclude that at least one pair of means is different from one another.

(b) The standardized variability among the group averages is higher than the estimate of the variability of the data within each group.

true.

(c) A post-hoc pairwise analysis will identify if there is at least one pair of means that are significantly different.

false. A post-hoc pairwise test will identify which pair are significantly different.

4. (5 points)

Determine if the following statements about statistical modeling are true or false, and explain your reasoning. If false, state how it could be corrected.

(a) Decreasing the significance level (α) will increase the probability of making a Type 1 Error.

false. Decreasing the confidence level $(1-\alpha)$ will increase the probability of making a type 1 error.

(b) With large sample sizes, even small differences between the null value and the observed point estimate will be identified as statistically significant.

true.

(c) Correlation is a measure of the association between any two variables.

true.

Section 2: Short answer questions

5. (4 points)

Briefly describe a benefit of analyzing the studentized residuals of a regression model rather than just analyzing the observed residuals.

analyzing the studentized residuals allows us to easily identify potential outliers using moderate threshold value of 2 and extreme threshold value of 3.

6. (3 points)

If you could only use one measure (among the studentized residuals, leverage values, and Cook's distance values) to identify potentially influential data points, which would you choose and why?

I would choose Cook's distance, because it accounts for how unusual a response is and the potential to affect the model (how unusual predictor values are), incorporating the benefits of leverage values & studentized residuals.

For questions 7-9 consider the following random single-serving samples of $n = 76$ breakfast cereals. We are going to model the average calories per serving (in g) (`calories`) as a linear function of the cereal manufacturer (a categorical variable with levels: G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina). Below is the R summary output for this one-way ANOVA model.

```
##
## Call:
## lm(formula = calories ~ Manufacturer, data = cereal_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.696  -8.696  -0.126   5.909  51.304
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     111.364      3.959  28.126  < 2e-16 ***
## ManufacturerK    -2.668      5.538  -0.482  0.63149
## ManufacturerN   -24.697      8.553  -2.887  0.00516 **
## ManufacturerP    -2.475      7.348  -0.337  0.73729
## ManufacturerQ   -16.364      7.667  -2.134  0.03633 *
## ManufacturerR     3.636      7.667   0.474  0.63678
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.57 on 70 degrees of freedom
## Multiple R-squared:  0.1618, Adjusted R-squared:  0.102
## F-statistic: 2.703 on 5 and 70 DF,  p-value: 0.02724
```

7. (3 points)

(a) What are the error degrees of freedom based on this model?

(b) What is the reference level?

$df = 76 - 5 - 1 = \boxed{70}$

The reference level is G = General Mills

8. (6 points)

Suppose the average amount of calories for all these samples is 106.97 over all 76 data points. What is the estimated group effect for Quaker Oats cereal brand?

$111.364 - 16.364 - 106.97 = \boxed{71.97 \text{ g calories per serving}}$

$\mu_j = \mu + \alpha_j$

$111.364 - 16.364 = 106.97 + \alpha_j$

9. (4 points)

Consider two additional numeric predictors: `sugars` (in g) and `protein` (in g). If we were to fit a regression model including each of the three predictor variables (including manufacturer) and an interaction between the two numeric variables, explain the meaning of the coefficient for the interaction term within the context of this data. (You should be able to answer this in no more than two sentences.)

The coefficient of the interaction term is the effect of sugars on calories depending on how much protein is in the cereal, allowing for simultaneous changes in manufacturer.

numeric – cost, % full-time staff

categorical – private/public

liberal arts/cc/technical/group

Section 3: Long answer questions

10. (9 points)

Suppose you have access to a data set on a random sample of undergraduate-only institutions in the US. The variables included in this data set are a numeric variable for the average cost of tuition each semester, a binary categorical variable distinguishing private institutions from public ones, a numeric variable for the percentage of full-time instructional staff employed at the institution, and a categorical variable indicating whether the school is a liberal arts college, a community college, a technical/vocational school, or if they are institutionally affiliated with certain groups (e.g. historically Black, women's only, tribal, etc).

State a research question that can be answered with the overall F-test for each of the following models, based on this data. (You do not need to use every variable, but you can.) Also provide a mathematical representation of the model and state the null hypothesis based on the notation you define for each model.

(a) a simple linear regression model;

(b) an ANOVA model;

(c) a multiple linear regression model (not SLR or ANOVA).

a) Is there a statistically significant linear association between the average cost of tuition each semester and the % of full time staff employed?

$y = \beta_0 + \beta_1 x_1 + \varepsilon$ where y = average cost of tuition each semester, x_1 = % of full time staff employed

$H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$

b) Is there a statistically significant difference in average cost of tuition across different categories of schools?

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ where y = average cost of tuition each semester

$x_1 \begin{cases} 1, \text{if community college} \\ 0, \text{if otherwise} \end{cases}$ $x_2 \begin{cases} 1, \text{if technical/vocational school} \\ 0, \text{if otherwise} \end{cases}$; $x_3 \begin{cases} 1 \text{ if group affiliated} \\ 0, \text{if otherwise} \end{cases}$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ H_A: at least one of $\beta_1, \beta_2, \beta_3 \neq 0$

c) allowing for simultaneous changes in % full time staff, what is the impact of an institution's status as private or public on the average cost of tuition each semester?

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ where y = average cost of tuition each semester

x_1 = % of full time staff employed, $x_2 \begin{cases} 1, \text{if private} \\ 0, \text{if otherwise} \end{cases}$

$H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$

11. (8 points)

Consider the ANOVA model for the cereal data you used in questions 7-8. Reference the R output on pg 5 and the plots on pg 10 to answer the following questions about this model.

(a) Check the conditions necessary for conducting a test to determine if the average calories (per serving) is significantly different for these six different cereal manufacturers. (You do not need to check the zero mean or linearity conditions but you do need to describe what it means for the group effects to be constant in this context.)

(b) Write out in words and in symbols the hypotheses that would be tested in part (a). (Clearly define your notation.)

(c) What can you conclude about the test in part (b)? Write a paragraph discussing your conclusions and reference any relevant statistics and/or plots as part of your discussion.

a) we can assume that the independence condition is met because the average calories in one cereal likely do not affect those in another. The samples were randomly collected, meeting that conditions. We have slight concerns about the normality condition, because we observe tailing on either end of the normal quantile plot. We have reason to believe that the constant variance condition may be violated, because the ranges of the box plot vary widely across groups, meaning that the group effects are not constant. This is especially apparent in the general mills & quaker groups.

b) H_0: There is no evidence of a linear relationship between calories and manufacturer.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$ where y = average calories per serving

$x_1 \begin{cases} 1, \text{if kellogs} \\ 0, \text{if otherwise} \end{cases}$ $x_2 \begin{cases} 1, \text{if Nabisco} \\ 0, \text{if otherwise} \end{cases}$ $x_3 \begin{cases} 1, \text{if post} \\ 0, \text{if otherwise} \end{cases}$ $x_4 \begin{cases} 1, \text{if quakeroats} \\ 0, \text{if otherwise} \end{cases}$ $x_5 \begin{cases} 1, \text{if Ralston Purina} \\ 0, \text{if otherwise} \end{cases}$

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_a: There is evidence of a linear relationship between calories & at least one pair of manufacturers. At least one of $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 \neq 0$

c) Based on the ANOVA test, we can conclude that there is evidence of a linear relationship between calories and at least one pairing of manufacturers at a 95% confidence level, because $p = 0.0272 < 0.05$. Because of the violation of the normality & constant variance condition, this test may not be reliable. Further analysis is needed to determine which pairwise comparisons are significant. The $R^2_{adjusted}$ for this model is 0.102, suggesting that the model isn't doing a great job of accounting for variability in the model.

12. (8 points)

Suppose two people are studying the historic data set about the amount of arsenic (Arsenic) in local wells. This data contains $n = 70$ observations from a random selection of well water samples from across the state. In addition to the levels of arsenic, the data also records the year the data was collected (Year) and the distance from the well to the nearest mining site (Miles).

Person A fits the following MLR model to the data:

$$Arsenic = \beta_0 + \beta_1 Year + \beta_2 Miles + \epsilon$$

and computes an adjusted R^2 value of 0.26.

Person B considers the following correlations:

$$Cor(Arsenic, Year) = \rho_1; \quad Cor(Arsenic, Miles) = \rho_2$$

and estimates each with their sample correlations $r_1 = 0.77$ and $r_2 = -0.34$. Are the two people's conclusions contradictory? Explain your answer.

While these results appear surprising, they are not necessarily contradictory. Adding a new variable to a linear model will never cause the overall R^2 to go down. However, the adjusted R^2 for a model will decrease if a newly added variable does not contribute much to a model. Additionally, there may be interaction occurring between the two predictor variables that is causing the model to do a worse job explaining variability, since the correlation between Arsenic & Year is positive & the correlation between Arsenic & Miles is negative.

Section 4: Extra credit opportunity

If the response rate to my end of the semester evaluation form (on Moodle under Week 13 and 14) is at least 85% of our class size (over both sections), two percentage points will be added to everyone's Test 3 grade (up to 100 total possible points). **Hint:** You may not know how to or want to contact everyone in my class but you do know your group mates pretty well.

Cereal ANOVA Model