902216353

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: Mackenzie Tucker

Swarthmore Username: mtucker1

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

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.

TRUE: If predictors are collinear, they explain the same information inside a y variable. Removing one form should not have significant influence on the other.

this model predicts an average increase of 5.7 units.

(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: the value of y_i cannot be changed. If all other variables are held constant & x_i increases by 1,

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

TRUE: for error $(n-k-1)$ and total $(n-1)$

* Note for model → df related to # of predictor terms

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 of the means are different from the "grand" mean.

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

TRUE (see model output?)

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

like Fisher's LSD

TRUE finds the smallest difference considered statistically significant

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.

TRUE - requires more evidence to reject null

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

TRUE - larger sample size, SD and variance tend to decrease, meaning that a small difference is more significant than in a small sample

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

TRUE → ranges from -1 to 1; the closer the correlation is to 1 → the higher the correlation

Correlation measures the strength of a relationship between two variables!

takes into account the standard deviation and leverage to give context about the point within the model.

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.

The studentized residuals allows for comparison between the observed and expected values in a regression model across many different predictor variables for a specic point. Observed residuals are more limited to just the observed & predicted value.

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?

Cook's distance specifically gives an estimate of the influence of a point. We could use these values to determine potentially influential data points

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
```

$\hat{\sigma}$

7. (3 points)

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

(b) What is the reference level?

a) $n-k-1 \rightarrow 76-5-1=70$

b) Reference Level: 5 (based on number of cereal brands studied)

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?

$y = \mu + \alpha_1 + \epsilon$

$\mu_{k-1} = \mu_k - \beta_{k-1} = \beta_0 - \beta_{k-1}$

$106.97 - 16.364 = \mu_{k-1}$

μ

106.97 ← given group mean

$\alpha_Q \rightarrow$ **-16.364 cals/serving** ← found as estimate in R

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.)

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2 \cdot x_3$

x_1: indicator var: cereal brand

x_2: sugars(g)

x_3: protein(g)

The interaction term is analyzing if the relationship between sugars and proteins per serving has an effect on total calories per serving. If the cereal brand predictor variable is held constant, then β_2, β_3, and β_4 represent the average effect sugars & proteins (including their additive effect) have on calories.

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) $y = \beta_0 + \beta_1 x_1 + \epsilon$ (error)

y: numeric variable for avg cost of tuition

x_1: % of full time instructional staff employed at the institution

Question: Is the % of full time instructional staff employed considered statistically significant when predicting average tuition?

$H_0: \hat{\beta}_1 = 0$

$H_A: \hat{\beta}_1 \neq 0$

b) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$

where $x_j = \begin{cases} 1, & \text{if observation in group} \\ 0, & \text{otherwise} \end{cases}$

y: numeric variable for average tuition cost

$x_1 \rightarrow$ liberal arts

$x_2 \rightarrow$ community

$x_3 \rightarrow$ technical/vocational

$x_4 \rightarrow$ affiliated

Question: Does the type of school (levels described) have an additional effect on the average cost of tuition at undergraduate-only institutions?

$H_0: \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = \hat{\beta}_4 = 0$

H_A: $\hat{\beta}_1 \neq \hat{\beta}_2 \neq \hat{\beta}_3 \neq \hat{\beta}_4 \neq 0$.

c) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

y: numeric variable for average tuition cost

x_1: % of full time instructional staff employed

x_2: categorical variable $\begin{cases} 1: \text{if public} \\ 0: \text{if private} \end{cases}$

x_3: $\begin{cases} 1: \text{if private} \\ 0: \text{if public} \end{cases}$

Question: Is the % of full time instructional staff employed and the type of school (private or public) statistically significant in predicting average cost of tuition?

$H_0: \hat{\beta}_1 = \hat{\beta}_2 = 0$ $H_A: \hat{\beta}_1 \neq \hat{\beta}_2 \neq 0$

* It is important to note that the variance assumption is questionable due to the residuals boxplot. Also, if β coefficient tests were completed, not all of the indicator levels should be considered significant

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.

8) Normality ?
Not very clear, but not enough data to necessarily disprove

→ a) Assumptions:

1) Zero Mean? ☑
2) Linearity? ☑
3) Randomness? ☑
A random single-serving sample was completed
4) Independence? ☑
Sample size appears to be relatively large; see effects confound for more explanation.
5) Effects constant? ☑
The treatment effect (a brand of cereal's impact on avg cals per serving) doesn't affect another treatment effect
6) Effects Additive? ☑
There are no interaction terms.
7) Constant Variance? ?
Questionable. Many box plots have vastly different spreads BUT centers are relatively similar.

b) $H_0: \mu_K = \mu_N = \mu_P = \mu_Q = \mu_R = \mu$

H_A $\mu_K \neq \mu_N \neq \mu_P \neq \mu_Q \neq \mu_R \neq \mu$

→ Are the group means different from the grand mean?

NOTATION:
- μ_K → group mean for Kellog
- μ_N → group mean for Nabisco
- μ_P → group mean for Post
- μ_Q → group mean for Quaker Oats
- μ_R → group mean for Ralston Purina
- μ → grand mean summarizing overal behavior of y

y is avg cals/serving

c) Based on the discussion on assumptions above, I'm relatively comfortable completing an inference test based on the hypotheses above. Since the overall p-value of 0.03 is less than the significance level of 0.05, we reject the null that all the group means are the same. The R^2 is ~0.16, meaning only 16% of the error in y is explained by 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).

Probably unrelated, but saw this:

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.

$0.77 + 0.34 = 0.43$

$\frac{0.43}{2} = 0.26$

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.

I don't believe that the two people's conclusions are necessarily contradictory, but they raise questions.

1) According to Person B, there is a negative correlation between Arsenic and distance from the well, but it's weaker than the positive correlation between Arsenic and the year.

2) According to person A, the predictors in the model only explain 26% of the error found in the response variable, amount of arsenic.

132) One would think that based on the higher correlation between Arsenic and year, a model using that term would have a higher R^2. However, it's possible that the while the variables are correlated, more information is needed to successfully predict Arsenic levels.

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.

Correlation is the strength of a relationship between two variables (for clarity!)

Cereal ANOVA Model