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: Raya Tuffaha

Swarthmore Username: rtuffah1

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

$$Y = \beta_0 + \underset{\text{sparrow}}{\beta_1 x_1} + \underset{\text{finch}}{\beta_2 x_2} + \underset{\text{age}}{\beta_3 x_3} + \underset{\text{sparrow age}}{\beta_4 x_1 x_3} + \underset{\text{finch age}}{\beta_5 x_2 x_3} + \epsilon,$$

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 - if predictors are collinear, that means there is some dependency & when one variable is removed, the coefficient of the dependent variable(s) will change

(b) If a regression model's first variable has a coefficient of $\widehat{\beta}_1 = 5.7$, 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 - it would increase by a factor of 5.7 units, not 5.7 units themselves

(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 of the means is different from the others

(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.

true

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 α will allow less room to mistakenly reject the null, which is a type one error. so it will decrease the chances

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

false - small sample sizes lend themselves to noticing smaller differences, and not every difference is statistically significant

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

false - measure of association between predictor & response

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.

Studentized residual plots are fitted on an additional level as compared to just the observed residuals – they balance extreme outliers or non-linearity. by using studentized residuals it becomes easier to identify distinct & likely statistically significant trends that could be misinterpreted in a normal resid plot.

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 uses the studentized residuals to calculate influential data points, which means it is starting from an already-fitted graph where the outliers have been slightly adjusted. This provides a clearer picture of what points are actually influential.

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?

ⓐ total df = $n - 1 = 76 - 1 = 75$

ⓑ ref level = 4 (5-1 predictors)

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?

$Y = 106.97 + \alpha_j + \epsilon$

the estimated group effect for Quaker is $Y = 106.97 + \alpha_{quaker} + \varepsilon$. because the Quaker intercept is -16.34, this means avg calories for Quaker will be 106.97 - (-16.34) higher than the avg for all samples.

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 interaction term between sugars and protein accounts for any dependence between those values beyond their individual impacts on the response variable. the coefficient on this interaction term shows the value of the numeric predictors on each other and is included to account for that relationship.

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).

ⓐ what is the relationship btwn full-time instructional staff employed & schools' tuitions below/above a certain value?

Percentage of full time $= \beta_0 + \beta_1$ (avg tuition) $+ \varepsilon$

H_0: there is no significant difference in % of full time staff based on avg tuition

ⓑ are institutions w/ affiliations more likely to have high tuition?

tuition $= \beta_0 + \beta_1 \overset{x_1}{(}$binary private/public$) + \beta_2 \overset{x_2}{(}$affiliation$) + \beta_3 x_1 x_2 + \varepsilon$

H_0: there is no significant difference in tuition cost regarding affiliation

ⓒ what type of school has the most full time staff?

% full time $= \beta_0 + \beta_1 \overset{x_1}{(}$category of school$) + \beta_2 \overset{x_2}{(}$affiliation$) + \beta_3 x_1 x_2 + \varepsilon$

H_0: there is no difference in % of full time staff between categories of school or affiliation.

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.

ⓐ normality → not met, the second plot shows significant outliers on either end of the graph. and the data curves around the line

constant group effects → not met, the spread of data in each boxplot is widely different although the means are at roughly similar values

ⓑ $H_0: \underset{G}{\beta_1} = \underset{K}{\beta_2} = \underset{N}{\beta_3} = \underset{P}{\beta_4} = \underset{Q}{\beta_5} = \underset{R}{\beta_6}$ there is no statistically significant difference between the average calories of each manufacturer

$H_a: \beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq \beta_5 \neq \beta_6$ there is a statistically significant difference between avg calories of each manufacturer
w/ at least one $\neq$

β = avg. cal.

ⓒ based on the page 10 graphs and my hypothesis, I can reject the null. the boxplots show inconstant variance and the quantile plot does not show normality, so I had to proceed with caution. yet despite the wide ranges of data spread between manufacturers, it was difficult to see a significant difference in means. but with a p-value of 0.02 and an F statistic of 2.7, I am inclined to believe there is a small but significant difference in average calories across cereal manufacturers

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:

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

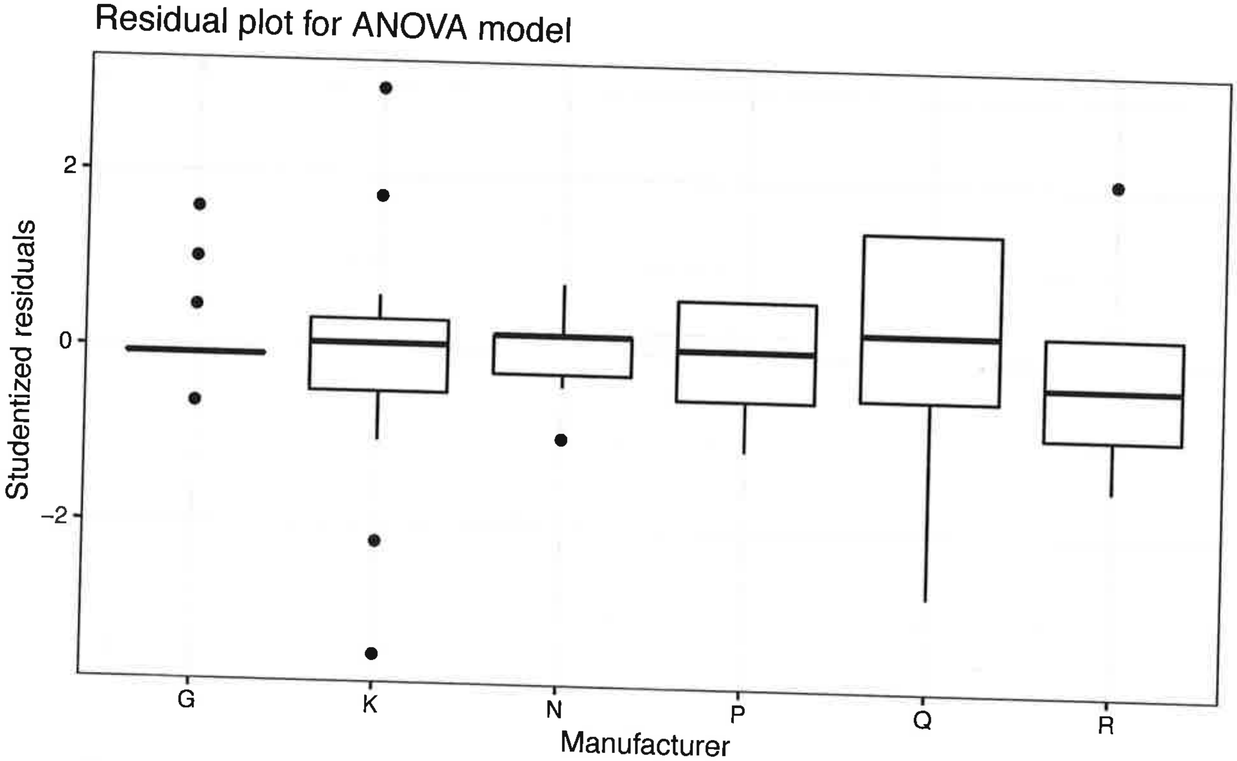
and estimates each with their sample correlations $r_1 = 0.77$ and $r_2 = -0.34$. Are the two people's conclusions contradictory? Explain your answer.

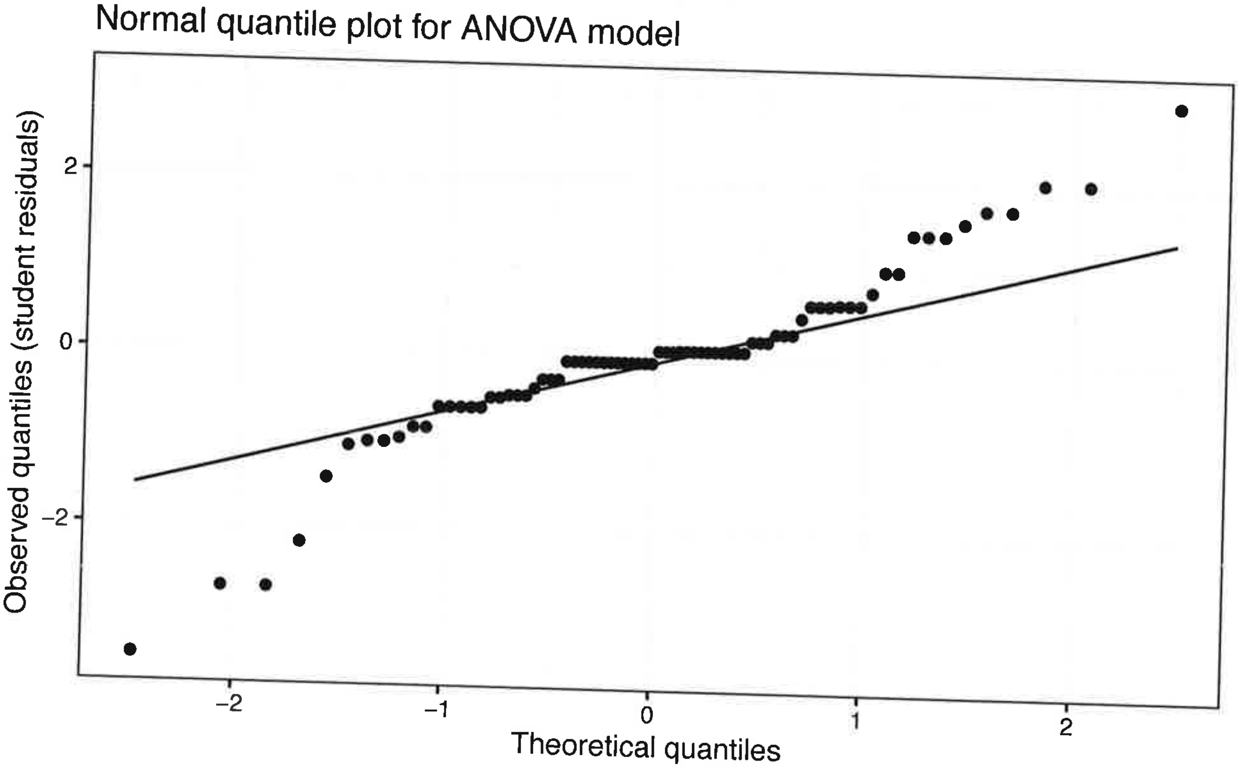
no, these are not contradictory answers. the correlations that Person B considers are more specific than the MLR Person A fits. this means that while the MLR shows the effects of Year & Miles on Arsenic together, the Correlation isolates Year effect on Arsenic and Miles effect on Arsenic separately. and, because r_1 has a fairly strong, positive value of 0.77 and r_2 has a weaker, negative value of -0.34, it makes sense that the MLR adj. R^2 is not very high but is a positive number, since both predictors have such different impacts on the response.

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


Residual plot for ANOVA model
Studentized residuals
G
K
N
P
Q
R
Manufacturer


Normal quantile plot for ANOVA model
Observed quantiles (student residuals)
Theoretical quantiles