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: Liam Mariscalco

Swarthmore Username: lmarisc1

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

2. C $H_0 : \beta_1 = 0$

3. D $H_0 : \beta_3 = 0$

4. B $H_0 : \beta_4 = \beta_5 = 0$

5. A $H_0 : \beta_1 = \beta_2 = 0$

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, as if there is multicollinearity and we remove a variable, the model may become more effective as the other variable's coefficient may change. It is not a certainty, so you cannot say that it will definitively have no impact

(b) Suppose a numerical variable x_1 has a coefficient of $\beta_1 = 2.5$ in the multiple regression model. Suppose also that the first observation has a value of $x_1 = 7.2$, the second observation has a value of $x_1 = 8.2$, and these two observations have the same values for all other predictors. The predicted value of the second observation will be 2.5 units higher than the prediction of the first observation based.

True

(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, as this only tells us that out of the 4 means, 2 are not equal

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

False, as if the means differ, the data differs

(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) If a given value (for example, the null hypothesized value of a parameter) is within a 95% confidence interval, it will also be within a 99% confidence interval.

False - it could be corrected by saying that if it was within a 99% CI, it would also be within a 95% CI

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

False, it could be corrected to say they may be statistically significant - it's not garenteed

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

True

Section 2: Short answer questions

5. (4 points)

State two reasons why we might consider transforming the response variable to fit an appropriate multiple linear regression model to some data.

One reason we would do this is if we had a high p-value and were unable to say that the model was statistically significant.

Another reason would be if we had bad confirmation of the conditions — i.e. Q-Q plot didn't show normality, residual vs fitted plot didn't show constant variance or linearity.

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 use leverage values, b/c this is the one that would most directly which data points have the ability to sway the fit of a MLR. This would also tell us which specific points were the ones in question, and show us the degree of their influence.

For questions 7-9 consider the following random sample of $n = 246$ online shoppers. We are going to model the average price (in US dollars) (`price`) as a linear function of the item's type (a categorical predictor with levels: trousers, skirts, blouses, on_sale). Below is the R summary output for this one-way ANOVA model.

price ~ type

```
##
## Call:
## lm(formula = price ~ type, data = retail_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.946  -8.946   0.893   6.054  35.054
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     41.946      1.512  27.750  < 2e-16 ***
## typeon_sale     -5.438      2.128  -2.555  0.01123 *
## typeskirts       9.161      2.138   4.285 2.64e-05 ***
## typetrousers     5.937      1.987   2.988  0.00309 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.31 on 242 degrees of freedom
## Multiple R-squared:  0.1913, Adjusted R-squared:  0.1813
## F-statistic: 19.09 on 3 and 242 DF,  p-value: 3.825e-11
```

7. (3 points)

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

(b) What is the reference level?

a. 242 degrees of freedom

b. $41.946 (i.e. intercept value)

8. (6 points)

Suppose the average price ~~number of plate appearances per game~~ is 44.63 over all 246 data points. What is the estimated group effect for clothing type trousers?

44.63 = 5.937 (Trousers) → $\frac{44.63}{5.937}$ = group effect for trousers

↳ rounds to 45 and 6

$\frac{45}{6}$ = 7.5 → group effect = ~7.5

$$\begin{array}{r} 7.5 \\ 6\overline{)45.00} \\ \underline{42\downarrow} \\ 3 \end{array}$$

9. (4 points)

Consider two additional numeric predictors: the amount of time the item has been available for purchase on this retailer's website, `release`, measured in weeks and the production cost associated with each item, `produce_cost`, measured in US dollars. If we were to fit a regression model including each of the three predictor variables (including `type`) 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 would tell us the impact on price that the combined effect of release date and production cost measures have. We want to create this interaction term b/c there is likely association between these 2 new variables and it would make the model more precise/effective.

Section 3: Long answer questions

10. (9 points)

Suppose you have access to a data set on a random sample of Swarthmore faculty. The variables included in this data set are a numeric variable for each person's age, a binary categorical variable distinguishing faculty who are tenured from those who are not, a numeric variable for each faculty member's starting salary, and a categorical variable indicating if the faculty member attended a liberal arts college, or a university, or entered the work force after graduating high school.

State a research question that can be answered with the overall F-test for each of the following models. 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.(SLR) How does the age of faculty members impact their starting salaries, and does their age have a statistically significant impact?

$Age = \beta_0 + \beta_1 \text{starting salary} + \varepsilon$ $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq$

b.(ANOVA) Do the tenure status and post-~~attend~~ HS graduation path of Swarthmore faculty have a statistically significant impact on their starting salary?

$\text{Start Salary} = \beta_0 + \beta_1 \text{Tenured} + \beta_2 \text{HS path} + \varepsilon$ $H_0: \mu_1 = \mu_2$ $H_A: \mu_1 \neq \mu_2$

~~Ho~~ For Tenured: =1 if Yes, =0 if No   For HS path: 1 = college, 0 = if otherwise

c.(MLR) Do the age, tenure status, and post HS grad plans have a statistically significant impact on starting salary?

$\text{StartSalary} = \beta_0 + \beta_1 \text{Tenured} + \beta_2 \text{Age} + \beta_3 \text{HS}_{\text{path}} + \varepsilon$ $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

For Tenured: =1 if yes, =0 if no   For HS path: =1 if college, =0 if otherwise $H_A: \beta_1 \neq \beta_2 \neq \beta_3 \neq 0$

11. (8 points)

Consider the ANOVA model for the retail 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 cost of the purchased items are significantly different for different types of clothing type. (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 know that our conditions for ANOVA are that the effects are constant/additive, and that there is a zero mean, linearity, follows a normal distribution, and there is constant variance. Based on the R output, we can see that all 3 types of residuals are significant at an alpha=.05 level given their low p-values. We can also use the Q-Q plot, which shows us that there is normality despite a few outliers on either end. Using the residual plot, we can see that there are constant group affects with their means all being very close to each other.

b. Our null hypothesis is that the group means for each type of clothing are all equal and that there is no difference in group means. Our alternative hypothesis is that there is a difference in the group means.

Say trousers = group 1, shirts = group 2, blouses = group 3, on-sale = group 4.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$, $H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \rightarrow \mu =$ group averages

c. Given the model's low p value, we can conclude that the model is affective and reject our null hypothesis in favor of the alternative that there is a difference in group means. We see that each type are also all statistically significant at alpha=.05 and confirm that they have a statistically significant impact. Our QQ plot shows there is normality and that residual plots show the group effects are constant. I would wanted to use a residual vs fitted plot to check for linearity and constant variance. The only concerning thing I noticed was a low R^2/adjusted R^2 value, showing that not a lot of the variability in the model is explained by the predictors, but this can be fixed by adding more variables.

at alpha = .05

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:

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

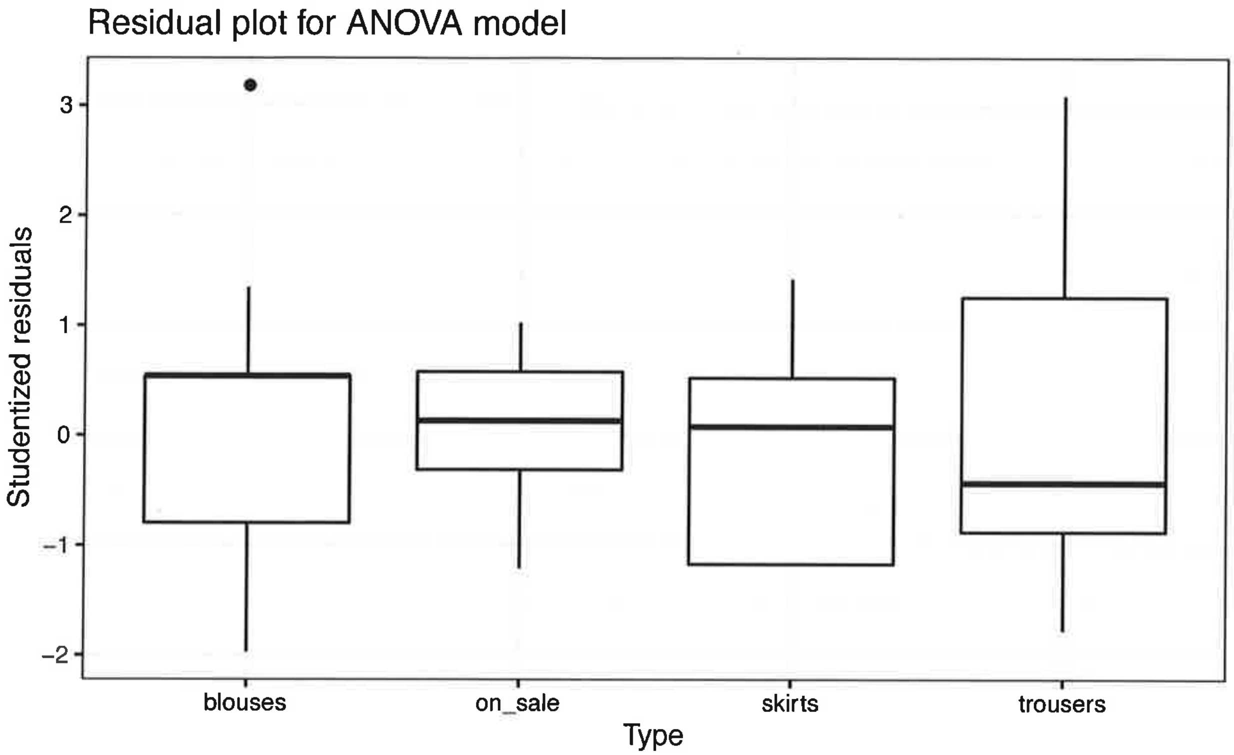
and estimates each with their sample correlations $r_1 = 0.77$ and $r_2 = -0.34$. Are the two people's conclusions contradictory? Explain your answer.

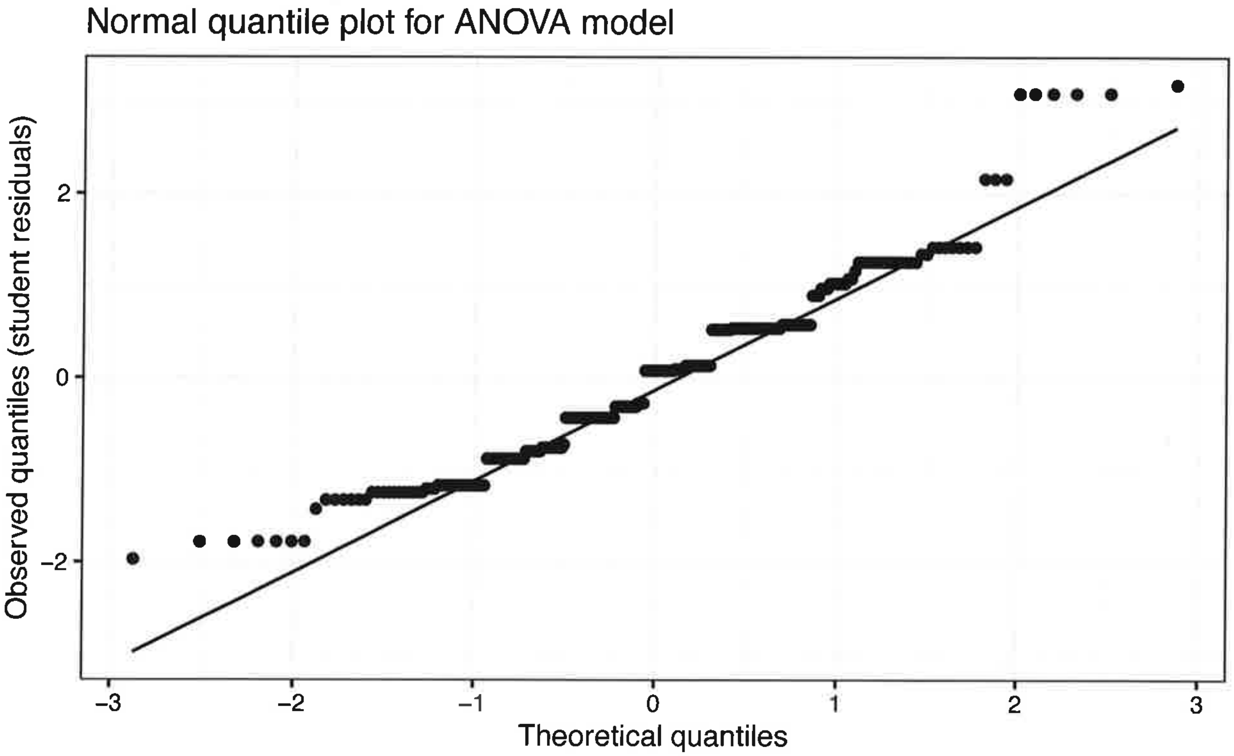
These two conclusions aren't necessarily contradictory. Our original r^2 shows us that only 26% of the variability in the model is explained by the two predictors. W/ the 1st correlation, we see that much more variability in Arsenic levels is explained by the year it was collected in comparison to both the Year and Miles model. The second correlation shows that even less variability is explained by Miles alone in comparison to Years and the model w/ both. They aren't contradictory, but show that a model w/ just a predictor variable of Years is more effective and has more variability explained than the other two models. As person A has an adjusted R^2 value, it shows us that adding Miles to the model w/ Years alone did not make that model more effective.

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.

Retail ANOVA Model





“