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: Elaine Kim

Swarthmore Username: ekim8 (ID#: 90221991)

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)

Weight ~ Type
↑
3 levels.

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? $x_3 = 0$

b) Does the effect of age on a bird's weight depend on what type of bird it is? x_3

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? $x_1 = 0, x_2 = 0$ $Y = \beta_0 + \beta_3 x_3 + \epsilon$

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 $H_0 : \beta_1 = \beta_2 = 0$

c) sparrow

$Y = \beta_0 + \beta_1 + \beta_2 x_3 + \beta_4 x_3 + \epsilon$

pigeon

$Y = \beta_0 + \beta_3 x_3 + \epsilon$

$0 = \beta_1 + \beta_4 x_3$

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. Even if there is no multicollinearity, removing one variable will lead to an adjustment to the values of the other coefficients. Removing one variable could also lead to a change in the R^2 value.

(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. We can only conclude that at least one of the means are significantly different from the rest.

(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, this information will be provided with the ANOVA F-test. The post-hoc pairwise analysis will tell you about the differences in each pair and will identify if they are each significant or not.

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.

True

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

False. This is a possibility, we cannot be certain that it will always be the case.

Corrected: With large sample sizes, even small differences between the null and observed point estimate can be identified as statistically significant.

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

False

Correction: correlation is a measure of the association between a predictor variable and a response variable.

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.

① If we found strong evidence against one or more of the assumptions, and a transformation helped fix this, then we may consider transforming the response variable. For example, if we saw some curvature in the residual plot, we may want to consider transforming the response variable

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?

(#5 continued)

② If transforming the response variable allowed for less potential influential points (observed through leverage values, standardized residuals, Cook's Distance, etc), then we may also want to transform the response variable

#6.

I would use Cook's distance values, because it takes into consideration both studentized residuals and leverage values. So, you get information about potential influential points for both predictor and response variables. You only get information about predictors with leverage values, and only information about response values with studentized residuals

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.

```
##
## 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. The total df is 246 − 1 = 245 degrees, and there is 4 − 1 degrees of df for the predictor terms. So, the remaining error df is 245 − 3 = 242 degrees.

(b) The reference level is 41.946 US dollars.

8. (6 points)

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

The average price for trousers = 41.946 − (5.438)(0) + (9.161)(0) + (5.937)(1) US dollars

The estimated group effect for clothing type trousers would be 41.946 + (5.937)(1) − 44.63 US dollars.

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.)

When allowing for change, and when holding the "type" constant, the average price for the item increases by the amount of the coefficient for the interaction term when release * produce_cost increases by 1.

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) One model could be $(\text{starting salary}) = \beta_0 + \beta_1(\text{age}) + \varepsilon$ We can answer if there is a statistically significant linear relationship between age and starting salary. The null hypothesis would be $H_0: \beta_1 = 0$

(b) One model could be $(\text{starting salary}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ $X_1 = \begin{cases} 1, & \text{if liberal arts} \\ 0, & \text{otherwise} \end{cases}$ $X_2 = \begin{cases} 1, & \text{if university} \\ 0, & \text{otherwise} \end{cases}$

We can answer if the mean starting salary for all "school" types are all the same, or if at least one is statistically different.

$H_0: \mu_1 = \mu_2 = \mu_3$ (1: liberal arts, 2: university, 3: high school)

(c) (on next page)

linearity — zero mean
randomness — normality
independence
constant variance

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.

(continuing #10)

(c) One model could be (starting salary) $= \beta_0 + \beta_1(\text{age}) + \beta_2 X_2 + \varepsilon$

$X_2 = \begin{cases} 1, & \text{if tenured} \\ 0, & \text{otherwise} \end{cases}$

We can use the F-test to answer if there is at least one statistically significant linear relationship between the response variable and predictor term.

$H_0: \beta_1 = \beta_2 = 0$

⊕ The additive effect condition would be assessed by seeing if there is any interaction between the groups.

#11.

(a) For constant effect, we would want the variance within each group to be similar. The sizes of the box plots are quite different, so I would say that there is some evidence against this assumption. The normality condition also appears to be violated, as the median lines for blouses, skirts, and trousers are skewed. The independence and randomness assumption would have to be assessed by looking at how the data was obtained. Although, I don't see how the price of one type of clothing would affect the other, so I presume that the indepence condition will probably be met.

(b) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_0: At least one $\mu_i \neq \mu_j$ $(i, j = 1, 2, 3, 4)$

μ is the average price, 1 is for on-sale, 2 is for skirts, 3 is for trousers, 4 is for blouses.

(c) on the back of page 9 !! ☆☆

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:

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

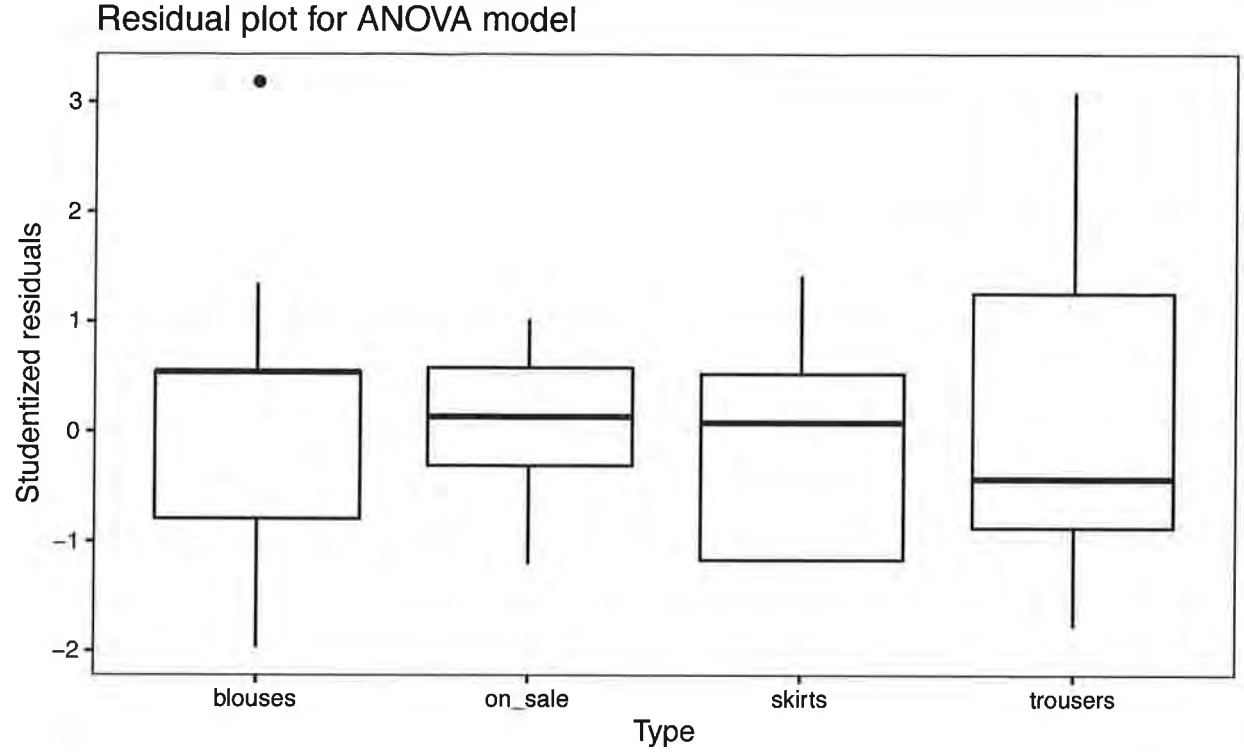
and estimates each with their sample correlations $r_1 = 0.77$ and $r_2 = -0.34$. Are the two people's conclusions contradictory? Explain your answer.

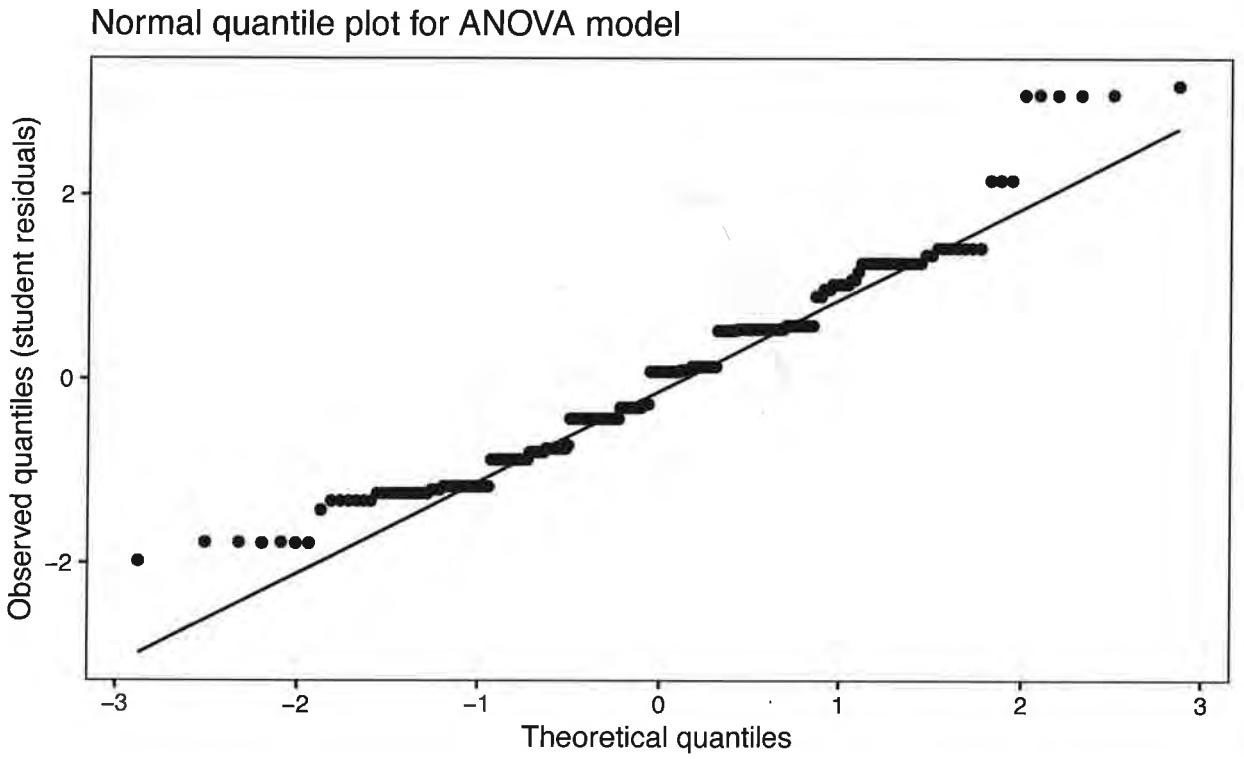
No, they are not contradictory. The adjusted R^2 value tells how well the model (which includes all predictor terms) does at explaining the variability with the response values. So, when Year and Miles were used together, they were able to explain about 26% of variability. Person B's data shows that Year has a stronger correlation with Arsenic than Miles does. The correlations account for individual predictor terms, not them as a group (like adj. R^2 did). Based on this data, I might next suggest excluding Miles from the model and seeing how the adjusted R^2 value changes.

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





"

Problem 11 (c).

From the one-way ANOVA output, we can see that the p-value is very close to 0, at $3.825e^{-11}$. So, we can conclude that we have statistically significant evidence to reject the null hypothesis in favor of the alternative. So, we could conclude that we have statistical evidence that there may be a difference in average price between at least two clothing types. However, one thing to note is that in part (a), we identified possible assumptions that may have not been met. So, I think it is difficult to say that this ANOVA F test result is completely reliable. Also, we can see that both the R^2 and R^2_{adj} values are quite low, indicating that our model does not do a good job of explaining the variability in "price" outcome. Based on these observations, we may want to go back and revise our model before making any conclusions.