902194628

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: Anika Rajamani

Swarthmore Username: arajamal

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.

No, the variables might not be closely related to each other, but it will still have an effect on the coefficient when one is removed.

(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.

7.2 × 2.5 = 18.00 (144 + 360)

8.2 × 2.5 = 20.50 (164 + 410)

True, because $\beta_1 x_1$ when $x_1 = 7.2$ is 18 and $\beta_1 x_1$ is 20.5 when $x_1 = 8.2$. If the values of the other predictors are the same, this 2.5 difference will be maintained

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

This statement is 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.

No, we can only conclude that at least one group has a different group mean. We only know that at least one is different. We do not know which ones

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

This statement is true.

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

No, we can do some post-hoc analysis such as Fisher's LSD to figure out which pairs have different means.

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.

No, 99% CI is more restrictive, so values near the bounds of the 95% CI will not be in the 99% CI

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

This statement is true → less variability in models with larger sample size.

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

Yes, correlation looks at how much the variables depend on each other, so it measures linear association.

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.

essentially, we look to transform when certain conditions are not met

We might consider a transformation if we see that the linearity condition is violated in the residuals. If there is an exponential trend, etc., we might need to transform the response to induce a linear relationship. We might also transform the response if the normality condition has been violated.

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 to identify potentially influential data points because it factors in both leverage & studentized results. Thus, since it takes both of these into account, it is a better holistic measure to use to identify influential points.

We might also transform the response to amplify the relationship between predictors & response if the β values are really really small and difficult to interpret.

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? $n-k-1 = 246-3-1 = 242$

(b) What is the reference level?

The reference level is typeon_sale.

8. (6 points)

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

$Y = \mu + \alpha_1 + \varepsilon$

$44.63 - (9.161 - 5.438)$

$44.63 - 3.723$

40.907

9.161
-5.438
3.723

44.630
$-\ 3.723$
40.907

5

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 coefficient of this interaction term would tell us how much price is jointly affected by simultaneous changes in release and produce_cost

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).

c) Does a faculty member's age and tenure status affect starting salary?

$\hat{Salary} = \beta_0 + \beta_1 Age + \beta_2 Tenure_{Yes} + \beta_3 Tenure_{No}$

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

H_a: at least one of $\beta_1, \beta_2, \beta_3 \neq 0$.

a) Does a faculty member's age affect their starting salary?

$\hat{Salary} = \beta_0 + \beta_1 \cdot Age$

$H_0: \beta_1 = 0$, $H_a: \beta_1 \neq 0$

b) Does a faculty member's tenure status affect their starting salary?

$Salary = \mu + \alpha_j + \epsilon$ where $\alpha_j = \begin{cases} 1 & \text{if tenured} \\ 2 & \text{not tenured} \end{cases}$

H_0: there is no difference in group effects between tenured & not tenured faculty members

looking at the box plots, my best educated guess is that there is some difference between on-sale & trousers or blouses & trousers.

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) To conduct this test, we need to check the constant variance condition. We can do this by looking at the box plots and seeing if the ranges for the different boxes are about aligned. There is some concern here about the constant variance condition because trousers, for example, has a very high upper tail as compared to on-sale. We also must assume randomness in data collection. We can check normality using the normal quantile plot and say that there is some concern of tailing in both the upper & lower tails. Constant group effects would mean that the size of each box in the box plots is about the same (one is not exponentially larger than the other).

$X_1 = \begin{cases} 1 & \text{on-sale} \\ 0 & \text{not} \end{cases}$ $X_2 = \begin{cases} 1 & \text{skirts} \\ 0 & \text{not} \end{cases}$ $X_3 = \begin{cases} 1 & \text{trousers} \\ 0 & \text{not} \end{cases}$

b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

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

H_a: at least one $\beta_i \neq 0$

c) Since our F-statistic has a p-value of 3.825e-11 in the ANOVA output, we can reject the null in favor of the alternative hypothesis in b. This just tells us that at least one group has a different group mean. In order to figure out which groups are significantly different from one another, we perform a Fisher's LSD on pairs of groups.

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:

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

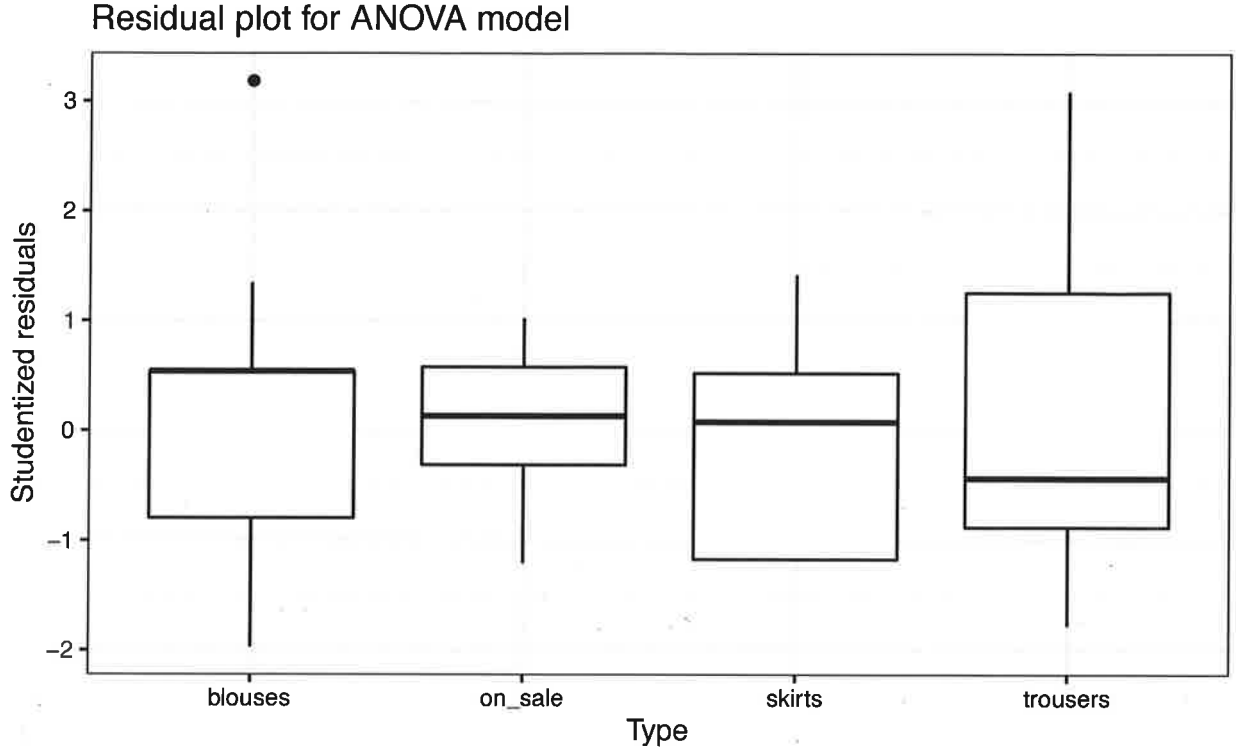
and estimates each with their sample correlations $r_1 = 0.77$ and $r_2 = -0.34$. Are the two people's conclusions contradictory? Explain your answer.

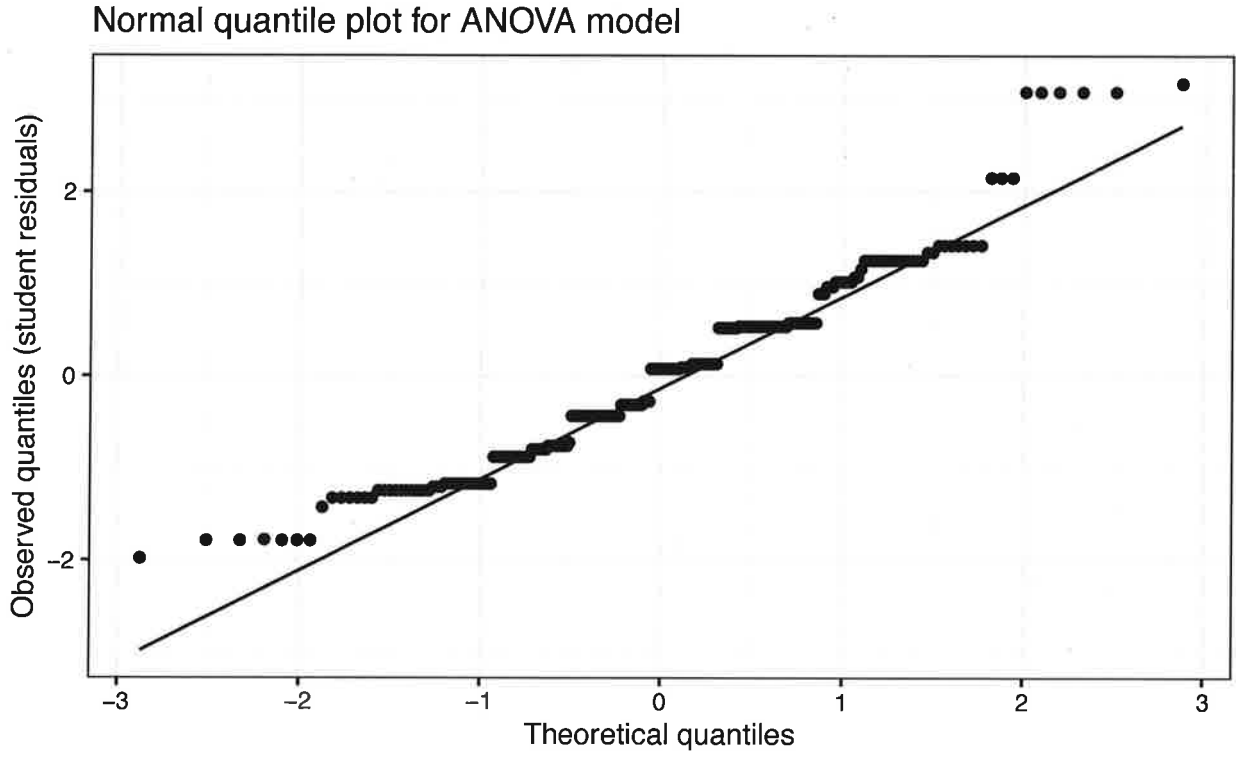
These people's conclusions are not necessarily contradictory. Due to the fact that Person B found a fairly strong positive relationship between Arsenic and year and weaker negative relationship between Arsenic and Miles, it is possible that R^2 did increase but adj R^2 decreased due to the addition of a conflicting, new predictor. However, it is also possible that these are indeed contradictory and that one or both correlations should be closer to 0 if adj R^2 is so small

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





"