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: Tyrique Arthur

Swarthmore Username: tarthur1

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: Removing one variable could reduce the predictive power of the 2nd variable; It likely would affect the coefficient

2.5(x1) 7.2 (1) 8.7 (2)

(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.

K=3 N=40 N=50

3/36/39 3/246/49

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 some of the means are different; not all

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

I know that it is when the variability ~~among~~ between the groups is higher than the variability within the groups. Assuming the statment is saying than then True.

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

True; if multiple group it will do an overall comparison between them

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.

H0: 20 CI = 18-22

True; as you increase CI level it simply increases the interval of the bounds; as such a 99% CI would likely still catch the H0 value

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

False; we would more likely see this with smaller sample size with 68, 95, 99.7% rule we would expect in the larger sample for most of the data to lie with 1 SD. It would then be unlikely for that small difference to be significant.

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

TRUE in MLR this is a measure of how the predictor variables correlate to the response variable. ~~It however can be interpreted~~ as

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.

- The relationship between x and y variables are not linear and perhaps curved
- Transforming perhaps with log might allow for a better fit reducing variability.

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?

only predictor

leverage

St Residuals

I would likely choose cooks distance as it incorporates both leverage and St Residuals which might increase accuracy

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
```

246 7

7. (3 points)

(a) What are the error degrees of freedom based on this model? DF of error $n - k - 1$

$246 - 1 - 1 = 244$

(b) What is the reference level?

The reference level is blouses

8. (6 points) price of each item

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?

~~The estimated group effect is 41.946 - 44.63~~ (group mean = grand mean)

$y = \mu + \alpha_T$

$y = 44.63 + 5.937$

estimated group effect = 44.63 + 5.937

β_i (model x gas)

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.)

Holding other variables constant, it ~~states~~ represents the average change in price assuming days available (release) is related to the production cost (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.

Tenure

= History

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) Is a person's age a significant predictor of starting salary

$Start\_Salary = B_0 + B_1(age)$

$H_0: B_1 = 0$

(b) Do the various level of History have a constant effect on starting salary

$Start\_Salary = B_0 + B_1(liberal) + B_2(university)$ 3 levels

$H_0: \alpha_1 = \alpha_2 = 0$ Workforce is reference

(c) Is a person's age, History, & tenure significant predictor of starting salary

$Starting\ salary = B_0 + B_1(age) + B_2(liberal) + B_3(university) + B_4(Tenure)$ Tenure = 1 O/W = 0

$H_0: B_1 = B_2 = B_3 = B_4 = 0$

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) Independ and random : Depend on design of experiment but they stated they took a "random sample" so these assumptions are likely met

Residuals are normally distributed : From the output on pg 9, for the most part it seems to be normally distributed except in perhaps skirt where it seems skewed Additionally trousers and maybe blouses could be argued for but as of now I can't say they have not been met (for these two)

equal variance : we might be able to argue equal variance between on_sale and skirt but relative to the other groups they are different. As such its likely that the equal variance has not been met.

Constant : The group effect of each level of clothes type is always the same (MET)

α_1 = Blouse
α_2 = on sale
α_3 = skirts
α_4 = trousers

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ (All treatment effects are equal to zero)

H_a : At least 1 $\alpha_i \neq 0$ (At least 1 treatment effect is $\neq 0$)

(C) As at least 1 of the necessary assumption has not been met / seems unlikely to have been met, I can't confirm the validity of this model. In the anova normal plot output although the point don't deviate completely from the line of normality we see some strange patterns; additionally at the tails we see skewness on both sides. For the Anova boxplot we also observed at least the equal variance assumption being violated.

NOTE: If however we disregard the fact that some assumptions may not have been met, from the R output on pg 5 we observe a F statistic of 19.09 and p value = $3.825e^{-11}$. From this we would reject the H_0 in favor of the H_a that at least one level of clothes has some effect on price.

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:

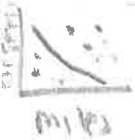
$$Cor(Arsenic, Year) = \rho_1; \quad Cor(Arsenic, Miles) = \rho_2$$

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.

For person B their correlation simply describes how the 2 variables are related (ex positive and linear). Person's A coefficient of determination is a measure of of correlation but specifically describes how much of the variability is explained in the model. It would be difficult to say whether they correlate or not because they are on a way measuring different things.

NOTE: If person's A fit has a positive coefficient for miles then it would however contradict with person B who had a - correlation.

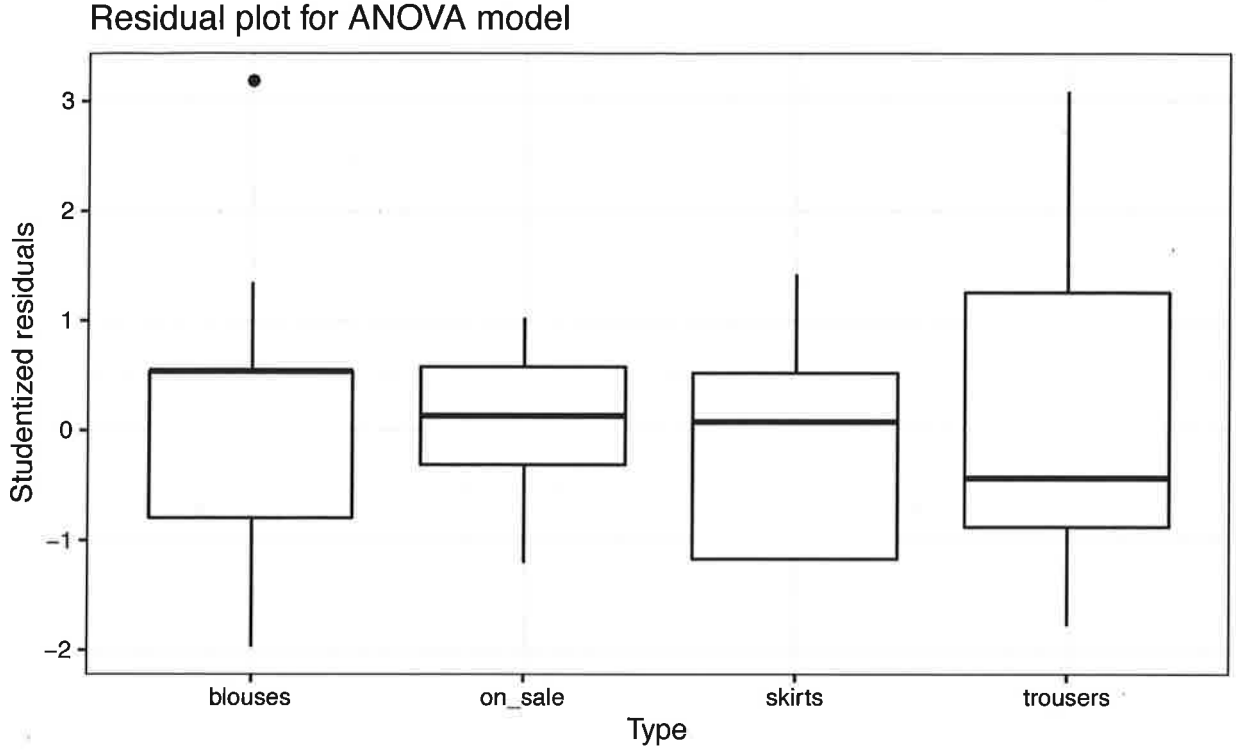
Also looking again we might be able to argue that they don't contradict. we see person A has a low Adj value and person B has a low correlation for miles. This low correlation could perhaps explain the high SSE in the model

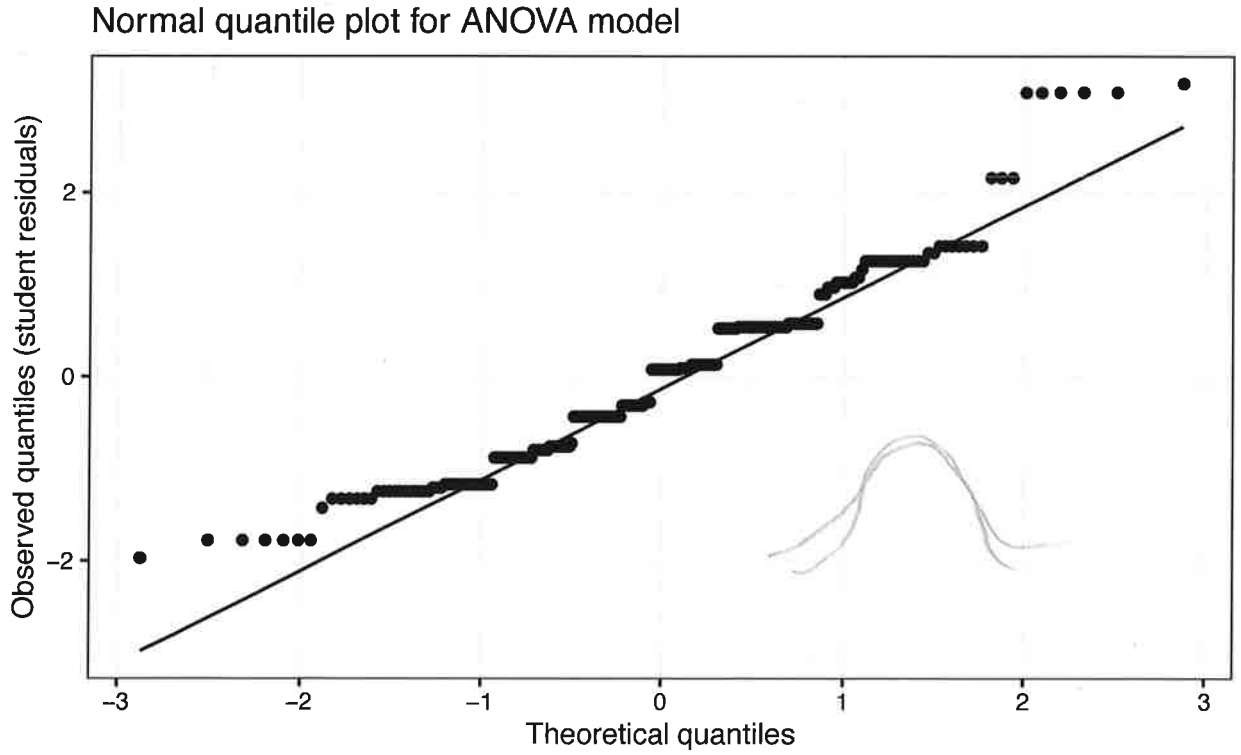


Section 4: Extra credit opportunity

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





"