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: Sumin Byun

Swarthmore Username: sbyun1

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

$$\text{weight (kg)} (Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon,$$

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.

If the predictors are collinear, it means those predictors are correlated, which means that one variable effects on the response variable depends on another variable. Therefore, if we remove one variable, how the other collinearied variable affects on the response variable will also be influenced, having an influence on the point-estimate of another variable's coefficient.

(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 cannot conclude if "all" the means are different from one another, because when we reject the H_0 in favor of H_A, H_A is that at least one pair of means are significantly different, not the all of the means.

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

(c) A post-hoc pairwise analysis will identify if there is at least one pair of means that are significantly different. False.
We measure "if" there is at least one pair of means that are significantly different by overall F-test, and after this, we do post-hoc pairwise analysis to see "which" pair of means is significantly different.

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 can not be guaranteed that a given value within a 95% confidence interval will also be within a 99% confidence interval all the time.

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

(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.

① When we make normal quantile plot to assess the normality assumption in multiple regression model, and, if the data points are not following the line linearly, such as having a curved pattern, then we can consider transforming the response variable (e.g. logarithm on y variable) to make the normal quantile plot looks more appropriate, making the normality assumption more appropriate.

② Also, in ANOVA, if the constant variance assumption is not met (when we look at the boxplot), we can collect more data and transform the response variable in order to make the constant variance assumption more appropriate.

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 values to identify potentially influential data points. Studentized residuals and leverage values are also a good measure of identifying potentially influential data points, but cook's distance considers both studentized residuals and leverage values, so if I could only use one measure, I would choose Cook's distance. Studentized residuals identify the potentially influential data points regarding the response variable, while leverage identify them regarding the predictor variables; but cook's distance identify them regarding both response and predictor variables

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.

m = 4

```
##
## 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
```

k = m-1 = 3

n = 246

Price = $\beta_0 + \beta_1$ Sale + β_2 Skirts + β_3 Trousers

7. (3 points)

(a) What are the error degrees of freedom based on this model? Since this model has 4 levels of item's type categorical predictor, k = m-1 = 3

Error degrees of freedom is n-k-1, which is 246-3-1 = 242.

(b) What is the reference level? blouses.

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?

ave price of each item

Estimated group effect for clothing type trousers is 5.937 related to 44.63

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 model would be "$\text{Price} = \beta_0 + \beta_1(\text{Type\_onsale}) + \beta_2(\text{Type\_skirts}) + \beta_3(\text{Type\_trousers}) + \beta_4(\text{release}) + \beta_5(\text{produce\_cost}) + \beta_6(\text{release})(\text{produce\_cost})$." Here, the interaction term coefficient β_6 indicates, given fixed values for other predictor variables (holding them constant), how both release and produce_cost, as one predictor term, will affect on the average price.

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.

→ 2 num. var. & 2 cat. var

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; → $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \varepsilon$, where Y is the numerical response variable for 'starting salary.' and X_1 is a numeric variable for each person's 'age'.

(b) an ANOVA model; $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \varepsilon$, where Y is numerical response variable for 'starting salary' and X_1 is the 'liberal arts college' level group of high school-attending categorical variable (X_1: 1 (liberal arts) / 0 (otherwise)) and X_2 is the 'University' level' group (X_2: 1 (University) / 0 (otherwise)) of the same categorical predictor; reference level of this variable is the group entered the work force after graduating high school.

(c) a multiple linear regression model (not SLR or ANOVA).

$$\hat{Y}_{\text{Salary}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_1 X_4 + \varepsilon,$$

where Y is the numerical response variable for 'starting salary', X_1 is the numerical predictor variable for 'age', X_2 is the 'liberal arts college' group of the categorical predictor variable of after-graduating attending, X_3 is that of the 'university' group), X_4 is the 'are tenured' group of the tenured binary categorical variable.

X_2: 1 (liberal arts colleges) / 0 (o/w)

X_3: 1 (university) / 0 (o/w)

X_4: 1 ("is tenured") / 0 (o/w)

And X_1X_4 is the interaction term of the tenured categorical variable and age numerical variable.

⇒ For all three models, $\hat{\beta}$(betas) are the estimated coefficients of each predictor term and β_0 is the intercept. Also, All three models are the estimated equations, so they are using $\hat{\beta}$ and $\hat{Y}$, ... notations, with the error term ε at the end. All $\hat{\beta}$(betas) are estimated values in these models.

<The research question for..>

① SLR model (a) is that if the age of Swarthmore faculty has a linear relationship with their starting salaries.
$H_0: \beta_1 = 0$

② ANOVA model (b) is that if the average starting salary of Swat faculty differs among the groups of whether attended a liberal arts college / university / or workforce after graduating.
$H_0: \beta_1 = \beta_2 = 0$

③ MLR model (c) is that if the starting salary of Swarthmore faculty has a correlation with their age, after-high-school-attending, or whether tenured or not
$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

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.

normality (✓)
constant variance (✓)
zero mean (✓)
independence
linearity (✓)
randomness (✓)

(a) In ANOVA, linearity assumption is the constant group effect assumption, which means the effect doesn't vary from the group over the other group. When estimating the effect size, we consider both estimated coefficient and standard error of that beta coefficient. In this ANOVA model, this assumption might not be met because the intercept coefficient, which is the reference group (item type_blouses), is a lot bigger compared to other three estimated coefficients, considering standard error of all four of them. For the normality assumption, normal quantile plot shows a slightly curved (V-shaped) pattern, rather than following the line linearly. This can be the red flag of normality assumption as well. We don't need to check zero mean assumption necessary in ANOVA, and we assume that the data is collected randomly. For the constant variance assumption, we look at the box plot, and it indicates that constant variance assumption might not be satisfied since the IQR range and especially the range of the four boxes are not constant: 'Trousers' group has a larger range compared to the other three. However, at the same time, this might not be too bad for constant variance to be met. Lastly, we can assume that the independence of error between and within groups.

(b) $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \varepsilon$, where $X_1 \begin{cases} 1 \text{ (on\_sale)} \\ 0 \text{ (o/w)} \end{cases}$, $X_2 \begin{cases} 1 \text{ (skirts)} \\ 0 \text{ (o/w)} \end{cases}$, $X_3 \begin{cases} 1 \text{ (trousers)} \\ 0 \text{ (o/w)} \end{cases}$, and error term is denoted as ε here, where Y is the estimated average price, $\hat{\beta}$'s are the estimated coefficients for each predictor term and $\hat{\beta}_0$ is the estimated intercept.

($\hat{\beta}_1$: 'on_sale' group; $\hat{\beta}_2$: 'skirts' group; $\hat{\beta}_3$: 'trousers' group; indicator variables of item's types.)

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
The average cost of the purchased items are not significantly different among different types of clothing type.

H_A: at least one $\beta \neq 0$.
The average cost of the purchased items is significantly different at least for one type of clothing type.

(c) Since p-value is less than 0.05 ($3.825e-11$), we can reject the null hypothesis in favor of H_A. It is a strong evidence that the average cost of the purchased item is different for at least one type of clothing type. However, since not all assumptions are met, this conclusion could not be valid. Also, adj. R^2 value is not that high (0.18), indicating that this model might not be a good model. for the further studies, we could transform the response variable or collect more data.

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:

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

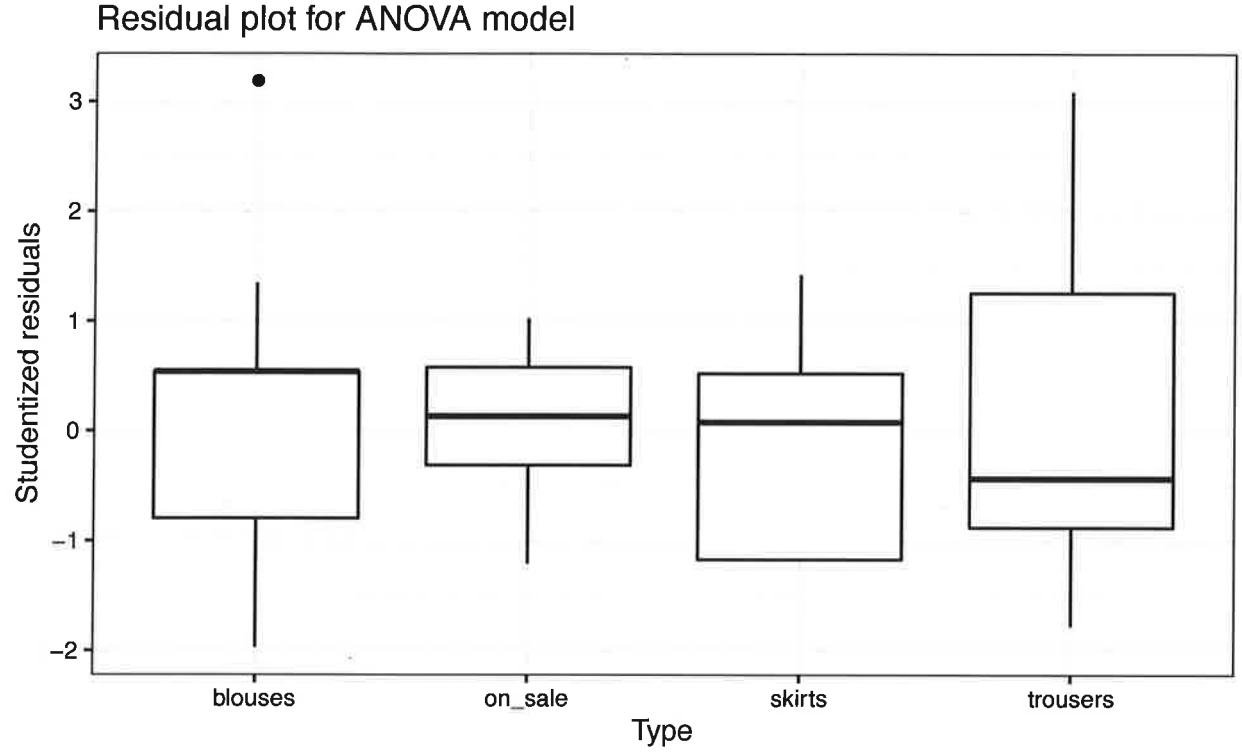
and estimates each with their sample correlations $r_1 = 0.77$ and $r_2 = -0.34$. Are the two people's conclusions contradictory? Explain your answer.

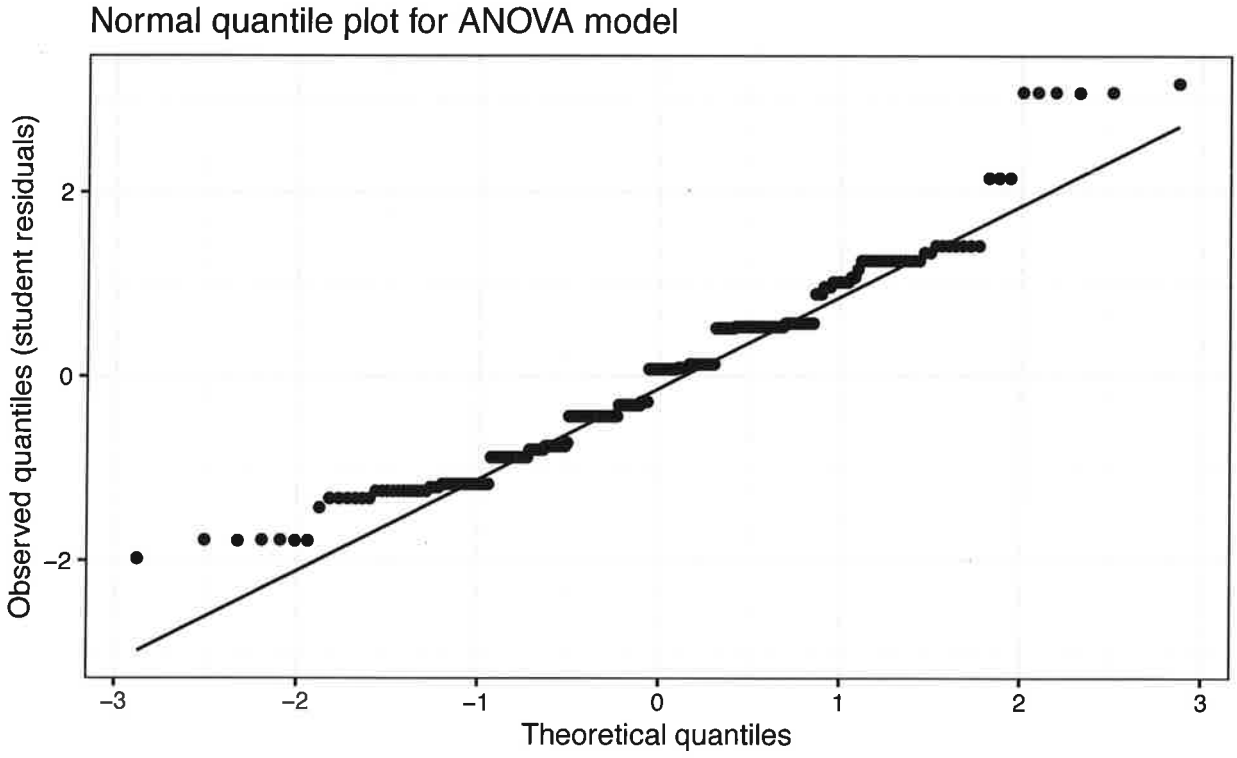
Two people's conclusions are not necessarily contradictory. Based on person B, it shows that 'Year' and 'Arsenic' have high correlation but not in 'Miles' and 'Arsenic' relationship. That might be why the person A's R^2 value is higher that person B's correlation for Arsenic & Miles but lower than Arsenic & Year.

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





"