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: Isabelle Courtney

Swarthmore Username: icourtn1

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?

$\beta_3 x_3$

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 $H_0 : \beta_1 = \beta_2 = 0$

$\beta_0 + \beta_1 x_1 + \beta_2 x_2$

nested F-test

sparrows: $\beta_1 x_1^{(1)} + \beta_3 x_3^{(1)} + \beta_4 x_1^{(1)} x_3^{(1)}$

pigeons: $\beta_3 x_3^{(1)}$

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. When one of the predictors is removed, the estimates for the other variable's coefficient will change in order to account for the effect the removed predictor had on the response. However, if they are perfectly collinear, the overall R^2 / p-value will not change

$\beta_1 x_1 + \ldots + \beta_4 x_1 x_2$

(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.

False, if there are interaction terms that include x_1, even if the values of the other predictors are the same, the difference in the predicted values would not be 2.5

(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, the overall F-test results only mean that at least one of the group means is different from the rest, not all of them

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

dfe / SSE, SSmod / dfMod

True, the F-statistic = $\frac{MSmod}{MSE}$ and the higher the F-statistic the more likely we are to reject the null.

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

False, it will identify which pairs of means are significantly different. To do just any pair of means, you can do an overall F-test.

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, a 99% confidence interval is bigger because there is more certainty that the value is in that range.

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

True, as n increases, SE decreases because $SE = \frac{\sigma}{\sqrt{n}}$ or similar. Then, to get a t-statistic, you divide $(\bar{x} - x_0)$ by SE. The smaller SE, the larger t-statistic, so the p-value will be smaller + more likely to be < 0.05,

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

False, it is a measure of association between two numeric variables.

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.

- If the transformed model meets the assumptions for MLR better, such as constant variance of errors
- If the original model has a p-value that is low but not significant, a transformed model could have a significant p-value

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, because for any data point, it combines the standardized residual and leverage value for that point. So, it is informative about both how unusual a data point is in its response value and in its predictor values.

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? 242

246 − 3 − 1 =

(b) What is the reference level?

type blouses

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?

$\alpha_t = \mu_t - \mu$ $\quad \mu_t = 41.946 + 5.937$

$= 41.946 + 5.937 - 44.63$

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.)

It shows how much a change in the value of one of the numeric predictors leads to a change in the effect of the other numeric predictor on the response. If that $\beta = 0$, then the interaction of those predictors does not change how those predictors effect the response.

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) Does a faculty member's age have a linear relationship with starting salary? model: $salary = \beta_0 + \beta_1 age + \varepsilon$

$H_0: \beta_1 = 0$

b) Is there a difference in the average age of tenured vs non tenured faculty?

model: $age = \beta_0 + \beta_1 x_1 + \varepsilon$ $\quad x_1 = \begin{cases} 1 & \text{if tenured} \\ 0 & \text{not tenured} \end{cases}$

$H_0: \beta_1 = 0$

c) Is there a linear relationship between starting salary and age and post-high school education they received?

model: $salary = \beta_0 + \beta_1 age + \beta_2 x_2 + \beta_3 x_3 + \beta_4 age\, x_2 + \beta_5 age\, x_3 + \varepsilon$

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ $\quad x_2 = \begin{cases} 1 & \text{university} \\ 0 & \text{else} \end{cases}$ $\quad x_3 = \begin{cases} 1 & \text{workforce} \\ 0 & \text{else} \end{cases}$

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) similar error variance across groups:
-for the group effects to be constant, the residuals should be similarly distributed across the groups: The residual boxplot shows that the error ranges for the groups are different, especially Trousers, and the medians of the residuals are not centered at zero for all groups: trousers is below, and blouses is above. Since being on-sale decreases price, we should check if the proportion of clothes on sale is the same for blouses, shirts, and trousers.
Normality of errors: the normal quantile plot shows that overall the errors are mostly normally distributed. Since there is only some deviation at the tails. But the boxplots show that the distribution of errors for the trousers group is skewed, which is not Normal.

b) $H_0: \mu_b = \mu_o = \mu_s = \mu_t$ where μ_b = mean of blouse prices
H_a: some $\mu_? \neq \mu_b$ μ_s = mean of shirt prices
μ_t = mean of trouser prices
μ_o = mean of on sale items prices

H_0: the average cost of items does not differ between the groups
H_a: at least one of the groups has a mean cost different from the rest

c) the p-value of the overall F-statistic is almost 0. So, there is strong statistical evidence that at least one of the group means is different from the rest. This does not indicate which of the group means are different from each other, just that at least one is. However, there are problems with the assumptions required for this inference. The boxplot of standardized residuals indicate that they do not vary similarly and Normally across each group.

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:

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

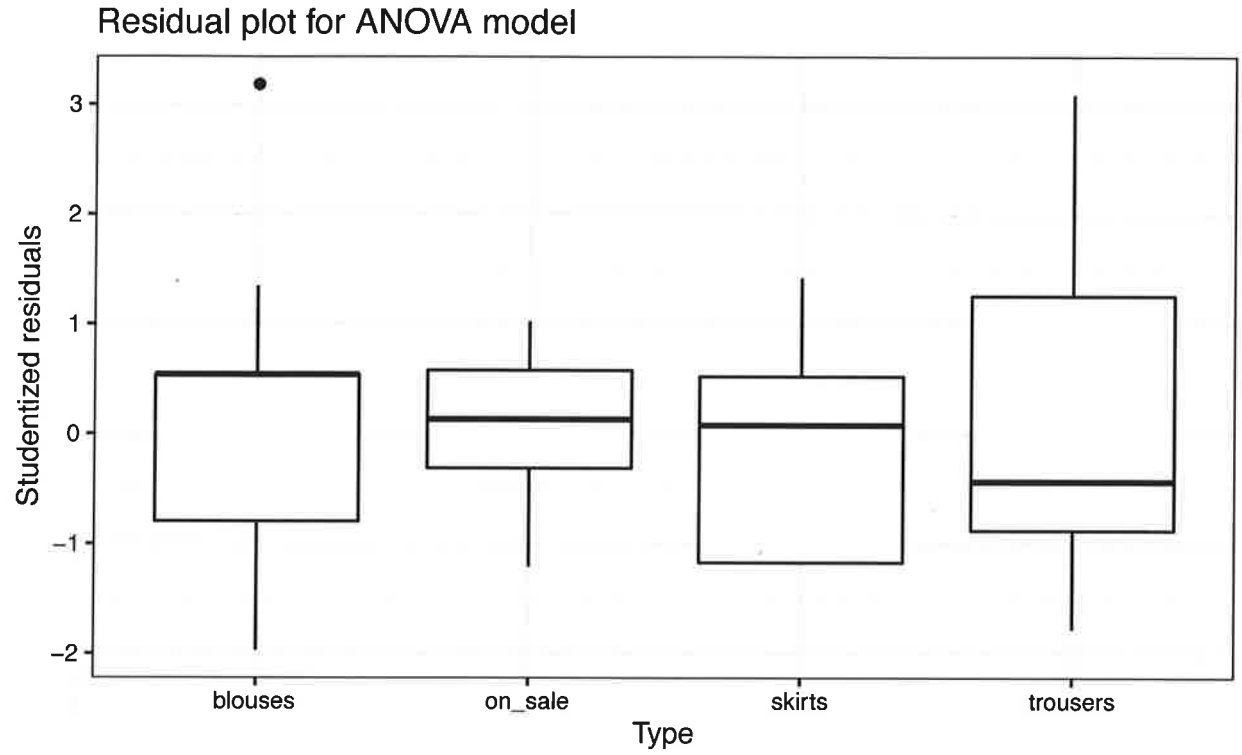
and estimates each with their sample correlations $r_1 = 0.77$ and $r_2 = -0.34$. Are the two people's conclusions contradictory? Explain your answer.

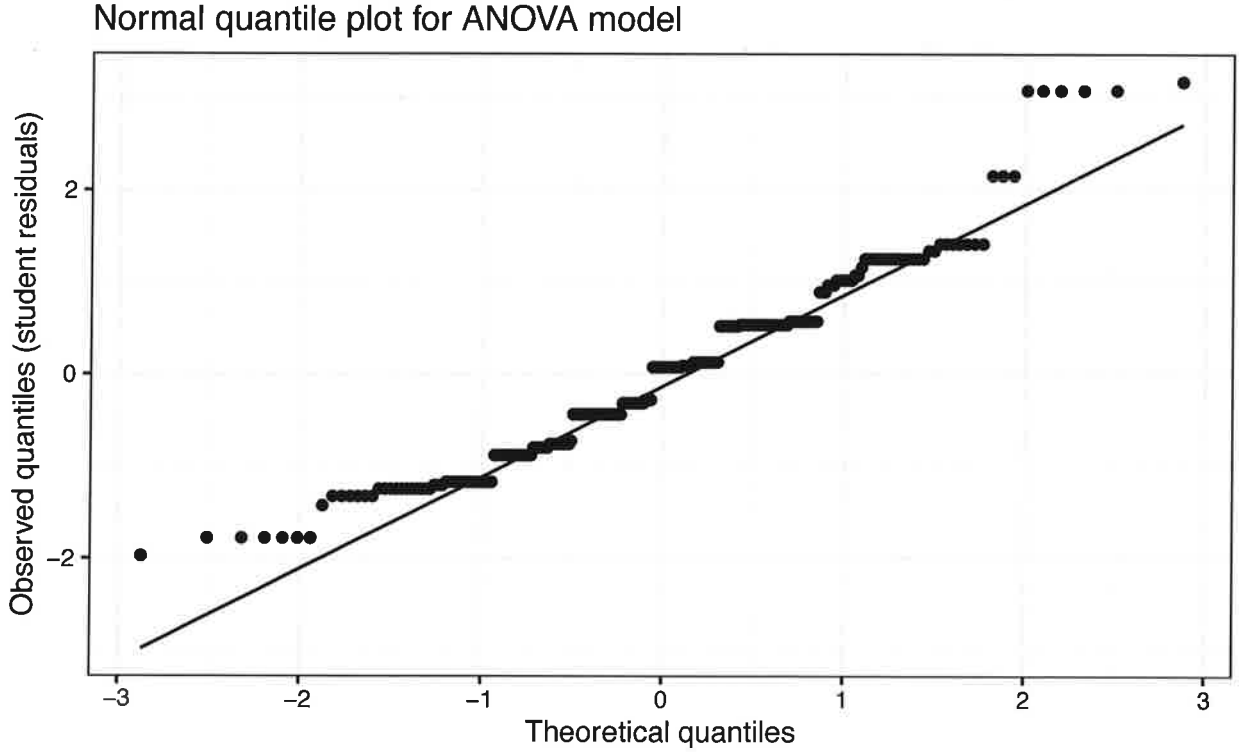
These conclusions are contradictory. Based on the correlation btw arsenic and year, an SLR with those variables would have an R^2 of $(0.77)^2 \approx 0.5$ to 0.6. So, the year explains at least half of the variability in arsenic levels. Doing the same for arsenic and miles, that $R^2 = (-0.34)^2 \approx 0.1$. So, the distance to a mining site explains about 1/10 of the variability in arsenic. Even if the variation based on miles is included in the variation explained by year, the total R^2 should not be lower than 0.5. Although adjusted R^2 penalizes for adding non-significant predictors, it should not decrease by that much.

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





"