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: Mih Nunta-aree

Swarthmore Username: pnuntaa1

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 1 2 3

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon,$$

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.

True: if say $x_1 = \tilde{\beta}_0 + \tilde{\beta}_1 x_2$ when $\tilde{\beta}_i$ is a constant we can not add any ~~explation~~ explanation to variability in $y = \beta_0 + \beta_1 x_1 + \epsilon$ even if we add x_2 in → $\tilde{\beta}_0$ & $\tilde{\beta}_{12}$ will be merged with β_0 & β_1 while ϵ remain the same.

(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 : let $\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$ $\therefore$ $\Delta\hat{y} = \beta_1 \Delta x_1 + \beta_2 \Delta x_2 + \dots \beta_k \Delta x_k$ for these 2 pairs
$= \beta_1(8.2 - 7.2) + \beta_2(0) + \dots + \beta_k(0) = 2.5 \times 1 = 2.5$

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

True : $Df_{resid} = Df_{total} - Df_{model} = n - k - 1$ $\therefore$ if n increase
$Df_{residual}$ increase

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 migh get 3 statistically indistinguisable means & 1 statistically different one

(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.

True :

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 : because 99%-CI will ... 95%-CI & added margins

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

True : $SE \approx \frac{s}{\sqrt{n}}$ → If n is very larg SE approach 0

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

True : +1 for total positive linear relation ship
-1 for total negative linear relation shp

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.

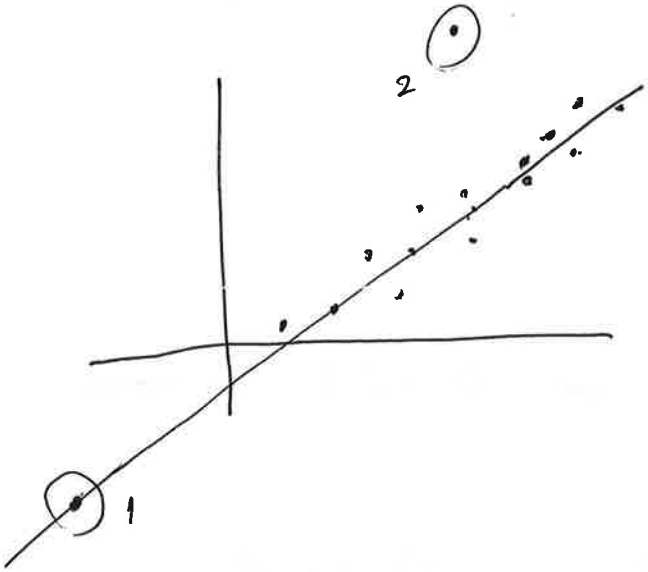
1. The data doesn't meet linearity assumption

2. The data doesn't meet constant variance assumption

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?

Personally, I would use Cook's distance because it already accounted for leverage and residuals. If we choose the other two we might notice some point that is uncommon but ~~does~~ they do not affect the model that much. For example, ② might have unusual residual & ① might have unusual leverage, but they don't affect the model that much.



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.

1 2 3 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
```

7. (3 points)

(a) What are the error degrees of freedom based on this model?

(b) What is the reference level?

a) $n = 246 \quad k = 3 \quad \therefore Df_{error} = 246 - 3 - 1 = 242$ ✱

b) $x_1 = \begin{cases} 1 & \text{trousers} \\ 0 & \text{otherwise} \end{cases} \quad x_2 = \begin{cases} 1 & \text{skirts} \\ 0 & \text{otherwise} \end{cases} \quad x_3 = \begin{cases} 1 & \text{trousers} \\ 0 & \text{otherwise} \end{cases}$ ∴ reference level is blouses ✱

8. (6 points)

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

$\alpha_{trouser} = \mu_{trouser} - \mu_{grand\ mean}$

$\approx (41.95 + 5.94) - (44.63)$

1.31

≈ 3.26 USD

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.)

release prod_cost

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 (X_4 * X_5) + \epsilon$ time

The coefficient explain how production cost affect the rate of change in price of the products.

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. ① ② ③

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 is on a different page.

a) Does age affect the starting salary of a faculty?

Starting Salary $= \beta_0 + \beta_1$ Age $+ \epsilon$

Null hypothesis: $\beta_1 = 0$ → Age doesn't affect starting salary of a faculty.

b) Does action after highschool affect a faculty's starting salary?

Starting Salary $= \beta_0 + \beta_1 I_{liberal} + \beta_2 I_{uni} + \epsilon$

Null hypothesis: $\beta_1 = \beta_2 = 0$ → the starting salary is not affected by faculty's action aft... highschool.

11. (8 points) ☆ C is on a different page.

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) There are 6 assumption necessary for doing F-test (Inference)

1) Zero mean
2) Constant variance
3) Linearity
4) Normality
5) Independence
6) Randomness

The term constant variance in this context means that the variability within each group is the same (at least similar). From what we can see in the box plot, it looks like this assumption is not met. The inter quartile range of trouser is a lot bigger than on-sale. Therefore, F-test might not be okay for this test.
The othe assumption (normality) is met though. The q-q plot of residuals look pretty normal, so there should nit be a problem.
We can't do anything about independance & randomness, but we have to assume that they are random & n is not too big to make the sample dependent.

b) $price = \beta_0 + \beta_1 I_{on\text{-}sale} + \beta_2 I_{skirt} + \beta_3 I_{trouser} + \epsilon$

The test will see if $\beta_1 = \beta_2 = \beta_3 = 0$ or not. The null hypothesis is the mean of every group is the same or $\beta_1 = \beta_2 = \beta_3 = 0$.
β_1 is group difference of on-sale & blouse
β_2 is group difference of skirt & blouse.
β_3 is group difference of trouser & blouse.

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.

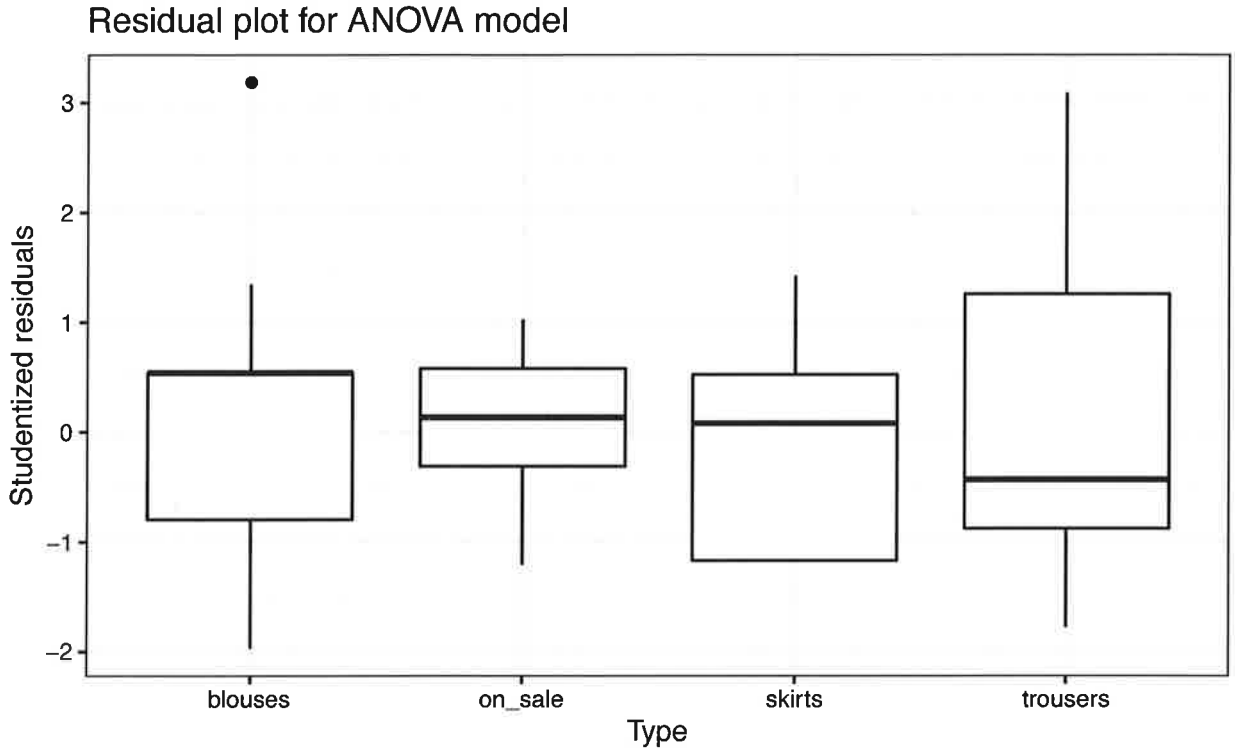
No, because ~~the correlation~~ adjust - R squared already adjust for some colinearity between predictor variables.

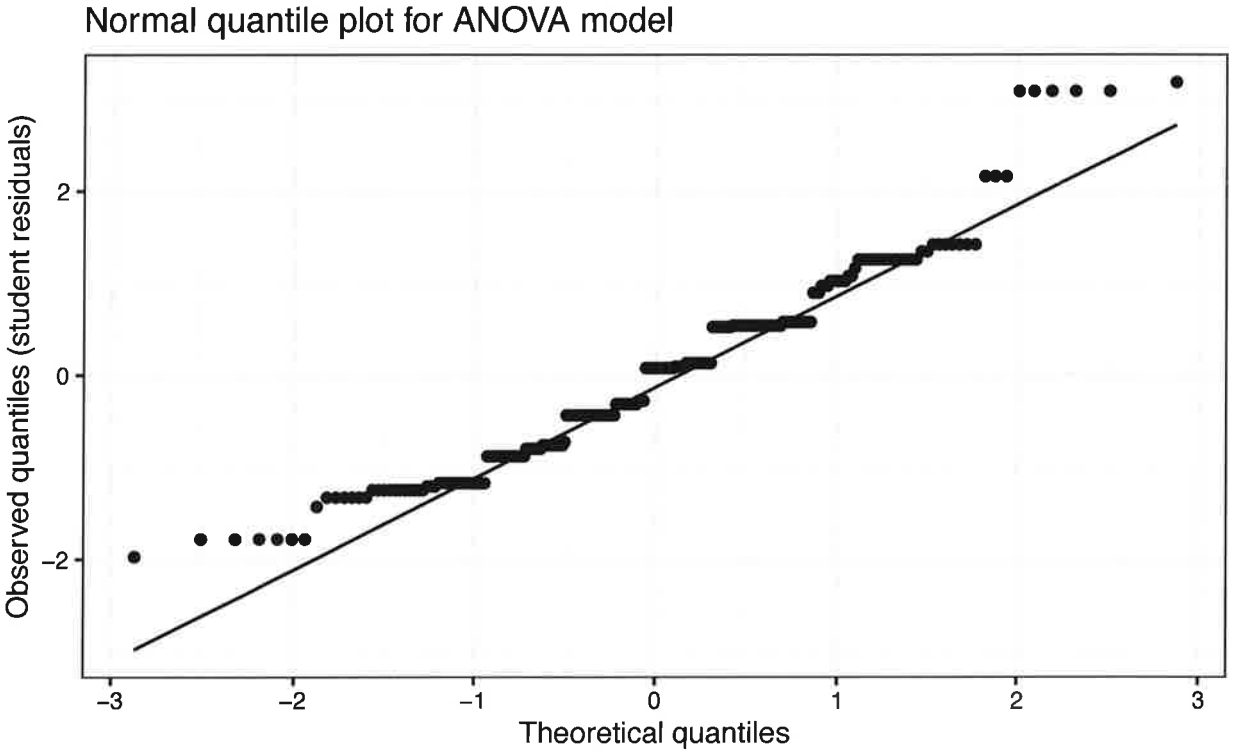
Therefore, $r_1 = 0.77^2 > R^2_{adjust} = 0.26$ is possible if Year & Miles are ~~strongly~~ linearly corelated. But if unadjusted R^2 of the model is less than r_1^2 or r_2^2 this will indicate that their finding are contradicting each other which is not the case in this problem.

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





"

Min Nunta-aree
pnuntaa1

10) C) Does starting salary have any relationship with age or being tenured?

$$\text{Starting Salary} = \beta_0 + \beta_1 \text{Age} + \beta_2 I_{\text{tenured}} + \epsilon$$

$H_0: \beta_1 = \beta_2 = 0$

→ Starting salary doesn't have linear relationship with either age or being tenured

---

11) C) Assuming the test is reliable (condition are met which from a) it likely isn't met), the F-test suggest that the probability of we getting sample this extreme due to random chance is $\sim 10^{-11}$ which is essentially 0.

This is lower than a predetermined $\alpha = 0.05$ (significant level 5%).

Therefore, it suggest we should reject the null hypothesis in favor of alternative hypothesis that the mean of some group is different.

However, like I said in the begining the box plot suggest that constant variance assumption is not met, making this test unreliable. This going along with just eyeing the box plot without using any test too because without using any test it is not clear that one or more group's mean is different from others'.

Thus, we might consider not trusting the test & still not reject the null hypothesis.