jmiao 1

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: Joanne Miao

Swarthmore Username: jmiao 1

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

$$Y = \beta_0 + \overset{\text{sparrow}}{\beta_1 x_1} + \overset{\text{finch}}{\beta_2 x_2} + \overset{\text{age}}{\beta_3 x_3} + \overset{\text{sparrow age}}{\beta_4 x_1 x_3} + \overset{\text{finch age}}{\beta_5 x_2 x_3} + \epsilon,$$

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

(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, ANOVA model & overall F-test only tells you if there is at least 2 means that are different from one another.

(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.

False, the analysis will identify 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, the more specific/narrow the interval, the more we can be confident that the true value is within the interval, so the values in a 99% CI will be within the 95% CI.

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

True, greater sample size increases degrees of freedom and decreases the p-value, leading to statistical significance even for small differences

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

False, correlation is a measure of the linear association between 2 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.

The residuals plot show that linearity assumption is not met (ex. the residuals curve downward as fitted values increase). The normal quantile plot shows that normality assumption is not met. Transforming the response variable might improve the residuals /normal quantile plot.

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 use Cook's distance values. Studentized residuals inform us how far away the data point is from the fitted line vertically (tells us about response variable). Leverage values tell us how far away the data point is from other data points horizontally (tells us about the predictor variable values). Cook's distance tells us about both; the Cook's distance equation contains leverage and standardized residuals.

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
##
```

Y= 41.946 - 5.438 Sale + 9.161 skirts + 5.937 trousers

```
## 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 degrees of freedom

(b) What is the reference level?

Reference level is blouses. It does not have its own coefficient.

8. (6 points)

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

5.937 is the estimated grp effect for trousers

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.)

$Y = \beta_0 + \beta_1 \text{ release} + \beta_2 \text{ pde-cost} + \beta_3 \text{ type} + \underline{\beta_4} \text{ release} * \text{pde-cost}$

Even when holding all other predictors constant, the effect of a 1 unit increase of release on price would depend on produce-cost (via interaction term coefficient) and the same is true for produce-cost.

For example, if we increase release by 1 unit, while holding all other predictors constant, the average increase in price is $\underset{\text{release coefficient}}{\beta_1} + \underset{\text{interaction term coefficient}}{\beta_4}$.

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. Is faculty age correlated with starting salary?

starting salary $= \beta_0 + \beta_1$ age

$H_0: \beta_1 = 0$

B. Does education type affect faculty starting salary?

liberal arts college, university, worked after graduating high school

starting salary $= \underset{\text{grand mean}}{\mu} + \alpha_{\text{liberal arts}} + \alpha_{\text{university}} + \alpha_{\text{high school}}$

$H_0: \alpha_{\text{liberal arts}} = \alpha_{\text{university}} = \alpha_{\text{high school}} = 0$

C. Does age and tenure status affect starting salary?

starting salary $= \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ tenure}$ — tenure $\begin{cases} 1, \text{ if faculty is tenured} \\ 0, \text{ if otherwise} \end{cases}$

$H_0: \beta_1 = \beta_2 = 0$

Constant, additive ✓

normal

indep + random

similar var

Y - zero mean

X - linearity

- constant spread

✓ - indep

- normal

✓ - random

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. Constant spread: met, ✓ though boxplots show that some clothing type have greater spread than other types, the maximum standard error is not greater than twice the minimum standard error

Independence and randomness: met, we are told it is a random sample

normality: not met, the normal quantile plot shows skewing at both ends and boxplots show that only on-sale have a roughly normal distribution of residuals

For the grp effect to be constant, changes in clothing type should lead to an additive change in avg price.

B. $H_0: \alpha_{sale} = \alpha_{skirts} = \alpha_{trousers} = 0$ Null hypothesis is that all group effects of clothing type is equal to zero.

H_A: some $\alpha_i \neq 0$ Alternative hypothesis is that there is at least one group effect that is not zero.

C. We reject the null hypothesis in favor of the alternative hypothesis because the p-value for the overall F-test is less than alpha value 0.05. The individual t-tests for each coefficient is statistically significant as well. We would like to conclude that clothing type does affect price, but unfortunately, the tests are not reliable as not all the necessary assumptions were met. Furthermore, the adjusted R^2 value of 0.1813 is quite small; very little of the variability in price is explained by the variability in the model. The main conclusion is that we need to fit a new model that meets all assumptions.

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, the two people's conclusions are not contradictory.

~~We do not know the coefficients of the MLR. If β_1 is positive, holding miles constant, an increase in Year would lead to an average increase in arsenic. This corresponds with the positive r_1. Similarly, if β_2 is negative, the interpretation of the MLR would correspond to r_2.~~

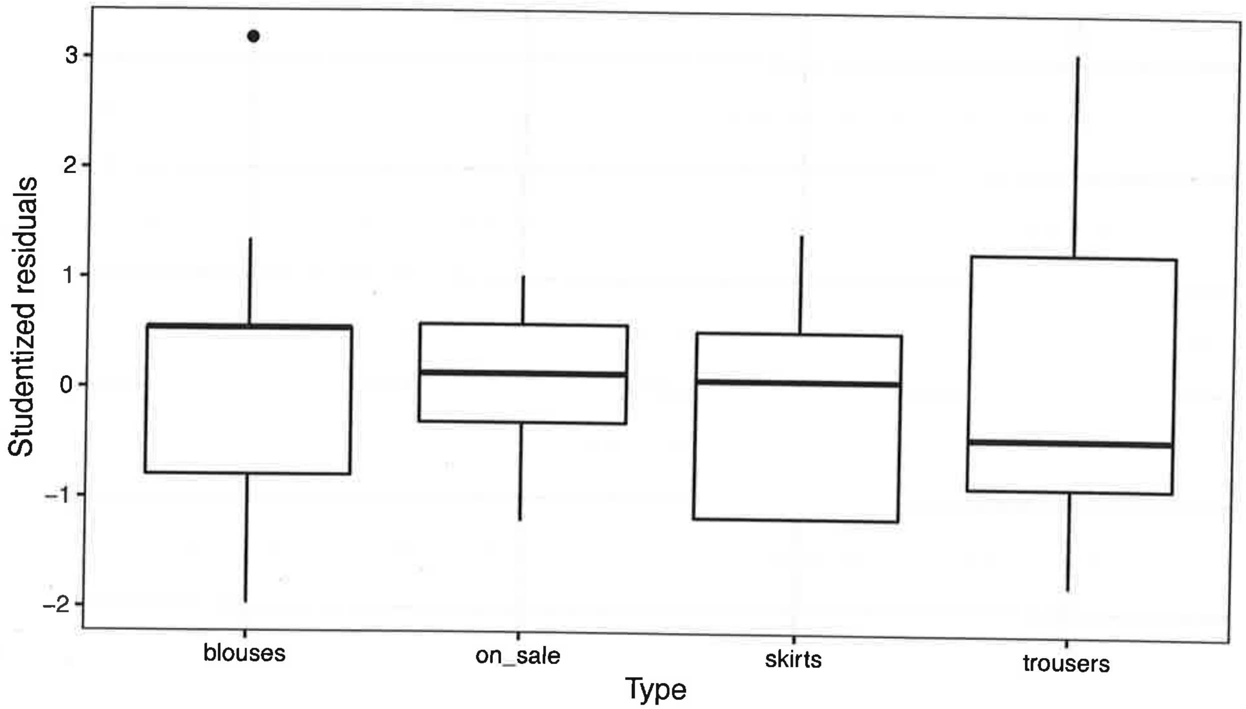
An adjusted R^2 of 0.26 is rather small and we would likely want to explore other models. The r_1 correlation of 0.77 is rather big while r_2 is much smaller; this suggests that ndividually, year is a better predictor of areenic than miles. Person 1 made a model with both year and miles (which predict aresenic not as well as year) so it is reasonable that the correlation strength (estimated from squarerooting adjusted R^2) is in btw $|r_1|$ and $|r_2|$.

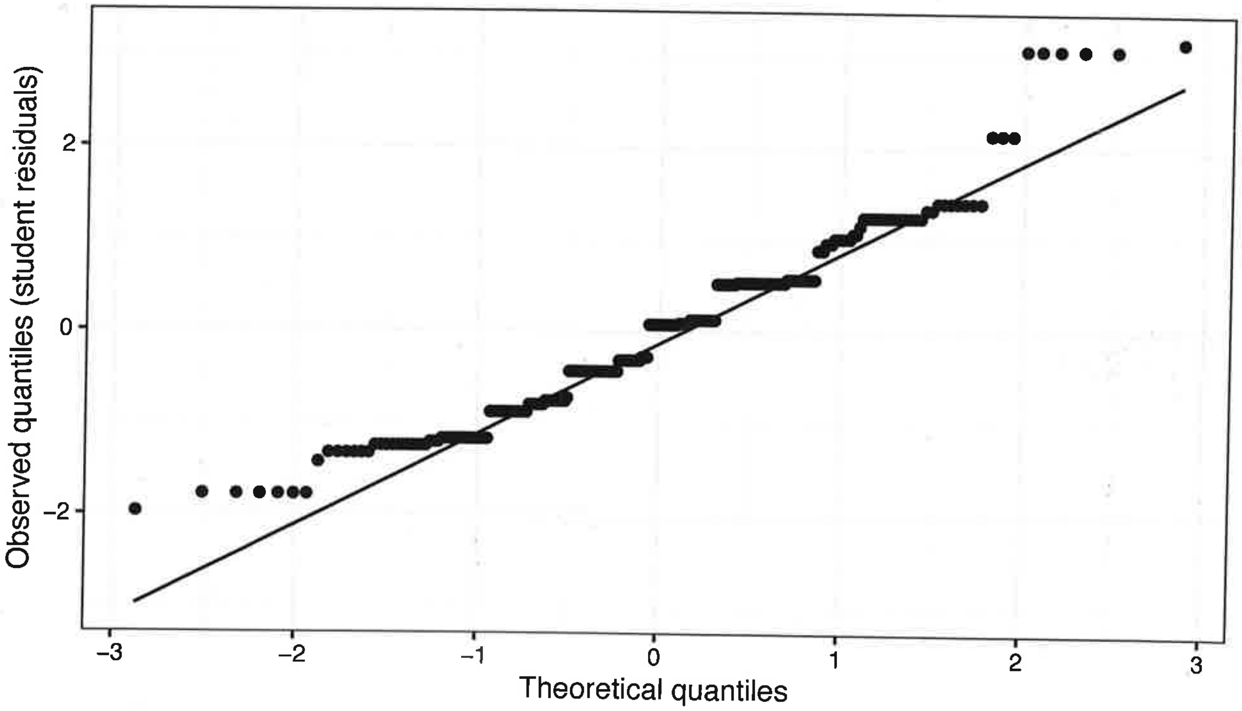
The 2 people's conclusions complement rather than contradict.

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


Residual plot for ANOVA model


Normal quantile plot for ANOVA model

“