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: Shikha Shrestha

Swarthmore Username: dshrest2 (90220375)

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. Everytime we add or remove a variable, no matter how correlated the predictors are, the coefficients of other variables are likely to change even by a small amount. Due to just random chance, adding a variable reduces the SSE because it would explain some relationship between the response and the predictor variable. Hence the coefficients will also change.

2.5(8.2) 2.5(7.2)

8.2 × 2.5 = 20.5 (8.2 × 2.5 = 8.2 × 2 + 8.2 × 0.5 = 16.4 + 4.1 = 20.5)

8.2 − 7.2 = 1.0

(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 everything else stays the same, the second observation will be 1 unit higher than the first prediction. $(2.5(8.2) - 2.5(7.2)) = 1$

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

True. If the predictor terms stay the same, the degrees of freedom increases with sample size.

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 answer would be atleast one mean is different than the others.

(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.

We can only use post-hoc if the overall F-test for model significance has a small enough p-value.

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. if our sample remains the same. The 99% CI includes all the values from 95% CI. However, if we choose another sample our mean and SE will be different and so will be the CI's.

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

True. For large sample sizes, the inference tests are more sensitive because the sample variance becomes smaller as sample size increases.

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

True. It measures strength of association between two 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.

→ We use transform the response variable in the following situations:

1) If there's evidence of non-linearity in residual plots.
2) If there's evidence of non-constant variance in residual plots.

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 the Cook's distance because it identifies influential data points based on both: studentized residuals and leverage values. Any data point, that has both high leverage and high residual is of considerable influence and we should be cautious about how it influences our analysis.

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)

$246 - k - 1 = 242$

(a) What are the error degrees of freedom based on this model? So, $246 - 1 - 242 = k$.

df for model = 3

(b) What is the reference level?

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?

$\alpha_{Blouse} = 44.63 - 41.946$

$= 2.684$

Then,

$\alpha_{Trousers} = 2.684 + 5.937$

$= 8.621$

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 coefficient of the interaction term explains the effect of release variable on the average price of the item based on produce_cost variable. It measures the strength of dependency between the two variables.

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).

Y = No. of research papers.

x_1 = age

x_2 = tenured or not (tenured $x_2 = 1$, not tenured $x_2 = 0$)

x_3 = starting salary

x_4 = college or uni or workforce

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

(a) Whether age plays statistically significant role in predicting research papers published by the professor.

$H_0: \beta_1 = 0$ → {The no. of papers published does not depend on age of the professors.}

$H_1: \beta_1 \neq 0$

(c) ← For MLR,

Whether age and tenure plays a statistically significant role in predicting number of research papers.

$H_0: \beta_1 = \beta_2 = 0$ → {The no. of paper published does not depend upon age or tenure}

H_1: Atleast β_1 or $\beta_2 \neq 0$

(b) ↑ ANOVA

Whether tenure plays a statistically significant role in predicting number of research papers

$H_0: \alpha_j = 0$ → {The number of papers published by tenured and non tenured professors is the same}

$H_1: \alpha_j \neq 0$ where α_j is group effect if group is tenured, 0 otherwise.

Linearity, group effect, constant variance, zero mean.

constant group effects: no interaction between two groups.

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) Conditions:

- Based on the boxplots, the height of the blouses boxplot seems to be a bit higher than rest of the group. However, constant variance assumption is reasonable.
- There seems to be some deviation from normality because the tails of the quantile plots are curled.
- The high outlier in blouse category might raise concern regarding constant variance.
- Since there is no interaction term between any of the groups (blouses, on-sale, skirts), we can say that the group effects are constant.
- It is reasonable to assume that price of the four groups are independent of each other.

b) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ { the average cost of the purchased items are the same for all the groups

H_1: atleast one of $\mu_1, \mu_2, \mu_3, \mu_4$ is not equal to others { the average cost of the purchased item is different for atleast one group.

c) The F-statistic is 19.09 with a significantly small p-value of 3.825e-11. Therefore, we reject the null hypothesis and conclude that the average cost of the purchased item is different for atleast one group. However, our conclusion comes with limitations since the quantile plot showed deviation from normality and the assumption of constant variance was mildly violated. Likewise we need to further investigate the high outlier in blouses category and how it influences our analysis. Likewise, the R^2 is only 0.18, so our model might not be very useful in explaining the relationship between price and items category.

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. With Person B's analysis, there seems to be a strong one-to-one association between Arsenic, Year and a bit less stronger relationship between Arsenic, Miles. However, the collinearity between Year and Miles is unaccounted for in Person A's model. If two predictor values are highly collinear, the adjusted R^2 value is likely to decrease. So including an interaction term of $\beta_1 \beta_2 (Year \cdot Miles)$ is likely to support Person B's analysis. When we are including two or more predictor variables into our model we need to consider how those variables might also be related. We need to consider whether the effect of distance from the well on arsenic level depend on the year the data was collected. Therefore, a possible relationship between the two predictor variables differentiates the conclusion of Person A and Person B.

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