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: Christopher Meng

Swarthmore Username: cmeng1

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 10.0

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

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. Collinear means that the predictors have a relationship to each other, so removing one variable will affect the magnitude of the other 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 can only conclude that at least two of the means are significantly different from each other, and we don't even know which two!

(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 overall ANOVA F-test does this. A post-hoc pairwise analysis, like Fisher's LSD, will tell us which pair or pairs of means are 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.

True.

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

As the sample size increases, the same magnitude of difference is easier to be (You need less strength of evidence to reject as sample size increases.)

False. While the general idea/pattern is true, it is not always true as it depends on exactly how large the sample size is and how small the difference is. I corrected the statement above.

(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.

One reason to transform the response variable is to better meet the regression assumptions, such as linearity, constant variance, and/or normality. Another reason to transform the response is to make the model more interpretable. For example, in our project, some beta coefficients were very small, and a transformation of the response is one solution to make the coefficients more interpretable.

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 to identify potentially influential data points, Cook's distance accounts for "unusualness" in terms of both residuals (the difference between the predicted and observed values) and the leverage (how extreme the predictor values are, which is a measure of how potentially influential the observation is on the fitted regression line) — it is a measure that encompasses both of the other two measures!

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 blouses 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
```

41.946
+ 5.937
47.883
−44.630
3.253

7. (3 points)

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

(b) What is the reference level?

The error degrees of freedom is 242, or the number of observations minus the number of groups (246-4=242).

The reference level is blouses because R encodes the beginning of the alphabet as the reference, and blouses does not have its own coefficient.

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?

The estimated group effect is the difference between the overall group mean and the group mean for trousers in this case. Adding the coefficient for trousers to intercept gives us the group mean for trousers, and finding the aforementioned difference gives us an estimated group effect of about $3.25 if I did my math right.

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.)

Given that the item type is held constant, the interaction term between release and produce-cost, $\hat{\beta}_{rp}$, tells us that for every additional week since the item was released, the effect of produce-cost on the average price of the item changes (i.e. increases/decreases depending on the sign) by $\hat{\beta}_{rp}$. OR for every additional dollar of production cost, the effect of release on the average price of the item changes by $\hat{\beta}_{rp}$.

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.

Age

Starting Salary

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; $\leftarrow$ Starting Salary $= \beta_0 + \beta_1 \text{Age} + \epsilon$ $H_0: \beta_1 = 0$

(b) an ANOVA model; $\leftarrow$ Age $= \beta_0 + \beta_1 \text{LiberalArts} + \beta_2 \text{University} + \epsilon$ $H_0: \beta_1 = \beta_2 = 0$

(c) a multiple linear regression model (not SLR or ANOVA) $\leftarrow$ Age $= \beta_0 + \beta_1 \text{Tenure} + \beta_2 \text{StartingSalary} + \epsilon$ $H_0: \beta_1 = \beta_2 = 0$

$$\text{LiberalArts} = \begin{cases} 1 & \text{if faculty attended liberal arts college} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{University} = \begin{cases} 1 & \text{if faculty attended a university} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Tenure} = \begin{cases} 1 & \text{if the faculty is tenured} \\ 0 & \text{otherwise} \end{cases}$$

(a) Does age of the faculty have a significant relationship with starting salary? The null hypothesis is that changes in age have no relationship w/starting salary.

(b) Are the average ages of faculty members grouped by post-high school graduation plans significantly different from each other (at least 2 of the 3 groups)? The null hypothesis is that the average ages are the same across the three possible post-high school graduation plans.

(c) Does the combination of tenure and starting salary have a significant relationship w/faculty age? The null hypothesis is that both predictors have no relationship with faculty age.

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) Beyond zero mean of errors and additive group effects, we need to consider if there are also constant group effects, which there will be as long as there are no confounding or "third" variables beyond those in our model that could systematically explain the average cost of item by clothing type. Constant variance is violated because the spread of trousers is much greater than the spread of observations by other clothing types (more than 2x greater!). Our book says that, if the greatest group SD is more than 2x the smallest group SD, then we are concerned about constant variance. The normality assumption is also possibly violated as the tails indicate that the studentized residuals are larger than we expect at both ends, which suggests that we have some right skew. We are told that the shoppers are a random sample, so the randomness condition is met. Finally, the independence assumption is in question because online shoppers individually are more likely to have completed their purchases that day from the same store. Additionally, clothing prices are in competition with each other, and may depend on time of year, such as the holiday season.

(b) $H_0: \mu_T = \mu_B = \mu_S = \mu_{OS}$, where $\mu_T, \mu_B, \mu_S, \mu_{OS}$ are the average prices of trousers, blouses, skirts, and onsale items respectively
H_A: at least two of the means are significantly different from each other
The null hypothesis is that all the group means of average item cost by clothing type are the same, while the alternative is what I wrote next to H_A.

(c) From pg 5, the F statistic is 19.09. With 3 numerator degrees of freedom and 242 denominator degrees of freedom, we return a p-value close to 0 and reject the null hypothesis that all the group means of average item cost by clothing type are the same under a significance level of 0.05. However, because many of our ANOVA assumptions are not met, we cannot be confident in the reliability of our statistical test, and we should be hesitant to draw (valid) conclusions. Descriptively, the coefficient estimates and standard errors on pg. 5 suggests that there are differences in price based on clothing type,

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. R^2 tells us the amount of variability in arsenic levels explained by our model, regardless of whether the model is MLR or SLR. In simple linear regression, R^2 happens to be exactly the sample correlation squared, but this is not the case for MLR. The sample correlation tells us the strength and direction of the relationship between one predictor and the response. Person A's R^2 for the MLR tells us the amount of variability in arsenic levels explained by both predictors, Year and Miles, together. Person A and Person B are measuring different things in different contexts (MLR vs SLR), so their responses are not contradictory but actually give us additional info about our model and data at hand.

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