902190128

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: Elizabeth Rosenthal

Swarthmore Username: erosent1

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 - the coefficient accounts represents the effect of variable 1 on Y when controlling for variable 2. If we remove variable 2, we are no longer controlling for it, so the coefficient will change.

$\hat{Y} = \hat{\beta}_0 + 2.5x_1 +$

$2.5(7.2) - 2.5(8.2) =$

(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.

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 only know that at least one mean is signif. diff. from the overall mean, not every single group mean

(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. But it will be more specific - it will tell us which pair(s) specifically are significantly different from one another

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. 99% confidence intervals are always larger than 95% CIs, and they are centered around the same #, so a number inside the bounds of a 95% CI will also be inside the bounds of a 99% CI

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

False. As sample size increases, it becomes more likely that the difference will be statistically significant, but there is ALWAYS a chance of getting a random sample where we don't see this difference (that chance just decreases a lot as sample size increases). With an infinitely large sample size, then this would be true

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

False. Correlation is a measure of the association between any two QUANTITATIVE 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.

1) The data is not normally distributed

2) The data is not linear (ie if it is curved)

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 it takes into account both studentized residuals and leverage. St.resids only tells us when a point is unusual with respect to the response variable, and leverage only tells us when the point is influential with respect to the predictors, so if we only look at one of these measures, we would be missing some important information Cook's Distance combines the two and tells us if the point is influential with respect to either/both* the predictors and the response

* Generally both but can be mainly just one if it's ~~such a large number for~~ so obscure for that one. That is to say, if one point has a very high leverage but normal stresid (or vice versa), it would have a high Cooks distance, but cooks distance can also detect points with a moderately high leverage and a moderately high stresid.

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

(b) What is the reference level? Blouses

8. (6 points)

Suppose the average number of ~~plate appearances~~ price per ~~game~~ item is 44.63 over all 246 data points. What is the estimated group effect for clothing type trousers? ↳ β_3

$\beta_0 = \mu_m$

$\beta_1 = \mu_m - \mu_1$

$\alpha_1 = \mu_1 - \mu_0$

$\alpha_1 \beta_0 - \beta_1 - \mu_0$

$\alpha_3 = 41.946 - 5.937 - 44.63$

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.)

$Price = \beta_0 + \beta_1 Type + \beta_2 Type + \beta_3 Type + \beta_4 Release + \beta_5 Cost + \beta_6 Release \cdot Cost$

The interaction term tells us whether the effect of release time on price is different for different values of product cost, holding product type constant
That is, does release have a different effect (SLOPE) on price depending on the product cost if product type stays the same

Section 3: Long answer questions

10. (9 points)

Age
Tenure
Salary
College

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 the professor's age significantly predict their starting salary
$Salary = \beta_0 + \beta_1 Age + \varepsilon$
$H_0: \beta_1 = 0$ or Age does not have a significant effect on Salary

b) Does the faculty member's path after high school predict their starting salary. That is, is the average starting salary for all faculty who either attended a liberal arts, attended a university, or entered the work force significantly different from the overall average starting salary
$Salary = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ where $X_1 \begin{cases} 1 = \text{liberal arts college} \\ 0 = \text{otherwise} \end{cases}$ $X_2 \begin{cases} 1 = \text{university} \\ 0 = \text{otherwise} \end{cases}$

$H_0: \beta_1 = \beta_2 = 0$ or there is no significant effect of path after high school on starting salary

SEE LAST PAGE

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) 1) Constant and Additive Effects. We know the effects are additive since there are no interaction terms. For effects to be constant, there must not be any other variables that should have been included in the model that could explain the effect of clothes type on price. That is, there isn't a major/important predictor missing from the model

2) Zero mean: We can assume this

3) Constant variance. Based on the side by side boxplots, there doesn't seem to be constant variance. The interquartile range of the on-sale group is much smaller than the range of the trousers group. However, it would be helpful to compare their standard deviations to see if this is a problem. As long as no standard deviation is not more than double another, we can probably say this assumption is met

4) Independence: I think its reasonable to assume that the price of one item will not affect the price of another, so the errors should be independent. However if the same person (or algorithm) determined the prices, then they might not be independent.

5) Randomness: This was a random sample, so the assumption is met

6) Normality: the normal quantile plot does not look usual, since the pattern seems to be steps of horizontal lines. However, the points do not fall that far from the diagonal line, and this obscure pattern may be due to the categorical nature of the variable (since there are a limited number of values the predictor could take), so we might be able to assume normality. However, the data in the parallel boxplots does not look symmetric, since most means are not near the center of the boxes. A transformation should probably be considered

b) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$. H_a: At least one $\mu_i \neq 0$ for $i = 1 \ldots 4$.

The null hypothesis is that there is no significant difference between the average cost of each group of clothing type. The alternative hypothesis is that at least one group mean is significantly different from the rest.

(SEE LAST PAGE)

* The Adjusted R^2 is saying that 26% of the variability in the data can be explained by the model including both Year and Miles. The correlation coefficients tell us that 59% of the variability can be explained by just Year and 11% by just Miles since Miles does not explain that much variability, so it might not be particularly helpful in the MLR (so the adjusted R^2 penalizes us for including it) & is decreasing

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:

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

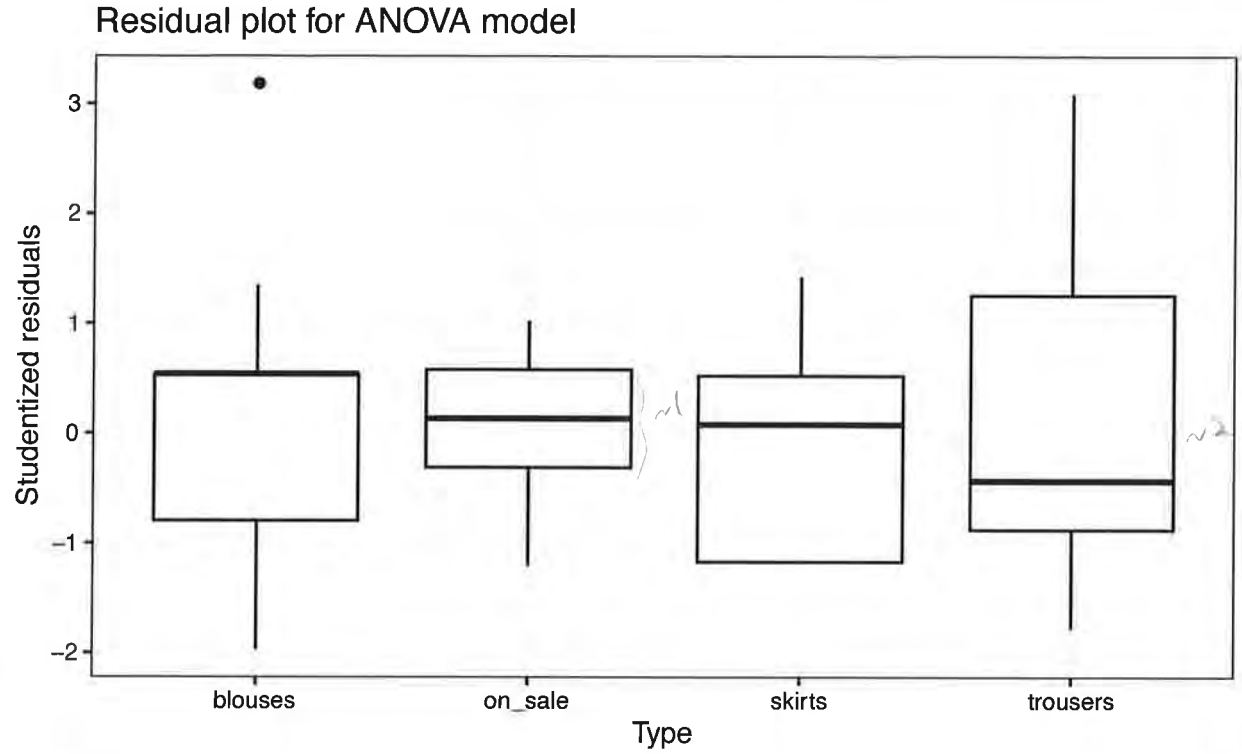
and estimates each with their sample correlations $r_1 = 0.77$ and $r_2 = -0.34$. Are the two people's conclusions contradictory? Explain your answer.

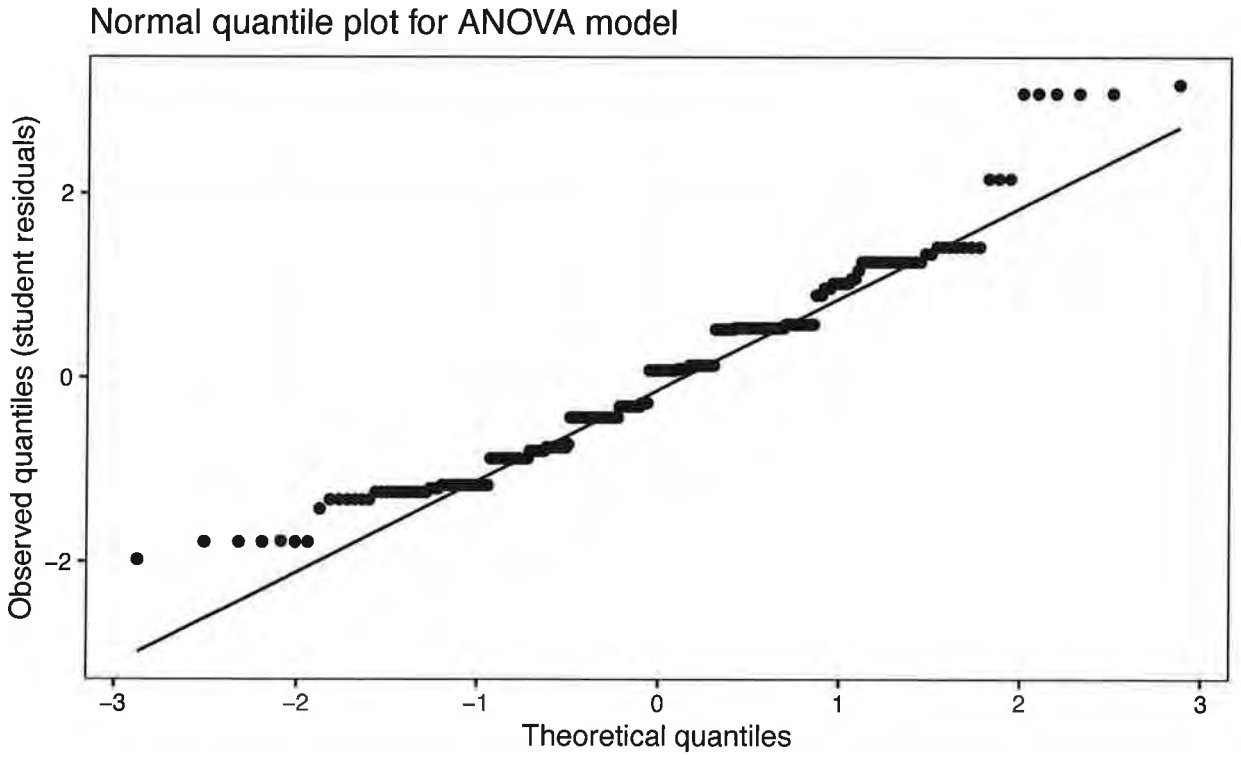
No, they are not contradictory. Person A did a MLR whereas person B effectively did 2 SLRs (since R^2 for a SLR is the square of the correlation coefficient). The MLR explains the effect of year and miles on arsenic levels when controlling for the other. The two SLRs do not. Moreover, the adjusted R^2 gives a penalty for including more predictors in the model, so if one predictor (likely Miles, since that individual correlation is low) does not explain much more variability in the data, then the adjusted R^2 will decrease * It's also important to consider whether these predictors are multicolinear. It's possible that early years had wells close to mining sites, since maybe there used to be more mining sites. (or maybe more mining sites opened recently, so years could then be negatively correlated with miles). This multicolinearity would affect the MLR but not the individual SLRs. Lastly, the MLR has fewer population parameters that need to be estimated (3 Bs and σ whereas the 2 SLRs have 4 Bs and 2 σ's to be estimated). The MLR only has one error term while the SLRs have 2 altogether.

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





“

QUESTION 10

c) Does a model with the professor's age and tenure status explain a significant amount in their starting salary. That is does age and tenure status predict their starting salary

$Salary = \beta_0 + \beta_1 Age + \beta_2 X_2$ where $X_2 \begin{cases} 1 = \text{tenure} \\ 0 = \text{otherwise} \end{cases}$

$H_0: \beta_1 = \beta_2 = 0$ or neither age nor tenure status has a significant effect on salary when controlling for the other

QUESTION 11

c) The p-value for the overall F-test is less than 0.05, so we can reject the null hypothesis and conclude that the price of at least one group is significantly different than the overall average price. However, the adjusted R^2 was very low (0.18)* and the assumptions might not have all been met, so this conclusion might not be very reliable

* meaning that the model does not explain very much of the variability in the data.

Next, we could want to find Fisher's LSD to see which groups are significantly different from one another. Since we don't have that information, we could look at the individual t-tests for the β-coefficients. Since they are all significant at an $\alpha = 0.05$ level, we can say that each $\beta_i \neq 0$. However, this still does not tell us which specific groups are different from one another. It is also hard to make guesses about which groups are different based on the boxplots, since the boxes seem to have a lot of overlap. My guess would be that blouses is significantly different from trousers since blouses has the highest mean and an outlier and trousers has the lowest mean but we would need to do a Fishers LSD test to confirm this.