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: Zhengfei Li (Alex)

Swarthmore Username: zli4

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

$$Y = \underset{P}{\beta_0} + \underset{S}{\beta_1 x_1} + \underset{f}{\beta_2 x_2} + \underset{age}{\beta_3 x_3} + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon,$$

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? $x_3=0$ $\beta_1=\beta_2=0$

b) Does the effect of age on a bird's weight depend on what type of bird it is? $\beta_4=\beta_5=0$

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? $\beta_1=0$

d) Given we are only comparing pigeons, is the effect of age on a bird's weight statistically significant? $X_1=X_2=0$ $\beta_3=0$

e) Is there statistically discernible evidence of a linear relationship between bird age and type and bird weight?

1. c) $H_0 : \beta_1 = 0$
2. a) $H_0 : \beta_1 = \beta_2 = 0$
3. d) $H_0 : \beta_3 = 0$
4. b) $H_0 : \beta_4 = \beta_5 = 0$
5. e) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

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. Because when collinearity is present, the "colinear predictor" have to allow changes in other variables when the variable charge. Therefore, the coefficient of the variable does not reflect its own relationship with the response (so does other variables). So when a predictor is removed, other variable's coefficient can charge a lot.

(b) If a regression model's first variable has a coefficient of $\hat{\beta}_1 = 5.7$, then if we are able to influence the data so that an observation will have a value of x_1 be one unit larger than it was before, the value of y_1 for this observation would increase by 5.7 units.

False. The change is also dependent on other terms related to x_1 e.g. whether there might be interaction term with x_1 involved like $x_1 x_3$.

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

Note: Predictors are held constant, k is the same.
Sample size increase, n increase.

3. (5 points) df(Error) = n - k - 1 increases

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. From ANOVA overall F-test, if we reject null hypothesis, we only kno that some mean is different from others but we don't knw whether all are different. Eg. We might have only one group mean significantly different but others the same

(b) The standardized variability among the group averages is higher than the estimate of the variability of the data within each group.

True.

Note: $F = \frac{MSMod}{MSE}$ need to be significantly larger than 1 to be "unusual".

(c) A post-hoc pairwise analysis will identify if there is at least one pair of means that are significantly different.

False. Because a post-hoc pairwise analysis focus on "Individual Error Rate" after conducting ANOVA F-test. Post-hoc pairwise analysis identifies where the difference is rather than whether there's difference. "Whether there's at least one

4. (5 points) difference" should focus on "family-wise error rate" and should be ANOVA F-test.

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) Decreasing the significance level (α) will increase the probability of making a Type 1 Error.

False. Type I error is error made when we reject null hypothesis when null hypothesis is true. Decreasing significance level would lead to stricter rule against rejecting null hypothesis and probability of making Type I error decreases.

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

False. To determine whether there's statistical significance depends on the value of difference and standard error, which is a measure of variance and degree of freedom. We cannot say about statistical significance without knowing variance.

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

True.

Section 2: Short answer questions

5. (4 points)

Briefly describe a benefit of analyzing the studentized residuals of a regression model rather than just analyzing the observed residuals.

Studentized residuals makes it more convient for us to identify unusual data points with cutoffs: studentized residual more extreme than ±2 is moderately unusual and studentized residual more extreme than ±3 is extremely unusual.

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 values. This is because Cook's distance values are a function of both leverage and standardised residuals.

$$D_i = \frac{(stdres_i)^2}{k+1}\left(\frac{h_i}{1-h_i}\right)$$

Therefore, Cook's distance value includes information about both leverages (which measures unusualness on predictor values) and unusualness of residuals.

For questions 7-9 consider the following random single-serving samples of $n = 76$ breakfast cereals. We are going to model the average calories per serving (in g) (`calories`) as a linear function of the cereal manufacturer (a categorical variable with levels: G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina). Below is the R summary output for this one-way ANOVA model.

```
##
## Call:
## lm(formula = calories ~ Manufacturer, data = cereal_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.696  -8.696  -0.126   5.909  51.304
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      111.364      3.959  28.126  < 2e-16 ***
## ManufacturerK     -2.668      5.538  -0.482  0.63149
## ManufacturerN    -24.697      8.553  -2.887  0.00516 **
## ManufacturerP     -2.475      7.348  -0.337  0.73729
## ManufacturerQ    -16.364      7.667  -2.134  0.03633 *
## ManufacturerR      3.636      7.667   0.474  0.63678
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.57 on 70 degrees of freedom
## Multiple R-squared:  0.1618, Adjusted R-squared:  0.102
## F-statistic: 2.703 on 5 and 70 DF,  p-value: 0.02724
```

7. (3 points)

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

(b) What is the reference level?

(a) 70. Error degrees of freedom can be read from summary output "Residual Standard error 18.57 on 70 degrees of freedom."

(b) General Mills is the reference level because the level is not included as a predictor term.

8. (6 points)

Suppose the average amount of calories for all these samples is 106.97 over all 76 data points. What is the estimated group effect for Quaker Oats cereal brand?

Let μ_Q denote group mean of Quaker Oats, α_Q denote group effect for Quaker Oats, and μ_{Grand} denote Grand Mean.

$\mu_Q = 111.364 - 16.364$ calories

$\alpha_Q = \mu_Q - \mu_{Grand} = 111.364 - 16.364 - 106.97$ calories

9. (4 points)

Consider two additional numeric predictors: **sugars** (in g) and **protein** (in g). If we were to fit a regression model including each of the three predictor variables (including manufacturer) 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 is the average effect of per unit increase in sugar on the average effect (slope) of per unit increase in protein on the average amount of calories in breakfast cereals.

Section 3: Long answer questions

10. (9 points)

Suppose you have access to a data set on a random sample of undergraduate-only institutions in the US. The variables included ~~in this~~ data set are a numeric variable for the average cost of tuition each semester, a binary categorical variable distinguishing private institutions from public ones, a numeric variable for the percentage of full-time instructional staff employed at the institution, and a categorical variable indicating whether the school is a liberal arts college, a community college, a technical/vocational school, or if they are institutionally affiliated with certain groups (e.g. historically Black, women's only, tribal, etc).

State ① a research question that can be answered with the overall F-test for each of the following models, based on this data. (You do not need to use every variable, but you can.) Also provide ② a mathematical representation of the model and state ③ 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) To which extent does the average cost of tuition each semester have an effect on the percentage of full-time instructional staff employed at the institution?

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$

where Y stands for the percentage of full-time instructional staff employed at the institution and X_1 stands for the average cost of tuition each semester.

$H_0: \beta_1 = 0$, which means that there's no effect of tuition on percentage of full time staff.

$H_A: \beta_1 \neq 0$, which means that there's statistical significant effect of tuition on percentage of full-time staff.

(b) To which extent does classification of schools affect average cost of tuition each semester?

$Y = \mu + \alpha_i + \varepsilon$

where Y is average cost of tuition each semester, μ is grand mean of average cost of tuition, α_i is the group effect where α_1 is group effect of liberal arts college, α_2 is group effect of community college, α_3 is group effect of techical/vocational school, and α_4 is group effect of institutionally affiliated schools.

$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, which means that none of the group effects are statistically significant.

H_A: at least one $\alpha_i \neq 0$, which means that at least one group effect is statistically significant.

(c) To which extent does private/public classification and average cost of tuition each semester can help predict the percentage of full-time instructional staff?

$Y = \beta_0 + \beta_1 \cdot \text{tuition} + \beta_2 \cdot \text{IPrivate} + \beta_3 \cdot \text{tuition} \cdot \text{IPrivate}$

Where Y is the percentage of full-time instructional staff employed at the institution, variable tuition stands for the average cost of tuition each semester and catagorical predictor term IPrivate $=1$ when it's a private institution and IPrivate $=0$ when it's a public institution.

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$, which mens that none of the predictors are statistically significant in predicting percent % of full time staff.

H_A: At least one $\beta_i \neq 0$, which means that at least one predictor term is statistically significant.

11. (8 points)

Consider the ANOVA model for the cereal 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 calories (per serving) is significantly different for these six different cereal manufacturers. (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) Randomness is met in the method of "Random Sampling of cereal". Independence assumption is reasonable because each sample is tested on their own and the amount of calorie does not affect each other.

For the model, just as suggested in question 9, Sugar and Protien level in the cereal can be lurking variables for prediction of calories level. Therefore, assumption of constant effect of different groups of manufactures can be resonable after assessing that lurking variables like Sugar and Protien does not take major effect on predicting amount of calories and group effect of manufacturers are reasonable to be believed as a constant value instead of depending on other factors.

Constant Variance is violated as we do the division of largest standard error of group effect to the smallest: $\frac{S_{max}}{S_{min}} = \frac{8.553}{3.959} > 2$ (data from R output). This means that constant variance is not acceptable due to the overly large difference in group variance.

Normality assumption is clearly violated according to the plots. The distributions of Group Manufacturer N, P, and Q are clearly assymmetric in the residual plot, which is one evidence against normality. Also, according to normal quantile plot, smaller residuals are smaller than expected and larger residuals are larger than expected. This provides evidence of very narrow peak, which is another evidence against normality.

(b) $Y = \mu + \alpha_i + \varepsilon$ where y stands for the amount of calories in cereal, μ stands for grand mean of calories of cereal, α_i stands for group effect of manufacturers.

$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_6 = 0$, which means that all group effects are not statistically significantly different from zero.

H_A: some $\alpha_i \neq 0$, which means that some group effect is statistically different from zero.

(c) There's no convincing conclusion to the hypothesis test in question (b). Refer to more detailed explanation in (a), assumptions of constant variance and normality is clearly violated and the probability model, which the test in (b) is based on, is not reasonable. Therefore, we cannot make conclusions from (b).

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.

The reason for the two findings is due to the additional variability explained by Miles not already explained by Year in Person A's model is too little.

The correlation between Arsenic and Year is strong with r_1^2 is 0.5929

$r_1^2 = 1 - \frac{SSE_{B1}}{SSTotal}$ in the model (B1) Arsenic $= \beta_0 + \beta_1$ Year $+ \varepsilon$. $\frac{SSE_{B1}}{SSTotal} = 0.4071$

When variable of Miles is added, we use adjusted R^2 $n=70$ $k=2$

$$R^2_{adj} = 1 - \frac{SSE_A/(n-k-1)}{SSTotal/(n-1)} = 1 - \frac{SSE_A/(70-2-1)}{SSTotal/(70-1)} = 1 - \frac{69}{67} \times \frac{SSE_A}{SSTotal}$$

$$\frac{69}{67} \times \frac{SSE_A}{SSTotal} = 0.74$$

Adding a variable can never decrease SSE. So $SSE_{B1} > SSE_A$.

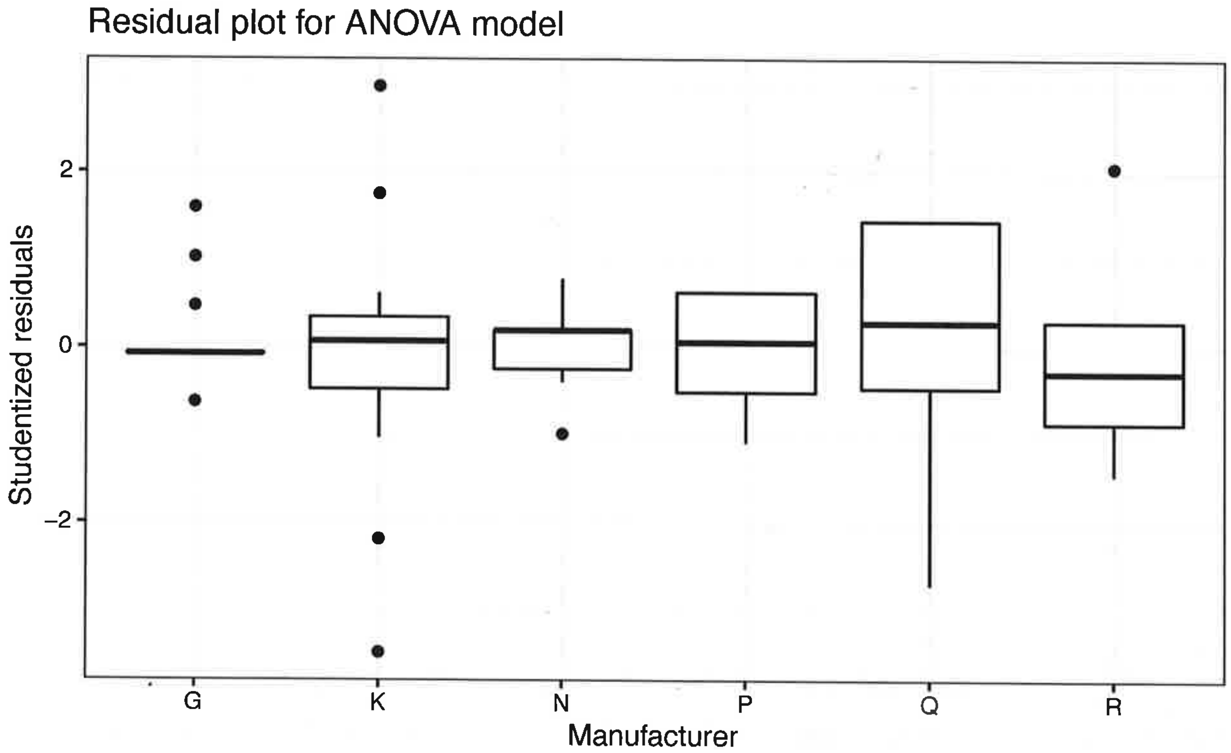
SSTotal is always the same if two people uses the same data set.

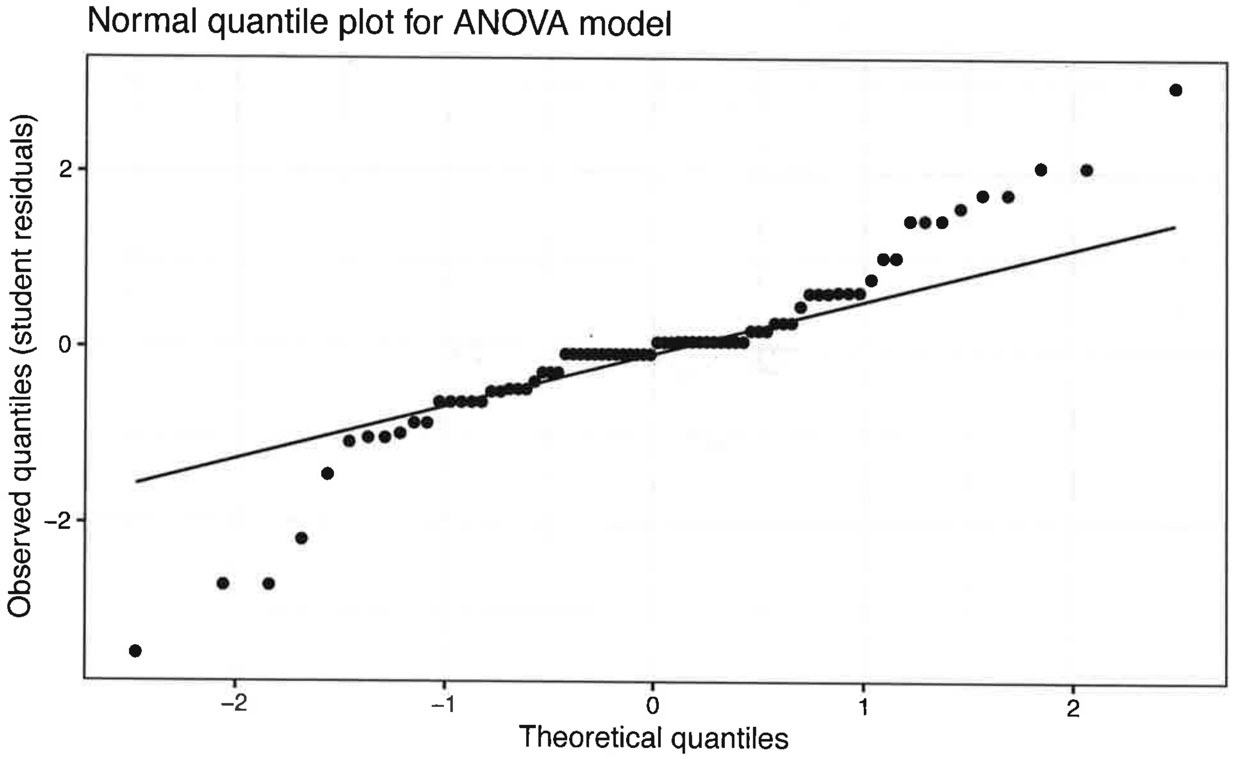
⇒ ANSWER on A SEPERATE SHEET

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.

Cereal ANOVA Model


Residual plot for ANOVA model


Normal quantile plot for ANOVA model

2114

12| No, the two conclusions are not contradictory.

This is because Person B's discussion of correlation is based on a simple linear regression. Correlation r (or determinant of correlation r^2) only depends on SSE and SSTotal acording to $r^2 = 1 - \frac{SSE}{SSTotal}$.

However, Person A's discussion is based on MLR. The adjusted R^2 used takes into account the number of predictors : $R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SSTotal/(n-1)}$.

Although adding a variable does not decrease SSE of a model (from model by person B to person A), R^2_{adj} can still be smaller than r^2 of Cor(Arsenic, Year) the increase in number of predictors.

Cause of such phenomenon is that the unique variability explained by Miles of Arsenic that is not explained already by Year is very little, provided correlation between Miles and Arsenic is not high either and posible colinearity with Years.

Therefore, SSE does not decrease much in addition of Miles but there's a predictor number increased. R^2_{adj} can be smaller than r^2 of individual predictors.