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: Jess Sang

Swarthmore Username: jsang1 P02186365

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.____C____ $H_0 : \beta_1 = 0$ $H_A : \beta_1 \neq 0$

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____ $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

reject

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 b/c once you remove that variable, the remaining one is going to over compensate for that missing effect from other variable.

(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, you need to also say holding everything else fixed.

(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; alt. hypothesis is that at least two group means are different than each other, so it may not be all means diff from each other

(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; in order for you to conduct post-hoc pairwise analysis a requirement is that ANOVA F-test p-value is significant, which means that a condition is that at least one pair of means are sig. diff.

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.

5%

$\frac{5}{100} = \frac{1}{20}$ type 1 error

$\frac{1}{100} = \frac{1}{100}$

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

false; if you have sig. level of 5%, it's like saying 1 in 20 times there's a false pos., so if you decrease sig level, the chance of getting false positive decreases.

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

true

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

true

Section 2: Short answer questions

5. (4 points)

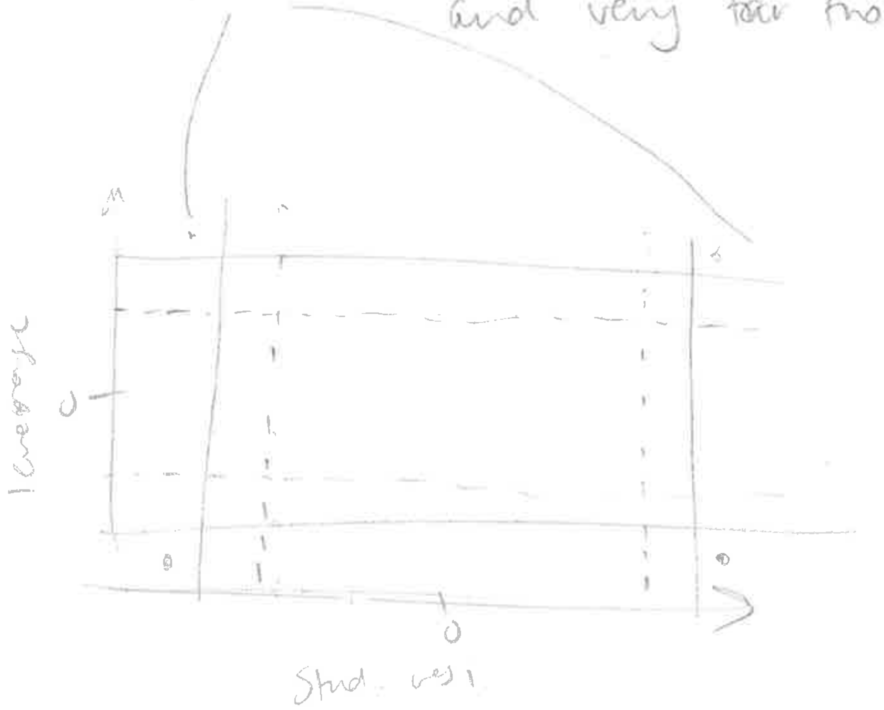
Briefly describe a benefit of analyzing the studentized residuals of a regression model rather than just analyzing the observed residuals.

the benefit of analyzing studentized residuals of a regression model rather than just observed is that we are able to put everything on the same scale and it makes comparability easy. For example if you had a simple linear regression and residuals, how would you know if this very large observed observed residual is that far away from center of residuals? with studentized, you can see how many standard deviations away that residual is to see if it is an outlier.

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 b/c it is a combination of stud. residuals and leverage. Since it takes into account both then that will tell us if it is really leveraged and very far from center 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? 70

(b) What is the reference level? general mills

	df
group	5
error	70
total	75

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?

$\mu = 106.97$ $\quad\quad$ $\mu + \alpha_Q = \mu_Q$

$\alpha_Q = \mu_Q - \mu$ $\quad\quad$ $\alpha_Q = \mu_Q - \mu$

$\alpha_Q = 95 - 106.97$ $\quad$ $\mu_Q = \beta_0 + \beta_4$

$\boxed{\alpha_Q = -11.97}$ $\quad$ $= 111.364 + (-16.364) = 95$

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 for this interaction term represents the additional effect of having an additional unit of both sugar and protein on the average calories per serving. The coefficients for each predictor alone are not sufficient in conveying their effect on average calories, so we include an interaction term to show that when you have an add. unit of both sugars and proteins in cereal, there is an even stronger effect from sugar + protein on # of cals, which would be the beta coefficient for that int. term.

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 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 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).

full model.

$$(Y)\ \text{graduation rate} = \beta_1 \overset{X_1}{\text{cost}} + \beta_2 X_2 + \beta_3 \overset{X_3}{\text{pct full time}} + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$$

(C)

where

$$X_2 = \begin{cases} 1, \text{ private} \\ 0, \text{ public} \end{cases} \quad X_4 = \begin{cases} 1, \text{ lib arts} \\ 0, \text{ other} \end{cases} \quad X_5 = \begin{cases} 1, \text{ comm coll.} \\ 0, \text{ other} \end{cases} \quad X_6 = \begin{cases} 1, \text{ tech school} \\ 0, \text{ other} \end{cases}$$

(a) SLR:

$$\underset{\text{graduation rate}}{Y} = \beta_1 \underbrace{X_1}_{\text{\% full time instructional staff}} + \varepsilon$$

null: $\beta_1 = 0$, which means the % of full time instruct. staff has no effect on grad. rate.

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

Question: is there a statistically significant relationship b/w % of full time instructional staff and the graduation rate?

null hypothesis:

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

all betas are equal to 0, which means none of the predictor variables has an effect on graduation rate.

H_A: not H_0 ↘ at least one $\beta_i \neq 0$

Question: Do any of these variables have a statistically significant relationship w/ grad rate?

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) zero mean ✓ linearity: going from kellog to quaker or quaker to general mills (i.e. going from one level to another) should not have non-constant and non-additive effects, like multiplicative effects or squares.

constant variance: not really met b/c the spread of data b/w groups is very diff, particularly that of Quaker, g mills, and nabisco as shown by distinctly diff. length boxes on the residual plot.

normality: not really met either with the dots straying from the line on both tails.

randomness. random samples so condition is met

independence: no reason to believe the errors from one cereal manufacturer would impact those of another so met

(b) $\underset{\text{aveg. calories}}{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$

$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$ or $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu$

H_A: at least one $\alpha_i \neq 0$ or at least two μ_i are diff from each other

Jess Sang

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.

From person A's output, the conclusion they would reach is that year and miles together account for 26% of the variability in the amount of arsenic in local wells. 0.26 is not bad for R², but there may be other variables that could be added into the model to improve it. From person B's correlations, it seems like year is very strongly positively correlated with arsenic while miles and arsenic are somewhat negatively correlated. While there is a stronger correlation b/w year and arsenic, I am inclined to think that has to do with time and the amount of arsenic last year having impact on arsenic this year. Therefore, I don't think the two people's conclusions are contradictory b/c although person B might feel the R^2 should be higher given the higher r_1, the higher r_1 is due to time, so it makes sense that the R^2 value is relatively low, but around the correlation of r_2.

also correlation is the square root of R^2, so if you do → $\sqrt{R^2}$ $\sqrt{.26} \approx .51$ → y

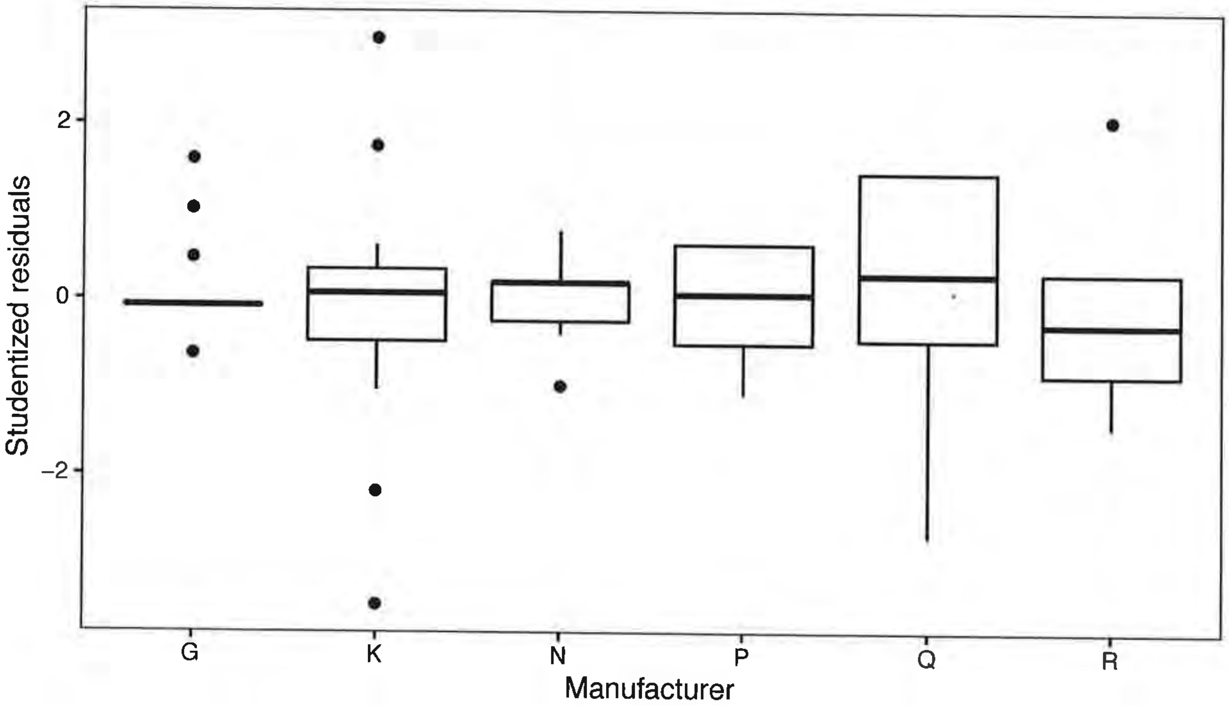
not part of answer

Section 4: Extra credit opportunity

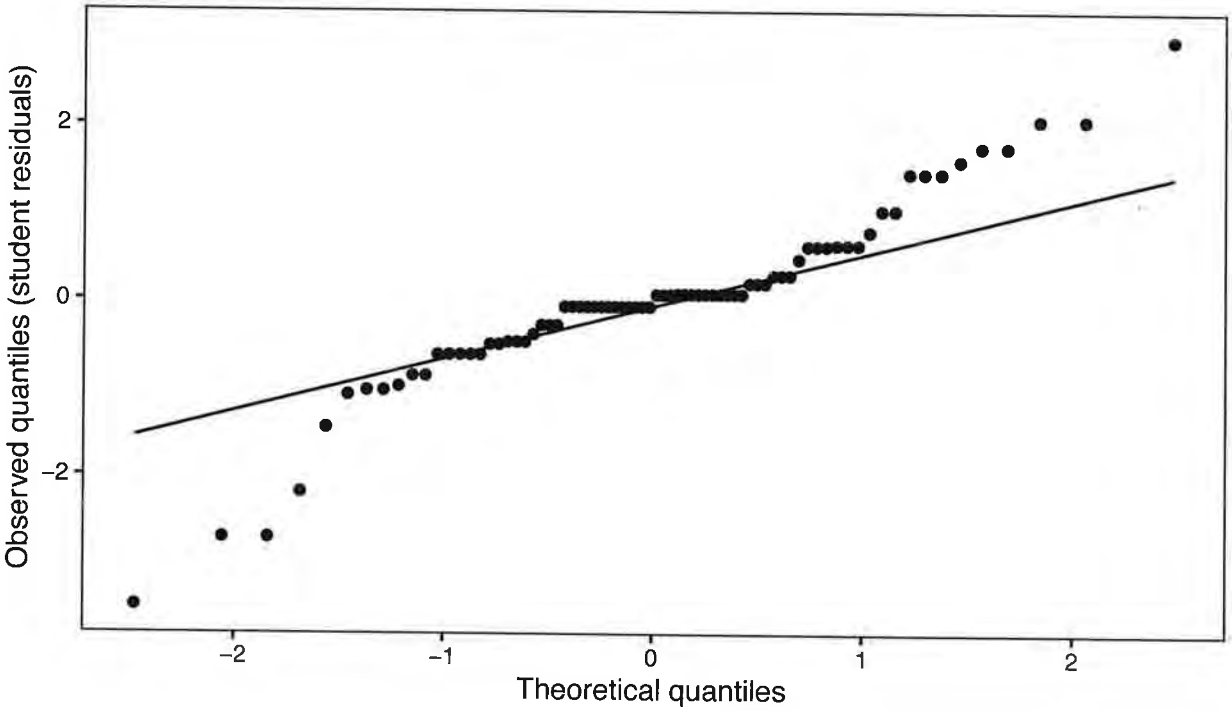
If the response rate to my end of the semester evaluation form (on Moodle under Week 13 and 14) is at least 85% of our class size (over both sections), two percentage points will be added to everyone's Test 3 grade (up to 100 total possible points). **Hint:** You may not know how to or want to contact 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



10. (cont.) Jess Sang

(b) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

grad. rate

$X_1 \begin{cases} 1 = \text{lib arts} \\ 0 = \text{other} \end{cases}$ $X_2 \begin{cases} 1 = \text{comm. college} \\ 0 = \text{other} \end{cases}$ $X_3 \begin{cases} 1 = \text{tech/voc. school} \\ 0 = \text{other} \end{cases}$

$H_0 =$ $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ or $H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$

$H_A =$ at least one $\alpha_i \neq 0$ or $H_A =$ at least two μ_i differ from each other

Question: is the grad. rate significanty ~~diff.~~ for these 4 different types of schools?

null: all ~~two~~ types of schools ~~have~~ have no effect on the grad rate or the mean grad rate of each type of school is not diff. from overall mean.

11. (cont.)

H_0: all cereal groups have no effect on avg. calories or all group means of each cereal are equal to overall mean

H_A: at least one cereal group has an effect on average cal. or at least two cereal groups have means diff from each other.

(c) from our summary output, while the p-value for our overall ANOVA F-test is statistically sig at $\alpha = 0.05$ b/c p-value $= 0.02724$, and would imply that we could reject our null hypothesis in favor of alternative, based off our residual plot and normal quantile plot, we have not met all conditions for our test to be reliable, namely, constant variance as shown by different length box plots and normality as shown by the dots straying the line towards the tails. Reliability depends on us meeting all model assumptions and is crucial for us to be able to use our test to reject or fail to reject the null.