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: Malavika Eby

Swarthmore Username: meby1

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

$$Y = \beta_0 + \overset{\text{sparrow}}{\beta_1 x_1} + \overset{\text{finch}}{\beta_2 x_2} + \overset{\text{age}}{\beta_3 x_3} + \overset{\text{sp age}}{\beta_4 x_1 x_3} + \overset{\text{finch age}}{\beta_5 x_2 x_3} + \epsilon,$$

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.

~~2~~ a) For newly hatched birds (of age zero months), is there a statistically discernible difference in the weights of these three different bird types?

~~4~~ b) Does the effect of age on a bird's weight depend on what type of bird it is?

~~2~~ 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?

~~3~~ 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. E $H_0 : \beta_3 = 0$

4. D $H_0 : \beta_4 = \beta_5 = 0$

5. B $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

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 → Even if 2 variables are collinear, there will likely be at least a minor change in the other variable coefficient when 1 variable is removed.

True

(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.

Yes, if keeping all other variables constant.

(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. $\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4$?

not necessarily

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 one group mean is different from the rest.

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

$\left(\text{high } F \rightarrow \frac{\text{big MSmodel}}{\text{MSE}}\right)$

True

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

False → By the ANOVA results, we already know that at least 1 pair of means is significantly diff. Post hoc analysis will tell us which group(s) is 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) Decreasing the significance level (α) will increase the probability of making a Type 1 Error.

False negative

True

(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.

False → Correlation is a measure of the linear association between 2 numerical variables.

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 are based off of the student's t distribution model + accounts for smaller sample size. Our data is usually not likely to be perfectly normal + may have heavier tails so unlike raw residuals, stud. residuals are informed of this in their calculation.

~~When we observe "raw" observed residuals, we might interpret them as large +/or unusual simply based on whether a number looks big. Studentized residuals are more statistically informed + alert us to whether a residual is actually minor or large relative to the already existing error within~~

Please ignore :)

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 to identify potentially influential data points. Cook's D is comprehensive in its detection of ~~unusual~~ influential points, taking into calculation both a measure of the distance between the points from its expected or estimated values & its observed values (stud. resids) as well as leverage, which provides insight about a point's potential influence on a regression. In my opinion, leverage would also adequately help us identify potentially influential data points as it reveals how some points can seem to hold more weight on the slope of a regression based on its residual & location near the upper or lower end of x-values.

Yet I would still prefer to take advantage of having the Cook's D measure for all my data points as I believe it provides me with more information & is based on more factors of unusuality & influence than merely the stud. residuals or leverage.

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
```

(76-5-1=70)

7. (3 points) $\boxed{70}$ deg. of freedom

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

(b) What is the reference level?

The reference level is General Mills.

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?

$$\mu + \alpha_j = \mu_j = \beta_0 + \beta_{Quaker}$$

$$111.364 + (-16.364) = 95$$

$$\begin{array}{r} 95 \\ -106.97 \\ \hline -11.97 \end{array}$$

$\hat{\alpha}_j = \boxed{-11.97}$ 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.)

→ manufacturer, sugars, protein

$$\left[\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_6 X_7 \right]$$

Keeping all other variables constant, we can expect the average calories per cereal serving to change by an average of β_8 calories for every unit increase in sugars or protein. Realistically, we cannot keep all other variables constant as when sugar level is manipulated, the $(\beta_6 X_6)$ term also contributes to the change in the response variable but we might still include the interaction term to preserve "built-in" collinearity between cereal sugar + protein levels.

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).

a.) Does the percentage of full-time instructional staff employed at an institution make a difference in the average cost of tuition/semester there?

$$\left(\underset{\text{estimated exp. avg cost of tuition}}{\hat{Y}} = \beta_0 + \beta_1 \underset{\text{\% full-time staff}}{X}\right)$$

$H_0 \rightarrow$ Percent of full-time instructional staff has no stat. sig. diff on avg. sem. tuition. $(\beta_1 = 0)$

$H_A \rightarrow \beta_1 \neq 0$

b.) Does an institution's standing as "public" versus "private" affect its average tuition per semester?

$$\left(\overset{\text{avg tuition}}{Y} = \underset{\text{grand avg}}{\mu} + \overset{\text{public effect}}{\alpha_1} + \overset{\text{priv effect}}{\alpha_2} + \overset{\text{error}}{\epsilon}\right)$$

$H_0 \rightarrow \alpha_1 = \alpha_2 = 0$ ("public" or "private" standing makes no effect on the avg. per-sem tuition)

$H_A \rightarrow$ some $\alpha_j \neq 0$

$H_0 \rightarrow \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (Tuition + identity make no difference in % full-time staff)

$H_A \rightarrow$ at least 1 $\beta_k \neq 0$

c.) Does an institution's categorical identity (liberal arts, community college, technical school, etc) or avg. semesterly tuition have an effect on the percent of full-time staff employed there?

$$\left(\underset{\text{estimated \% full-time staff}}{\hat{Y}} = \beta_0 + \beta_1 \underset{\text{tuition}}{X_1} + \overbrace{\beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}^{\text{indicator vars}}\right)$$

$X_2 = \begin{cases} 1 \rightarrow \text{liberal arts} \\ 0 \rightarrow \text{otherwise} \end{cases}$

$X_3 = \begin{cases} 1 \rightarrow \text{comm. college} \\ 0 \rightarrow \text{otherwise} \end{cases}$

$X_4 = \begin{cases} 1 \rightarrow \text{technical/vocational} \\ 0 \rightarrow \text{otherwise} \end{cases}$

(if all 0, then institutionally affiliated)

(SLR)

zero mean
CV
linearity
independence

Random
Normality

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.) ☑ Random group effects → It was most informative (no patterns = more random) to generate a residuals plot to test this condition the sampling process was randomized so we'll accept this assumption.

☑ Independence in group effects → Given that the production of cereal in one manufacturing plant doesn't affect the production process/ingredients/nutritional value of another, we can assume this condition is met.

☑ Constant variance → The calorie content for some cereals seem to have more variance than others (Q, possibly K, P, R) but for the most part they're quite similar & within the range of (-2, 2) so we can tentatively accept. (based on resid. plot)

Additive effects → There's no interaction term in this model so the effects are additive.

? Normality → We can detect a wave-like pattern in the normal q. plot & relatively large deviance at at low + high quantiles so this assumption is questionable but somewhat met at the center values.

b.) H_0 → There is no significant difference between the avg. calories/serving for the 6 different cereal manufacturers.

$$\hat{Y} = \underset{}{\beta_0} + \overset{(K)}{\beta_1 X_1} + \overset{(N)}{\beta_2 X_2} + \overset{(P)}{\beta_3 X_3} + \underset{(Q)}{\beta_4 X_4} + \underset{(R)}{\beta_5 X_5}$$

$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$

K → Kelloggs
N → Nabisco
P → Post
Q → Quaker Oats
R → Ralston Purina

H_A → At least 1 cereal manufacturer produces cereal w/ a significantly different avg. value of cal/serving.

$\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq \beta_5$

c.) The overall ANOVA test provides us with a small p-val of 0.027 (<0.05) which means we can reject our null hypothesis. There is less than a 5% chance that we could have collected our data set given that there's no sig. diff. between the average cal/serving for each cereal manufacturer. Looking at the p-val for each individual predictor term, the group effects of Kellogg's, Post & Ralston Purina do not seem significant (high p-values) in estimating avg cal/serving. This model accounts for about 10% of the variance in its results (moderately weak strength) as expressed by the R^2 adj.

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.

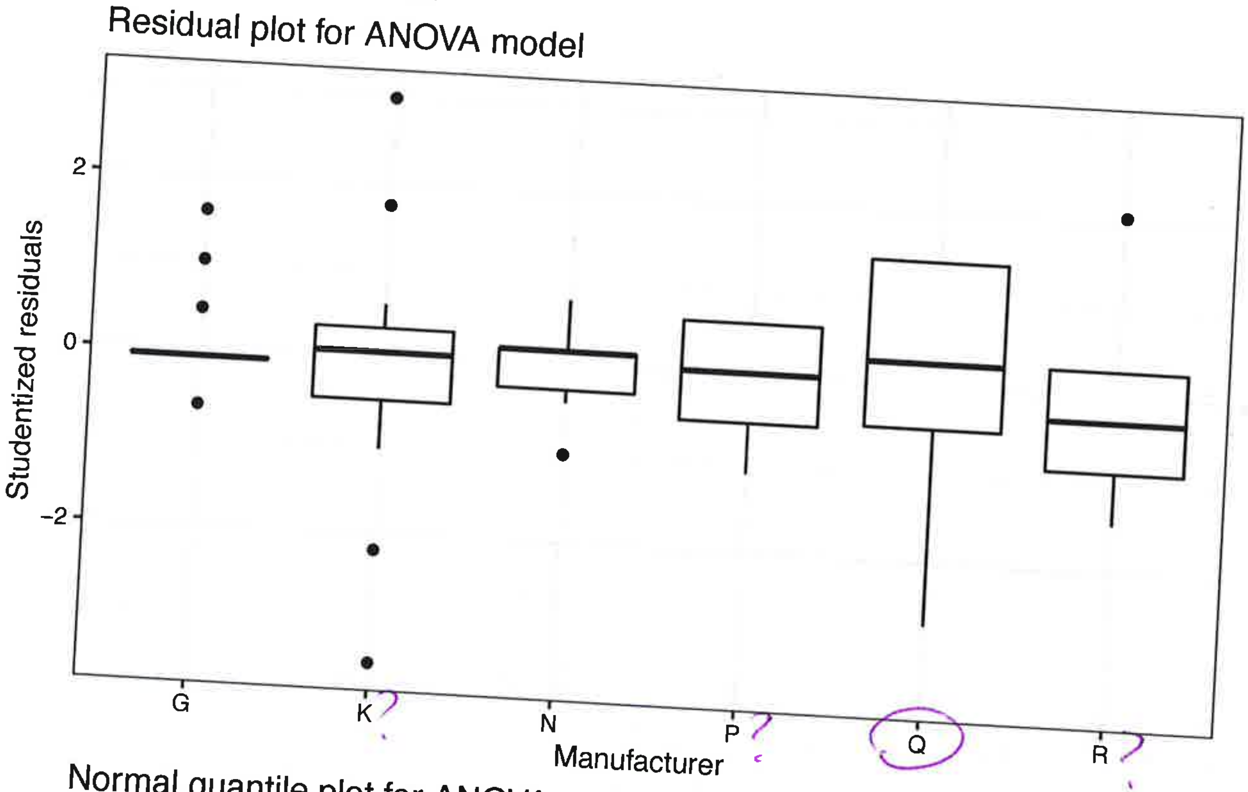
No, they're not contradictory. Person A's model observes the percent of variance in arsenic content accounted for by the combined effect of Year + distance from nearest mining site, the answer to which is 26%. Person B is modeling a SLR, not a MLR like Person A, + is looking at the correlation coefficient (r) instead of adjusted coefficient of determination (which is used in MLR to adjust for higher # of predictors artificially inflating R^2).

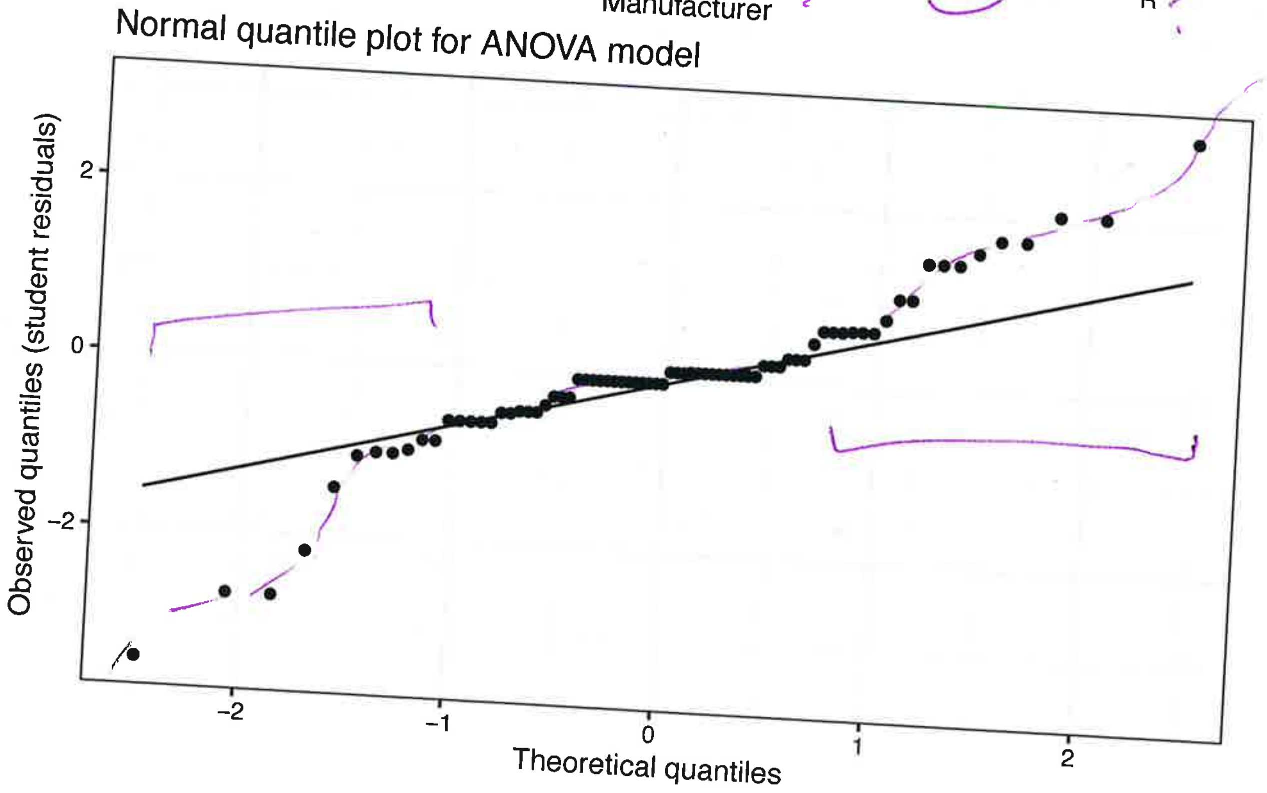
Though Year has a positive moderately strong correlation w/ arsenic content + Miles has a mildly weak negative correlation w/ arsenic content, when both variables are included as predictors in 1 model, it's possible for those combined effects to result in a relatively lower adj. R^2.

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
Studentized residuals
G
K
N
P
Q
R
Manufacturer
?
?
?


Normal quantile plot for ANOVA model
Observed quantiles (student residuals)
Theoretical quantiles