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: Xinxin Li

Swarthmore Username: xli5

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 $H_0 : \beta_1 = 0$

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. Since predictors are collinear, we can write a predictor as a function of some other collinear predictors, so removing will change coefficient.

eg. $X_1 = 2X_2 - 1$

$Y = -3X_2$

adding both sides: $Y + X_1 = -X_2 - 1 \Rightarrow Y = (-)X_2 - X_1 - 1$

changes from 2 to -1.

(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. It only means the y_1, on average, is expected to increase by 5.7 units.

(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 say at least one group has mean that is different from the others.

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

false. the variability of the data within each group should also be the standardized one, instead of the estimate.

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

false. It should be ANOVA F-test that does the job. post-hoc pairwise analysis checks which pair have significantly different means.

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. the probability of Type I error = α since Type I error is basically false alarm, so it will decrease that probability.

(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 measures the linear relationship only. Other types of association are not considered.

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.

For studentized residuals, we can use the generalized boundaries (>2 and >3) to get the extremes, which is relatively more precise. We can also compare different models in an easier way since different models may have different SE for errors. But ~~the~~ for the observed residuals, it is hard to make such comparison if SE (errors) differ a lot. It's also hard for us to discern the extremes.

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?

Cook's distance values.

It ~~measures both~~ takes into consideration of both studentized residuals ~~which~~ which measures the extremity of the response value and leverage which measures the extremity of the predictor values.

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
```

106.97
- 95
11 97

171
- 16
1

7. (3 points)

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

(b) What is the reference level?

a. df(error) = sample size - # of grps = 76 - 6 = 70

b. the Intercept value gives the measure for reference level, General Mills. since all the rest are ~~a~~ encoded in predictor terms.

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?

~~Or~~ $\bar{y}_G = 111.364$ $\hat{\beta}_Q = -16.364$ $\bar{y} = 106.97$ ← grand mean.

$\bar{y}_Q = \bar{y}_G + \hat{\beta}_Q = 111.364 - 16.364 = 95$

$\hat{\alpha}_Q = \bar{y}_Q - \bar{y} = 95 - 106.97 = -11.97$

estimated Group effect for Quaker Oats.

$\bar{y}_G$: mean for grap General Mills.

$\bar{y}_Q$: mean for Quaker Oats.

$\hat{\beta}_Q$: coefficient for Quaker Oats in the given model.

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.)

~~this is~~ If we hold one numeric (X_1) variable unchanged and raise the other numeric (X_2) variable by 1 unit, then the coefficient of the interaction term measures the extra effect of X_2 ~~of~~ on Y ~~when~~ that is caused by the interaction between X_1 and X_2.

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).

A B C D

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).

↗ on the separate paper.

a. Is there a statistically discernable evidence of a linear relationship between the average cost of tuition each semester (Y) and the percentage of full-time instructional staff employed at the institution (X) ~~for~~ for undergrad-only institutions in the US?

$$\text{average cost} = \beta_0 + \beta_1 \text{percent\_of\_staff} + \varepsilon$$

i.e. $Y = \beta_0 + \beta_1 X + \varepsilon$

$H_0: \beta_1 = 0$; so there is no linear relationship between Y and X.

b. Is there a statistically discernable ~~dif~~ difference between at lease ~~one~~ one institution type's mean of average cost of tuition each ~~e~~ semester and the mean of cost of other institution types? Types are A: liberal arts, B: community college, C: tech/vocational, D: institutionally affiliated.

grp mean of D — diffrac between grp means of A, B, C and grp D

$$\text{average\_cost} = \beta_0 + \beta_1 X_A + \beta_2 X_B + \beta_3 X_C + \varepsilon$$

X_A, X_B, X_C are set to 1 if the ~~gro~~ institution type ~~of~~ A is A, B, or C respectively. Otherwise, they are set to 0.

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$; so there is no difference among the grap effect of the 4 groups, and there is no difference ~~amo~~ among grap means.

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.)

on the separate paper. ← (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. From the residual boxplot, boxes for many groups have relatively different lengths, ~~some of them~~ so equal variance condition is not met. Normal quantile plots have points that lie below the line on the left and lie above the line on the right end, indication the distribution of ~~residuals~~ residuals have ~~long tail~~ heavy tails at the ends, so normality is ~~as~~ not met. Since the ~~as~~ data are taken from a SRS, we can assume the randomness and independence conditions are met, and we can also assume that each predictor is not a function of any other predictor terms or lurking variables, so the group effects are constant. Since ~~this as~~ is an one-way ANOVA model, we can assume the group effects are additive.

b. H_0: $\alpha_G = \alpha_K = \alpha_N = \alpha_P = \alpha_Q = \alpha_R = 0$.

where $\alpha_G \cdots \alpha_R$ are group effects of group (G) General Mills, K, N, P, Q, R respectively.

This means there's no difference among the ~~as~~ group means of average calories for the 6 groups.

H_A: at least one of the $\alpha_G, \alpha_K, \alpha_N, \alpha_P, \alpha_Q, \alpha_R \neq 0$.

So at least one group has group mean for average calories that is different from the rest of the groups.

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.

Not contradictory.

~~It is possible that there's collinearity between year and miles, which will make person A's model less valid, so the we cannot trust the R^2_{adj}.~~

~~If we are able to believe the R^2_{adj} in A is valid, then It means~~

From person A, we can say that MLR the model that considers Year and Miles can explain 26% of the variability of Arsenic. This seems to be a low value.

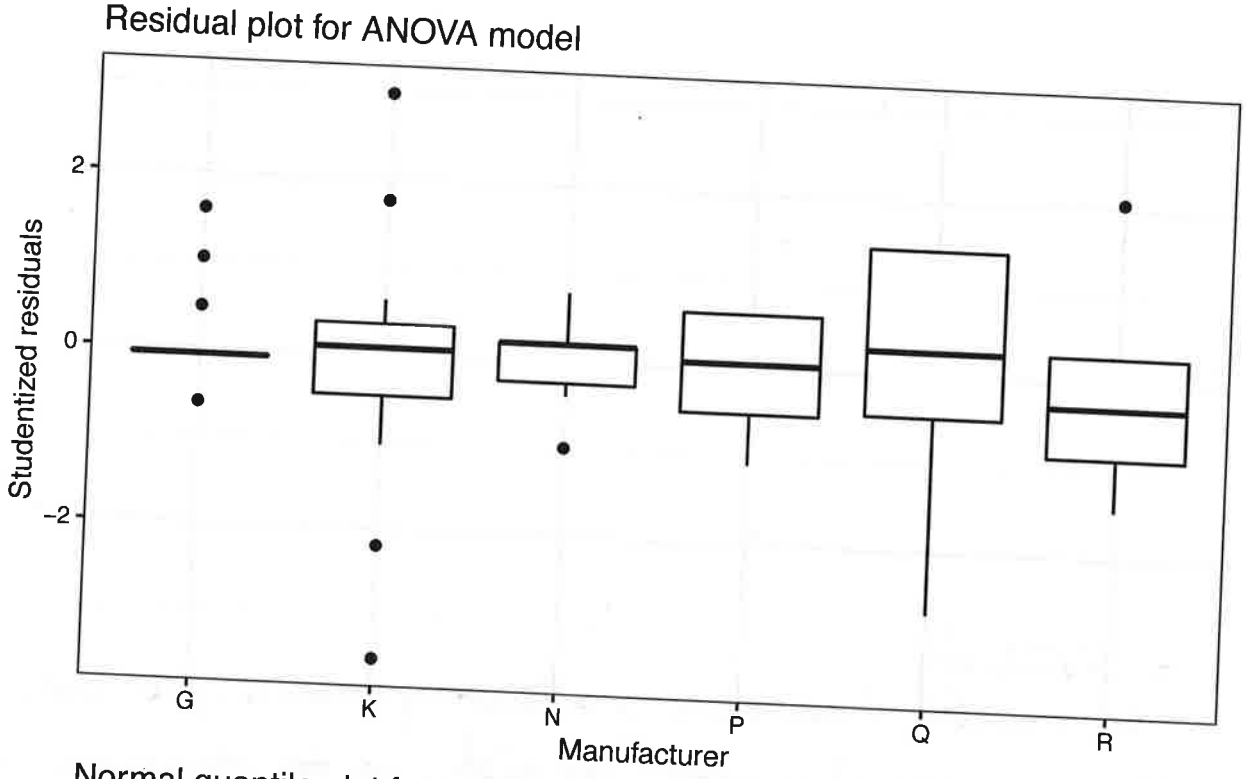
From person B: we can say that the SLR ~~with~~ for ~~Arsenic~~ predicting Arsenic using Year has $r^2 = 0.77^2$, which is quite high, so by Year itself, it can explain a large amount of variability of Arsenic. The SLR for predicting Arsenic using Miles, ~~though~~ has only ~~have~~ $r^2 = (-0.34)^2$ which is lower than R^2_{adj} in A, so Miles alone may no be effective in predicting Arsenic, and the linear relationship between Arsenic and Miles is also relatively weak.

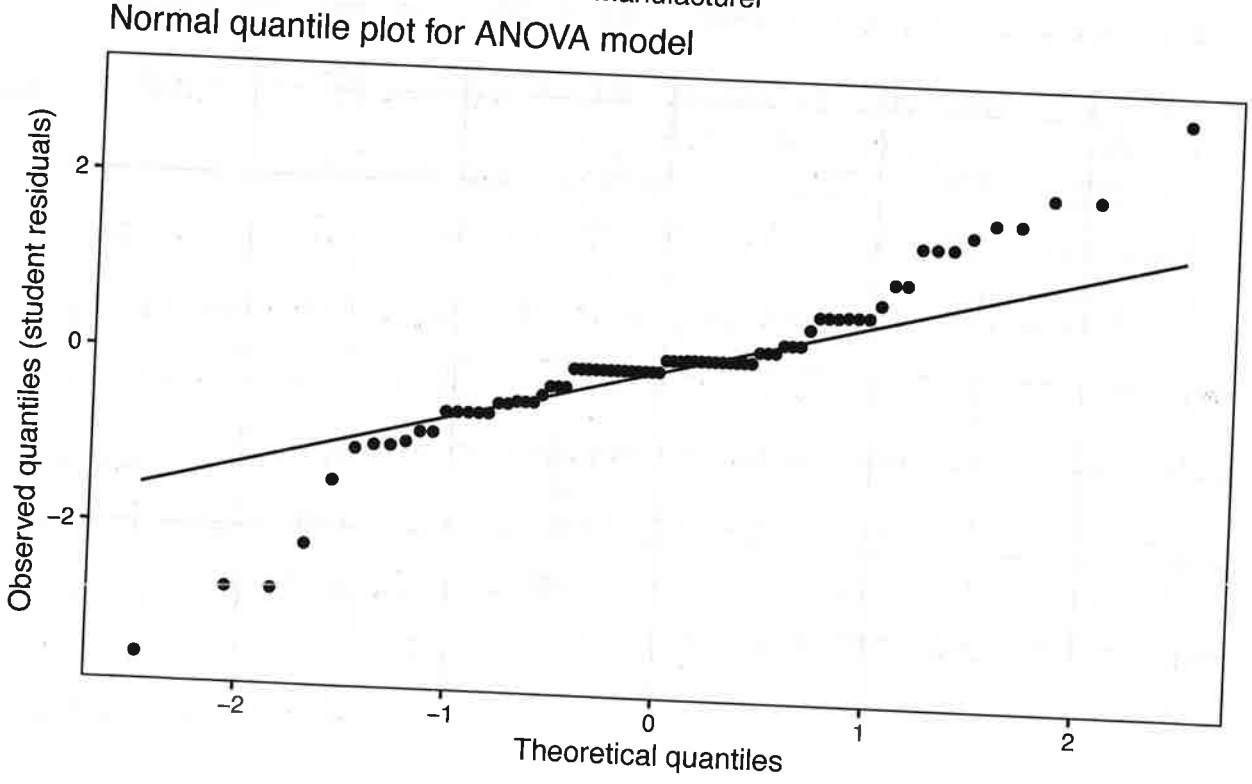
Section 4: Extra credit opportunity

(continued on the paper)

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





Xiwen Li

10. C. Is there a statistically discernable evidence of a linear relationship between the average cost (Y) and whether the institution is private or public (X_1) and the percentage of full-time staff (X_2)?

$$average\_cost = \beta_0 + \beta_1 ISPublic + \beta_2 percent\_staff + \varepsilon$$

ISPublic = 1 when the institution is public.

otherwise, ISPublic = 0.

$H_0: \beta_1 = \beta_2 = 0$.

there is no linear relationship between average cost and whether the institution is private or public and the percentage of full-time staff.

11. C. The ANOVA F-test has F-statistic = 2.703, which gives pvalue = 0.02724 if we take alpla = 0.05, then pvalue $< \alpha$, so we reject Ho, and there is statistically significant difference between at least one group mean for average calories and the group means for other groups.

However, since based on residual plots, equal-variance and normality conditions are not met, this conclusion based on ANOVA F-test may not be a valid conclusion.

If the conditions are met, since the data are from a SRS, so the conclusion can be generalized to the population of all cereals for these brands, but since there is no random assignment, causal conclusion cannot be made.

Xinxin Li

12. (continued).

Person A's model takes into consideration of a predictor variable that has a relatively ~~re~~ strong correlation ~~between~~ with Arsenic and a predictor that is weak, so the combining result gives an R^2_{adj} in between r_1^2 and r_2^2. So it is reasonable for A and B to get their conclusion from the same dataset.

It is also possible that there's multicollinearity between Miles and Years that make the result in MLR ~~for~~ less ~~relia~~ reliable, ~~one~~ or there is an interaction effect between Miles and Years that is not considered.

We may need to check residual plots for conditions and p-values ~~for~~ t-test for ~~one~~ the ~~model~~ MLR from A and SLR from B to get more reliable results.