lzarate1

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: Litzbeth Zarate-Hernandez

Swarthmore Username: lzarate1

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

$$Y = \beta_0 + \beta_1 \cancel{x}_1 + \beta_2 \cancel{x}_2 + \beta_3 x_3 + \cancel{\beta_4 x_1} x_3 + \cancel{\beta_5 x_2 x_3} + \epsilon,$$

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?

X_1 and $X_2 = \phi$

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

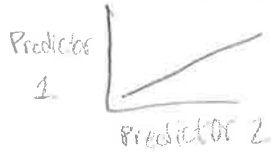
1. c $H_0 : \beta_1 = 0$
2. a $H_0 : \beta_1 = \beta_2 = 0$
3. d $H_0 : \beta_3 = 0$
4. b $H_0 : \beta_4 = \beta_5 = 0$
5. e $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

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 if they are collinear it means that the predictors are influenced by each other so if we remove 2 one variable it will have influence on the point estimate of the other variable's coefficient. This is where interaction term is needed

(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 we us the data points to calculate the coefficient of the variables, so if we influence therpoint, the coefficients will change.

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

Also in this scenario we haven't accounted for the other coefficien of the other variables

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 ... H_0 rejected

(a) We can then conclude that all the means are different from one another.

False we can conclude that at least one group mean 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.

True Fratio is large

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

True because alternative hypothesis, $\mu_i \neq 0$, at least one group mean 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.

True

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

False The larger the sample size, the small difference won't have a bigger effect on significance, rather small differences won't be ignored or large significant differences Caused by many points in sample will be statistically significant.

(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 allow us to see clearly outliers that are not as obvious in residuals plots because we are considering the estimated standard deviation of the error, $\hat{\sigma}_{(i)}$ when the i^{th} data point is deleted, so it tests each data point

$$\text{studres}_i = \frac{y_i - \hat{y}_i}{\sigma_{(i)}\sqrt{1-h_i}}$$

☆ **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 because it the formula takes into account the standardized residuals to check which values are too influential and the cuttoff is more strict (cutoff window is small because moderate values are >0.5 and extreme values are >1 while studentized residuals cutoff is >2 for moderate values and >3 for extreme values. While Leverage cutoff depends on the Predictor terms and sample size.

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
```

$\frac{7.667}{3.959}$ $\frac{8}{4} = 2$

7. (3 points)

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

(b) What is the reference level?

(a) According to the table, the error degrees of freedom is 70

(b) 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?

Manu. = manufacture

$\widehat{\text{calories}} = 111.364 - 2.668\ \text{Manu.K} - 24.697\ \text{Manu.N} - 2.475\ \text{Manu.P} - 16.364\ \text{Manu.Q} + 3636\ \text{Manu.R}$

106.97

$Y = \mu + \alpha_j$ where $j = K, N, P, Q, R, G$

~~$106.97 = 111.364 + 2.668\ \text{Manu.K} + 24.677\ \text{Manu.N} + 2.475\ \text{Manu.P} - 3636\ \text{Manu R}$~~

9. (4 points) $106.97 = \mu + \alpha_Q$ $\alpha_Q = 106.97 - \mu$ $\boxed{Q = \frac{106.97 - \mu}{-16.364}}$

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 the interaction term is the measure of the affect of sugar and protein combined. The interaction means that sugar has influence on the point estimate of protein's coefficient and vice versa so we have an interaction term inclued in the model to account for this.

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;

10c is on another page

(c) a multiple linear regression model (not SLR or ANOVA).

(a) a simple linear regression model question we could ask is, is there a relationship between average cost of tuition each semester and the percentage of full-time instructional staff employed. In other words, can we predict avg cost of tuition each semester & percentage of full-timed instructional staff employed.

Please refer to the other page for the rest of the Question/Answer:

(b) Is the average cost of tuition each semester significantly different for different type of institutions The types include liberal arts college, community college, technical/vocational school, or institutionally affiliated with certain groups?

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ There is no difference between the groups' means.

$H_A: \mu_i \neq 0$ at least one group mean is different.

Where μ_1 = liberal arts college, μ_2 = community college μ_3 = technical/vocational school, μ_4 = institutionally affilliated

model: $Y = \mu_j + \epsilon$ where $\mu_j = \mu + \alpha_j$ and $j = \underbrace{1,2,3,4}_{}$

numbers correspond to each of μ_i alread stated

Average cost of Tuition $= \mu_j + \epsilon$

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) The conditions neccssary include, additive & multiplicative, normality, variance among groups, and independence
From the studentized residials, we observe that there are notable skeweness from the cereal manufactures G, P, and Q.
So normality is not met. This skeweness also shows up in the normal quantile plot. The tails show skeweness. For indepence,
given the context of the research, the cereal type would effect each other
so independence is met. For variance it is met because the boxplot
shows the constant variance accross groups despite the oultiers from
Manofacture K but that is okay. Given that this randomly sample
We can assume linearity. If the group effects are constant, we can compare them to each other.

(b) Hypothesis:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = 0$

$H_a: \mu_i \neq 0$ i can be 1, 2, 3, 4, 5, 6

where μ_1: K, μ_2: N, μ_3: P, μ_4: Q, μ_5: R, μ_6: G

The null hypotlesis would be that the means of groups are not different.

groups = cereal manufacture

The alternative hypothesis would be that at least one group mean is different.

11c) is on anotler page

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 two people's conclusions are not necessarily contradicting because while Person A's $R^2 = 0.26$ (which is low) and Person B reported $r_1 = 0.77$ and $r_2 = -0.34$, we don't have information about how much variability is explained by Year vs Miles.
It may look like Miles have no clear correlation with levels of arsenic, but Miles may be able to explain for variability that Year can not explain.

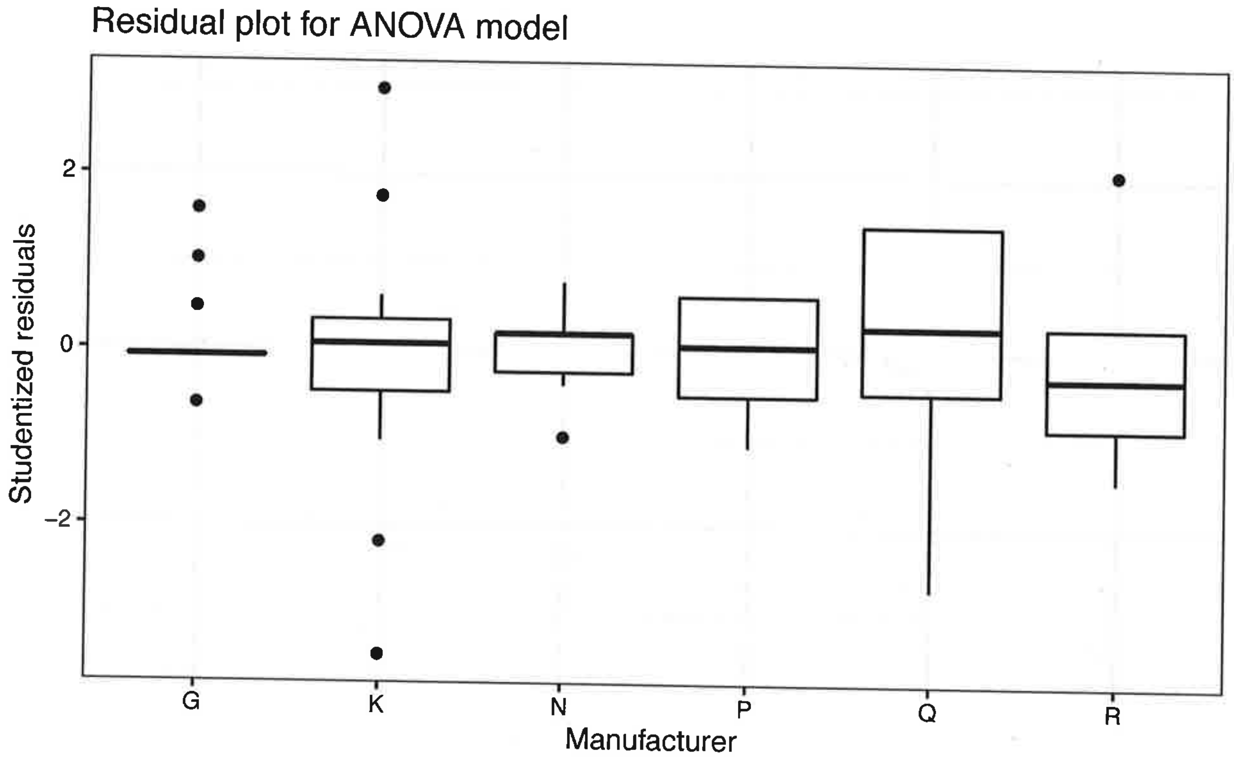
Also R^2 is the coefficient of multiple determination for a model, so the % of data points that can be predicted using the model while the sample correlation measures the relationship b/w the two variables.

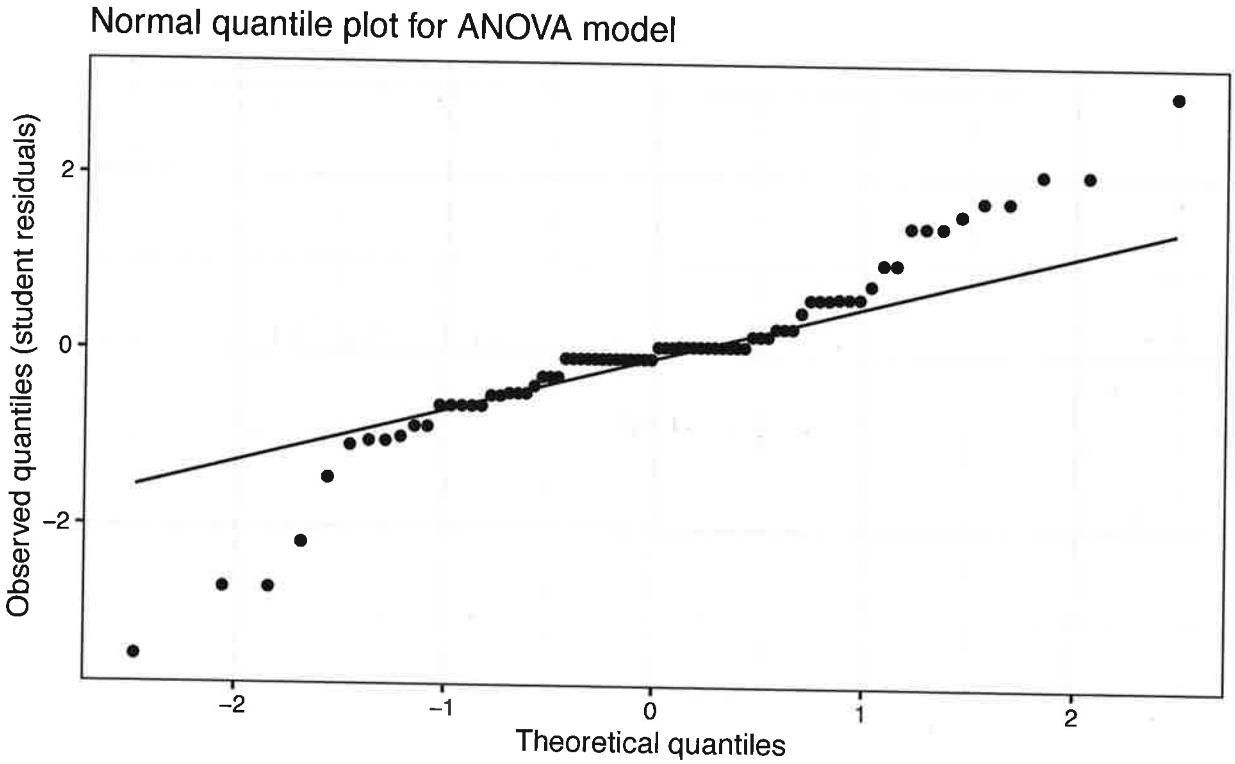
We would need to compare their (Miles and Year) residuals.

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





Izarate

10a) A simple linear regression model question we could ask is, is there a relationship between average cost of tuition each semester and the percentage of full-time instructional staff employed at the institution. In other words, can we predict average cost of tuition each semester using % of full-time instructional staff employed at the instiution?

Model: $Y = \beta_0 + \beta_1 X_1 + \epsilon$

response

Y = average cost of tuition each semester

X_1 = percentage of full-time instructional staff employed at the institution. Predictor

Hypotheses:

$H_0 : B_1 = 0$

Null hypothesis is: There is no relationship b/w the response and predictor

$H_A : B_1 \neq 0$

Alternative hypothesis is that there is a relationship b/w the response and predictor.

10 c) a multiple linear regression model research question could be, is there statistically discernible evidence of a linear relationship between average cost of tuition each semester, percentage of full-time instructional staff employed at the institution, type of institution, and whether it is private of public institution

By type, I am reffering wheller the school is a liberal arts college, a community college, a technical/vocational school, or if they are institutionally affiliated with certain groups.

model: Avg cost of tuition each semester = β_0 + β_1 Percentege of full time instruc. staff employed + β_2 type of institution + β_3 Private/Public + ϵ

Hypothesis:

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

H_a: some $\beta_i \neq 0$ i can be 1, 2, or 3

Null hypothesis is if there is no relationship between the predictors and response, thus all coefficients equal zero.

Alternative hypothesis if there is a relationship between the response and at least one predictor.

11c) F-statistic: 2.703 , P-value: 0.02724
Adjusted R-squared: 0.102

From the summary output for this one-way ANOVA model, we note that the F-statistic is 2.703 and the P-Value ≈ 0.027. so, P-value < $\alpha = 0.05$ and we have a large F-ratio, Thus we reject the null hypothesis in favor of the alternative. At least one group (one of the cereal manufactures) mean (avg calories) is different. But due to my discussion in assessing the conditions necessary to conduct a test in part a, This test conclusion is not trustworthy. It is also important to note that the ajusted R-square is 0.102, ver low, so the current model isn't the best.

Perhaps if we could apply a transformation to the model, to response, the normality, the distribution within the groups will look better.