902170687

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: Dulce Ventura

Swarthmore Username: dventur1

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? $\beta_1 = \beta_2 = 0$

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? $\beta_3 = 0$

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, removing one variable may, in fact, alter the point estimate because the model would be different + even if predictors are collinear they may still explain some different variability than each other

(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, this will only be the case if we control for all other predictor variables.

(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. $H_0: \alpha_1 = \ldots = 0$

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 conclude that at least one mean is different than the others

(b) The standardized variability among the group averages is higher than the estimate of the variability of the data within each group.
$F = \frac{MS_{mod}}{MSE}$
False, the F-test would be rejected if one of the estimates of variability with each group is greatly larger than the predicted one

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

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, Decreasing the significance level will decrease the probability of making a Type I Error.

(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.

Studentized residuals are a way of standardizing residuals that allows you to observe greater extreme (outlier) residuals than the observed residuals. Some points that are outliers in an observed residual aren't on the studentized residual plot because they ~~aa~~ are not as influential

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 takes into account both studentized residuals & leverage values when identifying potentially influential data points. I believe that you should look at both residuals & leverage, and Cook's distance uses both. because the residuals will inform you of errors in regards to how badly your model predicted your output. & the leverage will indicate influence in regards to your predictors. Extreme values in any one of them are points of worry, but ~~it~~ points are unusual in respect to both measures they ~~indice~~ may greatly influence your model.

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
```

G is the reference level

k = predictor terms

$k = 7$

$n = 76$

$76 - 7 - 1 = 68$

k = predictor terms

7. (3 points) error $n-k-1$
$n=76$

(a) What are the error degrees of freedom based on this model? 68

(b) What is the reference level? General mills is the reference level

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?

$\mu = 106.97$

$\alpha_K = 106.97 - 111.364$

$-16.364 = \beta_1$

$\beta_0 = \mu - \alpha_K$

$\beta_0 = 111.364 = 106.97 - \alpha_{1C}$

$\beta_1 = \alpha_K - \alpha_1$

$\alpha_1 = \alpha_K + 16.364$

$\alpha_1 = 106.97 - 111.364 + 16.364$

$\boxed{\alpha_1 = 11.97}$ The estimated group effect for quaker oats is 11.97 calories.

106.97
+ 16
122.97
- 111.00
11.97

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.)

calories ~ sugars + protein + manufacturer + protein · sugar

The coefficient of the interaction term between sugar & protein will inform us on how the dependency of the two variables on each other affect the average amount of calories when controlling for all other predictors.

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 number of full time instructional staff employed at an institution predict the average cost of a tuition per semester?
$y = \beta_0 + \beta_1 x_1 + \varepsilon_i$ where y is the average cost of tuition per semester, and x_1 is the number of full time instructional staff.
$H_0: \beta_1 = 0$ The null hypothesis is that the coefficient β_1 will equal 0, meaning that there is no linear relationship between x_1 & y

(b) Is there a significant difference between the type of school (i.e. liberal arts, community, etc) and the average cost of tuition per semester?
$y = \mu_j + \varepsilon_{ij}$, where y is the average cost of tuition per semester & $\mu_j = \mu + \alpha_j$ $j = 1, \ldots, m$ where m is the total amount of levels of the categorical variable indicating what type of institution a school is.
$H_0: \mu_1 = \ldots \mu_m$; where $\mu_j = \mu + \alpha_j$ the null hypothesis indicates that there is no difference between group means, other wise saying that there is no difference between the overall mean & group means

c) Does the # of full-time instructional staff, & whether an institution is a private or public predict the average cost per semester?
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$ where y = the average cost per semester, x_1 = the # of full time instructional staff & $x_2 = \begin{cases} 1, & \text{if a school is public} \\ 0, & \text{if other wise} \end{cases}$

$H_0: \beta_1 = \beta_2 = 0$

The null hypothesis is that β_1 & β_2 both $= 0$, meaning that there neither predictor has a linear relationship with y.

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) 1) 0 mean ✓
2) Linearity ✓
3) Constant variance — The constant variance of the anova model does not seem to be met given that the IQ ranges are not similar across all manufacturer groups
4) Independence — The independence sample might also not be met given that some of these manufactures may obtain ingredients from similar sources or be produced at similar factories.
5) Normality — The normal quantile plot indicates a lot of lifting at beginning & end tails, signifying that there are heavier tails. This might be indicative of a violation of the normality assumption
6) Randomness — There is random sampling of cereal selection. so no visble violation of randomness assumption.

b) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_m$; where $\mu_j = \mu + \alpha_j$ & $j = 1, \ldots, m$
$H_a: \mu_1 \neq \ldots \neq \mu_m$
μ_k is the group mean for manufacturer k, μ_2 for man. N, μ_3 for man. P, μ_4 for man. Q, μ_5 for man. R & μ_m for the reference level man G
The hypothesis states that there is no difference between the group means, they are all equal to each other.

c) The p-value of the anova test indicates significance at an α level 0.05 since the p-value is 0.027, with that we can reject the null in favor of the alternative. The p-value indicates that the likely chance of all of the means being the same is less than 0.05 meaning that it is likely that at least one of the group means is different. However, due to the violation of assumptions seen in the boxplot, we should not use this model as the group effects are not constant.

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.

I don't believe that their conclusions are contradictory because when the plotting individual potential predictors against the response variable we often see singular relationships. However, we can't use the two individually to fully know how they will interact in a regression model. Moreover, the beta coefficients of the model might see similar relationships between predictors & response variables as the sample correlations. In addition r & R^2 are different things. R^2 is telling us how much of the variability is explained by the model & tell us about the fit of our model rather than tell us how correlated the predictors are to the response variable.

The R^2 in this case tells us that the model may not be a great fit for the data as the model only predicts 26% of the variability. It does not say anything about arsenic & year or arsenic & miles individually.

(individual; using both predictors)

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