90220902

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: Oliver Clackson

Swarthmore Username: oclacks1

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

sparrow finch age type · age

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon,$$

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 $H_0 : \beta_1 = 0$ sparrow vs. reference
2. a $H_0 : \beta_1 = \beta_2 = 0$ indicators significant
3. d $H_0 : \beta_3 = 0$ age when bird type constant
4. b $H_0 : \beta_4 = \beta_5 = 0$ interaction
5. e $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ full model test

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. Regardless of whether two predictors contain similar information, the model containing just one of the two will have a different coefficient for either variable than the model containing both.

on average → prediction, not reality

(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. On average, a unit increase in x_1 is associated with an increase of 5.7 units in the response. In reality, a point of a unit greater x_1 value may have a lesser or greater effect on the response, when compared to a x_1 value that is a unit lower

(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 then conclude that at least one of the means 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

(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. Increasing the significance level will increase the probability of 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.

False. Correlation is a measure of the association between any two QUANTITATIVE 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.

In analyzing the studentized residuals of a regression model, one is able to determine how influential each individual datapoint is in relation to the predictors (leverage) as well as the response (obs - exp). When looking at mere residuals, only the latter can be gauged.

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, which is a function of both studentized residuals and leverage. As such, it encapsulates the information provided in the other two statistics, providing a summary of which points are influential in regards to the predictors and the response variable.

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
```

reference = General Mills

grand mean = 106.97

group effect (R) =

111.364 + (-16.364) = group mean

group effect = group mean − grand mean

7. (3 points)

(a) What are the error degrees of freedom based on this model? 70 df

(b) What is the reference level? 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?

$(111.364_{cal}$ – $16.364_{cal})$ – 106.97_{cal} = group effect for Q

reference mean; difference btwn quaker mean & reference; grand mean

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 average change in the effect of [redacted] a unit [redacted] increase in sugar on the response, per unit change in protein, while keeping manufacturer constant.

- average cost of tuition
  - private / public
- # teachers
- liberal arts / com. college / tech / historically affiliated

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) An inquiry that could be investigated using a SLR model, given our data, could concern whether the number of teachers (full-time instructional staff) employed at a university relates to its average per-semester tuition.

$$\text{Tuition} = \beta_0 + \beta_1(\text{Teachers}) + \epsilon$$

where β_1 represents the avg. effect of a unit increase in teachers on tuition

$$H_0: \beta_1 = 0$$

b) Does the average number of teachers employed by an institution vary across different university types (e.g., liberal arts, historically affiliated college)?

where i represents categorical level: liberal arts = 1, community college = 2, tech/vocational = 3, institutionally affiliated = 4

$$\text{Teachers} = \mu + \alpha_i + \epsilon$$

where μ is the overall average number of teachers at an institution and α_i is the difference between the overall mean and the mean for categorical level i

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

c) Can average cost of [undergraduate, per-semester] tuition be predicted using number of teachers employed by an institution and its status as a public or private university?

$$\text{Tuition} = \beta_0 + \beta_1(\text{Teachers}) + \beta_2(\text{Private}) + \epsilon$$

where β_1 is the average effect of a unit change in teachers on tuition* and β_2 is the average change in tuition when institution type changes from public to private*

* when keeping all other variables constant

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

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)

- Constant Variance – the group deviations are not massively different ✓
- Constant and Additive Effects – the group effects do not appear to scale in a multiplicative manner. While we cannot confirm whether or not they depend on a lurking variable, it appears as though they are constant. ✓
- Normality – there is large tailing at both the higher and lower quantiles of our residuals when comparing to expected values. Must be skeptical ✗
- Independence – We can assure that the production of one cereal does not interfere with the production of another. ✓
- SRS – our sample was selected randomly ✓

b)

H_0 : $\mu_K = \mu_N = \mu_P = \mu_Q = \mu_R = \mu_G$

H_A : Any of μ_i are not equal

where μ_K represents the average caloric content of Kelloggs, μ_N represents that of Nabisco, μ_P represents that of post, μ_Q represents that of Quaker Oats, μ_R represents that of Ralston Purina, and μ_G represents that of general mills.

c) Given that the p-value for the model ANOVA F-test assessing caloric content across cereal type was statistically significant (pval = .027), there is sufficient evidence to conclude that the average caloric content of ~~[illegible]~~ cereal ~~[illegible]~~ is not equal across the six included brands. This being said, we cannot, from this test, conclude that the averages for each brand are all different from one another. In order to investigate such, post-hoc pairwise comparisons between groups should be performed. It is also key to note that, due to our model's inability to sufficiently meet the ANOVA condition of Normality in errors, our ~~[illegible]~~ conclusions may not be entirely reliable.

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. The conclusions each person reaches concern different inquiries. Person A produces an MLR model with Arsenic as the response, while person B finds how each included predictor is correlated with Arsenic on its own. While the adjusted R^2 value for person A's model is lower than the R^2 for the model predicting Arsenic on Year, this is not an impossible occurrence. It may be that Year and Miles contain redundant information, and thus the inclusion of a second predictor is not necessary. When adjusting the R^2 value for an MLR model, the number of included predictors is taken into account. Therefore if the amount of response variability explained in a model remains similar when a new predictor is added, the value for R^2_{adj} will decrease. This could explain how the R^2 value for the model predicting Arsenic on Year could decrease when Miles is included as an additional predictor in the model.

divides SSE by df for error when calculating 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