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: Koji Flynn-Do

Swarthmore Username: Kflynn01@...

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

ref = pigeon.

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

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? → $x_3 = 0$, so $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ 2)

b) Does the effect of age on a bird's weight depend on what type of bird it is? int. terms, so $\beta_4 \neq \beta_5$ 4)

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? β_1 vs. ref 1)

d) Given we are only comparing pigeons, is the effect of age on a bird's weight statistically significant? $\beta_0 + \beta_3 x_3$, $x_1 = x_2 = 0$ 3)

e) Is there statistically discernible evidence of a linear relationship between bird age and type and bird weight? anything. 5)

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. If two preds are collinear, then they capture similar info. So removing one will ↑ magnitude of the other's coefficient.

(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. $\beta_1 = 5.7$ is the average, not precise, change associated w/a one unit ↑ in x_1. So it could be 5.7, or some other nearby value.

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

True. Resid df $= n - k - 1$, so $n \uparrow \rightarrow \uparrow$ resid df

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 conclude that one pair is diff – we don't know which yet

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

True. $MS_{Mod} > MS_{Error}$ by a bunch, because F-Stat is $\frac{MS_{mod}}{MS_{Error}}$

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

False. We already know this from ANOVA F-test. Post-hoc pairwise tells us which pairs are sig. diff.

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. Type I is incorr. rejecting H_0. Lower α means fewer null get rejected. So "increase" → "decrease"

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

True. Tiny differences, with sufficiently large sample sizes, will be picked up as statistically (but not necessarily practically) signif. Eg US birth by sex is 51-49 male, and we can detect tiny diff. be really big.

(c) Correlation is a measure of the association between any two variables.

False. Correlation must be between quantitative variables.

Section 2: Short answer questions

5. (4 points)

Unusual & influential data

Briefly describe a benefit of analyzing the studentized residuals of a regression model rather than just analyzing the observed residuals.

Stud. resids delete the data point in question (the i^{th}) then find resids. This prevents a single unusual/influential point from pulling the regression line unduly toward itself, which skews resids, but really is just about that one pt

6. (3 points)

Cook's

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. Recall, Cook's distance is a combination of both studentized resids ("how off is the prediction?") and leverage ("how much could this pt influence the overall regression?"). We want both high lev & prediction error, because that <u>together</u> is concerning.

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
```

$n-k-1$

$n = 76 \rightarrow$ error df $= 70$

$k = 5$

7. (3 points)

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

(b) What is the reference level?

a) Error df = 70. Recall error df is sample size minus levels minus one. Here: 76 − 5 − 1 = 70. Also shown in R output.

b) Reference: General Mills. First alphanumerically, also not shown as one of the coefficient vars.

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 − 16 = 95

Reformulate in terms of group effects: $Y = \mu + \alpha_j + \epsilon$, $j = 1, \ldots, m$

we want $\hat{\alpha}_j$. $\hat{Y} = 111.381 - 16.364$ } so $\hat{\alpha}_j = 95 - 106.97$.

$\mu / \bar{y} = 106.97$ }

That is, the estimated group effect, $\hat{\alpha}_j$, of Quaker Oats on calories is roughly ~ -12 calories per serving.

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.)

Now: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \text{sug.} + \hat{\beta}_2 \text{prot} + \hat{\beta}_3 \text{sug} \times \text{prot} + \text{indicators}$.

The interaction term, which can be represented as $\hat{\beta}_j$ sugar × protein, models the (potential) phenomenon wherein the relationship btwn sugar → calories does not just depend on sugar, but varies depending on how much protein is present.

interaction for diff slope dep on level of each

The coefficient $\hat{\beta}_j$ would then be indicative of how the sugar → cal relationship changes w/ different levels of protein, or how protein → cal changes w/ diff sugar levels, allowing for simultaneous changes in other variables (incl. indicators). See below for example

Toy eqs.

$\hat{\beta}_0 + \hat{\beta}_1 \text{sug} + \hat{\beta}_2 \text{prot} + \hat{\beta}_3 \text{sug} \times \text{prot} + \text{indic.}$

suppose prot = 5g. → $(\hat{\beta}_0 + 5\hat{\beta}_2) + (\hat{\beta}_1 + 5\hat{\beta}_3) \text{sug.} + \text{indic.}$

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; cost (full-time)

(b) an ANOVA model; cost (private/public)

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

ft staff (cost, private/public)

a) SLR Is there a statistically detectable linear relationship between the percentage of full-time instruct. staff employed at a college and the avg. cost of tuition?

Model: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, where Y is avg. cost of tuition and X_1 is % staff ft.

H_0: $\beta_1 = 0$

b) ANOVA Is there a statistically detectable difference in group means of avg. cost of tuition between private and public institutions?

Model $Y = \mu_j + \varepsilon$ where $\mu_j = \mu + \alpha_j$ and $j = 1, 2$, where Y is avg cost of tuition, μ_1 is group mean for private inst, μ_2 is group mean for public inst.

H_0: $\mu_1 = \mu_2$

c) MLR Is there a statistically detectable relationship b/twn the combination of avg. cost of tuition and private vs. public status on the % of full time instr. staff? That is, does a model w/ both variables included perform better in a statistically detectable way than just the mean?

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where X_1 is % full-time staff, X_2 is indicator var. $X_2 = 1$ if private, $X_2 = 0$ if public

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.

Effects
1. constant effects (exclusive)
2. additive (not multiplicative)

(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.) – not good on var!

ANOVA overall $\alpha_j = 0 \ \forall j$

(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) I don't think conditions are met. For one, it appears that variance is very different across groups, based on the boxplot. Eg. example, the spread of resids for Kelloggs is much wider than that of Nabisco or Post. Moreover, the normal quantile plot shows non-normal behavior at the tails. With respect to effects: if group effects are constant, that means they do not vary depending on other variables not included in the model, eg. depending on sugar content. We could check this by looking at other relationships between calories and variables.

b) We do an ANOVA overall F-test. First, the model in group effects form: $Y = \mu_j + \alpha_j + \varepsilon$ with α_j as the effect for each group, $j = 1, \ldots, 6$. 1 corresponds to General Mills, 2 to Kelloggs, etc. We want to know if any pair of avg calories by manufacturer is statistically detectable. Our null hypothesis is that none of the effects are non-zero, or $H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_6 = 0$. Our alternative would be that at least one α_j is non-zero, or H_a: at least one of $\alpha_1, \alpha_2, \ldots, \alpha_6 \neq 0$.

c) If we take the test literally, we can reject the null in favor of the alt at $\alpha = 0.05$. The p-val for the ANOVA overall is 0.02724, which is to say that the prob. of observing an F stat (which is based on the difference between groups and degrees of freedom) this extreme is 0.02724. So we would conclude that at least one α_j is non-zero.

However, we should not trust the test because, in my view, crucial conditions have not been met, as discussed in part (a). We might take these differences as suggestive and go try to design a better study, but we should not simply take the results of the test at face-value.

High corr → high adj R^2 for SLR
Like, just year w/ $r_1 = 0.77 \rightarrow r^2 = \sim 0.6$

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 think the conclusions are contradictory I don't think their figures are consistent. First, suppose we fit an SLRM w/just year. Then, since $r_1 = 0.77$, the R^2 for this model would be ~ 0.6. We know that adding more predictors can never ↓ the un-adjusted R^2, so going from this SLRM to Person A's MLRM can only ↑ R^2. That means the downward adjustment from including 1 more predictor must account for the difference between SLRM R^2 of ~ 0.6 and the MLRM adj. R^2 of 0.26.

Can the adjustment factor do such a thing? I think no. Recall: $R^2 = 1 - \frac{SSE}{SS_{Tot}}$, adj $R^2 = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SS_{Tot}}{n-1}}$.

Rewrite adj $R^2 = 1 - \frac{SSE}{SS_{Tot}} \times \frac{n-1}{n-k-1}$. So they are identical, except we multiply the $\frac{SSE}{SS_{Tot}}$ term by this $\frac{n-1}{n-k-1}$ term.

Here, $n=70$, $K=2$, so $\frac{n-1}{n-k-1} = \frac{69}{67}$, or ~ 1.03. For the SLR, $R^2 = \sim 0.6$, so $\frac{SSE}{SS_{Tot}} = \sim 0.4$.

Assuming we got no unadj R^2 improvement from adding Miles (which is implausible), that would mean our $\frac{n-1}{n-k-1}$ factor would need to be: $0.26 = 1 - 0.4 \times 1 \Rightarrow f = \frac{0.74}{0.4}$, or roughly 2.

(0.26 = adj R^2, 0.4 = $\frac{SSE}{SS_{Tot}}$, 1 = $\frac{n-1}{n-k-1}$)

But we just shown that $\frac{n-1}{n-k-1} = \frac{69}{67} = \sim 1.03$, not 2! So something is inconsistent in Person A & B's conclusions. They are contradictory.

I should note that this answer feels wrong. It feels like I should have identified some missed way to reconcile the conclusions, because why else would this be asked? So, maybe I'm wrong and actually you can't go from r to R^2 in the way I described. Or maybe A had Year coded as factors & not numbers. Or maybe I'm missing something obvious?

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