962200891

# 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:** Brandon Cramblit

**Swarthmore Username:** bcrambl1

**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

$$Y = \overset{\text{pigeon}}{\beta_0} + \overset{\text{sparrow}}{\beta_1 x_1} + \overset{\text{finch}}{\beta_2 x_2} + \overset{\text{age}}{\beta_3 \underset{0}{x_3}} + \overset{\text{sp age}}{\beta_4 x_1 x_3} + \overset{\text{finch age}}{\beta_5 x_2 x_3} + \epsilon,$$

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.

True,

(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. If the model has more than 1 predictor, then a unit increase in $x_1$ is associated with increase in $y_1$ by 5.7 units, when keeping all other predictors constant.

(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 conclude that mean for at least 1 group is different from the mean of other groups.

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

$\frac{MSM}{MSE}$ True, because then we can say the difference in mean is likely due to variability between groups.

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

False because ANOVA can come to conclusion stated above. Post-hoc pairwise analysis could tell which of the means are different from the other means, after ANOVA rejects the Ho that all means are the same across four groups.

**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 ($\alpha$) will increase the probability of making a Type 1 Error.

type I : reject true Ho

False. increasing $\alpha$ will increase the probability of making 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 can be measured between two quantitative variables, but not categorical 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.

For example, if we want to see which data point could be an influential point based on observed residuals, we will not know how large a residual ~~for~~ one datapoint is relative to the model and other datapoints' residuals. By studentizing the residuals, we can compare residuals with other datapoints and model, and indicate that the datapoint is moderately influential if the studentized residual >2 and extremely influential if >3.

**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 because it is a measure that indicates which datapoint may be influential based on both leverage values and studentized residuals. Moderate/extreme cook's distance could tell that the datapoint is influential on the basis of residuals and leverage.

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
```

**7.** (3 points)

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

(b) What is the reference level?

(a) $n-k-1 = 96-5-1 = 70$ = error df

(b) 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?

$Y_i = \mu + \alpha_Q + \varepsilon$ reference mean = 111.364

$\beta_0 = 111.364$

$\beta_Q = -16.364$

$\mu = 106.97$

$(111.364 - 16.364) - 106.97 = \alpha_Q$ = group effect for Quaker Oats cereal brand.

**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.)

Interaction between 2 numeric predictors: sugars and protein.

The average effect of 1 gram increase of cereal on average calories perserving increases by the value of the coefficient of interaction term, "sugar:protein", for 1 gram increase of protein in cereal.

## 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) SLR: $\widehat{Ave\_cost\_tuition} = \hat{\beta}_0 + \hat{\beta}_1(pub\_inst)$ where $pub\_inst = \begin{cases} 1 \text{ if public institution} \\ 0, \text{ otherwise} \end{cases}$

reference level = private institution

$\widehat{Ave\_cost\_tuition}$ represents predicted average tuition based on the type of institution,

Research question: Is there a significant difference in average cost of tuition each semester between public and private institutions?

$H_0: \beta_1 = 0$, There is no difference in average cost of tuition between public and private institutions,.

(b). ANOVA: $Y = \mu + \alpha + \epsilon$

Y represents the predicted cost of tuition based on types of school.

reseach question: Is at least one of the liberal arts, community, technical/vocational school or school affiliated with certain group differ in terms of tuition each semester?

$\alpha_{lib}$ = group effect for liberal arts college

$\alpha_{com}$ = group effect for community college

$\alpha_{t/v}$ = group effect for technical/vocational school.

$\alpha_{iacg}$ = group effect for school institutionally affiliated with certain group.

$H_0: \alpha_{lib} = \alpha_{com} = \alpha_{t/v} = \alpha_{iacg}$

there is no difference in average tuition each semester among the 4 types of schools.

(c) $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(lib) + \hat{\beta}_2(com) + \hat{\beta}_3(tec)$ where $lib = \begin{cases} 1 \text{ if liberal arts college} \\ 0 \text{ otherwise} \end{cases}$

$com = \begin{cases} 1 \text{ if community college} \\ 0 \text{ otherwise} \end{cases}$

$tec = \begin{cases} 1 \text{ if tech/vocational school} \\ 0 \text{ otherwise} \end{cases}$

reference level = institutionally affiliated with certain groups,

and $\hat{Y}$ represents the predicted average cost of tuition each semester based on the type of school.

research question: Is at least one of the liberal arts, community, technology/vocational school, or college that is institutionally affiliated with certain groups differ from one another in terms of average cost of tuition each semester?

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$, there is no difference in average cost of tuition each semester among the 4 types of schools.

**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) zero mean: ✓

constant variance is violated because boxplot shows the size of box are largely different, indicating that the variance of datapoints for different levels are not constant.

~~A~~ Additive and const effect is met because we donot have interaction term in the model and the datapoints donot seem to have multiplicative increase across different levels.

Model shows constant effect because the effect of one cereal type of calories per serving doesn't depend on another cereal type.

Independence: no reason to suspect dependence. No locational, temporal, or genetic issues, and no reason to suspect bias in sampling.

normality: Normal quantile plot shows that there is high distribution of large residuals and low distribution of low residuals than expected as the points are below the line in lower quantile and points are above the line in upper quantile, so residuals are not normally distributed.

randomness: No reason to suspect nonrandomness in the sample as the cereals sample was obtained through random sample.

(b) $\widehat{\text{calories per serving}} = \hat{\beta}_0 + \hat{\beta}_1(\text{Manufacture K}) + \hat{\beta}_2(\text{manufacture N}) + \hat{\beta}_3(\text{manufacture P}) + \hat{\beta}_4(\text{manufacture Q}) + \hat{\beta}_5(\text{manufacture K})$

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, where $\beta_1$ = coefficient for kellogs

There is no difference in average calories per serving among 6 types of cereals.

$\beta_2$ = coefficient for Nabisco, $\beta_3$ = coefficient for Post

$\beta_4$ = coefficient for Quaker oats

$\beta_5$ = coefficient for Ralstone Purina.

reference level = General Mills.

$H_A$: Not $H_0$.

at least one of the cereals have different average calories per serving from other cereals.

(c) P-value for Overall F-test is 0.02724, which is lower than 0.05, so at alpha level of 0.05, we conclude that there is evidence to sugest that at least one of the cereal manufacture has an average calories per serving different from other cereal manufacturer.

**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:

$$Cor(Arsenic, Year) = \rho_1; \quad Cor(Arsenic, Miles) = \rho_2$$

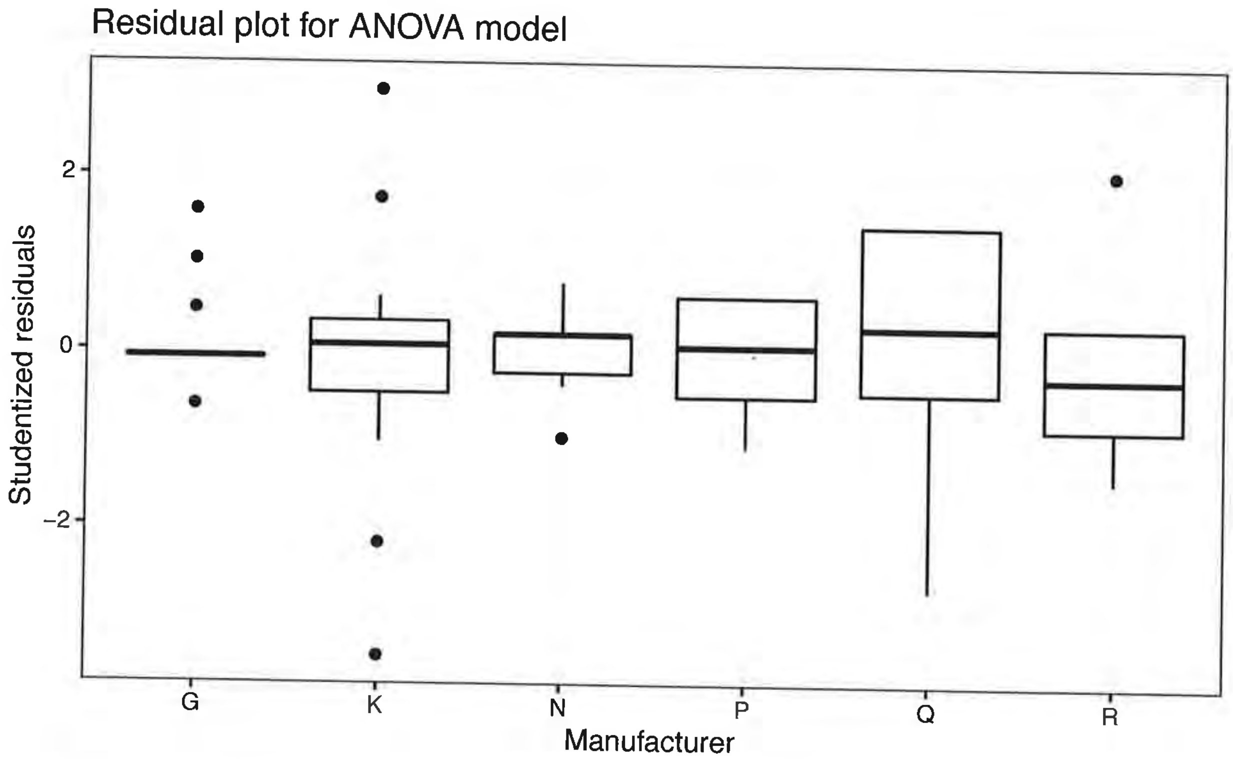
and estimates each with their sample correlations $r_1 = 0.77$ and $r_2 = -0.34$. Are the two people's conclusions contradictory? Explain your answer.

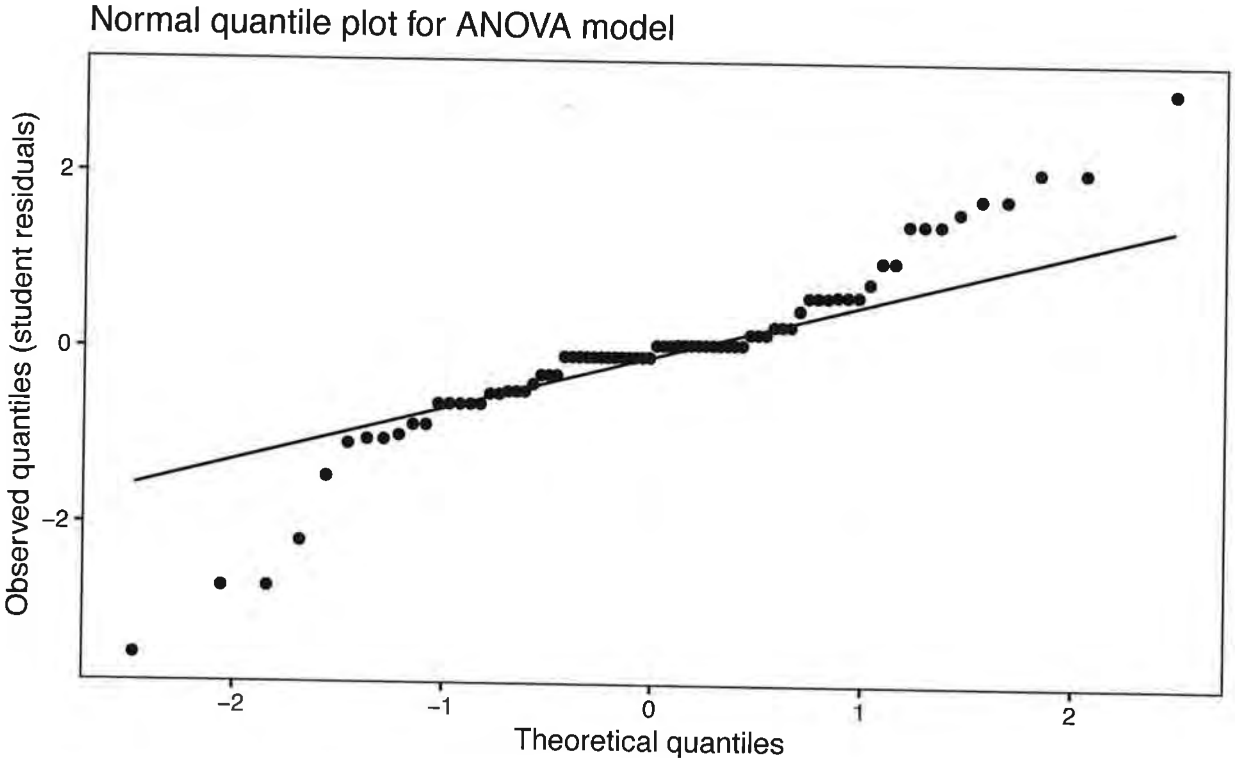
These two finding are not contradictory. $R^2$ od 0.26 indicates that 26% od variability in Arsenic in local wells can be explained by Year and distance from the well to the nearest mining site in miles. $r_1 = 0.77$ and $r_2 = -0.34$ only tells us that Year and Miles variable correlate with Arsenic, finding from person A could be used to say that there are other potential variables that could be included in the model to explain more od the 74% od variability in Arsenic. Any predictor could largely correlate with the response, but only explain a low fraction of variability in the response if there are other predictors that explains variability in the response.

## 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


Residual plot for ANOVA model
Studentized residuals
G
K
N
P
Q
R
Manufacturer


Normal quantile plot for ANOVA model
Observed quantiles (student residuals)
Theoretical quantiles