STAT 4010/5010, Statistical Methods and Applications II, Exam #2 Review

1. True or false:

   (a) As a reminder, the effects model for one-way ANOVA is given as:

   $$Y_{i,j} = \mu + \tau_j + \varepsilon_{i,j}, \text{ where } \varepsilon_{i,j} \overset{iid}{\sim} N(0, \sigma^2).$$

   In this model, μ can be interpreted as the mean of the response over all units in the sample.

   Solution: True. This is the effects formulation of one-way ANOVA¿

   (b) A one-way ANOVA model with a J-level factor is a multiple linear regression model with $J-1$ predictor/explanatory variables, and a continuous response.

   Solution: True. Having J predictors would result in a non-identifiable model.

   (c) Analysis of Covariance (ANCOVA) can help answer the question: are there differences, with respect to the population mean of a response variable, across groups, adjusting for several continuous variables thought to be correlated with the response?

   Solution: True. This is how we defined ANCOVA.

   (d) In the context of one-way ANOVA,

   $$\sum_{i=1}^{n_j} (Y_{i,j} - \bar{Y_{.j}})^2$$

   is a measure of within group variability in the sample.

   Solution: True. This measures the amount of variability around the (sample) mean in the j^{th} group.

   (e) A very small p-value suggests that the differences with respect to the mean of a continuous variable across groups of experimental units is very large.

   Solution: False P-values are affected by effect size (size of the differences across groups) and sample size.

   (f) The variance inflation factor for a linear model is defined as

   $$\frac{1}{1-R^2},$$

   where R^2 is the coefficient of determination of the regression model.

   Solution: False. $VIF_j = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of determination for a regression with $\mathbf{x_j}$ as the response and all other original predictor variables as predictors.

   (g) Very low pairwise correlations between predictors implies that multicollinearity is not an issue with the model.

   Solution: False. There may be more complicated dependencies.

(h) If the predictors are orthogonal (zero pairwise correlations), then $Var\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{i,j}-\bar{x}_j)^2}$, where $\bar{x}_j$ is the mean of the j^{th} predictor, and σ^2 is the regression error variance.

Solution: True. If the predictors are orthogonal then $R_j^2 = 0$, which makes $VIF_j = 1$.

2. Explain why the following situations might arise. There may be more than one correct answer. If you don't think that they can arise, explain why. If you think they always arise, prove it!

(a) A multiple linear regression is fit with $p = 75$ predictors and $n = 150$ data points. The full F-test p-value is greater than α but several of the predictors' t-test p-values are less than α.

Solution: This can happen when none of the predictors are linearly related to the response, but several of the individual t-tests result in type I errors (which is more likely than the individual rate of type I error!).

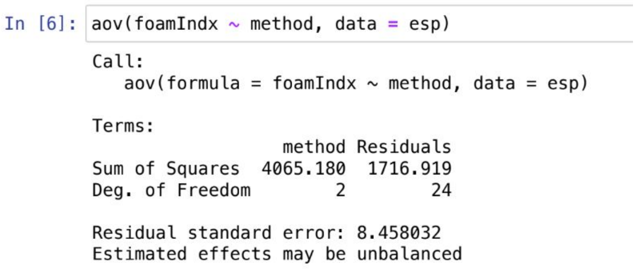
(b) Suppose that a One-Way ANOVA is conducted, with a factor of 5 levels. A contrast is constructed using $\mathbf{c} = (0, 1, -1/3, -1/3, -1/3)$ and $\mu = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$, and the associated null hypothesis is rejected.

Solution: $H_0 : \mu_2 - 1/3\mu_3 - 1/3\mu_4 - 1/3\mu_5 = 0 \iff H_0 : \mu_2 = 1/3(\mu_3 + \mu_4 + \mu_5)$. If we reject the null, then we have evidence that $H_1 : \mu_2 \neq 1/3(\mu_3 + \mu_4 + \mu_5)$

(c) A MLR plot of the residuals vs. fitted values shows something other than random scatter around zero.

Solution: Some regression assumption is violated. For example, if there is some nonlinear trend, there are likely missing predictors. If the plot shows a trumpeting pattern, then there is evidence of a non-constant response variance.

3. Using the code output below, calculate the (full) F-statistic.



```
In [6]: aov(foamIndx ~ method, data = esp)

        Call:
           aov(formula = foamIndx ~ method, data = esp)

        Terms:
                          method Residuals
        Sum of Squares  4065.180  1716.919
        Deg. of Freedom        2        24

        Residual standard error: 8.458032
        Estimated effects may be unbalanced
```

Solution:

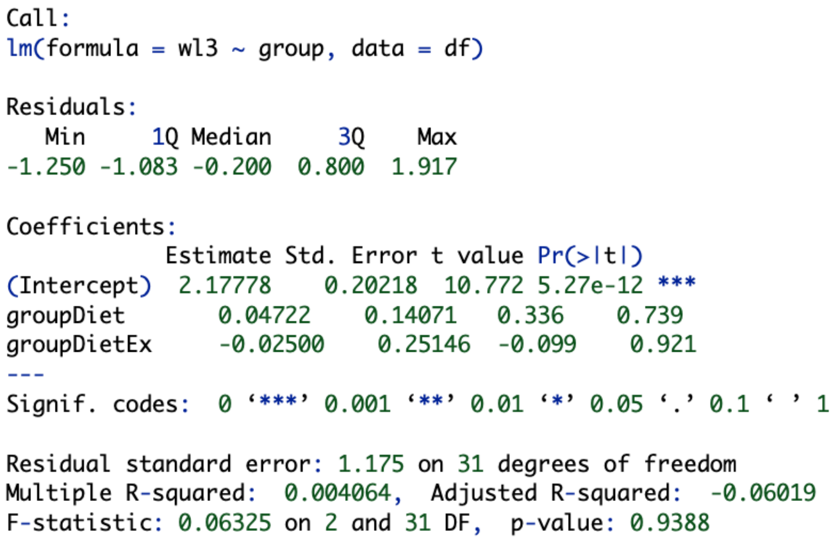
$$\frac{ESS/df_{ESS}}{RSS/df_{RSS}} = 28.41262$$

4. Researchers conducted an experiment to learn the effect of diet and exercise, on weight loss. In the experiment, $n = 34$ people were randomly assigned to one of three different groups (factor levels):

- a control group, where individuals were not given the prescribed diet or partaking in the exercise plan ($n_1 = 12$).

- a diet group, where individuals ate a prescribed low carb diet ($n_2 = 12$);
- a diet and exercise group, where individuals ate a prescribed low carb diet and engaged in at least 60 minutes of cardiovascular exercise four times per week ($n_3 = 10$).

Individuals were monitored, and their weight loss in pounds (response) was recorded after three months (i.e., a positive response measurement means that an individual lost weight) . The table below summarizes a one-way ANOVA regression model, where the response variable is wl3 (weight loss after three months) and the predictor is the group factor. Assuming that the model is correct, which of the following are correct? **It is possible that more than one is correct.**



```
Call:
lm(formula = wl3 ~ group, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-1.250 -1.083 -0.200  0.800  1.917 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.17778    0.20218  10.772 5.27e-12 ***
groupDiet       0.04722    0.14071   0.336    0.739    
groupDietEx    -0.02500    0.25146  -0.099    0.921    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.175 on 31 degrees of freedom
Multiple R-squared:  0.004064,	Adjusted R-squared:  -0.06019 
F-statistic: 0.06325 on 2 and 31 DF,  p-value: 0.9388
```

Solution in bold

(a) There is statistical evidence that there are population level differences with respect to the mean weight loss across diet and exercise groups.

(b) **There is statistical evidence that there are sample level differences with respect to the mean weight loss across diet and exercise groups.**

(c) **The average amount of weight lost in the control group sample was 2.17778 pounds.**

(d) **The average amount of weight lost in the diet group sample was 2.225 pounds.**

(e) The average amount of weight lost in the diet group sample was 0.04722 pounds.

(f) On average, in the sample, individuals in the diet and exercise group gained weight.

(g) **On average, in the sample, individuals in the control group did better (in terms of weight loss) than individuals in the diet and exercise group.**

5. Describe the difference between a designed experiment and an observational study. Articulate how you might learn about causality with each.

   Solution:

   A designed experiment is a study in which researchers have control over a ”treatment” variable (or set of treatments), and observe the changes in the response. We might contrast this with an observational study, in which researchers cannot control the administration of any treatments on units in a sample. Instead, researchers collect and analyze data without changing existing conditions. Experiments are helpful in learning about causality if we randomly assign the treatment to units in the sample. Learning about causality using observational data can be difficult, and is not possible without imposing additional assumptions on the data. (Here’s a good resource with a chapter on causal inference in observational studies: https://bit.ly/3d69NQz.)

6. Load the hotel data into R. Here’s a description of the data:

```
Dataset:  hotel_energy.csv

Source: Y. Xin, S. Lu, N. Zhu, W. Wu (2012). "Energy Consumption Quota of
Four and Five Star Luxury Hotels Buildings in Hainan Province, China,"
Energy and Buildings, Vol. 45, pp. 250-256.

Description: Energy Consumption and attributes for 19 Luxury Hotels.

Variables/Labels
Hotel Id   (hotel)
Energy Consumption in Kilowatt-hours   (enrgcons)
Area in square meters   (area)
Age in years     (age)
Number of guestrooms    (numrooms)
Occupancy Rate in percent   (occrate)
Effective number of guestrooms  (effrooms = numrooms*occrate/100)
```

   Solution in Jupyter Notebook

   (a) Perform simple linear regression using energy consumption as the response and area as the predictor. Report a summary of your model.

   (b) Find the p-value for the *upper-tailed* t-test associated with the slope parameter. (Note that this isn’t a test directly in the output)

   (c) Now perform multiple linear regression with number of guest rooms, area, and age as predictors (same response).

   (d) Plot the residuals against the fitted values (for the MLR model), and the fitted values against the observed response. Do these plot suggest that any regression assumptions are violated?

   (e) Look for evidence of successively correlated errors when ordering the data by the age variable.

(f) Using the multiple linear regression model, predict energy consumption at the sample mean values of guest rooms, area, and age.

(g) Compute the mean squared error for the model.