STATISTICS 110/201, FALL 2017 LECTURE B, FINAL EXAM

NAME:_____KEY_____ Homework code :_____ Seat: _____

Open notes, calculator required. Your exam should have 7 pages and an Appendix with R output, handed out separately. Make sure you have them all. Each part of each problem is worth 4 points unless specified otherwise. Use the back of the pages if you need more space, but *tell us to turn the page over and look.*

The following scenario is for Questions 1 to 8. R output is contained in the separate Appendix.
College departments often run multiple offerings of the same introductory course each quarter. An experiment is conducted during a particular quarter in which students are randomly assigned to one of three versions of the same course, all held at the same time, but each using a different method of teaching. The three methods are (1) Active Learning, (2) Lecture using slides, and (3) Lecture writing on the board. Fifty students were randomly assigned to each method. The Department would like to determine whether there are statistically significant differences in the students' average final exam scores after being taught with the three different methods. Assume students taking the course that quarter are representative of all students who ever take the course.

1. (2 points each) Specify each of the following for this scenario:

a. The response variable.

Final exam score.

b. The factor, and the number of levels it has.

Teaching method, 3 levels

c. Whether the factor is fixed or random. (You do not need to explain your answer.)

Fixed.

2. Using the R output in the Appendix, provide numerical values for each of the unfilled boxes in the ANOVA table below. A few values are filled in already. Boxes with 'X' don't need to be filled in.

One way to do these is using the calculations shown in the table, done is this order: Read df for Group and Residual, F and p-value all from output; then MSGroup = F×MSE; SSGroup = MSGroup×df; SSE = MSE × df; SSTotal = SSGroups + SSE. Other methods are okay and might differ slightly due to rounding.

	Df	**Sum Sq**	**Mean Sq**	**F-value**	**Pr(>F)**
Group	2	1296 (648×2)	648 (3.445×188.1)	3.445	0.03452
Residuals	147	27,651 (188.1×147)	188.1	X	X
Total	149	28,947 (1296+27,651)	X	X	X

3. In class we covered three versions of the population model for one-factor analysis of variance. This question explores those models.

a. For the cell means model, the population model is $Y_{ik} = \mu_k + \varepsilon_{ik}$. Define what the parameter μ_k represents in the context of this problem for each of k = 1, 2, 3.

μ_k = Population mean final exam score for all students who would ever take the course, if taught using method k, where:
Method 1 = Active learning
Method 2 = Lecture using slides
Method 3 = Lecture writing on the board.

b. Write the factor effects model. Population model: $Y_{ik} = \mu + \alpha_k + \varepsilon_{ik}$

List all of the parameters used in the model and express them in terms of the μ_k defined in Part (a).

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3}{3}$$

$\alpha_k = \mu_k - \mu$ for k = 1, 2, 3

Choose one of the parameters and explain in words what population quantity it represents, in the context of this problem.

μ is the mean final exam score for the population of all students who would ever take the course, averaged across the three teaching methods.

You could also choose one of the α_k. They represent the amount by which the population mean final exam scores for students taught with Method k exceeds the population mean final exam score for students taught with all methods combined. The difference is positive if the mean for Method k is greater than the overall mean, and negative if it is less than the overall mean.

c. Write the regression model, including the definition of any variables you use in the model.

Population model: $Y_{ik} = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_{ik}$

X_k = 1 *if the individual is taught with teaching method k, and 0 otherwise.*

List all of the parameters used in the model and express them in terms of the μ_k defined in Part (a).

$\beta_0 = \mu_1$, $\beta_2 = \mu_2 - \mu_1$, $\beta_3 = \mu_3 - \mu_1$

Choose one of the parameters and explain in words what population quantity it represents, in the context of this problem.

β_0 is the population mean final exam score for all students who would take the course with the Active learning method. The other β_k (k = 2 and 3) represent the difference between the mean for teaching method k and the mean for teaching method 1 (active learning).

4. (2 points) Of the three models defined in Question 3, which one is represented in the R output in the Appendix?

The regression model.

5. **a.** (2 points) Based on the results shown in the R output in the Appendix, which teaching method resulted in the highest average score on the final exam for the students in the sample: (1) Active Learning, (2) Lecture using slides, or (3) Lecture writing on the board?

Method 3, Lecture writing on the board.

b. (2 points) What was the sample mean on the final exam for the method you identified in Part (a)?

74.7 + 4.58 = 79.28

6. You should be able to use the R output in the Appendix to estimate the parameters for any of the versions of the model. Calculate the estimates of all of the parameters in the *factor effects* model.

$\hat{\mu}$ is the average of the 3 sample means; $\hat{\mu}_1 = 74.7,\ \hat{\mu}_2 = 74.7 - 2.52 = 72.18,\ \hat{\mu}_3 = 79.28$ so

$$\hat{\mu} = \frac{74.7+72.18+79.28}{3} = 75.39$$

$$\hat{\alpha}_1 = 74.70 - 75.39 = -0.69$$

$$\hat{\alpha}_2 = 72.18 - 75.39 = -3.21$$

$$\hat{\alpha}_3 = 79.28 - 75.39 = +3.89$$

7. Based on the output in the Appendix, give the value of the test statistic and p-value that would be used to test $H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$. (The notation here is the standard notation used in the version of the model that contains α_k.)

Test statistic value = *3.445*

p-value = *0.03452*

8. Based on the description of this study, could you conclude that the different teaching methods caused differences in the average final exam scores? Explain briefly.

Yes, because students were randomly assigned to the 3 teaching methods.

The following scenario is for Questions 9 to 13. Some R output is in the separate Appendix.
The data set **Cereal** accompanying the textbook contains data for 36 breakfast cereals. The response variable is Y = Calories per serving. The explanatory variables are Sugar = grams of sugar per serving, and Fiber = grams of fiber per serving. The Appendix provides a scatterplot of Sugar vs Fiber and regression results for the full model, which includes both Sugar and Fiber.

9. Does the estimated intercept have a useful interpretation in this example? If so, provide the numerical value and interpret it. If not, explain why not.

Yes, because the data set includes a case with Sugar = 0 and Fiber = 0, so it's in the range of the data. The estimated intercept is 109.3082. If a cereal has 0 grams of sugar and 0 grams of fiber per serving, it is predicted to have 109.3082 calories per serving.

10. (2 points each) Give numerical values for each of the following:

a. SSModel = *4567.2 + 4783.0 = 9350.2*

b. SS(Fiber | Sugar) = *4783.0*

c. The value of the test statistic for testing H_0: Reduced model with Sugar only vs H_a: Full model

Test statistic = *20.124*

11. The estimated coefficient for "Sugar" is 1.0050. Interpret that coefficient.

Adjusting for grams of fiber, for each additional gram of sugar in a serving of cereal, the calories per serving are predicted to increase by 1.0050.

12. Here are some results from fitting a simple linear regression model with Y = Calories, X = Sugar.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	87.4277	5.1627	16.935	<2e-16
Sugar	2.4808	0.7074	3.507	0.0013

Explain why the "Sugar" coefficient is statistically significantly different from 0 for this model, but not for the model with both Sugar and Fiber included, for which the *p*-value for Sugar was 0.134.

From the scatterplot, it's clear that Sugar and Fiber are (negatively) correlated. The model with both variables is testing whether sugar is useful in predicting calories <u>after</u> Fiber is already in the model. The single variable model is testing Sugar without accounting for Fiber.

13. (2 points each) Consider three possible models, defined as follows:

```
> Both <- lm(Calories ~ Sugar + Fiber, data = Cereal)
> Sug  <- lm(Calories ~ Sugar, data = Cereal)
> Fib  <- lm(Calories ~ Fiber, data = Cereal)
```

For each of the statistical terms listed, indicate whether the numerical value must be the same for all 3 models, for Sug and Fib only, or for none of the 3 models. Place an X in the appropriate column for each row.

Would be the same for:	All 3 models	Sug and Fib only	None of the models
Adjusted R^2			X
Degrees of freedom for SSE		X	
The estimate for β_0			X
SSTotal	X		
Hat values h_i			X
Predicted calories per serving for a cereal with Sugar = 0 and Fiber = 0			X

14. In a multiple regression scenario with a response variable Y and three explanatory variables labeled X_1, X_2, and X_3, the Variance Inflation Factor for the variable X_2 is 5. Explain how the other variables were used to calculate this VIF value.

Fit a model with X_2 = Response and explanatory variables X_1 and X_3. Let R^2 be the R-squared value for that model. Then VIF = 5 = 1/(1- R^2).

15. Consider a two-factor ANOVA situation with 2 levels for each factor and $n = 10$ observations in each combination of the factors. Suppose $\hat{\mu} = 70,\ \hat{\alpha}_1 = -3,\ \hat{\beta}_1 = +1,\ \hat{\gamma}_{11} = +2.$

a. (1 point each) Fill in the values for the cell means, showing your work in each cell.

	B_1	B_2
A_1	70 – 3 – 1 + 2 = 70	70 – 3 – 1 – 2 =64
A_2	70 + 3 + 1 – 2 = 72	70 + 3 – 1 + 2 = 74

b. (2 points) Would the values of SSA and SSB be <u>different</u> for the two commands shown below?

```
>summary(aov(Y ~ A*B))
>summary(aov(Y ~ B*A))
```

Circle your answer: YES **NO**

MULTIPLE CHOICE (2 pts each) *Circle the best choice*

1. In multiple regression, which of the following involves only the explanatory variables?
A. The variance inflation factors
B. SSTotal
C. The values of Cook's distance
D. None of the above. They all involve the response variable as well as the explanatory variables.

2. Suppose 3 explanatory variables (X1, X2, X3) are all correlated with each other. A regression model is fit using X1 and X2. If X3 is added to the model, which of the following will not change?
A. The leverage values
B. The value of the F statistic for the overall F test.
C. The value of adjusted R-squared
D. The value of SS(X1), when X1 is added to the model first

3. In two-factor analysis of variance with two levels for each factor, which of the following must be true about the interaction plot when interaction is present? (In what follows, "the lines" refer to the lines connecting the means for A1 and A2 for the two levels of Factor B.)
A. The lines must cross.
B. The lines are not parallel to each other, but they cannot cross.
C. The lines are not parallel to each other, and they may or may not cross.
D. Neither of the lines can be parallel to the horizontal axis.

4. In two-factor anova, $\sqrt{MSE}$ is used as an estimate of σ. In this context, what is σ?
A. The standard deviation of the population of Y values within each combination of the two factors.
B. The standard deviation of the population of all Y values.
C. The standard deviation of the mean of the Y values within each combination of the two factors.
D. The standard deviation of the mean of all Y values.

5. A scatterplot of Y versus X reveals a curved pattern. Under what circumstance would it be appropriate to transform the Y values either instead of or in addition to adding a quadratic term in the model?
A. If the curved pattern looks like a U shape with constant variance as X increases.
B. If the curved pattern looks like an inverted U shape with constant variance as X increases.
C. If the scatterplot indicates that the values of X are getting farther apart as X increases.
D. If the scatterplot indicates that the variance of the Y values is increasing as X increases.

6. Under what circumstances can a cause and effect conclusion generally be made based on a statistical study?
A. When participants are randomly sampled from the population.
B. When participants are randomly assigned to the treatments in an experiment.
C. When a randomized block design is used.
D. All of the above, because they all involve randomness.

Multiple choice continues on the next page.

The following scenario is for Multiple Choice questions 7 to 10:
A large medical organization runs hospitals in 3 different states in the Western United States. They conduct a study in which they randomly choose 4 of their hospitals in each state. In each hospital, they randomly choose 100 patients from each of 5 Age groups - under 21, 21 to 40, 41 to 60, 61 to 80, and over 80 years old, and collect data on how many nights they stayed in the hospital.

7. Which of the following is the list of factors for this study?
A. States, Hospitals, Age groups, Patients
B. States, Hospitals, Age groups
C. Hospitals, Age groups, Patients.
D. None of the above provides the correct list of factors.

8. Which of the following is true about the relationship between the factors?
A. Age groups are nested under Hospitals.
B. States are nested under the Western United States but crossed with Hospitals.
C. Hospitals are nested under States, but crossed with Age groups.
D. All factors are crossed with all other factors.

9. Which of the following is true about the factors?
A. The factors Hospital and State are both random.
B. State is a fixed factor, but Hospital is a random factor.
C. The factors Hospital and Patient are both random.
D. There are no random factors in this study.

10. Which of the following is true about "Patients" in this study?
A. Patients should be treated as a random factor.
B. Patients should be treated as a fixed factor.
C. Patients should not be treated as a factor because only one measurement was taken on each one.
D. Patients should be treated as a factor nested under hospital and age group.

APPENDIX: R Output for Statistics 110/201, Lecture B, Final Exam, Fall 2017

Output for Questions 1 to 8:

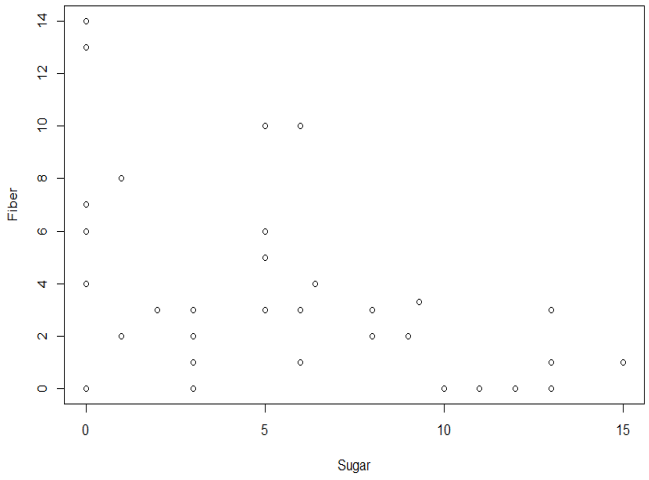
```
> FinalModel <- lm(Final ~ Method, data = classdata)
> summary(FinalModel)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   74.700      1.939  38.517   <2e-16 ***
Method2       -2.520      2.743  -0.919   0.3597
Method3        4.580      2.743   1.670   0.0971 .
---
Residual standard error: 13.71 on 147 degrees of freedom
Multiple R-squared:  0.04477, Adjusted R-squared:  0.03177
F-statistic: 3.445 on 2 and 147 DF,  p-value: 0.03452
```

Output for Questions 9 to 13:

The scatterplot shows values of two explanatory variables that might be useful for estimating calories per serving for 36 breakfast cereals:

Sugar = grams of sugar per serving

Fiber = grams of fiber per serving



Results for the Full model (Sugar and Fiber) for the Cereal data:

```
> Both <- lm(Calories ~ Sugar + Fiber, data = Cereal)
> summary(Both)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.3082     6.3913  17.103  < 2e-16
Sugar         1.0050     0.6546   1.535    0.134
Fiber        -3.7442     0.8346  -4.486 8.31e-05

Residual standard error: 15.42 on 33 degrees of freedom
Multiple R-squared: 0.5438, Adjusted R-squared: 0.5162
F-statistic: 19.67 on 2 and 33 DF,  p-value: 2.375e-06

> anova(Both)
Analysis of Variance Table

Response: Calories
          Df Sum Sq Mean Sq F value    Pr(>F)
Sugar      1 4567.2  4567.2  19.216 0.0001119 ***
Fiber      1 4783.0  4783.0  20.124 8.315e-05 ***
Residuals 33 7843.2   237.7
```