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

NAME:_________________________________ 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 6. R output is contained in the separate Appendix.
A large international company offers an intensive 10-week training course in a foreign language to its new employees. There are three instructors (A, B, and C) for the course, and the company wants to know if there is a difference among them in terms of effectiveness, as measured by scores on a final exam. An experiment is conducted in which 150 new employees are randomly assigned to the three instructors, with 50 assigned to each instructor. The same final exam is given to all of the 150 employees at the end of the course, and the scores are recorded. Assume that the 150 employees are representative of all employees who ever take the course. Define 3 indicator variables: InstructorA = 1 for employees taking the course with Instructor A and 0 otherwise; similarly for the indicator variables InstructorB and InstructorC.

1. Answer the following for this experiment.

a. (2 points) What is the response variable?

Final exam score.

b. There is one factor, “Instructor” with 3 levels. For Questions 2 to 6 Instructor will be considered to be a fixed effects factor. But for this question, describe a scenario for which Instructor would be considered to be a random effects factor in this experiment.

Instructor would be a random factor if the three instructors A, B, C were selected from a larger pool of potential instructors, and interest was not in the effectiveness of these 3 particular instructors. In that case, the company would be interested in variability among possible instructors.

2. The R output in the Appendix shows the results of using `aov` followed by `summary`. The model also could have been analyzed using the following commands:

```
> FinalLM <- lm(Final ~ InstructorB + InstructorC, data = FinalData2)
> anova(FinalLM)
```

The ANOVA Table below shows the structure that would result from doing that. Using the information in the Appendix and the information filled in below, provide numerical values for each of the unfilled boxes. Boxes with ‘X’ don’t need to be filled in.

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
InstructorB	1	1041.6	1041.6	5.7356	0.01788
InstructorC	1	198.82	198.82	1.0948	0.29714
Residuals	147	26696	181.6054	X	X

3. In class we covered three versions of the population model for one-factor analysis of variance. They were called the cell means model, the factor effects model, and the regression model. This question explores those models and the estimates of their parameters. When needed, you may use the indicator variables InstructorA, InstructorB, InstructorC defined in the introduction to this scenario. You do not need to redefine them.

a. (2 points) All three models include an error term. One condition is that all errors are independent and N(0, σ). Using the output in the Appendix, give a numerical value for the estimate of σ.

$\sqrt{MSE} = \sqrt{181.6} = 13.48$

b. Write the cell means model. Population model is $Y_{ik} = \mu_k + \varepsilon_{ik}$.

Provide numerical estimates for each of the parameters in the model.

They are the 3 sample means: $\hat{\mu}_1 = 75.8,\ \hat{\mu}_2 = 71.62,\ \hat{\mu}_3 = 78.62$ *so*

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

Provide numerical estimates for each of the parameters in the model.

$\hat{\mu}$ *is the average of the 3 sample means so* $\hat{\mu} = \frac{75.8 + 71.62 + 78.62}{3} = 75.35$

$$\hat{\alpha}_1 = 75.8 - 75.35 = 0.45$$

$$\hat{\alpha}_2 = 71.62 - 75.35 = -3.73$$

$$\hat{\alpha}_3 = 78.62 - 75.35 = 3.27$$

d. Write the regression model.

Population model is $Y_{ik} = \beta_0 + \beta_2 InstructorB + \beta_3 InstructorC + \varepsilon_{ik}$

Provide numerical estimates for each of the parameters in the model.

$\hat{\beta}_0 = 75.8,\ \ \hat{\beta}_2 = \hat{\mu}_2 - \hat{\mu}_1 = 71.62 - 75.8 = -4.18,\ \ \hat{\beta}_3 = \hat{\mu}_3 - \hat{\mu}_1 = 78.62 - 75.8 = 2.82$

e. For the regression model, choose any 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 Instructor A.
β_2 represents the difference between the mean for Instructor B and the mean for Instructor A.
β_3 represents the difference between the mean for Instructor C and the mean for Instructor A.

4. (2 points each) The goal of the experiment was to determine whether the three instructors differ in their effectiveness for teaching the foreign language. This problem addresses that question.

a. State the null and alternative hypotheses used to test whether the three instructors differ in their effectiveness. Specify which of the three versions of the model you are using, and use notation appropriate for that version in the hypotheses.

It is easiest to use the cell means model.
Hypotheses are H_0: $\mu_1 = \mu_2 = \mu_3$ and Ha: Not all μ_k are equal.

For the other versions of the model:
Factor effects model: H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$, and Ha: Not all α_k are 0.
Regression model: H_0: $\beta_2 = \beta_3 = 0$ and Ha: At least one of β_2 and β_3 is not 0.

b. Provide the numerical values for the test statistic and p-value used to test the hypotheses.

Test statistic value = *3.415*

p-value = *0.0355*

c. Give the conclusion for the test in the context of this situation using the 0.05 significance level.

Using the standard rejection criterion of $p < 0.05$, the null hypothesis can be rejected. Conclude that the population mean final exam score for at least one of the three instructors differs from the others. For the question of interest to the company, at least one of the instructors differs in effectiveness from the others.

5. The Appendix includes results of the Tukey procedure with overall 95% confidence level. For which pair(s) of instructors is there a statistically significant difference in mean final exam scores, if any? Explain how you know.

The mean final exam scores are significantly different for Instructors B and C. They are not significantly different for A and B or for A and C. You can tell in two ways. Two means are significantly different in the confidence interval for the difference does not cover 0, which is the case for Instructors B and C, but not for the other pairs. Another way to tell is that the means are significantly different if the p-value accompanying the difference is 0.05 or less. Again, that is true here only for the comparison of B and C, not for A and B or A and C.

6. Rounded to one decimal place, one of the Tukey confidence intervals is 0.6 to 13.4. Write a sentence or two interpreting that interval in the context of this experiment.

With 95% confidence that all three of the intervals cover the true population mean differences, we can say that the population mean final exam score for Instructor C is between 0.6 and 13.4 points higher than the population mean final exam score for Instructor B.

The following scenario is for Questions 7 to 12. Some R output is in the separate Appendix.
The data set **Houses** accompanying the textbook contains data for 20 houses that were sold in 2008 in a small Midwestern town. The response variable is Y = *Price* = the sales price in thousands of dollars. The two explanatory variables are *Size* = size of the house in 100s of square feet (ranging from 7.68 to 43.74), and *Lot* = size of the lot in hundreds of square feet (ranging from 73.61 to 253.51). The Appendix provides the following results:

- The correlation matrix for the 3 variables.
- Regression output for the model with both Size and Lot, called `Cost`
- Regression output for the model with Size only, called `CostSize`

7. For the model with both Size and Lot, 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.

No, it does not. The minimum house and lot sizes (Size and Lot) are given as 768 square feet and 7361 square feet, respectively. The intercept would only have a useful interpretation if 0 square feet was within the range of the data for both Size and Lot.

8. (2 points each) Give numerical values for each of the following. If you have to compute the value from other values you do not need to show your work, but could get partial credit if you do.

a. For the model `Cost,` SSModel = 40447 + 7601 = 48,048

b. For the model `Cost,` SS(Lot | Size) = 7601

c. For the model `CostSize,` SSModel = 40447

d. For the model `CostSize,` the predicted Price for a house with 1000 square feet =

$\hat{Y}$ = 64.544 + 4.82(10) = 112.744, but this is in thousands of dollars, so the predicted price is $112,744. (Given that "Price" was defined in thousands of dollars, it is acceptable to leave your answer as 112.744 as the predicted "Price.")

9. The estimated coefficient for "Size" in the model with both Size and Lots is 2.32. Interpret that coefficient value.

If two houses differ in size by 100 square feet (Size = 1) but they have the same lot size, the predicted sales price will be 2.32 thousand dollars (or $2320) higher for the larger house.

10. Calculate the VIF values for Size and Lot for the `Cost` model, which includes both variables.

First, find R^2 for the model with Size as the response and Lot as the explanatory variable. This is simply the correlation squared, or $(0.7668722)^2 = 0.59$. Then VIF $= \frac{1}{1-0.59} = 2.44$. The VIF value is the same for Size and Lot, because the correlation between them doesn't depend on which one is the explanatory variable and which is the response variable in the simple linear regression used to find R^2.

11. The results of the `Cost` model are shown using both the `anova` command and the `summary` command. With the `anova` command, the p-value corresponding to “Size” is 0.0005485, but with the `summary` command it is 0.2068. Explain why they are different. You can either explain in words, or explain by showing what null and alternative hypotheses are being tested in each case.

The anova command tests each variable in the order they are entered into the model, whereas the summary command tests each variable after all of the others are in the model. So in this case, the p-value of 0.0005485 is for the test of whether Size is a statistically significant predictor of price, with no other variables in the model. The p-value of 0.2068 is testing whether Size is a statistically significant predictor of price, after Lot is being used already to predict price.

12. (2 points each) Consider three possible models, defined as follows:

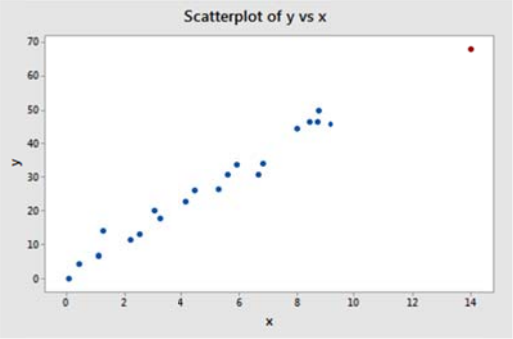
```
> Cost <- lm(Price ~ Size + Lot, data = Houses)
> CostSize  <- lm(Price ~ Size, data = Houses)
> CostLot  <- lm(Price ~ Lot, data = Houses)
```

The column on the left of the table below lists statistical terms that can be computed for each model. For each of the terms listed, indicate whether the numerical value must be the same for all 3 models, for CostSize and CostLot only, or for none of the 3 models. Put an X in the appropriate column for each row.

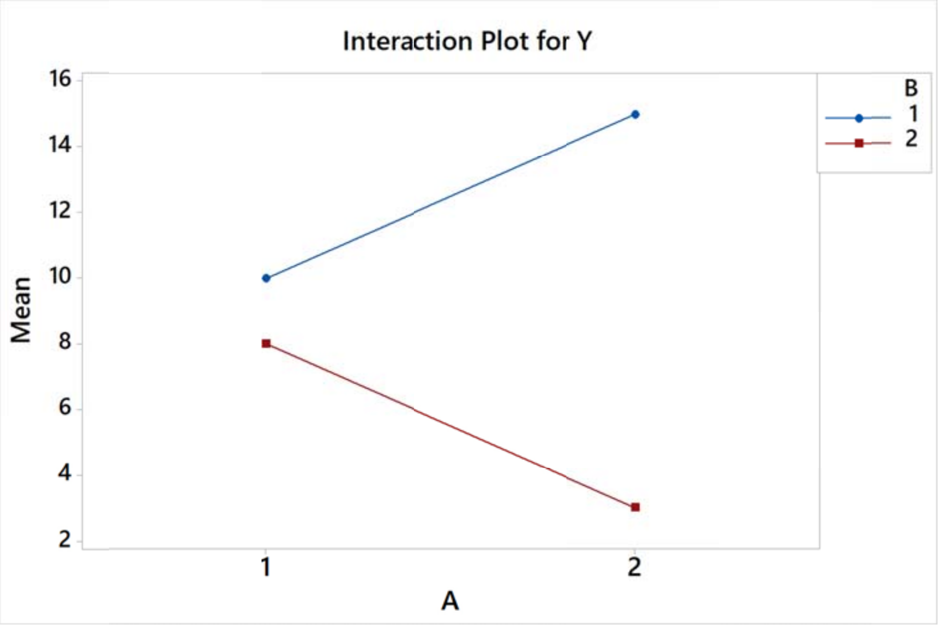
Would be the same for:	All 3 models	CostSize and CostLot only	None of the models
R^2			X
Degrees of freedom for SSModel		X	
MSE			X
Degrees of freedom for SSTotal	X		
$\hat{Y}$ when Size = 20 and Lot = 100			X
The value of Y for Case 1	X		
The value of Y for the case with the largest value of h_i for that model			X

13. In simple linear regression, can a case have high leverage but a small standardized residual? If your answer is yes, sketch a picture showing an example of such a point. If your answer is no, sketch a picture showing a point with high leverage and a large standardized residual.

Yes, it can happen. This picture, from Lecture 13, shows one example. The point in the upper right has high leverage and a small standardized residual.



14. (2 points each) A two-factor ANOVA situation with 2 levels for each factor and $n = 10$ observations in each combination of the factors resulted in the interaction plot shown below.



a. Is there a Factor A effect? Explain how you know.

No, there is no Factor A effect. The mean for A1 averaged over the two levels of B is the same as the mean for A2 averaged over the two levels of B. Th means are both about 9.

b. Is there a Factor B effect? Explain how you know.

Yes, there is a Factor B effect. The mean for B1 averaged over the two levels of A is about 12.5, whereas the mean for B2 averaged over the two levels of A is about 5.5. (You don't need to provide the values for the averages, you just need to say that they are clearly different, because the lines are so far apart.)

c. Is there an interaction? Explain how you know.

Yes, there is an interaction. You can tell because the lines are parallel. A more complete answer is that the difference between B1 and B2 is much larger for A2 then it is for A1.

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

1. Two methods for selecting which variables to include in multiple regression are "best subsets" and "forward selection." For a situation in which there are P possible explanatory variables, for which of the following model sizes will the two methods always choose the same model?

A. Only models with a single variable.
B. Only models with a single variable and the model with all P variables.
C. Only models with a single variable, models with two variables, and the model with all P variables.
D. None of the above. There is no model size where they are guaranteed to choose the same model.

2. In a multiple regression situation with three possible explanatory variables X1, X2 and X3, which of the following pairs of models could not be tested using a nested F test?

A. The model with X1 and X2 versus the model with X1 only.
B. The model with X1, X2, X3 versus the model with X1 and X3 only.
C. The model with X1 and X2 versus the model with X1 and X3.
D. All of the above could be tested using a nested F test.

3. In a multiple regression situation with Y = Sales price of a house, X_1 = hundreds of square feet and an indicator variable X_2 for whether the house has a pool, which of the following describes a model that includes the interaction between X_1 and X_2?

A. The linear relationship between Y and X_1 could have different intercepts for the two values of X_2, but cannot have different slopes.
B. The linear relationship between Y and X_1 cannot have different intercepts or different slopes for the two values of X_2.
C. The linear relationship between Y and X_1 may or may not have different intercepts for the two values of X_2, but definitely has different slopes.
D. The linear relationship between Y and X_1 must have different intercepts for the two values of X_2, but has the same slope.

4. In a multiple regression situation with P possible variables and no missing data, the choice of model can be based on various criteria. When trying to choose from a list of models, which of the following pairs of criteria will always result in the same model choice?

A. The model with the largest adjusted R^2 and the model with the smallest MSE.
B. The model with the smallest Cp and the model with the largest adjusted R^2.
C. The model with the smallest Cp and the model with the smallest MSE.
D. The model with Cp closest to p and the model with the largest adjusted R^2.

5. Suppose a one-factor analysis of variance model results in a *p*-value of 0.13 for testing the null hypothesis that the population means are equal for the levels of the factor. If a quantitative covariate is added to the model, which of the following must be true?

A. The *p*-value for testing the factor will stay the same or decrease.
B. The *p*-value for testing the factor will stay the same or increase.
C. SSE will stay the same or decrease.
D. SSE will stay the same or increase.

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

Output for Questions 1 to 6:

```
> tapply(FinalData2$Final, FinalData2$Instructor, mean)
    A     B     C 
75.80 71.62 78.62 

> FinalModel2 <- aov(Final ~ Instructor, data = FinalData2)
> summary(FinalModel2)
             Df Sum Sq Mean Sq F value Pr(>F)  
Instructor    2   1240   620.2   3.415 0.0355 *
Residuals   147  26696   181.6                 

> TukeyHSD(FinalModel2, ordered = T)
$Instructor
    diff        lwr       upr     p adj
A-B 4.18 -2.2014049 10.561405 0.2703864
C-B 7.00  0.6185951 13.381405 0.0277823
C-A 2.82 -3.5614049  9.201405 0.5489589
```

Output for Questions 7 to 12:

```
> cor(Houses)
          Price      Size       Lot
Price 1.0000000 0.6848219 0.7157072
Size  0.6848219 1.0000000 0.7668722
Lot   0.7157072 0.7668722 1.0000000
--------------------------------------------------
> Cost <- lm(Price ~ Size + Lot, data = Houses)
> anova(Cost)
Analysis of Variance Table

Response: Price
          Df Sum Sq Mean Sq F value    Pr(>F)
Size       1  40447   40447 18.0018 0.0005485
Lot        1   7601    7601  3.3831 0.0833990
Residuals 17  38196    2247                  
----------------------------------------------------
> summary(Cost)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.1216    29.7165   1.148   0.2668
Size          2.3232     1.7700   1.313   0.2068
Lot           0.5657     0.3075   1.839   0.0834

Residual standard error: 47.4 on 17 degrees of freedom
Multiple R-squared:  0.5571,      Adjusted R-squared:  0.505
F-statistic: 10.69 on 2 and 17 DF,  p-value: 0.000985
-------------------------------------------------
> CostSize <- lm(Price ~ Size, data = Houses)
> summary(CostSize)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   64.554     26.268   2.458 0.024362
Size           4.820      1.209   3.987 0.000864

Residual standard error: 50.44 on 18 degrees of freedom
Multiple R-squared:  0.469, Adjusted R-squared:  0.4395
F-statistic:  15.9 on 1 and 18 DF,  p-value: 0.0008643
```