COMM 581 - Assignment #6
Multiple Linear Regression – Categorical variables

Name: _____ **Total: 30 marks**
Due date: Thursday Oct. 20, 2016 (11pm)

Background:
You are analyzing data from a company, Spectra Technologies, to determine if they are giving equal pay to men and women after controlling statistically for their level of experience. If there is a difference in salary between men and women, the company needs to know how much of a difference there is, and if this difference changes depending on level of experience (if there is an interaction between gender and experience).

All graphs should be created using R, and all graphs discussed in your assignment should be included with your submission. You can use sum of squares and standard errors from the R output. Show calculations for test statistics, confidence intervals and measures of goodness of fit based on sums of squares.

You will be presenting this assignment as a report to the company, with your recommendations.

You need to include the following sections in your report:

1. **Introduction**. Explain the purpose of your analysis, and why equal pay for men and women after accounting for their experience might be important to this company. Explain why you will use certain models and what they represent. Explain the effect on salary that you might expect to see from experience and gender. Define any terms that you will use in your report. (~ 3 paragraphs) (**5 marks**)

2. **Methods & Results**. Describe the data used, and the stats package used (including the version, you can get this directly from R using `citation()`). Explain which models you used, which hypothesis tests you conducted, and the results of those tests. *E.g.*: I used a partial F test to test if the variable "experience" explains a significant amount of variation in the model. (**mark breakdown based on questions**)

3. **Discussion & Recommendations.** Based on your final model, what specific recommendations do you have to equalize salary for men and women if there are differences? Do you have any recommendations regarding hiring that could improve equality between men and women? What could you do to improve your model and therefore offer better recommendations? (~ 3 paragraphs) (**5 marks**)

4. **Appendix:** The R code for your analysis. Include this as part of your report PDF file. (**1 mark**)

   Formatting: Provide labels and captions for all your tables and figures. Captions should inform the reader about the results shown in the table or figure. (**0.5 marks**)

The following questions should be answered in your report. *Discussion questions are in italics and should be addressed in the discussion section.*

Summary statistics (present these data in one or more tables)

1. What is the range of values for salary and experience? What is the range for these variables for men and women? *Are the ranges different, and what could account for this difference?* (**1 mark**)
2. What is the mean value for salary and experience? What is the mean for these variables for men and women? *Are the means higher for men or women? What could account for this difference?* (**0.5 marks**)
3. Graph salary vs. experience, with two different plotting characters, one for men and one for women (remember to explain your plotting characters in the caption or as a legend). *Do you think the relationship between salary and experience is linear after including gender as a variable?* (**1 mark**)
4. *Describe your qualitative observations about the relationship between these variables from the graph*. Remember to discuss these in the Discussion section. (**2 marks**)

Developing a model

5. Write the model statement for the (**full) interaction model** including the interaction between gender and experience. Show the indicator variable for gender. How is the indicator variable coded for men and women? (**1 mark**)
6. Fit this model using R. Assess the assumptions of linearity (salary vs. experience graph, residual plot), equal variance (residual plot) and normality of errors (histogram, normality plot, normality tests). State any concerns you have and their consequences. (**2 marks**)
7. Note: The data was collected at one point in time, from the single location for the company, so you will not need to investigate if the assumption of independence of errors has been met. Experience was self-reported, so there is no reason to think that there is error associated with it, and salary was obtained from the company records.
8. Write the ANOVA table for this model using Type III SS and showing df and SS. See how to do this based on the class example. Sources of variation should be model, error and total. (**1 mark**)
9. Test the significance of the regression, showing the calculation of the F statistic based on sum of squares. (**1 mark**)
10. Test the significance of the variable gender, using a partial F-test. Remember that gender requires two variables in the model, so you need to **fit a model that does not have gender**, then compare the two models. Show all steps in your calculation of the partial F-statistic. *What does the result of this test mean for the model? Interpret what this result means in terms of the overall purpose of this study.* **(1 mark)**
11. Test the significance of the variable experience, using a partial F-test. You will need to **fit a model that does not have experience**, then compare the full model to this model. *What does the result of this test mean for the model? Interpret what this result means in terms of the overall purpose of this study.* **(1 mark)**

Option A: gender and experience are both required in the model (3 marks)

(MLR with categorical variable model)

12. a) Fit a new model that has the same slopes for men and women. Assess the assumptions of linearity, equal variance and normality.

    b) Use a partial F-test to determine if the slopes are the same or different. *What does the result of this test mean for the model? Interpret what this result means in terms of the overall purpose of this study.*

Option B: only gender is required in the model (ANOVA model) (3 marks)

12. a) Fit a new model with only gender. Assess the assumptions of equal variance and normality (you will no longer assess the assumption of linearity because there are no continuous explanatory variables in the model)

    b) Test the significance of the regression.

    c) What is the mean salary for men and for women? Calculate the confidence intervals for these means.

Option C: only experience is required in the model (SLR model) (3 marks)

12. a) Fit a new model with only experience. Assess the assumptions of linearity, equal variance and normality.

    b) Test the significance of the regression.

    c) What is the mean salary for men and for women? Calculate the confidence intervals for the intercept and slope.

For your final model:

13. Write the model equation with variable names (including any indicator variables). Write the model with the values of the co-efficients in the equation. What does each of these co-efficients represent? (**2 marks**)
14. If you chose option A or C, what is the equation to calculate predicted salary for men? For women? If you chose option B, what is the equation relating salary and experience? (**1 mark**)
15. Calculate the measures of goodness of fit: R^2 and root MSE. What represents a good fit for these measures? (**1 mark**)