COMM 581 - Assignment #8
Logistic Regression

Name: ____________________________ **Total: 20 marks**
Due date: Monday Nov. 14, 2015 (11pm)

Background:

A magazine company is trying to figure out who to target with email advertisements for a children's magazine: Kid Creative. When they send an email, they only want to advertise three magazines to each potential customer, so they want to maximize their probability of making a sale, thereby maximizing their overall revenue. If they advertise magazines that someone doesn't buy, they have wasted that opportunity. They send out some experimental email advertisements to people on their mailing list. The sample includes people who have already purchased a magazine from the magazine company. The magazine company purchases additional information on these customers from a third party credit agency to create a profile of these customers. They want to see which variables can be used to predict if someone will buy the magazine Kid Creative.

Please submit your R script file for this assignment as part of your assignment PDF. Clearly label each model that you used in the assignment. (1 mark)

Questions

1. What relationship do you expect between each potential explanatory variable and the response variable? Why? (You can say "no relationship", however, you must explain why.) (**2 marks**)

2. Which 3-5 potential explanatory variables do you think are most likely to be good predictors of whether or not someone will buy a children's magazine? (Which potential explanatory variables do you think will have the strongest relationship with the response variable?) (**0.5 mark**)

3. Fit a null model with the intercept only. What is the estimate of the intercept for this model? Convert this estimate from logit (log odds) units to probability units? (**0.25 marks**)

4. Using the original data, calculate the probability of a purchase. What do you notice about this value and the intercept from the null model? (**0.25 marks**)

5. Use R to obtain the log likelihood for the null model.

Single variable models

6. Fit a model with **income** only. Use R to obtain the log likelihood for this model.

7. Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R) for the model including **income**. Test the significance of the model (include all four steps of your hypothesis test). (**1 mark**)

8. Show the full calculation (based on log likelihood values from R) of the AIC value for the model including **income**. (**0.25 marks**)

9. Fit a model with **married** only. Use R to obtain the log likelihood for this model.

10. Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R) for the model including **married**. Test the significance of the model (include all four steps of your hypothesis test). (**1 mark**)

11. Show the full calculation (based on log likelihood values from R) of the AIC value for the model including **married**. (**0.25 marks**)

12. Fit 11 models, **each with one of the following explanatory variables**: income, gender, married, education, professional job, retired, unemployed, dual income, children, bought children's magazine previously, and bought parenting magazine previously. Record information about each model in the following table. Organize your models from lowest AIC value to highest AIC value. Example table below. (**4 marks**)

Explanatory variable	**AIC value**	**G statistic**	**p value**	**Include variable in model?**

13. What do you notice about the AIC values and the results of the likelihood ratio tests? Which AIC values indicate a better fit? (**0.5 marks**)

14. Out of all the possible explanatory variables, are there any that you think are redundant or correlated? Out of redundant variables, which one would you prefer to use? Why? (**1 mark**)

15. Using R, create a graph of purchase vs. income with a method to visualize married. What are some overall patterns that you see? How can these help you answer your research question? (**1 mark**)

Model A: Model with income and married:

16. Create a model with the following explanatory variables: income, married, the interaction between income and married. (buy ~ income*married)

17. Does this model meet the assumptions of generalized linear models? (**1 mark**)

18. What does the interaction between income and married allow in this model allow? (**0.25 marks**)

19. Test the likelihood of the whole model compared to the likelihood of the null model using a likelihood ratio test. Show your calculation of the test statistic using the log-likelihoods from R. Confirm your results using R. (**1 mark**)

20. Test each variable using a likelihood ratio test. This will require you to fit models that eliminate one variable. Show your calculation of the test statistics using the log-likelihoods from R. Confirm your results using R. (**1 mark**)

21. If both variables should remain in the model, test the interaction term only using a likelihood ratio test (you will have to fit a reduced Model A that excludes the interaction). Show your calculation of the test statistics using the log-likelihoods from R. Confirm your results using R. (**1 mark**)

Model B: Model with income, married and professional job:

22. Create a model with the following explanatory variables: income, married, the interaction between income and married, job, and the interaction between job and income. (buy ~ income*married + income*job)

23. What does the interaction between job and income allow in this model? (**0.25 marks**)

24. Compare this model (Model B) to Model A, which did not have job, using a likelihood ratio test. Does including professional job as a variable improve the model? (**1 mark**)

    You could continue trying different variables to add to your model. Choose the variables that explain the most variation on their own and add those next. For this assignment, we will stop developing models at this point.

Final model: With the variables that you have decided should be included

25. For your final model, calculate the Pseudo-R^2 and the scaled Pseudo-R^2. (**1 mark**)

26. Calculate the AIC value for your final model. Compare the AIC value of your final model to the AIC values for the models that had only one of the variables that are included in your final model (Single variable models). Put all of these models and AIC values in a table to make them easy to compare. How much does the AIC value improve from the single variable models of income, married and professional job to your final model? (**2 marks**)

27. Create a classification table for this data based on the final model. Include the full classification table in your output. Remember that you can output dataframes to a .csv file using write.csv().

28. For sensitivity, specificity, false negatives, false positives, which do you want to maximize or minimize, and which do you not care about? Why? (**2 marks**)

29. You decide to maximize the percentage of correct of predictions. Why? What probability cut-off should you use to maximize the percent of correct predictions? (**0.5 marks**)