

COMM 581 - Assignment #09
Poisson Regression

Name: _____
Due date: Monday Nov. 21, 2015 (11pm)

Total: 20 marks

Background:

At a high school, the administration is trying to determine which variables explain the number of awards each student receives. Number of awards is the response variable. There are two possible explanatory variables: math final exam score (continuous) and program (categorical with 3 levels: 1 = "General", 2 = "Academic" and 3 = "Vocational"). Based on the results of your analyses, the administration would like your opinion about how to distribute 5 new scholarships among the programs.

Questions

Include captions for each of your graphs (Figure 1, Figure 2, etc.) describing what is shown in the graph and how different category levels are represented. **(1 mark)**

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)**

1. What are your predictions (with explanation) regarding the relationships between explanatory variables and the response variable? **(1 mark)**
2. Import the data and change program to a factor using the factor command (change the labels to make them more informative).
3. Create a histogram of the responses. Do you think this follows a Poisson distribution? Why or why not? **(1 mark)**
4. Create a scatterplot of the response variable against the continuous explanatory variable with a lowess line. Does this indicate that math score would be a good explanatory variable? **(1 mark)**
5. Create a grouped bar plot that shows the frequency of different responses, grouped by program. Does it look like each category level follows a Poisson distribution? In the caption, explain how each category level is represented on the graph. **(1 mark)**
6. What is the mean of awards received per student for the different programs (general, academic, vocational)? Use the tapply function for this. Does this indicate that program would be a good explanatory variable? **(0.5 marks)**
7. Fit a null model. What is the value of the intercept? Backtransform this value using the appropriate backtransformation for Poisson regression. What does this value represent? **(0.5 marks)**

8. Fit a model with math score as the explanatory variable (Model A). Test the significance of the regression using a likelihood ratio test (include all four steps of your hypothesis test). Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R). **(1 mark)**
9. Fit a model with math score and program as explanatory variables (Model B); include the interaction term. Test the significance of the regression using a likelihood ratio test. Test the significance of the regression using a likelihood ratio test (include all four steps of your hypothesis test). Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R). **(1 mark)**
10. Test each variable in Model B using likelihood ratio tests (include all four steps of your hypothesis test). Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R). What do you conclude about the explanatory variables in Model B? Is model B a good model? **(2 marks)**
11. Fit a model with math score and program as explanatory variables excluding the interaction term (Model C). Test the significance of the interaction term using a likelihood ratio test comparing model B and C. **(1 mark)**
12. Write the equation of the final model with the values of the co-efficients. **(1 mark)**
13. Define the terms overdispersion and underdispersion (use graphs or sketches to supplement your explanation). **(1 mark)**
14. Calculate the residual deviance / degrees of freedom. Do you see any evidence for overdispersion or underdispersion in your final model? **(1 mark)**
15. Create a scatterplot of the response variable against the math score with program shown in different colors. Add the lines (in the matching color) of the model fit. **(1 mark)**
16. What are your observations from the graph? How do these observations relate to the co-efficients from the model? **(2 marks)**
17. The school is wondering about adding 5 new scholarships. Give your opinion about how the scholarships should be distributed among the programs. Discuss: how the program(s) you chose would benefit the most from the additional scholarships, and which criteria should be used for determining who receives the scholarship. **Justify your opinion based on the results of your analyses and how you define the primary objective when giving scholarships. (2 marks)**