Modeling Assignment 9: Poisson and ZIP Regression Models

Assignment Overview

In this assignment we will be fitting models and calculating the various summative statistics that are associated with Poisson and Zero-Inflated Poisson Regression. The data set for this assignment, STRESS, includes information from about 650 adolescents in the United States who were surveyed about the number of stressful life events they had experienced in the past year (STRESS). STRESS is also an integer variable that represents counts of stressful events. The dataset also includes school and family related variables, which are assumed to be continuously distributed. The variables in this data set are:

COHES = measure of how well the adolescent gets along with their family (coded low to high)
ESTEEM = measure of self-esteem (coded low to high)
GRADES = past year's school grades (coded low to high)
SATTACH = measure of how well the adolescent likes and is attached to their school (coded low to high)

There is no other information about this data or the variables.

Assignment Tasks

1. For the STRESS variable, make a histogram and obtain summary statistics. Obtain a normal probability (Q-Q) plot for the STRESS variable. Is STRESS a normally distributed variable? What do you think is its most likely probability distribution for STRESS? Give a justification for the distribution you selected.

2. Fit an OLS regression model to predict STRESS (Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Obtain the typical diagnostic information and graphs. Discuss how well this model fits. Obtain predicted values (Y_hat) and plot them in a histogram. What issues do you see?

3. Create a transformed variable on Y that is LN(Y). Fit an OLS regression model to predict LN(Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Obtain the typical diagnostic information and graphs. Discuss how well this model fits. Obtain predicted values (LN(Y)_hat) and plot them in a histogram. What issues do you see? Does this correct the issue?

4. Use the glm() function to fit a Poisson Regression for STRESS (Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Interpret the model's coefficients and discuss how this model's results compare to your answer for part 3). Similarly, fit an over-dispersed Poisson regression model using the same set of variables. How do these models compare?

5. Based on the Poisson model in part 4), compute the predicted count of STRESS for those whose levels of family cohesion are less than one standard deviation below the mean (call this the low group), between one standard deviation below and one standard deviation above the mean (call this the middle

group), and more than one standard deviation above the mean (high). What is the expected percent difference in the number of stressful events for those at high and low levels of family cohesion?

6. Compute the AICs and BICs from the Poisson Regression and the over-dispersed Poisson regression models from part 4). Is one better than the other?

7. Using the Poisson regression model from part 4), plot the deviance residuals by the predicted values. Discuss what this plot indicates about the regression model.

8. Create a new indicator variable (Y_IND) of STRESS that takes on a value of 0 if STRESS=0 and 1 if STRESS>0. This variable essentially measures is stress present, yes or no. Fit a logistic regression model to predict Y_IND using the variables using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Report the model, interpret the coefficients, obtain statistical information on goodness of fit, and discuss how well this model fits. Should you rerun the logistic regression analysis? If so, what should you do next?

9. It may be that there are two (or more) process at work that are overlapped and generating the distributions of STRESS(Y). What do you think those processes might be? To conduct a ZIP regression model by hand, fit a Logistic Regression model to predict if stress is present (Y_IND), and then use a Poisson Regression model to predict the number of stressful events (STRESS) conditioning on stress being present. Is it reasonable to use such a model? Combine the two fitted model to predict STRESS (Y). Obtained predicted values and residuals. How well does this model fit? HINT: You have to be thoughtful about this. It is not as straight forward as plug and chug!

10. Use the pscl package and the zeroinfl() function to Fit a ZIP model to predict STRESS(Y). You should do this twice, first using the same predictor variable for both parts of the ZIP model. Second, finding the best fitting model. Report the results and goodness of fit measures. Synthesize your findings across all of these models, to reflect on what you think would be a good modeling approach for this data.

Assignment Document

Results should be presented, labeled, and discussed in the numerical order of the questions given. Please use MS-WORD or some other text processing software to record and present your answers and results. The report should not contain unnecessary results or information. Tables are highly effective for summarizing data across multiple models. The document you submit to be graded MUST be submitted in pdf format. Please use the naming convention: ModelAssign9_YourLastName.pdf.