

|  |
| --- |
| Computational Assignment #6: Poisson and Zero-Inflated Poisson Regression  *MSDS 410* |

In this assignment we will be fitting models and calculating the various summative statistics that are associated with Poisson and Zero-Inflated Poisson Regression. In addition, we will be fitting logistic regression models and interpreting the results. Students are expected to show all work in their computations. A good practice is to write down the generic formula for any computation and then fill in the values need for the computation from the problem statement. Throughout this assignment keep all decimals to three places, i.e. X.xxx. Students are expected to use correct notation and terminology, and to be clear, complete and concise with all interpretations of results. This computational assignment is worth 50 points. The points associated with each problem are given with the specific question.

Any computations that involve “the log function”, denoted by log(x), ***are always meant to mean the natural log function (which will show as ln() on a calculator).*** The only time that you should ever use a log function other than the natural logarithm is if you are given a specific base.

For this assignment, we will be using the STRESS dataset. This includes information from about 650 adolescents in the US who were surveyed about the number of stressful life events they had experienced in the past year (STRESS). STRESS is 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. These variables 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)

Each problem is worth 5 points.

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.

Chart, histogram

Description automatically generatedText

Description automatically generated with medium confidence

Chart

Description automatically generated

The STRESS variable is not normally distributed, as it is skewed to the right a lot, due to such high amounts of individuals have little to no stress. This appears to have a Poisson distribution, as the events do not affect each other and the probability that an event occurs does not change overtime.

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?

Table

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

The residuals vs. fitted values plot does not seem to be truly random, and the Q-Q plot also has an almost linear trend with the theoretical quantities and the standardised resudals. This should not occur in a well fitting model. The scale-location plot also is not truly random, indicating this model does not fit well. This is further emphasized by the low R-value of the model.

Chart, histogram

Description automatically generated

This histogram seems to have a normal distribution, but we can see from the STRESS vs. Y\_hat plot how the linear model is not the proper way to go in this case:

Chart

Description automatically generated

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?

Text, letter

Description automatically generated Table

Description automatically generated

Chart

Description automatically generated

The residuals vs fitted plot still does not seem to show a random distribution of the relationship between the two, and there is still a trend in the plot.

Chart, line chart

Description automatically generated

The Q-Q plot for the theoretical quantities and the standardized residuals have a large gap in between two halves of the data, as is the case with the residuals vs. leverage plot.

Chart, scatter chart

Description automatically generated

The Cook’s line in splitting the data in half when finding outliers, indicating this is not a well-fitting model still.

Chart, histogram

Description automatically generated

The histogram is normally fitting, however. Still this model has an even lower R^2 value, and is not a good fit for this data.

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?

Text

Description automatically generated

Y= 2.73446 - 0.01292\*B1 - 0.02369\*B2 - 0.02347\*B3 - 0.01648\*B4

2.73446 is the intercept, - 0.01292 is the coefficient for COHES, - 0.02369 is the coefficient for ESTEEM,

-0.02347 is the coefficient for GRADES, and -0.01648 is the coefficient for SATTACH.

To find the effect of each coefficient on the variable, we must use e^(coefficient) to find the true effect in a Poisson distribution. So the coefficients have an effect of 0.987, 0.9765, 0.9768, and 0.98366 respectively on the variable in the above model, which appears to make sense in this case.

Text, letter

Description automatically generated

The over-dispersed Poisson regression model seems to be very similar to the original Poisson model.

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?

Table

Description automatically generated with medium confidence

A picture containing graphical user interface

Description automatically generated

There is about a 25.485% increase in low vs. high stress counts.

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?

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

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.

Chart, scatter chart

Description automatically generated

This plot still does not indicate a good fitting model, due to the excessive residuals not lining up with the predictions.

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?

Text

Description automatically generated

Y= 3.51673-0.02073\*B1-0.01887\*B2-0.02549\*B3-0.02773\*B4

3.51673 is the intercept, - 0.02073 is the coefficient for COHES, -0.01887 is the coefficient for ESTEEM,

-0.02549 is the coefficient for GRADES, and -0.02773 is the coefficient for SATTACH.

To find the effect of each coefficient on the variable, we must use e^(coefficient) to find the true effect in a Poisson distribution.

Graphical user interface, text, application

Description automatically generated

The pseudo r^2 still indicates poor fitment, and the p-value is extremely small in the Wald test. The AIC does seem relatively low however compared to previously.

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!

Table

Description automatically generated

Chart

Description automatically generated

The model does seem to have improved based on this plot, as a relationship now seems to be visible. The residuals are higher early on.

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.

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generatedText

Description automatically generated

The better fitting model is run using the “negbin” distribution and having the COHES and ESTEEM variables run as count model coefficients and the GRADES and SATTACH variables run as zero-inflation model coefficients. This improves the AIC value, lowering it to 2272.497 from 2288.802. The BIC also lowers from 2333.587 to 2303.847.

11. Conclusion.

This assignment has added more ways to conduct models to what we have learned, including the Poisson distribution and zero-inf. This helps add more ways to conduct better fitting models depending on the data on hand, especially if it is not normally distributed initially.