GLM for Count Data

Yi Chen

Teachers College, Columbia University

HUDM 5123

Lab Ten for Linear Models and Experimental Design

GLM for Count Data

In this assignment, we will replicate the analysis with the ` THIRTYDAYCIG.3`. We begin by quick and direct check to examine the sample mean and variable of the outcome variable. The mean of ` THIRTYDAYCIG.3` is 1.27 and the variance is 50.98. The variance is about 40 times bigger than the mean. This give us a crude idea of the overdispersion risk of this dataset.

Firstly, we will compare the models in Poisson regression with or without the covariate ` THIRTYDAYCIG.1`. Here, we list the model results.

* Without ` THIRTYDAYCIG.1`
* With ` THIRTYDAYCIG.1`

)

Based on the AIC and BIC, model with ` THIRTYDAYCIG.1` has better fit. Additionally, based on the likelihood ratio test, the comparation result between these two models is , P < 0.001. Based on the dispersion test, model with ` THIRTYDAYCIG.1` has the statistics 29.11 with p-value 0.004, while the model without ` THIRTYDAYCIG.1` has the statistics 35.35 with p-value 0.001. Both models fall to reject the hypothesis that the dispersion is greater than 1. Consequently, both models are acceptable in terms of dispersion. Overall, model with ` THIRTYDAYCIG.1` has better fit and covariate ` THIRTYDAYCIG.1` should be added.

Secondly, we take five different models for one and one: linear regression model, Poisson regression model, negative binomial regression model, and zero-inflation model. For the likelihood ratio test, each model is compared with the previous model. For example, Poisson regression is compared with linear regression, and negative binomial regression is compared with the Poisson regression. To simplify, we do not take other covariate for the Bernoulli process in the zero-inflation model.

Table 1: Model Comparation

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Likelihood Ratio Test | AIC | BIC |
| Linear regression model | Baseline | 4313.79 | 4431.88 |
| Poisson regression model | 95.55 (p<0.001) | 4216.24 | 4429.81 |
| Negative binomial regression model | 3325.7  (p<0.001) | 993.09 | 1011.19 |
| Zero-inflation model | 72.66  (p<0.001) | 926.44 | 958.09 |

Based on Table 1, the Zero-inflation model without predators for the Bernoulli process has the best model performance. We should pick this model. The model result is:

The effect from treatment is not significant. In Poisson regression, the interpretation of the coefficient is additive on the log scale and multiplicative on the original scale. In general, if we fix all the other effect, the log average of THIRTYDAYCIG.3 for the girls in the treatment group will be 0.52 lower than the log average of THIRTYDAYCIG.3 for the girl in the control group. Similarly, average of THIRTYDAYCIG.3 for the girls in the treatment group is about times the average of THIRTYDAYCIG.3 for the girls in the control group.