Data Analytics for Business

Homework 2 – Part 2

Count Models

The context and data for this exercise are based on a real research project on online user-generated-content (UGC). We would like to model online forum participants' posting behavior and investigate what factors would affect the number of posts each forum participant writes. The following data set consists of a daily log of individual forum users' posting records, along with the factors that could affect the number of posts they write each day, including the daily total number of posts on the forum (to account for the supply of competing content), the average number of times each post was read (to account for the readership or the demand, which provides motivation for users to write posts), the number of posts individual participants have written in the past (to account for past behaviors), and potential time effects.

The data file is "**Forum_Posts.csv**". It has the following variables:

Data Description

Column Name	Variable Description
posts	The daily number of posts each participant of the online forum wrote
totalPosts	The daily total number of posts on the forum (in 10,000s)
readingRate	The daily average number of times each post was read
postStock	A measure of each individual participant's total posts in the past (discounted by time lapses)
wknd	An indicator whether that day is a weekend day

Import the data into R and perform the following analysis.

Task 1: Poisson Regression Model

1. Estimate a Poisson regression model using this data set by regressing **posts** on all the other explanatory variables. Use the **glm()** function and produce the summary of the estimation results.

2. Examine the potential overdispersion by computing the estimate of σ^2 as we introduced in class. Compare the estimated $\hat{\sigma}^2$ to 1, and state your conclusion regarding whether the data are over-dispersed or not.

Task 2: Negative Binomial Model

3. Next, estimate a negative binomial model using this data set in a similar fashion, that is, by regressing **posts** on all the other explanatory variables. You need to first install the **MASS** package, and then use the **glm.nb()** function to do the estimation.

4. How would you interpret the estimated $\hat{\theta}$? What does the value imply about the potential overdispersion?

5. Compute and compare the AIC and BIC of both the Poisson regression model and the negative binomial model. Which model fits the data better?

6. Based on the model estimation results, predict the probability of any given number of posts using the Poisson regression model and the negative binomial model, respectively. Create a vector of 0:20 as **k**, and calculate the predicted probability of observing **k** number of posts based on the Poisson regression model and the negative binomial model (evaluated at the mean of the explanatory variables in the sample) as **p1** and **p2**, respectively. Plot **p1** and **p2** against **k** on the same plot, as we demonstrated in class. Discuss **how and why** the distributions of the Poisson model and the negative binomial model differ.