Home Work 3

Due: December 2, before 9:00 PM

1 Poisson Regression in Stan

In this exercise you are required to estimate a Poisson Regression model in Stan. Our interest here is in modeling count data (Number of Doctor's visits) using a number of explanatory variables. The purpose of this exercise is to enhance your ability to use RStan on real data. In order to do this exercise you need to understand what is a Poisson distribution and why we need to do a Poisson regression rather than a ordinary regression.

We now that the Poisson distribution is useful for modeling count data (the number of events that occur on an observation). The dependent variable is discrete and is incremented by integers. The probability mass function for a Poisson distribution is given by

$$P(Y = y) = \frac{\lambda^y \exp(-\lambda)}{y!}, \tag{1}$$

where $\lambda > 0$ is called the Poisson rate and is interpreted as the rate at which events happen. The mean and variance of the random variable Y are both given by λ.

Poisson Regression When interest is in finding out what determines how many events happen, then the rate λ can be written in terms of independent variables. As $\lambda > 0$, $\log(\lambda)$ is modeled as a linear combination of the independent variables. We therefore can write $\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$.

In this exercise, do the following

1. Read the data from the file PoissonData.csv. It has 5190 observations and 13 variables. The first column is a vector of 1's (the intercept) and the last (13th) column is your dependent variable.

2. Split the data *randomly* into two subsets. The first contains 90% of the observations and is the estimation sample (Data1). The remaining 10% is for holdout predictions (Data2).

3. Estimate the model using RStan on Data1 and make predictions on Data2.

4. Interpret the second Parameter coefficient. The second variable is a gender variable that takes 1 if Female and 0, otherwise.

5. Compute in-sample and holdout Mean Absolute Deviations between actual and predicted values

You need the attached files for doing the exercise. The file **PoissonData.csv** contains the data.

The R file, **Code_MNL.R**gives you a program that Khaled used for the multinomial logit. This will be useful in figuring out how to make predictions within RStan.The file ModelReg.stan is stan code for a simple regression that we saw in class. Note that a poisson distributed variable in the model block of Stan is represented as $y \sim poisson(lambda)$, just like how we use $y \sim normal(xbmat, sigma)$ for a normal in a ordinary regression. Note that you can reach out to me if you need help.