Assignment 3: Count Data Modeling

STAT247C/PH242C

Fall 2018

Goals of the Assignment:

In this assignment you will simulate longitudinal data with fixed covariates that can be summarized as count data, and analyze this data using a Poisson Regression Model and a Negative Biniomial Regression.

The context for this assignment is a study of school absences in a cohort of students with math scores recorded at baseline. The length of time a student was present in the study is recorded, as well as whether their gender, their math score, and the number of days they were absent from school during the study.

Generating and Summarizing a Random Sample

The following describes the data generating steps. The subsequent code implements these steps for a slightly different data generating distribution.

1. Set the seed to 242 (*set.seed(242)*) to make your "random" results reproducible.
2. Generate a sample of size $n = 500$ from random variable X that is discretely uniformly distributed as either 1, 2, 3.
3. For the same observations, generate a binary (0 = Male, 1 = Female) random variable called Z for gender, where $P(Z = 1) = 0.5$
4. Generate for each individual a random number of days in the study called T from a Poisson distribution with mean 15. Remove any individuals that get 0 days.
5. For each day in the study within each individual, generate a random Bernoulli random variable (whether a student is absent or not) with probability given by $\lambda(X, Z)$:

$$\lambda(X) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 I(X = 2) + \beta_2 I(X = 3) + \beta_3 * Z))}$$

Where:

$$(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.25, 1.0, 1.25, -0.25)$$

Analysis Tasks

1. Show the code used to generate the data. (Modify the example code below. You only need to change the coefficients in the code to match the $\beta_0, \beta_1, \beta_2, \beta_3$ above).

```r
rm(list=ls())
set.seed(242)
library(dplyr)
n = 500
#To get a discrete uniform we can use the ceiling function
X = as.factor(ceiling(3*runif(n)))

#Bernoulli is a binomial with 1 trial (size = 1), repeated n times
Z = rbinom(n, size = 1, p = .5)
```

```
#T is poisson with mean 15
T = rpois(n, lambda = 15)

beta0 = 1
beta1 = 1
beta2 = 1
beta3 = 1

#Create a data frame of the observed data called O and remove any observations for which
# T is 0
O = data.frame(X, Z, T)

O = filter(O, T != 0)

#The probability each of the outcome occurring on each day is given by prob_Y_XZ
prob_Y_XZ = plogis(beta0 + beta1*(X==2) + beta2*(X==3) + beta3*Z)

# For each individual i, the total number of events over all T_i days follows a binomial # distribution
# (recall that the binomial distribution is a sum of identical bernoulli random variables)
O$Y = rbinom(n, size = O$T, prob = prob_Y_XZ)

head(O)
```

```
##   X Z  T  Y
## 1 3 1  7  6
## 2 1 0  9  5
## 3 2 0 12 11
## 4 3 0 13 10
## 5 2 1 17 17
## 6 3 0 20 17
```

2. Ignoring the gender variable Z, fit the following model based on summary statistics of the data alone:

$$\lambda(X) = exp(\beta_0 + \beta_1 I(X = 2) + \beta_2 I(X = 3))$$

Show your calculations. Remember that you have to account for the differing amounts of follow up time T. This "by hand" fitting will involve using software to count the total number of events and the total time in the study for each math score group. This process is converting a "simple estimates" of means within groups into coefficients in a simple saturated model.

```
#Insert code here
```

3. Do the same as question 1, but by fitting a Poisson Regression model in R. Your answers should match. (See the *glm()* function, and investigate the *family* argument and how to use an *offset* term in the model formula).

```
#Insert code here
```

4. Do a test of the difference in mean absences of males (the *male* variable) versus females using a T-like statistic in the following 3 ways:

a. Do a two-sample t-test on the new "rate" outcome $Y_i^* = Y_i/T_i$. In other words, calculate Y_i^* for each individual, and then compare the Y^* values for males and females.

```
#Insert code here
```

b. Do the same comparison assuming the counts come from a Poisson distribution (that is, use Poisson regression). Hint: You will need to use an offset to treat the outcome as a rate. Use the *glm()* function.

```
#Insert code here
```

c. Do the same comparison using the counts come from a Negative Binomial distribution (that is, use Negative Binomial regression). Hint: You will need to use an offset to treat the outcome as a rate. See the *glm.nb()* function in the *MASS* package.

```
#Insert code here
```

d. Give potential reasons for any differences observed between a-c.