HW8

INTRODUCTION. For this week’s homework, we will again be using the 2016 BRFSS data. The dataset link is here: Our motivating question for this week’s exercise is what are the predictors of marijuana and hasish use rates. We will use the BRFSS data from Colorado citizens since they may be more likely to accurately report their marijuana or hasish use because marijuana is legal in that state. The question regarding marijuana or hasish use is as follows: “During the past 30 days, on how many days did you use marijuana or hasish?” The responses are coded as follows:

- 1-30 Number of days
- 88 None
- 77 Don’t know/Not Sure
- 99 Refused
- Blank Not asked or missing

The dataset includes 14,958 observations and is not a sample of the BRFSS data like previously but includes all respondents from Colorado. Of note, this is not ideal data for this type of problem because there are a lot of 0’s in the data (people who didn’t use any marijuana or hashish in the last 30 days); however, the dataset is sufficient to illustrate the major key points for Poisson and Negative Binomial regression. Ideally, these data might be modeled using zero inflated negative binomial or hurdle regression which is beyond the scope of this course. A reading suggestion is provided however at the end for those who want to know more and it also may be helpful someday when you are faced with this problem in your professional careers.

0. Import the data from the Github website (BRFSS.2016.co.xlsx) and load these libraries: MASS, lmtest, stargazer, sandwich, margins, readxl, ggplot2, and tidyverse for this homework.

1. Recode values for MARIJANA=88 to 0’s and make a complete dataset by excluding individuals with values for MARIJANA of “NA”, 77, and 99, and values for sex of “Refused”and for income of “NA”. Recode MARIJANA as an integer using the as.integer function. Code sex as 0 for female and 1 for male, income as 3=“Refused” and “Don’t know/Not Sure”, 2 as between <$10K-<25K, 1 as 25K-50K and 0 as >=50K. Code employed as 0-8 with “Employed for wages” coded as 0. Code age as a binary variables with 0 as <50 and 1 as >=50 years old. Convert all of the categorical variables to factor variables and label the levels. Example code for this that you may use is below:

dat$var < -factor(dat$var, levels = c(0,1), labels = c(“level 0 descriptor”, “level 1 descriptor”))

Check to make sure the code worked as you would expect. Note: Yes “MARIJANA” is spelled wrong in the BRFSS variable name.

Note the efficiency of the data management code will be reviewed for medals Recode:

2. Visualize data

a. Display a histogram of the number of counts of days marijuana or hasish was used in the last 30 days.
b. Display histograms of the number of counts of days marijuana or hasish was used in the last 30 days by sex, employed category, income category, and age category (hint use facet_grid)
c. Comment on what you see. Are there any differences in the patterns of marijuana or hasish use in the last 30 days between males and females, employment categories, income categories and/or age categories?

3. Run a Poisson regression model for age category, sex, income category, and employed category as predictors and Marijana/hashish use as the dependent variable (remember the predictors as factor variables).

4. Run a Negative Binomial regression model for sex, income category, and employed category as predictors and Marijana/hashish use as the dependent variable.

5. Compare models using the likelihood ratio test. Which model is better and on what basis do you make that conclusion?

6. Further compare two models to see if their coefficients and standard errors differ (hint use the stargazer function).

7. Compute robust standard errors and calculate incidence rate ratios and 95% CIs using the robust standard errors

8. Interpret the incidence rate ratio results for age category, sex, education category, and income category. Does the incidence rate of marijuana and hashish use in the last 30 days vary by the levels of these factors?

For extra learning, how well did our models do at predicting 0's (run code below and see the output)? Not very well! This argues that a zero inflated model or a hurdle model might be better. To read more see: https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf

```
mu <- predict(modP, type = "response") # predict expected mean count
exp <- sum(dpois(x = 0, lambda = mu))  # sum the probabilities of a 0 count for each mean
round(exp)                            # predicted number of 0's
sum(marj_complete$MARIJANA < 1)     # observed number of 0's


mu <- predict(modN, type = "response") # predict expected mean count
exp <- sum(dpois(x = 0, lambda = mu))  # sum the probabilities of a 0 count for each mean
round(exp)                            # predicted number of 0's
sum(marj_complete$MARIJANA < 1)     # observed number of 0's
```