HW7: Poisson and Negative Binomial Regression

Kim Johnson

2/20/2019

Introduction. This homework will use data from the 2016 GSS dataset. Information about this dataset can be found here:https://gssdataexplorer.norc.org/pages/show?page=gss%2Fabout. We will use three variables from this dataset, age, sex, and number of children. We will model the rate of children by sex and age. Note we will not perform all tests that we should do with any data analysis (e.g. testing functional form of predictors, outlier tests) for the sake of time, instead we will just emphasize what is unique to running these types of models.

Run the below code to create RR function for generating IRRs and 95% CIs for Poisson models only.

```
glm.RR <- function(GLM.RESULT, digits = 2) {

    if (GLM.RESULT$family$family == "binomial") {
        LABEL <- "OR"
    } else if (GLM.RESULT$family$family == "poisson") {
        LABEL <- "RR"
    } else {
        stop("Not logistic or Poisson model")
    }

    COEF      <- stats::coef(GLM.RESULT)
    CONFINT   <- stats::confint(GLM.RESULT)
    TABLE     <- cbind(coef=COEF, CONFINT)
    TABLE.EXP <- round(exp(TABLE), digits)

    colnames(TABLE.EXP)[1] <- LABEL

    TABLE.EXP
}
```

Import GSS data

```
library(foreign)
  read.dct <- function(dct, labels.included = "yes") {
      temp <- readLines(dct)
      temp <- temp[grepl("_column", temp)]
      switch(labels.included,
             yes = {
                 pattern <- "_column\\(([0-9]+)\\)\\s+([a-z0-9]+)\\s+(.*)\\s+%([0-9]+)[a-z]\\s+(.*)"
                 classes <- c("numeric", "character", "character", "numeric", "character")
                 N <- 5
                 NAMES <- c("StartPos", "Str", "ColName", "ColWidth", "ColLabel")
             },
             no = {
                 pattern <- "_column\\(([0-9]+)\\)\\s+([a-z0-9]+)\\s+(.*)\\s+%([0-9]+).*"
                 classes <- c("numeric", "character", "character", "numeric")
                 N <- 4
                 NAMES <- c("StartPos", "Str", "ColName", "ColWidth")
             })
      temp_metadata <- setNames(lapply(1:N, function(x) {
          out <- gsub(pattern, paste("\\", x, sep = ""), temp)
          out <- gsub("^\\s+|\\s+$", "", out)
          out <- gsub('\"', "", out, fixed = TRUE)
          class(out) <- classes[x] ; out }), NAMES)
      temp_metadata[["ColName"]] <- make.names(gsub("\\s", "", temp_metadata[["ColName"]]))
      temp_metadata
  }

  read.dat <- function(dat, metadata_var, labels.included = "yes") {
      read.fwf(dat, widths = metadata_var[["ColWidth"]], col.names = metadata_var[["ColName"]])
  }

GSS_metadata <- read.dct(
  "https://raw.githubusercontent.com/kijohnson/ADA_Spring_2019/master/Class%206/Class_6_data/GSS.dct")
GSS_ascii <- read.dat(
  "https://raw.githubusercontent.com/kijohnson/ADA_Spring_2019/master/Class%206/Class_6_data/GSS.dat",
  GSS_metadata)
attr(GSS_ascii, "col.label") <- GSS_metadata[["ColLabel"]]
GSS <- GSS_ascii

#recode sex as 1 for male and 0 for female
GSS$SEX[GSS$SEX==1]<-1
GSS$SEX[GSS$SEX==2]<-0

GSS$SEX<-factor(GSS$SEX, levels=c(0,1), labels=c("Female", "Male"))
```

Load libaries (note you may have to install associated packages)

```
# Load MASS for negative bin
library(MASS)
# Load ggplot for graphing
library(ggplot2)
# Load lmtest library for coeftest
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# Load sandwich library for robust estimator
library(sandwich)
```

```r
#load stargazer library to view a comparison of standard errors
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

1. First look at the distribution of reported number of children in the dataset using a histogram to see if it roughly follows a Poisson distribution.

2. Do males have less children on average than females? Determine the answer to this question prior to modeling and write the answer in a full sentence.

3a. Run two univariate poisson models to determine if the rate of children varies by SEX and AGE (and use the summary function to see the results), 3b. Describe how the rate of children varies by SEX and AGE using incidence rate ratios (hint: you can use the glm.RR function to get IRRs here).

4. We discussed in class that Poisson models are often inappropriate because the variance in the sample exceeds the mean in the sample. To check for overdispersion, we can run a negative binomial model and then use the LR test to see if adding an overdispersion paremeter improves the model fit. a. Run two negative binomial models, one for SEX and one for AGE, b. Check for overdispersion using the lrtest function to compare the Poisson and negative binomial models for both SEX and AGE to see which is a better fit. Interpret the output.

5. Use the stargazer function to compare SEs for SEX and AGE from the Poisson and negative binomial models. Interpret the output (in terms of how the SEs compare in size between Poisson and negative binomial regression).

6. Determine if AGE, *an independent predictor of the number of children*, improves the model fit for the negative binomial model estimating the effect of SEX on the rate of children. HINT: use the LR test to compare the negative binomial models for SEX with and without AGE included as a covariate. What conclusion can you make from the LR test results?

7. We learned in class that it is a good idea to use robust standard errors for both Poisson and negative binomial regression models. Use robust standard errors for your negative binomial model from #6 that includes SEX and AGE. Save the model results as an object called 'robust' and look at the results by printing 'robust'.

8. Calculate the IRR for the effect of sex on the rate of children from the negative binomial model that includes SEX and AGE as covariates. You can modify the below code to do this. Interpret the IRR for SEX from the model.