ADA project

Yandra Mariano

12/3/2021

Research question: Does smoking lead to poor mental health?

```
#install.packages("stargazer")  For model comparison
#install.packages("sandwich") For robust SE estimator
#install.packages("MASS") For negative binomial
#install.packages("lmtest") For model comparison
#install.packages("foreign") To import SAS BRFSS 2020 dataset

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
```

```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 4.0.5
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.5
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.5

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.5

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

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

```
library(foreign)

#importing dataset into R
BRFSS2020 <- read.xport("BRFSS2020.XPT")
```

Now I will begin cleaning my data for analysis:

```
#creating subdataset with variables of interest, coding RFSMOK3 into factor var
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df <- BRFSS2020 %>%
    select(SEXVAR, MENTHLTH, X_IMPRACE, X_RFSMOK3) %>%
    mutate(X_RFSMOK3 = recode_factor(.x = X_RFSMOK3,
                                     '1' = "No",
                                     '2' = "Yes")) %>%
  na.omit
```

```
## Warning: Unreplaced values treated as NA as .x is not compatible. Please specify
## replacements exhaustively or supply .default
```

```
# check new data
summary(object = df)
```

```
##      SEXVAR         MENTHLTH       X_IMPRACE     X_RFSMOK3
##  Min.   :1.000   Min.   : 1.00   Min.   :1.000   No :328369
##  1st Qu.:1.000   1st Qu.:15.00   1st Qu.:1.000   Yes: 52487
##  Median :2.000   Median :88.00   Median :1.000
##  Mean   :1.543   Mean   :61.18   Mean   :1.707
##  3rd Qu.:2.000   3rd Qu.:88.00   3rd Qu.:1.000
##  Max.   :2.000   Max.   :99.00   Max.   :6.000
```

```r
table(df$SEXVAR)
```

```
##
##      1      2
## 174120 206736
```

```r
table(df$X_IMPRACE)
```

```
##
##      1      2      3      4      5      6
## 289334  28201   9558   6460  34076  13227
```

```r
table(df$X_RFSMOK3)
```

```
##
##     No    Yes
## 328369  52487
```

```r
#changing to 0 (MENTHLTH) and N.A. (MENTHLTH & RFSMOK3)
df$MENTHLTH[
  df$MENTHLTH=="88"]<-0

df$MENTHLTH[df$MENTHLTH==77] <- NA
df$MENTHLTH[df$MENTHLTH==99] <- NA
df$X_RFSMOK3[df$X_RFSMOK3==9] <- NA

table(df$MENTHLTH)
```

```
##
##      0      1      2      3      4      5      6      7      8      9     10
## 241281  10624  18800  11992   6101  16080   1730   6246   1261    238  12080
##     11     12     13     14     15     16     17     18     19     20     21
##    106    872    129   2371  11503    176    154    249     24   6290    418
##     22     23     24     25     26     27     28     29     30
##    113     75     83   2292     69    155    616    377  21164
```

```r
table(df$X_RFSMOK3)
```

```
##
##     No    Yes
## 328369  52487
```

```r
# descriptive table including entire small dataset
library("tableone")
```

```
## Warning: package 'tableone' was built under R version 4.0.5
```

```
desc.table <- CreateTableOne(data = df)
print(desc.table, nonnormal = c('SEXVAR', 'X_IMPRACE', 'X_RFSMOK3'))
```

```
##
##                             Overall
##   n                         380856
##   SEXVAR (median [IQR])       2.00 [1.00, 2.00]
##   MENTHLTH (mean (SD))        3.94 (8.05)
##   X_IMPRACE (median [IQR])    1.00 [1.00, 1.00]
##   X_RFSMOK3 = Yes (%)        52487 (13.8)
```

Now I will begin the Poisson regression analysis:

```
#function to calculate IRR
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
}

# Check shape of distribution of counts of poor mental health days (1 - 30 days) using density plot
# poor mental health days distribution
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

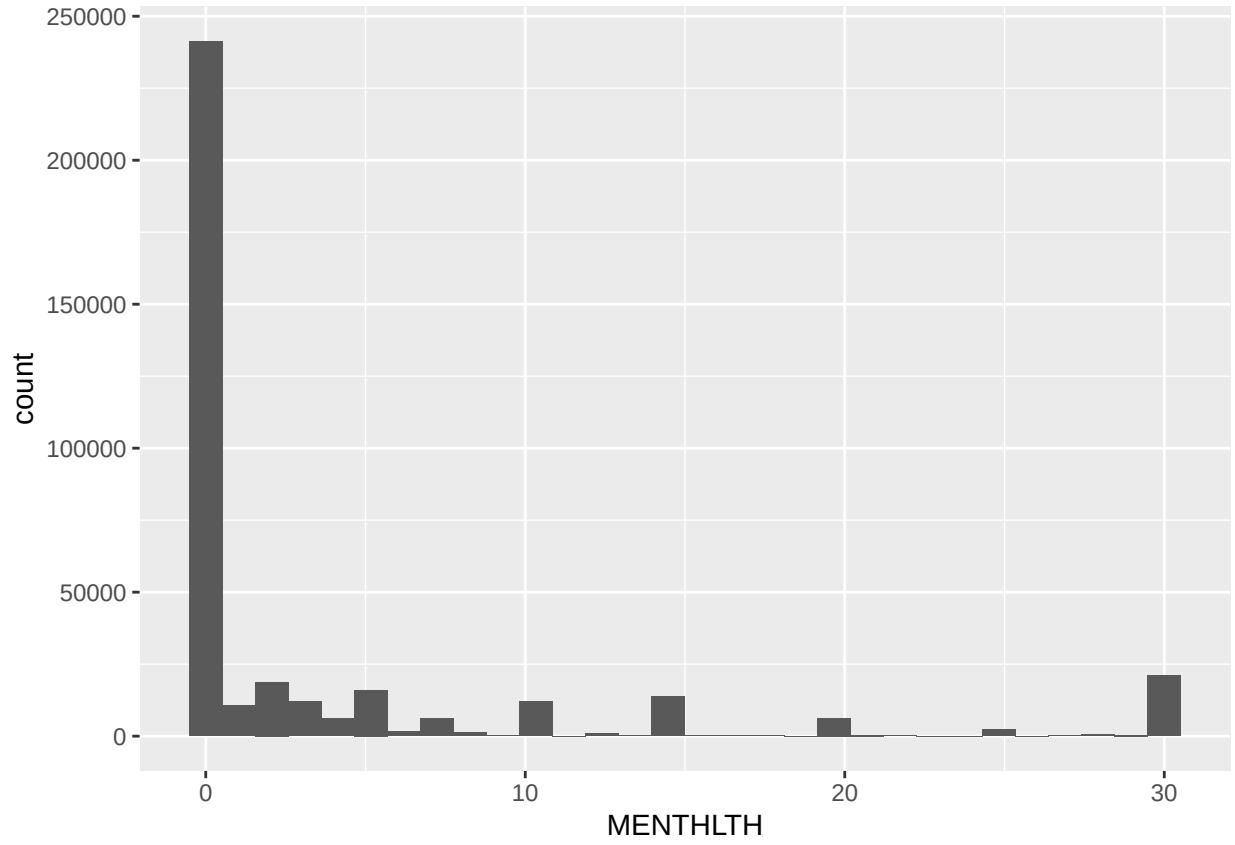
```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'purrr' was built under R version 4.0.5

## Warning: package 'stringr' was built under R version 4.0.5

## Warning: package 'forcats' was built under R version 4.0.5

## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
df %>%
  ggplot(aes(x = MENTHLTH)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 7187 rows containing non-finite values (stat_bin).
```



I want to determine whether the number of poor mental health days varies between the smoking status (mean count of cases) after adjusting for sex. Since MENTHLTH (# of days with poor mental health) are counts per person, I will not use an **offset*.

Poisson regression models

```
#without offset
model.0 <- glm(MENTHLTH ~ X_RFSMOK3 + SEXVAR, family ="poisson", data = df)
summary(model.0)
```

```
##
## Call:
## glm(formula = MENTHLTH ~ X_RFSMOK3 + SEXVAR, family = "poisson",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0367  -2.8676  -2.3370  -0.4646   9.4476
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.595436   0.002976   200.1   <2e-16 ***
## X_RFSMOK3Yes 0.683883   0.001941   352.4   <2e-16 ***
## SEXVAR       0.409184   0.001714   238.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4229726  on 373668  degrees of freedom
## Residual deviance: 4067026  on 373666  degrees of freedom
##   (7187 observations deleted due to missingness)
## AIC: 4568893
##
## Number of Fisher Scoring iterations: 6
```

Use the function glm.RR created above to get IRRs and 95% CIs

```
glm.RR(model.0, 3) # the second option in the function is the number of decimal places
```

```
## Waiting for profiling to be done...
```

```
##                 RR 2.5 % 97.5 %
## (Intercept)  1.814 1.803  1.824
## X_RFSMOK3Yes 1.982 1.974  1.989
## SEXVAR       1.506 1.501  1.511
```

Interpretation: The incidence rate of poor mental health days among smokers is 51% (95% CI 1.501-1.511) times higher than the incidence rate of poor mental health days experienced by non-smokers after adjusting for biological sex.

Running Negative binomial regression to check for overdispersion

```
#negative binomial model (no offset)
model.0nb <- glm.nb(MENTHLTH ~ X_RFSMOK3 + SEXVAR,control=glm.control(maxit=50), data = df)
summary(model.0nb)
```

```
##
## Call:
## glm.nb(formula = MENTHLTH ~ X_RFSMOK3 + SEXVAR, data = df, control = glm.control(maxit = 50),
##     init.theta = 0.1383303721, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0640  -0.9734  -0.9159  -0.1072   1.4274
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.595955   0.014797   40.27   <2e-16 ***
## X_RFSMOK3Yes 0.683594   0.012941   52.82   <2e-16 ***
## SEXVAR       0.408892   0.009008   45.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.1383) family taken to be 1)
##
##     Null deviance: 271749  on 373668  degrees of freedom
## Residual deviance: 266646  on 373666  degrees of freedom
##   (7187 observations deleted due to missingness)
## AIC: 1393131
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.138330
##           Std. Err.:  0.000473
##
##  2 x log-likelihood:  -1393122.603000
```

```
#run lrtest to compare models
lrtest(model.0, model.0nb)
```

```
## Likelihood ratio test
##
## Model 1: MENTHLTH ~ X_RFSMOK3 + SEXVAR
## Model 2: MENTHLTH ~ X_RFSMOK3 + SEXVAR
##   #Df   LogLik Df   Chisq Pr(>Chisq)
## 1   3 -2284443
## 2   4  -696561  1 3175764  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overdispersion seems to be present due to the significant p-value ($p<2.2e-16$) for the Likelihood ratio test comparing the Poisson and Negative Binomial Regression models.

New Interpretation: The incidence rate of poor mental health days among smokers is 51% (95% CI 1.501-1.511) times higher than the incidence rate of poor mental health days experienced by non-smokers after adjusting for biological sex.

I will now use robust standard errors to correct for SEs overdispersion. To get robust standard errors, I am using the code below

```
## Poisson model with SE estimated via robust variance estimator
coeftest(model.0, vcov = sandwich)

#You can get the robust standard errors shown in the table using the code below
cov.model.0 <- vcovHC(model.0, type="HC0") #type specifies variance estimator method, the vcovHC functi
std.err <- sqrt(diag(cov.model.0)) #estimate robust standard errors for each coefficient
std.err

#make a summary table of IRRs,  p-values and LL and UL confidence intervals
r.est2 <- cbind(IRR= exp(coef(model.0)), "Robust SE" = std.err,
"Pr(>|z|)" =round(2 *pnorm(abs(coef(model.0)/std.err),lower.tail=FALSE), 4),
LL = exp(coef(model.0) - 1.96 * std.err),
UL = exp(coef(model.0) + 1.96 * std.err))
options(digits=10)
r.est2
```

Below I further compare the estimates between the two models. As the results show, the Poisson regression estimates SEs that are usually smaller than those from the negbin. This implies that the Poisson regression leads to biased significance tests, and tends to make non-significant predictors significant.

Final Interpretation: The incidence rate of poor mental health days among smokers is 51% (95% CI 1.485-1.526) times higher than the incidence rate of poor mental health days experienced by non-smokers after adjusting for biological sex.

```
stargazer(model.0, model.0nb, title="Model Comparison",
          type="text",align=TRUE,single.row=TRUE, digits=6)
```

I will now check for effect modification of poor mental health days related to smoking by sex

```
#without offset (smoking status*sex)
model.sexint <- glm.nb(MENTHLTH ~ X_RFSMOK3 + SEXVAR + X_RFSMOK3*SEXVAR, control=glm.control(maxit=50),
summary(model.sexint)
```

```
##
## Call:
## glm.nb(formula = MENTHLTH ~ X_RFSMOK3 + SEXVAR + X_RFSMOK3 *
##     SEXVAR, data = df, control = glm.control(maxit = 50), init.theta = 0.1383303811,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0642  -0.9734  -0.9160  -0.1073   1.4272
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.596572   0.015844  37.652   <2e-16 ***
## X_RFSMOK3Yes         0.679333   0.041221  16.480   <2e-16 ***
## SEXVAR               0.408495   0.009718  42.033   <2e-16 ***
## X_RFSMOK3Yes:SEXVAR 0.002819   0.025889   0.109    0.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.1383) family taken to be 1)
##
##     Null deviance: 271749  on 373668  degrees of freedom
## Residual deviance: 266646  on 373665  degrees of freedom
##   (7187 observations deleted due to missingness)
## AIC: 1393133
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.138330
##           Std. Err.:  0.000473
##
##  2 x log-likelihood:  -1393122.591000
```

```
#Test the hypothesis with the lrtest
lrtest(model.0nb, model.sexint)
```

```
## Likelihood ratio test
##
## Model 1: MENTHLTH ~ X_RFSMOK3 + SEXVAR
## Model 2: MENTHLTH ~ X_RFSMOK3 + SEXVAR + X_RFSMOK3 * SEXVAR
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -696561
## 2   5 -696561  1 0.0119     0.9133
```

Interpretation: The interaction between smoking status and biological sex does not impact the relationship between smoking and poor mental health. (Not significant)