HW6: logistic regression

INSTRUCTIONS

In this homework, we will be working on analyzing the association between *seatbelt use* to determine if it varies according to sex, age, and BMI using the BRFSS. dataset. Don't forget to use reproducible research practice recommendations for annotating your code.

0. Conduct any data import and librarysteps needed to complete this homework using below code.

```
library(plyr)
library(ggplot2)
BRFSS<-read.csv(
"https://raw.githubusercontent.com/kijohnson/ADA_Spring_2019/master/BRFSS2017_10percent_v2.csv",
header=TRUE, sep=",", na.strings=FALSE)

#code to drop X column, which is not needed
col.dont.want<-"X"
BRFSS<-BRFSS[,!names(BRFSS)%in%col.dont.want, drop=F]
print(head(BRFSS))
```

Look at each variable used in the analysis to get a handle on observation numbers, missing data etc.

```
#use table function to look at categorical variables
table(BRFSS$seatbelt)#some missing/refused data
table(BRFSS$sex)#some refused data
#use summary function to look at continuous variables
summary(BRFSS$X_AGE80)#no missing data
summary(BRFSS$bmi)#no missing data
```

1. Make a new seatbelt variable categorizing seatbelt use into always vs. never/sometimes/nearly always with always being coded as 1 for logistic models and never/sometimes/nearly always being coded as 0. Check how many you have in each seatbelt category in your newly created variable and make sure that the number is correct.

2. An important first step in any regression modeling exercise is to know your data. One of the components of knowing your data includes data visualization and creating bivariate tables. Make (separate) plots to visually examine seat belt use by the continuous variables BMI and age. For sex, calculate the proportion of individuals in each category (use prop.table function, see https://www.statmethods.net/stats/frequencies.html for more informaton) who are seat belt wearers.

For comparing regression models, it is important to have the same number of observations. For this analysis we will do a complete data analysis. The below code creates an analytic dataset and removes any observations with missing/refused values for variables to be used in data analysis

```
#Complete case analysis exclusions of missing/refused data
myvars<-c("rowID", "X_AGE80", "age_cat", "sex", "bmi", "seatbelt_binary")
BRFSS_ex<-BRFSS[myvars]
BRFSS_ex<-na.omit(BRFSS_ex)
BRFSS_ex<-BRFSS_ex[which(BRFSS_ex$sex!="Refused"),]
BRFSS_ex<-BRFSS_ex[which(BRFSS_ex$age_cat!="Don't know/refused/missing"),]
```

3. Check whether running age and bmi as continous variables is appropriate for your regression models. Describe whether it is appropriate or not for each variable.

4. You make the decision based on these results to run a univariate logistic regression model for each risk factor of interest (age_cat, bmi, sex) and calculate ORs and 95% CIs.

5. Run a multivariate logistic regression model that includes age_cat, sex, and bmi in the model as predictors

6. Determine the top 5 influential observations using a Cook's Distance plot.

7. Exclude the top 5 influential observations and compare Betas between models with and without these observations.

8. Interpret your results in one paragraph.

Extra credit.

a. Determine the number of always seatbelt wearers vs. never/sometimes/always seatbelt wearers predicted by your model.
b. From these numbers (and the actual data) calculate and report the sensitivity and specificity of your model.
c. Is this a good model?