HW12

due 4/17/2019

Instructions. We will again use the BRFSS dataset on github for this homework assignment. We will impute missing data on diab_bin, ht_meters, and wt_kg and from that imputed data, calculate bmi. We will compare results from analyses run on the multiple imputation datasets to that run on the complete dataset. The packages and code for loading the dataset and creating the binary diabetes variable is provided below as well as some additional code with explanations. Please submit your homework by uploading the .RMD file or the HTML NB file to Canvas under the HW12 assignment.

Install libraries and packages

```
#install.packages("mice")
#install.packages('VIM')
#install.packages("lattice")
library(mice)
library(VIM)
library(lattice)
```

Let's go back to the BRFSS data and impute missing data for our logistic model that examined the association between diabetes and bmi adjusted for age and sex

```
#load BRFSS dataset (in class 4 folder) and discard variables that will not be used
BRFSS <- read.csv(
  "https://raw.githubusercontent.com/kijohnson/ADA_Spring_2019/master/BRFSS2017_10percent_v2.csv")

keeps<-c("diabetes", "sex", "X_AGE80", "ht_meters", "wtkg") #keep only these variables
BRFSS_k<-BRFSS[keeps] #drops variables that are not in the keeps list

BRFSS_k$diab_bin[
  BRFSS_k$diabetes=="No"]<-0 #Assign 0 to those who responded no to the diabetes question

BRFSS_k$diab_bin[
  BRFSS_k$diabetes=="Yes"]<-1 #Assign 1 to those who responded yes to the diabetes question

BRFSS_k$diab_bin<-as.factor(BRFSS_k$diab_bin)

class(BRFSS_k$diab_bin)
```

```
## [1] "factor"
```

```
#remove diabetes from dataset
drops <- c("diabetes")
BRFSS_k<-BRFSS_k[ , !(names(BRFSS_k) %in% drops)]
```

1. Examine the missing data pattern using the md.pattern function. Describe the different patterns.

2. Use margin plot to look at the missing data patterns for height and weight. Describe what you see.

3. Look at the age distribution (hint: set pos=2) for height, weight, and diabetes using pbox. Comment on what you see.

4. Perform multiple imputation using the code below on the BRFSS_k dataset (it might take a couple minutes, no worries). Look at the imputation results and answer the questions. a. How many datasets with imputed values were created? b. What method was used to impute numerical variables? c. Check the bmi calculation on the data where noted using the formula weight/height^2 for observation 12 (imputed). Show your work. Is the bmi value for observation 3 as expected from the height and weight values?

```
# Imputations
#make new variable setting NA
bmi<-NA

#add new empty bmi variable to BRFSS_k data set, this will ensure it isn't used to impute
#variable values
BRFSS_i<-cbind(BRFSS_k, bmi)

#create dry run with object name ini to set imputation settings--what regression method will be used
#to impute the value for the variable (this is a quick way instead of setting them yourself)
ini<-mice(BRFSS_i, maxit=0)
ini

#you can see height and weight have been used but we are only using the meth (aka method) output
#from the dry run. The real imputation is below.
complete(ini)
#assign the ini$meth to meth
meth<-ini$meth
#Assign the formula to calculate bmi, the tilda indicates a passive imputation method that is needed
#when you have transformed variables or variables that are derived from other variables.
meth["bmi"]<-"~I(wtkg/(ht_meters*ht_meters))"
#do the imputation using the methods assigned from the dry run and for bmi
imp <- mice(BRFSS_i, meth=meth, seed=10000) #set seed so you can reproduce exact results
#everytime you run the code.
#look at the imputation results
imp

#check bmi is correct using height and weight from this line
complete(imp)[is.na(BRFSS_k$ht_meters)|is.na(BRFSS_k$wtkg),]
```

5. Check to make sure that weight, height, and bmi values are plausible by looking at the first 10 observations with imputed data from height or weight. Qualitatively speaking, do the values for weight, height, and bmi look plausbile?

6. Compare the imputed to the non-imputed data for height and weight using stripplot to see how the distributions look. Do the imputed values fall randomly (i.e. no clear pattern evident) within the range of the non-imputed (i.e. non-missing) values?

7. Run a logistic model of the imputed data modeling the association between bmi and diabetes adjusted for sex and age. Report the ORs and 95% CIs for bmi, sex, and age. Hint: the glm model specification is glm(diab_bin ~ bmi +sex +X_AGE80, family="binomial"). Another hint: to get ORs and 95% CIs, you can either hand calculate them (not preferred) or find/write some code that will automatically calculate them (preferred and worth extra credit of 0.2 extra credit points toward the final grade). Report the OR for bmi and 95% CIs. Interpret the result.

8. Write and run code to get results for the complete dataset (i.e. not imputed). Hint: You will need to calculate bmi from ht_meters and wtkg before you run the logistic model. Comment on any differences between the ORs and the 95% CIs generated from the imputed vs the complete dataset.

Extra: code for conducting liklihood ratio tests with imputed datasets. Use cautiously, this is beta code. For more information: https://stefvanbuuren.name/fimd/sec-multiparameter.html

```
fit0 <-with(imp, exp=glm(diab_bin~ bmi, family="binomial")) #bmi alone
summary(fit0)
fit1 <-with(imp, exp=glm(diab_bin ~bmi +sex , family="binomial")) #bmi +sex
summary(fit1)
D3(fit1, fit0) #D3 is the function from mice for the likelihood ratio test.
#The p-value gives the result of the test.
#These data indicate that the model with sex has significantly
#better fit that the model without sex.
```