HW12

4/12/2018

Instructions. We will again use the BRFSS dataset on github for this homework assignment. The packages and code for loading the dataset and creating the binary diabetes variable is provided below. Please submit your homework by uploading the .RMD file or the HTML NB file to Blackboard under the HW12 assignment. You may also submit this for peer review.

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/Advanced-Data-Analysis/master/Class%204%20I

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

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

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

1. Examine the missing data pattern. 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 distribution of height and weight according to missing non-missing status by age (hint: set pos=3) using pbox. Comment on what you see.

4. Perform multiple imputation using the code below on the BRFSS_k dataset (it might take one or 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 3 (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
BRFSS_i<-cbind(BRFSS_k, bmi)

#create dry run to set imputation settings (this is a quick way instead of setting them yourself)
ini<-mice(BRFSS_i, maxit=0)

#you can see height and weight have been used but we are only using the meth output from the dry run. T
complete(ini)
#assign the ini$meth to meth
meth<-ini$meth
#for assign the formula to calculate bmi
meth["bmi"]<-"~I(weight/(height*height))"
#do the imputation using the methods assigned from the dry run and for bmi
imp <- mice(BRFSS_i, meth=meth)
#look at the imputation results
imp

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

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 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(diabetes_binary ~ bmi +sex +age, family="binomial"). Another hint: to get ORs and 95% CIs, you can either hand calculate them (not preferred) or find some code that will automatically calculate them (preferred and worth extra credit of 0.2 extra credit points toward the final grade).

8. Run the code below to compare the results from the non-imputed dataset to the imputed dataset. Comment on the differences between the ORs and the 95% CIs.

```
BRFSS_k$bmi<-BRFSS_k$weight/(BRFSS_k$height*BRFSS_k$height)
mylogit<-glm(diabetes_binary ~ bmi +sex +age, data=BRFSS_k, family="binomial")
  summary(mylogit)
  ORmodel<-exp(cbind(OR = coef(mylogit), confint(mylogit))) #calculate ORs and 95% CIs
  ORmodel #print ORs and 95% CIs
```