IDS 702 - Homework #5

Ana Belen Barcenas J.

11/11/2018

Missing Data - Multiple Imputation

2) Missing Data Mechanics

a) Create dataset with 30% age missing vales completely at random

```
treeage2 <- treeage

set.seed(5)
missings <- sample(1:nrow(treeage2),nrow(treeage2)*0.3)

for (i in 1:nrow(treeage2)){
  if (i %in% missings){
  treeage2[i,3] = NA}}

treeage2
```

```
##    number diameter age
## 1       1     12.0 125
## 2       2     11.4  NA
## 3       3      7.9  83
## 4       4      9.0  85
## 5       5     10.5  NA
## 6       6      7.9 117
## 7       7      7.3  69
## 8       8     10.2 133
## 9       9     11.7 154
## 10     10     11.3 168
## 11     11      5.7  NA
## 12     12      8.0  80
## 13     13     10.3 114
## 14     14     12.0  NA
## 15     15      9.2 122
## 16     16      8.5 106
## 17     17      7.0  NA
## 18     18     10.7  88
## 19     19      9.3  97
## 20     20      8.2  NA
```

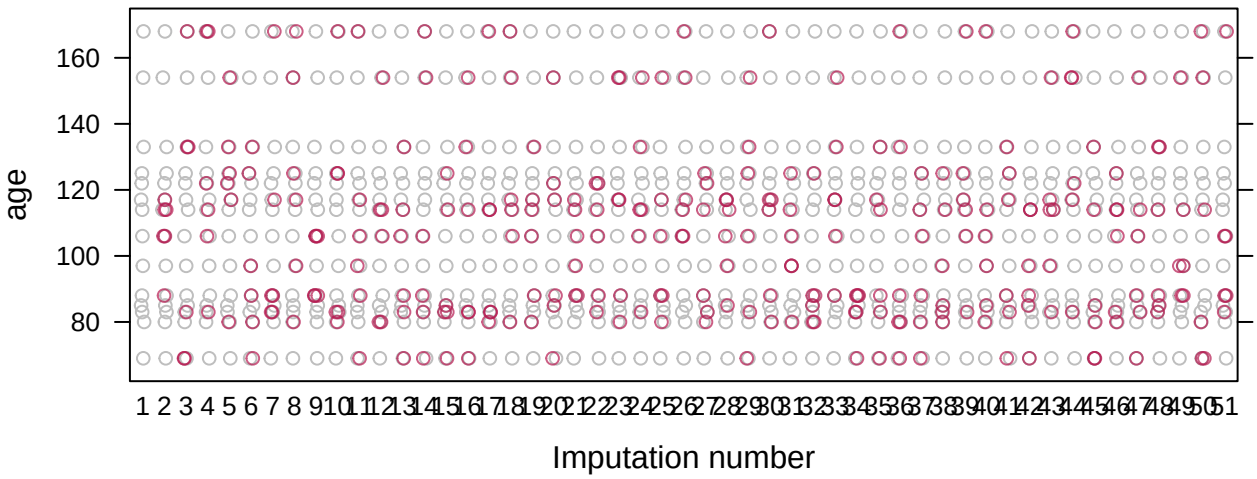
b) & c) Let's fill the missing values using the multiple imputation approach

```
treeage_MI = mice(treeage2, m=50)
```

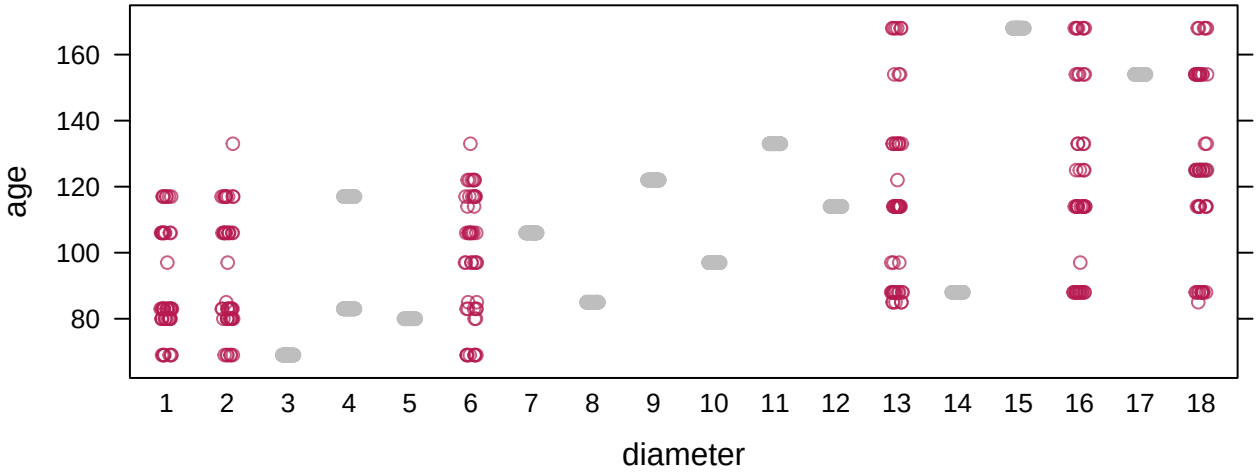
I decided to use the default model because otherwise I was getting negative values for age.

IMPUTATION DIAGNOSTICS

```
par(mfrow=c(1,2))
stripplot(treeage_MI, age~.imp, col=c("grey",mdc(2),pch=c(1,20)))
```




```
stripplot(treeage_MI, age~diameter, col=c("grey",mdc(2),pch=c(1,20)))
```



From the first plot we can observe that the imputed values (red dots) lies within the range of the observed values (grey dots). What suggests that the imputation is being conservative and the imputed values makes sense. From the second plot, the trend seems to be similar between imputed and observed values. What also suggests that the model is imputing values that make sense and follows the same positive relation with diameter as the observed values.

Now, let's analyze posterior predictive checks.

```
ppcheck = rbind(treeage2, treeage2)
ppcheck[21:40, 3] = NA
ppcheck_MI = mice(ppcheck, m=50)
```

Once we have the model, let's check the diagnostics for 3 completed datasets.

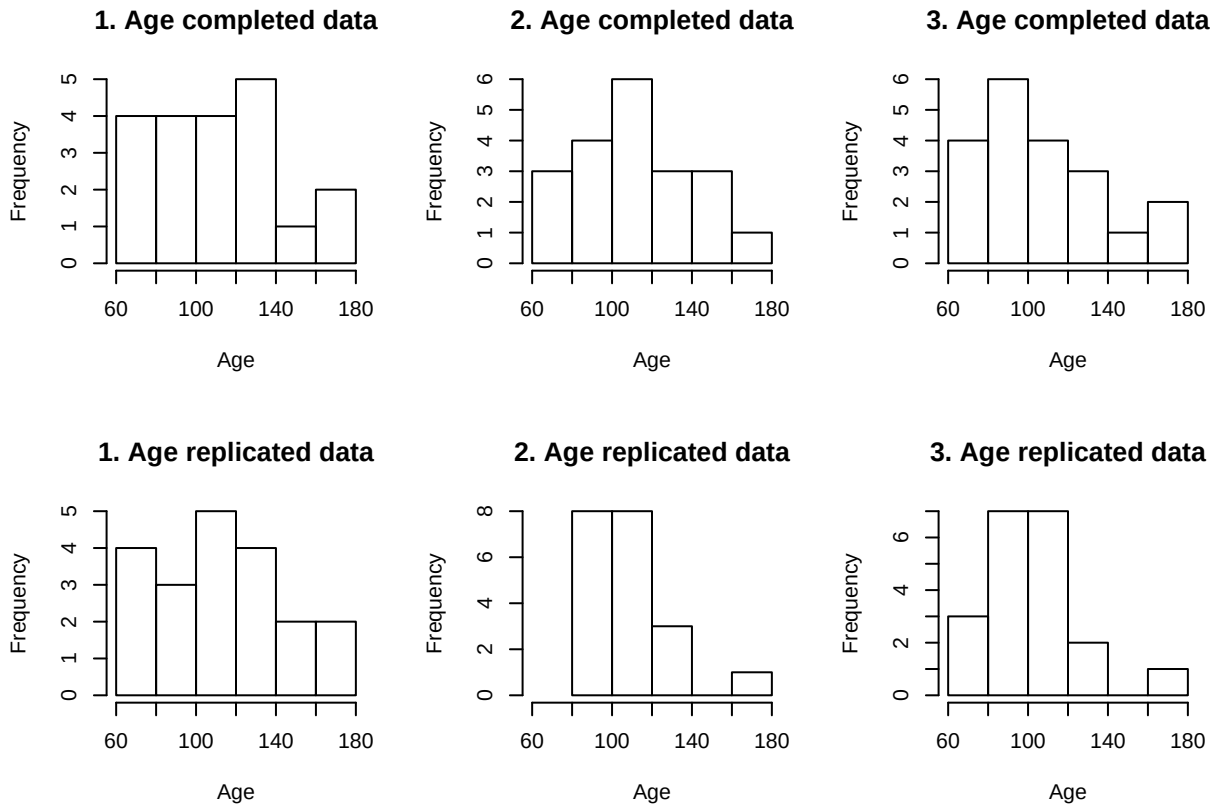
```
d1ppcheck = complete(ppcheck_MI, 4)
d2ppcheck = complete(ppcheck_MI, 2)
d3ppcheck = complete(ppcheck_MI, 3)
```

First, let's compare the marginal distribution of age between the observed data and the imputed data:

```
par(mfcol=c(2,3))
hist(d1ppcheck$age[1:20], xlim=c(60,180), breaks=6, xlab = "Age", main = "1. Age completed data")
hist(d1ppcheck$age[21:40], xlim=c(60,180), breaks=6, xlab = "Age", main = "1. Age replicated data")
```

```r
hist(d2ppcheck$age[1:20], xlim=c(60,180), breaks=6, xlab = "Age", main = "2. Age completed data")
hist(d2ppcheck$age[21:40], xlim=c(60,180), breaks=6, xlab = "Age", main = "2. Age replicated data")


hist(d3ppcheck$age[1:20], xlim=c(60,180), breaks=6, xlab = "Age", main = "3. Age completed data")
hist(d3ppcheck$age[21:40], xlim=c(60,180), breaks=6, xlab = "Age", main = "3. Age replicated data")
```



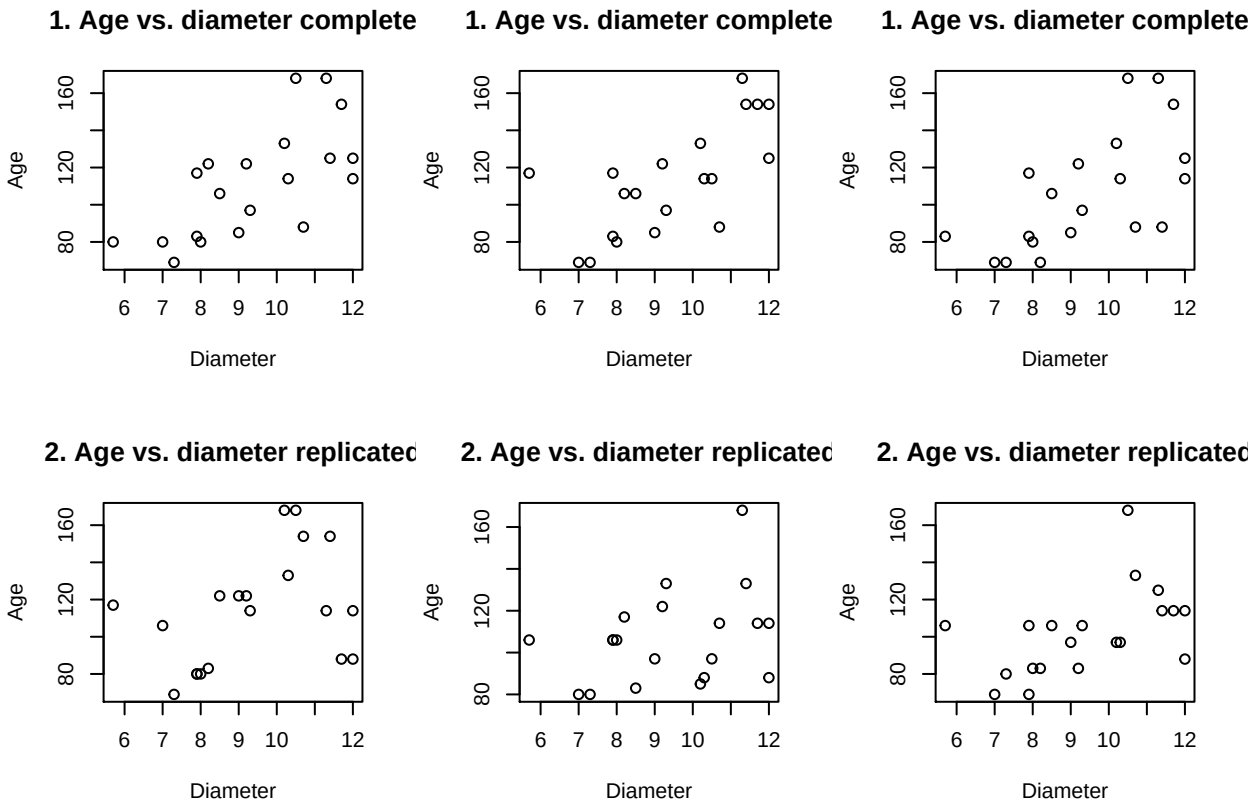
It seems that the replicated data does not differ too much from the observed data at least for this 3 cases.

Second, let's compare the relashionship between age and diameter of the observed and the imputed datasets:

```r
par(mfcol=c(2,3))
plot(d1ppcheck$age[1:20]~d2ppcheck$diameter[1:20], ylab = "Age", xlab = "Diameter", main = "1. Age vs. d
plot(d1ppcheck$age[21:40]~d2ppcheck$diameter[21:40], ylab = "Age", xlab = "Diameter", main = "2. Age vs

plot(d2ppcheck$age[1:20]~d2ppcheck$diameter[1:20], ylab = "Age", xlab = "Diameter", main = "1. Age vs. d
plot(d2ppcheck$age[21:40]~d2ppcheck$diameter[21:40], ylab = "Age", xlab = "Diameter", main = "2. Age vs

plot(d3ppcheck$age[1:20]~d2ppcheck$diameter[1:20], ylab = "Age", xlab = "Diameter", main = "1. Age vs. d
plot(d3ppcheck$age[21:40]~d2ppcheck$diameter[21:40], ylab = "Age", xlab = "Diameter", main = "2. Age vs
```



I would say that the trends look pretty similar between the observed and imputed datasets. The positive relation between age and diameter is clear in the replicated datasets. I feel comfortable with this model.

d) Regression of age on diameter using multiple imputation combining rules.

```
reg1 = with(data=treeage_MI, lm(age~diameter))
reg2 = pool(reg1)
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
summary(reg2, conf.int = T)
```

```
##               estimate std.error statistic       df    p.value      2.5 %
## (Intercept) 18.750947 32.364155 0.5793739 10.86613 0.57415181 -52.589295
## diameter     9.622653  3.440899 2.7965520 10.46426 0.01756695   2.001709
##                97.5 %
## (Intercept) 90.09119
## diameter    17.24360
```

According with the model fitted above, if the tree diameter increase by one, the average age of the tree will be 10.44 months greater. We are 95% confident that the average of the tree given the increase of one in the diameter will lie between 2.39 and 18.5 months. It is a wide range but the change is always greater than zero.

2) Multiple Imputation in NHANES data

Let's analyze the data and find what variables has missing values on it.

```
nhanes <- nhanes[,5:ncol(nhanes)]
summary(nhanes)
```

```
##     ridageyr        riagendr        ridreth2        dmdeduc
```

```
##  Min.   : 0.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:10.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :19.00   Median :2.000   Median :2.000   Median :1.000
##  Mean   :30.06   Mean   :1.509   Mean   :2.003   Mean   :1.767
##  3rd Qu.:49.00   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :85.00   Max.   :2.000   Max.   :5.000   Max.   :9.000
##                                                  NA's   :1744
##     indfminc          bmxwt           bmxbmi          bmxtri
##  Min.   : 1.000   Min.   :  2.40   Min.   :12.40   Min.   : 2.80
##  1st Qu.: 3.000   1st Qu.: 38.90   1st Qu.:19.72   1st Qu.: 9.90
##  Median : 6.000   Median : 63.70   Median :24.31   Median :14.40
##  Mean   : 7.587   Mean   : 60.64   Mean   :25.07   Mean   :16.55
##  3rd Qu.: 9.000   3rd Qu.: 80.80   3rd Qu.:29.21   3rd Qu.:22.18
##  Max.   :99.000   Max.   :209.10   Max.   :64.97   Max.   :44.60
##  NA's   :158      NA's   :593      NA's   :1435    NA's   :1536
##     bmxwaist         bmxthicr        bmxarml
##  Min.   : 32.00   Min.   :28.00   Min.   : 9.0
##  1st Qu.: 71.10   1st Qu.:46.50   1st Qu.:31.0
##  Median : 86.40   Median :51.00   Median :35.5
##  Mean   : 85.84   Mean   :51.62   Mean   :33.1
##  3rd Qu.:100.50   3rd Qu.:56.10   3rd Qu.:38.2
##  Max.   :170.70   Max.   :94.80   Max.   :48.0
##  NA's   :1725     NA's   :2928    NA's   :932
```

I have selected just those variables that are important for the analysis and also I deleted age variable since is highly correlated with age at screening (ridageyr).

Now, let's clean the data imputing NAs in cases where people answer "don't know" or "refused", mean center age variable, and convert as factors the categorical variables.

```
nhanes <- nhanes %>%
  mutate(dmdeduc = ifelse(dmdeduc > 3, NA, dmdeduc),
         indfminc = ifelse(indfminc > 13, NA, indfminc)) %>%
  mutate(ridreth2 = as.factor(ridreth2),
         dmdeduc = as.factor(dmdeduc),
         indfminc = as.factor(indfminc),
         riagendr = as.factor(riagendr)) %>%
  mutate(indfminc.2 = factor(indfminc,
                             levels = c('6', '1', '2', '3', '4', '5', '7', '8',
                                        '9', '10', '11', '12', '13')),
         age.c = ridageyr - mean(ridageyr),
         age.c2 = age.c**2)
```

It seems that the complete variables are ridageyr, riagendr, and ridreth2. The multiple imputation technique will take this variables to predict the ones with missing values. Let's create an m=10 imputed datasets using mice command.

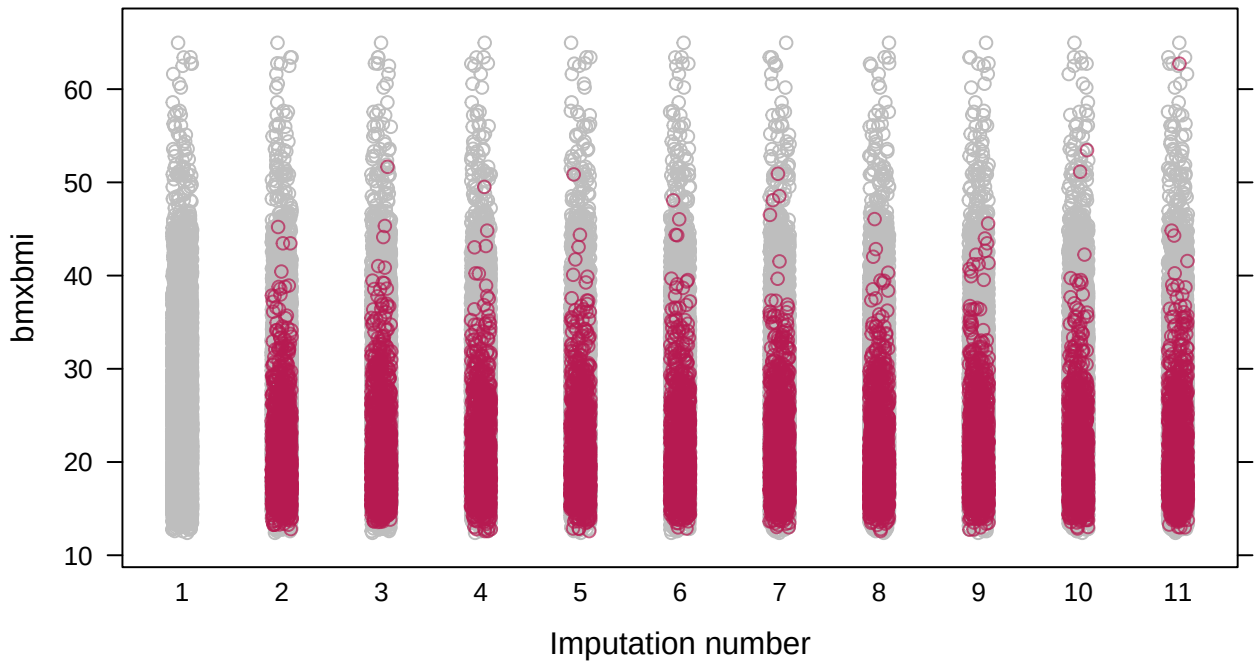
a) Let's fill missing values using a multiple imputation approach with m=10

```
nhanes_MI = mice(nhanes, m=10)
```

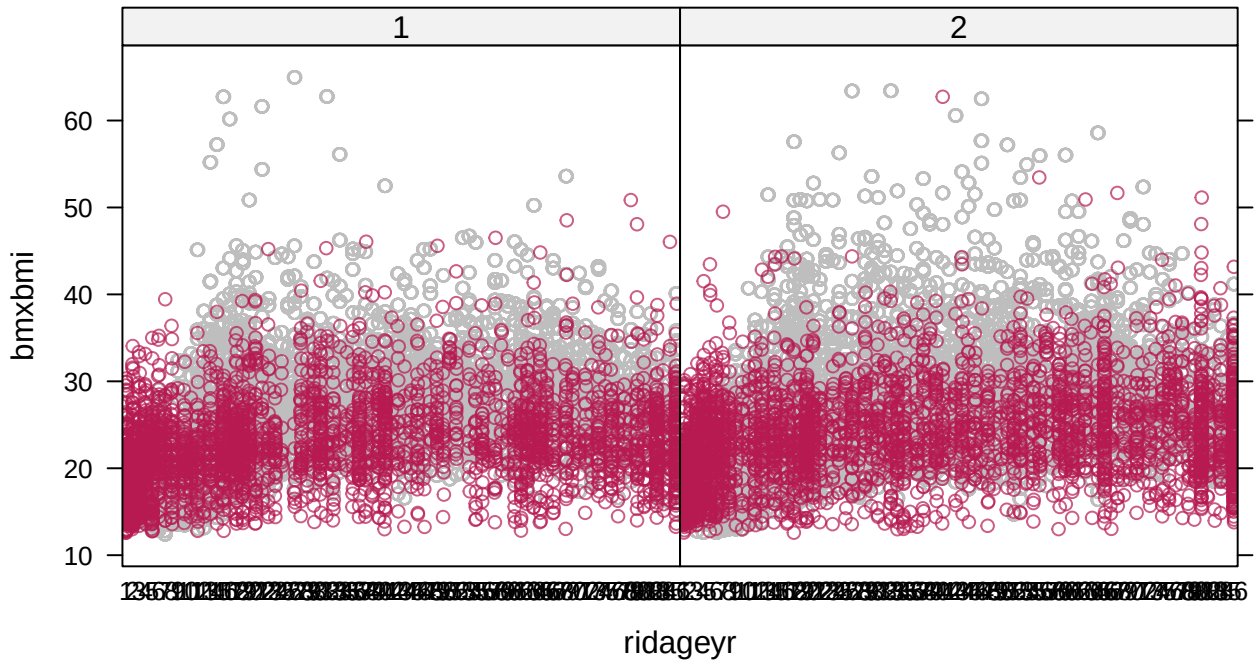
```
## Warning: Number of logged events: 36
```

To check the quality of the imputations, let's see some diagnostics of the completed datasets. Specially, let's see BMI by age and gender

```
par(mfrow=c(1,2))
stripplot(nhanes_MI, bmxbmi~.imp, col=c("grey",mdc(2),pch=c(1,20)))
```




```
stripplot(nhanes_MI, bmxbmi~ridageyr|riagendr, col=c("grey",mdc(2),pch=c(1,20)))
```



Both plots suggest that the imputation approach is creating values that looks pretty similar than the observed values. Now, let's look at some posterior predictive checks.

```
nhanes_pp = rbind(nhanes, nhanes)
nhanes_pp[10123:20244, 4:12] = NA
nhanes_ppMI = mice(nhanes_pp, m=10)
```

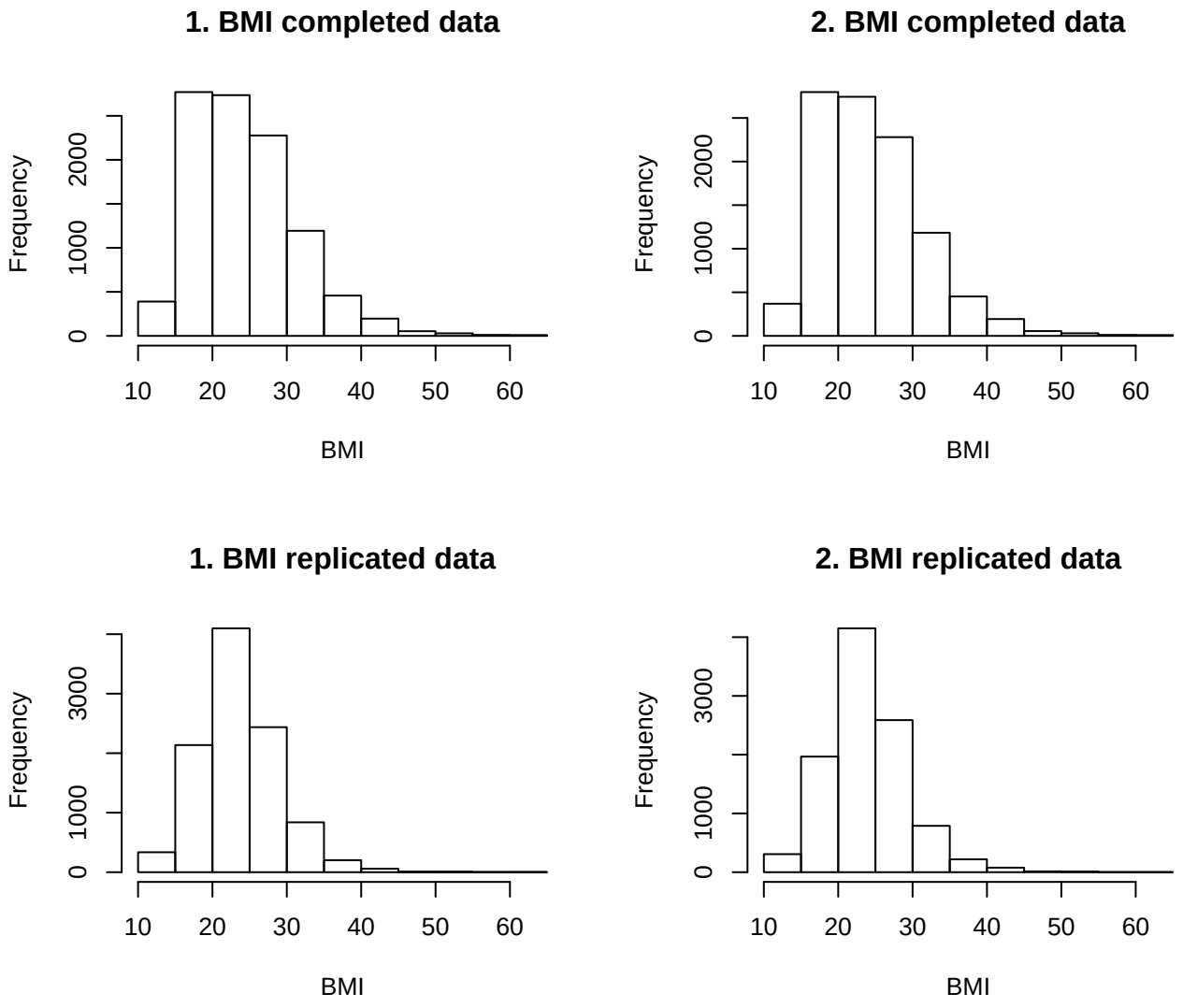
```
## Warning: Number of logged events: 40
```

Once we have the model, let's check the diagnostics for 2 completed datasets. First, we will analyze at the marginal distributions of the most important variable: BMI.

```
d1pp = complete(nhanes_ppMI, 1)
d2pp = complete(nhanes_ppMI, 2)
```

```
par(mfcol=c(2,2))
hist(d1pp$bmxbmi[1:10122], xlab = "BMI", main = "1. BMI completed data")
hist(d1pp$bmxbmi[10123:20244],  xlab = "BMI", main = "1. BMI replicated data")

hist(d2pp$bmxbmi[1:10122], xlab = "BMI", main = "2. BMI completed data")
hist(d2pp$bmxbmi[10123:20244], xlab = "BMI", main = "2. BMI replicated data")
```



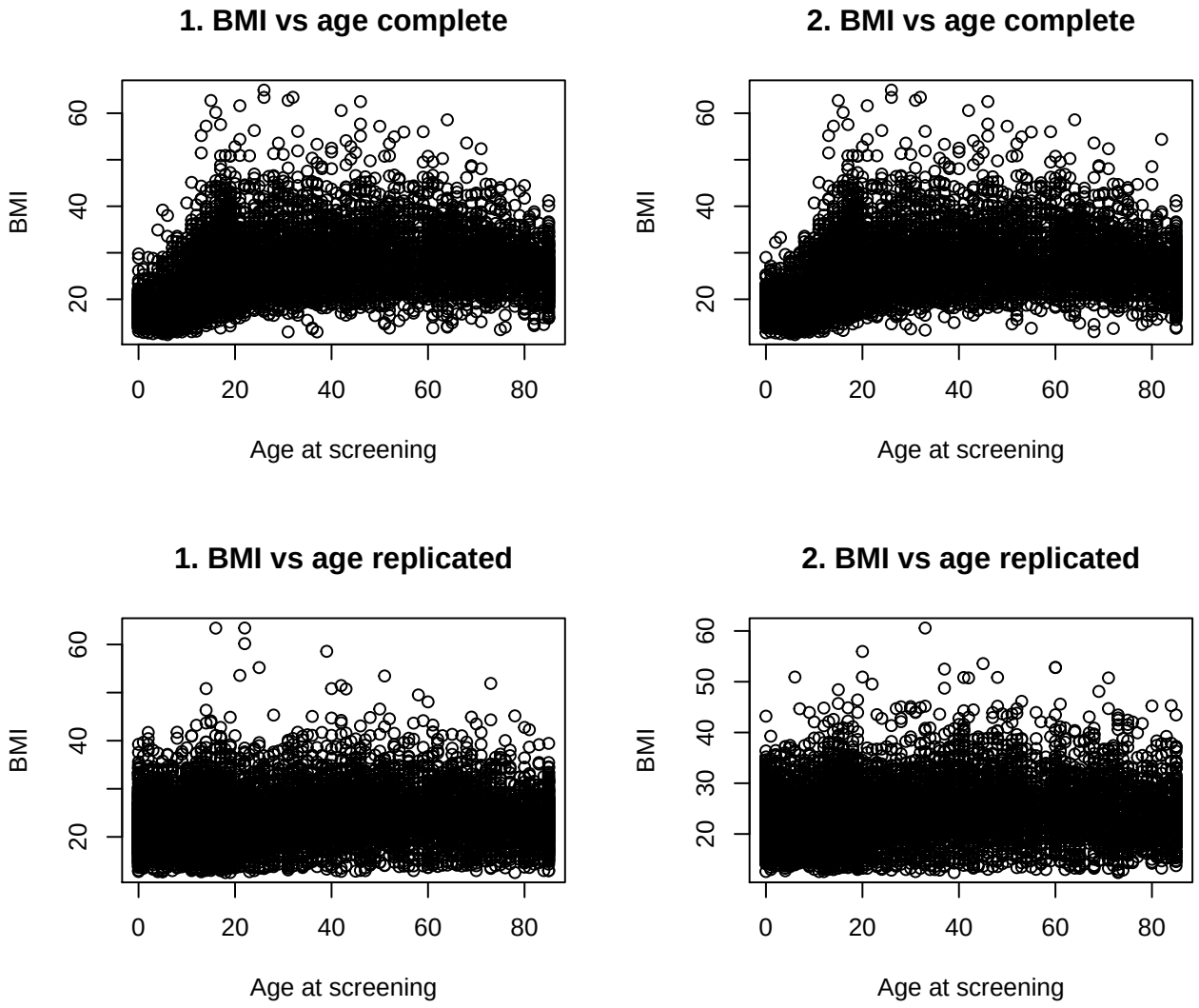
It seems that the distribution of the imputed values and the observed values are pretty similar. I feel comfortable with the performance of the imputation approach.

Second, let's analyze the relationship between BMI by age at screening (ridageyr) and BMI by gender (riagendr).

```
par(mfcol=c(2,2))
plot(d1pp$bmxbmi[1:10122]~d1pp$ridageyr[1:10122], ylab = "BMI", xlab = "Age at screening", main = "1. BM
plot(d1pp$bmxbmi[10123:20244]~d1pp$ridageyr[10123:20244], ylab = "BMI", xlab = "Age at screening", main

plot(d2pp$bmxbmi[1:10122]~d2pp$ridageyr[1:10122], ylab = "BMI", xlab = "Age at screening", main = "2. BM
```

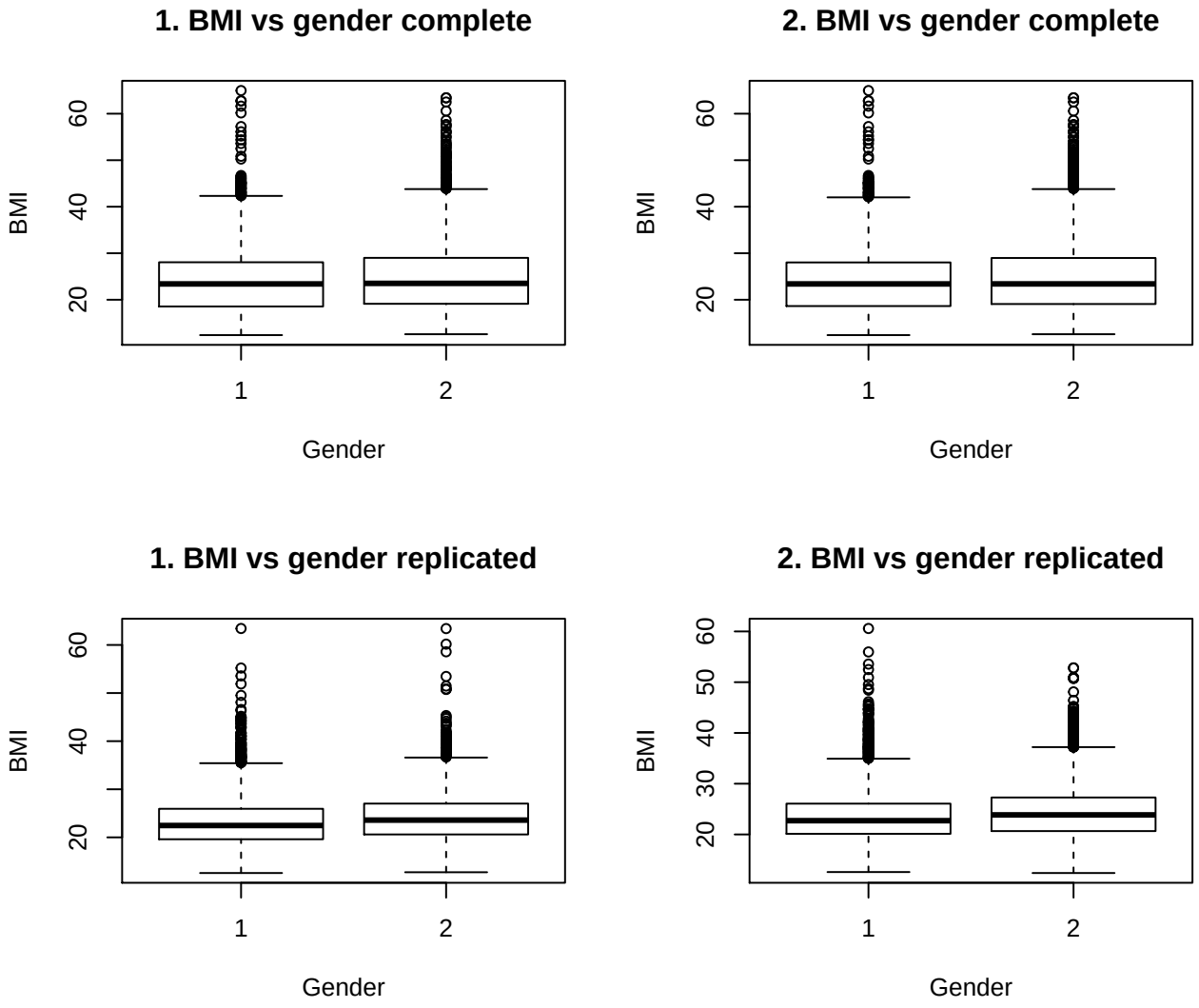
```
plot(d2pp$bmxbmi[10123:20244]~d2pp$ridageyr[10123:20244], ylab = "BMI", xlab = "Age at screening", main
```



The trend seems to be pretty similar. The complete data seems to have a cuadratic form and the imputed data does not follow that relationship as much as I would like. However I believe that the imputation is good enough.

```
par(mfcol=c(2,2))
plot(d1pp$bmxbmi[1:10122]~d1pp$riagendr[1:10122], ylab = "BMI", xlab = "Gender", main = "1. BMI vs gend
plot(d1pp$bmxbmi[10123:20244]~d1pp$riagendr[10123:20244], ylab = "BMI", xlab = "Gender", main = "1. BMI

plot(d2pp$bmxbmi[1:10122]~d2pp$riagendr[1:10122], ylab = "BMI", xlab = "Gender", main = "2. BMI vs gend
plot(d2pp$bmxbmi[10123:20244]~d2pp$riagendr[10123:20244], ylab = "BMI", xlab = "Gender", main = "2. BMI
```



From this plots I would say that the imputation models is well specified.

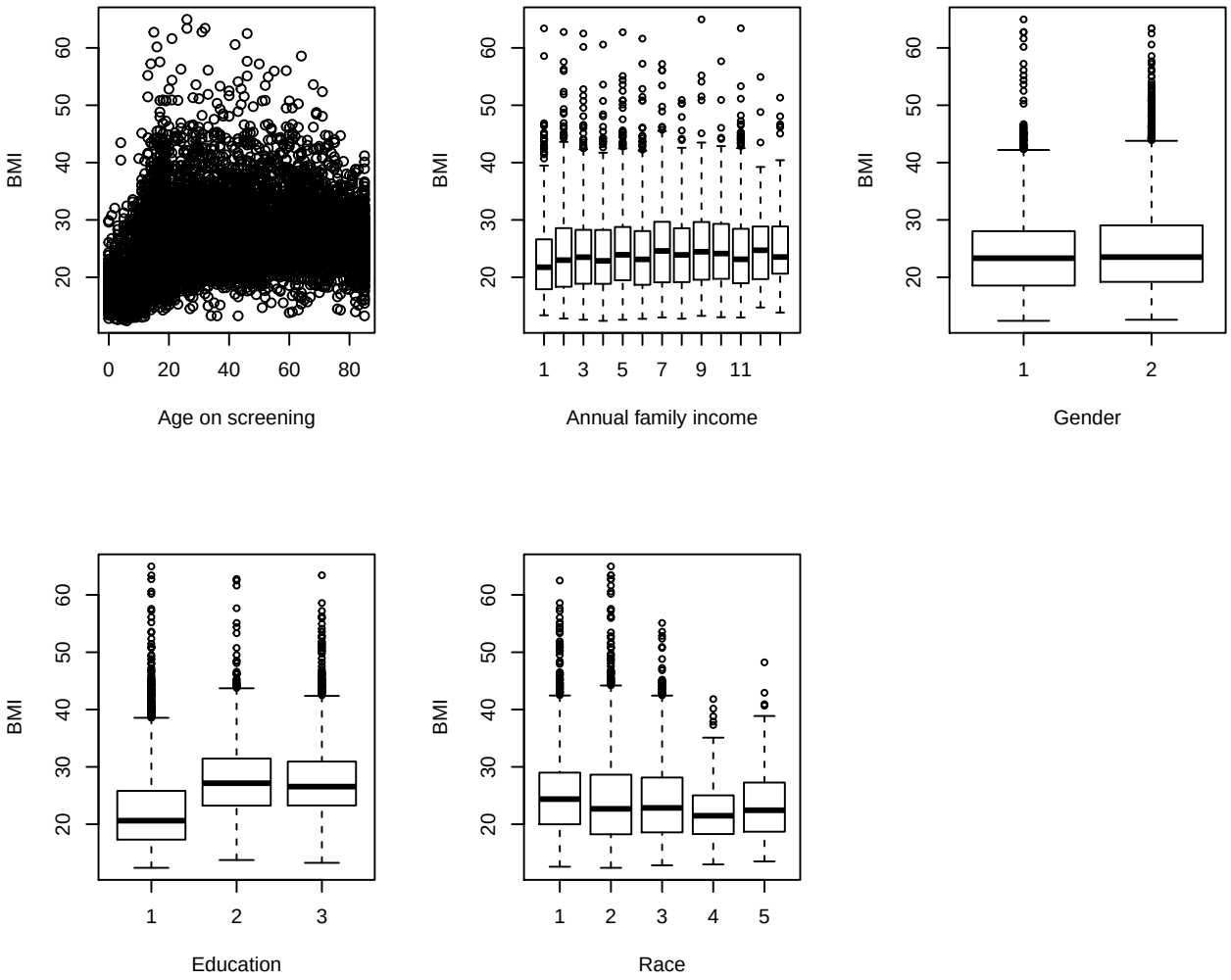
b) Let's run a model that predicts BMI from a subset of age, gender, race, education, and income using the multiple imputation combining rules.

First, let's do some EDA with one of the imputed datasets.

```
ex1 <- complete(nhanes_MI,1)
ex1
```

Let's see the individual relation between BMI and the covariates of interest

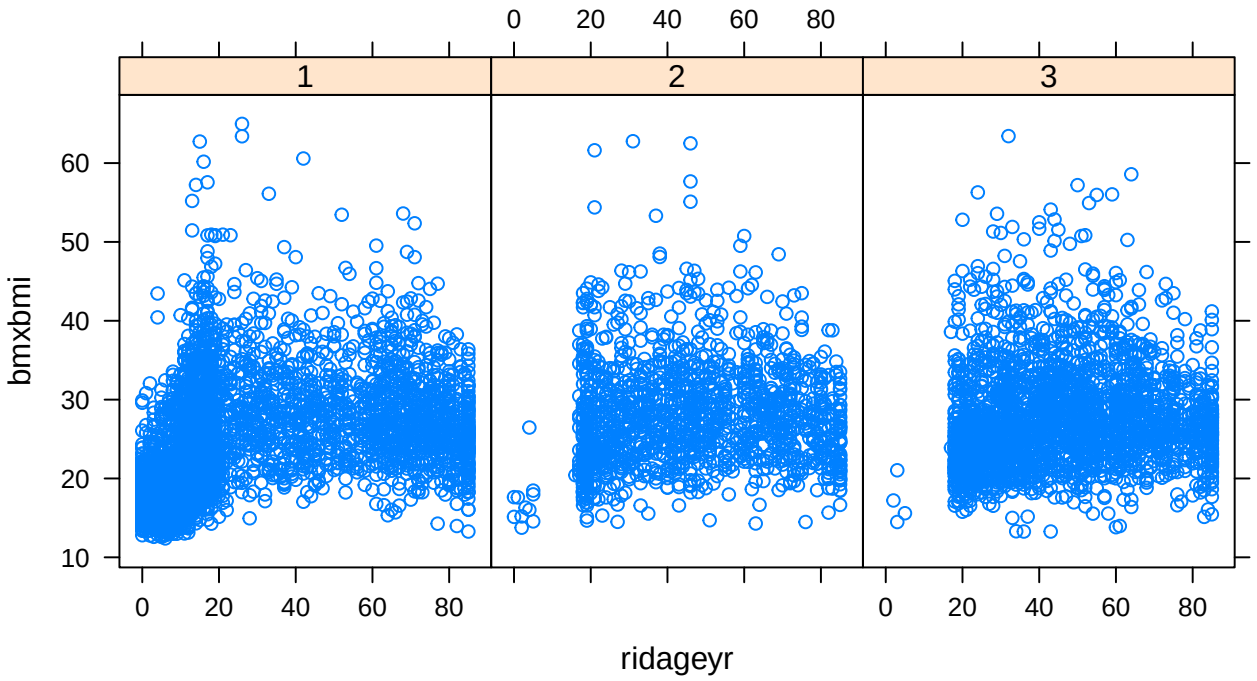
```
par(mfcol=c(2,3))
plot(y = ex1$bmxbmi, x = ex1$ridageyr, xlab = "Age on screening", ylab = "BMI")
boxplot(bmxbmi~dmdeduc, data = ex1, ylab = "BMI", xlab = "Education")
boxplot(bmxbmi~indfminc, data = ex1, ylab = "BMI", xlab = "Annual family income")
boxplot(bmxbmi~ridreth2, data = ex1, ylab = "BMI", xlab = "Race")
boxplot(bmxbmi~riagendr, data = ex1, ylab = "BMI", xlab = "Gender")
```



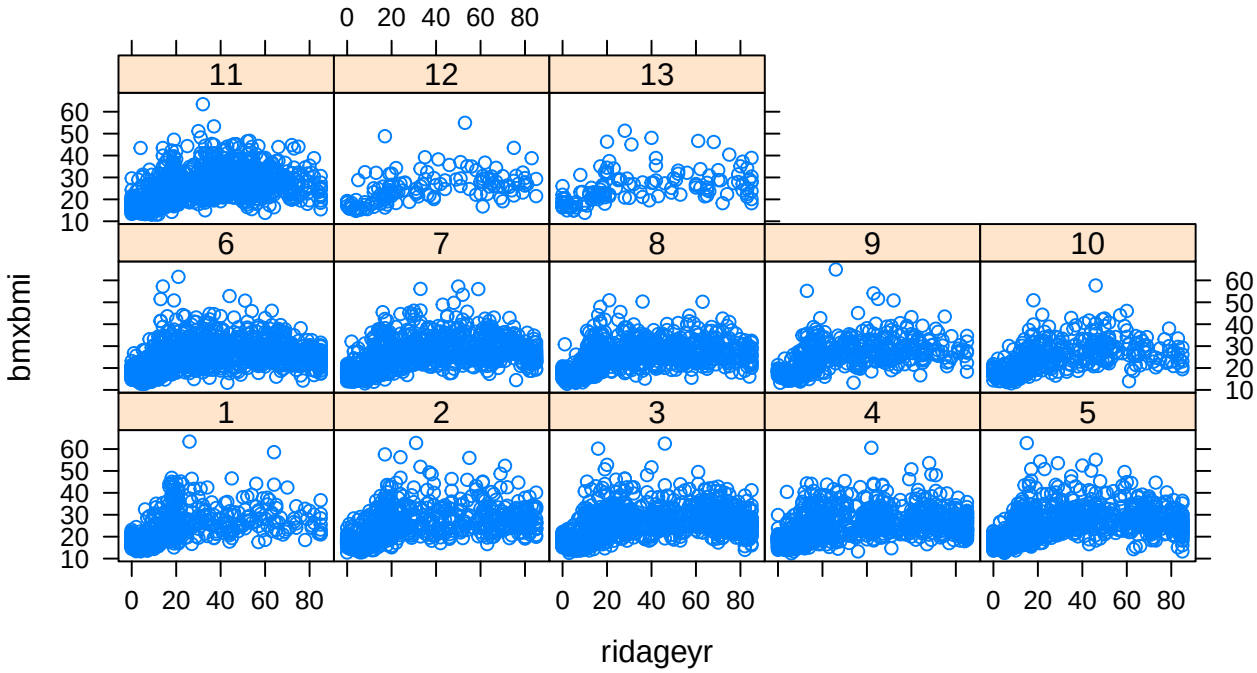
There seems to be a cuadratic relationship between BMI and age. I will specify age squared in the model (the variable is created on the beginning of the code to avoid errors). Also, I will use the age mean centered to facilitate interpretation.

Now, let's look for interaction effects between the variables that could have those effects based on intuition.

```
par(mfcol=c(1,2))
xyplot(bmxbmi~ridageyr |dmdeduc , data = ex1)
```




```
xyplot(bmxbmi~ridageyr |indfminc , data = ex1)
```



there is no evidence of interaction effects between age and income or age and education level. I will not employ interactions in the model.

Once I have determined the transformations, I will fit the model employing the 10 imputated datasets using combining rules.

```
bmireg1 = with(data=nhanes_MI, lm(bmxbmi~age.c + age.c2 + riagendr + ridreth2 + dmdeduc + indfminc ))
bmireg2 = pool(bmireg1)
summary(bmireg2, conf.int = T)
```

```
##                 estimate    std.error   statistic        df      p.value
## (Intercept) 26.815152847 0.2827671173  94.8312276 1466.2697 0.000000e+00
## age.c        0.218812416 0.0037465213  58.4041558 1244.7047 0.000000e+00
## age.c2      -0.004815593 0.0001175528 -40.9653489  989.0947 0.000000e+00
## riagendr2    0.692761765 0.1159912226   5.9725361  741.3374 2.680965e-09
## ridreth22    1.097333310 0.1515804085   7.2392819  886.1497 6.035172e-13
## ridreth23    0.693806686 0.1578079307   4.3965261 1012.5618 1.147625e-05
## ridreth24   -1.662712029 0.3279778595  -5.0695862 1257.8020 4.289394e-07
## ridreth25    0.115265503 0.3228793857   0.3569925 2413.6000 7.211287e-01
## dmdeduc2     0.680690878 0.1963402630   3.4668940  676.6610 5.357172e-04
## dmdeduc3     0.151952142 0.1762244166   0.8622650 1414.1405 3.886274e-01
## indfminc2    0.056228927 0.3232084507   0.1739711  326.4036 8.619028e-01
## indfminc3   -0.071511473 0.2883892163  -0.2479686  579.9858 8.041798e-01
## indfminc4   -0.365613167 0.3077750381  -1.1879234  378.2515 2.349805e-01
## indfminc5    0.036331522 0.3062308476   0.1186410  380.9929 9.055697e-01
## indfminc6   -0.514707040 0.2859349610  -1.8000843  529.9181 7.197215e-02
## indfminc7    0.186192274 0.2984139167   0.6239396  809.0207 5.327262e-01
## indfminc8   -0.495754139 0.3118768962  -1.5895828  745.1641 1.120599e-01
## indfminc9    0.224406641 0.3507504006   0.6397901  844.1881 5.223698e-01
## indfminc10  -0.273720034 0.3680000160  -0.7438044 1731.5253 4.570672e-01
## indfminc11  -0.704873154 0.2742819921  -2.5698849 1110.1128 1.023274e-02
## indfminc12  -0.486777498 0.5589381348  -0.8708969  579.4999 3.838971e-01
## indfminc13   0.362112274 0.5591902644   0.6475654  380.1873 5.173277e-01
##                    2.5 %       97.5 %
## (Intercept) 26.260481622 27.369824072
## age.c        0.211462222  0.226162611
## age.c2      -0.005046274 -0.004584911
## riagendr2    0.465051380  0.920472151
## ridreth22    0.799834834  1.394831786
## ridreth23    0.384138672  1.003474700
## ridreth24   -2.306155988 -1.019268070
## ridreth25   -0.517883971  0.748414978
## dmdeduc2     0.295181482  1.066200274
## dmdeduc3    -0.193737239  0.497641524
## indfminc2   -0.579605631  0.692063485
## indfminc3   -0.637925951  0.494903004
## indfminc4   -0.970777508  0.239551175
## indfminc5   -0.565782640  0.638445685
## indfminc6   -1.076412182  0.046998102
## indfminc7   -0.399564575  0.771949123
## indfminc8   -1.108016088  0.116507810
## indfminc9   -0.464038553  0.912851836
## indfminc10  -0.995491336  0.448051268
## indfminc11  -1.243042740 -0.166703568
## indfminc12  -1.584568916  0.611013919
## indfminc13  -0.737380653  1.461605202
```

Intercept: For a male with average age, less than highschool education, non-hispanic white and an annual familyn income between 0.00 and 4,999.00 USD, the average BMI is 26.8. We are 95% confident that the average BMI for a person with characteristics mentioned above will be between 26.25 - 27.35.

Age: From the age and age squared coefficients, we can interpret that the relationship between age and BMI is positive but has a decreasing slope because the age squared coefficient is negative. This means that the positive effect is higher in the first years and starts becoming flat when the person gets older.

Gender: For a person with average age, less than highschool education, non-hispanic white, and an annual family income between 0.00 and 4,999.00 USD, if the person is a women, the average BMI is 0.7 higher than if it is a male. We are 95% confident that the average BMI will change somewhere between 0.48 - 0.93 if it is a female.

Race: For a male with average age, less than highschool education, and an annual family income between 0.00 and 4,999.00 USD, being non-hispanic black instead of non-hispanic white represents an average increase in BMI of 1.12 (CI: 0.83-1.42); being mexican american instead of non-hispanic white represents an average increase of 0.71 (CI: 0.41-1.01); being other race instead of non-hispanic white represents an average decrease of -1.70 (CI: -2.34 - -1.07); and being other hispanic instead of non-hispanic white represents an averga increase of 0.08 (CI: -0.57 - 0.73).

Education: For a male with average age, non-hispanic white, and an annual family income between 0.00 and 4,999.00 USD, high school diploma instead of less than high school education, represents an average increase in BMI of 0.62 (CI: 0.24-1.00); having more studies than high school represent an average increase in BMI of 0.13 (CI: -0.22 - 0.47).