Data Imputation for Income

Shreya Prasad

December 1, 2018

```
#Reading in the data
fdata <- read.csv("chds_births.csv")
head(fdata)
```

```
##    wt gestation parity meth mage med mht mwt feth fage fed fht fwt marital
## 1 120       284      1    8   27   5  62 100    8   31   5  65 110       1
## 2 113       282      2    0   33   5  64 135    0   38   5  70 148       1
## 3 128       279      1    0   28   2  64 115    5   32   1  NA  NA       1
## 4 123        NA      2    0   36   5  69 190    3   43   4  68 197       1
## 5 108       282      1    0   23   5  67 125    0   24   5  NA  NA       1
## 6 136       286      4    0   25   2  62  93    3   28   2  64 130       1
##   income smoke time number
## 1      1     0    0      0
## 2      4     0    0      0
## 3      2     1    1      1
## 4      8     3    5      5
## 5      1     1    1      5
## 6      4     2    2      2
```

```
#General Model
m1 <- lm(wt ~ ., data = fdata)

m1
```

```
##
## Call:
## lm(formula = wt ~ ., data = fdata)
##
## Coefficients:
## (Intercept)    gestation       parity         meth         mage
##   -78.57994      0.43624      0.71464     -0.53185      0.01494
##         med          mht          mwt         feth         fage
##     0.98255      1.17614      0.02011     -0.36673      0.02465
##         fed          fht          fwt      marital       income
##    -1.01724     -0.14863      0.08742     -2.22637     -0.24949
##       smoke         time       number
##     3.01584     -0.80687     -2.44994
```

```
summary(m1)
```

```
##
## Call:
## lm(formula = wt ~ ., data = fdata)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -46.936 -10.546  -0.259   9.705  48.212
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -78.57994   25.70499  -3.057 0.002338 **
## gestation     0.43624    0.04201  10.385  < 2e-16 ***
## parity        0.71464    0.43667   1.637 0.102264
## meth         -0.53185    0.28619  -1.858 0.063623 .
## mage          0.01494    0.21755   0.069 0.945259
## med           0.98255    0.60918   1.613 0.107309
## mht           1.17614    0.30402   3.869 0.000122 ***
## mwt           0.02011    0.03777   0.533 0.594512
## feth         -0.36673    0.29257  -1.253 0.210534
## fage          0.02465    0.18041   0.137 0.891377
## fed          -1.01724    0.53226  -1.911 0.056477 .
## fht          -0.14863    0.28960  -0.513 0.607994
## fwt           0.08742    0.03491   2.504 0.012558 *
## marital      -2.22637    3.06339  -0.727 0.467661
## income       -0.24949    0.31386  -0.795 0.426988
## smoke         3.01584    1.81105   1.665 0.096403 .
## time         -0.80687    0.99487  -0.811 0.417678
## number       -2.44994    0.40534  -6.044 2.69e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.86 on 581 degrees of freedom
##   (637 observations deleted due to missingness)
## Multiple R-squared:  0.2854, Adjusted R-squared:  0.2644
## F-statistic: 13.65 on 17 and 581 DF,  p-value: < 2.2e-16
```

```
#Cleaning that data
#Calculat the number of missing valuesin each column
na_count <- sapply(fdata, function(y) sum(length(which(is.na(y)))))
na.count <- data.frame(na_count)
na.count
```

```
##           na_count
## wt               0
## gestation       13
## parity           0
## meth             1
## mage             2
## med              1
## mht             22
## mwt             36
## feth            31
## fage             7
## fed             13
## fht            492
## fwt            499
## marital          0
## income         124
## smoke           10
```

```
## time            10
## number          21
```

```
#Check the columns that have more than 10% of the data missing
count <- sapply(fdata, function(y) length(y))
na_percent <- (na_count/count)*100
na_percent <- data.frame(na_percent)
na_percent
```

```
##            na_percent
## wt         0.00000000
## gestation  1.05177994
## parity     0.00000000
## meth       0.08090615
## mage       0.16181230
## med        0.08090615
## mht        1.77993528
## mwt        2.91262136
## feth       2.50809061
## fage       0.56634304
## fed        1.05177994
## fht       39.80582524
## fwt       40.37216828
## marital    0.00000000
## income    10.03236246
## smoke      0.80906149
## time       0.80906149
## number     1.69902913
```

```
sapply(fdata, function(x) sum(is.na(x)))
```

```
##        wt gestation    parity      meth      mage       med       mht
##         0        13         0         1         2         1        22
##       mwt      feth      fage       fed       fht       fwt   marital
##        36        31         7        13       492       499         0
##    income     smoke      time    number
##       124        10        10        21
```

Visualizing the missing data

```
library(VIM)
```

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

```
## Loading required package: colorspace
```

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

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
## Warning: package 'data.table' was built under R version 3.4.4
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

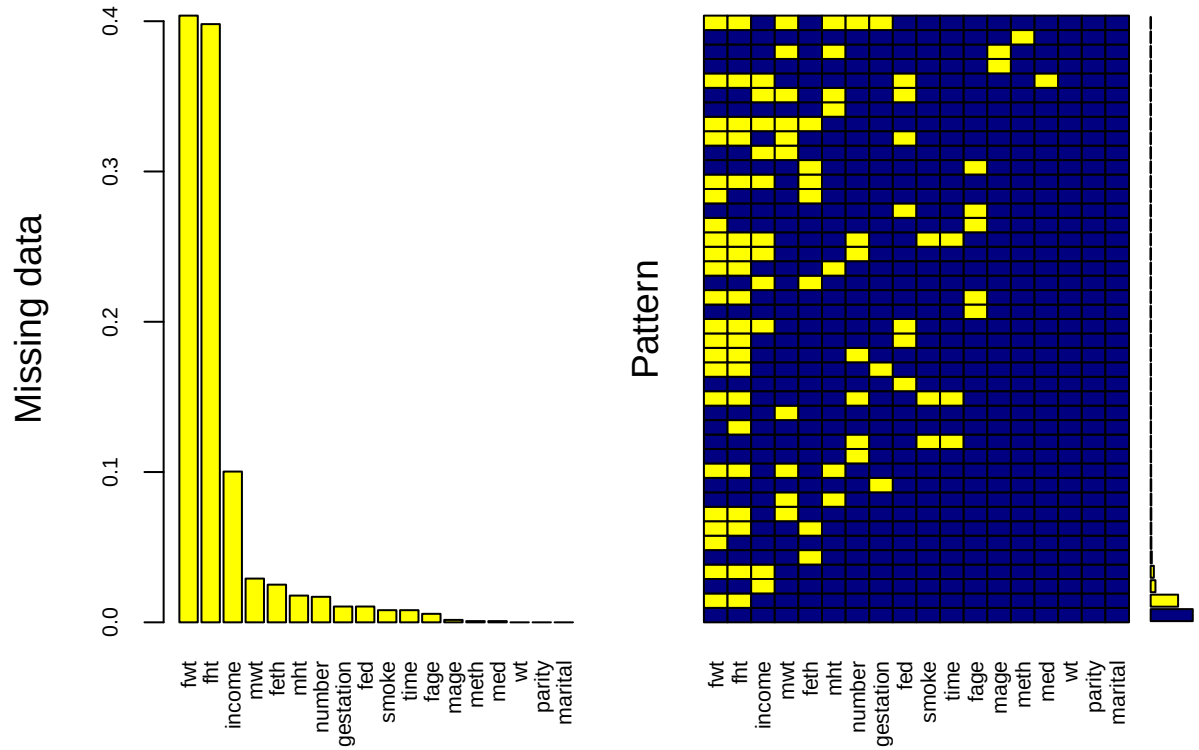
```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
miss_plot <- aggr(fdata, col=c('navyblue','yellow'),
                  numbers=TRUE, sortVars=TRUE,
                  labels=names(fdata), cex.axis=.7,
                  gap=3, ylab=c("Missing data","Pattern"))
```

```
## Warning in plot.aggr(res, ...): not enough vertical space to display
## frequencies (too many combinations)
```



Looking the the above data frame, we might want to remove or impute values into the following features of the data * fwt : Fathers Weight * fht : Fathers Height * income : Fathers Income

Since imputting may not work for fathers height and weight, we will remove those features from our data. Logically, it makes sense for the baby's weight to not depend on these features.

```
fdata$fht <- NULL
fdata$fwt <- NULL
colnames(fdata)
```

```
##  [1] "wt"        "gestation" "parity"    "meth"      "mage"
##  [6] "med"       "mht"       "mwt"       "feth"      "fage"
## [11] "fed"       "marital"   "income"    "smoke"     "time"
## [16] "number"
```

```
head(fdata)
```

```
##    wt gestation parity meth mage med mht mwt feth fage fed marital income
## 1 120       284      1    8   27   5  62 100    8   31   5       1      1
## 2 113       282      2    0   33   5  64 135    0   38   5       1      4
## 3 128       279      1    0   28   2  64 115    5   32   1       1      2
## 4 123        NA      2    0   36   5  69 190    3   43   4       1      8
## 5 108       282      1    0   23   5  67 125    0   24   5       1      1
## 6 136       286      4    0   25   2  62  93    3   28   2       1      4
##   smoke time number
## 1     0    0      0
## 2     0    0      0
## 3     1    1      1
## 4     3    5      5
## 5     1    1      5
## 6     2    2      2
```

Using the mice library

```
library(mice)
```

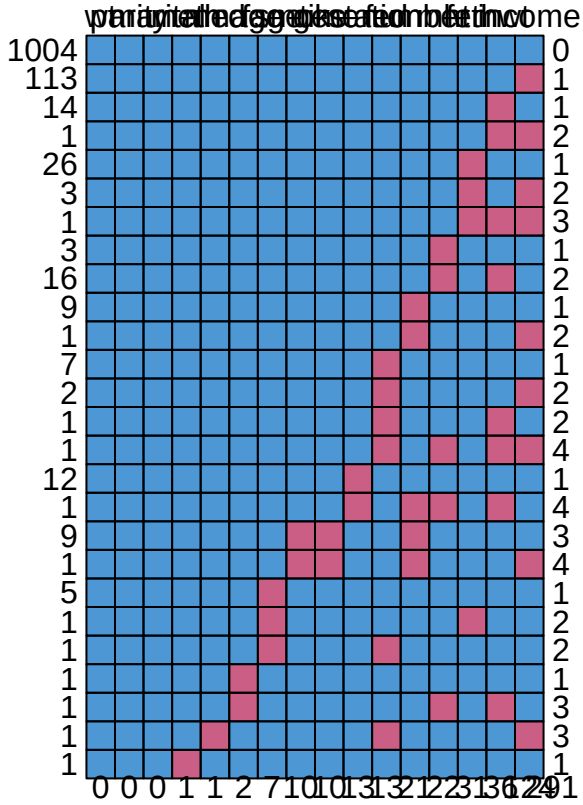
```
## Warning: package 'mice' was built under R version 3.4.4
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
md.pattern(fdata)
```

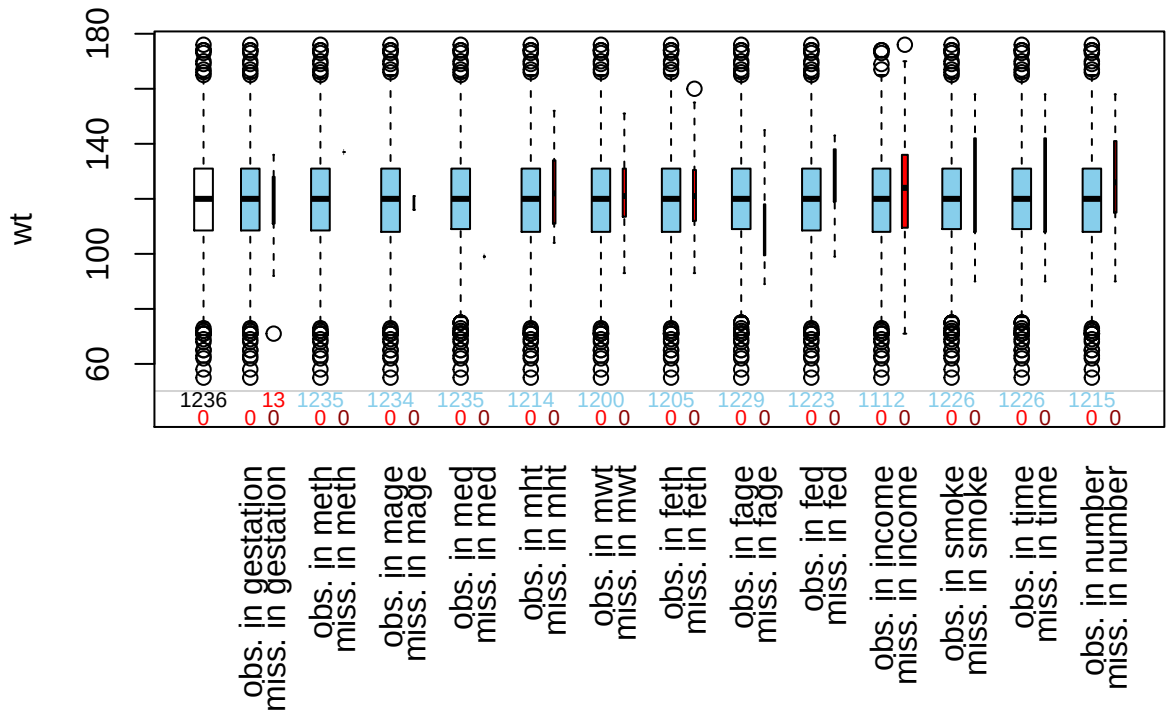



```
##      wt parity marital meth med mage fage smoke time gestation fed number
## 1004  1      1       1    1   1    1    1     1    1         1   1      1
## 113   1      1       1    1   1    1    1     1    1         1   1      1
## 14    1      1       1    1   1    1    1     1    1         1   1      1
## 1     1      1       1    1   1    1    1     1    1         1   1      1
## 26    1      1       1    1   1    1    1     1    1         1   1      1
## 3     1      1       1    1   1    1    1     1    1         1   1      1
## 1     1      1       1    1   1    1    1     1    1         1   1      1
## 3     1      1       1    1   1    1    1     1    1         1   1      1
## 16    1      1       1    1   1    1    1     1    1         1   1      1
## 9     1      1       1    1   1    1    1     1    1         1   1      0
## 1     1      1       1    1   1    1    1     1    1         1   1      0
## 7     1      1       1    1   1    1    1     1    1         1   0      1
## 2     1      1       1    1   1    1    1     1    1         1   0      1
## 1     1      1       1    1   1    1    1     1    1         1   0      1
## 1     1      1       1    1   1    1    1     1    1         1   0      1
## 12    1      1       1    1   1    1    1     1    1         0   1      1
## 1     1      1       1    1   1    1    1     1    1         0   1      0
## 9     1      1       1    1   1    1    1     0    0         1   1      0
## 1     1      1       1    1   1    1    1     0    0         1   1      0
## 5     1      1       1    1   1    1    0     1    1         1   1      1
## 1     1      1       1    1   1    1    0     1    1         1   1      1
## 1     1      1       1    1   1    1    0     1    1         1   0      1
## 1     1      1       1    1   1    0    1     1    1         1   1      1
## 1     1      1       1    1   1    0    1     1    1         1   1      1
## 1     1      1       1    1   0    1    1     1    1         1   0      1
```

```
## 1     1      1       1    0   1    1    1     1    1         1   1      1
##       0      0       0    1   1    2    7    10   10        13  13     21
##      mht feth mwt income
## 1004   1    1   1      1   0
## 113    1    1   1      0   1
## 14     1    1   0      1   1
## 1      1    1   0      0   2
## 26     1    0   1      1   1
## 3      1    0   1      0   2
## 1      1    0   0      0   3
## 3      0    1   1      1   1
## 16     0    1   0      1   2
## 9      1    1   1      1   1
## 1      1    1   1      0   2
## 7      1    1   1      1   1
## 2      1    1   1      0   2
## 1      1    1   0      1   2
## 1      0    1   0      0   4
## 12     1    1   1      1   1
## 1      0    1   0      1   4
## 9      1    1   1      1   3
## 1      1    1   1      0   4
## 5      1    1   1      1   1
## 1      1    0   1      1   2
## 1      1    1   1      1   2
## 1      1    1   1      1   1
## 1      0    1   0      1   3
## 1      1    1   1      0   3
## 1      1    1   1      1   1
##       22   31  36    124 291
```

From the pattern we can see that 1004 of the data points are complete, most values are missing for income

```
library(VIM)
pbox(fdata, pos = 1, int = FALSE, cex = 0.7)
```



We can see that the distributions for the missing data are not similar for all the covariates

```
#Imputting the data with "pmm" method as referenced from "https://stefvanbuuren.name/mice/"
imp <- mice(fdata, m = 5, maxit = 50, meth = 'pmm', seed = 500)
```

Checking the income imputation for all 5 datasets

```
imp$imp$income
```

This shows the imputation for each of the 5 iterations

```
head(complete(imp))
```

```
##    wt gestation parity meth mage med mht mwt feth fage fed marital income
## 1 120       284      1    8   27   5  62 100    8   31   5       1      1
## 2 113       282      2    0   33   5  64 135    0   38   5       1      4
## 3 128       279      1    0   28   2  64 115    5   32   1       1      2
## 4 123       275      2    0   36   5  69 190    3   43   4       1      8
## 5 108       282      1    0   23   5  67 125    0   24   5       1      1
## 6 136       286      4    0   25   2  62  93    3   28   2       1      4
##   smoke time number
## 1     0    0      0
## 2     0    0      0
## 3     1    1      1
## 4     3    5      5
```

```
## 5     1    1      5
## 6     2    2      2
```

Looking at the complete dataset #2

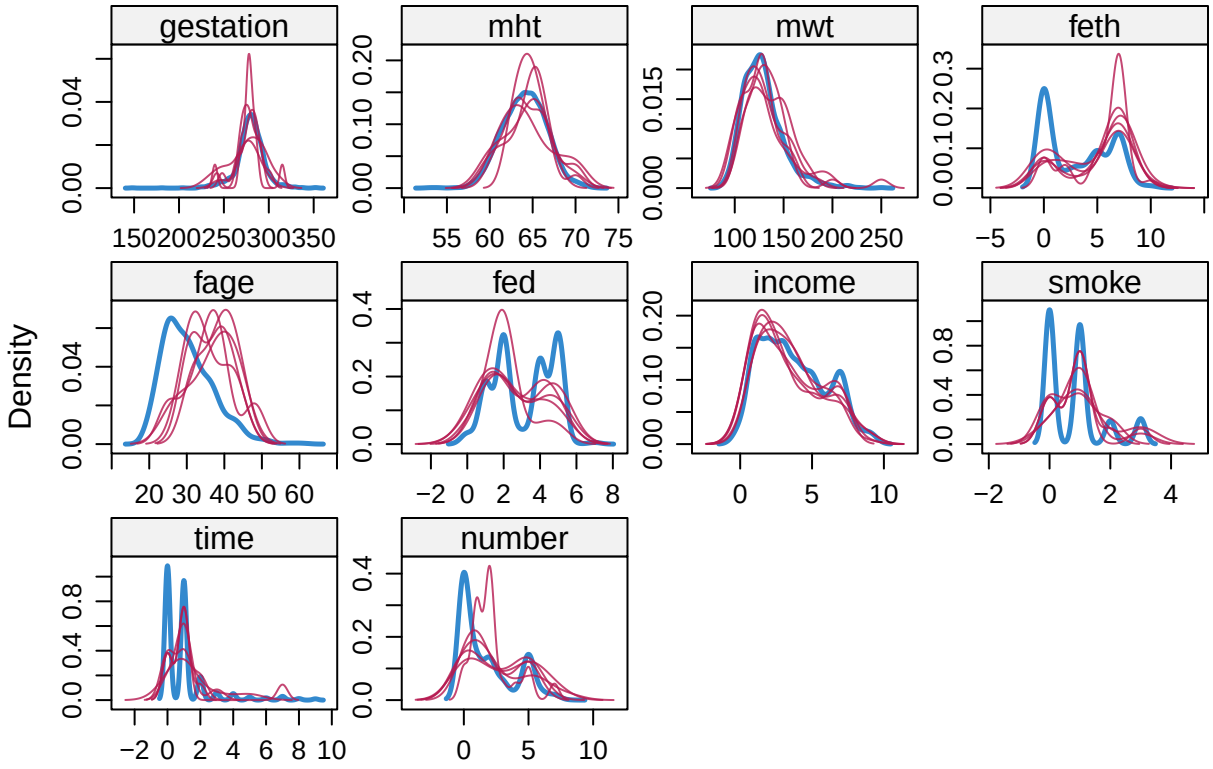
```
head(complete(imp,2))
```

```
##    wt gestation parity meth mage med mht mwt feth fage fed marital income
## 1 120       284      1    8   27   5  62 100    8   31   5       1      1
## 2 113       282      2    0   33   5  64 135    0   38   5       1      4
## 3 128       279      1    0   28   2  64 115    5   32   1       1      2
## 4 123       290      2    0   36   5  69 190    3   43   4       1      8
## 5 108       282      1    0   23   5  67 125    0   24   5       1      1
## 6 136       286      4    0   25   2  62  93    3   28   2       1      4
##   smoke time number
## 1     0    0      0
## 2     0    0      0
## 3     1    1      1
## 4     3    5      5
## 5     1    1      5
## 6     2    2      2
```

```
summary(imp)
```

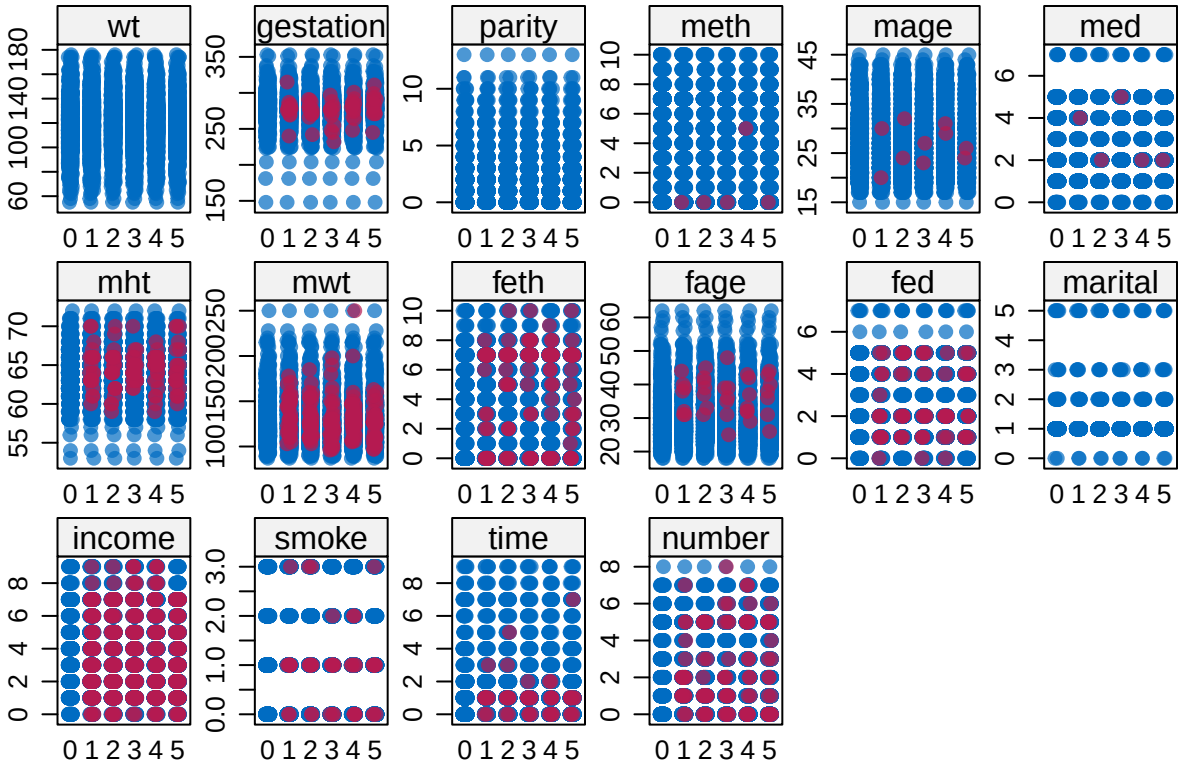
```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##        wt gestation    parity      meth      mage       med       mht
##        ""     "pmm"        ""     "pmm"     "pmm"     "pmm"     "pmm"
##       mwt      feth     fage       fed   marital    income     smoke
##     "pmm"     "pmm"     "pmm"     "pmm"        ""     "pmm"     "pmm"
##      time    number
##     "pmm"     "pmm"
## PredictorMatrix:
##           wt gestation parity meth mage med mht mwt feth fage fed marital
## wt         0         1      1    1    1   1   1   0    1    1   1       1
## gestation  1         0      1    1    1   1   1   1    1    1   1       1
## parity     1         1      0    1    1   1   1   1    1    1   1       1
## meth       1         1      1    0    1   1   1   1    1    1   1       1
## mage       1         1      1    1    0   1   1   1    1    1   1       1
## med        1         1      1    1    1   0   1   1    1    1   1       1
##           income smoke time number
## wt             1     1    1      1
## gestation      1     1    1      1
## parity         1     1    1      1
## meth           1     1    1      1
## mage           1     1    1      1
## med            1     1    1      1
```

```
densityplot(imp)
```



blue - observed obseravtions magenta - imputed values

```
stripplot(imp, pch = 20, cex = 1.2)
```



The imputed values for income fit very well.

Pooling

For the automatic forward selection model

```
modelFit1 <- with(data = imp, lm(wt ~ gestation + parity + meth + mage + income))
summary(modelFit1)
```

```
## # A tibble: 30 x 5
##    term         estimate std.error statistic  p.value
##    <chr>           <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)  -7.34      8.77      -0.837 4.03e- 1
##  2 gestation     0.452     0.0297    15.2   5.18e-48
##  3 parity        0.766     0.291      2.63  8.72e- 3
##  4 meth         -0.745     0.152     -4.91  1.03e- 6
##  5 mage          0.0824    0.101      0.815 4.15e- 1
##  6 income       -0.211     0.224     -0.943 3.46e- 1
##  7 (Intercept)  -8.47      8.78      -0.965 3.35e- 1
##  8 gestation     0.456     0.0298    15.3   1.24e-48
##  9 parity        0.772     0.291      2.65  8.09e- 3
## 10 meth         -0.735     0.152     -4.85  1.39e- 6
## # ... with 20 more rows
```

```
combine <- pool(modelFit1)
```

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

```
summary(combine)
```

```
##                estimate  std.error  statistic        df      p.value
## (Intercept) -8.09635135 8.78968688 -0.9211194 1204.4226 3.571712e-01
## gestation    0.45472097 0.02983967 15.2388075 1188.0471 0.000000e+00
## parity       0.76600626 0.29200164  2.6232944 1213.2300 8.817598e-03
## meth        -0.73080894 0.15189782 -4.8111878 1212.1511 1.688815e-06
## mage         0.07752513 0.10215645  0.7588863 1198.8733 4.480680e-01
## income      -0.17539411 0.22687227 -0.7730963  937.9803 4.396160e-01
```

All the pooled data sets

```
imp_1 <- data.frame(complete(imp,1))
imp_2 <- data.frame(complete(imp,2))
imp_3 <- data.frame(complete(imp,3))
imp_4 <- data.frame(complete(imp,4))
imp_5 <- data.frame(complete(imp,5))
```

Since the imputted values for income for all the imputed data sets follow a similar distribution, we choose imputed dataset #4, since it has the smallest difference in the means from the observed data

```
imp_list <- list(imp_1, imp_2, imp_3, imp_4, imp_5)
#al_dif <- for(ii in imp_list){
# dtset <- imp_list[ii]
# cm <- sapply(fdata, mean, na.rm = T) - sapply(dtset, mean)
# sm <- sum(cm)
#}

#cm
```

```
cm1 <- sapply(fdata, mean, na.rm = T) - sapply(imp_1, mean)
sum(cm1)
```

```
## [1] -0.2007378
```

```
cm2 <- sapply(fdata, mean, na.rm = T) - sapply(imp_2, mean)
sum(cm2)
```

```
## [1] -0.07209707
```

```
cm3 <- sapply(fdata, mean, na.rm = T) - sapply(imp_3, mean)
sum(cm3)
```

```
## [1] 0.02579937
```

```
cm <- sapply(fdata, mean, na.rm = T) - sapply(imp_5,mean)
sum(cm)
```

```
## [1] 0.02579937
```

```r
cm <- sapply(fdata, mean, na.rm = T) - sapply(imp_4,mean)
sum(cm)
```

```
## [1] -0.2169191
```

```r
#Get sample 4 income values

imputed_income <- imp_4$income
fdata$income <- imputed_income
```