MRD_Assignment4

Sebastián Soriano Pérez

11/11/2019

Question 1: MISSING DATA MECHANICS

- Create a dataset with 30% of the age values missing completely at random, leaving all values of diameter observed. Report the R commands you used to make the dataset. Also report the dataset values after you made the ages missing. (This is so we can tell which cases you made missing.)

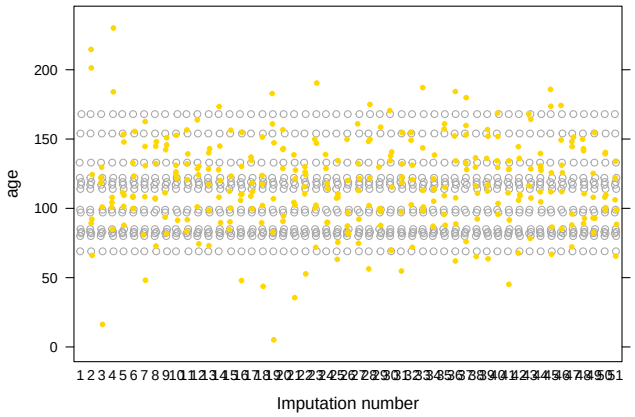
```
#Loading dataset and creating random missing values for age
trees <- read.csv('treeage.txt')
set.seed(1)
trees[sample(nrow(trees), 6),]$age <- NA
kable(trees, row.names = TRUE,
      label = "tables", format = "latex", booktabs = TRUE) %>% kable_styling(
        latex_options = "HOLD_position"
        )
```

	number	diameter	age
1	2	11.4	NA
2	2	11.4	119
3	3	7.9	83
4	4	9.0	85
5	13	10.3	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	4	9.0	NA
12	12	8.0	80
13	13	10.3	114
14	7	7.3	NA
15	15	9.2	122
16	19	9.3	NA
17	17	7.0	82
18	1	12.0	NA
19	19	9.3	97
20	20	8.2	99

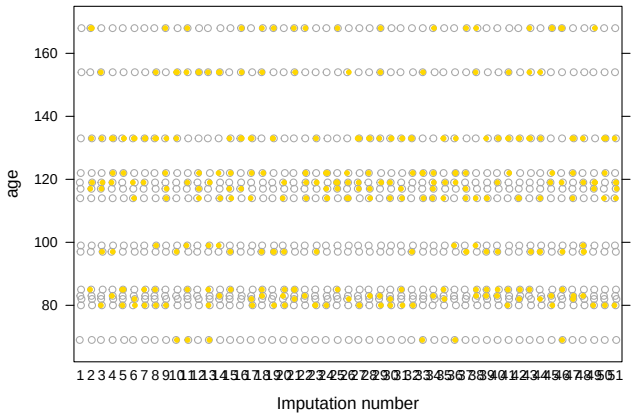
- Use a multiple imputation approach to fill in missing ages with the R software mice using a default application, i.e., no transformations in the imputation models. Create m = 50 imputed datasets. Use multiple imputation diagnostics to check the quality of the imputations of age, looking at both the marginal distribution of age and the scatter plot of age versus diameter. Run the diagnostics on at least two of the completed datasets. Turn in the graphical displays you made (showing results for at least two completed datasets) and your conclusions about the quality of the imputation model.

```
#Adding imputations with 'norm' and 'pmm'
trees_nrm <- mice(trees, m = 50, defaultMethod = c('norm'), print = F)
```

```
trees_pmm <- mice(trees, m = 50, defaultMethod = c('pmm'), print = F)
par(mfcol = c(1, 2))
stripplot(trees_nrm, age ~ .imp, col = c('darkgray', 'gold'), pch = c(1, 20))
```



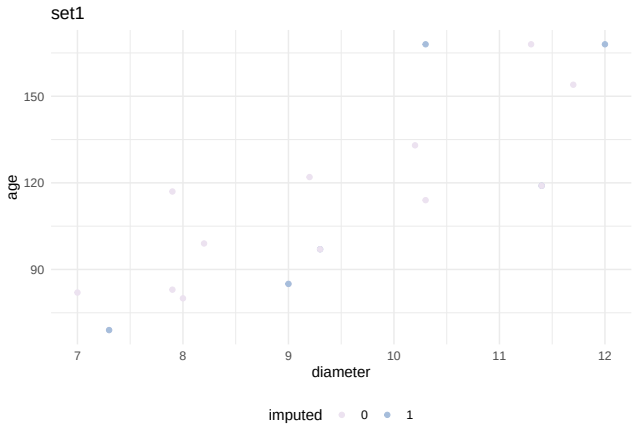

```
stripplot(trees_pmm, age ~ .imp, col = c('darkgray', 'gold'), pch = c(1, 20))
```



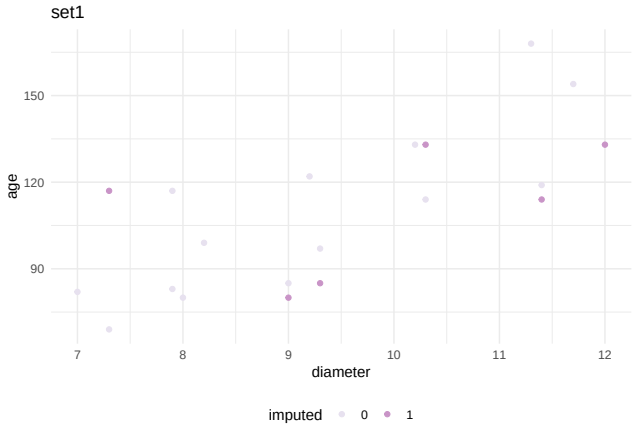
The pmm method produced all imputed values within the range of values for the actual remaining data. The norm method includes values that would seem to be outliers among the rest of the data. For this reason I will choose the pmm method.

```
#Plotting two random datasets for 'pmm'
#Adding imputed column to identify rows with imputed values
#set.seed(1233)
s1 <- as.integer(runif(1, 1, 50))
s2 <- as.integer(runif(1, 1, 50))

set1 <- complete(trees_pmm, s1)
set1$imputed <- 0
set1[is.na(trees$age),]$imputed <- 1
set1$imputed <- as.factor(set1$imputed)

set2 <- complete(trees_pmm, s2)
set2$imputed <- 0
set2[is.na(trees$age),]$imputed <- 1
set2$imputed <- as.factor(set2$imputed)

#Plotting both datasets
```

```
par(mfcol = c(1, 2))
ggplot(set1, aes(x = diameter, y = age, color = imputed)) + geom_point() +
  theme_minimal() + scale_color_brewer(palette = 10) +
  theme(legend.position = "bottom") + ggtitle('set1')
```




```
ggplot(set2, aes(x = diameter, y = age, color = imputed)) + geom_point() +
  theme_minimal() + scale_color_brewer(palette = 11) +
  theme(legend.position = "bottom") + ggtitle('set1')
```



The scatterplots for two random sets of imputations (set1 & set2) show that the pmm method seems to produce values that match the linear relationship present for the rest of the data between age and diameter

- Estimate a regression of age on diameter. Apply the multiple imputation combining rules to obtain point and variance estimates for the regression parameters that account for missing data. What can you conclude about the relationship between age and diameter?

```
lms_trees <- with(trees_pmm, lm(age ~ diameter))
lm_trees = pool(lms_trees)
kable(summary(lm_trees), row.names = TRUE,
      label = "tables", format = "latex", booktabs = TRUE) %>% kable_styling(
        latex_options = "HOLD_position"
        )
```

	estimate	std.error	statistic	df	p.value
(Intercept)	-18.90856	28.429361	-0.6651068	12.01758	0.5185430
diameter	13.80517	3.043758	4.5355662	11.51117	0.0007582

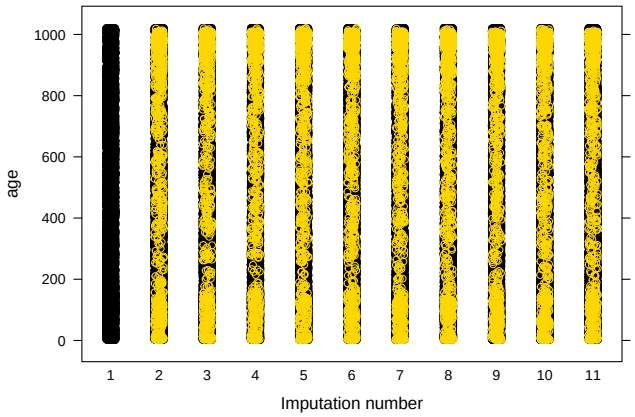
Although the model concludes that the intercept cannot be rejected to be zero, there exists significant correlation between diameter and age. For every unit increase in a tree's diameter, it could be expected that the tree would be 9.12 units older.

Question 2: MULTIPLE IMPUTATION IN NHANES DATA

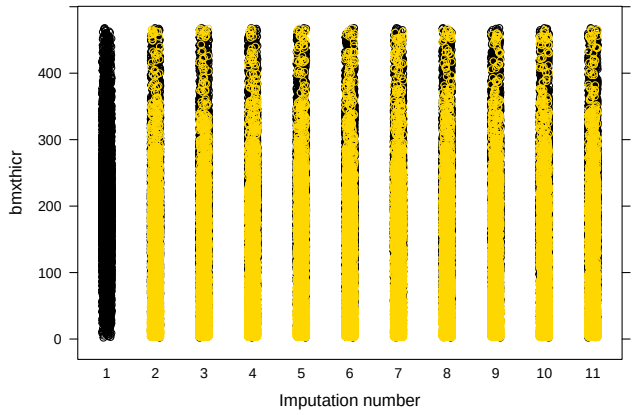
- Use a multiple imputation approach to fill in missing values with the R software mice using a default application (no transformations in the modeling).
  - Create m = 10 imputed datasets.
  - Use multiple imputation diagnostics to check the quality of the imputations, looking at both marginal distributions and scatter plots. Run the diagnostics on at least two of the completed datasets. Turn in plots for bmxbmi (BMI measurement) by age and bmxbmi by riagendr (gender).
  - What are your conclusions about the quality of the imputation model?

```
#Loading dataset and deleting unnecessary columns
nhanes = read.csv('nhanes.csv')
nhanes$wtmec2yr <- NULL
nhanes$sdmvstra <- NULL
nhanes$sdmvpsu <- NULL
nhanes <- replace_with_na_all(data = nhanes, condition = ~.x == '.')

nhanes_pmm <- mice(nhanes, m = 10, defaultMethod = c("pmm"), print = F)
```

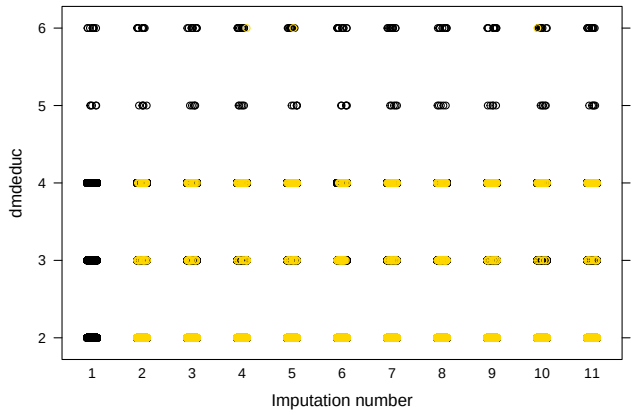
```
#Plotting imputation sets for age, bmxthicr, dmdeduc, and bmxbmi
stripplot(nhanes_pmm, age ~ .imp, col = c('black', 'gold'), pch = c(1, 1))
```



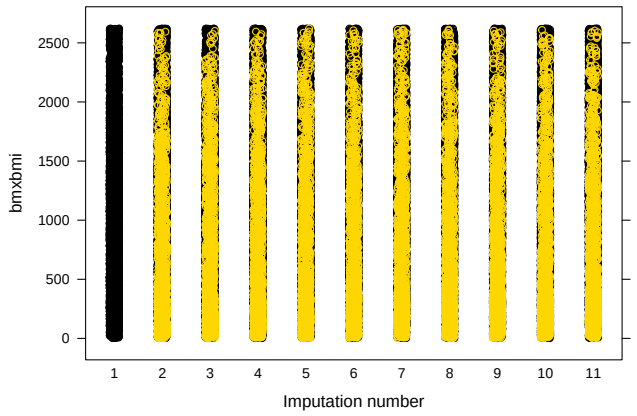

```
stripplot(nhanes_pmm, bmxthicr ~ .imp, col = c('black', 'gold'), pch = c(1, 1))
```



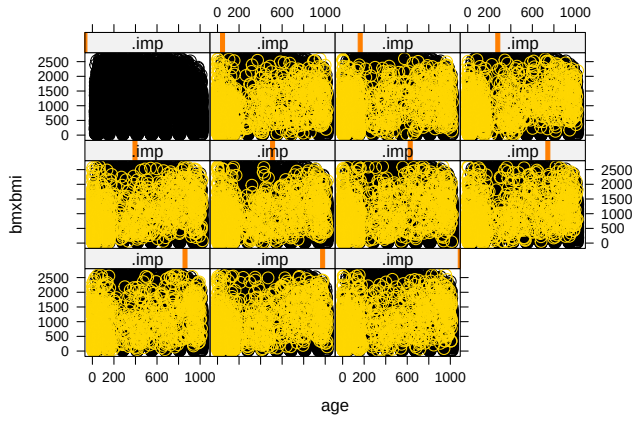
```
stripplot(nhanes_pmm, dmdeduc ~ .imp, col = c('black', 'gold'), pch = c(1, 1))
```



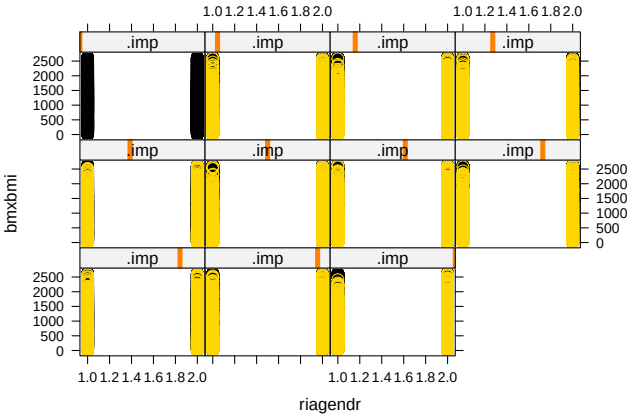

```
stripplot(nhanes_pmm, bmxbmi ~ .imp, col = c('black', 'gold'), pch = c(1, 1))
```




```
#Plotting bmxbmi by age and bmxbmi by riagendr
xyplot(nhanes_pmm, bmxbmi ~ age | .imp, pch = c(1, 1), cex = 1.4, col = c('black', 'gold'))
```




```
xyplot(nhanes_pmm, bmxbmi ~ riagendr | .imp, pch = c(1, 1), cex = 1.4, col = c('black', 'gold'))
```



The imputation model used was 'pmm'. It seems to do a very good job as most of the values are within the range of the existing data points. I will stick with this model. I tried the 'norm' model but there were too manby outliers imputed.

- Run a model that predicts BMI from some subset of age, gender, race, education, and income. Apply the multiple imputation combining rules to obtain point and variance estimates for the regression parameters that account for missing data. Interpret the results of your final model.

```
nhanes_6 = complete(nhanes_pmm, 6)
null_model = lm(bmxbmi ~ 1, data = nhanes_6)
full_model = lm(bmxbmi ~ age + ridageyr + riagendr + ridreth2 + dmdeduc + indfminc + age:dmdeduc +
                  riagendr:ridreth2, data = nhanes_6)
lm_nhanes_6 <- step(null_model, scope = formula(full_model), direction = 'both', trace = 0)
summary(lm_nhanes_6)
```

```
##
## Call:
## lm(formula = bmxbmi ~ ridageyr + age + dmdeduc + riagendr + ridreth2 +
##     indfminc + age:dmdeduc + riagendr:ridreth2, data = nhanes_6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1460.3  -359.2  -102.1   296.0  1933.6
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         19.59830   46.48747   0.422  0.67334
## ridageyr            19.63508    0.33774  58.136  < 2e-16 ***
## age                  0.22677    0.05441   4.168  3.1e-05 ***
## dmdeduc            267.14140   11.43507  23.362  < 2e-16 ***
## riagendr             2.78385   21.96323   0.127  0.89914
## ridreth2           -15.43823   15.51468  -0.995  0.31972
## indfminc             4.43507    1.35401   3.275  0.00106 **
## age:dmdeduc         -0.39345    0.02190 -17.968  < 2e-16 ***
## riagendr:ridreth2   21.93924    9.71396   2.259  0.02393 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 511.7 on 10113 degrees of freedom
## Multiple R-squared:  0.3701, Adjusted R-squared:  0.3696
## F-statistic: 742.8 on 8 and 10113 DF,  p-value: < 2.2e-16
```

```
lm_nhanes <- with(nhanes_pmm, lm(bmxbmi ~ ridageyr + age + dmdeduc + riagendr + ridreth2 +
                                   indfminc + age:dmdeduc + riagendr:ridreth2))
summary(pool(lm_nhanes))
```

estimate	std.error	statistic	df	p.value
-2.77	53.3	-0.052	156	0.959
19.6	0.362	54.1	509	0
0.246	0.0658	3.74	93	0.000314
270	14.2	19.1	77.9	0
10	22.7	0.442	1.74e+03	0.659
-11	16.3	-0.677	1.01e+03	0.499
4.13	1.43	2.88	706	0.00406
-0.399	0.028	-14.3	62.3	0
19.9	10.2	1.95	930	0.0515

All coefficients are significant in the final pooled model. dmdeduc is the largest coefficient, for which every unit increase represents a 262.3 increase in bmxbmi. On the other hand, every unit increase in ridreth2 represents an 11.1 decrease in bmxbmi.