Missing Data Imputation

Derek Wales

09NOV19

Loading the homework files.

Part One A: 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.)

Part One A: Done by taking a sample of 6 and then using match to find the corresponding location. After that, I replaced the appropriate values with NA (in R code below) (Varun Pasad, MIDS 2021).

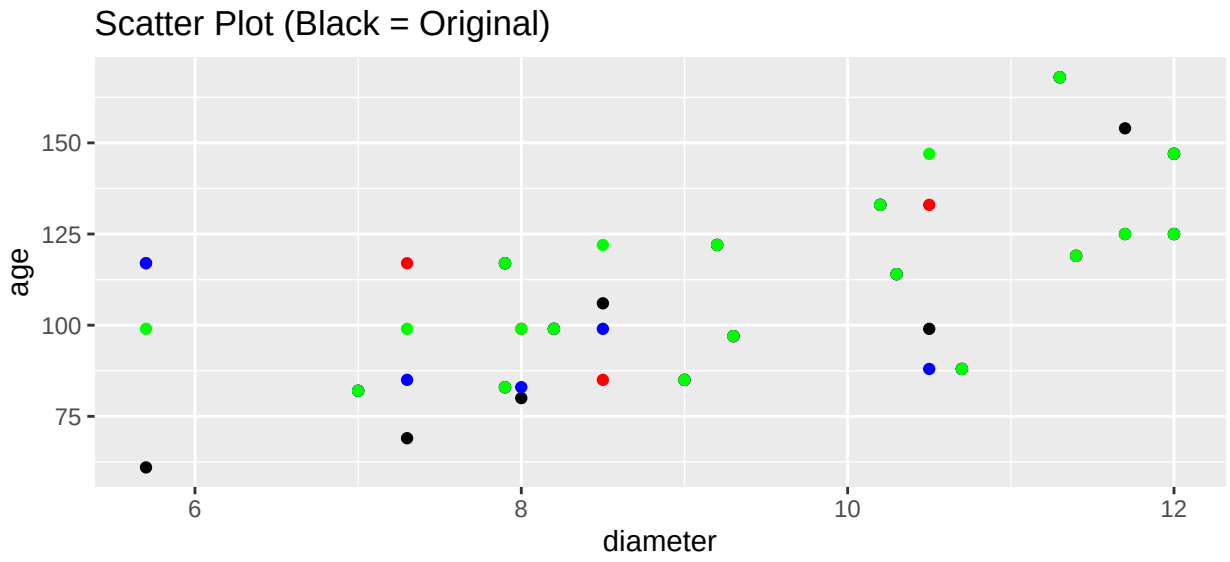
```
# EDA Revealed need for a log transform of the Data

#Part One A: Removing 30% of the age values (Varun Pasad, MIDS 2021).
rand_age <- sample(tree_df$age, 6)
rand_id <- match(rand_age,tree_df$age)

# Saving the dataframe as a random and 30% of values are replaced
tree_df_random <- tree_df
tree_df_random$age[rand_id] <- NA
```

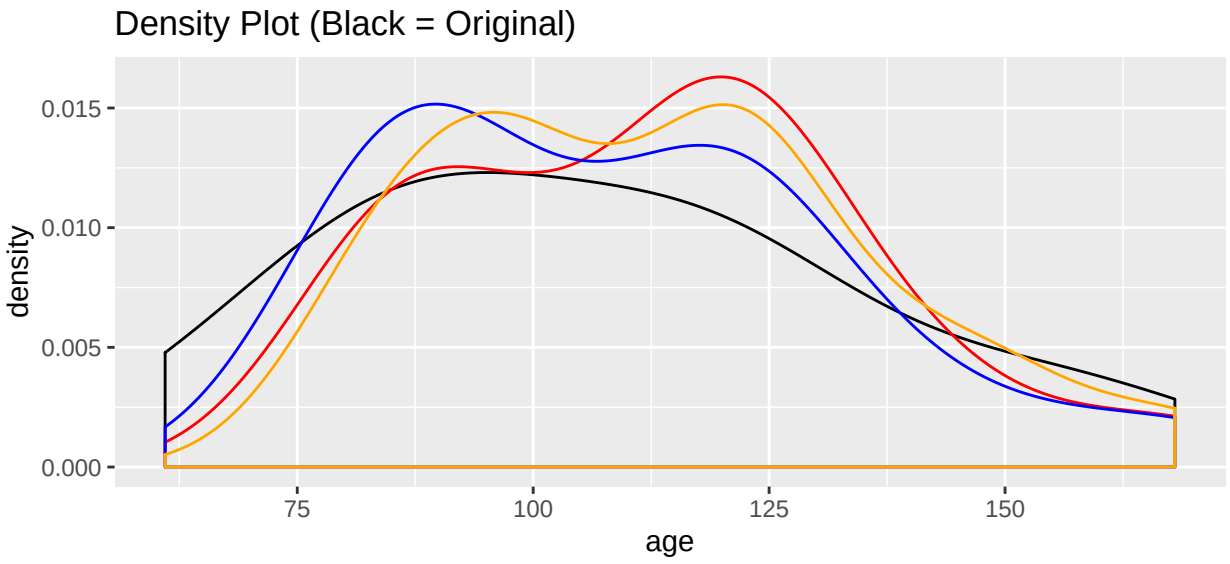
Part One B: Use a multiple imputation approach to fill in missing ages with the R software mice using a default application, i.e., no transformations in the imputation models. Create m = 50 imputed datasets. Use multiple imputation diagnostics to check the quality of the imputations of age, looking at 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.

This was done using a combination of a stripper, xy and density plots with the imputed datasets. Overlaying those with the observed values reveled that our imputed values were close to the actual values. Especially considering the limited information available.

Part One C: 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.

After completing three separate linear models with the imputed values, model validation revealed that the first model preformed the best in model validation, even though it suffered from having a small sample size.





Part One D: 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?

This was done using the 'pool' command in R. Which got a summary of all the imputed values. After which I built a linear determining the relationship between age and diameter.

This shows that a base tree age "starts" at the intercept listed below and incrementally increases by one year each time the diameter increases by the coefficient below (note specific values left out because they will be aggregated by the pool function and are slightly different each time).

```
##               estimate std.error statistic       df    p.value
## (Intercept) 32.706990 28.161547  1.161406 11.78554 0.26845930
## diameter     8.336631  2.924401  2.850713 11.93830 0.01466919
```

Part Two A: Use a multiple imputation approach to fill in missing values with the R software mice using a default application.

To do this I first had to replace the missing values across the dataframe with N/As (currently labeled with '.'). Additionally, many of variables were numeric values but factor variables (like race). All of which had to be correctly categorized before I could impute the missing values. Once that was complete, I could impute the required 10 datasets and determine the pattern of missing data.

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).

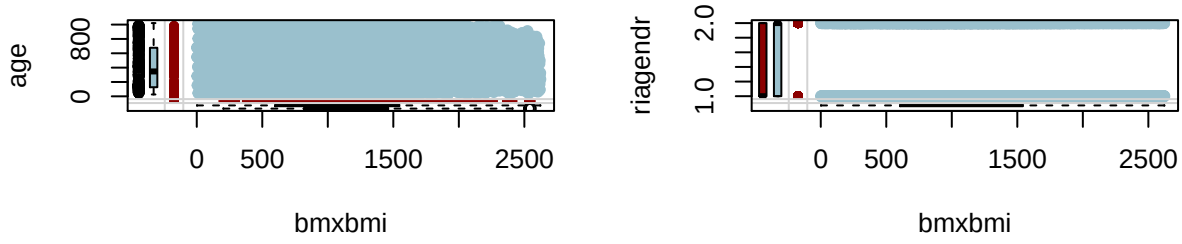
What are your conclusions about the quality of the impuation model?

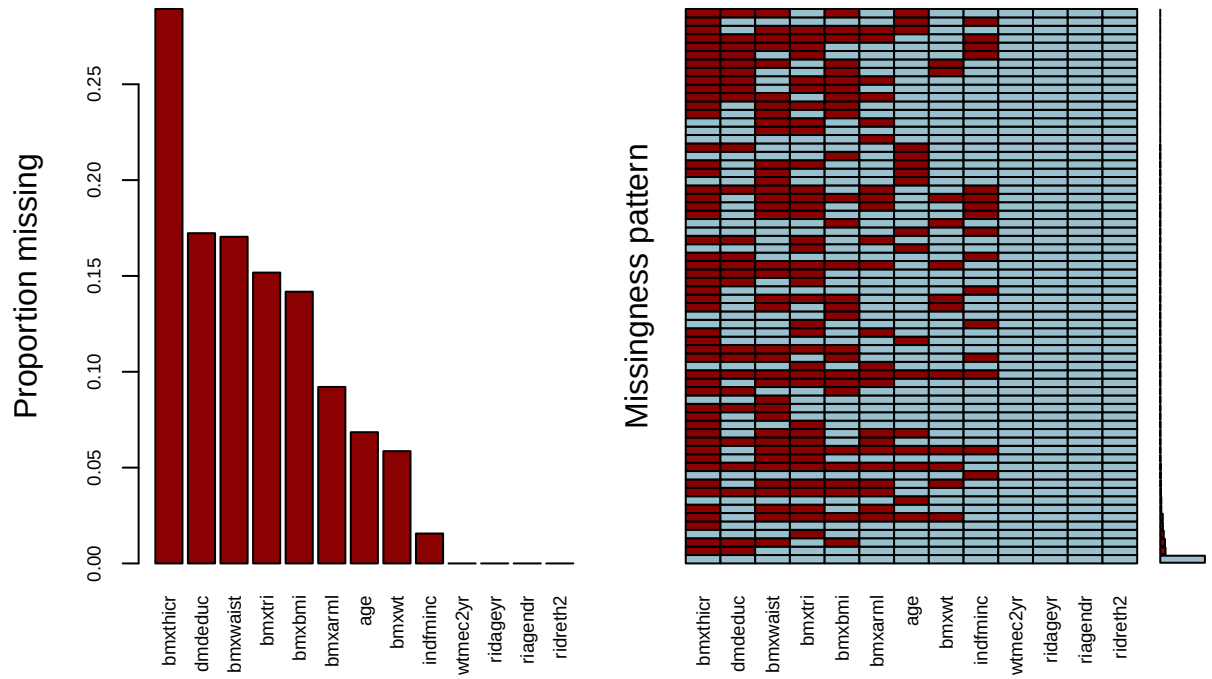
Looking at the imputed data, some categories are missing a much higher percentage of data than others with bmxthicr missing in almost 30% of the cases, while indfminc is missing less than 2%.

These were important factors to keep in mind prior to conducting analysis.

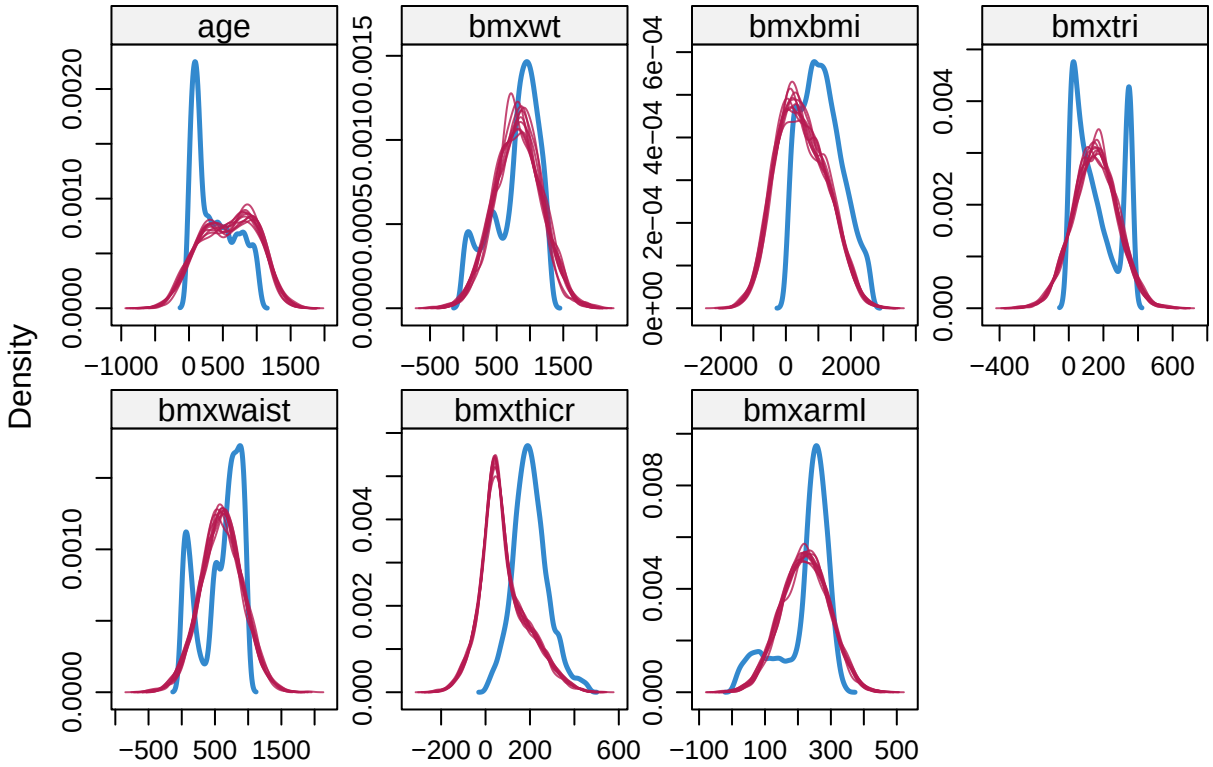
This is best represented by the density plot. Which would ideally show a perfect match between the imputed values and the observed values. This will result in less predictive accuracy and have higher variance.

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






```
##
##  Variables sorted by number of missings:
##  Variable      Count
##  bmxthicr 0.28927090
##   dmdeduc 0.17229796
##  bmxwaist 0.17042087
##    bmxtri 0.15174867
##    bmxbmi 0.14177040
##   bmxarml 0.09207666
##       age 0.06846473
##     bmxwt 0.05858526
##  indfminc 0.01560956
##  wtmec2yr 0.00000000
##  ridageyr 0.00000000
##  riagendr 0.00000000
##  ridreth2 0.00000000
```



Run a model that predicts BMI from some subset of age, gender, race, education, marital status, 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.

Looking at BMI across age, gender, race, education, marital status, and income we we can observe the following trends. Age does increase your BMI (aka the older you get, the more likely you are to gain weight). Additionally, men are on average heavier than females. I also noticed that BMI changes across race with blacks generally having lower BMI and Mexicans and Hispanics being heavier. Education and income are also significant in predicting an individual's BMI. With a lower income and education levels tending to be heavier.
