Methods and Data Analysis 4

Srishti Saha (ss1078)

11 November, 2019

Part 1

Creating missing values in the dataset

```
set.seed(6)
rand_sampl= sample(treeage_df$age,6)

treeage_df$age<-ifelse(treeage_df$age %in% rand_sampl,NA,treeage_df$age)
```

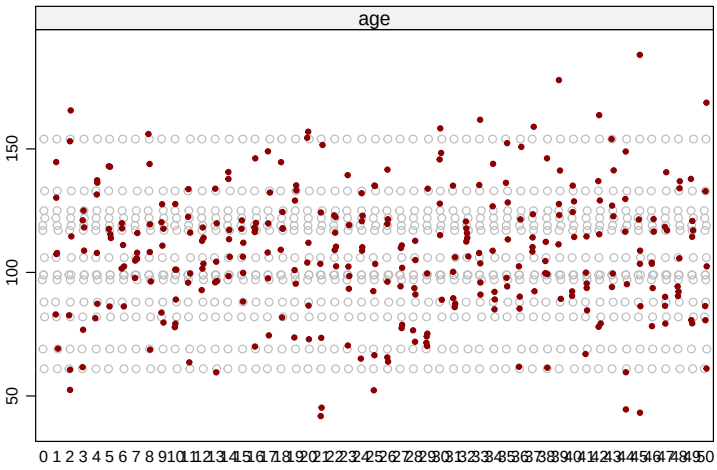
Using the above snippet, we get the dataset below:

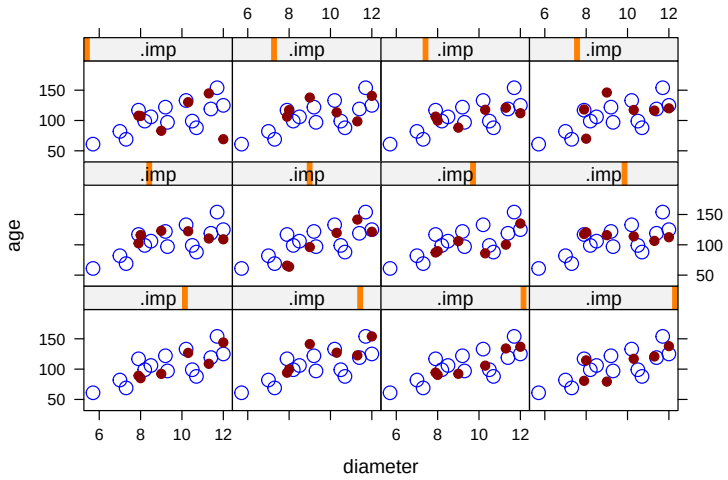
number	diameter	age
1	12.0	125
2	11.4	119
3	**7.9**	**NA**
4	**9.0**	**NA**
5	10.5	99
6	7.9	117
7	7.3	69
8	10.2	133
9	11.7	154
10	**11.3**	**NA**
11	5.7	61
12	**8.0**	**NA**
13	**10.3**	**NA**
14	**12.0**	**NA**
15	9.2	122
16	8.5	106
17	7.0	82
18	10.7	88
19	9.3	97
20	8.2	99

Imputation using mice

Let us now try to run imputation on m=50 datasets with method= 'norm' for continuous variables (setting seed=50). Let us then look at the plot imputed and observed values for the variable 'age'.

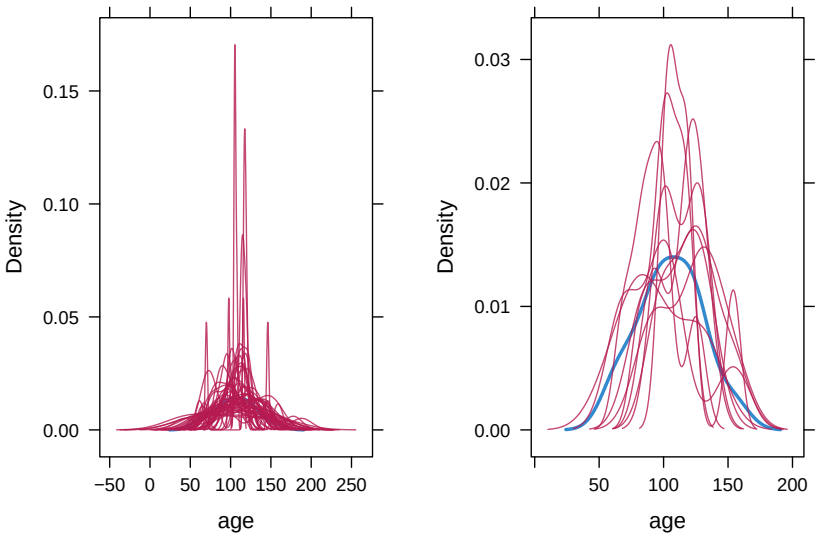
Evaluating imputations





We see that the shape and distribution of the grey and the dark red dots are similar in many imputed datasets. The matching shapes says that a lot of the imputations are rather plausible. Let us look at the xyplot of these imputations.

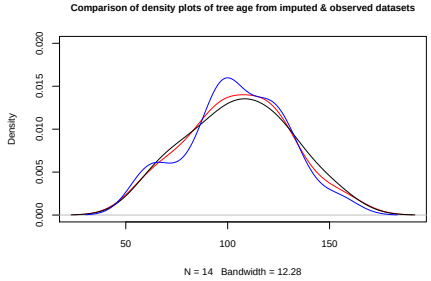
The scatter plots show age as a function of diameter, from the dataframe with missing values with the 10 imputed dataframes. Across all datasets, there seems to be a positive trend and a similar distribution with respect to diameter. Let us compare the overall results of imputation using norm versus PMM to check their general performance.



We will go ahead with norm because PMM datasets seems to have a lot of distortions in the imputed datasets.

Deep dive into two imputed datasets

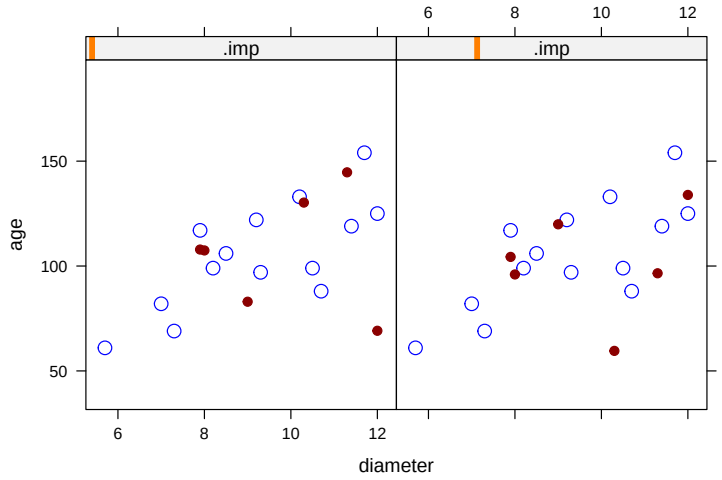
Let us now select two datasets from the imputed ones. By simple eyeballing, we have selected 1 and 13. In the below plot, **red line- observed values**, **black-dataset1** and **blue-dataset13**



The distribution of the observed data is almost perfectly normal. The other imputed datasets seem to have a similar distribution with some distortion to the kurtosis.

In the norm imputations, selecting 1 and 13 as the imputed dataset for diagnostics as they seem to preserve the distribution of the observed data the best. Let us now look at the dataset themselves and the scatter plots.

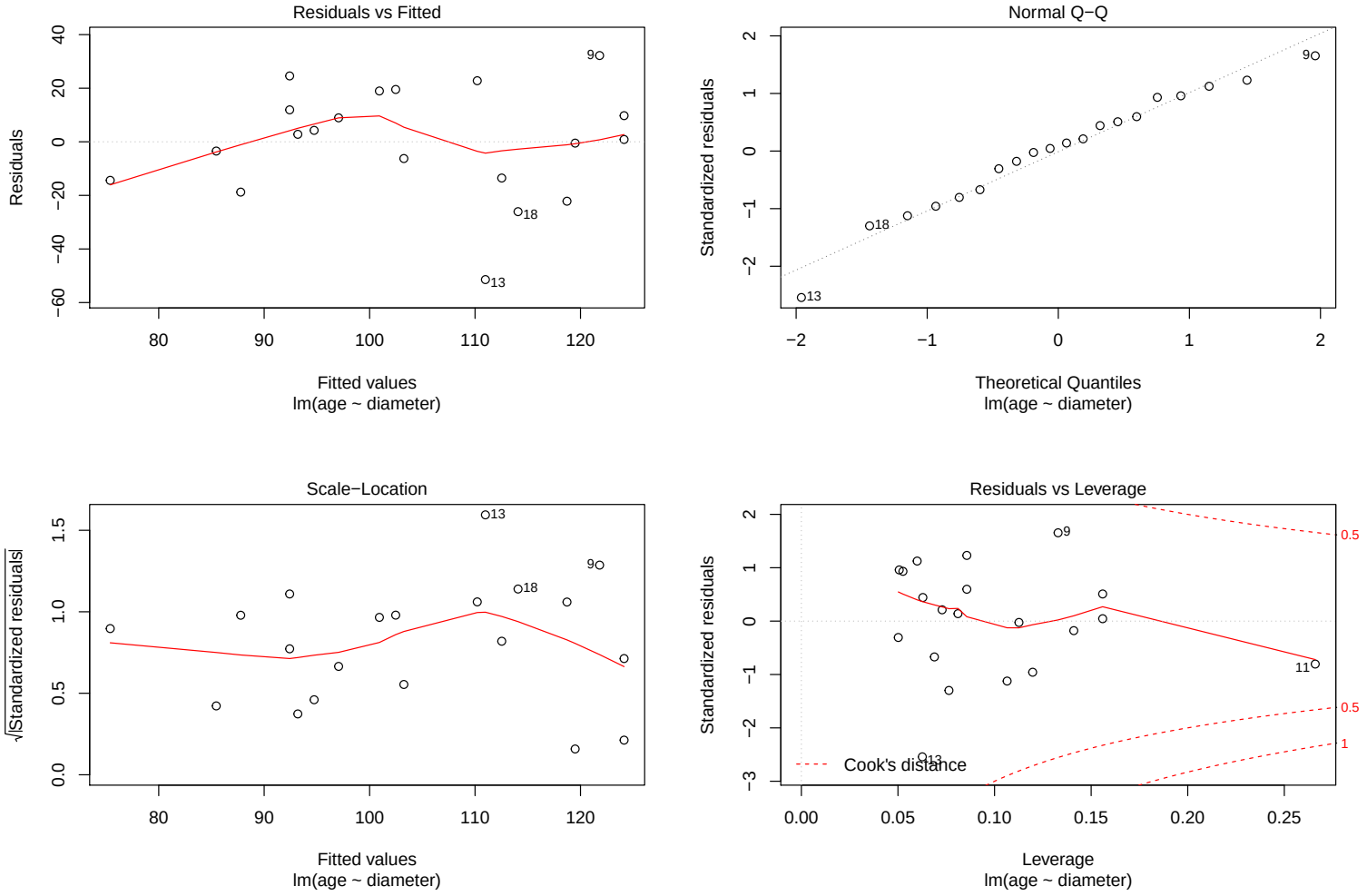
number	diameter	age_from_d13	age_from_d1	original_age
1	12.0	125.00000	125.00000	125
2	11.4	119.00000	119.00000	119
3	7.9	104.34615	107.85382	83
4	9.0	119.88745	82.98746	85
5	10.5	99.00000	99.00000	99
6	7.9	117.00000	117.00000	117
7	7.3	69.00000	69.00000	69
8	10.2	133.00000	133.00000	133
9	11.7	154.00000	154.00000	154
10	11.3	96.54050	144.66684	168
11	5.7	61.00000	61.00000	61
12	8.0	95.98321	107.42736	80
13	10.3	59.55069	130.24567	114
14	12.0	133.89089	69.14136	147
15	9.2	122.00000	122.00000	122
16	8.5	106.00000	106.00000	106
17	7.0	82.00000	82.00000	82
18	10.7	88.00000	88.00000	88
19	9.3	97.00000	97.00000	97
20	8.2	99.00000	99.00000	99



We see that the distributions of both the observed and the imputed age are following a similar upward trend (showing that age and distribution have a positive correlation.). Let us now try to fit a model with the above datasets.

Regression model

Let us consider dataset1 to fit the model.



From the above residual plots and summary plots of the model we can make the following observations:

- Linearity: Based on the residual plot, there is no evident pattern in the residual plot which indicates that the model will capture the pattern and the information within Y and X. Thus, the assumption of Linearity is met.

- Independence and Equal Variance: The residual versus fitted plot does not show an evident conical (spread out or converging) pattern. This might indicate that there is no heteroskedasticity in the dataset or the model.
- Normality: The Q-Q plot can be used to interpret Normality. If the standard residuals are linearity related to the theoretical quantiles (no curve), then the condition of Normality is met. Here, normality is met.

There are some limitations to the model. The only predictor for 'age' is 'diameter' which explains only about ~30% (R-squared from the 13th dataset is 28.9%) of the variation. Moreover, the data has been built on only 20 data points which might reduce the reliability of the model for inference.

term	estimate	std.error	statistic	p.value
(Intercept)	31.308301	25.085937	1.248042	0.2280073
diameter	7.735422	2.620643	2.951726	0.0085347

Thus, our inference of the relationship between age and diameter on the basis of this model can be considered legit. According to the model summary on the 13th imputed dataset, the intercept is 31.31 which implies that the baseline age is ~31 years (with diameter 0). However, this does not make intuitive sense. The coefficient of diameter is 7.74. This means that just on the basis of the diameter, for every unit increase in the diameter will lead to an increase of 7.74 units (years) in age. This model also confirms the positive relation between age and diameter. The standard error is 2.62.

Let us also at the multiple imputation inferences by combining the rule.

```
##             estimate std.error statistic    df p.value
## (Intercept)    17.10     26.80      0.64 11.37    0.54
## diameter        9.45      2.81      3.37 11.31    0.01
```

```
##           est      lo 95     hi 95 fmi
## R^2 0.4806991 0.08351167 0.7880986 NaN
```

The coefficient of diameter from the pooled results (9.45) are close to the results obtained from the 13th dataset. The intercept (baseline) is much lower than the model results of the 13th imputed dataset. This might mean that the imputation might not be completely reflective of the observed population and **there is a lot of variation across the imputed datasets.** The model results above show that **diameter is a significant variable for age (low p-value)**. The model has an R-squared value of 48.07%. This means that ~**48% of the variance in age is explained by diameter**.

Part 2

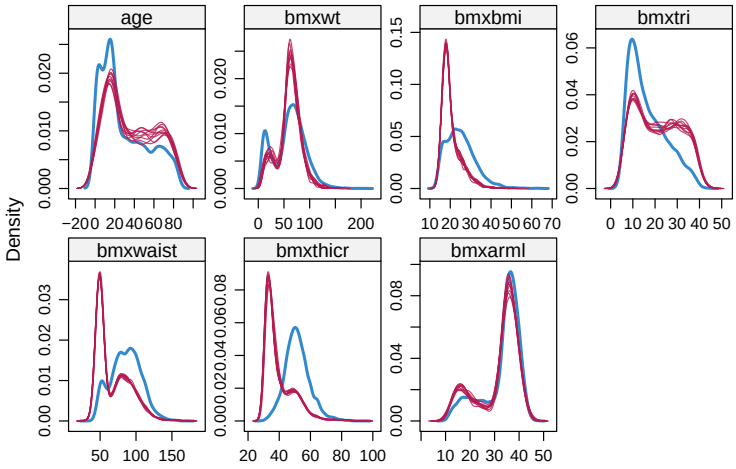
Data import and evaluation

There are NAs in age, dmdeduc, indfminc, bmxwt, bmxbmi, bmxtri, bmxwaist, bmxthicr, and bmxarml. There are variables like riagendr which are factor variables with 2 levels. There are also other factor variables with more than 2 levels.

Among the numerical variables, bmi is highly correlated (>0.5) to bmxwaist, bmxwt, bmxthicr and bmxtri. There is a moderate correlation (between 0.3 and 0.5) between bmxbmi and age and bmxbmi and bmxarml.

Imputation using mice

We will use pmm, logreg and polyreg for imputation in this case.



The imputations seem plausible in the way that there are no negative imputations and that for the categorical variable 'dmdeduc', the levels seem to have been preserved.

We see that the distributions of a few variables like:

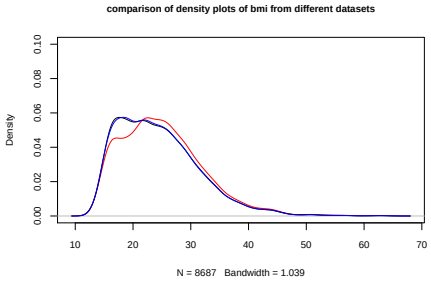
- bmxarml
- age

are fairly similar for the observed and the imputed data. For 'bmxwt', although the shape is similar, the density plot differs in kurtosis. For all other variables including 'bmxbmi', there is a significant distortion in the imputed values versus the observed values.

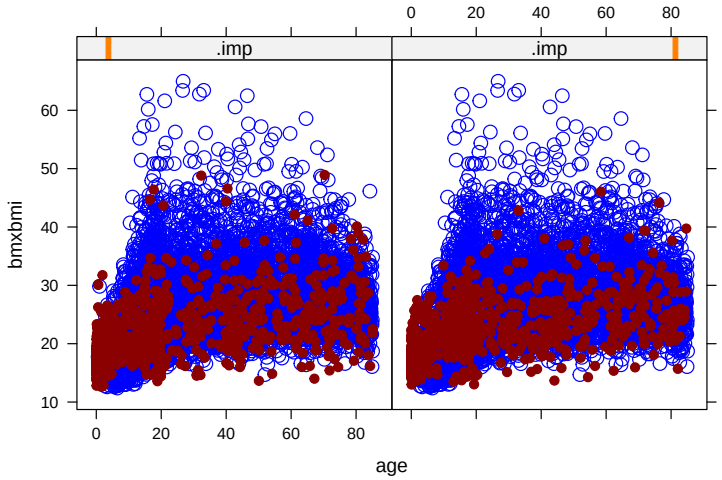
Diagnostics on 2 datasets

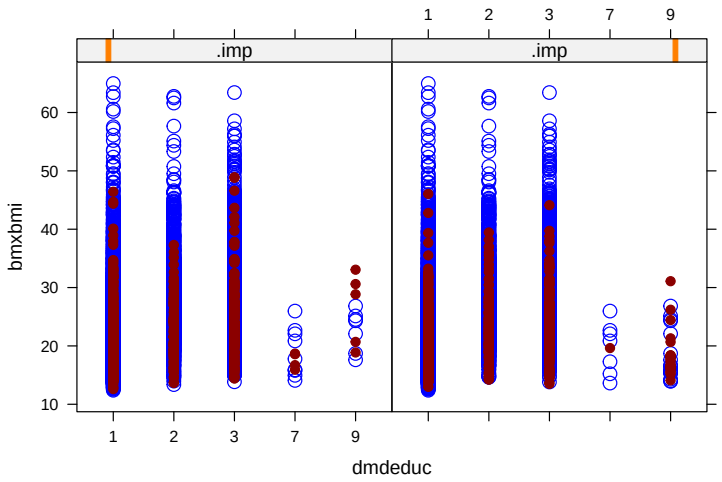
Let us select the 1st and 9th imputed dataset.

Let us compare the density plots of these datasets for bmxbmi.



Looks like the imputed dataset have the same distribution for bmi (in blue and black). This is heavily right skewed as compared to the more normal distribution of the observed cases of bmxbmi in the original dataset. Let us look at the scatterplots for the variables 'bmxbmi' by 'age' and 'riagendr'.





From the scatterplots of bmxbmi as a function of age, the distribution of the imputed values (red) and the observed values (blue) seem to be similar except a few outliers. The scatter plot of bmxbmi versus education (dmdeduc) reveals similar results.

Regression Model for bmxbmi

Since there is a very evident right-skew in the bmi variable for the imputed values, we should try transformations. With log transformation, the distribution seems to be more normal. It is important to note that even age follows a right skewed distribution and might violate the assumption of normality in a model. However, log transformation of the variable age yields worse results and will also distort the interpretation of the model. Hence, we will keep age as is.

We started with building a model with all main effects of the variables: lm(formula = log(bmxbmi) ~ age + riagendr + ridreth2 + indfminc + dmdeduc)

Education is significant; Age and gender are too. Ridreth3 (race) seems to be significant for all levels except 3. Since income does not seem to be significant for most levels, we can remove them. On doing a stepwise selection with BIC as the selection criterion, we see that income is not significant. So, the final model will have main effects of all variables but income.

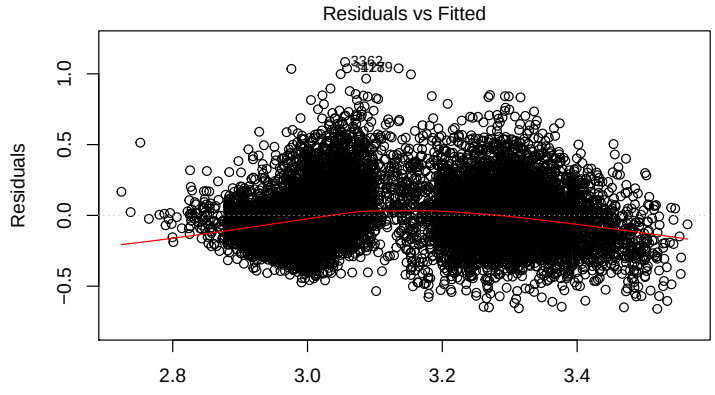
We also did a preliminary analysis on the interactions effects of race and education with age. There was no clear differentiating pattern in the EDA. We checked for variable significance by including these effects in the model. We added an interaction term between age and education as that was significant across the different levels of education (had low p-values). The final model is now:

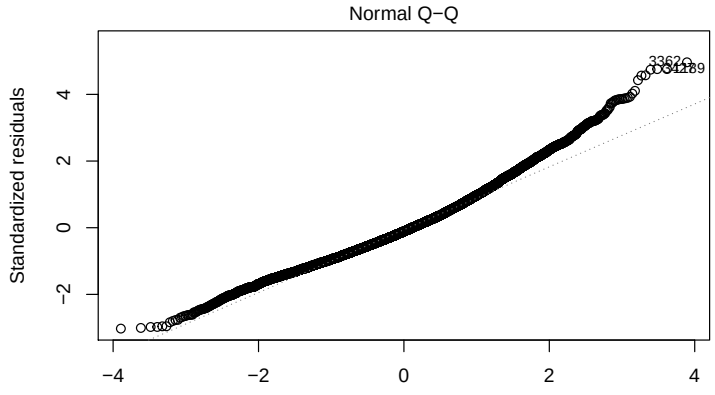
lm(formula = log(bmxbmi)~age+riagendr+ridreth2+dmdeduc)

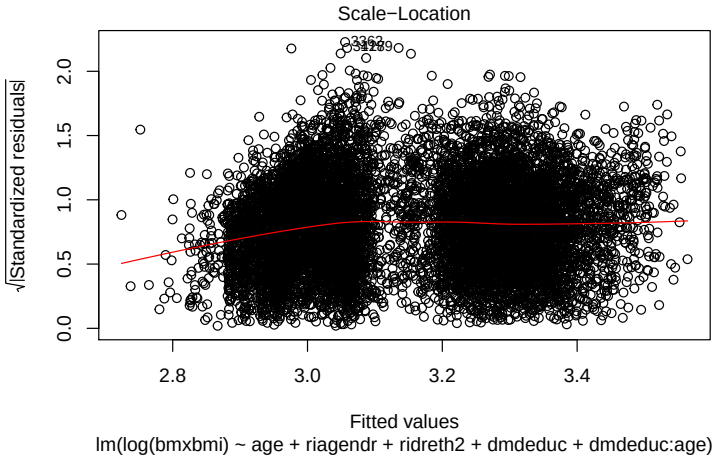
Table 3: Logistic Regression Summary

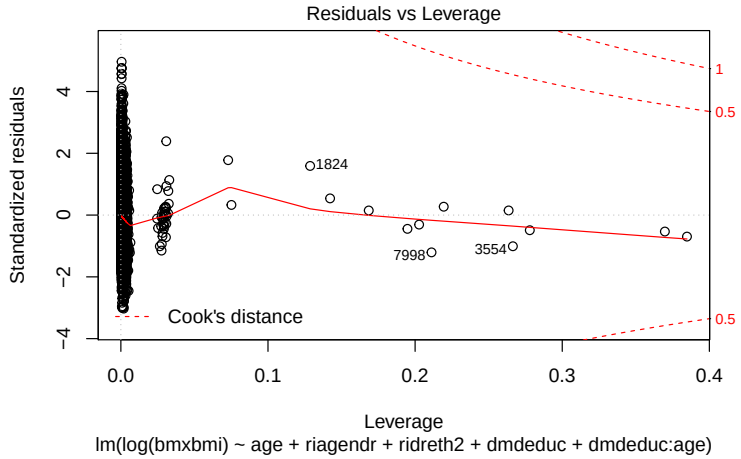
term	estimate	std.error	statistic	p.value
(Intercept)	2.8811926	0.0055932	515.1234840	0.0000000
age	0.0071167	0.0001302	54.6616492	0.0000000
riagendr2	0.0194179	0.0043504	4.4634316	0.0000082
ridreth22	0.0636291	0.0055919	11.3788433	0.0000000
ridreth23	0.0469889	0.0058388	8.0476465	0.0000000
ridreth24	-0.0595867	0.0124640	-4.7807044	0.0000018
ridreth25	0.0318960	0.0123916	2.5739913	0.0100674
dmdeduc2	0.2708035	0.0135130	20.0401573	0.0000000
dmdeduc3	0.2638016	0.0116829	22.5800616	0.0000000
dmdeduc7	-0.1613743	0.1190253	-1.3557979	0.1751938
dmdeduc9	-0.1026835	0.0384064	-2.6736007	0.0075162
age:dmdeduc2	-0.0045036	0.0002960	-15.2129831	0.0000000
age:dmdeduc3	-0.0044577	0.0002616	-17.0404894	0.0000000
age:dmdeduc7	-0.0020060	0.0024737	-0.8109583	0.4174087
age:dmdeduc9	-0.0016659	0.0013665	-1.2191019	0.2228340

[1] Adjusted R-square: 36.8%
[2] Residual standard error: 0.22









On investigating the residual plots of the above model, we see that the following observations can be made:

- Linearity- The residual plots show a very slight curvature. However, the pattern is very slight and might not be significant. We can consider linearity being maintained here.
- Independence and Equal Variance: The residual versus fitted plot does not show an evident conical (spread out or converging) pattern. This might indicate that there is no heteroskedasticity in the dataset or the model.

- Normality- The assumption of normality is not met as the QQ-plot shows a slight curvature.

All residuals seem to be similar across the different levels of the categorical variables. For age, the residual plot follows an almost linear trend with a very slight downward curve.

Interpretation of the model results.

The model response is log(bmxbmi). The model results reveal the following:

- All variables are significant. However, 2 levels of main effects of education (i.e. 7 and 9) did not come out to be significant owing to their high p-value.
- The most significant predictor is age based on the absolute value of the t-statistic.
- The intercept of the model is 2.88. This implies that the baseline of $log_e(bmi)$ is 2.88. In the exponential scale, this means the baseline 'bmi' for a non-hispanic white male person with age 0 and less than high school education, is 17.81.
- The most significant predictor is age. The coefficient for the same is 0.007. This means, for every unit increase in age, keeping all other factors constant, the bmi of a person will increase by 1.007 in exponential scale (and 0.007 in log scale).
- Similarly, the coefficient for gender level 2 (female) is 0.019. This means, compared to the baseline, with all other factors as constants, a female (white non-Hispanic person with less than high school education and age 0) will have a bmi 1.019 higher in exponential scale (or 0.021 in log scale).
- Getting a high school diploma, keeping all other factors constant, increases the bmi by 1.32 on the exponential scale (0.27 in log-scale).

The residual standard error of the model is 0.22. It denotes the difference between the predictions and the observed values. The error does not seem to be too high.

There are some limitations to the model. The assumption of normality is not completely met. Transforming the response variable by log reduces the interpretability of the model. The model has been built on imputed values which might not be the exact representation of the data. Let us once confirm these results against the pooled results.

Let us now look at the pooled results.

```
##               estimate std.error statistic      df p.value
## (Intercept)       2.88      0.01    498.44 2260.90    0.00
## age               0.01      0.00     51.29  669.54    0.00
## riagendr2         0.02      0.00      4.59 2107.70    0.00
## ridreth22         0.06      0.01     11.27 7584.55    0.00
## ridreth23         0.05      0.01      8.30 3124.14    0.00
## ridreth24        -0.06      0.01     -4.37  747.88    0.00
## ridreth25         0.03      0.01      2.27 1298.08    0.02
## dmdeduc2          0.27      0.02     15.86   60.44    0.00
## dmdeduc3          0.27      0.01     17.86   61.85    0.00
## dmdeduc7         -0.04      0.13     -0.31   85.42    0.75
## dmdeduc9         -0.05      0.09     -0.48   82.04    0.63
## age:dmdeduc2      0.00      0.00    -12.55   78.44    0.00
## age:dmdeduc3      0.00      0.00    -14.05   86.00    0.00
## age:dmdeduc7      0.00      0.00     -1.62  232.44    0.11
## age:dmdeduc9      0.00      0.00     -1.48  137.55    0.14
```

```
##          est     lo 95     hi 95 fmi
## R^2 0.368771 0.3529527 0.3845326 NaN
```

Even the pooled results show an equivalent R-squared of 36.8%. This indicates that 36.8% of the variation in the response variable is explained by the predictors. This matches our model results on the 9th imputed dataset. The intercept of the model on dataset number 9 and the pooled model is also the same i.e. 2.88. The significant variables are age, gender, race, education (except two levels- i.e. level 7 and 9) and interaction terms between age and education barring levels 7 and 9 for education. The most significant variable from the pooled model is age based on its t-statistic. The interpretation on the pooled model can be derived in a similar fashion as the one on the imputed dataset.