Assignment 4 (need grade)

Maobin Guo

Part I

Task 1.1

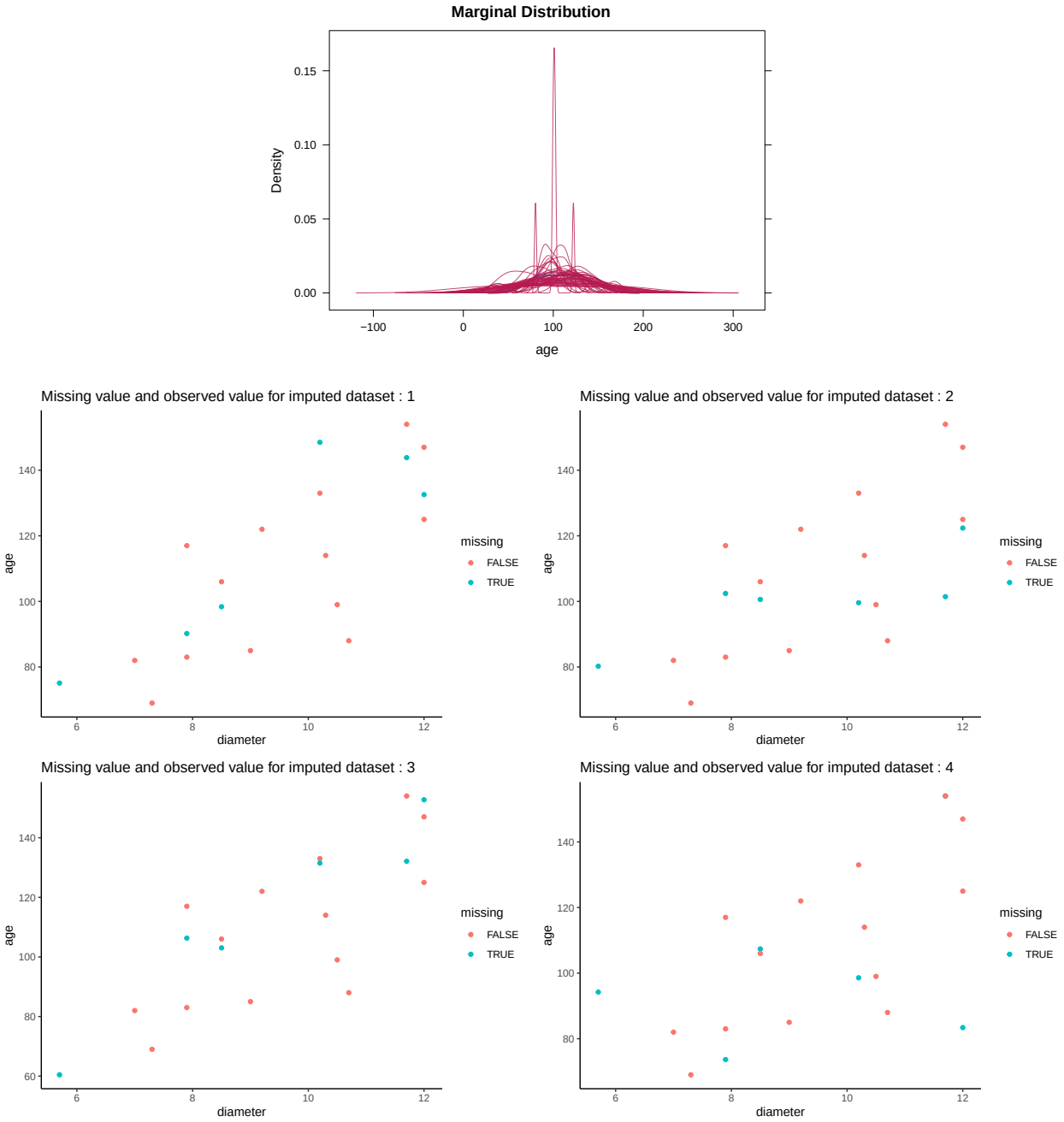
The following code is used to replace 30% age with missing value randomly.

```
missing_data[head(sample(20), 0.3 * nrow(df)), ]$age <- NA
```

This is the new dataset with missing value.

	diameter	age
1	12.00	125
2	11.70	
3	7.90	83
4	9.00	85
5	10.50	99
6	7.90	117
7	7.30	69
8	10.20	133
9	11.70	154
10	10.20	
11	8.50	
12	5.70	
13	10.30	114
14	12.00	147
15	9.20	122
16	8.50	106
17	7.00	82
18	10.70	88
19	12.00	
20	7.90	

Task 1.2



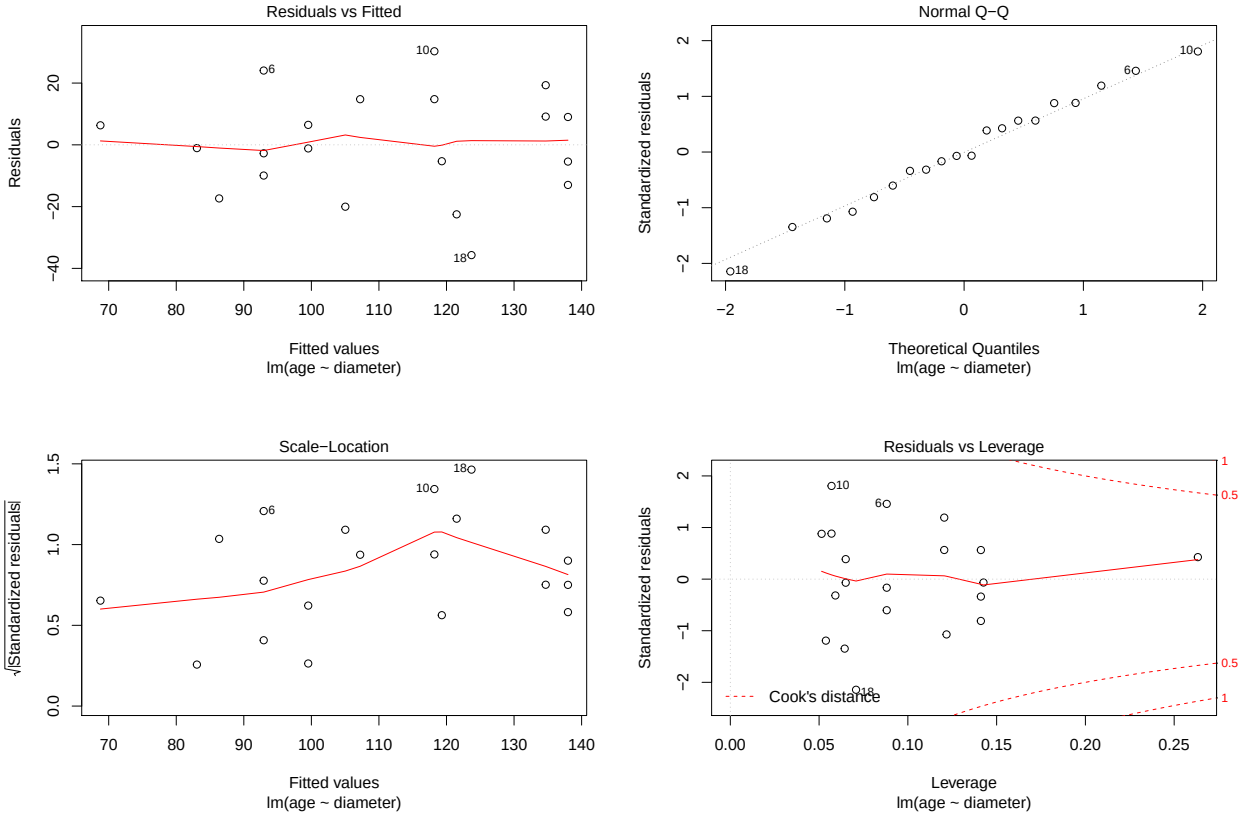
Conclusion: According to the scatter plot, the result is acceptable since the added data are in a reasonable range. Most of the imputed ages are larger than the real value for a small diameter data point. However, one of the imputed datasets predicts a very close value of real age. According to marginal distribution, most of the imputed data's distribution is similar to observed data, except for a few too centered distributions. Considering the tiny size of the dataset, the few exceptionals are acceptable.

Task 1.3

The model is :

$$age_i = \beta_1 * diameter_i + \beta_0 + \varepsilon_i; \varepsilon \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$$

Model Assessment The assessment plots of one of the imputed datasets are as follows. According to residual analysis, there is no obvious evidence indicate the assumptions of linear regression were broken. Moreover, there is not high influence data to concern.



Conclusion

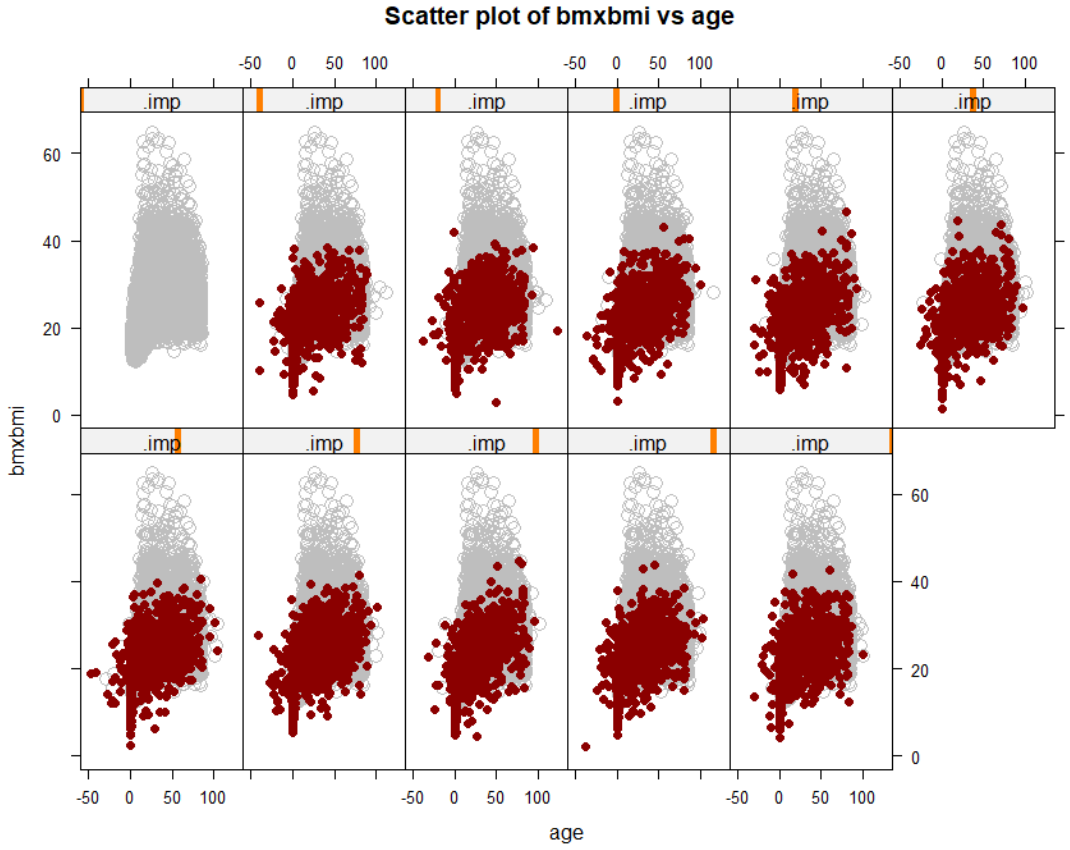
	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	2.51	32.86	0.08	7.17	0.94
2	diameter	11.08	3.44	3.22	6.96	0.01

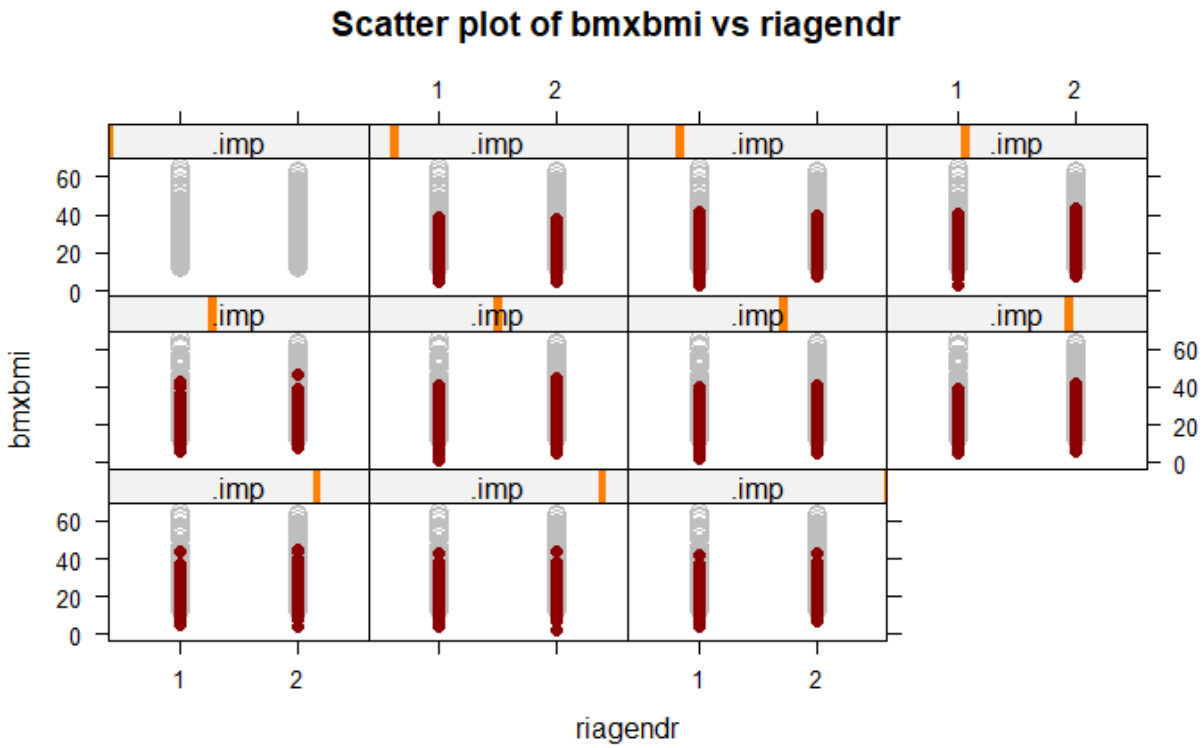
The table is the summary of the model applying on multiple imputation datasets. The diameter has positive effects on the age of trees because its p-value is significant. Suppose the diameter increase by 1 unit, the age of the tree would increase by 11 years. The intercept's p-value is near 1, which means it is not significant. If there is more data, then a better model can be obtained, in which intercept may be significant.

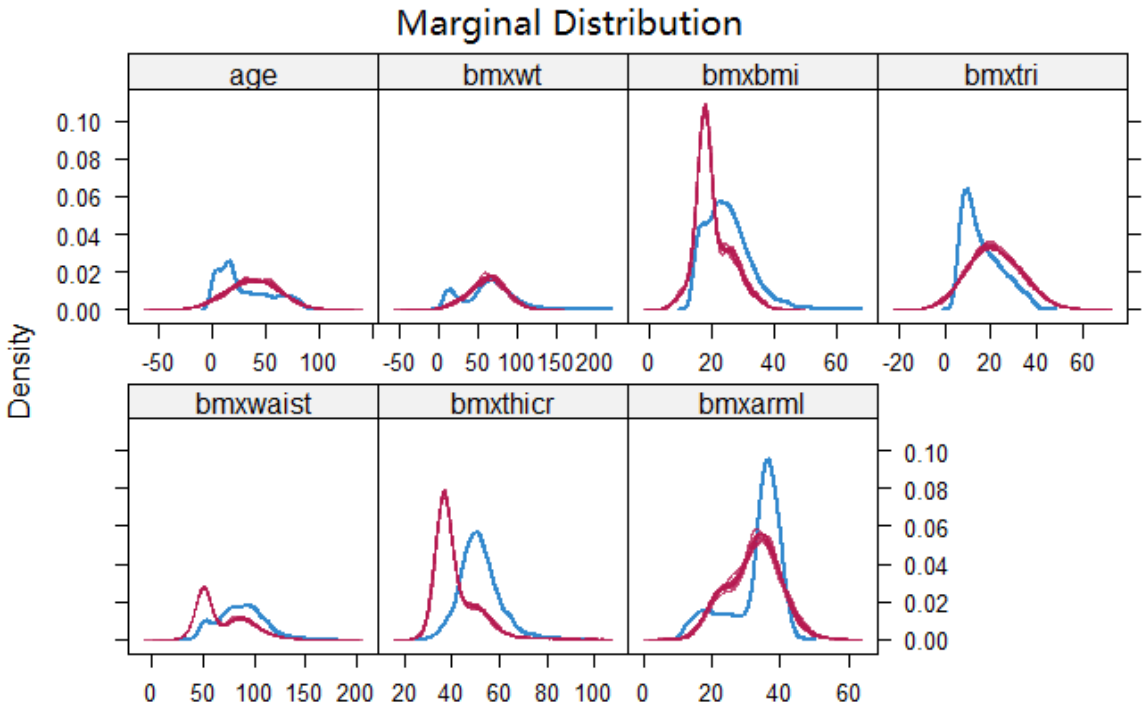
PART II

Task 2.1

Plots:







Conclusions:

According to the scatter plots of bmxbmi by age and riagendr, I do not think the imputation is great. Because the imputation data (red points) clusters at a subset of the observed data (gray points). Bmxbmi and bmxaml's marginal distributions show quite different distribution between imputation data and observed

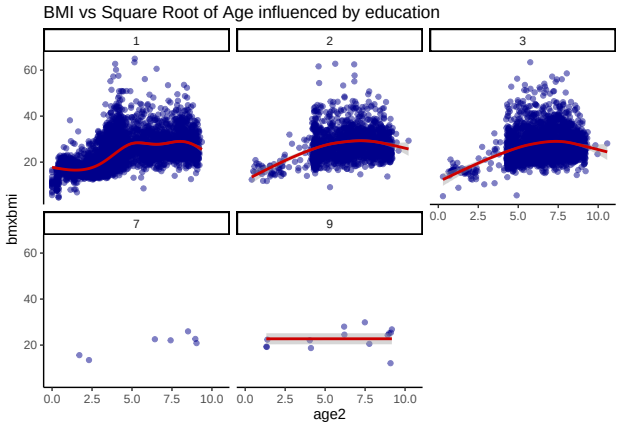
data. Considering the dataset's size, the differences are considerable. The quality of the imputation model is not satisfying.

Task 2.2

The model is

$$\begin{aligned} log(bmxbmi_i) =& \beta_1 * \sqrt{age}_i + \beta_2 * riagendr_i + \beta_3 * dmdeduc_i + \\ & \beta_4 * indfminc_i + \beta_5 * dmdeduc_i : \sqrt{age}_i + \beta_0 + \varepsilon_i; \\ & \varepsilon \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \end{aligned}$$

Model Exploration



In the model exploring stage, I tried to compare many models with ANOVA test on a complete data set and finally decided to use the above model. After that, the model was applied to another data set, and a similar conclusion was reached. Then I decided to use the model to predict BMI. In the model fitting step, I found the log transformation could cause the response variable closer to the normal distribution. Moreover, the ANOVA test suggests that the model will perform better after transforming age to the square root of age ($\sqrt{age}$). So I kept these two transformations in the final model.

The plot above illustrates the interaction between the square root of age and education versus the response variable. The BMI trend with the square root of age is different according to education. This interaction is also confirmed in the model fitting step.

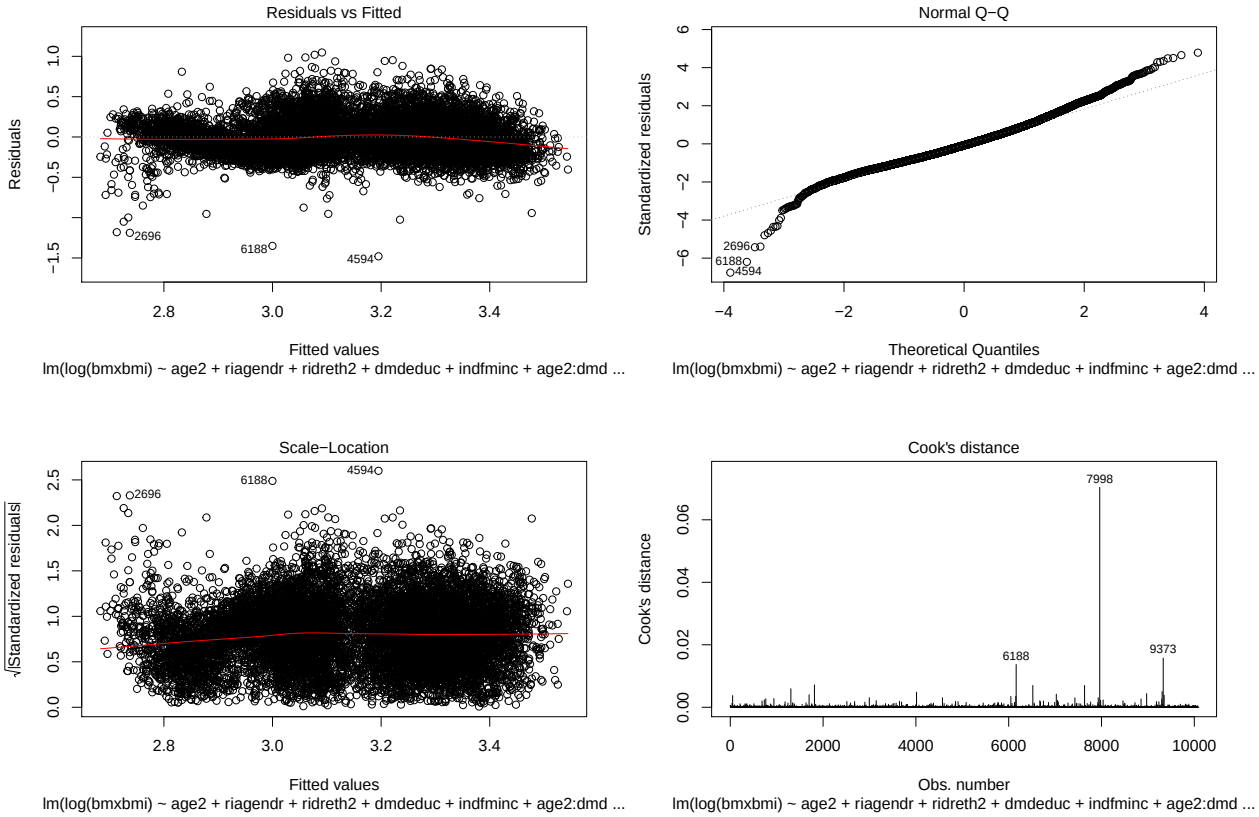
For the indfminc variable, I try to combine some different categories together, such as combining incomes of less than US$20,000 into a group and combining incomes of more than US$20,000 into another group. However, the ANOVA test suggested that this kind of merging will not improve the model's performance, so I finally gave up this kind of merging. I also used AIC, BIC did a model search in different directions, and the final result was the same as the above model. So I finally decided to use this model.

Model Assessment

From the residual plots, there are no obviously violations of assumptions:

1) Linearity: The residual versus predictor plot seems random.
2) Independence and Equal Variance: Absence of any pattern, randomness and wide-spread distributions over the spectrum support these assumptions.
3) QQ-plot supports the assumption of Normality generally as the plot is a straight line.

Also, according to Cook's distance, there is no high leverage data point, which is good news for the imputed dataset. To ensure no multicollinearity, VIF scores were generated, noticing that all variables had VIF value below 5.



	GVIF	Df	GVIF^(1/(2*Df))
age2	1.92	1.00	1.38
riagendr	1.01	1.00	1.00
ridreth2	1.22	4.00	1.03
dmdeduc	5852.09	4.00	2.96
indfminc	1.19	14.00	1.01
age2:dmdeduc	7125.69	4.00	3.03

Conclusion

The baseline values incorporated in the intercept are age = 0, male, Non-Hispanic white, less than high school education, and income below $4999. From the model summary, we found that age, gender, race, education, and the interaction between the square root of age and education are strongly associated (statistically significant) with the BMI. Almost all income categories are not statistically significant with BMI except indfminc9 ($55,000 to $64,999). Considering that no other income categories are statistically significant, and the coefficient of indfminc9 is small, I tend to consider income would not affect a person's BIM.

1. Controlling other factors, one year raise on the square root of age would increase his/her BMI by 8.4%
2. Controlling other factors, the gender of female would raise BMI by 2.1%
3. All categories of race are statistically significant; however, their coefficient is small. Controlling other factors, when the race was changed from Non-Hispanic whites to Non-Hispanic black, Mexican American, other race and other Hispanic, their BMI would increase by 7%, 5% -4.3% 2.5% respectively.
4. Controlling other factors, high school, and more than a high school diploma would increase BMI by 33% and 30%.

	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	2.70	0.01	246.49	10044.88	0.00
2	age2	0.08	0.00	63.49	10044.88	0.00
3	riagendr2	0.02	0.00	4.77	10044.88	0.00
4	ridreth22	0.07	0.01	12.07	10044.88	0.00
5	ridreth23	0.05	0.01	9.14	10044.88	0.00
6	ridreth24	-0.04	0.01	-3.46	10044.88	0.00
7	ridreth25	0.03	0.01	2.01	10044.88	0.04
8	dmdeduc2	0.29	0.02	13.17	10044.88	0.00
9	dmdeduc3	0.27	0.02	13.50	10044.88	0.00
10	dmdeduc7	-0.22	0.20	-1.11	10044.88	0.27
11	dmdeduc9	0.24	0.13	1.83	10044.88	0.07
12	indfminc2	-0.01	0.01	-0.81	10044.88	0.42
13	indfminc3	-0.01	0.01	-1.29	10044.88	0.20
14	indfminc4	-0.01	0.01	-1.25	10044.88	0.21
15	indfminc5	0.00	0.01	0.32	10044.88	0.75
16	indfminc6	-0.02	0.01	-1.60	10044.88	0.11
17	indfminc7	0.02	0.01	1.39	10044.88	0.16
18	indfminc8	-0.00	0.01	-0.35	10044.88	0.73
19	indfminc9	0.03	0.01	2.55	10044.88	0.01
20	indfminc10	0.01	0.01	0.53	10044.88	0.60
21	indfminc11	-0.00	0.01	-0.42	10044.88	0.67
22	indfminc12	-0.02	0.02	-0.98	10044.88	0.33
23	indfminc13	-0.00	0.02	-0.10	10044.88	0.92
24	indfminc77	-0.02	0.03	-0.88	10044.88	0.38
25	indfminc99	0.01	0.03	0.22	10044.88	0.83
26	age2:dmdeduc2	-0.04	0.00	-10.94	10044.88	0.00
27	age2:dmdeduc3	-0.04	0.00	-11.52	10044.88	0.00
28	age2:dmdeduc7	-0.01	0.03	-0.22	10044.88	0.82
29	age2:dmdeduc9	-0.06	0.02	-3.13	10044.88	0.00

5. At the significance level of $p<0.001$, keeping other variables constant for every increase in the square root of age, a person with a high school diploma his/her BMI tends to decrease by 3.8%.

6. At the significance level of $p<0.001$, keeping other variables constant for every increase in the square root of age, a person with more than a high school diploma his/her BMI tends to decrease by 3.6%.