702 Assignment IV

Question 1

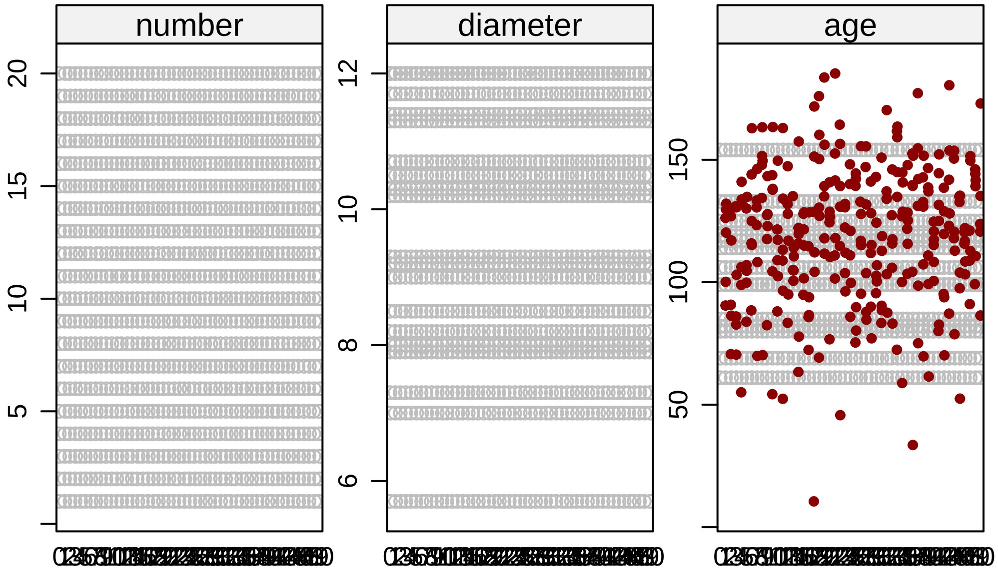
Data Inspection

After loading the tree.txt, we randomly assign 30% of the age data to be NaN values.

number	diameter	age
1	12.0	125
2	11.4	119
3	7.9	NA
4	9.0	85
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	80
13	10.3	114
14	12.0	NA
15	9.2	NA
16	8.5	106
17	7.0	82
18	10.7	NA
19	9.3	NA
20	8.2	99

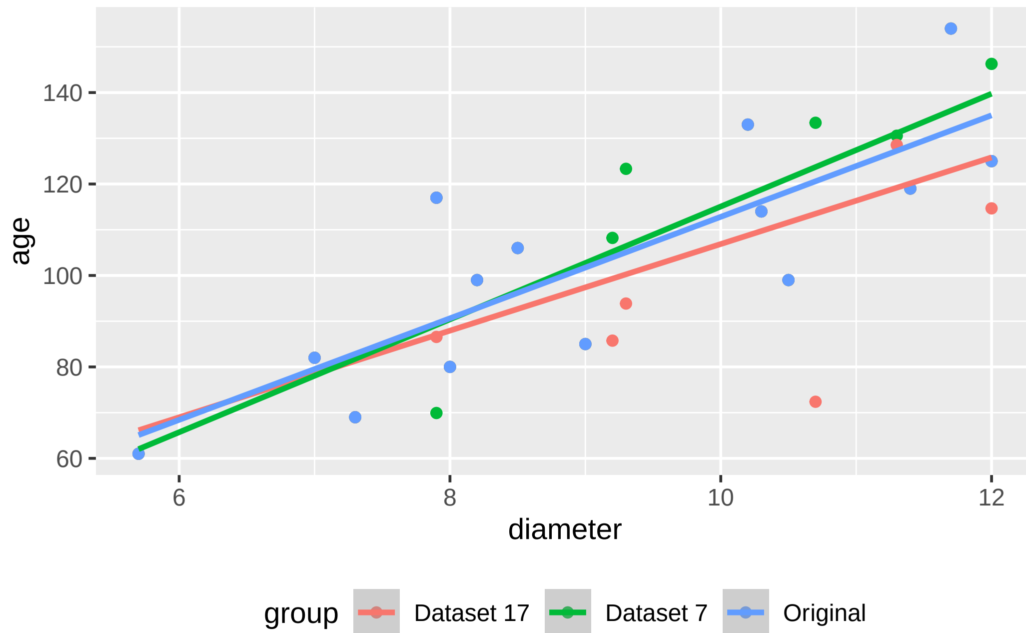
Imputation

First, we look at stripplot to examine the distribution of imputed values compared to the observed values. Using 'norm'argument, it turns out that our imputed values have much larger variance than the actual data.

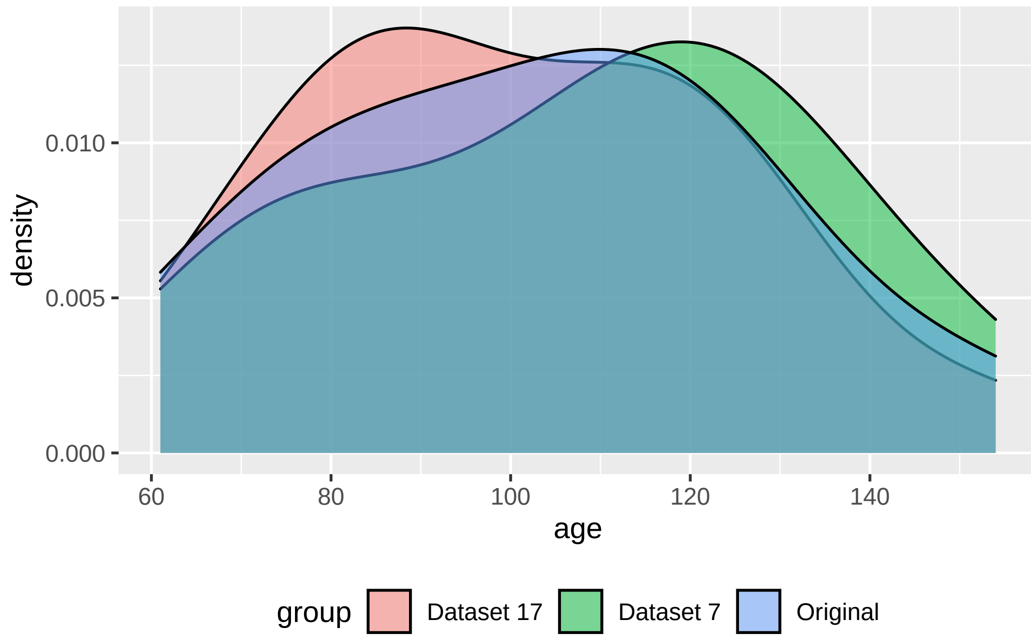


In order to examine the quality of imputation, we randomly select dataset 7 and 17 from the 50 imputed complete dataset.

Using scatter plot to examine the relationship between `diameter` and imputed `age` across dataset 7, 17, and original data, we observe a similar positive correlation between the two variables.



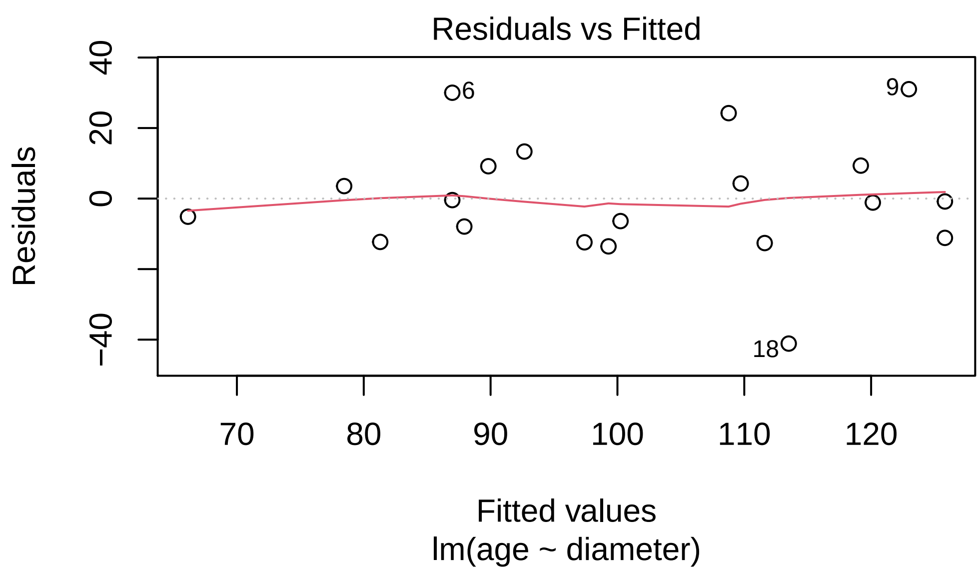
Using density plot, we can inspect the how much our imputed age values overlap with the observed values. Intuitively, the more the overlap there is, the better the quality our imputed values are, since it signals that they approximate the true values.

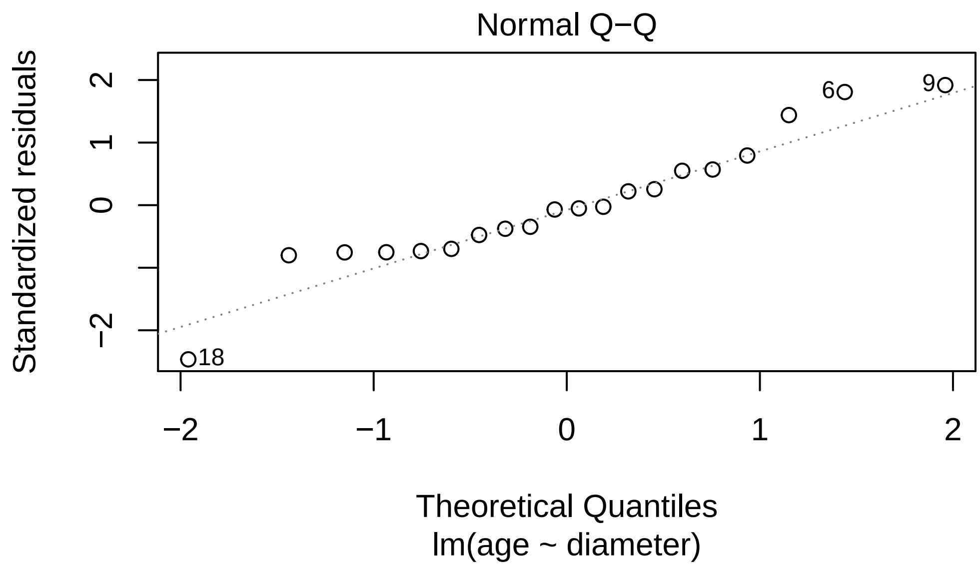


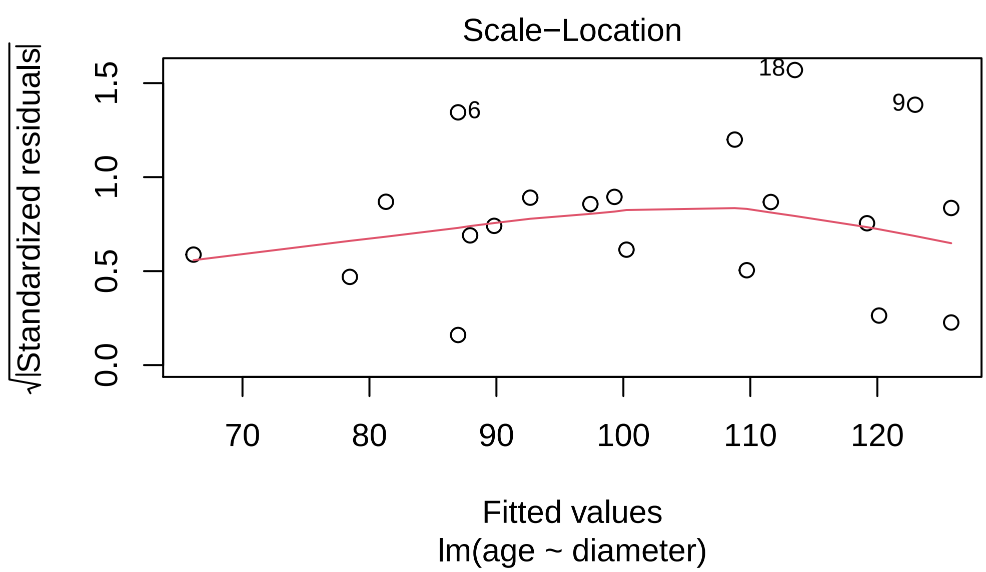
Via visual inspection, our two randomly drawn imputed `age` values well approximate the observed values, and the relationship between `age` and `diameter` is preserved.

Model

Before fitting a linear model on our full dataset, we can fit the regression model on one of the selected dataset to check for linear regression assumptions beforehand. This is done on dataset 17.







Based on the residual vs. fitted plot and QQ plot, we are confident that our assumptions of linearity, equal variance, and indepence are satisfied on dataset 17, showing that our imputed data are likely to meet linear regression assumptions.

Pooling

Table 2: Final Imputation Model

term	estimate	std.error	statistic	p.value	b	df	dfcom	fmi	lambda	m	riv	ubar
(Intercept)	3.5866	1.2505	2.8681	0.0143	0.3966	11.8769	18	0.3583	0.2587	50	0.3490	1.1593
age	0.0541	0.0117	4.6180	0.0007	0.0000	11.0366	18	0.4062	0.3076	50	0.4442	0.0001

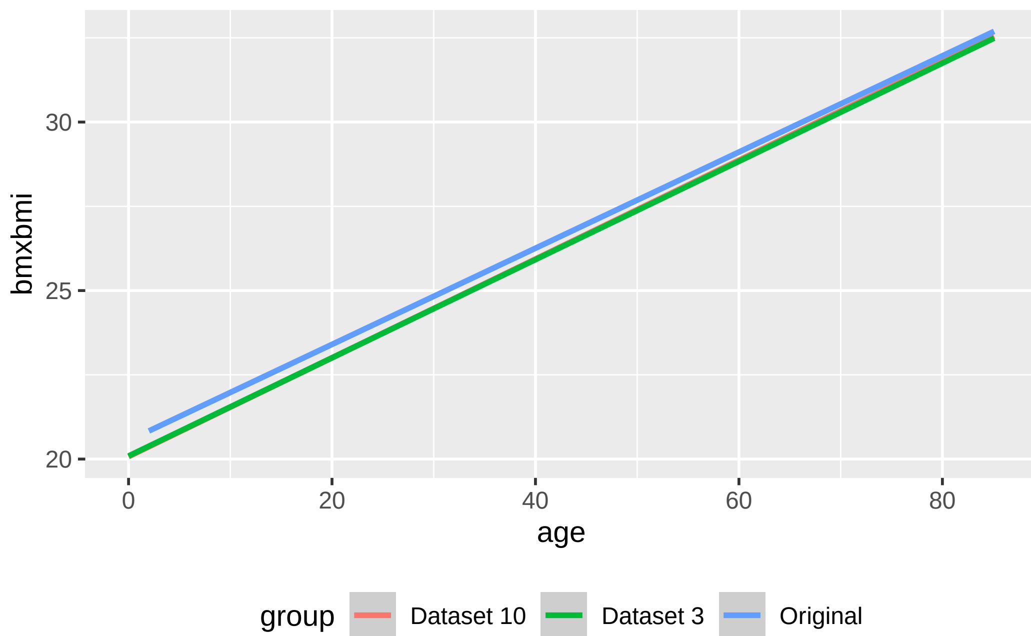
Using the multiple imputation combining rules to do the regression of age on diameter, we find that the coefficient of age (= 0.05) as a predictor for diameter is statistically significant ($p < 0.05$), with a confidence level of (0.03, 0.08). Our intercept is 3.59, which is also statistically significant ($p < 0.05$). It means that for any given tree at age 0, the predicted diameter is 3.59 cm. It shows that with every unit increase age, the diameter of the tree is predicted to increase by 0.05 cm. Overall, this regression model summary shows that there is a positive correlation between age and diameter, and this relationship is statistically significant given our pooled data. However, our limited sample size can potentially undermine the validity of our conclusion, and the quality of imputed values can be further investigated.

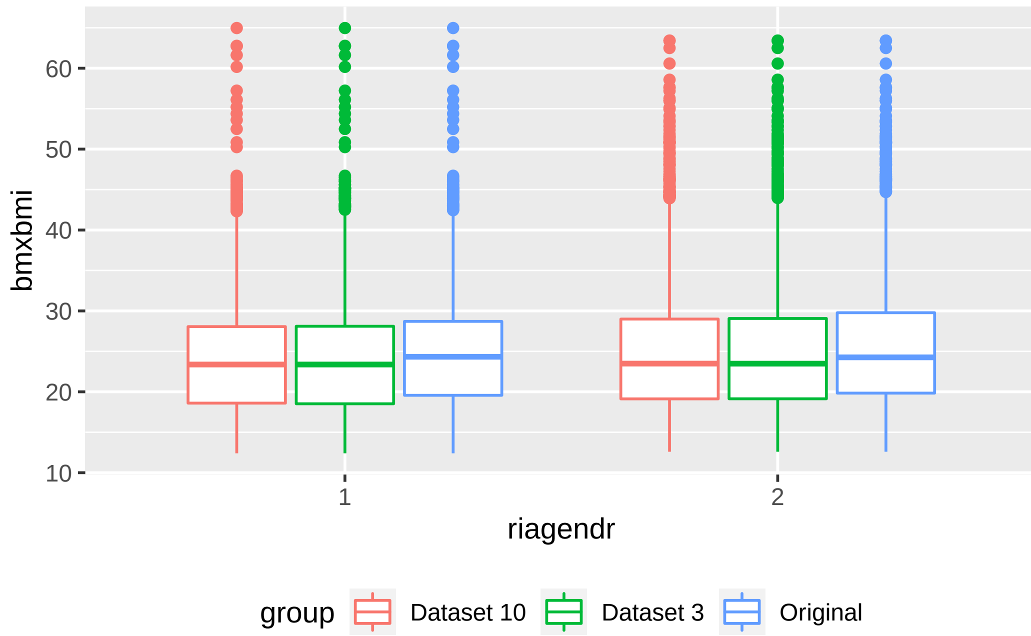
Question 2

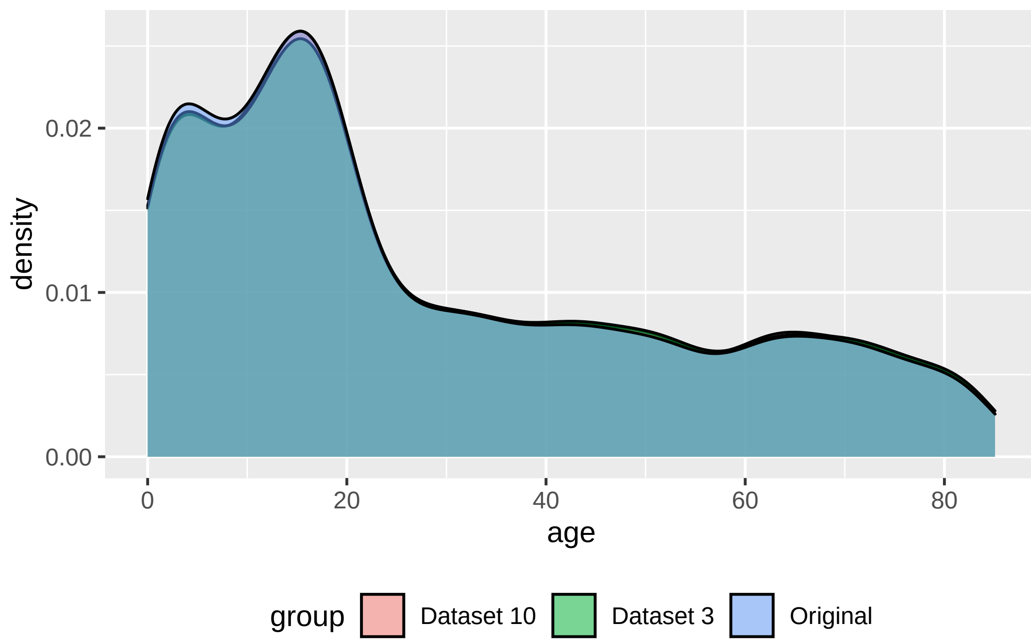
Imputation

We use the nhanes data to investigate the relationship between Body Mass Index and potential predictors. To deal with the missing values, we use multiple imputation approach, with 'pmm' argument to create 10 imputed datasets. In this case, we prefer 'pmm' instead of 'norm' as an effort to avoid incurring negative values of age in imputed datasets.

In order to examine the quality of imputed data, dataset 3 and dataset 10 were randomly chosen. By plotting the relationship between `bmxbmi` and `age` across dataset 3, dataset 10, and the original data, we find that the linear relationship is consistent. The distribution of `bmxbmi` across `riagendr` groups are also consistent across the three datasts. Besides, the density plot of `age` shows that our imputed `age` highly overlap with the observed values. As a result, we are confident about the quality of our imputed values.

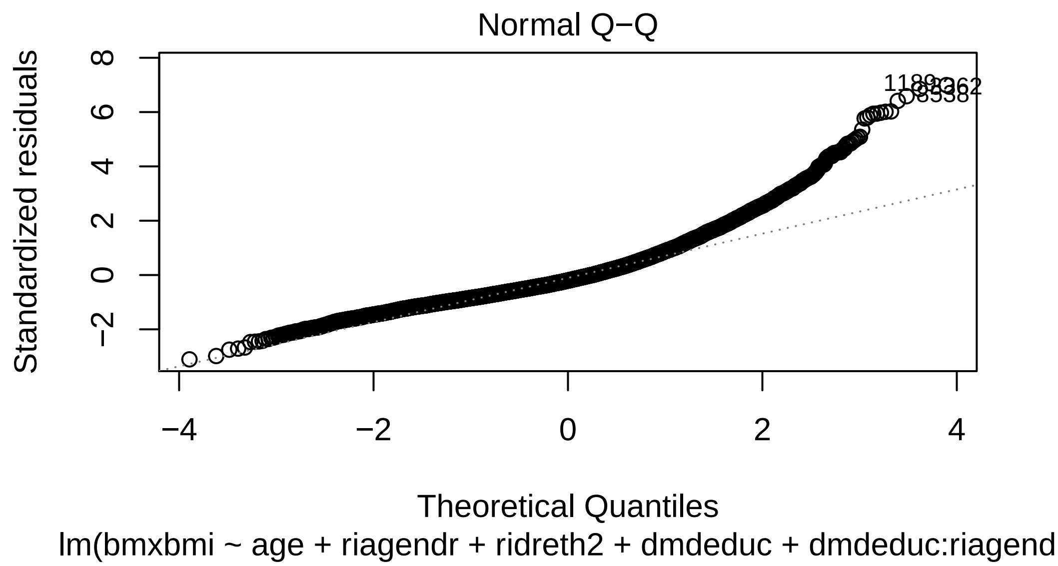




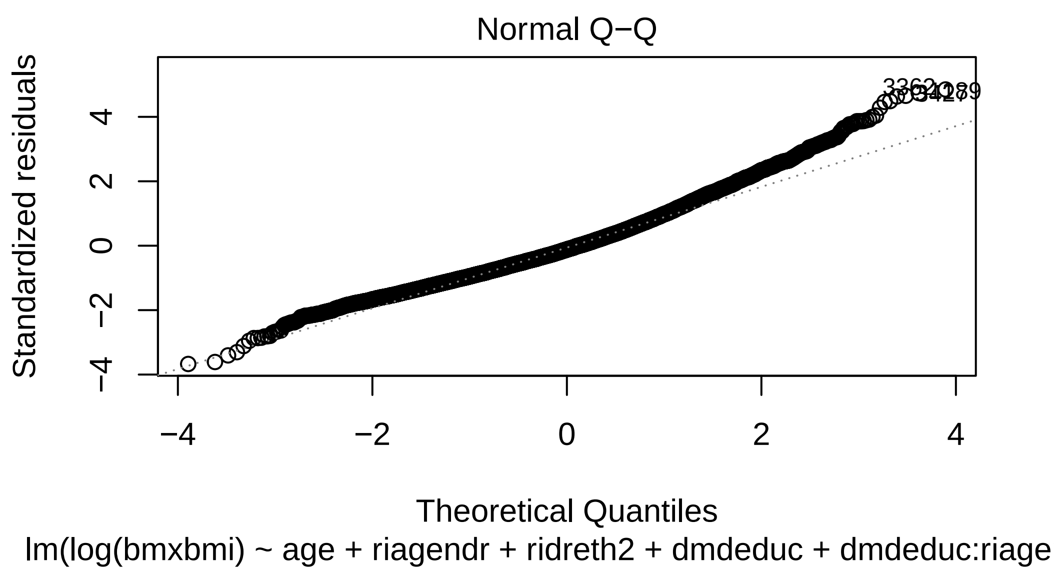


Model

Given our research interest, we then fit a linear model on imputed data 3 using `age`, `riagendr`, `ridreth2`, `dmdeduc` as predictors of `bmxbmi`. Since we are also interested in whether or not education and gender will have interaction effect on `bmxbmmi`, an interaction term `dmdeduc:riagendr` is also included. We find that regression assumptions, including equal variance and especially normality are violated. Thus, we consider transforming our response variable to log scale in order to mitigate the skewness.



After using log-scale `bmxbmi` as the response variable, we find that the issue of assumption violation is greatly alleviated. Thus, we decide to conduct our analysis on log-bmxbmi scale.



Before fitting the model, we use backward model selection method (metric = AIC) to aid model selection. It turns out that the the selection did not screen out any predictor using AIC metric, and we continue using the model on pooled dataset.

Thus, we first convert `bmxbmi` to log scale and conduct another multiple imputation.

Pooling

Then we use multiple imputation combining ruls to find point and variance estimates.

Table 3: Final Imputation Model

term	estimate	std.error	statistic	p.value	b	df	dfcom	fmi	lambda	m	riv	ubar
(Intercept)	2.8748	0.0065	442.9371	0.0000	0.0000	360.4742	10107	0.1593	0.1546	10	0.1829	0.0000
age	0.0059	0.0001	49.7535	0.0000	0.0000	311.2925	10107	0.1722	0.1669	10	0.2003	0.0000

term	estimate	std.error	statistic	p.value	b	df	dfcom	fmi	lambda	m	riv	ubar
riagendr2	0.0337	0.0060	5.6323	0.0000	0.0000	2323.5773	10107	0.0550	0.0541	10	0.0572	0.0000
ridreth22	0.0751	0.0060	12.4969	0.0000	0.0000	1875.9576	10107	0.0630	0.0620	10	0.0661	0.0000
ridreth23	0.0643	0.0061	10.5067	0.0000	0.0000	6062.9800	10107	0.0242	0.0239	10	0.0245	0.0000
ridreth24	-0.0467	0.0138	-3.3963	0.0007	0.0000	680.9034	10107	0.1131	0.1105	10	0.1243	0.0002
ridreth25	0.0409	0.0134	3.0606	0.0022	0.0000	1608.7523	10107	0.0693	0.0681	10	0.0731	0.0002
dmdeduc2	0.1360	0.0101	13.4271	0.0000	0.0000	1492.7837	10107	0.0724	0.0712	10	0.0767	0.0001
dmdeduc3	0.1253	0.0088	14.2647	0.0000	0.0000	1910.2861	10107	0.0623	0.0613	10	0.0653	0.0001
dmdeduc7	-0.1373	0.1178	-1.1648	0.2445	0.0013	724.9141	10107	0.1093	0.1069	10	0.1196	0.0124
dmdeduc9	-0.1231	0.0900	-1.3685	0.1741	0.0021	104.7429	10107	0.3041	0.2910	10	0.4104	0.0057
riagendr2:dmdeduc2	-0.0375	0.0138	-2.7206	0.0067	0.0000	828.1068	10107	0.1016	0.0994	10	0.1104	0.0002
riagendr2:dmdeduc3	-0.0233	0.0112	-2.0812	0.0376	0.0000	1927.9666	10107	0.0620	0.0610	10	0.0650	0.0001
riagendr2:dmdeduc7	-0.0922	0.1579	-0.5836	0.5596	0.0022	842.4569	10107	0.1006	0.0985	10	0.1092	0.0225
riagendr2:dmdeduc9	-0.0191	0.1182	-0.1618	0.8717	0.0035	117.3954	10107	0.2867	0.2747	10	0.3787	0.0101

Based on our model, we find that multiple predictors/levels of predictor are statistically significant for predicting `bmxbmi`. The intercept is 2.87, meaning that for a white male aged 0 with less than high school educcation, the predicted BMI is 17.67. Holding all other variables constant, one unit increase in age will increase predicted BMI by 1.00. Holding all other variables constant but changing the gender to female will increase predicted BMI by 3.4%. Holding all other variables constant, race being Black will increase predicted BMI by 8%.Holding all other variables constant but changing the gender to female will increase predicted BMI by 3.4%. Holding all other variables constant, race being Mexican American will increase predicted BMI by 7%. Holding all other variables constant, race being other race including multi-racial will decrease predicted BMI by 4%. Holding all other variables constant, race being other Hispanic will increase predicted BMI by 1.05. Holding all other variables constant, highest level of eucation being high school diploma will increase predicted BMI by 15%. Holding all other variables constant, highest level of eucation being more than high school will increase predicted BMI by 13%. The interaction effect is statistically significant only when holding all other variables constant and the gender is female and the highest level of education is high school diploma or more than high school. Speficially, when holding all other variables constant, if gender is female and the highest level of education is high school diploma, BMI value is predicted to decrease by 4%. When holding all other variables constant, if gender is female and the highest level of education is more than high school, BMI value is predicted to decrease by 2.3%.