Assignment 4

Anshupriya Srivastava

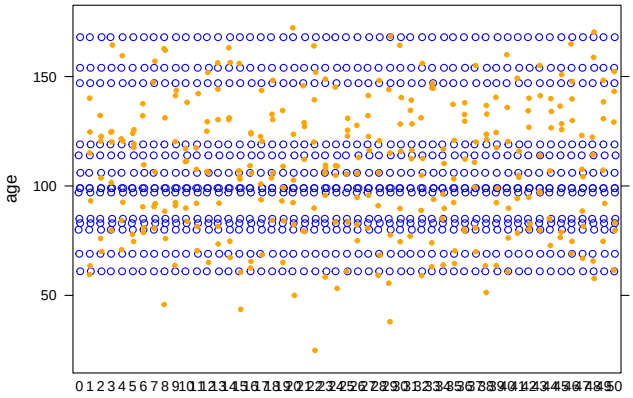
11/11/2019

Missing Data Mechanics

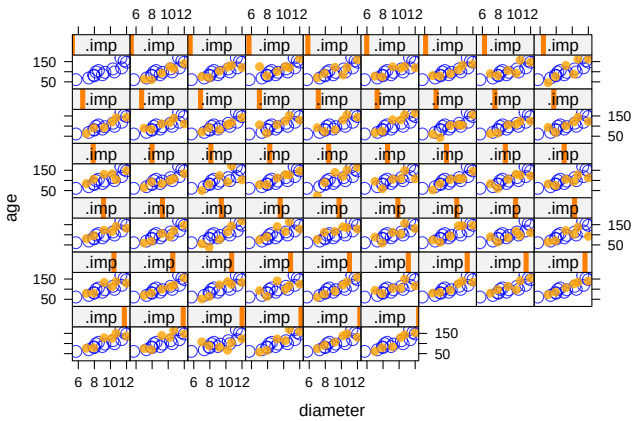
The given dataset contains the diameters and age of 20 trees. A dataset is created with 30% of the age values missing completely at random leaving all other values intact. The following command is used to create missing values -

```
set.seed(123)
ind <- floor(runif(6, 1, 20))
treeage$age[ind] <- NA
```

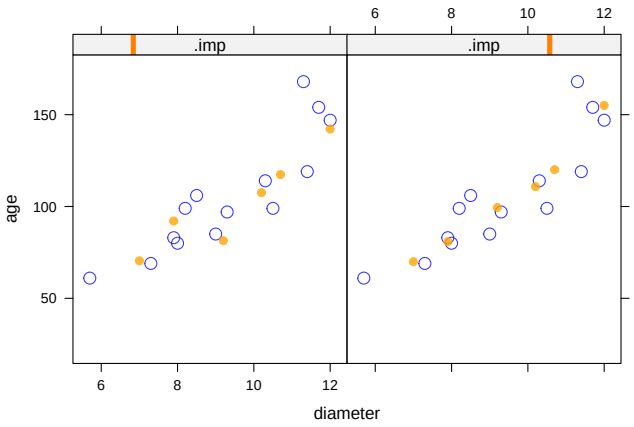
50 imputated datasets are created using the method "norm". The seed has been set to 100. The plot showing the imputed values and observed values is shown below.



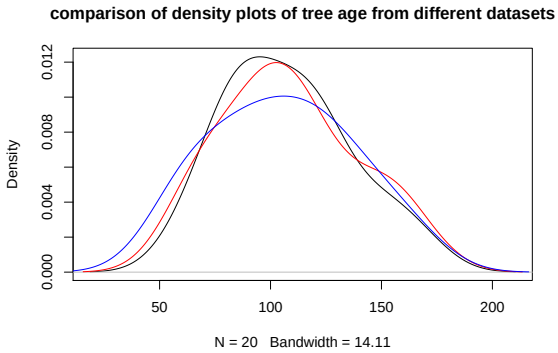
The shape and distribution of the imputed and observed data points look quite similar. It can be assumed that the generated values can be used in model generation. An xyplot showing the relationship between the age as a function of diameter further evalutes these imputations.



The scatterplot show a positive linear relationship between age and diameter. By examining all 50 plots, it seems like plots - 12 and 38 have the closest relationship with the observed points. Further evalutions will be performed on there plots. Scatterplots of these plots are shown below.



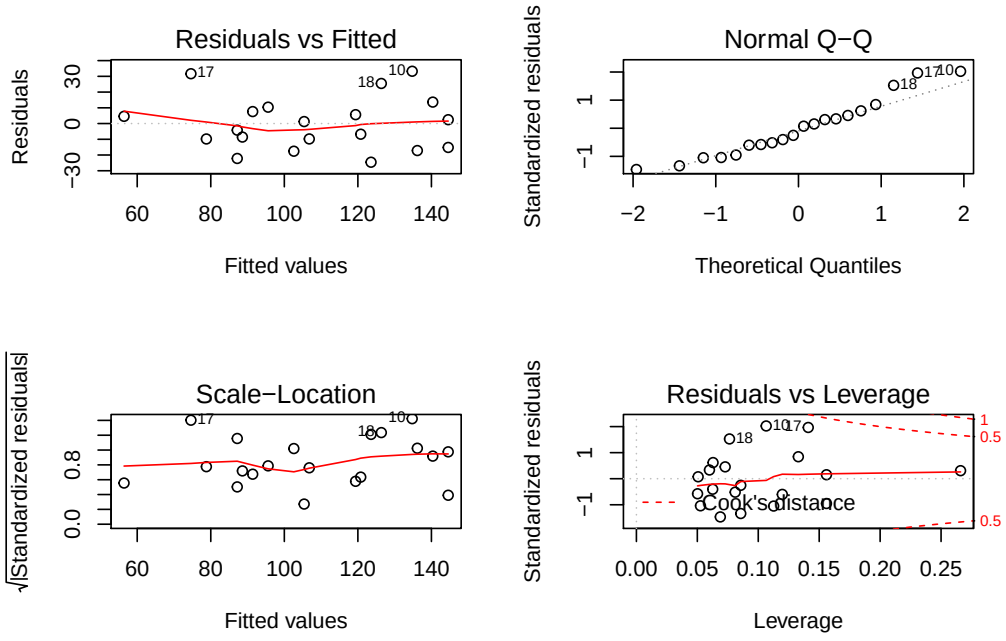
The density plots of the selected impuations are compared to original dataset.



It can be observed that age is normally distributed. The black line shows the distribution of the original dataset. Red and blue line shows the distribution of the imputed dataset selected in this analysis. The selected dataset seem to show a similar distribution. Dataset 12 seems more appropriate so a linear model -

$$age = \beta_0 + \beta_1 * diameter$$

is used to fit the dataset.



term	estimate	std.error	statistic	p.value
(Intercept)	-23.33876	20.85173	-1.119272	0.2777382
diameter	13.99214	2.17831	6.423393	0.0000048

Using the residual plots and summary plots we can infer the following:

- **Linearity**: There seems to be no evident pattern in the residual plot. This indicates that the model is meeting assumption linearity.
- **Independence and Equal Variance**: Since the residual versus fitted plot indicates no evident conical pattern (spread out or converge) there is no indication of heteroskedasticity in the dataset or the model.
- **Normality**: The Q-Q plot is used to interpret Normality. The relationship is approximately linear with the exception of three data points.
- **Leverage Plot**: The points are well inside the curves. No point has leverage on the overall dataset.

The model using the 12th dataset suggestes that the age of the tree when diameter is zero is a negative value of 23.339. The coefficient of the diameter is 13.992 which suggests that a unit increase in the diameter can increase the age by 13.992. The R-squared value for this model is 75.1% which indicates that 75% of the variance in age can be explained by the diameter. The model validates the **positive relationship** between age and diameter. Using multiple imputation inferences by combining the rule.

	estimate	std.error	statistic	df	p.value
(Intercept)	-32.22728	22.899305	-1.407348	11.49464	0.1857860
diameter	14.74684	2.422864	6.086532	11.18404	0.0000735

The pooled results suggest that the base value of age is -32.23 (when diameter = 0). The coefficient of the diameter is 14.75 which suggests that a unit increase in the diameter can increase the age by 14.75. This observation is quite close to one made using the 12th dataset. Also, the diameter is significant in estimating the value of age.

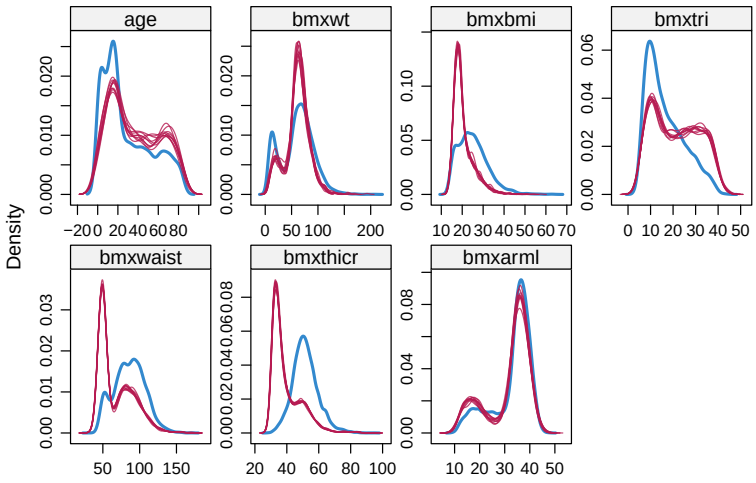
Multiple Imputations In Nhanes Data

The following variables contain missing values -

```
*age            *bmxbmi
*bmxtri         *dmdeduc
*bmxwaist       *bmxarml
*bmxthicr       *indfminc
*bmxwt
```

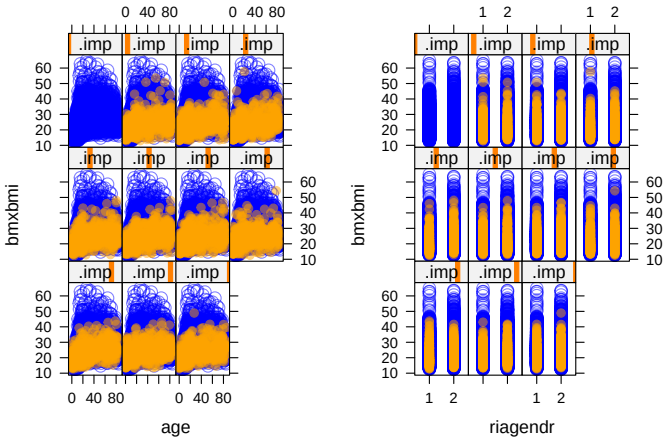
The dataset also contains variables which are factor with 2 or more levels. Thus for this analysis pmm or predictive mean matching will be used as the imputation technique. Also, the variables - "sdmvstra","sdmvpsu","ridageyr","wtmec2yr" are dropped from this analysis.

10 imputated datasets are created. The density plots of the datasets are shown below.



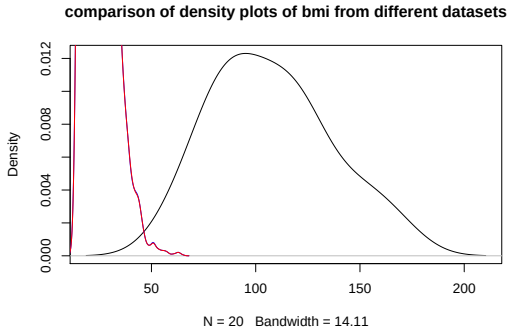
The density plot of the imputed values for age, bmsxarml are fairly similar to the observed. For bmxwt, though the shape is similar, there is difference in kurtosis. There seems to be some distortion for other variables like bmxbmi, there seems to be a certain amount of bias and the distribution is not very similar to the observed value.

The scatterplot of bmxbmi (BMI measurement) by age and bmxbmi by riagendr (gender) is observed below.

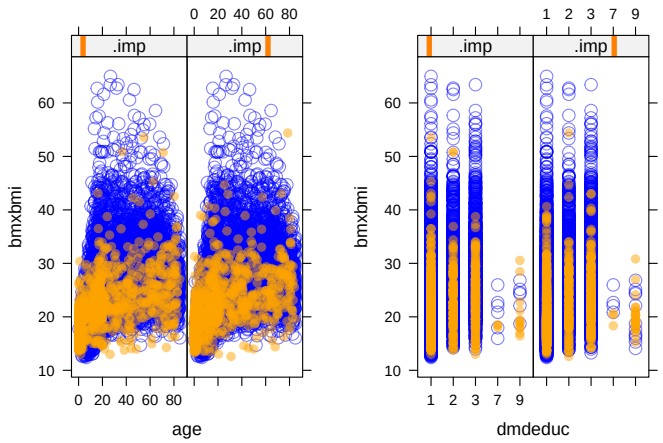


From the above graphs the 2nd and 8th imputed dataset seems like a good fit.

Comparing the density plots for the selected the datasets -



The imputated dataset (blue and red) seem to have similar distrubution. But the distribution is heavily right skewed as compared to normal distribution of the bmxbmi from the original dataset.



The impuated values seem quite similar to the original values as can be seen from the scatterplot above. This is can quantified by using the following model -

$$bmxbmi = \beta_0 + \beta_1 * age + \beta_2 * riagender + \beta_3 * ridreth2 + \beta_4 * indfminc + \beta_5 * dmdeduc + \beta_6 * bmxwt$$

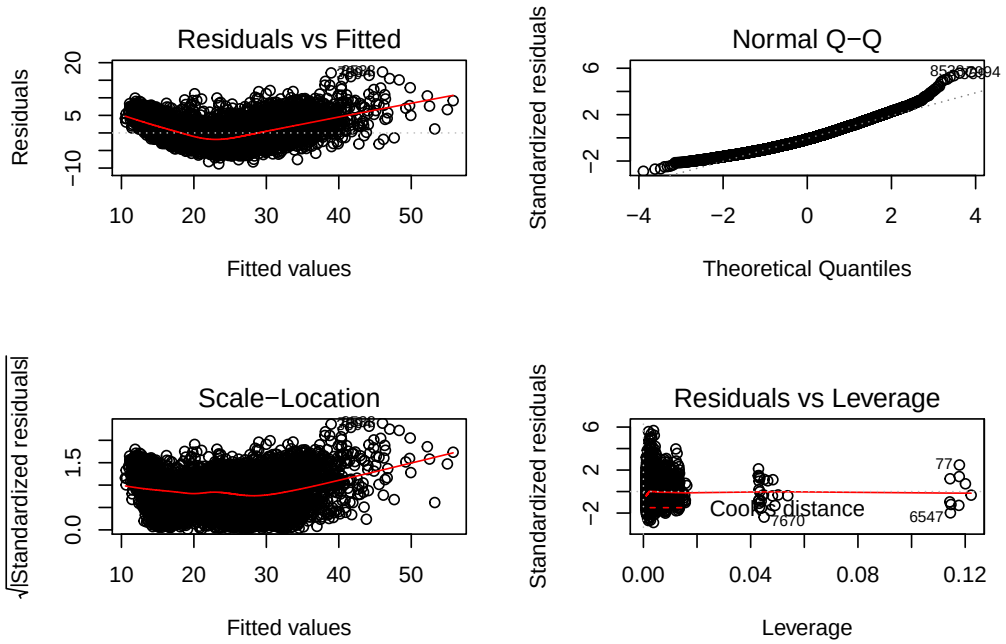
For this analysis the 2nd imputed dataset is used. A stepwise AIC shows that the above variables are significant. So, the next set of evalutions will be conducted on this model.

term	estimate	std.error	statistic	p.value
(Intercept)	10.6059539	0.1475618	71.8746380	0.0000000
age	-0.0076454	0.0016939	-4.5135593	0.0000064
riagendr2	2.1183553	0.0616606	34.3550904	0.0000000
ridreth22	0.0549064	0.0802219	0.6844312	0.4937186
ridreth23	0.7699825	0.0831204	9.2634625	0.0000000
ridreth24	-0.0969481	0.1752381	-0.5532365	0.5801137
ridreth25	0.8083114	0.1746570	4.6279926	0.0000037
indfminc2	0.0855022	0.1656163	0.5162668	0.6056794
indfminc3	0.1493827	0.1517670	0.9842894	0.3249968
indfminc4	-0.1538040	0.1588035	-0.9685173	0.3328093
indfminc5	-0.2503680	0.1583391	-1.5812142	0.1138604
indfminc6	-0.2191803	0.1499555	-1.4616354	0.1438723
indfminc7	-0.2417898	0.1585727	-1.5247878	0.1273434
indfminc8	-0.5099882	0.1650796	-3.0893469	0.0020114
indfminc9	-0.3100133	0.1862073	-1.6648828	0.0959672
indfminc10	-0.6416385	0.1995316	-3.2157245	0.0013053
indfminc11	-0.9332301	0.1471038	-6.3440229	0.0000000
indfminc12	-0.2461998	0.2966043	-0.8300612	0.4065238
indfminc13	0.0980599	0.2876723	0.3408735	0.7332059
indfminc77	-0.7520447	0.3499934	-2.1487394	0.0316788
indfminc99	-0.6149900	0.3621365	-1.6982273	0.0894957
dmdeduc2	-0.4245312	0.0977980	-4.3408970	0.0000143
dmdeduc3	-0.6682375	0.0882225	-7.5744566	0.0000000
dmdeduc7	2.5488624	1.0324572	2.4687343	0.0135756
dmdeduc9	1.8317550	0.6291515	2.9114689	0.0036052
bmxwt	0.2182446	0.0013125	166.2819540	0.0000000

The model shows that age is a significant variable. Along with age, dmededuc (education), weight, race

(Mexican American and Other Hispanic), gender and annual income levels (USD 45,000 to USD 54,999, USD 65,000 to USD 74,999 and USD 75,000 and Over) are also significant.

The model using the 2nd dataset suggestes that the bmi is 10.605954 at baseline where age = 0, gender = Male, race = Non-Hispanic white, education = Less than high school, income = USD 0 to USD 4,999 and weight = 0. The coefficient of the age is -0.007645 which suggests that a unit increase in the age can decrease the bmi by -0.007645. When gender = Female, the bmi increases by 2.118355. Education however shows a very interesting relationship with Bmi. At baseline, when education is less than high school, if there is a increase to level 2 (High school diploma (including GED) and level 3 (More than high school) the bmi reduces by -0.424531 and -0.668238 respectively. But in case a certain individual is in level 7 (they refused to answer this question) or level 9 (they don't know their level education) the bmi is higher by 2.548862 and 1.831755 respectively. The R-squared value is 81.16% which indicates that approximately 81% of the variance in bmi can be explained by this model.



Using the residual plots and summary plots we can infer the following:

- **Linearity**: There seems to be a curved pattern in the residual plot. This indicates that the model is not meeting the assumption of linearity.
- **Independence and Equal Variance**: Since the residual versus fitted plot indicates a evident pattern there is an indication of heteroskedasticity in the dataset or the model. This assumption is not met.
- **Normality**: The Q-Q plot is used to interpret Normality. The relationship is approximately linear with the exception of a few data points.
- **Leverage Plot**: The points are well inside the cooks. However, the points seem to be accumalated on one left side of the graph with the exception of a few points towards the right.

Thus, it is evident that linear model is not a good fit for this model.

Using multiple imputation inferences by combining the rule.

	estimate	std.error	statistic	df	p.value
(Intercept)	10.5746255	0.1557835	67.8802808	750.43883	0.0000000
age	-0.0071896	0.0018716	-3.8413588	245.55702	0.0001558
riagendr2	2.1292854	0.0623910	34.1280906	5108.81478	0.0000000
ridreth22	0.0853316	0.0827340	1.0313969	1660.95285	0.3025049

	estimate	std.error	statistic	df	p.value
ridreth23	0.7510024	0.0859689	8.7357479	1417.65921	0.0000000
ridreth24	-0.0685287	0.1869214	-0.3666176	524.09292	0.7140522
ridreth25	0.8256133	0.1849582	4.4637828	630.55539	0.0000095
indfminc2	0.1338255	0.1743588	0.7675295	848.70913	0.4429802
indfminc3	0.1596279	0.1624984	0.9823354	497.68517	0.3264119
indfminc4	-0.1469826	0.1650502	-0.8905330	1311.13135	0.3733432
indfminc5	-0.2431116	0.1667714	-1.4577541	774.31037	0.1453138
indfminc6	-0.2163269	0.1551536	-1.3942758	1544.03906	0.1634349
indfminc7	-0.2167033	0.1624385	-1.3340640	2549.33587	0.1823020
indfminc8	-0.5346892	0.1695303	-3.1539452	2254.82542	0.0016320
indfminc9	-0.3031526	0.1937120	-1.5649653	1270.66672	0.1178400
indfminc10	-0.6541024	0.2131791	-3.0683228	504.01248	0.0022685
indfminc11	-0.8980838	0.1520553	-5.9062962	1610.91969	0.0000000
indfminc12	-0.3239087	0.3344142	-0.9685854	176.96452	0.3340740
indfminc13	0.2823224	0.3089348	0.9138575	465.79826	0.3612645
indfminc77	-0.5947688	0.3919936	-1.5172922	241.36415	0.1305016
indfminc99	-0.4614131	0.3905580	-1.1814201	369.21842	0.2381963
dmdeduc2	-0.3909726	0.1143833	-3.4180905	119.88969	0.0008621
dmdeduc3	-0.6552471	0.1003119	-6.5320946	161.11537	0.0000000
dmdeduc7	0.6488308	1.5145542	0.4283972	34.33142	0.6710378
dmdeduc9	0.9327853	1.4418030	0.6469575	16.52569	0.5265444
bmxwt	0.2180011	0.0014244	153.0459430	344.56678	0.0000000

The pooled results suggest that the base value of bmi is 10.57 (at baseline of age = 0, gender = Male, race = Non-Hispanic white, education = Less than high school, income = USD 0 to USD 4,999 and weight = 0.). The coefficient of age is -0.01 which suggests that a unit increase in age can decrease the bmi by -0.01. In case of gender, it seems that when gender is female the bmi increaes by 2.13. These observation is quite close to one made using the 2nd dataset.