Case3 Part 1: Missing Data

Sonia Xu & Vinai Oddiraju

November 6, 2017

Preliminary Analysis

Preliminary analysis shows that the missing data is Missing at Random. Exploring the relationships between the features shows that PHC and $GREEN$ are the most correlated.

Dataset

The dataset is relatively simple. The response variable is labeled Y and records whether or not malaria was present in a patient's blood, with 1 being yes and 0 being no. *Age* denotes the age of the subject. The ages were distributed across the values 1, 2, 3, and 4. Given the discrete nature of these levels, we plan to treat *Age* as a categorical variable. The subjects were distributed uniformly in age, with the most frequent factor being 1. The table below displays the distribution.

```
table(gambia$AGE) %>% as.data.frame() %>% kable(col.names = c("Age", "Freq"))
```

Age	Freq
1	238
2	188
3	199
4	180

PHC and $BEDNET$ are binary variables; PHC records whether or not a public health center is nearby, and $BEDNET$ records whether or not a mosquito net was present on the child's bed at home.

$GREEN$ is a variable that shows the percentage of greenery around the child's home. In our sample, it has 5 distinct levels, with the factor 40.85 as the most frequently occurring level. The table below shows the distribution.

```
table(gambia$GREEN) %>% as.data.frame() %>% kable(col.names = c("Green", "Freq"))
```

Green	Freq
28.85	32
33.65	84
40.1	69
40.85	586
47.65	34

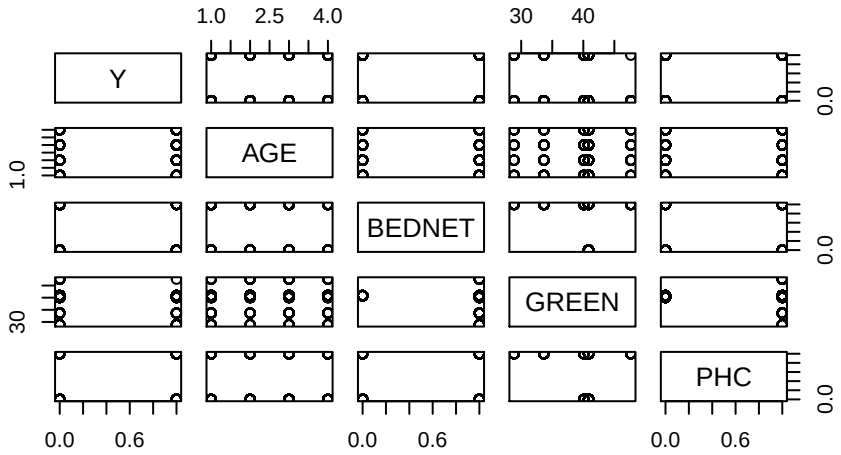
Missing Data: MAR

Initial exploratory data analysis reveals that the feature bednet contains missing data for 317 out of the 805 observations. To determine whether the data is Missing at Random or Missing Completely at Random, features from the missing dataset were compared in porportion to the features from the original dataset via

quantiles. If the dataset were Missing Completely at Random, the quantiles would match for each feature. However, as the table of quantiles below shows, the missing dataset does not match in quantiles for the features *Age* and *PHC*.

Feature	0%	25%	50%	75%	100%
Missing Age	1	2	3	4	4
Age	1	1	2	3	4
Missing Green	28.85	40.85	40.85	40.85	47.65
Green	28.85	40.85	40.85	40.85	47.65
Missing PHC	0	1	1	1	1
PHC	0	0	1	1	1

Correlation Exploration

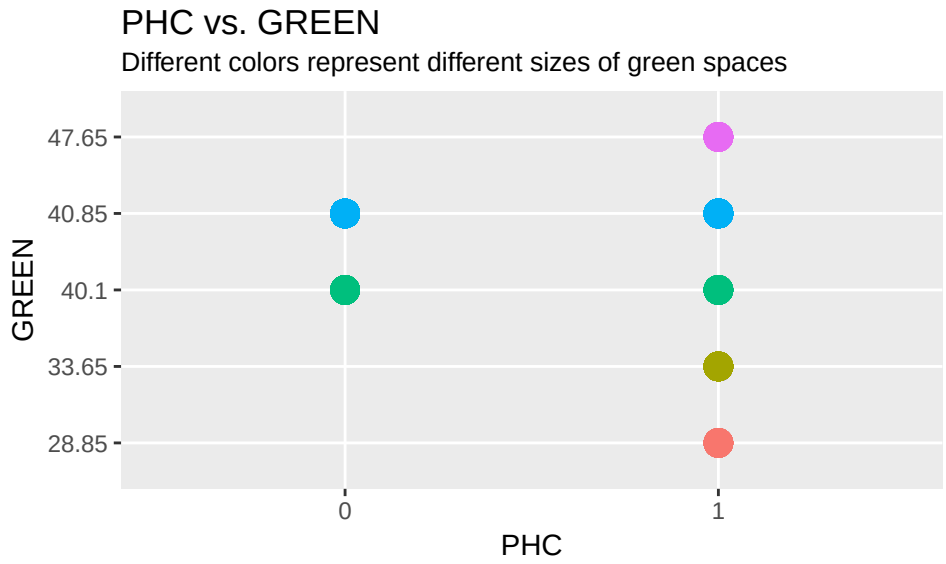


	Y	AGE	BEDNET	GREEN	PHC
Y	1.0000000	0.1750289	-0.0140220	-0.0254663	-0.1259620
AGE	0.1750289	1.0000000	-0.0530640	0.0563457	-0.1033950
BEDNET	-0.0140220	-0.0530640	1.0000000	-0.1818298	0.2912400
GREEN	-0.0254663	0.0563457	-0.1818298	1.0000000	-0.2211094
PHC	-0.1259620	-0.1033950	0.2912400	-0.2211094	1.0000000

Visually looking at the correlation plots for each of the features from the non-missing data, none of the features appear highly correlated. When comparing the correlation matrix, the largest correlations are between the features, *PHC* and *GREEN* (-0.221) and PHC and *BEDNET* (0.291). Similarly, looking at the missing data, the largest correlation is between *PHC* and *GREEN* (-0.137).

Relationship Between PHC and GREEN

Looking at the plot below, observations that live near green space measurements of 28.85, 33.65, or 47.65 are more likely to live close to a public health center, whereas obersevations that live near green space measurements of 40.1 or 40.85 do not. Perhaps, green spaces of 40.1/40.85 units are not close to the city, so they are further away from public health centers.



Next Steps

Future plans include exploring different models that can: 1. Imputate the missing data 2. Predict the presence of malaria

Imputating the Missing Data

In terms of imputating the missing data, we plan on clustering the data set on its other features via k-means, and then using these groups to predict $BEDNET$.

Predicting the Presence of Malaria

The presence of malaria is a binary response, so we plan on using a logit model to fit the final dataset.

Since we plan to build a predictive model, we plan to separate our data into a training set and a test set. This will help us test the validity of our model.