case3-2

Sarah Zimmermann, Wuming Zhang, Adam Wood

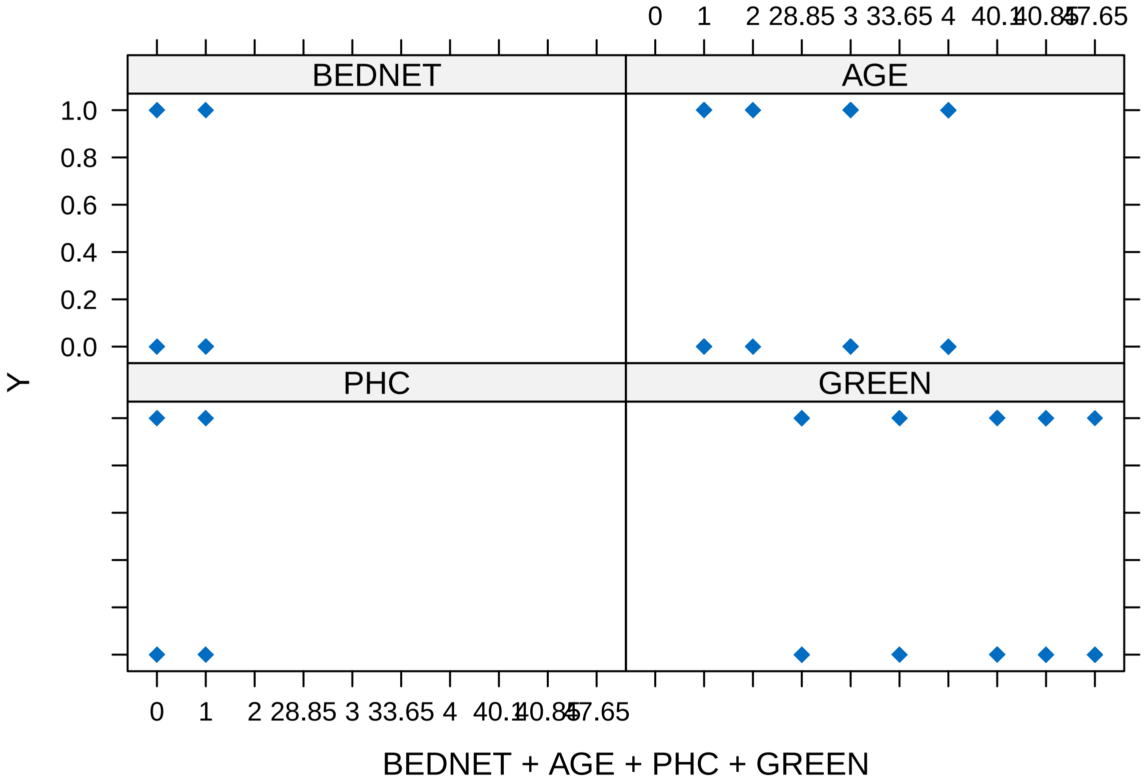
November 9, 2017

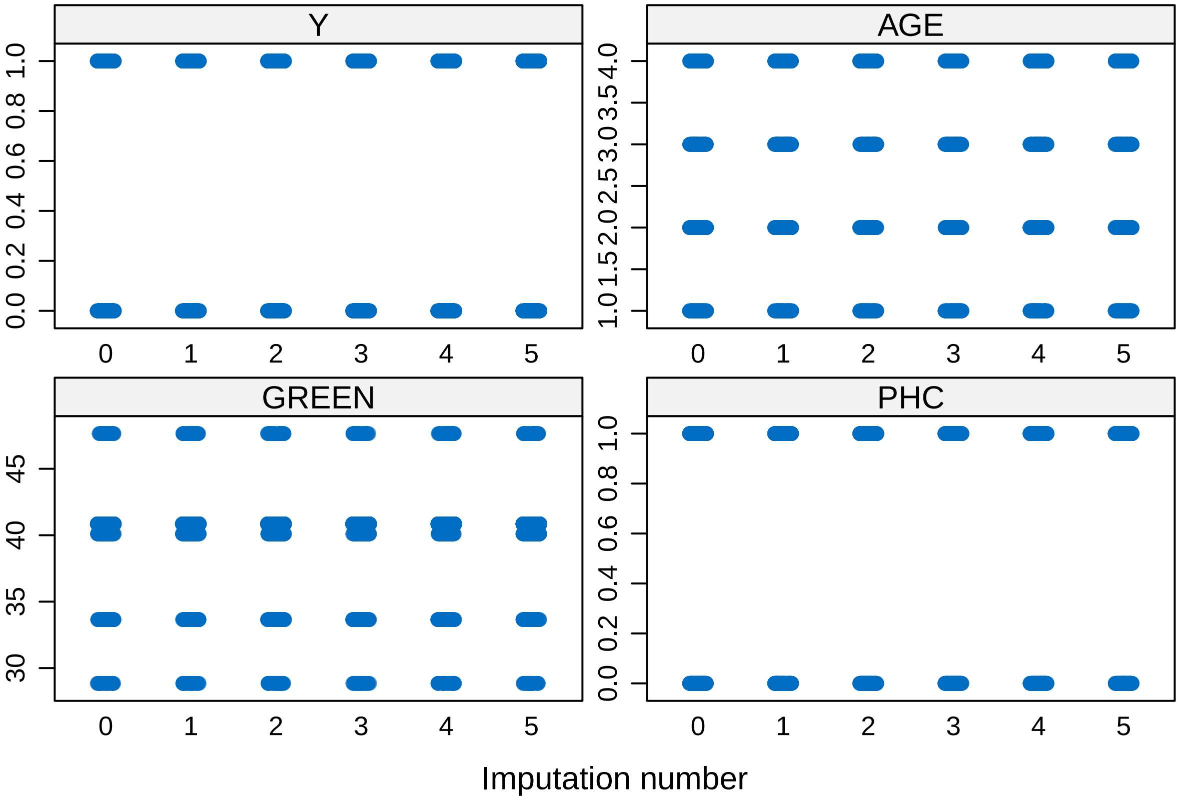
Data

Multiple Imputation/Chained Regression

The multiple imputation method using chained regression was utilized to impute missing data for the bednet variable in the *gambia* data set. The *mice* package provides several functions for easy implementation of multiple imputation. Functions **mice()**, **complete()**, and **with()** allow us to apply multiple imputation for missing data, replace missing data with imputed data, and use the imputed data set to create a model, respectively.

Inspection of Missing Data





In order to ensure the multiple imputation process functioned as expected, we analyze the distributions of the imputed data sets versus the originally observed data. The scatterplot of malaria presence vs. each of the predictor variables fails to provide any insight, and in cases where fewer variables are binary or few in unique value, it might be more useful. The strip plot also yields very little with respect to the relationship between imputed and observed data, but does illustrate that the *bednet* variable is the only variable with missing data (previously known). The density plot provides us with the most meaningful information regarding the imputed data sets and the observed data.

In our density plot, the 5 magenta density curves represent the 5 imputed data sets, while the blue density curve represents the density of observed data. In determining whether the imputed data for missing values in the bednet variable are plausible, we want to determine if the imputed data density curves follow a similar curvature to the observed data density curve. As illustrated, the 5 magenta curves follow a similar shape to the observed data curve. It is of note to mention that all 5 imputed data sets densities fall below the density of the observed data for the number of individuals without a bednet. 2 imputed data sets fall below the observed density for individuals with a bednet, 2 above the observed density, and 1 with a density of individuals with a bednet very close to the observed. It appears that each of the 5 imputed data sets seems plausible and have similar shape to the blue observed curve.

Model

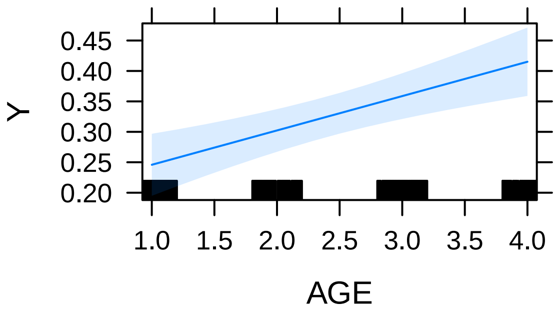
To alleviate dependance created by our choice in the **mice()** function seed selection, we use the multiple imputation method to create 50 imputed data sets. The density plot for these 50 data sets follow very similar shape to the observed density curve, and so we proceed with model fitting.

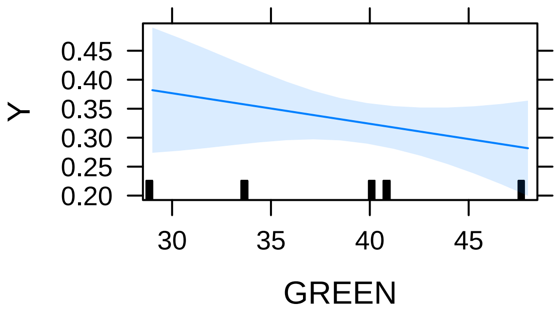
The summary of our model fit with 50 imputed data sets indicates that the variables *age* and *PHC* (presence of a health clinic in the village) are statistically significant with p-values under 0.01. Bednet use was not statistically significant, with a p-value over 0.1. This provides insight into the effects of bednet use on malaria presence in the individual, and indicates that bednet use is not as impactful on predicting cases of malaria as previously thought. Under the current model, age and presence of a health clinic in the area are the strongest variables in predicting malaria cases.

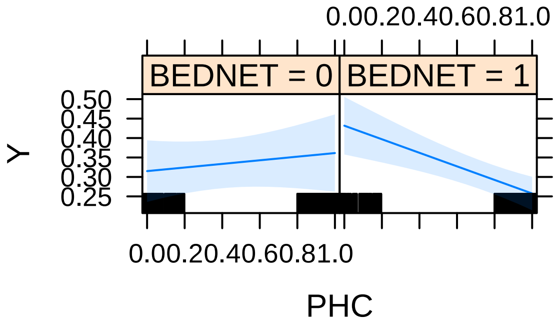
Model Diagnostics

Interaction Analysis

Since Malaria is a mosquito-borne disease and we suspect that the presence of bed net would be an influential factor. Some other factor, such as the presence of a public clinic, could potentially have some interaction with the bed net variable. In another word, the effect of public clinic to Malaria appearance could possibly vary with whether the bed net is used. We picked one of the imputed dataset and fit a logistic regression model with an interaction term of the bed net and public clinic. The resulting plots below show that the slopes of $Y\ PHC$ are quite different for different $BEDNET$ variable values.


AGE effect plot


GREEN effect plot


BEDNET*PHC effect plot

Model Selection

We then start off the model selection process by first construct a naive model with intercept only. We then compare our full model, $Y = \alpha_1 BEDNET + \alpha_2 AGE + \alpha_3 PHC + \alpha_4 GREEN + \alpha_5$, with this naive model using the likelihood ratio test. From the p-value, we can observe that the full model is not significant compared to the naive model. We also use similar technique to compare a model with an extra interaction variable but we found that it is not statistically significant as well. This appears to be different from what we expected, and we are planning to use cross-validation technique to further verify this finding. For now, we choose to use the model with all variables and an additional interaction term as our best model, that is, $Y = \alpha_1 BEDNET + \alpha_2 AGE + \alpha_3 PHC + \alpha_4 GREEN + \alpha_5 BEDNET * PHC + \alpha_6$.

residual v fitted and corresponding write up, model selection and the pool.compare p-value, explanation of estimates and thier implication on the study

Model Interpretation

Future Revisions

variable selection and cross validation and testing for interaction terms

Credits

References

https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/