case3-2

Sarah Zimmermann, Wuming Zhang, Adam Wood

November 9, 2017

Introduction

The data are from a study malaria presence in Gambia. Each observation corresponds to one child who has been tested for the presence of malaria. We are interested in the factors predictive of the malaria parasits found in the blood for n=805 children. The goal of the analysis is to perform inferences on the impact of age, presence of bednet, presence of clinic, and amount of surrounding greenery on a child having malaria, where greenery is a measure of how much vegetation is around the child's village based on satellite images.

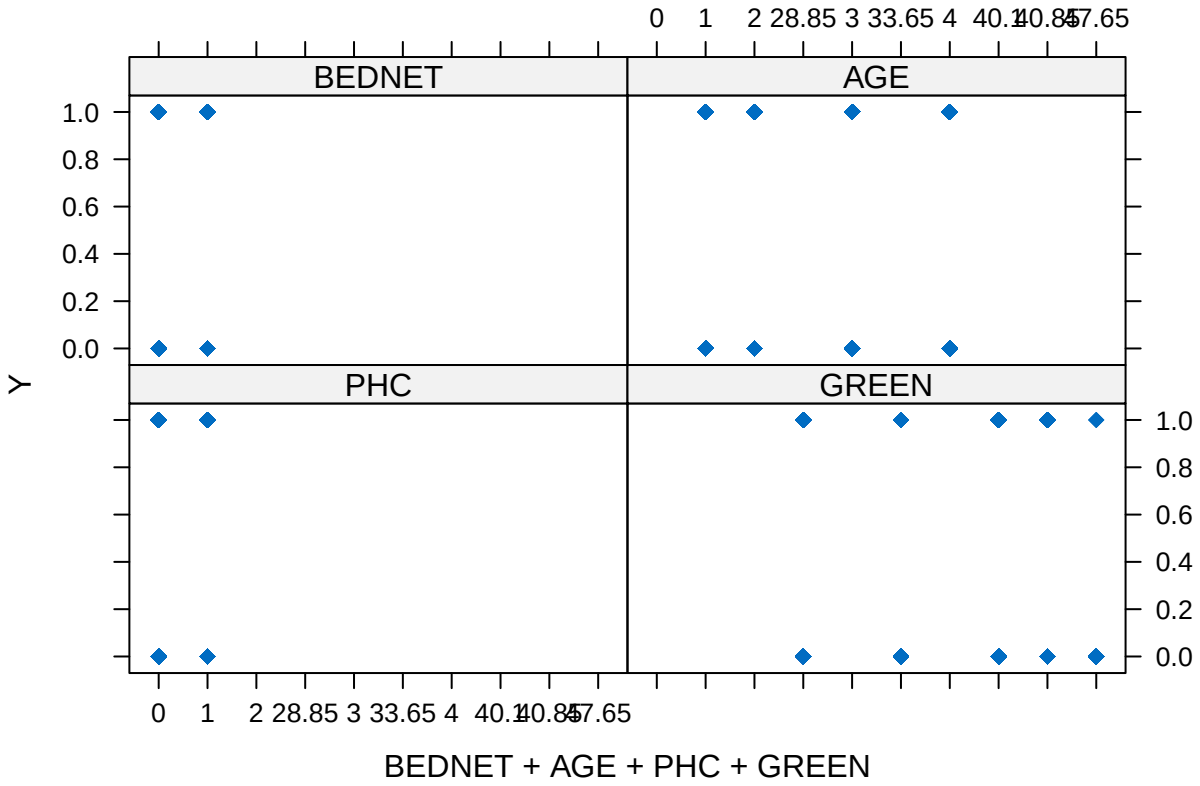
This week the team focused on imputation strategies and prelimnary model creation and diagnostic. We explored multiply imputation and fit general linear models with the imputed data sets. Our diagnoses and interpretations are found below.

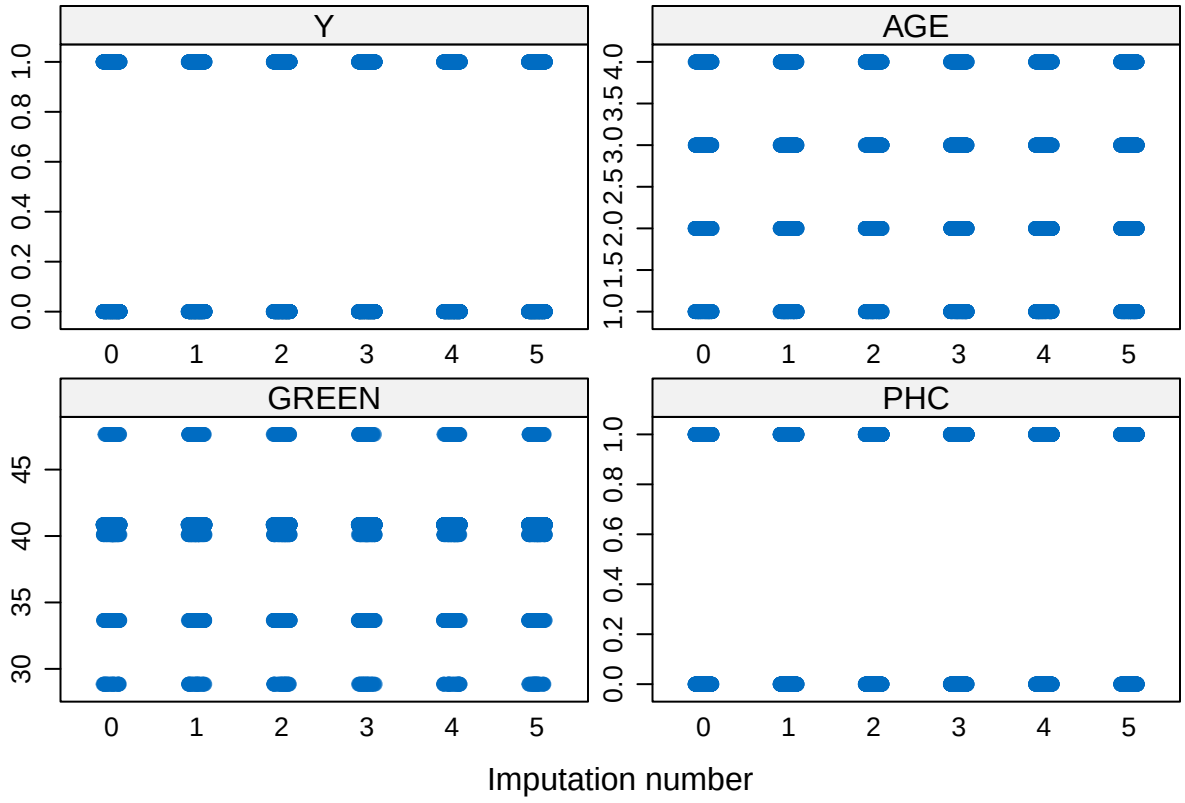
Multiple Imputation/Chained Regression

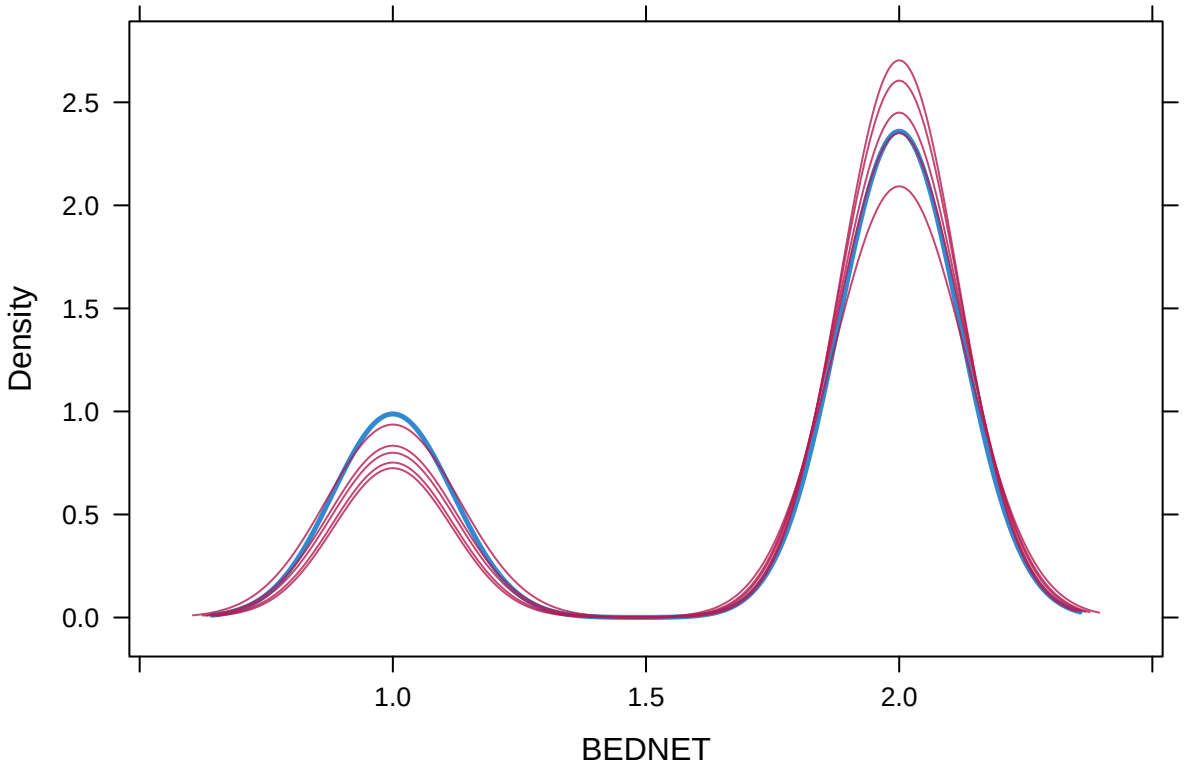
As an overview for our work, the multiple imputation method using chained regression was utilized to impute missing data for the bednet variable in the *gambia* data set. The *mice* package provides several functions for easy implementation of multiple imputation. Functions **mice()**, **complete()**, 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.

We imputed the data 5 times and iterated 50. With this new data set we must inspect if the imputation functioned as expected.

Inspection of Missing Data




Y
AGE
GREEN
PHC
0.0 0.2 0.4 0.6 0.8 1.0
1.0 1.5 2.0 2.5 3.0 3.5 4.0
30 35 40 45
0 1 2 3 4 5
Imputation number

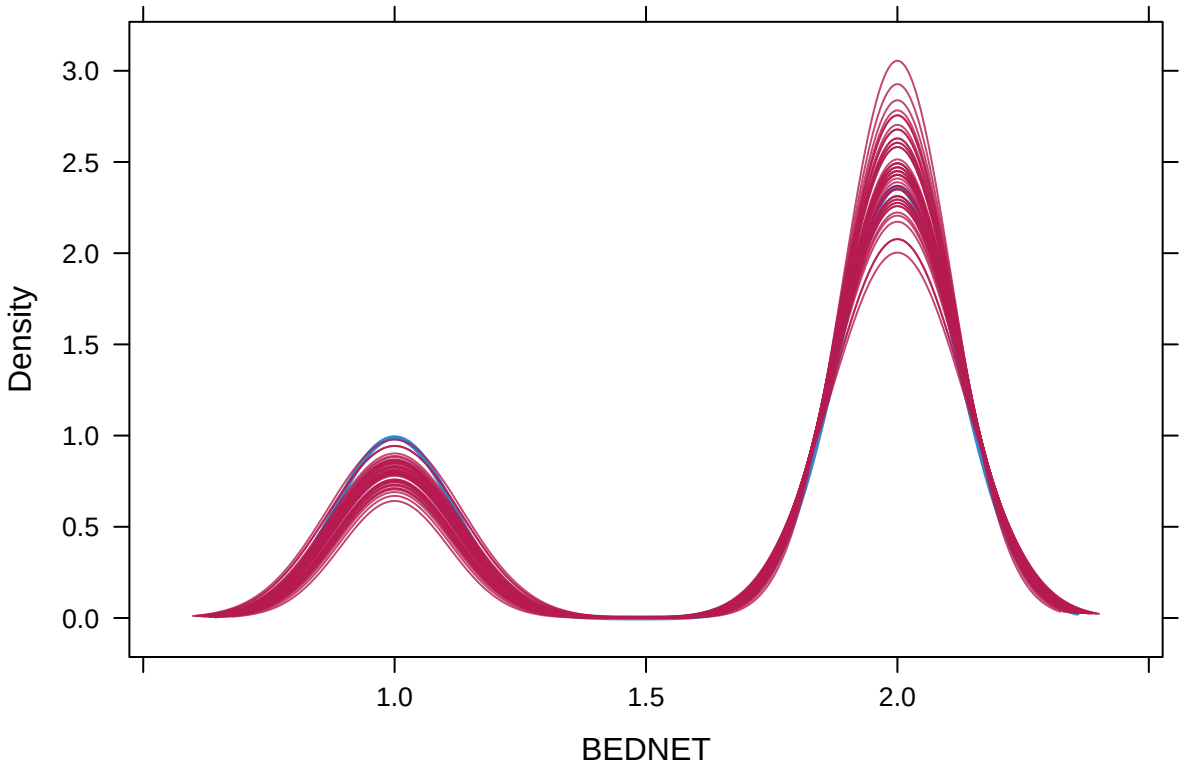


We analyze the distributions of the imputed data sets versus the originally observed data. The scatterplot of malaria presence vs. each of the predictor variables fails to provide any insight, and in cases where fewer variables are binary or few in unique value, it might be more useful. The strip plot 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.

We continue on my modeling the imputed data.

Model



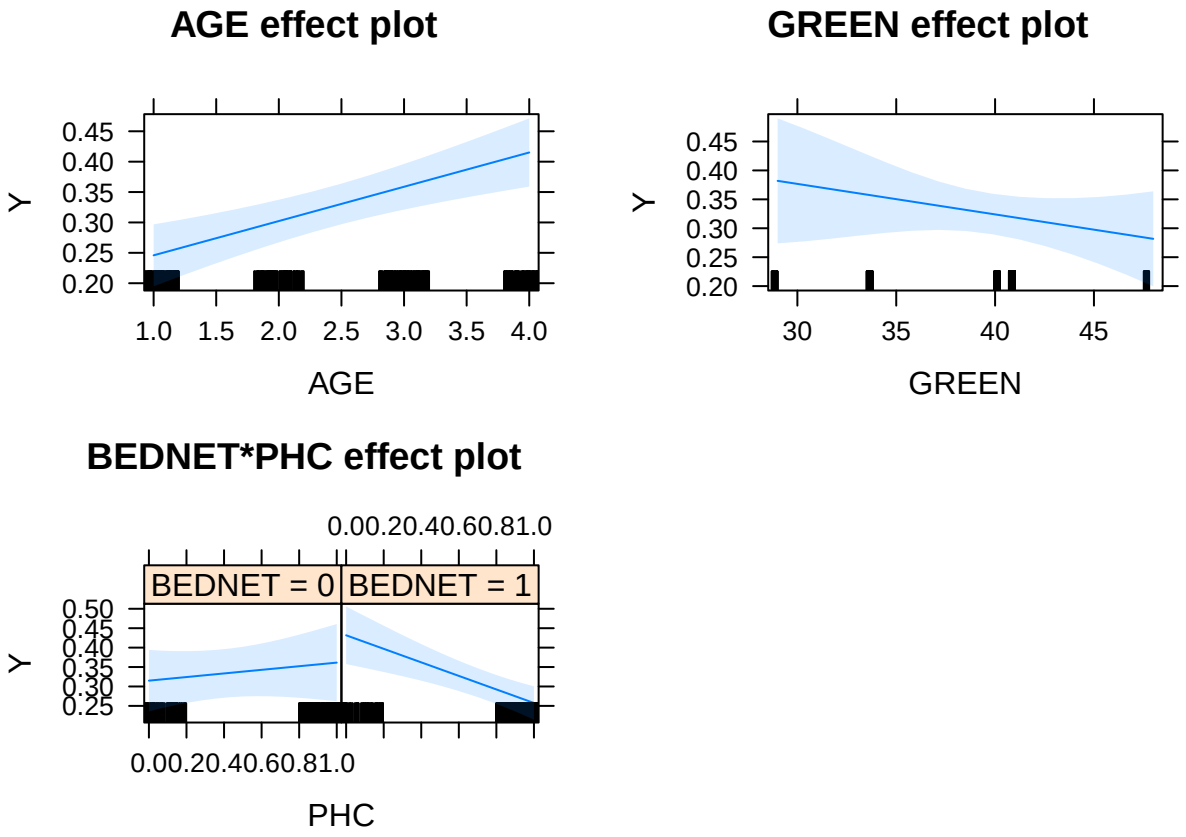
To alleviate dependance created by our choice in the **mice()** function seed selection, we use the multiple imputation method to create 50 imputed data sets. The density plot for these 50 data sets follow very similar shape to the observed density curve, and so we proceed with model fitting.

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 and Selection

Interaction Analysis

Since Malaria is a mosquito-borne disease, 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 other words, the effect of having a public clinic to malaria appearance could possibly vary with whether a bed net is used. We picked one of the imputed datasets and fit a logistic regression model with an interaction term between 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.

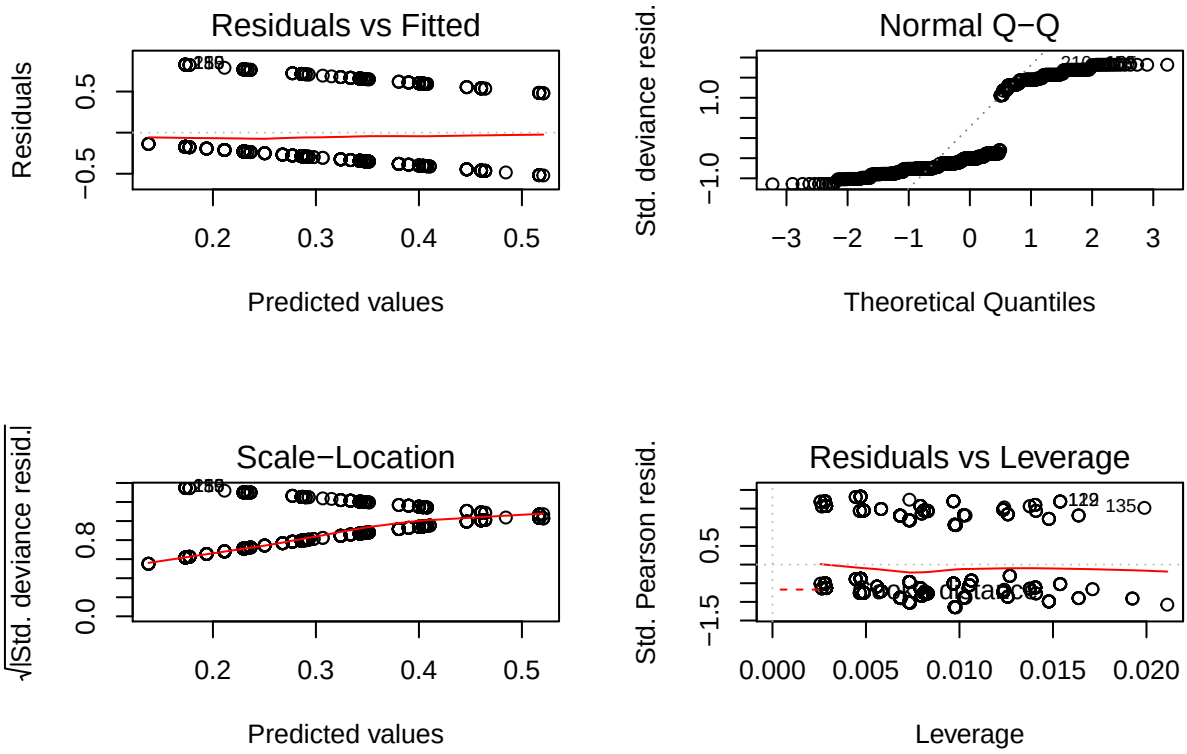


Model Selection

We then start off the model selection process by first constructing 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 a likelihood ratio test. From the p-value above, we can observe that the full model is not significant compared to the naive model. We also use a 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 to further verify this finding. For now, we choose to use the model with all variables and an additional interaction term as our most current model, that is, $Y = \alpha_1 BEDNET + \alpha_2 AGE + \alpha_3 PHC + \alpha_4 GREEN + \alpha_5 BEDNET * PHC + \alpha_6$.

Model Residuals and AIC

```
par(mfrow=c(2,2))
plot(model)
```




```
#stepAIC starting with the intercept only model
stepAIC(glm(Y~1, data = completedData), scope = ~ BEDNET + AGE + PHC + GREEN + BEDNET:PHC)
```

```
## Start:  AIC=1046
## Y ~ 1
##
##          Df Deviance    AIC
## + AGE     1   169.18 1034.8
## + PHC     1   170.51 1041.1
## <none>        171.98 1046.0
## + BEDNET  1   171.89 1047.6
## + GREEN   1   171.97 1048.0
##
## Step:  AIC=1034.77
## Y ~ AGE
##
##          Df Deviance    AIC
## + PHC     1   167.50 1028.8
## <none>        169.18 1034.8
## + BEDNET  1   169.08 1036.3
## + GREEN   1   169.16 1036.7
## - AGE     1   171.98 1046.0
##
## Step:  AIC=1028.75
## Y ~ AGE + PHC
##
```

```
##          Df Deviance    AIC
## <none>        167.50 1028.8
## + GREEN   1   167.36 1030.1
## + BEDNET  1   167.48 1030.7
## - PHC     1   169.18 1034.8
## - AGE     1   170.51 1041.1
```

```
##
## Call:  glm(formula = Y ~ AGE + PHC, data = completedData)
##
## Coefficients:
## (Intercept)          AGE          PHC
##     0.24241      0.05420     -0.09622
##
## Degrees of Freedom: 804 Total (i.e. Null);  802 Residual
## Null Deviance:       172
## Residual Deviance: 167.5     AIC: 1029
```

The residuals vs. fitted plot illustrates a dichotomous trend amongst residuals, with two separate, decreasing lines of residuals on both sides of the horizontal zero line. In evaluating the residuals for a logistic regression model, this is not surprising. If the observed value is 0, we will have negative residuals because we will predict higher than 0, and vice versa for observed values at 1.

The **stepAIC()** function allows us to evaluate our model using stepwise variable selection, with AIC as our criterion. The function's output, run over our current model, determined that the model with the lowest AIC is: $Y = \alpha_1 AGE + \alpha_2 PHC$. In future revisions, we will evaluate the consequences of using AIC as the primary variable selection criterion, as well as the consequences of variable selection processes in the context of this study.

Model Interpretation

The model after selection gives insight into the effect of age and presence of clinics on malaria. According to the model, all else held constant an increase in age will increase the log odds of malaria by .054. All else held constant the presence of the clinic in a village decreases the log odds of malaria by .096.

The output overall made sense to the team. According to the World Health Organization children under 5 in high-transmission areas of the world are the most vulnerable group. In 2015, over 69% of malaria deaths worldwide occurred in children under five years of age (http://www.who.int/malaria/areas/high_risk_groups/en/). The model indicating age is significant in malaria in child makes sense as it is backed by WHO and other health organizations. Additionally, the presences of clinics decreasing the log odds of malaria is logical because with the presence of medical personnel villages in the surrounding area have access to treatment and education for example which would logically help in prevention and care.

Although a fair model we will continue to make improvements for future models.

Future Revisions

In future revisions of this analysis, we look to examine the effects of choosing AIC as the primary variable selection criterion, as well as the effect of using a stepwise variable selection technique has on our analysis in this setting. We aim to utilize other techniques for model validation, i.e. cross validation, to further solidify our model selection. Other interaction terms will be considered and tested to determine their contributions or lack thereof to our current model.

Credits

This was a team effort. Adam contributed most heavily to the write up, model diagnostics (residuals and AIC), and evaluation of our model with imputed data using plots. Sarah wrote the code for the imputation, imputation diagnostics, and model as well as interpreting the final model after model selection.

References

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

http://freakonometrics.hypotheses.org/8210