# report2

*Nathaniel Brown, Huijia Yu, Angie Shen*

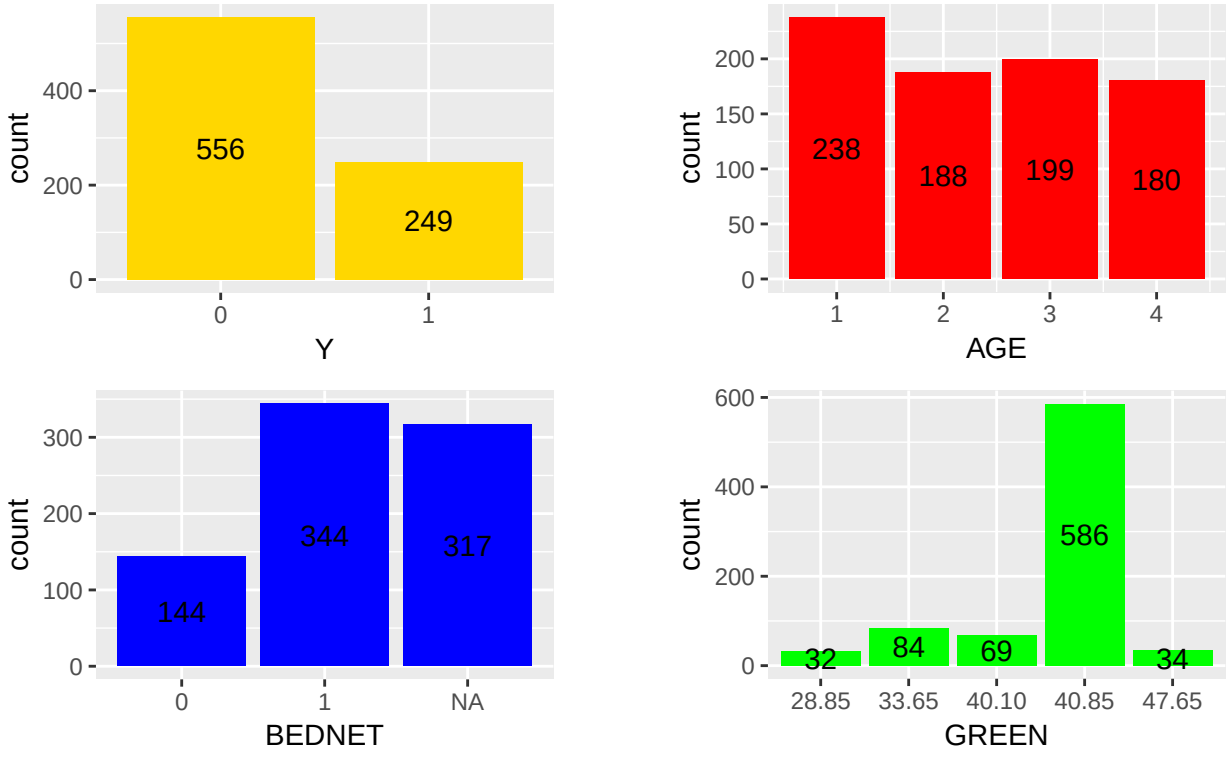
*November 6, 2017*

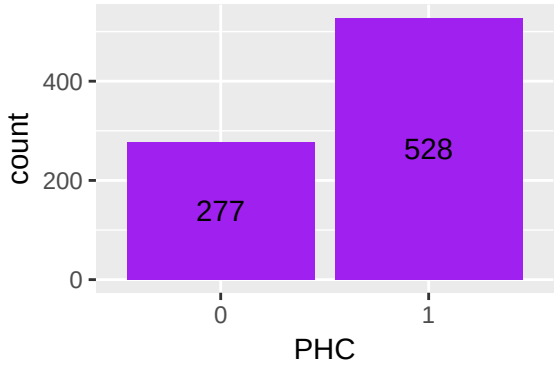
## Introduction

We are interested in predicting whether malaria parasites will be found in a child's blood based on his/her age, bednet use, the amount of greenery in his/her village, and the presence of a health clinic. However, almost 40% of the observations do not have bednet use reported, so removing those observations would cause us to lose too much information. Instead, we will impute bednet values to predict the outcome.

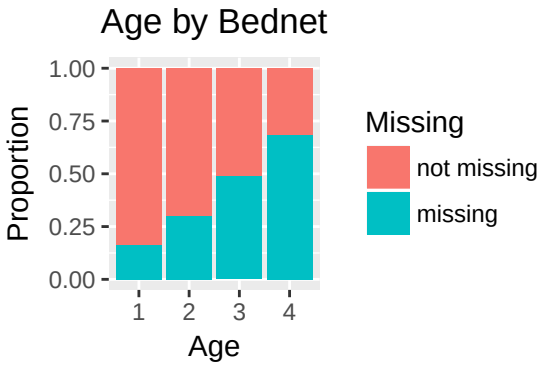
## Exploratory Data Analysis

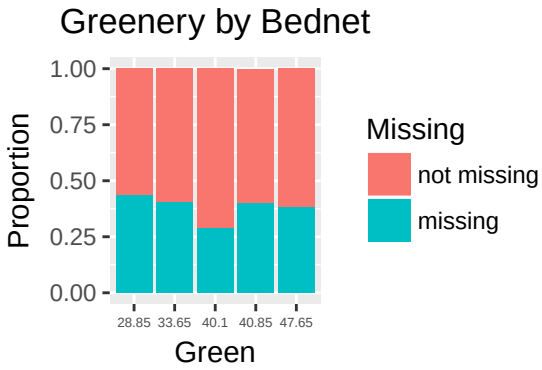
In the graphs below, we illustrate the frequency of each category within each the dataset:

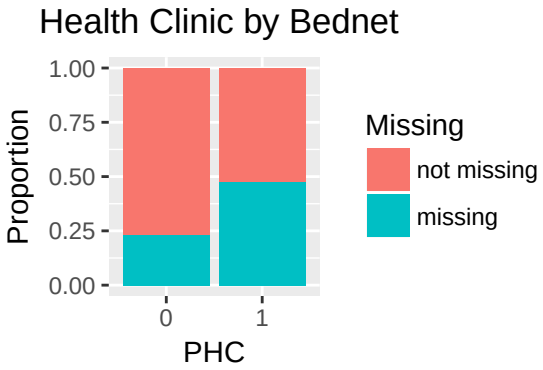


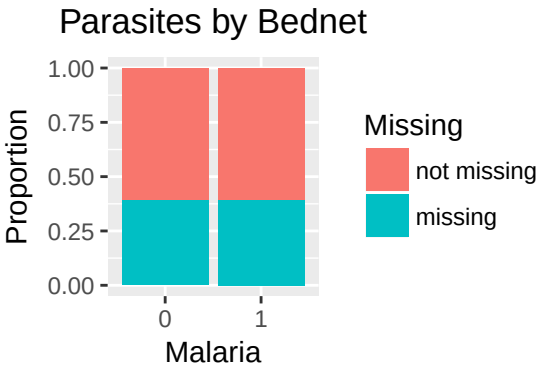



Below, we investigate the proportions of missing bednet responses by each level of the predictors.

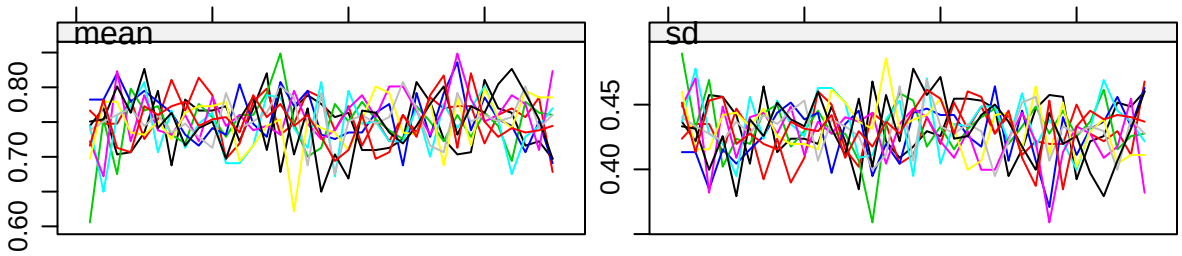


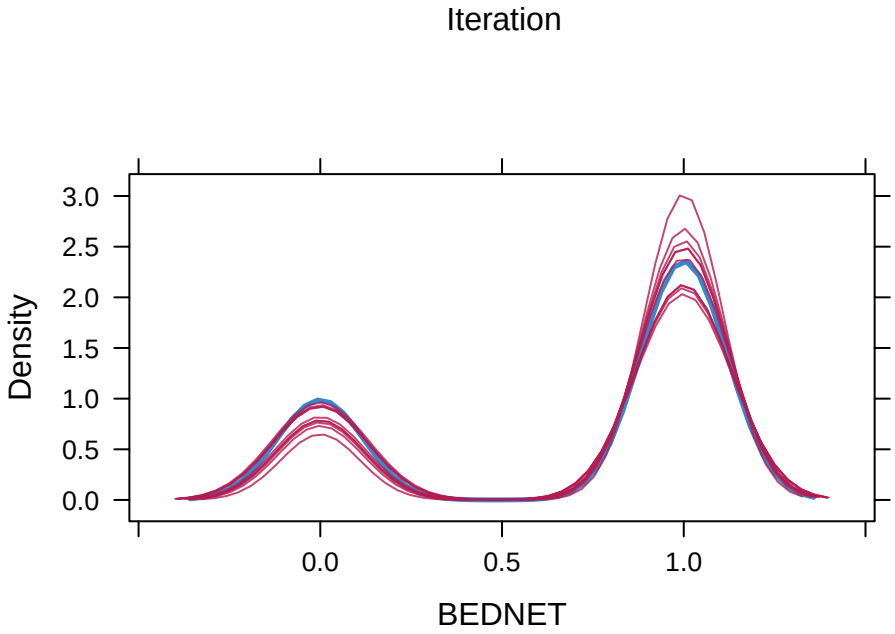



## Multivariate Imputation by Chained Equations (MICE)

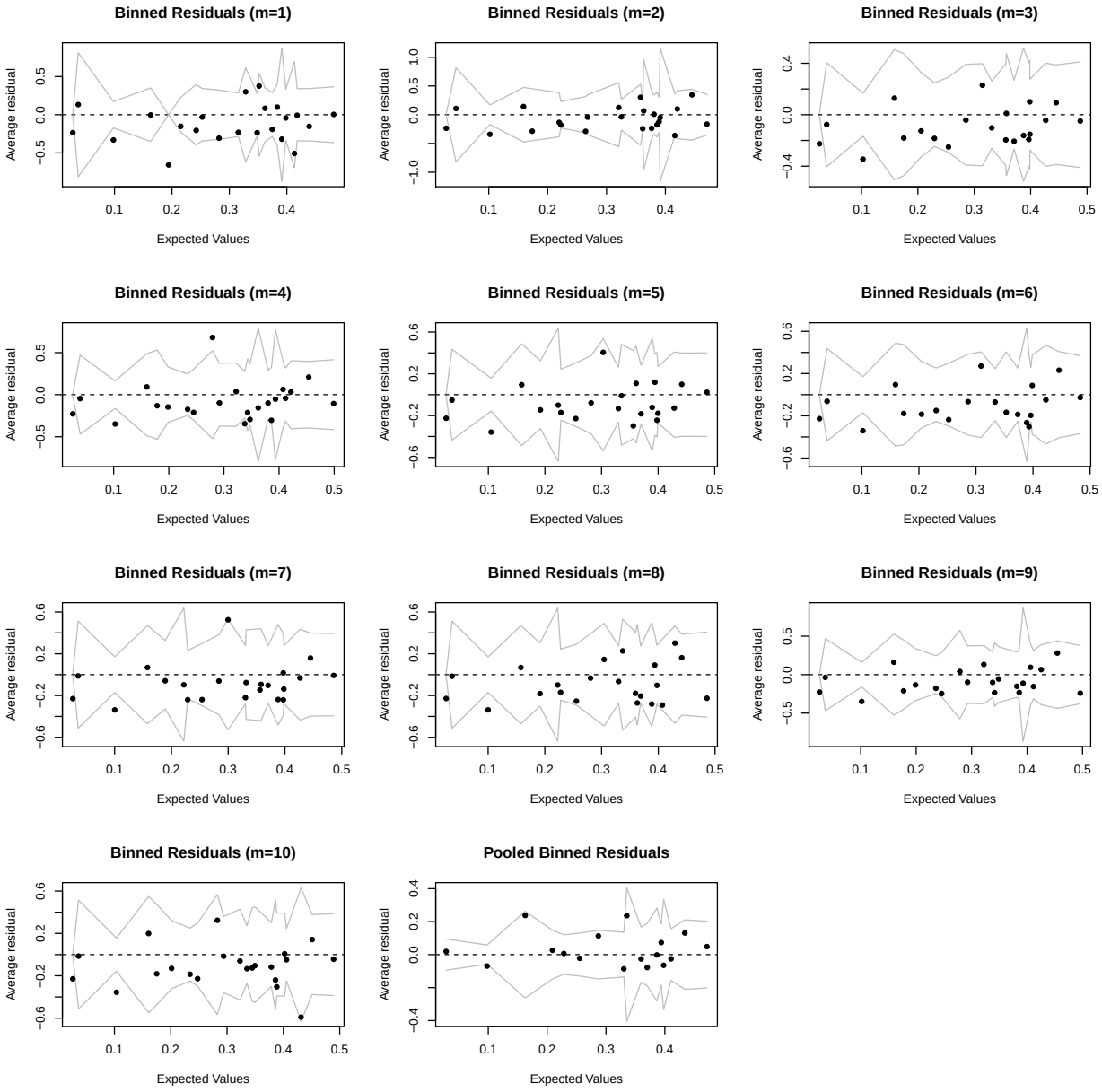



The trace lines appear to be stationary and free of trends, indicating convergence. We can see that the density distribution of each of the imputed datasets (in red) is congruent with the original one (in blue).

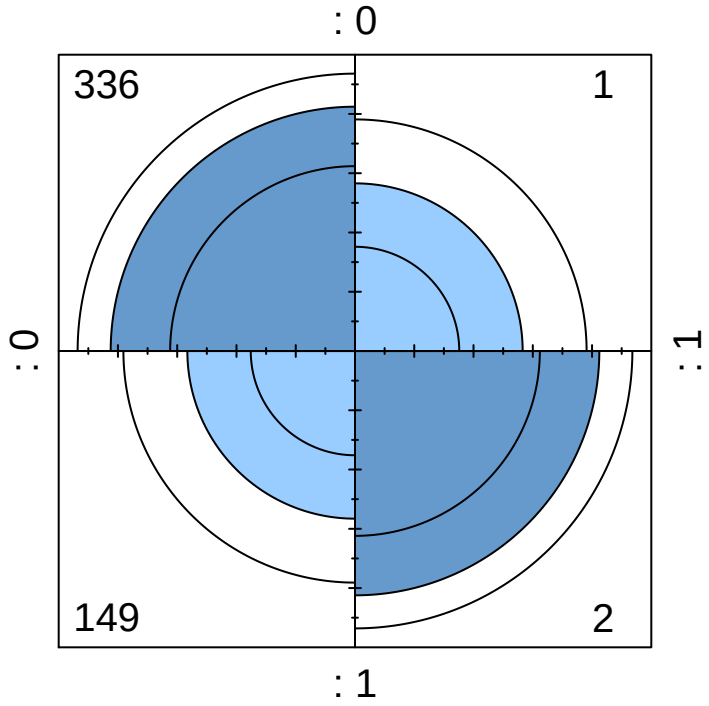
# Predicting Malaria Parasites

To predict whether someoneone will have malaria parasites, and to calculate the effect of each predictor (age, surrounding greenery, and accessible public health clinic), we build a logistic regression model on each generated dataset individually, and the pooled data. Below, we display the binned residual plots for each logistic regression model:

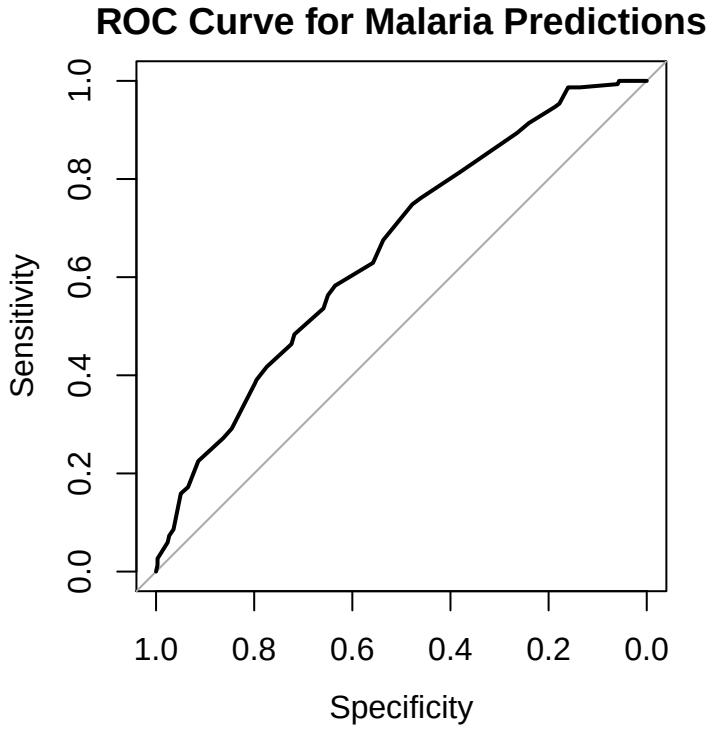



Neither the pooled nor the individual residuals display any trends or changes in variance, which means this model is appropriate and meets the assumptions for constant variance of the logit probabilities.

We can also check the accuracy of the model by looking at the confusion matrix and ROC curve for the pooled model.




The model is correct most of the time, but it has a lot of false negatives.




The ROC curve shows us that the predictions are not accurate. We can see whenever we observe a high true positive rate (sensitivity), there must also be a low true negative rate (specificity), and vise versa. for high sensitivity (true positive rate), we must have low specificity (true negative rate). The area under the curve is

0.6561, which is not much better than a random predictor at 0.5 (the dotted line on the plot).

## Inference on Coefficients

| | est | se | t | df | Pr(>\|t\|) | lo 95 | hi 95 | nmis |
|---|---|---|---|---|---|---|---|---|
| Intercept | -0.5775159 | 0.4508644 | -1.2809083 | 610.52067 | 0.2007122 | -1.4629492 | 0.3079174 | NA |
| AGE = 2 | 0.5097197 | 0.2270284 | 2.2451805 | 791.54172 | 0.0250319 | 0.0640708 | 0.9553686 | NA |
| AGE = 3 | 0.6854469 | 0.2210406 | 3.1009999 | 792.40345 | 0.0019970 | 0.2515525 | 1.1193413 | NA |
| AGE = 4 | 0.8033803 | 0.2251422 | 3.5683244 | 789.26637 | 0.0003810 | 0.3614320 | 1.2453286 | NA |
| GREEN = 33.65 | -1.6532851 | 0.4866846 | -3.3970359 | 792.92752 | 0.0007151 | -2.6086277 | -0.6979426 | NA |
| GREEN = 40.1 | -0.5487540 | 0.4539039 | -1.2089651 | 791.68124 | 0.2270372 | -1.4397514 | 0.3422435 | NA |
| GREEN = 40.85 | -0.5152966 | 0.3827774 | -1.3462044 | 784.31604 | 0.1786256 | -1.2666860 | 0.2360928 | NA |
| GREEN = 47.65 | -3.3189395 | 1.0794894 | -3.0745457 | 792.88005 | 0.0021805 | -5.4379346 | -1.1999445 | NA |
| PHC = 1 | -0.2765367 | 0.1779092 | -1.5543699 | 717.40916 | 0.1205372 | -0.6258216 | 0.0727482 | NA |
| BEDNET = 1 | 0.1520534 | 0.2330051 | 0.6525755 | 59.94121 | 0.5165250 | -0.3140355 | 0.6181423 | NA |

After fitting a logistic regression to each of the 10 generated datasets, we can see that the bednet variable is actually not significant at the $\alpha$=0.05 confidence level.

To further confirm that the bednet variable is not a significant predictor of malaria, we use a likelihood test comparing models with and without BEDNET. The p-value of this test is 0.5304, which means that the larger model with bednet as a predictor does not explain deviance significantly better than the smaller one.

## Contributions

Nathaniel made the Exploratory Data Analysis plots on the missingness of the bednet variable, and wrote the observations. Huijia worked on the Discussion of Approaches section. Angie worked on chained regression.

## References

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple Imputation by Chained Equations: What is it and how does it work? International Journal of Methods in Psychiatric Research, 20(1), 40-49. http://doi.org/10.1002/mpr.329