report1

Nathaniel Brown, Huijia Yu, Angie Shen

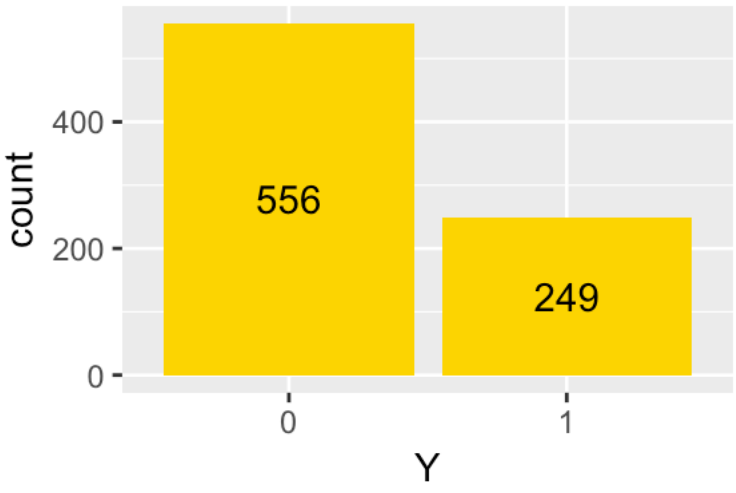
November 6, 2017

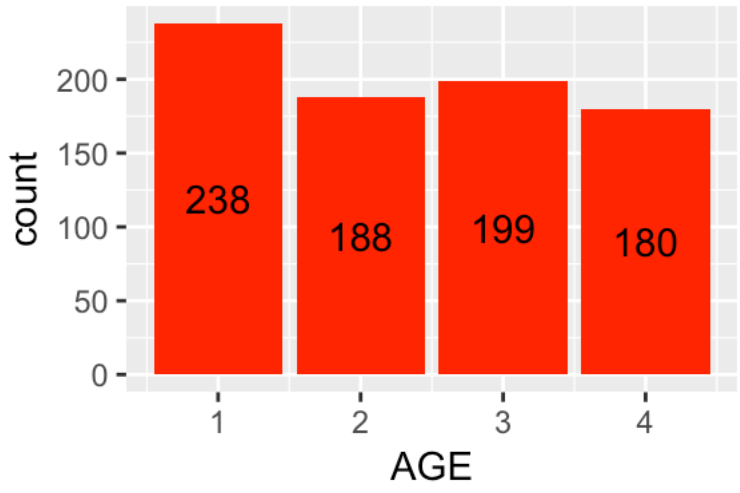
Introduction

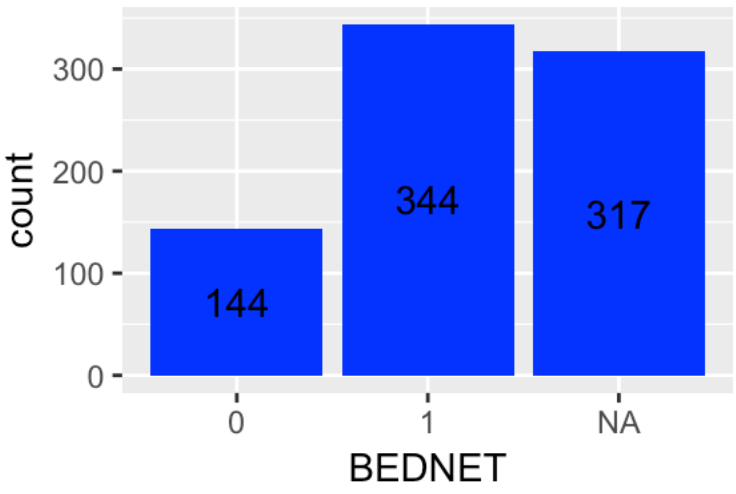
We are interested in predicting whether malaria parasites will be found in a child's blood based on his/her age, bednet use, the amount of greenery in his/her village, and the presence of a health clinic. However, almost 40% of the observations do not have bednet use reported, so removing those observations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

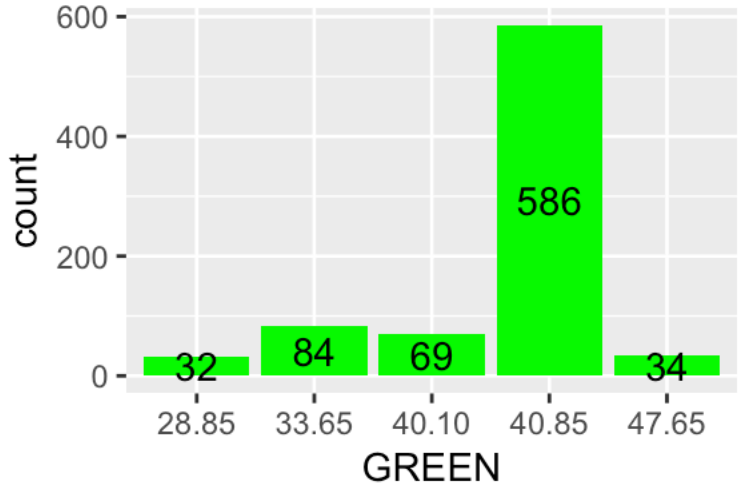
Exploratory Data Analysis

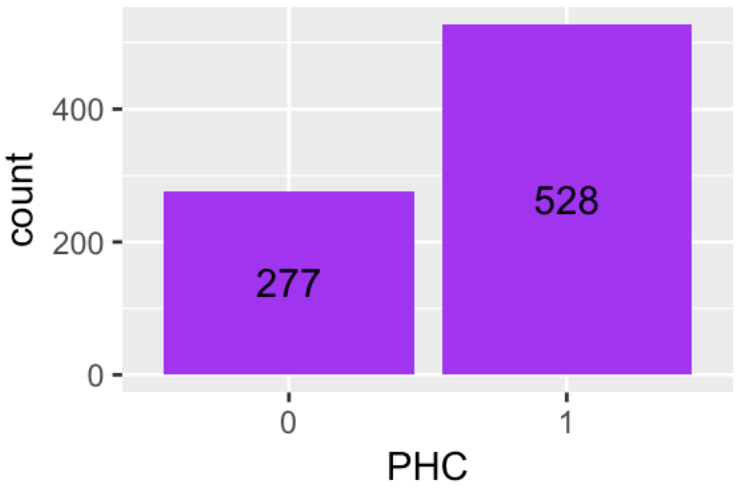
In the graphs below, we illustrate the frequency of each category within each the dataset:



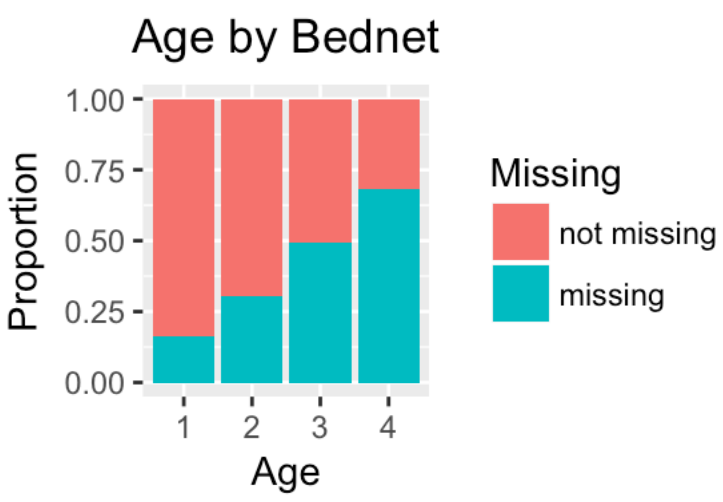


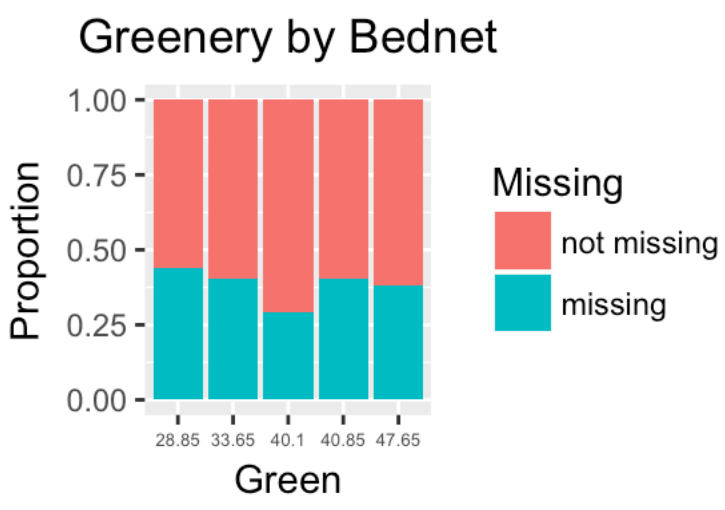


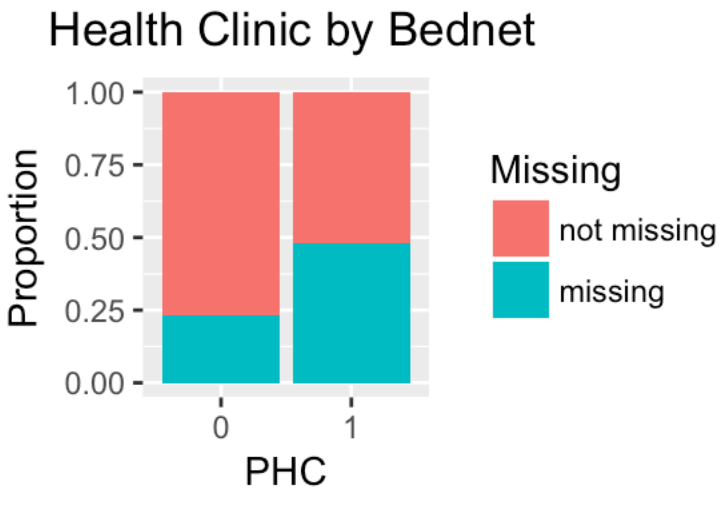


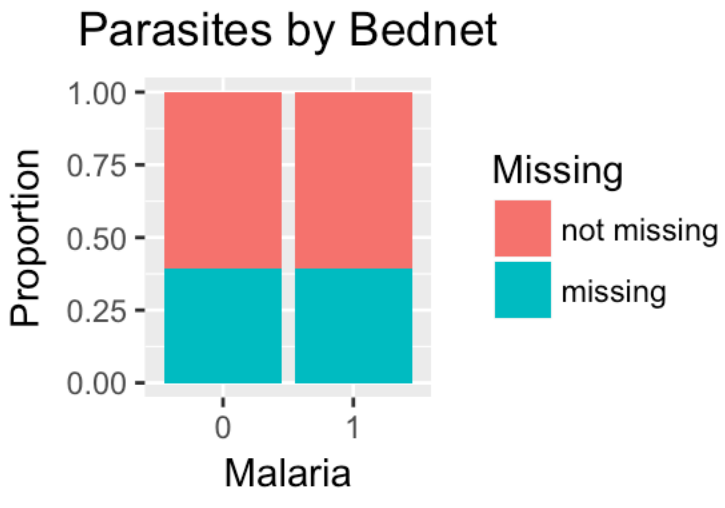


Below, we investigate the proportions of missing bednet responses by each level of the predictors.









There appears to be a relationship between missingness in the bednet variable and age. As Age increases, the proportion of missing bednet responses increases. The other predictor variables and the response do not have an obvious visible relationship with missing bednet. We will use logistic regression to formally test the null hypothesis that the bednet data is missing completely at random (MCAR) versus the alternative that it is missing at random (MAR).

Test for Missingness Mechanism

description of the three mechanisms

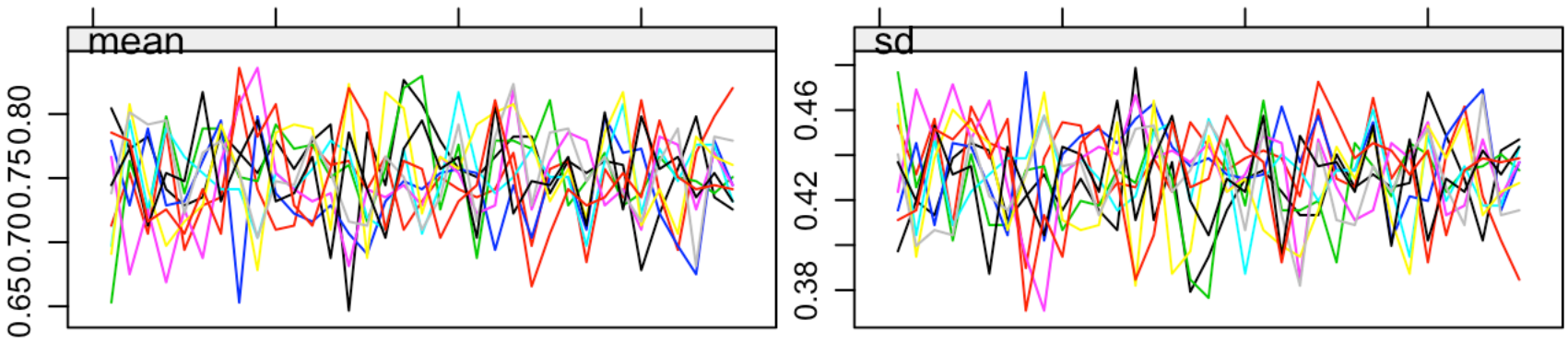
To test the assumption of the MCAR mechanism versus MAR, we build a logistic regression model to determine if our observed predictors (age, green, and public health clinic) are preditive of the missing bednet values. If this “full” model is not a better predictor than a “null” model with only an intercept, then we do not reject the null hypothesis that the missingness mechanism is MCAR.

Based on a Chi-Squared test comparing the deviance of the full model to the deviance of the null model, there is extremely strong evidence that the missingness mechanism is MAR instead of MCAR (p=0). There is no way of knowing whether it is MAR or NMAR, since NMAR implies that the missing mechanism depends on unobserved factors.

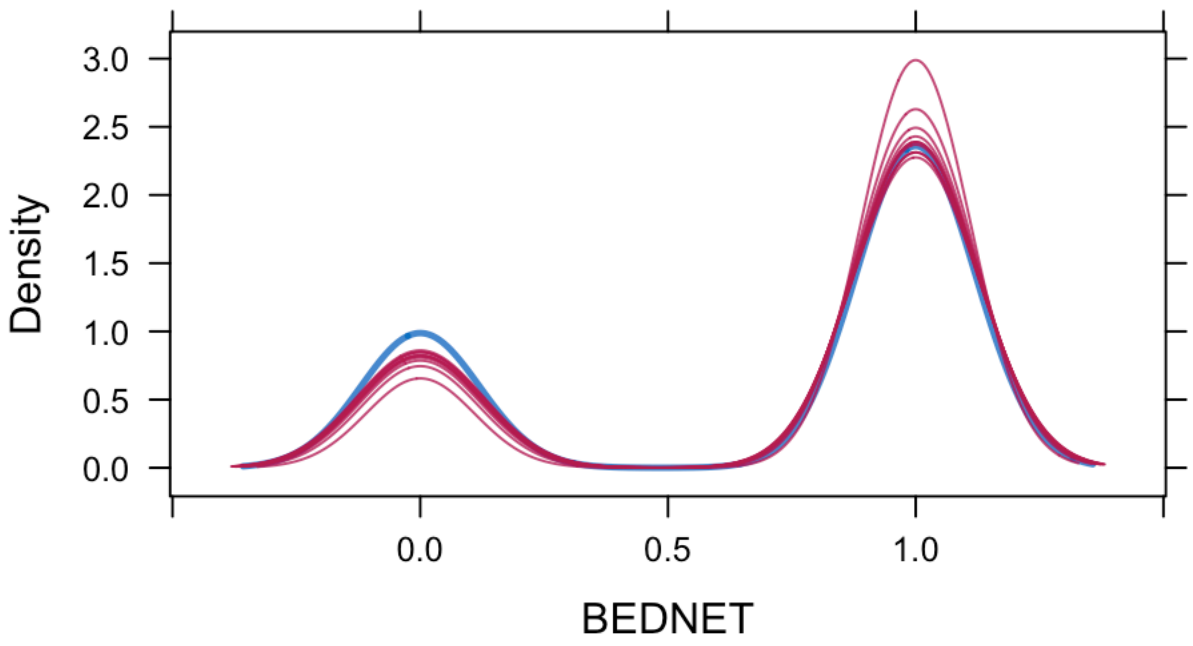
Discussion of Approaches

Since the data is not MCAR, we cannot use bootstrapping within the bednet variable. We will instead use chained regression using the MICE (Multivariate Imputation by Chained Equations) package to impute the missing values and calculate malaria prevalence.

Multivariate imputation by chained equations (MICE) has emerged as a principled method of dealing with missing data. In the MICE procedure a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data. This means that each variable can be modeled according to its distribution, with, for example, binary variables modeled using logistic regression and continuous variables modeled using linear regression. The entire imputation process is repeated to generate multiple imputed datasets. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations.



Iteration



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).

	est	se	t	df	Pr(>|t|)	lo 95	hi 95	nmis
(Intercept)	-0.5271690	0.9882332	-0.5334459	774.1181	0.5938780	-2.4671035	1.4127655	NA
AGE	0.2619201	0.0690935	3.7908044	797.6705	0.0001615	0.1262934	0.3975467	0
GREEN	-0.0170033	0.0232395	-0.7316588	793.6038	0.4645929	-0.0626214	0.0286147	0
PHC	-0.5090915	0.1721985	-2.9564232	751.5774	0.0032097	-0.8471387	-0.1710443	0
BEDNET	0.1140422	0.2081797	0.5478064	136.2558	0.5847211	-0.2976389	0.5257232	317

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 α=0.05 confidence level.

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 (http://doi.org/10.1002/mpr.329)