Case 3-1 Stat 440

Sarah Zimmermann, Adam Wood, Wuming Zhang

November 2, 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 greenry on a child having malaria, where greenery is a measure of how much vegetation is around the child's village based on satellite village. A summary of the variables is found below:

Variable Name	Short Description	Type
Y	indicator of whether malaria parasites were found in the blood of the child	binary
AGE	age of child in years	continuous
BEDNET	indicator of whether the child has a benet over his or her bed	binary
GREEN	a measure of how much greenery is around the child's village, derived from satellite images	continuous
PHC	indicator for the presence of a public health clinic in the child's village	binary

Data

```
## [1] 805   5
```

EDA

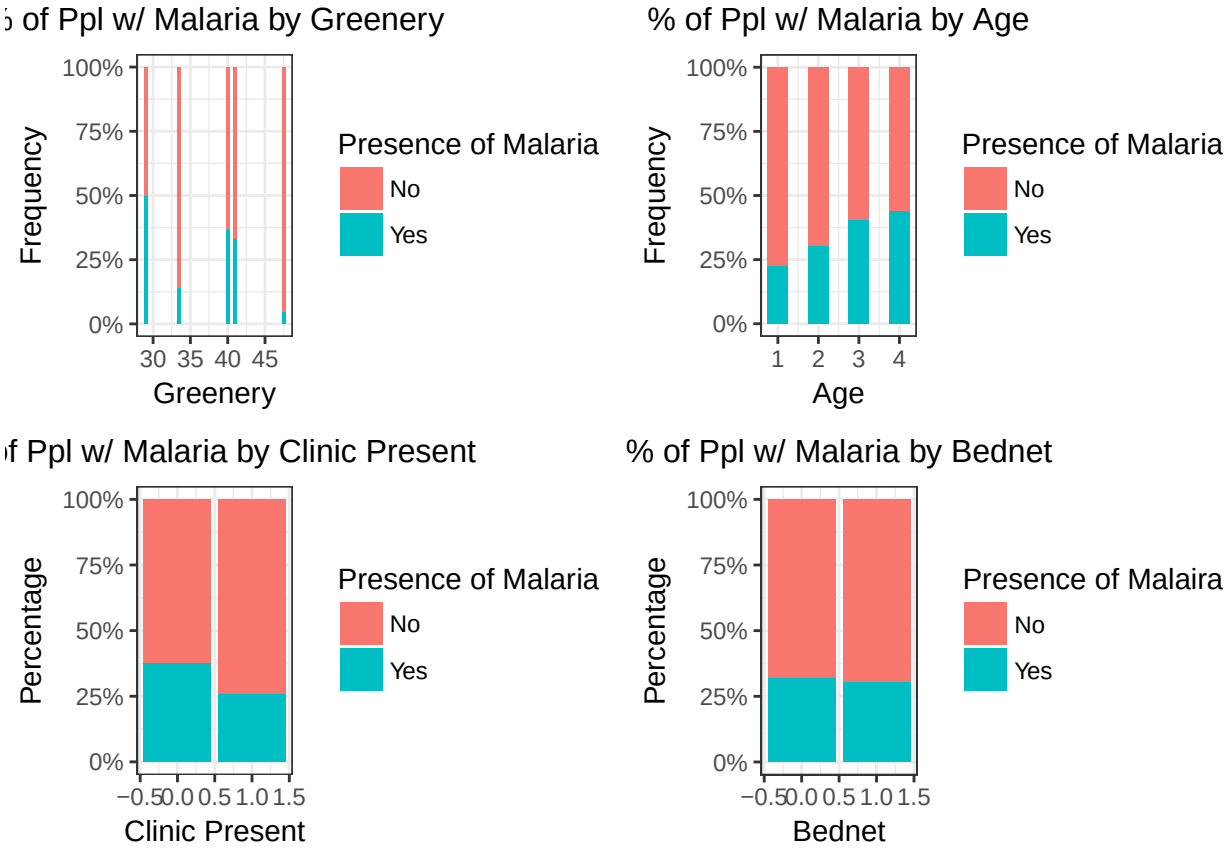
(Top Right) Here we visualize the percent of people with malaria based on the different greenery in the surrounding village. We see the percent of people with the disease varies with the amount of greenery. The relationship between the greenery and people sick is not a strictly positive or negative correlation because the percent of sick both increases and decreases as greenery increases. This is somewhat surprising as mosquitoes

(which transmit the disease to humans) are likely to live in places with greenery; however, this graphic does not suggest the more greenery the higher the percent of children with malaria.

(Top Left) Here we visualize the percent of people with and without malaria based across different ages. As age increases the percent of people in the age bracket sick increases.

(Bottom Right) Here we visualize the percent of people with and without malaria based on the presence of a clinic in the village. A higher percent of people are sick in villages without a clinic compared to those sick in a village with a clinic. This makes sense that the presence of malaria is smaller where there are clinics because this means children have access to medicine. Additionally, it is possible the clinics probably help educate and spread awareness about prevention therefore decreasing the presence of malaria.

(Bottom Left) Here we visualize the percent of people with and without malaria based on use of bednets. Of the child who do not have bednets there is not a significantly higher percent of children sick than those who used bed nets. This is surprising because bednets are a means of prevention and are recommended by agencies such as the CDC (https://www.cdc.gov/malaria/malaria_worldwide/reduction/itn.html). The data suggests, however, there is not much difference between the percent of people sick with and without bed use.



Approaches Outline

Missing Data Mechanisms

From initial exploration, we observe that we are missing 39.38% data for the explanatory variable BEDNET among all observations. We first suspect that this might simply be some random data collection issue. For instance, it could be that it is generally hard to collect data for people and missing data would occur very randomly; that is, the missing completely at random (MCAR) assumption. However, it is a bit suspicious

that we have no missing data for any other explanatory variable in the dataset. Therefore, another possibility is that, it is harder to collect the BEDNET data for certain group of the people in this study than the others; in another word, the missing at random (MAR) assumption.

To further explore whether certain groups are more likely to have missing value, we perform a logistic regression with missing value as th reponse variable and all other variables in the original dataset as the explanatory variables. We observe from the results below that, the coefficients for AGE and PHC variables are statistically significant. This means that whether missing data would occur has some correlation with some variables in the dataset. Therefore, we can reject the MCAR assumption and assume that only MAR assumption holds.

```
##
## Call:
## glm(formula = missing ~ Y + AGE + GREEN + as.factor(PHC), data = gambia)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7748  -0.3566  -0.1566   0.4128   0.9820
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.256489   0.186907  -1.372    0.170
## Y               -0.031669   0.033638  -0.941    0.347
## AGE              0.170278   0.013713  12.417  < 2e-16 ***
## GREEN            0.002550   0.004489   0.568    0.570
## as.factor(PHC)1  0.228686   0.033099   6.909 9.95e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1893755)
##
##     Null deviance: 192.17  on 804  degrees of freedom
## Residual deviance: 151.50  on 800  degrees of freedom
## AIC: 951.94
##
## Number of Fisher Scoring iterations: 2
```

Method proposal

A naive way of dealing with missing data is to perform listwise deletion, that is, to delete the entire observation if a missing data occurs. Although this method is very easy to implement, it would potentially reduce the statistical power of our analysis since we would need to reduce the size of the dataset by almost half. Additionaly, since the MCAR assumption does not hold, our estimates would be biased using this method.

In order to obtain unbaised estimates while properly using the MAR assumption, we plan to use the multiple imputation method. This method contains three main steps. First, we impute the missing values for m times using an appropriate model. Second, we perform our desired analysis on each of the m datasets. Last, we combine the m analysis into one final result using some proper method. Note that the multiple imputation method would bring us more variability thus improving the accuracy of our estimates. However, the implementation of this method could be much more challenging and we need to be very careful when choosing models.

Credit

Sarah wrote the introduction and created/analyzed bivariate graphs.

References

http://www.stat.columbia.edu/~gelman/arm/missing.pdf