case3-2

Sarah Zimmermann, Wuming Zhang, Adam Wood

November 9, 2017

Data

```
library(ggplot2)
library(devtools)
library(mice)
```

```
## Loading required package: lattice
```

```
library(lattice)
gambia= read.csv("gambiaMissing.csv")
#gambia$log.green = log(gambia$GREEN)
# may need log transformation for GREEN
dim(gambia)
```

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

Multiple Imputation/Chained Regression

```
tempData <- mice(gambia, m=5,maxit=50,meth='pmm',seed=500)
```

```
##
##  iter imp variable
##   1   1  BEDNET
##   1   2  BEDNET
##   1   3  BEDNET
##   1   4  BEDNET
##   1   5  BEDNET
##   2   1  BEDNET
##   2   2  BEDNET
##   2   3  BEDNET
##   2   4  BEDNET
##   2   5  BEDNET
##   3   1  BEDNET
##   3   2  BEDNET
##   3   3  BEDNET
##   3   4  BEDNET
##   3   5  BEDNET
##   4   1  BEDNET
##   4   2  BEDNET
##   4   3  BEDNET
##   4   4  BEDNET
##   4   5  BEDNET
##   5   1  BEDNET
##   5   2  BEDNET
##   5   3  BEDNET
```

```
##    5   4  BEDNET
##    5   5  BEDNET
##    6   1  BEDNET
##    6   2  BEDNET
##    6   3  BEDNET
##    6   4  BEDNET
##    6   5  BEDNET
##    7   1  BEDNET
##    7   2  BEDNET
##    7   3  BEDNET
##    7   4  BEDNET
##    7   5  BEDNET
##    8   1  BEDNET
##    8   2  BEDNET
##    8   3  BEDNET
##    8   4  BEDNET
##    8   5  BEDNET
##    9   1  BEDNET
##    9   2  BEDNET
##    9   3  BEDNET
##    9   4  BEDNET
##    9   5  BEDNET
##    10   1  BEDNET
##    10   2  BEDNET
##    10   3  BEDNET
##    10   4  BEDNET
##    10   5  BEDNET
##    11   1  BEDNET
##    11   2  BEDNET
##    11   3  BEDNET
##    11   4  BEDNET
##    11   5  BEDNET
##    12   1  BEDNET
##    12   2  BEDNET
##    12   3  BEDNET
##    12   4  BEDNET
##    12   5  BEDNET
##    13   1  BEDNET
##    13   2  BEDNET
##    13   3  BEDNET
##    13   4  BEDNET
##    13   5  BEDNET
##    14   1  BEDNET
##    14   2  BEDNET
##    14   3  BEDNET
##    14   4  BEDNET
##    14   5  BEDNET
##    15   1  BEDNET
##    15   2  BEDNET
##    15   3  BEDNET
##    15   4  BEDNET
##    15   5  BEDNET
##    16   1  BEDNET
##    16   2  BEDNET
```

```
##   16   3  BEDNET
##   16   4  BEDNET
##   16   5  BEDNET
##   17   1  BEDNET
##   17   2  BEDNET
##   17   3  BEDNET
##   17   4  BEDNET
##   17   5  BEDNET
##   18   1  BEDNET
##   18   2  BEDNET
##   18   3  BEDNET
##   18   4  BEDNET
##   18   5  BEDNET
##   19   1  BEDNET
##   19   2  BEDNET
##   19   3  BEDNET
##   19   4  BEDNET
##   19   5  BEDNET
##   20   1  BEDNET
##   20   2  BEDNET
##   20   3  BEDNET
##   20   4  BEDNET
##   20   5  BEDNET
##   21   1  BEDNET
##   21   2  BEDNET
##   21   3  BEDNET
##   21   4  BEDNET
##   21   5  BEDNET
##   22   1  BEDNET
##   22   2  BEDNET
##   22   3  BEDNET
##   22   4  BEDNET
##   22   5  BEDNET
##   23   1  BEDNET
##   23   2  BEDNET
##   23   3  BEDNET
##   23   4  BEDNET
##   23   5  BEDNET
##   24   1  BEDNET
##   24   2  BEDNET
##   24   3  BEDNET
##   24   4  BEDNET
##   24   5  BEDNET
##   25   1  BEDNET
##   25   2  BEDNET
##   25   3  BEDNET
##   25   4  BEDNET
##   25   5  BEDNET
##   26   1  BEDNET
##   26   2  BEDNET
##   26   3  BEDNET
##   26   4  BEDNET
##   26   5  BEDNET
##   27   1  BEDNET
```

```
##   27   2  BEDNET
##   27   3  BEDNET
##   27   4  BEDNET
##   27   5  BEDNET
##   28   1  BEDNET
##   28   2  BEDNET
##   28   3  BEDNET
##   28   4  BEDNET
##   28   5  BEDNET
##   29   1  BEDNET
##   29   2  BEDNET
##   29   3  BEDNET
##   29   4  BEDNET
##   29   5  BEDNET
##   30   1  BEDNET
##   30   2  BEDNET
##   30   3  BEDNET
##   30   4  BEDNET
##   30   5  BEDNET
##   31   1  BEDNET
##   31   2  BEDNET
##   31   3  BEDNET
##   31   4  BEDNET
##   31   5  BEDNET
##   32   1  BEDNET
##   32   2  BEDNET
##   32   3  BEDNET
##   32   4  BEDNET
##   32   5  BEDNET
##   33   1  BEDNET
##   33   2  BEDNET
##   33   3  BEDNET
##   33   4  BEDNET
##   33   5  BEDNET
##   34   1  BEDNET
##   34   2  BEDNET
##   34   3  BEDNET
##   34   4  BEDNET
##   34   5  BEDNET
##   35   1  BEDNET
##   35   2  BEDNET
##   35   3  BEDNET
##   35   4  BEDNET
##   35   5  BEDNET
##   36   1  BEDNET
##   36   2  BEDNET
##   36   3  BEDNET
##   36   4  BEDNET
##   36   5  BEDNET
##   37   1  BEDNET
##   37   2  BEDNET
##   37   3  BEDNET
##   37   4  BEDNET
##   37   5  BEDNET
```

```
##   38   1  BEDNET
##   38   2  BEDNET
##   38   3  BEDNET
##   38   4  BEDNET
##   38   5  BEDNET
##   39   1  BEDNET
##   39   2  BEDNET
##   39   3  BEDNET
##   39   4  BEDNET
##   39   5  BEDNET
##   40   1  BEDNET
##   40   2  BEDNET
##   40   3  BEDNET
##   40   4  BEDNET
##   40   5  BEDNET
##   41   1  BEDNET
##   41   2  BEDNET
##   41   3  BEDNET
##   41   4  BEDNET
##   41   5  BEDNET
##   42   1  BEDNET
##   42   2  BEDNET
##   42   3  BEDNET
##   42   4  BEDNET
##   42   5  BEDNET
##   43   1  BEDNET
##   43   2  BEDNET
##   43   3  BEDNET
##   43   4  BEDNET
##   43   5  BEDNET
##   44   1  BEDNET
##   44   2  BEDNET
##   44   3  BEDNET
##   44   4  BEDNET
##   44   5  BEDNET
##   45   1  BEDNET
##   45   2  BEDNET
##   45   3  BEDNET
##   45   4  BEDNET
##   45   5  BEDNET
##   46   1  BEDNET
##   46   2  BEDNET
##   46   3  BEDNET
##   46   4  BEDNET
##   46   5  BEDNET
##   47   1  BEDNET
##   47   2  BEDNET
##   47   3  BEDNET
##   47   4  BEDNET
##   47   5  BEDNET
##   48   1  BEDNET
##   48   2  BEDNET
##   48   3  BEDNET
##   48   4  BEDNET
```

```
##   48   5  BEDNET
##   49   1  BEDNET
##   49   2  BEDNET
##   49   3  BEDNET
##   49   4  BEDNET
##   49   5  BEDNET
##   50   1  BEDNET
##   50   2  BEDNET
##   50   3  BEDNET
##   50   4  BEDNET
##   50   5  BEDNET
```

```
#gambia$BEDNET = as.factor(gambia$BEDNET)
#tempData <- mice(gambia, m=5,maxit=50,meth='logreg',seed=500)
# should probably use logreg since the variable is binary?
summary(tempData)
```

```
## Multiply imputed data set
## Call:
## mice(data = gambia, m = 5, method = "pmm", maxit = 50, seed = 500)
## Number of multiple imputations:  5
## Missing cells per column:
##      Y    AGE BEDNET  GREEN    PHC
##      0      0    317      0      0
## Imputation methods:
##      Y    AGE BEDNET  GREEN    PHC
##  "pmm"  "pmm"  "pmm"  "pmm"  "pmm"
## VisitSequence:
## BEDNET
##      3
## PredictorMatrix:
##        Y AGE BEDNET GREEN PHC
## Y      0   0      0     0   0
## AGE    0   0      0     0   0
## BEDNET 1   1      0     1   1
## GREEN  0   0      0     0   0
## PHC    0   0      0     0   0
## Random generator seed value:  500
```

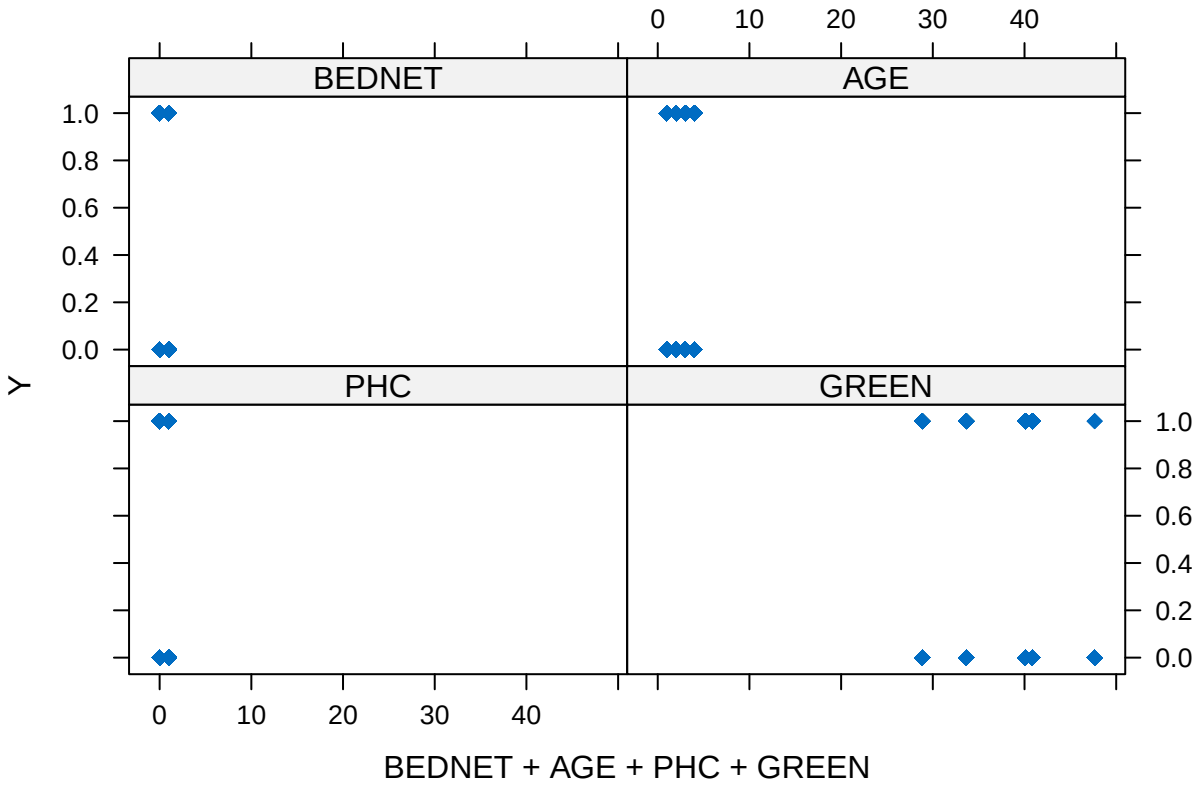
```
#where m is the number of imputed data sets, pmm is the method(in this case predictive mean matching)

completedData <- complete(tempData,1)
#this function returns the imputed data set, with the second parameter in the function determining whic
```

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.

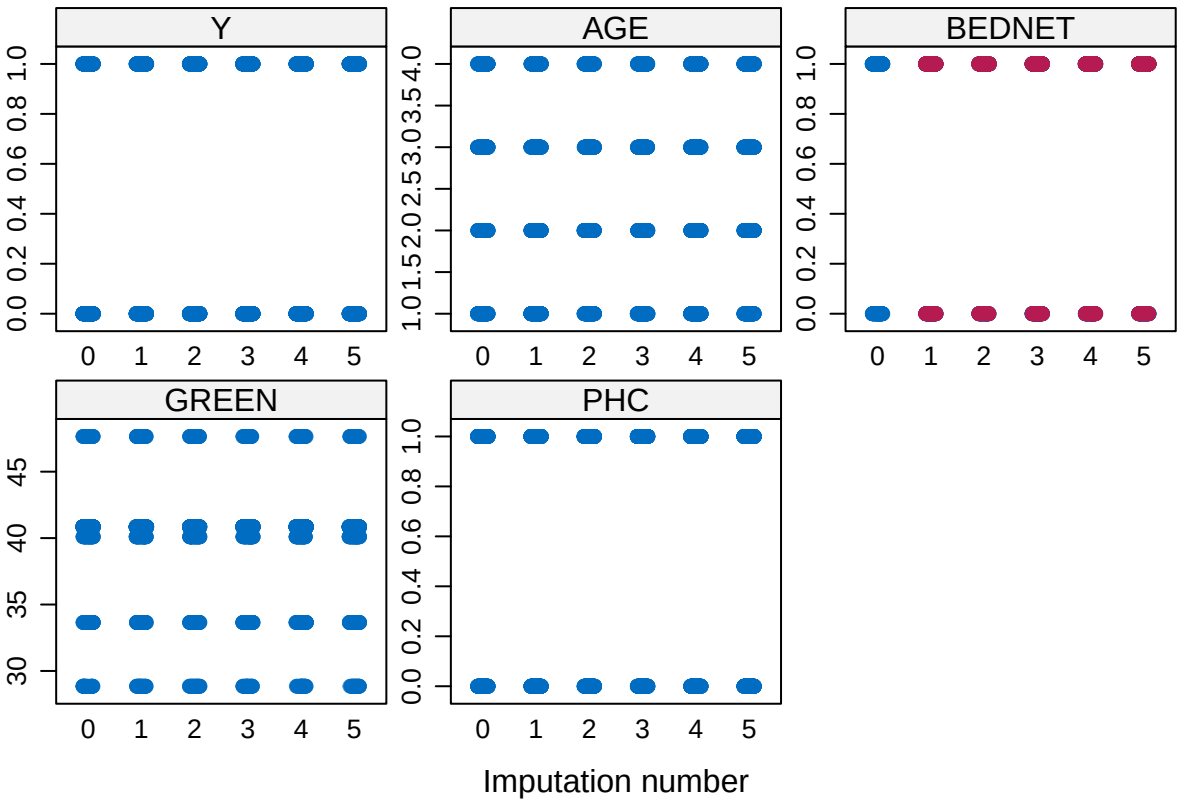
Inspection of Missing Data

```
xyplot(tempData,Y ~ BEDNET + AGE+ PHC+ GREEN,pch=18,cex=1)
```



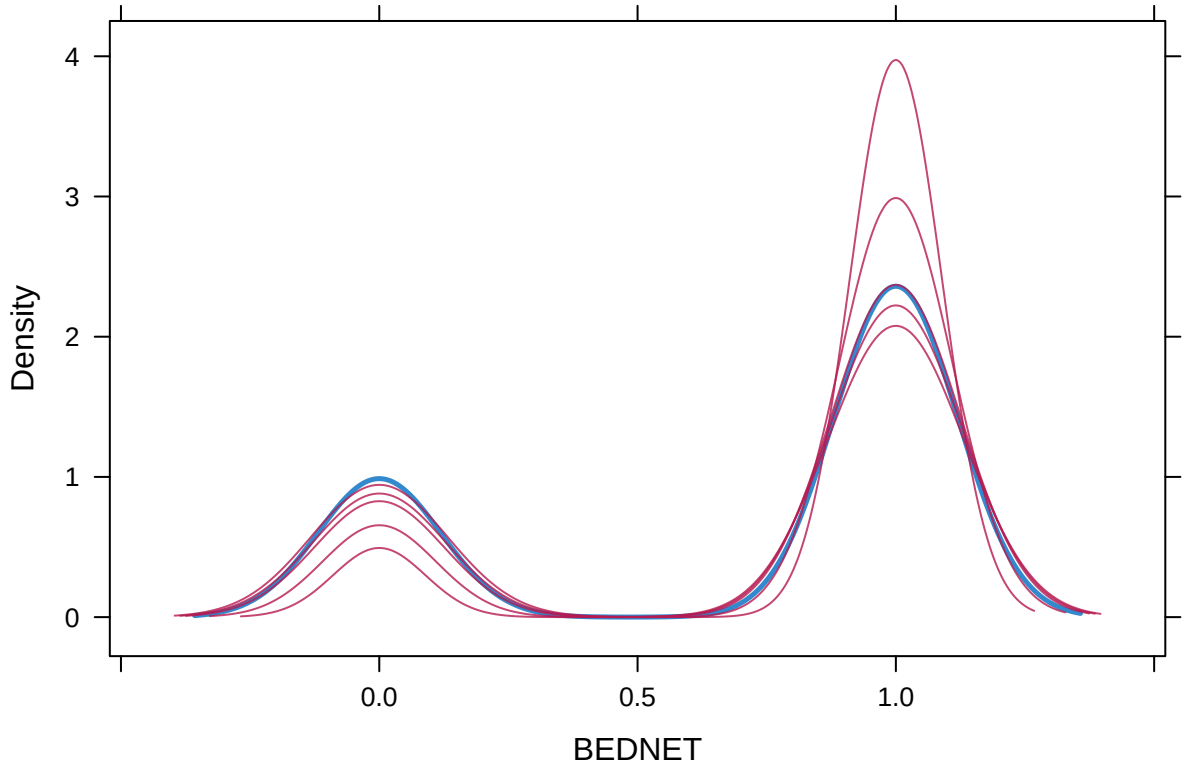

```
#unfortunately provides little information when comparing imputed to observed

stripplot(tempData, pch=20, cex=1.2)
```



```
#would be a useful plot for largely continuous data without so many binary or ordinal variables (both Y

densityplot(tempData)
```




```
#the densities for each imputed data set follow a similar shape to the observed density curve
```

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

```
modelFit1 <- with(tempData,glm(Y~ BEDNET+AGE+PHC+GREEN))
summary(pool(modelFit1))
```

```
##                      est          se          t         df     Pr(>|t|)
## (Intercept)  0.380529272 0.215807642  1.7632799 177.308292 0.0795758352
## BEDNET       0.009648102 0.075522704  0.1277510   7.051291 0.9019131167
## AGE          0.054493576 0.014386908  3.7877197 782.333848 0.0001636916
## PHC         -0.101808531 0.039416868 -2.5828671 104.887427 0.0111765039
## GREEN       -0.003554237 0.004867051 -0.7302649 540.431797 0.4655446865
##                    lo 95        hi 95 nmis        fmi     lambda
## (Intercept) -0.04535278  0.806411320   NA 0.13927164 0.12961726
## BEDNET      -0.16867172  0.187967922  317 0.79193725 0.74025294
## AGE          0.02625206  0.082735090    0 0.01033538 0.00780858
## PHC         -0.17996588 -0.023651185    0 0.19419213 0.17897204
## GREEN       -0.01311489  0.006006419    0 0.04982991 0.04632007
```

```
#takes the imputed data sets, fits a model to each data set, and pools together results

tempData2 <- mice(gambia,m=50,seed=245435)
```

```
##
##  iter imp variable
##   1   1  BEDNET
##   1   2  BEDNET
##   1   3  BEDNET
##   1   4  BEDNET
##   1   5  BEDNET
##   1   6  BEDNET
##   1   7  BEDNET
##   1   8  BEDNET
##   1   9  BEDNET
##   1   10  BEDNET
##   1   11  BEDNET
##   1   12  BEDNET
##   1   13  BEDNET
##   1   14  BEDNET
##   1   15  BEDNET
##   1   16  BEDNET
##   1   17  BEDNET
##   1   18  BEDNET
##   1   19  BEDNET
##   1   20  BEDNET
##   1   21  BEDNET
##   1   22  BEDNET
##   1   23  BEDNET
##   1   24  BEDNET
##   1   25  BEDNET
##   1   26  BEDNET
##   1   27  BEDNET
##   1   28  BEDNET
##   1   29  BEDNET
##   1   30  BEDNET
##   1   31  BEDNET
##   1   32  BEDNET
##   1   33  BEDNET
##   1   34  BEDNET
##   1   35  BEDNET
##   1   36  BEDNET
```

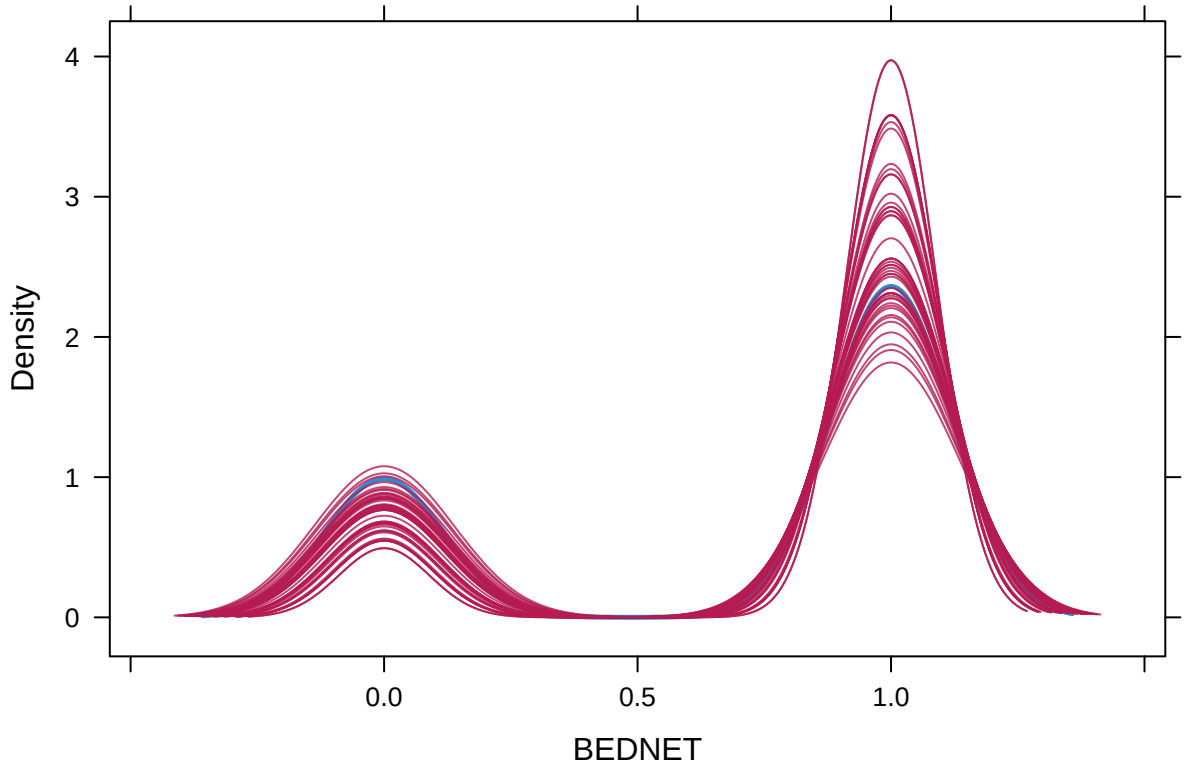
```
##   1   37  BEDNET
##   1   38  BEDNET
##   1   39  BEDNET
##   1   40  BEDNET
##   1   41  BEDNET
##   1   42  BEDNET
##   1   43  BEDNET
##   1   44  BEDNET
##   1   45  BEDNET
##   1   46  BEDNET
##   1   47  BEDNET
##   1   48  BEDNET
##   1   49  BEDNET
##   1   50  BEDNET
##   2   1  BEDNET
##   2   2  BEDNET
##   2   3  BEDNET
##   2   4  BEDNET
##   2   5  BEDNET
##   2   6  BEDNET
##   2   7  BEDNET
##   2   8  BEDNET
##   2   9  BEDNET
##   2   10  BEDNET
##   2   11  BEDNET
##   2   12  BEDNET
##   2   13  BEDNET
##   2   14  BEDNET
##   2   15  BEDNET
##   2   16  BEDNET
##   2   17  BEDNET
##   2   18  BEDNET
##   2   19  BEDNET
##   2   20  BEDNET
##   2   21  BEDNET
##   2   22  BEDNET
##   2   23  BEDNET
##   2   24  BEDNET
##   2   25  BEDNET
##   2   26  BEDNET
##   2   27  BEDNET
##   2   28  BEDNET
##   2   29  BEDNET
##   2   30  BEDNET
##   2   31  BEDNET
##   2   32  BEDNET
##   2   33  BEDNET
##   2   34  BEDNET
##   2   35  BEDNET
##   2   36  BEDNET
##   2   37  BEDNET
##   2   38  BEDNET
##   2   39  BEDNET
##   2   40  BEDNET
```

```
##   2   41  BEDNET
##   2   42  BEDNET
##   2   43  BEDNET
##   2   44  BEDNET
##   2   45  BEDNET
##   2   46  BEDNET
##   2   47  BEDNET
##   2   48  BEDNET
##   2   49  BEDNET
##   2   50  BEDNET
##   3   1  BEDNET
##   3   2  BEDNET
##   3   3  BEDNET
##   3   4  BEDNET
##   3   5  BEDNET
##   3   6  BEDNET
##   3   7  BEDNET
##   3   8  BEDNET
##   3   9  BEDNET
##   3   10  BEDNET
##   3   11  BEDNET
##   3   12  BEDNET
##   3   13  BEDNET
##   3   14  BEDNET
##   3   15  BEDNET
##   3   16  BEDNET
##   3   17  BEDNET
##   3   18  BEDNET
##   3   19  BEDNET
##   3   20  BEDNET
##   3   21  BEDNET
##   3   22  BEDNET
##   3   23  BEDNET
##   3   24  BEDNET
##   3   25  BEDNET
##   3   26  BEDNET
##   3   27  BEDNET
##   3   28  BEDNET
##   3   29  BEDNET
##   3   30  BEDNET
##   3   31  BEDNET
##   3   32  BEDNET
##   3   33  BEDNET
##   3   34  BEDNET
##   3   35  BEDNET
##   3   36  BEDNET
##   3   37  BEDNET
##   3   38  BEDNET
##   3   39  BEDNET
##   3   40  BEDNET
##   3   41  BEDNET
##   3   42  BEDNET
##   3   43  BEDNET
##   3   44  BEDNET
```

```
##   3   45  BEDNET
##   3   46  BEDNET
##   3   47  BEDNET
##   3   48  BEDNET
##   3   49  BEDNET
##   3   50  BEDNET
##   4   1  BEDNET
##   4   2  BEDNET
##   4   3  BEDNET
##   4   4  BEDNET
##   4   5  BEDNET
##   4   6  BEDNET
##   4   7  BEDNET
##   4   8  BEDNET
##   4   9  BEDNET
##   4   10  BEDNET
##   4   11  BEDNET
##   4   12  BEDNET
##   4   13  BEDNET
##   4   14  BEDNET
##   4   15  BEDNET
##   4   16  BEDNET
##   4   17  BEDNET
##   4   18  BEDNET
##   4   19  BEDNET
##   4   20  BEDNET
##   4   21  BEDNET
##   4   22  BEDNET
##   4   23  BEDNET
##   4   24  BEDNET
##   4   25  BEDNET
##   4   26  BEDNET
##   4   27  BEDNET
##   4   28  BEDNET
##   4   29  BEDNET
##   4   30  BEDNET
##   4   31  BEDNET
##   4   32  BEDNET
##   4   33  BEDNET
##   4   34  BEDNET
##   4   35  BEDNET
##   4   36  BEDNET
##   4   37  BEDNET
##   4   38  BEDNET
##   4   39  BEDNET
##   4   40  BEDNET
##   4   41  BEDNET
##   4   42  BEDNET
##   4   43  BEDNET
##   4   44  BEDNET
##   4   45  BEDNET
##   4   46  BEDNET
##   4   47  BEDNET
##   4   48  BEDNET
```

```
##   4   49  BEDNET
##   4   50  BEDNET
##   5   1  BEDNET
##   5   2  BEDNET
##   5   3  BEDNET
##   5   4  BEDNET
##   5   5  BEDNET
##   5   6  BEDNET
##   5   7  BEDNET
##   5   8  BEDNET
##   5   9  BEDNET
##   5   10  BEDNET
##   5   11  BEDNET
##   5   12  BEDNET
##   5   13  BEDNET
##   5   14  BEDNET
##   5   15  BEDNET
##   5   16  BEDNET
##   5   17  BEDNET
##   5   18  BEDNET
##   5   19  BEDNET
##   5   20  BEDNET
##   5   21  BEDNET
##   5   22  BEDNET
##   5   23  BEDNET
##   5   24  BEDNET
##   5   25  BEDNET
##   5   26  BEDNET
##   5   27  BEDNET
##   5   28  BEDNET
##   5   29  BEDNET
##   5   30  BEDNET
##   5   31  BEDNET
##   5   32  BEDNET
##   5   33  BEDNET
##   5   34  BEDNET
##   5   35  BEDNET
##   5   36  BEDNET
##   5   37  BEDNET
##   5   38  BEDNET
##   5   39  BEDNET
##   5   40  BEDNET
##   5   41  BEDNET
##   5   42  BEDNET
##   5   43  BEDNET
##   5   44  BEDNET
##   5   45  BEDNET
##   5   46  BEDNET
##   5   47  BEDNET
##   5   48  BEDNET
##   5   49  BEDNET
##   5   50  BEDNET
```

```
densityplot(tempData2)
```




```
modelFit2 <- with(tempData2,glm(Y~BEDNET+AGE+PHC+GREEN))
summary(pool(modelFit2))
```

```
##                    est          se          t        df     Pr(>|t|)
## (Intercept)  0.363786933 0.223053236  1.6309422 447.22763 0.1036064939
## BEDNET       0.027714221 0.082866855  0.3344428  53.81402 0.7393454571
## AGE          0.054865266 0.014371742  3.8175794 785.53008 0.0001453744
## PHC         -0.107272203 0.038695895 -2.7721856 500.15610 0.0057757808
## GREEN       -0.003409203 0.004917768 -0.6932419 694.84806 0.4883894048
##                   lo 95        hi 95 nmis        fmi     lambda
## (Intercept) -0.07457569  0.802149559   NA 0.18861581 0.18499539
## BEDNET      -0.13843697  0.193865413  317 0.79545079 0.78798740
## AGE          0.02665370  0.083076830    0 0.01547621 0.01297275
## PHC         -0.18329874 -0.031245669    0 0.16154545 0.15819937
## GREEN       -0.01306467  0.006246263    0 0.07072912 0.06805822
```

```
#50 imputed data sets instead of 5 to pool together and fit a 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

```
#plot(modelFit2)
#residuals v fitted plot if applicable, don't know how to plot resid v fit for 50 different data sets


#AIC or BIC or another model fit diagnostic
modelBase <- with(tempData2,glm(Y~1))
modelFit3 <- with(tempData2,glm(Y~BEDNET+AGE+PHC+GREEN+BEDNET:PHC))
summary(pool(modelFit3))
```

```
##                        est          se            t       df     Pr(>|t|)
## (Intercept)  0.3591190924 0.223268705  1.608461396 436.3666 0.1084573844
## BEDNET       0.1109555996 0.088555876  1.252944515  93.5243 0.2133495362
## AGE          0.0548788698 0.014401863  3.810539769 776.1749 0.0001496333
## PHC          0.0008491797 0.093090331  0.009122104 114.3251 0.9927376140
## GREEN       -0.0044717285 0.005041384 -0.887004089 608.2982 0.3754272487
## BEDNET:PHC  -0.1584296797 0.116399492 -1.361085666 104.3055 0.1764199736
##                     lo 95       hi 95 nmis        fmi     lambda
## (Intercept) -0.07969663 0.797934811   NA 0.19429504 0.19061069
## BEDNET      -0.06488583 0.286797028  317 0.61456028 0.60640492
## AGE          0.02660765 0.083150087    0 0.02247167 0.01995608
## PHC         -0.18355642 0.185254784    0 0.55023736 0.54243745
## GREEN       -0.01437236 0.005428902    0 0.11079891 0.10788013
## BEDNET:PHC  -0.38924628 0.072386925   NA 0.57927708 0.57128656
```

```
pool.compare(modelFit2, modelBase, data = tempData2, method = "likelihood")$pvalue
```

```
## [1] 0.1931898
```

```
# weird that the fitted model is not significant compared to a model with only intercept....
pool.compare(modelFit3, modelFit2, data = tempData2, method = "likelihood")$pvalue
```

```
## [1] 0.3746709
```

Credits

References

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