Case 3 Part 2

Sonia Xu, Vinai Oddiraju

November 13, 2017

Problem Overview:

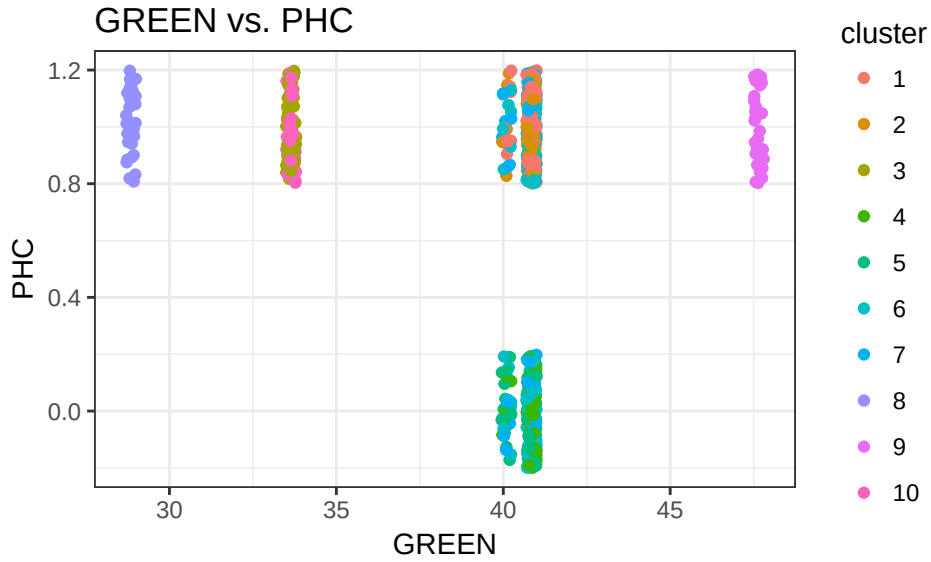
Data from a study of malaria prevalence in Gambia was collected to assess which characteristics were significant in predicting the presence of malaria parasites in a blood sample of a child. However, missing data within the feature, $BEDNET$, an indicator of whether the child has a net over his or her bed, has created a challenge for fully analyzing and understanding the dataset. The purpose of this paper is to determine the best way to imputate the missing data in $BEDNET$ in order to create the best predictive model. Once the best fit model has been determined, the model will be used to determine to what extent each feature affects the probability of a child containing any presence of malaria parasites in their blood sample.

Model Overview

Before creating a robust model that predicts the presence of malaria in child's blood stream, it is important to imputate the missing data well. Three different methods in missing were tested (Clustering & 2 methods within the MICE package) and compared to assess the best way to imputate the missing data.

K-Means Clustering Data Missing Data

To fill in the missing data in $BEDNET$, it is assumed that children who share similar features have a simlar $BEDNET$ history. Each observation was thus clustered on their other features (i.e. $GREEN$), and separated into groups via k-means clustering with 10 groups. 10 groups were chosen because this was the highest number of groups that maintained at least 30 observations for each group, which would ensure a large enough sample size for estimation. Missing $BEDNET$ observations were imputated with the mean of each cluster.



Above is graph of the different clusters–since the features are mainly categorical, some of the clusters were hidden, so the features were jittered to show the different clusters.

The gambia data set was split into test and training data to assess the fit of the model and missing dat imputation, where 70% of the full dataset was training and 30% was the test data. After iterating through the k-means clustering method to impute the missing data and then predicting the Y response via logit regression 100 times, the model predicted the correct test data on average, 63.4025318% of the time.

Assessing Missing Data Imputation with Clustering

To assess the fit of the missing data clusters, the true values of the non-missing $BEDNET$ observations were compared to the predicted values of $BEDNET$ obtained from the clusters. The clusters labeled the correct $BEDNET$ value 70.49% of the time.

Using Mice Package to Imputate Data

For the methods implemented through the "mice" package, 100 iterations of 5 imputations were run in order for the estimates to converge. The estimated malaria indicator values were then compared to the real malaria indicator values to test for accuracy. Below is an overview of each method:

Part I: Mice with Predictive Mean Matching ("PMM")

```
## [1] 0.3937888
```

```
## [1] 0
## [1] 0
## [1] 0.3937888
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
```

```
##     Y AGE GREEN PHC clust impBEDNET impBEDNET_combine BEDNET
## 488 1   1     1   1     1         1                 1      1   0
## 317 1   1     1   1     1         1                 1      0   1
##     0   0     0   0     0         0                 0    317 317
```

```
## [1] 0.6332917 0.6329167 0.6335000 0.6326667 0.6326667
```

Part II: Mice with Bayesian Linear Regression

```
## [1] 0.6323333 0.6315833 0.6327083 0.6319583 0.6333333
```

Between the two "mice" methods, both imputations schemes were similar, but the Bayesian linear regression method was slighlty better. For now, we will move forward with this imputation scheme. Depending on the diagnostics of future models, we may opt to try alternative imputation methods.

To better make use of our Bayesian linear regression method, we may conduct research to see if any sort of prior distribution on the presence of bed nets can be leveraged in our analysis. Hopefully this will lead to better performance of the imputation accuracy.

Example Model

```
FirstTrial = glm(data = gambia, Y ~ AGE + GREEN + PHC + Imp51, family = "binomial")
summary(FirstTrial)
```

```
##
## Call:
## glm(formula = Y ~ AGE + GREEN + PHC + Imp51, family = "binomial",
##     data = gambia)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1497  -0.9065  -0.7480   1.3566   1.8164
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.47319    0.97531  -0.485 0.627555
## AGE          0.26136    0.06904   3.786 0.000153 ***
## GREEN       -0.01770    0.02312  -0.766 0.443907
## PHC         -0.50152    0.17259  -2.906 0.003663 **
## Imp51        0.07185    0.18063   0.398 0.690802
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 995.87  on 804  degrees of freedom
## Residual deviance: 973.94  on 800  degrees of freedom
## AIC: 983.94
##
## Number of Fisher Scoring iterations: 4
```

The above model is neither a finalized version nor a fully diagnosed model, but it represents part of what we hope to achieve. In this case, we replaced the variable $BEDNET$ with an imputation scheme we were satisfied with for the time being. We still want to leverage a training set and a test set in the future, and we also want to continue to improve our imputation scheme for the $BEDNET$ variable.