Case 3 Part 2

Sonia Xu, Vinai Oddiraju

November 13, 2017

Contribution Notes

Both team members were present at class and met outside of class to collaborate on this report. Sonia comtributed most heavily to the clustering section and corresponding visuals and explanations. Vinai contributed most heavily to the implementation of the "mice" package and corresponding explanations. Both members contributed to exploratory data analysis and went over the other member's sections to check for accuracy.

Problem Overview:

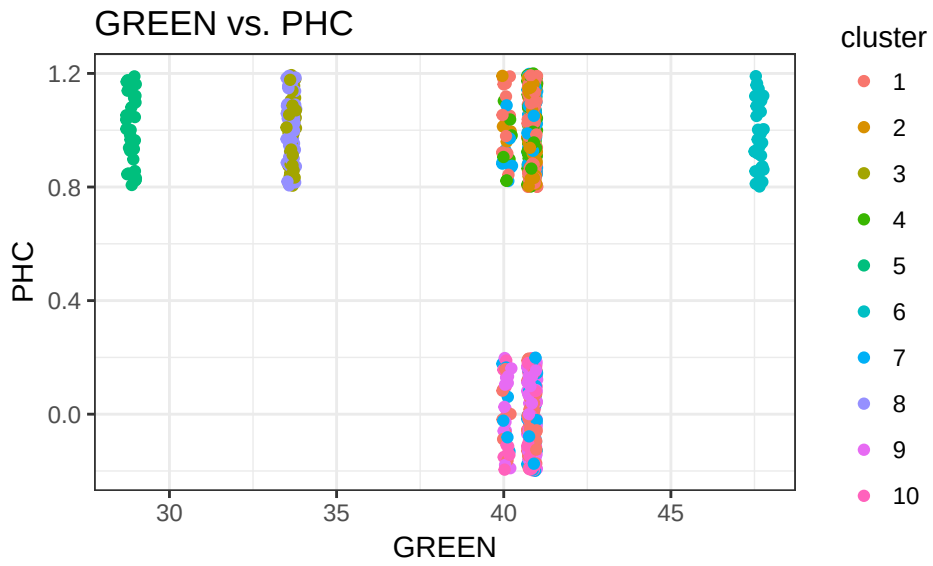
Data from a study of malaria prevalence in Gambia was collected to assess which characteristics were significant in predicting the presence of malaria parasites in a blood sample of a child. However, missing data within the feature, $BEDNET$, an indicator of whether the child has a net over his or her 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 by modeling the imputated data as a logistic regression. The fitted models were compared in prediction accuracy of the presence of malaria 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.Below is a table of that shows the count of children who have and do not have a $BEDNET$ after imputating the missing data via clustering.

No Bednet	165
Has Bednet	640

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, 63.3975373% 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 70.7% of the time.

Using Mice Package to Imputate Data

Two methods within the Mice Package were used to imputate the missing data in $BEDNET$. 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")

Below is a table of that shows the count of children who have and do not have a $BEDNET$ after imputating the data via predictive mean matching 5 times.

No Bednet	Has Bednet
220	585
213	592
228	577
206	599
184	621

After 100 iterations, the model with the missing data imputated via price mean matching predicted the true test data set on average, `r round(mean(apply(TEST_ACCURACY, 2, mean)) * 100,2)`% of the time.

Part II: Mice with Bayesian Linear Regression

Below is a table of that shows the count of children who have and do not have a $BEDNET$ after imputating the data via Bayesian linear regression 5 times.

Bad Value	No Bednet	Has Bednet	Bad Value
1	221	562	21
239	556	10	239
3	239	546	17
230	563	12	230
229	562	14	229

The table has values that are implausible (-1, 2), although a small subset of the data for some imputations.

After 100 iterations, the model with the missing data imputated via price mean matching predicted the true test data set on average, % of the time.

Between the two "mice" methods, both imputations schemes were similar, but the Bayesian linear regression method was slighlty better in predicting the true test data. However, as noted by the table, the imputated data produced values that were not possible (-1,2), which would make interpreting the data difficult. If the end goal is prediction accuracy, then Bayesian linear regression may be the best option. 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
## -1.1451  -0.8969  -0.7500   1.3503   1.7933
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.34434    0.96997  -0.355 0.722588
## AGE          0.25960    0.06906   3.759 0.000171 ***
## GREEN       -0.01914    0.02307  -0.830 0.406725
## PHC         -0.46745    0.16927  -2.762 0.005753 **
## Imp51       -0.05028    0.17454  -0.288 0.773300
## ---
## 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: 974.02  on 800  degrees of freedom
## AIC: 984.02
##
## 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.