**Predictive performance of a hybrid technique for the multiple imputation of survey data**

**Keywords:** Survey data; Multiple Imputation; Complex dependencies; Hybrid; Dirichlet process prior distributions, predictive performance.

1. **Introduction**

Analysis of data for scientific investigations becomes complicated, biased and less efficient in presence of missing information. In recent decades, lots of effort has been made in development of statistical methods to carter missing data. General reasons for the missing datasets include data entry errors, system failures, nonresponse etc. Missing categories are: (i) missing completely at random (MCAR), (ii) missing at random (MAR), (iii) missing not at random (MNAR) In many survey based studies, the logistic regression model is used to investigate the effect of various background characteristics (e.g. demographics, age, education, motherhood and recent births etc.) on a binary outcome variable such as breast feeding practices. This model can be difficult to apply when the confounding variables are missing. According to studies estimation of regression coefficients can be biased when ad hoc methods and complete case analysis for handling missing data are used. For handling missing data, Rubin[1] has purposed multiple-imputation (MI). Various MI approaches based on EM algorithm, fully Bayesian approach, maximum likelihood, a mixture of independent multinomial distributions approachand weighted estimating equation has been developed for fitting a logistic regression model when some covariate values are missing at random. A popular chained equations model MI approach called Multivariate Imputation by Chained Equations (MICE) fails to perform sometimes due to computational efficiency, complex dependency structure among categorical variables and high percentage of missing information in large scale survey data whereas, MI approach called Dirichlet Process Infinite Mixtures of Products of Multinomials (DPMPM) purposed by Si and Reiter[2], seems to perform very well for categorical variables having complex dependencies. Since the MI approach DPMPM is limited to categorical variables and complex dependencies structure can be problematic to be identified by MICE sometimes, it is not possible to get complete dataset when both approaches are applied separately.

**2. Methods**

We develop a Hybrid Multiple-Imputation (HMI) approach for handling data for the problem described above. We purpose to apply DPMPM MI approach to impute categorical variables having complex dependencies and for the same dataset regular MICE is used to create imputations for continuous variables where categorical variables are imputed beforehand. HMI method enables us to utilize properties of DPMPM MI approach and simplicity of MICE to obtain complete dataset where we have missing values not in categorical variables but in continuous variables too. Hence, providing a flexible and practical HMI approach to obtain complete data, which cannot be possible to obtain when both MI approaches are applied separately.

**2.1 Proposed hybrid architecture**

The proposed missing data imputation approach is a 3-stage approach. Step 1: Multiple complete versions (Imp.cat) of categorical variables were generated utilizing R package “NPBayesImpute” under a fully Bayesian joint modeling approach. Step 2: Multiple incomplete versions *(Miss.num)* of continuous variables were created and were combined with multiple complete versions of *(Imp.cat)* resulting in m datasets where values in the continuous variables may be missing and values in the categorical variables are imputed. Step 3: The continuous variables in the m datasets were imputed using R package “mice” such that the means of the draws from the posterior predictive distribution of the unobserved data depend on the data imputed by fully Bayesian joint modelling (DPMPM) multiple imputations. This process is repeated using different MICE algorithms.

|  |
| --- |
| **Algorithm 1:** Hybrid MI andholdout cross validation for a logistic regression model |
| Require: *P nxp* matrix with incomplete data1. *Miss.cat , Miss.num* ← Initial division of *p* variables into factor and numeric subsets
2. **for Z= 1, …,z do**
3. **for** *m= 1, …,M* **do**$ $
4. *Imp.*$P\_{cat\_{m}}^{z}$← Imputation using NPBayesImpute for  *Miss.cat*
5. *Imp.*$P\_{cat\_{m}}^{z}$$Miss.\_{num\_{m}}^{z}$← Combining Imp.$P\_{cat\_{m}}^{z}$and $Miss.\_{num\_{m}}^{z}$ to generate partially imputed dataset
6. $Imp\_{m}^{z}$← Imputing *Imp.*$P\_{cat\_{m}}^{z}$$Miss.\_{num\_{m}}^{z}$using MICE i.e. $f( Miss.\_{num\_{m}}^{z}⃒Imp.P\_{cat\_{m}}^{z})$
7. $Imp\_{m}^{z}$← Final imputed data set
8. $Imp\_{testing\_{m}}^{z}$, $Imp\_{training\_{m}}^{z}$ ← Divide matrix$Imp\_{m}^{z}$into testing and training subsets
9. $ P(y=1ǀ x\_{1,…,}x\_{p})\_{m }^{z}= 1/(1+e^{-(a+\sum\_{k}^{}(b \_{k\_{m}}^{z}x\_{k\_{m}}^{z} )}$ *)←* Train a GLM model on $Imp\_{training\_{m}}^{z}$
10. *Ҏmz* ← Make prediction on $Imp\_{testing\_{m}}^{z}$
11. $AUROC\_{m}^{z}$ ← AUROC curve based on *Ҏmz*
12. **end for**
13. $\overbar{AUROC}=\frac{\sum\_{m=1}^{M}(AUC\_{m }^{z})}{M}$ ←Pooled AUROC curve$ $
14. **end for**
 |

**2.2 Cross validation**

Holdout cross validation is used to access the predictive performance of a logistic regression model for binary response which is a generalized linear model (GLM).Test and train datasets are generated randomly using a *70%/ 30%* split.

* 1. **Statistical Analysis**

The method proposed is used to analyse data from a MICS 2014 women's data. The association between various factors and breast feeding practices among women in Punjab is investigated. The relationship between binary response (Ever breastfeed) and explanatory variables is modelled using a generalized linear model (GLM). The accuracy predictive distributional model is assessed by the area under the receiver operating characteristic (ROC) curve, known as (AUROC) and the results obtained under purposed and existing MI methods for large spectrum of data characteristics are compare.

**2.4 Evaluation of Performance**

The area under the receiver operating characteristic (ROC) curve, known as (AUROC) is used to compare different MI methods. Let Y be a random variable representing the classes (positive ⊕, negative$ ⊖ $). Ŷ represents the prediction for a randomly drawn sample. An ROC curve is made by plotting sensitivity ($\frac{TP}{TP + FN}) $on y-axis and 1- specificity ($\frac{TN}{TN + FP})$ on x-axis. Where TP =True positive = P (Ŷ = ⊕|Y = ⊕), FP=False positive = P (Ŷ = $⊕$|Y =$⊖$), TN=True negative = P (Ŷ = $⊖$ |Y =$⊖$) and FN=False negative = P (Ŷ$= ⊖$ |Y = ⊕).

The AUROC can be used to show the trade-off between the fractions (TP and FP) and the TP fraction at a specific FP fraction respectively. The AUROC is obtained by combining all the areas between x-axis and a line connecting two adjacent data points.

AUROC = $\sum\_{k=1}^{n}(X\_{k }-X\_{k-1 })\*(Y\_{k }-Y\_{k-1 })$.

**3. Results**

**Figure 1:** *Simulation study: Boxplots of pooled AUROC obtained by various MI methods for r=0.7.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.****imputations** | **Imp. Default**  | **Imp.Cart** | **Imp.PMM**  | **Imp.Hybrid.Cart**  | **Imp.Hybrid.PMM**  |
|  **Imp.2** | 15.12 mins | 15.74 mins | 25.52 mins | 13.51 mins | 15.50 mins |
| **Imp.5** | 36.47 mins | 38.08 mins | 59.48 mins | 30.99 mins | 36.45 mins |
| **Imp.10** | 1.13hours | 1.21 hours | 1.83 hours | 1.04 hours | * 1. ours
 |

**Table 1:** *Simulated data (r=0.7): Time taken for various MI methods.*



**Figure 2:** *Real data: Boxplots of pooled AUROC obtained by various MI methods.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.** **imputations** | **Imp.Default** | **Imp.Cart**  | **Imp.PMM** | **Imp.Hybrid.Cart** | **Imp.Hybrid.PMM** |
| **Imp.5** | 1.93 days | 1.88days | 1.80 days | 10.78 hours | 11.59 hours |

**Table 2: Real data:** *Time taken for various MI methods.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.****imputations** | **Imp.Default** | **Imp.Cart**  | **Imp.PMM** | **Imp.Hybrid.Cart** | **Imp.Hybrid.PMM** |
| **Imp.5** | 0.94 | 0.63 | 0.82 | 0.88 | 0.88 |

**Table 3:**  *Real data: Median pooled AUROC curve for various MI methods.*

4. **Conclusions**

We purposed a computational efficient hybrid MI method. Our purposed method makes it possible to MI both types of variables (categorical with large numbers of outcomes and continuous) in survey data in the presence of complex dependencies. This method combines MI by chained equations and mixtures of multinomial. To implement this method no knowledge of complicated models is required. The dependence among continuous and categorical variables can be made through an easy engine. Better predictive performance with minimum computational time as compared to the existing methods is partly achieved in simulation studies.

**5. References**

[1] D.B. Rubin, (1987), “Multiple Imputation for Nonresponse in Surveys”. 80, 353-362.

 [2] Si, Y., Reiter, J. P. (2013), “Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys,” Journal of Educational and Behavioral Statistics, 38, 499–521.