VOL. 13, NO. 4, FEBRUARY 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

THE USE OF FULLY CONDITIONAL SPECIFICATION OF MULTIPLE IMPUTATION AND INVERSE PROBABILITY WEIGHTING TO MODEL THE PULMONARY DISEASE OCCURRENCE IN SURVEY DATA WITH NON-RESPONSE

Aluko O. and Mwambi H.
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X01 Scottsville, Pietermaritzburg, South Africa
E-Mail: llaluko@gmail.com

ABSTRACT

Incomplete data is a frequent occurrence in many research areas especially cross sectional survey data in epidemiology, health and social sciences research. In this paper, the effect of missing observations were accounted for by using multiple imputation (MI) and inverse probability weighting (IPW) methods. Generally, multiple imputation has the ability to draw multiple values from plausible predictive distribution for the missing values. However, under the inverse probability weighting procedure the weights are the inverse of the predicted probabilities of response estimated from the missing ness models of incomplete variables. A simulation study is conducted to compare methods and demonstrate that a cross sectional survey data can be used to mitigate bias induced by missing data. The application and simulation results show the benefit of the IPW compared with the MI. The former performs well but not as the latter.

Keywords: fully conditional specification, multiple imputation, inverse probability weighting, non-response, survey data.

1. INTRODUCTION

Non-response in the population surveys are becoming a great challenge especially to the health sector. However, the common and earliest technique to handle such problem is the use of complete case analysis which most statistical software applications have capabilities to do. This technique is called the listwise deletion and many statisticians and clinicians adopt the approach simply because it deletes the items with incomplete information. This approach is valid when the missingness assumption is missing completely at random (MCAR) and the investigation of this method is detailed in[1] However, justifying this approach in real applications may be somehow complex. In fact, there are many ad hoc methods that handle the missing observations especially where missing values are substituted plausibly, such as last observation carried forward (LOCF), the mean and regression predictions (i.e. single imputation). One of the major difficulties of these methods is when there is high percentage of incomplete observations as explained by [1, 2]. Biased parameter estimates and loss of relationship among variables may be possible especially when the complete data does not give a true representation of the population target, and in such situation the MCAR assumption is violated. According to [2], the single imputation method and other related methods may produce uncommon small standard errors because uncertainty about the imputation values are not emphasized. The purpose why survey data have incomplete data are many [1, 2, 3, 4, 5]. Incomplete information arose when an element in the planned population is mistaken excluded on the survey's sampling frame, this results in what is called non-coverage. Explanations from authors in [1, 5] 6say that elements may have zero probabilities of being included in the sample population. Thus, total/unit non-response is defined as when a sampled person fails to take part in the survey. However, the occurrence of total non-response is recorded when an individual fails to actively participate in the survey due to some reasons ranging from the sensitivity of the questions, language of communication to unavailability on the day of the interview as stated by [7]. Household surveys collection can record success when all the target persons are present on the day scheduled for the interview. The failure of the selected individual to provide an accurate response(s) to one or many question(s) is called item non-response in a survey data. The partial non-response is simply defined as when a non-response falls between unit and item non-response. This is certain, for illustration, when a respondent cuts off the phone conversation in the middle of the interview or in a multiphase survey, the respondent provides data for some but not all phases of data collection [1, 4, 6].

The classification of missing data are under three missing data mechanisms namely: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Different techniques has been developed that handle non-response in a survey data. But the form of remedial measures depends on the mechanism that generated the missing data. Many of these techniques that handle the non-response range from the traditional method like deletion, weighting adjustments to modern one which is imputation method. For total non-response and non-coverage; weight adjustments are mostly appropriate. Individuals with complete data receives greater weights, so as to compensate for inadequacies coming from non-respondents. In the situation of non-coverage, the use of weighting adjustments become imperative as it handles external data sources especially when there is no complete information from the sampled individual. In item non-response, one of the appropriate ways to compensate for missing data is through the

multiple imputation. Incomplete observations are filled plausibly with imputed values. The two techniques just stated are used to compensate for partial non-response. In this research study, we embraced the use of multiple imputation and inverse probability weighting methods to handle the incomplete observations, for the purpose of obtaining estimates that are unbiased in modeling the prevalence of chronic obstructive pulmonary disease in three countries from Southern America, namely Argentina, Chile and Uruguay. We used socio-demographic risk factors and illnesses variables as covariates in the analysis. Multiple imputation is a Monte Carlo technique and utilizes a Bayesian inference paradigm as indicated by [5]. The complete datasets are then analyzed using an approach that makes the imputed datasets look like a real datasets as if they were observed from the non-respondents. Furthermore, unbiased parameter estimates and confidence interval are obtained from m complete datasets that incorporate missing data uncertainty. The parameter estimates and their variances that account for both the within-imputation and across-imputation variability; are derived from the mean of multiple imputed estimates obtained from the multiple analyses [3, 8, 9, 10, 11]. However, correct specification of the imputation model is important and is based on this MAR assumption. In addition, to account for variability due to the missing observations multiple imputation accounts for this in the variance formula.

One version of the Bayesian approach is that which employs the fully conditional specification (FCS) method to impute observations; from the posterior distribution of the missing data given the observed data as utilized by [12].

Inverse probability weighting (IPW) is one of the methods that can also reduce the estimation bias. In this method, obtaining the weight of the complete case is done by the inverse of the probability of being complete case. Another usefulness of the inverse probability weighting is for correcting the unequal sampling fractions. When a survey is conducted, the sample is expected to be representative, that is, everyone is equally likely to be sampled, few or no individuals with rare or unusual characteristics will be chosen [13]. However, interest may be on such individuals. In order to ensure that adequate numbers of individuals are sampled, sampling weights are employed. In the population, every individual is given a sampling weight and the probability that such individual is chosen is proportional to this weight. As outlined in [13], in such an approach the sample estimates of population parameters may be biased, because the sample is slightly different from the population.

In our study, we check that underestimation of the occurrence of chronic obstructive pulmonary disease may likely occur due to dependent-item non-response. Therefore, the incomplete data were adjusted and the occurrence of the disease in the survey data was also re-estimated. Using IPW method adjusts for non-response by specifying a regression model for the missingness mechanism given fully-observed covariates. To obtain valid inferences, this method relies on two assumptions: it assumes that the missingness process is independent of the fully-observed covariates; and relies on a correct specified regression model for the missingness process. In order to harness the strength of the second assumption; we compare it with the correct-specification of MI (FCS). When both models are correct, the advantage of MI(FCS) is that it is efficient than IPW because the former uses the entire sample while the latter uses the complete case.

2. MATERIAL AND METHODS

We used the research data obtained from "de Excelenciaen Salud Cardio-vascular para el Cono Sur" center for excellence in cardiovascular health for the southern cone (CESCAS). It is a country-level population-based household survey. The study was designed with the goal of examining the occurrence and to determine the dangerous factors, as well as the prevalence of cardiovascular and chronic obstructive pulmonary diseases in the general population.

The study was based on a sample of 8,000 non-institutionalized adult men and women between the ages of 35 and 74 years old (2000 per site) coming from Bariloche and Marcos Paz (Argentina), Temuco (Chile) and Canelones (Uruguay). In the study, specially trained interviewers conducted a household survey to uncover information about lifestyle (diet, physical activity, quality of life, smoking, alcohol consumption), socio-demographic data (age, sex, occupation, conditions of life), access to and utilization of health services (consultations, laboratory analysis, hospitalizations, etc.), risk factors and illnesses (high blood pressure, diabetes mellitus, cardiovascular and pulmonary problems, among others). Once the questionnaire was finished, the interviewers invited the participants to visit the assigned health centers to complete baseline evaluations (physical examination, blood test, electrocardiogram and spirometry). Two years from the initial visit participants were required to visit the clinic in the as-signed health centers. During the clinical visit, the activities conducted included; taking laboratory blood sample measurements (lipids total cholesterol, HDL-cholesterol, LDL-cholesterol and triglycerides, glucose and plasma creatinine), physical measurements (arterial tension, height, weight, waist and hip circumference), electrocardiogram (ECG). Between 5% and 10% of the samples selected at random were repeated with the purpose to quantify the variability among the samples.

In this study, the disease status is the outcome variable that has binary response with indication either a respondent's status of chronic bronchitis is positive, negative or missing. These two variables - demographic and lifestyle (were used as covariates with incomplete observations) are chosen as factors that can cause chronic obstructive pulmonary disease status. These risk factors included gender, marital status, agegroup, religion, blood cholesterol, asthma, chronic bronchitis, pneumonia and severe wheezing.

In Table-1, we display the response and covariates used and their percentages of missing

observations. The range of the incomplete observations highly vary from 0% to 87.80%.

2.1 METHODS

2.1.1 Multiple imputations

The estimate of target population is represented by θ. The statistic estimates θ if complete data were available is denoted by $\hat{\theta}=\hat{\theta}(Y_{obs}, Y_{mis})$ and the variance is represented by $U = U(Y_{obs}, Y_{mis}^{l})$. When Y_{mis} is accounted for we suppose to have $m \geq 2$ independent imputations, $Y_{mis}^{l}, \ldots, Y_{mis}^{m}$ and the estimates from the imputed datasets are calculated as $\hat{\theta}^{(l)} = \hat{\theta}(Y_{obs}, Y_{mis}^{l})$ and the estimated variances $U^{(l)} = U(Y_{obs}, Y_{mis}^{l})$, =1,…$m$.

The average mean estimate of θ is computed as follows:

$$\bar{\theta} = \frac{1}{m}\sum_{l=1}^{m} \hat{\theta}^{(l)} \quad (1)$$

In addition, we obtain the standard error of $\bar{\theta}$ as the square root of the total estimated variance given by

$$T = (1 + m^{-1})B + \bar{U} \quad (2)$$

the between-imputation variance (B) given by

$$B = \frac{\sum_{l=1}^{m}(\hat{\theta}^{(l)} - \bar{\theta})}{m-1}$$

and the within-imputation variance ($\bar{U}$) given by

$$\bar{U} = \frac{\sum_{l=1}^{m} U^{(l)}}{m}$$

The confidence interval (CI) for the population quantity, θ from the combined multiple imputed estimate is computed using its standard error of $\bar{\theta}$ and critical value from the student's t-distribution as CI(θ)= $t_{\widetilde{v,mi}}$ where $\tilde{v}_{mi}$, are the required degrees of freedom as detailed in [1].

2.1.2 Inverse probability weighting

The IPW also adjusts for item non-response by creating from the complete cases the so called pseudo-population. In this case individual weights are obtained by the inverse of the conditional probability of being observed given fully observed predictors. In the result of pseudo-population, the participants' responses with complete data represent themselves and those with the same features who had incomplete information on the variable of choice, [14]. In other words, in the absence of model mispecification and under the missing at random assumption, missing data information in the pseudo-population is a chance mechanism unrelated to the observed or unobserved information, as stated by [15]. In the complex sampling framework, the inverse probability weights are modified to adjust simultaneously for item non-response and the probability of being chosen into the study population respectively [14]. As earlier stated by [17] in the missing covariate context, the final weight W_i^* W_i for each individual i is constructed by multiplying the inverse probability $\bar{W}_i = \frac{1}{\pi_i(\widetilde{M}_i,\widehat{\alpha})}$ by the survey weight $W_{i,s}$. The maximum likelihood estimate predicting that the outcome is observed is represented by $\pi_i(\widetilde{M}_i, \widehat{\propto})$. That is $W_{i,s}^{*} = \widetilde{W}_i^{*} W_{i,s}$. According to [14], the resulting inverse probability weighted regression estimator is given by the weighted sample average. Using the IPW procedure discussed in [13], the analysis model is also fitted only to the complete cases, but some complete cases receive more weight than the other.

3. SIMULATION STUDY

The importance of the simulation study is to examine the techniques to handle incomplete observations, and explore the performance of such methods under different missing data conditions. In this study, the focus is on the intermittent missing data pattern.

3.1 Data generation, simulation designs and analysis of the simulated data

We simulated cross sectional binary datasets to mimic the original dataset and introduced different missing rates. For each of these different cases, we simulated 1000 datasets based on a logistic regression model scheme of equation (3) for sample sizes N= 100, 200, 500. The cross sectional binary outcomes were generated following a model with a linear combination of the predictor variable:

$$\text{logit}[P(Y_{ij} = 1)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) \quad (3)$$

where we assume an underlying response variable y of the type similar to the real application study. The response is defined based on some underlying assumptions that follow linear regression model. The null model effect is β_0, while β_1 and β_2 are the main effects for the variables x_1 and x_2 respectively for individual i with their interaction effect captured by β_3. We generated a dataset that assume a vector of covariates variables which combine both the binary and continuous variables. We simulated two different covariates: x_2 from Bernoulli distribution with probability of success equals to 0.5, x_1 from Uniform distribution. For the purpose of simulation study, we used the parameters, $\beta_0 = 1$; $\beta_1 = 1$; $\beta_2 = 0{:}07$ and $\beta_3 = 0{:}25$. In this current work, the simulation model is explicitly written as

$$\text{logit}[P(Y_{ij} = 1)] = \beta_0 + (-1)x_1 + 0.07x_2 + (-0.25)(x_1 * x_2) \quad (4)$$

When the logit link function is inverted it leads to conditional binary logistic regression which is the probability of the event occurring as a function of covariates, thus equation (3) can be written equivalently as

$$P[P(Y_{ij} = 1)] = \frac{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2))} \quad (5)$$

Full datasets were generated and later we assumed a MAR model of missingness. The missing rates were not the same.

For the purpose of comparison, a large number of imputations is advised (e.g., M=20) and supported by the current available computational power. In our study, we performed M=20 imputations. This relatively high value was chosen to account for the relatively large fraction of missing data and to limit the loss of power for testing any associations of interest [18]. In order to compare the performance of the methods, we used bias and mean squared error (MSE).

4. RESULTS

4.1 Simulation results

Table-2 shows the outcome of the simulation study for the MI and IPW methods in terms of bias and MSEs, under N=100, 200 and 500 sample sizes. The missing rates of 10%, 30% and 50% represent low, moderate and high missing entries. In the table, large bias and MSE are shown in bold. From this table, under the sample size of 100 we observed that for low and moderate missing rates MI produced more unbiased estimates than the IPW. In the case of MSEs, both methods were comparable, except for β_3 that produced high values for all levels of missing rates and methods. Furthermore, under the sample size of 200; with low missing rate the MI yielded more unbiased estimates than the IPW, except for the moderate missing rate. The performance of the MSEs were closer to each other. However, under the sample size of 500 for all levels of missing rates the MI produced more unbiased estimates, except for β_3 for each case. As we increased the sample size and the missing rate, it is observed that the values for the MSEs also reduced. The findings reveal that the MI performs better than the IPW. The results also mean that studies should be carefully planned and designed putting remedial measures to reduce the rate of missing values.

4.1.1 Application results

4.1.2 Results from the application analysis

In Table-3, we present the results from each method. The overall estimates from the MI and IPW methods are not very different, but the former produces smaller standard errors than the latter for all covariates. This displays the superior efficiency in the multiple imputation method over the inverse probability weighting method in real application. At the 5% level, covariates associated with single marital status and living with a partner/divorced/widowed are non-significant under the two methods, whereas the covariates associate with asthma, chronic bronchitis pneumonia and severe wheezing are significant under the multiple imputation method, but chronic bronchitis and severe wheezing are non-significant under the inverse probability weighting method.

4.1.3 Logistic regression results

In Table-4, we display the adjusted odds ratio estimates of the logistic regression models for the two methods. The purpose of reference level was to ensure estimation and interpretation. Furthermore, the odds ratios help to observe the multiplicative effect of each level and the possibility of having pulmonary disease as a predictor in relation to a reference level that controls for the effect of the other predictors in the model.

It is obtained from the results that the risk factors of pulmonary disease is slightly lower among singles (OR=0.595, 95% CI=0.312-1.134) under the MI than (OR=0.793, 95% CI=0.316-1.994) under the IPW. However, the effect is not significant, so the approaches agree. It is also less among the married (OR = 0.462, 95% CI = 0.293-0.729) under the MI and (OR= 0.459, 95% CI=0.217-0.927) under the IPW. This low trend is also recorded for living with a partner (OR=0.635 95% CI=0.338-1,195) under the MI and under the IPW (OR=0.741 95% CI=0.260-2.109). In addition, it is observed that the confidence interval between the singles and living with a partner is 1 because there is no significant difference that they will be diagnosed for pulmonary disease, except the married. The \No" response in asthma category are close to each other (OR=0.574, 95% CI=0,306-1,075) under the MI and (OR=0.517, 95% CI=0.335-0.797) under the IPW analysis controlling for other covariates in the model. The explanations are the same for religion and pneumonia, but different for severe wheezing.

However, the "No" response in the chronic bronchitis status are significant under the two techniques. But the approach is different in the blood cholesterol status as the both approaches are non-significant. Furthermore, in the age group status the risk factors of pulmonary disease increases with the age of both males and females. In addition, the age group sub-level of 55-64 years is significant under both approaches and ceased to be non-significant in other sub-levels. The findings support the general belief that both males and females are diagnosed with chronic obstructive pulmonary disease.

5. DISCUSSION AND CONCLUSIONS

The survey logistic regression is fitted and the results obtain show dissimilarities in the parameter estimates in the methods used. Nevertheless, multiple imputation may be preferred to inverse probability weighting, but it may not be well-suited to some conditions where large numbers of variables predict both the missing variables of interest and missingness itself [18]. The results reveal that the IPW performs so closely to the MI using FCS. The strength of this research lies on the use of the MI method for imputing missing values in chronic obstructive pulmonary disease study. Incomplete data are not avoidable, pervasive and could produce biased estimates, if it is not handled carefully. Using an appropriate statistical method to analyze disease measures of focus and its variability can validate the inferences. A possible area of extension is the use of sensitivity analysis to analyze binary cross sectional survey data.

Table-1. Frequencies and percentages of missing values in each variable.

Variable	Frequency of missing values	% of missing values
COPD	40	0.53
Gender	0	0.00
Marital Status	5	0.07
Agegroup	0	0.00
Religion	4	0.05
Blood Cholesterol	5	0.07
Asthma	36	0.48
Chronic Bronchitis	51	0.68
Pneumonia	63	0.84
Severe Wheezing	6612	87.8

Table-2. Bias and mean squared error (MSE) estimates for multiple imputation and inverse probability weighting methods, under MAR mechanism over 1000 samples: N=100,200 and 500 individuals.

Sample size	Missing rate	Par	MI		IPW	
			bias	MSE	bias	MSE
100	10%	$_0$	-0.1557	0.1184	0.1634	0.1215
		1	-2.0161	6.4840	-2.1418	6.4843
		$_2$	-0.0030	0.0020	-0.0031	0.0020
		3	0.0856	0.0634	0.0443	0.0466
	30%	$_0$	-0.0171	0.0787	0.0913	0.1152
		1	-2.1133	6.4499	-1.6610	5.4304
		$_2$	0.0006	0.0016	-0.0015	0.0020
		3	0.0962	0.0570	0.0301	0.0647
	50%	$_0$	0.1358	0.1448	0.1436	0.1465
		1	-2.7397	10.3119	-3.3642	13.7106
		$_2$	-0.0030	0.0028	-0.0030	0.0028
		$_3$	0.1387	0.0818	0.0708	0.0594
200	10%	$_0$	-0.0254	0.0685	-0.0887	0.0759
		1	-1.2168	3.6496	-1.6535	4.4260
		2	0.0009	0.0014	0.0021	0.0014
		$_3$	0.0244	0.0538	0.0014	0.0413
	30%	0	0.0214	0.0746	-0.0736	0.0821
		1	-1.7036	4.4089	-1.6581	4.2745
		$_2$	-0.0001	0.0015	0.0018	0.0016
		$_3$	0.1065	0.0152	0.0021	0.0370
	50%	$_0$	-0.0123	0.0957	0.0160	0.0905
		$_1$	-1.0318	2.6379	-1.0534	3.2890
		$_2$	0.0008	0.0020	-0.0003	0.0019
		3	0.1568	0.0582	0.0300	0.0529

500	10%	0	-0.0482	0.0446	-0.0489	0.0454
		1	-1.0479	2.3600	0.0541	1.1093
		2	0.0013	0.0009	0.0012	0.0009
		3	0.0457	0.0299	-0.0041	0.0272
	30%	0	-0.0405	0.0520	-0.0934	0.0575
		1	-1.5439	1.1148	0.7541	1.9190
		2	0.0012	0.0011	0.0022	0.0010
		3	0.1035	0.0345	-0.0249	0.0042
	50%	0	-0.1410	0.0714	-0.1590	0.0831
		1	-1.5502	1.1141	2.4613	1.6264
		2	0.0033	0.0011	0.0032	0.0012
		3	0.1430	0.0447	-0.6437	0.0439

Table-3. Mean and subgroup estimates, standard errors and pr>|t| of chronic obstructive pulmonary disease prevalence for (a) multiple imputation and (b) inverse probability weighting.

Variable	**(a)Multiple imputations analysis**			**(b)Inverse probability weighting**		
	Estimate	**S.E**	**P r > |t|**	**Estimate**	**S.E**	**P r > |t|**
Mean	-0.1802	0.4512	0.6897	-0.6980	0.6519	0.2846
Gender						
Male	Ref					
Female	0.3131	0.2000	0.1174	0.3772	0.3550	0.2883
Marital Status						
Single	-0.4595	0.3297	0.1635	-0.2322	0.4691	0.6207
Married	-0.6671	0.2322	0.0035	-0.7781	0.3822	0.0421
Separated	Ref					
Living with a partner/						
divorced/widowed	-0.3850	0.3270	0.2399	-0.3003	0.5333	0.5735
Agegroup						
34-44	Ref					
45-54	0.2832	0.3000	0.3452	0.4192	0.4953	0.3076
55-64	0.6101	0.3001	0.0421	0.5523	0.4622	0.2324
65-74	0.6734	0.3184	0.0346	0.7258	0.5103	0.1553
Belief						
Religion	Ref					
No Religion	-0.1021	0.3292	0.7565	0.3643	0.5394	0.4996
Blood Cholesterol						
Yes	Ref					
No	-0.3015	0.2031	0.1375	-0.3331	0.3048	0.2747
Asthma						
Yes	Ref					
No	-0.6993	0.2271	0.0021	-0.5549	0.3198	0.0831
Chronic Bronchitis						

Yes	Ref					
No	-1.5337	0.2188	<.0001	-1.5904	0.3338	<.0001
Pneumonia						
Yes	Ref					
No	-1.4732	0.1988	<.0001	-1.0454	0.3220	0.0012
Severe Wheezing						
Yes	Ref					
No	-0.7438	0.2929	0.0137	-0.7487	0.4080	0.0668

Table-4. Adjusted odds ratio estimates for the survey logistic regression model under (a) multiple imputation analysis (b) inverse probability weighting.

	(a) Multiple imputations analysis		**(b) Inverse probability weighting**	
Variable	OR	95%CL	OR	95%CL
Gender				
Male	Ref			
Female	1.315	(0.898,1.926)	1.458	(0.727,2.927)
Marital Status				
Single	0.595	(0.312,1.134)	0.793	(0.316,1.994)
Married	0.462	(0.293,0.729)	0.459	(0.217,0.972)
Separated	Ref			
Living with a partner				
/divorced/widowed	0.635	(0.338,1.195)	0.741	(0.260,2.109)
Agegroup				
34-44	Ref			
45-54	1.274	(0.717,2.263)	1.521	(0.575,4.020)
55-64	1.726	(0.977,3.050)	1.737	(0.701,4.302)
65-74	1.856	(1.020,3.378)	2.066	(0.759,5.626)
Belief				
Religion	Ref			
No Religion	0.968	(0.518,1.809)	1.439	(0.499,4.149)
Blood Cholesterol				
Yes	Ref			
No	0.752	(0.507,1.115)	0.717	(0.394,1.304)
Asthma				
Yes	Ref			
No	0.517	(0.335,0.797)	0.574	(0.306,1.075)
Chronic Bronchitis				
Yes	Ref			
No	0.194	(0.127,0.295)	0.204	(0.106,0.393)
Pneumonia				
Yes	Ref			
No	0.253	(0.173,0.370)	0.352	(0.187,1.066)

Severe Wheezing				
Yes	Ref			
No	0.485	(0.311,0.757)	0.473	(0.212,1.053)

ACKNOWLEDGEMENT

We acknowledge the National Heart Lung and Blood Institute for giving us the permission to use the data from the Center for excellence in cardiovascular health for the southern cone.

REFERENCES

[1] Rubin DB. 1987. Multiple Imputations for Non-response in Surveys. New York, USA: John Wiley and Sons, Ltd.

[2] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG *et al*. 2009. Multiple Imputation for missing data in epidemiology and clinical research: potential and pitfalls. BMJ. 29; 338:b2393.

[3] Baraldi AN, Enders CK. 2010. An Introduction to Modern Missing Data Analysis. Journal of School Psychology. 48(1): 5-37.

[4] Kalton G, Brick JM. 1996. Handling Missing Data in Survey Research. Statistical Methods in Medical Research. 1; 5: 215-38.

[5] Little RJ, Rubin DB. 1987. Statistical Analysis with Missing Data. 2nd Edition. New York, USA: Wiley Series in Probability and Statistics.

[6] Lohr S. 2010. Sampling: Design and analysis, Second Edition. Boston, UK: Cengage Learning.

[7] Chinomona A, Mwambi H. 2015. Multiple Imputation for Non-response When Estimating HIV Prevalence Using Survey Data. BMC Public Health. 16; 15:1059.

[8] Heeringa SG, West BT, Berglund PA. 2010. Applied Survey Data Analysis. New York, USA: Chapman and Hall/CRC Press.

[9] Pigott TD. 2001. A Review of Methods of Missing Data. Educational Research and Evaluation. 7: 353-83.

[10] Schefer JL. 1997. Analysis of Incomplete Multivariate Data. New York, USA: Chapman and Hall.

[11] Schefer JL, Olsen MK. 1998. Multiple Imputation for Multivariate Missing Data Problems: A Data Analyst's Perspective. Multivariate Behavioral Research. 1; 33(4): 545-71.

[12] Patricia AB. 2015. Multiple Imputation Using the Fully Conditional Specification Method: A Comparison of SAS, Stata, IVEware and R, USA, Paper 2081-2015; pp. 1-17.

[13] Seaman SR, White IR. Review of Inverse Probability Weighting for Dealing with Missing Data. Statistical Methods in Medical Research.2013 Jun; 22(3) 278-295.

[14] Kathleen EW, Eric JTT, Megan M. 2010. Adjustment for Missing Data in Complex Surveys Using Doubly Robust Estimation: Application to Commercial Sexual Contact among Indian Men. Epidemiology. 21(6): 863-871.

[15] Hernan MA, Hernandez-Diaz S, RobinsJM. 2004. A structural approach to selection bias. Epidemiology. 15(5): 615-625.

[16] Moore CG, Lipsitz SR, Addy CL *et al*. 2009. Logistic Regression with Incomplete Covariate Data in Complex Survey Sampling: Application of rewighting Estimating Equations. Epidemiology. 20(3): 382-390.

[17] Kombo AY, Mwambi H, Molenberghs G. 2016. Multiple Imputation for Ordinal Longitudinal Data with Monotone Missing Data Patterns. Journal of Applied Statistics. 44(2): 1-18.

[18] Doidge JC. 2016. Responsiveness-informed Multiple Imputation and Inverse Probability-weighting in cohort studies with Missing Data that are Non-monotone or Not Missing at Random. Statistical Methods in Medical Research. 16; 0(0) 1-15. DOI: 10.1177/0962280216628902.