Cihan University-Erbil Scientific Journal (CUESJ)

Research Article

Likelihood Approach for Bayesian Logistic Weighted Model: Missing Completely at Random Case

Dler H. Kadir*

Department of Statistics, College of Administration and Economics, Salahaddin University-Erbil, Kurdistan Region - F.R. Iraq

ABSTRACT

Increasing the response rate and minimizing non-response rates represent the primary challenges facing researchers performing longitudinal and cohort research, especially can be seen in the area of pediatric medicine. When there are missing data, complete case analysis makes findings bias. Inverse probability weighting (IPW) is one of the many available approaches for reducing bias using complete case analysis. Here, a complete case is weighted by probability inverse of complete cases. The data were collected from the neonatal intensive care unit at Erbil maternity hospital from 2012 to 2017. In total, 570 babies (288 male and 282 females) were born very preterm. The aim of this paper is to use IPW on the Bayesian logistic model developmental outcome. The mental development index approach was used for assessing the cognitive development of those born very preterm. Almost half of the information for the babies was missing, meaning that we do not know whether they have cognitive development issues. We obtained greater precision in results and standard deviation of parameter estimates which are less in the posterior weighted model in comparison with frequent analysis. Further, research is needed using methods such as bootstrapping, sandwich, resampling, and jackknife methods for dealing with missing data.

Keywords: Likelihood, logistic weighting, missing data, preterm infants

INTRODUCTION

Increasing the response rate and minimizing non-response rates represent the primary challenges facing researchers performing longitudinal and cohort research. This can especially be seen in the area of pediatric medicine, whereby birth cohorts are often utilized for epidemiological research and randomized clinical trials of parental interventions. The contrition of participants lowers the strength of the research, as well as leading to bias in the results.[1] A non-response usually has more medical and socioeconomic risks and can have systematic variations regarding interest disorders, causing a biased estimate of an adverse outcome.[2] It is also frequently the case that data are missing in social research. They present ambiguities in statistical analyses of a different type to that of the usual imprecisions of samples, which become lower with increases in sample sizes. Therefore, greater assumptions are required to permit inferences to be reached. Over the past 10 years, there has been much theoretical research into ways of analyzing missing data sets.

Missing data can be defined as a value that is not recorded for a variable in the observation of Interest. Almost all branches of scientific research face this issue and will occasionally have to deal with missing data. Frequently, the missing data appear as incomplete data on a subject. Usually, the following analysis uses only a subject with a complete case measurement (complete case analysis). This proves expensive not only with regard to reductions in sample sizes, as an unclear variance estimate, lower statistical power, and the parameters which are thought to be possibly biased until a complete case analysis represents a random sampling of the focus population. To ensure that bias is considered, it is important to understand the mechanisms and the patterns of the missing data relevant to the research. The missing data mechanisms, as stated by Rubin, indicate the link between the missing values and the observed data.[3]

It is obvious that bias cannot only occur when there are a systematic dropout and non-response, it can additionally stem from a different sampling probability resulting from the study design. A non-participants rate of, perhaps, 10% might not produce a stronger bias unless the non-response more powerfully relates to the parameters of interest.[4]

Corresponding Author:
Dler H. Kadir, Department of Statistics, College of Administration and Economics, Salahaddin University-Erbil, Kurdistan Region - F.R. Iraq.
E-mail: dler.kadir@su.edu.krd

Received: Jul 7, 2020
Accepted: Jul 25, 2020
Published: Aug 13, 2020

DOI: 10.24086/cuesj.v4n2y2020.pp9-12

INVERSE PROBABILITY WEIGHTING (IPW)

An IPW method will directly model the missingness instead of modeling missing data observations.[5] IPW is occasionally termed "Inverse Propensity Weighting." When a probability score is projected for all subjects, the interest covariate observed values are weighted through the inverse of the relevant probability scores. Response probability is able to be modeled with logistical regression models and uses the inverse of the probability scores as one of the factors of adjustment. The IPW adjustment permits a greater number of variables to be employed for predicting non-responses. The more appropriate variable set needs to be used to discover the model, which is the best fit for predicting non-response. This results in a "smooth" adjustment factor distribution, with no need for choosing an arbitrary cut-point.[6] Yet, the IPW may possess an extreme value, causing an adjustment factor which might possess a highly covariate weight, and thus, a high covariate weight-adjusted estimates. This issue can be resolved by trimming the adjustment factor or trimming non-response adjusted weights. However, such remedies might increase bias possibilities.

WEIGHTED POSTERIOR DISTRIBUTIONS

From a Bayesian perspective, we can gather the posterior distribution $\pi(\theta|x)$ through the combination of two types of information concerning the random variable θ. One source is given by the observed data which are summarized by the likelihood function, and the other information source is the previous information regarding its distribution $\pi(\theta)$. Weighted posterior distributions can be defined through the replacement of the likelihood function by its IPW counterpart, as discussed in the previous section. The weighted posterior distribution is as follows:

$$\pi^{IPW}(\theta|x) \propto \pi(\theta) \times L^{IPW}(x|\theta) \quad (1)$$

Next, we propose IPWs evaluated as in the above equation of the form $\text{IPW}(x_i)=\text{PW}(x_i;\hat{\theta}_{IPW},\hat{F}_n)$. It is apparent that from the part of $L^{IPW}(x|\theta)$, we can obtain first-order property of the actual function likelihood under the model assumption; therefore, this has validity for Bayesian estimates in a standard manner. One benefit of weighting is that it uses other pseudo-likelihood functions, leading to posterior distributions which belong to the same family of those obtained through the use of the genuine likelihood function. Therefore, the weighted posterior distributions vary from the genuine posterior distributions for the estimated values. It can seem that there is a conflict between the method and a proper Bayesian perspective. This is because the weighted likelihood function is not immediately driven by a probabilistic model; rather, it is driven by adaptive weights. The data still tell a story; however, some values are not consistent with the required models, and we are unable to simply delete outliers, yet it still contributes to the posterior estimate.[7]

PRETERM DATA FOR USING LOGISTIC REGRESSION

The data were collected from the neonatal intensive care unit at Erbil maternity hospital from 2012 to 2017. In total, 570 babies (288 males and 282 females) were born very preterm. We have considered the infants born before 28 weeks. The mental development index approach was used for assessing the cognitive development of those born very preterm. Almost half of the information for the babies was missing, meaning that we do not know whether they have cognitive development issues. Now let

$$R_i = \begin{cases} 1 & \textit{if the data was collelected with probability } \mathrm{P_i} \\ 0 & \textit{if the there is no response} \end{cases} \quad (2)$$

- The procedure, therefore, involves: Fitting a binary logistic regression, responding to the research under observation (1 if observed, 0 if not) with the more appropriate variable set as an explanatory variable.
- Obtaining the fitted probability for all infants, P_i, $i \in (1,\ldots,N)$
- Calculation of the IPW for all infants $IPW_i = 1/P_i$) and uses *IPW* to fit weighted Bayesian logistic regression models using WinBUGS software.

Through the assumption that the posterior weighting model is correct, we can obtain consistent parameter estimates to know the effect of the outcome model. Yet, the major issue in weighted data analysis is that the weight is not representative of the actual subject number; however, only an expected number might be applicable if the statistical weight features every detail regarding the sampling probability. Identical samples appear from simple random sampling (whereby every individual of the same sizes is able to be sampled with an equal probability).[4] An additional issue with using IPW is where a missingness predictor distribution in full cases varies from incomplete cases. The IPW will then greatly vary since complete case analyses, where the missingness predictor observation is nearer the center of the observation distributions in the incomplete case which might obtain a larger weight. This will, therefore, cause a larger standard error.[8]

When the weight can account for the missing data, parameters must be predicted. The complete case data variance estimator makes the assumption that the weight is known and ignores any uncertainty in estimations about them.[9] Seaman *et al*. recommend the use of sandwich estimators in accounting for uncertainties in the weights. In reality, the true asymptotic uncertainty is frequently more when a true weight is utilized than when they are estimated. Thus, ignoring uncertainties in a fixed weight might cause a standard error.[8]

Here, the weights value was altered in all iterations using Markov chain Monte Carlo. The weight calculated from the variable weights sampled from posterior distributions instead of being obtained from a fixed value. Therefore, in all iterations of the outcome models, various weights values were gained. Ignoring any uncertainty between variable and fixed weights was examined to determine if there were any issues. Consequently, uncertainties were included in the weight value using variable weights.

In addition, weights were standardized (and multiplied by the number of observations/total number of the complete population). This results from the total of the weights being equal to the sums of the sample sizes. If the weights are not standardized, the sum of the weight is then equal to the entire

population instead of the total amount of observations, meaning that uncertainty (i.e., standard deviation and standard error) in the model without standardization will be underestimated. Thus, in this project, the weight is standardized.

In these analyses, complete cases are weighted by the inverse probability of there being complete cases. Two logistic regression models were operated at the same time for both outcome and response. A covariate of mother birth age, sex, gestational age, and birth weight z-scores was used in the response model. Let, R_i denotes the outcomes (response (infants where the developmental questionnaire was responded to)/ non-response [infants where the developmental questionnaire was not responded to]). The outcome was modeled with the assumption of the Bernoulli distribution.

$$R_I \sim bernoulli(q_I) \quad (3)$$

$$\text{where, } \text{logit}\left(\frac{q_i}{1-q_i}\right) = a_0 + aX$$

When statistical models for weight have been recognized, it can then be used in developmental delay model analyses to be run alongside the dataset. Nonetheless, weighting is vital and, thus, the amount of information is required to account for the weight relying on particular parameters being considered. Inference is normally changed by multiplying the contributions of every infant to a statistic by its statistical weight. For modeling outcome and weight, each individual's outcome variable value was multiplied by the individual's weight. IPW for each of the infants is calculated using 1/probability of responses; the IPW was then standardized. Therefore, the likelihood functions created in WinBUGS and is thus:

$$L_i = P^{y*IPW}(1-P)^{y*IPW} \quad (4)$$

P is the probability of babies surviving with developmentally delay issues, y is the outcome variable, and IPW is the IPW. Five hundred seventy infants featured in the response model. Yet, only 235 babies were used in the outcome model analysis (those babies known to be alive).

We have used IPWs as adjustment factors for babies about which we do not have cognitive developmental delay information. We put weights on the likelihood function using WinBUGS software. We repeated the analysis using frequentist logistic regression using variable weights. We obtained greater precision in results and standard deviation of parameter estimates as it is shown in Table 1, which are less in the posterior weighted model in comparison with frequent analysis.

Table 1: Posterior odds ratio and the standard deviation gained from the binary model using variable weights

Estimates of parameters	Odds ratio	SD	95% Credible interval
Constant	0.54	0.11	0.34, 0.79
Gestational age (centered)[a]	0.79	0.06	0.67, 0.94
Mother's age	0.987	0.02	0.91, 1.02
Birth weight z-score	1.05	0.04	0.97, 1.14
Female	0.52	0.18	0.27, 0.98

[a]Gestational age centered around 28 weeks

DISCUSSION

We have developed a likelihood-based approach to place weights which were calculated from response models into outcome models using logistic regression. It is noticed that unweighted models produce biased results. We have realized that a weighted model would provide more precise results in terms of odds ratio and uncertainty.

IPW is one of the many available approaches for reducing bias using complete case analysis. Here, a complete case is weighted by probability inverse of complete cases. Even though weighting is often used in designing and analyzing surveys, using it in analyses of missing data not as well recognized, as the IPW fact parameter estimates can be inefficient in regard to probability-based analysis.[10] The resulting estimate is frequently sensitive to the exact type of models for the response probabilities.[11] In reality, the greatest concern of the data analysis is the efficacy of the IPW approach regarding the likelihood approach. Although some methods have been suggested for obtaining more efficient and robust estimates, this approach has not yet been effectively developed to handle more than one situation. Alternative methods could have been used in this analysis, for example, doubly robust IPW and multiple imputation.

In addition, one disadvantage of this paper is the assumption that the weight is known; we have ignored uncertainties in the weight. To calculate IPW estimator standard error, methods like robust standard error created by weighted models could have been used. Other methods such as bootstrapping, sandwich, resampling, and jackknife methods are also possibilities, although these methods require intensive computations.[4,12]

REFERENCES

1. D. Wolke, B. Sohne, B. Ohrt and K. Riegel. Follow-up of preterm children: Important to document dropouts. *Lancet*, vol. 345, no. 8947, p. 447, 1995.
2. S. Johnson, S. E. Seaton, B. N. Manktelow, L. K. Smith, D. Field, E. S. Draper, N. Marlow and E. M. Boyle. Telephone interviews and online questionnaires can be used to improve neurodevelopmental follow-up rates. *BMC Research Notes*, vol. 7, p. 219, 2014.
3. D. B. Rubin. Inference and missing data. *Biometrika*, vol. 63, no. 3, pp. 581-592, 1976.
4. M. Hofler, H. Pfister, R. Lieb and H. U. Wittchen. The use of weights to account for non-response and drop-out. *Social Psychiatry and Psychiatric Epidemiology*, vol. 40, no. 4, pp. 291-299, 2005.
5. L. Lazzeroni, N. Schenker and J. Taylor. Robustness of Multiple-imputation Techniques to Model Misspecification. United States: Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 260-265, 1990.
6. B. L. Carlson and S. Williams. A Comparison of Two Methods to Adjust Weights for Non-response: Propensity Modeling and Weighting Class Adjustments. United States: Proceedings of the Annual Meeting of the American Statistical Association, 2001.
7. C. Agostinelli and L. Greco. Weighted Likelihood in Bayesian Inference. New York: Proceedings of the 46th Scientific Meeting

of the Italian Statistical Society, 2012.
8. S. R. Seaman and I. R. White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, vol. 22, no. 3, pp. 278-295, 2013.
9. S. R. Seaman, I. R. White, A. J. Copas and L. Li. Combining multiple imputation and inverse-probability weighting. *Biometrics*, vol. 68, no. 1, pp. 129-137, 2012.
10. D. Clayton, D. Spiegelhalter, G. Dunn and A. Pickles. Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society*, vol. 60, no. 1, pp. 71-87, 1998.
11. R. J. Little and D. B. Rubin. Statistical Analysis with Missing Data. Chichester: Wiley, p. 5, 1987.
12. L. H. Curtis, B. G. Hammill, E. L. Eisenstein, J. M. Kramer and K. J. Anstrom. Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. *Medical Care*, vol. 45, no. 10, pp. S103-S107, 2007.