Proceedings of the 3rd International Conference on Statistics: Theory and Applications (ICSTA'21)
~~Prague, Czech Republic~~ Virtual Conference – July 2021
Paper No. 117
DOI: 10.11159/icsta21.117

Estimation of Parameters of Logistic Regression with Missing Covariates via Joint Conditional Likelihood Method

Phuoc-Loc Tran[1,3], Truong-Nhat Le[2,3], Shen-Ming Lee[3], and Chin-Shang Li[4]
[1]Department of Mathematics, College of Science, Can Tho University, Viet Nam
tploc@ctu.edu.vn
[2]Faculty of Mathematics and Statistics, Ton Duc Thang University, Viet Nam
letruongnhat@tdtu.edu.vn
[3]Department of Statistics, Feng Chia University, Taiwan, R.O.C.
smlee@mail.fcu.edu.tw
[4]School of Nursing, The State University of New York, University at Buffalo, Buffalo, NY, USA
csli2003@gmail.com

Extended Abstract

Missing data occur when variables in observations have no data. Missing values regularly appear in social sciences, health and other scientific studies because of many reasons including, e.g., forgetting to answer questions, the experimental layout, data collection conditions or experimental time too long. In practice, the lack of data occurring in regression models, e.g., logistic regression models, is inevitable, and it possibly leads to some potential threats to a valid inference or decision. Dealing with missing values, researchers have used some deletion methods, namely complete-case (CC), semiparametric inverse probability weighting (SIPW) ([1] and [2]) and validation likelihood (VL) ([3]) methods, etc., or some multiple imputation (MI) approaches ([4] and [5]) for analysis. However, some literatures show that the methods based on deleting the un-observed values may lead to reducing the efficiency of estimation and the variance type from Rubin's MI method ([4]) may be under-estimated ([6] and [7]). Wang et al. ([3]) proposed the joint conditional likelihood (JCL) method, which is a semiparametric approach and uses both the validation and non-validation data, to estimate the parameters of a logistic regression model with a covariate missing. Lee et al. ([8]) presented a semiparametric estimation method and used the VL method and JCL method for logistic regression with both outcome and covariates missing at random. Hsieh et al. ([9]) also applied the methods of Lee et al. ([8]) to deal with logistic regression with outcome and covariates missing separately or simultaneously. Jiang et al. ([10]) proposed a stochastic approximation version of the expectation-maximization (EM) algorithm, which is based on Metropolis-Hastings algorithm, to perform statistical inference for logistic regression with missing covariates. In this study, the JCL method is proposed to estimate the parameters of a logistic regression model when two covariate vectors are missing separately or simultaneously by using one validation and three non-validations data sets to improve estimation. The asymptotic results of the JCL estimators are established under the assumption that all observable covariate variables including surrogates are categorical. Simulation results show that the proposed method is the most efficient compared to the CC, SIPW, and VL methods. The proposed methodology is illustrated by a real data example.

References

[1] C. Y. Wang, S. Wang, L. P. Zhao, and S. T. Ou, “Weighted semiparametric estimation in regression analysis with missing covariate data”, *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 512-525, 1997.

[2] S. Wang, and C. Y. Wang, “A note on kernel assisted estimators in missing covariate regression”, *Statistics and probability letters*, vol. 55, no. 4, pp. 439-449, 2001.

[3] C. Y. Wang, J. C. Chen, S. M. Lee, and S. T. Ou, “Joint conditional likelihood estimator in logistic regression with missing covariate data”, *Statistica Sinica*, pp. 555-574, 2002.

[4] D. B. Rubin, “Inference and missing data”, *Biometrika*, vol. 63, no. 3, pp. 581-592, 1976.

[5] D. B. Rubin, *Statistical analysis with missing data.* Wiley, 1987.

[6] D. B. Rubin, and N. Schenker, “Multiple imputation for interval estimation from simple random samples with ignorable nonresponse”, *Journal of the American Statistical Association*, vol. 81, no. 394, pp. 366-374, 1986.

[7] P. Righi, S. Falorsi, A. Fasulo, “Methods for variance estimation under random hot deck imputation in business surveys”, *Rivista di statistica ufficiale*, (1-2), pp. 45-64, 2014.

[8] S. M. Lee, C. S., Li, S. H. Hsieh, and L. H. Huang, “Semiparametric estimation of logistic regression model with missing covariates and outcome”, *Metrika*, vol. 75, no. 5, pp. 621-653, 2012.

[9] S. H. Hsieh, C. S. Li, and S. M. Lee, “Logistic regression with outcome and covariates missing separately or simultaneously”, *Computational Statistics and Data Analysis*, vol 66, pp. 32-54, 2013.

[10] W. Jiang, J. Josse, M. Lavielle, and T. Group, “Logistic regression with missing covariates—Parameter estimation, model selection and prediction within a joint-modeling framework”, *Computational Statistics & Data Analysis*, vol. 145, 106907, 2020.