Letter to the Editor

Cluster vs. Robust Estimation of Risk Ratio using Expanded Logistic Regression

Dear Editor,

In their article, Dr. Janani *et al*. discussed some methods to obtain adjusted risk ratio (RR).[1] Among options, authors mentioned the method named "expanded logistic regression", which consists in changing the original dataset by duplicating data of each individual that developed the outcome.[2,3] In this new data-set, duplicated observations are identified as non-outcome. The probability of success in the original dataset will be equal to the odds of success in the modified dataset, therefore, a logistic regression model fitting to the new dataset results in risk ratio instead of an odds ratio.

This simple tool could be useful for calculating adjusted RRs even using not sophisticated software. The main problem with this method is that the confidence intervals are wider than those observed with the reference methods.[4]

It was suggested that robust standard errors (SE) are needed to account for the within-subject correlation resulted from the duplicated observations.[1] However, robust estimation of SE does not solve that problem because the dependence of duplicate observations persists.

Recently, Dwivedia *et al*. proposed the cluster option to correct SE inflation associated with the ELR.[5] Thus, each case and its duplicate would be considered within a cluster, which allows estimating RRs considering the dependence of these observations.

In order to represent the differences between robust estimation of SE and cluster option for logistic regression, this communication present an analysis comparing these two methods against log-binomial regression.

For these purposes, it was used a simulated database, whose design was already described in another manuscript.[4] Table 1 presents the RR estimations obtained from four methods: log-binomial regression (reference method), the ordinary ELR, ELR with robust SE (ELR-Robust) and ELR using the cluster option (ELR-Cluster).

RRs obtained from every method were similar each other. However, confidence intervals obtained with ordinary ELR were much wider than those observed with log-binomial regression. Robust estimation does not correct this problem; in fact, robust confidence intervals are identical to those obtained with the ordinary ELR (Table 1). On the other hand, ELR-Cluster allows obtaining confidence intervals similar to those obtained from log-binomial regression.

Thus, this simulation verifies that the cluster option would be the right strategy to correct the effects of data duplication. Thereby, it would be possible to adequately calculate the precision of RRs obtained in clinical studies. This is essential to estimate the effect of risk factors, as well as, the impact of health interventions.

Fredi Alexander Diaz-Quijano[1]
Author's affiliation: Departamento de Epidemiologia, Faculdade de Saúde Pública da Universidade de São Paulo. São Paulo, SP, Brazil.
•**Corresponding author and reprints**: Fredi Alexander Diaz-Quijano, Departamento de Epidemiologia, Faculdade de Saúde Pública da Universidade de São Paulo, Av. Dr. Arnaldo, 715, Cerqueira César, CEP 01246-904, São Paulo, SP, Brazil. Tel: +55 11 946590193 – 11 30617738. E-mail: frediazq@msn.com

References

1. Janani L, Mansournia MA, Nourijeylani K, Mahmoodi M, Mohammad K. Statistical Issues in Estimation of Adjusted Risk Ratio in Prospective Studies. *Arch Iran Med*. 2015; 18(10): 713 – 719.
2. Schouten EG, Dekker JM, Kok FJ, Cessie SL, Van Houwelingen HC, Pool J, et al. Risk ratio and rate ratio estimation in case-cohort designs: Hypertension and cardiovascular mortality. *Stat Med*.1993; 12(18): 1733 – 1745.
3. Yelland LN, Salter AB, Ryan P. Relative risk estimation in randomized controlled trials: a comparison of methods for independent observations. *Int J Biostat*. 2011; 7(1): 1 – 31.
4. Diaz-Quijano FA. A simple method for estimating relative risk using logistic regression. *BMC Med Res Methodol*. 2012; 12: 14.
5. Dwivedia AK, Mallawaarachchi I, Lee S, Tarwater P. Methods for estimating relative risk in studies of common binary outcomes. *J Appl Statistics*. 2014; 41(3): 484 – 500.

Reply,

Dear Editor,

We would like to thank Dr. Diaz-Quijano for taking time and reading our article.

There are two types of robust standard errors available in statistical software, one assumes that the observations are independent; the other does not make any assumption about independence within cluster.[1,2] Of course, only the latter is appropriate for clustered data and for this reason, it is used more commonly than the former. As explained by Dr. Yelland *et al*.[3] and also in our article,[4] we should account for within-subject correlation in

Table 1. Comparison of methods to estimating adjusted risk ratios based on logistic regression.

Method	Risk Ratio for each Predictor (95%CI)	
	Predictor A	Predictor B
Log-Binomial	1.91 (1.59 – 2.29)	3.08 (2.56 – 3.71)
Expanded Logistic Regression (ELR)	1.91 (1.44 – 2.53)	3.08 (2.36 – 4.03)
ELR-Robust	1.91 (1.44 – 2.53)	3.08 (2.36 – 4.03)
ELR-Cluster	1.91 (1.57 – 2.32)	3.08 (2.56 – 3.71)

95%CI: 95% confidence Interval.
ELR-Robust: Expanded Logistic Regression with robust estimation of standard errors.
ELR-Cluster: Expanded Logistic Regression using option of cluster.

constructing confidence intervals for double of cases method[5] which was named as "expanded logistic regression" method in our paper. One possibility is using the cluster robust standard error which was simply called "robust standard error" in our paper in accordance with most statistical literature.[6]

For example, we can use PROC GENMOD in SAS and to compute robust standard errors, we need to use REPEATED statement and specify a unique subject identifier (id) as the SUBJECT in the REPEATED statement. In R software, we can use geeglm() function in library (geepack) and specifying a unique subject identifier (id) in it to get robust standard errors. Similarly, one can use "vce (cluster id)" option in Stata. We know that this approach produces cluster robust standard errors. In our large simulation study, we also used cluster robust standard errors using geeglm() function of R package geepack to account for within-duplicated subject correlation in the augmented dataset.

Leila Janani PhD[1], Mohammad Ali Mansournia MD MPH PhD[1,2]

Authors' affiliations: [1]Assistant Professor of Biostatistics, Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran. E-mail: Leila_janani@yahoo.com. [2]Assistant Professor of Epidemiology, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran. E-mail: mansournia_ma@yahoo.com

•**Corresponding author and reprints:** Mohammad Ali Mansournia MD MPH PhD, [2]Assistant Professor of Epidemiology, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran. E-mail: mansournia_ma@yahoo.com

References

1. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions." Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1976; 1(1).
2. Rogers, William. Regression standard errors in clustered samples. *Stata Technical Bulletin.* 1994; 3:13.
3. Yelland LN, Salter AB, Ryan P. Relative risk estimation in randomized controlled trials: a comparison of methods for independent observations. *Int J Biostat*. 2011; 7(1): 1 – 31.
4. Janani L, Mansournia MA, Nourijeylani K, Mahmoodi M, Mohammad K. Statistical Issues in Estimation of Adjusted Risk Ratio in Prospective Studies. *Arch Iran Med*. 2015; 18(10): 713 – 719.
5. Schouten EG, Dekker JM, Kok FJ, Cessie SL, Van Houwelingen HC, Pool J, et al. Risk ratio and rate ratio estimation in case-cohort designs: Hypertension and cardiovascular mortality. *Stat Med*.1993; 12(18): 1733 – 1745.
6. Harrell Jr FE. Jr. Regression modeling strategies. 2001: 179 – 213.