ARCHIVES OF IRANIAN MEDICINE

Open Access

Research Methods

Risk Ratio Estimation in Longitudinal Studies

Seyedeh Solmaz Talebi, PhD[1]; Kazem Mohammad, PhD[1]; Aliakbar Rasekhi, PhD[2]; Mohammad Ali Mansournia, MD, MPH, PhD[1*]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[2]Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran

Abstract

Longitudinal studies are very common in medical, behavioral, and interventional sciences. One measure of effect of interest in longitudinal studies is risk ratio, naturally estimated by log-binomial regression which suffers from convergence problems. Odds ratio (OR) does not approximate risk ratio (RR) well when the outcome is common, so alternative methods have been introduced in cohort studies with one follow-up visit. In this paper, we illustrate two simple methods: the COPY method and the modified log-Poisson regression for RR estimation in longitudinal data setting. Our unpublished simulation study on RR estimation in longitudinal data setting suggests that the COPY method performs well in terms of closeness of the RR estimate and true RR (mean square error) and so we suggest this method for RR estimation in longitudinal data setting.

Keywords: COPY method, Generalized estimation equations, Longitudinal data, Modified log-Poisson regression, Risk ratio

Cite this article as: Talebi SS, Mohammad K, Rasekhi A, Mansournia MA. Risk ratio estimation in longitudinal studies. Arch Iran Med. 2019;22(1):46–49.

Received: June 13, 2018, Accepted: July 4, 2018, ePublished: January 1, 2019

Introduction

Because of their specific features, longitudinal studies are very common in medical, behavioral, and interventional sciences.[1] In longitudinal studies, the outcome may be measured in several follow-up visits.[2] The main aim of such research is to estimate the effect of an exposure or treatment on an outcome over time.[3] Due to repeated measurements of outcome, correlation is observed among responses of a participant at different visits, which should be accounted for in the analysis; otherwise the standard error of effect estimates will be underestimated leading to *P* values which are too small and confidence intervals which are too narrow.[4,5] One measure of effect of interest in longitudinal studies with binary outcomes is risk ratio (RR). The natural model for estimating adjusted RR is log-binomial regression. Unfortunately, this model suffers from convergence problems.[6] For rare outcomes (approximately less than 10%), RR can be approximated with odds ratio (OR), but for common outcomes, OR overstates RR. In this paper, we illustrate two simple alternative methods, the COPY method and the modified log-Poisson regression, for estimating adjusted RR in longitudinal data setting.

The GEE Method

In 1986, Liang and Zeger introduced an estimation method, called generalized estimation equations (GEE), for analyzing non-normal longitudinal data.[7] This method is based on the quasi-likelihood approach in which only the mean (the regression model) and variance of outcome is specified and knowing the outcome's distribution is not necessary. In GEE, the quasi-likelihood approach is generalized to allow for a working correlation structure for example via a matrix that explains how outcomes in different visits are correlated. Moreover, cluster robust standard errors, based on the empirical variability of data, are used to account for the within-participant correlation.[8,9]

The working correlation structure has different types such as exchangeable, auto-regressive (AR) and unstructured. Exchangeable correlation means that correlation between each two visits is equal for example, the correlation between visits 1 and 2 is equal to the correlation between visits 1 and 3 and is equal to the correlation between visits 1 and 4. AR correlation is suitable for situations that correlation between adjacent visits is equal, but it decreases over time. For example: correlation between visits 1 and 2 is equal to correlation between visits 2 and 3 or visits 3 and 4, but correlation between visits 1 and 3 is weaker than correlation between visits 1 and 2. In an unstructured correlation, there is no structure in correlation between different visits.[10] The GEE method is not sensitive to the incorrect selection of the correlation structure type: if a wrong correlation structure is selected, the estimation of parameters is still unbiased assuming the regression model is correct, but precision will be reduced.[4]

Risk Ratio Estimation in Longitudinal Studies

Binary outcomes are very common in longitudinal studies, and logistic regression is commonly used for estimating the effect of exposure/treatment. For cohort studies with only one outcome measurement, the logistic regression

*Corresponding Author: Mohammad Ali Mansournia, MD, MPH, PhD; Assistant Professor of Epidemiology, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran. P.O. Box: 14155-6446, Tehran, Iran. Tel: +98-21-88989127; Email: mansournia_ma@yahoo.com

model assumes that logarithm of odds is equal to the linear combination of variables:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Trt + \ldots + \beta_p C_p \qquad \text{Model 1}$$

where *P* represents the risk of outcome during the follow-up, Trt denotes the treatment variable, *C*s represent confounders, and *β*s represent regression coefficients. The exp(β_1) is the treatment OR i.e., the ratio of the odds in the treatment group to that in the control group. The logistic regression model 1 can be generalized to longitudinal data as model 2 below:

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 Trt + \beta_2 C_1 \ldots + \beta_{p-1} visit \qquad \text{Model 2}$$
$$+\beta_p visit * Trt$$

where P_j represents the risk of outcome in follow-up visit j (varies from 0 to J), *Trt* denotes treatment variable, visit is follow-up visit, *C*s represent confounders, and *β*s represent regression coefficients. The main difference between models 1 and 2 is inclusion of visit and interaction term between visit and treatment which means the effect of treatment depend on follow-up visit i.e., exp(β_1) is the treatment OR in the first follow-up visit (j = 0), exp(β_1 + β_p) is the treatment OR in the second follow-up visit (j = 1), exp(β_1 + 2β_p) is the treatment OR in the third follow-up visit (j = 0), and so on. Another difference is that GEE method should be used for model estimation to take into account the within-subject correlation.

Although logistic regression is a convenient model frequently used in practice, its result is summarized as OR which cannot be easily interpreted by non-experts. In fact, OR is sometime misinterpreted as RR, the ratio of the risk in the treatment group to that in the control group, in published papers.[11,12] Moreover, OR also suffers from a mathematical peculiarity known as non-collapsibility[13] so that adjusted OR can be different from unadjusted OR in the absent of confounder. Both issues are not important as long as the risk of outcome is low (say below 10%) in all strata of treatment and covariates. However, the rare outcome assumption is violated in many randomized clinical trials, so the effect measure of interest is generally RR in cohort studies.[14,15] OR is only useful if it approximate RR.

RR can directly be estimated using the log-binominal regression.[16] In log-binominal regression model, the logarithm of risk is assumed to be equal to the linear combination of variables:

$$\log(P) = \beta_0 + \beta_1 Trt + \beta_2 C_1 + \ldots + \beta_p C_p \qquad \text{Model 3}$$
$$0 < P \leq 1$$

where as in the logistic regression model 1, *P* represents the risk of outcome during the follow-up **,** *Trt* denotes the treatment variable, *C*s represent confounders, and *β*s represent regression coefficients. The exp(β_1) is the treatment RR. The log-binomial regression model 3 can be generalized to longitudinal data as model 4 below:

$$\log\left(P_j\right) = \beta_0 + \beta_1 Trt + \beta_2 C_1 \ldots + \beta_{p-1} visit \qquad \text{Model 4}$$
$$+\beta_p visit * Trt \qquad 0 < P \leq 1$$

as model 2, P_j represents the risk of outcome in follow-up visit j (varies from 0 to J), *Trt* denotes the treatment variable, visit is follow-up visit, *C*s represent confounders, and *β*s represent regression coefficients.

Log-binomial regression model such as models 3 and 4, suffers from a structural problem. The left-hand side of model 3 or 4 cannot take a positive value but the right-hand side is unbounded. Thus, fitting log-binomial regression may lead to non-convergence without providing any treatment effect estimate.[17] This event is more likely for high values in the right side**.**

The COPY Method

In 2003, Petersen and Deddens introduced the COPY method for solving the convergence problems inherent in the log-binomial regression.[18] In this method, the original data is augmented with 1 copy of data in which the outcome status is reversed and frequency weight "c-1" is assigned to the original data set and frequency weight 1 is assigned to data with reversed outcome. Thus augmented data set include c copies of original data with reversed outcome status in just 1 copy. Then, the log-binomial regression is fitted to this augmented data set and the standard error is corrected by multiplying the apparent standard error by the square root of c. Petersen and Deddens showed if c was large enough, the model estimates obtained from the COPY method are close to real values. Their simulation study showed that the RR estimate is closer to true RR with c = 1000 than c = 100. Another simulation study by Lumley et al suggested c = 25 is sufficient for solving convergence problems in log-binomial regression.[19] The COPY method can be easily generalized to longitudinal data by fitting model 4 mentioned above to the augmented data set. The GEE method should be used for model estimation to take into account the within-subject correlation.

Modified Log-Poisson Regression

In 2004, Zou proposed a simple and effective method for adjusted RR estimation in cohorts with one follow-up visit. In this method, known as modified log-Poisson regression, a log-Poisson regression model is used for RR estimation and the standard error is corrected using the robust sandwich approach. The robust standard error is based on the empirical variability of the outcome and corrects the apparent standard error suggested by Poisson distribution. Unlike log-binomial regression, this method does not have convergence problems[20] but may generate predicted probability greater than 1.[21] Modified log-Poisson regression can be easily generalized to longitudinal data. Again, the

GEE method should be used for model estimation to take into account within–subject correlation.

Application

We fit different regression models mentioned above and compared the results using longitudinal mental depression data with 3 follow-up visits.[22] The data set includes 340 depressed patients who, according to initial diagnosis, were divided into two groups of mild and severe. In each group, patients randomly received the standard and new treatments. The outcome was disease status (1: abnormal and 0: normal), and was recorded in the first, second and the fourth week after receiving treatment. The risks were equal to 21%, 14%, and 8%, respectively.

To estimate RR for treatment, we fitted logistic regression, log-binomial regression, and the COPY method with different c (25, 100 and 1000and modified the log-Poisson regression. We included treatment (new treatment: 1, standard: 0), initial diagnosis (sever: 1, mild: 0), logtime (logarithm of visit in week based on 2), and interaction term between logtime and treatment in the model. The R codes for all analyses are available upon request.

Results

Table 1 contains the results of fitting models with treatment as the sole predictor in the model. Comparison of logistic and log-binomial regression models suggests that the RR estimate of the treatment effect from the former exaggerates the effect estimate from the latter which is unsurprising given that the outcome is not uncommon (>10% in the first and second follow-up visits). Confidence interval is also wider for logistic regression model than log-binomial regression. As shown in Table 1, the number of copy should be at least 100 for RR estimate from COPY method closely approximates the estimate from log-binomial regression.

Table 1. Unadjusted RR Estimates for Treatment Using Different Methods

Models	Risk Ratio	95% CI
Logistic*	0.49	(0.38–0.63)
Log-binomial	0.68	(0.60–0.78)
COPY method (c = 25)	0.71	(0.63–0.79)
COPY method (c = 100)	0.69	(0.60–0.79)
COPY method (c = 1000)	0.68	(0.60–0.78)
Modified log-Poisson	0.68	(0.60–0.78)

* For logistic regression unadjusted OR was reported.

Table 2 presents the results of RR estimate from different models adjusted for initial diagnosis in different visits. The RR and OR are almost equal in the first week after receiving treatment, because treatment has almost no effect in that time. However, as expected, the OR estimate exaggerates the treatment effect in later visits in which the treatment effect appears. The results of COPY method (c = 1000) are close to log-binomial regression (Table 2).

Discussion

In this study, we illustrated several methods including logistic regression, log-binomial regression, the COPY method and modified log-Poisson regression for estimating RR in longitudinal data setting. The natural model for estimating RR is log-binomial regression but it suffers from convergence problems. Unfortunately, logistic regression is often used to estimate the treatment effect on common binary outcomes in longitudinal data analysis and the resulting OR estimate is misinterpreted as RR.

In our unpublished simulation study on RR estimation in longitudinal data setting, we compared modified log-Poisson regression, the COPY method, and several other methods. We concluded that the COPY method is superior to other methods in terms of closeness of the RR estimate and true RR (mean square error). However, there is no simulation study comparing the COPY method and modified log-Poisson regression in the longitudinal data setting or more generally in clustered data setting. There is one simulation study comparing these two methods in a cohort study with one follow-up visit. This study concluded that the COPY method is preferred for RR estimation because it does not produce probability greater than 1 and has the smallest bias and mean square error.[21]

There are many methods for RR estimation in longitudinal studies.[14] We choose the COPY method and modified log-Poisson regression because they have generally worked well in simulation studies and are simple to implement in statistical software. However, there are time-varying confounders in many longitudinal studies and all of conventional statistical methods including the COPY method and modified log-Poisson regression fail to provide unbiased RR estimate in this setting if time-varying confounders are effected by prior treatment.[23] Causal methods including inverse probability-of-treatment weighting and parametric g-formula should be

Table 2. Adjusted RR Estimates[a] for Treatment in Different Weeks Using Different Methods

Models	RR in 1st Week (95% CI[**])	RR in 2nd Week (95% CI)	RR in 4th Week (95% CI)
Logistic[b]	1.06 (0.67 – 1.67)	0.38 (0.29 – 0.51)	0.14 (0.09 – 0.22)
Log-binomial	1.03 (0.92 – 1.16)	0.57 (0.50 – 0.66)	0.32 (0.24 – 0.42)
COPY method (c=25)	1.03 (0.92 – 1.16)	0.62 (0.55 – 0.70)	0.37 (0.29 – 0.48)
COPY method (c=100)	1.03 (0.92 – 1.16)	0.58 (0.51 – 0.67)	0.33 (0.25 – 0.44)
COPY method (c=1000)	1.03 (0.92 – 1.16)	0.58 (0.50 – 0.66)	0.32 (0.25 – 0.43)
Modified log-Poisson	1.04 (0.91 – 1.19)	0.56 (0.48 – 0.66)	0.30 (0.22 – 0.41)

[a] The predictors in the model include treatment, initial diagnosis, logtime (logarithm of visit in week based on 2), and interaction term between logtime and treatment.

[b] For logistic regression adjusted OR was reported.

used in this setting to estimate RR.[23-30]

Recommendations for Researchers

The measure of effect of interest in longitudinal studies is adjusted RR. We recommend that researchers use the COPY method with number of copies greater than 100 for RR estimation in a longitudinal data setting.

Authors' Contribution

All authors were involved in the concept and design of the study. SST and MAM wrote the first draft of the manuscript, and all authors contributed to subsequent revisions of the manuscript and approved the final version.

Conflict of Interest Disclosures

The authors have no conflicts of interest.

Ethical Statement

Not applicable.

References

1. Pascarella ET. How college affects students: Ten directions for future research. J Coll Stud Dev. 2006;47(5):508-20. doi: 10.1353/csd.2006.0060.
2. Bijleveld CCJH, van der Kamp LJT, Mooijaart A, van der Kloot WA, van der Leeden R, van der Burg E. Longitudinal data analysis: Designs, models and methods. Thousand Oaks, CA: Sage Publications Ltd; 1998.
3. Vickers AJ. How many repeated measures in repeated measures designs? Statistical issues for comparative trials. BMC Med Res Methodol. 2003;3:22. doi: 10.1186/1471-2288-3-22.
4. Nakai M, Ke W. Statistical models for longitudinal data analysis. Appl Math Sci. 2009;3(40):1979-89.
5. Ballinger GA. Using generalized estimating equations for longitudinal data analysis. Organ Res Methods. 2004;7(2):127-50. doi: 10.1177/1094428104263672.
6. Barros AJ, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol. 2003;3:21. doi: 10.1186/1471-2288-3-21.
7. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13-22. doi: 10.1093/biomet/73.1.13.
8. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss--Newton method. Biometrika. 1974;61(3):439-447. doi: 10.2307/2334725.
9. Diaz-Quijano FA, Janani L, Mansournia MA. Cluster vs. Robust Estimation of Risk Ratio using Expanded Logistic Regression. Arch Iran Med. 2016;19(8):608-9. doi: 0161908/aim.0015.
10. Jang MJ. Working correlation selection in generalized estimating equations. University of Iowa; 2011. doi: 10.17077/etd.kj4igo6k.
11. Lipsitz SR, Fitzmaurice GM, Arriaga A, Sinha D, Gawande AA. Using the jackknife for estimation in log link Bernoulli regression models. Stat Med. 2015;34(3):444-53. doi: 10.1002/sim.6348.
12. Yu Y, Li H, Sun X, Su P, Wang T, Liu Y, et al. The alarming problems of confounding equivalence using logistic regression models in the perspective of causal diagrams. BMC Med Res Methodol. 2017;17(1):177. doi: 10.1186/s12874-017-0449-7.
13. Mansournia MA, Greenland S. The relation of collapsibility and confounding to faithfulness and stability. Epidemiology. 2015;26(4):466-72. doi: 10.1097/ede.0000000000000291.
14. 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):713-9. doi: 0151810/aim.0012.
15. Janani L, Mansournia MA, Mohammad K, Mahmoodi M, Mehrabani K, Nourijelyani K. Comparison between Bayesian approach and frequentist methods for estimating relative risk in randomized controlled trials: a simulation study. J Stat Comput Simul. 2017;87(4):640-51. doi: 10.1080/00949655.2016.1222610.
16. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol. 1986;123(1):174-84. doi: 10.1093/oxfordjournals.aje.a114212.
17. Savu A, Liu Q, Yasui Y. Estimation of relative risk and prevalence ratio. Stat Med. 2010;29(22):2269-81. doi: 10.1002/sim.3989.
18. Deddens JA, Petersen MR, Lei X. Estimation of prevalence ratios when PROC GENMOD does not converge. Paper presented at: Proceedings of the 28th annual SAS users group international conference, 2003; 270-28. Available from: http://www2.sas.com/proceedings/sugi28/270-28.pdf.
19. Lumley T, Kronmal R, Ma S. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. UW Biostatistics Working Paper Series. Working Paper 293. 2006:1-24.
20. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702-6. doi: 10.1093/aje/kwh090.
21. Luo J, Zhang J, Sun H. Estimation of relative risk using a log-binomial model with constraints. Comput Stat. 2014;29(5):981-1003. doi: 10.1007/s00180-013-0476-8.
22. Agresti A. Categorical data analysis. 2nd ed. John Wiley & Sons; 2003. doi: 10.1002/0471249688.
23. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. Bmj. 2017;359:j4587. doi: 10.1136/bmj.j4587.
24. Mansournia MA, Altman DG. Inverse probability weighting. BMJ. 2016;352:i189. doi: 10.1136/bmj.i189.
25. Mansournia MA, Danaei G, Forouzanfar MH, Mahmoodi M, Jamali M, Mansournia N, et al. Effect of physical activity on functional performance and knee pain in patients with osteoarthritis: analysis with marginal structural models. Epidemiology. 2012;23(4):631-40. doi: 10.1097/EDE.0b013e31824cc1c3.
26. Gharibzadeh S, Mohammad K, Rahimiforoushani A, Amouzegar A, Mansournia MA. Standardization as a Tool for Causal Inference in Medical Research. Arch Iran Med. 2016;19(9):666-70. doi: 0161909/aim.0011.
27. Shakiba M, Mansournia MA, Salari A, Soori H, Mansournia N, Kaufman JS. Accounting for Time-Varying Confounding in the Relationship Between Obesity and Coronary Heart Disease: Analysis With G-Estimation: The ARIC Study. Am J Epidemiol. 2018;187(6):1319-26. doi: 10.1093/aje/kwx360.
28. Almasi-Hashiani A, Nedjat S, Mansournia MA. Causal Methods for Observational Research: A Primer. Arch Iran Med. 2018;21(4):164-9.
29. Abdollahpour I, Nedjat S, Mansournia MA, Sahraian MA, Kaufman JS. Estimating the Marginal Causal Effect of Fish Consumption during Adolescence on Multiple Sclerosis: A Population-Based Incident Case-Control Study. Neuroepidemiology. 2018;50(3-4):111-8. doi: 10.1159/000487640.
30. Mansournia MA, Altman DG. Population attributable fraction. BMJ. 2018;360:k757. doi: 10.1136/bmj.k757.