International Journal of Epidemiology, 2022, 1–6
https://doi.org/10.1093/ije/dyac220
Education Corner

Education Corner

Reflection on modern methods: risk ratio regression—simple concept yet complex computation

Murthy N Mittinty[1,2]* and John Lynch[1,2,3]

[1]School of Public Health, University of Adelaide, Adelaide, SA, Australia, [2]Robinson Research Institute, University of Adelaide, Adelaide, SA, Australia and [3]Population Health Sciences, University of Bristol, Bristol, UK

*Corresponding author. School of Public Health, University of Adelaide, Adelaide 5005, SA, Australia.
E-mail: murthy.mittinty@adelaide.edu.au

Received 26 March 2021; Editorial decision 14 October 2022; Accepted 10 November 2022

Abstract

The risk ratio (RR) is the ratio of the outcome among the exposed to risk of the outcome among the unexposed. This is a simple concept, which makes one wonder why it has not gained the same popularity as the odds ratio. Using logistic regression to estimate the odds ratio is quite common in epidemiology and interpreting the odds ratio as a risk ratio, under the assumption that the outcome is rare, is also common. On one hand, estimating the odds ratio is simple but interpreting it is hard. On the other, estimating the risk ratio is challenging but its interpretation is straightforward. Issues with estimating risk ratio still remain after four decades. These issues include convergence of the algorithm, the choice of regression specification (e.g. log-binomial, Poisson) and many more. Various new computational methods are available which help overcome the issue of convergence and provide doubly robust estimates of RR.

Key words: Relative risk, regression, generalized linear models, epidemiology

Key Messages

- Estimating risk ratio (RR) using a simple cross-tabulation is easy. However, when it comes to estimating RR using regression, there is no one particular model.
- Use of log-binomial models with continuous covariates may lead to convergence issues.
- Computational methods such as combinatorial expectation maximiation allow convergence of generalized linear models using the binomial family and log link function. However, specification of starting values can be difficult.
- The binary regression method which allows direct modelling of risk ratios may be a better choice.

Introduction

Relative risk is a common term used in epidemiology to refer to risk and rate ratios.[1] The concept of risk ratio (RR), when introduced first to students, is taught using a simple 2×2 table and a hand calculator. The 2×2 table is created using a simple cross-tabulation of a binary exposure and a binary outcome. Using the information from this cross-tabulation, RR is estimated as the ratio of risk of the outcome among the exposed versus risk of the outcome among the unexposed. For example, let's say the outcome is low birthweight (Yes = 1 or No = 0) and the exposure is maternal smoking during pregnancy (Yes = 1 and No = 0). Risk ratio, in this example, is the ratio of the proportion of low-birthweight children among smokers to the proportion of low-birthweight children among non-smokers. In this form it is simple and easy to calculate.

Let's consider adjusting for one confounder, like gender which is binary; in this case, RR can be estimated within the stratum of gender. Now suppose there is a long list of confounders which includes age, education, income, pregnancy-related factors and others. To estimate RR in this case, one may need to use regression. Use of regression methods for estimating RR gained popularity when they became available in regular commercial and non-commercial statistical software. Even with this availability, it is still not free from problems, which has concerned researchers since the 1980s.[2] Other methods, such as logistic regression, gained immense popularity and have become essential tools in epidemiology due to the computational ease, and as the odds ratio (OR) can approximate the RR in the case of rare events. Evidence suggests that logistic regression is used to estimate the OR but is commonly interpreted as RR.[3] However, OR overestimates RR, whenever RR is greater than 1, and hence should not be interpreted as RR.[4–6]

If logistic regression is used to estimate RR under the rare disease assumption, then one must note that this assumes that the conditional probability of having an outcome, given the unexposed state (baseline prevalence, $p(Y = 1|X = 0) = p_0$) approaches zero (as shown in Supplementary Material, Section S1, available as Supplementary data at *IJE* online). Moreover, as suggested by the reviewer, relation between OR and RR can be derived as shown in Supplementary Section 1 using this derivation: if we assume $RR^{max} = 10$ (*upper bound*) and $p_0 = 0.001$, we have $\frac{OR}{RR} \leq 1.01$. Thus, if the RR^{max} is less than or equal to 10 and the baseline prevalence is 1 in 100-then the relative error OR/RR is 1%. With a prevalence of 1 in 10000 it is 0.1%; when the prevalence is very small but not zero, the approximation errors are small enough to be practically negligible. We assume the RR >1 but less than some maximum plausible value $RR^{max} > 1$.

Alternatively, let's examine this using a simple 2×2 table with four cells. Let these cells be labelled as *a, b, c* and *d*, where '*a*' is the count when the outcome is 1 and the exposure is 1, '*b*' is the count when the exposure is 1 and outcome is 0, '*c*' is the count when the outcome is 1 and the exposure is 0 and '*d*' is when both outcome and exposure are 0. Now to estimate RR, we use the formula $\frac{\frac{a}{a+b}}{\frac{c}{c+d}}$. If we rearrange the terms, we estimate the RR as $\frac{a*(c+d)}{c*(a+b)} = \frac{ac+ad}{ca+bc}$, whereas the OR is estimated as $\frac{ad}{bc}$. Again, from these formulae, one may note that RR does not equal (or even approximate) OR without some assumptions. One common assumption can be that the outcome is rare in both the groups of the exposure (if exposure is the only variable; else, the outcome of interest must be rare for all the levels of the covariates). Furthermore, let's put some numbers instead of *a*, *b*, *c*, *d*, say a = 1, b = 5, c = 1 and d = 11. In this case, the estimate of RR using the above formula equates to 12/6 = 2 which is the ratio of the marginal totals of the exposure when $X = 1(a + b)$ and $X = 0$ $(c + d)$. Now, if we estimate the OR (= 2.2), as shown in Supplementary Material, Section S2, (available as Supplementary data at *IJE* online) the OR equates to the ratio of not having the outcome when the exposure is absent versus not having the outcome when exposure is present. In this example, equating OR and RR may not be appropriate as they are estimating two different things. Moreover, both OR and RR are not estimating the risk of disease whenever the counts *a* and *c* are equal in a 2×2 table or a stratified 2×2 table. In summary, if the study outcome is common, interpretation of OR as an approximation to RR becomes unreliable.

Odds ratios may still be of interest because they are symmetrical, in the sense that the odds of having an outcome is the inverse of odds of not having the outcome (mathematically this might be interesting but practically, when the outcome is defined as death or survival, this property might not seem desirable), and when the covariate set is large it may be a preferred choice.[7] Moreover, in some case-control studies when studies use cumulative incidence sampling, OR maybe valuable.[8] On the other hand, RR is not symmetrical (with respect to relabelling of the outcome *Y*) but the size of the RR will not change if adjustment is made for a variable that is not a confounder. This is referred to as collapsibility. The collapsibility property implies that the risk ratio can be expressed as the ratio change in average risk due to exposure among the exposed.[7,9–12] It is for this reason RR, and for its ease of interpretation, maybe a preferred parameter of interest over OR.

Several methods have been proposed to estimate the RR. These include the Stratified Mantel–Haenszel method,[4] Cox regression,[3,13] adjustment to OR[14] (even

though this method was later noted to be biased[15]) and generalized linear models (GLM) with family binomial and link log, referred to as log-binomial.[16,17] However, the log-binomial method has the issue of convergence[2,3,16] in STATA, R, SAS, Splus or any other software. To overcome this issue, methods such as the COPY method,[2,13,16] modified Poisson,[18] marginal standardization,[16,17,19] binary regression models,[9] quasi-likelihood Poisson method[20] constrained optimization[21] and non-linear least squares[3] have been proposed. Some of the software available include libraries such as *logbin* (log binomial models)[21–23] and *brm* (binary regression model) in R.[9]

This raises the question: if RR is a simple concept, why don't regression methods, using maximum likelihood estimation (MLE) with standard Fisher scoring matrix, converge when estimating RR? Are there different computational methods? How should the results be presented? We provide some answers to these questions.

Why are there different methods?

Let Y be the binary outcome of interest, X the binary exposure and C be the vector of confounders. Y is 1, representing the occurrence of an event, and 0 represents the non-occurrence. Similarly, when the exposure equals 1, we say the individual is exposed/treated and 0 indicates those non-exposed/untreated. Confounders can be continuous, categorical or binary variables (examples include age, levels of gender). The success probability in RR regression is modelled as:

$$\log(P[Y_i=1|D_i=(X_i,C_i)])=\beta_0+\beta_1X_i+\beta_2C_{1i}+\cdots+\beta_pC_{pi} =(\boldsymbol{\beta D_i})$$

Denote $P[Y_i=1|D_i]=p_i, \forall\ i=1,2,.,n$, as the probability of having an outcome for n individuals in the data (D). The above equation can be rearranged in a matrix form and written as:

$$\log(p_i)=\boldsymbol{\beta D_i}$$

Using the relation between natural logarithms and exponentials, the above equation can be expressed as, $p_i=e^{\boldsymbol{\beta D_i}}$. Here, the parameter $\boldsymbol{\beta}$ is unknown and this vector needs to be estimated. To estimate the unknown parameter, we will use the Bernoulli likelihood function which is given by:

$$L(\beta)=\prod_{i=1}^{n}p_i^{Y_i}(1-p_i)^{(1-Y_i)}$$

Various methods to estimate/fit the model include the maximum likelihood estimating equation for β, obtained by taking the derivative of the logarithm of the above likelihood function ($L(\beta)$) and equating it to zero. Mathematically this is simplified as (for complete derivation, see Supplementary Section S3):

$$S(\beta)=\frac{\partial log\ L(\beta)}{\partial\beta_j}=\sum_{i=1}^{n}\frac{(Y_i-p_i)}{p_i(1-p_i)}\boldsymbol{d}_ip_i=\sum_{i=1}^{n}\frac{\boldsymbol{d}_i(Y_i-p_i)}{(1-p_i)} \quad (1)$$

where $\boldsymbol{d_i}$ is a realization of vector $\boldsymbol{D_i}$. This is an asymptotically efficient estimate: when the probability of success (p_i) is less than 1 then the MLE exists, and it is unique. However, when $p_i\approx 1$ then the estimating function will be dominated by observation i, and convergence issues persist. When the MLE does not converge, some software uses constrained optimization techniques as a default solution and thus attains convergence.[3]

In standard software, MLE is computed using methods like the Newton–Raphson method, iteratively reweighted least squares (IRWLS) and Fisher scoring.[21–23] However, these computational methods have issues when the probability approaches 1 in log-binomial models.[2,3,23] Alternatively, the modified Poisson regression method[15] has been proposed for estimating β and has gained attention. The MLE for the Poisson regression is given by:

$$S_{Poisson}(\beta)=\sum_{i=1}^{n}\boldsymbol{d}_i(Y_i-p_i)$$

As seen from above notation, the Poisson regression does not suffer from the issue of convergence as there is no denominator which may approach zero. Now the question is: how does one get rid of the denominator in Equation 1? To understand this, it is important to take a step back and revisit the concept of Maclaurin series. Using the Maclaurin series, $\frac{1}{1-p_i}$, in Equation 1 can be expressed as:

$$\frac{1}{1-p_i}=1+p_i+p_i^2+p_i^3+\cdots=\sum_{m=0}^{M}p_i^m. \quad (2)$$

Replace $\frac{1}{1-p_i}$ in Equation 1 as the weight, $w(p_i,M)$, then this can be re-expressed as:

$$S(\beta)=\sum_{i=1}^{n}\boldsymbol{d}_i(Y_i-p_i)w(p_i,M)$$

When $M=0$ in Equation 2, then the weight $w(p_i\ M)=1$. Hence, the RR estimated using a Poisson regression can be viewed as Maclaurin series truncated at $M=0$. However, with binary outcomes not all combinations of parameters lead to fitted means that are between

zero and one. It allows for higher values of M to be used. In 2014, Fitzmaurice *et al.*[24] proposed a method that uses $M = 20,\ 30,\ 40$ and 60, also known as the 'almost efficient estimation of RR'. Thus, all variants of the weighted regression, including when M equals 0, will only estimate RR almost efficiently, but not completely efficiently.

If using Poisson and interpreting results from this regression, then one must specify it as a truncated $(M = 0)$ Maclaurin series. If using higher terms, as done by Fitzmaurice *et al.*,[24] then we must say that exactly $(M = 60)$. When Poisson regression is applied to binomial data, the standard error for the estimated RR will be overestimated.[15] To overcome this issue one can use the sandwich estimation procedure to compute the robust error variance.[19,25,26] However, when the sample sizes are small, Poisson models do not work well because the sandwich estimators tend to underestimate the true standard errors (Department of Statistics, Rupert Carroll, unpublished observations).[27] Furthermore, one of the weakness of sandwich estimators is that their variance can be less efficient than the variance estimated from a parametric model (Department of Statistics, Rupert Carroll, unpublished observations). This weakness then impacts on the coverage probability, the probability that a confidence interval covers the true RR.[28] Moreover, the predicted probabilities using Poisson regression can lie outside of the range [0,1].[6,27] This happens because RR is variation dependent on the baseline probability $(p(Y = 1|X = 0))$. For example, if $RR = 2$, then $p(Y = 1|X = 1) = 2 * p(Y = 1|X = 0)$ and this indicates that $p(Y = 1|X = 0) \leq 0.5$. This is a restricted domain over which the quantities $(RR, p(Y = 1|X = 0))$ need to be compatible with a valid probability distribution. As can be observed from this example, it is not only for the Poisson regression; even in log-binomial models, with considerable numbersof covariates, finding MLE can be a problem as the parameter space is constrained and the log likelihood (Equation 1) needs to be maximized using constrained optimization.[3]

The next questions that naturally arise are: (i) how to achieve convergence in log-binomial models; and (ii) presenting results from multiple methods.

How to achieve convergence in log-binomial models?

With the log-binomial fitting procedure one can start by increasing the maximum number of iterations along with specification of starting values. The starting values can be set to the mean observed proportion for the intercept and all the rest of the parameters can be set to zero. However, if the standard default procedures (IRWLS algorithm, Newton–Raphson or Fisher scoring computational methods) are used in estimating the MLE of the log-binomial, then they may not converge. In such situations, computational methods like combinatorial expectation maximization (CEM), adaptive barrier method, parabolic expectation maximization (PEM)/quasi Newton methods may be used through packages such as *logbin* in R.[21–23] Use of these latter computational methods may allow convergence if the starting values are specified. [17] Coming up with a proper set of starting values can be tricky. If the starting values are appropriate, then there is a chance of attaining convergence, else not. For CEM, if the covariate set is large, then again there will be no convergence. With alternative methods, convergence still persists because of the constrained optimization.[9]

Alternatively, one can use the binary regression model (BRM) approach that overcomes the variance dependence and constrained optimization.[9] The BRM allows direct modelling of RR.[9] BRM uses two different regressions: (i) an outcome regression; and (ii) a propensity score regression of the exposure on the baseline covariates. Furthermore, the outcome regression uses two different models: (a) a target model for estimating RR directly; and (b) a nuisance model for the log odds product. Use of the log odds product allows estimation of RR either using an unconstrained MLE or semiparametric g-estimation methods. If the target model is correctly specified and either the log odds product model or the propensity score model is correctly specified, BRM yields a robust estimate.[9] Similar to *glm* methods, BRM also requires specification of starting values and may converge to local maxima.

Presenting results

For this demonstration, we use data from the National Health and Nutrition Examination Survey Follow-up Study to estimate the RR in the covariate-adjusted associational sense. These data are available as accompanying data to the book by Hernan and Robin.[29] We are primarily interested in the association between quitting smoking (Yes/No) and a dichotomized weight change (above and below median weight) between 1971 and 1982. Code that is required to run all analysis and reproduce the results presented here is available in the Supplementary Material, Section S4 (available as Supplementary data at *IJE* online). In our analysis we adjusted for sex, age, race, income, marital status, education, asthma and bronchitis. All analysis was conducted in R version 3.6.3 and Stata 15.1. Results from these methods are presented in Table 1.

Conclusion

In general, RR can be estimated using a hand calculator if presented as a simple 2 × 2 table. A common problem

Table 1 Risk ratio estimates for the association between quitting smoking and greater than median weight gain among 1629 men and women in the National Health and Nutrition Examination Survey Epidemiologic Follow-up Data between 1971 and 1982

Method	Estimation	Computation	Risk ratio (95% CI)
Mantel–Haenszel (Combined)			1.28 (1.17,1.41)
GLM (Family=binomial, link=log)	MLE	IRLS	NC
GLM (with defined starting values) (Family=binomial, link=log)	MLE	IRLS	1.29 (1.18,1.43)
GLM (Se = Sandwich) (Family=Poisson, Link=log)	MLE	IRLS	1.34 (1.22,1.48)
Binary regression model	MLE		1.36 (0.98,1.73)
Binary regression model	DR		1.39 (0.86,1.91)
Log-binomial	MLE	EM (CEM)	1.32 (1.20,1.45)
Log-binomial	MLE	AB	1.32 (1.20,1.45)

All models, except Mantel–Haenszel, adjusted for age, sex, race, income, marital status, education, asthma and bronchitis.

GLM, Generalised Linear Model; Se, standard error; MLE, Maximum Likelihood Estimation; DR, Doubly Robust Estimation; IRLS, Iterative Reweighted Least Squares; EM, Expectation Maximization; CEM, Combinatorial Expectation Maximization; AB, adaptive barrier; NC, non-convergence.

when using regression to estimate RR is lack of convergence of r, orthe MLE. With the provision of additional computational methods such as CEM, adaptive barrier or other methods as alternatives to standard methods such as IRWLS, we may overcome the issues of convergence in the log-binomial model if proper starting values are specified and the covariate set is small. When the covariate set is large, then one can use the BRM method which allows direct modelling of RR. Use of BRM for estimating RR may produce wider (conservative) confidence intervals for the RR. However, further research directly comparing the RR estimates between BRM and *glm* methods need to be conducted.

Ethics approval

All NHANES protocols were approved by the National Center for Health Statistics Research Ethics Review Board, and all participants provided documented consent: detailed information at [https://www.cdc.gov/nchs/nhanes/irba98.htm].

Supplementary Data

Supplementary data are available at *IJE* online.

Author contributions

M.N.M. and J.L. contributed equally to the conceptualization, writing, analysis and presentation of the document.

Funding

There is no funding for this research.

Acknowledgements

Authors would like to thank the anonymous reviewers and the editor for their constructive suggestion which improved the work. The first author would also like to thank Manasi Murthy Mittinty (University of Sydney) for stimulating discussions and feedback on the work.

Conflict of interest

None declared.

References

1. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008.
2. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol* 1986;**123**:174–84.
3. Lumley T, Kronmal R, Ma S. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. Working Paper 293. http://biostats.bepress.com/uwbiostat/paper293 (10 October 2022, date last accessed).
4. 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**:1–3.
5. Altman DG, Deeks JJ, Sackett DL. Odds ratios should be avoided when events are common. *BMJ* 1998;**317**:1318.
6. Greenland S. Cornfield, risk relativism, and research synthesis. *Stat Med* 2012;**31**:2773–77.
7. Cummings P. The relative merits of risk ratios and odds ratios. *Arch Pediatr Adolesc Med* 2009;**163**:438–45.
8. Blakely T, Pearce N, Lynch J. Case-control studies. *JAMA* 2019; **321**:806–07.
9. Richardson TS, Robins JM, Wang L. On modeling and estimation for the relative risk and risk difference. *J Am Stat Assoc* 2017;**112**:1121–30.
10. Guo J, Geng Z. Collapsibility of logistic regression coefficients. *J R Stat Soc Series B Stat Methodol* 1995;**57**:263–67.
11. Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. *Stat Sci* 1999;**14**:29–46.
12. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Rev* 1991; **59**:227–40.

13. Cummings P. Methods for estimating adjusted risk ratios. *Stata J* 2009;**9**:175–96.
14. Zhang J, Yu FK. What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998;**280**:1690–91.
15. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol* 2003;**157**:940–43.
16. Dwivedi AK, Mallawaarachchi I, Lee S, Tarwater P. Methods for estimating relative risk in studies of common binary outcomes. *J Appl Stat* 2014;**41**:484–500.
17. Dwivedi AK, Shukla R. Evidence-based statistical analysis and methods in biomedical research (SAMBR) checklists according to design features. *Cancer Rep* 2020;**3**:e1211.
18. Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol* 2004;**159**:702–06.
19. Naimi AI, Whitcomb BW. Estimating risk ratios and risk differences using regression. *Am J Epidemiol* 2020;**189**:508–10.
20. Carter RE, Lipsitz SR, Tilley BC. Quasi-likelihood estimation for relative risk regression models. *Biostatistics* 2005;**6**:39–44.
21. Marschner IC, Gillett AC. Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics* 2012;**13**: 179–92.
22. Marschner IC. Relative risk regression for binary outcomes: methods and recommendations. *Aust N Z J Stat* 2015;**57**:437–62.
23. Donoghoe MW, Marschner IC. logbin: an R package for relative risk regression using the log-binomial model. *J Stat Softw* 2018; **86**:1–22.
24. Fitzmaurice GM, Lipsitz SR, Arriaga A, Sinha D, Greenberg C, Gawande AA. Almost efficient estimation of relative risk regression. *Biostatistics* 2014;**15**:745–56.
25. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol* 2004;**160**:301–05.
26. Chen W, Qian L, Shi J, Franklin M. Comparing performance between log-binomial and robust Poisson regression models for estimating risk ratios under model misspecification. *BMC Med Res Methodol* 2018;**18**:1–2.
27. Zhu C, Blizzard L, Stankovich J, Wills K, Hosmer DW. Be wary of using Poisson regression to estimate risk and relative risk. *Biostat Biom Open Access J* 2018;**4**:1–3.
28. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc* 2001;**96**: 1387–96.
29. Hernan MA, Robins J. *Causal Inference: What If*. Boca Raton, FL: Chapman & Hill/CRC, 2020.