American Journal of Epidemiology

Vol. 189, No. 6
DOI: 10.1093/aje/kwaa044
Advance Access publication:
March 27, 2020

The *AJE* Classroom

Estimating Risk Ratios and Risk Differences Using Regression

Ashley I. Naimi* **and Brian W. Whitcomb**

* Correspondence to Dr. Ashley I. Naimi, Department of Epidemiology, University of Pittsburgh, 130 DeSoto Street, 1503 Public Health Building, Pittsburgh, PA 15261 (e-mail: ashley.naimi@pitt.edu).

Initially submitted March 3, 2020; accepted for publication March 17, 2020.

Generalized linear models (GLMs) are often used with binary outcomes to estimate odds ratios. Though not as widely appreciated, GLMs can also be used to quantify risk differences, risk ratios, and their appropriate standard errors (1). Here, we illustrate how GLMs can be used to quantify these latter effect measures, and we demonstrate how to obtain valid standard errors.

Logistic regression for binary outcomes are often implemented via GLM software routines (e.g., PROC GENMOD in SAS (SAS Institute, Inc., Cary, NC), or the *glm* functions in Stata (StataCorp LP, College Station, TX) and R (R Foundation for Statistical Computing, Vienna, Austria) by selecting the binomial distribution and the logistic link function. Such a formulation can generate a conditionally adjusted odds ratio for the exposure-outcome association, which is often not the most intuitive measure of choice. GLMs can also be used to quantify conditionally adjusted risk ratios and risk differences using a binomial distribution and strategically selected link functions, but convergence problems can arise.

Here, we use publicly available data to demonstrate different techniques for estimating the risk difference and risk ratio for the relation between quitting smoking and weight gain. We highlight strategies that can be used to avoid convergence problems but that require special considerations for estimating standard errors.

GENERALIZED LINEAR MODELS

GLMs consist of a family of regression models that are fully characterized by a selected distribution and a link function. The distribution determines the nature of the conditional mean and variance of the outcome under study, whereas the link function determines how the exposure and confounders relate to the conditional mean.

There are a wide variety of distributions and link functions available in standard statistical software programs that fit GLMs. In this article, we consider a binary outcome Y with probability $P = P(Y = 1)$, and focus attention on 3 link functions: 1) logit (i.e., $\log\{P(Y = 1)/[1 - P(Y = 1)]\}$); 2) log (i.e., log(P)); and 3) identity (i.e., P). A common misconception is that to use GLMs correctly, one must choose the distribution that best characterizes the data, as well as the canonical link function corresponding to this distribution. For example, if the outcome is binary, one must choose the binomial distribution with the logit link. Although the binomial distribution and logit link work well together for binary outcomes, they do not easily provide contrasts like the risk difference or risk ratio, because of the selected link function. Alternative specifications of the distribution and link function for GLMs can address this limitation.

Link functions and effect measures

There is an important relation between the chosen link function and the interpretation of the coefficients from a GLM. For models of a binary outcome and the logit or log link, this relation stems from the properties and rules governing the natural logarithm. The quotient rule states: $\log(X/Y) = \log(X) - \log(Y)$.

Because of this relation, the natural exponent of the coefficient in a logistic regression model yields an estimate of the odds ratio. However, by the same reasoning, exponentiating the coefficient from a GLM with a log link function and a binomial distribution (i.e., log-binomial regression) yields an estimate of the risk ratio. Alternately, for GLM models with a binomial distribution and identity link function, because logarithms are not used, the unexponentiated coefficient yields an estimate of the risk difference.

Unfortunately, using a binomial distribution can lead to convergence problems with the log() or identity link functions for reasons that have been explored (2). This will occur when, for example, the combined numerical value of all the independent variables in the model is very large. This can result in estimated probabilities that exceed 1, which violates the very definition of a probability (binomial) model (probabilities can only lie between 0 and 1) and, hence, convergence problems. We show how these problems can be overcome.

Table 1. Data on the Association Between Quitting Smoking and Greater Than Median Weight Gain Among 1,507 Men and Women in the National Health and Nutrition Examination Survey Epidemiologic Follow-up Study Data Between 1971 and 1982

Method	Risk Difference	95% CI	Risk Ratio	95% CI
GLM[a]	0.14	0.09, 0.20	1.32	1.19, 1.46
Marginal standardization[b]	0.14	0.09, 0.21	1.31	1.18, 1.46

Abbreviations: CI, confidence interval; GLM, generalized linear model.
[a] Using a conditionally adjusted regression model without interactions. Gaussian distribution and identity link were used to obtain the risk difference. A Poisson distribution and log link were used to obtain the risk ratio. 95% confidence intervals were obtained via the sandwich variance estimator.
[b] The 95% confidence intervals were obtained using the bias-corrected and accelerated bootstrap confidence interval estimator.

Data

For this article, we use data from the National Health and Nutrition Examination Survey Epidemiologic Follow-up Study, available as a companion data set to a forthcoming book by Hernán and Robins (3). We are interested primarily in the covariate adjusted association (on the risk difference and risk ratio scales) between quitting smoking and a greater than median weight change between 1971 and 1982. Code needed to run all analyses and reproduce our results is available on GitHub (4).

In our analyses, we regress an indicator of greater than median weight change against an indicator of whether the person quit smoking. We adjust for exercise status, sex, age, race, income, marital status, education, and indicators of whether the person was asthmatic or had bronchitis. All analyses are conducted in R, version 3.6.2.

GLMs for risk differences and ratios

For our analyses of the aforementioned data using GLM with a binomial distributed outcome with a log link function to estimate the risk ratio and identity link function to estimate risk difference, an error is returned. Instead, one may resort to using different distributions that are more compatible with the link functions that return the association measures of interest.

For the risk ratio, one may use a GLM with a Poisson distribution and log link function. Doing so will return an exposure coefficient whose natural exponent can be interpreted as a risk ratio. However, the model-based standard errors (i.e., the standard errors one typically obtains directly from the GLM output) are no longer valid. Instead, one should use the robust (or sandwich) variance estimator to obtain valid standard errors (the bootstrap can also be used) (5).

For the risk difference, one may use a GLM with a Gaussian (i.e., normal) distribution and identity link function, or, equivalently, an ordinary least squares estimator. Doing so will return an exposure coefficient that can be interpreted as a risk difference. However, once again, the robust variance estimator (or bootstrap) should be used to obtain valid standard errors.

The risk ratio and difference, as well as the 95% sandwich variance confidence intervals obtained for the relation

Table 2. Possible Strategies for Using GLMs to Quantify Risk Differences, Risk Ratios, and Odds Ratios[a]

Odds Ratio	Risk Ratio	Risk Difference
GLM family = binomial	GLM family = binomial	GLM family = binomial
GLM link = logistic	GLM link = log	GLM link = identity
Standard errors = model based	Standard errors = model based	Standard errors = model based
	GLM family = Poisson	GLM family = Gaussian
	GLM link = log	GLM link = identity
	Standard errors = sandwich	Standard errors = sandwich
		Least squares regression
		Standard errors = sandwich

Abbreviation: GLM, generalized linear model.
[a] The top row provides the first steps that should be pursued when seeking to estimate measures of effect. If convergence problems ensue, one can then proceed to options in the second row for the risk ratio, and the second or third row for the risk difference. To use marginal standardization, one can rely on the procedures described in the text.

between quitting smoking and greater than median weight change are provided Table 1.

Unfortunately, use of a Poisson or Gaussian distribution for GLMs for a binomial outcome can introduce different problems. For one, using the log or identity links does not guarantee that probabilities from these models are bounded between 0 and 1. Second, performance of the robust variance estimator is notoriously poor with small sample sizes. Finally, the interpretation of the risk differences and ratios becomes more complex when the exposure interacts with other variables in the model. For instance, in the National Health and Nutrition Examination Survey Epidemiologic Follow-up Study data, the association between quitting smoking and weight gain seems to interact with baseline exercise status. Thus, to properly interpret the association, exercise status should be considered.

Marginal standardization

A final approach to obtaining risk differences and ratios from GLMs that are not subject to the aforementioned limitations is to use marginal standardization, which is equivalent to g-computation when the exposure is measured at a single time point (6). This process can be implemented by fitting a separate logistic model in each exposure stratum (i.e., exposed and unexposed), regressing the binary outcome against all confounder variables. But instead of reading the coefficients the model, one can obtain risk differences and ratios by using these models to generate predicted risks for each individual under "exposed" and "unexposed" scenarios in the data set. This stratified modeling approach avoids the exposure effect homogeneity assumption across levels of the confounders. To obtain predictions under the exposed scenario, we use the model fit to the exposed individuals to generate predicted outcomes in the entire sample. To obtain predictions under the unexposed scenario, we repeat the same procedure but with the model fit among the unexposed. One can then average the risks obtained under each exposure scenario and take their difference and ratio to obtain the risk differences and ratios of interest. To obtain standard errors, the entire procedure must be bootstrapped (see the code available at the GitHub link). These marginal risk differences and ratios, as well as their bootstrapped confidence intervals are presented in Table 2.

When predicted risks are estimated using a logistic model, relying on marginal standardization will not result in probability estimates outside the bounds (0, 1). And because the robust variance estimator is not required, model-based standardization will not be as affected by small sample sizes. However, the bootstrap is more computationally demanding than alternative variance estimators, which may pose problems in larger data sets.

CONCLUSION

Risk differences and risk ratios are often of interest for epidemiologic research, but methods to obtain these measures of association have not been widely recognized. We have described how GLMs can be easily implemented to address this gap. Here, we showed how these effects can be estimated with conditional GLMs as well as with marginal standardization. In Table 2, we provide a workflow summarizing the strategies outlined here to facilitate their estimation.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania (Ashley I. Naimi); and Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, Amherst, Massachusetts (Brian W. Whitcomb).

Conflict of interest: none declared.

REFERENCES

1. 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(4):301–305.
2. Williamson T, Eliasziw M, Fick GH. Log-binomial models: exploring failed convergence. *Emerg Themes Epidemiol*. 2013;10(1):14.
3. Hernán MA, Robins J. *Causal Inference*. In press. London, UK: Chapman and Hall.
4. Naimi AI. Code for paper entitled: Estimating Risk Ratios and Risk Differences Using Regression. https://github.com/ainaimi/GLM_RD_RR. Accessed April 30, 2020.
5. Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7):702–706.
6. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173(7):731–738.