How to Estimate Risk Ratios

Statnews #92

Created winter 2018. Last updated August 2020

Introduction

Risk ratios provide statisticians with a way to compare the risk of an event between two groups. A risk ratio can be defined as the ratio of the probability of the event occurring in a treatment group to the probability of the event occurring in a control group. For more information on the similarities and differences between risk ratios and odds ratios, see Statnews #90.

Estimating Risk Ratios

There are several ways to estimate risk ratios. Formulas can be used to convert odds ratios to risk ratios, but these formulas can produce biased results. Calculating risk ratios directly from generalized linear models and their extensions is preferable (Schmidt, 2008).

A generalized linear model requires a probability distribution from the exponential family and a continuous link function. Logistic regression is one example of a generalized linear model, which is estimated by specifying a binomial distribution and a logit link function. If a log link is used with a binomial distribution, risk ratios can be then obtained by exponentiating the coefficients. (Note that fitting a binomial GLM with a log link is not the same as logistic regression.) However, this type of model may have convergence issues. As an alternative, one can model binary data using a generalized linear model with a Poisson distribution and a log link. Exponentiating the coefficients will again give you estimates of the risk ratio. An important property of the Poisson distribution is that the variance of the Poisson is equal to the mean.

When a Poisson regression is applied to binomial data, the model will be under-dispersed, causing the standard errors to be overestimated (Zhu, 2004). To reduce the bias, there are several other models that can be estimated. The quasi-Poisson distribution can be used to estimate a scale parameter, which allows the variance to be a multiple of the mean. Another possible solution is to estimate a generalized estimating equation (GEE), with clustering at the residual level, to estimate a scale parameter. More information about generalized estimating equations can be found in Statnews #76, #88, and our Fall 2009 Stats Happening.

Example calculations

To show how the estimated risk ratios and their standard errors differ depending on the model, we will estimate the models described above using simulated data In this simulated dataset, information on 100 individuals such as their gender, age, and whether the individual was in the control group or treatment group is given. Table 1 displays the data for the first six simulated individuals. Table 2 displays the number of subjects in each treatment group with and without disease.

Table 1: First six rows of the simulated dataset.

outcome	age	treatment	gender
0	25	1	1
1	25	0	0
0	25	1	1
0	25	1	0
0	25	1	0
1	25	0	0

Table 2: Number of subjects in each group with and without disease.

	No disease	Disease
Control group	24	23
Treatment group	40	13

We would like to estimate the risk ratio for each of these variables (treatment group, age, and gender) while controlling for the other variables. Estimates of risk ratios, with standard errors in parentheses, are provided in Table 3 for each of the models described previously. As an example, R code for fitting each of these models is given below.

```
library(geepack)
# Binomial model with log link
model.bin<-glm(outcome~treatment+gender+age, data=dat, family=binomial(log))
# Poisson
model.pos <- glm(outcome~treatment+gender+age, data=dat, family=poisson(log))
# Quasi-Poisson
model.qpos<-glm(outcome~treatment+gender+age, data=dat, family=quasipoisson())
# Poisson GEE
model.gee <- geeglm(outcome~treatment+gender+age, data=dat, id=id,
corstr="independence",
                    family=poisson())
```

The control group is the reference level for the treatment variable, female is the reference level for the gender variable. For a continuous variable like age, the risk ratio is the probability of disease at age $a + 1$ divided by the probability of disease at age a, where a is a fixed number of years.

Table 3: Risk ratios and standard errors (in parentheses).

	Binomial with log link	Poisson with log link	Quasi-Poisson with log link	Poisson GEE
Treatment	0.4624 (0.1258)	0.4743 (0.1653)	0.4743 (0.1355)	0.4743 (0.1308)

	Binomial with log link	Poisson with log link	Quasi-Poisson with log link	Poisson GEE
Gender	0.5224 (0.1452)	0.5475 (0.1943)	0.5475 (0.1592)	0.5475 (0.155)
Age	0.9922 (0.0161)	0.9992 (0.0227)	0.9992 (0.0186)	0.9992 (0.0166)
Scale Parameter			0.672	0.6449 (0.1777)

From the table above, we can see that the estimates of the risk ratios are similar when a Poisson model or a binomial model is used, but the standard errors vary depending on the model. The scale parameter estimated in the quasi-Poisson model is 0.672. Since this value is less than 1, the Poisson model has under-dispersion. The square root of the scale parameter is 0.82, which indicates that the standard errors of the Poisson model have been decreased by a factor of 0.82 for the quasi-Poisson model. The estimates, standard errors, and scale parameter of the Poisson GEE model are very similar to those from the quasi-Poisson model.

Presenting risk ratios is, in most situations, an appropriate alternative to reporting odds ratios; however, one must be careful to estimate and label them appropriately. If you need help conducting this type of analysis, please do not hesitate in contacting CSCU.

Author: Stephen Parry

References

Schmidt & Kohlmann. When to use the odds ratio or the relative risk? Institute for Community Medicine. (2008).

Woodward, Mark. Epidemiology: Study Design and Data Analysis. (1999).

Zou, Guangyong. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. American Journal of Epidemiology. (2004).

Naimi, A. I., & Whitcomb, B. W. (2020). Estimating Risk Ratios and Risk Differences Using Regression. American Journal of Epidemiology.