How can I estimate relative risk using glm for common outcomes in cohort studies?

Authored by

stats writer

July 1, 2024

Estimating relative risk is a commonly used statistical method in cohort studies to analyze the relationship between a potential risk factor and a specific outcome. This can be achieved using the generalized linear model (glm) approach. The glm framework allows for the calculation of relative risk by estimating the risk ratio or risk difference between two groups, while accounting for potential confounding factors. This method is particularly useful in cohort studies, where individuals are followed over time, as it can provide valuable insights into the potential impact of a risk factor on a particular outcome. By using glm, researchers can assess the relative risk of a potential risk factor and its significance in a more comprehensive and robust manner, aiding in the interpretation and understanding of the results in cohort studies.

How can I estimate relative risk using glm for common outcomes in cohort studies? | Stata FAQ

Credits

This page was developed and written by Karla Lindquist, Senior Statistician in the Division of Geriatrics at UCSF. We are very grateful to Karla for taking the time to develop this page and giving us permission to post it on our site.

Introduction

Binary outcomes in cohort studies are commonly analyzed by applying a logistic regression model to the data to obtain odds ratios for comparing groups with different sets of

characteristics. Although this is often appropriate, there may be situations in which it is more desirable to estimate a relative risk or risk ratio (RR) instead of an odds ratio (OR). Several articles in recent medical and public health literature point out that when the outcome event is common (incidence of 10% or more), it is often more desirable to estimate an RR since there is an increasing differential between the RR and OR with increasing incidence rates, and there is a tendency for some to interpret ORs as if they are RRs (-). There are some who hold the opinion that the OR should be used even when the outcome is common, however (). Here the purpose is to demonstrate methods for calculating the RR, assuming that it is the appropriate thing to do.

There are several options for how to estimate RRs directly in Stata. Two of
these methods will be demonstrated here using hypothetical data created for this
purpose. Both methods use command `glm`. One estimates the RR with a log-binomial regression model, and the other uses a
Poisson regression model with a robust error variance.

Example Data: Odds ratio versus relative risk

A hypothetical data set was created to illustrate two methods of estimating relative risks using Stata. The outcome generated is called lenses, to indicate if the hypothetical study participants require corrective lenses by the time they are 30 years old. Assume all participants do not need them at a baseline assessment when they are 10 years old. Assume none of them have had serious head injuries or had brain tumors or other major health problems during the 20 years between assessments. Suppose we wanted to know if requiring corrective lenses is associated with having a gene which causes one to have a lifelong love and craving for carrots (assume not having this gene results in the opposite), and that we screened everyone for this carrot gene at baseline (carrot = 1 if they have it, = 0 if not). We also noted their gender (= 1 if female, = 2 if

male), and what latitude
of the continental US they lived on the longest (24 to 48 degrees north). All
values (N=100) were assigned using a random number generator. The data set is
eyestudy.dta in Stata 8 format. Here's a quick description of the variables.

```
use https://stats.idre.ucla.edu/stat/stata/faq/eyestudy, clear
describe
```

Contains data from https://stats.idre.ucla.edu/stat/stata/faq/eyestudy.dta
obs: 100
vars: 5
size: 900 (99.9% of memory free)

---

storage display value
variable name type format label variable label

---

id byte %8.0g
carrot byte %8.0g
gender byte %8.0g
latitude byte %8.0g
lenses byte %8.0g

Sorted by:

```
summarize
```

Variable | Obs Mean Std. Dev. Min Max

----------+----------

id | 100 50.5 29.01149 1 100

carrot | 100 .51 .5024184 0 1

gender | 100 1.48 .5021167 1 2

latitude | 100 35.97 7.508184 24 48

lenses | 100 .53 .5016136 0 1

We have an overall outcome rate of 53%. So if we want to talk about whether the carrot-loving gene, gender, or latitude is associated with the risk of requiring corrective lenses by the age of 30, then relative risk is a more appropriate measure than the odds ratio. Here is a simple crosstab

of carrot and lenses, which will allow us to calculate the unadjusted OR and RR

by hand.

```
tabulate carrot lenses
```

| lenses

carrot | 0 1 | Total

----------+----------------------+----------

0 | 17 32 | 49

1 | 30 21 | 51

----------+----------------------+----------

Total | 47 53 | 100

It is interesting that fewer people with the carrot-loving gene needed corrective lenses (especially since these are fake data!). The OR and RR for those without the carrot gene vs. those with it are:

OR = (32/17)/(21/30) = 2.69

RR = (32/49)/(21/51) = 1.59

We could use either command `logit` or command `glm` to calculate
the OR. Since command `glm` will be used to calculate the RR, it will also be used to calculate the OR for comparison purposes (and it gives the same results as command `logit`). Here is the logistic regression with just carrot as the predictor:

```
glm lenses ib1.carrot, fam(bin) nolog
```

Generalized linear models No. of obs = 100
Optimization : ML Residual df = 98
Scale parameter = 1
Deviance = 132.366467 (1/df) Deviance = 1.350678
Pearson = 100 (1/df) Pearson = 1.020408

Variance function: V(u) = u*(1-u)
Link function : g(u) = ln(u/(1-u))

AIC = 1.363665
Log likelihood = -66.1832335 BIC = -318.9402

| OIM
lenses | Coef. Std. Err. z P>|z|
---+---
0.carrot | .9891975 .4135528 2.39 0.017 .1786489 1.799746
_cons | -.3566749 .2845213 -1.25 0.210 -.9143265 .2009766

```
glm lenses ib1.carrot, fam(bin) nolog eform
```

Generalized linear models No. of obs = 100

Optimization : ML Residual df = 98
Scale parameter = 1
Deviance = 132.366467 (1/df) Deviance = 1.350678
Pearson = 100 (1/df) Pearson = 1.020408

Variance function: V(u) = u*(1-u)
Link function : g(u) = ln(u/(1-u))

AIC = 1.363665
Log likelihood = -66.1832335 BIC = -318.9402

--
| OIM
lenses | Odds Ratio Std. Err. z P>|z|
-------------+--
0.carrot | 2.689076 1.112075 2.39 0.017 1.195601 6.048112
_cons | .7 .1991649 -1.25 0.210 .4007865 1.222596
--

The `eform` option gives us the same OR we calculated by hand above for those without the carrot gene versus those with it. Now this can be contrasted with the two methods of calculating the RR described below.

Relative risk estimation by log-binomial regression

With a very minor modification of the statements used above for the logistic regression, a log-binomial model can be run to get the RR instead of the OR. All that needs to be changed is the link function between the covariate(s) and outcome. Here it is specified as log instead of logit:

```
glm lenses ib1.carrot, fam(bin) link(log) nolog
```

Generalized linear models No. of obs = 100
Optimization : ML Residual df = 98
Scale parameter = 1
Deviance = 132.366467 (1/df) Deviance = 1.350678
Pearson = 100.0000007 (1/df) Pearson = 1.020408

Variance function: V(u) = u*(1-u)
Link function : g(u) = ln(u)

AIC = 1.363665
Log likelihood = -66.1832335 BIC = -318.9402

--

| OIM
lenses | Coef. Std. Err. z P>|z|

------------+--

0.carrot | .4612188 .1971117 2.34 0.019 .0748869 .8475507

_cons | -.8873031 .1673655 -5.30 0.000 -1.215333 -.5592728

--

```
glm lenses ib1.carrot, fam(bin) link(log) nolog eform
```**Generalized linear models No. of obs = 100**

Optimization : ML Residual df = 98

Scale parameter = 1

Deviance = 132.366467 (1/df) Deviance = 1.350678

Pearson = 100.0000007 (1/df) Pearson = 1.020408

Variance function: V(u) = u*(1-u)

Link function : g(u) = ln(u)

AIC = 1.363665

Log likelihood = -66.1832335 BIC = -318.9402

--

| OIM

lenses | Risk Ratio Std. Err. z P>|z|

------------+--

0.carrot | 1.586006 .3126203 2.34 0.019 1.077762

2.333923

_cons | .4117647 .0689152 -5.30 0.000 .2966111 .5716246

Now the `eform` option gives us the estimated RR instead of the OR, and it also matches what was calculated by hand above for the RR. Notice that the standard error (SE) for the beta estimate calculated here is much smaller than that calculated in the logistic regression above (SE = 0.414), but so is the estimate itself (logistic regression beta estimate = 0.989), so the significance level is very similar (logistic regression p = 0.017) in this case. One of the criticisms of using the log-binomial model for the RR is that it produces confidence intervals that are narrower than they should be, and another is that there can be convergence problems (,). This is why the second approach is also presented here.

Relative risk estimation by Poisson regression with robust error variance

Zou () suggests using a "modified Poisson" approach to estimate the relative risk and confidence intervals by using robust error variances. Using a Poisson model

without robust error variances will result in a confidence interval that is too wide. The robust error variances can be estimated by using the `robust` option, as Zou cleverly points out. Here is how it is done:

```
glm lenses ib1.carrot, fam(poisson) link(log) nolog vce(robust)
```

Generalized linear models No. of obs = 100
Optimization : ML Residual df = 98
Scale parameter = 1
Deviance = 64.53613549 (1/df) Deviance = .658532
Pearson = 46.99999999 (1/df) Pearson = .4795918

Variance function: V(u) = u
Link function : g(u) = ln(u)

AIC = 1.745361
Log pseudolikelihood = -85.26806774 BIC = -386.7705

--
| Robust
lenses | Coef. Std. Err. z P>|z|
-------------+--
0.carrot | .4612188 .1981048 2.33 0.020 .0729406 .849497
_cons | -.8873032 .1682086 -5.28 0.000 -1.216986 -

.5576203

```
glm lenses ib1.carrot, fam(poisson) link(log) nolog vce(robust) eform
```

Generalized linear models No. of obs = 100
Optimization : ML Residual df = 98
Scale parameter = 1
Deviance = 64.53613549 (1/df) Deviance = .658532
Pearson = 46.99999999 (1/df) Pearson = .4795918

Variance function: V(u) = u
Link function : g(u) = ln(u)

AIC = 1.745361
Log pseudolikelihood = -85.26806774 BIC = -386.7705

| Robust
lenses | IRR Std. Err. z P>|z|
---+---
0.carrot | 1.586006 .3141953 2.33 0.020 1.075667 2.33847
_cons | .4117647 .0692624 -5.28 0.000 .2961213 .57257

Again, the `eform` option gives us the estimated RR, and it

matches exactly what was calculated by the log-binomial method. In this case, the SE for the beta estimate and the p-value are also exactly the same as in the log-binomial model. This may not always be the case, but they should be similar. The SE calculated without the
robust option is 0.281, and the p-value is 0.101, so the robust method is quite different
(see the output below).

```
glm lenses ib1.carrot, fam(poisson) link(log) nolog
```

Generalized linear models No. of obs = 100
Optimization : ML Residual df = 98
Scale parameter = 1
Deviance = 64.53613549 (1/df) Deviance = .658532
Pearson = 46.99999999 (1/df) Pearson = .4795918

Variance function: V(u) = u
Link function : g(u) = ln(u)

AIC = 1.745361
Log likelihood = -85.26806774 BIC = -386.7705

--

| OIM

lenses | Coef. Std. Err. z P>|z|

----------+----------

0.carrot | .4612188 .2808363 1.64 0.101 -.0892103 1.011648

_cons | -.8873032 .2182179 -4.07 0.000 -1.315002 -.459604

Adjusting the relative risk for continuous or categorical covariates

Adjusting the RR for other predictors or potential confounders is simply done by adding them to the model statement as you would in any other procedure. Here gender and latitude will be added to the model:

```
glm lenses ib1.carrot ib2.gender latitude, fam(poisson) link(log) nolog vce(robust)
```

Generalized linear models No. of obs = 100

Optimization : ML Residual df = 96

Scale parameter = 1

Deviance = 63.7617568 (1/df) Deviance = .664185

Pearson = 46.7434144 (1/df) Pearson = .4869106

Variance function: V(u) = u

Link function : g(u) = ln(u)

AIC = 1.777618

Log pseudolikelihood = -84.8808784 BIC = -378.3346

| Robust

lenses | Coef. Std. Err. z P>|z|

---+---

0.carrot | .4832204 .1963616 2.46 0.014 .0983587 .8680821

1.gender | .2052008 .1857414 1.10 0.269 -.1588456 .5692472

latitude | -.0100092 .0128142 -0.78 0.435 -.0351246 .0151061

_cons | -.6521218 .4929193 -1.32 0.186 -1.618226 .3139822

```
glm lenses ib1.carrot ib2.gender latitude, fam(poisson) link(log) nolog vce(robust) eform
```

Generalized linear models No. of obs = 100

Optimization : ML Residual df = 96

Scale parameter = 1

Deviance = 63.7617568 (1/df) Deviance = .664185

Pearson = 46.7434144 (1/df) Pearson = .4869106

Variance function: V(u) = u

Link function : g(u) = ln(u)

AIC = 1.777618

Log pseudolikelihood = -84.8808784 BIC = -378.3346

```
------------------------------------------------------------------------------
             |               Robust
     lenses |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0.carrot     |   1.621287   .3183586     2.46   0.014     1.103358    2.382337
1.gender     |   1.227772    .228048     1.10   0.269     .8531281    1.766936
   latitude  |   .9900407   .0126866    -0.78   0.435     .9654851    1.015221
      _cons  |   .5209393    .256781    -1.32   0.186     .1982501    1.368865
------------------------------------------------------------------------------
```

We have also requested the RRs for gender and latitude in the estimate statement. In this case, adjusting for them does not reduce the association between having the carrot-loving gene and risk of needing corrective lenses by age 30.

One should always pay attention to goodness of fit statistics and perform other diagnostic tests.

References

1. 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(10):940-3.
2. Zou G. A
Modified Poisson Regression Approach to Prospective Studies with Binary Data.
Am J Epidemiol 2004; 159(7):702-6.
3. Sander Greenland ,
Model-based
Estimation of Relative Risks and Other Epidemiologic Measures in Studies of
Common Outcomes and in Case-Control Studies,
American Journal of Epidemiology 2004;160:301-305
4. Cook TD. Up with
odds ratios! A case for odds ratios when outcomes are common. Acad Emerg Med
2002; 9:1430-4.
5. Spiegelman, D. und Hertzmark,

Easy SAS

Calculations for Risk or Prevalence Ratios and Differences, E American

Journal of Epidemiology, 2005, 162, 199-205.