The Stata Journal (2011)
11, Number 2, pp. 290–298

Estimating adjusted risk ratios for matched and unmatched data: An update

Peter Cummings
Department of Epidemiology
School of Public Health and
Harborview Injury Prevention and Research Center
University of Washington
Seattle, WA
peterc@u.washington.edu

Abstract. The Stata 11 `margins` command makes it easier to estimate adjusted risk ratios, and the new robust variance option for `xtpoisson, fe` provides correct confidence intervals for adjusted risk ratios from matched-cohort data.

Keywords: st0162_1, conditional Poisson regression, margins, matched-cohort design, risk ratio, standardization, xtpoisson

1 Introduction

Previously, I reviewed Stata commands that allow estimation of adjusted risk ratios from unmatched (Cummings 2009) and matched (Cummings and McKnight 2004) cross-sectional, cohort, or clinical trial data. This update shows how the `margins` command in Stata version 11 and the robust variance option (`vce(robust)`) for conditional Poisson regression (`xtpoisson, fe`) in Stata version 11.1 make it easier to estimate adjusted risk ratios with appropriate confidence intervals.

2 Estimating risk ratios in unmatched data

I will use data from table 5.3 in Newman's (2001, 98 and 126) textbook for 192 women who were diagnosed with breast cancer in Canada and were followed for five years. The goal is to estimate the risk ratio for death at five years among women who had low estrogen receptor levels in their breast cancer tissue compared with women who had high receptor levels, adjusted for cancer stage (I, II, or III). The data are tabulated below:

stage	low	died	count
1	0	0	50
1	0	1	5
1	1	0	10
1	1	1	2
2	0	0	57
2	0	1	17
2	1	0	13
2	1	1	9
3	0	0	6
3	0	1	9
3	1	0	2
3	1	1	12

Previously, I discussed seven methods for estimating adjusted risk ratios (Cummings 2009): 1) Mantel–Haenszel and inverse-variance stratified methods; 2) generalized linear regression with a log link and binomial distribution; 3) generalized linear regression with a log link, normal distribution, and robust variance estimator; 4) Poisson regression with a robust variance estimator; 5) Cox proportional hazards regression with a robust variance estimator; 6) standardized risk ratios from logistic, probit, complementary log-log, and log-log regression; and 7) a substitution method. Here I discuss only standardized risk ratios from regression models (Lane and Nelder 1982; Flanders and Rhodes 1987; Greenland 2004; Rothman, Greenland, and Lash 2008, 442–446; Localio, Margolis, and Berlin 2007) to show how these can be estimated using the `margins` command.

After fitting a regression model for binomial outcomes (logistic, probit, log-log, or complementary log-log regression), we can first estimate the average risk of death that would be expected if all 192 women had low estrogen receptor tumors and they had the distribution of cancer stages for all women in the observed data. A second average risk can be estimated assuming all 192 women had a high estrogen receptor tumor. The risk ratio is the first average risk divided by the second, and the standard error for this risk ratio can be estimated using the delta method. This risk ratio is said to be standardized to the distribution of the variables used to estimate the average risks, which is cancer stage in this example. First, let me show how this can be done using the `predictnl` command after logistic regression:

```
. logistic died low stage2 stage3, nolog
Logistic regression                               Number of obs   =        192
                                                  LR chi2(3)      =      42.27
                                                  Prob > chi2     =     0.0000
Log likelihood = -92.939847                       Pseudo R2       =     0.1853

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         low |   2.508065   .9916923     2.33   0.020     1.155507    5.443836
      stage2 |   3.109772    1.44851     2.44   0.015     1.248087    7.748406
      stage3 |    18.8389   11.03231     5.01   0.000     5.978343    59.36498
------------------------------------------------------------------------------

. #delimit ;
delimiter now ;
. predictnl lnrr = ln(
>   sum(1/(1+exp(-(_b[_cons]+_b[stage2]*stage2+_b[stage3]*stage3+_b[low])))) /
>   sum(1/(1+exp(-(_b[_cons]+_b[stage2]*stage2+_b[stage3]*stage3))))),
>   se(lnrr_se);
. #delimit cr
delimiter now cr
. di "Risk ratio = " exp(lnrr[_N]) _skip(4) /*
>         */ "95% CI  " exp(lnrr[_N] - invnormal(1-.05/2)*lnrr_se[_N]) /*
>         */ ", "       exp(lnrr[_N] + invnormal(1-.05/2)*lnrr_se[_N])
Risk ratio = 1.6755988    95% CI  1.0935712, 2.567397
. replace low = 0
(48 real changes made)
. predict risk0
(option pr assumed; Pr(died))
. summ risk0, meanonly
. scalar avrisk0 = r(mean)
. replace low = 1
(192 real changes made)
. predict risk1
(option pr assumed; Pr(died))
. summ risk1, meanonly
. scalar avrisk1 = r(mean)
. scalar rr = avrisk1/avrisk0
. di "Risk1 = " avrisk1 "  Risk0 = " avrisk0 "  Risk ratio = " rr
Risk1 = .40087947  Risk0 = .2392455  Risk ratio = 1.6755988
```

While the above commands do the job, they are rather busy; the estimation of the log of the risk ratio by `predictnl` is cumbersome and possibly prone to typing errors. Below I show how the same results can be obtained using new features of Stata 11: 1) factor-variable designators, 2) `margins`, and 3) `nlcom` after `margins`.

```
. logistic died i.(low stage), nolog
Logistic regression                                   Number of obs   =        192
                                                      LR chi2(3)      =      42.27
                                                      Prob > chi2     =     0.0000
Log likelihood = -92.939847                           Pseudo R2       =     0.1853

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.low |   2.508065   .9916923     2.33   0.020     1.155507    5.443836
             |
       stage |
          2  |   3.109772    1.44851     2.44   0.015     1.248087    7.748406
          3  |    18.8389   11.03231     5.01   0.000     5.978343    59.36498
------------------------------------------------------------------------------

. margins low, post
Predictive margins                                    Number of obs   =        192
Model VCE    : OIM
Expression   : Pr(died), predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         low |
          0  |   .2392455   .0332082     7.20   0.000     .1741586    .3043324
          1  |   .4008795   .0659652     6.08   0.000       .27159    .5301689
------------------------------------------------------------------------------

. nlcom (lnrr: ln(_b[1.low]/_b[0.low])), post
       lnrr:  ln(_b[1.low]/_b[0.low])

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnrr |   .5161706   .2177193     2.37   0.018     .0894487    .9428925
------------------------------------------------------------------------------

. display "Risk ratio = " exp(_b[lnrr]) _skip(3) /*
>         */ "95% CI =  " exp(_b[lnrr]-invnormal(1-.05/2)*_se[lnrr]) /*
>         */ ", "         exp(_b[lnrr]+invnormal(1-.05/2)*_se[lnrr])
Risk ratio = 1.6755988   95% CI =  1.0935712, 2.567397
```

In the output above, the use of `i.` in the regression command told Stata to treat both `low` and `stage` as factor (indicator) variables. This is necessary so that `margins` can recognize the factor variable `low`. The `post` option after `margins` was needed so that the estimated risks would be available to the `nlcom` command. The predicted average risks were reported with their standard errors and confidence intervals. Finally, I had `nlcom` estimate the log of the risk ratio, and I then used the `display` command to report the risk ratio with its confidence interval. Not only are the commands in Stata 11 easier to use, but they report more information compared with the Stata 10 commands.

I used `nlcom` to estimate the log of the risk ratio, but in the Stata manual section called [R] **margins postestimation** (StataCorp 2009, 1008–1009), the command was used to estimate the risk ratio directly. Let us see what happens if I follow that example:

```
. logistic died i.(low stage), nolog
  (output omitted)
. margins low, post
  (output omitted)
. nlcom (rr: _b[1.low]/_b[0.low]), post
         rr:  _b[1.low]/_b[0.low]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rr	1.675599	.3648101	4.59	0.000	.9605841	2.390614

Something seems amiss; the risk ratio estimate of 1.68 is identical to what I obtained earlier, but the p-value is smaller, the confidence interval bounds have moved toward 0, and despite a p-value < 0.001, the 95% confidence interval includes the null risk ratio of 1. What happened?

First, the p-value that `nlcom` estimated was for a test that the risk ratio of 1.68 was equal to 0; `nlcom` treats 0 as the default null hypothesis, even though we customarily wish to compare an estimated risk ratio with a null value of 1.

Second, after I had `nlcom` estimate the log of the risk ratio, I used the standard error of the log of the risk ratio to estimate confidence-interval endpoints for the log risk-ratio; those endpoints were then exponentiated to obtain the confidence intervals for the risk ratio itself. This transform-the-endpoints method produces intervals that are symmetric around the log of the risk ratio, which is desirable because the log of the risk ratio ranges from minus infinity to plus infinity, and the null estimate of no association is a log risk-ratio of 0. When I used `nlcom` to estimate standard errors and confidence intervals for the risk ratio directly, the intervals were symmetric around the risk ratio, which is not desirable because the risk ratio has an asymmetrical range from 0 to plus infinity with a null estimate of no association equal to 1. This use of `nlcom` can produce biased interval coverage and can even generate a negative lower confidence bound for a risk ratio; negative risk ratios are impossible, because risks range from 0 to 1.

If we have many observations, the confidence intervals from both of the methods described above will tend to agree. In statistical jargon, the two methods are asymptotically equivalent. But as a general rule, I think it is best to instruct `nlcom` to estimate the log of the risk ratio, not the risk ratio itself, and to take the small extra step of writing a command to estimate the confidence interval using the transform-the-endpoints method. The authors of the Stata manuals are well aware of these issues, which are nicely explained in a section called *nlcom versus eform* in the manual entry that describes the `nlcom` command (StataCorp 2009, 1207–1208).

3 Estimating risk differences in unmatched data

Adjusted risk differences can also be estimated in Stata. One method is to use the `binreg` command with the `rd` option. The other is to obtain a standardized risk difference after regression:

```
. logistic died i.(low stage), nolog
  (output omitted)
. margins low, post
  (output omitted)
. nlcom (rd: _b[1.low] - _b[0.low]), post
         rd:  _b[1.low] - _b[0.low]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rd	.161634	.0745364	2.17	0.030	.0155454	.3077226

4 Better confidence intervals for risk ratios from matched-cohort data

Stata's command for conditional Poisson regression (`xtpoisson, fe`) can estimate adjusted risk ratios from matched-cohort data, but because risk ratios are for binomial outcomes, not count outcomes, the usual variance estimator will produce standard errors, p-values, and confidence intervals that are too large (Cummings, McKnight, and Greenland 2003; Cummings, McKnight, and Weiss 2003; Cummings and McKnight 2004). A robust variance estimator can correct these problems (Wooldridge 2010, 762–764), and this option was introduced with Stata 11.1 on 3 June 2010.

To illustrate, I will use data regarding 311 Australian twin pairs (Lynskey et al. 2003); the exposure was use of cannabis prior to age 17 years, and the outcome was later use of cocaine. The user-written `csmatch` command (Cummings and McKnight 2004) can produce the counts of twin pairs with the correct risk ratio and confidence interval from Mantel–Haenszel methods:

```
. use twin.dta, clear
. csmatch cocaine exposed, group(id)
```

	Not exposed		
Exposed	Outcome=1	Outcome=0	Total
Outcome = 1	61	88	149
Outcome = 0	21	141	162
Total	82	229	311

Cohort matched-pair risk ratio	[95% Conf.	Interval]
1.81707	1.50999	2.18661

Below are results from `xtpoisson, fe` with the conventional variance estimator:

```
. xtpoisson cocaine exposed, fe irr i(id) nolog
note: 141 groups (282 obs) dropped because of all zero outcomes
Conditional fixed-effects Poisson regression    Number of obs      =       340
Group variable: id                              Number of groups   =       170
                                                Obs per group: min =         2
                                                               avg =       2.0
                                                               max =         2
                                                Wald chi2(1)       =     18.87
Log likelihood  = -107.97754                    Prob > chi2        =    0.0000
```

cocaine	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
exposed	1.817073	.2498494	4.34	0.000	1.387814	2.379104

Last are results from `xtpoisson, fe` with the robust variance option:

```
. xtpoisson cocaine exposed, fe irr i(id) nolog vce(robust)
note: 141 groups (282 obs) dropped because of all zero outcomes
Conditional fixed-effects Poisson regression    Number of obs      =       340
Group variable: id                              Number of groups   =       170
                                                Obs per group: min =         2
                                                               avg =       2.0
                                                               max =         2
                                                Wald chi2(1)       =     39.98
Log pseudolikelihood  = -107.97754              Prob > chi2        =    0.0000
                                    (Std. Err. adjusted for clustering on id)
```

cocaine	IRR	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
exposed	1.817073	.171627	6.32	0.000	1.509991	2.186606

All three methods correctly estimate the risk ratio, and with the robust variance option, `xtpoisson` estimates correct standard errors, p-values, and confidence intervals.

For a stiffer test, I simulated data for 10,000 cars. Each car crashed with three occupants, each of whom had the same risk of death, but this risk varied from car to car; risk was greater with faster crash speed, and seat belt use was less common in crashes with faster speed. Comparing a belted occupant with an unbelted occupant, the risk ratio for death was set as 0.4. In 50,000 simulations, `xtpoisson, fe` estimated this risk ratio as 0.4003. With the conventional variance estimator, the 95% confidence interval included the true value of 0.4 in 97.58% of the simulated samples: 95% confidence interval of 97.44% to 97.71% for the coverage estimate. With the robust variance estimator, the coverage was a more accurate 95.12%: 95% confidence interval of 94.93% to 95.31% for this coverage estimate.

5 Summary

Recent advances in Stata 11 and 11.1 make it easier to estimate adjusted risk ratios with approximately correct confidence intervals in unmatched and matched-cohort data.

6 Acknowledgment

This work was supported by grant R49/CE000197 from the Centers for Disease Control and Prevention, Atlanta, GA.

7 References

Cummings, P. 2009. Methods for estimating adjusted risk ratios. *Stata Journal* 9: 175–196.

Cummings, P., and B. McKnight. 2004. Analysis of matched cohort data. *Stata Journal* 4: 274–281.

Cummings, P., B. McKnight, and S. Greenland. 2003. Matched cohort methods for injury research. *Epidemiologic Reviews* 25: 43–50.

Cummings, P., B. McKnight, and N. S. Weiss. 2003. Matched-pair cohort methods in traffic crash research. *Accident Analysis and Prevention* 35: 131–141.

Flanders, W. D., and P. H. Rhodes. 1987. Large sample confidence intervals for regression standardized risks, risk ratios, and risk differences. *Journal of Chronic Diseases* 40: 697–704.

Greenland, S. 2004. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case–control studies. *American Journal of Epidemiology* 160: 301–305.

Lane, P. W., and J. A. Nelder. 1982. Analysis of covariance and standardization as instances of prediction. *Biometrics* 38: 613–621.

Localio, A. R., D. J. Margolis, and J. A. Berlin. 2007. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology* 60: 874–882.

Lynskey, M. T., A. C. Heath, K. K. Bucholz, W. S. Slutske, P. A. F. Madden, E. C. Nelson, D. J. Statham, and N. G. Martin. 2003. Escalation of drug use in early-onset cannabis users vs. co-twin controls. *Journal of the American Medical Association* 289: 427–433.

Newman, S. C. 2001. *Biostatistical Methods in Epidemiology*. New York: Wiley.

Rothman, K. J., S. Greenland, and T. L. Lash. 2008. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins.

StataCorp. 2009. *Stata 11 Base Reference Manual*. College Station, TX: Stata Press.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

About the author

Peter Cummings is professor emeritus in the Department of Epidemiology in the School of Public Health and is a faculty member at the Harborview Injury Research and Prevention Research Center, University of Washington, Seattle, WA. He lives in Bishop, CA.