The Stata Journal

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

The Stata Journal publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those teaching, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the Stata Journal, including information for authors, see the web page

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates:

Subscriptions mailed to US and Canadian addresses:

3-year subscription (includes printed and electronic copy)	$153
2-year subscription (includes printed and electronic copy)	$110
1-year subscription (includes printed and electronic copy)	$ 59
1-year student subscription (includes printed and electronic copy)	$ 35

Subscriptions mailed to other countries:

3-year subscription (includes printed and electronic copy)	$225
2-year subscription (includes printed and electronic copy)	$158
1-year subscription (includes printed and electronic copy)	$ 83
1-year student subscription (includes printed and electronic copy)	$ 59
3-year subscription (electronic only)	$153

Back issues of the Stata Journal may be ordered online at

http://www.stata.com/bookstore/sj.html

The Stata Journal is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the Stata Journal, StataCorp, 4905 Lakeway Drive, College Station TX 77845, USA, or email sj@stata.com.

The Stata Journal (2004)
4, Number 3, pp. 274–281

Analysis of matched cohort data

Peter Cummings Barbara McKnight

School of Public Health and Community Medicine
University of Washington
Seattle, WA

Abstract. Matching is occasionally used in cohort studies; examples include studies of twins and some studies of traffic crashes. Analysis of matched cohort data is not discussed in many textbooks or articles and is not mentioned in the Stata manuals. Risk ratios can be estimated using matched-pair cohort data with Stata's `mcc` command. We describe a new command, `csmatch`, which can produce these risk ratios and is often more convenient. We briefly review flexible regression methods that can estimate risk ratios in matched cohort data: conditional Poisson regression and some versions of Cox regression.

Keywords: st0070, csmatch, cohort study, conditional Poisson regression, matching, matched-pair, matched cohort study, risk ratio, odds ratio

1 Introduction

When conducting a case–control study, investigators sometimes match each case to a control on a factor that might confound the association of the study exposure with the study outcome. If the matching is exact, accounting for the matching in the analysis will eliminate confounding by the matching variable. The analytic method commonly used to account for the matching in case–control studies is conditional logistic regression; the estimated odds ratios this method produces will approximate risk ratios if the outcome is sufficiently rare, as it usually is in a case–control study (Koepsell and Weiss 2003, 203–205). Mantel–Haenszel methods (Mantel and Haenszel 1959) can be used to analyze matched case–control data. For a dichotomous exposure, matched case–control data can be displayed in a 2×2 contingency table in which the cell frequencies are counts of pairs:

	Controls	
Cases	Exposed	Not exposed
Exposed	A	B
Not exposed	C	D

Only the counts in cells B and C—the cells discordant on both exposure and outcome—are needed to estimate the odds ratio, a confidence interval, and a p-value. The odds ratio for the outcome (becoming a case) among those exposed, compared with those not exposed, is B/C. The p-value for a test that this odds ratio differs from 1 is derived from the well-known McNemar test (McNemar 1947), which uses only information from cells B and C. This method is implemented in the Stata `mcc` command.

Matching also can be used in a cohort study to prevent confounding of the rate ratio, risk ratio, or hazard ratio. If each exposed subject is matched with a subject not exposed, with regard to the potential confounding variable, confounding will be avoided, provided that there is no loss to follow-up. Unlike in a case–control study, there is no need to account for the matching in the analysis to avoid bias. However, an analysis that does account for the matching may offer an advantage in some studies: a matched-pair analysis only requires data from matched pairs in which one or both had the study outcome (Rothman and Greenland 1998, 283–285; Cummings, McKnight, and Weiss 2003; Cummings, McKnight, and Greenland 2003). The matched-pair risk ratio can be estimated even if the analyst has no information regarding the pairs in which no subject had the outcome.

Although the Stata manuals do not mention the matched cohort design, Stata's `mcc` command provides output for the matched-pair risk ratio. The method implemented in the command is an extension of Mantel–Haenszel methods (Nurminen 1981; Rothman and Greenland 1998, 283–285). The `mcc` command's output swaps exposure and outcome labels if the data are from a matched-pair cohort study; those exposed are labeled "cases", those not exposed are labeled "controls", those with the outcome are labeled "exposed", and those without the outcome are labeled "unexposed". The `mcc` command also has the disadvantage of requiring that each study pair be in the same row of a Stata data file. This data format is different from that required by conditional logistic regression and other commands, which require that each study subject be in their own row of data; a `group` option is used to identify which subjects are in the same pair. To shift from conditional logistic regression to the `mcc` command, the data file must be reformatted.

We have written a Stata command, `csmatch`, that estimates the matched risk ratio. When applied to matched-pair cohort data, this command correctly labels exposure and outcome. Furthermore, the `csmatch` command has a `group()` option that allows the user to analyze data in the same format that Stata uses for regression commands.

Like Stata's `mcc` command, `csmatch` is chiefly useful for teaching purposes to illustrate how matched cohort data may be analyzed. More flexible methods for matched cohort analyses are available in Stata and these will be discussed in the last section of this article.

2 The csmatch command

2.1 Syntax

`csmatch` *depvar expvar* [`if` *exp*] [`in` *range*], `group(`*varname*`)` [`level(.#)` `personvar(`*varlist*`)` `pairvar(`*varlist*`)`]

2.2 Description

`csmatch` estimates the risk ratio for the outcome, *depvar*, given the exposure, *expvar*; *depvar* and *expvar* must be binary and coded as 0 or 1.

2.3 Options

`group(`*varname*`)` specifies the identifier variable (numeric or string) for the matched pairs. The data must be organized so that there is one record for each person, i.e., two records for each pair.

`level(.#)` specifies the confidence level, as a fraction, for the estimates. Unlike many Stata commands, `level()` must be a fraction between 0 and 1, such as .95, not a percentage, such as 95%. The default is `level(.95)`.

`personvar(`*varlist*`)` specifies a list of potential confounding variables that are specific to a person or individual, such as age or sex. These must be numeric.

`pairvar(`*varlist*`)` specifies a list of variables that are the same for each member of a pair but may differ between pairs. If you studied vehicle occupants paired in their cars, examples might include speed or crash angle. These must be numeric.

3 An example of a matched-pair cohort study

To study risk factors for drug use, investigators (Lynskey et al. 2003) used data from a registry of Australian twins to estimate the association of early cannabis use with the use of other drugs at a later age. They selected all twin pairs ($n = 311$) among whom one had used cannabis before age 17 years (the exposed twin) and the other had not. By studying pairs of twins, they eliminated confounding by all genetic, family, and environmental factors that were shared by the twins. Data on subsequent cocaine use for the 311 pairs may be analyzed using the immediate version of Stata's `mcc` command:

(*Continued on next page*)

```
. mcci 61 88 21 141
```

Cases	Controls: Exposed	Controls: Unexposed	Total
Exposed	61	88	149
Unexposed	21	141	162
Total	82	229	311

```
McNemar's chi2(1) =     41.18    Prob > chi2 = 0.0000
Exact McNemar significance probability       = 0.0000

Proportion with factor
        Cases       .4790997
        Controls    .2636656     [95% Conf. Interval]

        difference  .2154341      .1509335   .2799346
        ratio       1.817073      1.509991   2.186606
        rel. diff.  .2925764      .2174201   .3677328
        odds ratio  4.190476       2.58041   7.104024   (exact)
```

The outcome of cocaine use was more common among those who used marijuana when they were young (149/311 = .48) compared with those who did not use marijuana (82/311 = .26); these risks are given in the output under the heading "Proportion with factor". The matched-pair odds ratio for cocaine use among those who used marijuana before they reached 17 years, compared with those who did not use marijuana, is simply B/C = 88/21 = 4.2. The investigators (Lynskey et al. 2003) used conditional logistic regression to adjust for alcohol use, tobacco use, and other variables and reported an adjusted odds ratio of 4.06.

If we use `csmatch` to analyze these matched-pair cohort data, we obtain

```
. csmatch cocaine exposed, group(id)
```

Exposed	Not exposed: Outcome=1	Not exposed: 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
```

The cell counts are the same in the `mcci` and `csmatch` tables, and the labels for exposure and outcome are now correct for a cohort study in the `csmatch` table.

In a case–control study, each case is matched with a noncase, so there can be no pairs in which both have the outcome; pairs in which both have the outcome appear in cell A of a 2×2 contingency table of pair counts using the format of the `csmatch` command. In a cohort study, the risk of the outcome might be great, resulting in some pairs that both have the outcome. This was the situation in the Australian twin study, where there were 61 pairs in which both twins had the outcome of using cocaine. Because the underlying risk of the outcome was large, the risk ratio (1.8) in this cohort study was closer to 1 than the odds ratio (4.2).

The risk ratio of 1.8 is given in the output of the `mcc` command; it is labeled "ratio" in the output and is exactly the same as the risk ratio given by the `csmatch` command. The confidence intervals are the same for the two commands because they use the same matched-pair variance estimator. The `mcc` command produces the risks themselves, and the matched-pair risk ratio is the same as the unmatched ratio of the risks: $.479/.264 = 1.8$. The `csmatch` command counts the pairs in which no twin had the outcome but makes no use of this count to calculate the risk ratio. A matched-pair cohort study that lacks information for cell D cannot produce correct risk estimates for those exposed, $(A + B)/(A + B + C + D)$, and those not exposed, $(A + C)/(A + B + C + D)$. It can produce the risk ratio, however, as the denominator of the two risks is the same, $(A + B + C + D)$, and the risk ratio therefore reduces to $(A + B)/(A + C)$, which does not require information from cell D (Cummings, McKnight, and Weiss 2003; Cummings, McKnight, and Greenland 2003). In the twin study example, the matched-pair risk ratio was $149/82 = 1.8$. This risk ratio can be obtained, even if the 141 pairs in which no one went on to cocaine use are omitted from the analysis; without those pairs, however, the risk of future cocaine use in those exposed and not exposed to early marijuana use cannot be estimated.

4 Other Stata commands for matched cohort data

The `csmatch` and `mcc` commands are chiefly useful for teaching purposes or for preliminary analysis of matched data. In an actual analysis of matched cohort data, the investigator will usually desire a more flexible analytic method that can adjust for additional confounding variables and assess the evidence regarding statistical interaction. In Stata, two flexible options are available. Conditional Poisson regression can be used; the method produces risk ratios from matched cohort data using Stata's `xtpois, fe` command, with grouping on the matched sets defined by the `i(`*varname*`)` option. Here is the output for the twin study data:

```
. xtpois cocaine exposed, fe nolog irr i(id)
note: 141 groups (282 obs) dropped due to all zero outcomes
Conditional fixed-effects Poisson regression    Number of obs      =       340
Group variable (i): 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

The risk ratio produced by conditional Poisson regression is just the same as the risk ratio produced by the `mcc` and `csmatch` commands. But the 95% confidence interval has expanded from $1.51 - 2.19$ to $1.39 - 2.38$, since the standard errors are derived

differently under this model. The regression model output was obtained using only the 170 pairs with at least one member who had the outcome of cocaine use.

The same results can be obtained from the Cox proportional hazards model using Stata's `stcox` or `cox` commands, with stratification on the matched pairs. If follow-up time is equal for members of each matched set, the Breslow and Efron methods for accounting for ties in follow-up time will produce risk ratios, while the exact marginal and exact partial methods will produce odds ratios (Cummings, McKnight, and Weiss 2003). Here is output when time is set to 1 for all study subjects and the Breslow method is used:

```
. gen byte time = 1
. cox time exposed, strata(id) hr nolog dead(cocaine)
Stratified Cox regr. -- Breslow method for ties
Entry time 0                                    Number of obs   =        622
                                                LR chi2(1)      =      19.71
                                                Prob > chi2     =     0.0000
Log likelihood = -150.25951                     Pseudo R2       =     0.0616

------------------------------------------------------------------------------
        time | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exposed |   1.817073   .2498494     4.34   0.000     1.387814    2.379104
------------------------------------------------------------------------------
                                                              Stratified by id
```

The results above produce a hazard ratio, standard error, and confidence interval that are the same as the risk ratio, standard error, and confidence interval from the conditional Poisson method; the likelihood functions for the two methods are the same for matched pairs. To produce the matched odds ratio, we can use the exact partial (or marginal) method of accounting for ties in follow-up time:

```
. cox time exposed, strata(id) hr nolog dead(cocaine) exactp
Stratified Cox regr. -- exact partial likelihood
Entry time 0                                    Number of obs   =        622
                                                LR chi2(1)      =      44.27
                                                Prob > chi2     =     0.0000
Log likelihood = -53.416308                     Pseudo R2       =     0.2930

------------------------------------------------------------------------------
        time | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exposed |   4.190476   1.017714     5.90   0.000      2.60338    6.745112
------------------------------------------------------------------------------
                                                              Stratified by id
```

The hazard ratio of 4.2 is the same as the odds ratio produced by the `mcc` command and is the same as the odds ratio produced by conditional logistic regression, which uses the same likelihood for these data. Note that the standard errors for these regression models are different from the standard errors computed in the `mcc` command.

```
. clogit cocaine exposed, group(id) or nolog
note: multiple positive outcomes within groups encountered.
note: 202 groups (404 obs) dropped due to all positive or
      all negative outcomes.
Conditional (fixed-effects) logistic regression   Number of obs   =        218
                                                  LR chi2(1)      =      44.27
                                                  Prob > chi2     =     0.0000
Log likelihood = -53.416308                       Pseudo R2       =     0.2930
```

cocaine	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
exposed	4.190476	1.017704	5.90	0.000	2.603392	6.745082

We have reported in simulation studies that estimates of variance and confidence intervals for the risk ratio are correct using the `mcc` and `csmatch` commands but too large using the conditional Poisson or Cox methods (Cummings, McKnight, and Weiss 2003). In large datasets, the difference between the two confidence intervals is often of little practical importance. Unbiased variance estimates and confidence intervals can be obtained using bootstrap methods.

5 Saved Results

`csmatch` saves results in `r()`:

Scalars

`r(prct)`	count of matched pairs discordant on exposure in the estimation sample
`r(rr)`	risk-ratio estimate
`r(vlrr)`	variance ln risk ratio

6 References

Cummings, P., B. McKnight, and S. Greenland. 2003. Matched cohort methods in 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.

Koepsell, T. D. and N. S. Weiss. 2003. *Epidemiologic Methods: Studying the Occurrence of Illness*. Oxford: Oxford University Press.

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.

Mantel, N. and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies. *Journal of the National Cancer Institute* 22: 719–748.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12: 153–157.

Nurminen, N. 1981. Asymptotic efficiency of general noniterative estimation of common relative risk. *Biometrika* 68: 525–530.

Rothman, K. and S. Greenland. 1998. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott–Raven.

About the Authors

Peter Cummings is a Professor in the Department of Epidemiology in the School of Public Health and Community Medicine and a core faculty member at the Harborview Injury Research and Prevention Research Center, University of Washington, Seattle, WA.

Barbara McKnight is a Professor in the Department of Biostatistics in the School of Public Health and Community Medicine at the University of Washington, Seattle, WA.