

## Title

**clogit** — Conditional (fixed-effects) logistic regression

## Syntax

```
clogit depvar [indepvars] [if] [in] [weight], group(varname) [options]
```

| <i>options</i>                        | description  |
|---------------------------------------|--|
| Model                                 |  |
| * <u>group</u> ( <i>varname</i> )     | matched group variable   |
| <u>offset</u> ( <i>varname</i> )      | include <i>varname</i> in model with coefficient constrained to 1  |
| <u>constraints</u> ( <i>numlist</i> ) | apply specified linear constraints   |
| <u>collinear</u>                      | keep collinear variables   |
| SE/Robust                             |  |
| <u>vce</u> ( <i>vcetype</i> )         | <i>vcetype</i> may be <u>oim</u> , <u>robust</u> , <u>cluster</u> <i>clustvar</i> , <u>opg</u> , <u>bootstrap</u> ,<br>or <u>jackknife</u> |
| <u>nonest</u>                         | do not check that panels are nested within clusters  |
| Reporting                             |  |
| <u>level</u> (#)                      | set confidence level; default is <u>level</u> (95)   |
| <u>or</u>                             | report odds ratios   |
| Max options                           |  |
| <u>maximize_options</u>               | control the maximization process; seldom used  |

\*group(*varname*) is required.

bootstrap, by, jackknife, nestreg, rolling, statsby, stepwise, svy, and xi are allowed; see [U] **11.1.10 Prefix commands**.

Weights are not allowed with the bootstrap prefix.

vce(*vcetype*), nonest, and weights are not allowed with the svy prefix.

fweights, iweights, and pweights are allowed (see [U] **11.1.6 weight**), but they are interpreted to apply to groups as a whole, not to individual observations. See *Use of weights* below.

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

## Description

`clogit` fits what biostatisticians and epidemiologists call conditional logistic regression for matched case–control groups (see, for example, Hosmer and Lemeshow 2000, chap. 7) and what economists and other social scientists call fixed-effects logit for panel data (see, for example, Chamberlain 1980). Computationally, these models are the same.

See [R] **aslogit** if you want to fit McFadden’s choice model (McFadden 1974). Also see [R] **logistic** for a list of related estimation commands.

## Options

### Model

`group(varname)` is required; it specifies an identifier variable (numeric or string) for the matched groups. `strata(varname)` is a synonym for `group()`.

`offset(varname)`, `constraints(numlist)`, `collinear`; see [R] **estimation options**.

### SE/Robust

`vce(vctype)` specifies the type of standard error reported, which includes types that are derived from asymptotic theory, that are robust to some kinds of misspecification, that allow for intragroup correlation, and that use bootstrap or jackknife methods; see [R] *vce\_option*.

`nonest`, available only with `vce(cluster clustvar)`, prevents checking that matched groups are nested within clusters. It is the user's responsibility to verify that the standard errors are theoretically correct.

### Reporting

`level(#)`; see [R] **estimation options**.

`or` reports the estimated coefficients transformed to odds ratios, i.e.,  $e^b$  rather than  $b$ . Standard errors and confidence intervals are similarly transformed. This option affects how results are displayed, not how they are estimated. `or` may be specified at estimation or when replaying previously estimated results.

### Max options

*maximize\_options*: `difficult`, `technique(algorithm_spec)`, `iterate(#)`, `[no]log`, `trace`, `gradient`, `showstep`, `hessian`, `shownrtolerance`, `tolerance(#)`, `ltolerance(#)`, `gtolerance(#)`, `nrtolerance(#)`, `nonnrtolerance`, `from(init_specs)`; see [R] **maximize**. These options are seldom used.

Setting the optimization type to `technique(bhhh)` resets the default `vctype` to `vce(opg)`.

## Remarks

Remarks are presented under the following headings:

*Matched case-control data*  
*Use of weights*  
*Fixed-effects logit*

`clogit` fits maximum likelihood models with a dichotomous dependent variable coded as 0/1 (more precisely, `clogit` interprets 0 and not 0 to indicate the dichotomy). Conditional logistic analysis differs from regular logistic regression in that the data are grouped and the likelihood is calculated relative to each group; i.e., a conditional likelihood is used. See *Methods and Formulas* at the end of this entry.

Biostatisticians and epidemiologists fit these models when analyzing matched case-control studies with 1 : 1 matching, 1 :  $k_{2i}$  matching, or  $k_{1i}$  :  $k_{2i}$  matching, where  $i$  denotes the  $i$ th matched group for  $i = 1, 2, \dots, n$ , where  $n$  is the total number of groups. `clogit` fits a model appropriate for all these matching schemes or for any mix of the schemes since the matching  $k_{1i}$  :  $k_{2i}$  can vary from group to group. `clogit` always uses the true conditional likelihood, not an approximation. Biostatisticians and epidemiologists sometimes refer to the matched groups as “strata”, but we will stick to the more generic term “group”.

Economists and other social scientists fitting fixed-effects logit models have data that look exactly like the data biostatisticians and epidemiologists call  $k_{1i} : k_{2i}$  matched case–control data. In terms of how the data are arranged,  $k_{1i} : k_{2i}$  matching means that in the  $i$ th group the dependent variable is 1 a total of  $k_{1i}$  times and 0 a total of  $k_{2i}$  times. There are a total of  $T_i = k_{1i} + k_{2i}$  observations for the  $i$ th group. This data arrangement is what economists and other social scientists call “panel data”, “longitudinal data”, or “cross-sectional time-series data”.

So, no matter what terminology you use, the computation and the use of the `clogit` command is the same. The following example shows how your data should be arranged to use `clogit`.

### ▷ Example 1

Suppose that we have grouped data with the variable `id` containing a unique identifier for each group. Our outcome variable, `y`, contains 0s and 1s. If we were biostatisticians,  $y = 1$  would indicate a case,  $y = 0$  would be a control, and `id` would be an identifier variable that indicates the groups of matched case–controls subjects.

If we were economists,  $y = 1$  might indicate that a person was unemployed at any time during a year and  $y = 0$  that a person was employed all year, and `id` would be an identifier variable for persons.

If we list the first few observations of this dataset, it looks like

```
. use http://www.stata-press.com/data/r10/clogitid
. list y x1 x2 id in 1/11
```

|     | y | x1 | x2 | id   |
|-----|---|----|----|------|
| 1.  | 0 | 0  | 4  | 1014 |
| 2.  | 0 | 1  | 4  | 1014 |
| 3.  | 0 | 1  | 6  | 1014 |
| 4.  | 1 | 1  | 8  | 1014 |
| 5.  | 0 | 0  | 1  | 1017 |
| 6.  | 0 | 0  | 7  | 1017 |
| 7.  | 1 | 1  | 10 | 1017 |
| 8.  | 0 | 0  | 1  | 1019 |
| 9.  | 0 | 1  | 7  | 1019 |
| 10. | 1 | 1  | 7  | 1019 |
| 11. | 1 | 1  | 9  | 1019 |

Pretending that we are biostatisticians, we describe our data as follows. The first group (`id = 1014`) consists of four matched persons: 1 case ( $y = 1$ ) and three controls ( $y = 0$ ), i.e., 1 : 3 matching. The second group has 1 : 2 matching, and the third 2 : 2.

Pretending that we are economists, we describe our data as follows. The first group consists of 4 observations (one per year) for person 1014. This person had a period of unemployment during 1 year of 4. The second person had a period of unemployment during 1 year of 3, and the third had a period of 2 years of 4.

Our independent variables are `x1` and `x2`. To fit the conditional (fixed-effects) logistic model, we type

```
. clogit y x1 x2, group(id)
note: multiple positive outcomes within groups encountered.
Iteration 0:  log likelihood = -123.42828
Iteration 1:  log likelihood = -123.41386
Iteration 2:  log likelihood = -123.41386
Conditional (fixed-effects) logistic regression  Number of obs   =      369
                                                  LR chi2(2)         =        9.07
                                                  Prob > chi2        =       0.0107
Log likelihood = -123.41386                    Pseudo R2         =       0.0355
```

| y  | Coef.    | Std. Err. | z    | P> z  | [95% Conf. Interval] |
|----|----------|-----------|------|-------|----------------------|
| x1 | .653363  | .2875215  | 2.27 | 0.023 | .0898312 1.216895    |
| x2 | .0659169 | .0449555  | 1.47 | 0.143 | -.0221943 .1540281   |



### □ Technical Note

The message “note: multiple positive outcomes within groups encountered” at the top of the clogit output for the previous example merely informs us that we have  $k_{1i} : k_{2i}$  matching with  $k_{1i} > 1$  for at least one group. If your data should be  $1 : k_{2i}$  matched, this message tells you that there is an error in the data somewhere.

We can see the distribution of  $k_{1i}$  and  $T_i = k_{1i} + k_{2i}$  for the data of the previous example by using the following steps.

```
. by id, sort: gen k1 = sum(y)
. by id: replace k1 = . if _n < _N
(303 real changes made, 303 to missing)
. by id: gen T = sum(y < .)
. by id: replace T = . if _n < _N
(303 real changes made, 303 to missing)
. tab k1
```

| k1    | Freq. | Percent | Cum.   |
|-------|-------|---------|--------|
| 1     | 48    | 72.73   | 72.73  |
| 2     | 12    | 18.18   | 90.91  |
| 3     | 4     | 6.06    | 96.97  |
| 4     | 2     | 3.03    | 100.00 |
| Total | 66    | 100.00  |        |

```
. tab T
```

| T     | Freq. | Percent | Cum.   |
|-------|-------|---------|--------|
| 2     | 5     | 7.58    | 7.58   |
| 3     | 5     | 7.58    | 15.15  |
| 4     | 12    | 18.18   | 33.33  |
| 5     | 11    | 16.67   | 50.00  |
| 6     | 13    | 19.70   | 69.70  |
| 7     | 8     | 12.12   | 81.82  |
| 8     | 3     | 4.55    | 86.36  |
| 9     | 7     | 10.61   | 96.97  |
| 10    | 2     | 3.03    | 100.00 |
| Total | 66    | 100.00  |        |

We see that  $k_{1i}$  ranges from 1 to 4 and  $T_i$  ranges from 2 to 10 for these data.



□ **Technical Note**

For  $k_{1i} : k_{2i}$  matching (and hence in the general case of fixed-effects logit), `clogit` uses a recursive algorithm to compute the likelihood, which means that there are no limits on the size of  $T_i$ . However, computation time is proportional to  $\sum T_i \min(k_{1i}, k_{2i})$ , so `clogit` will take roughly 10 times longer to fit a model with 10 : 10 matching than one with 1 : 10 matching. But `clogit` is fast, so computation time becomes an issue only when  $\min(k_{1i}, k_{2i})$  is around 100 or more. See *Methods and Formulas* for details.

□

**Matched case–control data**

Here we give a more detailed example of matched case–control data.

▷ **Example 2**

Hosmer and Lemeshow (2000, 25) present data on matched pairs of infants, each pair having one with low birthweight and another with regular birthweight. The data are matched on age of the mother. Several possible maternal exposures are considered: race (three categories), smoking status, presence of hypertension, presence of uterine irritability, previous preterm delivery, and weight at the last menstrual period.

```
. use http://www.stata-press.com/data/r10/lowbirth, clear
(Applied Logistic Regression, Hosmer & Lemeshow, pp. 262-265)
. describe
Contains data from http://www.stata-press.com/data/r10/lowbirth.dta
  obs:                112                Applied Logistic Regression,
                                         Hosmer & Lemeshow, pp. 262-265
  vars:                11                21 Jan 2007 11:48
  size:               1,792 (99.8% of memory free)
```

| variable name | storage type | display format | value label | variable label                   |
|---------------|--------------|----------------|-------------|----------------------------------|
| pairid        | byte         | %8.0g          |             | Case-control pair id             |
| low           | byte         | %8.0g          |             | Baby has low birth weight        |
| age           | byte         | %8.0g          |             | Age of mother                    |
| lwt           | int          | %8.0g          |             | Mother's last menstrual weight   |
| smoke         | byte         | %8.0g          |             | Mother smoked during pregnancy   |
| ptd           | byte         | %8.0g          |             | Mother had previous preterm baby |
| ht            | byte         | %8.0g          |             | Mother has hypertension          |
| ui            | byte         | %8.0g          |             | Uterine irritability             |
| race1         | byte         | %8.0g          |             | mother is white                  |
| race2         | byte         | %8.0g          |             | mother is black                  |
| race3         | byte         | %8.0g          |             | mother is other                  |

Sorted by:

We list the case–control indicator variable, `low`; the match identifier variable, `pairid`; and two of the covariates, `lwt` and `smoke`, for the first 10 observations.

```
. list low lwt smoke pairid in 1/10
```

|     | low | lwt | smoke | pairid |
|-----|-----|-----|-------|--------|
| 1.  | 0   | 135 | 0     | 1      |
| 2.  | 1   | 101 | 1     | 1      |
| 3.  | 0   | 98  | 0     | 2      |
| 4.  | 1   | 115 | 0     | 2      |
| 5.  | 0   | 95  | 0     | 3      |
| 6.  | 1   | 130 | 0     | 3      |
| 7.  | 0   | 103 | 0     | 4      |
| 8.  | 1   | 130 | 1     | 4      |
| 9.  | 0   | 122 | 1     | 5      |
| 10. | 1   | 110 | 1     | 5      |

We fit a conditional logistic model of low birthweight on mother’s weight, race, smoking behavior, and history.

```
. clogit low lwt smoke ptd ht ui race2 race3, strata(pairid) nolog
Conditional (fixed-effects) logistic regression   Number of obs   =       112
                                                LR chi2(7)       =       26.04
                                                Prob > chi2      =       0.0005
Log likelihood = -25.794271                    Pseudo R2       =       0.3355
```

|       | low | Coef.     | Std. Err. | z     | P> z  | [95% Conf. Interval] |
|-------|-----|-----------|-----------|-------|-------|----------------------|
| lwt   |     | -.0183757 | .0100806  | -1.82 | 0.068 | -.0381333 .0013819   |
| smoke |     | 1.400656  | .6278396  | 2.23  | 0.026 | .1701131 2.631199    |
| ptd   |     | 1.808009  | .7886502  | 2.29  | 0.022 | .2622828 3.353735    |
| ht    |     | 2.361152  | 1.086128  | 2.17  | 0.030 | .2323796 4.489924    |
| ui    |     | 1.401929  | .6961585  | 2.01  | 0.044 | .0374836 2.766375    |
| race2 |     | .5713643  | .689645   | 0.83  | 0.407 | -.7803149 1.923044   |
| race3 |     | -.0253148 | .6992044  | -0.04 | 0.971 | -1.39573 1.345101    |

We might prefer to see results presented as odds ratios. We could have specified the `or` option when we first fitted the model, or we can now redisplay results and specify `or`:

```
. clogit, or
Conditional (fixed-effects) logistic regression   Number of obs   =       112
                                                LR chi2(7)       =       26.04
                                                Prob > chi2      =       0.0005
Log likelihood = -25.794271                    Pseudo R2       =       0.3355
```

|       | low | Odds Ratio | Std. Err. | z     | P> z  | [95% Conf. Interval] |
|-------|-----|------------|-----------|-------|-------|----------------------|
| lwt   |     | .9817921   | .009897   | -1.82 | 0.068 | .9625847 1.001383    |
| smoke |     | 4.057862   | 2.547686  | 2.23  | 0.026 | 1.185439 13.89042    |
| ptd   |     | 6.098293   | 4.80942   | 2.29  | 0.022 | 1.299894 28.60938    |
| ht    |     | 10.60316   | 11.51639  | 2.17  | 0.030 | 1.261599 89.11467    |
| ui    |     | 4.06303    | 2.828513  | 2.01  | 0.044 | 1.038195 15.90088    |
| race2 |     | 1.770681   | 1.221141  | 0.83  | 0.407 | .4582617 6.84175     |
| race3 |     | .975003    | .6817263  | -0.04 | 0.971 | .2476522 3.838573    |

Smoking, previous preterm delivery, hypertension, uterine irritability, and possibly the mother’s weight all contribute to low birthweight. `race2` (mother black) and `race3` (mother other) are

statistically insignificant when compared with the `race1` (mother white) omitted group, although the `race2` effect is large. We can test the joint statistical significance of `race2` and `race3` by using `test`:

```
. test race2 race3
( 1) race2 = 0
( 2) race3 = 0

      chi2( 2) =    0.88
      Prob > chi2 = 0.6436
```

For a more complete description of `test`, see [R] `test`. `test` presents results in coefficients rather than odds ratios. Jointly testing that the coefficients on `race2` and `race3` are zero is equivalent to jointly testing that the odds ratios are 1.

Here one case was matched to one control, i.e., 1 : 1 matching. From `clogit`'s point of view, that was not important— $k_1$  cases could have been matched to  $k_2$  controls ( $k_1 : k_2$  matching), and we would have fitted the model in the same way. Furthermore, the matching can change from group to group, which we have denoted as  $k_{1i} : k_{2i}$  matching, where  $i$  denotes the group. `clogit` does not care. To fit the conditional logistic regression model, we specified the `group(varname)` option, `group(pairid)`. The case and control are stored in separate observations. `clogit` knew that they were linked (in the same group) because the related observations share the same value of `pairid`. ◀

## □ Technical Note

`clogit` provides a way to extend McNemar's test to multiple controls per case (1 :  $k_{2i}$  matching) and to multiple controls matched with multiple cases ( $k_{1i} : k_{2i}$  matching).

In Stata, McNemar's test is calculated by the `mcc` command; see [ST] `epitab`. The `mcc` command, however, requires that the matched case and control appear in one observation, so the data will need to be manipulated from 1 to 2 observations per stratum before using `clogit`. Alternatively, if you begin with `clogit`'s 2-observations-per-group organization, you will have to change it to 1 observation per group if you wish to use `mcc`. In either case, `reshape` provides an easy way to change the organization of the data. We will demonstrate its use below, but we direct you to [D] `reshape` for a more thorough discussion.

In the example above, we used `clogit` to analyze the relationship between low birthweight and various characteristics of the mother. Assume that we now want to assess the relationship between low birthweight and smoking, ignoring the mother's other characteristics. Using `clogit`, we obtain the following results:

```
. clogit low smoke, strata(pairid) or
Iteration 0:  log likelihood = -35.425931
Iteration 1:  log likelihood = -35.419283
Iteration 2:  log likelihood = -35.419282
Conditional (fixed-effects) logistic regression   Number of obs   =       112
                                                    LR chi2(1)      =        6.79
                                                    Prob > chi2     =       0.0091
                                                    Pseudo R2      =       0.0875
Log likelihood = -35.419282
```

|  | low   | Odds Ratio | Std. Err. | z    | P> z  | [95% Conf. Interval] |
|--|-------|------------|-----------|------|-------|----------------------|
|  | smoke | 2.75       | 1.135369  | 2.45 | 0.014 | 1.224347 6.176763    |

Let us compare our estimated odds ratio and 95% confidence interval with that produced by `mcc`. We begin by reshaping the data:

```
. keep low smoke pairid
. reshape wide smoke, i(pairid) j(low 0 1)
Data                long   ->   wide
-----
Number of obs.      112   ->    56
Number of variables    3   ->    3
j variable (2 values)  low   -> (dropped)
xij variables:
                    smoke   ->  smoke0 smoke1
```

We now have the variables `smoke0` (formed from `smoke` and `low = 0`), recording 1 if the control mother smoked and 0 otherwise; and `smoke1` (formed from `smoke` and `low = 1`), recording 1 if the case mother smoked and 0 otherwise. We can now use `mcc`:

```
. mcc smoke1 smoke0
```

| Cases     | Controls |           | Total |
|-----------|----------|-----------|-------|
|           | Exposed  | Unexposed |       |
| Exposed   | 8        | 22        | 30    |
| Unexposed | 8        | 18        | 26    |
| Total     | 16       | 40        | 56    |

```
McNemar's chi2(1) = 6.53 Prob > chi2 = 0.0106
Exact McNemar significance probability = 0.0161
Proportion with factor
Cases      .5357143
Controls   .2857143 [95% Conf. Interval]
difference .25      .0519726 .4480274
ratio      1.875    1.148685 3.060565
rel. diff. .35      .1336258 .5663742
odds ratio 2.75    1.179154 7.143667 (exact)
```

Both methods estimated the same odds ratio, and the 95% confidence intervals are similar. `clogit` produced a confidence interval of [1.22, 6.18], whereas `mcc` produced a confidence interval of [1.18, 7.14].



## Use of weights

With `clogit`, weights apply to groups as a whole, not to individual observations. For example, if there is a group in your dataset with a frequency weight of 3, there are a total of three groups in your sample with the same values of the dependent and independent variables as this one group. Weights must have the same value for all observations belonging to the same group; otherwise, an error message will be displayed.

## ▷ Example 3

We use the example from the above discussion of the `mcc` command. Here we have a total of 56 matched case–control groups, each with one case matched to one control. We had 8 matched pairs in which both the case and the control are exposed, 22 pairs in which the case is exposed and the control is unexposed, 8 pairs in which the case is unexposed and the control is exposed, and 18 pairs in which they are both unexposed.

With weights, it is easy to enter these data into Stata and run `clogit`.

```
. clear
. input id case exposed weight
      id      case  exposed  weight
1.  1  1  1  8
2.  1  0  1  8
3.  2  1  1  22
4.  2  0  0  22
5.  3  1  0  8
6.  3  0  1  8
7.  4  1  0  18
8.  4  0  0  18
9.  end

. clogit case exposed [w=weight], strata(id) or
(frequency weights assumed)
Iteration 0:  log likelihood = -35.425931
Iteration 1:  log likelihood = -35.419283
Iteration 2:  log likelihood = -35.419282

Conditional (fixed-effects) logistic regression   Number of obs   =       112
LR chi2(1)                                       =         6.79
Prob > chi2                                       =       0.0091
Pseudo R2                                         =       0.0875

Log likelihood = -35.419282
```

| case    | Odds Ratio | Std. Err. | z    | P> z  | [95% Conf. Interval] |
|---------|------------|-----------|------|-------|----------------------|
| exposed | 2.75       | 1.135369  | 2.45 | 0.014 | 1.224347 6.176763    |

◀

**Fixed-effects logit**

The fixed-effects logit model can be written as

$$\Pr(y_{it} = 1 \mid \mathbf{x}_{it}) = F(\alpha_i + \mathbf{x}_{it}\beta)$$

where  $F$  is the cumulative logistic distribution

$$F(z) = \frac{\exp(z)}{1 + \exp(z)}$$

$i = 1, 2, \dots, n$  denotes the independent units (called “groups” by `clogit`), and  $t = 1, 2, \dots, T_i$  denotes the observations for the  $i$ th unit (group).

Fitting this model by using a full maximum-likelihood approach leads to difficulties, however. When  $T_i$  is fixed, the maximum likelihood estimates for  $\alpha_i$  and  $\beta$  are inconsistent (Andersen 1970 and Chamberlain 1980). This difficulty can be circumvented by looking at the probability of  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})$  conditional on  $\sum_{t=1}^{T_i} y_{it}$ . This conditional probability does not involve the  $\alpha_i$ , so they are never estimated when the resulting conditional likelihood is used. See Hamerle and Ronning (1995) for a succinct and lucid development. See *Methods and Formulas* for the estimation equation.

▷ Example 4

We are studying unionization of women in the United States by using the union dataset; see [XT] xt. We fit the fixed-effects logit model:

```
. clear all
. set memory 4000
(4000k)
. use http://www.stata-press.com/data/r10/union
(NLS Women 14-24 in 1968)
. clogit union age grade not_smsa south black, group(idcode)
note: multiple positive outcomes within groups encountered.
note: 2744 groups (14165 obs) dropped due to all positive or
all negative outcomes.
note: black omitted due to no within-group variance.
Iteration 0: log likelihood = -4521.3385
Iteration 1: log likelihood = -4516.1404
Iteration 2: log likelihood = -4516.1385
Iteration 3: log likelihood = -4516.1385
Conditional (fixed-effects) logistic regression   Number of obs   =   12035
                                                    LR chi2(4)      =    68.09
                                                    Prob > chi2     =    0.0000
                                                    Pseudo R2      =    0.0075
Log likelihood = -4516.1385
```

| union    | Coef.    | Std. Err. | z     | P> z  | [95% Conf. Interval] |
|----------|----------|-----------|-------|-------|----------------------|
| age      | .0170301 | .004146   | 4.11  | 0.000 | .0089042 .0251561    |
| grade    | .0853572 | .0418781  | 2.04  | 0.042 | .0032777 .1674368    |
| not_smsa | .0083678 | .1127963  | 0.07  | 0.941 | -.2127088 .2294445   |
| south    | -.748023 | .1251752  | -5.98 | 0.000 | -.9933619 -.5026842  |

We received three messages at the top of the output. The first one, “multiple positive outcomes within groups encountered”, we expected. Our data do indeed have multiple positive outcomes (`union = 1`) in many groups. (Here a group consists of all the observations for a particular individual.)

The second message tells us that 2744 groups were “dropped” by `clogit`. When either `union = 0` or `union = 1` for all observations for an individual, this individual’s contribution to the log-likelihood is zero. Although these are perfectly valid observations in every sense, they have no effect on the estimation, so they are not included in the total “Number of obs”. Hence, the reported “Number of obs” gives the effective sample size of the estimation. Here it is 12,035 observations—only 46% of the total 26,200.

We can easily check that there are indeed 2,744 groups with `union` either all 0 or all 1. We will generate a variable that contains the fraction of observations for each individual who has `union = 1`.

(Continued on next page)

```
. by idcode, sort: generate fraction = sum(union)/sum(union < .)
. by idcode: replace fraction = . if _n < _N
(21766 real changes made, 21766 to missing)
. tabulate fraction
```

| fraction         | Freq. | Percent | Cum.   |
|------------------|-------|---------|--------|
| 0                | 2,481 | 55.95   | 55.95  |
| .0833333         | 30    | 0.68    | 56.63  |
| .0909091         | 33    | 0.74    | 57.37  |
| .1               | 53    | 1.20    | 58.57  |
| (output omitted) |       |         |        |
| .9               | 10    | 0.23    | 93.59  |
| .9090909         | 11    | 0.25    | 93.84  |
| .9166667         | 10    | 0.23    | 94.07  |
| 1                | 263   | 5.93    | 100.00 |
| Total            | 4,434 | 100.00  |        |

Since  $2481 + 263 = 2744$ , we confirm what `clogit` did.

The third warning message from `clogit` said “black omitted due to no within-group variance”. Obviously, race stays constant for an individual across time. Any such variables are collinear with the  $\alpha_i$  (i.e., the fixed effects), and just as the  $\alpha_i$  drop out of the conditional likelihood, so do all variables that are unchanging within groups. Thus they cannot be estimated with the conditional fixed-effects model.

There are several other estimators implemented in Stata that we could use with these data:

```
cloglog ... , vce(cluster idcode)
logit ... , vce(cluster idcode)
probit ... , vce(cluster idcode)
scobit ... , vce(cluster idcode)
xtcloglog ... , i(idcode)
xtgee ... , i(idcode) family(binomial) link(logit) corr(exchangeable)
xtlogit ... , i(idcode)
xtprobit ... , i(idcode)
```

See [R] **cloglog**, [R] **logit**, [R] **probit**, [R] **scobit**, [XT] **xtcloglog**, [XT] **xtgee**, [XT] **xtlogit**, and [XT] **xtprobit** for details.

◀

## Saved Results

clogit saves the following in `e()`:

### Scalars

|                              |   |                           |                                     |
|------------------------------|---|---------------------------|-------------------------------------|
| <code>e(N)</code>            | number of observations  | <code>e(df_m)</code>      | model degrees of freedom            |
| <code>e(N_drop)</code>       | number of observations dropped<br>due to all positive or negative<br>outcomes | <code>e(r2_p)</code>      | pseudo- <i>R</i> -squared           |
| <code>e(N_group_drop)</code> | number of groups dropped<br>due to all positive or negative<br>outcomes       | <code>e(ll)</code>        | log likelihood                      |
| <code>e(k)</code>            | number of parameters  | <code>e(ll_0)</code>      | log likelihood, constant-only model |
| <code>e(k_eq)</code>         | number of equations in <code>e(b)</code>                                      | <code>e(N_clust)</code>   | number of clusters                  |
| <code>e(k_eq_model)</code>   | number of equations in model<br>Wald test                                     | <code>e(chi2)</code>      | $\chi^2$                            |
| <code>e(k_dv)</code>         | number of dependent variables   | <code>e(p)</code>         | significance                        |
|                              |   | <code>e(rank)</code>      | rank of <code>e(V)</code>           |
|                              |   | <code>e(ic)</code>        | number of iterations                |
|                              |   | <code>e(rc)</code>        | return code                         |
|                              |   | <code>e(converged)</code> | 1 if converged, 0 otherwise         |

### Macros

|                          |                                       |                            |   |
|--------------------------|---------------------------------------|----------------------------|---|
| <code>e(cmd)</code>      | <code>clogit</code>                   | <code>e(vce)</code>        | <i>vce</i> type specified in <code>vce()</code> |
| <code>e(cmdline)</code>  | command as typed                      | <code>e(vcetype)</code>    | title used to label Std. Err.                   |
| <code>e(depvar)</code>   | name of dependent variable            | <code>e(opt)</code>        | type of optimization                            |
| <code>e(group)</code>    | name of <code>group()</code> variable | <code>e(ml_method)</code>  | type of ml method                               |
| <code>e(wtype)</code>    | weight type                           | <code>e(user)</code>       | name of likelihood-evaluator program            |
| <code>e(wexp)</code>     | weight expression                     | <code>e(technique)</code>  | maximization technique                          |
| <code>e(title)</code>    | title in estimation output            | <code>e(crittype)</code>   | optimization criterion                          |
| <code>e(clustvar)</code> | name of cluster variable              | <code>e(properties)</code> | <code>b V</code>                                |
| <code>e(offset)</code>   | offset                                | <code>e(predict)</code>    | program used to implement <code>predict</code>  |
| <code>e(chi2type)</code> | LR: type of model $\chi^2$ test       |                            |   |

### Matrices

|                          |                                     |                   |   |
|--------------------------|-------------------------------------|-------------------|---|
| <code>e(b)</code>        | coefficient vector                  | <code>e(V)</code> | variance-covariance matrix of the<br>estimators |
| <code>e(ilog)</code>     | iteration log (up to 20 iterations) |                   |   |
| <code>e(gradient)</code> | gradient vector                     |                   |   |

### Functions

|                        |                         |
|------------------------|-------------------------|
| <code>e(sample)</code> | marks estimation sample |
|------------------------|-------------------------|

## Methods and Formulas

clogit is implemented as an ado-file.

Breslow and Day (1980, 247–279), Collett (2003, 251–267), and Hosmer and Lemeshow (2000, 223–259) provide a biostatistical point of view on conditional logistic regression. Hamerle and Ronning (1995) give a succinct and lucid review of fixed-effects logit; Chamberlain (1980) is a standard reference for this model. Greene (2003, chap. 21) provides a straightforward textbook description of conditional logistic regression from an economist’s point of view, as well as a brief description of choice models.

Let  $i = 1, 2, \dots, n$  denote the groups and let  $t = 1, 2, \dots, T_i$  denote the observations for the  $i$ th group. Let  $y_{it}$  be the dependent variable taking on values 0 or 1. Let  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})$  be the outcomes for the  $i$ th group as a whole. Let  $\mathbf{x}_{it}$  be a row vector of covariates. Let

$$k_{1i} = \sum_{t=1}^{T_i} y_{it}$$

be the observed number of ones for the dependent variable in the  $i$ th group. Biostatisticians would say that there are  $k_{1i}$  cases matched to  $k_{2i} = T_i - k_{1i}$  controls in the  $i$ th group.

We consider the probability of a possible value of  $\mathbf{y}_i$  conditional on  $\sum_{t=1}^{T_i} y_{it} = k_{1i}$  (Hamerle and Ronning 1995, equation 8.33; Hosmer and Lemeshow 2000, equation 7.4),

$$\Pr(\mathbf{y}_i \mid \sum_{t=1}^{T_i} y_{it} = k_{1i}) = \frac{\exp(\sum_{t=1}^{T_i} y_{it} \mathbf{x}_{it} \boldsymbol{\beta})}{\sum_{\mathbf{d}_i \in S_i} \exp(\sum_{t=1}^{T_i} d_{it} \mathbf{x}_{it} \boldsymbol{\beta})}$$

where  $d_{it}$  is equal to 0 or 1 with  $\sum_{t=1}^{T_i} d_{it} = k_{1i}$ , and  $S_i$  is the set of all possible combinations of  $k_{1i}$  ones and  $k_{2i}$  zeros. Clearly, there are  $\binom{T_i}{k_{1i}}$  such combinations, but we need not count all these combinations to compute the denominator of the above equation. It can be computed recursively.

Denote the denominator by

$$f_i(T_i, k_{1i}) = \sum_{\mathbf{d}_i \in S_i} \exp\left(\sum_{t=1}^{T_i} d_{it} \mathbf{x}_{it} \boldsymbol{\beta}\right)$$

Consider, computationally, how  $f_i$  changes as we go from a total of 1 observation in the group to 2 observations to 3, etc. Doing this, we derive the recursive formula

$$f_i(T, k) = f_i(T - 1, k) + f_i(T - 1, k - 1) \exp(\mathbf{x}_{iT} \boldsymbol{\beta})$$

where we define  $f_i(T, k) = 0$  if  $T < k$  and  $f_i(T, 0) = 1$ .

The conditional log-likelihood is

$$\ln L = \sum_{i=1}^n \left\{ \sum_{t=1}^{T_i} y_{it} \mathbf{x}_{it} \boldsymbol{\beta} - \log f_i(T_i, k_{1i}) \right\}$$

The derivatives of the conditional log-likelihood can also be computed recursively by taking derivatives of the recursive formula for  $f_i$ .

Computation time is roughly proportional to

$$p^2 \sum_{i=1}^n T_i \min(k_{1i}, k_{2i})$$

where  $p$  is the number of independent variables in the model. If  $\min(k_{1i}, k_{2i})$  is small, computation time is not an issue. But if it is large—say, 100 or more—patience may be required.

If  $T_i$  is large for all groups, the bias of the unconditional fixed-effects estimator is not a concern, and we can confidently use `logit` with an indicator variable for each group (provided, of course, that the number of groups does not exceed `matsize`; see [R] `matsize`).

## References

- Andersen, E. B. 1970. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B* 32: 283–301.
- Breslow, N. E., and N. E. Day. 1980. *Statistical Methods in Cancer Research: The Analysis of Case–Control Studies, Vol. 1*. Lyon: IARC.
- Chamberlain, G. 1980. Analysis of covariance with qualitative data. *Review of Economic Studies* 47: 225–238.
- Collett, D. 2003. *Modelling Binary Data*. 2nd ed. London: Chapman & Hall/CRC.
- Greene, W. H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall.
- Hamerle, A., and G. Ronning. 1995. Panel analysis for qualitative variables. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, ed. G. Arminger, C. C. Clogg, and M. E. Sobel, 401–451. New York: Plenum.
- Hosmer, D. W., Jr., and S. Lemeshow. 2000. *Applied Logistic Regression*. 2nd ed. New York: Wiley.
- Kleinbaum, D. G., and M. Klein. 2002. *Logistic Regression: A Self-Learning Text*. 2nd ed. New York: Springer.
- Long, J. S., and J. Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. 2nd ed. College Station, TX: Stata Press.
- McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka, 105–142. New York: Academic Press.

## Also See

- [R] **clogit postestimation** — Postestimation tools for clogit
- [R] **asclogit** — Alternative-specific conditional logit (McFadden’s choice) model
- [R] **logistic** — Logistic regression, reporting odds ratios
- [R] **mlogit** — Multinomial (polytomous) logistic regression
- [R] **nlogit** — Nested logit regression
- [R] **ologit** — Ordered logistic regression
- [R] **scobit** — Skewed logistic regression
- [SVY] **svy estimation** — Estimation commands for survey data
- [XT] **xtgee** — Fit population-averaged panel-data models by using GEE
- [XT] **xtlogit** — Fixed-effects, random-effects, and population-averaged logit models
- [U] **20 Estimation and postestimation commands**