STATA TECHNICAL BULLETIN

January 2000

STB-53

A publication to promote communication among Stata users

Editor

H. Joseph Newton
Department of Statistics
Texas A & M University
College Station, Texas 77843
409-845-3142
409-845-3144 FAX
stb@stata.com EMAIL

Associate Editors

Nicholas J. Cox, University of Durham
Francis X. Diebold, University of Pennsylvania
Joanne M. Garrett, University of North Carolina
Marcello Pagano, Harvard School of Public Health
J. Patrick Royston, Imperial College School of Medicine

Subscriptions are available from Stata Corporation, email stata@stata.com, telephone 979-696-4600 or 800-STATAPC, fax 979-696-4601. Current subscription prices are posted at www.stata.com/bookstore/stb.html.

Previous Issues are available individually from StataCorp. See www.stata.com/bookstore/stbj.html for details.

Submissions to the STB, including submissions to the supporting files (programs, datasets, and help files), are on a nonexclusive, free-use basis. In particular, the author grants to StataCorp the nonexclusive right to copyright and distribute the material in accordance with the Copyright Statement below. The author also grants to StataCorp the right to freely use the ideas, including communication of the ideas to other parties, even if the material is never published in the STB. Submissions should be addressed to the Editor. Submission guidelines can be obtained from either the editor or StataCorp.

Copyright Statement. The Stata Technical Bulletin (STB) and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp. The contents of the supporting files (programs, datasets, and help files), may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the STB.

The insertions appearing in the STB may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the STB. Written permission must be obtained from Stata Corporation if you wish to make electronic copies of the insertions.

Users of any of the software, ideas, data, or other materials published in the STB or the supporting files understand that such use is made without warranty of any kind, either by the STB, the author, or Stata Corporation. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the STB is to promote free communication among Stata users.

The *Stata Technical Bulletin* (ISSN 1097-8879) is published six times per year by Stata Corporation. Stata is a registered trademark of Stata Corporation.

Contents of this issue

References

Cleves, M. 1999. sg120: Receiver Operating Characteristic (ROC) analysis. *Stata Technical Bulletin* 52: 19–31.

Dorfman D. D. and E. Alf. 1969. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology*. 6: 487–496.

sg124	Interpreting logistic regression in all its forms

William Gould, Stata Corporation, wgould@stata.com

Abstract: The interpretation of logistic regression in all of its forms—ordinary, conditional, ordered, and multinomial—is explained. In all cases, exponentiated coefficients can be interpreted as some form of odds ratio. Guidance is provided on the accuracy of interpreting odds ratios as risk ratios.

Keywords: logit, logistic regression, conditional logistic regression, MacFadden choice model, fixed-effects logistic regression, ordered logistic regression, multinomial logistic regression; case–control, matched case–control; odds, odds ratios, conditional odds ratios, risk ratios.

Contents:

1. Ordinary (binary-outcome) logistic regression

The logistic regression or logit model is

$$\text{odds}(y_j \neq 0) = \exp(\mathbf{x}_j\mathbf{b} + b_0)$$

The Stata commands `logit` and `logistic` both report binary-outcome (ordinary) logistic regression estimates. Exponentiated coefficients have the interpretation of odds ratios (ORs). `logit` reports coefficients and `logistic` reports the exponentiated coefficients. For instance, `logit` might report a coefficient of .5 and `logistic` would correspondingly report $\exp(.5) = 1.6487$, labeling that result an odds ratio.

`logit` will report exponentiated coefficients (specify option `or`) and `logistic` will report unexponentiated coefficients (specify option `coef`), which command you use to estimate your model makes no difference.

1.1 Odds

Let p be the probability of an event. $o = p/(1-p)$ is called the odds of the event. Either way of expressing likeliness works equally well:

Probability p	Corresponding odds o
.0	.00
.1	.11
.2	.25
.3	.42
.4	.67
.5	1.00
.6	1.50
.7	2.33
.8	4.00
.9	9.00
1.0	∞

When probabilities are small, $p/(1-p)$ approximately equals p because $1-p$ is approximately 1. For unlikely events, epidemiologists often ignore that formal definition of the odds and talk about "risk" as if $o = p$. That works reasonably well for $p < .1$:

Probability p	Corresponding odds o
.1	.11...
.01	.0101...
.001	.001001...
.0001	.00010001...

1.2 Odds ratios

The exponentiated coefficient in an ordinary logistic regression has the interpretation

$$\frac{\text{odds(if the corresponding variable is incremented by 1)}}{\text{odds(if variable not incremented)}}$$

or, equivalently,

$$\frac{\text{P(event}\,|\,x+1)\,/\,\big(1-\text{P(event}\,|\,x+1)\big)}{\text{P(event}\,|\,x)\,/\,\big(1-\text{P(event}\,|\,x)\big)}$$

For instance, consider the model

```
. logit outcome female age
```

If the exponentiated coefficient on female is 1.5, then the odds of the event are 50 percent greater when female == 1 than when female == 0.

If the exponentiated coefficient on age is .5, then the odds of the event halve as age increases by 1 and they halve at every age. If age is measured in years, the odds halve for each yearly increase in age. If age is measured in 5-year spans (variable age is true age divided by 5), then the odds halve for each 5-year increment in age.

For unlikely events, epidemiologists will sometimes speak about odds ratios as if they were relative risks (risk ratios), just as they sometimes speak about odds as if they were risks. The approximation is reasonably accurate for $p < .1$:

Table of actual Risk Ratios given
Odds Ratios and P(event)

P(event$\|x$)	.25	.50	.75	1.00	1.50	2.00	4.00
.2	.2941	.5556	.7895	1.000	1.364	1.667	2.500
.1	.2702	.5263	.7692	1.000	1.429	1.818	3.077
.01	.2519	.5025	.7519	1.000	1.493	1.980	3.883
.001	.2502	.5003	.7502	1.000	1.499	1.998	3.988
.0001	.2500	.5000	.7500	1.000	1.500	2.000	3.999

Note: Let $p = \text{P(event}|x)$ and $q = \text{P(event}|x+1)$.
The odds ratio is then $o = (q/(1-q))/(p/(1-p))$.
Thus, given p and o, the value of q can be solved for:
$q = c/(1+c)$ where $c = op/(1-p)$.
The risk ratio is then q/p.

1.3 Constancy of the odds ratios

It is a remarkable property of logistic regression that the odds ratio of an effect is constant regardless of the values of the covariates. For instance, say you estimate the following logistic regression model:

$$-13.70837 + .1685\,x_1 + .0039\,x_2$$

The effect on the odds of a 1-unit increase in x_1 is $\exp(.1685) = 1.18$, meaning the odds increase by 18 percent. Incrementing x_1 increases the odds by 18 percent regardless of the value of x_2—it does not matter whether $x_2 = 0$ or $x_2 = 1000$. For every observation in the dataset, incrementing x_1 has the same multiplicative effect on the odds.

1.4 Interpreting logit output

Do not confuse coefficients with exponentiated coefficients. Look at the headers above the coefficient table. If it says `Coef`, then coefficients are being reported:

```
Logit estimates                                   Number of obs   =         74
                                                  LR chi2(2)      =      35.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -27.175156                       Pseudo R2       =     0.3966
------------------------------------------------------------------------------
         dom |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   .1685869   .0919174      1.834   0.067       -.011568    .3487418
      weight |   .0039067   .0010116      3.862   0.000        .001924    .0058894
       _cons |  -13.70837   4.518707     -3.034   0.002      -22.56487   -4.851864
------------------------------------------------------------------------------
```

If it says `Odds Ratio`, then exponentiated coefficients are being reported:

```
Logit estimates                                   Number of obs   =         74
                                                  LR chi2(2)      =      35.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -27.175156                       Pseudo R2       =     0.3966
------------------------------------------------------------------------------
         dom | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   1.183631   .1087963      1.834   0.067       .9884987    1.417283
      weight |   1.003914   .0010156      3.862   0.000       1.001926    1.005907
------------------------------------------------------------------------------
```

When exponentiated coefficients are reported, Stata does not report the intercept (labeled `_cons` in the prior output). A coefficient for the intercept is nonetheless estimated.

After seeing the output one way, you can see it the other if you wish. Learn about `logit`'s `or` and `logistic`'s `coef` options.

1.5 Actions of predict after logit and logistic

If you use `predict` after `logit` or `logistic`, you obtain probabilities, not odds. If you want the odds, you can calculate them for yourself:

```
. predict p
. gen odds = p/(1-p)
```

1.6 Demonstration

It is well worth demonstrating that you can obtain the same odds ratios that Stata reports. Try the following experiment:

```
. use auto, clear
. logistic foreign mpg weight
. predict double p
. replace mpg = mpg + 1
. predict double q
. gen double or = (q/(1-q)) / (p/(1-p))
. summarize or
```

You will observe that `or` has the same value reported by `logistic` as the odds ratio for `mpg`, namely .8448578. You will observe that the variance of variable `or` is zero (except for roundoff error), meaning `or` is constant regardless of the value of variable `weight`.

It is also instructive to compare the odds ratio to the risk ratio:

```
. gen double rr = q/p
. summarize or rr
```

The risk ratio will be .8879504, a little larger than the odds ratio of .8448578. Moreover, the risk ratio will not be constant, varying in this dataset between .8450 and .9816. In this dataset, the average probability of the event is .2973, which you can obtain by typing `summarize foreign`.

2. Conditional logistic regression

The conditional logistic regression model is

$$\text{odds}(y_j) \quad \textit{are proportional to} \quad \exp(\mathbf{x}_j\mathbf{b})$$

and notice that this definition includes, as a special case, the definition of ordinary logistic regression,

$$\text{odds}(y_j) = \exp(\mathbf{x}_j\mathbf{b} + b_0)$$

The exponentiated conditional logistic regression coefficients have the same odds-ratio interpretation as ordinary logistic estimates.

Stata's `clogit` command estimates conditional logistic regressions. Specify option `or` to obtain the exponentiated coefficients.

Conditional logistic regression differs from ordinary logistic regression in that the data are divided into groups and, within each group, the observed probability of positive outcome is either predetermined due to the data construction (such as matched case–control) or in part determined because of unobserved differences across the groups. Thus, the likelihood of the data depends on the conditional probabilities—the probability of the observed pattern of positive and negative responses within group conditional on that number of positive outcomes being observed. Terms that have a constant within-group effect on the unconditional probabilities—such as intercepts and variables that do not vary—cancel in the formation of these conditional probabilities and so remain unestimated.

2.1 Derivation of model

Models are typically asserted and it is their properties that are derived, but conditional logistic regression really is ordinary logistic regression applied to a particular data problem.

In the conditional logistic problem, the data occur in groups:

```
group 1:
        obs. 1  outcome=1   x1 = ...  x2 = ...
        obs. 2  outcome=0   x1 = ...  x2 = ...
group 2:
        obs. 3  outcome=1   x1 = ...  x2 = ...
        obs. 4  outcome=0   x1 = ...  x2 = ...
        obs. 5  outcome=0   x1 = ...  x2 = ...
group 3:
        ...
.
.
Group k:
        ...
```

We wish to condition on the number of positive outcomes within group. That is, we seek to fit a logistic model that explains why observation 1 had a positive outcome in group 1 *conditional on* one of the observations in the group having a positive outcome.

In biostatistical applications, this arises, for example, because researchers collect data on the sick and infected (the so-called "positive" outcomes) and then match those cases with controls who are not sick and infected. Thus, the number of positive outcomes is not a random variable. Within each group, there had to be the observed number of positive outcomes because that is how the data were constructed.

Economists refer to this same model as the McFadden choice model. In this model, an individual is faced with an array of choices and must choose one.

The estimator is also known as the fixed-effects logistic regression estimator for reasons that will become more obvious shortly.

Regardless of the justification, we are seeking to fit a model that explains why observation 1 had a positive outcome in group 1, observation 3 in group 2, and so on.

We assume the unconditional probability of a positive outcome is given by the standard logistic equation,

$$\text{odds}(y_j) = \exp(\mathbf{x}_j\mathbf{b} + b_0)$$

or equivalently,

$$\text{P(positive outcome)} = \exp(\mathbf{x}_j\mathbf{b} + b_0)/(1 + \exp(\mathbf{x}_j\mathbf{b} + b_0)) \tag{1}$$

Equation (1) is not the appropriate probability for our data because it does not account for the conditioning. In the first group, for instance, we want

$$\text{P(obs. 1 positive and obs. 2 negative} \,|\, \text{one positive outcome)}$$

and that is easy enough to write down in terms of the unconditional probabilities. It is

$$\frac{\text{P(1 positive)}\,\text{P(2 negative)}}{\text{P(1 positive)}\,\text{P(2 negative)} + \text{P(1 negative)}\,\text{P(2 positive)}} \tag{2}$$

From now on, when we write P(1 positive) and P(2 negative), etc., we will mean the probability that observation 1 had a positive outcome, the probability that observation 2 had a negative outcome, and so on.

Substituting equation (1) into (2), we obtain

$$\text{P(1 positive and 2 negative} \,|\, \text{one positive outcome)} = \frac{\exp(\mathbf{x}_1\mathbf{b})}{\exp(\mathbf{x}_1\mathbf{b}) + \exp(\mathbf{x}_2\mathbf{b})} \tag{3}$$

So that is the model we seek to fit or, at least, that is the term for group 1 and there are similar terms for all the other groups. (We have ignored the possibility of multiple positive outcomes within group, but that just adds complication.)

What is important to note in comparing equations (1) with (3)—in comparing ordinary logistic regression with conditional logistic regression—is that the logistic intercept b_0 cancelled. Whatever the value of b_0, it makes no difference in terms of the conditional outcomes that were observed and so cannot be estimated. Also note that b_0 could vary by group and it would still cancel. Thus, the conditional logistic estimator is often used to estimate the fixed-effects logistic model.

Finally note that, in equation (3), any variable that is constant within group will similarly cancel from both the numerator and denominator and so its effect cannot be estimated.

For a more thorough discussion of the conditional logistic derivation and its implications, see Gould (1999).

Groups that contain all-positive or all-negative outcomes provide no information because the conditional probability of observing such groups is 1 regardless of the values of the parameters **b**. Thus, when Stata encounters such groups, it reports that so many groups were dropped "due to all positive or negative outcomes".

2.2 Interpreting clogit output

By default, `clogit` reports coefficients. Specify option `or` if you want exponentiated coefficients (odds ratios) reported. Do not confuse coefficients with exponentiated coefficients. Look at the headers above the coefficient table. If it says `Coef.`, then coefficients are being reported. If it says `Odds Ratio`, then exponentiated coefficients are being reported. In neither case is an intercept reported because the intercept remains unestimated.

2.3 Actions of predict after clogit

The default calculation by `predict` following `clogit` estimation is the conditional probability of a positive outcome given a single positive outcome within group. This is not the same probability that `predict` calculates following estimation by `logit` or `logistic`. The overall probability of a positive outcome cannot be calculated because the intercepts of the logit model remain unestimated.

2.4 Equivalency of conditional and ordinary logistic regression

Ordinary and conditional logistic regression produce the same result when there is only one group, however, conditional logistic regression still leaves the intercept unestimated.

Try the following experiment:

```
. use auto, clear
. gen grp = 1
. logit foreign mpg weight
. clogit foreign mpg weight, group(grp)
```

The coefficient reported by `logit` and `clogit` will be the same. `logit`, however, will report an intercept and `clogit` will not.

Results are similarly the same in the exponentiated coefficient (odds ratio) metric:

```
. logistic foreign mpg weight
. clogit foreign mpg weight, group(grp) or
```

3. Ordered logistic regression

In ordered logistic regression, there are multiple outcomes and we will label them 1, 2, 3, ..., k. The outcomes are ordered from weak to strong, mild to severe, etc. The model is

$$\text{odds}(\text{outcome more severe than } i) = \exp(\mathbf{x}_j\mathbf{b} + b_{0i})$$

Exponentiated ordered-logistic regression coefficients can be interpreted as odds ratios. In the ordered logistic case, it is the ratio, given a one-unit increase in the covariate, of the odds of being in a higher rather than a lower category.

Stata's `ologit` command estimates ordered-logistic regression. There is currently no option to report exponentiated coefficients; you must exponentiate the coefficients for yourself.

3.1 Odds ratios

Let there be k ordered outcomes, numbered 1, 2, 3, ..., k.

In ordered logistic regression, the exponentiated coefficients are the ratios, for a one-unit increase in the covariate, of the odds of outcome k to outcomes below k, and simultaneously for outcomes $k-1$ and above to outcomes below $k-1$, and simultaneously for outcomes $k-2$ and above to outcomes below $k-2$, and so on.

That is, you have ordered outcomes

```
1        2       3        4       5     ...
```

and, just to fix ideas, let's pretend we have exactly 5 outcomes. If you were to calculate the particular odds ratio comparing outcome 5 to all the outcomes below it,

```
 1        2       3        4       5
------------------------------     ---
             |                      |
             +----------------------+
                        |
                        or = odds ratio for a 1-unit increase in x
```

you would obtain the same value as if you calculated the odds ratio comparing outcomes 4 and 5 to all outcomes below them,

```
 1        2       3        4       5
--------------------      -----------
          |                      |
          +----------------------+
                     |
                     or = odds ratio for a 1-unit increase in x
```

which would be the same value as the odds ratio comparing outcomes 3, 4, and 5 to the outcomes below that,

```
 1        2       3        4       5
------------     --------------------
     |                      |
     +----------------------+
                |
                or = odds ratio for a 1-unit increase in x
```

and so on.

3.2 Demonstration

To demonstrate this, use the auto data and estimate the model

```
. ologit rep78 mpg weight
```

Variable `rep78` takes on 5 (ordered) outcomes, 1 = Poor through 5 = Excellent.

Try the following:

```
. use auto, clear
```

```
. keep if rep78!=.
. ologit rep78 mpg weight
. predict double(p1 p2 p3 p4 p5)
. replace mpg = mpg + 1
. predict double(q1 q2 q3 q4 q5)
. gen double o5 = (q5/(q1+q2+q3+q4))/(p5/(p1+p2+p3+p4))
. gen double o4 = ((q5+q4)/(q1+q2+q3))/((p5+p4)/(p1+p2+p3))
. gen double o3 = ((q5+q4+q3)/(q1+q2))/((p5+p4+p3)/(p1+p2))
. gen double o2 = ((q5+q4+q3+q2)/(q1))/((p5+p4+p3+p2)/(p1))
. summarize o5 o4 o3 o2
. display exp(_b[mpg])
```

To explain,

1. First, we estimate the model `ologit rep78 mpg weight`.
2. We `predict p1 p2 p3 p4 p5`, obtaining the 5 predicted probabilities, observation by observation, of being in each of the `rep78` categories. Of course, $\texttt{p1} + \texttt{p2} + \texttt{p3} + \texttt{p4} + \texttt{p5} = 1$.
3. We add 1 to `mpg` and `predict q1 q2 q3 q4 q5`, thus obtaining the predicted probabilities when mpg is incremented by 1.
4. We `gen o5 = (q5/(q1+q2+q3+q4)) / (p5/(p1+p2+p3+p4))`. The numerator, `q5/(q1+q2+q3+q4)`, are the odds of being in the group $\texttt{rep78} == 5$ given `mpg` is incremented by 1. The denominator, `p5/(p1+p2+p3+p4)`, are the same odds when `mpg` is not incremented. New variable `o5` is the odds ratio.
5. We `gen o4 = ((q5+q4)/(q1+q2+q3))/((p5+p4)/(p1+p2+p3))`. This is just like the calculation above except now we are taking ratios for outcomes $\texttt{rep78} \geq 4$.
6. We `gen o3` and `gen o2`, taking ratios of outcomes $\texttt{rep78} \geq 3$ and $\texttt{rep78} \geq 2$.
7. We `summarize` all the `o` variables. We will discover that each is a constant and that they are all equal to each other!

We will also find that they are equal to $\exp(\texttt{\_b[mpg]})$ from the `ologit` output. In this case, we will have to calculate $\exp()$ for ourselves because `ologit` does not have an odds-ratio option (although it should).

3.3 Calculating confidence intervals for the odds ratios

`ologit` does not report exponentiated coefficients. In the demonstration above, we obtain the following `ologit` output:

```
. ologit rep78 mpg weight
Iteration 0:   log likelihood = -93.692061
Iteration 1:   log likelihood = -86.794936
Iteration 2:   log likelihood = -86.513907
Iteration 3:   log likelihood = -86.513267
Iteration 4:   log likelihood = -86.513267
Ordered logit estimates                           Number of obs   =         69
                                                  LR chi2(2)      =      14.36
                                                  Prob > chi2     =     0.0008
Log likelihood = -86.513267                       Pseudo R2       =     0.0766
------------------------------------------------------------------------------
       rep78 |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   .1170693   .0712216      1.644   0.100      -.0225224    .2566611
      weight |  -.0003287   .0004849     -0.678   0.498       -.001279    .0006217
-------------+----------------------------------------------------------------
       _cut1 |  -2.300255   2.918117             (Ancillary parameters)
       _cut2 |  -.5423416    2.85169
       _cut3 |   1.794106   2.830686
       _cut4 |   3.410261   2.872927
------------------------------------------------------------------------------
```

Obtaining the odds ratio is easy enough:

For `mpg`, OR $= \exp(.1170693) = 1.124$.

For `weight`, OR $= \exp(-.0003287) = .99967$.

One way to obtain the 95% confidence intervals is to exponentiate the reported coefficient confidence intervals:

For `mpg`, OR $= \exp(.1170693) = 1.124$, and the 95% confidence interval is $[\exp(-.0225224), \exp(.2566611)] = [\,.978, 1.293\,]$.

For `weight`, OR $= \exp(-.0003287) = .99967$, and the 95% confidence interval is $[\exp(-.001279), \exp(.0006217)] = [.99872, 1.0006]$.

We could also obtain a standard error for the odds ratio using the delta rule—see Sribney and Wiggins (1999) for a description—which in this case results in $\text{SE}(\text{OR}) = \exp(b) \times \text{SE}(b)$:

For `mpg`, OR $= \exp(.1170693) = 1.124$, and the standard error is $1.124 \times .071226 = .0800$.

For `weight`, OR $= \exp(-.0003287) = .99967$, and the standard error is $.99967 \times .0004849 = .0004847$.

An easy way to obtain these results is to use `lincom` with the `or` option:

```
. lincom mpg, or
 ( 1)  mpg = 0.0
------------------------------------------------------------------------------
   rep78 | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   1.124197   .0800671      1.644   0.100       .9777293    1.292607
------------------------------------------------------------------------------

. lincom weight, or
 ( 1)  weight = 0.0
------------------------------------------------------------------------------
   rep78 | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   .9996714   .0004847     -0.678   0.498       .9987218    1.000622
------------------------------------------------------------------------------
```

3.4 Actions of predict after ologit

`predict` after `ologit` calculates probabilities, not odds. If you want the odds, you can calculate them for yourself:

```
. predict p1 p2 p3 p4 p5
. gen odds5 = p5/(p1+p2+p3+p4)
. gen odds4 = (p5+p4)/(p1+p2+p3)
. gen odds3 = (p5+p4+p3)/(p1+p2)
. gen odds2 = (p5+p4+p3+p2)/(p1)
```

3.5 Equivalency of ordered and ordinary logistic regression

When there are two outcomes, both ordered and ordinary logistic regression will produce the same results except for a confusing sign reversal. Try the following experiment:

```
. use auto, clear
. logit foreign mpg weight
. ologit foreign mpg weight
```

The coefficients on `mpg` and `weight` will be the same in both sets of output. `logit`, however, will report a coefficient for `_cons` of 13.70837 while `ologit` will report the ancillary parameter `_cut1` being -13.70837.

This difference is caused by how the parameters are used in their respective models. In `logit`, the interpretation is

$$\text{P}(\texttt{foreign}_j) = \text{P}(\mathbf{x}_j\mathbf{b} + b_0 + \textit{noise} > 0)$$

and in `ologit`, the interpretation is

$$\text{P}(\texttt{foreign}_j) = \text{P}(\mathbf{x}_j\mathbf{b} + \textit{noise} > \texttt{\_cut1})$$

Thus, `_cut1` $= -b_0$.

4. Multinomial logistic regression

In multinomial logistic regression there are k outcomes 1, 2, ..., k. One of the outcomes is arbitrarily chosen as the "base outcome" as we will choose outcome k. The model is

$$\begin{aligned}
\text{odds}(y_j = 1 \,|\, y_j = 1 \text{ or } y_j = k) &= \exp(\mathbf{x}_j\mathbf{b}_1 + c_1) \\
\text{odds}(y_j = 2 \,|\, y_j = 2 \text{ or } y_j = k) &= \exp(\mathbf{x}_j\mathbf{b}_2 + c_2) \\
&\vdots \\
\text{odds}(y_j = k-1 \,|\, y_j = k-1 \text{ or } y_j = k) &= \exp(\mathbf{x}_j\mathbf{b}_{k-1} + c_{k-1}) \\
\text{odds}(y_j = k \,|\, y_j = k \text{ or } y_j = k) &= 1
\end{aligned}$$

Note that if there are $k = 2$ outcomes, the model reduces to ordinary logit.

Exponentiated coefficients have two interpretations, that of (1) relative risk ratios (RRR) and (2) conditional odds ratios (COR). Both interpretations are relative to the base group, which in our case, we arbitrarily set to k. Had we chosen a different base group, the coefficients and their interpretations would change but the predictions of the model would not change.

Stata's `mlogit` command estimates multinomial logistic regressions. `mlogit` chooses a base group on its own—not necessarily k—but its choice may be overridden using its `basecategory()` option.

Specify option `rrr` to obtain the exponentiated coefficients, which will be labeled RRRs.

4.1 Relative Risk Ratio (RRR) interpretation

Relative risk is defined

$$r_1 = \mathrm{P}(y = 1) \,/\, \mathrm{P}(y = \text{base category})$$
$$r_2 = \mathrm{P}(y = 2) \,/\, \mathrm{P}(y = \text{base category})$$

and so on. Remember that odds are defined as

$$o_1 = \mathrm{P}(y = 1) \,/\, (1 - \mathrm{P}(y = 1))$$
$$o_2 = \mathrm{P}(y = 2) \,/\, (1 - \mathrm{P}(y = 2))$$

and so on. In the case of two outcomes, $1 - \mathrm{P}(y = 1) = \mathrm{P}(y = \text{base category})$ and odds and relative risks are equal. If there were more than two categories, however, they would differ. For instance,

Category	Probability	Odds	Rel. Risk rel. to category 3
1	.3	.429	1.5
2	.5	1.000	2.5
3 (base)	.2	.250	1.0

Exponentiated coefficients in ordinary logit are odds ratios—the ratio of the odds for a one-unit increase in x to the odds when x is unchanged:

$$\text{OR} = \frac{\mathrm{P}(y = 1 \,|\, x + 1) \,/\, (1 - \mathrm{P}(y = 1 \,|\, x + 1))}{\mathrm{P}(y = 1 \,|\, x) \,/\, (1 - \mathrm{P}(y = 1 \,|\, x))}$$

Exponentiated coefficients in multinomial logistic regression are relative risk ratios—the ratio of the relative risk for a one-unit increase in x to the relative risk when x is unchanged:

$$\text{RRR} = \frac{\mathrm{P}(y = 1 \,|\, x + 1) \,/\, \mathrm{P}(y = \text{base category} \,|\, x + 1)}{\mathrm{P}(y = 1 \,|\, x) \,/\, \mathrm{P}(y = \text{base category} \,|\, x)}$$

The RRR = OR when $\mathrm{P}(\text{base category}) = 1 - \mathrm{P}(y = 1)$, that is, when there are two outcomes. When there are more than two outcomes, ORs and RRRs are different. For instance, let's pretend

Category i	$\mathrm{P}(y = i \mid x)$	$\mathrm{P}(y = i \mid x + 1)$
1	.3	.4
2	.5	.3
3 (base)	.2	.3

Then the ORs and RRRs are

Category	OR	RRR
1	1.56	.89
2	.43	.40
3 (base)	1.71	1.00

Note how difficult RRRs can be to interpret. The probability of $y = 1$ increases and yet the RRR falls because the probability of the base category increases, too, and it increased even more.

4.2 Conditional Odds Ratio (COR) interpretation

(The author of this insert thanks Roland Perfekt of the Southern Swedish Regional Tumour Registry in Lund, Sweden, for pointing out this interpretation.)

The exponentiated coefficients from multinomial logistic regression can just as well be given the interpretation of Conditional Odds Ratios, which are defined as

$$\text{COR}_1 = \frac{\text{odds}(y = 1 \,|\, x+1 \text{ and } (y = 1 \text{ or } y = \text{base category}))}{\text{odds}(y = 1 \,|\, x \text{ and } (y = 1 \text{ or } y = \text{base category}))}$$

$$\text{COR}_2 = \frac{\text{odds}(y = 2 \,|\, x+1 \text{ and } (y = 1 \text{ or } y = \text{base category}))}{\text{odds}(y = 1 \,|\, x \text{ and } (y = 1 \text{ or } y = \text{base category}))}$$

and so on. Although we listed RRR ahead of COR, CORs are perhaps a more natural interpretation. Since CORs and RRRs are both equal to the same exponentiated coefficients, whether one uses CORs or RRRs is just a matter of taste.

In the description of RRRs, we offered the following hypothetical set of results

Category i	P$(y = i \mid x)$	P$(y = i \mid x+1)$
1	.3	.4
2	.5	.3
3 (base)	.2	.3

Category	OR	RRR
1	1.56	.89
2	.43	.40
3 (base)	1.71	1.00

The RRRs in this table could just as well be labeled CORs and, if we do that, the interpretation is easier. We previously noted that the probability of $y = 1$ increases and yet the RRR falls because the probability of the base category increases and the base increased even more.

Said in the COR way, we would merely note the odds of being in 1 versus the base (category 3) fall when x increases.

4.3 Demonstrations

Try the following:

```
. use auto, clear
. drop if rep78==. | rep78==1 | rep78==2
. gen outcome = 1 if rep78==3
. gen outcome = 2 if rep78==4
. gen outcome = 3 if rep78==5
```

So far, we have just created a 3-outcome problem (and we will ignore its ordered nature). Next, we estimate our model:

```
. mlogit outcome mpg foreign, base(3) rrr
```

We can now obtain the probabilities `p1`, `p2`, and `p3`. We will then increment `mpg` by 1 and obtain in `q1`, `q2`, and `q3` the probabilities associated with a 1-unit increase in the mpg.

```
. predict double(p1 p2 p3)
. replace mpg = mpg + 1
. predict double(q1 q2 q3)
```

Now we will obtain the RRR for `outcome` $==1$:

```
. gen double rrr = (q1/q3) / (p1/p3)
```

We will next obtain the COR for `outcome` $==1$, first obtaining the conditional probabilities:

```
. gen double p1g13 = p1/(p1+p3)
. gen double q1g13 = q1/(q1+q3)
. gen double cor = (q1g13/(1-q1g13)) / (p1g13/(1-p1g13))
```

Finally, we can compare results

```
. summarize rrr cor
```

You will obtain a mean of .8422838 for both RRR and COR, both with standard deviations 0 save for roundoff error. (The standard deviation being zero is important; that's what demonstrates that RRRs and CORs are independent of the other values of the covariates.)

`mlogit` will have reported .8422838 for the RRR in its original output.

4.4 Interpreting mlogit output

`mlogit` reports coefficients or, if you specify option `rrr`, exponentiated coefficients (RRRs or CORs). Do not confuse coefficients with exponentiated coefficients. Look at the headers above the coefficient table. If it says `Coef.`, then coefficients are being reported. If it says `RRR`, then exponentiated coefficients are being reported.

4.5 Actions of predict after mlogit

`predict` after `mlogit` calculates probabilities, not odds or relative risks:

```
. predict p1 p2 p3 p4 p5
```

If you want other values, you can calculate them from the probabilities.

4.6 Equivalency of multinomial and ordinary logistic regression

When there are two outcomes, multinomial and ordinary logistic regression produce the same results. Try the following experiment:

```
. use auto, clear
. logit foreign mpg weight
. mlogit foreign mpg weight
```

If you prefer results exponentiated, type

```
. logistic foreign mpg weight
. mlogit foreign mpg weight, rrr
```

This is perhaps one more reason to prefer the COR to RRR interpretation of exponentiated coefficients in the multinomial logistic model; it is more obvious that the CORs are ORs when there are only two outcomes.

5. References

Gould, W. 1999. Within-group collinearity in conditional logistic regression. In *Stata FAQs.* Available http://www.stata.com/support/faqs/stat/clogitcl.html. College Station, TX: Stata Corporation.

Sribney, W. and V. Wiggins. 1999. Standard errors, confidence intervals, and significance tests for ORs, HRs, IRRs, and RRRs. In *Stata FAQs.* Available http://www.stata.com/support/faqs/stat/2deltameth.html. College Station, TX: Stata Corporation.

sg125	Automatic estimation of interaction effects and their confidence intervals

Jokin de Irala-Estévez, University of Navarre, Pamplona, Spain, jdeirala@unav.es
Miguel Angel Martínez, University of Navarre, Pamplona, Spain

Interaction or effect modification refers to the biological situation where the effect of a putative causal factor under study is modified by another factor; see Hosmer and Lemeshow (1989). Effect modification is identified in multivariate analyses by testing the statistical significance of biologically sound interactions between the variables in a main-effects model. This is performed by including and evaluating the significance of second or higher order terms involving the two or more variables that are postulated to possibly modify their respective effects.

The consequence of the identification of variables as significant-effect modifiers is that the effect on the outcome of one of those variables will depend on the values taken by the other variable(s) involved in the interaction. This implies that the coefficients of the models obtained by any statistical packages cannot be directly interpreted without performing further calculations. The only model coefficients that can be directly used to estimate odds ratios are those not included in interaction terms. The remaining odds ratios and their corresponding confidence intervals have to be estimated across the different levels of the other variables of the interaction term (across different categories if the variable is qualitative or across a series of values, sometimes the minimum, mean, and maximum, if it is quantitative). The major difficulty in this process lies in the correct estimation of the variance of each of these odds ratios.