Analyzing Rare Events with Logistic Regression

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised November 18, 2024

This handout draws very heavily from Paul Allison's blog entry http://www.statisticalhorizons.com/logistic-regression-for-rare-events ; the help file for the Joseph Coveney's user-written firthlogit program; Heinz Leitgöb's **The Analysis of Rare Events** (https://methods.sagepub.com/foundations/analysis-of-rare-events); and his earlier working paper **The Problem of Rare Events in Maximum Likelihood Logistic Regression - Assessing Potential Remedies** (https://www.europeansurveyresearch.org/conf/uploads/494/678/167/PresentationLeitg_b.pdf). Also, Political Scientist Gary King has some papers on this, and also a very old Stata program called relogit (I might read the papers but would probably not use his software). See http://gking.harvard.edu/relogit.

I. The problem, as described by Allison

Here is most of Paul Allison's February 13, 2012 blog entry on Logistic Regression for Rare Events (http://www.statisticalhorizons.com/logistic-regression-for-rare-events)

Prompted by a 2001 article by King and Zeng, many researchers worry about whether they can legitimately use conventional logistic regression for data in which events are rare. Although King and Zeng accurately described the problem and proposed an appropriate solution, there are still a lot of misconceptions about this issue.

The problem is not specifically the *rarity* of events, but rather the possibility of a small number of cases on the rarer of the two outcomes. If you have a sample size of 1000 but only 20 events, you have a problem. If you have a sample size of 10,000 with 200 events, you may be OK. If your sample has 100,000 cases with 2000 events, you're golden. There's nothing wrong with the logistic *model* in such cases. The problem is that maximum likelihood estimation of the logistic model is well-known to suffer from small-sample bias. And the degree of bias is strongly dependent on the number of cases in the less frequent of the two categories. So even with a sample size of 100,000, if there are only 20 *events* in the sample, you may have substantial bias.

What's the solution? King and Zeng proposed an alternative estimation method to reduce the bias. Their method is very similar to another method, known as penalized likelihood, that is more widely available in commercial software. Also called the Firth method, after its inventor, penalized likelihood is a general approach to reducing small-sample bias in maximum likelihood estimation. In the case of logistic regression, penalized likelihood also has the attraction of producing finite, consistent estimates of regression parameters when the maximum likelihood estimates do not even exist because of complete or quasi-complete separation.

Unlike exact logistic regression (another estimation method for small samples but one that can be very computationally intensive), penalized likelihood takes almost no additional computing time compared to conventional maximum likelihood. In fact, a case could be made for *always* using penalized likelihood rather than conventional maximum likelihood for logistic regression, regardless of the sample size.

II. Some possible solutions, as discussed by Leitgöb

There was a paper by Heinz Leitgöb on rare events ("The Problem of Rare Events in Maximum Likelihood Logistic Regression - Assessing Potential Remedies") at the 2013 European Survey Research Association Meetings. See the last paper in the session at

http://www.europeansurveyresearch.org/conference/programme?sess=68&day=4

or else go directly to the paper at

http://www.europeansurveyresearch.org/conf/uploads/494/678/167/PresentationLeitg_b.pdf?

Leitgöb notes that in logistic regression, Maximum Likelihood Estimates are consistent but only asymptotically unbiased, i.e. estimates can be biased when there are rare events. He compares three methods for dealing with rare events.

- Exact logistic regression (Stata built-in command: `exlogistic`)
  - This only works when N is very small (< 200)
  - Works best when covariates are discrete (preferably dichotomous) and the number of covariates is very small
  - Requires a great deal of memory and hence usually won't work with bigger problems
- Bias Correction method proposed by King and Zeng (Stata command: `relogit`. Get it and papers related to it at http://gking.harvard.edu/relogit). This seems to have been very popular with political scientists but it may not be the best approach. However, see their papers for examples of the problem.
- Penalized Maximum Likelihood Estimation proposed by Firth (Stata program: Joseph Coveney's `firthlogit`, available from SSC)

Leitgöb does Monte Carlo simulations. Here (verbatim) are his conclusions (#e refers to the number of events, i.e. the number of cases where the outcome variable equals 1)

- MLEs are systematically biased away from 0 as n and #e are getting small -> underestimation of the "true" Pr ($y = 1 | \mathbf{x}$)
- In samples with n > 200 and/or in cases with "many" covariates and/or non-discrete covariates exact logistic regression will blow up working memory
- The correction method proposed by King/Zeng is somewhat overcorrecting bias in MLEs as n is getting small (<200)
- PMLEs seem unbiased, even in cases with small n and very few #e.
- Further advantages: PMLE is always converging and solves the "problem of separation" (Heinze/Schemper 2002)
- Recommendations: Try to keep n large and apply PMLE when estimating logistic regression models (with rare events data)!

Note that Leitgöb's results are consistent with Allison's belief that the `firthlogit` method is best. We will therefore examine that method further.

III. Penalized Maximum Likelihood Estimation (the Firth Method, estimated by the Joseph Coveney's `firthlogit` program.)

This is copied verbatim from the help section for `firthlogit`:

> Firth (1993) suggested a modification of the score equations in order to reduce bias seen in generalized linear models. Heinze and Schemper (2002) suggested using Firth's method to overcome the problem of "separation" in logistic regression, a condition in the data in which maximum likelihood estimates tend to infinity (become inestimable). The method allows convergence to finite estimates in cases of separation in logistic regression.
>
> … When the method is used in fitting logistic models in datasets giving rise to separation, the affected estimate is typically approaching a boundary condition. As a result, the likelihood profile is often asymmetric under these conditions; Wald tests and confidence intervals are liable to be inaccurate. In these circumstances, Heinze and coworkers recommend using likelihood ratio tests and profile likelihood confidence intervals in lieu of Wald-based statistics. Calculation of likelihood ratio test statistics with the method is done differently by Heinze and coworkers from what is conventionally done: instead of omitting the variable of interest and refitting the reduced model, the coefficient of interest is constrained to zero and left in the model in order to allow its contributing to the penalization. The test statistic is then computed as twice the difference in penalized log likelihood values of the unconstrained and constrained models by lrtest in a manner directly analogous to that of conventional likelihood ratio tests.
>
> The penalization that allows for convergence to finite estimates in conditions of separation also allows convergence to finite estimates with very sparse data. In these circumstances, the penalization tends to over-correct for bias.

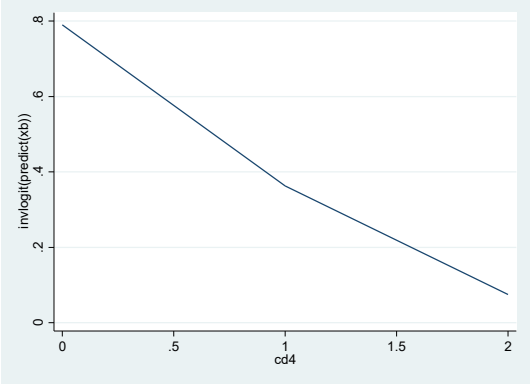
Here is an example. Note that only 14 cases are hiv-positive. Note also how the constraint command is used when estimating the constrained model; do NOT follow the usual approach of simply dropping constrained variables from the model!!!

```
. * Example: We want to contrast a full model that
. * includes both cd4 and cd8 with a constrained model
. * that only has cd8. We do NOT use the usual procedure;
. * instead we include both variables in both models, but
. * in the 2nd model we constrain the effect of cd4 = 0.
. webuse hiv1, clear
(prospective study of perinatal infection of HIV-1)
```

```
. tab1 hiv

-> tabulation of hiv

 1=positive |
       HIV; |
 0=negative |
        HIV |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         33       70.21       70.21
          1 |         14       29.79      100.00
------------+-----------------------------------
      Total |         47      100.00

. firthlogit hiv cd4 cd8

[iteration log deleted]

                                                  Number of obs   =         47
                                                  Wald chi2(2)    =       8.69
Penalized log likelihood = -18.984802             Prob > chi2     =     0.0130

------------------------------------------------------------------------------
         hiv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cd4 |  -2.213772   .7516798    -2.95   0.003    -3.687037   -.7405064
         cd8 |   1.417397   .7435227     1.91   0.057    -.0398809    2.874675
       _cons |   .4379538   .6436374     0.68   0.496    -.8235523     1.69946
------------------------------------------------------------------------------

. estimates store Full
. constraint 1 cd4 = 0
. firthlogit hiv cd4 cd8, constraint(1)

[iteration log deleted]

                                                  Number of obs   =         47
                                                  Wald chi2(1)    =       0.22
Penalized log likelihood = -25.933807             Prob > chi2     =     0.6365

 ( 1)  [xb]cd4 = 0
------------------------------------------------------------------------------
         hiv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cd4 |          0  (omitted)
         cd8 |   .2195384   .4645075     0.47   0.636    -.6908796    1.129957
       _cons |  -.9781207   .4925857    -1.99   0.047    -1.943571   -.0126705
------------------------------------------------------------------------------

. estimates store Constrained
. lrtest Full Constrained

Likelihood-ratio test                                 LR chi2(1)  =     13.90
(Assumption: Constrained nested in Full)              Prob > chi2 =    0.0002
```

Note: as of this writing, after `firthlogit` the `margins` command uses the `xb` option, not `pr`. (This may change in the future.) You can use the `expression` option to get around this, e.g.

```
. * Use margins and mcp with the equivalent of pr option
. est restore Full
(results Full are active now)

. margins , at(cd4 = (0 1 2)) expression(invlogit(predict(xb)))

Predictive margins                              Number of obs     =         47
Model VCE    : OIM

Expression   : invlogit(predict(xb))

1._at        : cd4             =           0

2._at        : cd4             =           1

3._at        : cd4             =           2

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .7898742   .1024176     7.71   0.000     .5891393     .990609
          2  |   .3628859   .0727902     4.99   0.000     .2202197    .5055522
          3  |   .0745141   .0453178     1.64   0.100    -.0143071    .1633353
------------------------------------------------------------------------------

. mcp cd4, margopts(expression(invlogit(predict(xb))))
```



Additional comments on `firthlogit`:

- Ancillary files can be downloaded with `firthlogit` that provide additional examples
- `firthlogit` also does not support the use of the `svy`: prefix. We don't know if this limitation can be overcome or not; it may be inherent in PMLE.
- `firthlogit` author Joseph Coveney and I spent some time a few years ago trying to broaden the command but it turned out not to be a very straightforward process.