Epidemiology/Biostats VHM 812/802 Course
Winter 2014, Atlantic Veterinary College, PEI
Javier Sanchez

Lecture 5b: Exact Logistic Regression

Index Page

Dataset: low birth weight – lbw.dta -

- effect of history of premature labor and smoking and birth weight.

```
  obs:            27
 vars:             3                          31 Jan 2013 12:59
 size:            81                          (_dta has notes)
------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
low             byte    %8.0g                 Low birth weight
smoke           byte    %8.0g      smoke      Smoking status during
pregnancy
ptl             byte    %8.0g      ptl        History of premature
labor
------------------------------------------------------------------------------
```

Logistic regression

- logistic regression
  - testing and inference methods based on large sample size
    - parameter estimates -> normally distributed
    - Wald tests -> follows a normal distribution
    - LRT -> follows a chi square distribution

Fisher test

- Similar to Chi2 test but more accurate for small sample size

	Smoke		
LBW	0	1	
0	19	4	23
1	2	2	4
	21	6	27

- knowing that 4 out 27 women are LBW+ and 2 out 6 are SMOKE=1, **what is the probability that 2 SMOKE=1 would be among the 4 LBW+ and 2 SMOKE=0 among the 4 LBW+?**

- exact probability->hypergeometric distribution
  - conditional probability of observing the values a, b, c, d conditionally on the observed marginals (i.e., assuming the row and column totals are given).
    - if we assume that
      - Pr(LBW+|SMOKE=1)= p
      - Pr(LBW-|SMOKE=1)= p
      - Pr(enter sample) independently of disease status
    - $$p=\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}=\frac{(a+c)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$
- Stata
  - -hypergeometricp function
    - hypergeometricp(N,K,n,k)
      - N = sample size
      - K = subjects with attribute of interest (eg SMOKE=1)
      - n = subjects with outcome (event) of interest (eg LBW+)
      - k = # of successes out of K

```
. di hypergeometricp(27,6,4,2)
.17948718
```

- p-value
  - constructs a statistical distribution
  - compute sufficient statistic (suff)
    - eg. number of possible allocations of 23 0s and 4 1s to 27 subjects with the resulting value of $obs_{suff} = \sum_{i=1}^{27} LOW_I x PTL_I$ = 2
    - possible values are 0,1,2,3 and 4
  - create distribution of "j" sufficient statistics

suff.	counts	prob. H_o true	
0	5985	0.341	pr. obs. 0 PTL+ and 4 PTL- in LBW+
1	7980	0.455	pr. obs. 1 PTL+ and 4 PTL- in LBW+
2	3150	0.179	pr. obs. 2 PTL+ and 2 PTL- in LBW+
3	420	0.024	pr. obs. 3 PTL+ and 1 PTL- in LBW+
4	15	0.001	pr. obs. 4 PTL+ and 0 PTL- in LBW+

  - compute p-values for $B_i=0$ (eg obs. = exp.)
  - sum probabilities over values of suff that are as likely or less likely than the observed value of suff.
  - observed suff. # cases with risk factor of interest = 2
  - p-value = 0.179+0.024+0.001=0.204

```
. tab low ptl, exact

           | History of premature
 Low birth |         labor
    weight |      None        One |     Total
-----------+----------------------+----------
         0 |        19          4 |        23
         1 |         2          2 |         4
-----------+----------------------+----------
     Total |        21          6 |        27

           Fisher's exact =                 0.204
   1-sided Fisher's exact =                 0.204
```

- CONCLUSION: there is no evidence that having history of pre-term delivery will increase the risk of having a low weight birth

Exact logistic regression

- follows Fisher's idea
  - estimates coefficients and confidence intervals for each parameter separately
  - conditioned on the other predictors
  - CML = conditional maximum likelihood estimates
  - computer intensive algorithms
    - simple models
    - or adjust SE in ordinary logistic regression (it might gives similar results)

Example – LWB dataset

```
. exlogistic low ptl

Enumerating sample-space combinations:
observation 1:   enumerations =          2
...
observation 27:  enumerations =          5

Exact logistic regression                       Number of obs =         27
                                                Model score   =   2.018634
                                                Pr >= score   =     0.2043
------------------------------------------------------------------------------
         low | Odds Ratio       Suff.  2*Pr(Suff.)     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ptl |   4.402267           2      0.4085      .2507705    79.01123
------------------------------------------------------------------------------
```

Logistic

```
. logistic low ptl

Logistic regression                             Number of obs   =         27
                                                LR chi2(1)      =       1.81
                                                Prob > chi2     =     0.1791
Log likelihood = -10.423421                     Pseudo R2       =     0.0797

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ptl |       4.75   5.421312     1.37   0.172     .5072157    44.48304
       _cons |   .1052632   .0782518    -3.03   0.002     .0245188    .4519108
------------------------------------------------------------------------------
```

Logistic with robust SE

```
. logistic low ptl, robust

Logistic regression                             Number of obs   =         27
                                                Wald chi2(1)    =       1.79
                                                Prob > chi2     =     0.1803
Log pseudolikelihood = -10.423421               Pseudo R2       =     0.0797

------------------------------------------------------------------------------
             |               Robust
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ptl |       4.75   5.524584     1.34   0.180      .486056    46.41955
       _cons |   .1052632   .0797424    -2.97   0.003     .0238477    .4646294
------------------------------------------------------------------------------
```

- confidence interval wider than standard logistic regression (even with robust SE)
  - uncertainty due to small sample size

- median unbiased estimates – MUE - (Stata)
  - $suff_{obs} = suff_{min}$ -> LL = - ∞
  - $suff_{obs} = suff_{max}$ -> UL = + ∞

```
. tab low smoke

           | Smoking status during
 Low birth |       pregnancy
    weight |        no        yes |     Total
-----------+----------------------+----------
         0 |        17          6 |        23
         1 |         0          4 |         4
-----------+----------------------+----------
     Total |        17         10 |        27

Exact logistic regression                       Number of obs =         27
                                                Model score   =   7.686957
                                                Pr >= score   =     0.0120
------------------------------------------------------------------------------
         low | Odds Ratio       Suff.  2*Pr(Suff.)     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   12.30305*          4      0.0239      1.361276       +Inf
------------------------------------------------------------------------------
(*) median unbiased estimates (MUE)


. logistic low smoke
note: smoke != 1 predicts failure perfectly
      smoke dropped and 17 obs not used
Logistic regression                             Number of obs   =         10
                                                LR chi2(0)      =       0.00
                                                Prob > chi2     =          .
Log likelihood = -6.7301167                     Pseudo R2       =     0.0000

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |          1  (omitted)
       _cons |   .6666667   .4303315    -0.63   0.530     .1881311    2.362419
```

Stata code

```
*5bl-Exact logistic regression
*vhm812, 2014 - Javier Sanchez

cd c:\vhm812\data
use "lbw.dta", clear
keep if age >=30
keep low ptl smoke

*create a 2by2 table with chi2
tab low ptl, chi
*2by2 with Fisher's exact
tab low ptl, exact

*binomial coefficient - factorial
di (comb(21, 2)*comb(6,2))/comb(27,4)
*hypergeometric distribution
di hypergeometricp(27,6,4,2)

*same as Fisher's exact test
exlogistic low ptl
*logistic
logistic low ptl

*logistic with robust SE
logistic low ptl, robust

* show problem with perfect prediction
*create a 2by2 table with chi2
tab low smoke, chi
exlogistic low smoke
logistic low smoke
```