Small-sample for LogisticalRegressionaland DetermineSample Size

Dian-Yan Wu,Ming-Hsuan Tsai, Yu-Ching Chang

Dept. of Applied Math National Dong Hwa University

June7, 2012

E ect of Sparse Data

- Sparse Data
  - Data collection is di cult

    MLE : in nite or not existing
  - Cell counts = 0 : sampling zeros

E ect of Sparse Data

____ :

All cell counts 6= 0) 9 nite est.

- Sampling zeros
  - marginal tables with in nite est. may have zero cell counts
  - Bias est. of odds ratio
    , ! add $\frac{1}{2}$ to cell counts
  - MLE of odds ratio = ¥
    , ! 9 CI with nite lower bound
  - Fitting processes do not converge
    , ! modify cell counts to6= 0 add a small constant e.g 10 ^{8}

Example : Clinical Trial with Sparse Data

- table 5.7

		Response(Y)		YZ marginal	
Center(Z)	Treatment(X)	Success	Failure	Success	Failure
1	Active drug	0	5	0	14
	Placebo	0	9		
2	Active drug	1	12	1	22
	Placebo	0	10		
3	Active drug	0	7	0	12
	Placebo	0	5		
4	Active drug	6	3	8	9
	Placebo	2	6		
5	Active drug	5	9	7	21
	Placebo	2	12		
XY	Active drug	12	36		

Example : Clinical Trial with Sparse Data

$$[\quad(\quad = 1)] = \quad a + b_x + b_k^z$$

- For , association
  - $\hat{b_1^z}$ and $\hat{b_3^z}$ + 26
    standard errors + 4 10^5 (non-informative)
  - odds ratios is not useful
    , ! success proportions
- For , association
  - log( ) = 1.55 $\hat{q}$ + 4.7
    standard error = 0.7
- For , association
  - can not use C-M-H test

Chi-squared Test

- G-O-F statistics $\overset{H_0:M_0}{!}\ c^2_{df}$, as large

  : (# par's $_s$) (# par's $_0$)

- Approximation depends on $\frac{n}{\#\ of\ cells}$
- When df >

G^2 statistic for small sample and sparse tables

- If $\frac{n}{\#\ of\ cells} < 5$, ^{2} ! c^2 but not good
  - tted value < 0.5 , p-value is too large
    pass G-O-F test too easy
  - $0.5 <$ tted value < 5 , p-value is too small
    fail G-O-F test too easy

Conditional Logistic Regression and Exact Inference

- Inference about logistic regression
  - approximate normal or chi-squared
  - improve as sample size %
  - what if n is small even very small
    , ! exact sampling distribution

Conditional Likelihood

- Exact inference approach

Conditional Likelihood

- Exact inference approach
  , ! conditional likelihood function
  > eliminate the other parameters

Conditional Likelihood

- Exact inference approach
  , ! conditional likelihood function
  > eliminate the other parameters
- Technique
  - conditional probability distribution ! potential sample
  -)
  - the distribution and the conditional likelihood function depend on the parameter of interest

Conditional Likelihood

- Exact inference approach
  , ! conditional likelihood function
  > eliminate the other parameters
- Technique
  - conditional probability distribution ! potential sample
  -)
  - the distribution and the conditional likelihood function depend on the parameter of interest
- MLE . conditional MLE

Conditional Likelihood

- Exact inference approach

Small-Sample Tests for Contingency Tables

- Logistic regression with a single variable:
  logit[$p(\ \ )$]=$a + b$

  Hypothesis of independent: $\ \ _0 : b = 0$
  , ! Fisher's exact test

Small-Sample Tests for Contingency Tables

- Logistic regression with a single variable:
  logit[$p(\ \ )$]=$a + b$

  Hypothesis of independent: $\ \ _0 : b = 0$
  , ! Fisher's exact test

- Logistic regression with two variables:
  logit[$p(\ \ )$]=$a + b\ \ + b_k^Z$

  To test $\ \ _0 : b = 0$

Example:Promotion Discrimination

Table 5.8 Promotion Decisions by Race and by Month

	July	August	September
Race	Yh2((932(N2((1(o((932((Y)83(h2((933(N(o((932((Y)83(h2((		

Marginal Table

Race	Yes	No	Total
Black	$0(\sum_k _{11k})$ 22	22 = $_1$	
White	10	42 $_2$	52 =
Total	10 = $_1$	64 = $_2$	74 = n

= 0 _ ($_1$ $_2$) = 0 = $_1$ ^ $_1$ = 10

5 $\sum_k _{11k}$ 5 $\sum_k _{11k}$ (10, 64, 22)

Calculate P-value

- To test $H_1 : b < 0, q < 1$
  - P-value $= P(\sum_k n_{1k} \leq 0) + 0.022$

Sample size and power for logistic regression

Why should we need to determine sample size before conducting a test?

Sample size and power for logistic regression

Why should we need to determine sample size before conducting a test?

, ! Since sample size a ects making a conclusion.

If sample size is too small, then the data might be non-informative.

therefore, focusing on sample size before data collection.

To compute sample size formula from a test

- $_0 : p_1 = p_2$. $_1 : p_1 \neq p_2$
  , ! Let $q = p_1 \quad p_2$, $_0 : q = 0$. $_1 : q \neq 0$

$$f(\ _1, \ _2) = \begin{cases} 1 & j\frac{Y_1}{n_1} \quad \frac{Y_2}{n_2} j > \\ 0 & \end{cases} \quad . \ .$$

Note: $\frac{Y_1}{n_1}$: p_1's MLE
$\frac{Y_2}{n_2}$: p_2's MLE

Using sample size to get an approximate sample size

- Ex: To test $_0 : p_1 = p_2$
  $a = 0.05 =) \quad _{\frac{a}{2}} = 1.96$
  $b = 0.1 \quad =) \quad _b = 1.28$
  $p_1 = 0.2$ and $p_2 = 0.3$

$$_1 = \quad _2 = \frac{(1.96+1.28)^2[(0.2)(0.8)+(0.3)(0.7)]}{(0.2 \quad 0.3)^2}$$

$$+ 388.4112 + 389$$

- For a logistic regression model :
  $(p) = b_0 + b_1 x$ w) x is quantitative

- Of interest is the dependence of X and Y ( ) to test $H_0 : b_1$ 0

Another sample size formula

An approximate sample size formula for a one-sided test is

$$n = \frac{[z_a + z_b \exp(\quad l^2/4)]^2(1 + 2\bar{p}\bar{d})}{\bar{p}l^2} \qquad (1)$$

where $d = [1+(1+l^2)\exp(5l^2/4$

Sample size's example

Ex:
Y:severe heart disease w) prob$=p$
X:cholesterol level for a middle aged population

Sample size in multiq 0.59776 2 0.7 rgnR 0.8ression

Sample size in multiple logistic regression

- For a multiple logistic regression model with k predictors :
  , ! logit(p) = $b_0 + b_1$ $_1 + b_2$ $_2 + \ldots + b_k$ $_k$
) Let R be twe multiple correlation between twe predictor X of interest and twe otwers in twe model
) a conservative sample size determining is $\frac{n}{1 \quad R^2}$ wwere n from (1)
) For previous example, n $\frac{612}{1 \quad 0.4^2}$ + 729) Hence, if R = $\frac{1}{2}$ + 0.7 twen n ! 2n