Lecture 19: Conditional Logistic Regression

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Conditional Logistic Regression

Purpose

1. Eliminate unwanted nuisance parameters
2. Use with sparse data

Prior to the development of the conditional likelihood, lets review the unconditional (regular) likelihood associated with the logistic regression model.

- Suppose, we can group our covariates into J unique combinations
- and as such, we can form J (2×2) tables

First, consider the j^{th} (2×2) table:

TABLE j (or stratum '$W = j$')

		Variable (Y): 1	Variable (Y): 2	
Variable (X)	1	Y_{j11}	Y_{j12}	Y_{j1+}
	2	Y_{j21}	Y_{j22}	Y_{j2+}
		Y_{j+1}	Y_{j+2}	Y_{j++}

We can write the (2×2) data table of probabilities for stratum $W = j$ as

		Y		
		1	2	
X	1	p_{j1}	$1 - p_{j1}$	1
	2	p_{j2}	$1 - p_{j2}$	1

Unconditional Likelihood

- Let p_j be the probability that a subject in stratum j has a success.
- Then the probability of observing y_j events in stratum j is

$$P(Y_j = y_j) = \left(\begin{array}{c} n_j \\ y_j \end{array} \right) p_j^{y_j} (1 - p_j)^{n_j - y_j}$$

- However, we know p_j is a function of covariates
- Without loss of generality, assume we are interested in two covariates, x_{j1} and x_{j2}, such that

$$p_j = \frac{e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}}}{1 + e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}}}$$

Then, the likelihood function can be written as

$$\begin{aligned}
L(\vec{y}) &= \prod_{j=1}^{J} \left(\begin{array}{c} n_j \\ y_j \end{array} \right) p_j^{y_j} (1-p_j)^{n_j - y_j} \\
&= \prod_{j=1}^{J} \left(\frac{e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}}}{1+e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}}} \right)^{y_j} \left(\frac{1}{1+e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}}} \right)^{(n_j - y_j)} \times \left(\begin{array}{c} n_j \\ y_j \end{array} \right) \\
&= \prod_{j=1}^{J} \frac{\left(e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}} \right)^{y_j}}{\left(1+e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}} \right)^{n_j}} \times \left(\begin{array}{c} n_j \\ y_j \end{array} \right) \\
&= \frac{e^{\beta_0 t_0 + \beta_1 t_1 + \beta_2 t_2}}{\prod_{j=1}^{J} \left(1+e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}} \right)^{n_j}} \times \prod_{j=1}^{J} \left(\begin{array}{c} n_j \\ y_j \end{array} \right)
\end{aligned}$$

where

$$t_k = \sum_{j=1}^{J} y_j x_{jk}$$

and

$$t_0 = \sum_{j=1}^{J} y_j$$

- The functions t_0, t_1, and t_2 are sufficient statistics for the data.
- Suppose we want to test $\beta_2 = 0$ using a likelihood ratio test.
- Then

$$L(\beta_0, \beta_1) = \frac{e^{\beta_0 t_0 + \beta_1 t_1}}{\prod_{j=1}^{J} \left(1 + e^{\beta_0 + \beta_1 x_{j1}}\right)^{n_j}} \times \prod_{j=1}^{J} \begin{pmatrix} n_j \\ y_j \end{pmatrix}$$

- and the LRT would equal

$$\begin{aligned}
\Lambda(\vec{y}) &= -2(lnL(\beta_0, \beta_1) - ln(L(\beta_0, \beta_1, \beta_2))) \\
&= -2ln \frac{L(\beta_0, \beta_1)}{L(\beta_0, \beta_1, \beta_2)} \\
&= -2ln \frac{\frac{e^{\beta_0 t_0 + \beta_1 t_1}}{\prod_{j=1}^{J} \left(1 + e^{\beta_0 + \beta_1 x_{j1}}\right)^{n_j}} \times \prod_{j=1}^{J} \begin{pmatrix} n_j \\ y_j \end{pmatrix}}{\frac{e^{\beta_0 t_0 + \beta_1 t_1 + \beta_2 t_2}}{\prod_{j=1}^{J} \left(1 + e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}}\right)^{n_j}} \times \prod_{j=1}^{J} \begin{pmatrix} n_j \\ y_j \end{pmatrix}}
\end{aligned}$$

- Which is distributed **approximately** chi-square with 1 df.

- When you have small sample sizes, the chi-square approximation is not valid.
- We need a method to calculate the **exact** $p-$value of $H_0 : \beta_2 = 0$ from the exact null distribution of $\Lambda(\vec{y})$.
- Note that the distribution of $\Lambda(\vec{y})$ depends on the exact distribution of $\vec{y}$.
- Thus, we need to derive the exact distribution of $\vec{y}$.

Under, H_0, $\beta_2 = 0$, so we will begin with

$$P(\vec{y}) = L(\beta_0, \beta_1) = \frac{e^{\beta_0 t_0 + \beta_1 t_1}}{\prod_{j=1}^{J} \left(1 + e^{\beta_0 + \beta_1 x_{j1}}\right)^{n_j}} \times \prod_{j=1}^{J} \left(\begin{array}{c} n_j \\ y_j \end{array} \right)$$

- We cannot use the above probability statement because it relies on the population parameters (i.e., unknown constants) β_0 and β_1.
- That is, β_0 and β_1 are a nuisance to our calculations and as such are called **nuisance parameters**
- Similar to what we did in the Fisher's Exact Test, we are going to condition out β_0 and β_1

Using Bayes' Law,

$$P(\vec{y}|t_0,t_1) = \frac{\frac{e^{\beta_0 t_0+\beta_1 t_1}}{\prod_{j=1}^{J}\left(1+e^{\beta_0+\beta_1 x_{j1}}\right)^{n_j}} \times \prod_{j=1}^{J}\begin{pmatrix} n_j \\ y_j \end{pmatrix}}{\sum_{\vec{y}\in\Gamma} \frac{e^{\beta_0 t_0+\beta_1 t_1}}{\prod_{j=1}^{J}\left(1+e^{\beta_0+\beta_1 x_{j1}}\right)^{n_j}} \times \prod_{j=1}^{J}\begin{pmatrix} n_j \\ y_j \end{pmatrix}}$$

where Γ is the set of all vectors of y such that

$$\sum_{j=1}^{J} y_j = t_0$$

$$\sum_{j=1}^{J} y_j x_{1j} = t_1$$

and $0 \leq y_j \leq n_j$ and y_j is an integer.

Then, $P(\vec{y}|t_0, t_1)$ can be simplified to

$$
\begin{aligned}
P(\vec{y}|t_0,t_1) &= \frac{\frac{e^{\beta_0 t_0+\beta_1 t_1}}{\prod_{j=1}^{J}\left(1+e^{\beta_0+\beta_1 x_{j1}}\right)^{n_j}} \times \prod_{j=1}^{J}\begin{pmatrix} n_j \\ y_j \end{pmatrix}}{\sum_{\vec{y}\in\Gamma} \frac{e^{\beta_0 t_0+\beta_1 t_1}}{\prod_{j=1}^{J}\left(1+e^{\beta_0+\beta_1 x_{j1}}\right)^{n_j}} \times \prod_{j=1}^{J}\begin{pmatrix} n_j \\ y_j \end{pmatrix}} \\
&= \frac{\frac{e^{\beta_0 t_0+\beta_1 t_1}}{\prod_{j=1}^{J}\left(1+e^{\beta_0+\beta_1 x_{j1}}\right)^{n_j}} \times \prod_{j=1}^{J}\begin{pmatrix} n_j \\ y_j \end{pmatrix}}{\frac{e^{\beta_0 t_0+\beta_1 t_1}}{\prod_{j=1}^{J}\left(1+e^{\beta_0+\beta_1 x_{j1}}\right)^{n_j}} \times \sum_{\vec{y}\in\Gamma}\prod_{j=1}^{J}\begin{pmatrix} n_j \\ y_j \end{pmatrix}} \\
&= \frac{\prod_{j=1}^{J}\begin{pmatrix} n_j \\ y_j \end{pmatrix}}{\sum_{\vec{y}\in\Gamma}\prod_{j=1}^{J}\begin{pmatrix} n_j \\ y_j \end{pmatrix}}
\end{aligned}
$$

which does not contain any unknown parameters and we can calculate the exact $p-$value.

Let λ by any value such that $\Lambda(\vec{y}) = \lambda$ and denote

$$\Gamma_\lambda = \{\vec{y} : \vec{y} \in \Gamma \text{ and } \Lambda(\vec{y}) = \lambda\}$$

Then

$$P(\Lambda(\vec{y}) = \lambda) = \sum_{\vec{y} \in \Gamma_\lambda} P(\vec{y}|t_0, t_1)$$

To obtain the exact $p-$value for testing $\beta_2 = 0$ we need to:

1. Enumerate all values of λ
2. Calculate $P(\Lambda(\vec{y}) = \lambda)$
3. $p-$value equals the sum of the as extreme or more extreme values of $P(\Lambda(\vec{y}) = \lambda)$

Or use SAS PROC LOGISTIC or LogXact

Example–Mantel (1963)

- Effectiveness of immediately injected penicillin or 1.5 hour delayed injected penicillin in protecting against $\beta-$hemolytic Streptococci:

Penicillin Level	DELAY	RESPONSE Died	Cured
1/8	None	0	6
	1.5 hr	0	8
1/4	None	3	3
	1.5 hr	1	6
1/2	None	6	0
	1.5 hr	2	4
1	None	5	1
	1.5 hr	6	0
4	None	2	0
	1.5 hr	5	0

- Note, there are many 0 cells in the table; may have problems with the large sample normal approximations.
- We want to estimate the (common) OR between Delay and Response, given strata (Penicillin).
- We can consider the data as arising from $J = 5$, (2×2) tables, where $J = 5$ penicillin levels.
- In the k^{th} row of (2×2) table j, we assume the data have the following logistic regression model:

$$\text{logit}\left(P[\text{Cured}|\text{penicillin}=j, \text{DELAY}_k]\right) = \mu + \alpha_j + \beta \text{DELAY}_k,$$

where

$$\text{DELAY}_k = \begin{cases} 0 \text{ if none} \\ 1 \text{ if 1.5 hours} \end{cases}.$$

- There are problems with the unconditional (usual) MLE, as we'll see in the computer output.

- This is equivalent to the logistic model

$$\text{logit}\{P[\text{Cured}|\text{penicillin=}j, \text{DELAY}]\} = \mu + \alpha_1 z_1 + ... + \alpha_4 z_4 + \beta \text{DELAY}_k,$$

  where

$$z_j = \begin{cases} 1 \text{ if penicillin=}j \\ 0 \text{ if otherwise} \end{cases}.$$

- Note, we have not taken the ordinal nature of penicillin into account
- We are considering penicillin (and the associated dose) as a nuisance parameter and are not interested in drawing inference about penicillin
- We are interested in learning about the timing to give **any dose** of penicillin

SAS Proc Logistic

- Using SAS Proc Logistic, we get the following results:

```
data cmh;
 input pen delay response count;
cards;
0.125 0 0 0  /* response: 1 = cured, 0 = died */
0.125 0 1 6  /* delay: 0 = none, 1 = 1.5 hrs */
0.125 1 0 0
0.125 1 1 8
0.250 0 0 3
0.250 0 1 3
0.250 1 0 1
0.250 1 1 6
0.500 0 0 6
0.500 0 1 0
0.500 1 0 2
0.500 1 1 4
1.000 0 0 5
1.000 0 1 1
1.000 1 0 6
1.000 1 1 0
4.000 0 0 2
4.000 0 1 0
4.000 1 0 5
4.000 1 1 0
;
```

```
proc logistic descending;
 class pen(PARAM=ref);
  model response = pen delay ;
 freq count;
run;

/* SELECTED OUTPUT */
                    Model Convergence Status

       Quasi-complete separation of data points detected.

WARNING: The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning.
Results shown are based on the last maximum likelihood iteration. Validity
the model fit is questionable.
```

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-13.7143	163.8	0.0070	0.9333
pen	0.125	1	25.1650	199.4	0.0159	0.8996
pen	0.25	1	13.6947	163.8	0.0070	0.9334
pen	0.5	1	11.9500	163.8	0.0053	0.9419
pen	1	1	10.0640	163.8	0.0038	0.9510
delay		1	1.8461	0.9288	3.9508	0.0468

WARNING: The validity of the model fit is questionable.

Odds Ratio Estimates

Effect			Point Estimate	95% Wald Confidence Limits	
pen	0.125	vs 4	>999.999	<0.001	>999.999
pen	0.25	vs 4	>999.999	<0.001	>999.999
pen	0.5	vs 4	>999.999	<0.001	>999.999
pen	1	vs 4	>999.999	<0.001	>999.999
delay			6.335	1.026	39.115

Possible remedies

There are (at least) two possible remedies to this problem:

1. Assign a score w_j to penicillin level j and use this with logistic regression, or
2. Since we are interested in

$$\beta = \log(OR^{\text{RESPONSE,DELAY}}),$$

an alternative is to eliminate the nuisance parameters (the effects of PENICILLIN LEVEL) by using conditional logistic regression.

Assigning Scores

Revised data step:

```
data cmh;
/* delay = 0 if None,  1 if 1.5 hr */
/* response = 1 if Cured,  0 if Died */
 input pen delay response count;
 if pen = 0.125 then pen1 = 1; else pen1=0;
 if pen = 0.250 then pen2 = 1; else pen2=0;
 if pen = 0.500 then pen3 = 1; else pen3=0;
 if pen = 1.000 then pen4 = 1; else pen4=0;
 pen_cont = pen;
 cards;
 ... (same as before)

proc logistic descending;
 model response = pen_cont delay /aggregate scale=1;
 freq count;
run;
```

Selected Results

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.2328	0.8051	7.6910	0.0055
pen_cont	1	-6.8591	1.9781	12.0238	0.0005
delay	1	1.6984	0.8623	3.8792	0.0489

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
pen_cont	0.001	<0.001	0.051
delay	5.465	1.008	29.620

Note: This model converged

Approach 2 - Elimination of Nuisance Parameters

- For the prospective study we have, the rows (y_{j1+} and y_{j2+}) of each (2×2) table are fixed, and we have two independent binomials:
  1) Row 1:

$$Y_{j11} \sim Bin(y_{j1+}, p_{j1}),$$

  where

$$p_{j1} = P[\text{Cured}|\text{penicillin=}j, \text{DELAY} = 1]$$

  and
  2) Row 2:

$$Y_{j21} \sim Bin(y_{j2+}, p_{j2}),$$

  where

$$p_{j2} = P[\text{Cured}|\text{penicillin=}j, \text{DELAY} = 0]$$

- In terms of the logistic model,

$$\text{logit}\left(P[\text{Cured}|\text{penicillin=}j, \text{DELAY} = 1]\right) = \text{logit}(p_{j1}) = \mu + \alpha_j + \beta,$$

and

$$\text{logit}\left(P[\text{Cured}|\text{penicillin=}j, \text{DELAY} = 0]\right) = \text{logit}(p_{j2}) = \mu + \alpha_j$$

- Then the log-odds ratio for the j^{th}, (2×2) table is

$$\begin{aligned} \log \frac{p_{j1}/(1-p_{j1})}{p_{j2}/(1-p_{j2})} &= \text{logit}(p_{j1}) - \text{logit}(p_{j2}) \\ &= [\mu + \alpha_j + \beta] - [\mu + \alpha_j] \\ &= \beta, \end{aligned}$$

and, the odds ratio is

$$(OR^{\text{RESPONSE,DELAY}}) = \frac{p_{j1}/(1-p_{j1})}{p_{j2}/(1-p_{j2})} = e^{\beta}$$

Conditional Logistic Regression

- The conditional likelihood is

$$L^c(\vec{y}) = \frac{\prod_{j=1}^{J} \binom{n_j}{y_j}}{\sum_{\vec{y}\in\Gamma} \prod_{j=1}^{J} \binom{n_j}{y_j}}$$

  which is the product of the probability functions over the J tables or strata.

- We can find the conditional MLE, $\hat{\boldsymbol{\beta}}_{CMLE}$ by maximizing $L^c(\beta)$.

- As with a single table, in 'large' samples, $\hat{\boldsymbol{\beta}}$ will be approximately unbiased, approximately normal, and have variance equal to the negative inverse of the (expected) second derivative.

What we mean by LARGE SAMPLES

For proper properties of the conditional likelihood $L^c(\beta)$, we need 'large' samples.

There are two kinds of 'large samples' that allow $\hat{\boldsymbol{\beta}}_{CMLE}$ to be approximately normal (by the central limit theorem).

1. We can have the within stata sample size (y_{j++}) large (with the number of strata J being small).
2. As the number of strata, J, becomes large (we can $y_{j++} < \infty$, and, actually, we can have y_{j++} as small as 2, which we will discuss later).

As a result, we can have both (y_{j++}) large and J large, and this would be a large sample as well.

- This is in contrast to the usual (unconditional MLE); if you look at the logistic regression model

$$\text{logit}\{P[\text{Died}|\text{penici=}j, \text{DELAY}_k]\} =$$

$$\mu + \alpha_j + \beta\text{DELAY}_k,$$

  in order to be able to estimate each α_j, we need each y_{j++} to be large

- In particular, if J is large, using unconditional logistic regression for

$$\text{logit}(p_{jk}) = \mu + \alpha_j + \beta x_k,$$

  we will have a large number of parameters (α_j's) to estimate, and we cannot unless each y_{j++} is also large.

- If y_{j++} is also large, then

$$\widehat{\beta}_{CMLE} \approx \widehat{\beta}_{MLE},$$

  but the MLE is preferable since it is simpler computationally, and has nice properties.

Mantel-Haenzsel Estimate

- One other possibility is the Mantel Haenszel estimate of a common odds ratio for a set of J, (2×2) tables,
- As with the conditional MLE, the Mantel-Haenszel estimator of the common odds ratio is (asymptotically unbiased) if either (y_{j++}) is large or J is large or both.
- However, the Mantel-Haenzel estimate does not generalize to data with many covariates, whereas the conditional likelihood does.

CMH Using PROC FREQ

However, for the data studied today, we can calculate the CMH (or just the Mantel-Haenzsel)

```
proc freq;
 tables pen*delay*response/cmh;
 weight count;
run;
```

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control	Mantel-Haenszel	4.6316	0.9178	23.3731
(Odds Ratio)	Logit **	3.9175	0.6733	22.7925

** These logit estimators use a correction of 0.5 in every cell of those tables that contain a zero. Tables with a zero row or a zero column are not included in computing the logit estimators.

Conditional Logistic Regression

- We can use SAS Proc Logistic as before to do conditional logistic regression.
- SAS Proc Logistic will give us the conditional logistic regression estimate of the odds ratio, and an exact 95% confidence interval for the odds ratio using the conditional likelihood.

SAS Proc Logistic

```
proc logistic descending;
 class pen(PARAM=ref);
  model response = pen delay ;
  exact delay / estimate = both /*both = logor & or */;
 freq count;
run;

/* SOME OUTPUT */
             Conditional Exact Tests

                                        --- p-Value ---
Effect   Test           Statistic    Exact      Mid

delay    Score             4.0880   0.0587   0.0384
         Probability       0.0407   0.0587   0.0384
```

Exact Parameter Estimates

Parameter	Estimate	95% Confidence Limits		p-Value
delay	1.6903	-0.2120	4.1588	0.0937

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		p-Value
delay	5.421	0.809	63.995	0.0937

Effectiveness of immediately injected penicillin or 1.5 hour

OR of Delayed injected penicillin in protecting against Streptococci

Variable	ODDS Ratio	95% Confidence Limits Lower	Upper
COND. LOGIT	5.421	0.809	63.995
MANTEL-HA	3.918	0.673	22.793
LOGISTIC*	6.335	1.026	39.115
LOGISTIC**	5.465	1.008	29.620

* From Proc Logistic with warning:
 ‘The validity of the model fit is questionable.’
  Uses dose as a class
**Using dose with scores

- From each, we see that odds of being cured if there is no delay is about 5 times of what it is if there is a delay.
- Although we get an estimate from (ordinarly) logistic regression, it is not as stable as the other methods.
- Note, these strata sizes were small, but there are situations when they are even smaller (matched pairs, discussed later).
- First though, let's look at a couple of other models.
- In particular, let's test for interaction.

Penicillin Data

- If we fit the model with interaction between delay and penicillin, (with 4 dummy variables for penicillin and the 4 for the interaction), and the unconditional had the same problems as above. The conditional also has problems:
- Proc Logistic gave me the following:

```
proc logistic descending;
 class pen(PARAM=ref);
  model response = pen delay pen*delay ;
  exact pen*delay /estimate;
 freq count;
run;
```

SOME OUTPUT

Exact Conditional Analysis

Conditional Exact Tests

Effect	Test	Statistic	--- p-Value --- Exact	Mid
delay*pen	Score	7.0666	0.1354	0.0990
	Probability	0.0729	0.1354	0.0990

Exact Parameter Estimates

Parameter	Estimate	95% Confidence Limits		p-Value
delay*pen 0.125	.#	.	.	.
delay*pen 0.25	.#	.	.	.
delay*pen 0.5	.#	.	.	.
delay*pen 1	.#	.	.	.

NOTE: # indicates that the conditional distribution is degenerate.

- Proc Logistic can give us an exact conditional Score test (with $p-$value = .1354), even though it cannot give us estimates of the interaction terms.
- Thus, I fit a little simpler model, with the main effects of penicillin fitted using dummy variables, but the interaction with penicillin continuous (w_j):

$$\text{logit}\left(P[\text{Cured}|\text{penici}=j, \text{DELAY}=k]\right) = \mu + \alpha_j + \beta_1 x_k + \beta_{12} x_k w_j,$$

where

$$x_k = \begin{cases} 0 \text{ if DELAY = none } (k=1) \\ 1 \text{ if DELAY = 1.5 hours } (k=2) \end{cases}.$$

and w_j equals the actual dose (0.125, .25, .5, or 1), treated as continuous.

- We are interested in testing for homogeneity of the odds ratio across the J stratum.

```
proc logistic descending;
 class pen(PARAM=ref);
  model response = pen delay pen_cont*delay ;
  exact pen_cont*delay /estimate;
 freq count;
run;
```

Exact Parameter Estimates

Parameter	Estimate	95% Confidence Limits		p-Value
delay*pen_cont	-4.3902	-15.4187	2.8438	0.2708

- When W (pennicilin) is treated as continuous,

$$\text{logit}\left(P[\text{Cured}|\text{penici}=j, \text{DELAY}=k]\right) = \mu + \gamma_1 w_j + \beta_1 x_k + \beta_{12} x_k w_j$$

the unconditional logistic regression also converged, and we get the following comparison

95% Confidence Limits

PEN*DELAY		
METHOD	Parameter Estimate	P Value
COND	-4.3902	0.2708
UNCOND	-3.4990	0.4288

- Although both algorithms converged, the estimates are pretty large, and the confidence intervals are very wide (not shown).
- However, we have indication that the common odds ratio assumption is valid.

LogXact

- All of the models could have been fit in LogXact, a software package specializing in conditional (or exact) logistic regression
- I noted small variations in the parameter estimates between SAS and LogXact.
- LogXact is mouse driven and is easy to use
- SAS is easier to do model building activities – creation of derived variables, “selection =” routines and “copy and paste” model code