Lecture 24: Log-linear Models -Sparse Tables

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Sparse Tables

- Consider an arbitrary contingency table
- We could have a table that cross classifies students in BMTRY 711 for an arbitrary year on
  1. Race: White, black, and other
  2. Gender: Male and female
  3. Year: 1^{st}, 2^{nd}, and other
- In theory, all combinations are possible
- But in practice some combinations are unobserved
- What do the “zero” cells say about the relationship of race, gender and year?

Sampling Zeros

- In our example (or current class),
- We have no White Males of any year
- However, “in theory” we would expect some in a given year
- That is, $P(\text{white male in } 1^{st} \text{ year}) > 0$ or

$$\mu_{\text{white male in } 1^{st} \text{ year}} > 0$$

- When we would expect some observations in the ikj cell but fail to sample any, we have **sampling zeros**

Structural Zeros

- In some cases,

$$\mu_{ijk} = 0$$

- or the probability of observing a specific combination is zero
- When $\mu_* = 0$ for a specific cell in the table, we have a **structural zero**
- For example, in an oncology study that enrolls a cohort of individuals, you would expect lung cancer in males and females; however, prostate cancer can biologically occur only in males.
- Thus, a marginal table summed across all other factors could yield a similar table

	Cancer Type			
	Lung	Prostate	Ovarian	Other
Male	$\mu > 0$	$\mu > 0$	n/a	$\mu > 0$
Female	$\mu > 0$	n/a	$\mu > 0$	$\mu > 0$

- “n/a” is used here to distinguish a sampling zero from a structural zero

A lot about nothing

- With sampling zeros, a larger (or different) sample may allow for observed values where the present structural zeros may exists
- Note that **0** is a valid Poisson response with probability $\exp(\mu_*)$
- As such, it contributes to the likelihood function
- However, no matter how large the sample, structural zeros will always remain
- Thus, we have constraints that $\mu_* = \widehat{\mu_*} = n_* = 0$
- Contingency tables with structural tables are called **incomplete tables**
- We need to take the constraints into account when estimating the model parameters

Sparse Tables and effects on G^2

- A small sample size (and hence a sparse table) affect the asymptotic convergence of chi-square based tests
- If the total sample size (n) divided by the total number of cells (N) is less than 5 ($n/N < 5$), then the chi-square approximation of G^2 is generally poor
- Pearson's X^2 may perform better, but a guideline generally isn't accepted
- In the context of log-linear models, the delta-deviance (or delta-G^2) with small degrees of freedom generally are better approximated by a chi-square distribution

Solutions to sampling zeros

- In the 2×2 table we discussed at the start of the semester add .5 to all of the cells
- It was discussed (but not proven) that this actually produces less bias than the “unadjusted” odds ratio
- For a generalized table, this approach may not always be your best bet
- For example, in a table with $N = 30$ cells, adding 1/2 to each of these cells may in fact add too much data to the table
- One approach that is generally recommended is to perform **sensitivity analyses** to check the robustness of the results

Example

- Consider a multicenter clinical trial in with subjects are randomized to either Active drug or placebo for the treatment of fungal infections.
- A binary response of a “success” “or failure” is recorded for each subject

		Response	
Center	Tx	Success	Failure
1	A	0	5
	P	0	9
2	A	1	12
	P	0	10
3	A	0	7
	P	0	5
4	A	6	3
	P	2	6
5	A	5	9
	P	2	12

- Note all of the zeros.

SAS

```
options nocenter;
data one;
 input center tx $ success fail;
count = success;
outcome = 1;
output;
count=fail;
outcome = 2;
output;
drop success fail;
 cards;
1   A   0   5
1   P   0   9
2   A   1   12
2   P   0   10
3   A   0   7
3   P   0   5
4   A   6   3
4   P   2   6
5   A   5   9
5   P   2   12
;
run;
```

Proc Logistic

```
proc logistic data=one;
freq count;
 class center tx /param=ref;
model outcome(ref=’2’) = center tx;
run;

From LOG:
NOTE: PROC LOGISTIC is modeling the probability that outcome=1.
WARNING: There is possibly a quasi-complete separation of
         data points. 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 of the model fit is questionable.
```

```
              Odds Ratio Estimates

                    Point          95% Wald
Effect           Estimate      Confidence Limits

center 1 vs 5      <0.001      <0.001    >999.999
center 2 vs 5       0.113       0.012       1.041
center 3 vs 5      <0.001      <0.001    >999.999
center 4 vs 5       2.895       0.733      11.442
tx     A vs P       4.693       1.186      18.564
```

Proc GENMOD

```
proc genmod data=one;
 class center tx outcome;
 model count = center|tx|outcome@2 /dist=poi link=log;
run;

FROM LOG:
WARNING: The negative of the Hessian is not positive definite.
         The convergence is questionable.
WARNING: The procedure is continuing but the validity of the model
         fit is questionable.
WARNING: The specified model did not converge.
WARNING: Negative of Hessian not positive definite.
```

```
Parameter                DF   Estimate      Error          Limits
***** Something fishy?*********
center*outcome   1   1    1   -25.4133   159175.4    -312004   311952.7
center*outcome   1   2    0     0.0000     0.0000     0.0000     0.0000
center*outcome   2   1    1    -2.1802     1.1327    -4.4003     0.0399
center*outcome   2   2    0     0.0000     0.0000     0.0000     0.0000
center*outcome   3   1    1   -25.3866   145462.6    -285127   285076.1
```

Deleting sites

- Clinics 1 and 3 only have failures
- Thus, they do not affect the OR of treatment by success
- One solution to the convergence problems is to model the data without sites 1 and 3

```
proc logistic data=one;
 where center in (2,4,5);
 freq count;
 class center tx /param=ref;
 model outcome(ref=’2’) = center tx;
run;
```

Selected results

```
                   Model Convergence Status

         Convergence criterion (GCONV=1E-8) satisfied.


              Odds Ratio Estimates

                    Point            95% Wald
Effect           Estimate      Confidence Limits

center 2 vs 5       0.113       0.012       1.041
center 4 vs 5       2.895       0.733      11.442
tx     A vs P       4.693       1.186      18.564
```

Note: This OR for the treatment is the same for the model with 5 sites

CMH Estimator

```
proc freq data=one;
 tables center*tx*outcome / cmh;
 weight count;
run;
proc freq data=one;
 where center in (2,4,5);
tables center*tx*outcome / cmh;
 weight count;
run;
```

Selected Results

```
ALL SITES**************
            Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                        Value     95% Confidence Limits
---------------------------------------------------------------------------
Case-Control      Mantel-Haenszel              4.7151       1.1840      18.7768
  (Odds Ratio)    Logit **                     3.9677       1.0978      14.3395

ONLY SITES 2,4,5************
            Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                        Value     95% Confidence Limits
-------------------------------------------------------------------------
Case-Control      Mantel-Haenszel              4.7151       1.1840      18.7768
  (Odds Ratio)    Logit **                     3.9677       1.0978      14.3395
```

That is, sites with only one type of response do not contribute to the OR estimate

Loglinear Model

Dropping sites 1 and 3 also make the loglinear model converge

```
proc genmod data=one;
 where center in (2,4,5);
 class center tx outcome;
 model count = center|tx|outcome@2 /dist=poi link=log type3;
run;
----------
        LR Statistics For Type 3 Analysis

                                Chi-
Source                 DF     Square    Pr > ChiSq

center                  2       4.76        0.0926
tx                      1       2.86        0.0911
center*tx               2       1.44        0.4860
outcome                 1      22.57        <.0001
center*outcome          2      12.20        0.0022
tx*outcome              1       5.49        0.0192
```

However, this model (the homogeneous association model) is the same as a logistic regression model

Log-linear results

```
                               Analysis Of Parameter Estimates

                                         Standard   Wald 95% Confidence
Parameter                DF   Estimate      Error          Limits

Intercept                 1     2.5148     0.2785     1.9689     3.0606
center           2        1    -0.2270     0.4206    -1.0513     0.5973
center           4        1    -0.7597     0.4722    -1.6852     0.1658
center           5        0     0.0000     0.0000     0.0000     0.0000
tx               A        1    -0.3588     0.4208    -1.1834     0.4659
tx               P        0     0.0000     0.0000     0.0000     0.0000
outcome          1        1    -2.0223     0.6700    -3.3354    -0.7092
outcome          2        0     0.0000     0.0000     0.0000     0.0000
tx*outcome       A   1    1     1.5460     0.7017     0.1708     2.9212
(the common odds ratio of tx with outcome exp(1.5460) = 4.69)
center*outcome   2   1    1    -2.1802     1.1327    -4.4003     0.0399

center*outcome   4   1    1     1.0631     0.7011    -0.3110     2.4373

center*tx        2   A    1     0.5682     0.5938    -0.5956     1.7319

center*tx        4   A    1    -0.2280     0.6771    -1.5550     1.0990
```

Structural Zeros

- By default, if you have data with a zero count in GENMOD, it will be considered a sampling zero
- By default, PROC CATMOD will consider it a structural zero
- To make GENMOD consider the zeros as structural, delete the observations with zero
- To make CATMOD consider the zeros as sampling, add a small weight (like the .5 approach, but much smaller like 10-6) to the count
- As stated before, I tend to use GENMOD more than CATMOD

```
data two;
 set one;
 if count = 0 then delete;
run;
proc genmod data=two;
 class tx outcome center;
 model count = tx|outcome|center@2 /link=log dist=poi type3;
run;
```

```
                               Analysis Of Parameter Estimates

                                         Standard   Wald 95% Confidence
Parameter                DF   Estimate      Error          Limits

Intercept                 1     2.5084     0.2796     1.9604     3.0565
tx               A        1    -0.3435     0.4213    -1.1692     0.4822
tx               P        0     0.0000     0.0000     0.0000     0.0000
outcome          1        1    -1.9695     0.6710    -3.2847    -0.6543
outcome          2        0     0.0000     0.0000     0.0000     0.0000
tx*outcome       A   1    1     1.4696     0.7164     0.0654     2.8738
tx*outcome       A   2    0     0.0000     0.0000     0.0000     0.0000
tx*outcome       P   1    0     0.0000     0.0000     0.0000     0.0000
tx*outcome       P   2    0     0.0000     0.0000     0.0000     0.0000
center           1        1    -0.3112     0.4351    -1.1640     0.5415
center           2        1    -0.2059     0.4221    -1.0332     0.6215
center           3        1    -0.8990     0.5274    -1.9327     0.1347
center           4        1    -0.7655     0.4738    -1.6941     0.1632
center           5        0     0.0000     0.0000     0.0000     0.0000
```

With Structural Zeros (model converged)

```
                                         Standard   Wald 95% Confidence
Parameter                DF   Estimate      Error          Limits

tx*center        A   1    1    -0.2443     0.6990    -1.6143     1.1257
tx*center        A   2    1     0.5258     0.6007    -0.6515     1.7031
tx*center        A   3    1     0.6800     0.7213    -0.7338     2.0938
tx*center        A   4    1    -0.2098     0.6738    -1.5305     1.1108
tx*center        A   5    0     0.0000     0.0000     0.0000     0.0000
tx*center        P   1    0     0.0000     0.0000     0.0000     0.0000
tx*center        P   2    0     0.0000     0.0000     0.0000     0.0000
tx*center        P   3    0     0.0000     0.0000     0.0000     0.0000
tx*center        P   4    0     0.0000     0.0000     0.0000     0.0000
tx*center        P   5    0     0.0000     0.0000     0.0000     0.0000
outcome*center   1   2 *   1    -1.9850     1.1586    -4.2559     0.2859
outcome*center   1   4 *   1     1.0533     0.6968    -0.3124     2.4190
outcome*center   1   5 *   0     0.0000     0.0000     0.0000     0.0000
outcome*center   2   1    0     0.0000     0.0000     0.0000     0.0000
outcome*center   2   2    0     0.0000     0.0000     0.0000     0.0000
outcome*center   2   3    0     0.0000     0.0000     0.0000     0.0000
outcome*center   2   4    0     0.0000     0.0000     0.0000     0.0000
outcome*center   2   5    0     0.0000     0.0000     0.0000     0.0000
```

*** outcome by center for centers 1 and 3 not estimated

Limitations

- We have forced the outcomes of treatment by center to be zero for sites 1 and 3
- This is technically not correct (Sites 1 and 3 could have a success in repeated sampling)
- However, we have constrained them to be zero to make the model converge
- We estimate the Common OR of treatment and outcome to be

$$\exp(1.4696) = 4.34$$

- This result is consistent with other results (CMH, Logistic, and log-linear with Sites 1 & 3 eliminated)
- The synergy of these different methods suggests that the treatment is beneficial in reducing the fungal infection