36-720: Zero Cells

Brian Junker

September 24, 2007

- Zero Counts in Tables
- Fixed (Structural) Zeros and Incomplete Tables
- Sampling (Random) Zeros and Small Samples

Zero Counts in Tables

Sometimes we see tables that have 0 counts in them. For example, consider the tables:

	$Z=1$		$Z=2$	
	$Y=1$	$Y=2$	$Y=1$	$Y=2$
$X=1$	58	69	11	34
$X=2$	12	41	43	–

	$Z=1$		$Z=2$	
	$Y=1$	$Y=2$	$Y=1$	$Y=2$
$X=1$	0	69	11	34
$X=2$	12	41	43	0

We have to think carefully about zeros because:

- SE's, Pearson (and other) residuals, G^2-tests, X^2-tests, etc. all depend on asymptotics ($n \to \infty$) for their distribution theory; and $0 \neq \infty$.
- Our main modeling tool is the *log-linear* model, $\log m = X\beta$, and while $n_{ijk\ldots} = 0$ is ok, we cannot deal with $m_{ijk\ldots} \leq 0$.

How we deal with 0's in tables depends on why they are there!

Fixed (Structural) Zeros and Incomplete Tables

Fixed zeros are cells in the table that are *forced* to be there, due to (a) the nature of the things being cross-classified; or (b) the method of observation. For example,

- For c objects we can form a $c \times c$ table of preference data, where n_{ij} is the number of respondents that prefer object i to object j. Clearly n_{ii} are *fixed zeros*.
- A capture-recapture experiment might take X = "captured in first try" and Y = "captured in second try". Clearly X = "no" and Y = "no" is a *fixed zero*.
- An auto-insurance study might take X ="has drivers' license", Y = "has made insurance claim", Z = "has had license suspended", etc. Clearly X = "yes" and Z = "yes" is a *fixed zero*, regardless of Y.

What to do with fixed zeros

- The term "fixed zeros" is a misnomer: These aren't actually "zero cells", they are cells for which an observation cannot be made (i.e. a kind of *missing data*).
- A table with a fixed zero is an "*incomplete table*" and instead of putting a zero, we should just indicate no observation is possible, e.g.:

	$Z = 1$		$Z = 2$	
	$Y = 1$	$Y = 2$	$Y = 1$	$Y = 2$
$X = 1$	58	69	11	34
$X = 2$	12	41	43	–

- In the Poisson log-linear modeling framework, fixed zeros are easy to deal with: just as we omit them in the table, we should omit them in the model.
- Calculating MLE's will not be a problem. However for G^2 and other goodness of fit tests, be sure to check *degrees of freedom*, because the missing cells do not contribute to saturated model df's.

Example 1

Consider a capture-recapture experiment where $X = 1$ if captured on the first try, $Y = 1$ if captured on the second try.

	$Y = 1$	$Y = 0$
$X = 1$	11	34
$X = 0$	43	–

if we attempt to check the model of independence [X][Y], we get

```
> n <- c(11,34,43)
> x <- c(1,1,0)
> y <- c(1,0,1)
> summary(fit <- glm(n ~ x + y,family=poisson))
[...]
    Null deviance:  2.1354e+01  on 2  degrees of freedom
Residual deviance: -2.8399e-29  on 0  degrees of freedom
```

This should have been a 1 df test, but because the saturated model only has 2 df instead of 3, it is a 0 df test (i.e. [X][Y] is already saturated for the incomplete table!).

We cannot test independence but we can assume it, fit the model and estimate the empty cell:

```
> unlist(predict(fit,newdata=data.frame(n=0,x=0,y=0),
+ type="response",se.fit=T))
           fit         se.fit residual.scale
     132.90909       50.36126        1.00000
```

This is the same answer that we would have gotten by setting

$$\frac{n_{11}n_{00}}{n_{10}n_{01}} = 1$$

and solving for n_{00}: $\hat{n}_{00} = 1 \cdot \frac{(43)(34)}{(11)} = 1 \cdot \frac{n_{10}n_{01}}{n_{11}} = 132.91$

(Note:
There is some issue about whether the SE above will be valid for the incomplete table under *multinomial sampling*, since we are in effect "conditioning on the observed counts", instead of conditioning on n_{++}, but it will still be valid asymptotically. See for example Darroch, Fienberg, Glonek & Junker, 1993, *JASA*, and the references therein).

Example 2

Now consider the incomplete 3-way table

	$Z = 1$		$Z = 0$	
	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$
$X = 1$	58	69	11	34
$X = 0$	12	41	43	–

Here X represents persons listed in the US census, Y represents persons found in a "post-enumeration survey", and Z represents persons found in administrative records (drivers' licenses, tax records, etc.). If we try to fit [X][Y][Z],

```
> n <- c(58,69,12,41,11,34,43)
> x <- rep(c(1,1,0,0),2)[-8]
> y <- rep(c(1,0),4)[-8]
> z <- c(rep(1,4),rep(0,4))[-8]
> summary(fit <- glm(n ~ x + y + z,family=poisson))
    Null deviance: 81.717  on 6  degrees of freedom
Residual deviance: 54.830  on 3  degrees of freedom
```

we see that R got the df right (should have been 4 df but only 3 because of the missing cell), but the fit is terrible.

We try again with [XY][XZ][YZ] and we discover

```
> summary(fit <- glm(n ~ x*y*z - x:y:z,family=poisson))
    Null deviance: 8.1717e+01  on 6  degrees of freedom
Residual deviance: 2.4425e-15  on 0  degrees of freedom
```

so this is the saturated model for the incomplete table. Again, we can estimate the missing cell

```
> unlist(predict(fit,newdata=data.frame(n=0,x=0,y=0,z=0),
+ type="response",se.fit=T))
          fit         se.fit residual.scale
     381.7123      203.0747         1.0000
```

and this is again equivalent to assuming that the odds ratios are equal in the two tables conditional on Z

$$\frac{n_{110}n_{000}}{n_{100}n_{010}} = \frac{n_{111}n_{001}}{n_{101}n_{011}}$$

and estimating $\hat{n}_{000} = \frac{n_{111}n_{001}}{n_{101}n_{011}} \cdot \frac{n_{100}n_{010}}{n_{110}} = \frac{(58)(41)}{(69)(12)} \cdot \frac{(34)(43)}{(11)} = 381.71$

Notes

- *It does not always make sense to estimate the missing cell!* For example in a forced choice experiment where n_{ij} is the number of respondents who prefer Cola i to Cola j we may observe a table like

	Cola A	Cola B	Cola C
Cola A	–	5	6
Cola B	8	–	12
Cola C	12	15	–

  In this case we might try fitting the [1][2] model, but it wouldn't make sense to estimate n_{ii} since this represents an impossible response!
- *Although notation like [X][Y][Z] is still useful it does not quite have the usual interpretation!*. Knowing that $X = 0$ in the capture-recapture experiment, for example, is informative about Y, since the table is not rectangular!
  For this reason, [X][Y][Z] is sometimes called the model of *quasi-independence* for an incomplete 3-way table (and similarly for other models for other incomplete tables).

Sampling (Random) Zeros and Small Samples

Fixed zeros occur because the cell probability $p_{ijk\cdots}$ really is zero.

Sampling zeros occur when $p_{ijk\cdots} > 0$ but no observations occurred in that cell. If $n_{+++\cdots}$ were increased, we would eventually see counts in that cell!!

It seems as though sampling zeros should be a small nuisance that we could avoid by increasing the sample size. *But. . .*

- For many seemingly innocent problems, sampling zeros are virtually unavoidable.

  For example consider cross-classifying the responses on a 10-item True/False test. There are $2^{10} = 1024$ *cells. Some (many!) cells will have probabilities at or below 0.001, and we will need* a lot *of students to be confident of observations in every cell!*
- Sampling zeros immediately suggest that asymptotics (G^2, Pearson residuals, etc.) may be problematic (especially for higher-order models).
- Sampling zeros have to be included in the model. This can play havoc with MLE calculations for log-linear models.

Example 3

Consider the table

	$Z=1$		$Z=2$	
	$Y=1$	$Y=2$	$Y=1$	$Y=2$
$X=1$	0	69	11	34
$X=2$	12	41	43	0

- Consider the saturated model [XYZ]. Ignoring lower-order terms for simplicity, we may parametrize this log-linear model as

$$\log m_{ijk} = u_{123(ijk)}$$

for finite u_{ijk}. But at the MLE, we must have observed = expected sufficient statistics,

$$n_{ijk} = \hat{m}_{ijk} \ ,$$

so that, e.g., $\hat{u}_{123(111)} = \log \hat{m}_{111} = \log n_{111} = \log 0 = -\infty$.

Similarly, any model for which an observed minimal sufficient statistic equals zero will fail to have an MLE.

- Now consider the model [XY][XZ][YZ]. The minimal sufficient statistics are

```
> n <- c(0,69,12,41,11,34,43,0)
> apply(array(n,c(2,2,2)),c(1,2),sum)  # n_{ij+}
     [,1] [,2]
[1,]   11   55
[2,]  103   41
> apply(array(n,c(2,2,2)),c(1,3),sum)  # n_{i+k}
     [,1] [,2]
[1,]   12   54
[2,]  110   34
> apply(array(n,c(2,2,2)),c(2,3),sum)  # n_{+jk}
     [,1] [,2]
[1,]   69   45
[2,]   53   43
```

yet there is still no MLE for this model. The reason (Haberman, 1973; Rinaldo, 2005) is that there is no "nearby table" with all positive entries and the same minimal sufficient statistics: If we add a little to n_{111} to make it positive, we will have to take away from n_{222} to preserve the sufficient statistics, but this will make n_{222} negative.

Note that computer software does not usually detect non-existence of MLE's!

```
> n <- c(0,69,12,41,11,34,43,0)
> x <- rep(c(1,1,0,0),2)
> y <- rep(c(1,0),4)
> z <- c(rep(1,4),rep(0,4))
>
> summary(fit <- glm(n ~ x*y*z,family=poisson))
[...]
    Null deviance: 1.9205e+02  on 7  degrees of freedom
Residual deviance: 3.0331e-10  on 0  degrees of freedom
[...]
Number of Fisher Scoring iterations: 21
>
> summary(fit <- glm(n ~ x*y*z - x:y:z,family=poisson))
[...]
    Null deviance: 1.9205e+02  on 7  degrees of freedom
Residual deviance: 4.4414e-10  on 1  degrees of freedom
[...]
Number of Fisher Scoring iterations: 21
```

Neither of the above models "has" MLE's!!! The output gives us two "hints"

- A large number of iterations in the modified Newton-Raphson algorithm (Fisher Scoring) before the software "declares" convergence;
- Point estimates (not shown above) look strange and SE's (also not shown) are enormous.

but is not able to detect by itself that anything is wrong.

On the other hand the model of independence [X][Y][Z] does possess MLE's for this table (check!), and so we can fit

```
> summary(fit <- glm(n ~ x + y + z,family=poisson))
[...]

    Null deviance: 192.05  on 7  degrees of freedom
Residual deviance: 155.30  on 4  degrees of freedom
[...]
Number of Fisher Scoring iterations: 5
```

The fewer number of iterations is typical when the MLE exists.

Checking existence of MLE's

Haberman (1973, *Ann. Stat.*) gives the following results, for log-linear models ($\log m = X\beta$):

- The log-likelihood is a strictly concave function of $\log m$;
- If all individual cell counts $n_{ijk\ldots} > 0$ then the MLE exists.
- If the MLE of m exists, it is unique and satisfies $n^T X = m^T X$;
- If the MLE of m exists, then the minimal sufficient statistics will all be strictly positive;
- The MLE of m exists, if and only if there is a table of real numbers δ such that $\delta^T X = 0$, and $n + \delta$ has strictly positive entries.

Rinaldo (2005; `http://www.stat.cmu.edu/~arinaldo/Thesis/`) gives a geometric interpretation, shows how to "hunt" for tables with/without MLE's, and shows how to calculate "extended MLE's", corresponding testing procedures, etc., when the usual MLE's do not exist.

Dealing with sampling zeros in practice

- *Some authors* [e.g. Christensen, Ch. 8] recommend
  - Identify all the cells for which the model constraints imply $\hat{m}_{ijk\ldots} = 0$;
  - Treat these cells as fixed zeros (drop them from the model); and analyze the resulting incomplete table.

  This is practical but somewhat unappealing (since the dropped cells aren't really fixed zeros!).
- *Another traditional strategy* is to add a little ε to all cells (say, $\varepsilon = 0.01$ if the counts are "small", $\varepsilon = 0.5$ or 1.0 if the counts are "large") and work with that table instead. The MLE's for the new table are like posterior modes for a Bayesian analysis with a certain Dirichlet prior on the cell probabilities.
- *Lower-order models* can usually be fitted as well, as in the above example. (Reason: lower order minimal sufficient margins sum over more cells, and hence, impose less contraints on the individual table cells);
- *Rinaldo (2005)* provides an alternative to all this, in principle: Compute extended MLE's, and work with them instead.
  - Rinaldo's "extended MLE" is what the `glm()` function is trying to converge to, when the MLE doesn't exist.