Calculation of Degrees of Freedom in Sparse Complete Contingency Tables

Elmar Weixlbaumer
Abteilung für Statistik
University of Economics and Business Administration
A-1090 VIENNA
Austria
email: weix@nelly.mat.univie.ac.at

1. Introduction

Adjustments to degrees of freedom in comparison of complete contingency tables to non-sparse tables are due to parameter inestimability because of marginal zeros. This paper presents GLIM macros for the calculation of the correct degrees of freedom in sparse tables. The algorithm used in the macros is based on the methods that Haslett (1985) introduced. The macros are applicable to a test which is not conditioned on certain table margins.

2. Theory

For multiway tables the logarithm of the probability in each cell may be modelled by a linear combination of sets of u-terms, which parallel the terms in linear models. The number of parameters associated with each set of u-terms is the product of one less than the number of levels for each of the variables in that u-term. Under the log transformation, a fitted probability of zero for a cell corresponds to $-\infty$, but this does not necessarily imply that the parameter is inestimable. If some parameters are inestimable, the degrees of freedom relative to the saturated model are calculated by adding the number of inestimable parameters $z_A(u_q)$ for all u_q not in the unsaturated hierarchical model and subtracting this from the degrees of freedom for this model applied to a non-sparse table. (For notational concepts see Bishop *et al* 1975.) If the difference between two nested hierarchical unsaturated models is calculated, all $z_A(u_q)$ for all u_q that are in one model but not in the other have to be subtracted from the degrees of freedom for testing the same model difference in a non-sparse table.

In Haslett (1990) the number of inestimable parameters $z_A(u_q)$ for a term u_q, where q is a set of variables, is determined by

$$z_A(u_q) = z(u_q) - z_S(u_q). \tag{1}$$

$z_S(u_q)$ is the number of seed zeros in a set of marginal sums C_q, i.e., zeros which are not a consequence of zeros in some $C_{q'}$ for all $q' \subset q$. For example in the table $C_{12} = (0, 0, x_{12}, x_{22})$ the zeros in C_{12} are not seed zeros, because they depend on a zero in $C_2 = (0, x_{12} + x_{22})$ which is a seed zero. $z(u_q)$ is the loss of independent parameters in u_q and is calculated by

$$z(u_q) = z(u_q)^* - Z^*_{k-1} + Z^*_{k-2} - \quad + (-1)^{k-1} Z^*_1, \tag{2}$$

where $z(u_q)^*$ is the number of zeros found in the set C_q and Z^*_l is the sum of the number of zeros in all sets $C_{q'}$ containing exactly l of the k variables in q. A proof for this given in Haslett (1990).

3. Algorithm

Step 1 Test if the table is completely empty or contains no zeros.
Step 2 Set d = number of variables in the model.
Step 3 Determine the number of zeros in each set C_{q_d}, where q_d contains exactly d variables.
Step 4 By considering all subconfigurations of order $d-1$, determine $z_S\left(u_{q_d}\right)$, the number of seed zeros in C_{q_d} for all q_d.
Step 5 Determine $z\left(u_{q_d}\right)$ from (2).
Step 6 Calculate $z_A\left(u_{q_d}\right)$ via (1).
Step 7 Decrease d and repeat Steps 3 to 7 until $d=0$.
Step 8 Calculate degrees of freedom as in Section 2.

4. Example of usage of the macros

The macros below carry out all the steps described in Section 3 except for the last (the calculation of degrees of freedom). Consider the 2^5 table C_{12345}:

$$(0,6,5,3,\ \ 12,9,31,0,0,0,0,6,17,2,3,6,8,1,9,2,0,11,7,9,4,15,2,12,3,4,6)$$

The sixth zero in this table does not depend on a zero in a table C_q, with $q \subset \{1,2,3,4,5\}$ and is therefore a seed zero. Other seed zeros can be found in C_{1345} = (0, 6, 5, 3, 7, 29, 11, 34, 15, 12, 16, 11, 14, 3, 15, 13) and in C_{123} = (14, 53, 0, 28, 24, 20, 30, 25). These zeros cause some parameters to be inestimable. In fact in each of the sets C_{12345}, C_{1234} and C_{1235} exactly one parameter is inestimable.

To achieve these results by using the macros below, the following steps are necessary. First the macros have to be loaded by an input statement.

```
$input 'uep.mac'$
```

This causes the following output if the macros could be loaded from file `uep.mac`, otherwise an error message is printed.

```
********** Macro UEP has been loaded **********
```

Now the table C_{12345} is named `Y` and the factors are named `A`, `B`, `C`, `D` and `E`. These factors have to be collected in a list, here called `FACS`. The GLIM statements for these assignments and the calculation of inestimable parameters would then be:

```
$var 32 Y $
$ass Y=0, 6, 5, 3, 1, 12,  9, 31, 0, 0, 0, 0,  6, 17, 2, 3,
      6, 8, 1, 9, 2,  0, 11,  7, 9, 4,15, 2, 12,  3, 4, 6 $
$gfac 32 A 2 B 2 C 2 D 2 E 2 $
$list FACS=A+B+C+D+E $
$use UEP Y FACS $
```

The macro `UEP` then produces the following output:

```
 Inestimable parameters in configuration
 ---------------------------------------
 1                E,D,C,B,A
 1                D,C,B,A
 1                E,C,B,A

 Number of cells in Y: 32
 Factors in FACS: { A,B,C,D,E }
 Number of zeros: 6       .
 Linearly independent parameters in the saturated model  29
 Inestimable parameters in the saturated model: 3
```

So the saturated model contains 29 linearly independent parameters. If the model with $u_{12345} = 0$ is tested against the saturated model with no condition on marginal sums, the number of inestimable parameters not in the model (= 1) have to be subtracted from 32. Degrees of freedom are then $(32 - 1) - 31 = 0$, as this model for a non-sparse table with these dimensions would contain 31 parameters.

5. Macros

```
$SUBFILE UEP!

!-------------+-----------------------------------------------------
! Author: Elmar Weixlbaumer, Vienna, April 1995
! Version 1.0 for GLIM 4.00
! Main macros:
! UEP Calculates the number of inestimable parameters of a sparse
! complete contingency table with no condition on marginal sums.
! Formal arguments: %1 The name of the variate
!                   %2 The name of the list of factors
! Output: The number of inestimable parameters in each
! configuration,the number of cells, zeros, the factors,
! inestimable and linearly independent parameters in the
! saturated model.

! Example of use: $ass y=1 2 3 0 0 0 7 8 $
! $gfac 8 a 2 b 2 c 2 $list list=a+b+c $
! $use uep y list $
!-------------------------------------------------------------
! $PRI ;'********** Macro UEP has been loaded ***********';$

$MAC UEP!
$INP 'uep.mac' macros$!
$WAR OFF!
$PRI '*** Calculation of inestimable parameters in sparse tables
    ' ***'
$PRI '*** Written by Elmar Weixlbaumer, Vienna, 1995 ***'
$TID TMPL1 TMPLN1
$DEL y_ LIST_ llist ly zs za z
$DEL SUBSET param TMPL2 TMPL3 v1 v2
$NUM zs za z llist param ly $!
$CAL ly=%len(%1) : llist=%len(%2) : param=ly : %a=%if(%a2,0,1) $!
$FAU %a ' Wrong usage, correct: use UEP variate LIST ' $!
$CAL %a=%if(%typ(%1)==11,0,1)$!
$FAU %a ' Wrong usage, correct: use UEP variate LIST ' $!
$CAL %a=%if(%typ(%2)==39,0,1)$!
$FAU %a ' Wrong usage, correct: use UEP variate LIST ' $!
$CAL %a=%cu(%1>0) : %a=(%a==0)$!
$FAU %a ' The table contains only zeros. ' $!
$CAL %z=%cu(%1==0) : %a=(%z==0)$!
$FAU %a ' The table doesn't contain zeros. '$!
$PRI ; ' Inestimable parameters in configuration' $!
$PRI '---------------------------------------' $!
$VAR ly y_$
$LIS LIST_=%2 $!
$CAL y_=%1 : %n=llist $!
$WHI %n loop9 $!
```

```
$PRI ;'Number of cells in ' *n %1 ': ' *i ly
 'Factors in ' *n %2 ': { ' %2 ' }';
 'Number of zeros: ' *i %z ;
 'Linearly independent parameters in the
  *i param $!
$CAL param=ly-param$!
$PRI 'Inestimable parameters in the saturated
      *i param ; $!
$TID TMPL1 TMPLN1 $!
$DEL y_ LIST_ llist ly zs za z $!
$DEL SUBSET param TMPL2 TMPL3 v1 v2 $!
$DEL UEP CALC LOOP1 LOOP2 LOOP3 LOOP4 LOOP5 LOOP6
$DEL LOOP9 LOOP10 LOOP11 LOOP12 PRTPARAM$!
$END!
$RETURN!

!---------------------- Main calculation
!
$SUBFILE macros!
!
$MAC calc!
$CAL zs=0 : z=0 : %b=%len(SUBSET) $!
$LIS %b TMPL1 $LIS %b TMPLN1 $!
$TAB the y_ total for SUBSET into v2 by TMPL1 $!
$CAL %a=%cu(v2==0) : %c=%len(v2)*(%a>0) $!
$WHI %c loop1 $!
$CAL %q=1 : %s=(%a>0) :%t=%len(SUBSET) $!
$WHI %s loop6 $!
$CAL za=z-zs : %y=(za>0) $!
$SWI %y prtparam$!
$CAL param=param-za $!
$TID TMPL1 TMPLN1 $DEL v2 TMPL2 v1 $!
$END!
!
!----------------- Calculation of seed zeros

$MAC loop1 !
$CAL %d=(v2(%c)==0) $SWI %d loop2 $CAL %c=%c-1$!
$END!

$MAC loop2 !
$CAL %e=%b $WHI %e loop3 $!
$CAL %e=%b : %h=1 $WHI %e loop4 $!
$CAL zs=zs+(%h>0) $!
$END!

$MAC loop3 !
$NUM TMPLN1[%e] $CAL TMPLN1[%e]=TMPL1[%e](%c) : %e=%e-1 $!
$END!
!
$MAC loop4 !
$LIS TMPL2=SUBSET-SUBSET[%e] : TMPN2=TMPLN1-TMPLN1[%e] $!
$CAL v1=y_ : %f=%b-1 $WHI %f loop5 $!
$CAL %g=%cu(v1) : %h=%h*%g : %e=%e-1$!
$END!

$MAC loop5 !
$CAL v1=v1*(TMPL2[%f]==tmpn2[%f]) : %f=%f-1 $!
$END!

!---------------------- Calculation of z -----------------

$MAC loop6 !
$DEL TMPL3 $LIS TMPL3 $!
$CAL %r=%t : %u=%q $WHI %r loop7 $!
$DEL v2 $!
$TAB the y_ total for TMPL3 into v2 $!
$CAL %v=%cu(v2==0) : %v=(-1)**(%t-(%len(TMPL3)))*%v $!
$CAL z=z+%v :%q=%q+1 : %s=(%q/=2**%t) $!
$END!
```

```
!
$MAC loop7 !
$CAL %p=%tr(%u/(2**(%r-1)))$!
$SWI %p loop8 $CAL %r=%r-1 $!
$END!
!
$MAC loop8 !
$CAL %u=%u-2**(%r-1) $
$LIS TMPL3=TMPL3+SUBSET[%r] $!
$END
!                    .
                        Calculation of subsets ---

$MAC loop9 !
$CAL %o=1 : %k=1 $WHI %k loop10 $CAL %n=(%n-1) $!
$END!

$MAC loop10 !
$DEL SUBSET $LIS SUBSET $!
$CAL %j=llist : %l=%o $WHI %j loop11 $!
$CAL %m=%if(%len(SUBSET)==%n,1,0) $!
$SWI %m calc $!
$CAL %o=%o+1 : %k=(%o/=2**llist) $!
$END!
!
$MAC loop11 !
$CAL %i=%tr(%l/(2**(%j-1))) $SWI %i loop12 $CAL %j=%j-1 $!
$END!

$MAC loop12 !
$CAL %l=%l-2**(%j-1) $LIS SUBSET=SUBSET+LIST_[%j] $!
$END!

!---------------------------- Output

$MAC prtparam !
$PRI ' ' *i za,3 '   ' SUBSET $!
$END!

$RETURN
$FINISH
```

6. Remarks on the macros

6.1. Error messages

If input is not in the form `$USE UEP Y LIST $`, where `Y` is the variate and `LIST` is a list which contains all the factors, an error message will be printed. Error messages will also be printed if `Y` does not contain any zeros or `Y` is completely empty.

6.2. Deletion of variables and lists

The macros include several lines with directives to delete the intermediary variables to free memory. If it is of interest to save these results, not all of the `$DEL` and `$TIDY` directives can be deleted as some of them are important for calculations.

6.3. Deletion of macros

The file `uep.mac` is divided into two subfiles. The macro `UEP` loads the second half of the file for calculations and deletes it afterwards to free memory. If this is not necessary all `$SUBFILE` and `$RETURN` directives have to be deleted. The last `$FINISH` statement has then to be altered to a `$RETURN` statement. The first line of macro `UEP` (`$INP 'uep.mac' MACROS $`) and the last two (`$DEL LOOP1 ...$`) can then be deleted.

7. References

Bishop Y M M, Fienberg S E and Holland P W (1975) *Discrete Multivariate Analysis* MIT Press, Cambridge.

Haslett S (1985) An algorithm for degrees of freedom calculations in sparse contingency tables *Generalized Linear Models: Lecture Notes in Statistics* (ed R Gilchrist, B Francis and J Whitaker) Springer-Verlag.

Haslett S (1990) Degrees of freedom and parameter estimability in hierarchical models for sparse complete contingency tables *Computational Statistics and Data Analysis* 9 179–195.