A MATRIX APPROACH TO LOG-LINEAR DATA ANALYSIS USING SAS/IML®

Paul Johnson, University of Wales, Aberystwyth, Wales, Great Britain

ABSTRACT

This work involves a matrix algebra approach to log-linear data analysis using SAS/IML®.

Consider the log-linear model $\log(m) = \mu = X\beta$. The Maximum Likelihood Estimator for m , $\hat{m}$, is found. Suppose we are interested in the log odds ratios and the variance/covariance matrix associated with the log odds ratios. The terms get very messy. It makes great sense therefore to revert to the matrix approach in obtaining these estimates.
SAS/IML ,the matrix language procedure, can be used in calculating the log-odds ratio estimates and their respective variance estimates. By using SAS/IML the user has far more power in choice of the design matrix, contrasts and the estimate counts. The estimate counts may take into account any prior information available to produce pseudo-Bayesian estimates. This method is in particular very useful when zero cells are contained within the contingency table.
Adjusted residuals, leverages and Cook's distances to detect for the presence of outliers and/or influential points can be calculated for the expected cell counts using Pseudo-Bayesian Estimators. SAS/IML can be used to perform such computation.

INTRODUCTION

Consider the log-linear model $\log(m) = \mu = X\beta$.
The Maximum Likelihood Estimator for m , $\hat{m}$, is found. For illustrative purposes suppose there are 4 variables of interest A, B, C and D.
Let there be I levels of A, J levels of B, 2 levels of C and L levels of D.

$\hat{m}$ is a column vector of estimated counts and contains the $\hat{M}_{ijkl}$ values.

Suppose we are interested in the log odds ratio

$$\ln(\psi_l) = \ln\left[\frac{M_{11l} * M_{IJl}}{M_{1Jl} * M_{I1l}}\right] \text{ and } \ln(\hat{\psi}_l) = \ln\left[\frac{\hat{M}_{11l} * \hat{M}_{IJl}}{\hat{M}_{1Jl} * \hat{M}_{I1l}}\right].$$

$$\Rightarrow \ln(\psi_l) = \ln(M_{11l}) + \ln(M_{IJl}) - \ln(M_{1Jl}) - \ln(M_{I1l}) \text{ and}$$
$$\ln(\hat{\psi}_l) = \ln(\hat{M}_{11l}) + \ln(\hat{M}_{IJl}) - \ln(\hat{M}_{1Jl}) - \ln(\hat{M}_{I1l}).$$

"." indicates that whether we were to choose the (k=1) level for the variable C or the (k=2) level, since the level of C remains the same when comparing the A x B combinations for each level of D, the odds ratios will be equal for either the (k=1) level or the (k=2) level.

The table of variance and covariance estimates along with the model restrictions equation can be used to obtain the $\widehat{\text{Var}}[\ln(\hat{\psi}_l)]$ values for $l = 1, 2, \ldots\ldots, L$.
This table can be obtained from the PROC CATMOD procedure of SAS/STAT®. The terms get very messy. It makes great sense therefore to revert to the matrix approach in obtaining these estimates. Christensen (1987) shows a matrix approach to linear models. Christensen (1990) shows a matrix approach to log-linear models.

Given the $\hat{M}_{ijkl}$ values $\hat{L}_{ijkl} = LN(\hat{M}_{ijkl})$ values can be computed.
These are the estimated Loge cell counts.

Let the vector $\underline{\hat{l}}$ represent the vector of these estimated Loge cell counts. From these contrasts can be computed.

Contrasts in the $\hat{L}_{ijkl}$'s correspond to vectors $\underline{\rho}$ in which the elements of $\underline{\rho}$ add up to zero i.e., $\underline{\rho}'\underline{J} = 0$ where $\underline{J} = [1,1,1.....1]'$.

Log odds' ratio estimates are themselves contrasts of the estimated loge cell counts. Now consider the variance of the contrast $\underline{\rho}'\underline{\hat{l}}$.
Christensen (1990) shows:

$$\text{Var}(\underline{\rho}'\underline{\hat{l}}) = \underline{\rho}'X(X'D(m)X)^{-1}X'\underline{\rho}$$

where $D(m) = \left[d_{ij}\right]$ where $d_{ii} = M_{ijkl}$, $d_{ij} = 0$ $i' \neq j'$.

The estimated variance is given by :

$$\widehat{\text{Var}}(\underline{\rho}'\underline{\hat{l}}) = \underline{\rho}'X(X'D(\hat{m})X)^{-1}X'\underline{\rho}$$

where $D(\hat{m}) = \left[d_{ij}\right]$ where $d_{ii} = \hat{M}_{ijkl}$, $d_{ij} = 0$ $i' \neq j'$.

From the estimated loge cell counts the vector $\underline{\hat{l}}$ can be constructed.
The estimated variances of the estimated log cell counts equals
$X(X'D(\hat{m})X)^{-1}X'$.

PROC IML ,the matrix language procedure, can be used in calculating the log-odds ratio estimates and their respective variance estimates. By using PROC IML the user has far more power in choice of the design matrix, contrasts and the estimate counts $\hat{M}_{ijkl}$. The estimate counts may take into account any prior information available to produce pseudo-Bayesian estimates. This method is in particular very useful when zero cells are contained within the contingency table.
For any log-linear analysis PROC CATMOD procedure creates response profiles only for those profiles that are actually observed.
SAS recommends that if you use the standard response functions and there are zero frequencies then one should use maximum-likelihood estimation rather than weighted least squares to analyze the data.
The maximum likelihood method estimates the parameters of the linear model so as to maximize the value of the joint multinomial likelihood function of the responses. The weighted-least squares method minimizes the weighted residual sum of squares for the model.

In general zero valued cells are troublesome because, paradoxically, some zeros are smaller than others when we compute rates. For example it may be misleading to report 0/100 and 0/1 as being equal to zero. We may also observe 0/0. Now 0/100,0/1 and 0/0 carry quite different information. The basic problem here is one of simultaneously estimating a large number of parameters (the expected cell frequencies).One way to provide such estimates is to assume an underlying parametric model for the expected cell frequencies, where the number of parameters in the model is typically much smaller than the total number of cells and then to estimate the parameters. The structural zeros are automatically treated as missing values and are deleted by PROC CATMOD. If these cells are to be treated as non missing often 0.5 is added to each cell.

A striking conclusion made in "Discrete Multivariate Analysis" by Yvonne Bishop ,Stephen Fienberg and Paul Holland is that the method proposed using pseudo-Bayesian estimation is "clearly" superior to the generally accepted practice of adding 0.5 to the count in each cell of a large sparse table.

To obtain pseudo-Bayes estimates for an observed array $\{ x_{ijkl} \}$ with

$N = \sum_{ijkl} x_{ijkl}$ the procedural steps are:

1. Select a prior array of probabilities $\{ \lambda_{ijkl} \}$ which may be based on external information or the data values themselves.
$\{ i = 1,2,....., I; j = 1,2,....., J; k = 1, 2; l = 1,2,....., L \}$.
2. Compute the weighting factor

$$\hat{K} = \frac{N^2 - \sum_{ijkl} x_{ijkl}^2}{\sum_{ijkl} (x_{ijkl} - N\lambda_{ijkl})^2} = \frac{N^2 - \sum_{ijkl} x_{ijkl}^2}{\sum_{ijkl} x_{ijkl}^2 - 2N \sum_{ijkl} x_{ijkl}\lambda_{ijkl} + N^2 \sum_{ijkl} \lambda_{ijkl}^2}$$

3. Compute the cell estimates

$$\hat{M}^{*}_{ijkl} = N\hat{p}^{*}_{ijkl} = \frac{N(x_{ijkl} + \hat{K}\lambda_{ijkl})}{N + \hat{K}}.$$

Johnson (1994) provides for a software program, written in C, to calculate the pseudo-Bayesian estimator for missing cell data in contingency tables.
PROC IML can then be used to calculate the variance estimates of the contrast $\underline{\rho}'\hat{\underline{l}}$.

$$\widehat{\text{Var}}^{*}(\underline{\rho}'\hat{\underline{l}}) = \underline{\rho}'\mathbf{X}(\mathbf{X}'\mathbf{D}(\hat{\mathbf{m}})\mathbf{X})^{-1}\mathbf{X}'\underline{\rho}$$

where $\mathbf{D}(\hat{\mathbf{m}}) = \left[d_{ij} \right]$ where $d_{ii} = \hat{M}^{*}_{ijkl}$, $d_{ij} = 0$ $i' \neq j'$.

* indicates the use of the pseudo-Bayesian estimators for estimating cell counts.

RESIDUALS, LEVERAGE POINTS AND OUTLIERS

Christensen (1990) gives a theorem by which is shown that for multinomial, or product-multinomial sampling, if the log-linear model $\log(\mathbf{m}) = \mu = \mathbf{X}\beta$ holds, then as $N \rightarrow \infty$,

$$N^{-1/2}(\mathbf{n}_N - \hat{\mathbf{m}}_N) \xrightarrow{L} N(\mathbf{0}, \mathbf{D}^{*}(\mathbf{I} - \mathbf{A}))$$

where $\mathbf{D}(\mathbf{m}^{*}) = \left[d_{ij} \right]$ where $d_{ii} = M^{*}_{ijkl}$, $d_{ij} = 0$ $i' \neq j'$

and $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{D}(\mathbf{m}^{*})\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}(\mathbf{m}^{*})$,

and M^{*}_{ijkl} indicates these are population cell count values under the assumption of a conditional independence model.
$\mathbf{m}$ is the vector $(M_{1111}, \ldots\ldots\ldots M_{IJKL})'$.

For large samples this leads to the approximation:

$$(\mathbf{n} - \hat{\mathbf{m}}) \sim N(\mathbf{0}, \mathbf{D}(\hat{\mathbf{m}})(\mathbf{I} - \mathbf{A}(\hat{\mathbf{m}})))$$

where $\mathbf{D}(\hat{\mathbf{m}}) = \left[d_{ij} \right]$ where $d_{ii} = \hat{M}_{ijkl}$, $d_{ij} = 0$ $i' \neq j'$

and $\mathbf{A}(\hat{\mathbf{m}}) = \mathbf{X}(\mathbf{X}'\mathbf{D}(\hat{\mathbf{m}})\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}(\hat{\mathbf{m}})$,

$\hat{M}_{ijkl}$ represents the ijkl 'th estimated cell count.

These $(\mathbf{n} - \hat{\mathbf{m}})$ are the residuals. There are two methods of standardizing residuals. One is a crude standardization where the matrix $\mathbf{A}(\hat{\mathbf{m}})$ is ignored. Thus the crude standardized residuals are defined as:

$$\tilde{r}_i = \frac{n_i - \hat{m}_i}{\sqrt{\hat{m}_i}}$$

These residuals are often called standardized residuals or Pearson residuals. These residuals have the advantage that they do not require the computation of the diagonal elements of a complicated matrix depending on $\mathbf{X}$. The sum of the squared crude residuals is the Pearson test statistic for lack of fit.

Taking the matrix $\mathbf{A}(\hat{\mathbf{m}})$ into account the corrected standardized residuals are defined as:

$$r_i = \frac{n_i - \hat{m}_i}{\sqrt{\hat{m}_i(1 - \hat{a}_{ii})}}$$

where $\hat{a}_{ii}$ is the i'th diagonal element of the square matrix $\mathbf{A}(\hat{\mathbf{m}})$.
These adjusted residuals can be calculated for the expected cell counts using the Pseudo-Bayesian Estimators and using PROC IML to perform for the computation.
The leverage measures how far the variables of a particular case are from the average of all of the cases. Recall for log-linear models the estimated leverage of the i'th case is denoted by

$\hat{a}_{ii}$, the i'th diagonal element of the square matrix $\mathbf{A}(\hat{\mathbf{m}})$, defined above.
Outliers are data points, in the log-linear case cells, that split off or are very different from the rest of the data. An influential point is one in which the point has a large effect on the outcome of the log-linear models that are fit to the data. Cook's distance is a measure which can be used to detect whether or not particular cells are influential points or outliers. Cook's distance for the i'th case is a measure of the influence the i'th case has on the fit of the model. Since we are fitting log-linear models each cell is dropped from the table, the remaining cells are fit and an expected cell count for the dropped cell is estimated.
This is done in turn for each cell.
A natural version of Cook's distance that is appropriate for log-linear models is:

$$C_q(\mathbf{X}'\mathbf{D}(\hat{\mathbf{m}})\mathbf{X}, df) = \frac{(\hat{\beta} - \hat{\beta}_{(q)})'\mathbf{X}'\mathbf{D}(\hat{\mathbf{m}})\mathbf{X}(\hat{\beta} - \hat{\beta}_{(q)})}{df}$$

where cell q may be an outlier and the log-liner model is $\mu = \mathbf{X}\beta$.

$\hat{\beta}_{(q)}$ is the vector of estimated coefficients with the q'th cell deleted.

A one step version of Cook's distance is:

$$C^{1}_{q}(\mathbf{X}'\mathbf{D}(\hat{\mathbf{m}})\mathbf{X}, df) = \frac{(\hat{\beta} - \hat{\beta}^{1}_{(q)})'\mathbf{X}'\mathbf{D}(\hat{\mathbf{m}})\mathbf{X}(\hat{\beta} - \hat{\beta}^{1}_{(q)})}{df}$$

where $\hat{\beta}^{1}_{(q)}$ is a one-step approximation to $\hat{\beta}_{(q)}$.

A computational formula for $C^{1}_{q}(X'D(\hat{m})X,df)$ is

$$C^{1}_{q}(X'D(\hat{m})X,df) = \frac{1}{df}\,\frac{\hat{a}_{qq}}{(1-\hat{a}_{qq})^{2}}\,\frac{(x_q-\hat{m}_q)^{2}}{\hat{m}_q}.$$

Cook and Weisberg (1982) indicate that a distance measure greater than 1 would generally be considered large. SAS/IML can be used to obtain the residuals, leverages and Cook's distances.

SAMPLE PROGRAM

```
/*  Set of SAS/IML Subroutines in order to estimate the adjusted
    residuals, leverages and Cook's distances to detect for the presence
    of outliers and/or influential points for the  expected cell counts
    using pseudo-Bayesian estimators.                                  */

proc iml;

  x = {1 1 0  1  0  1  1  0  0  1  0  0  0  1
       0  1  0  0  0  0  0  1  0  1  0  0  0
       .  .  .  .  .  .  .  .  .  .  .  .  .
       .  .  .  .  .  .  .  .  .  .  .  .  .
       1  1  1  1  1 -1 -1 -1 -1 -1 -1 -1 -1
      -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 };

    print x;
e = {101.75 31.87 21.21 21.26 11.42 3.87 6.5 5.94 32.87 5.11 23.23
23.08 2.87 0.85 17.69 12.92 11.93 4.78 3.48 7.93 4.64 1.95 6.36 9.42
167.79 92.74 23.52 12.6 12.97 7.74 4.97 2.42 25.41 35.79  17.84 33.09
1.53 4.1 9.4 12.75 18.18 9.91 4.82 8.69 4.87 2.78 6.05 7.11 883.22
612.73 63.0 81.14 68.83 51.58 13.4 15.72 610.44 243.99 115.37 180.55
37.04 28.21 61.0 70.15 198.05 94.08 24.33 99.4 53.43 26.61 30.84
81.9};
o = {100.75 31.60 22.01 18.91 12.43 4.14 5.70 8.29 31.60 4.92 21.24
26.16 4.14 1.04 19.68 9.84 11.65 4.92 4.92 9.84 4.92 1.81 4.92 7.51
176.63 92.20 20.20 10.88 4.14 8.29 8.29 4.14 25.90 33.93 18.91 26.16
1.04 5.96 8.29 19.68 18.91 11.65 4.14 10.88 4.14 1.04 6.73 4.92 875.39
613.55 65.52 85.21 76.66 50.76 10.88 11.65 611.22 246.04 116.29
184.40 36.26 26.16 60.09 66.30 197.61 92.20 23.57 95.31 53.87 28.49
31.60 85.99};
d = diag(e);
c = diag(x*inv(x`*d*x)*x`*d);
b=j(1,72,1);
a=b*c;
leverage=a;
print leverage;
r1=o-e;
r2=b-a;
r3=e#r2;
r4=sqrt(r3);
r5=sqrt(e);
residual=r1/r4;
approx=r1/r5;
print residual;
print approx;
r6=r3#r2;
r7=r1#r1;
r8=r7#a;
cook1=r8/r6;
cook=cook1/50;
print cook;
run;
```

REFERENCES

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1977). Discrete Multivariate Analysis - Theory and Practice. M.I.T.

Christensen, R. (1987). Plane Answers to Complex Questions: The Theory of Linear Models. New York, Springer-Verlag.

Christensen, R. (1990). Log-Linear Models. New York, Springer-Verlag.

Cook, R.D., and Weisberg, S. (1982). Residuals and Influence in Regression. New York, Chapman and Hall.

Johnson, P. (1994). "A Pseudo-Bayesian Estimator for Missing Cell Data in Contingency Tables," Internet Biomedical Gopher, University of Texas Medical School, Houston.

SAS Institute. Inc. (1993). SAS/IML Software: Usage and Reference, Version 6, 1st ed. SAS Institute. Inc., Cary, NC.

SAS Institute. Inc. (1993). SAS/STAT User's Guide, Version 6, 4th ed., Volume 1 and Volume 2. SAS Institute. Inc., Cary, NC.

SAS, SAS/IML, and SAS/STAT are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are registered trademarks or trademarks of their respective companies.