A GENERALIZED DIRICHLET MULTINOMIAL MODEL FOR CATEGORICAL DATA WITH EXTRA VARIATION

Jeffrey R. Wilson
Arizona State University

Summary

The generalized-multinomial model proposed by Tallis (1962) for correlated multinomials is generalized to account for extra variation by allowing the vectors of proportions to vary according to a Dirichlet Distribution. The model allows for a second order of pairwise correlation among units, a type of assumption found reasonable in some biological data, Kupper and Haseman (1978). An alternative derivation allowing for two kinds of variations with more practical applications is proposed. Asymptotic normal properties of parameter estimators are derived, allowing the use of Wald statistics for testing hypotheses.

Key Words: Generalized Multinomial; Wald Statistics; Dependent Multinomials.

1. Introduction

The problem considered here for the variation among proportions is analogous to the randomized block design with random components for interval level data. The model presented in this paper allows for the analysis of variation among replicates and among units for a given replicate. Ignoring either level of variation leads to underestimation of the true standard errors of estimated proportions.

Such problems for quantitative data have been addressed by Healy (1972) and Cochran (1943). They examined the analysis of variance for percentages based on unequal numbers through a non parametric analysis. In this paper the Dirichlet Multinomial distribution is used to include these two types of variations.

It is shown here that the results obtained using Dirichlet Multinomial models are similar to the results obtained when one considers the Generalized Multinomial distribution with a Dirichlet prior. Tallis (1962) proposed the use of the generalized multinomial model for dependent multinomials. The model is extended to allow for a second random component. The models considered here can be viewed as multivariate extensions of the Beta-binomial and correlated binomial models considered by Kupper and Haseman (1978) and Crowder (1978) for binary data.

2. Generalized Multinomial Model

Consider a system of J units which are simultaneously observed at n different times. At each time, each unit is classified as being in one of I mutually exclusive states. Let the random variable X_{ijt} take the value 1, if at time t the j-th unit is observed to be in the i-th state and zero otherwise. The probability that X_{ijt} take the value 1 is assumed to be π_i for each unit j and time point t. Furthermore, observations taken at different time points are assumed to be independent and $\underset{\sim}{X}_j = (X_{1j}, X_{2j}, \ldots, X_{Ij})'$, the vector of counts for the j-th unit has a multinomial distribution with probability vector $\underset{\sim}{\pi} = (\pi_1, \pi_2, \ldots, \pi_J)'$ and sample size n. However, responses given by the J units at a particular time point may be correlated, producing a set of J correlated multinomial random vectors, $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_J$.

Tallis (1962) developed a model for this situation referred to as a generalized multinomial distribution in which a single parameter, ρ is used to reflect the common dependency between any two of the dependent multinomial random vectors. The distribution of the category total $X_{ij} = \sum_{t=1}^{n} X_{ijt}$ is binomial with sample size n and parameter π_i, for each unit. Tallis formalized the dependencies among unit totals for the i-th category by specifying the joint moment generating function as

(2.1)
$$G_i(\underset{\sim}{\mu}) = \rho \sum_{k=1}^{n} p_{ik}\left(\prod_{j=1}^{n} e^{u_j}\right)^k + (1-\rho)\prod_{j=1}^{J} P(e^{u_j})$$

where $P(e^{u_j}) = \sum_{k=0}^{n} p_{ik}e^{ku_j}$ and $\underset{\sim}{\mu} = (\mu_1, \mu_2, \ldots, \mu_J)'$.

The parameter ρ appearing in (2.1) is the correlation coefficient between X_{ij} and $X_{ij'}$ for any $j \neq j'$. When $\rho \neq 0$, (2.1) is a linear combination of moment generating functions for perfectly correlated X_{ij}'s, with weights ρ and (1-ρ), respectively. Altham (1978) proposed a similar model for a joint moment generating function for correlated binary variables.

Consider the overall vector of category totals $\underset{\sim}{X} = \sum_{j=1}^{J} \underset{\sim}{X}_j$. From the moment generating function in (2.1) it can be shown that $E(X) = Jn\underset{\sim}{\pi}$ and $V(\underset{\sim}{X}) = Jn\{1+(J-1)\rho\}M_{\underset{\sim}{\pi}}$ for the generalized multinomial model, where $M_\pi = \mathrm{diag}(\underset{\sim}{\pi}) - \underset{\sim}{\pi}\underset{\sim}{\pi}'$ and $\mathrm{diag}(\underset{\sim}{\pi})$ is a diagonal matrix. Consequently $\hat{\underset{\sim}{\pi}} = (Jn)^{-1}\underset{\sim}{X}$ is an unbiased estimator for π. Tallis (1962) proposed estimators for ρ, but he did not discuss techniques for making inferences about $\underset{\sim}{\pi}$. We consider here a technique for making such inferences.

One approach is to use the limiting normal distribution of $\underset{\sim}{X}$ as $n\to\infty$. At time t consider a vector of dimension IJ, denoted by $\underset{\sim}{X}_{t(J)} = \underset{\sim}{1}_J \boxtimes \underset{\sim}{X}_{jt}$ where $\underset{\sim}{1}_J$ is a J dimensional vector of ones, ⊠ denotes direct product between matrices, and $\underset{\sim}{X}_{jt} = (X_{1jt}, X_{2jt}, \ldots, X_{Ijt})'$.

Define $\underset{\sim}{X}_{(J)} = \sum_{t=1}^{n} \underset{\sim}{X}_{t(J)}$. Since the $\underset{\sim}{X}_{t(J)}$ vectors are independent and the first and second moments of $\underset{\sim}{X}_{t(J)}$ are finite, the multivariate Central Limit theorem implies that

$$n^{-\frac{1}{2}}(\underset{\sim}{X}_{(J)} - n\mu) \to N_{IJ}(0, \underset{\sim}{\Sigma}), \text{ as } n\to\infty,$$

where $\underset{\sim}{\mu} = \underset{\sim}{1}_J \boxtimes \underset{\sim}{\pi}$, $\Sigma = M_\pi \boxtimes Q$ and Q is a square matrix of dimension J with ones on the diagonal and ρ as each off diagonal element. Now $\underset{\sim}{X} = C\underset{\sim}{X}_{(J)}$ where $C = \underset{\sim}{1}_J' \boxtimes I_m$ and I is the identity matrix of dimension mI. Then, by the reproductive property of the multivariate normal distribution, Anderson (1958),

(2.2)
$$n^{-\frac{1}{2}}(\underset{\sim}{X} - nJ\underset{\sim}{\pi}) \sim N_I(\underset{\sim}{0}, J\{1+(J-1)\rho\}M_\pi).$$

Given a consistent estimator for ρ, chi-square tests involving sufficiently smooth functions of π can be obtained from Wald statistics as

(2.3)
$$X_W^2 = nJ\{1+(J-1)\rho\}^{-1}(g(\hat{\pi})-g(\pi))'[D\hat{M}_{\pi}D']^{-}[g(\hat{\pi})-g(\pi)]$$

where D is the matrix of first partial derivatives of g evaluated at π, and $[D\hat{M}_{\pi}D]$ is a generalized inverse of $D\hat{M}_{\pi}D$. The degrees of freedom correspond to the rank of $DM_{\pi}D'$.

3. Dirichlet-Multinomial Model

An alternative derivation of the generalized multinomial model is obtained from the Dirichlet-Multinomial which will be presented here as a random time effects model. At time t, observe independent multinomial responses for each of the J units, each with parameters $\pi_t = (\pi_{1t}, \pi_{2t}, \ldots, \pi_{It})'$ and sample size 1. Furthermore assume the observations taken at different time points are independent. The probability vector π_t is is assumed to fluctuate across time according to a dirichlet distribution with mean vector $\pi = (\pi_1, \pi_2, \ldots, \pi_I)'$ and scaling parameter α.

For this model the sum of the vector of counts, X has a Dirichlet-multinomial distribution and the estimator $\hat{\pi}$ has first moment π and covariance matrix $V(\hat{\pi})=(J+\alpha)(1-\alpha)^{-1}(Jn)^{-1}M_{\pi}$. This Dirichlet-Multinomial model with time effects is related to the generalized multinomial model through the equation $\alpha = \rho^{-1}(1-\rho)$. Thus when the dependency constant ρ is 1 the Dirichlet parameter α is 0 and we have J identical units. When ρ approaches 0, α approaches infinity and we have the case of J distinct units. The dirichlet distribution provides a convenient model for describing variation among vectors of proportions since it has relatively simple mathematical properties. The Dirichlet Multinomial model has been studied by Mosimann (1962) and Good (1965). Brier (1980) used the model to analyze sample proportions obtained from two-stage cluster samples. Koehler and Wilson (1986) generalized some of Brier's techniques and provided extension for comparing vectors of proportions for two-stage cluster samples taken from several populations.

4. Generalized Dirichlet-Multinomial Model

In this section a generalized Dirichlet-Multinomial model is developed for which the observed vectors of counts may be correlated as in the generalized multinomial model. Suppose J units are randomly selected from a population for which the vectors of proportions are distributed with respect to a Dirichlet distribution with parameter σ and $\pi = (\pi_1, \pi_2, \ldots, \pi_I)'$.

As in the generalized multinomial model, the X_j vectors are identically distributed and are not independent. The observations taken at time t on the J individuals are equally pairwise correlated as measured by the parameter ρ. The vector of total counts $X = \sum_{j=1}^{J} X_j$ for the generalized Dirichlet-multinomial model has mean vector $E(X) = N\pi$ and covariance matrix $V(X) = NC\{1+\rho(J-1)\}M_{\pi}$, where N=nJ is the total number of observations and $C = (n+\sigma)(1+\sigma)^{-1}$. Using an argument similar to the one in Section 2, it can be shown that $n^{-\frac{1}{2}}(X-N\pi) \sim N_I(0, JC\{1+(J-1)\rho\}M_{\pi})$ and test of hypotheses about π or vector functions $g(\pi)$, where g is a continuous function with second partial derivatives, can be obtained using the large sample chi-square distributions for the Wald statistic

(4.1)
$$N\{C\{1+(J-1)\rho\}^{-1}(g(\hat{\pi}) - g(\pi))'(DM_{\pi}D')^{-}(g(\hat{\pi}) - g(\pi))$$

where $[DM_{\pi}D']^{-}$ denotes the generalized inverse of $DM_{\pi}D'$, with degrees of freedom equal to rank of $DM_{\pi}D'$. Since C is greater than 1, the test statistic will be smaller than the case for the generalized multinomial model. This reflects the greater imprecision in the estimation for π due to variation in vectors of proportion among individuals. The consequence, of ignoring this extra variation is an inflation of the type I error levels for such tests.

An alternative derivation to the generalized Dirichlet Multinomial model is the following. Suppose that J independent multinomial units are initially selected from a larger population and these may respond with a random vector π_t at particular time point. Thus at time t assume the conditional distribution of π_{jt} is Dirichlet (β, π_t) and the marginal distribution of π_t is Dirichlet (α, π). This model accounts for the extra variation due to time and due to the sampled units. Under this model $E(X) = Jn\pi$ and $V(X) = \frac{(J+\alpha)}{(1+\alpha)}\frac{(n+\beta)}{(1+\beta)} M_{\pi}$. The generalized Dirichlet Multinomial model and the two way model are related through the equation $(n+\sigma)(1+\sigma)^{-1}\{1+(J-1)\rho\} = (n+\beta)(J+\alpha)(1+\beta)^{-1}(1+\alpha)^{-1}$ which results in the same relationship as in the generalized multinomial model and the alternative derivation given to it previously.

5. Estimation of Intra time Correlation

Tallis (1964) considered two methods for estimating the common parameter ρ, but an alternative method is considered here. For any given set of J multinomials a sample correlation matrix, R of dimension J can be obtained. Let the elements of R be denoted by $r_{jj'}$ and define an estimate of ρ as

(5.1)
$$\hat{\rho} = 2J^{-1}(J-1)^{-1}\sum_{j<j'}^{J}\sum r_{jj'}.$$

Once a consistent estimate of ρ and a consistent estimator of C are obtained, the extra variation factor can be computed in the use of the test statistics. Consistent estimates of C can be computed using methods of Brier (1980) in estimating the clustering effect or the regression methods to do the same as in Koehler and Wilson (1986). One simple estimate of C which is easily computed, through most statistical computer packages, is $C = X^2_{MI}/(I-1)(J-1)$ where X^2_{MI} is the Pearson statistic value for testing independence in an IxJ two dimensional contingency table.

6. Test of the Model Assumptions

In using the generalized multinomial model there are two basic assumptions: a) the correlations between the units X_j, and $X_{j'}$ are constant for any $j \neq j'$ and b) the X_j's $j=1,2,\ldots,J$; are identically multinomially distributed. Test statistics are now presented to assess the validity of these assumptions. Large sample tests for the Dirichlet distribution assumption were given by Wilson (1986) and by Koehler and Wilson (1986).

To test that the correlation coefficient is constant in each of the populations but not necessarily the same across populations, one can use the following test procedure as shown by Lawley (1963).

Define

$$r^2_{jj'} = \Big[\sum_{i=1}^{I} (X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_{j'})\Big] \Big/ \sum_{i=1}^{I} (X_{ij} - \bar{X}_j)^2 \sum_{i=1}^{I} (X_{ij'} - \bar{X}_{j'})^2 \qquad 2$$

where $\bar{X}_j = I^{-1} \sum_{i=1}^{I} X_{ij}$.

Define $\bar{r}_i = (J-1)^{-1} \sum_{j \neq i'}^{J} r_{ji'}$ as the average of the off diagonal elements in the i-th column of R.

Define

Define $\hat{r} = 2[J(J-1)]^{-1} \sum_{j<j'}^{J} r_{jj'}$ as the overall average of the off diagonal elements and let $w = (J-1)^2 [1-(1-\hat{r})^2][J-(J-2)(1-\hat{r})^2]^{-1}$.

Then a test statistic

(6.1)

$$T = (J-1)(1-\hat{r})^{-2}\Big[\sum_{j<j'}^{J}\sum (r_{jj'} - \hat{r})^2 - w \sum_{i=1}^{J} (\bar{r}_i - \hat{r})^2\Big]$$

is approximately distributed as a chi-square random variable with $2^{-1}(J+1)(J-2)$ degrees of freedom.

The test for homogeneity of several multinomial distributions is equivalent to testing the hypothesis H_o: $\pi_j = \pi_o$ (unknown vector) $j = 1, 2, \ldots, J$; where $E(X_j) = n\pi_j$.

Let $\hat{\pi}_j = n^{-1} X_j$ and $\hat{\pi}_o = J^{-1} \sum_{i=1}^{I} \hat{\pi}_j$.

The lack-of-fit test statistic is

$$X^2_{LF} = N(\hat{\pi}^{(J)} - \hat{\pi}_o^{(J)})' \{[(1-\rho) I_m + \rho J_m] M_{\hat{\pi}_o}\}^- (\hat{\pi}^{(J)} - \hat{\pi}_o^{(J)})$$

where I_m is the identify matrix of dimension I and J_m is the matrix of ones and $\hat{\pi}^{(J)} - \hat{\pi}_o^{(J)}$ is a ones and concaternation of the vectors $\hat{\pi}_j - \hat{\pi}_o$ for j=1, 2, ...,J. When ρ is zero the lack of fit test is equivalent to the usual Pearson chi-square square test of independence. By use of (3.1) and limiting theorems of quadratic forms given by Stroud (1971), the asymptotic distribution of X^2_{LF} is a chi-square random variable with (I-1)(J-1) degrees of freedom under H_o.

To investigate the homogeneity aspect in the Generalized Dirichlet-Multinomial model we compare $C^{-1} X^2_{LF}$ with a chi-square random variable with (I-1)(J-1) degrees of freedom.

References

Altham, P. M. E. (1978). Two generalizations of the binomial distribution. Applied Statistics 27, 162-167.

Anderson, T. W. (1958). An introduction to multivariate statistical analysis. John Wiley and Sons, New York.

Brier, S. S. (1980). Analysis of contingency tables under cluster sampling. Biometrika 67, 591-596.

Cochran, W. G. (1943). Analysis of variance for percentages based on unequal numbers. Journal of the American Statistical Association 38, 287-301.

Crowder, M. J. (1978). Beta-binomial Anova for proportions. Applied Statistics 27, 34-37.

Good, I. J. (1965). Estimation of Probabilities. MIT Press, Cambridge, Massachusetts.

Healy, M. J. R. (1972). Animal litters as experimental units. Applied Statistics, 21, 155-159.

Koehler, K. J. and Wilson, J. R. (1986). Chi-square tests for comparing vectors of proportions for several cluster samples. Communication in Statistics (in Press).

Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated bionomial model for the analysis of certain toxicological experiments. Biometrics 34, 69-76.

Lawley, D. N. (1963). On testing a set of correlation coefficients for equality. Annals of mathematical statistics 34, 149-151.

Moore, D. S. (1977). Generalized inverses, Wald's method and construction of chi-squared tests of fit. J. Amer. Statist. Assoc. 72: 131-137.

Moseman, J. E. (1962). On the compound multinomial distribution, the multivariate -distribution, and correlation among proportion. Biometrika 49: 65-82.

Stroud, T. W. F. (1971). On obtaining large sample tests from asymptotic normal estimations. Annals of Mathematics Statistics Associations, 72, 881-885.

Tallis, G. M. (1962). The use of a generalized multinomial distribution in the estimation of correlation in discrete data. J. R. Statistical Soc., Series B, 24, 530-534.

Tallis, G. M. (1964). Further Models for estimating correlation indiscrete data. J. R. Statist. Soc., Series B, 26, 82-85.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math., Soc., 54, 426-482.

Wilson, J. R. (1986). Approximate distribution and test of fit for the clustering effects in the dirichlet multinomial model. Communication in Statistics, Vol. A15, No. 4, Theory and Methods.