FITTING CATEGORICAL DATA MODELS USING DATA FROM COMPLEX SAMPLE SURVEYS WITH ITEM NONRESPONSE

Louis Rizzo
The University of Iowa, Iowa City, Iowa 52242

Key words: Dirichlet-multinomial, cluster samples, maximum likelihood

The problem of categorical data analysis when one has a complex survey sample is an important problem as many response variables from these surveys are in fact categorical. The problem has received extensive attention over the past fifteen years. One can trace two main lines of development: the first is a design-based approach developed by J. N. K. Rao, A. J. Scott, R. Fay, D. R. Thomas and others. For a summarization of this approach the author recommends Rao and Thomas [1981] in Skinner, Holt, and Smith [1989], which will provide a complete bibliography. The second line of development is a model-based approach, which was developed in Altham [1976], Cohen [1976], Brier [1980], and Holt and Ewings in Chapter 13 of Skinner, Holt, and Smith [1989], with related papers of Beitler and Landis [1985], Harville and Mees [1989], Mak [1988], and Choi and McHugh [1989].

This paper will continue development of the model-based approach, while taking into account the sample design through pseudo-likelihood methods. The Brier [1980] Dirichlet-multinomial model is the main departure point. This model will be developed to take into account pseudo-likelihood weighting and nonresponse. Full maximum likelihood will be the estimation approach rather than method of moments which tends to be the dominant approach in the model-based papers. This will also allow for easy incorporation of nonresponse effects, and use of maximum likelihood guarantees asymptotic efficiency. An IMSL-type Fortran program is available free from the author to implement the methodology (mail net lrizzo@ stat.uiowa.edu).

1. An Example

The complex survey used as an example in this paper is the National Educational Longitudinal Survey of 1980, sponsored by the U.S. Dept. of Education and carried out by the National Opinion Research Center. The purpose of the survey was to collect information from a cohort of high school sophomores and seniors, and then follow up at intervals into the future. The sample design was primarily developed such that every sophomore or senior in the U.S. would have an equal probability of selection. Certain sub-groups of interest, such as certain minorities, were given higher probabilities of selection.

The sample was a cluster sample, where schools were clusters. The probabilities of selection were somewhat proportional to the size of the two classes, and equal samples of size 40 were taken randomly within each class of a selected school (unless the total size of the class was less than 40 in which case the whole class was the sample). Information was collected at the school level, as well as from the students.

Nonresponse at the school level was dealt with through an adjustment in the probability of selection by multiplying by an estimated probability of response. These adjusted weights will be treated as constants. Unit nonresponse at the student level is ignored: a complete case analysis is done (no information is available on nonresponding students). Item nonresponse at the student level will be dealt with explicitly in this paper.

The population of interest will be restricted to New England, and the responding number of clusters is 52. The average responding subsample size (among both sophomores and seniors) was 52.75, with a minimum of 12 and a maximum of 69.

The first and simplest example to be presented is of a 2×2 categorical table with one margin being sex of the student and the other margin response to a question about the importance to the student of success in their future career ($0 =$ not very important, $1 =$ very important). This question was one of a series of questions about attitudes towards the future, and it is of interest to know if these attitudes differ across categories of students, sex being one category. The null hypothesis of interest would be that of independence between the two margins. We have then four cross categories with cell numbers respectively 1, 2, 3, 4 corresponding to pairs (M,1), (F,1), (M,0), (F,0). Thus the first cell corresponds to males who view success as important, cell 4 are females who do not view success as important, etc.

2. The Model and Full Data Pseudo Log Likelihood

Let y_i be the cell indicator for a particular student. Then in the general population $Pr(y_i = k) = p_k \quad k = 1, 2, \ldots, r$. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{r-1}) = (p_1, \ldots, p_{r-1})$ with $p_r = 1 - \mathbf{1}'\Delta$. Let $\boldsymbol{\eta}$ be a $(r-1)$-vector of $\{\log \lambda_k\}$. Then we assume $\boldsymbol{\eta} = X\boldsymbol{\beta}$, where X is a $(r-1)$ by d design matrix, $\boldsymbol{\beta}$ a d-vector of parameters. In our example a model of interest is that of independence between sex and success, in which case $d = 2$, and the rows of X are [0,0], [1,0], and [0,1]. The y_i's are assumed independent in the population.

Our sample is not a simple random sample from the population. The y_i's in the *responding sample* do not have the simple distribution $Pr(y_i = k) = \lambda_k$, with y_i's independent. The clustering and nonresponse has the effect of making the distribution of the y_i's we actually observe different from the full population distribution. (See Little & Rubin [1987], or Sugden & Smith [1984], for example). A way of adjusting for this is to specify the relationship between the response variable and either the design covariates or the probabilities of selection treated as covariates, and then in effect predicting the y_i's for the nonsampled population. The design covariates are whatever covariates were used in specifying the sample design - in this example size of school (sophomore & senior class) and strata determined probability of selection, where strata were determined by type of school, percentages of particular minorities, and geographical location.

The justification for this predictive approach based on using design covariates and/or probabilities of selection is given in a number of sources, including Rizzo [1992], Chambers [1986], Pfeffermann [1988], Skinner, Holt, and Smith [1989], and Sugden and Smith [1984].

The most general model of this kind would be based on a conditioning on the cluster indicator. If I_C is an indicator for cluster C and $\mathbf{Y}_c$ are the response vectors for elements subsampled within the cluster we can write $Pr(y_i = k|I_c) = p_{kc}$. The substantive modeling assumption in this paper is then to assume these p_{kc}'s are drawn from a Dirichlet distribution with mean vector $\mathbf{p}$, as in Brier [1980], i.e.,

$$f(\mathbf{p}_c|\mathbf{p},k) = \frac{\Gamma(k)}{\prod_{j=1}^{r}\Gamma(kp_j)} \prod_{j=1}^{r}(p_{jc})^{kp_j-1} \quad (2.1)$$

where $\mathbf{p}_c = (p_{1c}, \dots, p_{rc})$, $\mathbf{p} = (p_1, \dots, p_r)$, and k is a parameter which essentially determines the variability of the $\boldsymbol{\lambda}_c$'s around $\boldsymbol{\lambda}$. The Dirichlet assumption at the population level seems reasonable: we have unimodality, a variability in λ_{kc} proportional to $\lambda_{kc}(1-\lambda_{kc})$ and some skewness to the middle (symmetry for $\boldsymbol{\lambda} = \frac{1}{2}$). (See Figure 2).

Assuming no nonresponse the actual student sample in the longitudinal survey will be a simple random sample from the school. This combined with the model independence of the y responses gives independent responses for the students overall within school. Let $(x_{1c}, \dots, x_{kc})$ be the counts of each cell in the cluster subsample. Then the distribution of $(x_{1c}, \dots, x_{kc})$ is multinomial given $\mathbf{p}_c$. The overall distribution of $(x_{1c}, \dots, x_{rc})$ is Dirichlet multinomial:

$$f(\mathbf{x}_c|\mathbf{p},k) = \left(n_c x_{1c} \cdots x_{rc} \right) \frac{\Gamma(k)}{\Gamma(n_c+k)} \times \prod_{j=1}^{r}\Gamma(kp_j) \prod_{j=1}^{r}\Gamma(x_{jc}+kp_j) \quad (2.2)$$

If the clusters are *iid* we could write the log-likelihood as $\sum_{c=1}^{m} \ell n f(\mathbf{x}_c|\mathbf{p},k)$.

However we do not have an independent identically distributed sample: the probability of selection of each cluster is π_c (the π_c are unequal). In addition we adjust this probability by an estimated probability of response p_c, giving an overall probability of inclusion for the cluster of $\pi_c^* = \pi_c p_c$. (π_c^* is provided in the data file from the sampling organization in the school example). p_c will be treated as a fixed value though it is an estimate, as is commonly done.

The approach in this paper to adjusting for differing π_c^*'s is to use pseudo maximum likelihood, i.e., we use a weighted log-likelihood $\sum_{c=1}^{m} w_c \ell n f(\mathbf{x}_c|\boldsymbol{\lambda},k)$, where w_c is proportional to π_c^{*-1}. There are a variety of ways of justifying this: for example the weighted sample log-likelihood is an asymptotically design unbiased estimator of the full population log likelihood (Kish and Frankel [1974], Godambe and Thompson [1986]). One can also justify p-weighting from a more model-based viewpoint: the reader is referred to Rizzo [1992], section 3.4.4 of Skinner, Holt, and Smith [1989], and Smith [1989]. Rizzo [1992] also discusses an alternative to p-weighting of adding π_c^* as a covariate to the model.

3. The Observed Data Pseudo Log-Likelihood

The full data pseudo log likelihood is given in Section 2 is $\sum_{c=1}^{m} w_c \; \ell n \; f(\mathbf{x}_c|\boldsymbol{\lambda},k)$. We deal with missing clusters through weighting alone, i.e., by adjusting w_c to take into account an estimated probability of response for the responding clusters. There also is non-response within the clusters: using Rubin and Little's notation let $\mathbf{x}_{c,obs}$ be the observed data in cluster c. Then our observed data pseudo log likelihood is $\sum_{c=1}^{m} w_c \ell n f(\mathbf{x}_{c,obs}|\boldsymbol{\lambda},k)$. To illustrate in this case we can look at one of the clusters from the NELS survey with one margin sex (S0=Male, S1=Female), the other margin the importance of future success to the student (I0=Not very important, I1=very important).

	I0	I1	
S0	1	16	0
S1	2	23	1
	1	2	

The 2×2 table within the cluster was as follows: The 2×2 table corresponds to complete observations, the two margins to partially missing observations: we had one student who gave her sex, but did not give a success answer, and 3 students who gave success answers, but not their sexes.

For the complete data the probability is

$$f(\mathbf{x}_c|\boldsymbol{\lambda},k) = \binom{42}{1\;2\;16\;23} \frac{\Gamma(k)}{\Gamma(42+k)\prod_{j=1}^{4}\Gamma(kp_j)} \times \Gamma(1+kp_1)\Gamma(2+kp_2)\Gamma(16+kp_3)\Gamma(23+kp_4)$$

For the missing data margins define $\mathbf{x}_c^{(1)} = (x_{c1}^{(1)}, x_{c2}^{(1)}) = (0,1)$ and $\mathbf{x}_c^{(2)} = (x_{c1}^{(2)}, x_{c2}^{(2)}) = (1,2)$. The superscript indicates missingness pattern. Let $n_c^{(1)} = 1$, $n_c^{(2)} = 3$ be the totals of each cluster falling into each missingness pattern. Since we assume the data is missing at random, the $n_c^{(j)}$'s as random variables have distributions independent of k and $\mathbf{p}$. Thus we will condition on the $n_c^{(j)}$'s as ancillary, and consider them as fixed constants.

It is clear that conditional on $\mathbf{p}_c$, k, and $n_c^{(1)}$ that $x_{c1}^{(1)}$ is $Bin(n_c^{(1)}, p_{c1}^{(1)})$, where $p_{c1}^{(1)} = p_{c1} + p_{c3}$, the sum of the probabilities of the cells corresponding to

$x_{c1}^{(1)}$'s margin. Likewise $x_{c1}^{(2)}$ is $Bin(n_c^{(2)}, p_{c2}^{(2)})$, where $p_{c1}^{(2)} = p_{c1} + p_{c2}$. What are the distributions of $(p_{c1}^{(1)}, p_{c2}^{(1)})$, $(p_{c1}^{(2)}, p_{c2}^{(2)})$? We know the p_{ci}'s are Dirichlet with parameters $(k, p_1, \ldots, p_k)$. The $p_{ci}^{(j)}$'s are summations of particular p_{cj}'s, and these summations must also have Dirichlet distributions. For example $(p_{c1}^{(1)}, p_{c2}^{(1)}) = (p_{c1} + p_{c3}, p_{c2} + p_{c4})$ has Dirichlet distribution with parameters $k, (p_1 + p_3, p_2 + p_4)$. (See Wilks [1962], p. 181). This convenient property makes the Dirichlet desirable, if it is an adequate approximation of truth.

It follows that $\mathbf{x}_c^{(1)} = (x_{c1}^{(1)}, x_{c2}^{(1)})$ will be Dirichlet multinomial with parameters $n_c^{(1)}$, k, $(p_1^{(1)}, p_2^{(1)})$, where $p_1^{(1)} = p_1 + p_3, p_2^{(1)} = p_2 + p_4$:

$$\begin{aligned} f(x_{c1}^{(1)}, x_{c2}^{(1)} | \boldsymbol{\lambda}, k, n_c^{(1)} = 1) \\ = \frac{\Gamma(k)}{\Gamma(1+k)\prod_{j=1}^{2}\Gamma(kp_j^{(1)})} \\ \times \Gamma(0 + kp_1^{(1)})\Gamma(1 + kp_2^{(1)}) \end{aligned}$$

$$\begin{aligned} f(x_{c1}^{(2)}, x_{c2}^{(2)} | \boldsymbol{\lambda}, k, n_c^{(2)} = 3) \\ = 3\frac{\Gamma(k)}{\Gamma(3+k)\prod_{j=1}^{2}\Gamma(kp_j^{(2)})} \\ \times \Gamma(1 + kp_1^{(2)})\Gamma(2 + kp_2^{(2)}) \end{aligned}$$

The overall probability of $\mathbf{x}_{c,obs}$, the observed data given in the table above, is a product of the complete data likelihood and the two likelihoods given above. In general, suppose we have L missingness patterns possible, in each cluster. Then the overall pseudo log-likelihood of the observed data in the clusters will be

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}, k) &= \sum_{m=1}^{c} w_c \sum_{\ell=0}^{L} \mathcal{L}_c^{(\ell)}(\boldsymbol{\lambda}, k) \\ &= \sum_{\ell=0}^{c} w_c \sum_{\ell=0}^{L} \Bigg[\log\left(n_c^{(\ell)} \mathbf{x}_c^{(\ell)} \right) + \log\Gamma(k) \\ &\quad - \log\Gamma(n_c^{(\ell)} + k) - \sum_{j=1}^{r_\ell} \log\Gamma(kp_j^{(\ell)}) \\ &\quad + \sum_{j=1}^{r_\ell} \log\Gamma(x_{cj}^{(\ell)} + kp_j^{(\ell)}) \Bigg] \end{aligned} \tag{3.1}$$

$\mathcal{L}_c^{(\ell)}$ is the likelihood function for the data from one cluster c and missingness pattern ℓ. The case $\ell = 0$ corresponds to the complete data in each cluster, thus $\mathbf{p}^{(0)}$ is just $\mathbf{p}$, $\mathbf{x}_c^{(0)}$ is the vector of complete data counts for the cluster, etc. r_ℓ is the number of cells in the margin corresponding to the missingness pattern: in our example $r_1 = r_2 = 2$, and $r_0 = 4$. If $n_c^{(\ell)}$ is zero for any c, ℓ then $\mathcal{L}_c^{(\ell)}(\boldsymbol{\lambda}, k)$ will be zero rather than the formula as given above (this is presumed in the formula for $\mathcal{L}(\boldsymbol{\lambda}, k)$ to avoid clutter).

The $\mathbf{p}_j^{(\ell)}$'s are obviously straightforward linear functions of $\boldsymbol{\lambda}$ (see section 2). As a function of β, each $\lambda_i = e^{\mathbf{x}_i\beta}/1 + \sum_{i=1}^{r-1} e^{\mathbf{x}_i\beta}$, from the model, as is standard in log-linear models, and the $\mathcal{L}$ can be written as a function of β based on (3.1).

4. Finding the Maximum Likelihood Estimators

Before discussing the maximum likelihood estimators, an initial discussion of why the EM algorithm was not used here is probably necessary. If we used the EM algorithm we would maximize an expectation of the complete data log-likelihood over the distribution of the missing data given the observed data. In this situation the observed data $\mathcal{L}$ is a product of Dirichlet multinomials, thus it is not more complicated in form than the complete data log-likelihood. Thus the M step in the EM algorithm is no easier than a direct maximization of $\mathcal{L}$. In addition there is no simple formula for the expectation of the complete data log-likelihood given the missing data, making the E step somewhat computationally intensive. In short, the EM algorithm in this situation was much less efficient computationally, and no simpler conceptually.

To get the maximum likelihood methods we need the score function and higher-order derivatives of $\mathcal{L}$ with respect to k and β. The derivatives with respect to k are easy to obtain. Let $\psi(x) = \frac{\partial}{\partial x}\log\Gamma(x)$. (The ψ function and its derivatives are given in a number of software packages, e.g., NAG Fortran Library). Then

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial k} = \sum_{m=1}^{c}\sum_{\ell=0}^{\mathcal{L}} \Bigg[\psi(k) - \psi(n_c^{(\ell)} + k) \\ - \sum_{j=1}^{r^{(\ell)}} p_j^{(\ell)} \psi(kp_j^{(\ell)}) \\ + \sum_{j=1}^{r^{(\ell)}} p_j^{(\ell)} \psi(x_{c\ell}^{(\ell)} + kp_j^{(\ell)}) \Bigg] \end{aligned} \tag{4.1}$$

$$\begin{aligned} \frac{\partial^2\mathcal{L}}{\partial k^2} = \sum_{m=1}^{c}\sum_{\ell=0}^{\mathcal{L}} \Bigg[\psi'(k) - \psi'(n_c^{(\ell)} + k) \\ - \sum_{j=1}^{r^{(\ell)}} [p_j^{(\ell)}]^2 \psi'(kp_j^{(\ell)}) \\ + \sum_{j=1}^{r^{(\ell)}} [p_j^{(\ell)}]^2 \; \psi'(x_{c\ell}^{(\ell)} + kp_j^{(\ell)}) \Bigg] \end{aligned} \tag{4.2}$$

The next step is to get the first and second partial derivatives of $\mathcal{L}$ with respect to β, the main parameter of interest. To facilitate this matrix derivative theory will be used. A major reference for this theory is Magnus and Neudecker [1988]. We will briefly discuss what is necessary in developing our estimators.

Suppose we have a matrix function F mapping $n \times q$ matrix X into an $m \times p$ matrix $F(X)$. Then the matrix derivatives DF is an $mp \times nq$ matrix for which the $(i,j)^{th}$ element of DF is the partial derivative of the i^{th} element of vec $F(X)$ with respect to the j^{th} element of vec (X). Two properties of matrix derivatives proven in Magnus and Neudecker [1988] used below are a chain rule and product rule: if $F(X) = F_1(F_2(X))$, then $DF = (DF_1) * (DF_2)$, and if $G(X) = AXB$, where A and B are constant matrices, then $DG = B' \otimes A$. Simpler versions of the product rule for vectors can be obtained by replacing A or B with an appropriate identity matrix.

We will proceed by finding $\frac{\partial \mathcal{L}_c^{(\ell)}}{\partial \boldsymbol{\beta}}$ for each ℓ and c. Let $\boldsymbol{\lambda}^{(\ell)} = (p_1^{(\ell)}, \dots, p_{r_\ell - 1}^{(\ell)})$, i.e., the margin cell probabilities except the last: $p_{r_\ell}^{(\ell)} = 1 - p_1^{(\ell)} - \cdots - p_{\ell-1}^{(\ell)}$. Then $\boldsymbol{\lambda}^{(\ell)} = Z^{(\ell)}\boldsymbol{\lambda}$, where $Z^{(\ell)}$ is a $(r_\ell - 1) \times (r-1)$ matrix of 1's and 0's. Note that $Z^{(0)}$ is an identity matrix. Let $\mathcal{L}_c^{(\ell)} = F_1(F_2(\boldsymbol{\lambda}))$, where $F_2(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{(\ell)}$. Then from (3.1), DF_1 is a $(1 \times r_\ell - 1)$ vector with j^{th} element $-k\psi(kp_j^{(\ell)}) + k\psi(x_{cj}^{(\ell)} + kp_j^{(\ell)}) + k\psi(kp_{r_\ell}^{(\ell)}) - k\psi(x_{cr_\ell}^{(\ell)} + kp_{r_\ell}^{(\ell)})$, i.e., $\frac{\partial \mathcal{L}_c^{(\ell)}}{\partial \boldsymbol{\lambda}^{(\ell)}}$ as a row vector. DF_2 is just $Z^{(\ell)}$ from the product rule.

Now let $\boldsymbol{\lambda} = F_3(\boldsymbol{\eta})$, where $\boldsymbol{\eta} = X\boldsymbol{\beta}$. Then $\lambda_i = e^{\eta_i}/1 + \sum e^{\eta_i}$ from Section 2, and we obtain $DF_3 = \Delta_\lambda - \boldsymbol{\lambda}\boldsymbol{\lambda}'$ (differentiating with respect to η), where Δ_λ is a diagonal matrix of λ_i's, from direct computation of the partials. Finally $\boldsymbol{\eta} = F_4(\boldsymbol{\beta}) = X\boldsymbol{\beta}$, and $DF_4 = X$. We then have $\mathcal{L}_c^{(\ell)} = F(\beta) = F_1(F_2(F_3(F_4(\beta))))$, with

$$DF = \left[\frac{\partial \mathcal{L}_c^{(\ell)}}{\partial \boldsymbol{\lambda}^{(\ell)}}\right]' Z^{(\ell)}[\Delta_\lambda - \boldsymbol{\lambda}\boldsymbol{\lambda}']X \tag{4.3}$$

from the chain rule.

To get the second derivative of $\mathcal{L}_c^{(\ell)}$ with respect to $\boldsymbol{\beta}$ we just differentiate DF above as a column vector (DF is a row vector in (4.3)). $Z^{(\ell)}$ and X are both constant, so as a function of $\boldsymbol{\beta}$, DF' is $G(\boldsymbol{\beta}) = X'[G_1(\boldsymbol{\beta})]Z^{(\ell)'}[G_2(\boldsymbol{\beta})]$ where $G_1(\boldsymbol{\beta}) = \Delta_\lambda - \boldsymbol{\lambda}\boldsymbol{\lambda}'$, $G_2(\boldsymbol{\beta}) = \frac{\partial \mathcal{L}_c^{(\ell)}}{\partial \boldsymbol{\lambda}^{(\ell)}}$. To differentiate G we have a essentially a product of two functions. For matrix derivatives this derivative operates in the same way as a simpler scalar function product $(f_1(x)f_2(x))' = f_1'(x)f_2(x) + f_1(x)f_2'(x)$. Combining this with the product rule above we get $DG = \left[\left[\frac{\partial \mathcal{L}_c^{(\ell)}}{\partial \boldsymbol{\lambda}^{(\ell)}}\right]' Z^{(\ell)} \otimes X'\right] DG_1 + [X'G_1(\beta)Z^{(\ell)'}]DG_2$.

DG_2 can be derived in a manner similar to DG using the chain rule: we obtain $DG_2 = [\frac{\partial^2 \mathcal{L}_c^{(\ell)}}{\partial \boldsymbol{\lambda}^{(\ell)} \partial \boldsymbol{\lambda}^{(\ell)'}}] \times Z^{(\ell)}[\Delta_\lambda - \boldsymbol{\lambda}\boldsymbol{\lambda}']X$. The $(r_e - 1) \times (r_e - 1)$ second derivative matrix $\frac{\partial^2 \mathcal{L}_c^{(\ell)}}{\partial \boldsymbol{\lambda}^{(\ell)} \partial \boldsymbol{\lambda}^{(\ell)T}}$ has as diagonal elements $-k^2\psi'(kp_i^{(\ell)}) + k^2\psi'(y_{ij} + kp_i^{(\ell)}) - k^2\psi'(kp_{r_\ell}^{(\ell)}) + k^2\psi'(y_{ir_\ell} + kp_{r_e}^{(\ell)})$, and off-diagonal elements $-k^2\psi' \times (kp_{r\ell}^{(\ell)}) + k^2\psi'(y_{ir_\ell} + kp_{r_\ell}^{(\ell)})$. (See 3.1 and the discussion of $\frac{\partial \mathcal{L}_c^{(\ell)}}{\partial \boldsymbol{\beta}}$).

For DG_1 we need to differentiate $\Delta_\lambda - \boldsymbol{\lambda}\boldsymbol{\lambda}'$ as a function of $\boldsymbol{\beta}$. We can write $\Delta_\lambda - \boldsymbol{\lambda}\boldsymbol{\lambda}'$ as $H_2(H_1(\boldsymbol{\beta})) - H_3(H_1(\boldsymbol{\beta}))$, where $H_1(\boldsymbol{\beta}) = \boldsymbol{\lambda}$. We know from work preceding (4.3) that $DH_1 = [\Delta_\lambda - \boldsymbol{\lambda}\boldsymbol{\lambda}']X$. H_2 maps $\boldsymbol{\lambda}$ into Δ_λ; if for example $\boldsymbol{\lambda}$ is 3×1 DH_2 will be the 9×3 matrix $\begin{bmatrix} 100 \cdot 000 \cdot 000 \\ 000 \cdot 010 \cdot 000 \\ 000 \cdot 000 \cdot 001 \end{bmatrix}'$, obtained by direct partial differentiation. H_3 maps $\boldsymbol{\lambda}$ into $\boldsymbol{\lambda}\boldsymbol{\lambda}'$. From Magnus and Neudecker [1988], p. 182, if $F(X) = XX'$, where X is $n \times q$, then $DG = (I_{n^2} + K_{n,n})(X \otimes I_n)$, where $K_{n,n}$ is the matrix that satisfies $K_{n,n}[vec(Z)] = \vec{(}Z^T)$ for any $n \times n$ matrix Z. Thus $DH_3 = (I_{(r_e-1)^2} + K_{r_e-1,r_e-1})(\boldsymbol{\lambda} \otimes I_{r_e-1})$, and the second derivative is

$$\begin{aligned} DG = {} & X'[\Delta_\lambda - \boldsymbol{\lambda}\boldsymbol{\lambda}']Z^{(\ell)'}\left[\frac{\partial^2 \mathcal{L}_c^{(\ell)}}{\partial \boldsymbol{\lambda}^{(\ell)} \partial \boldsymbol{\lambda}^{(\ell)'}}\right] \\ & \times Z^{(\ell)}[\Delta_\lambda - \boldsymbol{\lambda}\boldsymbol{\lambda}']X \\ & + \left[[\frac{\partial \mathcal{L}_c^{(\ell)}}{\partial \boldsymbol{\lambda}^{(\ell)}}]' Z^{(j)} \otimes X'\right] \\ & \times \{DH_2 + [DH_3]\}[\Delta_\lambda - \boldsymbol{\lambda}\boldsymbol{\lambda}']X] \end{aligned} \tag{4.4}$$

Obtaining these second partials is do-able at least in theory without the matrix derivative theory, but it is far less pleasant–you need to do them one by one.

The derivative $\frac{\partial^2 \mathcal{L}}{\partial K \partial \beta}$ has a derivation very similar to the above work and will not be given. To actually find the maximum likelihood estimators of k and $\boldsymbol{\beta}$ a Newton-Raphson method is used. Generally getting the expectation of the second derivative matrix is computationally inefficient, thus using the observed Fisher information matrix is much better if it is positive definite.

5. Properties of the Pseudo Maximum Likelihood Estimator

Let γ be $(k, \boldsymbol{\beta})$, $\hat{\gamma}$ the pseudo MLE, and $\mathcal{L}_c = \sum_{e=0}^{\mathcal{L}} \mathcal{L}_c^{(\ell)}$. The "pseudo" score function can be written as $\sum_{c=1}^m w_c \frac{\partial \mathcal{L}_c}{\partial \gamma}$, with $\frac{\partial \mathcal{L}_c}{\partial \gamma}$ derived in section 4. The second derivative of $\mathcal{L}_c$ will be $\sum_{c=1}^m w_c \frac{\partial^2 \mathcal{L}_c}{\partial \gamma \partial \gamma'}$. To derive the asymptotic variance of $\hat{\gamma}$, we use the usual expansion

$$-\sum_{c=1}^m w_c \frac{\partial \mathcal{L}_c}{\partial \gamma}\bigg|_\gamma \cong \sum_{c=1}^m w_c \left[\frac{\partial^2 \mathcal{L}_c}{\partial \gamma \partial \gamma'}\bigg|_\gamma (\hat{\gamma} - \gamma) \right. \tag{5.1}$$

The variance of the pseudo score function is $-\sum_{c=1}^{m} w_c^2 \mathcal{E}\left[\frac{\partial^2 \mathcal{L}_c}{\partial\gamma\partial\gamma'}\right]$, since the $\frac{\partial\mathcal{L}_c}{\partial\gamma}$ are ordinary score functions and are independent. The asymptotic variance of $\hat{\gamma}$ will be

$$\text{Var}(\hat{\gamma}) \sim -\left[\sum_{c=1}^{m} w_c \mathcal{E}[\frac{\partial^2 \mathcal{L}_c}{\partial\gamma\partial\gamma'}]\right]^{-1} \times \sum_{c=1}^{m} w_c^2 \mathcal{E}[\frac{\partial^2 \mathcal{L}_c}{\partial\gamma\partial\gamma'}]\left[\sum_{c=1}^{m} w_c \mathcal{E}[\frac{\partial^2 \mathcal{L}_c}{\partial\gamma\partial\gamma'}]\right]^{-1} \qquad (5.2)$$

The asymptotic framework that is most workable presupposes that m, the number of clusters, goes to infinity, while the n_c, $n_c^{(\ell)}$ remain bounded (by an M, for example, which is not dependent on m). A rigorous proof of asymptotic normality will not be given here, but an outline of the proof will be described.

The critical element is proving the asymptotic normality of the pseudo score function $\sum_{c=1}^{m} w_c \frac{\partial\mathcal{L}_c}{\partial\gamma}\Big|_{\gamma}$. We need to use a Lindeberg-Feller type argument: we would require for example that $a/n \leq w_c \leq b/n$ for $a, b > 0$, so that no w_c is "large," violating uniform asymptotic negligibility. In practice a w_c can be large relative to the others if the probability of selection of a cluster is very small - remember that w_c is the inverse probability of inclusion. Existence of very small probabilities of selection will cause deviance from the normal distribution even with large m. If n_c is bounded, then $\frac{\partial\mathcal{L}_c}{\partial\gamma}$ has finite support, therefore it is sufficient to show it is bounded above to bound in any moments. To bound in the first derivative and higher order derivatives in a neighborhood of γ (see Serfling [1980], p. 144) we require that k be strictly positive and $\mathbf{p}_c$ contains no zeroes. The ψ function and all derivatives thereof can be shown to be bounded on intervals of the form $[\eta_1, \eta_2]$ where $\eta_1 > 0$: thus if its arguments in (4.1), etc. are kept away from zero the overall derivatives will be bounded above for any order. The partial derivatives of $\mathcal{L}_c$ of any order with respect to γ can easily be seen to be linear combinations of the ψ function quantities and their derivatives. Variance matrices need to be positive definite and second derivative matrices negative definite. With these kinds of conditions it is not difficult to construct a proof of asymptotic normality of $\hat{\gamma}$ along the lines of Serfling [1980], pp. 144–148.

6. Results for the 2 x 2 Example

The model of interest is the independence model - where sex and views toward future success are independent. This model as well as the saturated model was fit and compared. Using the starting values from Chapter 6, both model iterations converged (a small number of iterations were necessary (< 10): a few seconds on the SUN Sparc stations). For the saturated model the maximum likelihood estimator for k has 553 and the maximum likelihood estimators for the cell probabilities were:

		Success	
		NI	VI
Sex	M	.069	.420
	F	.087	.424

To test independence a Wald Test was used: in the saturated model we have parameters $(\beta_1, \beta_2, \beta_3)$ with $p_1 = \frac{e^{\beta_1}}{1+\sum_{i=1}^{3} e^{\beta_i}}$, etc. Under the independence model $\beta_1 = \beta_2 + \beta_3$, thus we can test $a'\beta = 0$, where $a = [1, -1, -1]$, using $(a'\hat{\beta})^2/\text{Var}(a'\hat{\beta})$, $(\text{Var}(\hat{\beta})$ is a submatrix of (5.2) above), which we refer to a χ_1^2 distribution. The value of the test statistic was .448, with a p-value of $\sim$ .5. Thus we accept the independence model, sex being independent of views on future success is consistent with the data.

For this model the estimator of k is 54.5, with estimators for the cell probabilities of success, i.e., we have a ratio of 1.13 of girls vs. boys, and a ratio of 5.17 of those viewing success as important against those not viewing it as not very important. The correlation between the two β parameters in the asymptotic variance matrix was only $-.05$, thus we can for simplicity assume independence and set up two standard deviation confidence intervals for these ratios: for girls vs. boys the C.I. for the ratio is [.977, 1.305], and for VI's vs. NI's the C.F. for the ratio is [4.20, 6.31]. (Note: the true ratio of girls vs. boys is not necessarily 1 because of variable drop-out rates, etc.). The asymptotic standard deviation for k was 24.3, reflecting a great deal more variability for what is essentially a kind of variance component.

		Success	
		NI	VI
Sex	M	.076	.393
	F	.086	.444

Suppose we look at the distribution of $\boldsymbol{\lambda}_c$ under the model using the maximum likelihood estimators under the independence model. Then the marginal distribution of λ_{1c} is Beta $(kp_1, k(1-p_1))$ and that of λ_{3c} is Beta $(kp_3, k(1-p_3))$ from basic results regarding the Dirichlet (Wilks [1977]).

7. Discussion and Extensions

In arguing for this methodology there are three questions that need to be discussed: (1) Why use a model-based approach? (2) Why use this particular model? (3) With this model why use maximum likelihood as an estimation procedure? Question (1) is pertinent because of the extensive development of a design-based approach to categorical data models. There is a strong argument for the value of model-free inference especially in situations in which many different users with many different presuppositions (i.e., models) will be using and critiqueing the results. This approach by using pseudo-likelihood methods does allow one to say the sample pseudo-log likelihood is an asymptotically design unbiased estimator of the true population log-likelihood, but in general the basis for inference is the model distribution, not the sample design, thus in general design effects are not directly relevant to this type of inference (we do care about what Skinner, Holt, and Smith [1987] call "model effects" (Chapter 2)). On the positive side of the ledger for a more model-based approach, the models allow for fuller "exploration" of the data, as argued by Holt and Ewings [1987], they allow for an easier dealing with non-response, as seen in this paper, and on a practical side do not require knowledge of design effects for cell probabilities (which are not available for this survey).

The second question we pose is assuming we want a more model-based approach why use the Dirichlet multinomial model? The author feels that it is easier to proceed by dealing with each cluster as an independent multinomial sample rather than attempting to throw everything together, and take account of intracluster correlation by defining covariance components, as is done in most of the model-based papers. This approach does not allow one to define a full likelihood easily, and gets very messy when one tries to deal with differing cluster sizes, and non-response. Another approach is to use a normal prior for β. This is harder to work with, requiring use of an EM mechanism, and in general is less efficient computationally. A product multinomial, with a Dirichlet "prior" on the cluster parameters, allows a very elegant likelihood approach to non-response differing cluster sizes, and asymptotic theory.

The third question is why use maximum likelihood? Brier [1980] in the original paper uses method of moments. However Brier does not cover either differing cluster sizes or non-response. Method of moments is very messy when one tries to generalize to these situations. However with a technique like Magnus and Neudecker's matrix derivatives finding maximum likelihood estimates in the most general situation is fairly straightforward, and assuming we have no sparse cells or disproportionately large clusters the asymptotic efficiency of the direct maximum likelihood estimators can be established (we lose some of this efficiency by using a pseudo-likelihood with the weights based on inverse probabilities of selection).

References

Altham, P. (1976), "Discrete Variable Analysis for Individuals Grouped Into Families," *Biometrika* **63**, 263–269.

Beitler, P. and Landis, J. (1985), "A Mixed Effects Model for Categorical Data," *Biometrics* **41**, 991–1000.

Brier, S. (1980), "Analysis of Contingency Tables Under Cluster Sampling," *Biometrika* **67**, 591–596.

Chambers, R. L. (1986), "Design-Adjusted Parameter Estimation," *J. Roy. Stat. Soc.* A **149**, 161–173.

Choi, J. and McHugh, R. (1989), "A Reduction Factor in Goodness-of-Fit and Independence Tests for Clustered and Weighted Observations," *Biometrics* **45**, 979–996.

Cohen, J. (1976), "The Distribution of the X^2 Statistic Under Clustered Sampling from Contingency Tables," *J. Amer. Stat. Assoc.* **71**, 665–670.

Godambe, V. P. and Thompson M. R. (1986), "Parameters of Superpopulation and Survey Populations: Their Relationships and Estimation," *Inter. Stat. Rev.* **54, 2**, 127–138.

Harville, D. and Mees, R. (1984), "A Mixed-Model Procedure for Analyzing Ordered Categorical Data," *Biometrics* **40**, 393–408.

Holt, D. and Ewings, P. D. (1987), "Logistic Models for Contingency Tables" in *Analysis of Complex Surveys*, eds. C. Skinner, D. Holt and T. M. F. Smith; John Wiley & Sons.

Kish, L. and Frankel, M. R. (1974), "Inference from Complex Samples," *J. Roy. Stat. Soc.* **B, 36**, 1–37.

Little, R. J. A. and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Mak, T. K. (1988), "Analysing Intraclass Correlation for Dichotomous Variables," *Appl. Stat.* **37**, 344–352.

Magnus, J. E. and Neudecker, H. (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, New York.

Pfeffermann, D. (1988), "The Effect of Sampling Design and Response Mechanisms on Multivariate Regression-Based Predictors," *J. Amer. Stat. Assoc.* **83**, 824–833.

Rao , J. N. K. and Thomas, D. R. (1987), "Chi-Squared Tests for Contingency Tables," in *Analysis of Complex Surveys*, eds, C. Skinner, D. Holt and T. M. F. Smith; John Wiley & Sons, New York.

Rizzo, L. (1992), "Conditionally Consistent Estimators Using Only Probabilities of Selection in Complex Sample Surveys," *J. Amer. Stat. Assoc.* , to appear Dec. 1992.

Serfling, R. (1980), *Asymptotic Theorems of Mathematical Statistics*, John Wiley & Sons.

Skinner, B. J., Holt, D. and Smith, T. M. F., *Analysis of Complex Surveys*, John Wiley & Sons.

Smith, T. M. F. (1988), "To Weight or Not to Weight, That is the Question," in *Bayesian Statistics 3*, eds. I. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, Elsevier Science Publishers (North- Holland).

Sugden, R. A. and Smith, T. M. F. (1984), "Ignorable and Informative Designs in Survey Sampling Inference," *Biometrika* **71**, 495–506.

Wilks, S. (1977), *Mathematical Statistics*, John Wiley & Sons.