KM 15(1984)
pag 28 - 44

ANOTA: ANALYSIS OF TABLES

by

Wouter J. Keller*) and Albert Verbeek**)

Abstract

This paper discusses a quick-and-easy method for the analysis of contingency tables with one categorical variable, Y, to be explained or predicted, and the others, $X_1,\ldots,X_M$, explaining or predicting. The method seeks to optimize ease of interpretation and ease of computation rather than comprehensiveness and mathematical sophistication. The results of ANOTA are as easy to interpret as the results of multiple regression. ANOTA directly translates the bivariate tables $Y \times X_m$ into tables of the same size of regression coefficients, where the effect of category j of X_m on category i of Y is standardized for the effects of the other X_ℓ, ($\ell \neq m$). Standard errors are also provided, allowing to judge the accuracy of the regression coefficients. The computational requirements are so limited that ANOTA may be run on a small micro-computer.

*) Netherlands Central Bureau of Statistics and Department of Actuarial Sciences and Econometrics, Vrije Universiteit Amsterdam.
**) Netherlands Central Bureau of Statistics and Social Science Faculty, University of Utrecht.

Correspondence to: CBS, Department for Statistical Methods, P.O. Box 959, 2270 AZ VOORBURG. Tel. 070-694341. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the Netherlands Central Bureau of Statistics.

1. Introduction

ANOTA deals with the contingency table analogue of multiple regression analysis. There is a categorical Y-variable to be explained and one or more explaining variables $X_1, X_2, \ldots, X_M$, which are also categorical. Existing models such as logit- and probit-models are not quite satisfactory for Y-variables with more than two categories, the logit- or probit-transform hampers the interpretation of the linear parameters, and the computational requirements are substantial. In these respects ANOTA is far superior. Of course this superiority has its price. The ANOTA-model may exhibit lack-of-fit (but which model may not?), the estimated proportions may be out of the 0-1-range, and we have also sacrified some statistical efficiency (variance of the estimators) to a gain in ease of use.

This paper is composed as follows. Section 2 gives an example. Section 3 will introduce the notation and set the stage. In section 4 the formal ANOTA-model is introduced, while section 5 presents the estimation method and algorithm, and proposes an estimator for the variance-covariance matrix of the regression coefficients. Section 6 concludes with some discussion.

For the ease of interpretation of ANOTA we refer to section 2. The computational requirements are so limited that ANOTA may be run on small microcomputers. An APL-prototype is currently used for experimenting and a compact BASIC program is being planned.

2. An example

To provide the reader with some feeling of what ANOTA does, this section treats a very simple example. In a CBS household survey, called the Life Situation Survey 1977 (see CBS, 1978), questions were asked, among others, about Satisfaction (S), Income (C) and Education (E). The three relevant bivariate frequency tables are displayed in tables 1, 2, and 3, where also the categories are described. The relevant sample size is 4108 (all figures are unweighted sample frequencies; we will assume that the sample design is simple random sampling).

Table 1. Satisfaction (S) by Income (C)

Satisfaction	Income[a]				Total
	<21	21-40	>40	unknown	
not too satisfied	132	78	13	41	264
rather satisfied	208	198	46	87	539
satisfied	631	773	192	261	1 857
very satisfied	282	485	152	169	1 088
extremely satisfied	103	155	51	51	360
Total	1 356	1 689	454	609	4 108

a) Dfl. 1 000 per annum, 1977

Table 2. Satisfaction (S) by Education (E)

Satisfaction	Education				Total
	low	medium	high	unknown	
not too satisfied	175	54	22	13	264
rather satisfied	304	140	59	36	539
satisfied	1 159	452	169	77	1 857
very satisfied	632	291	115	50	1 088
extremely satisfied	222	90	36	12	360
Total	2 492	1 027	401	188	4 108

Table 3. Income (C) by Education (E)

Income[a]	Education				Total
	low	medium	high	unknown	
<21	1 037	196	59	64	1 356
21-40	912	546	154	77	1 689
>40	146	152	133	23	454
unknown	397	133	55	24	609
Total	2 492	1 027	401	188	4 108

a) Dfl. 1 000 per annum, 1977

Tabel 4. Satisfaction (S) by Income (C) and Education (E) as deviations from average, in %
(standard errors in parentheses)

Satisfaction	Average	Income[a]				Education			
		<21	21-40	>40	unknown	low	medium	high	unknown
not too satisfied	6.4	3.3	-1.8	-3.6	0.3	0.6	-1.2	-0.9	0.5
	(0.4)	(0.6)	(0.4)	(0.8)	(0.9)	(0.3)	(0.6)	(1.1)	(1.8)
rather satisfied	13.1	2.2	-1.4	-3.0	1.2	-0.9	0.5	1.6	6.0
	(0.5)	(0.8)	(0.6)	(1.4)	(1.3)	(0.4)	(0.9)	(1.7)	(2.8)
satisfied	45.2	1.3	0.6	-2.9	-2.3	1.3	-1.2	-3.1	-4.2
	(0.8)	(1.1)	(0.9)	(2.2)	(1.9)	(0.6)	(1.3)	(2.3)	(3.5)
very satisfied	26.5	-5.7	2.2	7.0	1.3	-1.1	1.9	2.2	0.1
	(0.7)	(0.9)	(0.8)	(2.1)	(1.7)	(0.6)	(1.2)	(2.1)	(3.1)
extremely satisfied	8.8	-1.2	0.4	2.5	-0.4	0.1	0.0	0.2	-2.4
	(0.4)	(0.6)	(0.5)	(1.4)	(1.0)	(0.4)	(0.8)	(1.4)	(1.8)
Total	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

a) Dfl. 1 000 per annum, 1977

Let us consider Satisfaction as dependent and Income and Education as predictor (independent) variables. A good way to represent the sample information with respect to this view is displayed in table 4, where for each category of a predictor variable (C or E) the distribution over the categories of the dependent variable (S) are given as deviations of the average proportions of the categories of S in the sample. From this table we concluded that more Income or more Education will in general increase the chances on a positive Satisfaction score and decrease the chances on a negative score. Besides the average proportions and the deviations of proportions, also their standard errors, based on a multinomial sampling process, are shown.

However, if we look at the distribution of the predictor variables C and E, as displayed in table 5, it will be clear that these two variables are not independent: more Education in general means more Income and vice versa. So, now we are in doubt if the higher Satisfaction scores for higher Education is caused by Education itself or by Income, in view of the relatively higher Income in the categories of higher Education. This question is completely analogous to the problem behind multiple regression models.

ANOTA supplies us with a simple answer: see table 6. In this table, the effects of the predictor variables on the dependent variable are displayed in the same way as in table 5, but now corrected for the interdependencies between the predictor variables. To be more precise, the effect of, say, Education on Satisfaction is computed as if the Income distribution per Education category is the same as the average distribution in the sample. In other words, it is the net effect of Education on Satisfaction under constancy of Income; or it is the effect of Education after removal of the Income-effect. The interpretation is exactly the same as the interpretation of regression coefficients in multiple regression analysis (with dummy variables) or as the interpretation of effects in analysis of variance.

Looking at table 6, we see that, after correcting for the interdependencies between Income and Education, the effect of Education on Satisfaction is changed in sign with respect to table 4: now more Education means less Satisfaction. The positive effect of Income on Satisfaction is accentuated in the ANOTA result. Note that we may read the ANOTA table two ways: in one column the 'standardized' distribution, expressed as a devia-

tion from the average, can be read, while one row gives the regression coefficients explaining the proportion in that row as a function of the predictor variables.

Table 5. Income (C) by Education (E) as deviations from average, in %

Income[a)]	Average	Education			
		low	medium	high	unknown
<21	33.0	8.6	-13.9	-18.3	1.0
21-40	41.1	-4.5	12.0	- 2.7	-0.2
>40	11.1	-5.2	3.7	22.1	1.2
unknown	14.8	1.1	-1.9	- 1.1	-2.1
Total	100.0	0.0	0.0	0.0	0.0

a) Dfl. 1 000 per annum, 1977

3. Notation

Since matrix algebra eases the notation and derivation of the regression coefficients considerably, we will mainly use matrix algebra when dealing with theory, but also use scalar notation when it helps the interpretation.

We first need some notation which draws heavily on the theory of the linear model. The N×I-indicator matrix Y contains the scores of the N individuals (cases) in the sample on the I categories of the dependent variable: entry y_{ni} equals '1' if individual n scores in category i of the dependent variable, and '0' else. We assume each individual to score in exactly one category (so $Y\iota=\iota$, with $\iota=(1,1,\ldots,1)'$). The sample frequency distribution of Y is given by the vector $f_Y=(f_1^Y,f_2^Y,\ldots,f_I^Y)'$ of length I. So we have $f_Y=Y'\iota$. (Subsequently, we will use the terms 'frequency' and 'frequency distribution' for counts, and 'proportions' for fractions.)

The scores of the N observations on the m-th predictor variable are collected in an $N\times K_m$ indicator matrix X_m, with the m-th variable having K_m categories (m=1,...,M), and element (n,k) of X_m (i.e. x_{nk}^m) equal to '1' if individual n scores on category k of the m-th predictor variable and '0'

Tabel 6. ANOTA: regression coefficients of Satisfaction (S) by Income (C) and Education (E), in % (standard errors in parentheses)

Satisfaction	Average	Income				Education			
		<21	21-40	>40	unknown	low	medium	high	unknown
not too satisfied	6.4	3.3	-1.8	-3.6	0.3	0.0	-0.4	0.4	0.5
	(0.4)	(0.6)	(0.4)	(0.8)	(0.9)	(0.3)	(0.6)	(1.1)	(1.8)
rather satisfied	13.1	2.7	-1.6	-4.1	1.3	-1.5	1.3	3.0	6.1
	(0.5)	(0.8)	(0.6)	(1.4)	(1.3)	(0.4)	(0.9)	(1.7)	(2.8)
satisfied	45.2	0.9	0.7	-2.0	-2.5	1.2	-1.1	-2.5	-4.3
	(0.8)	(1.1)	(0.9)	(2.3)	(1.9)	(0.7)	(1.4)	(2.4)	(3.5)
very satisfied	26.5	-5.6	2.2	6.9	1.3	-0.2	0.6	-0.3	0.1
	(0.7)	(1.0)	(0.8)	(2.1)	(1.7)	(0.6)	(1.2)	(2.2)	(3.1)
extremely satisfied	8.8	-1.3	0.5	2.8	-0.4	0.4	-0.3	-0.6	-2.4
	(0.4)	(0.6)	(0.5)	(1.4)	(1.0)	(0.4)	(0.8)	(1.4)	(1.8)
Total	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

else. Each individual scores exactly on one category of each predictor variable (so $X_m 1=1$). Besides M true predictor variables, an additional predictor variable (called 'the constant') with index m=0 is introduced, which has only one category (so $K_0=1$) on which everyone scores: $X_0=1$ (or $x^0_{n1}=1$ for all n).

The scores on all M+1 predictor variables are collected in one N×K matrix $X=(X_0,X_1,\ldots,X_M)$ with $K=\Sigma_{m=0}^{M} K_m$. The sample frequency distribution of the m-th predictor variable X_m is given by the vector $f_m=(f^m_1,f^m_2,\ldots,f^m_{K_m})'$ of length K_m. So we have $f_m=X'_m 1$. Concatenating these distribution vectors for all M+1 predictor variables, we get $f=(f'_0,f'_1,\ldots,f'_M)'=X'1$.

With these indicator matrices Y and X spelled out, we are ready to translate tables into matrices. The I×K_m-table of frequencies (= number of individuals) of the dependent variable against the m-th predictor variable can now be written as $Y'X_m$ as can easily be confirmed. Analogously, the K_ℓ×K_m-table X_ℓ×X_m simply becomes $X'_\ell X_m$. The total scores on the dependent variable equal $Y'X_0=Y'1=f_Y$, while the I×K-matrix Y'X contains all the relevant scores on the Y variable in the categories of all the X variables (subsequently, we will use Y, X_m and X to denote either the indicator matrices or the variables themselves). The K×K-matrix X'X contains all the crossings of $(X_0,\ldots,X_M)$ by $(X_0,\ldots,X_M)$. The diagonal matrix $X'_m X_m$ with the vector of frequencies f_m on the diagonal is located as submatrix around the diagonal of X'X.

Finally, we arrive at the proportion of scores on Y for each category in X (see e.g. table 4) by computing the table Y'XW, where W equals the K×K diagonal matrix with the reciprocals of the frequencies in f (=X'1) on the diagonal. Each column of Y'XW contains I proportions, adding up to one, and corresponds to the conditional distribution of Y. Analogously, X'XW represents the supermatrix with outside the diagonal blocks the tables $X'_\ell X_m W_m$, which represent the proportions of scores on X_ℓ for each category of X_m.

Instead of matrix notation we will sometimes use scalar notation for frequencies and proportions. Frequencies will be denoted by f. Subscripts refer to categories, superscripts to variables. E.g., f^Y_i is the frequency in category i of the dependent variable Y, $f^{YX_1}_{ij}$ ($=f^{Y1}_{ij}$ for short) equals the

frequency in cell (i,j) of the table Y by X_1, $f^{X_1X_2}_{jk} = f^{12}_{jk}$ the frequency in cell (j,k) in table $X_1 \times X_2$, etc. Note that by the special nature of X_0 (i.e. the constant), we have $f^{Y0}_{il}=f^{Y}_{i}$ and $f^{0m}_{1j}=f^{m}_{j}$. Proportions, denoted by p, are always with respect to the total score on all but the first variable simultaneously. So $p^{YX_1X_2}_{ijk}=f^{Y12}_{ijk}/f^{12}_{jk}$ ($=p^{Y12}_{ijk}$ for short), while $p^{12}_{jk}=f^{12}_{jk}/f^{2}_{k}$. If only one variable is indicated, the proportion is with respect to the total sample size, i.e. $p^{Y}_{i}=f^{Y}_{i}/N$ (one might notice that $f^{0}_{1}=N$, so $p^{Y0}_{il}=p^{Y}_{i}=f^{Y}_{i}/f^{0}_{1}$). This completes our notation.

4. The model

The model is a direct derivate from the well-known linear model for regression analysis or analysis of variance:

$$\mathbf{E}y = Xb \ ,$$

where y is an N-vector of dependent scores, X an N×K-matrix of predictor scores, b a K-vector of regression coefficients and $\mathbf{E}y$ the expectation of y. We make two amendments. First, we generalize to the multivariate linear model

$$\mathbf{E}Y = XB \tag{4.1}$$

with Y an N×I-matrix, X as above and B correspondingly a K×I-matrix. Second, Y is now a matrix of (dependent) scores (0 or 1) satisfying $Y\iota=\iota$ (each row contains exactly one 1); hence each row of $\mathbf{E}Y$ is a vector of probabilities, adding to 1.

In order to interpret equation (4.1), we consider the scalar representation of an arbitrary element in the i-th column of (4.1). We have

$$p^{YX_1..X_M}_{ij_1..j_M} = b^{Y}_{i} + \sum_{m=1}^{M} b^{YX_m}_{ij_m} \tag{4.2}$$

with $b^{YX_m}_{ij_m}=b^{Ym}_{ij_m}$ equal to the (i,j)-th element of B' with j the column corresponding to the j_m-th category of the predictor variable X_m, and b^{Y}_{i} standing for the 'constant' b^{Y0}_{il}. So, our model says that the cell probabil-

ities are equal to the sum of regression coefficients which depend only on the bivariate indices. For our simple example of section 2, we have in an obvious notation

$$p_{ijk}^{SCE} = b_i^S + b_{ij}^{SC} + b_{ik}^{SE}$$

saying that the probability p_{ijk}^{SCE} of a score on category i of Satisfaction, given the predictor scores j of Income and k of Education, equals a constant depending only on i, b_i^S, plus a regression coefficient b_{ij}^{SC} reflecting the effect of the j-th category of Income on the i-th score of Satisfaction, plus a regression coefficient b_{ik}^{SE} for the effect of $k \in E$ on $i \in S$.

5. The ANOTA estimator for B

To estimate B, we consider the equation X'Y=X'XB, which is equivalent to

$$Y'XW = B'X'XW \tag{5.1}$$

because the diagonal matrix W is invertible (assuming that all categories of all X_m do have observations). It is well-known from OLS theory that X'Y=X'XB is consistent, i.e. B can be solved from it or from (5.1). This will be the ANOTA estimator. Formally this is just the OLS estimator of model (4.1), but that is merely coincidental. Much more important is the following, very simple and interesting interpretation of (5.1): the left hand side corresponds to the set of tables $Y \times X_m$ in column percentages (as table 4); X'XW can be seen as a normalising constant, eliminating the interactions between the predictor variables X_m; and B' has exactly the same size as Y'XW. In fact, each element of B', say b_{ij}^{Ym}, can be interpreted as a normalized version of the corresponding element p_{ij}^{Ym} of Y'XW, eliminating the effects of X_ℓ, $\ell \neq m$. This can be seen as follows: Take one row Y'XW of the left hand side of equation (5.1); this corresponds with a row in table 4. For this row, say the i-th, the scalar representation of y'XW=b'X'XW (see (5.1)) equals

$$p_{ij}^{Ym} = \sum_{\ell=0}^{M} \sum_{k=1}^{K_\ell} b_{ik}^{Y\ell} p_{kj}^{\ell m} . \tag{5.2}$$

Now since $p_{jj}^{mm}=f_{jj}^{mm}/f_j^m=1$, and also $p_{1j}^{Om}=f_{1j}^{Om}/f_j^m=1$, we might write

$$p_{ij}^{Ym} = b_i^Y + b_{ij}^{Ym} + \sum_{\substack{\ell=1 \\ \ell \neq m}}^{M} \sum_{k=1}^{K_\ell} b_{ik}^{Y\ell} p_{kj}^{\ell m} \qquad (5.3)$$

with $b_i^Y = b_{i1}^{YO}$. From this equation we see that our model implies that the true proportion p_{ij}^{Ym} equals the regression constant b_i^Y (which is independent of the choice of the category j of the predictor variable m), plus the regression coefficient b_{ij}^{Ym} corresponding to the 'cell' on the left hand side, plus a weighted sum of the remaining regression coefficients $b_{ij}^{Y\ell}$ with weights, $p_{kj}^{\ell m}$, equal to the sample proportions of X_ℓ conditional on the category j of X_m. In other words, the observed proportion p_{ij}^{Ym} equals a weighted average of the corresponding regression coefficients, with weights reflecting the bivariate distributions between X_m and X_ℓ for $\ell=1,\ldots,M$. For our example of section 2, equation (5.3) becomes with $X_m=E$, in an obvious notation,

$$p_{ij}^{SE} = b_i^S + b_{ij}^{SE} + \sum_{k=1}^{K_C} b_{ik}^{SC} p_{kj}^{CE}$$

showing the effect of the bivariate distribution between the predictor variables, p_{kj}^{CE}, as weight for the regression coefficients b_{ik}^{SC} in the decomposition of p_{ij}^{SE}.

As with ANOVA, some restrictions on the b's are necessary, in order to allow for unique identification of them. Since $X_m{}^1=1$, there are at least M different linear combinations, c, of the columns of X such that Xc=0. So for each i we need at least M additional restrictions on b. We propose

$$\sum_{k=1}^{K_\ell} b_{ik}^{Y\ell} p_k^\ell = 0 \qquad \text{for } \ell=1,\ldots,M;\ i=1,\ldots,I\ , \qquad (5.4)$$

with $p_k^\ell = f_k^\ell / f_1^O = f_k^\ell / N$, for $k=1,\ldots,K_\ell$, the sample proportions of X_ℓ. The interpretation of (5.4) is simple: The average regression coefficient corresponding to a predictor variable X_ℓ (excluding $\ell=0$, the constant) is zero when weighted with the sample proportions of X_ℓ. As a consequence, we have for the constant

$$b_i^Y = p_i^Y \ , \qquad (5.5)$$

as can easily be verified by multiplication of (5.3) by p_j^m and summing over j, using (5.4). In matrix notation we may formulate (5.4) as

$$RB = 0 \ , \qquad (5.6)$$

with R an M×K-matrix with $R=(R_0,R_1,\ldots,R_M)$ and the m-th row of R_m equal to f_m (for m=1,...,M) and zero elsewhere. The restrictions just identify the b's if all vectors $(p_1^\ell,\ldots,p_{K_\ell}^\ell)$ are non-zero and rank X=K-M, cf. Verbeek and Denteneer (1984).

Using the normalization (5.4) we could provide an alternative characterization of b_{ij}^{Ym}: if all predictor variables X_m and X_ℓ (ℓ=1,...,M) are mutually independent, so $p_{kj}^{\ell m}=p_k^\ell$, then we have, using (5.4) and (5.5).

$$p_{ij}^{Ym} = p_i^Y + b_{ij}^{Ym} \ .$$

Then, the regression coefficient b_{ij}^{Ym} simply equals the deviation $(p_{ij}^{Ym}-p_i^Y)$ of the conditional proportion p_{ij}^{Ym} with respect to the mean proportions p_i^Y (see table 4). In other words, the b_{ij}^{Ym} represent the effect of X_m on the sample proportions of Y as deviations from the mean proportions if the distributions of X_ℓ ($\ell\neq m$) given X_m are equal to the unconditional sample distributions of X_ℓ. This is also clear from the decomposition of p_{ij}^{SE} in our example above: if $p_{kj}^{CE}=p_k^C$ (the distribution p_{kj}^{CE} equals the marginal one, p_k^C), we have $b_{ij}^{SE}=p_{ij}^{SE}-p_i^S$ and table 6 would become identical to table 4. Therefore, the regression coefficients b_{ij}^{Ym} might also be interpreted as 'standardized deviations' (on standardization, see also Israëls and De Ree, 1983).

The computation of regression coefficients B can be organized as follows. Provided that the rank of X is K-M (i.e., no zero frequencies in f_k^m, k=1,...,K_m; m=0,...,M and no collinearity between $X_1,\ldots,X_M$, we can obtain B as unique solution of (see (5.1) and (5.6))

$$\begin{bmatrix} WX'Y \\ 0 \end{bmatrix} = \begin{bmatrix} WX'X \\ R \end{bmatrix} B \ .$$

So

$$B = \begin{bmatrix} WX'X \\ R \end{bmatrix}^{-} \begin{bmatrix} WX'Y \\ 0 \end{bmatrix} , \tag{5.7}$$

where A^{-} denotes any g-inverse of A (i.e. $AA^{-}A=A$), cf. Rao & Mitra, 1971. The variance-covariance matrix of the i-th column of B equals

$$Var(b) = \begin{bmatrix} WX'X \\ R \end{bmatrix}^{-} \begin{bmatrix} Var(WX'y) & 0 \\ 0 & 0 \end{bmatrix} (X'XW,R')^{-} , \tag{5.8}$$

where, as before, y denotes a column of Y (the i-th, say) corresponding to b.

To obtain the variance-covariance matrix of WX'y we will assume that the sample was drawn from a finite population with equal probabilities and with replacement. Without loss of generality, we will concentrate on the variance with respect to the distribution of Y conditional on $X_1 \times \ldots \times X_M$. After all, the ANOTA estimator B is unbiased, so

$$\begin{aligned} Var\ B &= \mathbf{E}\ Var(B|X_1 \times \ldots \times X_M) + Var\ \mathbf{E}(B|X_1 \times \ldots \times X_M) = \\ &= \mathbf{E}\ Var(B|X_1 \times \ldots \times X_M) , \end{aligned} \tag{5.9}$$

which is naturally estimated by $\hat{Var}(B|X_1 \times \ldots \times X_M)$. So writing $f_{ij_1..j_M} = f^{YX_1..X_M}_{ij_1..j_M}$ for the frequencies in the M+1 dimensional table, the I-vector $(f_{1j_1..j_M}, \ldots, f_{Ij_1..j_M})$ has a multinomial distribution and for different vectors $(j_1, \ldots, j_M)$ these distributions are independent.

To compute the variance-covariance matrix of the i-th row of the original table, i.e. Var(WX'y), we need to examine the components of WX'y, i.e. $f^{Ym}_{ij}/f^{m}_{j.}$. The covariances of two components, say $f^{Y\ell}_{ij}$ and f^{Ym}_{ik}, fall into three classes: $\ell=m$ and $j\neq k$, $\ell=m$ and $j=k$, and $\ell\neq m$.

The first case, $\text{cov}(f_{ij}^{Ym}, f_{ik}^{Ym})$, is the easiest: f_{ij}^{Ym} and f_{ik}^{Ym} are sums of binominally distributed terms, $f_{ij}^{Ym}=\Sigma f_{ij_1..j..j_M}$ and $f_{ik}^{Ym}=\Sigma f_{ij_1..k..j_M}$, with all terms $f_{ij_1..j..j_M}$ and $f_{ij_1..k..j_M}$ mutually independent. Note that terms are independent within each sum but also between sums. So f_{ij}^{Ym} and f_{ik}^{Ym} are independent and have covariance zero.

The next case is var f_{ij}^{Ym}. Again f_{ij}^{Ym} is a sum of independent binomially distributed terms. Note though that f_{ij}^{Ym} itself is generally not binomially distributed, since the success probabilities are not equal. So

$$\text{var } f_{ij}^{Ym} \doteq f_j^m p_{ij}^{Ym}(1 - p_{ij}^{Ym}) , \tag{5.10}$$

where the right hand side is a systematic (but slight) overestimate of the left hand side. (One may note that the assumption that the $f_{ij_1..j_M}$ are independently Poisson distributed would circumvent these problems.)

Finally $\text{cov}(f_{ij}^{Y\ell}, f_{ik}^{Ym})$ with $\ell \neq m$. Here the expansions for $f_{ij}^{Y\ell}$ and f_{ik}^{Ym} have $f_{ijk}^{Y\ell m}$ in common, while all other terms are independent. So

$$\text{cov}(f_{ij}^{Y\ell}, f_{ik}^{Ym}) = \text{var } f_{ijk}^{Y\ell m} , \tag{5.11}$$

and by the same approximation as above we may write

$$\text{cov}(f_{ij}^{Y\ell}, f_{ik}^{Ym}) \doteq f_{jk}^{\ell m} p_{ijk}^{Y\ell}(1 - p_{ijk}^{Y\ell}) . \tag{5.12}$$

Since $p_{ijk}^{Y\ell m}$ is unknown (we assume that only bivariate information is available), we use the ANOTA-model to estimate it:

$$p_{ijk}^{Y\ell m} = b_i^Y + b_{ij}^{Y\ell} + b_{ik}^{Ym} . \quad (\ell \neq m) \tag{5.13}$$

This completes our discussion of the computation of Var(b).

6. Discussion

Obvious competitive estimators for the OLS estimator studied here are GLS (see e.g. Grizzle e.a., 1969), and ML (see e.g. Fienberg, 1980). In statistical asymptotic efficiency these two are equivalent, optimal and superior to OLS. But al three are consistent; OLS and GLS even being unbiased. The loss of efficiency of OLS is discussed e.g. in Rao and Mitra (1971). Moreover, ML estimates always lead to estimated probabilities that are positive and add to 1, while probabilities estimated by OLS and GLS can be negative or larger than one. But if the model is correct, and all true probabilities are positive, the probability of negative estimates vanishes asymptotically. Computationally OLS estimates have the advantage that they are much easier to obtain than the two competitors, and that they only depend on the bivariate tables. The latter point reduces the data entry and allows ANOTA on data from which only bivariate (or higher dimensional) tables are given, but not the entire matrix.

We have mentioned above that ANOTA may lead to negative estimates for probabilities. Small negative estimates might be acceptable, interpreting them as minor anomalies resulting from sampling errors and imperfect fit of the model. But large negative values will in general be unacceptable. The classical method to avoid these is to transform the observed proportions from the $[0,1]$-range to the $(-\infty,\infty)$-range, e.g. via the logit or probit transformation. Note, however, that this leads to a different class of models (appropriately called generalized linear models, cf. McCullagh and Nelder, 1983, or the original paper: Nelder and Wedderburn, 1972). In general there is no prior reason to believe that the ANOTA-model is more or less close to the true model than, say, a logit model. These transformations are introduced for convenience only, not because their models are 'more likely'. The advantage of logit and probit models that all estimated probabilities are positive is paid for dearly by much more difficulty in the interpretation of the coefficients. Our personal point of view is that sometimes this is acceptable, and sometimes it is not. If the ANOTA-model fits well and all estimated probabilities are positive, ANOTA will be easier to explain to a broad public than logit or probit analysis.

Table 7 crudely summarizes the advantages and disadvantages of ANOTA as compared to logit-analysis.

Tabel 7. ANOTA compared with logit-analysis

	ANOTA	Logit-analysis
interpretation of parameters	very simple	hampered by the transformation
(no. of categories of Y) > 2	trivial	non-trivial
data requirements	bivariate tables only	full table (or almost)
computational requirements	solving linear system (5.7)	iterative solution of a more complex system
statistical efficiency	not fully efficient	efficient
estimated proportions in 0-1-range	not guaranteed	guaranteed
there exists a saturated model	no	yes

References

CBS (Netherlands Central Bureau of Statistics), 1978, De Leefsituatie van de Nederlandse bevolking 1977 (Well-being of the population in the Netherlands 1977) (Staatsuitgeverij, The Hague).

Fienberg, S., 1980, The analysis of cross-classified categorical data (MIT Press, Cambridge).

Grizzle, J.E., Starmer, C.F. and G.G. Koch, 1969, Analysis of categorical data by linear models. Biometrics 25, pp. 489-504.

Israëls, A.Z. and S.J.M. de Ree, 1983, Standaardisatietechnieken (Standardization technics). In: CBS-Select 2 (Staatsuitgeverij, The Hague).

McCullagh P. and J.A. Nelder, 1983, Generalized linear models (Chapman and Hall, London).

Nelder, J.A. and R.W.M. Wedderburn, 1972, Generalized linear models. Journal of the Royal Statistical Society A 135, pp. 370-384.

Rao, C.R. and S.K. Mitra, 1971, Generalized inverse of matrices with its applications (Wiley, New York).

Verbeek, A. and D. Denteneer, 1984, Parametrization of singular linear models (in preparation).

Ontvangen: 18-5-84