ANOTA – Analysis Of Tables

Jelke Bethlehem, January 2006

1. Introduction

1.1 About ANOTA

ANOTA (ANalysis Of TAbles) is a statistical technique to explore possibly existing relationships between categorical (nominal) variables. One of the variables is assigned a special role. It is called the dependent variable. All other variables, the explanatory variables, are considered to be predictors of the dependent variable.

ANOTA resembles linear regression analysis. The main difference is that the dependent variable in linear regression analysis must be numerical. In ANOTA dependent as well as explanatory variables are categorical. The estimated coefficients have the same interpretation as regression coefficients. They measure the effect of the categories of the explanatory variables on the categories of the dependent variable. The coefficients are corrected for possible effects of other explanatory variables and therefore present 'pure' effects. In terms of the Nelder & Wedderburn (1972) generalized linear models, ANOTA analyses a linear model with main effects only, and with identity link function. So the main effects can be interpreted as contributions to proportions (rather than to transformed proportions).

Due to the specific nature of the model, it is not necessary to have the raw data. If suffices to input all possible bivariate tables. This reduces the amount of data which have to be processed.

1.2 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 Survey on Well-being of the Population in The Netherlands 1977 (see CBS, 1978), questions were asked, among others, about Satisfaction, Income, and Education. The three relevant bivariate frequency tables are displayed in tables 1.2.1, 1.2.2 and 1.2.3, 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 1.2.1. Satisfaction (S) by Income (I)

Satisfaction	Income *)				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	1857
Very satisfied	282	485	152	169	1088
Extremely satisfied	103	155	51	51	360
Total	1356	1689	454	609	4108

*) Dutch Guilders x 1000 per annum, 1977

Table 1.2.2. 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	1159	452	169	77	1857
Very satisfied	632	291	115	50	1088
Extremely satisfied	222	90	36	12	360
Total	2492	1027	401	188	4108

Table 1.2.3. Income (I) by Education (E)

Income *)	Education				Total
	Low	Medium	High	Unknown	
< 21	1037	196	59	64	1356
21 -40	912	546	154	77	1689
> 40	146	152	133	23	454
Unknown	397	133	55	24	609
Total	2492	1027	401	188	4108

*) Dutch Guilders x 1000 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 1.2.4.

This table displays for each category of a predictor variable (Income or Education) the distribution over the categories of the dependent variable (Satisfaction) as deviations from the average proportions of the categories of Satisfaction 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.

Table 1.2.4. Satisfaction by Income and Education as deviations from average, in % (standard errors in parentheses).

Satisfaction	Average	Income *)			
		<21	21 - 40	> 40	Unknown
Not too satisfied	6.4 (0.4)	3.3 (0.6)	-1.8 (0.4)	-3.6 (0.8)	0.3 (0.9)
Rather satisfied	13.1 (0.5)	2.2 (0.8)	-1.4 (0.6)	-3.0 (1.4)	1.2 (1.3)
Satisfied	45.2 (0.8)	1.3 (1.1)	0.6 (0.9)	-2.9 (2.2)	-2.3 (1.9)
Very satisfied	26.5 (0.7)	-5.7 (0.9)	2.2 (0.8)	7.0 (2.1)	1.3 (1.7)
Extremely satisfied	8.8 (0.4)	-1.2 (0.6)	0.4 (0.5)	2.5 (1.4)	-0.4 (1.0)
Total	100.0	0.0	0.0	0.0	0.0

*) Dutch Guilders x 1000 per annum, 1977

Table 1.2.4. Satisfaction by Income and Education as deviations from average, in % (standard errors in parentheses), continue).

Satisfaction	Average	Education			
		Low	Medium	High	Unknown
Not too satisfied	6.4 (0.4)	0.6 (0.3)	-1.2 (0.6)	-0.9 (1.1)	0.5 (1.8)
Rather satisfied	13.1 (0.5)	-0.9 (0.4)	0.5 (0.9)	1.6 (1.7)	6.0 (2.8)
Satisfied	45.2 (0.8)	1.3 (0.6)	-1.2 (1.3)	-3.1 (2.3)	-4.2 (3.5)
Very satisfied	26.5 (0.7)	-1.1 (0.6)	1.9 (1.2)	2.2 (2.1)	0.1 (3.1)
Extremely satisfied	8.8 (0.4)	0.1 (0.4)	0.0 (0.8)	0.2 (1.4)	-2.4 (1.8)
Total	100.0	0.0	0.0	0.0	0.0

However, if we look at the distribution of the predictor variables Income and Education, as displayed in table 1.2.5, 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 whether the higher Satisfaction scores for higher Education are 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 explanation problem behind multiple regression models.

Table 1.2.5. Income(I) by Education(E) as deviations from average, in %.

Income *)	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

*) Dutch Guilders x 1000 per annum, 1977

ANOTA supplies us with a simple answer: see table 1.2.6. This table contains the effects of the predictor variables on the dependent variable. They are displayed in the same way as in table 1.4, 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 1.2.6, 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 1.2.4: 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 deviation 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 1.2.6. Satisfaction by Income and Education as deviations from average, in % (standard errors in parentheses).

Satisfaction	Average	Income *)			
		<21	21 – 40	> 40	Unknown
Not too satisfied	6.4 (0.4)	3.3 (0.6)	-1.8 (0.4)	-3.6 (0.8)	0.3 (0.9)
Rather satisfied	13.1 (0.5)	2.2 (0.8)	-1.4 (0.6)	-3.0 (1.4)	1.2 (1.3)
Satisfied	45.2 (0.8)	1.3 (1.1)	0.6 (0.9)	-2.9 (2.2)	-2.3 (1.9)
Very satisfied	26.5 (0.7)	-5.7 (0.9)	2.2 (0.8)	7.0 (2.1)	1.3 (1.7)
Extremely satisfied	8.8 (0.4)	-1.2 (0.6)	0.4 (0.5)	2.5 (1.4)	-0.4 (1.0)
Total	100.0	0.0	0.0	0.0	0.0

*) Dutch Guilders x 1000 per annum, 1977

Table 1.2.6. Satisfaction by Income and Education as deviations from average, in % (standard errors in parentheses), continued.

Satisfaction	Average	Education			
		Low	Medium	High	Unknown
Not too satisfied	6.4 (0.4)	0.6 (0.3)	-1.2 (0.6)	-0.9 (1.1)	0.5 (1.8)
Rather satisfied	13.1 (0.5)	-0.9 (0.4)	0.5 (0.9)	1.6 (1.7)	6.0 (2.8)
Satisfied	45.2 (0.8)	1.3 (0.6)	-1.2 (1.3)	-3.1 (2.3)	-4.2 (3.5)
Very satisfied	26.5 (0.7)	-1.1 (0.6)	1.9 (1.2)	2.2 (2.1)	0.1 (3.1)
Extremely satisfied	8.8 (0.4)	0.1 (0.4)	0.0 (0.8)	0.2 (1.4)	-2.4 (1.8)
Total	100.0	0.0	0.0	0.0	0.0

1.3 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.

Let the *dependent variable* in the model have q categories. For each category of this variable there is a dummy variable which assumes the value 1 if the particular observation belongs to that category, and otherwise it assumes the value 0. The $n \times q$-matrix of scores of the dummy variables is denoted by Y, where n is the sample size. The columns of Y sum up to the sample frequencies of the categories of the dependent variable.

Suppose there are m *predictor variables* in the model. Let the i-th predictor variable have c_i categories. For each category of this variable there is a dummy variable which assumes the value 1 if the particular observation belongs to that category, and otherwise it assumes the value 0. The $n \times c_i$-matrix of scores on the dummy variables corresponding to the i-th predictor variable is denoted by X_i. The columns of X_i sum up to the sample frequencies of the categories of the i-th predictor variable.

Beside m true predictor variables, an additional predictor variable (called 'the constant') is included in the model. This special predictor variable has only one category, on which everyone scores. The $n \times 1$-matrix of scores on this variable is denoted by X_0. So X_0 is a single column consisting of 'ones'.

The scores on all m +1 predictor variables are collected in one $n \times p$-matrix $X = (X_0, X_1, ..., X_m)$ with $p = 1 + c_1 + ... + c_m$.

With the indicator matrices Y and X spelled out, we are ready to translate tables into matrices. The $q \times c_i$-table of frequencies (= number of individuals) of the dependent variable against the i-th predictor variable can now be written as $Y'X_i$, as can easily be confirmed (Y' denotes the transpose of Y). Analogously, the $c_i \times c_j$-table of frequencies of predictor variable i by predictor variable j simply becomes $X_i'X_j$. The score frequencies on the categories of the dependent variable equal $Y'X_0 = f(Y)$, while the $q \times p$-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_i and X to denote either the indicator matrices or the variables themselves).

The $p \times p$-matrix $X'X$ contains all the crossings of $(X_0, X_1, ..., X_m)$ by $(X_0, X_1, ..., X_m)$. The diagonal matrix $X_i'X_i$ with the vector of frequencies $f(X_i)$ on the diagonal is located as submatrix around the diagonal of $X'X$. The vector containing the frequency distribution of all explanatory variables, i.e. the diagonal of $X'X$, is denoted by $f(X)$. This completes our notation.

1.4 The model

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

$$E(y) = Xb$$

where y is an *n*-vector of dependent scores, *X* an *n*×*p*-matrix of predictor scores, *b* a *p*-vector of regression coefficients, and *E*(*y*) the expectation of *y*. This model is amended in a number of ways.

In the first place, the model is generalized to a multivariate linear model

$$E(Y) = XB \tag{2.1}$$

with *Y* an *n*×*q*-matrix, and *B* now a *p*×*q*-matrix of regression coefficients.

In the second place, *X* is assumed to represent the scores on the m predictor variables as described in the previous section. Hence, each row of *X* contains exactly *m*+1 times the value '1', and *p*-*m*-1 times the value '0'.

In the third place, *Y* is now a matrix of scores on a dependent variable only assuming the values 0 or 1, with each row containing exactly one '1'. Hence each row of *E*(*Y*) is a vector of probabilities, adding up to 1.

In order to interpret equation (2.1), we consider the scalar representation of an arbitrary element in the *k*-th column of (2.1). We have

$$p_k(j_1, j_2, ..., j_m) = b_{k0} + \sum_{i=1}^{m} b_{ki}(j_i) \tag{2.2}$$

with b_{k0} standing for the 'constant term' in the model, i.e. the *k*-th element in the first row of *B*, and with $b_{ki}(j_i)$ equal to the *k*-th element of *B* in the row corresponding to the j_i-th category of the predictor variable X_i. So, our model says that the cell probabilities are equal to the sum of regression coefficients which depend only on the bivariate indices.

For our simple example of chapter 1, we have in an obvious notation

$$p_k(i, j) = b_{k0} + b_{kI}(i) + b_{kE}(j),$$

saying that the probability $p_k(i,j)$ of a score on category *k* of Satisfaction, given the predictor scores *i* of Income and *j* of Education, equals a constant b_{k0} depending only on *k*, plus a regression coefficient $b_{kI}(i)$ reflecting the effect of the *i*-th category of Income on the *k*-th score of Satisfaction, plus a regression coefficient $b_{kE}(j)$ for the effect of the *j*-th category of Education on the *k*-th category of Satisfaction.

1.5 Estimation

To estimate B, we consider the (normal) equation

$$Y'X = B'X'X\ . \tag{2.3}$$

By solving this equation, with the theory of ordinary least squares theory (OLS), the ANOTA estimator is obtained. Formally this is just the OLS estimator of model (2.1), but more important is the following, very simple and interesting interpretation of (2.3): the left hand side corresponds to the set of tables $Y \times X_i$ (for i = 1, 2, ..., m). $X'X$ corresponds to the set of tables $X_i \times X_j$ (for i, j = 1, 2, ..., m), and can be seen as a normalising constant, eliminating the interactions between the predictor variables.

As with ANOVA, some restrictions on B are necessary, in order to allow for unique identification of the coefficients. Since the columns of each matrix X_i sum up to a column vector of ones, at least m additional restrictions on B are needed for each category k of the dependent variable. Keller et al. (1985) propose

$$\sum_{j=1}^{c_i} b_{ki}(j) f_j(X_i) = 0 \tag{2.4}$$

for k = 1, 2, ..., q and i = 1, 2, ..., m. The quantity $f_j(X_i)$ denotes the sample frequency of category j of the i-th predictor variable. The interpretation of (2.4) is simple: The average regression coefficient corresponding to a predictor variable (excluding the constant) is zero when weighted with the sample proportions of its categories. It can be shown that, as a consequence of this set of restrictions, the coefficients corresponding to the constant can be written as

$$b_{k0} = f_k(Y)/n, \tag{2.5}$$

i.e. they are equal to the sample proportions in the categories of the dependent variable. In matrix notation we may formulate (2.4) as

$$RB = 0 \tag{2.6}$$

with R an $m \times p$-matrix with $R = (R_0, R_1, ..., R_m)$. The i-th row of the $m \times c_i$-matrix R_i is equal to the vector $f(X_i)$ of sample proportions of the i-th predictor, and all other rows are zero (for i = 1, 2, ..., m). The restrictions just identify the coefficients in B if all vectors $f(X_i)$ are non-zero and the rank of X is equal to p - m. In that case B is obtained as the unique solution of

$$X'Y = (X'X + R'R)B \tag{2.7}$$

The variance-covariance matrix of b, the k-th column of B, is obtained by solving

$$(X'X + R'R)V(b)(X'X + R'R) = V(X'y), \tag{2.8}$$

where y denotes the column of Y corresponding to b.

To estimate the variance-covariance matrix of $X'y$ it is assumed that the sample was drawn from a finite population with equal probabilities and with replacement. So y is a random n-vector of independently (but not identically) distributed zeros or ones, and $V(y)$ is diagonal. For an observation in category j_1 of the first predictor, category j_2 of the second predictor, ..., and category j_m of the m-th predictor, the corresponding value on the diagonal is

$$p_k(j_1, j_2, ..., j_m)\big(1 - p_k(j_1, j_2, ..., j_m)\big) \tag{2.9}$$

For simplicity V(y) is approximated by

$$V(y) = p_Y(k)\big(1 - p_Y(k)\big)I_n, \tag{2.10}$$

where I_n is an $n \times n$ identity matrix. Here $p_Y(k)$ is the unconditional probability that an observation falls into category k of the dependent variable. This approximation conveniently ignores the heteroscedasticity of y, and thus leads to a great simplification.

Certainly for $p_k(j_1, j_2, ..., j_m)$ in the range (0.15, 0.85) the heteroscedasticity is rather limited: the standard deviation only ranges from 0.36 to 0.50. Now (2.10) is estimated by substitution of the sample value $f_Y(k) / n$ for $p_Y(k)$, and (2.8) simplifies to

$$(X'X + R'R)\hat{V}(b)(X'X + R'R) = \hat{V}(X'y), \tag{2.11}$$

where

$$\hat{V}(y) = \frac{f_Y(k)}{n}\left(1 - \frac{f_Y(k)}{n}\right)I_n, \tag{2.12}$$

For a discussion and comparison of ANOTA with other estimation procedures, and other models, see Keller et al. (1985).

2. 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).

Keller, W.J., A. Verbeek and J.G. Bethlehem, 1985, *ANOTA: Analysis of tables*. Internal CBS report (Netherlands Central Bureau of Statistics, Voorburg, The Netherlands).

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