KM 39 (1991)
pg.5-17

LINEAR MODELS AND NOMINAL VARIABLES

Jan Lammers and Ben Pelzer *

Abstract

This article proposes a method for the analysis of nominal variables with linear models in which the dependency of a nominal variable upon one or more nominal and/or interval variables is expressed. The program being used is RENOVA and is available for main frame and pc. The interpretation of results is very easy and parallels the normal interpretation of contingency tables. It extents the latter by providing p.e. standard errors. It enables also to construct a two-dimensional table with frequencies that are controlled for the influence of other variables. Parameters are estimated by ordinary least squares and for the calculation of standard errors it is possible to take into account the heterogeneity of variances of the error terms.

Introduction

There is a long tradition of contingency table analysis in social science research. In the fifties and sixties, the analysis amounted to reading tables using simple rules of Lazarsfeld's elaboration technique. Its popularity was based on the easy interpretability of the results. In the seventies and eighties, formal analysis with loglinear models steadily became more popular. Their attractiveness was due to many aspects, such as statistical tests of significance of many kinds of relationships and determination of the goodness of fit of the model. A serious drawback, however, was the difficult interpretation of the model parameters.

Linear models for the analysis of contingency tabels have never gained this kind of popularity. Attempts have been made by Andrews and Messenger (1973), Boyle (1970) and more recently by Keller, Verbeek and Bethlehem (1985) and Israëls (1987). Use of these models seems to be even suspicious because of not taking the statistical demands seriously. Their unpopularity is however in sharp contrast to the popularity of the intuitive inspection of tables using Lazarsfeld's method, whereas the latter corresponds roughly with an analysis based on linear models.

* The address of both authors is: Department of Methodology, Faculty of Social Sciences, University of Nijmegen, Postbus 9108, 6500 HK Nijmegen, tel. 080-612025/080-612943.

In this article we demonstrate and discuss an analysis of tables using linear models. The authors have developed a computer program for mainframe and personal computer, called RENOVA (Lammers, Pelzer, 1990). It is very similar to the ANOTA program of Keller, Verbeek and Bethlehem but contains some important extensions[1]. As a consequence, the characterization these authors give of their program ANOTA, can be applied to RENOVA. Compared to logit- and probit-analysis it allows for an easier interpretation of the results. However, the price payed for that is that proportions can be estimated outside the range of 0-1 and the parameters are estimated less efficiently. In other words it drops some statistical accuracy to gain ease in use. The article is organised as follows. In the following section we give an idea of uses of RENOVA by presenting some results. Section 2 deals with the formal aspects of the analysis. Finally, in section 3 we discuss the problems that arise in analyses of this type and what to do about these.

1. Uses of RENOVA

To give an idea of what RENOVA does, we use data from the project 'Religion in Dutch society' (a.o. Felling, 1985). In this survey 2966 randomly selected people are asked: is a woman better suited to raise children than a man? The respondent could agree (+), disagree (-) or do neither (0). We wish to analyze the dependence of the answers to this question upon respondents marital status (M), sex (S) and age (A). We regard the dependent variable (Y) as nominal with three categories. The independent variables M and S are treated as nominal variables with 4 (unmarried, married, divorced, widow(-er)) and 2 (man, woman) categories respectively

---

[1] RENOVA and ANOTA differ in the following aspects:

1. In ANOTA only nominal variables can be entered in the model. In RENOVA it is possible to enter nominal and interval variables as dependent or as independent variable.
2. In order to reach column independency of the design matrix ANOTA uses the sample probabilities as general reference for all nominal predictors. In RENOVA for every predictor one can choose between a general reference (the sample probability), the probability of a particular category or the probability of the directly preceding or following category.
3. ANOTA calculates the standard errors of the regression coefficients without taking into account differences of variance of the error term (that is heteroscedasticity). In RENOVA it is possible to input the suitable data for the calculation of the standarderrors based on heteroscedasticity.

and A is considered as an interval variable measured in years.

Table 1 Bivariate regression effects of each predictor on Y (Is a woman better suited to raise children than a man?).

Y	Gen.	Marital status				Sex		Age
	%	unm	mar	div	wid	man	woman	
+	44.67	-13.92*	3.78*	4.55	21.64*	ref	-19.07*	1.10*
		(1.57)	(.62)	(4.24)	(4.95)		(1.80)	(.06)
0	17.67	1.38*	-.19	-5.17*	.23	ref	3.09*	-.13*
		(1.22)	(.48)	(3.30)	(3.85)		(1.40)	(.05)
-	37.66	12.54*	-3.60*	.62	-21.87*	ref	15.98*	-.97*
		(1.53)	(.61)	(4.14)	(4.83)		(1.76)	(.06)

In parentheses the standard errors.
An asterisk indicates that the deviation from the reference is significant at the 5% level.

Table 1 gives the bivariate raw regression effects of each predictor on Y. In the second column the sample percentages of the responses are given; 44.67% of the sample agrees, 17.67% is indifferent and 37.66% disagrees with the statement. Columns 3 up to 6 contain the deviations from these percentages within categories of the marital status. For example the percentage of unmarried persons who agree with the statement is 13.92% lower than the general percentage, that is 30.75%. The deviations for the other response categories of Y are 1.38% and 12.54%. Their sum is zero. The categories of sex (columns 7 and 8) are not compared with the general percentage. For this variable the category man is taken as reference group. Among women the percentage of agreement is 19.07% lower than the percentage among men and this difference is significant. The regression effect of age in the last column parallels the usual meaning; it indicates the increase or decrease of the percentage of the response category for an increase of one year of age. All these effects are significant.

The interpretation of the results for the nominal predictors so far is very similar to the usual way of reading a contingency table. The similarity of the two methods is seen in the fact that the table of observed frequencies can be derived from the parameters of marital status or sex with Y. The regression effects reflect nothing more than the over- or underrepresentations of a category of the nominal variable

with reference to the sample or to a reference category. The extra information contained in table 1 is the standard errors.

In table 2 the regression effects are controlled for the other independent variables. Column 2 contains again the general percentages in the sample. From these, the percentages in the categories of marital status deviate. The deviation for example for unmarried persons is now only 2.62% lower than the general percentage of 44.67%. This deviation is controlled for the other variables and is not statistically significant.

Table 2 Multiple regression effects of each predictor on Y (Is a woman better suited to raise children than a man?).

Y	Gen.	Marital status				Sex		Age
	%	unm	mar	div	wid	man	woman	
+	44.67	-2.62	.69	-1.66	7.95	ref	-20.13*	1.04*
		(1.71)	(.62)	(4.03)	(4.89)		(1.72)	(.07)
0	17.67	-.03	.21	-4.41	1.80	ref	3.14*	-.13*
		(1.41)	(.51)	(3.32)	(4.03)		(1.41)	(.06)
-	37.66	2.65	-.90	6.07	-9.76*	ref	16.99*	-.91*
		(1.53)	(.62)	(3.99)	(4.84)		(1.70)	(.07)

In parentheses the standard errors.
An asterisk indicates that the deviation from the reference is significant at the 5% level.

The other effects can be interpreted in a similar way. Controlling for the other predictor variables, the difference between men and women turns out to be even greater than in table 1. The effects of age do not change very much. It is also clear from table 2 that sex and age are more important in explaining the answers for the different levels of Y than marital status. This can roughly be derived from the strengths of the effects in table 2. But for establishing the relative importance of nominal and interval variables table 2 is less suitable because it contains the unstandardized regression effects. The proportion variance of Y a predictor is directly responsible for, is a better measure for this. Table 3 gives the percentages of variance explained by each predictor after elimination of the effects of the other predictors.

In the last column of table 3 we see that an extreme answer is generally explained more than an indifferent answer. If we examine the explanatory power of each

predictor separately, the same conclusion can be drawn. Age is the most important predictor, followed by sex. Marital status hardly matters at all.

Table 3 Percentage variance of Y, explained by each predictor directly and the percentage of totally explained variance.

Y	Mar. Status	Sex	Age	Total
+	.17	4.08*	8.83*	13.80*
0	.07	.17*	.25*	.47*
-	.30*	3.06*	7.07*	11.13*

* significant at 5% level

Table 3 gives information about each level of Y. It could be desirable to know to what extent Y as a whole is influenced by each predictor. For nominal predictors, this can be done by reconstructing the observed table using the uncontrolled effects. This is done in tables 4 and 6. Tables 5 and 7 contains the reconstructed tables based on the controlled effects.

Table 4 and 5 Observed and controlled percentages of answers +, 0 and - in categories of marital status.

Y	Observed table Marital status				Controlled table Marital status			
	unm	mar	div	wid	unm	mar	div	wid
+	30.7	48.5	49.2	66.3	42.0	45.4	43.0	52.6
0	19.0	17.5	12.5	17.9	17.6	17.9	13.2	19.5
-	50.2	34.1	38.3	15.8	40.3	36.8	43.7	27.9
total	(735)	(2008)	(128)	(95)	(735)	(2008)	(128)	(95)

chi-square 101.34
degrees of freedom 6
Cramers V .1307

chi-square 9.50
degrees of freedom 6
Cramers V .0408

From the tables 4 through 7 it is clear again that the association between marital status and Y is quite sensitive to controlling for the other independent variables, whereas the relation between sex and Y is not. In the former case Cramers' V falls down from .1307 to .0408 and in the latter case this measure remains approximately constant at values of .1963 and .2075. The disappearance of the association between marital status and Y could be due to the fact that age and sex antecede

Table 6 and 7 Observed and controlled percentages of answers +, 0 and - in categories of sex.

Y	Observed table Sex		Controlled table Sex	
	man	woman	man	woman
+	54.8	35.7	55.3	35.2
0	16.0	19.1	16.0	19.1
-	29.2	45.2	28.7	45.7
total	(1397)	(1569)	(1397)	(1569)

chi-square 114.26 chi-square 127.40
degrees of freedom 2 degrees of freedom 2
Cramers V .1963 Cramers V .2075

both marital status and Y. We could therefore proceed in taking marital status as the dependent variable and analyzing the effects of sex and age on this variable. In fact this would be what is done in a usual path analysis. For the treatment of interval as well as nominal variables in a path analysis we refer to Israëls (1987). It suffices to say here that such an analysis can be done with RENOVA.

2. The RENOVA model

RENOVA is the abbreviation of REgression analysis with NOminal VAriables. The underlying model is a linear regression model in which interval and nominal variables can be entered as independent and as dependent variable. When the dependent variable Y is of nominal level, the model is a multivariate linear model. Y is conceived as a set of dummy variables with one dummy variable Y_j for each category. Y_j is 1, if the respondent belongs to category j and zero if he or she does not. If the dependent variable is at the interval level, the model reduces to a univariate model for Y.

On the predictor side of the model, interval and nominal variables may be used. Nominal predictors are dummified in the same way as Y. To continue with the example of the previous section with a nominal dependent variable Y, two nominal variables (marital status with dummies M1, M2, M3 and M4 and sex with dummies S1 and S2) and one interval variable (age), the linear model for Y_j is as

follows:

$$Y_{ji} = b_{j0} + b_{j1}M1_i + b_{j2}M2_i + b_{j3}M3_i + b_{j4}M4_i + b_{j5}S1_i + b_{j6}S2_i + b_{j7}A_i + e_{ji} \quad (1)$$

or in matrix form:

$$y_j = X b_j + e_j \quad (2)$$

where y_j is a vector of scores on Y_j, X is a (n*p) matrix with scores of n individuals on p predictors (the dummies and interval variables), b_j the vector with p parameters and e_j the vector with error terms. It is usually assumed that each error term is normally distributed with an expectation of zero and a variance that is the same for all error terms (assumptions of normality and homoscedasticity). In the case of a nominal dependent variable, model (1) is a probability model. For the expectation of Y_{ji}, $E(Y_{ji})$, is the probability p_{ji} of belonging to category j of Y for individual i and is equal to the structural part of the model because the expectation of the error term is zero.

In model (1) the dummies of marital status are a perfect linear combination of each other and so are the dummies of sex. Thus the model cannot be estimated. There are several ways to solve this problem. Most commonly one dummy of each nominal predictor is eliminated with the consequence that the category of which the dummy is eliminated, becomes the reference category for that variable. In this example it seems a good choice for sex. Whichever dummy is eliminated, the parameter will indicate the difference of predicted Y_j between both sexes. We take the first dummy and remove the first vector of matrix X for sex and the corresponding parameter of b_j in (2). For marital status, the use of a reference category is less desirable. This variable has four categories, none of which is a logical choice for reference category. In such cases, it is preferable to use not a particular category but the sample as a reference. This is realized by imposing the following restriction on the parameters. In equation form for a variable with K categories (from k to K) the restriction is:

$$\sum_{k=1}^{K} p_k b_{jk} = 0 \quad \text{or} \quad \sum_{k=1}^{K} f_k b_{jk} = 0. \tag{3}$$

The restriction specifies that the weighted sum of the parameters equals zero. The weights are the proportions p of individuals in the categories of the predictor. Because the relative frequencies are the best estimates of these proportions, the frequencies of the categories can be taken as weights as well. If the number of nominal variables of which the parameters are restricted in this manner, is r, the restrictions can be added to model (1) as follows:

$$\begin{pmatrix} y_j \\ \overline{0} \end{pmatrix} = \begin{pmatrix} X \\ \overline{R} \end{pmatrix} b_j + \begin{pmatrix} e_j \\ \overline{0} \end{pmatrix} \tag{4}$$

In (4) 0 is a (rx1) null vector and R is a (rxp) matrix with, on each row, the proportions (or frequencies) of the categories in the suitable places and zeros elsewhere. In the previous example with restrictions only on the parameters of marital status, R is the following matrix:

$$R = (0\ f_{m1}\ f_{m2}\ f_{m3}\ f_{m4}\ 0\ 0)$$

The first zero pertains to the intercept, the second and third zero to the parameters of sex and age. After imposing this restriction on the parameters of marital status and after elimination of one dummy of sex, model (2) is estimable. We have, however, modified the model a little further to let the estimate of the intercept be equal to the sample mean. This can be realized by subtracting the means of sex and age from the original scores. For the OLS-estimate of the intercept in this example, this would be:

$$\hat{b}_{j0} = \bar{Y}_j - (\hat{b}_{j1}\bar{M}1 + \hat{b}_{j2}\bar{M}2 + \hat{b}_{j3}\bar{M}3 + \hat{b}_{j4}\bar{M}4) - \hat{b}_{j6}\bar{S}2 - \hat{b}_{j7}\bar{A} \tag{5}$$

The term in parentheses is zero because the mean of a dummy is the proportion of individuals in the corresponding category and because of the restriction on the parameters of marital status. The other terms, except the mean of Y_j, are zero

because the means of the normalized dummy and interval variable are zero. So the intercept is the mean of Y_j. In order to compute the parameters of the model we can premultiply the equation on both sides with the transpose of the designmatrix.

$$(X^t|R^t)\begin{pmatrix} y_j \\ \hline 0 \end{pmatrix} = (X^t|R^t)\begin{pmatrix} X \\ \hline R \end{pmatrix} b_j + (X^t|R^t)\begin{pmatrix} e_j \\ \hline 0 \end{pmatrix} \tag{6}$$

Assuming that the error terms are uncorrelated with the X-variables, it follows that

$$\hat{b}_j = (X^tX + R^tR)^{-1} X^t y_j \tag{7}$$

Vector $\hat{b}_j$ contains the OLS estimators of b_j of which the restricted parameters satisfy restriction (3). The results for the example were given in Table 2 above. For the calculation of the standard errors of $\hat{b}_j$ we will replace $(X^tX + R^tR)^{-1}$ by the letter K for simplicity's sake. We then obtain

$$\begin{aligned} \hat{b}_j &= KX^t y_j = KX^t(Xb_j + e_j) \\ &= KX^tXb_j + KX^te_j \end{aligned} \tag{8}$$

Because vector b_j in the first term on the right side contains the population parameters, the variance of $\hat{b}_j$ does not depend on the first term. In the second term e_j is stochastic, so the variance-covariance matrix of $\hat{b}_j$ can be written as

$$cov(\hat{b}_j) = cov(KX^te_j) = KX^t\ cov(e_j)\ XK = KX^t\ E(e_je_j^t)\ XK \tag{9}$$

Matrix $E(e_je_j^t)$ is of the order n*n with the variances of e_{ji} in the diagonal and the covariances in the off-diagonal positions. Because of random sampling it is reasonable to assume that the covariances are zero. If we assume moreover, as is

usually done in OLS, that the variances are equal (homoscedastic), $cov(\hat{b}_j)$ can further be simplified to:

$$cov(b_j) = \sigma^2 K X^t X K \tag{10}$$

The symbol σ^2 denotes the variance of e_{ji}. This quantity is unknown, but is usually estimated by:

$$\hat{\sigma}^2 = \frac{y_j^t y_j - \hat{b}_j^t X^t X \hat{b}_j}{n - p - 1} \tag{11}$$

where n gives the sample size and p the number of independent vectors of the design matrix X. However the assumption of homoscedasticity is not always met. Some consequences of this will be discussed below.

Once the model is estimated, the proportion of explained variance can be calculated as an indication of how much the predicted scores deviate from the observed scores. This statistic is of value for the whole model, but can also be useful for a part of the model. Especially the proportion of variance explained by a particular variable can be interesting. This proportion is calculated with the estimated model after having set to zero the effects of all other variables. With the predicted scores, thus calculated, the proportion of explained variance that is directly attributed to a particular predictor is obtained[2].

3. Discussion

[2] For an interval variable it can be shown that the resulting proportion equals the squared standardized regression coefficient. For a nominal variable an analogue of this squared standardized coefficient exists (Eisinga, Scheepers, Snippenburg, 1991). To get this coefficient, first a compound variable is constructed using as weights the unstandardized regression effects of the dummy variables of the nominal variable. In the second step the dummies of the nominal variable are replaced by the compound variable and the standardized regression coefficient of the compound is calculated. For this coefficient it is also true that the square can be interpreted as the proportion of variance which the nominal predictor is directly responsible for.

There are especially three assumptions, usually made in regression analysis, which are problematic if the dependent variable is categorical: linearity, normality and homoscedasticity. To start with the last two, it can easily be proven that both assumptions are violated in the case of a categorical dependent variable. Violation of the assumption of normality is the least problematic. It is often mentioned in statistical literature that results of tests of significance of the b parameters are fairly robust against violation of this assumption.

The presence of heteroscedasticity is more serious. Pretending that it does not exist, could lead to standard errors of parameters which are quite different from those calculated on the basis of heteroscedasticity. Findings can therefore wrongly be said to be statistically significant. One can expect, however, that if the probabilities of belonging to a particular category of Y, vary between .20 and .80, the damage caused by heteroscedasticity is minor. Apart from that, it is also possible to estimate the standard errors taking into account the unequal variances of e_{ji}'s. In the variance-covariance matrix of $\hat{b}_j$ (see (9)) estimates of the variances can be placed on the diagonal of matrix $E\left(e_j e_j^t\right)$. It can easily be shown that the variance of e_{ji} is $p_{ji}(1-p_{ji})$. So, having an estimate of p_{ji} is sufficient. This estimate can be obtained with the estimated model. The computer program RENOVA is suited to calculate the standard errors in this way. Thus, RENOVA gives the OLS estimator of the model parameters and can produce its standard error taking into account heteroscedasticity[3].

The assumption of linearity is perhaps most often discussed (Aldrich and Nelson, 1984). The disadvantage of the linear model is that it often occurs that probabilities are predicted outside the range 0-1. Competitors of the linear model are logit and probit models. They never give such disputable results. Predictions outside the

---

[3] The OLS estimators in the case of a categorical variable Y are unbiased and consistent, but not efficient. That is, firstly the mean of all estimates obtained with samples of the same size is the population parameter (unbiased), secondly increasing the size of the sample, the variance of the estimator will become smaller (consistent), but thirdly the variance of the estimator is, given the sample size, not as small as possible (Gujurati, 1983). The WLS estimator possesses this last property, but has the disadvantage that the estimated p_{ji}'s do not necessarily sum up to 1 over all categories of the dependent variable.

0-1 range can be brought about by sampling inaccuracy. They can also arise as a consequence of misspecification and in that case they say something about the quality of the model. Cases with a predicted probability outside the 0-1 range can be sought. One may consider to eliminate them or other cases that are responsible for those deviances, from analysis. After having eliminated the 9 out of 2966 cases which are responsible for such undesired predicted probabilities in the aforehand example, the estimated parameters hardly changed. However, the purpose of analysis is often to map the differential influence of a set of independent variables on a dependent variable and not to predict probabilities. We believe therefore that for the purpose of acquiring insight into the relative importance of variables, a linear model is a useful and easy tool.

References

Aldrich, J.H., Nelson, F.D., *Linear probability, logit and probit models*, a Sage University paper, Series: Quantitative Applications in the Social Sciences, no 45, 1984.

Andrews, D.F., Messenger, R.C., *Multivariate nominal scale analysis*, Institute for Social Research, Ann Arbor, Michigan, 1973.

Boyle, R.P., *Path analysis and ordinal data*, American Journal of Sociology, 75(1969-1970), 461-480.

Eisinga, R., Scheepers, P., Snippenburg, L. van, *The standardized effect of a compound of dummy variables or polynomial terms*, Quality & Quantity, 25: 103-114, 1991.

Felling, A., Schreuder, O., *Religion in Dutch society 1985. Documentation of a national survey on religious and secular attitudes in 1985*, Steinmetz Archive, Amsterdam, 1987.

Gujurati, D., *Basic Econometrics*, International Student Edition, McGraw-Hill Inc., 1983.

Israëls, A.Z., *Path analysis for mixed qualitative and quantitative variables*, Quality & Quantity, 21:91-102, 1987.

Keller, W.J., Verbeek, A., *ANOTA: Analysis of tables*, Kwantitatieve Methoden,

15(1984), 28-44.

Lammers, J., Pelzer, B., *Regressie analyse met nominale variabelen, achtergrond en beschrijving van een programma,* SWI-reeks, vakgroep Methoden, FSW, Katholieke Universiteit, Nijmegen, 1990.

ontvangen 3 - 6 -1991
geaccepteerd 25- 11-1991