Correspondence Analysis of Controlled Frequencies

Piet Verschuren*
Ben Pelzer
Fred Huijgen

ABSTRACT

In correspondence analysis similarities or correlations in categorical data are translated into distances between dots in a scatterplot, making possible very easy and clear interpretation. However, a limitation of this technique thus far in analysing causally related variables, is that it offers no information about the effects a variable has without the aid of other variables in the model. For this information we have to control for the effects of these other variables. By means of an example of industrial relations it is shown that this may be done by using a linear model and the computer program RENOVA.

INTRODUCTION

For the analysis of quantitative data of ordinal and interval level the social science researcher can choose out of a wide range of techniques. In the last few decades the number of techniques for the analysis of nominal data has considerably increased. These techniques may be divided in techniques for statistical testing on the one hand, and for description on the other. An important technique in the first category is loglinear regression. The results of this technique are slightly more difficult to interpret than those of ordinary least squares analysis. From the descriptive techniques we mention homogeneity analysis, nonlinear principal components analysis and correspondence analysis. In this article we discuss one of the main variants of the latter technique.

In this article correspondence analysis is promoted as a suitable alternative for other multivariate techniques, provided that the right type of inputdata is used. It will be argued that nominal variables are gaining interest of researchers and of the users of social science research. Together with its graphical output, that is highly appreciated by most people, this is very much in favor of correspondence analysis (section 2). However, a limitation of correspondence analysis applied to variables that are supposed to be causally related, is that it offers no information about the effects a variable has of its own, without the aid of other variables in the model. We will scrutinize under what conditions adjusting for the influence of other variables can be neglected or is important (section 3). Next it is shown how the problem of adjusting for the influence of variables may be solved by combining correspondence analysis with an analysis of a linear model and the use of the computer program RENOVA. This procedure is illustrated by an example in the field of industrial relations (section 4).

---

* Corresponding author: Dr. Piet J.M. Verschuren, Vakgroep Methoden, Katholieke Universiteit, Postbus 9104, 6500 HE Nijmegen. Tel.: 024 3612709
Email: p.verschuren@maw.kun.nl

WHY CORRESPONDENCE ANALYSIS?

The last decade there has been a growing interest in nominal data. As a consequence of internationalization and globalization during the last decade, new uses of nominal data come into being. That is, as a consequence of this trend we want to compare different countries and cultures. We also may want to know to what extend results of social science research are different for different cultures. For instance we want to know whether a correlation between wage and output of industrial workers, is constant over different countries and different cultures. This comes down to introducing 'cultural unit' as a nominal variable.

Another instance of growing interest in nominal data is the rise of practice oriented research. Instead of theoretical knowledge with a high degree of generalizability, in practice oriented research we want knowledge that is specific for different groups. An example is marketing research. Many industrial and agricultural products, and all kinds of services as well, aim at different target groups. For that reason target group may be an important nominal variable in this type of research.

A last but not least reason for an interest in categorical data is the feeling of many researchers that background variables such as age and gender, although they often make possible a high degree of explained variance, are theoretically empty. For instance, age in its bare essence is an interval variable that has a strong correlation with many other variables. But as such it is an empty variable. An alternative that is much more interesting from a theoretical and practical point of view, is the variable 'stages of life'. For instance, we may divide the life of human beings in five episodes: childhood, adolescence, career building, consolidation of the career and old-age. Thus an interval variable changes in a nominal one. Another example is gender. This variable also often gives a high proportion of explained variance in a wide variety of dependent variables. But instead of gender it may be very fruitful to think about differences between the sexes, such as differences in interests and capabilities. So the dichotomous variable gender may be split up in several nominal variables of theoretical and practical interest.

Besides this growing importance of nominal variables there is a second reason for using correspondence analysis as a technique for data analysis. The output of correspondence analysis is a graphical display that, even in case of complex contingency tables, permits rapid interpretation and understanding. In the last few decades there has been a tremendous rise of visualisation and visual information. The reason for this is that most people are prepared for visual rather than for textual information. Moreover, visual information may be processed more rapid than textual information. In most graphical displays we see at a single glance the way the data are related to each other, and we more easily see regularities such as patterns and profiles in the data.

The graphical output and the easy way it can be interpreted, is the main advantage of correspondence analysis. Greenacre: '...hundreds of researchers were introduced to the method and became familiar with its ability to communicate complex tables of numerical data to nonspecialists through the medium of graphics' (Greenacre, 1993, preface).

ADJUSTING FOR THE INFLUENCE OF OTHER VARIABLES

Normally the input of a correspondence analysis is a contingency table with data that are not adjusted for the influence of other variables in the analysis or the model. In many cases this is no problem or even preferable. The first case to be mentioned is when our aim is prediction. For instance, we may want to predict how much time people spend each day at looking tv. We may do this by simply regressing tv exposure (Y) on variables (X_i) that correlate positively or negatively with exposure, as is visualized in model (a) of figure 1. The parameters linking exposure to each individual predictor is not our main interest. We rather are interested in R-square, the coefficient of multiple determination, indicating the proportion of explained variance in Y that can be attributed to the X variables. For calculating R-square we only need bivariate correlation coefficients and we need not to account for the influence of the other predictors in the model.

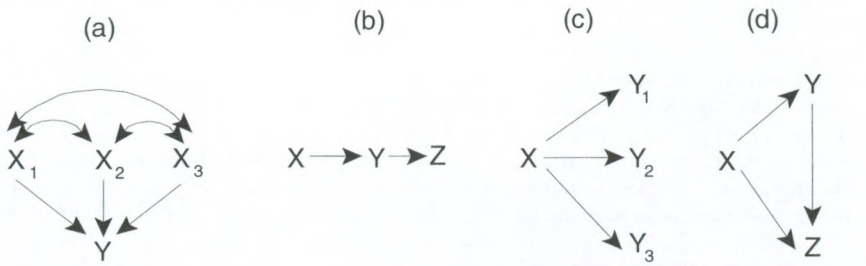

Figure 1. Four basic types of causal models.

Causal chains is another category of models where there is no need for adjusted data (case b in figure 1). In this type of model the effect parameters are bivariate. For instance, in calculating the effect of Y on Z, we need not to partial out X. The same goes for models of type (c). Most measurement models have this form, where X is a theoretical variable to be measured, and Y_1, Y_2 and Y_3 are indicators of X. Here too the parameters linking the indicators with the theoretical variables are bivariate; there is no need for adjusted data.

However, in causal models where there is more than one causal pathway from one variable to another, we need multivariate effectparameters, which means that we have to partial out the other variables. In principle this is the case in models of type d. For instance, in order to know what in this model is the effect of X on Z without the help of Y, we have to partial out the latter. But also in models of type (a) we may have to partial out other variables. This is the case when we want to know the causal effects of each individual X variable on the Y variable. This is just the difference between explanation in the sense of prediction of Y (former use of model (a)) and explanation in terms of causal effects of X variables (latter case).

One might argue that in case a model of type (d) is used as an impact model, i.e. a model that indicates that we have to manipulate X in order to affect Z, the only thing that matters is the amount of change in Z that can be invoked by (manipulating) X. From a standpoint of effectively intervening in reality this argument is valid. However, if after the intervention an evaluation points out that Z did not change very much, it is very important to know whether the reason for a poor effectivety is either a weak direct effect of X on Z, or a weak causal path leading from X to Z via Y. In the latter case we may take measures in order to reinforce this causal pathway. This indicates that for impact models and practice oriented research too it may be very important to adjust for the influence of other variables in the model. As to correspondence analysis we may conclude from this paragraph that in case of models of type (d) and eventually (a), we prefer the input of contingency tables with frequencies that are adjusted for the influence of other variables. In the next and last section will be demonstrated how this can be done.

THE LINEAR MODEL AND RENOVA

In this section we will show how correspondence analysis in an efficient way can be based on controlled contingency tables. Simple correspondence analysis (CA) can be used to analyze and visualize the relation between two categorical variables. The researcher is free to consider one of them, say Y, to be causally dependent on the other one, say X. When the Y variable is thought to be dependent on several X variables simultaneously, a specific version of CA can be used called Composite CA (Israels, 1987). However, in this version the influence of one X is not controlled for the influence of the other X's in the model. This may be a problem, as this case is an instance of model (a) in section 2, interpreted as a causal rather than as a predictive model. As will be shown, this control in principle can be simply achieved by using a crossed contingency table, where an independent variable is crossed with all the other independent variables in the model. However, a difficulty is that the number of cells in a crossed table increases very quickly with the number of variables and the number of categories per variable. A more simple contingency table may be produced as input for a controlled correspondence analysis.
The procedure that we propose for doing this adequately will be elaborated by means of an example in the field of industrial relations (table 1). The data in this table are taken from the EPOC research project in which 5786 establishments of business companies in 10 member states of the European Union participated (EPOC Research Group, 1997).

This large scale comparative survey was initiated by the European Foundation for the Improvement of Living and Working Conditions. One of the research questions was which variables (X's) are responsible for changes in the number of employees during the last three years (Y). Of all features that turned out to be significantly related to change in employee number, we selected 'country' (X_1) and 'unionization' (X_2), the latter being the percentage of union-membership per establishment. For the sake of simplicity we only selected Spain and Sweden for the first variable, whereas we divided the second one into three categories (see table 1). The count columns show the numbers of

X	Spain		Sweden		0% union members		1-69% union members		70-100% union memb.		Total	
Y	Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %
number												
increased	139	36%	193	27%	48	61%	124	35%	160	24%	664	30%
same	137	35%	316	44%	21	27%	141	40%	291	43%	906	41%
reduced	115	29%	207	29%	10	13%	86	25%	227	33%	645	29%
Total	391	100%	716	100%	79	100%	351	100%	678	100%	2215	100%

Table 1: Observed frequencies of number of employees (Y) against country (X_1) and unionization (X_2).

establishments for each row-column combination, while under the name Col % the corresponding column percentages are specified.
In this example with one Y-variable and two X-variables, in a standard composite CA the composite table simply consists of the two contingency-tables of Y against X_1 and X_2 respectively (table 1). By means of Composite CA we produced the two-dimensional scatterplot presented in figure 2. For determining the coordinates of the asterisks in this plot we used the canonical normalization option of ANACOR in SPSS (SPSS,1990), which is proposed by Gifi (1990).

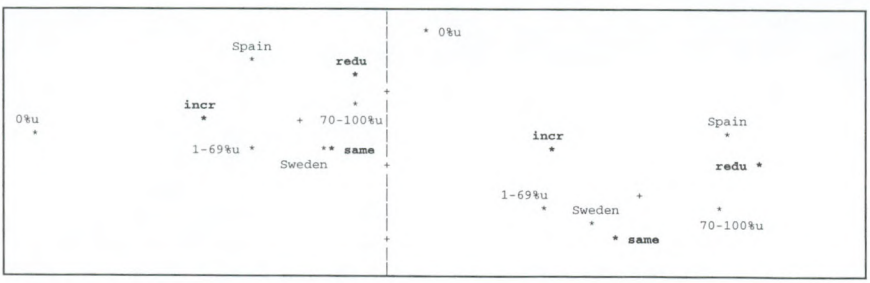

Figure 2: Composite CA plot for table 1 (uncontrolled)

Figure 3: composite CA plot for table 4 (controlled)

In this figure all the X- and Y-categories are represented by an asterisk. The distance between the asterisks indicates similarity of the respective column-proportions in the contingency tables; the smaller the distance the higher the similarity. For instance, the column-proportions of Sweden and of '70-100% union members' in the broad columns two and five of table 1 are more alike than those of 'Spain' and '70-100% union members'. Correspondingly in figure 1 'Sweden' is more close to '70-100%' than 'Spain' is. Moreover 'Sweden' and 'same' are in the same direction from the origin +, while 'Spain' and 'same' lie in the opposite direction. This corresponds with the fact that in table 1 'Sweden' has relatively more companies in 'same' (44%) than

the total sample has (41%), while 'Spain' has less companies (35%) in 'same'. To put it simply, figure 2 shows which X-categories 'belong to' which Y-category in the sense of having a relatively high proportion of companies in that Y-category. In figure 2 the X-categories '0% union members', '1-69% union members' and 'Spain' all belong to the Y-category 'increase in employee-number'. They all are left of the origin +. 'Sweden' and '70-100% union members' on the contrary belong to 'same employee number' and/or 'reduced in number', as they all are right of the origin +. In conclusion, the figure shows that Spain has done better during the last three years, as it has more 'increase' and less 'same' than Sweden.
However, instead of looking at the gross difference between both countries we might ask how different the Y-scores for both countries are, comparing establishments that have the same degree of unionization. That is, we want to know the effects of country on number of employees, controlled or adjusted for the influence of unionization. A straightforward way to get these controlled effects is to construct a table of

	0% union members				1-69% union members				70-100% union members			
	spain		sweden		spain		sweden		spain		sweden	
number	Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %
increased	32	54%	16	82%	102	36%	22	34%	4	9%	156	25%
same	19	31%	3	14%	109	38%	32	49%	9	20%	281	45%
reduced	9	15%	1	4%	75	26%	11	17%	31	71%	195	31%

Table 2: Number of employees against country, crossed with unionization.

(changes in) the number of employees (Y) against country (X_1), one for each of the levels of unionization (X_2) (table 2). Row 1 of table 2 shows that Sweden has more increase than Spain. At the same time in row 3 we see that Sweden has less reduction than Spain. So controlling for unionization reveals that Sweden is doing *better* than Spain as to employment, whereas in the uncontrolled case we found the opposite. This difference shows the importance of partialling out the influence of the other variables.
However, as already mentioned, in case of many independent variables, each containing several categories, the number of cells may increase very quickly. For instance, when we have an Y variable and three X variables all having three categories, we obtaint a crossed contingency table having 3x3x3=27 columns. As each column is represented by an asterisk in the CA plot, this makes the plot very complex and its interpretation difficult. It is true that using this complex contingency table has the advantage that we get insight in all eventual interactions between the variables. But if there is little reason for expecting interactions, we may prefer a more simple plot, where we get an impression of the *global* effect of the X variables on Y, instead of the detailed effects of each X within each combination of categories of the other X's. In doing so we suppose that each X has the same or nearly the same effect on Y within each combination of categories of the other X's, which in many cases will be a reasonable assumption. In that case we need a contingency table with only 3+3+3=9

columns, where the cellfrequences are controlled for the effects of the other X variables in the model. Because of a modest number of variables and of categories, in our example there only is a small reduction of columns; 3x2=6 against 3+2=5 columns respectively (compare tables 2 and 3).

X	Spain		Sweden		0% union members		1-69% union members		70-100% union memb.		Total	
Y	Count	Col %	Count	Col%	Count	Col%	Count	Col%	Count	Col	Count	Col
number												
increased	95	24%	236	33%	50	64%	138	39%	143	21%	663	30%
same	128	33%	325	45%	25	32%	161	46%	266	39%	906	41%
reduced	167	43%	155	22%	3	4%	51	15%	268	40%	645	29%
Total	391	100%	716	100%	79	100%	351	100%	678	100%	2214	100%

Table 3: Controlled frequencies of number of employees (Y) against country (X_1) and unionization (X_2).

Such a condensed table can be adequately produced with the aid of a method proposed by Lammers and Pelzer (1992), where in principle there is no limitation to the number of X-variables to be included in the analysis. This procedure, that can be carried out by using a computer-programma called RENOVA., is shown in the next section.For more details the reader is referred to Lammers and Pelzer (1992). Applying this procedure to our example we get a table of Y against X_1 controlled for X_2 as is presented in the left part of table 3. This table shows that higher unionization goes together with less employment (less increase, more the same, more reduction). This is the same conclusion as can be drawn from the right part of the table 1.
Just like we did for table 1 we performed a Composite CA on table 3. The resulting scatterplot is presented in figure 3. Comparison of figures 2 and 3 shows that Sweden has moved a little towards 'increased', while Spain lies clearly in the same direction as 'reduction'. This corresponds to our previous conclusion that Sweden is doing better as to employment than Spain, provided that we control for the influence of unionization. We also see that the distances from the three unionization-levels to the origin are greater in figure 3, corresponding to the fact that the relation between growth and unionization has become stronger after controlling for the influence of country.

CONSTRUCTING CONTROLLED FREQUENCIES.

Suppose we have a variable Y with three categories and three predictor variables, say A, B and C, with A and B both being nominal-scale variables with three categories each and C being an interval-scale variable. We would like to produce the crosstable of Y against A controlled for the influences of B and C. In order to calculate the cell-frequencies in this crosstable we employ a lineair regression model for each categorie of Y. The three model-equations are:

$$\begin{aligned} Y_1 &= b_1 + b_2 A_2 + b_3 A_3 + b_4 B_2 + b_5 B_3 + b_6 C + e_1 \\ Y_2 &= b_7 + b_8 A_2 + b_9 A_3 + b_{10} B_2 + b_{11} B_3 + b_{12} C + e_2 \\ Y_3 &= b_{13} + b_{14} A_2 + b_{15} A_3 + b_{16} B_2 + b_{17} B_3 + b_{18} C + e_3 \end{aligned} \quad (1)$$

In the equations Y_1, Y_2, Y_3 are 0-1 dummy-variables indicating whether a respondent belongs (1) to a Y-categorie or not (0). In the same way A_2, A_3 and B_2, B_3 are dummy-variables indicating the A and B categories where respondents belong to or not. Since A has three categories only two dummy-variables (we arbitrarily chose A_2 and A_3) can be entered as predictor-variable in the above equations; the same applies to B. The terms e_1, e_2 and e_3 are random error-terms with expectations assumed to be zero.

If the fist equation in (1) is a correct specification of the true relation between Y and A, B and C respectively, the expectation of Y_1 for any given set of values of A, B and C should equal

$$E(Y_1 \mid A, B, C) = b_1 + b_2 A_2 + b_3 A_3 + b_4 B_2 + b_5 B_3 + b_6 C \quad (2)$$

Apart form this the expectation of variable Y_1 given the values of A, B and C is defined as

$$E(Y_1 \mid A, B, C) = 0 * P(Y_1 = 0 \mid A, B, C) + 1 * P(Y_1 = 1 \mid A, B, C) = P(Y_1 = 1 \mid A, B, C) \quad (3)$$

where $P(Y_1 = 0 \mid A, B, C)$ and $P(Y_1 = 1 \mid A, B, C)$ stand for the probability of Y_1 being 0 and 1 respectively, given the values of A, B and C. Combining (2) and (3) we obtain

$$P(Y_1 = 1 \mid A, B, C) = b_1 + b_2 A_2 + b_3 A_3 + b_4 B_2 + b_5 B_3 + b_6 C$$

This equation shows that the expected value of Y_1 as can be computed from the first equation in (1) under the condition that the values of A, B and C and of the b parameters are known, in fact stands for the probability of Y_1 being 1 (i.e. of Y being 1). Because of this the first equation in (1) is called a probability-model and because of the linearity the full name is 'linear probability model'. In the same way the second and third equation in (1) specify the probabilities of Y_2 being 1 and of Y_3 being 1 (i.e. of Y being 2 and Y being 3). Further on we shall spend a few words on problems that may arise when assuming a linear model for probabilities and when applying ordinary least squares (OLS) estimation (as we do) to such a model.

In order to estimate the b-parameters in (1) we apply OLS for each of the three equations seperately. It can be proven that the resulting three $\hat{b}$-estimates for the same predictor-variable (or dummy) sum to zero, e.g. for A_2 we have $\hat{b}_2 + \hat{b}_8 + \hat{b}_{14} = 0$ and for C we have $\hat{b}_6 + \hat{b}_{12} + \hat{b}_{18} = 0$. In OLS the intercept b_1 of the first equation is estimated as

$$\hat{b}_1 = \bar{Y}_1 - \hat{b}_2 \bar{A}_2 - \hat{b}_3 \bar{A}_3 - \hat{b}_4 \bar{B}_2 - \hat{b}_5 \bar{B}_3 - \hat{b}_6 \bar{C}$$

For given values of A, B and C we can now compute $\hat{Y}_1$, that is the predicted probability of Y being 1:

$$\begin{aligned} \hat{Y}_1 &= \hat{b}_1 + \hat{b}_2 A_2 + \hat{b}_3 A_3 + \hat{b}_4 B_2 + \hat{b}_5 B_3 + \hat{b}_6 C \\ &= \bar{Y}_1 - \hat{b}_2 \bar{A}_2 - \hat{b}_3 \bar{A}_3 - \hat{b}_4 \bar{B}_2 - \hat{b}_5 \bar{B}_3 - \hat{b}_6 \bar{C} + \hat{b}_2 A_2 + \hat{b}_3 A_3 + \hat{b}_4 B_2 + \hat{b}_5 B_3 + \hat{b}_6 C \end{aligned} \quad (4)$$

We now define $\hat{Y}_1^*$ as the probability of Y being 1 predicted solely by variable A but controlled for the effects of B and C. The value of $\hat{Y}_1^*$ can be computed from (4) by simply setting to zero the b effects of all variables other then the dummies of A:

$$\begin{aligned} \hat{Y}_1^* &= \bar{Y}_1 - \hat{b}_2 \bar{A}_2 - \hat{b}_3 \bar{A}_3 - 0*\bar{B}_2 - 0*\bar{B}_3 - 0*\bar{C} + \hat{b}_2 A_2 + \hat{b}_3 A_3 + 0*B_2 + 0*B_3 + 0*C \\ &= \bar{Y}_1 - \hat{b}_2 \bar{A}_2 - \hat{b}_3 \bar{A}_3 + \hat{b}_2 A_2 + \hat{b}_3 A_3 \end{aligned} \quad (5)$$

For each A-categorie we can now compute the predicted probability of Y being 1, solely based on the fact that the respondent falls in that A-categorie. Indicating this probability as $\hat{Y}^*_{1|A=i}$, $i = 1, 2$ *or* 3 we can write the following equations:

$$\begin{aligned} \hat{Y}^*_{1|A=1} &= \bar{Y}_1 - \hat{b}_2 \bar{A}_2 - \hat{b}_3 \bar{A}_3 \\ \hat{Y}^*_{1|A=2} &= \bar{Y}_1 - \hat{b}_2 \bar{A}_2 - \hat{b}_3 \bar{A}_3 + \hat{b}_2 \\ \hat{Y}^*_{1|A=3} &= \bar{Y}_1 - \hat{b}_2 \bar{A}_2 - \hat{b}_3 \bar{A}_3 + \hat{b}_3 \end{aligned}$$

In a similar way we can compute the values of $\hat{Y}^*_{2|A=1}$, $\hat{Y}^*_{2|A=2}$ and $\hat{Y}^*_{2|A=3}$ and also the values of $\hat{Y}^*_{3|A=1}$, $\hat{Y}^*_{3|A=2}$ and $\hat{Y}^*_{3|A=3}$. We can collect all these probabilities in a crosstable of Y against A:

		A 1	2	3
Y	1	$\hat{Y}^*_{1\,|\,A=1}$	$\hat{Y}^*_{1\,|\,A=2}$	$\hat{Y}^*_{1\,|\,A=3}$
	2	$\hat{Y}^*_{2\,|\,A=1}$	$\hat{Y}^*_{2\,|\,A=2}$	$\hat{Y}^*_{2\,|\,A=3}$
	3	$\hat{Y}^*_{3\,|\,A=1}$	$\hat{Y}^*_{3\,|\,A=2}$	$\hat{Y}^*_{3\,|\,A=3}$
total		1	1	1

All probabilities in the table are based on the effects of variable A alone. These effects are multiple effects, meaning that they are controlled for the influences that B and C may have on Y. Once we have computed the probabilities in the table we can simply convert them to frequencies by multiplying each probability by the total number of respondents in the category of A that applies. Since the probabilities in each column of the table sum to zero, the frequencies in each column sum to the total number of respondents in that A-categorie. The frequencies calculated this way are

based solely on the effect that variable A has on Y controlled for the influences of B and C.

There are certain drawbacks in specifying probabilities by means of linear equations as we do in (1). A detailed description of these problems can be found in Aldrich and Nelson (1984). One of the problems arises from the fact that the expected probablities as predicted by the model can become negative or larger than one, and such 'impossible' probabilities could never be predicted by any realistic probability model. These out-of-range probabilities typically occur when the relation between the Y variabele and the predictor-variabels is not linear. Application of a linear model might then produce biased parameter-estimates. However, when the underlying relation is (almost) linear, most of the predicted probablities will be within the 0-1 range. Only a few will lie outside that range due to sampling error. However, these few out-of-range probabilities will be very close to 0 or 1. In such a situation the linear probablity model is still appropriate.

Two other problems are attached to the linear probability model. In order to test the significance of estimated parameters the assumption is often made that the terms e_1, e_2 en e_3 are normally distributed. This assumption is violated here since e_1 (and the same applies to e_2 and e_3) can take only two different values for given values of A, B and C. Despite the violation it can be proven that the OLS-estimates of the b-parameters in (1) are unbiased and asymptotically normally distributed. The other problem has to do with heterogeneity of the variance of the error terms, i.e. the variance of say e_1 is not the same for every combination of values of A, B and C. As a consequence the estimates of the standard errors of the OLS $\hat{b}$-parameters in (1) are uncorrect but the bias is very small when the predicted probabilies are in the range 0.15-0.85 (Keller, 1985). If they are in a wider range one might consider methods of correcting for the heterogeneity described by Keller (1985).

Concluding the linear model can be an attractive and easy to interpret alternative in contrast to other probability models such as the logistic and the probit model. However, for the linear model to be appropriate, by far the majority of the predicted probabilities should be within the 0-1 range while the few that probably are not should be close to 0 or 1.

CONCLUSIONS

In causal models where there is more than one causal pathway from an independent X variable to a dependent Y variable we have to adjust for the effect of the other X variables in the model, in order to get a correct impression of the effects of the X variables in the model. The example shows that if we do not adjust, we can easily obtain biased estimates of the parameters, which may lead to wrong conclusions. In case of correspondence analysis on categorical data the adjustment may be done in an adequate way by means of a linear model and the computer program RENOVA.

LITERATURE

Aldrich, J.H. and Nelson, F.D. 1984. *Linear Probability, Logit, and Probit Models.* Sage, London.

Gifi, A. 1990. *Nonlinear Multivariate Analysis.* Wiley, Chichester.

Israëls, A.Z. 1987. *Eigenvalue Techniques for Qualitative Data.* DSWO Press, Leiden, Holland.

Keller, W.J., Verbeek, A. and Bethlehem, J. 1985. ANOTA: Analysis of tables. *Symposium statitische software.* Debets, P., Meurs, A. van, Veling, S.H.J., Verbeek, A. Technisch Centrum FSW, Universiteit van Amsterdam.

Lammers, J. and Pelzer, B. 1992. Linear Models and Nominal Variables. *Kwantitatieve Methoden*, 39, 5-17.

Without author name. 1990. *SPSS categories.* SPSS Incorporation, Chicago.

Ontvangen: 3 maart 1999
Geaccepteerd: 2 december 1999