138

DISCRETE NORMAL LINEAR REGRESSION MODELS

Jan de Leeuw
Department of Data Theory
University of Leiden
Middelstegracht 4, 2312 TW Leiden
The Netherlands

0. ABSTRACT

In this paper we continue our study of the Pearsonian approach to discrete multivariate analysis, in which structural properties of the multivariate normal distribution are combined with the essential discreteness of the data into a single comprehensive model. In an earlier publication we studied these 'block-multinormal' methods for covariance models. Here we propose a similar approach for the regression model with fixed regressors. Likelihood methods are derived and applied to some examples. We review the related literature and point out some interesting possible generalizations. The effect of continuous misspecification of a discrete model is studied in some detail. Relationships with the optimal scaling approach to multivariate analysis are also investigated.

1. INTRODUCTION

In this paper we are interested in various generalizations of the familiar normal linear regression model. These generalizations are intended to be more realistic, at least in most social science situations that we are familiar with, than the classical model. Our starting point is that observed variables are inherently discrete (cf also De Leeuw, 1983). This is true, of course, for both dependent and independent variables, but only the discreteness of the dependent variables has far-reaching consequences for the regression model.

Although observed variables are always discrete, continuous variables can be used in regression modelling for at least two purposes. In the first place the model can stipulate that we are really interested in the relationship between unobservable 'true' continuous variables, of which we have observed discreticized versions. This is the *Pearsonian approach* to discrete variables in multivariate analysis. It is also known as the *treshold approach*, or the *grouped continuous approach*. General discussions and criticisms of this approach, as well as alternatives to this approach, are given in Fienberg (1980), De Leeuw (1983), Winship and Mare (1983), Anderson (1984). Although *Yulean* alternatives (i.e. based on the work of G. Udny Yule) to discrete multivariate analysis are sometimes undoubtedly more natural, we shall restrict our attention in this paper to grouped continuous models. Continuous variables can, in the second place, also be used for computational purposes. A statistical model for a discrete observed phenomenon may be computationally too demanding. In such cases we often approximate the discrete model by a continuous one to simplify the computations. In fact, this is the way the normal distribution originally appeared on the statistical scene. Thus continuous variables have the dual role of modelling the latent or underlying process, and of serving as computational tools.

In this paper we shall study the question in how far such continuous approximations, in our case the classical linear regression estimates, continue to function well in more general discrete models. For completeness we also review and/or derive 'optimal' statistical procedures for discrete regression models. We use the convention of underlining random variables (Hemelrijk, 1966). The abbreviation i.i.d. is used for 'independent and identically distributed'. The notation $\underline{x} \sim N(\mu,\sigma^2)$ is short for 'random variable $\underline{x}$ has a univariate normal distribution with mean μ and variance σ^2'.

2. MODELS

The classical normal linear regression model makes the following assumptions about the sequence of random variables $\underline{y}_i$ (i=1,2,...).

$$\underline{y}_i = x_i'\beta + \underline{\varepsilon}_i. \tag{1a}$$

$$\underline{\varepsilon}_i \text{ i.i.d.} \tag{1b}$$

$$\underline{\varepsilon}_i \sim N(0,\sigma^2). \tag{1c}$$

In (1a) the x_i is a sequence of fixed vectors, say with m elements. The parameters in model (1), which must be estimated, are the m elements of β and the single additional parameter σ^2. Statistical theory for model (1) is entirely routine.

The first generalization we discuss is the *discrete normal linear regression model*. It supposes that the classical model (1) is true for unobserved variables $\underline{y}_i$, which are then discreticized or rounded to produce the observed $\underline{z}_i$. We define the model by four assumptions (2a)-(2d). Assumptions (2a), (2b), and (2c) are the same as (1a), (1b), and (1c). The remaining assumption is

$$\underline{z}_i = \phi(\underline{y}_i), \text{ with } \phi \text{ a monotone step-function.} \tag{2d}$$

In model (2) we assume the number of steps of the step-function to be known. This is entirely natural, because the number of steps is equal to the number of possible values of our discrete observed variable. We also assume, and this is less natural, that the location of the steps are known. Thus there are t known real numbers $\alpha_1 < \alpha_2 < \ldots < \alpha_t$ such that (2d) can also be written as: $\underline{z}_i = s$ if $\alpha_{s-1} < \underline{y}_i \leq \alpha_s$. Here $s \leq t + 1$, and we have used $\alpha_0 = -\infty$ and $\alpha_{t+1} = +\infty$ for notational convenience. We can also eliminate the unobserved variables from the model, and write it directly in terms of the observed variables $\underline{z}_i$. Thus $\underline{z}_i$ is a sequence of discrete random variables, taking values 1,2,...,t+1. We assume

$$\underline{z}_i \text{ independent.} \tag{3a}$$

$$\text{prob}(\underline{z}_i = s) = \Phi((\alpha_s - x_i'\beta)/\sigma) - \Phi((\alpha_{s-1} - x_i'\beta)/\sigma). \tag{3b}$$

Here we have used Φ for the standard normal cumulative distribution function. Also remember that in the discrete normal linear regression model the α_s are given real numbers.

The next generalization is now rather obvious. We assume that the number of steps is still known, but the location of the steps is no longer known. Thus the α_s are t additional parameters, of which it is only known that they are increasing. The resulting *monotone discrete normal linear regression model* is defined by assumptions (4a) and (4b), with (4a) identical to (3a), and with (4b) given by

$$\text{prob}(\underline{z}_i = s) = \Phi(\alpha_s - x_i'\beta) - \Phi(\alpha_{s-1} - x_i'\beta). \tag{4b}$$

In (4b) we have absorbed the parameter σ from (3b) into α and β. The α_s are now unknown, except for their order. Comparing (3b) and (4b) again, we can also say that in the earlier model the α_s are known *up to a linear transformation* (the extra parameter σ), while in the later model the α_s are known *up to a monotone transformation*. If we use terminology from the optimal scaling tradition (De Leeuw, Young, and Takane, 1976, Gifi, 1981, Young, 1981) then the dependent variable is measured on an interval scale for model (3) and on an ordinal scale for model (4).

One further generalization of (4) deserves to be mentioned. In (4) we have used the normal c.d.f. Φ. Models similar to (4), but with other functions in stead of Φ have been proposed by many authors (compare the next section for a partial review). The interesting generalization of (4) assumes that

$$\text{prob}(\underline{z}_i = s) = F(\alpha_s - x_i'\beta) - F(\alpha_{s-1} - x_i'\beta), \tag{5b}$$

with F any c.d.f. This could be called the *monotone discrete nonparametric linear regression model*. We shall make some remarks about possible techniques based on this model in a later section. In the bulk of the paper we concentrate on models (3) and (4), however.

3. SOME HISTORY

The history of the discrete normal linear regression models is very complicated, and our account of it is probably incomplete. We could start the history with Pearson's work on biserial correlation, but this really deals with the situation in which both dependent and independent variables are stochastic. Thus it adresses a somewhat different problem. Perhaps the first appearance of discrete normal linear regression is in *probit analysis*. This was already used by the early psychophysicists Müller and Urban around 1900. The model was called the phi-gamma hypothesis. It only had a single regressor, the physical value of the stimulus. Efficient estimates were computed by weighted linear regression on the transformed proportions, using the 'Müller-Urban weights'. References are in Guilford (1954, chapter 6). Probit analysis is usually attributed to Bliss and Finney, who used the method in bioassay since the thirties. References are in the latest edition of Finney's book on probit analysis (1971). These methods were introduced in econometrics by Tobin (1958), and, in a simultaneous equation context, by Zellner and Lee (1965). Probit models are also popular in the area of modelling discrete choice, although logit models are definitely preferred for choice modelling. An authorative review is given by Manski and McFadden (1981). Probit models were introduced into psychometric test theory by Ferguson and Lawley in the early forties. References are in Lord and Novick (1968, chapter 16, sections on the normal ogive model).

It seems appropriate at this point to point out that probit analysis is a very special case of the discrete normal linear regression model. In the first place the dependent variable in probit analysis is binary: present/absent, dead/alive, correct/wrong, and so on. Extensions to ordered dependent variables with more than two response categories were carried out in biometrics by Aitchison and Silvey (1957), Ashford (1959), Gurland, Lee, and Dahm (1960), and in psychometrics by Samejima (1969). These extensions are all equivalent to our models (3) and (4), although in the biometric literature there is usually only a single regressor and in the psychometric literature this regressor is an unobservable latent trait. There is another, more fundamental sense, in which probit analysis is special. In the classical 'bioassay' problem the independent variable assumes only a small number of values (dosages). For each possible value of x_i there are a large number of replications. In cases such as these it is better to model replications more explicitly. We do this by double indexing. The $\underline{z}_{ij}$ with the same first index, i.e. with the same x_i, are now i.i.d. The $\underline{z}_{ij}$ with a different first index are merely independent. In a particular experiment we have $i=1,\ldots,n$ and $j=1,\ldots,m_i$. We can estimate the probability that $\underline{z}_i = s$ by relative frequencies. And using these probabilities we can apply, for example, the minimum normit chi square method. Or, which is the same thing, weighted least squares with Müller-Urban weights. Thus the probit model, and this is true for the logit and other similar methods as well, is a model which is basically designed for observed proportions. The number of observations is large compared with the number of possible values of the independent variables. There are many replications 'within cells'. The data can be collected in the form of a contingency table with the values of the independent variables defining the rows, and the t+1 possible responses the columns. In the typical social science regression situation we have in mind in this paper the number of values of the independent variables is much too large to make this a practical way of organizing the data. There are too many rows, most cells are empty, the other cells have only a single observation. This makes it impossible to apply minimum normit chi square methods, for example. Maximum likelihood methods (Ashford, 1959, McCullagh, 1980) can still be applied, as we show below. But in probit-like models, and this class includes the discrete choice models of Manski and McFadden and the regression models for ordinal data discussed by McCullagh (1980) and Anderson (1984), the basic statistical reasoning is quite different. In probit models we use asymptotic statistics based on $m_i \to \infty$ for all i. In discrete regression models we use $n \to \infty$. This last form of asymptotic statistics leads to nonstandard results.

It has been pointed out by Amemiya (1973, page 998) that the literature on the estimation of parameters for truncated or censored normal distributions is more closely related to the discrete normal regression problem than the probit literature. Maximum likelihood estimation in truncated normal distributions has been studied in the early fifties by Cohen, Halperin, Hald. This literature is reviewed by Kendall and Stuart (1967, chapter 32). In a biometric context the censored normal distribution was used in survival time research as early as Bliss and Stevens (1937). In the econometric literature the pioneering paper was Tobin (1958). Tobin studied a regression model, which was of the form

$$\underline{z}_i = \max(0, x_i'\beta + \underline{\varepsilon}_i). \tag{6}$$

In fact Tobin considered an apparently more general model, but Amemiya (1973) showed that the additional generality was only apparent. Amemiya also showed that the maximum likelihood estimates behaved properly in Tobin's model, and he provided a convenient consistent initial estimate of the parameters. In Amemiya (1974) the model was extended to multivariate and simultaneous equation models. Rosett and Nelson (1975) extended the Tobin-model to the case in which there are upper limit, lower limit, and non-limit observations. Moreover they made the important extension to the case in which the non-limit values are unknown, which is, in fact, the three-category discrete normal linear regression model. In all papers discussed so far the limit-values (i.e. the α_s of the discrete regression model) are supposed to be known real numbers. The same thing is true in a remarkable biometric paper by Sampford and Taylor (1959), who propose a factorial analysis of variance for censored survival times. The likelihood equations in the paper are solved by a version of the EM-algorithm, that we shall discuss and use more extensively further on.

The monotone discrete normal linear regression model was discussed, in full, for the first time by McKelvey and Zavoina (1975). They define the model, give the likelihood function, differentiate it twice, and discuss a computer program based on Newton's method. Nelson (1976) imbeds the monotone model in a general class of limited dependent variable models, for which he has written a general computer program. Although it may very well be the case that programs based on Newton's method are the most efficient for estimation of parameters, it is also true that the EM-algorithm (Dempster, Laird, Rubin, 1976) provides a lot of insight in the estimation problem and at the same time a very simple algorithm. It has been used for the discrete normal linear regression model, with known bounds, by Wolynetz (1979) and Stewart (1983). A very important theoretical result is that the likelihood function, both for the discrete and for the monotone discrete normal linear regression model, is concave in its parameters. This result was first mentioned by Haberman (in the discussion of McCullagh, 1980), and it was proved, independently it seems, by Burridge (1981) and Pratt (1981). In fact both authors proved it for the more general monotone discrete model given by our (5b), with F not necessary cumulative normal but only restricted to have a strictly log-concave density. Robustness, misspecification, and bias in the Tobit model have recently been investigated by Greene (1981), Nelson (1981), Arabmazar and Schmidt (1982). The book by Manski and McFadden (1981) contains many additional references to closely related work by Lee, Heckman, Hausman, Amemiya, Robinson, Schmidt, and others.

4. ESTIMATION

The technique we use for parameter estimation in this paper is the EM-algorithm, which was introduced in full generality in the statistical literature by Dempster, Laird, and Rubin (1976). In many older papers versions of the EM-algorithm are derived simply by setting the derivatives of the likelihood function equal to zero. The form of the likelihood equations often suggests an iterative algorithm, which is subsequently tried out and found to be convergent. The general theory of the EM-algorithm automatically provides us with a convergence proof, and with information on the speed of convergence (Louis, 1982, Wu, 1983, Boyles, 1983). We shall first adapt the general EM-framework to our discrete regression models.

The function we want to maximize is

$$\underline{L}(\theta) = \sum_{i=1}^{n} \sum_{s=1}^{t+1} \underline{g}_{is} \ln \pi_{is}(\theta), \tag{7}$$

with

$$\pi_{is}(\theta) = \int_{I_s} \phi_i(\theta,x)\,dx. \tag{8}$$

In (7) $\underline{g}_{is}$ indicates what interval observation i is in. Thus $\underline{g}_{is}$ is either zero or one, and for each i only one $\underline{g}_{is}$ is non-zero. I_s in (8) is the interval corresponding with category s, and ϕ_i is the density in cell i, which depends on parameters θ. If $\tilde{\theta}$ is another parameter value, then the Kullback-Leibler inequality (or, if you prefer, the concavity of the logarithm) gives the result

$$\underline{L}(\theta) \geq \underline{L}(\tilde{\theta}) + \{\underline{K}(\theta,\tilde{\theta}) - \underline{K}(\tilde{\theta},\tilde{\theta})\}, \tag{9}$$

with

$$\underline{K}(\theta,\tilde{\theta}) = \sum\sum \underline{g}_{is}\, E_{\tilde{\theta}}(\ln \phi_i(\theta,x) \mid I_s). \tag{10}$$

The M-step of the EM-algorithm maximizes $\underline{K}(\theta,\tilde{\theta})$ over θ, where $\tilde{\theta}$ is the previous value of the parameters. The E-step computes the conditional expectation of the log-density, given that the observation is in I_s, that is required in (10). Observe that if the density is log-concave in the parameters, then $\underline{K}(\theta,\tilde{\theta})$ is concave in θ. Inequality (9) is the key to the convergence proof for the EM-algorithm. If the new estimate of θ is θ^{+}, then

$$\underline{L}(\theta^{+}) \geq \underline{L}(\tilde{\theta}) + \{\underline{K}(\theta^{+},\tilde{\theta}) - \underline{K}(\tilde{\theta},\tilde{\theta})\} > \underline{L}(\tilde{\theta}) + \{\underline{K}(\tilde{\theta},\tilde{\theta}) - K(\tilde{\theta},\tilde{\theta})\} = \underline{L}(\tilde{\theta}). \tag{11}$$

Thus we increase the likelihood, and if the conditions on the density are sufficient to guarantee that the map $\tilde{\theta} \rightarrow \theta^{+}$ is continuous, it follows that all accumulation points of the EM-sequence are stationary points of the likelihood function. The concavity result of Haberman, Burridge, and Pratt, mentioned in the previous section, actually shows that the EM-sequence converges to the unique maximum likelihood estimate. The reasoning above assumes that all parameters are under the integration sign, and does not apply directly to the monotone discrete model. It is easy, however, to adapt it to this case by a simple linear change of variables. In our actual computations, however, we solve for the regression parameters for fixed boundaries by EM. We then solve for the boundaries for fixed regression parameters by Newton's method. Alternating these steps gives a convenient convergent algorithm, which is perhaps not optimal in an overall sense, but which has very simple subproblems that fit very conveniently in a matrix-oriented programming language such as APL.

In the case of discrete normal models the conditional expectation in (10) has a simple explicit form. It was already derived in this context by Wolynetz (1979) and Stewart (1983). First define $\tilde{w}_{is} = (\alpha_s - x_i'\tilde{\beta})/\tilde{\sigma}$. Also

$$\tilde{\kappa}_{is} = (\phi(\tilde{w}_{is}) - \phi(\tilde{w}_{i,s-1}))/(\Phi(\tilde{w}_{is}) - \Phi(\tilde{w}_{i,s-1})), \tag{12}$$

$$\tilde{\tau}_{is} = 1 - (\tilde{w}_{is}\phi(\tilde{w}_{is}) - \tilde{w}_{i,s-1}\phi(\tilde{w}_{i,s-1}))/(\Phi(\tilde{w}_{is}) - \Phi(\tilde{w}_{i,s-1})), \tag{13}$$

$$\tilde{y}_{is} = x_i'\tilde{\beta} - \tilde{\sigma}\tilde{\kappa}_{is}. \tag{14}$$

Then

$$\underline{K}(\theta,\tilde{\theta}) = -\tfrac{1}{2}\{n \ln \sigma^2 + \sigma^{-2}(\tilde{\sigma}^2 \sum_{is}\sum \underline{g}_{is}(\tilde{\tau}_{is} - \tilde{\kappa}_{is}^2) + \sum_i (x_i'\beta - \sum_s \underline{g}_{is}\tilde{y}_{is})^2\}. \tag{15}$$

It is immediately clear from (14) that

$$\beta^{+} = (X'X)^{-1}X'\tilde{y} = \tilde{\beta} - \tilde{\sigma}(X'X)^{-1}X'\tilde{\kappa}. \tag{16}$$

Moreover, using some obvious additional notation,

$$(\sigma^2)^{+} = n^{-1}\{\tilde{\sigma}^2 \sum_i \tilde{\upsilon}_i + \sum_i (x_i'\beta^{+} - \tilde{y}_i)^2\}. \tag{17}$$

This is different from the update-formula of Wolynetz and Stewart, who have

$$(\sigma^2)^{+} = \sum_i (x_i'\beta^{+} - \tilde{y}_i)^2 / \sum_i (1 - \tilde{\upsilon}_i). \tag{18}$$

Although (18), if convergent, leads to the maximum likelihood solution, it does not seem to follow directly from the usual EM-steps.

The algorithm described by (16) and (17) is exceedingly simple to carry out. It consists of a sequence of simple OLS regressions, which are guaranteed to converge to the maximum likelihood solution. Convergence may be slow, but because the iterations are very simple it does not matter if we have to perform a lot of them. Numerical comparions with the Newton-Raphson method are in Burridge (1981).

5. BIAS IN OLS ESTIMATES

Suppose the discrete regression model is misspecified as a continuous model, and parameters are estimated by ordinary least squares. Then clearly the resulting estimates will not be consistent. To study inconsistency in this context we assume, following Greene (1981) and Stewart (1983), the slightly different model

$$\underline{y}_i = \mu + \underline{x}_i'\beta + \underline{\varepsilon}_i, \tag{19a}$$

$$(\underline{x}_i, \underline{\varepsilon}_i) \text{ i.i.d.} \tag{19b}$$

Model (2) differs from (19) in two important respects. In the first place the disturbances in (2) are assumed to be normally distributed, and in the second place the regressors in (2) are fixed. In (19) we assume random regressors, and no normality. We do assume, in addition,

$$E(\underline{\varepsilon}_i) = 0 \text{ and } E(\underline{\varepsilon}_i \underline{x}_i') = 0 \text{ for all } i. \tag{19c}$$

And, of course,

$$\underline{z}_i = \phi(\underline{y}_i), \text{ with } \phi \text{ a monotone step function.} \tag{19d}$$

Model (2) and model (19) are different, although it will be impossible to distinguish them on the basis of empirical data alone. We shall study bias and inconsistency of OLS estimation in model (19), assuming that the results are of relevance also to model (2). This somewhat devious path is followed because model (19) is much simpler, and because 'usually' results derived for 'structural' models such as (19) are also relevant for their 'functional' versions such as (2).

The OLS estimates of β in model (19) are of the form

$$\hat{\beta} = S_{XX}^{-1} S_{XG} \eta, \tag{20}$$

where S_{XX} is the observed covariance matrix of the regressors, where S_{XG} is the observed covariance between the regressors and the column of the indicator matrix $\underline{G}$ with elements $\underline{g}_{is}$, and where η contains scores for the categories. Column s of S_{XG} converges in probability to $\pi_s(E_s(\underline{x}) - m_x)$, where $E_s(\underline{x})$ is the conditional expectation of $\underline{x}$ if $\underline{y}$ is in I_s, where π_s is the content of I_s, and m_x the expected value of x. Now assume that *the regression of $\underline{x}$ on $\underline{y}$ is linear*. Then column s of S_{XG} converges to $\xi_s \Sigma\beta$, where $\Sigma = \text{plim } S_{XX}$ and $\xi_s = \pi_s(\overline{y}_s - \overline{y})/\sigma_y^2$. Here $\overline{y}_s$ is the conditional expectation of y, given that it is in interval I_s. Observe that the ξ_s sum to zero. It follows that $\text{plim } S_{XX}^{-1} S_{XG} = \beta\xi'$, and thus $\text{plim } \hat{\beta} = (\xi'\eta)\beta$. This generalizes results of Stewart (1983), who assumes that the regressors are normally distributed. Our results can be used to derive simple consistent estimates of β by using the fact that $S_{XX}^{-1} S_{XG}$ converges to a rank-one matrix. We can also use the results on optimal quantization, reviewed in Gifi (1981, section 12.3.3), to bound the bias $\xi'\eta$, and to show that the bias disappears if we let $t \to \infty$ (and choose the scores in appropriate ways). Detailed results will be published at another occasion.

OLS-bias can also be studied in a somewhat different way, following Don (1981) and Dempster and Rubin (1983). We write the likelihood equation for β_j in the form, using $\underline{z}$ for a standard normal variable,

$$\sum\sum \underline{g}_{is} E\{\underline{z} \mid (\alpha_{s-1} - x_i'\beta)/\sigma < \underline{z} < (\alpha_s - x_i'\beta)/\sigma\} x_{ij} = 0. \tag{21}$$

Now let $\eta_s = \frac{1}{2}(\alpha_s + \alpha_{s-1})$ and let $\delta_s = \alpha_s - \alpha_{s-1}$. If δ_s is small compared to σ, then the conditional expectation in (21) can be replaced by its Sheppard-approximation. If we substitute this in (21) we obtain

$$\sum (1 - \underline{\delta}_i^2/12\sigma^2)(\underline{\eta}_i - x_i'\beta)x_{ij} = 0, \tag{22}$$

where $\underline{\eta}_i$ is the midpoint and $\underline{\delta}_i$ the length of the interval that observation i is in. If we solve (22) for β we have applied a form of Sheppard's correction or a correction for continuity. A similar development is possible if we want to estimate the variance parameter σ. We write the likelihood equation in the form

$$\sum\sum \underline{g}_{is} E\{\underline{z}^2 \mid (\alpha_{s-1} - x_i'\beta)/\sigma < \underline{z} < (\alpha_s - x_i'\beta)/\sigma\} = n. \tag{23}$$

If we substitute the approximation of the conditional expectation, then this becomes

$$\frac{1}{12} \sum \underline{\delta}_i^2 + \sum (1 - \frac{1}{6}\underline{\delta}_i^2/\sigma^2)(\underline{\eta}_i - x_i'\beta)^2 = n\sigma^2. \tag{24}$$

If we want to solve (22) and (24) we have to decide what to do which the two extreme categories first, because δ_s and η_s are not defined for those categories. If all δ_s are equal, excepting the end-categories of course, then (22) shows that the approximate maximum likelihood estimates are the least squares estimates using the midpoints. This is different from Don (1981) and Dempster and Rubin (1983), but they work with model (19) in which also the regressors are random or are rounded.

We briefly summarize our results on bias. If the intervals are small compared with the variance of the disturbances, then (22) and (24) show that OLS on the midpoints will not be far off. These equations also show how OLS estimates can be corrected, but we do not know if these corrections are really useful. In the examples we have analyzed, which are reported below, the approximation conditions are clearly not met. In model (19), which can be used as another type of approximation to model (2), we have seen that OLS is consistent 'up to a scale factor' provided the regression of the predictors on the criterion is linear. For completeness we also mention a result of Ruud (1983), who analyzes (19) completed with distributional assumptions for $\underline{x}_i$ and $\underline{\varepsilon}_i$. Ruud proves, that if the regression of the components of $\underline{x}_i$ on $\underline{y}_i$ is linear, then the maximum likelihood estimates of β are consistent 'up to a scale factor' even if the distribution of $\underline{\varepsilon}_i$ is incorrectly specified. Sampford and Taylor (1959) point out, correctly, that the maximum likelihood estimate of the variance parameter is biased and inconsistent in the continuous regression model. In the discrete models this bias persists, and they discuss several correction methods. Of course the bias is most serious if the number of parameters is large compared with the number of observations. This will be the case in analysis of variance and analysis of covariance type situations, with a single observation in each cell. In the more usual regression situations the bias will be quite small.

6. RELATIONSHIPS WITH OPTIMAL SCALING

We have already shown that the EM algorithm for discrete normal linear regression models amounts to performing a sequence of ordinary linear regressions on data transformed by using the E-step. The transformations are given explicitly by equation (14). This particular algorithm is useful because it shows in what way maximum likelihood estimation is connected with optimal scaling and partial least squares theory.

In the optimal scaling approach to discrete linear regression or monotone linear regression (Kruskal, 1965, De Leeuw, Young, and Takane, 1976, Young, De Leeuw, Takane, 1976, Gifi, 1981) two types of projection algorithms are alternated iteratively. In the first projection the currently optimally scaled data are projected on the subspace defined by the regression model. This is OLS estimation of the regression parameters. In the second projection the predicted or expected values are projected on the cone of feasible quantifications, which gives new optimal quantifications for the current regression estimates. And so on. The technique is not derived from a particular probability model. It simply starts from the fact that the predicted values are in a given subspace (spanned by the regressors). The feasible quantifications (which must be monotone with the data) are in a convex cone. We are looking for two vectors, one in the cone and one in the subspace, with an angle between them that is as small as possible. Different 'measurement levels' define different cones of feasible quantifications, but the basic idea is the purely geometrical one outlined above.

The optimal scaling approach, which has been extended to many nonlinear models by Gifi (1981), is quite similar to the PLS-approach of Wold. The PLS approach is explained in Jöreskog and Wold (1982). The basic idea is that a simultaneous equation system is constructed using latent variables. Each latent variable has various observed indicators associated with it. The PLS method consists of two steps: in the first step a 'proxy' is computed for each latent variable by making suitable linear combinations of the indicators, and in the second step these currently optimal estimates of the latent variables are used to fit the (recursive) simultaneous equations. Again these two steps are alternated iteratively, until convergence is obtained.

Thus we can say that optimal scaling and PLS define their models in terms of unobserved variables. In optimal scaling the quantification is unknown, or only partially known. In PLS the latent variables are unobserved, but partially known through their indicators. It is clear that the EM-algorithm for discrete regression can be interpreted in exactly the same way as the alternating least squares methods for optimal scaling. The 'optimal' transformation (for current best regression estimates) is now given by the E-step. It is not dictated by purely geometrical considerations any more, but it is influenced strongly by the choice of the probability model. As a consequence it is, in a well defined sense, optimal for the given model, and suboptimal for other models. The optimal scaling transformation (which is often 'monotone regression') may not be optimal in the statistical sense for any particular model, although it is of course optimal in the geometrical sense. According to PLS-adepts the estimates they compute have two major advantages over maximum likelihood estimates for the simultaneous equation system with latent variables. The first advantage is that PLS also estimates the latent variables, and not only the structural parameters of the covariance matrix. It follows directly from the general theory of the EM-algorithm that this is merely a question of algorithm choice. We can compute maximum likelihood estimates by using the EM-method, which computes 'proxies' for the latent variables in the E-step. It can be argued that the EM-proxies are even better than the PLS-proxies, because using EM-proxies leads to consistent estimation of the structural parameters, while using PLS-proxies does not (Dijkstra, 1983). The second acclaimed advantage of PLS is that it may not be optimal in terms of statistical efficiency, but it is optimal in terms of prediction efficiency. This is a very complicated claim, whose precise meaning is not clear to me, but it may be possible to translate this claim in terms of robustness or in terms of the geometry of least squares.

Be this as it may, purely computational considerations suggest that the optimal data transformations from the E-step of the EM-algorithm may have advantages over the cone-projection steps of optimal scaling algorithms as well. Cone projection works on the outside of the cone, and large cones contain many possibilities for 'degenerate' solutions. It seems that, even in nonlinear cases such as multidimensional scaling or principal component analysis, the EM-transformation is less susceptible to degeneracy. Moreover it is statistically optimal in at least one well-defined model, which may be interesting in cases in which we are fairly confident that this model makes sense.

7. SOME EXAMPLES

The model example we use is the serological readings example from Fisher's 'Statistical methods for research workers'. In the 1970 edition it is example 46.2, and it occurs on page 291-300. 'Twelve samples of human blood tested with twelve different sera gave reactions represented by the five symbols –, ?, w, (+), and +.' (l.c., pag 291). Thus the data are a 12 x 12 ANOVA table, with one qualitative observation in each cell. The categories are supposed to be ordered, but this is all we know. Fisher analyzes the data by, essentially, correspondence analysis. We shall analyze them by using our regression methods.

For the analysis we have used a provisory computerprogram MONREG, written in APL. It fits both the discrete and the monotone normal linear regression model by using the Newton-Raphson method to maximize the (concave) log-likelihood function. We do not present the results in detail, but we outline some of the tentative conclusions from the analysis. The log-likelihood for the OLS-solution, using scores 1-5 for the categories, is -74.2155. The discrete linear normal model, with the data intervals equally spaced, category s being interpreted as between $s - \frac{1}{2}$ and $s + \frac{1}{2}$, has a maximum log-likelihood of -68.4316. The estimates of the regression parameters are very close to the OLS estimates, the residual variance is about 2/3 of the OLS residual variance. If we assume that the ML estimates for the discrete linear model are the true values, then we can compute bias and dispersion of OLS estimates. The bias is small, and the dispersion of the OLS estimates is smaller than that of the ML estimates. In fact the mean square error of the OLS estimates is less than the approximate mean square error of the ML estimates. Thus the biased OLS estimates outperform the ML estimates in this respect. The maximum log-likelihood for the monotone model is -57.4829. The estimates of the regression parameters are not very well behaved in this model. The maximum is along a direction of recession of the likelihood function, which some parameters tending to infinity. The information matrix is singular at the maximum likelihood estimate, which causes the Newton method to converge only linearly, and which gives estimates of the sampling variances which are difficult to interpret.

Of course the Fisher example is somewhat special, because the number of regression parameters is large compared to the number of observations. We have also analyzed other examples, varying from ordinary bivariate linear regression to covariance analysis. The Newton algorithm is usually well behaved, and even if some parameters tend to infinity the maximum of the likelihood is still well defined. It is clear from our examples that the concavity of the log-likelihood is a very important property. In many cases it would be difficult to recognize convergence or to know when to stop if concavity would not obtain.

Many things must still be investigated. The results on OLS bias in section 5 have not been investigated for practical usefulness. The precise condition under which parameter estimates tend to infinity have not been established. The nature of the 'optimal' transformation found by the ML-methods has not been studied in any detail. The conditions under which MSE of OLS is less than MSE of ML could also be made more precise, perhaps. All these loose ends call for further investigation. It is of course very important to find out if and in how far our results can be generalized to nonlinear models such as principal component analysis and multidimensional scaling. Optimal scaling methods of the alternating least squares type have been extended to many nonlinear models (Gifi, 1981), and we can perhaps hope that the discrete ML methods generalize equally easily.

8. SOME GENERALIZATIONS

We have already mentioned generalization to nonlinear models in the previous section. This is straightforward, and theoretically not exceptionally interesting. Generalizations such as these are mainly of an algorithmic nature. In this section we point out two types of generalizations which are of a more theoretical nature. They also have major consequences for the algorithms, but these consequences tend to go somewhat deeper.

To discuss the first generalization we remember that the regressors are mapped linearly into the real line, by using a weighted combination. The real line is partioned into intervals, and the probabilities of the responses are integrals over the intervals. These intervals correspond with parallel strips in the regressor space. The generalization is to map the regressors into higher dimensional space, by making more than one linear combination, and to define response probabilities by integrating over non-parallel regions that partition this space. We can think of quadrants, or rectangles, for instance. This is the natural multidimensional generalization of the discrete regression models we have studied. It is consequently not at all true that the grouped continuous approach is limited to one-dimensional models, as Anderson (1984) maintains.

The second generalization steers away from the cumulative normal in (3b) or (4b), using a general F as in (5b). Thus we do not substitute another fixed cdf, such as the logistic or double exponential (discussed extensively by McCullagh,, 1980), but we leave F free. It follows form the properties of general moment problems (a recent review is given in the book by Krein and Nudelman, 1977) that we can suppose without loss of generality that F is a step function. Alternatively we can approximate F by using splines. It is useful to use integrated B-splines for this purpose, because they have the properties of distribution functions, and because B-splines are log-concave on the interval in which they are positive. And log-concavity of the density guarantees concavity of the log-likelihood, according to the result of Burridge and Pratt we discussed earlier. The log-concavity of B-splines is proved in the basic paper by Curry and Schoenberg (1966, theorem 3). Integrated B-splines have been used in data analysis for a rather similar purpose by Winsberg and Ramsay (1981).

It is clear that our suggested generalizations are tentative. The results of our simpler program, with linear models and a completely determined univariate cdf, will indicate if the thought of generalization is at all useful. If the monotone normal linear regression model already has too many parameters to be well-conditioned, then obviously in such cases further generalization would be useless. Much additional work with simpler models is needed, before these generalizations can be attempted. Perhaps most urgently we need to study the speed of convergence of the maximum likelihood estimates to their theoretical limit distributions. If this is already very slow in the simpler models, then the theoretical consistency and efficiency of ML are not very important from a practical point of view. In this case there is room for theoretically less satisfactory, but practically perhaps preferable alternatives. It could very well be that OLS estimation is one such useful alternative.

9. REFERENCES

Aitchison, J. and Silvey, S.D. The generalization of probit analysis to the case of multiple responses. Biometrika, 1957, 44, 131-140.

Amemiya, T. Regression analysis when the dependent variable is truncated normal. Econometrica, 1973, 41, 997-1016.

Amemiya, T. Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. Econometrica, 1974, 42, 999-1012.

Anderson, J.A. Regression and ordered categorical variables. Journal of the Royal Statitiscal Society(B), 1984, 46, 1-30.

Arabmazar, A. and Schmidt, P. An investigation of the robustness of the Tobit estimator to non-normality. Econometrica, 1982, 50, 1055-1063.

Ashford, J.R. An approach to the analysis of data for semi-quantal responses in biological assay. Biometrics, 1959, 15, 573-581.

Bliss, C.I. and Stevens, W.L. The calculation of the time-mortality curve. Annals of Applied Biology, 1937, 24, 815-852.

Boyles, R.A. On the convergence of the EM algorithm. Journal of the Royal Statistical Society, 1983, 45, 47-50.

Burridge, J. A note on maximum likelihood estimation for regression models using grouped data. Journal of the Royal Statistical Society(B), 1981, 43, 41-45.

Curry, H.B. and Schoenberg, I.J. On Polya frequency functions IV: The fundamental spline functions and their limits. Journal d'Analyse Mathematique, 1966, 17, 71-107.

De Leeuw, J. Models and methods for the analysis of correlation coefficients. Journal of Econometrics, 1983, 22, 113-138.

De Leeuw, J., Young, F.W., & Takane, Y. Additive structure in qualitative data: an alternating least squares method with optimal scaling features. Psychometrika, 1976, 41, 471-503.

Dempster, A.P., Laird, N.M., & Rubin, D.B. Maximum likelihood from incomplete data using the EM algorithm. Journal of the Royal Statistical Society(B), 1977, 39, 1-38.

Dempster, A.P. and Rubin, D.B. Rounding error in regression: the appropriateness of Sheppard's corrections. Journal of the Royal Statistical Society(B), 1983, 45, 51-59.

Don, F.J.H. A note on Sheppard's corrections for grouping and maximum likelihood estimation. Journal of Multivariate Analysis, 1981, 11, 452-458.

Dijkstra, T. Some comments on maximum likelihood and partial least squares. Journal of Econometrics, 1983, 22, 67-90.

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

Finney, D. Probit analysis. Cambridge, Cambridge University Press, 1971.

Fisher, R.A. Statististical methods for research workers. (14e edition). Edinburgh, Oliver and Boyd, 1970.

Gifi, A. Nonlinear multivariate analysis. Leiden, Department of Data Theory, 1981.

Greene, W.H. On the asymptotic bias of the ordinary least squares estimator in the Tobit model. Econometrica, 1981, 49, 505-513.

Guilford, J.P. Psychometric methods. (2e edition). New York, McGraw-Hill, 1954.

Gurland, J., Lee, I., & Dahm, P.A. Polychotomous quantal response in biological assay. Biometrics, 1960, 16, 382-398.

Hemelrijk, J. Underlining random variables. Statistica Neerlandica, 1966, 20, 1-8.

Jöreskog, K.G. and Wold, H.O.A. Systems under indirect observation. Amsterdam, North Holland Publishing Company, 1982.

Kendall, M.G. and Stuart, A. The advanced theory of statistics. Volume II, 2e edition, London, Griffin, 1967.

Krein, M.G. and Nudelman, A.A. The Markov moment problem and extrremal problems. Providence (RI), American Mathematical Society, 1977.

Kruskal, J.B. Analysis of factorial experiments by estimating monotone transformations of the data. Journal of the Royal Statistical Society, 1965, 27, 251-263.

Lord, F.M. and Novick, M.R. Statistical theories of mental test scores. Reading (Ma), Addison-Wesley, 1967.

Louis, T.A. Finding the observed information matrix when using the EM Algorithm. Journal of the Royal Statistical Society(B), 1982, 44, 226-233.

McCullagh, P. Regression models for ordinal data. Journal of the Royal Statistical Society, 1980, 42, 109-142.

McKelvey, R.D. and Zavoina, W. A statistical model for the analysis of ordinal level dependent variables. Journal of Mathematical Sociology, 1975, 4, 103-120.

Manski, C.F. and McFadden, D. Structural analysis of discrete data with econometric applications. Cambridge (Ma), MIT Press, 1981.

Nelson, F.D. On a general computer algorithm for the analysis of models with limited dependent variables. Annals of economic and social measurement, 1976, 5, 493-509.

Nelson, F.D. A test for misspecification in the censored normal model. Econometrica, 1981, 49, 1317-1329.

Pratt, J.W. Concavity of the log likelihood. Journal of the American Statistical Association, 1981, 76, 103-106.

Rosett, R.N. and Nelson, F.D. Estimation of the two-limit probit regression model. Econometrica, 1975, 43, 141-146.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, no 17, 1969.

Sampford, M.R. and Taylor, J. Censored observations in randomized block experiments. Journal of the Royal Statistical Society, 1959, 21, 214-237.

Stewart, M.B. On least squares estimation when the dependent variable is grouped. Review of economic studies, 1983, 50, 737-753.

Tobin, J. Estimation of relationships for limited dependent variables. Econometrica, 1958, 26, 24-36.

Winsberg, S. and Ramsay, J.O. Analysis of pairwise preference data using integrated B-splines. Psychometrika, 1981, 46, 171-186.

Winship, C. and Mare, R.D. Structural equations and path analysis for discrete data. American Journal of Sociology, 1983, 89, 54-110.

Wolynetz, M.C. Maximum likelihood estimation in a linear model from confined and censored data. Applied Statistics, 1979, 28, 196-206.

Wu, C.F.J. On the convergence properties of the EM algorithm. The Annals of Statistics, 1983, 11, 95-103.

Young, F.W., De Leeuw, J., & Takane, Y. Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. Psychometrika, 1976, 40, 505-529.

Zellner, A. and Lee, T.H. Joint eestimation of relationships involving discrete random variables. Econometrica, 1965, 33, 382-394.

Ruud, P.A. Sufficient conditions for the consistency of maximum likelihood estimates despite misspecification of distribution in multinomial discrete choice models. Econometrica, 1983, 51, 225-228.

Added in proof

Both Theo Dijkstra and Tom Wansbeek have pointed out to me that the results of Green, Stewart, and Ruud, which were generalized in section 5, have also been generalized, in a very similar way, by Chung and Goldberger (Proportional projections in limited dependent variable models, Econometrica, 1984, 52, 531-534).