Explaining Individual Response using Aggregated Data

Bram van Dijk*
Econometric Institute
Tinbergen Institute
Erasmus University Rotterdam

Richard Paap
Econometric Institute
Erasmus University Rotterdam

ECONOMETRIC INSTITUTE REPORT EI 2006-05

Abstract

Empirical analysis of individual response behavior is sometimes limited due to the lack of explanatory variables at the individual level. In this paper we put forward a new approach to estimate the effects of covariates on individual response, where the covariates are unknown at the individual level but observed at some aggregated level. This situation may, for example, occur if the response variable is available at the household level but covariates only at the zip-code level.

We describe the missing individual covariates by a latent variable model which matches the sample information at the aggregate level. Parameter estimates can be obtained using maximum likelihood or a Bayesian approach. We illustrate the approach estimating the effects of household characteristics on donating behavior to a Dutch charity. Donating behavior is observed at the household level, while the covariates are only observed at the zip-code level.

JEL Classification: C11, C51
Keywords: aggregated explanatory variables, mixture regression, Bayesian analysis, Markov Chain Monte Carlo

*We thank Dennis Fok, Philip Hans Franses, Rutger van Oest, Björn Vroomen and participants of seminars at the Institute of Advanced Studies in Vienna, the Université Catholique de Louvain in Louvain-la-Neuve, NAKE research day in Amsterdam, and Facultes Universitaires Notre-Dame de la Paix in Namur, for helpful comments. Corresponding author: Tinbergen Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, e-mail: avandijk@few.eur.nl, phone: +31-10-4088943, fax: +31-10-4089162.

1 Introduction

Empirical analysis of individual behavior is sometimes limited due to the lack of explanatory variables at the individual level. There may be various reasons why individual-level explanatory variables are not available. When using individual revealed preference data, information about explanatory variables may simply not be available as databases cannot be properly linked. When using surveys, one may be confronted with a missing question concerning an important explanatory variable. It may also be the case that respondents interpreted the relevant question the wrong way which makes the explanatory variable unusable.

In some cases it may be possible to obtain information on explanatory variables at some aggregated level. For example, if the zip code of households is known, one may obtain aggregated information on household characteristics, like income and family size, at the zip-code level. This zip-code level information is usually obtained through surveys. The aggregated information of the variables is summarized in marginal probabilities which reflect the probability that the explanatory variable lies in some interval (income, age) or category (gender, religion) for a household in that zip-code region.

The goal of the current paper is to estimate the effects of covariates on individual response when the covariates are unobserved at the individual level but observed at some aggregated level. This problem is related to the literature on ecological inference, see, for example, Wakefield (2004) for an overview. The main difference with regular ecological inference problems is that we observe individual responses, whereas in ecological inference one also has to rely on aggregated information on the response variable. The extra information on individual responses may help us to overcome certain identification issues in ecological inference.

There are several studies in economics which link individual and aggregated data, see, for example, Imbens and Lancaster (1994) and van den Berg and van der Klaauw (2001). The difference of these studies with our problem is that they assume that both individual-level data and aggregated data is available. The aggregated data is assumed to be more reliable and is used to put restrictions on the individual-level data. Our problem

bears more similarities with symbolic data analysis, see Billard and Diday (2003) for an overview. Symbolic data analysis also deals with aggregated explanatory variables and dependent variables at an individual level. The motivation for the use of aggregated data is however different. Aggregation is pursued to summarize large datasets. Therefore the form of the aggregated information is different and represents, for example, intervals instead of marginal probabilities.

As far as we know, the only paper that comes close to our situation is Steenburgh *et al.* (2003). The motivation in this paper is however different from ours. They use zip-code information to describe unobserved heterogeneity in the individual behavior of households instead of estimating the effects of covariates on behavior.

To deal with our specific problem we propose in this paper a new approach to estimate the effects of covariates on individual response, where the covariates are unknown at the individual level but observed at some aggregated level. We add to the model describing the individual responses a latent variable model for the explanatory variables. This latent variable model describes the missing explanatory variables in such a way that it matches the sample information at the aggregated level. In case of one explanatory variable, the model simplifies to a mixture regression. A simple simulation experiment shows that this new approach outperforms in efficiency the standard method, where one replaces the missing explanatory variables by the observed marginal probabilities at the aggregated level.

Parameter estimates of the response model can be obtained using Simulated Maximum Likelihood [SML] or a Bayesian approach. Given the computational burden of SML, the latter approach may be more convenient. To obtain posterior results, one can use a Gibbs sampler with data augmentation. The unobserved explanatory variables are sampled alongside the model parameters. Conditional on the sampled explanatory variables, one can use a standard Markov Chain Monte Carlo [MCMC] sampler for the model describing individual response.

The outline of the paper is as follows. In Section 2 we provide a simple introduction into the problem and perform a small simulation experiment to illustrates the merits of our approach. In Section 3 we generalize the discussion to a more general setting. Parameter

estimation is discussed in Section 4. In Section 5 we illustrate our approach estimating the effects of household characteristics on donating behavior to a Dutch charity. We use aggregated information on household characteristics at the zip-code level to explain the individual response of households to a direct mailing by the charity. Finally, Section 6 concludes.

2 Preliminaries

To illustrate the benefits of our new approach, we start the discussion with a simple example. We consider a linear regression model with only one explanatory variable. The explanatory variable x_i can only take the value 0 or 1, for example, a gender dummy. Let the observed response of individual i, y_i, be described by

$$y_i = \alpha + \beta x_i + \varepsilon_i, \tag{1}$$

where α is an intercept parameter and where β describes the effect of the 0/1 dummy variables x_i on y_i for $i = 1, \ldots, N$. The error term ε_i is assumed to be normally distributed with mean 0 and variance σ^2. We assume that x_i is unobserved at the individual level but that we have aggregated information on x_i, for example, at the zip-code level. This aggregated information is summarized by $p_i = \Pr[x_i = 1]$ for $i = 1, \ldots, N$.

A simple approach to estimate β is to regress y_i on p_i instead of x_i. The error term of this regression equals

$$\eta_i = (x_i - p_i)\beta + \varepsilon_i. \tag{2}$$

The OLS estimator is consistent if $\mathrm{E}[p_i\eta_i|p_i] = 0$. As

$$\begin{aligned} \mathrm{E}[p_i\eta_i|p_i] &= \mathrm{E}\left[p_i \times ((x_i - p_i)\beta + \varepsilon_i)|p_i\right] = \mathrm{E}\left[p_i(x_i - p_i)\beta|p_i\right] + \mathrm{E}[p_i\varepsilon_i|p_i] \\ &= \mathrm{E}\left[p_i x_i \beta|p_i\right] - \mathrm{E}\left[p_i^2\beta|p_i\right] + \mathrm{E}[p_i\varepsilon_i|p_i] \end{aligned} \tag{3}$$

this condition is fulfilled if $\mathrm{E}[p_i\varepsilon_i|p_i] = 0$ and $\mathrm{E}[x_i|p_i] = p_i$. Although this OLS estimator is consistent, it is clear from (2) that the error term is heteroskedastic, and hence the OLS estimator is not efficient. Hence, one may opt for a GLS estimator.

An alternative approach to use the aggregated information to estimate β is to consider a mixture regression, see Quandt and Ramsey (1978); Everitt and Hand (1981);

Titterington *et al.* (1985). To describe the response variable y_i we consider a mixture of two regression models where in the first component the x_i variable is 1 and in the second component x_i equals 0. The mixing proportion is p_i which is known but may be different across individuals. Hence, the distribution of y_i is given by

$$y_i \sim \begin{cases} \mathrm{N}(\alpha+\beta,\sigma^2) & \text{with probability } p_i \\ \mathrm{N}(\alpha,\sigma^2) & \text{with probability } (1-p_i). \end{cases} \tag{4}$$

The parameters α and β can be estimated using maximum likelihood [ML]. ML estimates can easily be obtained using the EM algorithm of Dempster *et al.* (1977).

To illustrate the efficiency gain of the mixture approach we perform a simulation study. For $N = 1,000$ individuals we simulate 0/1 x_i values according to $\Pr[x_i = 1] = p_i$. We use different simulation schemes for p_i. We either allow the value of p_i to be different across individuals, or we impose that groups of individuals have the same value for p_i corresponding to the idea that these individuals live the same zip-code region. Furthermore, we allow the range of possible values for p_i to be different. We sample p_i from $U(0.2, 0.4)$ or $U(0.01, 0.99)$. The values of y_i are generated according to $y_i = 1 + 2x_i + \varepsilon_i$ with $\varepsilon_i \sim \mathrm{N}(0,\sigma^2)$ and $\sigma^2 = 1$.

We estimate the β parameter using four approaches. In the first approach we estimate β using a linear regression model where we include the true x_i as explanatory variables. In practice this solution is of course not feasible but it allows us to compute the efficiency loss due to using explanatory variables at an aggregated level. In the second approach we consider an OLS estimator in a linear regression model with p_i as explanatory variable. The third approach uses a GLS estimator in the same linear regression model. The GLS weights are based on (3) and are computed using the true value of β and σ^2. In practice these parameters are of course unknown but the simulation results already show that accounting for heteroskedasticity using the true values does not compensate the efficient loss of the OLS estimator. In the last approach we consider the mixture solution as in (4).

Table 1 displays the efficiency loss in the estimator for β for the last three estimation approaches compared to using full information. Simulation results are based on 1000 replications. The efficiency loss is computed using the root mean squared error of the

Table 1: Efficiency loss of using aggregated data with respect to using full information for the three estimators

Distribution of p_i	Number of p_i[a]	Efficiency Loss OLS	GLS	Mixture
$U(0.20, 0.40)$	1,000	90.5%	90.5%	33.3%
$U(0.01, 0.99)$	1,000	50.4%	49.8%	24.1%
$U(0.20, 0.40)$	100	90.4%	90.4%	32.4%
$U(0.01, 0.99)$	100	52.5%	52.2%	23.0%
$U(0.20, 0.40)$	10	92.4%	92.3%	31.5%
$U(0.01, 0.99)$	10	62.4%	62.1%	23.0%
$U(0.20, 0.40)$	2	96.6%	96.6%	33.5%
$U(0.01, 0.99)$	2	73.9%	73.9%	31.3%

[a] Number of different p_i values drawn from the uniform distribution. Number of individuals is 1,000.

estimates as all estimators are consistent. Several conclusions can be drawn from the table. First of all, the mixing approach outperforms the other two estimators. Secondly, the GLS estimator hardly improves upon the OLS estimator, indicating that heteroskedasticity is not the main cause of the efficiency loss of the OLS estimator. Thirdly, all estimators perform better when the range in possible values of p_i is larger, which is not a surprise as a large variation in p_i provides more information about the slope parameter. Finally, the estimators perform better when there are less individuals with the same p_i value. The mixing approach however seems hardly affected by the number of individuals with the same value for p_i.

To illustrate the effects of the efficiency loss, we display in Figure 1 the density of $\hat{\beta}$ for the full information estimator, the OLS estimator and the mixture approach, where we use the simulation settings as in the first line of Table 1. The graph clearly illustrates the superiority of the mixture approach.

As already indicated by our simulation results, a GLS estimator does not compensate the efficiency loss due to aggregation of the explanatory variables. A second reason why the GLS estimator is not useful, is that constructing a feasible GLS estimator is often not possible if we have more than one explanatory variable. Consider, for example, the case

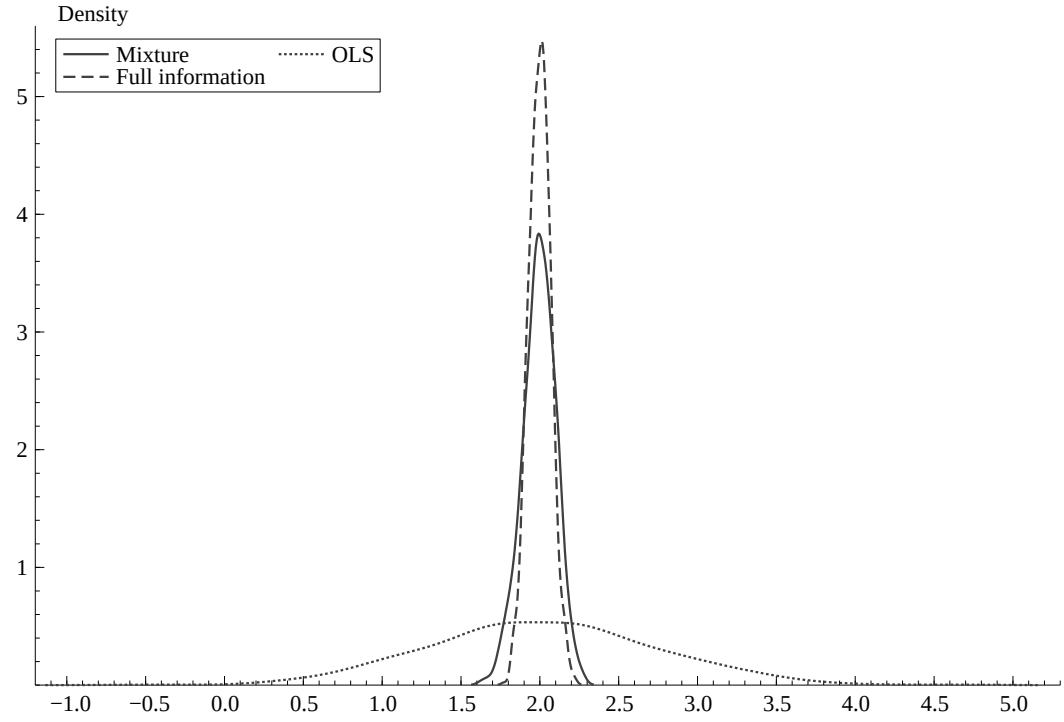

Figure 1: Density plots of the three estimators for β

with k explanatory variables which are unobserved at the individual level

$$y_i = \alpha + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \tag{5}$$

where x_{ij} are unobserved 0/1 dummy variables. Assume that we have aggregated information summarized in k marginal probabilities $\Pr[x_{ij} = 1] = p_{ij}$. It is straightforward to extend the proof above to show that OLS estimator for β_j where the x_{ij} are replaced by p_{ij} is consistent. If we replace x_{ij} by p_{ij}, the error term becomes

$$\eta_i = \sum_{j=1}^{k} (x_{ij} - p_{ij})\beta_j + \varepsilon_i. \tag{6}$$

Although the OLS estimator is consistent, it is impossible to estimate the variance of η_i, because the covariance matrix of x_i is unknown. As in practice we often only observe the marginal probabilities $\Pr[x_{ij} = 1] = p_{ij}$ and not the joint probabilities it is not feasible to estimate these covariances.

To obtain a more efficient estimator for the β parameters we extend in the next section the mixture approach to more than one explanatory variable. The proposed approach

uses the information in the individual responses to estimate the unobserved correlations between the covariates.

3 Model specification

In this section, we generalize the discussion in the previous section in several ways. First, we relax the assumption that the model for y_i is a linear regression model. Secondly, we allow for m explanatory variables summarized in the m-dimensional vector X_i. Finally, we allow for other type of explanatory variables like ordered and unordered categorical variables and continuous variables. The vector of explanatory variables is written as $X_i = (X_i^{(1)\prime}, X_i^{(2)\prime}, X_i^{(3)\prime}, X_i^{(4)\prime})'$, where $X_i^{(1)}$ contains the binary explanatory variables, $X_i^{(2)}$ the ordered categorical explanatory variables, $X_i^{(3)}$ the unordered explanatory variables and $X_i^{(4)}$ the continuous explanatory variables.

We will use the general model specification

$$y_i = g(X_i\beta, \varepsilon_i), \tag{7}$$

where y_i is the observed dependent variable, β is an m-dimensional vector with the parameters of interest, ε_i is a random term, and g is some (non)linear function. The distribution of ε_i is known and depends on the unknown parameter vector θ.

This general model can be a linear regression model, but also a limited dependent variable model or any other nonlinear model. If the X_i variables are observed, parameter estimation is usually standard. In our case, the X_i variables are unobserved at the individual level but we have sample information at some aggregated level. To estimate the model parameters β and θ we propose a latent variable model to describe the joint distribution of the X_i variables. Some of the parameters of this latent variable model are fixed to match the available sample information at the aggregated level. In the following subsections we describe the latent variable model for the different types of explanatory variables.

3.1 Binary explanatory variables

Assume that $X_i^{(1)}$ consists of k binary variables. The joint distribution of $X_i^{(1)}$ is discrete with 2^k mass points which sum up to 1. If we observe these $2^k - 1$ mass points at some aggregated level, we can follow the mixture approach of Section 2 to estimate the β parameters. In practice, however, we typically observe the k marginal probabilities denoted by $P_i^{(1)} = (p_{i1}^{(1)}, \ldots, p_{ik}^{(1)})'$. Romeo (2005) proposes a method to estimate the joint discrete distribution from the marginal probabilities. However, he assumes that the joint distribution is known at an aggregated level. Since we do not have this joint distribution at an aggregated level, his method is not feasible for our problem at hand.

The k marginal probabilities plus the fact that probabilities sum up to 1 leave us with $2^k - (k+1)$ degrees of freedom on the 2^k mass points, unless we assume that the explanatory variables are independent. To facilitate modeling the joint distribution of $X_i^{(1)}$, we introduce a latent continuous random vector $X_i^{(1)*} = (x_{i1}^{(1)*}, \ldots, x_{ik}^{(1)*})'$ with

$$\begin{array}{lll} x_{ij}^{(1)} = 1 & \text{if} & x_{ij}^{(1)*} > 0 \\ x_{ij}^{(1)} = 0 & \text{if} & x_{ij}^{(1)*} \leq 0 \end{array} \tag{8}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, k$, see also Joe (1997) for a similar approach. A convenient distribution for $X_i^{(1)*}$ is a multivariate normal. The variance of $x_{ij}^{(1)*}$ is set equal to 1 for identification. We impose that the mean of $x_{ij}^{(1)*}$ equals $\Phi^{-1}(p_{ij}^{(1)})$ for $j = 1, \ldots, k$ and $i = 1, \ldots, N$, where Φ denotes the CDF of the standard normal distribution. It holds that $\Pr[x_{ij}^{(1)} = 1] = \Pr[x_{ij}^{(1)*} > 0] = \Phi(\Phi^{-1}(p_{ij}^{(1)})) = p_{ij}^{(1)}$, and hence these restrictions match the marginal distribution of the $X_i^{(1)}$ variables. Hence, we assume that

$$X_i^{(1)*} \sim \mathrm{N}\left(\Phi^{-1}(P_i^{(1)}), \Omega_1\right), \tag{9}$$

where Ω_1 is a $k \times k$ positive definite symmetric matrix with ones on the diagonal. This leaves us with $\frac{1}{2}k(k-1)$ free parameters, the sub-diagonal elements of Ω_1. Although we loose some flexibility by assuming this structure, the correlation parameters do get an intuitive interpretation as they are related to correlations between the explanatory variables. The model for $X_i^{(1)}$ is in fact a multivariate probit [MVP] model, see Ashford and Sowden (1970); Amemiya (1974); Chib and Greenberg (1998). In our case however,

the aggregated data provides restrictions on the intercepts and only the sub-diagonal elements of Ω_1 have to be estimated.

3.2 Ordered categorical explanatory variables

The setup for the binary variables can easily be extended to ordered categorical variables. If we have one ordered categorical variable with r categories, the $X_i^{(2)}$ vector in (7) contains $r-1$ 0/1 dummies, leaving one category, say the last one, as a reference category. Denote the $r-1$ dummies by $x_{i1}^{(2)}, \ldots, x_{ir-1}^{(2)}$. We typically observe the marginal distribution of the categories at some aggregated level which we denote by the r probabilities $P_i^{(2)} = (p_{i1}^{(2)}, \ldots, p_{ir}^{(2)})'$.

If we only have one ordered categorical explanatory variable in our model, we can use the simple mixture approach in Section 2 to estimate the effects of the r categories. In practice, one usually has a combination of several binary and ordered categorical variables and one has to deal with correlation between these variables. To describe correlations between several categorical variables, it is convenient to introduce a normal distributed random variable $x_i^{(2)*}$ to describe the distribution of the categorical variable in the following way

$$\begin{array}{rlll} x_{i1}^{(2)} = 1 & \text{if} \quad x_i^{(2)*} \leq q_{i1} & \text{and} & x_{i1}^{(2)} = 0 \quad \text{otherwise} \\ x_{i2}^{(2)} = 1 & \text{if} \quad q_{i1} < x_i^{(2)*} \leq q_{i2} & \text{and} & x_{i2}^{(2)} = 0 \quad \text{otherwise} \\ \vdots & & & \\ x_{ir-1}^{(2)} = 1 & \text{if} \quad q_{ir-2} < x_i^{(2)*} \leq q_{ir-1} & \text{and} & x_{ir-1}^{(2)} = 0 \quad \text{otherwise.} \end{array} \tag{10}$$

For identification we impose that the variance of $x_i^{(2)*}$ is 1 such that

$$x_i^{(2)*} \sim \mathrm{N}(0,1). \tag{11}$$

To match sample probabilities $P_i^{(2)}$, the limit points $q_{i1} \ldots q_{ir-1}$ are set equal to

$$q_{ij} = \Phi^{-1}\left(\sum_{l=1}^{j} p_{il}^{(2)}\right), \; i = 1, \ldots, N, \; j = 1, \ldots, r-1. \tag{12}$$

The proposed model for $X_i^{(2)}$ is in fact the ordered probit model of Aitchison and Silvey (1957).

The equations (10)–(12) provide the latent variable model for the case of one ordered categorical explanatory variable. In case one has more categorical variables one can easily extend the current solution with more latent $x_{ij}^{(2)*}$ variables and allow them to correlate using a covariance matrix Ω_2 with ones on the diagonal. One can also correlate the resulting $X_i^{(2)*}$ variables with the latent random variables for the binary variables $X_i^{(1)*}$ to describe correlations between binary and ordered categorical explanatory variables.

3.3 Unordered categorical explanatory variables

One may also encounter an explanatory variable which is categorical with, say, r categories, but that there is no natural ordering. We assume here that an individual can only belong to one category. If (s)he can belong to several categories one can apply the approach in Section 3.1. To model the effects of such a variable on y_i we include $r-1$ 0/1 dummy variables $x_{i1}^{(3)}, \ldots, x_{ir-1}^{(3)}$ in $X_i^{(3)}$, leaving the rth category as reference. We observe the marginal probabilities of the r categories at some aggregate level which we denote by $P_i^{(3)} = (p_{i1}^{(3)}, \ldots, p_{ir}^{(3)})'$.

To deal with the unordered categorical variable we build upon the multinomial probit [MNP] literature, see, for example, Hausman and Wise (1978) and Keane (1992). We introduce $r-1$ normally distributed variables $X_i^{(3)*} = (x_{i1}^{(3)*}, \ldots, x_{ir-1}^{(3)*})$ with

$$\begin{array}{rl} x_{i1}^{(3)} = 1 & \text{if} \quad x_{i1}^{(3)*} > \max(x_{i2}^{(3)*}, \ldots, x_{ir-1}^{(3)*}, 0) \quad \text{and} \quad x_{i1}^{(3)} = 0 \quad \text{otherwise} \\ \vdots & \\ x_{ir-1}^{(3)} = 1 & \text{if} \quad x_{ir-1}^{(3)*} > \max(x_{i1}^{(3)*}, \ldots, x_{ir-2}^{(3)*}, 0) \quad \text{and} \quad x_{ir-1}^{(3)} = 0 \quad \text{otherwise,} \end{array} \tag{13}$$

which means that $x_{i1}^{(3)} = \ldots = x_{ir-1}^{(3)} = 0$ if $\max(x_{i1}^{(3)*}, \ldots, x_{ir-1}^{(3)*}) \leq 0$. Hence, the vector $X_i^{(3)*}$ correspond exactly to the utility differences in MNP models. The distribution of $X_i^{(3)*}$ is given by

$$\begin{pmatrix} x_{i1}^{(3)*} \\ \vdots \\ x_{ir-1}^{(3)*} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_{i1}^{(3)*} \\ \vdots \\ \mu_{ir-1}^{(3)*} \end{pmatrix}, \begin{pmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{2} \\ \frac{1}{2} & \cdots & \frac{1}{2} & 1 \end{pmatrix} \right), \tag{14}$$

where $\mu_i^{(3)*} = (\mu_{i1}^{(3)*}, \ldots, \mu_{ir-1}^{(3)*})'$ represents the mean of $X_i^{(3)*}$. The covariance matrix has the same correlation structure as in an MNP model with an identity matrix for the

individual utilities. The observed probabilities imply $r-1$ restrictions on the distribution parameters of $X_i^{(3)*}$. To match the sample data with the model we have to solve $\mu_i^{(3)*}$ from

$$\begin{aligned} \Pr[x_{i1}^{(3)*} > x_{i2}^{(3)*}, \ldots, x_{i1}^{(3)*} > x_{ir-1}^{(3)*}, x_{i1}^{(3)*} > 0] &= p_{i1}^{(3)} \\ &\vdots \\ \Pr[x_{ir-1}^{(3)*} > x_{i1}^{(3)*}, \ldots, x_{ir-1}^{(3)*} > x_{ir-2}^{(3)*}, x_{ir-1}^{(3)*} > 0] &= p_{ir-1}^{(3)} \\ \Pr[x_{i1}^{(3)*} \leq 0, \ldots, x_{ir-1}^{(3)*} \leq 0] &= p_{ir}^{(3)}. \end{aligned} \tag{15}$$

Note that the last restriction is automatically satisfied if the first $r-1$ restrictions hold. Unfortunately, there is no closed form expression for the probabilities from the LHS of (15) and hence we have to use numerical methods. If r is small, numerical integration techniques can be used to evaluate the probabilities. For larger values of r the probabilities can be evaluated using the Stern simulator (Stern, 1992) or the Geweke-Hajivassiliou-Keane [GHK] simulator (Börsch-Supan and Hajivassiliou, 1993; Keane, 1994). The values of $\mu_i^{(3)*}$ can be found using a numerical solver. Notice that the values of $\mu_i^{(3)*}$ have to be determined only once before parameter estimation.

The equations (13) and (14) provide the latent variable model in case of one unordered categorical explanatory variable. In case one has more categorical variables one can easily extend the current solution in a similar way as discussed before. One can also correlate the $X_i^{(3)*}$ variables with the $X_i^{(1)*}$ and $X_i^{(2)*}$ variables in a straightforward manner.

3.4 Continuous explanatory variables

It may also be the case that the unobserved explanatory variable $x_i^{(4)}$ is a continuous variable. Even when the $x_i^{(4)}$ variable is continuous, the known distribution of the aggregated data $P_i^{(4)}$ is usually discrete. One may consider this variable as an ordered categorical variable. However, Breslaw and McIntosh (1998) argue that adding 0/1 dummies to (7) may not be the best option because it can lead to biased results. Instead, they opt for imputing the real continuous variable to the response equation (7). Using the approach in Section 3.2 this implies that we have to add the underlying latent variable for $x_i^{(4)}$, that is $x_i^{(4)*}$ or a function of $x_i^{(4)*}$ to $X_i^{(4)}$ in (7) instead of a set of 0/1 dummies. Note that the

assumption of a normal distribution may be harmful if the real distribution is different. It is of course also possible to assume another distributions for the latent variable $x_i^{(4)*}$. Correlating the $X_i^{(4)*}$ variable with $X_i^{(1)*}$, $X_i^{(2)*}$ and $X_i^{(3)*}$ can be done in a straightforward manner.

To summarize this section. The explanatory variables X_i which are missing at the individual level are described by the latent variable $X_i^* = (X_i^{(1)*\prime}, X_i^{(2)*\prime}, X_i^{(3)*\prime}, X_i^{(4)*\prime})'$. This latent variable has a multivariate normal distribution. The mean of this distribution is determined by the marginal probabilities at the aggregate level. The covariance matrix of X_i^* is denoted by Ω. The free elements in this matrix describe the correlations between the latent variables X_i^*. The models for the X_i^* variables together with (7) provide the complete model specification.

4 Parameter estimation

To estimate the model parameters of the model proposed in the previous section, we can choose for maximum likelihood or a Bayesian approach. In this section we discuss both approaches and their relative merits.

We first derive the likelihood function. Let the density function of the data y_i for the model in (7) conditional on the missing variables X_i be given by

$$f(y_i|X_i;\beta,\theta), \tag{16}$$

where β and θ denote the model parameters. To derive the unconditional density of y_i we have to sum over all possible values of X_i, which we will denote by the set χ. Hence, the density of y_i given the observed marginal probabilities P_i is given by

$$f(y_i|P_i;\beta,\theta,\rho) = \sum_{X_i \in \chi} \Pr[X_i|P_i;\rho] f(y_i|X_i;\beta,\theta), \tag{17}$$

where $\Pr[X_i|P_i;\rho]$ denotes the probability of observing X_i given the data at the aggregated level which we denote by $P_i = (P_i^{(1)\prime}, P_i^{(2)\prime}, P_i^{(3)\prime}, P_i^{(4)\prime})'$. These probabilities depend on the unknown parameters ρ which summarizes the free elements of the covariance matrix of the latent variables Ω, as discussed in the previous section. Hence, the log likelihood

function is given by

$$\mathcal{L}(y|P;\beta,\theta,\rho) = \sum_{i=1}^{N} \ln f(y_i|P_i;\beta,\theta,\rho), \tag{18}$$

where $y = (y_1,\ldots,y_N)'$ and $P = (P_1,\ldots,P_N)'$. The parameters β, θ and ρ have to be estimated from the data.

4.1 Maximum likelihood estimation

A maximum likelihood estimator can be obtained by maximizing the log likelihood function (18) with respect to (β,θ,ρ). To evaluate the log likelihood function we need to evaluate the probabilities $\Pr[X_i|P_i;\rho]$. Unfortunately, there is no closed form expression to compute these probabilities. For small dimensions one may use numerical integration techniques but in general we have to use simulation methods to evaluate the probabilities. This implies that we end up with a Simulated Maximum Likelihood [SML] estimator, see Lerman and Manski (1981). The probabilities $\Pr[X_i|P_i;\rho]$ can be evaluated using the Stern simulator (Stern, 1992) or the GHK simulator (Börsch-Supan and Hajivassiliou, 1993; Keane, 1994).

The SML estimator is consistent if the number of observations and the number of simulations goes to infinity. Given the literature on SML in MNP models (see for example, Geweke *et al.*, 1994), one may expect that obtaining an accurate values of the probabilities $\Pr[X_i|P_i;\rho]$ is computationally intensive, especially when the dimension of the latent X_i^* is large and/or the number of observations N is large. Note that the number of probabilities $\Pr[X_i|P_i;\rho]$ which need to be evaluated grows exponentially with the number of variables in X_i.

4.2 Bayesian analysis

The model can also be analyzed in a Bayesian framework. To obtain posterior results for the model parameters, we propose a Gibbs sampler (Geman and Geman, 1984) with data augmentation, see Tanner and Wong (1987). The latent X_i^* variables are simulated along side the model parameters (β,θ,ρ). The main advantage of this Bayesian approach is that it does not require the evaluation of the complete likelihood function. If suffices to evaluate the likelihood function conditional on the latent X_i^* which determine X_i.

We focus in this section on the sampling of the latent variable X_i^*. We assume that if we know the X_i^* and hence the X_i variables, a MCMC sampling scheme to simulate from the posterior distribution of the model parameters β and θ is available. Hence, we do not discuss simulating from the full conditional distribution of β and θ as this is model specific. We do however discuss simulating from the full conditional distribution of the ρ parameters as this is part of the model for the latent variable X_i^*.

4.2.1 Sampling of X_i^*

To sample X_i^* we have to derive its full conditional density which is given by

$$f(X_i^*|y_i, P_i; \beta, \theta, \rho) \propto f(y_i|X_i; \beta, \theta) f(X_i^*|P_i; \rho) f(\rho), \tag{19}$$

where $f(y_i|X_i; \beta, \theta)$ is given in (16), where $f(X_i^*|P_i; \rho)$ denotes the density of X_i^* implied by the latent variable model for X_i^*, and where $f(\rho)$ denotes the prior distribution for ρ. Given the structure of the latent variable model, X_i^* is a multivariate normal distribution with a mean μ_i which depends on P_i and a covariance matrix Ω where the free elements are denoted by ρ, that is,

$$f(X_i^*|P_i; \rho) = \phi(X_i^*; \mu_i(P_i), \Omega(\rho)), \tag{20}$$

where ϕ denotes the multivariate normal density function. Sampling the complete X_i^* vector at once is very difficult. Therefore, we sample the individual elements of X_i^* separately from their full conditional distribution. Let us consider the jth element of X_i^* denoted by x_{ij}^*. The full conditional density of x_{ij}^* is given by

$$f(x_{ij}^*|y_i, X_{i,-j}^*, P_i; \beta, \theta, \rho) \propto f(y_i|x_{ij}, X_{i,-j}; \beta, \theta) f(x_{ij}^*|X_{i,-j}^*, P_i; \rho), \tag{21}$$

where $X_{i,-j}^*$ and $X_{i,-j}$ denote the vector X_i^* and X_i without x_{ij}^* and $x_{i,-j}$, respectively.

The full conditional density of x_{ij}^* consists of two parts. The second part $f(x_{ij}^*|X_{i,-j}^*, P_i; \rho)$ is the conditional density of one of the elements of X_i^* which is of course a normal density with known mean, say, $\bar{\mu}_{ij}$, and variance, say, $\bar{s}_j^2$, which are functions of $\mu_i(P_i)$ and $\Omega(\rho)$. The first part $f(y_i|x_{ij}, X_{i,-j}; \beta, \theta)$ is a function of X_i and can take a discrete number of values depending on the value of x_{ij}^*. If we for simplicity assume that x_{ij}^* corresponds to a

binary explanatory variable, x_{ij} can take two values depending one whether x_{ij}^* is larger or smaller than 0.

To sample from the full conditional posterior distribution of x_{ij}^* we use the inverse CDF technique. In Appendix A we derive the inverse CDF of x_{ij}^*. The simulation scheme for $x_{ij}^{(1)*}$ corresponding to the binary variable $x_{ij}^{(1)}$ can be summarized as follows

1. Draw $u \sim U(0,1)$

2. Set

$$x_{ij}^{(1)*} = \begin{cases} \Phi^{-1}\left(\frac{u}{c_i k_{i0}}\right)\bar{s}_j + \bar{\mu}_{ij} & \text{if } u \leq c_i k_{i0}\Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right) \\ \Phi^{-1}\left(\frac{u}{c_i k_{i1}} + \frac{k_{i1}-k_{i0}}{k_{i1}}\Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right)\right)\bar{s}_j + \bar{\mu}_{ij} & \text{if } u > c_i k_{i0}\Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right) \end{cases},$$

where $k_{i0} = f(y_i|x_{ij}^{(1)} = 0, X_{i,-j}; \beta, \theta)$, $k_{i1} = f(y_i|x_{ij}^{(1)} = 1, X_{i,-j}; \beta, \theta)$, and c_i is the integrating constant of the full conditional distribution of $x_{ij}^{(1)*}$ given by

$$c_i = \left((k_{i0} - k_{i1})\Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right) + k_{i1}\right)^{-1}. \tag{22}$$

Appendix A also provides the derivations of the inverse CDF in case x_{ij}^* is associated with an ordered or an unordered categorical variable.

4.2.2 Sampling of ρ

To complete the Gibbs sampler, we need to sample the parameters in ρ from their full conditional posterior distribution. The vector ρ contains the free elements of the covariance matrix of X_i^* which is denoted by Ω. As discussed in Section 3, identification requires several restrictions on the covariance matrix Ω. For example, all diagonal elements of Ω are equal to 1 and hence Ω is a correlation matrix. Furthermore, the correlations between elements of the same unordered categorical variable are set equal to 1/2. Hence, the full conditional distribution of Ω is not an inverted Wishart distribution.

There are some algorithms to draw a correlation matrix, see, for example, Chib and Greenberg (1998), Manchanda *et al.* (1999), and Liechty *et al.* (2004). In this paper we follow Barnard *et al.* (2000). They suggest sampling one correlation at a time from their full conditional posterior distribution using a griddy-Gibbs sampler, see Ritter and Tanner (1992).

Suppose we want to draw the jth correlation in ρ denoted by ρ_j. Denote the vector ρ without ρ_j as ρ_{-j}. Furthermore, let $X^* = (X_1^*, \ldots, X_N^*)'$ and $\mu = (\mu_1, \ldots, \mu_N)'$, where μ_i denotes the mean of X_i^* and is a function of P_i, for $i = 1, \ldots, N$. The full conditional posterior density of ρ_j is given by

$$\begin{aligned} f(\rho_j|\rho_{-j}, X^*) &\propto f(X^*|\rho_j, \rho_{-j}) f(\rho_j|\rho_{-j}) \\ &\propto \prod_{i=1}^{N} \phi(X_i^*; \mu_i, \Omega(\rho_j, \rho_{-j})) f(\rho_j|\rho_{-j}) \\ &\propto |\Omega(\rho_j, \rho_{-j})|^{-\frac{N}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}((X^* - \mu)'(X^* - \mu)\Omega(\rho_j, \rho_{-j})^{-1})\right] f(\rho_j|\rho_{-j}). \end{aligned} \tag{23}$$

Barnard *et al.* (2000) show how to determine the range of values for which ρ_j leads to a positive definite matrix. Within this range we can define a set of grid points to evaluate the kernel (23) for the griddy-Gibbs sampler.

As correlations in Ω which are related to the jth explanatory variable are not identified if $\beta_j = 0$, we have to impose an informative prior for the parameters in ρ. We use a (truncated) standard normal prior for the parameters in ρ, that is,

$$f(\rho_j) \propto \exp(\rho_j^2). \tag{24}$$

Hence, we concentrate the probability mass around zero.

5 Application

To illustrate our approach, we consider in this section an application where we analyze the characteristics of households who donate to a large Dutch charity in the health sector. Households receive a direct mailing form the charity with a request to donate money. The household may not respond and donate nothing or respond and donate a positive amount. We have no information about the characteristics of the households apart from their zip code. At the zip-code level we know aggregated household characteristics.

Our sample contains 10,000 households which are randomly selected. The mailing took place in February 1997. The response rate is 39.4%. The average donation is 3.21 euros and the average donation conditional on response is 8.15 euros. We match these data with aggregated data at the zip-code level (4 digits) from Statistics Netherlands (CBS).

Table 2: Available explanatory variables at the zip-code level

Variable	Type	Description
Church	Binary	Goes to church every week
Not-active	Binary	Not active in labor force
Reference: Family with kids		
Single	Unordered (3 cat.)	Lives alone
Family no kids	Unordered (3 cat.)	Family without kids
Reference: Average income		
Income low	Ordered (3 cat.)	Income in lowest 40% nationally
Income high	Ordered (3 cat.)	Income in highest 20% nationally
Reference: Age between 25 and 44		
Age 0-24	Ordered (4 cat.)	Age between 0 and 24
Age 45-64	Ordered (4 cat.)	Age between 45 and 64
Age 65+	Ordered (4 cat.)	Age over 65
Urbanization	Observed	Measure for the degree of urbanization

Table 2 shows the relevant aggregated data at the zip-code level. As can be observed from the table we have aggregated data for different types of explanatory variables, that is, for binary, unordered, and ordered categorical variables. Note that we only know urbanization level at the zip-code level. As it is the same for each individual in the zip-code region, this variable is treated as an observed variable.

To describe donating behavior we consider a censored regression (Tobin, 1958) because the donated amount is censored at 0. We use the log of (1 + amount) as dependent variable which leads to the following model specification

$$\begin{array}{lll} \ln(1+y_i) = x_i'\beta + \varepsilon_i & \text{if} & x_i'\beta + \varepsilon_i > 0 \\ \ln(1+y_i) = 0 & \text{if} & x_i'\beta + \varepsilon_i \leq 0 \end{array} \tag{25}$$

with $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. As explanatory variables we take the variables displayed in Table 2.

To estimate the effects of the covariates on response, we use two approaches . First, we follow the simple regression approach of Section 2, which means that we replace the unknown household characteristics by their sample averages at the zip-code level. The parameters of (25) are estimated using ML. Although we have only shown in Section 2 that OLS in a linear regression model provides consistent estimates, simulations suggest that this results carries over to the ML estimator in a censored regression model. Secondly, we use the mixture approach to estimate the censored regression parameters, where we opt for

Table 3: Posterior means, posterior standard deviations, and HPD regions of the model parameters for the mixture approach together with ML results for the simple approach

	mixture approach			simple approach	
	mean	st.dev.	95% HPD	ML	s.e.[a]
Intercept	-0.36	0.008	(-0.37,-0.34)	-2.19	0.82
Urbanization	-0.00	0.008	(-0.02, 0.01)	0.38	0.34
Church	0.33	0.003	(0.32, 0.34)	0.15	0.24
Not-active	0.70	0.005	(0.69, 0.71)	0.24	0.70
Single	-1.14	0.010	(-1.16,-1.12)	0.57	0.56
Family no kids	2.07	0.008	(2.06, 2.09)	2.82	1.17
Income low	-1.83	0.005	(-1.84,-1.82)	-1.30	1.11
Income high	0.59	0.003	(0.58, 0.59)	1.18	0.83
Age 0-24	0.73	0.006	(0.72, 0.74)	2.62	1.35
Age 45-65	-0.57	0.003	(-0.58,-0.57)	-0.35	1.24
Age 65+	0.22	0.004	(0.21, 0.23)	1.60	1.00
σ	0.06	0.001	(0.06, 0.06)	2.26	0.02

[a] Heteroskedasticity consistent standard errors, see White (1982).

a Bayesian approach. We use an uninformative prior for β and σ^2, that is, $p(\beta, \sigma^2) \propto \sigma^{-2}$ and the informative prior (24) for the ρ parameters. We use a total of 110,000 draws, of which the first 50,000 were used as burn-in period. Furthermore, we only used every 12th draw to obtain an approximately random sample of 5,000 draws.

Table 3 displays the estimation results for both approaches. It is clear from the table that the posterior standard deviations of the mixture approach are much smaller than the standard errors of the ML estimator, where the unknown household characteristics are replaced by their sample averages at the zip-code level. Although the number of observations is very high, the estimated standard errors are very large. This illustrates the huge efficiency gain of using our method. This efficiency gain enables us to identify more significant influences from household characteristics. When using the simple approach, only *Family no kids* is identified as having a significant impact on the donating behavior. But, using our mixture approach it becomes clear that many other household characteristics also influence this decision. Being Religious and not being active in the labor force has a positive effect. Note that latter group also contain retired people. Being single has a negative effect, while families without children tend to donate more. Household with

Table 4: Posterior means of the correlations between unobserved variables with posterior standard deviations in parenthesis

	Church	Not-active	Single	Family no kids	Income	Age
Church	1 (-)					
Not-active	-0.19 (0.05)	1 (-)				
Single	-0.24 (0.05)	0.68 (0.02)	1 (-)			
Family no kids	-0.23 (0.03)	0.15 (0.03)	0.50 (-)	1 (-)		
Income	-0.08 (0.04)	-0.25 (0.03)	0.43 (0.02)	0.75 (0.02)	1 (-)	
Age	-0.12 (0.06)	-0.82 (0.02)	-0.24 (0.03)	0.23 (0.03)	0.67 (0.02)	1 (-)

higher income tend to donate more, while the effect of age is nonlinear. The highest posterior density [HPD] interval shows that urbanization grade has no influence on donating behavior.

Table 4 displays the estimated correlation matrix of X_i^*. The diagonal elements are put equal to 1 for identification. The correlation between the variables *Single* and *Family no kids* is fixed at 1/2, because they belong to the same unordered categorical variable. Most of the posterior means of the correlations are more than two times larger than their posterior standard deviation, illustrating the importance of our approach. The correlations usually have the expected sign. For example, there is a negative correlation between being not active and income, and a negative correlation between being single and age.

6 Conclusions

In this paper we have developed a new approach to estimate the effects of explanatory variables on individual response where the response variable is observed at the individual level but the explanatory variables are only observed at some aggregated level. This approach can, for example, be used if information about individual characteristics is only available at the zip-code level. To solve the limited data availability, we extend

the model describing individual responses with a latent variable model to describe the missing individual explanatory variables. The latent variable model is of the probit type and matches the sample information of the explanatory variables at the aggregated level. Parameter estimates for the effects of the explanatory variables in the individual response model can be obtained using maximum likelihood or a Bayesian approach.

A simulation study shows that our new approach clearly outperforms a standard approach in efficiency. The efficiency loss which is due to aggregation is about 50% smaller than for the standard method. We illustrated the merits of our approach by estimating the effects of the household characteristics on donating behavior to a Dutch charity. For this application we used data of donating behavior at the household level, while the covariates were only observed at the zip-code level.

There are several ways for future research. It may be interesting to investigate whether the proposed method can be used to deal with nonresponse in survey data. Another topic for future research is to consider the complement case where explanatory variables are observed at the individual level but that the response variable is only observed at some aggregated level.

A Derivation of full conditional distributions

In this appendix we give a short derivation for the full conditional posterior distributions in case the missing explanatory variable is a binary variable, an ordered categorical variable and an unordered categorical variable. As starting point we take the general form of the full conditional density of x_{ij}^* given in (21).

A.1 Binary variable

In case x_{ij} is a binary variable, the posterior distribution of $x_{ij}^{(1)*}$ can be summarized by

$$f(x_{ij}^{(1)*}|y_i, X_{i,-j}^*, P_i; \beta, \theta, \rho) = \\ c_i(k_{i0}\phi(x_{ij}^{(1)*}; \bar{\mu}_{ij}, \bar{s}_j^2)I[x_{ij}^{(1)*} \leq 0] + k_{i1}\phi(x_{ij}^{(1)*}; \bar{\mu}_{ij}, \bar{s}_j^2)I[x_{ij}^{(1)*} > 0]), \quad (26)$$

where X_{i-j}^* denotes X_i^* without $x_{ij}^{(1)*}$, $k_{i0} = f(y_i|x_{ij}^{(1)} = 0, X_{i,-j}; \beta, \theta)$, $k_{i1} = f(y_i|x_{ij}^{(1)} = 1, X_{i-j}; \beta, \theta)$, $\phi(x; m, s)$ denotes the normal density function with mean m and variance s evaluated in x, $\bar{\mu}_{ij}$ denotes the conditional mean of $x_{ij}^{(1)*}|X_{i,-j}^*, \rho$, and where $\bar{s}_j^2$ denotes the conditional variance of $x_{ij}^{(1)*}|X_{i,-j}^*, \rho$. The integrating constant c_i follows from

$$\begin{aligned} c_i^{-1} &= \int_{-\infty}^{0} k_{i0}\phi(x_{ij}^{(1)*}, \bar{\mu}_{ij}, \bar{s}_j^2)dx_{ij}^{(1)*} + \int_{0}^{\infty} k_{i1}\phi(x_{ij}^{(1)*}, \bar{\mu}_{ij}, \bar{s}_j^2)dx_{ij}^{(1)*} \\ c_i^{-1} &= k_{i0}\Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right) + k_{i1}\left(1 - \Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right)\right) \\ c_i &= \left((k_{i0} - k_{i1})\Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right) + k_{i1}\right)^{-1}, \end{aligned} \quad (27)$$

where Φ denotes the standard normal distribution function. Hence, the full conditional density of $x_{ij}^{(1)*}$ is given by

$$f(x_{ij}^{(1)*}|y_i, X_{i,-j}^*, P_i; \beta, \theta, \rho) = \begin{cases} c_i k_{i0}\phi(x_{ij}^{(1)*}, \bar{\mu}_{ij}, \bar{s}_j) & \text{if } x_{ij}^{(1)*} \leq 0 \\ c_i k_{i1}\phi(x_{ij}^{(1)*}, \bar{\mu}_{ij}, \bar{s}_j) & \text{if } x_{ij}^{(1)*} > 0 \end{cases} \quad (28)$$

and the distribution function reads

$$F(x_{ij}^{(1)*}|y_i, X_{i,-j}^*, P_i; \beta, \theta, \rho) = \begin{cases} c_i k_{i0}\Phi\left(\frac{x_{ij}^{(1)*} - \bar{\mu}_{ij}}{\bar{s}_j}\right) & \text{if } x_{ij}^{(1)*} \leq 0 \\ c_i\left[(k_{i0} - k_{i1})\Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right) + k_{i1}\Phi\left(\frac{x_{ij}^{(1)*} - \bar{\mu}_{ij}}{\bar{s}_j}\right)\right] & \text{if } x_{ij}^{(1)*} > 0. \end{cases} \quad (29)$$

To draw from this distribution we use the inverse CDF technique, which means that we draw $u \sim U(0,1)$ and solve $x_{ij}^{(1)*}$ from $F(x_{ij}^{(1)*}|y_i, X_{i,-j}^*, P_i; \beta, \theta, \rho) = u$. The inverse function of (29) is given by

$$x_{ij}^{(1)*}(u) = \begin{cases} \Phi^{-1}\left(\frac{u}{c_i k_{i0}}\right)\bar{s}_j + \bar{\mu}_{ij} & \text{if } u \leq c_i k_{i0}\Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right) \\ \Phi^{-1}\left(\frac{u}{c_i k_{i1}} + \frac{k_{i1}-k_{i0}}{k_{i1}}\Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right)\right)\bar{s}_j + \bar{\mu}_{ij} & \text{if } u > c_i k_{i0}\Phi\left(\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right). \end{cases} \tag{30}$$

A.2 Ordered categorical variable

For an ordered categorical variable with r categories, we have to sample the variable $x_{ij}^{(2)*}$. If we assume that r is the reference category, the $x_{ij}^{(2)*}$ variable determines the $r-1$ 0/1 dummy variables $x_{i1}^{(2)}, \ldots, x_{ir-1}^{(2)}$. Let $P_i^{(2)} = (p_{i1}^{(2)}, \ldots, p_{ir}^{(2)})'$ denote the observed marginal probabilities that the individual belongs to the r categories. The threshold levels q_{it} are equal to $\Phi^{-1}(\sum_{l=1}^{t} p_{il}^{(2)})$ for $i = 1, \ldots, N$ and $t = 1, \ldots, r-1$. Let $\bar{\mu}_{ij}$ denote the conditional mean of $x_{ij}^{(2)*}|X_{i,-j}^*, \rho$ in the latent model and let $\bar{s}_j^2$ denote the conditional variance of $x_{ij}^{(2)*}|X_{i,-j}^*, \rho$, where $X_{i,-j}^*$ denotes X_i^* without $x_{ij}^{(2)*}$.

Sampling of $x_{ij}^{(2)*}$ proceeds in the same way as for the binary variables except for the fact that there are now r possible values for $f(y_i|X_i; \beta, \theta)$ instead of only 2. The r values for $f(y_i|X_i; \beta, \theta)$ are given by

$$\begin{aligned} k_{i1} &= f(y_i|X_{i,-j}, x_{i1}^{(2)} = 1, x_{i2}^{(2)} = 0, \ldots, x_{ir-1}^{(2)} = 0; \beta, \theta) \\ &\vdots \\ k_{ir-1} &= f(y_i|X_{i,-j}, x_{i1}^{(2)} = 0, \ldots, x_{ir-2}^{(2)} = 0, x_{ir-1}^{(2)} = 1; \beta, \theta) \\ k_{ir} &= f(y_i|X_{i,-j}, x_{i1}^{(2)} = 0, \ldots, x_{ir-1}^{(2)} = 0; \beta, \theta), \end{aligned}$$

where $X_{i,-j}$ denotes X_i without $x_{i1}^{(2)}, \ldots, x_{ir-1}^{(2)}$. The integrating constant now equals

$$c_i = \left(\sum_{t=1}^{r-1}(k_{it} - k_{it+1})\Phi\left(\frac{q_{it} - \bar{\mu}_{ij}}{\bar{s}_j}\right) + k_{ir}\right)^{-1}. \tag{31}$$

Straightforward derivation leads to following inverse CDF

$$x_{ij}^{(2)*}(u) = \begin{cases} \Phi^{-1}\left(\frac{u}{c_i k_{i1}}\right)\bar{s}_j + \bar{\mu}_{ij} & \text{if } u \leq \bar{u}_1 \\ \Phi^{-1}\left[\frac{u}{c_i k_{i2}} - \frac{k_{i1}-k_{i2}}{k_{i2}}\Phi\left(\frac{q_{i1}-\bar{\mu}_{ij}}{\bar{s}_j}\right)\right]\bar{s}_j + \bar{\mu}_{ij} & \text{if } \bar{u}_1 < u \leq \bar{u}_2 \\ \vdots & \vdots \\ \Phi^{-1}\left[\frac{u}{c_i k_{ir}} - \sum_{t=1}^{r-1}\frac{k_{it}-k_{it+1}}{k_{ir}}\Phi\left(\frac{q_{it}-\bar{\mu}_{ij}}{\bar{s}_j}\right)\right]\bar{s}_j + \bar{\mu}_{ij} & \text{if } \bar{u}_{r-1} < u, \end{cases} \tag{32}$$

where

$$\begin{aligned}
\bar{u}_1 &= c_i k_{i1} \Phi\left(\frac{q_{i1} - \bar{\mu}_{ij}}{\bar{s}_j}\right) \\
\bar{u}_2 &= c_i k_{i2} \Phi\left(\frac{q_{i2} - \bar{\mu}_{ij}}{\bar{s}_j}\right) + c_i (k_{i1} - k_{i2}) \Phi\left(\frac{q_{i1} - \bar{\mu}_{ij}}{\bar{s}_j}\right) \\
&\vdots \\
\bar{u}_{r-1} &= c_i k_{ir-1} \Phi\left(\frac{q_{ir-1} - \bar{\mu}_{ij}}{\bar{s}_j}\right) + c_i \sum_{t=1}^{r-2} (k_{it} - k_{it+1}) \Phi\left(\frac{q_{it} - \bar{\mu}_{ij}}{\bar{s}_j}\right).
\end{aligned}$$

A.3 Unordered categorical variable

For an unordered categorical variable with r categories, we add $r-1$ 0/1 dummies in $X_i^{(3)}$, say, $x_{i1}^{(3)}, \ldots, x_{ir-1}^{(3)}$. The $r-1$ normal distributed random variables which belong to this unordered categorical variable are denoted by $(x_{i1}^{(3)*}, \ldots, x_{ir-1}^{(3)*})'$.

Suppose that we want to sample $x_{ij}^{(3)*}$ from its full conditional posterior distribution. The full conditional posterior density is given by (21). The second part $f(x_{ij}^{(3)*}|X_{i,-j}^*, P_i; \rho)$, where X_{i-j}^* denotes X_i^* without $x_{ij}^{(3)*}$, is a normal density with known mean, say $\bar{\mu}_{ij}$, and standard deviation, say $\bar{s}_j$. Conditional on $X_{i,-j}^*$ the first part $f(y_i|x_{ij}, X_{i,-j}; \beta, \theta)$ can take two values. Define $x_{il}^{(3)*} = \max(x_{i1}^{(3)*}, \ldots, x_{ij-1}^{(3)*}, x_{ij+1}^{(3)*}, \ldots, x_{ir-1}^{(3)*})$. The two possible values for $f(y_i|x_{ij}, X_{i,-j}; \beta, \theta)$ are given by

$$\begin{aligned}
k_{i0} &= f(y_i|x_{i1}^{(3)} = 0, \ldots, x_{il-1}^{(3)} = 0, x_{il}^{(3)} = 1, x_{il+1}^{(3)} = 0, \ldots x_{ir-1}^{(3)} = 0, X_{i,-j}; \beta, \theta) I[x_{il}^{(3)*} > 0] + \\
&\qquad f(y_i|x_{i1} = 0, \ldots, x_{ir-1} = 0, X_{i,-j}; \beta, \theta) I[x_{il}^{(3)*} \leq 0] \\
k_{i1} &= f(y_i|x_{i1}^{(3)} = 0, \ldots, x_{ij-1}^{(3)} = 0, x_{ij}^{(3)} = 1, x_{ij+1}^{(3)} = 0, \ldots x_{ir-1}^{(3)} = 0, X_{i,-j}; \beta, \theta),
\end{aligned}$$

where $X_{i,-j}$ denotes X_i without $x_{i1}^{(3)}, \ldots, x_{ir-1}^{(3)}$. The full conditional posterior density of $x_{ij}^{(3)*}$ is given by

$$c_i(k_{i0}\phi(x_{ij}^{(3)*}; \bar{\mu}_{ij}, \bar{s}_j) I[x_{ij}^{(3)*} \leq \max(x_{il}^{(3)*}, 0)] + k_{i1}\phi(x_{ij}^{(3)*}; \bar{\mu}_{ij}, \bar{s}_j) I[x_{ij}^{(3)*} > \max(x_{il}^{(3)*}, 0)]). \tag{33}$$

The integrating constant c_i is given by

$$c_i = \left((k_{i0} - k_{i1}) \Phi\left(\frac{\max(x_{il}^{(3)*}, 0) - \bar{\mu}_{ij}}{\bar{s}_j}\right) + k_{i1}\right)^{-1}. \tag{34}$$

Straightforward derivation leads to the following inverse CDF

$$x_{ij}^{(3)*}(u) = \begin{cases} \Phi^{-1}\left(\frac{u}{c_i k_{i0}}\right)\bar{s}_j + \bar{\mu}_{ij} & \text{if } u \leq c_i k_{i0}\Phi\left(\frac{\max(x_{il}^{(3)*},0)-\bar{\mu}_{ij}}{\bar{s}_j}\right) \\ \Phi^{-1}\left(\frac{u}{c_i k_{i1}} + \frac{k_{i1}-k_{i0}}{k_{i1}}\Phi\left(\frac{\max(x_{il}^{(3)*},0)-\bar{\mu}_{ij}}{\bar{s}_j}\right)\right)\bar{s}_j + \bar{\mu}_{ij} & \text{if } u > c_i k_{i0}\Phi\left(\frac{\max(x_{il}^{(3)*},0)-\bar{\mu}_{ij}}{\bar{s}_j}\right). \end{cases} \tag{35}$$

References

Aitchison, J. and S. Silvey (1957), The generalization of probit analysis to the case of multiple responses, *Biometrika*, **44**, 131–140.

Amemiya, T. (1974), Bivariate probit analysis: Minimum chi-square methods, *Journal of the American Statistical Association*, **69**, 940–944.

Ashford, J. and R. Sowden (1970), Multi-variate probit analysis, *Biometrics*, **26**, 536–546.

Barnard, J., R. McCulloch, and X. Meng (2000), Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage, *Statistica Sinica*, **10**, 1281–1311.

Billard, L. and E. Diday (2003), From the statistics of data to the statistics of knowledge: Symbolic data analysis, *Journal of the American Statistical Association*, **98**, 470–487.

Börsch-Supan, A. and V. Hajivassiliou (1993), Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models, *Journal of Econometrics*, **58**, 347–368.

Breslaw, J. and J. McIntosh (1998), Simulated latent variable estimation of models with ordered categorical data, *Journal of Econometrics*, **87**, 25–47.

Chib, S. and E. Greenberg (1998), Analysis of multivariate probit models, *Biometrika*, **85**, 347–361.

Dempster, A., N. Laird, and R. Rubin (1977), Maximum Likelihood estimation from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, series B*, **39**, 1–38.

Everitt, B. and D. Hand (1981), *Finite mixture distributions*, Chapman and Hall, London.

Geman, S. and D. Geman (1984), Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geweke, J., M. Keane, and D. Runkle (1994), Alternative computational approaches to inference in the multinomial probit model, *The Review of Economics and Statistics*, **76**, 609–632.

Hausman, J. and D. Wise (1978), A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and hetrogeneous preferences, *Econometrica*, **46**, 403–426.

Imbens, G. and T. Lancaster (1994), Combining micro and macro data in microeconometric models, *Review of Economic Studies*, **61**, 655–680.

Joe, H. (1997), *Multivariate models and dependence concepts*, Chapman and Hall, London.

Keane, M. (1992), A note on identification in the multinomial probit model, *Journal of Business & Economic Statistics*, **10**, 193–200.

Keane, M. (1994), A computationally practical simulation estimator for panel data, *Econometrica*, **62**, 95–116.

Lerman, S. and C. Manski (1981), On the use of simulated frequencies to approximate choice probabilities, in C. Manski and D. McFadden (eds.), *Structural analysis of discrete data with econometric applications*, MIT press, Cambridge, MA.

Liechty, J., M. Liechty, and P. Müller (2004), Bayesian correlation estimation, *Biometrika*, **91**, 1–14.

Manchanda, P., A. Ansari, and S. Gupta (1999), The "shopping basket": A model for multicategory purchase incedence decisions, *Marketing Science*, **18**, 95–114.

Quandt, R. and J. Ramsey (1978), Estimating mixtures of normal distributions and switching regressions, *Journal of the American Statistical Association*, **73**, 730–738.

Ritter, C. and M. Tanner (1992), Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler, *Journal of the American Statistical Association*, **87**, 861–868.

Romeo, C. (2005), Estimating discrete joint probability distributions for demographic characteristics at the store level given store level marginal distributions and a city-wide joint distribution, *Quantitative Marketing and Economics*, **3**, 71–93.

Steenburgh, T., A. Ainslie, and P. Engebretson (2003), Massively categorical variables: Revealing the information in zip codes, *Marketing Science*, **22**, 40–57.

Stern, S. (1992), A method for smoothing simulated moments of discrete probabilities in multinomial probit models, *Econometrica*, **60**, 943–952.

Tanner, M. and W. Wong (1987), The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, **82**, 528–540.

Titterington, D., A. Smith, and U. Makov (1985), *Statistical analysis of finite mixture distributions*, Wiley, New York.

Tobin, J. (1958), Estimation of relationships for limited dependent variables, *Econometrica*, **26**, 24–36.

van den Berg, G. and B. van der Klaauw (2001), Combining micro and macro unemployment duration data, *Journal of Econometrics*, **102**, 271–309.

Wakefield, J. (2004), Ecological inference for 2 x 2 tables, *Journal of the Royal Statistical Society, series A*, **167**, 385–445.

White, H. (1982), Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 1–25.