Explaining individual response using aggregated data

Dijk, Bram van; Paap, Richard

Postprint / Postprint
Zeitschriftenartikel / journal article

Explaining individual response using aggregated data

Bram van Dijk, Richard Paap

PII: S0304-4076(08)00061-4
DOI: 10.1016/j.jeconom.2008.05.008
Reference: ECONOM 3041

To appear in: *Journal of Econometrics*

Received date: 17 February 2006
Revised date: 24 April 2008
Accepted date: 16 May 2008

Please cite this article as: van Dijk, B., Paap, R., Explaining individual response using aggregated data. *Journal of Econometrics* (2008), doi:10.1016/j.jeconom.2008.05.008

Explaining Individual Response using Aggregated Data

Bram van Dijk*
Econometric Institute
Tinbergen Institute
Erasmus University Rotterdam

Richard Paap
Econometric Institute
Erasmus University Rotterdam

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 when 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 analysis. 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

*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. For survey data, there may be a missing explanatory variable due to a missing question in the survey or a question which is interpreted the wrong way by the respondents.

In some cases it is possible to obtain information on explanatory variables at some aggregated level. For example, if the zip code of households is known, it may be possible to 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. There are several studies in economics which try to link individual and aggregated data, see, for example, Imbens and Lancaster (1994) and van den Berg and van der Klaauw (2001). In contrast to our situation, these studies 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. The situation of missing individual covariates is related to 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 ecological inference also relies on aggregated information on the response variable. The extra information on individual responses may help us to overcome certain identification issues in ecological inference.

As far as we know, the only paper that comes close to our situation is Steenburgh et al. (2003). The motivation for the use of aggregated data in this paper is however different

from ours. The authors use zip-code information to describe unobserved heterogeneity in the individual behavior of households instead of estimating the effects of covariates on behavior. Our problem also bears 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.

In this paper we develop a new approach to estimate the effects of covariates on individual response when the covariates are unknown at the individual level but observed at some aggregated level in the form of marginal probabilities. We extend the standard individual response model with a latent variable model describing the missing 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 standard mixture regression with known mixing proportions. A simple simulation experiment shows that this new approach outperforms in efficiency the standard method, where we replace the missing explanatory variables by the observed marginal probabilities at the aggregated level.

Parameter estimates of the individual 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, we use a Gibbs sampler with data augmentation. The unobserved explanatory variables are sampled alongside the model parameters. Conditional on the sampled explanatory variables, a standard Markov Chain Monte Carlo [MCMC] sampler can be used 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 in general consistent, it is clear from (2) that the error term is heteroskedastic, and hence the OLS estimator is not efficient. Hence, a GLS estimator may improve upon OLS estimates.

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) and

Titterington et al. (1985). To describe the response variable y_i we consider a mixture of two linear 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 in 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 $\varepsilon_i \sim \mathrm{N}(0, 1)$.

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 1,000 replications. The efficiency loss is computed using the root mean squared error [RMSE] of the estimates as all estimators are consistent. Several conclusions can be drawn from

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. Total number of individuals is 1,000.

the table. First of all, the mixture 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 mixture approach however seems hardly affected by the number of individuals with the same value for p_i.

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 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 and show that the OLS estimator for β_j is consistent when the

Table 2: Average RMSE of forecasting y_i using the parameter estimates of 4 methods and using either individual, aggregated or a Bayesian update of the individual data in the forecasts

	using x_i	using p_i	using $\widehat{x_i \mid y_i}$
Full information	1.00	1.35	1.18
OLS	1.06	1.35	1.22
GLS	1.06	1.35	1.22
Mixture	1.00	1.35	1.18

[a] 1,000 out-of-sample observations.

x_{ij} are replaced by p_{ij}. In this case, 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.

Before we turn to our solution to this problem, we first consider forecasting. Forecasting individual response when only aggregated explanatory variables are available is hampered because of two reasons. First, the effects of the explanatory variables can be estimated less precise compared to the case where individual data is available. The second reason is that the lack of out-of-sample explanatory variables at the individual level introduces more uncertainty in our forecast. To assess the out-of-sample forecasting performance of the four estimation methods, we simulate another set of 1,000 y_i values for each replication. We predict the value of y_i using the estimates of α and β obtained in the first part of the simulation for each of the four estimation procedures.

Table 2 displays the average RMSE for each of the four estimation procedures. We only show the results where we simulate p_i from $U(0.2, 0.4)$ and where we draw a distinct value for each individual. The other cases show similar results. We make a distinction between three situations. The second column displays the results when we assume that the out-of-sample x_i are known. In this case the full information approach and the mixture

approach have similar average RMSE while the OLS and GLS approach perform worse. In case we only use aggregated out-of-sample information, all approaches perform the same, see third column of Table 2. Hence, the loss in forecast precision due to having aggregated out-of-sample information outweighs the efficiency loss in parameter estimation. The final column shows the results in case we only simulate new y_i values using the same x_i values of the original sample. This allows us to estimate the value of x_i given the in-sample information via Bayesian updating. Note that this is only possible in the case of a panel data set and time-invariant x_i variables. Again, the full information approach and the mixture approach have similar average RMSE while the OLS and GLS approach perform worse.

We can conclude from the simulation experiments in this section that the mixture approach is preferred when we want to estimate the effects of explanatory variables which are only observed at the aggregated level on individual response. In the next section we extend the mixture approach to situation of more than one explanatory variable. The information in the individual responses helps 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 types of explanatory variables like ordered and unordered categorical variables. The vector of explanatory variables is written as $X_i = (X_i^{(1)\prime}, X_i^{(2)\prime}, X_i^{(3)\prime})'$, where $X_i^{(1)}$ contains the binary explanatory variables, $X_i^{(2)}$ the ordered categorical explanatory variables and $X_i^{(3)}$ the unordered 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. We assume that ε_i is independent of X_i.

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 know the marginal distribution of each X_i, which may or may not vary across individuals. To estimate the model parameters β and θ we extend (7) with a latent variable model describing 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 three different types of explanatory variables. Note that we only discuss them separately to facilitate the exposition. The different types of variables can easily be combined in one multivariate model.

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 of which the associated probabilities sum up to 1. If we observe these 2^k 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. 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.

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 distribution function 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. In sum, we assume that

$$X_i^{(1)*} \sim \mathrm{N}\left(\Phi^{-1}(P_i^{(1)}), \Omega_{11}\right), \tag{9}$$

where Ω_{11} 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, that is, the sub-diagonal elements of Ω_{11}. 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) and Chib and Greenberg (1998). The aggregated data provides the values of the intercepts such that only the sub-diagonal elements of Ω_{11} 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, we usually have a combination of several binary and ordered categorical variables and hence we need to deal with correlation between these variables. To describe correlations between several categorical variables, it is convenient to introduce a normal distributed random variable $x_i^{(2)*}$ and 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), see also Cowles (1996) for a Bayesian estimation procedure.

The equations (10)–(12) provide the latent variable model for the case of one ordered categorical explanatory variable. In case we have more categorical variables it is easy to extend the current solution with more latent $x_{ij}^{(2)*}$ variables and allow them to correlate using a covariance matrix Ω_{22} with ones on the diagonal. It is also possible to 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

We may also encounter an explanatory variable which is categorical with, say, r categories, but without a natural ordering in the categories. We assume here that an individual can only belong to one category. If (s)he can belong to several categories we 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

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

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 $X_i^{(3)*}$. As individuals can only belong to one of the categories, we cannot identify the covariance matrix of $X_i^{(3)*}$ and have to fix its elements. For simplicity we take the implied covariance matrix of an MNP model where the individual utilities have a covariance equal to 1/2 times the identity matrix. If we take category r as the base category we end up with same covariance structure as above. The positive correlations are caused by the fact that the value of $x_{ij}^{(3)*}$ is influenced by both $p_{ij}^{(3)}$ and the probability of belonging to the reference category $p_{ir}^{(3)}$. If the reference has a vary small probability, all $x_{ij}^{(3)*}, j = 1, \ldots, r-1$ should have a high value.

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 (1992) simulator 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 there are more categorical variables it is easy to extend the current solution in a similar way as discussed before. It is also possible to 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

So far, we only used discrete explanatory variables. Dealing with the case where continuous variables are not observed at the individual level but at some aggregated level is not easy in practice. It is not enough to know the average value of the continuous variable at some aggregate level (e.g. the average value in each zip-code region) unless we make very strong assumptions. To deal with a continuous variable, we need to know the marginal distribution of the variable at the aggregated level. In case of a discrete variable, this distribution is represented by a few probabilities. In case of a continuous variable we need to know the type of distribution and the values of the parameters of the distribution. If the continuous variable is however divided in several intervals and we know the probability distribution over these intervals we can model it like an ordered categorical explanatory variables, see Section 3.2 and Section 5 for an example.

To summarize this section. The explanatory variables X_i which are missing at the individual level are described by the latent variable $X_i^* = (X_i^{(1)*\prime}, X_i^{(2)*\prime}, X_i^{(3)*\prime})'$. 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

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12}' & \Omega_{13}' \\ \Omega_{12} & \Omega_{22} & \Omega_{23}' \\ \Omega_{13} & \Omega_{23} & \Omega_{33} \end{pmatrix}, \tag{16}$$

where the matrices Ω_{11} and Ω_{22} contain ones on the diagonal and Ω_{33} is equal to the covariance matrix given in (14) in case of just one unordered variable. If there are more unordered variables, Ω_{33} contains as many blocks of the covariance matrix from (14) on the diagonal. The remaining elements of Ω are free and describe the correlations between the latent variables X_i^*. We summarize the free elements of Ω in the vector ρ. 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

$$f(y_i|X_i;\beta,\theta), \tag{17}$$

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

$$f(y_i|P_i;\beta,\theta,\rho) = \sum_{X_i\in\chi} \Pr[X_i|P_i;\rho] f(y_i|X_i;\beta,\theta), \tag{18}$$

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 $P_i = (P_i^{(1)\prime}, P_i^{(2)\prime}, P_i^{(3)\prime})'$. These probabilities depend on the unknown parameter ρ which summarizes the free elements of the covariance matrix Ω as discussed in the previous section. Hence, the log likelihood function is given by

$$\mathcal{L}(y|P;\beta,\theta,\rho) = \sum_{i=1}^{N} \log f(y_i|P_i;\beta,\theta,\rho), \tag{19}$$

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 (19) 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 it is possible to 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 (1992) simulator or the GHK simulator (Börsch-Supan and Hajivassiliou, 1993; Keane, 1994).

The SML estimator is only 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), we expect that obtaining 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 $\Pr[X_i|P_i;\rho]$ we need to evaluate 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, an 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 ρ as this is part of the model for the latent variable X_i^*.

4.2.1 Sampling of X_i^*

Because X_i is a deterministic function of X_i^* we only need to sample X_i^*. The full conditional density of X_i^* is given by

$$f(X_i^*|y_i, P_i; \beta, \theta, \rho) \propto f(y_i|X_i(X_i^*); \beta, \theta) f(X_i^*|P_i; \rho), \tag{20}$$

where $f(y_i|X_i(X_i^*); \beta, \theta)$ is given in (17) with $X_i(X_i^*)$ the deterministic mapping of X_i^* to X_i given in (8), (10) and (13). The function $f(X_i^*|P_i; \rho)$ denotes the density of X_i^* implied by the latent variable model for X_i^*. Given the structure of the latent variable model, X_i^* has a multivariate normal distribution with a mean μ_i which is determined by P_i and a covariance matrix Ω of which the free elements are denoted by ρ, that is,

$$f(X_i^*|P_i; \rho) = \phi(X_i^*; \mu_i(P_i), \Omega(\rho)), \tag{21}$$

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

$$f(x_{ij}^*|y_i, P_i, X_{i,-j}^*, \beta, \theta, \rho) \propto f(y_i|x_{ij}(x_{ij}^*), X_{i,-j}(X_{i,-j}^*); \beta, \theta) f(x_{ij}^*|X_{i,-j}^*, P_i; \rho), \tag{22}$$

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 $f(y_i|x_{ij}(x_{ij}^*), X_{i,-j}(X_{i,-j}^*); \beta, \theta)$ is a function of $X_i(X_i^*)$ and can take a discrete number of values depending on the value of x_{ij}^*.

In case x_{ij}^* corresponds to a binary explanatory variable, x_{ij} can take two values depending on whether x_{ij}^* is larger or smaller than 0. We can sample x_{ij}^* from in one step from its full conditional posterior distribution using the inverse CDF method but it is computationally more efficient to sample x_{ij}^* in two steps. In the first step, we determine whether x_{ij}^* is larger or smaller than 0, that is whether x_{ij} is 1 or 0, respectively. From

(22) it follows that

$$\Pr[x_{ij} = 1|y_i, P_i, X^*_{i,-j}, \beta, \theta, \rho] = \frac{k_{ij1}\left(1 - \Phi\left[\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right]\right)}{k_{ij0}\Phi\left[\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right] + k_{ij1}\left(1 - \Phi\left[\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right]\right)}, \tag{23}$$

where $k_{ij0} = f(y_i|x_{ij} = 0, X_{i,-j}(X^*_{i,-j}); \beta, \theta)$ and $k_{ij1} = f(y_i|x_{ij} = 1, X_{i,-j}(X^*_{i,-j}); \beta, \theta)$.

In the second step, we sample $x^*_{ij}|x_{ij}$ from a truncated normal distribution with mean $\bar{\mu}_{ij}$ and variance $\bar{s}^2_j$. We sample x^*_{ij} either positive or negative, depending on the whether x_{ij} is 1 or 0, respectively. For this truncated sampling we use the efficient accepting algorithm in Geweke (2005, pp. 113), see also Geweke (1991).

The other types of variables can be sampled in a similar manner. Appendix A also provides the sampling schemes 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, see (16). As discussed in Section 3, identification requires several restrictions on the covariance matrix Ω. In the first place, 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 exists several algorithms to sample 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 $\mu(P) = (\mu_1(P_1), \ldots, \mu_N(P_N))'$, where $\mu_i(P_i)$ denotes the mean of X^*_i determined by P_i for $i = 1, \ldots, N$. The full condi-

tional posterior density of ρ_j is given by

$$\begin{aligned} f(\rho_j|y,P,X^*,\beta,\theta,\rho_{-j},) &\propto f(X^*|P,\rho_j,\rho_{-j})f(\rho_j|\rho_{-j}) \\ &\propto \prod_{i=1}^{N}\phi(X_i^*;\mu_i(P_i),\Omega(\rho_j,\rho_{-j}))f(\rho_j|\rho_{-j}) \\ &\propto |\Omega(\rho_j,\rho_{-j})|^{-\frac{N}{2}}\exp\left(-\frac{1}{2}(X^*-\mu(P))'\Omega(\rho_j,\rho_{-j})^{-1}(X^*-\mu(P))\right)f(\rho_j|\rho_{-j}), \end{aligned} \tag{24}$$

where $f(\rho_j|\rho_{-j})$ denotes the prior density of the jth element of ρ, conditional on all other elements of ρ. 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 (24) 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 suggest to impose an informative prior for the parameters in ρ. We use a truncated normal prior with variance ω^2 for the parameters in ρ, that is,

$$f(\rho) \propto I[\Omega(\rho) = \text{PD}]\prod_j \exp(-\rho_j^2/(2\omega^2)), \tag{25}$$

where $I[\Omega(\rho) = \text{PD}]$ is an indicator function which is 1 if $\Omega(\rho)$ is positive definite and 0 otherwise. 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 from 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 from the database. The mailing we consider took place in February 1997. The response rate is 39.0%. The average donation is 3.39 euros and the average donation conditional on response is 8.68 euros. We match these data with aggregated data at the zip-code level (4 digits) from Statistics Netherlands (CBS). Table 3 shows the relevant aggregated data at the zip-code

Table 3: 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

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

$$\log(1+y_i) = \begin{cases} x_i'\beta + \varepsilon_i & \text{if} \quad x_i'\beta + \varepsilon_i > 0 \\ 0 & \text{if} \quad x_i'\beta + \varepsilon_i \leq 0, \end{cases} \tag{26}$$

with $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. As explanatory variables we take the variables displayed in Table 3.

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 (26) 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 result 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 a Bayesian approach. For ρ we take the informative prior (25) with $\omega^2 = 1/4$.

Table 4: 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	s.d.	95% HPD	ML	s.e.[a]
Intercept	-1.77	0.05	(-1.87,-1.68)	-2.89	0.81
Urbanization	0.05	0.03	(-0.01, 0.11)	0.63	0.35
Church	0.64	0.02	(0.59, 0.69)	0.24	0.25
Not-active	-0.72	0.01	(-0.75,-0.69)	-1.03	0.74
Single	3.63	0.04	(3.55, 3.71)	0.52	0.55
Family no kids	3.61	0.04	(3.53, 3.70)	3.51	1.20
Income low	-0.01	0.03	(-0.07, 0.05)	0.41	1.17
Income high	0.71	0.02	(0.68, 0.75)	1.60	0.87
Age 0-24	-0.01	0.03	(-0.07, 0.06)	2.96	1.35
Age 45-65	-3.60	0.09	(-3.77,-3.42)	-1.42	1.28
Age 65+	0.72	0.02	(0.68, 0.75)	1.80	1.05
σ	0.17	0.00	(0.17, 0.18)	2.36	0.02

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

For β we use a normal prior with mean 0 and standard deviation 0.5 and for σ^2 we use an inverted Gamma-2 prior with parameters 12.5 and 50. These priors help to obtain a smoother convergence of the MCMC sampler. Posterior results turn out not to be sensitive to moderate changes of this prior specification.

We use a total of 120,000 draws, which took about six hours on an Pentium 4, 2.8 Ghz processor. The first 20,000 draws were used as burn-in period. Furthermore, we only used every 20th draw to obtain a random sample of 5,000 draws. Our code is tested using the approach of Geweke (2004).

Table 4 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 large, the estimated standard errors of the simple approach are still substantial. 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* and *Age 0-24* are identified as

Table 5: Posterior means of the correlations between unobserved variables with posterior standard deviations in parentheses

	Church	Not-active	Single	Family no kids	Income	Age
Church	1 (-)					
Not-active	0.19 (0.10)	1 (-)				
Single	-0.34 (0.10)	-0.14 (0.14)	1 (-)			
Family no kids	-0.72 (0.09)	0.19 (0.08)	0.50 (-)	1 (-)		
Income	0.06 (0.09)	-0.54 (0.16)	0.33 (0.10)	-0.27 (0.10)	1 (-)	
Age	0.33 (0.10)	0.27 (0.06)	0.29 (0.08)	-0.19 (0.06)	-0.17 (0.06)	1 (-)

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 has a positive effect, while not being active in the labor force has a negative effect. Both single households and families without children tend to donate more. Household with 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.

There are two differences in the results of the two methods. First, both find that families without children donate more than families with children, however, the ML results suggest that single households donate about the same as families with children while the mixture approach suggest that their donating behavior is more comparable to families without children. The second difference in results concerns the effect of age. The main difference is the level of the reference category, as all parameters have a higher value in the ML estimates. Moreover, according to the ML results individuals younger than 25 are the most lucrative group, while the mixture approach suggests that this group consists of individuals over 65.

Table 5 displays the estimated correlation matrix of X_i^*. The diagonal elements are fixed 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. Many

of the posterior means of the correlations are more than two times larger than their posterior standard deviation, illustrating the importance of our approach. In general, the correlations have the expected sign. For example, there is a negative correlation between being not active in the labor force and income, and a positive correlation between being religious 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.

Acknowledgements

We thank three anonymous reviewers, 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, the NAKE research day in Amsterdam, Facultes Universitaires Notre-Dame de la Paix in Namur, and EEA/ESEM 2007 in Budapest for helpful comments. All estimation results are obtained using Ox version 4.00 (Doornik, 2002).

A Derivation of full conditional distributions

In this appendix we provide the simulation schemes for missing ordered categorical variables and unordered categorical variables. As starting point we take the general form of the full conditional density of x_{ij}^* given in (22).

A.1 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 set 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 mean of $x_{ij}^{(2)*}|X_{i,-j}^*, \rho$ in the latent model and let $\bar{s}_j^2$ denote the 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, that is,

$$\begin{aligned}
k_{ij1} &= f(y_i|X_{i,-j}(X_{i,-j}^*), x_{i1}^{(2)} = 1, x_{i2}^{(2)} = 0, \ldots, x_{ir-1}^{(2)} = 0; \beta, \theta) \\
&\vdots \\
k_{ijr-1} &= f(y_i|X_{i,-j}(X_{i,-j}^*), x_{i1}^{(2)} = 0, \ldots, x_{ir-2}^{(2)} = 0, x_{ir-1}^{(2)} = 1; \beta, \theta) \\
k_{ijr} &= f(y_i|X_{i,-j}(X_{i,-j}^*), x_{i1}^{(2)} = 0, \ldots, x_{ir-1}^{(2)} = 0; \beta, \theta),
\end{aligned}$$

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

First we draw $x_{i1}^{(2)}, \ldots, x_{ir-1}^{(2)}$ using the fact that they are mutually exclusive and

$$\Pr[x_{it}^{(2)} = 1, x_{i,-t}^{(2)} = 0|y_i, P_i, X_{i,-1}(X_{i,-j}^*), \beta, \theta, \rho] = \frac{k_{ijt}(\bar{p}_{it} - \bar{p}_{it-1})}{\sum_{l=1}^{r} k_{ijl}(\bar{p}_{il} - \bar{p}_{il-1})} \tag{27}$$

for $t = 1, \ldots, r-1$, where $\bar{p}_{it} = \Phi\left(\frac{q_{it} - \bar{\mu}_{ij}}{\bar{s}_j}\right)$ with $\bar{p}_{i0} = 0$ and $\bar{p}_{ir} = 1$, and where $x_{i,-t}^{(2)}$ denotes $x_{i1}^{(2)}, \ldots, x_{ir-1}^{(2)}$ without $x_{it}^{(2)}$.

Next, we sample $x_{ij}^{(2)*}|x_{i1}^{(2)}, \ldots, x_{ir-1}^{(2)}$ from a truncated normal distribution with mean $\bar{\mu}_{ij}$ and variance $\bar{s}_j^2$. If $x_{it}^{(2)} = 1$ we sample $x_{ij}^{(2)*}$ between q_{it-1} and q_{it} for $t = 1, \ldots, r-1$, where $q_{i0} = -\infty$. If $x_{i1}^{(2)} = \ldots = x_{ir-1}^{(2)} = 0$, we sample $x_{ij}^{(2)*}$ larger than q_{ir-1}. We again use the acceptance sampling algorithm of Geweke (2005, pp. 113).

A.2 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 (22). 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 variance, say $\bar{s}_j^2$. Conditional on $X_{i,-j}^*$ the first part $f(y_i|x_{ij},X_{i,-j}(X_{i,-j}^*);\beta,\theta)$ 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 are given by

$$\begin{aligned}k_{ij0} &= f(y_i|x_{i1}^{(3)}=0,\ldots,x_{il-1}^{(3)}=0,x_{il}^{(3)}=1,x_{il+1}^{(3)}=0,\ldots x_{ir-1}^{(3)}=0,X_{i,-j}(X_{i,-j}^*);\beta,\theta)\times\\ &\qquad I[x_{il}^{(3)*}>0]+f(y_i|x_{i1}=0,\ldots,x_{ir-1}=0,X_{i,-j}(X_{i,-j}^*);\beta,\theta)I[x_{il}^{(3)*}\leq 0]\\ k_{ij1} &= f(y_i|x_{i1}^{(3)}=0,\ldots,x_{ij-1}^{(3)}=0,x_{ij}^{(3)}=1,x_{ij+1}^{(3)}=0,\ldots x_{ir-1}^{(3)}=0,X_{i,-j}(X_{i,-j}^*);\beta,\theta),\end{aligned}$$

where $X_{i,-j}$ denotes X_i without $x_{i1}^{(3)},\ldots,x_{ir-1}^{(3)}$.

In the cases of a binary or ordered categorical variable the distribution of $x_{ij}^*|X_i$ is a univariate truncated normal. However, for an unordered variable this is not the case anymore, as its distribution also depends on the value of $x_{il}^{(3)*}$. The sampling of $x_{i1}^{(3)*},\ldots,x_{ir-1}^{(3)*}$ given $x_{i1}^{(3)},\ldots,x_{ir-1}^{(3)}$ becomes non-standard. Therefore, it is more efficient to use the inverse CDF method to draw from $x_{ij}^{(3)*}$ and $x_{ij}^{(3)}$ simultaneously.

The full conditional posterior density of $x_{ij}^{(3)*}$ is given by

$$\begin{aligned}c_{ij}(k_{ij0}\phi(x_{ij}^{(3)*};\bar{\mu}_{ij},\bar{s}_j)I[x_{ij}^{(3)*}\leq\max(x_{il}^{(3)*},0)]+\\ k_{ij1}\phi(x_{ij}^{(3)*};\bar{\mu}_{ij},\bar{s}_j)I[x_{ij}^{(3)*}>\max(x_{il}^{(3)*},0)]),\end{aligned}\quad(28)$$

where the integrating constant c_{ij} equals

$$c_{ij}=\left((k_{ij0}-k_{ij1})\Phi\left(\frac{\max(x_{il}^{(3)*},0)-\bar{\mu}_{ij}}{\bar{s}_j}\right)+k_{ij1}\right)^{-1}.\quad(29)$$

Straightforward derivation leads to the following inverse CDF

$$x_{ij}^{(3)*}(u)=\begin{cases}\Phi^{-1}\left(\frac{u}{c_{ij}k_{ij0}}\right)\bar{s}_j+\bar{\mu}_{ij} & \text{if } u\leq\bar{u}\\ \Phi^{-1}\left(\frac{u}{c_{ij}k_{ij1}}+\frac{k_{ij1}-k_{ij0}}{k_{ij1}}\Phi\left(\frac{\max(x_{il}^{(3)*},0)-\bar{\mu}_{ij}}{\bar{s}_j}\right)\right)\bar{s}_j+\bar{\mu}_{ij} & \text{if } u>\bar{u},\end{cases}\quad(30)$$

where $\bar{u} = c_{ij} k_{ij0} \Phi \left(\frac{\max(x_{il}^{(3)*}, 0) - \bar{\mu}_{ij}}{\bar{s}_j} \right)$.

References

Aitchison J., Silvey S.D., 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.R., Sowden R.R., 1970. Multi-variate probit analysis. Biometrics 26, 536–546.

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

Billard L., Diday E., 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., Hajivassiliou V.A., 1993. Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. Journal of Econometrics 58, 347–368.

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

Cowles M.K., 1996. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. Statistics and Computing 6, 101–111.

Dempster A.P., Laird N.M., Rubin R.B., 1977. Maximum Likelihood estimation from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38.

Doornik J.A., 2002. Object-Oriented Matrix Programming using Ox. 3rd edn. Timberlake Consultants Press, Londen.

Everitt B.S., Hand D.J., 1981. Finite Mixture Distributions. Chapman and Hall, London.

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

Geweke J., 1991. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In: Keramidas E.M. (Ed.), Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface. Interface Foundation of North America, Fairfax, VA, pp. 571–578.

Geweke J., 2004. Getting it right: Joint distribution tests of posterior simulators. Journal of the American Statistical Association 99, 799–804.

Geweke J., 2005. Contemporary Bayesian Econometrics and Statistics. Wiley, New York.

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

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

Imbens G.W., Lancaster T., 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.P., 1992. A note on identification in the multinomial probit model. Journal of Business & Economic Statistics 10, 193–200.

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

Lerman S.R., Manski C.F., 1981. On the use of simulated frequencies to approximate choice probabilities. In: Manski C., McFadden D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT press, Cambridge, pp. 305–319.

Liechty J.C., Liechty M.W., Müller P., 2004. Bayesian correlation estimation. Biometrika 91, 1–14.

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

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

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

Romeo C.J., 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.J., Ainslie A., Engebretson P.H., 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.A., Wong W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–540.

Titterington D.M.T., Smith A.F.M., Makov U.E., 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.J., van der Klaauw B., 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.