ISSN 1032-3813
ISBN 0 7326 0786 8

MONASH UNIVERSITY

AUSTRALIA

USING THE EM ALGORITHM WITH COMPLETE, BUT SCRAMBLED, DATA

Guyonne Kalb

Working Paper 5/96
June 1996

DEPARTMENT OF ECONOMETRICS

Using the EM algorithm with complete, but scrambled, data

Guyonne Kalb

Department of Econometrics,

Monash University,

Clayton, VIC 3168,

Australia.

E-mail: Guyonne.Kalb@BusEco.monash.edu.au

Telephone: (03) 99055180

Fax: (03) 99055474

May 1996

Abstract

Consider two sets of records from the same survey. One preserves full detail about a few questions under focus (on labour supply), but contains almost no other variables. The other set contains very little information about the question of interest, but has complete information on the remaining variables. Unfortunately, the key that would allow the two sets to be matched is not available. However, the structure of the record sets does allow a partial matching. In order to extract the maximum amount of information about the question of interest, the use of statistical inference is required.

In this paper the EM algorithm, which has been used successfully with censored and incomplete data sets, is adapted to the problem of scrambled data. The performance of the method is assayed using an artificially constructed data set. The relevance of the results for a real world labour market problem is explored.

JEL Classifications: C13, C15, C81

I would like to thank the participants of a conference and seminars where a previous version of the paper was presented and in particular Denzil Fiebig and Bill Griffiths for their comments, Gael Martin for the many references and explanations and my supervisors Alan Powell and Tim Fry for the helpful critical comments on this paper at all stages.

1 Introduction to the problem

It can be very difficult to get the "perfect" data set. In fact we are often quite happy with any data set that has some of our most wanted variables in it. Sometimes our perfect data set does not exist and at other times the perfect data set contains so much information that individuals could be identified. In that last case the data often cannot be released with the full information in it.

In our case the nearly perfect data set does exist and we have this data set available in two separate files[1].

Each of the records in both files is identified by an identity number uniquely allocated to each household, a family number, an income unit number and a person number. Unfortunately the household identity numbers do not match, so these numbers cannot be used to connect the two files.

However, there is some common information available in both data sources to connect the two files. Both files have some information on household composition and on the working hours of its members.

Obviously the most uncommon cases will be matched more easily, since there are not many similar cases that make the choice difficult. So a household that has a special composition or that has members that work a less common number of hours and/or have a second job, is more easily matched.

In our case *exact* additional information is available for only a small percentage (6.1%) of the income units.

We would like to use the exact information together with the information available for non-exact matches to make optimal use of the data

To do this, we need to find a way to use the information in the additional file without making too improbable assumptions.

We will present our problem in more general terms in the following section before starting to explore a possible solution. In the third section the method will be tried on a simple artificial example and results obtained by using this method will be compared to the results that would be obtained from a complete data set. The final

[1] See appendix A.1 for more detailed information on these data files.

section will contain a discussion on the usefulness of the proposed method for our problem.

2 Estimating with scrambled data

2.1 General presentation of the problem

Presenting the problem in a general notation results in the following:

Our *main data set* consists of N observations on records m_1, m_2, ..., m_N. Each record m_i is a vector of K_Z realizations of variables belonging to the vector Z and K_X realizations of variables belonging to another vector X. The latter consists of exogenous variables to be used as explanators of endogenous variables of interest. For the time being, we restrict our attention to the case of one scalar endogenous variable Y*. Note that the main data set contains no information on the values of Y*.

Our *supplementary data set* consists of N observations, s_1, s_2, ..., s_N, on K_Z variables contained in Z (the same as above; these variables in Z are called matching variables) and on a scalar variable Y. Notationally the symbols Y and Y* relate to the same variable. They are only distinguished because of the scrambling of the data.

Realizations of Y can be mapped exactly to realizations of Z. However, what is required is a mapping between realizations of Y and X, the variables of interest. The main data set allows exact mapping between realizations of X and Z. To implement an indirect (and as it turns out, inexact) mapping from Y to X via Z, the concept of a matching group is used.

A pair of incomplete data sets is defined as a main data set together with its supplementary data set.

Matching groups are defined for both the main and supplementary data set as follows. First Z is identified as the set of matching variables (i.e. the set that will be used to implement the indirect mapping). The records m_1, m_2, ..., m_N from the main data set are partitioned into groups g_1^m, ..., g_G^m such that the realizations on all the variables in Z are the same for all records m_i assigned to the same group g_j^m. These groups g_j^m are called *matching groups*. Note that if n_j is the number of records

belonging to matching group j and G is the total number of matching groups then $\sum_{j=1}^{G} n_j = N$.

For example, if our main set consists of just the records m_1, ..., m_4 as set out in scheme 2.1, then the following groups can be formed from these records:
$g_1^m = \{m_1, m_3\}$, $g_2^m = \{m_2\}$ and $g_3^m = \{m_4\}$.

Scheme 2.1 An example of data

main data set					**supplementary data set**			
	values of Z			values of X		values of Z		value of Y
	Z_1	Z_2	Z_3	X_1 X_2		Z_1	Z_2 Z_3	Y
m_1 = (	1,	3,	4;	100, 1000)	s_1 = (	1,	3, 4;	98)
m_2 = (	1,	3,	5;	50, 70)	s_2 = (	1,	3, 5;	55)
m_3 = (	1,	3,	4;	200, -30)	s_3 = (	-2,	5, 7;	12)
m_4 = (	-2,	5,	7;	10, 250)	s_4 = (	1,	3, 4;	205)

In the case of the supplementary data set, the same values of Z appear (but not necessarily in the same order as in the main data set). The supplementary data set has to consist of the same number of records as the main data set; an example is given in scheme 2.1.

In this case the matching groups for the supplementary data set are the following: $g_1^s = \{s_1, s_4\}$, $g_2^s = \{s_2\}$ and $g_3^s = \{s_3\}$. Note: (i) the number of matching groups for the supplementary data set is necessarily the same as it is for the main data set; (ii) that if the ordering of matching groups for both sets is generated by sorting according to the values of their matching variables (in Z), then the Z-parts of records m_i contained within any given g_j^m are necessarily the same as the Z-parts of the records s_i contained within the corresponding g_j^s (i.e. they match). This can be easily seen from scheme 2.2, where the records have been ordered according to ascending values on the Z variables. Records with the same values on the Z variables are now grouped together in both data sets. Forming groups starting from the first record and following the order of the records, we get for the main data set: $g_1^m = \{m_1\}$, $g_2^m = \{m_2, m_3\}$ and $g_3^m = \{m_4\}$ and for the supplementary data set:
$g_1^s = \{s_1\}$, $g_2^s = \{s_2, s_3\}$ and $g_3^s = \{s_4\}$. A possible match for m_2 is either s_2 or s_3.

Scheme 2.2 The data set from scheme 2.1 after reordering

main data set					**supplementary data set**			
	values of Z			values of X		values of Z		value of Y
	Z_1	Z_2	Z_3	X_1 X_2		Z_1	Z_2 Z_3	Y
m_1 = (	-2,	5,	7;	10, 250)	s_1 = (	-2,	5, 7;	12)
m_2 = (	1,	3,	4;	100, 1000)	s_2 = (	1,	3, 4;	98)
m_3 = (	1,	3,	4;	200, -30)	s_3 = (	1,	3, 4;	205)
m_4 = (	1,	3,	5;	50, 70)	s_4 = (	1,	3, 5;	55)

As noted above, we are interested in the effect that X has on Y*. Z is only relevant to determine the group of records where the correct match can be found. It defines the values for Y* we can choose from. In the example above the matching group g_1^m from the main data set corresponds to g_1^s from the supplementary data set. The two groups with only one record in them can be exactly matched. This means that X_1 = 50, X_2 =70 is combined with Y = 55 and that X_1 = 10, X_2 =250 is combined with Y = 12. The other two records belong to a common matching group, so no unique choice can be made in that case. The following choice of assignment is possible:

(100,1000) ⟷ 98
(200,-30) ⟷ 205

OR

(100,1000) ⟷ 205
(200,-30) ⟷ 98

The values for the variable Y corresponding to the correct (but unknown) assignment are y^{*}[2]. So y consists of the same values as y*, but possibly in a different and incorrect order. Above the possible values of $Y^*|X=(100,1000)$ are Y=98 and Y=205. The same values of Y are candidates for $Y^*|X=(200,-30)$.

The central question is whether the information in the additional file is useful even though we cannot always find the correct matches. Most of the time only a group of possible values, in which the correct match must lie, can be determined.

---

[2] Upper case letters X, Y and Z will be used to indicate the random variables and the lower case x, y and z will be used to indicate its realizations for a particular individual.

2.2 General approach to estimation

Y^* is an unobserved or latent variable whenever an observation belongs to a matching group with more than one element. The exact value of the realization of Y^* is unknown, but it is known that only certain realizations could have taken place. From now on this set of possible realizations will be called *A*. In the numerical example in scheme 2.1 the true realization of Y^* associated with X=(100,1000) could only have been 98 or 205, so *A* ={98, 205}. Likewise the true value of Y^* associated with X = (200, -30) must also lie in *A* ={98, 205}. Furthermore it is known that each element in *A* can only be used once. So either X=(100,1000) is combined with Y^*=98 and X=(200,-30) is combined with Y^*=205 or X=(100,1000) is combined with Y^*=205 and X=(200,-30) is combined with Y^*=98.

The principle of estimation to be adopted is maximum likelihood. Thus ideally the aim would be to find values of a parameter vector θ describing the relationship between Y^* and X, for which the likelihood $L(\theta|y^*)$ of the realizations of Y^* conditional on X is a maximum. Because not all of the realizations y^* on Y^* are known, the likelihood as defined above cannot be computed. However it is known that $y^* \in A$ and that each of the elements of *A* can only be used for one of the observations from the corresponding matching group in the main file.

The combinations of y^* and x_i that can be made for observation i are not independent from the other combinations made in the same matching group. This does not mean that the actual realizations y_i^* in group j are dependent, but only that the choice we can make for y_i^* given the observation on A_j depends on the choice that is made for the other elements of matching group j. The dependence is caused by the fact that values y^* are observed in groups and the correct assignment of these values y^* to the observations on X is unknown. So instead of setting up the likelihood per observation as is usually done, here the likelihood should be set up per matching group. In that way we can look at all possible joint realizations of Y_{1j}^*, $Y_{2j}^*,\ldots, Y_{n_jj}^*$ in the matching group given the values in A_j. All the different permutations of $\{y_1, y_2,\ldots, y_{n_j}\}$ over the observations x_i from the corresponding matching group are included in the likelihood. These different permutations are possible combinations of y^* and x, given the elements of each matching group. One of these combinations is the unknown true combination. So instead of

maximizing the likelihood based on the exact realizations of Y* the likelihood based on a summation of possible outcomes for Y* is going to be maximized.

Suppose we are interested in the following simple model:

$Y^* = X\beta + u \qquad u \sim N(0,\sigma^2)$

Write $\theta' = (\beta', \sigma^2)$.

Information on Y* and X is located in separate data sets, which can both be divided into corresponding matching groups.

The likelihood function can be constructed by taking the joint probability density functions for observations per matching group:

$$L(\theta \mid A_1,\ldots,A_G) = \prod_{j=1}^{G} \sum_{i=1}^{n_j} \sum_{\substack{k=1 \\ k\neq i}}^{n_j} \ldots \sum_{\substack{p=1 \\ p\neq i,k,\ldots}}^{n_j} \text{pdf}\left(Y_{1j}^* = y_{ij}, Y_{2j}^* = y_{kj}, \ldots, Y_{n_j j}^* = y_{pj} \mid \theta, X_{1j},\ldots,X_{n_j j}\right) =$$

$$\prod_{j=1}^{G} \sum_{i=1}^{n_j} \sum_{\substack{k=1 \\ k\neq i}}^{n_j} \ldots \sum_{\substack{p=1 \\ p\neq i,k,\ldots}}^{n_j} \text{pdf}\left(Y_{1j}^* = y_{ij} \mid \theta, X_{1j}\right) \ldots \text{pdf}\left(Y_{n_j j}^* = y_{pj} \mid \theta, X_{n_j j}\right) \qquad (2.1)$$

where

$A_j = \{y_{1j}, y_{2j}, \ldots, y_{n_j j}\}$ for j=1,...G

pdf = probability density function

Taking logarithms to obtain the log likelihood:

$$l(\theta \mid A_1,\ldots,A_G) = \sum_{j=1}^{G} \log\left\{ \sum_{i=1}^{n_j} \sum_{\substack{k=1 \\ k\neq i}}^{n_j} \ldots \sum_{\substack{p=1 \\ p\neq i,k,\ldots}}^{n_j} \text{pdf}\left(Y_{1j}^* = y_{ij} \mid \theta, X_{1j}\right) \ldots \text{pdf}\left(Y_{n_j j}^* = y_{pj} \mid \theta, X_{n_j j}\right) \right\} \qquad (2.2)$$

The idea of summing over all possible outcomes given the observed data is similar to integration in the case of grouped data. When the dependent variable Y* is only

known in a grouped version $y_i^* \in [l_i, u_i]$, then instead of:

$$L(\theta|y^*) = \prod_{i=1}^{N} pdf(y_i^* | \theta, x_i)$$

the following is taken:

$$L(\theta|y) = \prod_{i=1}^{N} \int_{l(y_i)}^{u(y_i)} pdf(y^* | \theta, x_i) dy^* \quad , \text{where y indicates to which group } y^* \text{ belongs.}$$

In the grouped data case all we know about y_i^* is that it lies in between $l(y_i)$ and $u(y_i)$ and in the above likelihood it is all the information that is used.

In the case of scrambled data we have a limited discrete number of possible outcomes for each matching group. So instead of integrating over possible outcomes within given intervals i (as is done with grouped data), we sum the probability density functions over all possible outcomes within the matching groups. For each matching group the sum is taken of the probability density functions of all possible permutations formed from $\{y_1,..., y_{n_j}\}$ and $\{x_1,..., x_{n_j}\}$. In case of double values in the set A_j the summation does not contain the double permutations. I.e. these are the permutations that give the same values on $\{Y_{1j}^*,..., Y_{n_j j}^*\}$ as a previous permutation. Only one of these permutations will appear in the sum.

The fact that a group of possible values for Y* instead of one true value is observed, complicates the likelihood function.
Looking at the problem of scrambled data in a less technical way, it shows similarities to the missing data problem or to the censored data problem. In our case for quite a few records the value of some of the variables is not known, but we do know the possible values. In the case of missing data there is no information at all on the missing variables. They can have any value. For censored data the exact values of a variable below or above a certain value are not known. In these two cases one value of the observed variable represents a range of possible values for

the latent variable. Comparing our problem to these last two cases it is obvious that more information is available on our "missing" variables.

A way of handling the missing data or censored data problem is to use the EM algorithm to estimate models instead of maximizing the likelihood directly (Ruud (1991)).

Because of the similarities between these two problems and our problem of scrambled data and since the likelihood function might be difficult computationally, an investigation into the EM algorithm seems worthwhile.

2.3 Description of the EM algorithm

Before adapting the EM algorithm to our specific problem, an outline of the method will be given. First the notation: y are the observed values related to the latent endogenous variable, y^* are the unobserved realizations of the latent variable Y^* and $pdf(y^*|\theta)$ is the value of the probability density function for Y^* at y^*, indicating how likely it is for y^* to be the true realization of Y^*, where θ is the parameter vector characterizing the probability distribution of Y^* conditional on the set X of exogenous variables.

Our objective is to maximize the likelihood based on the observations y ($L(\theta|y)$). There is however no direct theoretical relationship between the observed variable y and the exogenous variables X. From economic theory, only a relationship between Y^* (the underlying latent variable) and X can be derived, hence only $pdf(y^*|\theta)$ can be constructed directly from the theory.

$L(\theta|y)$ can be decomposed into[3]:

$$L(\theta|y) = \int_{A(y)} pdf(y^*|\theta)dy^* = pdf(y^*|\theta) / Pr(y^*|y,\theta) \qquad (2.3)$$

where

$$Pr(y^*|y,\theta) = \frac{pdf(y^*|\theta)}{\int_{A(y)} pdf(y^*|\theta)\,dy^*}$$

for $y^* \in A(y)$ and A(y) is the range of values the latent variable y^* can take given that we observe a value y.

---

[3] See Dempster, Laird and Rubin (1977).

Taking logarithms:

$$l(\theta|y) = \log(L(\theta|y)) = \log(pdf(y^*|\theta)) - \log(Pr(y^*|y,\theta)) \tag{2.4}$$

Multiplying by $Pr(y^*|y,\varphi)$ and integrating over the set A(y) of all possible values for the latent variable y* observationally equivalent to the realized value of the observed variable y, the following can be obtained:

$$\int_{A(y)} l(\theta|y) Pr(y^*|y,\varphi)\, dy^* = \int_{A(y)} \log(pdf(y^*|\theta)) Pr(y^*|y,\varphi)\, dy^* - \int_{A(y)} \log(Pr(y^*|y,\theta)) Pr(y^*|y,\varphi)\, dy^* \tag{2.5}$$

where φ is defined over the same domain as θ. Equation 2.5 is valid for any value of the parameter φ.

Keeping in mind that $Pr(y^*|y,\varphi)$ integrated over all possible values of y* will result in the value 1, we can write this expression as:

$$l(\theta|y) = \int_{A(y)} \log(pdf(y^*|\theta)) Pr(y^*|y,\varphi)\, dy^* - \int_{A(y)} \log(Pr(y^*|y,\theta)) Pr(y^*|y,\varphi)\, dy^* \tag{2.6}$$

This is often written more compactly as:

$$l(\theta|y) = Q(\theta, \varphi ; y) - H(\theta, \varphi ; y) \tag{2.7}$$

Dempster, Laird and Rubin (1977) show that optimizing Q with respect to θ in an iterative way results in an estimated θ, that also is a stationary point of $l(\theta|y)$. Their EM algorithm to maximize Q starts from some starting value θ^0 for φ. Then $Q(\theta,\theta^0,y)$ is maximized with respect to θ. The resulting value θ^1 is then used in the next step where $Q(\theta,\theta^1,y)$ is again maximized with respect to θ. This procedure is repeated until $\theta^q = \theta^{q-1}$ in iteration q.

By using the following lemma[4]:

$$\int f \log \frac{f}{g} \geq 0 \quad \text{when} \quad \int (f-g) \geq 0 \tag{2.8}$$

they show that if Q increases in value by choosing some $\theta^q \neq \theta^{q-1}$, then the value of $l(\theta|y)$ will increase as well when θ^{q-1} is substituted by θ^q. Apply Rao's lemma to $H(\theta,\theta^{q-1},y)$ and it will follow that $\theta = \theta^{q-1}$ is the maximum for H[5]. So any value $\theta \neq \theta^{q-1}$ will result in a lower or equivalent value for H than θ^{q-1} would. Therefore, if Q can be increased by selecting a new value for θ, $l(\theta|y)$ will increase as well.
It is easy to see that when both Q and H are maximized by a value $\theta^q = \theta^{q-1}$, the first derivatives of Q and H must be zero and so the first derivative of $l(\theta|y)$ must be zero as well. $\theta = \theta^q$ therefore is a stationary point of $l(\theta|y)$.

Once θ is estimated, the standard deviation can be calculated as well. The variance-covariance matrix of the estimator θ^q of θ can be derived from equation 2.6.
The second derivatives of $l(\theta|y)$ with respect to θ can be written as:

$$\frac{\partial^2 l(\theta|y)}{\partial \theta^2} = \left.\frac{\partial^2 Q(\theta,\varphi;y)}{\partial \theta^2}\right|_{\varphi=\theta} - \left.\frac{\partial^2 H(\theta,\varphi;y)}{\partial \theta^2}\right|_{\varphi=\theta} \tag{2.9}$$

Louis (1982) shows that this can also be written as:

$$\frac{\partial^2 l(\theta|y)}{\partial \theta^2} = \left.\frac{\partial^2 Q(\theta,\varphi;y)}{\partial \theta^2}\right|_{\varphi=\theta} + \text{var}\left[\frac{\partial \log(\text{pdf}(y^*|\theta))}{\partial \theta}\right] \tag{2.10}$$

The variance is taken with respect to $\Pr(y^*|y,\theta)$.

---

[4] See lemma 1e.6 in Rao (1973)

[5]

$$H(\theta^q,\theta^q,y) - H(\theta,\theta^q,y) = \int_A [\log(\Pr(y^*|\theta^q,y)) - \log(\Pr(y^*|\theta,y))]\Pr(y^*|\theta^q,y)\,dy^* \geq 0$$

$$\text{if} \int_A [\Pr(y^*|y,\theta^q) - \Pr(y^*|y,\theta)]dy^* \geq 0 \quad \text{and} \quad \Pr(y^*|y,\theta^q) > 0$$

This is valid for any value of θ.

Thus

$$\mathrm{Var}(\theta^q) = \left[-E \frac{\partial^2 l(\theta|y)}{\partial \theta^2} \Big|_{\theta=\theta^q} \right]^{-1}$$

$$= \left[-E \left[\frac{\partial^2 Q(\theta, \varphi; y)}{\partial \theta^2} \Big|_{\varphi=\theta^q} + \mathrm{var} \left[\frac{\partial \log(\mathrm{pdf}(y^*|\theta))}{\partial \theta} \right] \right] \Big|_{\theta=\theta^q} \right]^{-1} \qquad (2.11)$$

The variance-covariance matrix can be used to calculate the standard deviation of our estimated parameters in the usual way.

2.4 The EM algorithm applied to the scrambled data

The general method of the EM algorithm can be easily translated to our specific problem.

Our main data set consists of N observations, which can be grouped into matching groups. The records in a matching group of the main data set have exactly matching records in the corresponding matching group of the supplementary data set. However, the information to make the right assignments between the two groups is not available.

So in our supplementary data set a group of n_j records with values $\{y_{1j},..., y_{n_j j}\}$ (elements of the same matching group j) is observed. It is known that this group contains the true value y_i^* for the first record of the corresponding matching group in the main data set, so $y_i^* \in \{y_{1j},..., y_{n_j j}\} = A_j$. It is also known that this matching group contains n_j-1 other records which have a $y_{i'}^* \in A_j$ and there is a one to one relation between these n_j y_i^* and the values y_{1j} to $y_{n_j j}$. When n_j is bigger than 1, this unique one to one relation is unknown to us. It is only known that our unobserved variable Y^* has a realization that is equal to one of y_{1j} or y_{2j} or ... or $y_{n_j j}$. The realizations y to be included in A_j are determined by the matching group j an observation i belongs to.

Since in our case A_j is a discrete set of values, instead of an integration a summation over all values in A_j will be performed. Furthermore since the one to one relation has to hold, the expression of interest is not the probability density function

of each observation separately, but the joint probability density function for all observations in one matching group. The joint density functions will be summed over all possible combinations which can be formed from the two corresponding groups. Later we will sum over all matching groups. From equation 2.4 it then follows that:

$$\sum_{i=1}^{n_j}\sum_{\substack{k=1\\k\neq i}}^{n_j}\cdots\sum_{\substack{p=1\\p\neq i,k,\ldots}}^{n_j}\left[l\left(\theta\middle|A_j,X_{1j},\ldots,X_{n_jj}\right).\Pr\left(Y_{1j}{}^{*}=y_{ij},\ldots,Y_{n_jj}{}^{*}=y_{pj}\middle|X_{1j},\ldots,X_{n_jj},A_j,\varphi\right)\right]=$$

$$\sum_{i=1}^{n_j}\sum_{\substack{k=1\\k\neq i}}^{n_j}\cdots\sum_{\substack{p=1\\p\neq i,k,\ldots}}^{n_j}\left[\log\left(\mathrm{pdf}\left(Y_{1j}{}^{*}=y_{ij}\middle|X_{1j},\theta\right)\right)+\ldots+\log\left(\mathrm{pdf}\left(Y_{n_jj}{}^{*}=y_{pj}\middle|X_{n_jj},\theta\right)\right)\right].$$

$$\Pr\left(Y_{1j}{}^{*}=y_{ij},\ldots,Y_{n_jj}{}^{*}=y_{pj}\middle|X_{1j},\ldots,X_{n_jj},A_j,\varphi\right) \quad -$$

$$\sum_{i=1}^{n_j}\sum_{\substack{k=1\\k\neq i}}^{n_j}\cdots\sum_{\substack{p=1\\p\neq i,k,\ldots}}^{n_j}\left[\log\left(\Pr\left(Y_{1j}{}^{*}=y_{ij},\ldots,Y_{n_jj}{}^{*}=y_{pj}\middle|X_{1j},\ldots,X_{n_jj},A_j,\theta\right)\right)\right].$$

$$\Pr\left(Y_{1j}{}^{*}=y_{ij},\ldots,Y_{n_jj}{}^{*}=y_{pj}\middle|X_{1j},\ldots,X_{n_jj},A_j,\varphi\right) \qquad (2.12)$$

This can be simplified in the same way as for the continuous case:

$$l\left(\theta\middle|A_j,X_{1j},\ldots,X_{n_jj}\right)=$$

$$\sum_{i=1}^{n_j}\sum_{\substack{k=1\\k\neq i}}^{n_j}\cdots\sum_{\substack{p=1\\p\neq i,k,\ldots}}^{n_j}\left[\log\left(\mathrm{pdf}\left(Y_{1j}{}^{*}=y_{ij}\middle|X_{1j},\theta\right)\right)+\ldots+\log\left(\mathrm{pdf}\left(Y_{n_jj}{}^{*}=y_{pj}\middle|X_{n_jj},\theta\right)\right)\right].$$

$$\Pr\left(Y_{1j}{}^{*}=y_{ij},\ldots,Y_{n_jj}{}^{*}=y_{pj}\middle|X_{1j},\ldots,X_{n_jj},A_j,\varphi\right) \quad - \qquad (2.13)$$

$$\sum_{i=1}^{n_j}\sum_{\substack{k=1\\k\neq i}}^{n_j}\cdots\sum_{\substack{p=1\\p\neq i,k,\ldots}}^{n_j}\left[\log\left(\Pr\left(Y_{1j}{}^{*}=y_{ij},\ldots,Y_{n_jj}{}^{*}=y_{pj}\middle|X_{1j},\ldots,X_{n_jj},A_j,\theta\right)\right)\right].$$

$$\Pr\left(Y_{1j}{}^{*}=y_{ij},\ldots,Y_{n_jj}{}^{*}=y_{pj}\middle|X_{1j},\ldots,X_{n_jj},A_j,\varphi\right)$$

Using the same notation as for the continuous case:

$$l(\theta\,|\,A_j) = Q(\theta,\varphi\,;A_j) - H(\theta,\varphi\,;A_j) \qquad (2.14)$$

The question is whether the EM algorithm will be applicable for the discrete case as well. Rao (1973) has shown that a similar formula as for integrals holds for summations of converging sequences of positive numbers as well, from which

$$\sum f \log\frac{f}{g} \geq 0 \quad \text{when} \quad \sum (f-g) \geq 0 \tag{2.15}$$

can be derived[6]. So in the same way as in the continuous case, if Q can be increased by selecting a new value for θ, $l(\theta|A_j)$ will increase as well.
It is easy to see that when both Q and H are maximized by a value $\theta^q = \theta^{q-1}$, $\theta = \theta^q$ is a stationary point of $l(\theta|A_j)$.

The variance-covariance matrix of the estimator θ^q can be calculated in the same way as for the continuous case and the same formulas are derived (see equation 2.11).

Applying the EM algorithm to the problem of the scrambled data sets and assuming a distributional form: $pdf(y^*|\theta,X) = f(y^*|\theta,X)$, an expression for Q can be derived. Assuming that we know $f(y^*|\theta,X)$ we can construct $Pr(Y_{1j}^*, Y_{2j}^*, \ldots, Y_{n_j j}^* | \theta, A_j)$:

$$\begin{aligned}
&Pr(Y_{1j}^* = y_{ij}, \ldots, Y_{n_j j}^* = y_{pj} | A_j, \theta) \\
&= \frac{pdf\left(Y_{1j}^* = y_{ij} | X_{1j}, \theta\right) \ldots pdf\left(Y_{n_j j}^* = y_{pj} | X_{n_j j}, \theta\right)}{\sum_{i=1}^{n_j} \sum_{\substack{k=1 \\ k\neq i}}^{n_j} \ldots \sum_{\substack{p=1 \\ p\neq i,k,\ldots}}^{n_j} pdf\left(Y_{1j}^* = y_{ij} | X_{1j}, \theta\right) \ldots pdf\left(Y_{n_j j}^* = y_{pj} | X_{n_j j}, \theta\right)} \\
&= \frac{f\left(y_{ij} | \theta, X_{1j}\right) \ldots f\left(y_{pj} | \theta, X_{n_j j}\right)}{\sum_{i=1}^{n_j} \sum_{\substack{k=1 \\ k\neq i}}^{n_j} \ldots \sum_{\substack{p=1 \\ p\neq i,k,\ldots}}^{n_j} f\left(y_{ij} | \theta, X_{1j}\right) \ldots f\left(y_{pj} | \theta, X_{n_j j}\right)} \qquad \text{for } \{y_{ij}, \ldots, y_{pj}\} = A_j \\
&= 0 \qquad \text{elsewhere}
\end{aligned} \tag{2.16}$$

[6] See appendix A.2 for the proof of this.

This conditional probability density function is of major importance in the EM algorithm. In the function Q the contribution $\log(f(\theta,y^*))$ of an unobserved latent variable y^* to the log likelihood function is replaced by its expectation over the set of values in which its true value is known to lie. In the present case this leads to:

$$Q(\theta,\theta^{q-1};A_j)=\sum_{i=1}^{n_j}\sum_{\substack{k=1\\k\neq i}}^{n_j}\cdots\sum_{\substack{p=1\\p\neq i,k,..}}^{n_j}\left\{\log\left[f\left(y_{ij}\middle|\theta,X_{1j}\right)\right]+\log\left[f\left(y_{kj}\middle|\theta,X_{2j}\right)\right]+\ldots+\log\left[f\left(y_{pj}\middle|\theta,X_{n_jj}\right)\right]\right\}\cdot$$

$$\frac{f\left(y_{ij}\middle|X_{1j},\theta^{q-1}\right)\ldots f\left(y_{pj}\middle|X_{n_jj},\theta^{q-1}\right)}{\sum_{i=1}^{n_j}\sum_{\substack{k=1\\k\neq i}}^{n_j}\cdots\sum_{\substack{p=1\\p\neq i,k,..}}^{n_j}f\left(y_{ij}\middle|X_{1j},\theta^{q-1}\right)\ldots f\left(y_{pj}\middle|X_{n_jj},\theta^{q-1}\right)}$$

(2.17)

Summing over all matching groups j, we get the following expression:

$$Q(\theta,\theta^{q-1};A_1,\ldots,A_G)=\sum_{j=1}^{G}\sum_{i=1}^{n_j}\sum_{\substack{k=1\\k\neq i}}^{n_j}\cdots\sum_{\substack{p=1\\p\neq i,k,..}}^{n_j}\left\{\log\left[f\left(y_{ij}\middle|\theta,X_{1j}\right)\right]+\log\left[f\left(y_{kj}\middle|\theta,X_{2j}\right)\right]+\ldots+\log\left[f\left(y_{pj}\middle|\theta,X_{n_jj}\right)\right]\right\}\cdot$$

$$\frac{f\left(y_{ij}\middle|X_{1j},\theta^{q-1}\right)\ldots f\left(y_{pj}\middle|X_{n_jj},\theta^{q-1}\right)}{\sum_{i=1}^{n_j}\sum_{\substack{k=1\\k\neq i}}^{n_j}\cdots\sum_{\substack{p=1\\p\neq i,k,..}}^{n_j}f\left(y_{ij}\middle|X_{1j},\theta^{q-1}\right)\ldots f\left(y_{pj}\middle|X_{n_jj},\theta^{q-1}\right)}$$

(2.18)

The Q-function has to be maximized with respect to θ, where θ^{q-1} is given. It has been shown that iteratively maximizing this function using the previous optimal values θ^{q-1} leads to convergence[7]. We only need an arbitrary value θ^0 to start the process and the iterations are finished when $\theta^q = \theta^{q-1}$. Since the θ that is found in this way is not necessarily a maximum, second order conditions should be checked. To be reasonably certain that the maximum found is global, it is necessary to try some different starting values as well.

---

[7] See Dempster, Laird and Rubin (1977).

The part where the expected value of the log likelihood function is computed, using the conditional probabilities that are calculated based on the previous optimal values for the parameters, is called the Expectation step. Q is defined in this first step, while the optimization of Q takes place in the second step, which is called the Maximization step.

Looking at formula (2.18) it can be seen that to calculate Q, a summation over $n_j!$[8] terms has to be performed for each matching group j. When the same value occurs more than once in a matching group this decreases to $\frac{n_j!}{\prod_{i=1}^{v_j} m_i!}$ terms, where v_j is the number of values occurring more than once and m_i is the number of times a value occurs.

For $n_j = 10$ and no double values in the matching group, the summation would already consist of 3,628,800 terms. In our case n_j can be much bigger than 10, so exact calculation of Q does not seem to be feasible for practical reasons like computing time.

The exact calculation in the E-step can be replaced by a Monte Carlo implementation of the E-step (Tanner 1993). A Monte Carlo implementation means that from the distribution of Y* given the data and the current estimate of the parameter vector, a sample of size T with values $(y_{1j,1}^*,\ldots, y_{n_jj,1}^*)$, ..., $(y_{1j,T}^*,\ldots, y_{n_jj,T}^*)$ will be drawn for all matching groups j.

From this sample the different moments of interest can be approximated.

The procedure works as follows:

First, $(y_{1j,1}^*,\ldots, y_{n_jj,1}^*)$, ..., $(y_{1j,T}^*,\ldots, y_{n_jj,T}^*)$ is drawn from $Pr(Y_{1j}^*, Y_{2j}^*,\ldots, Y_{n_jj}^* \mid A_j, \theta^i)$.

So the values for all observations in one matching group have to be drawn simultaneously from the joint distribution function. The values y* are chosen from the set A_j.

---

[8] where n_j is the size of the matching group.

$Q(\theta,\theta^i;A_j)$ is then approximated in the following manner:

$$\sum_{j=1}^{G} \frac{1}{T}\sum_{t=1}^{T} \left\{\log\left[f\left(y_{1j,t}^{*}\middle|\theta,X_{1j}\right)\right]+\ldots+\log\left[f\left(y_{n_j j,t}^{*}\middle|\theta,X_{n_j j}\right)\right]\right\} \tag{2.19}$$

This function can be maximized with respect to θ, after which the new Q can be calculated using the new parameter estimates to obtain the probabilities $Pr(Y_{1j}^*,Y_{2j}^*,\ldots,Y_{n_j j}^* \mid A_j, \theta^{i+1})$. If T is chosen large enough the approximation to Q will be sufficiently accurate and convergence of the EM algorithm will occur. After convergence ($\theta^q = \theta^{q-1}$) the information matrix can be calculated by using:

$$\sum_{j=1}^{G} \left.\frac{\partial^2 l\left(\theta\middle|A_j,X_{1j},\ldots,X_{n_j j}\right)}{\partial\theta^2}\right|_{\theta^q} =$$

$$\sum_{j=1}^{G} \frac{1}{T}\sum_{t=1}^{T} \left.\frac{\partial^2\left\{\log\left[f\left(y_{1j,t}^{*}\middle|\theta,X_{1j}\right)\right]+\ldots+\log\left[f\left(y_{n_j j,t}^{*}\middle|\theta,X_{n_j j}\right)\right]\right\}}{\partial\theta^2}\right|_{\theta^q} +$$

$$\sum_{j=1}^{G} \frac{1}{T-1}\sum_{t=1}^{T} \left(\left.\frac{\partial\left\{\log\left[f\left(y_{1j,t}^{*}\middle|\theta,X_{1j}\right)\right]+\ldots+\log\left[f\left(y_{n_j j,t}^{*}\middle|\theta,X_{n_j j}\right)\right]\right\}}{\partial\theta}\right|_{\theta^q}\right)^2 -$$

$$\sum_{j=1}^{G} \frac{T}{T-1}\left(\frac{1}{T}\sum_{t=1}^{T} \left.\frac{\partial\left\{\log\left[f\left(y_{1j,t}^{*}\middle|\theta,X_{1j}\right)\right]+\ldots+\log\left[f\left(y_{n_j j,t}^{*}\middle|\theta,X_{n_j j}\right)\right]\right\}}{\partial\theta}\right|_{\theta^q}\right)^2 \tag{2.20}$$

A problem with this approach is that there are still $n_j!$ different combinations of y* and x-values for which the probability has to be calculated, since these $n_j!$ probabilities make up the discrete probability function of Y*. All elements of this probability function have to be calculated before observations can be sampled from the function.

However in the bayesian literature much use is made of importance sampling[9], which under certain conditions allows one to draw from a more simple distribution than the actual one. In order to get good results the approximating distribution has to be sufficiently similar to the original distribution. This means that the approximating distribution should have large probability where the actual distribution has large probability, so that all important combinations will be drawn from the importance function. The number of drawings T from the importance function $\gamma(Y_{1j}^*, Y_{2j}^*, \ldots, Y_{n_j j}^* \mid A_j, \theta^q)$ can be less when this function is more similar to the actual distribution.

After sampling from the simple distribution each drawing is weighted by the ratio of: the probability of this drawing according to the actual distribution and the probability according to the simple distribution. When calculating weighted averages over this sample, the constant terms in both probabilities will cancel out. This means in this case that the computation intensive denominator, consisting of the sum over all possible combinations of x and y* values, can be left out of the calculations.

The importance function considered here is constructed from a sequence of probabilities. The procedure for generating data for Y* goes as follows:

Choose one record on X in matching group j from the main data set to start from and calculate probabilities of observing each of the values in the set A_j given a value for θ and X:

$$\Pr\left(Y_{1j}^* = y_{ij} \middle| X_{1j}, \theta^q\right) = \frac{f\left(y_{ij} \middle| X_{1j}, \theta^q\right)}{\sum_{k=1}^{n_j} f\left(y_{kj} \middle| X_{1j}, \theta^q\right)} \quad , \text{for } y_{ij} \in A_j \tag{2.21}$$

Draw a value for Y_{1j}^* from this discrete distribution. Then go to the next record on X and repeat this procedure after removing the value $y_{1j,t}^* = y_{t_1 j}$ that is drawn in the previous step from the set A_j of available values. t_1 indicates the t^{th} simulated y* value belonging to X_{1j} and its value lies in between 1 and n_j.

---

[9] See Kloek and Van Dijk (1978 and 1980).

The probability distribution function then looks like:

$$\Pr\left(Y_{2j}^{*} = y_{ij} \middle| X_{2j}, \theta^{q}\right) = \frac{f\left(y_{ij} \middle| X_{2j}, \theta^{q}\right)}{\sum_{\substack{k=1 \\ k \neq t_1}}^{n_j} f\left(y_{kj} \middle| X_{2j}, \theta^{q}\right)} \quad , \text{for } y_{ij} \in A_j \setminus \{y_{t_1 j}\} \tag{2.22}$$

where A \ B means the set A after deletion of the elements in set B.

From this distribution function $y_{t_2 j}$ will be drawn. In the next step $y_{t_1 j}$ and $y_{t_2 j}$ will be excluded from the set of possible values for Y*.
After each draw the set of attainable values for Y* contains one value less. Following this procedure a value for Y* will be drawn for all records on X in the matching group in a sequential way. For each series of draws of y*-values the procedure will start from a different record on X. This is done, since the probability of certain combinations occurring is dependent on the order of observations X_{ij} ($i=1,\ldots,n_j$) for which y*-values are drawn. The last observation on X has to be combined with the last remaining value from A_j no matter how unlikely that combination is. Starting from the same observation on X over and over again could disadvantage certain combinations of values y* and x. By alternating the starting point it is hoped that all combinations that would occur with high probability according to the actual probability distribution, will also be drawn from this much simpler importance function.

Although an attempt is made to get as close as possible to drawing from the actual probability distribution function, we know that this will never exactly be the case. To correct for this drawing from an incorrect distribution, the simulated contributions to the log likelihood have to be weighted. The appropriate weights can be found by dividing the value of the actual probability density function by the approximated

value at the simulated data point[10]:

$$w_{tj} = \frac{f\left(Y_{1j}^* = y_{t_1 j} \middle| X_{1j}, \theta^q\right) \ldots f\left(Y_{n_j j}^* = y_{t_{n_j} j} \middle| X_{n_j j}, \theta^q\right)}{\frac{f\left(Y_{1j}^* = y_{t_1 j} \middle| X_{1j}, \theta^q\right)}{\sum_k f\left(Y_{1j}^* = y_{kj} \middle| X_{1j}, \theta^q\right)} \cdot \frac{f\left(Y_{2j}^* = y_{t_2 j} \middle| X_{2j}, \theta^q\right)}{\sum_{\substack{k \\ k \neq t_1}} f\left(Y_{2j}^* = y_{kj} \middle| X_{2j}, \theta^q\right)} \cdot \ldots .1} \tag{2.23}$$

Using these weights, $Q(\theta,\theta^q;A_j)$ is now approximated by:

$$\sum_{j=1}^{G} \frac{1}{\sum_{t=1}^{T} w_{tj}} \sum_{t=1}^{T} w_{tj} \left\{ \log\left[f\left(y_{t_1 j} \middle| \theta, X_{1j}\right)\right] + \ldots + \log\left[f\left(y_{t_{n_j} j} \middle| \theta, X_{n_j j}\right)\right]\right\} \tag{2.24}$$

and the information matrix can be approximated by using:

$$\begin{aligned}
&\sum_{j=1}^{G} \left. \frac{\partial^2 l\left(\theta \middle| A_j, X_{1j}, \ldots, X_{n_j j}\right)}{\partial \theta^2} \right|_{\theta^q} = \\
&\sum_{j=1}^{G} \frac{1}{\sum_{t=1}^{T} w_{tj}} \sum_{t=1}^{T} w_{tj} \left. \frac{\partial^2 \left\{ \log\left[f\left(y_{t_1 j} \middle| \theta, X_{1j}\right)\right] + \ldots + \log\left[f\left(y_{t_{n_j} j} \middle| \theta, X_{n_j j}\right)\right]\right\}}{\partial \theta^2} \right|_{\theta^q} + \\
&\sum_{j=1}^{G} \frac{1}{\sum_{t=1}^{T} w_{tj}} \sum_{t=1}^{T} w_{tj} \left(\left. \frac{\partial \left\{ \log\left[f\left(y_{t_1 j} \middle| \theta, X_{1j}\right)\right] + \ldots + \log\left[f\left(y_{t_{n_j} j} \middle| \theta, X_{n_j j}\right)\right]\right\}}{\partial \theta} \right|_{\theta^q} \right)^2 - \\
&\sum_{j=1}^{G} \left(\frac{1}{\sum_{t=1}^{T} w_{tj}} \sum_{t=1}^{T} w_{tj} \left. \frac{\partial \left\{ \log\left[f\left(y_{t_1 j} \middle| \theta, X_{1j}\right)\right] + \ldots + \log\left[f\left(y_{t_{n_j} j} \middle| \theta, X_{n_j j}\right)\right]\right\}}{\partial \theta} \right|_{\theta^q} \right)^2
\end{aligned} \tag{2.25}$$

[10] Note that only the numerator of both density functions appears in (2.23). The denominators are constants and can be left out of the formula.

2.5 Usefulness of the EM algorithm

By using the EM algorithm in this case we can use the information we have (exact matches and groups of possible values for the observations that could not be matched) to estimate the parameters of a model. We would expect the usefulness of this method to depend on the size of the matching groups (smaller groups give more accurate information, since there is less ambiguity).

An important question is whether this approach is feasible.

The following procedure could be used: divide the data into matching groups and choose starting values. Use these starting values to calculate the probabilities for each of the possible values per observation, taking into account the one to one relation of observations on X and observations on Y*. A sample of y*-values can be drawn from this discrete probability function for each observation on X. A weighted average calculated over the sample as described in section 2.4 can be used as an approximation to Q, the expected value of the log likelihood. The approximate Q function can then be maximized with respect to θ. The maximization will result in new values for θ that can then be used to calculate the probabilities again. For the larger matching groups this means a lot of computations since we have to calculate probabilities for each of the possible values per observation. For each iteration all these values have to be calculated again. Furthermore it could be necessary to increase the number of Monte Carlo replications T for the larger matching groups in order to get a realistic approximation to the part of Q that is contributed by those matching groups.

Although for large matching groups the number of calculations can be a computational burden, the calculations themselves are quite straightforward and easy to perform.

To give us more insight into the way the procedure works and the results it generates, a simple comparative Monte Carlo experiment will be performed and described in the next section.

3 Comparative Monte Carlo experiment

Suppose there is a simple relation between Y^* and X: $\ln(Y^*) = \beta X + u$, where u is an error term with a known distribution P_u. For the examples in this paper the normal distribution with mean zero and a variance term σ^2 is chosen, but theoretically any distribution could be used.

In our experiment each observation on X_i has one or more (say n_j) possible values $(y_{j1}, y_{j2}, \ldots, y_{jn_j})$ for Y^*_i. If there is only one possible value the usual contribution to the likelihood function can be maintained. In case of two or more possible values for Y^*_i this variable could be called latent, since no precise value is observed. In this case a set of values is observed instead of one value. However there is only one true value for the latent variable. The exact value is unknown for most observations. It is only known that this latent variable is an element of the set of observed values $\{Y_{j1}, Y_{j2}, \ldots, Y_{jn_j}\}$.

In this section two different situations will be distinguished. The case where the variables in Z (the matching variable(s)) do not involve the endogenous variable Y^* will be described in subsection 3.1. The case where one of the matching variables in Z is a categorized version of the dependent variable will be analysed in subsection 3.2.

3.1 Matching and dependent variables are non overlapping

3.1.1 Design of the simulation experiment

To compare the results obtained by using the EM algorithm on the scrambled data (formed by two incomplete data sets) with the results that would be obtained if the complete data set was available for estimation, a small data set is generated. The artificial data consist of variables X, Y^* and Z. X is generated by drawing a sample (of the desired size) from a normal distribution. Y^* is generated from the equation $\ln(Y^*) = \beta_1 + \beta_2 X + u$, where u represents white noise and β_1 and β_2 are known

parameters[11]. A value for the β's and X and a distribution for u have to be chosen, so that values for Y can be generated randomly. Z is generated by assigning group numbers from 1 to 45 to each observation. This artificially created data will be used to generate scrambled data by separating the data into two (incomplete) data sets: one with only X and Z and another with only Y (= Y* in a different order) and Z.
We know the original equation underlying these data and we can try to estimate the equation from the two separate data sets using the adapted EM algorithm. To have a standard of comparison the equation will be estimated from the complete data set as well. In order to get some feel for the quality of the estimates from the scrambled data as compared to the estimates from the complete data, the process of generating data will be repeated a thousand times, so an average value for the parameters can be estimated. The only change made in each replication is that a new sample of white noise values is drawn. The values of the variables X and Z will remain the same throughout the replications.
Some variations on the case described before will also be simulated and compared to the basic version.
The basic situation looks as follows:
To generate the two data sets the following equation is used:

$$\ln(Y^*) = 2 + 0.5^*X + u, \text{ where } u \sim N(0,1) \qquad (3.1)$$

$$\text{So here } f(Y^*,\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(\ln(Y^*) - \theta'\begin{pmatrix}1\\X\end{pmatrix})^2\right)$$

where $\theta = (\theta', \sigma^2)$ (= (2, 0.5, 1) in this case).

X is generated from N(3,4). A sample of 100 observations is generated by creating 100 values for X and Z, and by calculating the corresponding values for Y*.
Z is a vector that defines the matching groups. The values in Z are assigned in such a way as to form the desired matching group sizes. The data set created in this way is our basic data set that will be used in our simulations.

---

[11] Since our variable of interest is hours of labour supply, we want to generate a Y that looks like labour supply. Assuming a lognormal distribution for Y takes into account that negative hours cannot be observed.

For each replication an error term u (coming from the standard normal) will be added to ln(Y*) for all observations in the sample. This gives us the data set on which the estimations will be performed. Thus the only difference between each replication is the value of the error term u.

After this data set with X, Y* and Z is generated, two new files are created. One contains Z and Y while the other contains Z and X. The common factor between the two data sets is Z. Some of the records can be reconnected, but others have two or more possible records in the other file to which they can be connected. The drawn sample contains 45 different values for observations on Z. This means 45 groups containing from 1 to 7 different values for Y can be formed.

Besides this basic situation some interesting alternatives are also analysed:
a) generated data sets where the error term has a smaller variance.
b) generated data sets with more than 100 observations.
c) generated data sets where the average size of the matching groups is increased.

Besides the basic situation where u is the standard normal, an analysis of the effect of a different variance for the distribution of u will be performed. The amount of variance is an indication of the size of R^2. A large variance will cause a small R^2, since the error term will disguise the true relation between Y* and X to a greater extent. It is expected that the estimation of the parameters of a strong relation between Y* and X will suffer less from being scrambled than a weaker relation.
The second alternative explores the effect of increasing the size of the data set, while all other factors remain constant. The data set that we are planning to work with will have the considerable size of 4330 income units. This means that size is of major importance for our problem.
The last alternative is explored because it is expected that a pair of incomplete data sets that have small matching groups will perform better in the estimation procedure than a pair having an equal total number of observations but with large matching groups. In the first case more exact information is available, since within the matching groups less ambiguity surrounds the variable of interest. By increasing the average size of the matching groups and keeping all other factors constant the

effect of the average size of matching groups on the quality of the estimates can be investigated formally.

3.1.2 The simulation results

In tables A.3.1 and A.3.2 of appendix A.3 the results of the different Monte Carlo experiments are shown. In this section the properties of the estimated parameters for the different situations will be discussed; e.g. what is the bias, what is the mean square error and how close are the estimates to the actual values.

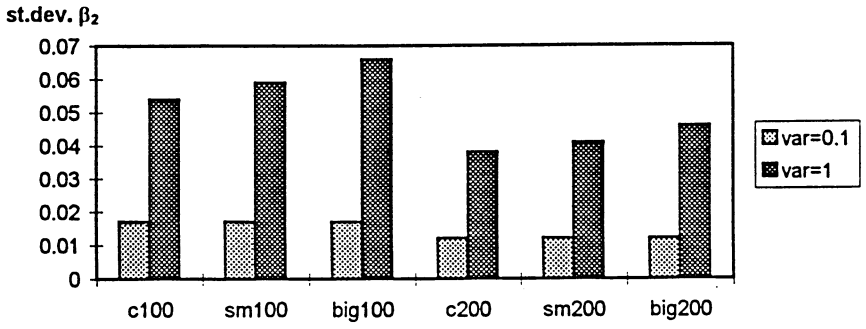
Before going into the different alternatives that are simulated, we want to make a general remark on the estimated standard deviations. For comparison in all simulations both the empirical standard deviation and the mean of the estimated standard deviations (using (2.25)) are calculated. The empirical standard deviation for a parameter estimate is defined over the 1000 replicated estimates of that parameter. In table A.3.4 both standard deviations are presented. From this table it is clear that in all cases the average estimated standard deviation is very close to the empirical version.

a Varying the variance of the error term

Comparing columns eight and nine with columns four and five in table A.3.1 in appendix A.3 it can be seen that a larger variance of the error term results in an increase in the standard deviations of our estimates from the scrambled data. In the complete data situation the standard deviations increase as well with an increase of the variance but at a lesser degree. In all cases the complete data provides estimates with smaller standard deviations. The *difference* in size of the standard deviations between estimates from complete data and estimates from scrambled data increases as the variance of the error term increases.

These results are illustrated in figure 3.1 for the estimated parameter of the X-variable (β_2). When the variance of the error term equals 0.1, using complete data or scrambled data with small matching groups or scrambled data with big matching groups gives similar results on the standard deviations. Increasing the variance to 1 will increase the standard deviations of the parameters more for the scrambled data than for the complete data, especially when the matching group size is larger.

Figure 3.1 Changing the variance of the errorterm



c = complete data

sm = scrambled data with small matching groups

big = scrambled data with big matching groups

100 = sample size is 100 observations

200 = sample size is 200 observations

The same conclusions can also be drawn from the percentage of cases in which the estimated values for the parameters are within 5% of their true values (near1 to near3 in the table A.3.2). The percentage is smaller for the scrambled data. For increasing variance of the error term this percentage will decrease both for the complete and the scrambled data, but the decrease will be relatively larger for the scrambled data.

In both the complete and scrambled data situation the estimates remain unbiased (see table A.3.2).

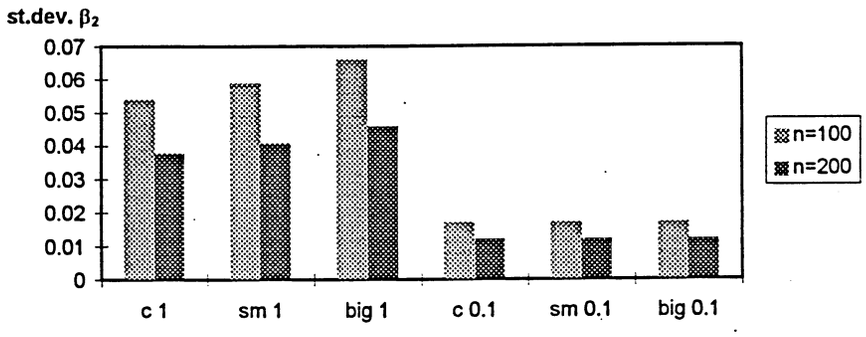
b Varying the sample size

The upper part of table A.3.1 contains the simulations where the sample size is one hundred observations and the lower part contains the simulations where the sample size is two hundred observations. A larger sample size decreases the standard deviation for both complete and scrambled data. This is true for the cases of both smaller and larger matching groups. More information improves the quality of the estimates. This is what one would expect from a consistent estimator.

In figure 3.2 the standard deviation of β_2 can be compared for the different cases. With an increase in the sample size, the performance of the estimators (measured

by the standard deviation of the parameters) in the complete and the scrambled case seems to improve at about the same rate for all three estimators.

Figure 3.2 Changing the sample size



c = complete data

sm = scrambled data with small matching groups

big = scrambled data with big matching groups

1 = variance of the error term is 1

0.1 = variance of the error term is 0.1

The Monte Carlo results suggest that the adapted EM algorithm produces consistent estimates under the experimental design adopted above. That is, the sample means of the parameters seem to be correctly located at the values adopted for the underlying data-generating mechanism, and the precision of their location appears to improve with increasing sample size (see table A.3.2). Moreover the standard deviation of the sampling distribution appears to be declining with sample size.

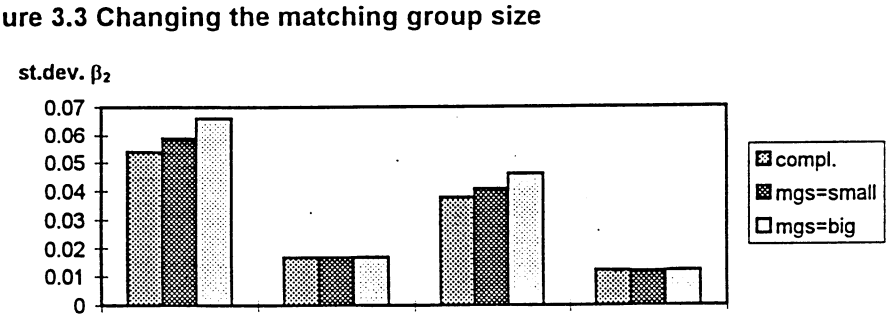
c Varying the matching group size

Comparing column four with five and column eight with nine in tables A.3.1 and table A.3.2 indicates the effect of an increase in average matching group size. Columns five and nine contain the simulations where the group size has increased. It can be seen that when matching groups are larger on average the standard deviation of the estimates increases. In figure 3.3 this is illustrated for β_2. Especially for larger variances of the error term it is obvious from this figure that a larger matching group size increases the standard deviation of the estimates.

Intuitively this seems reasonable, since larger matching groups mean the available information is more ambiguous. In a larger matching group there is a larger number

of different combinations of y and x values that can be made, given the available information.

Figure 3.3 Changing the matching group size



v = variance of the error term

n = the sample size

mgs = matching group size (on average)

In this contrived example the amount of precision in estimation lost as a result of an increase in the size of matching groups seems modest. In a more realistic situation, this may not be the case.

When estimating parameters with scrambled data, average group size seems to be important, although the effects observed here are small.

Summarizing the results we can say that the effects of variance of the error term, sample size and matching group size on the quality of the estimates are all as expected. Increasing the variance and matching group size have a negative effect on the quality and increasing the sample size has a positive effect.

3.2 Part of the matching and dependent variables are overlapping

3.2.1 Design of the simulation experiment

In the same way as described in 3.1 another data set can be generated. The difference is that in this case one of the matching variables will be a categorized version of the dependent variable. We are interested in the difference in quality of the results that using the additional information on the more detailed values of the

dependent variable will give us, compared to the results from the grouped data alone.
Apart from this dependence between one of the matching variables and the dependent variable, the same basic situation as in the previous section and some variations on that basic situation will be explored. However an additional estimation will be analysed. Besides the estimation using the complete data set and the estimation using the two separate data sets generated from the complete data, also estimates from just the data set containing the categorized version of the dependent variable will be explored.

Our basic situation will be generated in the same way as in 3.1, the only difference being an additional matching variable in the form of a categorized version of our dependent variable Y*. This causes our matching groups to be different for each replication, since the matching groups depend on the values of our matching variables and one of the matching variables changes for each replication with the value of Y*.

We will only investigate two alternatives besides the basic situation:
a) more matching groups, keeping the width of the intervals of the dependent variable constant.
b) fewer matching groups by increasing the width of the intervals.

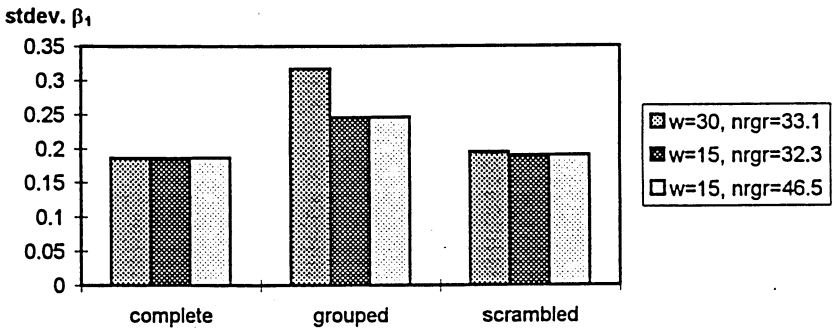
Here we are only interested in the difference between estimating from the grouped data (with no further information) and estimating from the scrambled data. Does the additional information in the scrambled data as compared with the grouped data give better results?

3.2.2 The simulation results

All the results that are discussed in this section can be found in table A3.3. For all the three cases mentioned before, estimation using the information on the possible values within a group (method A, say) provides better results than estimation using just the grouped data (method B, say).
The fewer elements in the matching groups the better the results for method A. So using method A instead of B gets more worthwhile when there is more additional

information in the scrambled data. This can be seen by comparing column 8 with column 6. Here there is only a slight difference in the quality of the estimates in the two cases. Since the width of the interval of the grouped dependent variable remains the same, column 7 and column 9 show identical results.
This means that changing the average matching group size (while keeping the width of the intervals, in which the categorized version of the dependent variable can be divided, constant) will improve the quality of the estimates when using method A.

Figure 3.4 Changing the width and the size of groups



w = width of the interval of the grouped dependent variable
nrgr = average number of matching groups in the data sets of the Monte Carlo experiment

Changing the number of groups by changing the width of the intervals of the dependent variable from 15 to 30, will affect both the results from method A and method B. Comparing column 6 with column 4 and column 7 with column 5 shows that method A suffers less from the increase in interval width than method B. The standard deviations for the wider intervals are slightly larger and the nearness gets slightly less when using method A. Method B shows a larger deterioration in the quality of the estimation results. Figure 3.4 shows this clearly as well.
It can be seen that changing the number of groups via width of the categorized dependent variable has a larger impact on the quality of estimation than changing the number of groups via one of the other matching variables.
From columns 4 and 8 the importance of the maximum width of the groups (indicating the range of values a variable within one matching group can take) for the quality of the estimates from scrambled data can be seen even more clearly. The average number of groups in column 8 is slightly smaller than the average

number of groups in column 4. This means that the average matching group size in column 8 is slightly bigger. Nevertheless the standard deviations and mean square errors in column 8 are slightly smaller. This indicates that the narrowing of the categories in itself has a positive effect on the quality of the estimates.
Further evidence for this can be found by comparing the results here with the results in table A.3.1. There the same basic data sets and errors are used, but the division in matching groups is not determined by a categorized version of the dependent variable Y. This means that the range of possible values for Y in a matching group can be much wider than in the simulations of table A.3.3.
The number of groups in the case of the small matching groups is 45 for samples of 100 observations. The standard deviations and the mean square errors in the upper part of column 4 of table A.3.1 and A.3.2 are larger than the standard deviations and mean square errors in columns 4 and 8 of table A.3.3, even though the average matching groups are bigger in size in case of the latter table.
Summarizing we can say that the performance of method A relative to method B improves in situations where the intervals, in which the dependent variable is grouped, are wider and where the scrambled data contains more additional information (i.e. the matching groups are smaller on average). In addition we note that a narrower range of possible values for Y within a matching group has a positive effect on the quality of the estimates.

4 Conclusion and perspective for further research

This paper has proposed a method for using scrambled data as part of the process of estimating a behavioural equation of interest. Monte Carlo trials for a simple linear model indicate that the method works quite well. The main difference between estimating from complete data and estimating from scrambled data lies in the size of the precision of the estimators. Standard deviations are bigger for the scrambled data. A higher noise to signal ratio will increase the size of the standard deviation more for scrambled data than for complete data.
However the simulations seem to suggest that our estimators will still be consistent. Furthermore the Monte Carlo trials suggest that the corresponding estimators of the standard deviations (of estimated parameters) are unbiased.

All of the above implies that the EM algorithm can be useful to get information from a scrambled data set. The next step is to apply this method to our real data set, which of course is the main goal of this study.

In the simulations we find that a larger average matching group size will increase the standard deviation of the estimates. In the data set of interest some matching groups consist of a few hundred observations. Hopefully the negative effect flowing from such a large group size will be counteracted by the positive effect of a large sample size ($N \approx 4000$). Further, it is possible that the availability in the data set of interest of a matching variable (actual hours worked), that is a categorized version of the endogenous variable under investigation, could enhance the precision of estimation in one of the models. At least we are confident that the estimated standard deviations will correctly indicate the accuracy of our estimates.

References

Dempster, A.P., Laird, N., and Rubin, D.B. (1977), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society B*, vol. 39, pp 1-38.

Hammersley, J.M. and Handscomb, D.C. (1964), *Monte Carlo Methods,* Methuen's Monographs on applied probability and statistics.

Kloek, T. and H.K. van Dijk (1978), "Bayesian estimates of equation system parameters: an application of integration by Monte Carlo", *Econometrica*, vol. 46, no.1, pp 1-19.

Louis, T.A. (1982), "Finding the observed information matrix when using the EM algorithm", *Journal of the Royal Statistical Society,* series B, vol. 44, pp 226-233.

Rao, C.R. (1973), *Linear Statistical Inference and Applications,* New York, Wiley.

Ruud, P.A. (1991), "Extensions of estimation methods using the EM algorithm", *Journal of Econometrics*, vol. 49, pp. 305-341.

Tanner, M.A. (1993), *Tools for Statistical Inference; Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer series in statistics, Springer-Verlag

Van Dijk, H.K. and T. Kloek (1980), "Further experience in bayesian analysis using Monte Carlo integration", *Journal of Econometrics*, vol. 14, pp 307-328.

Appendix A.1: Details on the data set of interest

A.1.1 Interesting features of the data

The data come from the unit record file of the ABS income distribution survey 1986. The main part of it comes from the regular file. It contains many background characteristics for all individual persons in the surveyed household who were 15 years or older at the moment of the survey. These background characteristics consist for example of age (in classes), highest educational qualification, field of education, occupation and industry. In addition to these characteristics we also have very detailed information on income received by each person. Even more important for our research is that there is information on number of hours worked. In the general file this has been aggregated into intervals (e.g., 0 to 9 hours per week, 10 to 19 hours per week, 20 to 24 hours per week, 25 to 29 hours per week, etc. till the last category of more than 50 hours per week) even though it has been asked in terms of an exact number of hours. Other information asked in the survey but not included in the general unit record file consists of preferred hours of work, reasons for working a different number of hours from the preferred number and questions on the desire to work if someone had not been looking for a job during the last four weeks.

This additional information can be very useful for research on labour supply. Therefore the ABS was approached to get these additional data. Although the information was not readily available any more (the survey was in 1986), the ABS allowed us to copy the necessary file from another university.

A.1.2 Common information available for matching

The main file includes observations on "hours worked in main job", "hours worked in second job" and "hours worked in all jobs". These variables have been aggregated into 10 classes, 4 classes and 10 classes respectively. In the smaller file with additional information the exact number of hours is available, so the information on hours can be used to connect the records. This information is available for all members of the household over 15 years of age. Although the household identity numbers are different, both files contain exactly the same households. Since the number of individual records for a particular household is the same in both files, this

information can also be used in an effort to connect the two files. The records have to be matched on a household basis, so groups of records representing a household will be formed and these groups have to be matched to the corresponding group in the other file.

To find the matching records the records in each file were ordered firstly according to the number of records per household, secondly according to household identity number (to create the groups per household), thirdly according to hours worked in main, second and all jobs of the head of the household and fourthly according to family number, income unit number and person number. In this way the data are grouped in a similar way in both data files.

Households with a unique combination of the above variables and hours worked of the other household members can be exactly matched. If more than one household shares the same recorded values of these characteristics, a choice will be necessary when attempting to link the two files.

After the matching procedure is finished about 17 % of the individual records is exactly matched.

For the analyses we only want to use households that contain at most one income unit and consist of a head and a partner with or without dependants. The data are also rearranged into records per income unit instead of individual records and detailed information is only kept on the head and the partner, besides the information on the total income unit. Only a few characteristics on the dependants are kept.

Starting from a total data set of 17,714 individual records, 4330 income units remain to be analysed after the above selection. From these income units 262 were exact matches; thus the number of exact matches dropped from about 17 % to 6.1 %.

This seems a dramatic drop. However there are some explanations. The first one is that the income unit of interest is one of the most common and therefore probably more difficult to be matched exactly. The second explanation is that we changed from records per individual to records per income unit. Income units with more records were easier to match, but all these separate individual records were transformed into just one income unit record. This causes the expected percentage of exactly matched individuals to be higher than the expected percentage of exactly matched income units.

A.1.3 Different types of variables in the supplementary data file

The possibilities might differ for the different variables in the additional data file. Actual hours worked is the most accurate variable since a grouped variant of the actual hours worked was used to do the matching. So we know that the chosen values are at least in the right range.

For all other variables we have no idea whether the chosen values are the right ones. In fact, the values we have to choose from can be very different, since the matching variables and the additional variables in the extra data file are probably not (very strongly) correlated. So we have a group of values that can be very different and only one of the values is the right one but we do not know which one that is. The data base reveals that the number of records in the group from which one record has to be chosen can range from 1 to 965. However, the number of different values for the variable of interest that these records represent, will be at most 90.

Appendix A.2 Rao's theorem in the discrete case

Following the same steps as in 1e.6 from Rao (1973):

Given a function f and g representing a probability mass function, $y^* \in A$ can be defined so that $f(y^*)>0$.

Then if $\sum_{y^* \in A} f(y^*) \geq \sum_{y^* \in A} g(y^*)$ it can be proven that .

$$\sum_{y^* \in A} f(y^*) \log\left(\frac{f(y^*)}{g(y^*)}\right) \geq 0 \qquad (A.1)$$

Proof

Use

$$\log(x) = (x - 1) - \frac{(x - 1)^2}{2r^2} \qquad r \in (1, x) \qquad (A.2)$$

So

$$\sum_{y^* \in A} f(y^*) \log\left(\frac{g(y^*)}{f(y^*)}\right) = \sum_{y^* \in A} f(y^*) \left[\left(\frac{g(y^*)}{f(y^*)} - 1\right) - \frac{\left(\frac{g(y^*)}{f(y^*)} - 1\right)^2}{2r^2} \right] \quad r \in \left(1, \frac{g(y^*)}{f(y^*)}\right) \qquad (A.3)$$

This is equivalent to:

$$\begin{aligned} &\sum_{y^* \in A} (g(y^*) - f(y^*)) - \sum_{y^* \in A} f(y^*) \left[\frac{(g(y^*) - f(y^*))^2}{2r^2 f(y^*)^2} \right] = \\ &\sum_{y^* \in A} g(y^*) - \sum_{y^* \in A} f(y^*) - \sum_{y^* \in A} f(y^*) \left[\frac{(g(y^*) - f(y^*))^2}{2(r^*)^2} \right] \leq 0 \qquad r^* \in \left(f(y^*), g(y^*)\right) \subset (0,1) \end{aligned} \qquad (A.4)$$

However, if $\sum_{y^* \in A} f(y^*) \log\left(\frac{g(y^*)}{f(y^*)}\right) \leq 0$, it is easy to see that $\sum_{y^* \in A} f(y^*) \log\left(\frac{f(y^*)}{g(y^*)}\right) \geq 0$.

Appendix A.3 Tables comparing the simulation results

Table A.3.1 Average parameter estimates and average standard deviations for the estimates for a simple linear model using complete and scrambled data sets with different matching group sizes. Sample size and variance is varied in the experiment (1000 replications per case).

	n=100, σ^2=1				n=100, σ^2=0.1			
	theor.[a]	compl.[b]	scrambled[c] small groups	scrambled big groups	theor.	compl.	scrambled small groups	scrambled big groups
1	2	3	4	5	6	7	8	9
β_1 (1)	2.0	2.007	2.006	2.003	2.0	2.002	2.002	2.002
β_2 (2)	0.5	0.499	0.499	0.500	0.5	0.500	0.500	0.500
σ^2 (3)	1.0	0.980	0.977	0.971	0.1	0.098	0.098	0.098
stdev1	0.190	0.187	0.198	0.216	0.060	0.059	0.059	0.060
stdev2	0.055	0.054	0.059	0.066	0.017	0.017	0.017	0.017
stdev3	0.141	0.139	0.156	0.182	0.014	0.014	0.015	0.017
R^2 [d]	0.440	0.444			0.889	0.891		
	n=200, σ^2=1				n=200, σ^2=0.1			
1	2	3	4	5	6	7	8	9
β_1 (1)	2.0	1.998	2.000	1.998	2.0	1.999	2.000	2.000
β_2 (2)	0.5	0.500	0.500	0.501	0.5	0.500	0.500	0.500
σ^2 (3)	1.0	0.989	0.990	0.986	0.1	0.099	0.099	0.099
stdev1	0.132	0.131	0.140	0.152	0.042	0.041	0.042	0.042
stdev2	0.038	0.038	0.041	0.046	0.012	0.012	0.012	0.012
stdev3	0.100	0.099	0.114	0.134	0.010	0.010	0.011	0.012
R^2	0.462	0.465			0.897	0.898		

[a] Values of the parameters used in the data generation for Monte Carlo work

[b] Maximum likelihood estimates based on complete data (i.e. data that is not scrambled)

[c] EM estimates from the scrambled data

[d] Coefficient of determination

Table A.3.2 Average bias, average mean square errors and a nearness measure for the estimated parameters from table A.3.1.

	n=100, σ^2=1				n=100, σ^2=0.1			
	theor.[a]	compl.[b]	scrambled[c] small groups	scrambled big groups	theor.	compl.	scrambled small groups	scrambled big groups
1	2	3	4	5	6	7	8	9
bias1		0.007	0.006	0.003		0.002	0.002	0.002
bias2		-0.001	-0.001	-0.000		-0.000	-0.000	-0.000
bias3		-0.020	-0.023	-0.029		-0.002	-0.002	-0.002
mse1	0.036	0.034	0.039	0.049	0.004	0.003	0.003	0.004
mse2	0.003	0.003	0.003	0.005	0.000	0.000	0.000	0.000
mse3	0.020	0.018	0.024	0.035	0.000	0.000	0.000	0.000
near1[d]		39.4%	38.5%	32.6%		92.1%	91.9%	91.3%
near2		34.6%	33.4%	26.4%		86.1%	86.8%	85.5%
near3		26.8%	22.5%	20.1%		26.8%	25.0%	23.4%
	n=200, σ^2=1				n=200, σ^2=0.1			
1	2	3	4	5	6	7	8	9
bias1		-0.002	-0.000	-0.002		-0.001	-0.000	-0.000
bias2		0.000	-0.000	0.001		0.000	0.000	0.000
bias3		-0.011	-0.010	-0.014		-0.001	-0.001	-0.001
mse1	0.017	0.018	0.021	0.025	0.002	0.002	0.002	0.002
mse2	0.001	0.001	0.002	0.002	0.000	0.000	0.000	0.000
mse3	0.010	0.009	0.013	0.018	0.000	0.000	0.000	0.000
near1		54.5%	50.9%	46.0%		98.0%	98.0%	97.5%
near2		49.1%	45.6%	40.2%		96.0%	96.3%	96.0%
near3		40.5%	36.2%	28.6%		40.5%	37.4%	31.2%

[a] Calculated theoretical values for a specific data set

[b] Maximum likelihood estimates based on complete data (i.e. data that is not scrambled)

[c] EM estimates from the scrambled data

[d] Percentage of estimates within 5 % of the actual value

Table A.3.3 Estimated parameters, standard deviations, biases, mean square errors and indicators of the nearness to the actual value from the complete, the grouped[a] and the scrambled data sets (1000 replications for each case).

			average nr. of groups=33.1, width[b] =30		average nr. of groups=46.5, width=15		average nr. of groups=32.3, width=15	
	theor.[c]	compl.[d]	scrambled[e] (method A)	grouped (methodB)	scrambled (method A)	grouped (method B)	scrambled (method A)	grouped (method B)
1	2	3	4	5	6	7	8	9
β_1 (1)	2.0	2.007	2.006	1.992	2.009	1.998	2.007	1.998
β_2 (2)	0.5	0.499	0.499	0.502	0.498	0.501	0.498	0.501
σ^2 (3)	1.0	0.980	0.978	0.998	0.981	0.992	0.980	0.992
stdev1	0.190	0.187	0.194	0.317	0.190	0.246	0.191	0.246
stdev2	0.055	0.054	0.057	0.081	0.056	0.068	0.056	0.068
stdev3	0.141	0.139	0.151	0.229	0.144	0.184	0.145	0.184
bias1		0.006	0.006	-0.008	0.009	-0.002	0.007	-0.002
bias2		-0.001	-0.001	0.002	-0.002	0.001	-0.002	0.001
bias3		-0.020	-0.022	-0.002	-0.019	-0.008	-0.020	-0.008
mse1	0.036	0.034	0.037	0.097	0.035	0.059	0.036	0.059
mse2	0.003	0.003	0.003	0.006	0.003	0.005	0.003	0.005
mse3	0.020	0.018	0.022	0.058	0.019	0.035	0.020	0.035
near1 [f]		39.4%	38.7%	24.9%	38.2%	32.5%	39.2%	32.5%
near2		34.6%	35.2%	25.5%	35.2%	30.8%	33.8%	30.8%
near3		26.8%	25.7%	16.6%	27.3%	21.1%	26.0%	21.1%
R^2 [g]	0.440	0.444						

[a] The dependent variable is only available in grouped form and is one of the matching variables for the scrambled data.

[b] Width of the intervals in which exp(dependent variable) is divided.

[c] Values of the parameters used in the data generation for Monte Carlo work

[d] Maximum likelihood estimates based on complete data (i.e. data that is not scrambled)

[e] EM estimates from the scrambled data

[f] Percentage of estimates within 5 % of the actual value

[g] Coefficient of determination.

Table A.3.4 Comparison of empirical and estimated standard deviations of the parameters[a] estimated from scrambled data for different sample sizes, matching group sizes and variances of the error term.

	β_1		β_2		σ^2	
	est. st.d.	emp. st.d.	est. st.d.	emp. st.d.	est. st.d.	emp. st.d.
small groups						
n=100, σ^2=0.1	0.059	0.059	0.017	0.017	0.015	0.015
n=100, σ^2=1	0.198	0.197	0.059	0.058	0.156	0.154
n=200,σ^2=0.1	0.042	0.043	0.012	0.012	0.011	0.011
n=200, σ^2=1	0.140	0.145	0.041	0.041	0.114	0.112
bigger groups						
n=100, σ^2=0.1	0.060	0.059	0.017	0.017	0.017	0.017
n=100, σ^2=1	0.216	0.222	0.066	0.068	0.182	0.184
n=200, σ^2=0.1	0.042	0.043	0.012	0.012	0.012	0.012
n=200, σ^2=1	0.152	0.159	0.046	0.046	0.134	0.135

[a] This table is formed by using the results from the same simulations as are used to construct table A.3.1 and A.3.2. The columns with the estimated standard deviations (est. st.d.) contain numbers that can also be found in tables A.3.1 and A.3.2.