The problem of not observing small expenditures in a consumer expenditure survey

Melenberg, B.; Alessie, R.J.M.; Gradus, R.H.J.M.

Published in:
Journal of Applied Econometrics

Publication date:
1990

JOURNAL OF APPLIED ECONOMETRICS, VOL. 5, 151–166 (1990)

THE PROBLEM OF NOT OBSERVING SMALL EXPENDITURES IN A CONSUMER EXPENDITURE SURVEY

ROB ALESSIE, RAYMOND GRADUS and BERTRAND MELENBERG
Department of Econometrics, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands

SUMMARY

In consumer expenditure surveys one often faces the problem that full information on small consumption expenditures is not available. Suppose a panel of households is available, which is divided into two subsamples, say, A and B. For subsample B all expenditures are registered, but from subsample A only expenditures above some fixed amount, say Z dollars. Suppose, furthermore, that in some economic analysis the use of the sum of all expenditures of each single observation in the sample is required. It is then evident that the use of the (observed) sum of expenditures *above* Z dollars instead of the (unobserved) sum of all expenditures in case of observations in subsample A will lead to underestimation of the sum of *all* expenditures. To correct for this underestimation one could, for instance, make use of blow-up factors, computed with subsample B, or one could construct a Tobit model explaining the sum of expenditures below Z dollars and use this model, after estimating it with subsample B, to predict the sum of expenditures below Z dollars in subsample A. In this paper we propose an alternative method to correct for the underestimation. The method consists of constructing a model, which explains the sum of expenditures below Z dollars, by explicitly taking into account that each one of these is below Z dollars. This model is estimated on the subsample B data and can then be used to compute the expected values of the sum of expenditures below Z dollars made by households in subsample A. We apply this and the other two methods to a Dutch panel where the above-mentioned situation actually occurred. The sample consists of two subsamples; for one subsample all expenditures are registered, but for the other subsample only expenditures above 10 Dutch guilders.

1. INTRODUCTION

In consumer expenditure surveys one often faces the problem that information on all consumption expenditures is not available. Especially collecting complete information on small expenditures is a difficult task. In the 1980–1981 Consumer Expenditure Survey of the Netherlands, conducted by the Netherlands Central Bureau of Statistics, for example, information on small expenditures is only gathered during a so-called registration month, once a year. Afterwards, the values of annual expenditures are obtained by inflating the monthly figures. In another Dutch panel, the so-called Expenditure Index, conducted by a private marketing research agency (INTOMART), the sample is divided into two subsamples. The respondents in the first subsample, for example 'A', are only asked to give information on large expenditures, defined as expenditures in excess of 10 Dutch guilders. The other subsample, for example 'B', is asked to give information on small expenditures also (expenditures less than or equal to 10 Dutch guilders). The consequence of this procedure is that we do not know the values of the small expenditures of the households in subsample A. Table I clearly shows that using expenditures only in excess of 10 Dutch guilders to determine

0883–7252/90/020151–16$08.00

Received October 1987
Revised May 1989

Table I. Mean level of the sum of expenditures below 10 Dutch guilders and of expenditures in excess of 10 Dutch guilders by households in subsample B in April 1984

Consumption category	EXP > 10*	EXP ⩽ 10	TOTEXP	SHARE
1 *Food*	453	47	500	0·09
2 *Clothing and footwear*	216	4	220	0·02
3 *Housing*, including rents and interest payments on and redemptions of mortgage payments	630	0	630	0·00
4 *Domestic decoration*, including furniture, expenditures on do-it-yourself articles, and on gardening	404	10	414	0·02
5 *Recreation, entertainment*, including holiday expenditures	189	11	200	0·06
6 *Vehicles*, including purchases of cars, bicycles, etc	48	1	49	0·02
7 *Transportation*, including expenditures on fuel, and public transportation	96	2	98	0·02
8 *Insurance*	140	0	140	0·00
9 *Appliances*, including electric appliances, such as hi-fi equipment, washing machines, and other personal expenditures	150	11	161	0·07
10 *Other expenditures*, including medical expenditures, gifts and donations	162	7	169	0·04
Total expenditure	2487	136	2623	0·05

*EXP > 10 = Expenditures (EXP) in excess of Dfl 10.
EXP ⩽ 10 = Expenditures below (or equal to) Dfl 10.
TOTEXP = Total expenditures.
SHARE = Ratio of expenditures below Dfl. 10 over total expenditures.
In April 1984 the Dutch guilder/dollar exchange rate was 2·97, and the Dutch guilder/pound sterling exchange rate was 4·27.

the total values of the expenditures would lead to considerable underestimation for several expenditure categories in case of subsample B. It is evident that such underestimation would also occur from subsample A.

To correct for this underestimation one could, for instance, make use of simple blow-up factors, calculated on the basis of the B subsample. One could also formulate a Tobit model, where the latent variable corresponding to the sum of the small expenditure, Y^*, depends on some household characteristics. The actual sum of small expenditures, Y, then equals Y^* if $Y^* > 0$; otherwise, $Y = 0$. Once such a model has been estimated on the basis of the data of the B subsample, one could use the model to calculate the expected values of the sum of small expenditures for each household in the A subsample.

In this paper we propose and describe an alternative approach to correct for the underestimation by taking into account that in the sum of small expenditures each single expenditure lies in the interval (0;10]. The model consists of two parts. The first part, which we will call the 'Count Model', describes on a household level the number of expenditures, N, below 10 Dutch guilders by using a probability distribution defined on the non-negative integers. The second part, which we will call the 'Amount Model', explains the amount of the expenditures below Dfl 10 by using a probability distribution (conditional upon N, the number of small expenditures) with each single expenditure defined on the (0;10] interval. Household characteristics are included by parameterization of the parameters of the particular probability distribution chosen. We will refer to this model as a 'CA (Count–Amount) model'. This model bears some resemblance to that of Robin (1987). A CA model can be chosen quite general. However, we will restrict ourselves to the case of stochastic independence between the Count and the Amount model. Some other simplifications will also be made.

We will estimate and compare these three models to correct for the underestimation. A comparison of the performance of the three methods is conducted by splitting the B subsample randomly into two further subsamples: one part containing approximately 90 per cent of the observations, the other part the remainder. On the basis of the 90 per cent subsample we re-estimated the three models. These estimation results are then used to correct for the underestimation in the remaining 10 per cent subsample by predicting the sum of small expenditures. These comparisons indicate that correction based on Tobit, as well as on the chosen version of the CA method, do quite well; at least, when compared to simply blowing up the data. The comparison between Tobit and the CA approach, on the basis of the prediction measures we used, shows that Tobit performs slightly better in the particular subdivision of subsample B we considered. However, if we use the actual number of small expenditures in the comparison between the CA and the Tobit-correction method, the correction based on the CA model performs much better. This suggests that the CA model may be improved by incorporating more information about the number of small expenditures.

In Section 2 we present the three correction methods and discuss some estimation strategies. In Section 3 we pay some attention to the data of the expenditure index and we present the empirical application of the models. In Section 4 we compare the performance of the three correction methods. Section 5 concludes.

2. THE MODEL

In this section we will first present the CA correction method and the version of it we are going to use. Next, we will discuss some estimation methods. Finally, we describe the blow-up method and the correction by application of a Tobit model.

Let us begin with the CA correction method. Each consumption category will be considered separately. As a consequence, we will not use indices to refer to a particular good. We first consider a particular househould h in a particular period t. Let $N(=N_{ht})$ be a random variable defined on the non-negative integers $\{0, 1, 2, \ldots\}$. Let Y_i $(=Y_{i,ht})$, $i = 1, 2, 3, \ldots$, be random variables defined on $(0;10]$. Assume $(N, Y_1, Y_2, \ldots)$ to be defined on $\{0, 1, 2, \ldots\} \times (0;10]^{\{1,2,\ldots\}}$. Then we define $Y(=Y_{ht})$, the sum of the expenditures below 10 Dutch guilders of a particular consumption good by household h in period t, by[1]

$$Y = \sum_{i=1}^{N} Y_i. \tag{1}$$

In this formula one should interpret N as a variable representing the number of small expenditures, and Y_i, $i \leqslant N$, as the amount of the ith small expenditure. The modelling of the distribution of N, $\Pr\{N=n\}$, will be referred to as the Count model; that of the distribution of $(Y_1, Y_2, \ldots)$ given N, $\Pr\{(Y_1, Y_2, \ldots) \in B \mid N\}$, as the Amount model. The combination of these two as in (1) leads to a CA (Count–Amount) model. We will restrict ourselves to the case where N, Y_1, $Y_2, \ldots$ are stochastically independent of one another. Moreover, we will assume that Y_i has the same distribution as Y_j, $i \neq j$, $i, j \in \{1, 2, \ldots\}$. As a consequence, in the Amount model we need to consider only the modelling of one Y_i, for which we take Y_1.

[1] Cf., for example, Feller (1968), ch. XII, dealing with compound distributions.

We work with some particular distributions which we will describe now. We begin with the Count model. As a starting point we first consider the Poisson (λ) model, i.e.

$$\Pr\{N=n\} = e^{-\lambda}\,\frac{\lambda^n}{n!}.$$

The parameter λ ($=\lambda_{ht}$) is specified as follows

$$\lambda = \exp(X'\beta), \tag{2}$$

where

$$X\ (=X_{ht}) = \text{vector of exogenous variables,}$$
$$\beta\ (=\beta_t) = \text{vector of parameters.}$$

The Poisson distribution is restrictive in several ways (see Cameron and Trivedi (1986), Gourieroux, Monfort, and Trognon (1984b)). For instance, the assumption, that the conditional mean and variance of N given X are equal, may be too strong. One way to relax this restriction is to allow for unobserved heterogeneity in λ by replacing (2) by the following equation

$$\lambda = \exp(X'\beta)\exp(\varepsilon), \tag{3}$$

where ε $(=\varepsilon_{ht})$ is a random variable representing unobserved heterogeneity, with $E(\exp \varepsilon \mid X) = 1$. Equation (3) implies that N given X and ε is Poisson (λ) distributed. Since ε is an unobservable random variable we must integrate it out to obtain the conditional distribution of N given X. Cameron and Trivedi (1986), among others, show that if $\lambda \sim$ gamma (ϕ, ν) (where $(\phi, \nu)' = (\phi_{ht}, \nu_{ht})'$), with $\phi = \exp(X'\beta)$, then $N \mid X \sim$ negative binomial (ϕ, ν), i.e.

$$\Pr\{N=n \mid X\} = \frac{\Gamma(n+\nu)}{\Gamma(n+1)\Gamma(\nu)}\left(\frac{\nu}{\nu+\phi}\right)^{\nu}\left(\frac{\phi}{\nu+\phi}\right)^{n}. \tag{4}$$

Notice that $E(N \mid X) = \phi = \exp(X'\beta)$ and $\operatorname{var}(N \mid X) = \phi + \phi^2/\nu$. The negative binomial distribution will be abbreviated as NEGBIN distribution. Concerning the parametrization of ν, we consider three possibilities:

(a) $\nu = \alpha^{-1}\phi = \alpha^{-1}\exp(X'\beta) \qquad \phi = \exp(X'\beta)$ (5)

(b) $\nu = \alpha^{-1}\phi^2 = \alpha^{-1}(\exp(X'\beta))^2 \qquad \phi = \exp(X'\beta)$ (6)

(c) $\nu = \exp(\gamma_0 + X'\gamma) \qquad \phi = \exp(X'\beta)$ (7)

With regard to (5) and (6), the parameterization of the NEGBIN distribution coincides with the NEGBINI and NEGBINII parameterization of Cameron and Trivedi (1986), respectively. The NEGBINI and NEGBINII distribution approach the Poisson distribution if $\alpha \downarrow 0$. In the sequel the NEGBIN distribution with ν parameterized as in (7) will be called the NEGBINA distribution. Notice that the NEGBINI and NEGBINII distributions are special cases of the NEGBINA distribution.[2]

Next, we turn to the Amount model. Y_1 is assumed to follow the beta-distribution with probability density function $f(\cdot)$ $(=f_{ht}(\cdot))$ given by:

$$f(v) = \frac{1}{10}\,\frac{1}{B(p,q)}\left(\frac{v}{10}\right)^{p-1}\left(1-\frac{v}{10}\right)^{q-1} I_{(0;10]}(v), \tag{8}$$

[2] To get from the NEGBINA distribution the NEGBINI distribution, just define $1/\alpha = \exp(\gamma_0)$ and set $\gamma = \beta$; to get the NEGBINII distribution also define $1/\alpha = \exp(\gamma_0)$ and set $\gamma = 2\beta$.

with $p > 0$, $q > 0$, and $I_{(0;10]}(v)$ an indicator function defined by

$$I_{(0;10]}(v) = 1 \text{ if } 0 < v \leqslant 10, \\ = 0 \text{ otherwise.} \tag{9}$$

We further parameterize the parameters p $(= p_{ht})$ and q $(= q_{ht})$ as follows:

$$p = \exp(Z'\phi_1), \\ q = \exp(Z'\phi_2). \tag{10}$$

where:

Z $(= Z_{ht})$ = vector of exogenous variables,
$\phi_1(=\phi_{1t})$ = vector of parameters,
$\phi_2(=\phi_{2t})$ = vector of parameters.

Given the above modelling, we can derive for each consumption category a log-likelihood function of the observations. We assume independence across households and across time. So, for a particular consumption category, the log-likelihood function is

$$\log L = \sum_{t\in T}\sum_{h\in H} \log P\{N_{ht} = n_{ht}\} + \sum_{t\in T}\sum_{h\in H}\sum_{i\in\{1,\ldots,n_{ht}\}} \log f_{ht}(y_{iht})\cdot I_{\{1,2,\ldots\}}(n_{ht}), \tag{11}$$

H stands for the set of households, T is the set of periods, n_{ht} is the number of small expenditures by household h in period t, and y_{iht} is the (positive) amount of the ith small expenditure by household h in period t, $i = 1, \ldots, n_{ht}$. As long as the sets of parameters of the first and of the second part of the right-hand side of this formula are disjoint (which we will assume), maximizing $\log L$ as a function of the parameters can be achieved by maximizing each part separately.

Concerning the Count model, two estimation methods are considered:

1. maximum likelihood (ML)
2. pseudo-maximum likelihood (PML)

(see Gourieroux, Monfort, and Trognon, 1984a,b).

The PML estimators are obtained by maximizing a likelihood function associated with some family of probability distributions, which does not necessarily contain the true distribution. Gourieroux, Monfort, and Trognon (1984a) show that under certain assumptions the PML method will, in the case of a linear exponential family, yield consistent estimators of the parameters appearing in the first-order moment of the true distribution. Since, in that case, the PML method only assumes a correctly specified mean, the ML method will give more efficient estimates than the PML method, if the distribution of $N|X$ is correctly specified. However, the PML method is likely to be more robust against misspecification of the distribution. In the empirical application we shall consider the PML estimators associated with the Poisson family. A consistent estimate of the variance–covariance matrix of the PML estimates is calculated by using the results of Gourieroux, Monfort, and Trognon (1984a). See also Cameron and Trivedi (1986, p. 37). Note that PML estimates can also be derived from the NEGBINA distribution with the parameter vector γ being given (see Gourieroux, Monfort, and Trognon, 1984b).

Once the model has been estimated with data from subsample B we can use the model to predict the total amount of small expenditures by households in subsample A. This prediction can be carried out by taking the expectation of the total value of the small expenditures Y_{ht}, where $Y_{ht}=\Sigma_{j\in 1}^{N} Y_{j,ht}$. We have (with $Y=Y_{ht}$, $Y_{j,ht}=Y_j$, and $N=N_{ht}$)

$$\begin{aligned} E(Y) &= E(E(Y\mid N)) = E\left(E\left(\sum_{j=1}^{N} Y_j \mid N\right)\right) = E\left(\sum_{j=1}^{N} E(Y_j\mid N)\right) \\ &= E\left(\sum_{j=1}^{N} E(Y_j)\right) = E\left(\sum_{j=1}^{N} E(Y_1)\right) = E(Y_1)E\left(\sum_{j=1}^{N} 1\right) = E(\mathrm{N})\cdot E(Y_1). \end{aligned} \tag{12}$$

The prediction of Y_{ht} based on this method is thus given by:

$$\hat{Y}_{ht}^{CA} = E(N_{ht})\cdot E(Y_{1,ht}). \tag{13}$$

As an alternatives to this model, we consider two other approaches to correct for the underestimation. The first one is simply the use of blow-up factors. By dividing the sample into, say, D disjoint cells according to some classification rule, a blow-up factor for each cell $d=1,\dots,D$ in subsample B (and for each good) can be calculated as follows:

$$BU_d = \left(\sum_{h\in H_d}\sum_{t\in T} K_t(YT_{ht})\right)\Big/\left(\sum_{h\in H_d}\sum_{t\in T} K_t(YA_{ht})\right), \tag{14}$$

with $K_t(x_h)=x_h$, if household h is in the sample in period t,
$=0$, otherwise,

and where:

$BU_d =$ the blow-up factor of subgroup d,
$YT_{ht} =$ the total sum of expenditures of household h,
$YA_{ht} =$ the sum of expenditures of household h above Dfl 10,
$H_d =$ the set of households belonging to subsample d.

For household h' in subsample A we can blow up its expenditures above Dfl 10 by the factor BU_d, if this household belongs to cell d, to obtain an estimate for Y_{ht}. Let $\hat{Y}_{ht}^{BU}$ denote the estimate of Y_{ht} on the basis of the blow-up factor. Then,

$$\hat{Y}_{ht}^{BU} = BU_d\cdot YA_{ht} - YA_{ht}. \tag{15}$$

Thus, blow-up is used to get a prediction of the total sum of expenditures ($BU_d\cdot YA_{ht}$), which is corrected by the total sum of expenditures above 10 Dutch guilders (YA_{ht}) to get an estimate of the sum of small expenditures ($\hat{Y}_{ht}^{BU}$).

The second alternative we consider is an application of the Tobit model. Let Y^* ($={Y^*}_{ht}$) denote a latent variable generated by

$$Y^* = V\eta + \zeta, \tag{16}$$

with: V ($=V_{ht}$) = vector of exogenous variables,
η ($=\eta_t$) = vector of parameters,
ζ ($=\zeta_{ht}$) = random error term, normally distributed with zero expectation and variance σ_ζ^2.

Then Y ($=Y_{ht}$) is modelled by

$$\begin{aligned} Y &= Y^*, \quad \text{if } Y^*>0, \\ Y &= 0, \quad\ \ \text{if } Y^*\leqslant 0. \end{aligned} \tag{17}$$

A prediction of Y_{ht} can now be obtained by taking the expectation of Y_{ht} with respect to the current distribution:

$$\hat{Y}^T_{ht} = V'_{ht}\eta_t + \phi(-V'_{ht}\eta_t/\sigma_\xi)/(1 - \Phi(-V'_{ht}\eta_t/\sigma_\xi)), \tag{18}$$

with $\phi(\cdot)$ the standard normal density function and $\Phi(\cdot)$ the standard normal distribution function.

3. DATA AND ESTIMATION RESULTS

In this section we first shortly describe the data we are going to use. Then we present the model selection and the resulting paramater estimates of the various models.

Data

The data stem from a panel survey conducted in the Netherlands between April 1984 and December 1986. In this study we only use data of the period April 1984–December 1984. Information about expenditures on different categories is collected on a monthly basis, while data on background variables such as net household income and family size are gathered once a year. The data set consists of about 800 households per month, of which approximately 10 per cent is in the B subsample. Not all households are registered each month, so the total number of different households participants in the sample during at least one period is about 1500.

From a first inspection of the data of subsample B (see Table I), it appears that expenditures below 10 guilders occur most frequently in the following consumption categories:

1. Food
2. Clothing and footwear
3. Domestic decoration
4. Recreation, entertainment
5. Appliances
6. Other expenditures

In the case of the other consumption categories (Housing, Vehicles, Transportation, and Insurance), such expenditures are rarely found in the B subsample. Therefore, we do not estimate the models for these consumption categories.

The following regressors appear in all the various models,

1. C = Constant term
2. URB = Degree of Urbanization
   1 = rural municipality
   2 = commuter towns and small towns
   3 = medium sized cities (30,000–100,000 inhabitants)
   4 = large cities (over 100,000 inhabitants)
3. CHILD = 1 if household contains children younger than 5 years of age
   = 0 otherwise
4. FS = Family size
5. WORKH = 1 if head of household does not have a paid job
   = 0 otherwise
6. AGEP = Age of the spouse (age of the head of the household if there is no spouse.)
7. SOC = Social group
   1 = upper class

2 = upper middle class
3 = middle class
4 = lower middle class
5 = lower class

8. SINGLE = 1 if household consists of a single person
= 0 otherwise

Table II gives sample summary statistics of these variables. In fact, we use the same exogenous variables in the various models. Since these variables do not change over the sample period we thus have

$$X_{ht} = Z_{ht} = V_{ht} = X_h = Z_h = V_h. \tag{19}$$

Table II. Summary statistics of the exogenous variables (B subsample, number of observations: 688)

	Mean	St. Dev.
URB	2·38	1·06
CHILD	0·22	0·42
FS	2·92	1·39
WORKH	0·37	0·48
AGEP	42·11	15·04
SOC	3·24	1·01
SINGLE	0·15	0·36

Model Selection—The CA Model

Now we turn to the selection of the models. We begin with the Count part of the CA model. The NEGBIN models (see formulae (4)–(7)) are chosen as a starting point of analysis. To begin with, we assume that the parameter vectors of the NEGBIN models do not change over time (the sample period). The resulting ML and PML estimates of the NEGBINI, NEGBINII and NEGBINA models are presented in Table III.

In case of the NEGBINA model, only the estimates of the parameters appearing in the first-order moment of the negative binomial distribution are presented. In order to calculate the expected value of the small expenditures (cf. formula (14)), we need only these parameters. A comparison of the ML and PML estimates of the parameters, which appear in the first-order moment of the NEGBINA model, shows that these are similar in sign and magnitude. This observation applies especially to the significant parameters. This is an encouraging result, because both the ML and PML estimates of these parameters are consistent if the distribution is correctly specified. The ML and PML estimates could, however, differ considerably if the first-order moment of the negative binomial distribution is not correctly specified. In section 2 we noted that NEGBINI and NEGBINII models follow from the NEGBINA model by imposing suitable restrictions. By means of a likelihood test, we check whether we may impose such restrictions. This test statistic, which under the null hypothesis is asymptotically distributed as $\chi^2(7)$, is (at the 5 per cent level) significant for both the NEGBINI and NEGBINII model in case of Food, Recreation and Other expenditures (cf. row 'LR' in Table III). For these consumption categories we select the NEGBINA model, and for the other categories we choose the NEGBINI model, because from the values of the Akaike information criterion (cf. row AIC or Table III) we can conclude that, in terms of this criterion, the NEGBINI model shows a better performance than the NEGBINII model for Clothing, Footwear and a similar performance for Domestic decoration and Appliances.

The age of the spouse, AGEP, has a significant (at the 5 per cent level) positive influence on the expected number of expenditures below Dfl 10 (see equation (6)), except for Recreation, Entertainment and Food. In the Netherlands, partners of lower age more frequently have a paid job and, consequently, have less leisure than older partners. Therefore, our intuition is that most households with partners of low age spend relatively little time on shopping, but buy these goods in large quantities in supermarkets and department stores. The signs of the parameters, corresponding to AGEP, conform to this intuition.

The expected number of expenditures below Dfl 10 on Recreation, Entertainment decreases significantly if small children are present in the household, and increases if the household consists of a single person. The last result would conform to a lifestyle of singles, especially those of young age, who spend a fair amount of time away from the home. Social class (SOC)—a variable which is strongly correlated with income—plays a significant role in explaining the number of expenditures below Dfl 10 for Clothing, Footwear and Domestic decoration. The higher the social class, the higher the expected number of small expenditures will be.

The degree of urbanization (URB) has a significant negative influence, and family size (FS) a significant positive influence on the expected number of expenditures below Dfl 10 for all consumption categories—except for Clothing and Footwear in the case of URB and Domestic decoration in the case of FS.

Finally, the results suggest that the dummy variable WORKH, which indicates whether the head of the household works, has no explanatory power for the number of expenditures below Dfl 10 for any category, except for Other expenditures.

For the selected models we have carried out a likelihood ratio test of the hypothesis, that the parameters corresponding to all explanatory variables except the constant term are equal to zero. The test statistic is asymptotically distributed as χ^2 (14) for Food, Recreaction, and Other expenditures and χ^2 (7) for the other consumption categories. In all cases we must reject the null hypothesis (see row LC in Table III).

In the preceding section it was noticed that the Poisson model is nested in the NEGBIN model. A likelihood ratio test indicates that the Poisson model must be rejected for all consumption categories (see row LRP in Table III). This is not a surprising result, because the Poisson model is very restrictive in several ways.

Next, we have tested, by means of a likelihood ratio test, whether the parameter vector is varying with time (months). The results, summarized in Table IV, suggest that the same model may be valid in all months.

In Section 2 we have assumed that the amount of the ith expenditure Y_i follows a beta $((0,10],p,q)$ distribution (see equation (9)), where p and q are parameterized as

$$p_{ht} = \exp(X'_h\phi_{1t}); \quad q_{ht} = \exp(X'_h\phi_{2t}).$$

(Notice that we put $Z_{ht} = X_h$.) The 144 parameters for each consumption category are estimated by means of maximum likelihood.

We tested the hypothesis that the parameters corresponding to the non-constant terms are all equal to zero by means of a likelihood-ratio test. Under this hypothesis the likelihood-ratio test is asymptotically χ^2_{126} distributed. The results are presented in Table V. Clearly, the hypothesis is rejected (at the 5 per cent level) without any exception. Next, we tested the hypothesis that the parameters do not vary over the sample period, again by means of a

Table III. ML and PML estimates of β for the NEGBIN models (asymptotic t-values in parentheses)

	Food				Clothing, Footwear				Appliances			
	ML NEGBINI	ML NEGBINII	ML NEGBINA	PML	ML NEGBINI	ML NEGBINII	ML NEGBINA	PML	ML NEGBINI	ML NEGBINII	ML NEGBINA	PML
Constant	2·234	2·115	2·926	2·810	−0·710	−0·555	−0·712	−0·896	0·236	0·284	0·571	0·489
	(9·68)	(13·02)	(6·252)	(10·87)	(−1·68)	(−1·69)	(−1·41)	(−1·91)	(0·768)	(1·26)	(1·48)	(1·27)
URB	−0·145	−0·108	−0·136	−0·104	−0·121	−0·109	−0·093	−0·081	−0·148	−0·118	−0·132	−0·119
	(−3·86)	(−4·43)	(−2·05)	(−2·96)	(−1·83)	(−2·15)	(−1·13)	(−1·17)	(−3·29)	(−3·50)	(−2·52)	(−2·13)
Child	−0·104	−0·055	−0·310	−0·234	−0·075	−0·164	−0·002	0·043	−0·183	−0·103	−0·342	0·335
	(−1·05)	(−0·87)	(−1·46)	(−2·15)	(−0·44)	(−1·13)	(−0·007)	(0·22)	(−1·345)	(−1·08)	(−2·08)	(−2·44)
FS	0·1145	0·089	0·073	0·082	0·218	0·176	0·239	0·229	0·205	0·174	0·176	0·172
	(3·22)	(4·02)	(0·99)	(2·36)	(3·63)	(3·65)	(3·29)	(4·18)	(4·56)	(5·36)	(2·88)	(3·87)
WORKH	0·020	−0·010	0·0032	0·069	−0·086	−0·052	−0·002	−0·096	−0·320	−0·010	−0·065	−0·049
	(0·24)	(−0·17)	(0·02)	(0·69)	(−0·54)	(−0·425)	(−0·01)	(−0·60)	(−0·29)	(−0·12)	(−0·42)	(−0·43)
AGEP	0·0088	0·008	−0·002	0·0003	0·012	0·009	0·007	0·012	0·010	0·008	0·007	0·007
	(3·35)	(4·56)	(−0·290)	(0·009)	(2·11)	(2·235)	(0·86)	(2·27)	(2·57)	(2·93)	(1·19)	(1·63)
SOC	−0·075	−0·030	−0·084	−0·115	−0·171	−0·146	−0·139	−0·146	−0·045	−0·034	−0·072	−0·049
	(−2·40)	(−1·31)	(−1·70)	(−2·55)	(−2·98)	(−3·25)	(−2·07)	(−2·26)	(−1·073)	(−1·14)	(−0·102)	(−1·32)
SINGLE	0·182	0·168	0·052	0·042	0·034	0·066	−0·0167	−0·003	0·075	0·151	−0·048	−0·125
	(1·61)	(2·26)	(0·29)	(0·33)	(0·16)	(0·42)	(−0·065)	(−1·51)	(0·59)	(1·42)	(−0·250)	(−0·79)
α	19·448	0·169	—	—	1·070	0·00076	—	—	2·72	0·006	—	—
	(11·12)	(7·34)	—	—	(6·41)	(5·30)	—	—	(10·69)	(7·81)	—	—
Log L	15,446·19	15,451·11	15,458·70	—	−593·94	−594·77	−589·24	—	119·49	119·45	125·2	—
LR	25·02	15·18	—	—	9·40	11·06	—	—	11·42	11·50	—	—
LRP	6121·64	6131·48	6146·66	—	184·58	182·92	193·88	—	848·68	848·60	860·1	—
LC	—	—	71·94	—	49·44	—	—	—	60·92	—	—	—
AIC	−30,874·38	−30,884·22	−30,885·4	—	1205·88	1207·54	1210·48	—	−220·98	−220·90	−218·9	—

Constant	0·559	0·623	0·467	0·397	0·587	0·514	0·879	0·754	1·363	1·197	1·191	1·091
	(1·46)	(2·38)	(0·80)	(1·23)	(1·94)	(2·201)	(2·11)	(2·50)	(4·25)	(5·49)	(2·98)	(2·36)
URB	−0·499	−0·316	−0·541	−0·559	−0·199	−0·152	−0·198	−0·206	−0·290	−0·194	−0·294	−0·232
	(−9·50)	(−7·38)	(−8·41)	(−8·07)	(−4·30)	(−4·23)	(−3·69)	(−4·52)	(−5·60)	(−5·47)	(−4·52)	(−2·75)
Child	−0·104	−0·002	−0·244	−0·266	0·131	0·135	0·069	0·060	0·424	−0·258	−0·492	−0·598
	(−0·656)	(−1·98)	(−1·34)	(−1·50)	(0·94)	(1·26)	(0·38)	(0·42)	(−3·01)	(−2·62)	(−2·61)	(−3·28)
FS	0·133	0·083	0·183	0·167	0·030	0·027	0·033	0·017	0·257	0·158	0·327	0·343
	(2·551)	(2·07)	(2·23)	(3·53)	(0·65)	(0·74)	(0·53)	(0·38)	(5·17)	(4·72)	(5·21)	(6·06)
WORKH	0·222	0·112	0·356	0·353	0·042	0·010	0·229	0·094	−0·125	−0·203	0·187	0·268
	(1·699)	(1·16)	(1·98)	(2·78)	(0·37)	(0·11)	(1·59)	(0·80)	(−1·07)	(−2·36)	(1·33)	(1·43)
AGEP	0·015	0·011	0·011	0·015	0·015	0·013	0·008	0·012	−0·0025	0·0015	−0·013	−0·012
	(3·321)	(0·326)	(1·59)	(3·47)	(3·80)	(4·37)	(1·26)	(3·73)	(−0·68)	(0·60)	(−2·79)	(−2·65)
SOC	−0·120	−0·117	−0·073	−0·081	−0·185	−0·152	−0·181	−0·174	(−0·188)	−0·134	−0·076	−0·111
	(−2·69)	(−3·35)	(−1·23)	(−1·65)	(−4·39)	(−4·64)	(−3·27)	(−4·54)	(−4·05)	(−4·29)	(−1·44)	(−1·84)
SINGLE	0·159	0·153	0·119	0·071	0·061	0·102	0·022	−0·049	0·922	0·629	−0·928	0·818
	(1·019)	(1·29)	(0·55)	(0·42)	(0·43)	(0·94)	(0·124)	(−0·36)	(5·66)	(5·67)	(4·50)	(4·34)
α	1·966	0·003	—	—	1·543	0·00023	—	—	4·754	0·012	—	—
	(7·794)	(6·61)	—	—	(7·56)	(6·06)	—	—	11·85	(8·26)	—	—
Log L	−305·80	−319·60	−296·64	—	−458·73	−458·49	−454·57	—	757·45	755·55	776·34	—
LR	18·32	45·92	—	—	8·32	7·84	—	—	37·78	41·58	—	—
LRP	447·84	420·29	466·16	—	329·22	329·7	337·54	—	1647·60	1643·80	1685·38	—
LC	—	—	146·98	—	63·0	—	—	—	—	—	143·42	—
AIC	629·60	657·20	625·28	—	935·46	934·88	941·14	—	−1496·90	−1493·10	−1520·68	—

Log L = Log-likelihood (with some constant terms disregarded).
LR = Likelihood ratio test of the hypothesis, that the NEGBINA distribution reduces to the NEGBINI or the NEGBINII distribution (depending on the model selection).
LRP = Likelihood ratio test of the hypothesis, that the number of small purchases is Poisson distributed.
LC = Likelihood ratio test of the hypothesis, that the parameters corresponding to all explanatory variables except the constant term are equal to zero.
AIC = Akaike's information criterion.

Table IV. Likelihood ratio test of the hypothesis that the parameters of the NEGBIN model do not vary with time (months)

1. Food	76·38
2. Clothing and footwear	59·98
3. Domestic decoration	54·24
4. Recreation and entertainment	118·12
5. Other expenditures	101·10
6. Appliances	47·26

The test statistic is asymptotically χ^2_{72} distributed* in the case of the categories Food, Recreation and Other expenditures, and asympotically χ^2_{128} distributed† for the other categories

*NEGBIN has $8+1$ parameters per period (8 regressors and α), so the degress of freedom become $9\times 9-9\times 1=72$.

†NEGBINA has 16 parameters per period, so the degrees of freedom become $16\times 9-16\times 1=128$.

Table V. Test of the null hypothesis that the parameters belonging to the non-constant exogenous variables are equal to zero

1. Food	354·1
2. Clothing and footwear	205·8
3. Domestic decoration	168·0
4. Recreation	320·8
5. Other expenditures	223·6
6. Appliances	263·3

The likelihood-ratio test statistic is asymptotically χ^2_{126} distributed

Table VI. Test of the null-hypothesis that the parameters ϕ_{1t} and ϕ_{2t} are constant over time

1. Food	196·4
2. Clothing and footwear	221·6
3. Domestic decoration	174·4
4. Recreation	261·3
5. Other expenditures	195·0
6. Appliances	201·5

The likelihood-ratio test statistic is asymptotically χ^2_{128} distributed

likelihood-ratio test. Now, this test statistic is asymptotically χ^2_{128} distributed. The result are given in Table VI. Again, we reject the hypothesis in all cases on the basis of these results. Consequently, we choose as the Amount model the one with ϕ_1 and ϕ_2 non-constant over time and varying with the exogenous variables in X_h.

Blow Up

In the case of blow-up—cf. (14) and (15)—we made use of the following classification rule of the B subsample.[3]

$d=1$ if URB = 1 or 2 and SOC = 1 or 2,
$d=2$ if URB = 1 or 2 and SOC = 3 or 4 or 5,
$d=3$ if URB = 3 or 4 and SOC = 1 or 2,
$d=4$ if URB = 3 or 4 and SOC = 3 or 4 or 5.

We calculated the blow-up factors for the B-subsample (see Table VII).

[3] The choice of this classification is based on the good performance of the variables URB and SOC in the Tobit model (and CA model).

Table VII. Blow-up factors, subsample B

	Food	Clothing and footwear	Domestic decoration	Recreation	Other expenditure	Appliances
BU_1	1·12	1·02	1·08	1·03	1·02	1·05
BU_2	1·12	1·02	1·07	1·06	1·05	1·07
BU_3	1·23	1·01	1·02	1·07	1·01	1·06
BU_4	1·11	1·02	1·03	1·04	1·03	1·07

Tobit

We estimated Tobit by making use of the same exogenous variables as in the case of the CA model. First, we estimated the parameters under the hypothesis that these are constant over the sample period. The estimation results are given in Table VIII. We observe from these estimates that the effects we found in case of the Count model are quite similar to the effects we find here. The influence of the variable AGEP on the sum of small expenditures is now somewhat more often significantly positive. The effects of the variables CHILD and SINGLE are comparable to those found in case of the Count model. A similar observation holds true for the variables SOC, URB, and FS. Also, the variables WORKH again has hardly any explanatory power, with the exception being Recreation instead of Other expenditures. The interpretation in the case of the *number* of small expenditures also seems valid for the *sum* of small expenditures.

We tested the hypothesis that the parameters of the Tobit are constant over the sample period by means of a likelihood-ratio test, which, in this case, is asymptotically distributed as χ^2_{72} (see Table IX). The results prevent us from rejecting the hypothesis in all cases. So, as Tobit model, we take the one with constant parameters over time.

Table VIII. ML estimates of η for the Tobit models (asymptotic *t*-values in parentheses)

	Food	Clothing and footwear	Domestic decoration	Recreation	Appliances	Other expenditure
Constant	50·25	−9·073	6·564	5·176	1·392	−6·887
	(32·62)	(−2·05)	(1·44)	(6·83)	(0·28)	(−1·30)
URB	−7·034	−0·958	−2·897	−3·899	−5·854	−1·961
	(−2·72)	(−1·35)	(−4·22)	(−3·42)	(−8·98)	(−2·54)
Child	−9·337	−1·336	1·638	−11·551	−1·030	−2·558
	(−1·34)	(−0·67)	(0·76)	(−3·53)	(−0·50)	(−1·05)
FS	6·990	2·567	0·358	6·113	1·863	3·510
	(2·82)	(3·42)	(0·48)	(5·56)	(2·59)	(4·05)
WORKH	−0·152	−1·003	−0·209	−2·820	4·018	−1·141
	(−0·03)	(−0·58)	(−0·12)	(−1·06)	(2·20)	(−0·56)
AGEP	0·459	0·129	0·247	−0·078	0·225	0·219
	(2·61)	(2·19)	(3·93)	(−0·95)	(3·97)	(3·23)
SOC	−7·309	−1·897	−2·836	−2·771	−1·709	−0·255
	(−3·76)	(−2·76)	(−4·06)	(−2·36)	(−2·46)	(−0·33)
SINGLE	5·579	0·522	0·632	16·687	0·242	−1·438
	(0·75)	(0·23)	(0·29)	(4·42)	(0·11)	(−0·54)
σ	59·26	14·80	16·72	25·97	15·12	18·87
Log L·	−3589·98	−1417·25	−2094·20	−2189·02	−1806·61	−2331·96

Log L = log-likelihood.

Table IX. Test of the null hypothesis that the parameter vector η_t is constant over time

1. Food	56·3
2. Clothing and footwear	64·6
3. Domestic decoration	68·5
4. Recreation	92·8
5. Other expenditures	60·3
6. Appliances	39·4

The likelihood-ratio test statistic is asymptotically χ^2_{72} distributed

4. COMPARISON OF THE THREE METHODS

In this section, we compare the performances of the three models, whose estimation results we presented in the previous section. The comparison is conducted by randomly splitting up the B subsample into two further subsamples, one containing approximately 90 per cent of the observations and the other one containing the remaining observations. On the basis of the 90 per cent subsample, we re-estimated the models chosen in the previous section. Next, we used the obtained estimation results to calculate for each observation, in the 10 per cent subsample, the estimation of the sum of the small expenditures by applying formulae (13), (15), and (18). Let H_{10} denote the set of households in the 10 per cent subsample. Then, the estimation results (for each consumption good) are compared on the basis of the mean square error (MSE) and the mean absolute deviation (MAD):

$$\text{MSE} = (1/N) \sum_{h \in H_{10}} \sum_{t \in T} (K_t(Y_{ht}^M) - K_t(Y_{ht}))^2, \qquad \text{M} = \text{CA, BU, T,}$$
$$\text{MAD} = (1/\text{N}) \sum_{h \in H_{10}} \sum_{t \in T} | K_t(Y_{ht}^M) - K_t(Y_{ht}) |, \qquad \text{M} = \text{CA, BU, T.} \tag{20}$$

As in (14) $K_t(x_h) = x_h$ if household h is in the sample in period t, $K_t(x_h) = 0$, otherwise. The results are presented in Table X for the MSE and in Table XI for the MAD. It turns out that Tobit and CA correction method both predict much better than the correction method based

Table X. Values of the MSE for the various goods and the three models CA (Count–Amount), BU (blow-up factors), and T (Tobit)

	Model		
Good	BU	T	CA
1. Food	3845·6	3149·5	3160·3
2. Clothing and footwear	40·7	24·8	24·9
3. Domestic decoration	440·6	116·2	120·5
4. Recreation	530·7	194·2	216·9
5. Other expenditures	181·8	83·8	87·9
6. Appliances	337·2	182·6	188·7

Table XI. Values of the MAD for the various goods and the three models CA (Count–Amount), BU (blow-up factors), and T (Tobit)

	Model		
Good	BU	T	CA
1. Food	47·8	45·2	45·5
2. Clothing and footwear	4·5	3·9	4·0
3. Domestic decoration	11·3	8·3	8·5
4. Recreation	12·9	11·0	11·5
5. Other expenditures	7·8	6·6	6·6
6. Appliances	13·0	10·1	10·2

Table XII. Values of the MSE and MAD for the various goods and the CA model with N_{ht} substituted for $E(N_{ht})$ in the formula of $\hat{Y}_{ht}^{CA}$

Good	MSE	MAD
1. Food	1046·1	14·6
2. Clothing and footwear	4·9	1·0
3. Domestic decoration	12·0	2·0
4. Recreation	47·6	3·8
5. Other expenditures	20·5	2·1
6. Appliances	20·9	2·8

on blow-up factors. A comparison between Tobit and CA correction method indicates that they perform quite similarly, but in all cases Tobit is slighly better than the CA method. An important advantage of the CA model, however, is that more information on the number of small expenditures (which can easily be obtained through additional questions in the questionnaire) can immediately be incorporated into the analysis. In Table XII we present the MSE and MAD when using the observed number of small expenditures instead of its expectation, i.e. instead of $E(N_{ht})$ we use N_{ht} in formula (13). The table clearly suggests that better estimates of the number of small expenditures will strongly improve the performance of the CA model, so that it will outperform the Tobit correction procedure.

5. SUMMARY AND CONCLUSIONS

In this paper we presented a two-stage model, consisting of a Count part and an Amount part (together referred to as a CA model) explaining the sum of small expenditures in one subsample (subsample B) and that can be used to predict the sum of small expenditures in the other subsample (subsample A). Compared to simply blowing up the data or using a Tobit model to correct for the underestimation that occurs once one neglects the small expenditures in the A subsample, a CA method differs mainly in that it takes explicitly into account that in the sum of small expenditures each single expenditure lies in the (0;10] interval. In this paper, we chose a very specific form of a CA model: we assumed the Count model and the Amount model to be stochastically independent, and in the Amount model the single expenditures are also assumed stochastically independent and identically distributed.

Moreover, we chose for the Count model various forms of the negative binomial distribution, and for the single expenditures in the Amount model we chose a beta-distribution.

Despite these restrictions, the chosen CA correction method nevertheless does not perform much worse than the correction method based on the Tobit model. We conjecture that an improvement of the CA method can be achieved by including more suitable explanatory variables, e.g. variables containing more information about the number of small expenditures. On the basis of our results we therefore recommend inclusion of such information in surveys of the kind considered here. It may also be interesting to investigate other probability distributions for the number of small expenditures, like the double hurdle model and with zero (WZ) model proposed by Mullahy (1986). These are topics for future research.

ACKNOWLEDGEMENTS

The authors would like to express their thanks to the anonymous referees, Arie Kapteyn, Marcel Kerkhofs, Theo Nijman, Jean Marc Robin, and Arthur van Soest for their valuable comments, and to Jan-Berno van Lochem and Ger Kock for their excellent research assistance. The authors also wish to thank the Netherlands Organization for Scientific Research (NWO) and Rob Alessie also thanks the NMB bank for financial support.

REFERENCES

Cameron, A., and P. Trivedi (1986), 'Econometric models based on count data', *Journal of Applied Econometrics*, **1**, 29–53.

Feller, W. (1968), *An Introduction to Probability Theory and its Applications*. Volume 1, John Wiley, New York.

Gourieroux, C., A. Monfort, and A. Trognon (1984a), 'Pseudo maximum likelihood methods: theory', *Econometrica*, **52**, 681–700.

Gourieroux, C., A. Monfort, and A. Trognon (1984b), 'Pseudo maximum likelihood methods: applications to Poisson-models', *Econometrica*, **52**, 701–720.

Mullahy, J. (1986), 'Specification and testing of some modified count models', *Journal of Econometrics*, **33**, 341–365.

Robin, J. M. (1987), 'A new model to estimate Engel curves from budget expenditure data', mimeo, ENSAE., Paris.

LINKED CITATIONS

You have printed the following article:

The Problem of not Observing Small Expenditures in a Consumer Expenditure Survey
Rob Alessie; Raymond Gradus; Bertrand Melenberg
Journal of Applied Econometrics, Vol. 5, No. 2. (Apr. - Jun., 1990), pp. 151-166.
Stable URL:
http://links.jstor.org/sici?sici=0883-7252%28199004%2F06%295%3A2%3C151%3ATPONOS%3E2.0.CO%3B2-B

References

Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests
A. Colin Cameron; Pravin K. Trivedi
Journal of Applied Econometrics, Vol. 1, No. 1. (Jan., 1986), pp. 29-53.
Stable URL:
http://links.jstor.org/sici?sici=0883-7252%28198601%291%3A1%3C29%3AEMBOCD%3E2.0.CO%3B2-M

Pseudo Maximum Likelihood Methods: Theory
C. Gourieroux; A. Monfort; A. Trognon
Econometrica, Vol. 52, No. 3. (May, 1984), pp. 681-700.
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198405%2952%3A3%3C681%3APMLMT%3E2.0.CO%3B2-N

Pseudo Maximum Likelihood Methods: Applications to Poisson Models
C. Gourieroux; A. Monfort; A. Trognon
Econometrica, Vol. 52, No. 3. (May, 1984), pp. 701-720.
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198405%2952%3A3%3C701%3APMLMAT%3E2.0.CO%3B2-4