ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 61 – 2001

Infrequency of Purchase: A Model for Clothing Consumption with Panel Data

Juan A. SANCHIS-LLOPIS *

ABSTRACT. – Microeconomic survey data sets offer important advantages for the analysis of consumer demand, although one of the problems associated with these data is the existence of zeros in the expenditure records of households. One of the reasons behind a zero record is infrequency of purchase. This paper focuses on the analysis of Infrequency of Purchase Models (IPMs). We develop an IPM for panel data and compare it with the IPM for cross-sectional data. The estimation of these models using panel data might shed some light on the restrictions that one has to make in order to estimate them using cross-sectional data. Finally, we test for individual heterogeneity and compare the two sets of estimation results. The data are drawn from the *Spanish Family Expenditure Survey* (ECPF) for the period 1985 to 1991.

Infréquence d'achat : un modèle de demande de consommation de vêtements avec des données de panel

RÉSUMÉ. – L'utilisation des enquêtes des données microéconomiques offre d'importants avantages pour l'analyse de la demande de consommation, tout en faisant apparaître un problème associé à l'existence de zéros pour la dépense de certains biens des ménages. On peut justifier l'existence de zéros au fait que certains biens peuvent être soumis *à l'infréquence d'achat*. Cet article présente un cadre d'analyse des modèles d'*infréquence d'achat*. Nous proposons un modèle d' *infréquence d'achat* sur données de panel et nous comparons ce modèle avec des modèles d'*infréquence d'achat* sur données de coupes transversales. Dans l'article, nous utilisons des données de *l'Enquête sur le Budget des Familles Espagnoles* (ECPF), couvrant la période de 1985-1991.

* J.A. SANCHIS-LLOPIS: University College London and Universitat de Valencia.

I would like to thank H. BIERENS, S. HOCHGUERTEL, M.J. LEE, T. MAGNAC, I. PRESTON, M. E. ROCHINA and specially, R. BLUNDELL, J. M. LABEAGA, and two anonymous referees for many valuable comments and suggestions. I would also like to thank some participants in a seminar at *Tilburg University* (1996), in the *XXI Simposium de Análisis Económico in Barcelona* (1996), in the *7th Panel Data Conference in Paris* (1997) and in the *ESEM98 held in Berlin* (1998). This paper is part of the first chapter of my thesis dissertation and finance from the ESRC and the Bank of Spain is gratefully acknowledged. All remaining errors are my own responsibility.

1 Introduction

Microeconomic survey data sets offer important advantages for the analysis of consumer demand. However, one of the major problems associated with microbudget surveys is the existence of zeros in the expenditure records of households. This is so because a household may be observed to spend nothing over the interview period, on a commodity that nevertheless consumes. It is clear that some commodities, *e.g.* clothes, are *"durable"* (or *"semi-durable"*) relative to a typical interview period. Then, in short duration surveys, consumption and purchases can differ markedly and there is a problem in fitting a model of consumption directly to data on purchases. As a result, much of the variation in reported expenditures may reflect stochastic elements in purchasing behaviour rather than differences in the underlying consumption.

In these surveys, zeros might arise for any of the following reasons. First, because the household is maximising utility at zero consumption with its current budget, *i.e.*, for its level of income or/and at current prices the household chooses not to consume (corner solution); second, because the household would not participate in the consumption of some commodities at any prices and income, for example non smokers with tobacco (non-participation); third, because no purchase has been made during the monitoring period of the survey although the household is a regular consumer of the goods (infrequency of purchase); and, fourth, misreporting due to any reason.

The incidence of zeros in the expenditure of some commodities has been typically modelled as a Tobit model (TOBIN [1958]). The underlying assumption in the Tobit model is that the decision to purchase and the amount of acquisition are driven by the same variables. This is a quite restrictive assumption if we take into account the different reasons that may cause zeros in expenditure surveys.

CRAGG [1971] developed several models for limited dependent variables that are extensions of the multiple Tobit analysis model. They differ from Tobit by allowing the determination of the size of the variable when it is not zero to depend on different parameters or variables from those determining the probability of being zero. As pointed out by CRAGG [1971], in some situations the decision to purchase and the amount of the acquisition may not be so intimately related (as it is assumed in the Tobit model).

In the 80s, an alternative modelling strategy arose, essentially in order to describe the short-run fluctuations of demands that are observed in the data and cannot be accounted for by price and income variations. The so-called Infrequency of Purchase Model (IPM), put forward by DEATON and IRISH [1984], KAY, KEEN, and MORRIS [1984], PUDNEY [1985], and BLUNDELL and MEGHIR [1987], among others, introduces a distinction between consumption and purchase behaviour and stresses the fact that a zero purchase may reflect either a zero consumption (as in the previous corner solution models) or a transitory abstention that may occur for various other reasons. The IPMs are econometric models that aim at providing the best possible estimation of expected demand functions given the available information about the underlying process of purchase renewals.

KAY, KEEN and MORRIS [1984] stress that the direct information on expenditures provided by household budget surveys may be of more value, since a household's chosen level of expenditure presumably reflects its own evaluation of its long-term economic position. Their paper is concerned only with the infrequency of purchase issue, *i.e.*, the relationship between the recorded expenditure and consumption of individual households. DEATON and IRISH [1984] consider a model where the standard Tobit specification is supplemented by the operation of a binary censor. Taking the Tobit specification as the starting point they add a second censoring process that randomly replaces a fraction of the observations generated by the Tobit model with zeros. BLUNDELL and MEGHIR [1987] claim that the IPM is a bivariate limited dependent variable model which may be used as an alternative to the Tobit model in the analysis of household consumption. They call their model *bivariate* as there is a separate process determining the zero-one discrete behaviour from that determining the continuous observations. The IPM provides a separate family of alternative models where zero values occur due to *durability* of a commodity in comparison to the duration of the survey period from which the expenditure observations are drawn.[1] In each case, allowing for these separate processes relaxes one of the strong assumptions underlying the standard Tobit model.

A natural extension of the two-regime IPMs, when the survey does not only record expenditures but also the number of purchases made during the recording period, consists of conditioning observed expenditures on the actual number of purchases or any function of it (see, for example, MEGHIR and ROBIN [1992] and ROBIN [1993]). The advantage of this approach is that if demand depends nonlinearly on total outlay, conditioning on purchase frequencies may substantially reduce this bias due to measurement error, as originally noticed by KEEN [1986].

As pointed out by ROBIN [1993], the main drawback of these IPMs is that given the decision to consume the decision to purchase is *ad hoc* and does not result from a properly defined individual decision process. Everyday consumption and purchases fluctuate according to a dynamic statistical process that should be the result of an underlying economic optimisation. Therefore, the identification of consumption functions requires panel data and, with cross-sections, only the expected consumer expenditures given present observables are identifiable.

It is worth mentioning that all the above methods have been applied to cross-sectional data. With panel data, however, there is the possibility to go beyond the one-snapshot nature of the information available in cross-sectional data in the sense that one can track household expenditure records over time for those who respond more than once.[2] An important implication with panel data is that if it is observed an alternation of positive and zero records over the number of interviews that a household undergoes, then one can infer that the household is a consumer of the good under consideration. One can also infer that zeros are

1. The question remaining once one of these broad specifications has been chosen relates to the feasibility of estimation and the provision of diagnostic tests for some of the underlying distribution assumptions. Both of these points are tackled in some detail in their paper and suggest the potential usefulness of applications of these models and tests to other data sets and preference specifications.
2. The data set we use in this paper is the *Spanish Family Expenditure Survey* (ECPF), from which it is possible to construct a real panel of eight periods for about 2,500 households.

due to, depending on the reference period and on the variability of the household budget between interviews, either infrequency of purchase or a switching regime between corner solutions and positive consumption.[3] For some goods (*i.e.* clothing), we can rule out the second possibility given that a person may not buy any clothes during a particular week, but nevertheless does not go naked. In this study, we are mainly interested in the case of zeros due to infrequency of purchase. Thus, we would rule out other possibilities by focusing on this sort of problem. As a result, it might be worth investigating the expenditure record of a group of households that can be tracked over time (panel data) since the probability of erroneous inference decreases with the number of periods we observe for every household.[4] Furthermore, using panel data would improve those estimations, given that the selection of those households who purchased at least once during the survey would not eliminate all the zero values of the number of times that a household purchases. Some other advantages of using panel data are that it is possible to control for individual heterogeneity and to introduce price variation.

Departing from the IPMs for cross-sectional data we have developed a IPM for panel data. We use a two stage panel data estimator in order to estimate the consumption equation for clothing. First, the probability purchase is estimated following CHAMBERLAIN [1980]. Second, in the demand equation, we correct the positive purchases using the previously estimated probabilities in order to get an estimate of clothing consumption. Once the infrequency has been taken into account we can estimate the panel data model by using the positive purchases and accounting for unobserved heterogeneity.

An application for all these models is provided using the *Spanish Family Expenditure Survey* (ECPF), and our results suggest that both price variation and individual heterogeneity play a role in the estimation of these models.

The layout of the paper is as follows. In the next section, we describe the statistical models. In section 3, the empirical application of each model is given using individual survey data drawn from the *Spanish Family Expenditure Survey* (ECPF). Finally, section 4 concludes.

2 Statistical Models

The model that has been used to account for censoring in commodity demand is the Tobit model (TOBIN [1958]). The underlying assumption with this model is that the same stochastic process determines **both** the value of continuous observations on the dependent variable and the discrete switch at zero. That is, a zero realisation of the dependent variable represents either a corner solution or a negative value for the underlying latent dependent variable. This restricts other quite reasonable determinants of zero observa-

3. However, in reality, changes in taste can induce participation changes.
4. In table 1 can be appreciated the dramatic decrease of the proportion in zeros in the expenditure of clothing as more periods of the survey are taken into account.

tions such as infrequency of purchase and non-participation or misreporting in commodity demand.[5]

The Tobit specification rests on strong distributional assumptions which should be tested if this model is to be used as a modelling framework for individual behaviour (see CHESHER and IRISH [1987], GOURIEROUX *et al.* [1987], BLUNDELL and MEGHIR [1987], among others). It is crucial to notice that as soon as it is recognised that zero observations on the dependent variable may be generated through infrequency of purchase then the censoring rule that underlies the Tobit model is **invalid** as a description of the censoring process and the sample log-likelihood function is misspecified. As a response to these problems, and following the literature, the bivariate models (or two regime IPMs) that we are going to consider provide a framework for dealing with these additional censoring rules.

The two-regime IPMs are obtained by conditioning expenditures on the alternative: purchasing or not. To deal with additional censoring rules, DEATON and IRISH [1984] develop a simple frequency of purchase model in which it is assumed that consumers correctly report purchases, but in which purchases themselves are made at intervals which may be longer than the period of the survey. To simplify the model they consider that the period of the survey is a fraction p of the purchase period. In this case, a purchase of y_i / p_i (where y_i is expenditure of household i) is observed with probability p_i during the survey, while with probability $(1 - p_i)$ no purchases are observed. This model is called the p-Tobit model.

DEATON and IRISH [1984] stress that with microdata, misreporting and infrequency of purchase are not distinguishable, at least with constant p. With non-constant p, identification would require some separation of the variables influencing consumption from those influencing p_i. An extension of this model, where the p_i are individual specific, can be found in BLUNDELL and MEGHIR [1987]. They develop a particular form of the IPM where zero values for the dependent variable cannot be attributed to corner solutions. This is a consumer demand model where consumption is always positive but recorded expenditures are often zero.[6] In this type of model, the latent variable is never directly observed. A positive expenditure will represent a purchase of stock whose services will be consumed over future periods typically longer than the period of observation. The IPM that BLUNDELL and MEGHIR [1987] present can be viewed as a snapshot of the dynamic process determining stock accumulation and consumption of services.

Finally, a generalisation of these models could be obtained by conditioning expenditures on the purchase frequency variable.[7] ROBIN [1993] points out that the identification of consumption functions requires panel data and that, with cross sections, only the **expected** consumer expenditures given present observables are identifiable. In general, however, it is hard, if possible at all, to

---

5. This restriction has been recognised, among others, by ATKINSON *et al.* [1984], DEATON and IRISH [1984] and BLUNDELL and MEGHIR [1987].

6. The example they take is clothing expenditure. Consumption services from this commodity are very rarely zero but surveys with short recording periods have some zero expenditures due to the durability of the clothing items and the resulting infrequency of purchase by consumers. Therefore, if we have a pure infrequency of purchase problem, once we account for the infrequency we have solved the problem of zeros in the sample.

7. Instead of conditioning on an indicator function, one could condition on any measurable function of purchase frequencies, see ROBIN [1993] and MEGHIR and ROBIN [1992], for some applications of this extension.

derive a closed form for these expectations both because this requires information on the joint process of the set of demand and explanatory variables, and because of the non-linearities arising from the qualitative structure of the model. Moreover, even if this time-aggregation were possible, it is well known that not all the preference parameters would be identifiable. And furthermore, with the type of data available in micro-surveys, in order to identify the nature of the observed zeros we would need prior information.

With panel data it is possible to find out more about the nature of the zero records. In the particular case of clothing expenditure it seems (see table 1) that the frequency of purchase story is the most reasonable. From this table we may also infer an informal test that gives evidence on the fact that during the recording period everybody consumes clothes but not everybody spends on clothing. As we take into account more periods of the sample there is a dramatic drop in the proportion of zeros (in the expenditure of adult clothing), as one would expect if infrequency of purchase is the main reason behind zero reports on clothing expenditure.[8] Thus, accounting for infrequency of purchase would solve the problem of zeros in a model for clothing consumption.[9]

TABLE 1
Proportion of Zeros on Adult Clothing Expenditure

Period (quarter)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Number and proportion of zeros in each quarter (whole sample)	1536 (19%)	1531 (17.6%)	1537 (17,5%)	1575 (17,9%)	1272 (18,3%)	940 (18,2%)	690 (18,5%)	420 (17,6%)
Number and proportion of zeros remaining after each quarter (whole sample)	1536 (19%)	496 (5.7%)	209 (2.38%)	106 (1.2%)	58 (0.83%)	23 (0.44%)	12 (0.32%)	6 (0.25%)
Number and proportion of zeros in each quarter (panel)	264 (17%)	240 (15.5%)	234 (15.1%)	260 (16.8%)	263 (17.0%)	250 (16.1%)	283 (18.2%)	262 (17.0%)
Number and proportion of zeros remaining after each quarter (panel)	264 (17%)	65 (4.19%)	21 (1.35%)	8 (0.52%)	5 (0.32%)	2 (0.13%)	1 (0.06%)	0 (0.0%)

Notes:

- ECPF, 1985.1 to 1991.4.

- Whole sample: the number of observations is 53.091 and the number of zeros is 9.600. Therefore, the proportion of zeros is 18.1%.

*- Panel sample: the number of observations is 12.400 (1550*8) and the number of zeros is 2.056. Therefore the proportion of zeros is 16.58%.*

8. There is a detailed description of the data in the Data appendix.
9. Consumption is a continuous variable although it is a latent variable.

In order to use the advantages of panel data, the extension we propose is to develop a commodity demand model for panel data that follows the developments for cross-sectional data of KAY, KEEN and MORRIS [1984], KEEN [1986], PUDNEY [1985 and 1989] and BLUNDELL and MEGHIR [1987]. Following BLUNDELL and MEGHIR [1987] we also choose a commodity (clothing) for which it is unlikely that any zero expenditures represent a corner solution.

The motivation of developing an IPM for panel data is to compare this model with the IPM with cross-sectional data in order to shed some light on the existence of individual heterogeneity and to assess the possibility of identifying those models using cross-sectional data. Apart from the advantages already mentioned, when dealing with panel data, it is possible to introduce price variation. In the specification below, the first equation tries to capture the decision of purchase and the second the consumption of clothing.[10]

Formally, we define consumption of individual i in period t for $i = 1,\dots,N$ and $t = 1,\dots,T$, as y_{it}^*. Consumption is such that, conditional on $z_i = (x_{it}',\dots,x_{iT}')$ and unobserved heterogeneity α_i (which can be a vector) we have,

(1) $$E(y_{it}^*|z_i,\alpha_i) = x_{it}'\beta + \alpha_i^{(1)}$$

where $\alpha_i^{(1)}$ is a subset of α_i.

We assume that individual purchases and consumption are equal on average, *i.e.*,

(2) $$E(y_{it}^*|z_i,\alpha_i) = E(y_{it}|z_i,\alpha_i)$$

Purchase occurrences are driven by the variable D_{it} in the following way,

(3) $$y_{it} > 0 \Leftrightarrow D_{it} > 0$$

and

(4) $$D_{it} = x_{it}'\theta + \alpha_i^{(2)} + \nu_{it}$$

where ν_{it} is a zero-mean error term assumed independent of z_i and α_i, and where $\alpha_i^{(2)}$ is a second component of α_i.

Finally, unobserved and observed heterogeneity components are correlated as follows,

(5) $$\alpha_i^{(j)} = z_i\pi^{(j)} + \eta_i^{(j)}, \quad j = 1,2.$$

with $E\left(\eta_i^{(j)}|z_i\right) = 0$.

Defining y_{it} as the observed expenditure we have by the law of iterated expectations

(6) $$E(y_{it}|z_i) = E(y_{it}|z_i,y_{it} > 0)p_{it}$$

given that $E(y_{it}|z_i,y_{it} \leqslant 0)(1 - p_{it}) = 0$.
In equation (6) $p_{it} = \Pr(y_{it} > 0|z_i)$ is the probability of purchase.[11]

10. We have to introduce individual unobserved heterogeneity in both equations.
11. In the conditioning set, we only use the sub-set of all the strictly exogenous regressors of the model. Also, note that the first element in each x_{it} is a one.

Therefore, we have

$$E(y_{it}|z_i,y_{it} > 0)p_{it} = E(y^*_{it}|z_i) \tag{7}$$

As noticed by BLUNDELL and MEGHIR [1987], since $p_{it} < 1$ (7) implies that observed expenditure when positive will, on average, exceed the level of consumed services. It is interesting to discuss the nature of the above assumption. It is possible to some extent to check the truthfulness of the assumption through the inspection of the real data. For example, following DEATON and IRISH [1984], if the assumption is true then those households who only buy four times out of eight quarters should buy twice as much as those who buy the eight periods. Taking the empirical individual probabilities one can assess the extent of this assumption.[12] Moreover, if we aggregate the data over time per individual it is possible to relax the assumption that $E(y^*_{it}|z_i,\alpha_i) = E(y_{it}|z_i,\alpha_i)$ (as we present below). Relaxing the above assumption does not prevent us from implementing a consistent estimator. Table 1 presents evidence about this point; aggregation over time (individual means) eliminates the zero records.

Now, since, using (4) and (5) we have

$$D_{it} = x'_{it}\theta + z_i\pi^{(2)} + \eta_i^{(2)} + \nu_{it} = x'_{i1}\delta_{t1} + ... + x'_{it}\delta_{tt} + ... + x'_{iT}\delta_{tT} + \eta_i^{(2)} + \nu_{it} \tag{8}$$

given that z_i includes x_{it} among its elements, $\delta_{tr} = \pi_r^{(2)}$ if $r \neq t$ and $\delta_{tt} = \pi_t^{(2)} + \theta$. Assuming that $\eta_i^{(2)} + \nu_{it}$ is distributed according to a standard $N(0,1)$ distribution, we have that[13]

$$\Pr\{D_{it} > 0|z_i\} = \Phi(x'_{i1}\delta_{t1} + ... + x'_{iT}\delta_{tT}) = \Phi(z_i\delta_t) \tag{9}$$

Moreover,

$$E(y^*_{it}|z_i) = x'_{it}\beta + E(\alpha_i^{(1)}|z_i) = x'_{it}\beta + \pi^{(1)\prime}z_i \tag{10}$$

Note that we write x_{it} as the explanatory variables in equations (8) and (10). We use these variables in both equations to simplify notation although some exclusion restrictions, *i.e.* some elements of x_{it} will be omitted in equation (10), in order to identify both sets of parameters.

To estimate the above model we use two different approaches. Firstly, we present an estimator based on the aggregation over time per individual, of our data set. Secondly, we present a two step procedure that follows CHAMBERLAIN's [1980] approach in the treatment of the individual effects.

As table 1 shows, aggregating over time solves the problem of zero records. Therefore, an interesting exercise is to specify the model of interest in terms of

12. To make this comparison it is necessary to select homogeneous households: same composition, income, labour supply, etc.
13. As pointed out by WOOLDRIDGE [1995], although normality is assumed, the temporal dependence in $(\nu_{it} : t = 1,...,T)$ could be entirely unrestricted. In our specification, without loss of generality, we assume that $\sigma_t^2 = 1$. Therefore, the probability of purchase, p_{it}, can be estimated as $\Phi(z_i\delta_t)$, where $\Phi(\cdot)$ represents the standard normal distribution function.

the unit means, *i.e.*, aggregating the data set over time for each individual.[14] The idea in using the time-aggregated model is to avoid the problem of zero records in the dependent variable, although we have to assume strong exogeneity of the regressors.

The previous identifying assumption (*i.e.*, $E(y_{it}|z_i,\alpha_i) = E(y^*_{it}|z_i,\alpha_i)$) can be relaxed given that for this estimator we only need to assume that

$$(11) \qquad E\left(\sum_{t=1}^{T} y_i|z_i,\alpha_i\right) = E\left(\sum_{t=1}^{T} y^*_i|z_i,\alpha_i\right)$$

where, as before, y^*_i is consumption and y_i is expenditure. Therefore on average zero records compensate positive purchases, and we do not need to assume (calculate) any probability of purchase. The aggregation of purchases, taking into account the zeros of the sample, can be equated to the average consumption. This assumption is weaker than the previous one in the sense that, if the reason behind the zero records is infrequency then the longer the time span the more likely the assumption to hold. A simple inspection of table 1 confirms the intuition behind (11).

To get consistency when we aggregate the model across time, for each household, we need to make some assumptions. First, we assume asymptotics on N (observations or individuals). Second, we impose independence between the frequency of purchase and consumption. Finally, we assume strict exogeneity. Which means that there is no feedback in the aggregated model.

It should be stressed that this model is robust to any frequency of purchase structure given that we do not assume anything about the underlying probability of purchase. The only required assumption is time stationarity of the frequency of purchase (or time stationarity of the probability of purchase per individual).[15] A final remark on this estimator is that it does allow for heteroskedasticity and therefore it is possible to test for heteroskedasticity.

We turn now to the other estimation approach. The estimation of the model presented in equation (10) leads to the following procedure:

(i) For each $t = 1,2,...,T$ estimate equation (9) by standard probit. For $D_{it} > 0$ obtain the probability of purchase as $p_{it} = \Phi(z_i\delta_t)$.[16]

(ii) To estimate equation (10) we use a two step procedure on the subsample of positive observations y_{it}. We have that

$$(12) \quad \Phi(z_i\delta_t)y_{it} = x'_{it}\beta + \pi^{(1)\prime}z_i + u_{it} = x'_{i1}\gamma_{t1} + ... \\ + x'_{it}\gamma_{t1} + ... + x'_{iT}\gamma_{tT} + u_{it} = z_i\gamma_t + u_{it} \quad \text{cov}(u_{it},z_i) = 0$$

where $\gamma_{tr} = \pi^{(1)}_r$ if $r \neq t$ and $\gamma_{tt} = \pi^{(1)}_t + \beta$.

---

14. Thus, if the panel data set is composed of N individuals and T time periods, aggregating over time yields a data set of N observations.

15. It does not seem to be very strong to assume that the pattern of purchase of a household is quite stable in the long run, *i.e.*, the distribution of the purchase probability of clothing is constant over time for each household.

16. If one is interested in recovering the structural parameter vectors, θ and $\pi^{(2)}$, this is possible using a two step procedure proposed by CHAMBERLAIN [1980] that is described as follows. In the first step, a Probit is estimated in each t, and in a second step, one can recover θ and $\pi^{(2)}$ by minimum distance. As our main interest is to estimate the probabilities of purchase, we do not recover the structural parameters and thus, we do not implement the minimum distance procedure.

In the first step, we estimate the above model for each t to get the reduced form parameters (or unrestricted parameters) for the sub-sample of positive observations y_{it}. We do LS regression of y_{it} on z_i to get the reduced form parameters γ_t. In a second step, in order to recover the structural parameters, $\psi = (\beta', \pi^{(1)\prime})'$, we use a minimum distance procedure that imposes the cross-equation restrictions.[17] Wave t provides an estimator $\hat{\gamma}_t$ for the parameter vector γ_t. Defining $\hat{\gamma} \equiv (\hat{\gamma}_1', \ldots, \hat{\gamma}_T')'$ and $\gamma \equiv (\gamma_1', \ldots, \gamma_T')'$, the cross equation restrictions to be exploited in the minimum distance procedure are $\gamma = R \cdot \psi$, where γ is the reduced form parameters stacked vector for all waves, $\psi = (\beta', \pi^{(1)\prime})'$ is the vector of structural parameters we want to recover in the minimum distance step and R is the matrix of restrictions that relates the reduced form parameters to the structural ones (see CHAMBERLAIN [1980]).

(iii) Estimate the asymptotic variance of $\hat{\psi}$, taking into account that we have previously estimated the probabilities of purchase.[18]

It is interesting to stress the nature of a pure infrequency of purchase model. Once we account for the probability of purchase and correct the positive observations of expenditure by this probability we obtain a measure of the *consumption* made by each individual. This allows us to drop all those observations in which we observe a zero, as in the remaining observations we directly account for the individual infrequency of purchase.[19] The advantage of having a panel data is twofold. First, we can directly assess if it is realistic to assume that everybody consumes clothing (see table 1); second, we can control for unobserved heterogeneity.

3 Empirical Application

3.1 Data and Variables

The data for this study are drawn from the *Spanish Family Expenditure Survey* (ECPF). This is a quarterly data set for 1985.1 to 1991.4. The ECPF is a quaterly cross-section survey of about 3,200 households that collects detailed information on household characteristics, income and expenditures. We construct a balanced panel data set in which it is possible to track the same household across time for a period of up to 8 quarters. This panel is composed of 12,400 observations (1,550 households over 8 quarters). We select all households staying in the sample for the maximum period of eight quarters and containing one married couple and possibly other adults. The head of the household is older than 20, but less than 65.

17. Not that the estimation of the structural parameter ψ requires at least two waves.
18. In appendix 1, we derive the asymptotic variance-covariance matrix.
19. A similar approach is taken in MEGHIR and ROBIN [1992] although they are interested in other goods.

We can think of the model to be estimated, as an equation from a complete demand system.[20] As pointed out by BLUNDELL and MEGHIR [1987], household's clothing expenditures recorded in a survey of limited duration seem to be one of the most appropriate commodity for illustrating infrequency of purchase models. The consumption of services from clothing made by households must be positive, but after aggregating across all adult clothing expenditures in a household, some zero expenditures are still likely (see table 1). In our case, we focus on the demand for adult clothing. The endogenous variable in all estimated models is the share of adult clothing consumption; it represents 8.44 % of total expenditure as an average for all individuals in the sample, and 10.36 % if we only take into account the subsample of consumers with positive observations.

An empirical issue for the estimation of model (10) is the consideration of some exclusion restrictions in order to estimate the parameters of both the purchase occurrence equation and the demand equation. As pointed out by BLUNDELL and MEGHIR [1987], the Engel curve, seen as an income-consumption equation, can be thought of as a reduced form from a set of structural relationships describing all household expenditures, savings and even labour market decisions. Therefore, a variety of economic and demographic variables can be expected to enter the determination of consumption. The probability of purchase, on the other hand, may depend more directly on variables determining the relative time and money costs of purchase and on general economic and demographic factors.[21] It would also be interesting to depart from a structural model for both the consumption policy and the purchase policy, in the same lines than MEGHIR and ROBIN [1992]. In practice, we follow BLUNDELL and MEGHIR [1987] approach and introduce a set of demographic variables in the selection equation that differs from the one used in the demand equation. We also introduce quarter dummies in our selection equation, but not in the demand equation.

As explanatory variables we use the logarithm of total real income (**lrtinc**) to account for the effect of income. We use quarterly prices for clothing. This variable is defined in terms of the logarithm of the relative price, *i.e.*, the ratio of the price index of clothing to the price index of other goods (**lpcloth**). It is important to notice that the variability with which this variable has been introduced makes it to capture seasonal effects. In order to capture other possible seasonal and time effects we also introduce quarter and year dummies (**q2, q3, q4** and **d86, d87, d88, d89, d90, d91**).

The rest of the variables included in our empirical specification are the following: family composition variables (**n1, n2, n3, n4, na1, na2, no**) which take into account the number of children, adults and elderly people in the hou-

20. The demand system we have in mind is the Almost Ideal Demand System (AIDS) derived by DEATON and MUELLBAUER [1980], in which the variables to be explained, the expenditure shares, are related linearly to the logarithm of prices and the logarithm of total real expenditure.

21. Looking at studies that have modelled expenditures using the *British Family Expenditure Survey* (British FES) (*e.g.*, ATKINSON and STERN [1980]; ATKINSON, GOMULKA and STERN [1988]; DEATON and IRISH [1984]; KAY, KEEN and MORRIS [1984]; BLUNDELL and MEGHIR [1987]) and using other data sets (*e.g.*, MEGHIR and ROBIN [1992], GARCÍA and LABEAGA [1993]), it is important to notice that family composition and age variations may be important determinants of consumption. And following the approach by BLUNDELL and MEGHIR [1987], we also enter education and labour market variables as variables that may influence the level of consumption and the degree of *"stocking-up"* through current or future income expectations.

sehold;[22] education of the head of the household (**ed4, ed0**); place of residence of the household (**drura, dcity**); age of the head of the household and his wife (**ag1, ag2**); labour status of the head of the household and his wife (**dunem, demploy, dpensi, dwwife**) and employment category of the head of the household (**bcollar, wcollar, unskill, highrank**). The descriptive statistics of these variables, for both the whole sample and the subsample of consumers, and a glossary of variable definitions are provided in the Data Appendix.

3.2 Results and Diagnostics

The empirical applications that we report in this section (see table 2) refer to the three statistical models that have been presented in the previous section. The first column refers to the maximum likelihood estimates of the infrequency of purchase model using the whole sample as a cross-section. In the second column, we report the estimates for the time aggregated model. Finally, the third column of the table presents the results of the IPM using panel data. We also report price and income elasticities of clothing consumption for all the models estimated (see table 3).[23]

In the estimation of the IPM for cross-sectional data, we replicate the approach of BLUNDELL and MEGHIR [1987] for the demand of clothing.[24] In the consumption equation, we observe that price increases contribute to an increase of the share of expenditure on adult clothing. Rises in income increase the share spent on adult clothing as well. A striking result is that age (both for the head of the household and his wife) does not seem to affect the share spent on adult clothing. This result seems to be counterintuitive. The presence of children in the family, as could be expected, has important effects on the share of adult clothing. The fact that the head of the household is unemployed does not seem to affect the share of clothing consumption, while if the wife is working or the head of the household is a "*blue collar*" worker raises the share spent on adult clothing. Looking at the locational variables, households living in rural areas have a bigger share spent on adult clothing, and the opposite seems to happen to households living in big cities. Regarding education, it may be expected that the higher the education level of the head of the household the lower the share spent on clothing. This seems to be confirmed by our results. To test for independence and heteroskedasticity, we use the approach derived in BLUNDELL and MEGHIR [1987]. We use the score test following the methodology developed by GOURIEROUX, MONFORT, RENAULT and TROGNON [1987]. Computation of the score test statistics in our empirical illustration follows CHESHER's HR^2 formulation [1983]. The independence and heteroskedastic tests give quite acceptable results (see bottom of table 2). As a summary, the overall properties of this model of consumption are quite reasonable. The standard errors reported in column 1 of table 2 have been calculated using the asymptotic variance developed in the appendix.

22. The family composition variables used in the frequency of purchase equations are: D1, D2, D3, D4, DA1, DA2, Do, D2q and D2qq. In the data appendix, there is a description of these variables.
23. In order to save some space in the paper, the frequency of purchase estimates for both the cross-section specification and for the panel data estimates (using a probit specification) are not reported in the paper although are available from the author upon request.
24. We have selected adult clothing in order to avoid other problems related to children clothing, *i.e.* the arrival of a new child in the family, the existence and sex of other children, etc.

TABLE 2
Demand Estimates

Variable	IPM with cross-sectional data[(1)]		Time aggregated model[(2)]		Panel data estimator [(3)]	
	Method of Estimation					
Intercept	– 0.5087	(0.0069)	–		– 0.1940	(0.0450)
ag1	0.0001	(0.0001)	0.0002	(0.0004)	– 0.0006	(0.0021)
ag2	0.0002	(0.0041)	0.0001	(0.0003)	0.0014	(0.0021)
n1	– 0.0125	(0.0017)	– 0.0116	(0.0025)	– 0.0085	(0.0045)
n2	– 0.0187	(0.0013)	– 0.0179	(0.0023)	– 0.0114	(0.0045)
n3	– 0.0198	(0.0015)	– 0.0181	(0.0017)	– 0.0159	(0.0035)
n4	0.0062	(0.0016)	0.0097	(0.0025)	– 0.0027	(0.0033)
na1	– 0.0226	(0.0123)	0.0007	(0.0023)	– 0.0011	(0.0033)
na2	– 0.1111	(0.0159)	– 0.0085	(0.0029)	– 0.0031	(0.0050)
lpcloth	0.0459	(0.0145)	– 0.1371	(0.0701)	0.0323	(0.0128)
drura	0.0088	(0.0018)	0.0063	(0.0026)	0.0058	(0.0019)
dcity	– 0.0148	(0.0034)	– 0.0170	(0.0041)	– 0.0153	(0.0031)
ed0	0.0091	(0.0021)	0.0063	(0.0034)	– 0.0031	(0.0056)
drent	0.0133	(0.0025)	0.0101	(0.0038)	0.0182	(0.0170)
dunem	– 0.0033	(0.0035)	– 0.0008	(0.0064)	– 0.0079	(0.0035)
dwwife	0.0039	(0.0021)	0.0049	(0.0032)	0.0059	(0.0021)
bcollar	0.0135	(0.0043)	0.0136	(0.0075)	– 0.0013	(0.0091)
wcollar	– 0.0024	(0.0020)	–		– 0.0056	(0.0050)
unskill	0.0032	(0.0028)	–		0.0044	(0.0056)
highrank	– 0.0084	(0.0055)	–		– 0.0329	(0.0123)
lrtinc	0.0478	(0.0059)	0.0317	(0.0035)	0.0233	(0.0035)
q2	– 0.0122	(0.0021)	–		– 0.0104	(0.0022)
q3	– 0.0176	(0.0021)	–		– 0.0153	(0.0021)
q4	– 0.0087	(0.0024)	–		– 0.0042	(0.0025)
d86	0.0176	(0.0083)	–		–	
d87	0.0151	(0.0036)	–		–	
d88	0.0101	(0.0034)	–		–	
d89	0.0078	(0.0034)	–		–	
d90	0.0053	(0.0035)	–		–	
sigma	0.0765	(0.0005)	–		–	
N	12400		1550		10041	
Log-likelihood	– 1.6797		–		–	

Notes:

(1) Dependent variable: share of adult clothing times the probability of purchase. Test for independence $\chi^2_{(1)} = 2.749$. Test for heteroskedasticity of the errors $\chi^2_{(6)} = 12.233$.

(2) Dependent variable: time averaged share of adult clothing. Estimates obtained using the between-groups estimator.

(3) Dependent variable: share of adult clothing times the probability of purchase. Estimates using the panel data estimator.

(4) Asymptotic standard errors in parenthesis.

We turn now to the estimates of the frequency of purchase equation.[25] The frequency of purchase probability estimates have plausible properties although the model is overparameterised. Income influences positively the probability of purchase, and educational variables, family composition variables and locational variables seem to have effects in the correct direction. The age of the head of the household and the age of his wife have opposite effects, the older the wife the less is the probability of purchase. We have estimated this model under the assumption of independence and normality of the disturbance term. We test for normality of the error in the binary censor for the frequency of purchase model. The result we get from this test is that we can not reject the H_0 of normality.[26]

The second step in our empirical application is the estimation of the time aggregated model (column 2 of table 2).[27] The aim of estimating this model is to compare it with the IPM estimated using cross-sectional data. To do this, we use a variation of the Hausman test (see MROZ [1987]). Assuming strict exogeneity all over, the null hypothesis of the test is that the IPM is the true model (as we use probabilities of purchase in the IPM, the estimates are more efficient), while the time aggregated model is consistent. The result of the test is $\chi^2_{(17)} = 114.18$, and therefore we reject the null hypothesis. This could imply that the assumption of strict exogeneity is not appropriate or that the IPM specification is not the correct one.

The third step in our analysis is the estimation of the IPM with panel data (column 3 of table 2). An important issue when using panel data is that we control for individual heterogeneity and we can include price variation across time, allowing to estimate more reliably the elasticities.

In the estimation of the demand equation with panel data, we previously estimate the frequency of purchase equation (the decision to purchase) for each period t using a probit model.[28] As a second step we estimate the demand for clothing equation. Overall the results we get are quite plausible and are in line with the predictions of the economic theory.[29]

With these results, we have tested for the correlation between the fixed effects and the explanatory variables. If α_i and z_i are correlated, then assuming $\pi^{(1)} = \pi^{(2)} = 0$ yields an unconsistent estimator. Thus, we can implement a Wald test to analyse this correlation. In this paper, we have separately analy-

---

25. We have modelled the purchase probability equation using quarter and year dummies, income, labour status of the head of the household and his wife, education level of the husband, place of residence of the family, age and age squared for both the head of the household and his wife, property status of the house and a set of family composition dummies. For the sake of brevity, we do not report these results although they are available from the author upon request.
26. The normality test is based on the BERA and JARQUE [1982] methodology which can be regarded as components of White's Information Matrix test (WHITE [1982b]).
27. This estimator is robust to any frequency of purchase model as, after time aggregation over individuals, practically all zeros disappear from the sample (see table 1); therefore there is no need to specify the purchase probability model.
28. We do not report the results of the estimation of the frequency of purchase equation in order to save some space, although these are available upon request.
29. For the sake of brevity, the estimates of the fixed effects variables are not reported in table 2, although are available upon request.

sed $\pi^{(1)} = 0$ and $\pi^{(2)} = 0$. The results of these tests are $\chi^2_{(153)} = 382.50$ and $\chi^2_{(157)} = 348.88$,[30] respectively, suggesting that it is crucial to control for unobserved individual heterogeneity.[31]

Finally, another diagnostic that has been carried out is the efficiency gains when estimating the IPM using the information provided in the frequency of purchase equation. To estimate our model we need an identification assumption, *i.e.*, $E(y^*_{it}|z_i,\alpha_i) = E(y_{it}|z_i,\alpha_i)$. It follows then that regressing y_{it} on z_i also yields a consistent estimate.[32] Comparing both estimators, it is possible to assess the efficiency gains obtained given that the estimator using our panel data model should be more efficient as it uses more structure.[33] Indeed, the panel data estimator is more efficient being the efficiency gain around a 10.7 %, on average.

A further step in our analysis is to compare the IPM with panel data estimates, $\hat{\beta}_{PD}$, with the IPM with cross sectional data estimates, $\hat{\beta}_{CS}$. The motivation of this test is based on the idea that under the hypothesis of no individual effects, both models are consistently estimated but the estimates from the IPM with panel data are inefficient (there is no need of estimating the model with panel data if we do not have fixed effects). However, under the alternative hypothesis of correlated individual effects, only the IPM with panel data estimates remain consistent. Under H_0 we have the chi-squared test based on the Wald criterion $W = \chi^2_{(J)} = [\hat{\beta}_{PD} - \hat{\beta}_{CS}]'\Sigma^{-1}[\hat{\beta}_{PD} - \hat{\beta}_{CS}]$, where $\hat{\Sigma} = \mathrm{Var}[\hat{\beta}_{PD} - \hat{\beta}_{CS}]$.[34] Under the null hypothesis, W is asymptotically distributed as a chi-squared with J degrees of freedom. In order to carry out the test, we need to calculate the asymptotic variances of the estimators from the infrequency of purchase models. In the appendix, there is a description of the asymptotic variance for both estimators ($\hat{\beta}_{PD}$ and $\hat{\beta}_{CS}$). The result we get from this test is $\chi^2_{(24)} = 1271.05$, and therefore we clearly reject the null hypothesis, suggesting that there are fixed effects which have to be taken into account in the estimation of these models, or, in other words, we reject the IPM estimated with cross-sectional data against the IPM estimated with panel data.

Finally, in table 3, we report the estimated values of income and price elasticities together with their standard values, for all models estimated. The income elasticities we get seem quite reasonable (and always significant) although it seems that the estimate from the IPM with cross-sectional data is the most precise (the lowest standard error). In the case of the price elasticities all the estimates are negative (as predicted by the economic theory) although we have a different range of estimates.

30. We have also tested for the correlation between regressors and unobserved heterogeneity in each of the t probits estimated and the results are similar to the one reported.

31. We have also tested for the existence of individual effects by comparing the estimates of the time aggregated model and the estimates using the panel data estimator following the suggestion of one referee. The result of this test is $\chi^2_{(20)} = 74.55$ and therefore we reject the H_0.

32. A similar estimator has been proposed by KEEN [1986] in a cross sectional context.

33. This point has been already made by MEGHIR and ROBIN [1992], and it is also in line with a suggestion made by one of the referees.

34. As before, the formulae used to construct the covariance matrix for $(\hat{\beta}_{PD},\hat{\beta}_{CS})$ follow MROZ's [1987] approach.

TABLE 3
Price and Income Elasticities

Model	**Price Elasticity**	**Income Elasticity**
IPM with cross-section data	– 0.55 (0.13)	1.46 (0.018)
Time Aggregated Model	– 2.37 (0.70)	1.37 (0.037)
IPM with the panel data estimator	– 0.67 (0.13)	1.23 (0.034)

Notes:
(1) Standard errors in parenthesis.

4 Summary and Conclusion

This paper discusses several classes of bivariate limited dependent models which may be used as alternatives to the Tobit model in the analysis of household consumption behaviour for goods that present some infrequency in their purchase. The IPM provides a separate family of alternative models where zero values occur due to the "*durability*" of a commodity for which the expenditure observations are drawn. Most of the empirical applications of these models have used cross-sectional data although it is crucial to control for individual heterogeneity and price variation in this context. In our application, we propose a panel data model which takes into account these questions. We estimate this model using a panel data estimator.

We have estimated these models using data from the *Spanish Family Expenditure Survey* (ECPF) for the period 1985 to 1991. The main aspect we exploit is to compare the estimation results when estimating infrequency of purchase models using cross-sectional data and panel data and we also carry out some diagnostic tests. We develop an IPM for panel data and a time averaged model which, under some restrictions, is robust to any specification of the infrequency of purchase equation. We also test these models against the cross-section model in our empirical application. We have also tested for the exogeneity of the fixed effects and the results we get seem to confirm that it is important to account for unobserved individual effects. The estimated income and price elasticities seem to be affected when we account for individual effects. ■

• References

AMEMIYA T. (1981). – « Qualitative Response Models: A Survey », *Journal of Econometrics*, 19, pp. 481-536.

ATKINSON A. B., STERN N. (1980). – « On the Switch from Direct to Indirect Taxation », *Journal of Public Economics*, 14, pp. 195-224.

ATKINSON A. B., GOMULKA J., STERN N. (1988). – « Spending on Alcohol: Evidence from the Family Expenditure Survey 1970-1983 », *The Economic Journal* 100, pp. 808-827.

BERA A.K., JARQUE C.M. (1982). – « Efficient Specification Tests for Limited Dependent Variable Models », *Economics Letters* 9, pp. 153-160.

BLUNDELL R., MEGHIR C. (1987). – « Bivariate Alternatives to the Tobit Model », *Journal of Econometrics*, 34, pp. 179-200.

CHAMBERLAIN G. (1980). – « Analysis of Covariance with Qualitative Data », *Review of Economics Studies*, 47, pp. 225-238.

CHESHER A. (1983). – « Computing Score Tests Statistics: An Expository Note », *Mimeo*, University of Bristol, Bristol.

CHESHER A., IRISH M. (1987). – « Residual Analysis in the Grouped and Censored Normal Linear Model », *Journal of Econometrics*, 34, pp. 33-61.

DEATON A., IRISH M. (1984). – « Statistical Models for Zero Expenditures in Households Surveys », *Journal of Public Economics*, 23, pp. 59-80.

DEATON A., MUELLBAUER J. (1980). – *Economics and Consumer Behaviour*, New York: Cambridge University Press.

GARCÍA J., LABEAGA J.M. (1992). – « A Cross-Section Model with Zeros: An Application to the Demand for Tobacco », Universitat Pompeu Fabra, Economics *Working Paper* 26.

GOURIEROUX C., MONFORT A., RENAULT E., TROGNON A. (1987). – « Simulated Residuals », *Journal of Econometrics* 34, pp. 201-225.

CRAGG J. G. (1971). – « Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods », *Econometrica*, 39, pp. 829-844.

HAM J. C. (1982). – « Estimation of a Labour Supply Model with Censoring Due to Unemployment and Underemployment », *Review of Economics Studies* 49, pp. 335-354.

HAUSMAN J. (1978). – « Specification Tests in Econometrics », *Econometrica*, 46, pp. 1251-1272.

KAY J.A., KEEN M.J., MORRIS C.N. (1984). – « Estimating Consumption from Expenditure Data », *Journal of Public Economics*, 23, pp. 169-181.

KEEN M.J. (1986). – « Zero Expenditures and the Estimation of Engel Curves », *Journal of Applied Econometrics*, 1, pp. 277-286.

LEE M.J. (1996). – *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variables Models*, Springler.

LEE L-F., MADDALA G.S., FROST R. (1980). – « Asymptotic Covariance Matrices of Two-Stage Probit and Two-Stage Tobit Methods for Simultaneous Equation Models with Selectivity », *Econometrica*, 48, pp. 491-503.

MEGHIR C., ROBIN J. M. (1992). – « Frequency of Purchase and the Estimation of Demand Systems », *Journal of Econometrics*, 53, pp. 53-85.

MROZ T.A. (1987). – « The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions », *Econometrica*, 55, pp. 765-799.

PUDNEY S.E. (1985). – « Frequency of Purchase and Engel Curve Estimation », London School of Economics, *Discussion Paper* A56, Econometrics Programme.

PUDNEY S.E. (1989). – *Modelling Individual Choice: The Econometrics of Corners, Kinks and Holes*, Oxford: Basil Blackwell.

ROBIN J. M. (1993). – « Econometric Analysis of the Short-run Fluctuations of Households' Purchases », *Review of Economics Studies*, 60, pp. 923-934.

TOBIN J. (1958). – « Estimation of Relationships for Limited Dependent Variables », *Econometrica*, 26, pp. 24-36.

WHITE H. (1982a). – « Maximum Likelihood Estimation of Misspecified Models », *Econometrica*, 48, pp. 1-25.

WHITE H. (1982b). – « Instrumental Variables Regression with Independent Observations », *Econometrica*, 50, pp. 483-499.

WOOLDRIDGE J. M. (1995). – « Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions », *Journal of Econometrics*, 68, pp. 115-132.

APPENDIX 1

1 The Asymptotic Variance of the IPM Estimator

We represent the demand equation by the regression function:

(A.1) $$y(\theta)_i = x_i\beta + u_i \qquad i = 1,\ldots,N.$$

where we denote the parameter vector of the frequency of purchase equation by θ, and

(A.2) $$y(\theta)_i = y_i\Phi(z_i\theta)$$

where y_i is the share of adult clothing expenditure over total expenditure of the household and $\Phi(z_i\theta)$ is the probability of purchase. To compute the asymptotic standard errors we need the derivatives of $y(\theta)_i$ with respect to θ

(A.3) $$\frac{\partial y(\theta)}{\partial\theta_k} = y_i\phi(z_i\theta)z_{ik}$$

The estimator of θ (the Probit estimate) has an asymptotic covariance matrix V_θ. The error term of the model, when we condition on consistent parameter estimates $\hat{\theta}$, can be approximated to the first order by

(A.4) $$u_i^* = u_i + (\partial y(\theta)/\partial\theta)(\hat{\theta} - \theta) \equiv u_i + Q_i'(\hat{\theta} - \theta)$$

hence $E(u^*(u^*)') = \Sigma + QV_\theta Q'$.[35] Then, we have[36]

(A.5) $$\begin{aligned}(\hat{\beta} - \beta) &= \left(\sum_{i=1}^{N} D_i x_i x_i'\right)^{-1} \sum_{i=1}^{N} D_i x_i u_i^* \\ &\overset{D}{=} \left(\sum_{i=1}^{N} D_i x_i x_i'\right)^{-1} \sum_{i=1}^{N} D_i x_i[u_i + Q_i'(\hat{\theta} - \theta)]\end{aligned}$$

where D_i is an indicator function that takes value 1 if the individual has a positive purchase and zero otherwise. Then, the asymptotic covariance matrix of β,

35. In general, there should exist a correlation between u_i and Q_i in expression (A.4). However, since the right IPM is the one given in equation (4) of section 2, then it follows that the correlation between the error term from the second stage and the error term from the first stage is equal to 0 and therefore the relation between is 0. I should thank one of the referees for clarifying this aspect of the formula for the asymptotic variance.

36. The expression $A \overset{D}{=} B$ means that A has the same asymptotic distribution as B.

ignoring covariance terms (LEE *et al.* [1980, p. 500]), yields,

$$A\,\mathrm{Var}(\beta) = \left(\sum_{i=1}^{N} D_i x_i x_i'\right)^{-1} \tag{A.6}$$

$$\sum_{i=1}^{N} D_i x_i [\mathrm{var}(u_i) + Q_i' \mathrm{var}(\hat{\theta}) Q_i] \sum_{i=1}^{N} D_i x_i' \left(\sum_{i=1}^{N} D_i x_i x_i'\right)^{-1}$$

$$= \left(\sum_{i=1}^{N} D_i x_i x_i'\right)^{-1}$$

$$\left(\sum_{i=1}^{N} D_i \ \mathrm{var}(u_i) x_i x_i' + \sum_{i=1}^{N} D_i x_i (Q_i' \ \mathrm{var}(\hat{\theta}) Q_i) \sum_{i=1}^{N} D_i x_i'\right)$$

$$\left(\sum_{i=1}^{N} D_i x_i x_i'\right)^{-1}$$

$$= \left(\sum_{i=1}^{N} D_i x_i x_i'\right)^{-1} (B_1 + B_2 \ \mathrm{var}(\hat{\theta}) B_2') \left(\sum_{i=1}^{N} D_i x_i x_i'\right)^{-1}$$

where $B_1 = \sum_{i=1}^{N} D_i \sigma^2 x_i x_i'$ and $B_2 = \sum_{i=1}^{N} D_i x_i Q_i'$. In the case at hand, we take u_i^2 as σ^2, where σ^2 is the variance we have estimated from the IPM with cross-sectional data (that is a Tobit model modified by the fact that we multiply the dependent variable by a an estimated probability). And V_θ is the asymptotic covariance matrix of θ. The asymptotic variance of θ is as follows

$$V_\theta = \left[\sum_{i=1}^{N} \frac{[\phi(\theta' z_i)]^2}{\Phi(\theta' z_i)[1 - \Phi(\theta' z_i)]} z_i z_i'\right]^{-1} \tag{A.7}$$

that is the inverse of the information matrix from the probit model, AMEMIYA [1981].

2 The Asymptotic Variance of the IPM with Panel Data Estimator

To calculate the correct variance-covariance matrix of the parameter estimates we have to take into account that: first, we have estimated in a first stage the probabilities of purchase and, second, we have obtained the structural parameter vector of the equation of interest using a two step procedure. In this procedure, we apply a minimum distance approach in the last step.

The minimum distance estimator is obtained by minimising $(\hat{\gamma} - R \cdot \varepsilon)'$ $W^{-1}(\hat{\gamma} - R \cdot \varepsilon)$ with respect to γ, where W is a positive definite matrix. The optimal choice for W corresponds to the variance $V(\hat{\gamma} - R \cdot \varepsilon) = V(\hat{\gamma} - \gamma)$. To get the consistent estimate W for the matrix $\hat{W}$ we need the influence function for $\hat{\gamma}_t$. Define $e_{it} = \Phi(z_i\hat{\delta}_t)y_{it} - z_i\hat{\gamma}_t$. The sample moment condition for $\psi = (\beta',\pi^{(1)\prime})'$ in the second step estimation procedure is

$$\frac{1}{N}\sum_{i=1}^{N} d_{it}\left\{\Phi(z_i\hat{\delta}_t)y_{it} - z_i\hat{\gamma}_t\right\}z_i = 0 \tag{A.8}$$

where d_{it} takes value 1 when y_{it} is observed. Expression (A.8) is the first order condition of a two stage *extremum estimator* with finite dimensional first stage parameters. The so called *delta method* yields[37]

(A.9)
$$\sqrt{N}(\hat{\psi}_t - \psi_t) =^p \frac{1}{\sqrt{N}}\sum_{i=1}^{N} E^{-1}(d_{it}z_iz_i')\{d_{it}e_{it}z_i + A_t\Lambda_{it}\} \equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\rho_{it}$$

where

$$A_t \equiv E(d_i\phi(z_i\hat{\delta}_t)y_tzz') \tag{A.10}$$

and $\phi(z_i\hat{\delta}_t)$ is the derivative of $\Phi(z_i\hat{\delta}_t)$ with respect to $\hat{\delta}_t$. From the probit estimation we observe that

(A.11)
$$\sqrt{N}(\hat{\delta}_t - \delta_t) =^p \frac{1}{\sqrt{N}}\sum_{i=1}^{N} I_{\delta_t}^{-1}\left\{z_i\frac{[d_{it} - \Phi(z_i\delta_t)]\phi(z_i\delta_t)}{\Phi(z_i\delta_t)[1-\Phi(z_i\delta_t)]}\right\} \equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Lambda_{it}$$

where I_{δ_t} is the probit information matrix for δ_t.[38]

In expression (A.9), $A_t\Lambda_{it}$ is the effect of the first stage probit estimation on the second stage. In (A.9) we have that the influence function for ψ_t is ρ_{it}. Define $\rho_t = (\rho_{1t},...,\rho_{it},...,\rho_{Nt})$ and $\rho \equiv (\rho_1',...,\rho_t',...,\rho_T')'$, then $W = E(\rho\rho')/N$. The T positive definite block-on-diagonal matrices in W are equal to $E(\rho_\tau\rho_\tau')/N$, for $\tau = 1,...,T$, respectively. These matrices are the corresponding variance-covariance matrices of the reduced form parameters for each wave of the panel. The $T(T-1)/2$ *distinct* block-off-diagonal matrices in W are equal to $E(\rho_\tau\rho_s')/N$, for the distinct combinations of panel waves we can have in a panel of length T and being $\tau \neq s$. These matrices are the variance-covariance matrices between the reduced form parameter estimates in two different waves. Estimates for all these matrices are obtained by replacing the parameters with their corresponding estimates and the expectations involved by their sample analogous.

Finally, with an estimate of W at hand whe can provide the closed form solution to the minimisation problem in

$$\hat{\psi} = (R'\cdot\hat{W}^{-1}R)^{-1}(R'\cdot\hat{W}^{-1}\hat{\gamma}) \tag{A.12}$$

where R is the cross-equation restrictions matrix. The asymptotic distribution for the minimum distance estimator is

$$\sqrt{N}(\hat{\psi} - \psi) \xrightarrow{d} N(0,(R'\cdot\hat{W}^{-1}R)^{-1}). \tag{A.13}$$

---

37. In the derivation of the asymptotic variance we follow LEE's [1996] approach for the two-stage extremum estimators with finite dimensional first-stage nuisance parameters.
38. Note that $=^p$ denotes convergence in probability.

DATA APPENDIX

The data used in this study is a sample of 12,400 observations (1,550 households during eight periods), from the *Spanish Family Expenditure Survey* (ECPF). Means, standard deviations and a glossary of definitions of the variables are presented in the following tables.

A.1 Glossary of Variable Definitions

Age

ag1: age of the head of the household,
ag2: age of the wife.

Household Composition

n1: number of children $\leqslant$ 4 years of age,
n2: number of children > 4 and $\leqslant$ 8,
n3: number of children > 8 and $\leqslant$ 13,
n4: number of children > 13 and $\leqslant$ 17,
na1: number of adults $\geqslant$ 18 and $\leqslant$ 24,
na1: number of adults > 24 and $\leqslant$ 64,
no: number of elderly people > 64.

Education of the Head of the Household

ed0: illiterate or no educational background,
ed1: primary education,
ed2: secondary education,
ed3: pre-university studies,
ed4: university studies.

Size of Town of Residence

drura: family is living in a town of less than 10,000 inhabitants,
dcity: family is living in a town of more than 500,000 inhabitants.

Educational Dummies

dana: head of the household is illiterate or has no educational background,
duni: head of the household has a university degree.

Labour Status

demploy: head of the household is working (full-time or part-time),
dunem: head of the household is unemployed,
dpensi: head of the household is pensioner,
dwwife: wife is working.

Occupational Dummies

bcollar:	head of the household is a "*blue collar*" worker,
wcollar:	head of the household is a "*white collar*" worker,
unskill:	head of the household is a unskilled worker,
highrank:	head of the household is in a high rank job.

Property Regime of the House of the Household

drent: house where the household lives is rented,
dseh: the household owns the house.

Quarter Dummies

q1: quarter 1,
q2: quarter 2,
q3: quarter 3,
q4: quarter 4.

Year Dummies

d85: year 1985,
d86: year 1986,
d87: year 1987,
q88: year 1988,
d89: year 1989,
d90: year 1990,
d91: year 1991.

Total Expenditure

lrtexp:logarithm of total real expenditure.

Price

lpcloth: logarithm of real price of clothing (deflated by index price of "*other non-durables goods*").

Family Composition Dummies

D1: =1 if n1 > 0, = 0 otherwise,
D2: =1 if n2 > 0 and n2 = 0, = 0 otherwise,
D3: =1 if n3 > 0 and n1 = n2 = 0, = 0 otherwise,
D4: =1 if n4 > 0 and n1 = n2 = n3 = 0, = 0 otherwise,
DA1: =1 if na1 > 0, = 0 otherwise,
DA2: =1 if na2 > 0 and na1 = 0, = 0 otherwise,
Do: =1 if no > 0, = 0 otherwise,
D2q: =1 if D2 or D3 = 1, = 0 otherwise,
D2qq: =1 if D1 or D2 or D3 = 1, = 0 otherwise.

A.2 Descriptive Statistics

	All Observations		*Positive Observations*	
Variable	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
Share adult clothing	0.086	0.090	0.104	0.089
ag1	46.244	11.131	46.787	10.976
ag2	43.469	11.366	43.959	11.192
n1	0.259	0.529	0.238	0.513
n2	0.261	0.516	0.232	0.488
n3	0.437	0.672	0.416	0.656
n4	0.256	0.501	0.277	0.516
na1	0.374	0.648	0.403	0.666
na2	2.043	0.544	2.054	0.543
no	0.146	0.407	0.151	0.413
lpcloth	0.073	0.015	0.073	0.015
drura	0.319	0.466	0.324	0.468
dcity	0.066	0.248	0.063	0.242
ed0	0.231	0.421	0.238	0.426
ed1	0.035	0.183	0.034	0.182
drent	0.127	0.333	0.123	0.329
dseh	0.085	0.279	0.090	0.286
demploy	0.789	0.408	0.790	0.407
dunem	0.065	0.246	0.061	0.240
dpensi	0.143	0.350	0.145	0.352
dwwife	0.194	0.396	0.197	0.397
bcollar	0.035	0.184	0.038	0.192
wcollar	0.393	0.488	0.389	0.487
unskill	0.118	0.323	0.116	0.320
highrank	0.023	0.151	0.024	0.152
lrtexp	12.829	0.486	12.876	0.473
lrtinc	12.814	0.486	12.830	0.483