ECONOMICS WORKING PAPERS SERIES

Economics Working Paper 26

# A Cross-Section Model with Zeros: an Application to the Demand for Tobacco*

Jaume Garcia†
and
José M. Labeaga‡

November 1992

UNIVERSITAT POMPEU FABRA

*Balmes, 132*
*Telephone (343) 484 97 00*
*Fax (343) 484 97 02*
*08008 Barcelona*
*e- mail econwp@upf.es*

## Abstract

The purpose of this paper is to emphasize the importance of the double-hurdle models in the estimation of tobacco demand equations. The data used in the empirical application are drawn from the Spanish Family Expenditure Survey. The paper shows that the Tobit, P-Tobit and first hurdle dominance models are restrictive and tends towards the existence of separate and non-independent individual decisions on participation and consumption. These results are confirmed by several tests for misspecification. Although bivariate normality is not completely fulfilled, we interpret that it is more an indication of over-rejection of the specification than non-normality.

## 1 Introduction

Microeconomic data sets offer important advantages for the analysis of consumer demand. The introduction of demographic variables into the specification of the equations accounts for the existence of heterogenous preferences among the individuals and allows explicit estimation of the effects on consumption of family size, occupation or other variables. In these databases we, generally, dispose of the consumption of commodities with a high disaggregated level, such as tobacco or alcohol. However, this could introduce a problem because we frequently observe that expenditure is zero for an important part of the sample. In some cases, absence of consumption arises from infrequence of purchase or false reporting. In others, the individual has decided not to consume. We may, therefore, wonder how to deal with the existence of zero observations for a frequently high number of consumers to get consistent estimates of the parameters.

Some papers in last years have tried to derive coherent models in order to deal with zero expenditures. Wales and Woodland (1983) develop it using Kuhn-Tucker (KT) conditions and Lee and Pitt (1986) by means of virtual prices, both in the context of complete demand systems assuming random preferences. In single equation applications, some authors try to model consumer behaviour, both for the making of decisions, as is the case of discrete choice models, and for those situations in which the dependent variable is truncated or censored. But, sometimes the standard specifications do not distinguish among the reasons that generate zeros and some generalizations of them had also been developed.

In this paper, we will deal with discrete choice, censored regression and alternatives to the censored regression models, as those of Cragg (1971), Amemiya (1984), Deaton and Irish (1984), Atkinson et al. (1984, 90), Blundell and Meghir (1986, 87) or Jones (1989), that take the decision making process into account with the aim of estimating an equation of demand for tobacco for the Spanish economy and derive the income and price elasticities. The data base used comes from the Spanish Family Expenditure Survey (EPF), a cross-section carried out between April of 1980 and March of 1981. In the estimation of the models, we pay special attention to the treatment of zero expenditures and we try to relax the observability and distributional restrictions. We will endeavour to analyze the factors which characterize smokers and the determinants of the quantity demanded.

The outline of the paper is as follows: in Section 2 we present a suitable economic framework for zero expenditures and corner solutions. In Section 3 we describe the statistical models and their relationship. The empirical results of this cross-section study are discussed in Section 4 where we also report an overview of the testing of the alternative specifications and the distributional assumptions, an evaluation of values of the income and price elasticities and some results concerning possible fiscal measures that the government could carry out, especially with the advent of the European Single Market Act. The paper ends with a summary of the main conclusions.

## 2 An economic model for the demand for tobacco and zero expenditures

In most of the applied work which deals with the problem of zero expenditures, the link between the economic model and the econometric specification has merely been to add random errors to the demand function in a way such that, with suitable hypothesis over the distribution of the disturbances, non-negativity is statistically imposed rather than derived from the optimization problem. Sometimes, this is not a good way of explaining the process generating zeros and, other times, we could derive the same econometric models through utility maximization.

Zero expenditures could arise for, at least, three reasons: too short recording periods that generate the well-known infrequency of purchase problem, misreporting or non-participation in the consumption of the good at any given value of prices and income i.e., corner solutions. The simplest approach to model zeros is to assume that the individual confronts the following optimization problem:

$$\underset{Y}{\text{Max}} \left\{ U(Y, \gamma) \;/\; P'Y = X \right\} \tag{1}$$

where Y is a vector of K commodities, P their corresponding price vector, X total expenditure and $\gamma$ a vector of unknown parameters. The individual maximizes a well-behaved utility function subject to the usual budget constraint. This leads to traditional demand functions.

The non-consumption of goods for some individuals is not introduced explicitly in (1) but could be considered as an ad hoc solution in the stochastic specification of the model, e.g. Tobit type models. However, it is more attractive to derive models in which the non-negativity constraints are not imposed. If we assume non-random preferences, the introduction of non-negativity constraints into the model supposes to change (1) to:

$$\underset{Y}{\text{Max}} \left\{ U(Y, \gamma) \;/\; P'Y = X,\; Y > 0 \right\} \tag{2}$$

These models have not, normally, been used in applied work due to the difficulties either in the solution of (2) or in the evaluation of multivariate integrals for computing the likelihood function (lf).

Assuming random preferences, we could incorporate the possibility of non-consumption into the structure of the model using the KT conditions as Wales and Woodland (1983) do. They derive the demand equations maximizing a stochastic direct utility function subject to the budget and non-negativity constraints. The problems of this direct approach are: first, it requires specification of the utility function and limits the use of some flexible demand systems and, second, it is difficult to deal with the KT conditions.

The approach of Lee and Pitt (1986) circumvents the above problems. The use of the indirect utility function (or cost function) allows us to express the restrictions in terms of prices rather than quantities using the virtual prices approach[1]. We can write the utility maximization problem as:

$$\underset{Y}{\text{Max}} \left\{ U(Y, \gamma, \varepsilon) \;/\; P'Y = X,\; Y > 0 \right\} \qquad (3)$$

where $\varepsilon$ is a K-vector of random errors.

We could define the solution of (3) as the notional demands ($Y_j^*$). They are latent variables in econometric terms. Their observed counterpart are the observed demands ($Y_j$) which correspond to the solution of the restricted optimization problem (2). We observe zero consumption of some goods in a way such that notional and observed demands are not the same. However, we could define a price vector so that, $Y_j = D_j(r^*)$; ($j = 1, \ldots, K$). $r$ is the vector of normalized prices $P/X$ and $r^*$ its corresponding virtual price vector. Neary and Roberts (1980) prove that these prices exist and are always positive with adequate hypothesis over the preferences. Their economic interpretation is the usual of a reservation price. An individual demands the good j if the market price equals the virtual price.

We are not analyzing the demand for K goods but only tobacco consumption. In this case, we can make a partition in the demand vector Y, say $Y_1$ tobacco and $Y_2$ the rest of the commodities. The possible demand regimes are: i) $Y_1 > 0$, $Y_2 > 0$; ii) $Y_1 = 0$, $Y_2 > 0$, (where we, obviously, discard the demand regime which corresponds to the non-consumption of the rest of the goods, except tobacco). The statistical structure which derives from this interpretation is no more than the usual Tobit model (Tobin, 1958).

The third form of dealing with zero consumption assuming random preferences is through the existence of a discrete random preference scheme. We suppose that the individuals have the same preferences as in the previous cases. We also assume that their utility function can be expressed as:

$$U = I^* V_1(Y_1, Y_2, \gamma, \varepsilon) + (1 - I^*) V_2(Y_2, \gamma, \varepsilon) \qquad (4)$$

---

[1] The concept of virtual price is due to Rothbarth (1941). The derivation of the model is based on the theory of consumer demand under rationing. See, for example, Deaton (1981).

being $V_1$ the utility function representing the preferences of smokers or potential smokers and $V_2$ that of non-smokers, $Y_1$, $Y_2$, $\gamma$ and $\varepsilon$ are defined as above and $I^*$ is the indicator of participation in the consumption (discrete preference parameter).

For non-smokers, $I^* = 0$, so $Y_1$ does not affect the preferences of non-smokers as a result of their rational choice. For smokers (or potential smokers), $I^* = 1$ and their optimal consumption is the solution of problem (2). The solution to this problem can be modeled by a Tobit and, the assumptions over the discrete preference parameter lead us to specifications like those proposed by Cragg (1971). The second process picks up a typical corner solution and, the first, those zeros which could arise for economic or non-economic decisions i.e., non-participation in consumption. It is possible to derive several structures for the model, assuming different hypothesis over the distribution of the discrete preference parameter[2].

We are not interested in this paper in the estimation of a complete demand system but a single equation, but we wish to make use of the theoretical advantages of the system to which this equation belongs. Deaton and Muellbauer (1980) propose a flexible functional form for the cost function which leads to the Almost Ideal Model (AIM). In this demand system, the variables to be explained are the expenditure shares which are related linearly with the logarithm of prices and total real expenditure. This relation is deduced on the basis of the optimization of a cost function that is also linear in its logarithmic form. In order to define an analogous model with random preferences and derive the notional demand functions, let us start off from a cost function in logarithmic form:

$$\text{Log } C(u, P, \gamma, \varepsilon) = \alpha_0 + \sum_j \left[\alpha_j + \varepsilon_j\right] \log p_j + \frac{1}{2} \sum_j \sum_k \xi^*_{jk} \log p_j \log p_k + u \beta_0 \prod_j p_j^{\beta_j} \tag{5}$$

where, u is the utility level, P the price vector, $\gamma$ is a vector of unknown parameters and $\varepsilon$ a vector of random components. $\alpha_j$, $\beta_j$ y $\xi^*_{jk}$ are parameters.

---

[2] For instance, the P-Tobit model of Deaton and Irish (1984), the generalized sample selection model of Heckman (1979) or this same model applied to tobacco consumption that is the first hurdle dominance model of Jones (1989).

To make C(u, P, $\gamma$, $\varepsilon$) homogeneous in price, it is necessary that:

$$\sum_j \alpha_j = 1, \sum_j \beta_j = 0 \quad y \quad \sum_j \xi^*_{jk} = 0 \qquad (j, k = 1, \ldots, K) \qquad (6)$$

In addition, given that the sum of shares is unity, we need a restriction for the random components $\sum_j \varepsilon_j = 0$.

We could be derived the shares for each commodity from the first order conditions of minimization as a function of prices and the utility level. After substituting u, we obtain the demand equations:

$$Y^*_j = \alpha_j + \sum_k \xi_{jk} \log p_k + \beta_j \log \left(\frac{X}{P}\right) + \varepsilon_j \qquad (7)$$

where $p_k$ is the price of good k, X is total expenditure, $\xi_{jk} = 0.5(\xi^*_{jk} + \xi^*_{kj})$ and P is the price index defined as:

$$\text{Log } P = \alpha_0 + \sum_j \left[\alpha_j + \varepsilon_j\right] \log p_j + \frac{1}{2} \sum_j \sum_k \xi_{jk} \log p_j \log p_k \qquad (8)$$

A problem appears in expression (8) with the introduction of the random components, but in the empirical application we can substitute Log P by some previously-selected fixed price index[3] and express the AIM through K linear equations in prices and total expenditure. Although this analysis was originated in the rationing literature where the derivation of the virtual prices is difficult, as Deaton (1981) pointed out, here they can be solved easily from the specific demand regimes either in the univariate Limited Dependent Variable (LDV) models or in the bivariate models, where we also have to look at the sign of the discrete preference parameter to establish the switching conditions (see Lee and Pitt, 1986).

Using the above considerations, and assuming that the prices which the individuals face with are the same, it could be thought that the parameters in the demand equations will be constant for all individuals, except that, following Pollack and Wales (1981), we allow the socioeconomic characteristics to affect the cost function multiplicatively and, as a result, affect the share additively through

---

[3] In the estimations we replace Log P by the log of the Retail Price Index.

the parameter $\alpha_j$[4]. On the other hand, the stochastic terms permit the existence of unexplained differences in tastes or preferences.

## 3 Econometric treatment and hypothesis testing

Our main concern throughout this paper will be the estimation of an equation of demand for tobacco in Spain using cross-section data from the EPF. The particular characteristics of the good we are analyzing make us pay special attention to the reasons why zero expenditures arise. With the availability of microdata there are two possibilities in analyzing the behaviour of the individuals in the decision making process: first, why one decides to be a smoker and second, the quantity one decides to consume.

Let us suppose that from the optimization problem we derive an equation which relates tobacco consumption to the explanatory variables through the following demand function and an observation rule:

$$Y_i^* = \beta' X_i + \varepsilon_i \qquad (9)$$

$$I_i^* = \alpha' Z_i + v_i \qquad (10)$$

where $Y_i^*$ is the value which corresponds to the latent variable (notional demand), $I_i^*$ is a non-observable variable which determines whether the individual i is a smoker or not (discrete preference parameter), $X_i$ and $Z_i$ are vectors of conditioning variables (economic and socio-demographic characteristics of individual i) and $\varepsilon_i$ and $v_i$ are non-observable random variables. We omit the j-subscript because we are now only analyzing a single equation.

Discrete choice models allow us to analyze situations in which only the decision to participate is considered. We could determine the probability that an individual belongs to each of the groups considered (smokers non-smokers in our case), being the observation rule $Y_i = 1(I_i^* > 0)$, where 1(A) is the indicator function of event A. The lf for this model is easily derived and we do not repeat it here.

The second aspect to consider is the analysis of the quantity demanded. A commonly used specification when dealing with individual data with censored problems is the Tobit model. The censoring mechanism for an

---

[4] It is assumed that the family characteristics have the same impact on the share of expenditure of tobacco regardless of the level of income and prices.

equation such as (9) which relates notional with observed demands is $Y_i = \max (Y_i^*, 0)$, that acts: whenever $Y_i$ is not observed it is replaced by zero, otherwise it is observed and replaced by its value. So, zero expenditures arise if the household does not purchase the good, but we do not know the specific reason for this. In other words, since the same relation and, consequently, the same factors determine whether one decides to smoke and if so how much, zero observations under the Tobit interpretation would correspond to virtual prices for the good, which will be lower than their market value. Consequently, this is a typical corner solution. Sometimes, it is not a valid (or unique) argument for zeros. The lf for this model is:

$$L_T = \prod_1 P\left(\varepsilon_i > -\beta' X_i\right) \; f\left(Y_i / \varepsilon_i > -\beta' X_i\right)$$

$$\prod_0 \left[1 - P\left(\varepsilon_i > -\beta' X_i\right)\right] \qquad (11)$$

where $\varepsilon_i$ is normally distributed with zero mean and variance $\sigma_\varepsilon^2$. f is the pdf corresponding to the normal random variable and $\prod_1$ and $\prod_0$ denote the product over positive and zero observations, respectively. We obtain consistent estimates of the parameters of the model ($\beta$, $\sigma_\varepsilon^2$) maximizing (11)[5].

The most common source of zeros in demand analysis is probably infrequency of purchase, but none of the above models take it into account. This type of model is dealt with by using the probability of making a purchase to link observed expenditure to underlying consumption and to introduce in this form a source of censoring. The P-Tobit model by Deaton and Irish (1984) is a good example but, it is more applicable to the durable good demand. In the case of tobacco infrequency is, probably, a minor problem. The main problem may be the non consideration of goods by some individuals, in such a way that they do not consume them at any price. It is possible to treat this decision under model (9)-(10) by introducing a new scheme of censoring, once again derived from the maximization of a random utility function such as (4), $Y_i = 1 (I_i^* > 0) Y_i^*$, in which positive expenditure is observed only after two decisions of the individual, first he wishes to participate ($I_i^* > 0$) and, second, he really participates ($Y_i^* > 0$). This, and other related structures, were proposed by Cragg (1971), who

---

[5] The solution of the first order conditions system of equations for the discrete choice and Tobit models can easily be obtained by means of any of the algorithms available in standard statistical-econometric packages.

called them double-hurdle models, although he did not draw them from the formal choice theory. Atkinson et al. (1984, 90), Blundell and Meghir (1987), Blundell et al. (1987, 89) and Jones (1989) are examples of its use in some interesting economic fields such as tobacco and clothing demand or female labour supply.

The reasons for the separation of these decision processes are: first, an individual may be a non-smoker and values of the exogenous variables (price, income, etc) will therefore not exist for which the consumer may purchase and, second, the individual could be a potential smoker, but for certain levels of the relevant variables he may decide not to consume. We can, therefore, suppose a group of variables which influence the decision to smoke or not and another group (there could be variables common to both) which determines the quantity that a potential smoker will eventually consume. So, it is clear that two 'hurdles' must be overcome before observing a positive consumption.

The variety of models derived depends on the type of assumptions on the joint distribution of the error terms entering equations (9) and (10) and the dominance concept introduced by Jones (1989). Under the hypothesis of independence between $\varepsilon_1$ and $v_1$, we have the double-hurdle model applied by Cragg to the demand for durables. In tobacco demand analysis this model has been used by Atkinson et al. and Jones with two different data sources for the British economy. But, given the relationship which exists between the two processes of decision making carried out by individuals, it seems adequate to think that the unobserved factors in both equations could generate not independent errors, so we could suppose that $(\varepsilon_1, v_1)$ is distributed as a bivariate normal random variable with zero means, unit variances and coefficient of correlation $\rho$. This is the model we wish to test against the Tobit, P-Tobit and independent double-hurdle specifications. We can write the lf for the double-hurdle independent model as:

$$L_{DH} = \prod_1 P\left( v_1 > -\alpha' Z_1 \right) P\left( \varepsilon_1 > -\beta' X_1 \right) \; f\left( Y_1 / \varepsilon_1 > -\beta' X_1 \right)$$

$$\prod_0 \left[ 1 - P\left( v_1 > -\alpha' Z_1 \right) P\left( \varepsilon_1 > -\beta' X_1 \right) \right] \qquad (12)$$

If we relax the assumption of independence, the lf becomes:

$$L_{DHD} = \prod_{1} P\left( v_1 > -\alpha' Z_1 \right) \; P\left( \varepsilon_1 > -\beta' X_1 \; / \; v_1 > -\alpha' Z_1 \right)$$

$$f\left( Y_1 \; / \; \varepsilon_1 > -\beta' X_1 \; , \; v_1 > -\alpha' Z_1 \right) \qquad (13)$$

$$\prod_{0} \left[ 1 - P\left( v_1 > -\alpha' Z_1 \right) P\left( \varepsilon_1 > -\beta' X_1 \; / \; v_1 > -\alpha' Z_1 \right) \right]$$

where all the terms are defined as before. The maximization of (12) and (13) give us, again, consistent estimates of the vector of parameters[6]. Equation (12) is constructed on the basis of two unidimensional random variables, but in order to optimize (13) we need to evaluate a bivariate normal distribution.

Independence is normally assumed to simplify the likelihood expressions, although, another possibility of different and easier statistical structures for (9) and (10) is the dominance concept. First hurdle dominance implies that participation dominates consumption, that is, once an individual decides to smoke, consumption takes place and he is not observed at a corner solution. The statistical implication of this concept is that $P\left(Y_1^* < 0 / Y_1^* = 1\right) = 0$ and, as a result, the lf corresponds to that of Heckman's generalized sample selection model:

$$L_H = \prod_{1} P\left( v_1 > -\alpha' Z_1 \right) \; f\left( Y_1 \; / \; v_1 > -\alpha' Z_1 \right)$$

$$\prod_{0} \left[ 1 - P\left( v_1 > -\alpha' Z_1 \right) \right] \qquad (14)$$

Complete dominance simplifies even further the model so that first hurdle dominance and independence are assumed. In this case, the model can be estimated separately, a Probit for the participation and OLS for the consumption equations. In this case, the lf takes the form:

$$L_D = \prod_{1} P\left( v_1 > -\alpha' Z_1 \right) \; f\left( Y_1 \right) \; \prod_{0} \left[ 1 - P\left( v_1 > -\alpha' Z_1 \right) \right] \qquad (15)$$

---

[6] The maximum likelihood estimation of the parameters of these models and those of the Tobit model have been carried out using the subroutine of optimization E04LBF of the NAG Library, which requires provision of first and second derivatives of the likelihood function. For the latter, we make use of the approximation proposed by Berndt et al. (1974).

Among the above models we can establish immediate relationships. A first approach to set for the adequacy of the Tobit model is to compare the results of the Probit with those of the Tobit divided by the estimated standard deviation. These coefficients need to be the same if the Tobit is a good specification to explain the reasons generating zeros. If it is not true, then this is an informal diagnostic that the process generating zeros are not gathered adequately by the Tobit specification. On the other hand, given that the lf of the Tobit model is no more that the sum of the lf's of the Truncated and Probit models, we could carry out a test over the adequacy of the Tobit comparing, by means of an LR test, the values of the function of the three models. It is also possible to use a Hausman type test (Ruud, 1986).

On the other hand, the double-hurdle independent is nested within the dependent version of the model. Under independence, the probability of consuming is not affected by the condition of being potential smoker and the conditional pdf reduces to the marginal pdf for observed shares, so (13) reduces to (12). An LR test, which has a $\chi^2$ distribution with one degree of freedom, could be carried out for testing the independence assumption. We could also check this hypothesis conducting a Hausman test. An alternative form of taking into account the decision making process is the infrequency of purchase model reckoned by Deaton and Irish. They consider the probability of being a smoker constant for all individuals (obviously for identification reasons). They also assume independence between the two random terms, although in this model dependence only changes the constant. So, this specification is nested within the double-hurdle model. Under the null, we have parameters that are not identified and the LR test has not a $\chi^2$ distribution.

There exists another relevant sequence of tests from the dependent double-hurdle to the Tobit model. It should be noted that the double-hurdle independent model simply supposes a generalization of the Tobit model. It is important to emphasize that if $F(\alpha' Z_1) = 1$, zero observations only appear in the decision about the quantity to consume and, as a result, the first hurdle is irrelevant. Therefore, it is important to test whether the slopes of the participation equation are zero. In this last diagnostic and the LR between the double-hurdle independent and the Tobit model, some of the parameters under the null are not identified. As a result, the tests have not a $\chi^2$ distribution, except conditional on fixed values of the $\alpha$'s, and the LR is no more than an intuitive diagnostic.

We can also check independence together with dominance. With dominance, the independence assumption implies that the marginal pdf is the same that the pdf conditional on positive observations and (14) reduces to (15). We could compare these equations by means of an LR test which is distributed as a $\chi^2$ with one degree of freedom. The existence of sample selection implies that complete dominance is restrictive. It could be checked by the t-test of the correction parameter (Heckman, 1979) or an equivalent Lagrange Multiplier (LM) diagnostic developed by Melino (1982). Finally, the fact that $F(\beta' X_1/\sigma_\varepsilon) \neq 0$ makes the dominance

assumption restrictive.

Although tobacco is not, probably, a good for which infrequency of purchase is a great problem, we would like to compare the P-Tobit and Tobit models. The lf of the P-Tobit model simplifies in two cases: first, if the zeros are only caused by the binary censor OLS for $Y_1 > 0$ are ML, consistent and fully efficient and, second, if the binary censor does not act the lf of the P-Tobit reduces to (11). We can check this hypothesis conducting a score test developed by Deaton and Irish that allows us the comparison without estimating the model under the alternative. It has again a $\chi^2$ distribution.

In addition to the set of diagnostics carried out for model specification, one important matter remains; it has to do with the testing of the restrictions which are assumed for the distribution of the random errors. The validity of the results in LDV models depends on the fulfillment of the normality and homoscedasticity hypothesis. We conduct LM tests in order to estimate the model only under the null. The normality tests in the Probit and Tobit models are based on Bera et al. (1984). In order to establish the alternative hypothesis, they make use of the fact that the normal is nested within the Pearson family of distributions. The bivariate normality tests are based on Lee (1984). In this case, under the alternative, the joint and marginal distributions are of the Gram-Charlier type. Finally, the heteroscedasticity tests are based on Blundell and Meghir (1986).

## 4 Empirical application

*Data and variables*

It is obvious that, even in spite of the statistical advantages of microeconomic data, we encounter certain shortcomings which may arise by the type of data or the nature of the good we are trying to study and cannot be ignored. First of all, the problem of non-response. Secondly, the survey only considers households or family dwellings and excludes collective dwelling-places (such as hospitals, prisons, barracks, hotels, etc.). As a result, the expenditure data is in relation to households and not to individuals. Thirdly, the collection of information only covers a seven-day period, with a 41.2% zero observations in tobacco. Fourthly, problems of not revealing information about goods such as the one we are dealing with, or alcoholic drinks, could exist in the information gathered.

The endogenous variable in all the models estimated is the share of tobacco consumption. It represents 1.2% of total expenditure as an average for all individuals in the sample, and 2.1% if we only take the subsample of consumers with positive observations into account. According to National Accounting data, tobacco consumption represented 1.2% of private national expenditure in 1980.

As explanatory variables, we weigh up the effect of income by means of

the logarithm of the total real expenditure (LINC). It is possible to pick up price effects, despite the fact that the EPF covers a time span of one year, so that we can distinguish between the data collected according to the quarter (and even the week) when the interview takes place. Although the lack of variation in this variable for such a short period of time could be put forward as an argument, it is necessary to examine Table A.2 in the Data Appendix in order to rule out this possibility. This variable is defined in terms of the logarithm of the relative price, i.e. the ratio of the price index of tobacco to the retail price index (LPRICE). We are assuming weak separability between tobacco and other goods, although the inexistence of direct substitutes for tobacco permits the only introduction of the own price. The variability with which this variable is used makes the price to pick up season effects. So, we could think in LPRICE as a group of quarterly dummies with restrictions.

Regarding the rest of the variables of the X or Z vectors of equations (9) and (10), it should be pointed out that most of them can only be defined for the head of the household. The personal characteristics used are occupation (NONMAN and FORCES), employment situation (UNEM and NEARN), education (ED1 to ED5), size of the town of residence (DM1 to DM5), region of residence (REG2 to REG17), age and age squared (AGEHH and AGEHH2) and household size (NA1, NA2 and NA3). The means and standard deviations of the variables, for the whole sample and the subsample of consumers, are shown in the Data Appendix.

*Results and diagnostics*

Tables 1 and 2 contain ML estimations of the Probit, Tobit and double-hurdle models. In addition, we present the consumption equation corresponding to the complete dominance model. The Probit is presented with a view to making comparisons as an intuitive test of the adequacy of the Tobit model, so we do not comment on the results. In the consumption equations, we observe that price rises contribute to an increase of the share of expenditure on tobacco (ceteris paribus). Likewise, rises in income reduce the share spent on tobacco (except in the first hurdle dominance model), although the values make the quantity consumed fall in the former case and rise in the latter.

The presence of additional contributors to household income, and the fact that the head of the household is unemployed raises the probability of smoking and the share of tobacco consumed. If the head of the household has a non-manual occupation it is reduced. The coefficient which corresponds to members of the armed forces has a positive sign in every case, but the t-Student test indicates that the fact of being able to accede to the same goods at different prices has a limited importance in a rise in the probability of smoking and in the share consumed.

Even when the family size variable is expected to influence positively the probability and the share, its disaggregation allows us to verify that it is in the case of younger family members where the effect is

more pronounced. Proof of this can also be found in the fact that the coefficient which corresponds to the age of the head of the household behaves in a negative and parabolic way; i.e. the greater the age, the smaller the share of expenditure, and the more the age rises, the more the share falls.

Regarding education, it is to be expected, that the higher the educational level of the head of the household, the lower will be the probability of smoking. This circumstance is reflected in the fact that variables ED4 and ED5 (individuals with pre-university or university studies) have a negative sign, while for illiterate individuals or those without an educational background or with only primary level studies (ED1 and ED2) the coefficient is positive. It actually happens, except for the double hurlde independent model.

In the participation equation, we have assumed, in accordance with the beliefs of Atkinson et al. and Jones and because of purely subjective issues that those variables related to education and occupation, in addition to those of income and age could be important in the decision to be smokers. The inclusion of these variables can be justified mainly by sociological factors, such as habit (income and age), restrictions to smoke for security or hygiene reasons (occupation) and information on damages or health risks (education and age). The non-introduction of other variables as determining factors in the decision to be a smoker has found support in intuitive and empirical aspects, because we have proved alternative specifications, even dropping out some of the ones previously mentioned from the first hurdle[7]. In a first step, we implement a model in which the random term in the participation equation is distributed normally and independently from the disturbance in the second hurdle. We estimate another version of the double-hurdle model in which the restriction of independence is not imposed.

We can see, with reference to the variables introduced at the first hurdle, that income positively influences the decision to participate and thus a rise in consumer income will produce a rise in the probability of an individual being a potential consumer. The effects of the educational variables show that individuals of a higher level of education are less likely to be smokers than those with low levels of education. Given the arguments advanced in favour of the introduction of the age variable, we could expect, a priori, a reduction in the probability of being a smoker with a rise in age, as happens. No great empirical evidence is provided by the models to support the introduction of the occupational dummy.

Although the importance of total expenditure in determining both the

---

[7] We have examined several alternative specifications for the first hurdle. Among them, it is important to emphasize that in both versions of the double-hurdle model, the exclusion of the income variable significantly retards the convergence of the lf. In some cases, it is impossible to reach the convergence criteria.

probability and the share, the introduction of the variable according to different intervals does not produce any important change in its coefficient. It does not affect the results for the other estimates either. On the other hand, the coefficients which are clearly defined in the Tobit model (those which present t-Student statistics over 4) generally stay quite stable in the estimation of the consumption equation of the double-hurdle models, while for income, price and education variables, the differences are pointed out. The consideration of two processes of decision rather than one and the high degree of correlation between income and education and, also, between the errors are, probably, explanation for these differences.

The intuitive diagnostic which compares the coefficients of the Probit model and the ratio $\beta/\sigma_{\varepsilon}$ which corresponds to the Tobit model leads us to think that alternative specifications could better bring together the reasons why zero observations arise. The formal LR test, among the Tobit, Probit and Truncated models, also confirms the inadequacy of this specification (see Table 3). The Tobit is, also, rejected vs. the P-Tobit. Although we do not include the results, they show a estimated value of p out of the interval [0, 1] and tending to infinity. The square root of the numerator of the score test proposed by Deaton and Irish is positive (1,816.9) and thus, the rejection of the Tobit is in the opposite direction to that predicted by the test indicating, probably, misspecification of the first hurdle and the existence of another source of censoring additional and different to that of the Tobit and P-Tobit. It is important to notice that given the significance of some of the variables in the first hurdle, the P-Tobit model as well as the one in which only a constant appears in the first decision are restrictive as the LR test carried out confirms. On the other hand, given that the binary censor do not explain all the zeros, OLS with positive observations is not consistent.

Although we get a more significant improvement when moving from the Tobit to the double-hurdle independent model, the restrictiveness of the independence hypothesis is confirmed by the LR and Hausman tests. In addition, the value (and t-ratio) of the coefficient of correlation allows to reject the restricted model and it reveals, as expected a priori, a high positive correlation between the errors.

The significance of the correction parameter and the LR and LM tests in the first hurdle dominance model, indicate that the participation decission cannot be neglected and that zeros affect the behaviour of smokers (and so rejects the univariate and the complete dominance models). The magnitude of the price parameter reflects the great importance of the habit for consumers, but neither the coefficient of total expenditure in the second hurdle is estimated efficiently nor it has the correct sign, probably because we are ignoring one possible source for zeros.

We have already mentioned that P-Tobit and double-hurdle with only a constant in the first hurdle are restrictive models. Given the endogeneity of the participation equation, the complete dominance model

is also restrictive. There will, therefore, be a percentage of non-smokers supplied by the probabilities estimated for the first hurdle and another percentage of potential smokers whose consumption is zero when they look at the values of price, income and other variables and whose probabilities are determined in the second hurdle. Therefore, it is important to answer the question of how important is the percentage of individuals that do not surpass the first hurdle (to be able to reject the univariate models) and how important is the percentage of potential smokers that do not consume (to be able to reject the dominance assumption). Table 4 brings together these values.

The probability of not being a smoker for the independent model is 36.23% for a household in the base situation, with 13.61% of the individuals falling at the first hurdle. These values become 31.62% and 17.75% for the dependent model, so the probability of being a potential smoker without actually consuming are 22.52% and 13.87% respectively. Where the head of the household is 25 years old, the probability of consuming rises 22% and 16%, respectively, with respect to the base situation. There is a 7% the probability of falling at the first hurdle. Among those individuals of 65, the probability of consuming falls to 60% and 40%, and in this group, 65% and 45% are non-smokers. The fall in the probability of smoking when passing from lower to higher educational levels is around 20%. The probability of not being smoker rises between 30% and 40% when we compare the latter group with the former. Educational background is therefore an important factor in the two decision making process of individuals. These results show that the generalized sample selection or first hurdle dominance model is not adequate, given the value of the probability of observing a zero expenditure once the individual has decided to be a potential smoker. This adds additional empirical evidence to the above comments.

The most suitable specification for the explanation of consumer behaviour regarding demand for tobacco seems to be the unrestricted bivariate model (from the point of view of the economic results and specification testing). However, in order to confirm these results we have to look at the tests for the distribution of the errors. We can observe that the equation which takes into account the smoking decision perform well against the normality tests. The consumption equation presents problems of skewness and heteroscedasticity. Also when we allow for correlation in the double-hurdle model, the assumption of bivariate normality is not completely fulfilled. However, it is probably that the diagnostics on these models are more an indication of over-rejection of the specifications than non-normality. To solve the problem of skewness, the Symmetrically Censored Least Squares estimator proposed by Powell (1986) could be adequate but, it is no more than an alternative for the Tobit model. We could think in semiparametric alternatives as those of Newey et al. (1990) for the bivariate models.

*Elasticities and effects of changes in prices*

The estimated values of the elasticities together with their standard errors are reported in Table 5. These elasticities have been evaluated

under two different assumptions: a fixed effects and a random effects hypothesis (Atkinson et al., 1989). These two interpretations do not produce significant changes in the values in any of the models. However, we find important differences when comparing the results of the Tobit, generalized sample selection, double-hurdle independent and double-hurdle dependent models. The results for these last models seem to be more in accordance with those of other authors that use microdata.

In addition, we will try to analyze the reactions of consumers to variations in the price of tobacco. It is different to evaluate the changes derived by a modification of any of the explanatory variables either for the unconditional or conditional expected values of the dependent variable and, as a consequence the implications of changes in exogenous variables such as prices and income. These effects can be broken down into two different forms: first, it could change the quantity above the limit (or the share) weighted by the probability of smoking and second, it could modify the probability of smoking weighted by the expected value of the endogenous variable conditional on positive observations[8].

Let us suppose that the government introduces a 25% rise in the real price of tobacco[9]. The effects of this rise on the quantity demanded for different types of individuals are shown in Table 5. This change in price affects the quantity of tobacco demanded in a very different way depending on the type of household under consideration. The reduction is greater for those households where the head is unemployed or in families with young employed and unemployed heads than in the base situation and it is more important in the case of those households whose head is illiterate or has no educational background than in the other cases. It also affects much more households with low levels of income. These reductions range form 0% to 18% and from 9.28% to 16.70% for a household in the base situation. Therefore, price rises, for example by means of an indirect tax, could be, in some cases, an effective tool in reducing the quantity of tobacco consumed.

## 5 Concluding comments

In this paper, we have estimated several models to explain the demand for tobacco using data from the 1980-81 Spanish EPF. The main aspect from both the economic and statistical viewpoints lies in the explanation of the reasons why zero expenditures arise. First, we try

---

[8] In the case of the double-hurdle models the probability of being a smoker could also change. For the applications to the Tobit model, see McDonald and Moffit (1979) and Maddala (1983).

[9] The average increment in the price of tobacco prior to 1993 is approximately 25%. Since we assume a rise in the real price, we are also assuming that the Retail Price Index remains constant.

to identify the reasons why individuals decide to be smokers or not, then we are worried about the quantity to consume. The problem in the Tobit model is that the same effects are assumed to exist for variables explaining probabilities and shares. To overcome this problem, we propose the estimation of models in which the individuals take on two decisions; one equation in which they consider to be smokers or not, and another equation in which those who have decided to be smokers choose the quantity, with and without a restriction of independence between the errors.

We have carried out tests for specification and distributional assumptions in the estimated models. The relaxation of the observability assumptions produces an improvement in the value of the log likelihood when moving from the Tobit to the double-hurdle independent model. The fact of not imposing the restriction of independence between the error terms, although less important, also improves the values of the If and the LM tests. But, the fulfillment of the distributional assumptions is not satisfactorily reached in the consumption equation and, probably, it is an indication of over-rejection of the specifications more than non-normality.

In the economic respect, variables such as income and education are importantin deciding to be a smoker or to consume tobacco. The income effects are different in both decisions and, so the univariate models do not appear to be suitable for registering consumer behaviour in the decision making process. In the case of the educational variables, we can observe that the higher the level of education of the head of the household, the higher the consciousness of the risks of smoking and lower the probability of smoking and the share of expenditure on tobacco. Regarding the price variable, we have assumed that it is not relevant in deciding the probability of smoking, but it is an important factor in the consumption equation.

As a last step, we have considered the effectiveness of the fiscal policy to reduce tobacco consumption. We have two objectives in mind. First, the interest of the authorities in implementing certain health and publicity measures to mitigate the social costs of tobacco smoking. Second, the imminent entry of the European Single Market Act. The results lead us to conclude that the effects of a rise in the price of tobacco over the reduction of the quantity demanded (or the probability of consuming)are more important than might be expected for an addict good.

**Table 1. Estimated coefficients for univariate models**

| Explanatory variable | OLS ($Y_i>0$) | Probit | Tobit |
|---|---|---|---|
| CONSTANT | - 0.1247 | - 2.2773 | 0.0224 |
| | (47.5) | (13.6) | (7.12) |
| LPRICE | 0.0113 | - 0.2395 | 0.0038 |
| | (10.2) | (3.48) | (2.73) |
| LINC | - 0.0114 | 0.3105 | - 0.0017 |
| | (39.9) | (18.4) | (5.30) |
| UNEM | 0.0023 | 0.2484 | 0.0063 |
| | (3.62) | (5.68) | (8.43) |
| NONMAN | - 0.0001 | - 0.0258 | - 0.0010 |
| | (0.19) | (1.07) | (2.17) |
| FORCES | 0.0008 | 0.1106 | 0.0011 |
| | (0.62) | (1.28) | (0.60) |
| NA1 | 0.0033 | 0.2747 | 0.0067 |
| | (15.3) | (19.3) | (25.6) |
| NA2 | 0.0017 | 0.2411 | 0.0058 |
| | (6.88) | (16.5) | (21.2) |
| NA3 | - 0.0001 | 0.1720 | 0.0040 |
| | (0.41) | (10.1) | (12.1) |
| AGEHH | - 0.0005 | - 0.0072 | - 0.0002 |
| | (6.65) | (1.77) | (2.40) |
| AGEHH2/100 | - 0.0004 | - 0.0055 | - 0.0002 |
| | (5.74) | (1.45) | (2.00) |
| RG2 | - 0.0046 | - 0.3602 | - 0.0097 |
| | (6.14) | (8.25) | (10.7) |
| RG3 | - 0.0034 | 0.0794 | - 0.0016 |
| | (4.05) | (1.40) | (1.56) |
| RG4 | - 0.0026 | - 0.0592 | - 0.0024 |
| | (2.48) | (0.91) | (1.87) |
| RG5 | - 0.0046 | - 0.1783 | - 0.0057 |
| | (5.59) | (3.61) | (5.86) |
| RG6 | - 0.0052 | - 0.0184 | - 0.0049 |
| | (5.51) | (0.29) | (3.29) |
| RG7 | - 0.0045 | - 0.0830 | - 0.0046 |
| | (8.62) | (2.49) | (7.07) |
| RG8 | - 0.0043 | - 0.3122 | - 0.0080 |
| | (7.50) | (8.91) | (11.1) |
| RG9 | - 0.0023 | - 0.0352 | - 0.0022 |
| | (2.93) | (0.72) | (2.40) |
| RG10 | - 0.0056 | - 0.4558 | - 0.0115 |
| | (8.18) | (11.6) | (14.4) |

| | | | |
|---|---|---|---|
| RG11 | - 0.0030 | - 0.1806 | - 0.0049 |
| | (3.88) | (3.75) | (5.02) |
| RG12 | - 0.0038 | - 0.0500 | - 0.0037 |
| | (6.31) | (1.31) | (5.07) |
| RG13 | - 0.0003 | - 0.1918 | - 0.0033 |
| | (0.26) | (2.93) | (2.73) |
| RG14 | - 0.0039 | - 0.3174 | - 0.0082 |
| | (3.32) | (4.34) | (4.95) |
| RG15 | - 0.0051 | - 0.1620 | - 0.0062 |
| | (4.17) | (2.17) | (3.72) |
| RG16 | - 0.0039 | - 0.1168 | - 0.0046 |
| | (6.49) | (3.05) | (5.94) |
| RG17 | - 0.0045 | - 0.3318 | - 0.0087 |
| | (6.27) | (7.54) | (9.06) |
| DM1 | - 0.0016 | - 0.0200 | - 0.0011 |
| | (3.15) | (0.66) | (1.79) |
| DM2 | - 0.0015 | - 0.0268 | - 0.0003 |
| | (3.45) | (1.02) | (0.66) |
| DM3 | - 0.0011 | - 0.0405 | - 0.0013 |
| | (2.73) | (1.61) | (2.47) |
| DM5 | 0.0007 | - 0.0013 | - 0.0003 |
| | (1.17) | (0.04) | (0.40) |
| ED1 | 0.0049 | - 0.0664 | 0.0007 |
| | (6.93) | (1.70) | (1.07) |
| ED2 | 0.0007 | 0.0342 | 0.0011 |
| | (1.78) | (1.47) | (2.39) |
| ED4 | 0.0002 | - 0.00007 | - 0.0004 |
| | (0.40) | (0.00) | (0.63) |
| ED5 | 0.0006 | - 0.2204 | - 0.0036 |
| | (1.03) | (5.99) | (4.31) |
| NEARN | 0.0015 | 0.0512 | 0.0015 |
| | (6.71) | (1.20) | (1.64) |
| $\sigma_u$ | 0.0165 | --- | 0.0243 |
| Log-Likelihood | -26,678.3 | -14,288.0 | -39,912.0 |
| $R^2$ / Pseudo-$R^2$ | 0.18 | 0.11 | 0.11 |
| % correctly predicted | --- | 67.86 | 67.86 |

Note. T-Student tests are in parenthesis.

Tests on Tobit model (d.f. in parenthesis)

| | | | |
|---|---|---|---|
| Skewness (1) | 104.95 | Normality (2) | 430.24 |
| Kurtosis (1) | 33.47 | Heteroscedasticity (6) | 86.31 |

**Table 2. Estimated coefficients for bivariate models**

| Explanatory variable | Heckman's model | Double-hurdle independent | Double-hurdle dependent |
|---|---|---|---|
| First hurdle | | | |
| CONSTANT | - 2.5730 (20.7) | - 6.7892 (25.5) | - 7.1178 (24.2) |
| LINC | 0.4286 (30.6) | 1.0843 (34.5) | 1.1165 (30.6) |
| AGEHH | - 0.0135 (19.1) | - 0.0176 (11.8) | - 0.0171 (11.3) |
| NONMAN | 0.0229 (1.04) | 0.0196 (0.42) | - 0.0163 (0.34) |
| ED1 | - 0.0003 (0.01) | - 0.0779 (1.42) | - 0.0044 (0.08) |
| ED2 | 0.1064 (4.99) | 0.0975 (2.22) | 0.0935 (2.13) |
| ED4 | - 0.0760 (2.77) | - 0.1101 (1.28) | - 0.0868 (0.91) |
| ED5 | - 0.3236 (9.15) | - 0.3701 (3.04) | - 0.3551 (2.60) |
| Second hurdle | | | |
| CONSTANT | - 0.0247 (0.69) | 0.1216 (36.6) | 0.0914 (26.8) |
| LPRICE | 0.0113 (4.87) | 0.0077 (5.66) | 0.0081 (6.03) |
| LINC | 0.0041 (1.11) | - 0.0119 (34.1) | - 0.0085 (23.9) |
| UNEM | 0.0024 (1.80) | 0.0047 (6.38) | 0.0046 (6.31) |
| NONMAN | 0.0008 (0.65) | - 0.0004 (0.87) | - 0.0003 (0.57) |
| FORCES | 0.0009 (0.36) | 0.0010 (0.66) | 0.0013 (0.82) |
| NA1 | 0.0035 (7.58) | 0.0064 (26.6) | 0.0065 (27.0) |
| NA2 | 0.0021 (3.84) | 0.0050 (18.1) | 0.0051 (18.7) |
| NA3 | - 0.00003 (0.05) | 0.0020 (5.95) | 0.0020 (6.20) |
| AGEHH | - 0.0008 (4.76) | - 0.0007 (8.33) | - 0.0006 (7.93) |
| AGEHH2/100 | 0.0002 (1.37) | 0.0006 (6.98) | 0.0005 (5.71) |

| | | | |
|---|---|---|---|
| RG2 | - 0.0049 | - 0.0091 | - 0.0085 |
| | (3.12) | (10.1) | (9.80) |
| RG3 | - 0.0035 | - 0.0026 | - 0.0026 |
| | (2.01) | (2.63) | (2.70) |
| RG4 | - 0.0025 | - 0.0029 | - 0.0030 |
| | (1.17) | (2.32) | (2.42) |
| RG5 | - 0.0044 | - 0.0067 | - 0.0066 |
| | (2.63) | (7.07) | (6.97) |
| RG6 | - 0.0052 | - 0.0049 | - 0.0047 |
| | (2.56) | (3.39) | (3.26) |
| RG7 | - 0.0046 | - 0.0052 | - 0.0053 |
| | (4.19) | (8.27) | (8.43) |
| RG8 | - 0.0042 | - 0.0076 | - 0.0074 |
| | (3.45) | (10.9) | (10.7) |
| RG9 | - 0.0023 | - 0.0025 | - 0.0024 |
| | (1.47) | (2.51) | (2.55) |
| RG10 | - 0.0056 | - 0.0108 | - 0.0107 |
| | (3.94) | (14.0) | (14.0) |
| RG11 | - 0.0028 | - 0.0048 | - 0.0048 |
| | (1.71) | (5.14) | (5.14) |
| RG12 | - 0.0039 | - 0.0041 | - 0.0042 |
| | (3.10) | (5.48) | (5.65) |
| RG13 | - 0.0003 | - 0.0029 | - 0.0027 |
| | (0.12) | (2.37) | (2.22) |
| RG14 | - 0.0040 | - 0.0072 | - 0.0072 |
| | (1.57) | (4.84) | (4.82) |
| RG15 | - 0.0049 | - 0.0067 | - 0.0064 |
| | (1.95) | (4.16) | (4.01) |
| RG16 | - 0.0038 | - 0.0048 | - 0.0047 |
| | (2.98) | (6.40) | (6.32) |
| RG17 | - 0.0045 | - 0.0081 | - 0.0081 |
| | (2.93) | (9.07) | (9.08) |
| DM1 | - 0.0019 | - 0.0016 | - 0.0019 |
| | (1.78) | (2.71) | (3.17) |
| DM2 | - 0.0015 | - 0.0010 | - 0.0010 |
| | (1.73) | (1.90) | (2.03) |
| DM3 | - 0.0011 | - 0.0016 | - 0.0015 |
| | (1.34) | (3.17) | (3.15) |
| DM5 | 0.0004 | 0.0006 | 0.0005 |
| | (0.34) | (0.82) | (0.79) |
| ED1 | 0.0038 | 0.0064 | 0.0047 |
| | (1.88) | (7.97) | (6.62) |
| ED2 | 0.0044 | 0.0008 | 0.0010 |
| | (3.12) | (1.50) | (2.32) |

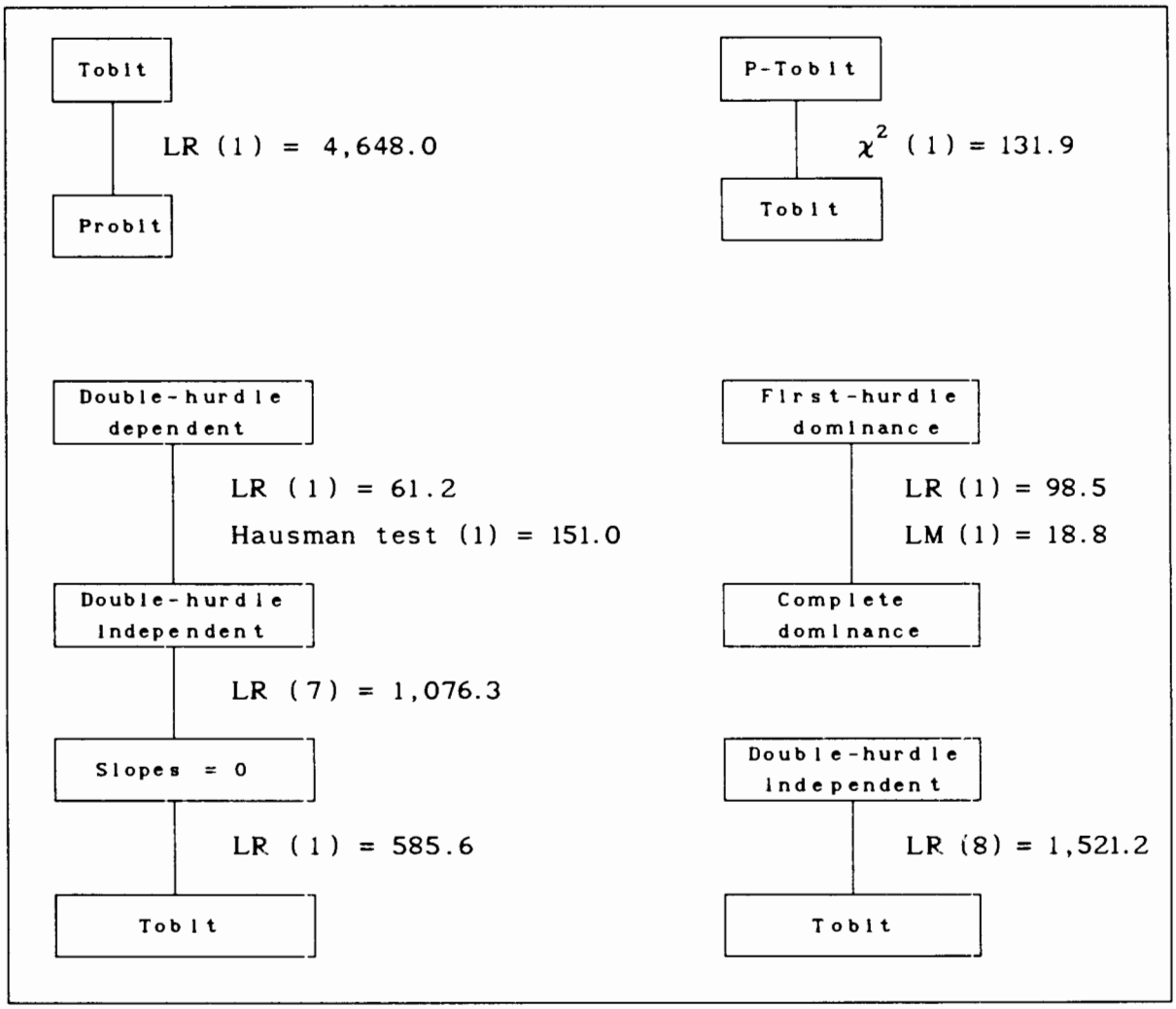
| | | | |
|---|---|---|---|
| ED4 | - 0.0026 | 0.0010 | - 0.0005 |
| | (1.78) | (1.62) | (0.81) |
| ED5 | - 0.0110 | 0.0002 | - 0.0012 |
| | (3.39) | (0.22) | (1.53) |
| NEARN | 0.0015 | 0.0010 | 0.0011 |
| | (3.25) | (1.23) | (1.34) |
| $\lambda$ | 0.0654 | --- | --- |
| | (4.33) | | |
| $\sigma_u$ | 0.0164 | 0.0205 | 0.0211 |
| $\rho$ | --- | --- | 0.5793 |
| | | | (17.4) |
| Log-Likelihood | - 41,481.4 | - 38,169.3 | - 38,138.7 |
| % correctly predicted | 66.32 | 68.88 | 68.59 |

Note. T-Student tests are in parenthesis.

Tests on double-hurdle models (d.f. in parenthesis)

| | Independent model | Dependent model |
|---|---|---|
| First hurdle | | |
| Skewness (1) | 3.09 | 2.91 |
| Kurtosis (1) | 0.38 | 0.35 |
| Normality (2) | 3.96 | 3.81 |
| Second hurdle | | |
| Skewness (1) | 42.07 | 31.11 |
| Kurtosis (1) | 14.81 | 9.57 |
| Normality (2) | 47.32 | 43.16 |
| Heteroscedasticity (6) | 38.44 | 37.42 |
| Bivariate normality (9) | --- | 50.12 |

**Table 3. Sequence of tests for model specification**




Notes

1. LR: Likelihood Ratio test. LM: Lagrange Multiplier test. $\chi^2$: Chi-squared test.
2. Degrees of freedom are in brackets.

**Table 4. Evaluation of probabilities**

| | Tobit | Heckman's model | Double-hurdle independent | | Double-hurdle dependent | |
|---|---|---|---|---|---|---|
| | | | (1) | (2) | (1) | (2) |
| Base situation | 58.98 | 58.76 | 63.77 | 13.61 | 68.38 | 17.75 |
| Changes with respect to the base situation | | | | | | |
| UNEM = 1 | 68.67 | 58.76 | 68.95 | 13.61 | 72.78 | 17.75 |
| NEARN = 2 | 63.70 | 58.76 | 66.76 | 13.61 | 70.62 | 17.75 |
| AGEHH = 25 | 73.05 | 71.43 | 77.67 | 6.27 | 79.08 | 7.56 |
| AGEHH = 65 | 49.57 | 51.04 | 55.58 | 19.59 | 61.69 | 26.26 |
| ED1 = 1 | 60.10 | 58.75 | 68.49 | 17.13 | 66.70 | 17.86 |
| ED5 = 1 | 52.98 | 45.93 | 55.35 | 22.32 | 60.64 | 28.27 |
| Bottom decile of total exp. | 56.48 | 37.91 | 45.01 | 54.48 | 43.29 | 44.76 |
| Top decile of total exp. | 61.53 | 68.74 | 57.22 | 2.60 | 63.14 | 2.60 |

Notes

1. Base situation: Household with four members, the head is a non-manual worker, aged 50, who has had a secondary education and who lives in an Andalusian town of between 50,000 and 500,000 inhabitants and has no additional contributors to household income. Income and prices take on their mean values. The changes with respect to the base situation are not cumulative.
2. Columns under heading (1) present the probability of smoking. Columns under heading (2) bring together the probability of being a non-smoker. (All the values are percentages).

**Table 5. Price and income elasticities and effects on changes in prices**

| a. Univariate models | | | |
|---|---|---|---|
| | OLS | Tobit (1) | Tobit (2) |
| Income elasticity | 0.46 (0.03) | 0.91 (0.02) | 0.91 (0.01) |
| Price elasticity | - 0.46 (0.01) | - 0.81 (0.08) | - 0.81 (0.03) |

| b. Bivariate models | | | | | | |
|---|---|---|---|---|---|---|
| | Heckman's model | | Double-hurdle independent | | Double-hurdle dependent | |
| | (1) | (2) | (1) | (2) | (1) | (2) |
| Income elasticity | 1.20 (0.21) | 1.11 (0.10) | 0.35 (0.02) | 0.43 (0.01) | 0.52 (0.02) | 0.59 (0.01) |
| Price elasticity | - 0.45 (0.13) | - 0.46 (0.06) | - 0.58 (0.08) | - 0.63 (0.03) | - 0.54 (0.08) | - 0.61 (0.03) |

| c. Effects of changes in prices | | | |
|---|---|---|---|
| | Tobit | Double-hurdle independent | Double-hurdle dependent |
| Base situation | 9.28 | 16.70 | 16.53 |
| Changes respect to the base sit. | | | |
| UNEM = 1 | 14.84 | 17.42 | 17.23 |
| AGE = 25 | 15.83 | 17.41 | 17.22 |
| AGE = 65 | 0.00 | 16.41 | 16.22 |
| ED1 = 1 | 11.02 | 17.56 | 17.24 |
| ED5 = 1 | 0.00 | 16.74 | 16.28 |
| Bottom decile of total exp. | 11.98 | 15.31 | 14.41 |
| Top decile of total exp. | 7.21 | 1.00 | 5.68 |

Notes

1. The values of the elasticities are calculated with the random effects interpretation (columns under heading 1) and with the fixed effects interpretation (columns under heading 2).
2. Standard errors are in brackets.
3. Base situation: The same as in Table 4. Changes are not cumulative.
4. The quantities above are percentages, evaluated for the expected value of the endogenous variable for a change in the real price of tobacco of 25%.

## DATA APPENDIX

The data are a sample of 23669 households from the 1980-81 Spanish EPF, 13918 observations with $Y_1 > 0$ and 9751 observations with $Y_1 = 0$, whose means, standard deviations and definition of the variables are presented in the following Table.

**Table A.1 Means and standard deviations of the explanatory variables**

| | All observations | | Positive observ. | |
|---|---|---|---|---|
| Variable | Mean | S. D. | Mean | S. D. |
| Share | 0.012 | 0.017 | 0.021 | 0.036 |
| LINC | 8.085 | 0.718 | 8.252 | 0.619 |
| LPRICE | - 0.352 | 0.137 | - 0.354 | 0.137 |
| AGEHH | 50.530 | 15.100 | 47.600 | 13.720 |
| Occupation. | | | | |
| UNEM | 0.044 | 0.206 | 0.054 | 0.225 |
| FORCES | 0.011 | 0.106 | 0.014 | 0.119 |
| NONMAN | 0.576 | 0.494 | 0.649 | 0.478 |
| Household composition. | | | | |
| NA1 | 0.368 | 0.693 | 0.476 | 0.770 |
| NA2 | 1.596 | 0.900 | 1.802 | 0.792 |
| NA3 | 0.538 | 0.772 | 0.427 | 0.725 |
| NEARN | 2.557 | 0.776 | 2.674 | 0.826 |
| Education. | | | | |
| ED1 | 0.070 | 0.254 | 0.049 | 0.216 |
| ED2 | 0.256 | 0.436 | 0.245 | 0.430 |
| ED4 | 0.131 | 0.337 | 0.149 | 0.356 |
| ED5 | 0.068 | 0.252 | 0.070 | 0.255 |
| Size of town of residence. | | | | |
| DM1 | 0.122 | 0.328 | 0.107 | 0.309 |
| DM2 | 0.168 | 0.374 | 0.164 | 0.370 |
| DM3 | 0.184 | 0.387 | 0.185 | 0.388 |
| DM5 | 0.120 | 0.325 | 0.122 | 0.327 |
| Region of residence. | | | | |
| RG2 (Aragón). | 0.055 | 0.228 | 0.046 | 0.210 |
| RG3 (Asturias). | 0.029 | 0.168 | 0.033 | 0.178 |
| RG4 (Baleares). | 0.020 | 0.141 | 0.020 | 0.149 |
| RG5 (Canarias). | 0.037 | 0.188 | 0.037 | 0.188 |
| RG6 (Cantabria). | 0.022 | 0.148 | 0.025 | 0.156 |
| RG7 (Castilla-León). | 0.141 | 0.348 | 0.141 | 0.348 |
| RG8 (Cast.-La Mancha). | 0.100 | 0.300 | 0.093 | 0.291 |
| RG9 (Cataluña). | 0.039 | 0.194 | 0.040 | 0.196 |
| RG10 (C. Valenciana). | 0.067 | 0.249 | 0.056 | 0.230 |
| RG11 (Extremadura). | 0.054 | 0.225 | 0.057 | 0.231 |

| | | | | |
|---|---|---|---|---|
| RG12 (Galicia). | 0.076 | 0.265 | 0.077 | 0.266 |
| RG13 (Madrid). | 0.019 | 0.137 | 0.019 | 0.136 |
| RG14 (Murcia). | 0.015 | 0.123 | 0.015 | 0.122 |
| RG15 (Navarra). | 0.015 | 0.120 | 0.015 | 0.120 |
| RG16 (País Vasco). | 0.075 | 0.263 | 0.076 | 0.266 |
| RG17 (La Rioja). | 0.051 | 0.220 | 0.049 | 0.216 |

Definition of the variables:

LINC: Logarithm of total real expenditure.
LPRICE: Logarithm of real price of tobacco.
AGEHH: Age of the head of the household.

Occupational dummies:

UNEM: 1 if the head of the household is unemployed.
FORCES: 1 if the head works in the armed forces.
NONMAN: 1 if the head is a non-manual worker.

Household composition:

NA1: Number of members between 17 and 24 years.
NA2: Number of members between 25 and 60.
NA3: Number of members over 60.
NEARN: Number of additional contributors to household income.

Educational dummies:

EDI=1, (I=1, ..., 5) if the head of the household is illiterate or has no educational background, he has completed primary education, secondary studies, pre-university studies or university studies respectively, 0 otherwise.

Size of town of residence:

DMI=1, (I=1, ..., 5) if the family lives in a town of under 2000, between 2000-10000, 10000-50000, 50000-500000 and over 500000 inhabitants respectively, 0 otherwise.

Region of residence:

RGI=1, (I=1, ..., 17) if the family resides in any of the specific regions presented in the table above. For avoiding the multicolinearity problem, RG1 has been omitted in the estimation. It corresponds to Andalusia.

**Table A.2 Retail Price Index and Price Index for Tobacco**

| Trim. | Black tobacco | Virginia tobacco | Total tobacco | Retail price index |
|---|---|---|---|---|
| II/80 | 122.2 | 152.4 | 129.8 | 195.6 |
| III/80 | 122.2 | 152.3 | 129.8 | 202.1 |
| IV/80 | 162.2 | 189.1 | 169.0 | 209.3 |
| I/81 | 182.2 | 207.6 | 188.6 | 218.2 |

Note. Base year: 1976.

## REFERENCES


AMEMIYA, T. (1984) "Tobit Models: A Survey", *Journal of Econometrics* **24**, 3-61

ATKINSON, A.B., J. GOMULKA and N.H. STERN (1984), "Household Expenditure on Tobacco 1970-1980: Evidence from the Family Expenditure Survey", ERSC Programme on Taxation, Incentives and the Distribution of Income, DP 57.

ATKINSON, A.B., J. GOMULKA and N.H. STERN (1989), "Spending on Alcohol: Evidence from the Family Expenditure Survey 1970-1983", *The Economic Journal* **100**, 808-827

ATKINSON, A.B., J. GOMULKA and N.H. STERN (1990), "Household Expenditure on Tobacco 1970-1983: Evidence from the Family Expenditure Survey ", ESRC Programme on Taxation, Incentives and the Distribution of Income, DP 134.

BERA, A.K., C.M. JARQUE and L.F. LEE (1984), "Testing for the Normality Assumption in Limited Dependent Variable Models", *International Economic Review* **25**, 563-578

BERNDT, E.R., B.H. HALL, R.E. HALL and J.A. HAUSMAN (1974), "Estimation and Inference in Non-Linear Structural Models", *Annals of Economic and Social Measurement* **3**, 653-665

BLUNDELL, R.W. and C. MEGHIR (1986), "Selection Criteria for a Microeconometric Model of Labour Supply", *Journal of Applied Econometrics* **1**, 55-81

BLUNDELL, R.W. and C. MEGHIR (1987), "Bivariate Alternatives to the Tobit Model", *Journal of Econometrics* **34**, 179-200

BLUNDELL, R.W. J. HAM and C. MEGHIR (1987), "Unemployment and Female Labour Supply" Suplement to *The Economic Journal* **97**, 44-64

BLUNDELL, R.W. J. HAM and C. MEGHIR (1989), "Unemployment, Discouraged Workers and Female Labour Supply", University College London, Discussion Paper 02.

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

DEATON, A.S. (1981), "Theoretical and Empirical Approaches to Consumer Demand Under Rationing", in A.S. Deaton (ed.), *Essays in the Theory and Measurement of Consumer Behavior*, Cambridge: Cambridge University Press, 55-72

DEATON, A.S. and M. IRISH (1984), "Statistical Models for Zero Expenditures in Household Budgets", *Journal of Public Economics* **23**, 59-80

DEATON, A.S. and J.N.J. MUELLBAUER (1980), "An Almost Ideal Demand System", *American Economic Review* **70**, 312-326

HECKMAN, J.J. (1979), "Sample Selection Bias as a Specification Error", *Econometrica* **47**, 153-161

JONES, A.M. (1989), "A Double-Hurdle Model of Cigarette Consumption", *Journal of Applied Econometrics* **4**, 23-39

LEE, L.F. (1984), "Tests for the Bivariate Normal Distribution in Econometric Models with Selectivity", *Econometrica* **52**, 843-864

LEE, L.F. and M.M PITT (1986), "Microeconometric Demand Systems with Binding Non-Negativity Constraints: The Dual Approach", *Econometrica* **54**, 1237-1242

MADDALA, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

McDONALD, J.F. and R.A. MOFFITT (1980), "The Uses of Tobit Analysis", *The Review of Economics and Statistics* **62**, 318-321

MELINO A. (1982), "Testing for Sample Selection Bias", *Review of Economic Studies* **49**, 151-153

NEARY, J.P. and K.W.S. ROBERTS (1980), "The Theory of Household Behavior Under Rationing", *European Economic Review* **13**, 25-42

NEWEY, W.K., J.L. POWELL and J.R. WALKER (1990), "Semiparametric Estimation of Selection Models: Some Empirical Results", *American Economic Review Papers and Proceedings* **80**, 324-328

POLLAK, R.A. and T.J. WALES (1981), "Demographic Variables in Demand Analysis", *Econometrica* **49**, 1533-1551

POWELL, J.L. (1986), "Simmetrically Trimmed Least Squares Estimation for Tobit Models", *Econometrica* **54**, 1435-1460

ROTHBARTH, E. (1941), "The Measurement of Changes in Real Income Under Conditions of Rationing", *Review of Economic Studies* **8**, 100-107

RUUD P.A. (1984), "Tests of Specification in Econometrics", *Econometric Reviews* **3**, 211-242

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

WALES T.J. and A.D. WOODLAND (1983), "Estimation of Consumer Demand Models with Binding Non-Negativity Constraints", *Journal of Econometrics* **21**, 263-285

## RECENT WORKING PAPERS

1. Albert Marcet and Ramon Marimon
Communication, Commitment and Growth. (June 1991)

2. Antoni Bosch
Economies of Scale, Location, Age and Sex Discrimination in Household Demand.(June 1991)

3. Albert Satorra
Asymptotic Robust Inferences in the Analysis of Mean and Covariance Structures. (June 1991)

4. Javier Andrés and Jaume Garcia
Wage Determination in the Spanish Industry. (June 1991)

5. Albert Marcet
Solving Non-Linear Stochastic Models by Parameterizing Expectations: An Application to Asset Pricing with Production. (July 1991)

6. Albert Marcet
Simulation Analysis of Dynamic Stochastic Models: Applications to Theory and Estimation. (November 1991)

7. Xavier Calsamiglia and Alan Kirman
A Unique Informationally Efficient and Decentralized Mechanism with Fair Outcomes. (November 1991)

8. Albert Satorra
The Variance Matrix of Sample Second-order Moments in Multivariate Linear Relations. (January 1992)

9. Teresa Garcia-Milà and Therese J. McGuire
Industrial Mix as a Factor in the Growth and Variability of States'Economies. (Jannuary 1992)

10. Walter Garcia-Fontes and Hugo Hopenhayn
Entry Restrictions and the Determination of Quality. (February 1992)

11. Guillem López and Adam Robert Wagstaff
Indicadores de Eficiencia en el Sector Hopitalario. (March 1992)

12. Daniel Serra and Charles ReVelle
The PQ-Median Problem: Location and Districting of Hierarchical Facilities. Part I (April 1992)

13. Daniel Serra and Charles ReVelle
The PQ-Median Problem: Location and Districting of Hierarchical Facilities. Part II: Heuristic Solution Methods. (April 1992)

**14.** Juan Pablo Nicolini
Ruling out Speculative Hyperinflations: a Game Theoretic Approach. (April 1992)

**15.** Albert Marcet and Thomas J. Sargent
Speed of Convergence of Recursive Least Squares Learning with ARMA Perceptions. (May 1992)

**16.** Albert Satorra
Multi-Sample Analysis of Moment-Structures: Asymptotic Validity of Inferences Based on Second-Order Moments. (June 1992)

**Special issue** Vernon L. Smith
Experimental Methods in Economics. (June 1992)

**17.** Albert Marcet and David A. Marshall
Convergence of Approximate Model Solutions to Rational Expectation Equilibria Using the Method of Parameterized Expectations.

**18.** M. Antònia Monés, Rafael Salas and Eva Ventura
Consumption, Real after Tax Interest Rates and Income Innovations. A Panel Data Analysis. (December 1992)

**19.** Hugo A. Hopenhayn and Ingrid M. Werner
Information, Liquidity and Asset Trading in a Random Matching Game. (February 1993)

**20.** Daniel Serra
The Coherent Covering Location Problem. (February 1993)

**21.** Ramon Marimon, Stephen E. Spear and Shyam Sunder
Expectationally-driven Market Volatility: An Experimental Study. (March 1993)

**22.** Giorgia Giovannetti, Albert Marcet and Ramon Marimon
Growth, Capital Flows and Enforcement Constaints: The Case of Africa. (March 1993)

**23.** Ramon Marimon
Adaptive Learning, Evolutionary Dynamics and Equilibrium Selection in Games. (March 1993)

**24.** Ramon Marimon and Ellen McGrattan
On Adaptive Learning in Strategic Games. (March 1993)

**25.** Ramon Marimon and Shyam Sunder
Indeterminacy of Equilibria in a Hyperinflationary World: Experimental Evidence. (March 1993)

**26.** Jaume Garcia and José M. Labeaga
A Cross-Section Model with Zeros: an Application to the Demand for Tobacco. (March 1993)