ESTUDIOS SOBRE LA ECONOMIA ESPAÑOLA

**The probability that a smoker
does not purchase tobacco: A note**

Daniel Miles

EEE 15

Fundación de Estudios de Economía Aplicada

<http://www.fedea.es/hojas/publicado.html>

The Probability that a Smoker does not Purchase Tobacco: A Note

Daniel Miles*

February 18, 1999

Abstract

In this note we are concerned with the estimation of the probability that a smoker does not purchase tobacco during a survey. Usually, tobacco demand has been estimated using limited dependent variable models under the assumption that an important proportion of smokers declared a zero expenditure in tobacco. However, if the probability of non purchasing by a smoker is negligible, zeros could be ignored and the demand equation could be estimated on positive expenditure data using traditional estimation methods. Here we estimate this probability and find that it is extremely small. A novelty of this work is the use of data on the quantity and frequency of tobacco purchases during the week of the survey, instead of the more commonly used expenditure data.

*Research funded by Spanish Dirección General de Enseñanza Superior, referencia PB-95-0292. This research is part of the fourth chapter of my doctoral dissertation at Universidad Carlos III de Madrid. I thank the comments of Miguel A. Delgado. Mail: dmiles@uvigo.es

1. INTRODUCTION

The existence of an important proportion of zeros in the tobacco expenditure data introduces some difficulties when estimating tobacco demand using household budget surveys. In first place, with survey data it is not possible to identify whether the household who reported a zero expenditure in tobacco is a nonsmoker or a smoker who did not purchase. In second place, zeros introduce a censoring problem invalidating traditional estimation methods.

To take into account both facts mentioned above, tobacco demand has been commonly estimated using limited dependent variable models e.g. Tobit or Double Hurdle (see Cragg, 1971; Jones, 1989; Pudney, 1989; García and Labeaga, 1996, among others). First, these models assume that a relatively important number of smokers reported a zero expenditure in tobacco during the survey, although these models are not able to verify this assumption. Second, to interpret the censoring at zero, limited dependent variable models assume the same decision process at zero for nonsmokers and smokers, for whom a zero can only correspond to a corner solution (Pudney, 1989; Blundell and Meghir, 1987, among others).

Notice that if all zeros were from nonsmokers or, if the probability that a smoker does not purchase tobacco during the survey were negligible, then zeros could be ignored and the tobacco demand estimation could be feasibly achieved by traditional methods. In this note we estimate the probability that a smoker reports a zero expenditure in tobacco, an issue that has not yet been addressed in applied literature, finding that it is very small. This fact empirically contradicts the first assumption supporting the use of limited dependent variable models to estimate tobacco demand using budget survey data.

Also, it does not seem reasonable to assume that a smoker will not purchase tobacco because he is in a corner solution. Notice that a corner solution implies that the

smoker is confronted to a relative price-income situation at which he finds optimal not to smoke, despite of his addiction and habits (Becker et al., 1994). In recent years, the zero expenditures reported on commodities which households are actually consuming have been interpreted as arising from the infrequency of purchases. In this note we assume that zeros of smokers arise because of the infrequency of tobacco purchases, i.e., some smokers did not purchase during the survey because they had consumed from their stocks. Hence, we discuss the tobacco stockpiling behavior in order to establish the percentage of smokers who had accumulated tobacco the weeks before to the survey, because these are the potential smokers reporting a zero.

At last, tobacco demand has been usually estimated using household tobacco expenditure data. A novelty of this paper is the use of data on the quantity of cigarettes bought and on the frequency of tobacco purchases by the household during the week of the survey. These data were recovered from the Spanish Expenditure Survey 1990-91 (EPF; Cardelús et al. 1995). We have selected a subsample with an employed household between 15 and 65 years and considered exclusively cigarette consumption, leaving a total of 10.009 observations.

The note is organized in three sections. First, we discuss whether there is an important proportion of smokers who stockpiled tobacco. Second, using Robin's (Robin, 1993) approach, we estimate the probability that a smoker did not purchase during the survey. We find that this probability is extremely small. Section three concludes.

2. STOCKPILING BEHAVIOR

There are smokers who buy tobacco regularly and, therefore, will report a positive tobacco expenditure with probability one¹. But also, there are smokers who stockpile,

¹Notice that we identify smoker with a smoker household, as is common in applied economic literature.

consuming during the survey from their stocks. In consequence, those smokers who stockpiled the week before to the survey are the potential smokers reporting a zero.

If the cigarette purchasing process was stationary and the stockpiling type of smokers was uniformly distributed along the weeks of the year, the percentage of smokers who stockpiled during the week before to the survey (those who declared a zero) could be approximated by those who had stockpiled during the survey. Therefore, in this section, to approximate those smokers who reported a zero we will use the percentage of smokers who stockpiled during the survey.

First, in the subsample considered, 32% of the households reported a zero expenditure in cigarettes. That is, only three of ten households declared a zero expenditure in tobacco. For those who declared a positive expenditure, in the following table we present the joint frequency of the quantity of packs of cigarettes bought and number of purchases.

Insert Table 1

Less than 3% of the smokers with a positive expenditure went only once to the tobacco shop buying 8 or more packs of cigarettes. Notice, also, that more than 90% of the smokers bought tobacco more than once. That is, most smokers went regularly to the tobacco shop. Hence, if the stockpiling behavior is characterized as buying large quantities of tobacco in few visits to the shop, not too many smokers had accumulated tobacco during this survey.

In Table 2 we present the numbers of packs of cigarettes per purchase.

Insert Table 2

Only a small percentage of smokers bought a relatively important quantity of cigarettes per purchase, i.e., 4.78% of the smokers bought an average of 5 or more packs of cigarettes per purchase. Both tables above suggest that smokers visit the tobacco shop regularly and purchase small quantities of tobacco per purchase. That is, only

a relatively small percentage of smokers appears to stockpile.

Naturally, stockpiling cigarettes should be defined in terms of the level of consumption. Let Q_h be the total packs bought by smoker h during the week of the survey and N_h the number of times he went to the shop. Then, $(20Q_h/N_h)$ is the consumption of cigarettes in the interpurchase period, being 20 the number of cigarettes per pack. If we assume that purchases are equally spaced between the days of the week, the number of purchases per day is given by $(N_h/7)$ and the cigarette consumption per day is given by $(20Q_h/N_h)(N_h/7)$ (see Kay, Keen and Morris, 1984). In Table 3 we present this approximation for those smokers with a positive expenditure and with 2 to 7 purchases.

Insert Table 3

Observe that smoker households consume an average of one pack per day (20 cigarettes), with a maximum of approximately 4 packs and a minimum of six cigarettes a day. This suggests that those who bought cigarettes more than once during the week of survey had consumed them during that week, not stockpiling tobacco.

Also, from the Spanish National Health Survey (1993), non habitual smokers consume an average of 4 cigarettes a day, and consequently a pack of cigarette lasts, at most, one week. If in Table 1 we use this level of consumption for those smokers that only bought once a week we find that, at most, a 4% of the total smokers declaring a positive expenditure had stockpiled tobacco during the week of the survey².

The discussion above suggests that not many smokers had stockpiled during the week of the survey. This implies, following the previous reasoning, that not too many smokers reported a zero expenditure in tobacco during the survey.

²Note that in Table 1, 4.14% of smokers that went once to the tobacco shop bought one pack. With an average consumption of 4 cigarettes a day, this one pack lasts less than a week.

3. PROBABILITY OF NON-PURCHASING

Using Robin's results, in this section we estimate the probability that a smoker declares a zero expenditure in tobacco.

Let N_h be the number of purchases per week by household h , where $N_h > 0$ if the household bought tobacco and $N_h = 0$ if it did not. Following Robin, let ε_h represent unobservable characteristics that determine whether household h is a smoker or not. That is, given a set of observable characteristics, z_h , let S be the set of the unobservable characteristics, ε_h , for a smoker, defined by $S(z_h) = \{\varepsilon \mid E(N_h \mid z_h, \varepsilon_h = \varepsilon) > 0\}$.

If a household declares a positive expenditure in tobacco, then it is a smoker with probability one, so $\Pr(N_h \mid z_h, \varepsilon_h \in S(z_h), N_h > 0) = \Pr(N_h \mid z_h, N_h > 0)$, is identifiable using the set of positive observations of N_h .

Now, assuming a parametric distribution function of N_h such that its complete distribution is recoverable from its truncated one, the probability of not purchasing being a smoker, $\Pr(N_h = 0 \mid z_h, \varepsilon_h \in S(z_h))$, becomes identifiable (Robin, 1993; Flinn and Heckman, 1982). As $\Pr(N_h = 0 \mid z_h)$ is identifiable from the proportion of zero expenditures on total observations, then we can identify the probability of smoking

$$\Pr(\varepsilon_h \in S(z_h) \mid z_h) = \frac{\Pr(N_h > 0 \mid z_h)}{\Pr(N_h > 0 \mid z_h, \varepsilon_h \in S(z_h))}. \quad (1)$$

Given the information on quantity or frequency of purchases, to estimate the probabilities above we need to assume a recoverable parametric density. Following Robin (1993), we assume that the purchasing process is distributed as a Negative Binomial where its truncated density is given by

$$\Pr(N_h = n_h \mid z_h, N_h > 0) = \frac{\Gamma(n_h + \delta)}{\Gamma(n_h + 1)\Gamma(\delta)} (\lambda_i/\delta)^{n_h} (1 + \lambda_i/\delta)^{-(n_h + \delta)} \left(1 - (1 + \lambda_i/\delta)^{-\delta}\right)^{-1}.$$

Notice that the truncated density introduces the correction $\Pr(N_h > 0 \mid z_h, \varepsilon_h \in S(z_h)) = 1 - (1 + \lambda_i/\delta)^{-\delta}$, the probability that a smoker purchases tobacco (see, Groger and

Carson, 1991; Creel and Loomis 1990). As usual, $\lambda_i = \exp(z_h' \beta)$, where β are the parameters of interest. The estimation results are presented in the appendix.

Then, the estimate of the probability that a smoker household purchases tobacco is given by

$$\hat{P}(N_h > 0 \mid z_h, \varepsilon_h \in S(z_h)) = 1 - \frac{1}{n} \sum_{i=1}^n \left(1 + \hat{\lambda}_i / \hat{\delta}\right)^{-\hat{\delta}} \quad (2)$$

where $\hat{\lambda}_i$ and $\hat{\delta}$ are the estimated parameters from the truncated binomial density function (see Robin, 1993). With this estimation and given that $P(N_h > 0 \mid z_h)$ is identifiable from the proportion of positive to total observations, in Table 4 we estimate the probability that a smoker does not purchase using equation (1). We have applied this last approach using both variables, the number of purchases and the quantity of packs bought as dependent.

Insert Table 4

The first column is the estimate of the probability that a smoker buys tobacco during the week of the survey. When the number of purchases is used as the dependent variable in Robin's approach, 97.8% of the smokers will buy during the week of the survey or 98.8% if we use the number of packs bought as the dependent variable. These suggest that practically all the smokers had purchased during the survey. In the second column, we present the proportion of positive observations to total observations in the sample. The third column is the probability of being a non-smoker, obtained as the complement of equation (1), from where it results that nearly one third of the sample is non-smoker. The fourth column states the number of zeros in the sample. Finally, the fifth column is the difference between columns four and three, which is the proportion of zeros that correspond to smokers. As can be seen, it is very small, i.e., less than 2% of the zeros correspond to smokers. That is, these estimations suggest that the proportion of smoker households that report a zero expenditures in tobacco during the survey is extremely small.

CONCLUDING REMARKS

In this note we have estimated of the probability that a household smoker does not purchase tobacco during a survey, finding that it is extremely small. This suggests that it seems feasible to estimate the household demand for tobacco using only positive expenditure observations, which allows for a direct interpretation of the estimated demand parameters..

REFERENCES

- [1] Becker, G., Grossman, M. and Murphy, K. (1994) "An Empirical Analysis of Cigarette Addiction", *The American Economic Review*, 84(3), 396-418.
- [2] Blundell, R. and Meghir, C. (1987) "Bivariate Alternatives to the Tobit Model", *Journal of Econometrics*, 34, 179-200.
- [3] Cardelús, M., Arévalo, R. and Ruiz Castillo, J. (1995) "La Encuesta de Presupuestos Familiares de 1990-91", Documento de Trabajo 95-07(05), Universidad Carlos III de Madrid.
- [4] Cragg, J. (1971) "Some Statistical Models for Limited Dependent Variables with Application to the Demand of Durable Goods", *Econometrica*, 39, 829-44.
- [5] Creel, M. and Loomis, J. (1990) "Theoretical and Empirical Advantages of Truncated Count Data Estimators for Analysis of Deer Hunting in California", *Journal of Agricultural Economics*, 72, 434-41.
- [6] Garcia, J. and Labeaga, J. (1996) "Alternative Approaches to Modelling Zero Expenditure: An Application to Spanish Demand for Tobacco", *Oxford Bulletin of Economics and Statistics*, 58(3), 489-506..
- [7] Grogger, J. y Carson (1991) "Models for Truncated Counts", *Journal of Applied Econometrics*, 6, 225-38.
- [8] Jones, A. (1989) "A Note on Computation of the Double-Hurdle Model with Dependence with an Application to Tobacco Expenditure", *Bulletin of Economic Research* 44(1), 67-74.
- [9] Kay, J., Keen, M. and Morris, C (1984) "Estimation Consumption from Expenditure Data", *Journal of Public Economics*, 23, 161-81.

- [10] Lawless, J. (1987) "Negative Binomial and Mixed Poisson Regression", *The Canadian Journal of Statistics*, 15(3), 209-25.
- [11] Miles, D. (1998) "Especificación e Inferencia en Modelos Económicos para Curvas de Engel" unpublished PhD dissertation, Universidad Carlos III de Madrid.
- [12] Mullahy, J. (1997) "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior", *The Review of Economics and Statistics*, 586-93.
- [13] Pudney, S. (1989) "*Modelling Individual Choice: the Econometrics of Corners, Kinks and Holes*", Oxford: Basil Blackwell.
- [14] Robin, JM. (1993) "Ecomometric Analysis of the Short-run Fluctuations of Household's Purchases", *Review of Economic Studies*, 60, 923-34.
- [15] Winkelmann, R. (1997) *Econometric Analysis of Count Data*, Second Edition, Springer.

TABLE 1: Joint Frequency of Quantity and Number of Purchases for Positive Expenditure .

Number of Packs per Week	Number of Purchases per Week								Total Packs
	1	2	3	4	5	6	7	>7	
1	4.14								4.14
2	1.03	5.75							6.78
3	0.19	4.36	4.15						8.70
4	0.12	2.99	1.16	4.70					8.97
5	0.06	1.25	0.62	1.64	3.14				6.71
6	0.03	0.41	0.53	1.31	0.94	4.33			7.55
7	0.01	0.10	0.10	0.83	0.57	1.41	9.65		12.67
8-10	2.32	0.09	0.19	0.84	0.87	2.50	4.17	6.39	17.37
11-12	0.09	0.9	0.57	0.09	0.20	0.94	1.08	3.95	7.83
13-15	0.00	0.16	0.74	0.54	0.45	0.59	3.59	5.88	11.95
16-20	0.00	0.22	0.07	0.19	0.17	0.37	0.65	3.80	5.47
>20	0.00	0.01	0.03	0.06	0.01	0.00	0.01	1.74	1.86
Total Purchases	8.0	16.2	8.2	10.2	6.4	10.1	19.2	21.7	

TABLE2: Positive Expenditure Subsample: Distribution of the ratio between quantity and number of purchases per week.

N. Purchases	Mean	Q ₂₅	Q ₅₀	Q ₇₅	Máx.	% Packs/Purchases		
						1-2	3-4	≤5
1	4.00	1.00	1.00	10.0	12.0	64.6	3.9	31.5
2	1.96	1.00	1.50	2.00	12.5	80.7	11.5	7.8
3	1.82	1.00	1.00	2.00	8.00	79.2	10.6	10.2
4	1.51	1.00	1.25	1.75	6.25	88.4	10.8	0.8
5	1.43	1.00	1.20	1.60	4.60	86.4	13.4	0.2
6	1.38	1.00	1.17	1.50	3.33	90.6	9.40	0
7	1.32	1.00	1.00	1.57	3.14	95.2	4.80	0
8>	1.25	1.00	1.12	1.38	2.89	95.9	4.10	0
Total	1.69	1.00	1.17	1.67	12.5	87.6	7.62	4.78

Note: The mean quantity of packs per purchase is defined as the ratio between the total quantity of packs bought per week to the total number of purchases. Q₂₅,Q₅₀,Q₇₅ represent the first, second and third quartil.

TABLE 3: Consumption of Cigarettes per day, defined as the ratio between the number of packs bought and the number of purchases, times 20, the number of cigarettes per pack.

Number Purchases	Mean	Q ₂₅	Q ₅₀	Q ₇₅	Min	Max
2	11	6	9	11	6	71
3	16	9	9	17	9	68
4	17	11	14	20	11	71
5	20	14	17	23	14	66
6	24	17	20	26	17	57
7	26	20	20	31	20	63
Total	23	11	20	31	6	71

Note: Round to nearest integer.

Table 4 Estimation of the Probability of Non-purchasing of Smokers .

Dep. Variable	Probability of Purchase of a Smoker ¹	Purchase in the Sample ²	Probability NonSmoker ³	Zeros in the Sample ⁴	Probability Smokers Zero ⁵
Purchase	97.80%	67.80%	30.67%	32.20%	1.53%
Quantity	98.78%	67.80%	31.36%	32.20%	0.84%

Note: Purchase or Quantity refers to the dependent variable used for estimating the negative binomial for the application of Robint's approach. 1. $\left(1 - \sum_{i=1}^n \left(1 + \hat{\lambda}_i / \hat{\delta}\right)^{-\hat{\delta}} / n\right)$;
2. $\sum_{i=1}^n I(N_h > 0) / n$; 3. $\left(1 - \frac{\widehat{\Pr}(N_h > 0 | z_h)}{\widehat{\Pr}(N_h > 0 | z_h, \varepsilon_h \in C(z_h))}\right)$; 4. $\sum_{i=1}^n I(N_h = 0) / n$;
5. $\left[\left(\sum_{i=1}^n I(N_h = 0) / n\right) - \left(1 - \frac{\widehat{\Pr}(N_h > 0 | z_h)}{\widehat{\Pr}(N_h > 0 | z_h, \varepsilon_h \in C(z_h))}\right)\right]$

Table A1 Negative Binomial

Number of Purchases							Number of Packs					
	Total Sample		Positive Sample				Total Sample		Positive Sample			
Negative Binomial					Negative Truncated Binomial		Negative Binomial				Negative Truncated Binomial	
Variable	Coef.	ST	Coef.	ST	Coef.	ST	Coef.	ST	Coef.	ST	Coef.	ST
Constant	.7940	.2405	1.342	.1662	1.290	.1803	1.039	.2446	1.637	.1695	1.606	.1768
Size Household	.2027	.0171	.1179	.0123	.1267	.0131	.1728	.0172	.0886	.0123	.0921	.0127
Child less 8 Years	-.1657	.0253	-.1018	.0181	-.1104	.0195	-.1252	.0254	-.0642	.0181	-.0668	.0188
Child 8-17 Years	-.1892	.0202	-.1105	.0145	-.1191	.0155	-.1519	.0203	-.0760	.0146	-.0790	.0152
Larger than 25	.0936	.0192	.0575	.0143	.0607	.0151	.0585	.0191	.0219	.0142	.0225	.0147
Age Partner	-.0130	.0014	-.0055	.0010	-.0060	.0010	-.0123	.0014	-.0046	.0010	-.0049	.0010
Town	.0700	.0220	.0614	.0157	.0659	.0170	.0661	.0224	.0520	.0158	.0541	.0165
Madrid	.0819	.0504	.0328	.0359	.0359	.0386	.1054	.0513	.0657	.0369	.0680	.0382
Catalunya	-.2047	.0435	-.0814	.0303	-.0884	.0331	-.1482	.0454	-.0256	.0320	-.0267	.0334
South	.1531	.0275	.0703	.0197	.0757	.0212	.1465	.0277	.0645	.0197	.0669	.0205
North	-.0607	.0307	-.0391	.0219	-.0423	.0238	-.0979	.0310	-.0758	.0221	-.0791	.0232
East	-.0399	.0345	-.0294	.0244	-.0316	.0265	-.0123	.0345	-.0048	.0241	-.0049	.0251
Service	.2854	.0511	.1320	.0362	.1446	.0397	.2192	.0510	.0652	.0353	.0684	.0370
Industry	.2140	.0557	.0972	.0395	.1064	.0433	.1926	.0560	.0717	.0389	.0751	.0408
Construction	.2409	.0588	.1039	.0417	.1135	.0457	.2126	.0590	.0753	.0412	.0787	.0431
Manual	-.2752	.0659	-.1543	.0459	-.1671	.0498	-.2707	.0664	-.1516	.0464	-.1576	.0483
Blue Collar	.3918	.0607	.2144	.0422	.2317	.0457	.3678	.0609	.1896	.0423	.1970	.0441
Study	.1690	.0364	.0970	.0259	.1062	.0283	.0941	.0371	.0237	.0264	.0248	.0276
Women not W	-.0593	.0257	-.0489	.0182	-.0533	.0197	-.0292	.0260	-.0256	.0183	-.0267	.0191
Alcohol	.8E-04	.1E-04	.3E-04	.1E-04	.4E-04	.1E-04	.1E-03	.1E-04	.5E-04	.1E-04	.5E-04	.1E-04
Outside Food	.0483	.5E-02	.0069	.0035	.0071	.0037	.0511	.0052	.0079	.0035	.0081	.0036
LogIncome	-.0110	.0230	-.0014	.0159	-.0021	.0172	.0156	.0235	.0205	.0163	.0213	.0171
Alfa	.8396	.0191	5.757	.1868	4.441	.1693	.6669	.0143	4.284	.1021	3.784	.1057

Note: Size Household: number of members of in the household; Child less 8 years: number of children less than 8; Children 8-17 years: number of children between 9 and 17 years; Larger than 25: number of children larger than 25; Age Partner: age of the second member of the household; town: 1 if town is larger than 50000 habs., 0 otherwise; Madrid, Catalunya, North, East, takes value 1 if household lives in any of these regions; Service, Industry, Construction, Manual, Blue Collar takes value 1 if the head of the household works in any of these sectors or categories; study: if the head of the household has a minimum education; Women not W: partner does not work; Alcohol: expenditure in alcohol; outside food: number of times the household eat outside the house during the survey; logincome: log of total monetary household income per capita per week