Salvado, João Cotter

Article

The Determinants of Health Care Utilization in Portugal: An Approach with Count Data Models

Swiss Journal of Economics and Statistics

Provided in Cooperation with:
Swiss Society of Economics and Statistics, Zurich

Suggested Citation: Salvado, João Cotter (2008) : The Determinants of Health Care Utilization in Portugal: An Approach with Count Data Models, Swiss Journal of Economics and Statistics, ISSN 2235-6282, Springer, Heidelberg, Vol. 144, Iss. 3, pp. 437-458, https://doi.org/10.1007/BF03399261

The Determinants of Health Care Utilization in Portugal: An Approach with Count Data Models

João Cotter Salvado[a]

JEL-Classification: I10, I12, C13, C25, C52
Keywords: health care utilization, count data models, Portugal

1. Introduction

This work models the demand for health services in Portugal measured as counts of utilization through the number of visits to a doctor. At the present time, it is important to understand the decision-making process to obtain a better evaluation of the forces that cause the increase in health care expenditures (Pohlmeier and Ulrich, 1995).

Portugal has highly state-financed health-care system like most nations, and subsequently the results obtained with this work will be appropriate to use in other countries with policy purposes.

Having a clear idea of both, health care demand determinants and the magnitude of their effects will help the debate on the delivery and organization of health services to efficiently attain a healthier society (Sarma and Simpson, 2006). To be precise, with these results we will be able to identify which factors are associated with more or less consumption of health services, the magnitude of the effect of those factors and finally we will be able to know how to act according to the results. This is the main motivation for this work.

In the Health Economics literature the first health care demand model was formulated by Grossman (1972). It is based on the traditional consumer theory and considers the individual as a sole agent or a prime decision maker in the process of health care utilization.

a Faculdade de Economia, Universidade Nova de Lisboa. Address: Campus de Campolide, 1099-032 Lisboa, Portugal. Tel: +351 21 380 16 00 Fax: +351 21 387 09 33. E-mail: jcotter@fe.unl.pt.
I am very grateful to Pedro Portugal who helped me throughout this paper as my master advisor. Additionally, I would like to thank to Pedro Pita Barros, Mafalda Sampaio, Miguel Câmara and Pedro Robalo for their important suggestions and corrections.

The second approach of this process is to assume that the demand for health services is made in two stages: the contact decision and the intensity of use decision. This model developed by Zweifel (1981) is based on the principal-agent framework.

This work will take into account these two approaches through the use of specific instruments. The Poisson and Negative Binomial procedures are appropriate to the first approach and the Hurdle Model is appropriate to the second approach.

Furthermore, our work will present another procedure. Instead of assuming a distinction between users and non-users made with the Hurdle Model, this procedure relies on a Latent Class analysis, distinguishing the agents into frequent and non-frequent users of health care services. This last approach is seen to dominate all the previously mentioned approaches (Deb and Triverdi, 2002).

2. Literature Review

Usually the regressors included in the estimation of health care utilization equation can be organized in four types: demographic, socioeconomic, health status and health insurance variables.

Among demographic variables, age and gender have a strong and significant effect in health care utilization. Age usually assumes a quadratic relationship with health care demand. Regarding the impact of gender women consume more health services than men (Windmeijer and Santos Silva, 1997; Pohlmeier and Ulrich, 1995; Cameron et al., 1988; Lourenço and Ferreira, 2005). Marital status does not appear to be relevant (Windmeijer and Santos Silva, 1997; Cameron et al., 1988), and education seems to have a negligible effect on health care utilization (Pohlmeier and Ulrich, 1995; Deb and Triverdi, 2002).

In the group of socioeconomic variables, family income is either negatively correlated with the demand of health care services (Pohlmeier and Ulrich, 1995) or not significant (Cameron et al., 1988) and being unemployed increase the demand although is generally not significant (Windmeijer and Santos Silva, 1997; Pohlmeier and Ulrich, 1995).

In most studies, health status variables are significant and seem to explain with much more relevancy health care utilization (Cameron et al., 1988). The number of visits to a doctor increases in the presence of chronic diseases and short-term or long-term disability (Pohlmeier and Ulrich, 1995; Cameron et al., 1988; Lourenço and Ferreira, 2005; Deb and Triverdi, 1997, 2002) and decreases with the individual health status.

For the health insurance variables, the magnitude of the effects of different health insurance cover is considerable in the demand of health services (Cameron et al., 1988). Privately insured individuals are associated with less general doctor visits and higher specialist visits (Pohlmeier and Ulrich, 1995).

The remainder of the work is structured as follows: in section 2 we will present the econometric methods used and their more relevant applications. In section 3 we briefly present and describe the data used in this work. In section 4 we analyse and discuss the results obtained, and in section 5 we present the conclusions and final remarks.

3. Econometric Specifications

In this work, we will use the number of visits to a doctor as dependent variable. This variable only assumes non-negative integer values; hence, the specified regression models have taken this into account.

Regression models for counts, like for other limited dependent variables, are nonlinear with many properties and features related to discreteness and nonlinearity (Cameron and Triverdi, 2005). They assume a dependent variable resulting from an underlying discrete probability function (Pohlmeier and Ulrich, 1995).

Linear regression models are not well suited for analysing count data due to two main reasons: they do not guarantee that the predicted value of the dependent variable, given the values of the independent variables, is nonnegative, and they do not take into account important functional forms.

We will begin by explaining the basic models for count data and then extend the analysis to the Hurdle Model and the advanced approach based on Latent Class theory.

Basic Models for Count Data

Our first estimation is the conventional model for count data, the Poisson regression model. This is the oldest parametric count data model and was widely used in several applications. It derives from the Poisson distribution and allows the intensity parameter to depend on regressors. The conditional mean and the conditional variance equal the single parameter μ that fully characterizes the model.

The earliest empirical application is the study of Bortkiewicz (1837) of the annual number of deaths associated with being kicked by mules in the Prussian army.

This is an attractive method for econometric applications mainly because it takes into account the integer properties of count data, it accommodates counts that are aggregated over time periods (Hausman et al., 1984) and the elasticities of the independent variables take a simple form (Grogger and Carson, 1991).

The number of doctor consultations, defined as y, takes one of the values 0,1, 2,… and for a sample of N individuals we observe y_i, with $i = 0,\ldots,N$. The Poisson regression model specifies that y_i given the regressors or set of characteristics x_i is Poisson distributed with density:

$$f(y_i|\mathbf{x}_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{1}$$

where $\mu_i = \exp(x_i'\beta)$.

If the Poisson distribution is adequate, and assuming we have a random sample of y_i and x_i, the maximum likelihood procedure produces a consistent, asymptotically efficient and normal estimator of β (Verbeek, 2004).

As mentioned before, in this regression model y_i has conditional mean and conditional variance equal to $\mu_i = \exp(x_i'\beta)$.

This feature is known as equidispersion and implies a clear restriction in the estimation because in economic count data, namely on health services utilization, the variance usually exceeds the mean. This situation arises as a consequence of the unobserved heterogeneity in the mean event rate of the Poisson parameter across the sample and it is usually called overdispersion (Wedel et al., 1993). Thus, the ordinary Poisson model usually is not suitable.

Despite having this problem, the Poisson regression model can estimate consistently the conditional mean, even if the Poisson distribution is not valid. The first order conditions of the maximum likelihood problem are still valid; we need only to adjust the way in which standard errors are computed. This procedure is known as pseudo-maximum likelihood approach (Gourieroux et al., 1984).

A convenient approach to estimate the variance is the Huber-White sandwich estimator for the variance instead of the traditional calculation:

$$Var_{PML}\left(\hat{\beta}_P\right) = \left(\sum_{i=1}^{n}\mu_i x_i x_i'\right)^{-1}\left(\sum_{i=1}^{n}(y_i-\mu_i)^2 x_i x_i'\right)\left(\sum_{i=1}^{n}\mu_i x_i x_i'\right)^{-1} \tag{2}$$

Even with its robustness, simple computation and attainment of satisfactory results, the pseudo-maximum likelihood approach has the disadvantage of not allowing the computation of conditional probabilities (Verbeek, 2004), e.g. the probability of having one or two visits to the doctor in a certain period of time.

An alternative solution for this problem is to apply full maximum likelihood analysis to the Negative Binomial (NB) regression model, the standard parametric model that accounts for overdispersion.

This is another common regression model applied in count data studies of medical care utilization. Cameron et al. (1988) used it to study the relationship between demand for health care and health insurance in Australia.

The NB regression models can be interpreted as continuous mixture models that can correct the problem of overdispersion and the generating process assumption in the Poisson regression model.

The distribution of y, defined as the number of doctor consultations, can be derived as a compound Poisson process where the Poisson distribution parameter μ is specified to be generated by a gamma distributed random variable (Pohlmeier and Ulrich, 1995).

The density function for the negative binomial model is given by:

$$f(y_i|\mathbf{x}_i) = \frac{\Gamma(y_i+\psi_i)}{\Gamma(\psi_i)\Gamma(y_i+1)}\left(\frac{\psi_i}{\mu_i+\psi_i}\right)^{\psi_i}\left(\frac{\mu_i}{\mu_i+\psi_i}\right)^{y_i} \tag{3}$$

where

$$\Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt$$

is the gamma function, $\psi_i = (1/\alpha)\mu_i^k$ is the "precision parameter" with constant k, $\alpha > 0$ is the overdispersion parameter and $\mu_i = \exp(x_i'\beta)$.

With this specification, the mean function is $E[y_i \mid x_i] = \mu_i$ and the variance is given by $Var[y_i \mid x_i] = \mu_i = \alpha\mu_i^{2-k}$.

Specifying different values for k we get different variants of NB. The Negative Binomial 1 (NB1) model is obtained by specifying $k = 1$ and the Negative Binomial 2 (NB2) model is obtained by setting $k = 0$. With $\alpha = 0$ we get the standard Poisson Model.

Due to the additional parameter, the NB distribution is parametrically richer than the Poisson distribution (Pohlmeier and Ulrich, 1995). Nevertheless, concentration in one single parameter allows for a more parsimonious specification

and as a consequence the choice between the two is not easy when we consider extensions of the basic models.

The Poisson Hurdle or Two-Part Model

The idea behind the Hurdle Model (HM) is precisely that the utilization of health care services can be divided into two stages. The first stage can be seen as the contact decision and the second stage as the intensity of utilization by the patient.

This model is a finite discrete mixture that combines two generating processes with different probability functions: one for the zeros and a one for the positive integers (Cameron and Triverdi, 1998).

Several applications of this method to health care demand have been used. See for example Pohlmeier and Ulrich (1995) and Gerdtham (1997) to the German and Swedish cases, respectively.

For the first part it is usually used a binary model specification (Logit or Probit), Poisson or NB, while the most common specifications for the second part are the Poisson or NB. In this work we will use the Logit model for the contact decision and for the second part a zero-truncated Poisson.

It can be argued that the models should not consider two sources of heterogeneity (the overdispersion parameter in the NB and the finite mixture) and as a consequence one should use Poisson in the second stage (Bago d'Uva, 2005).

Formally the density function for our model is:

$$f(y_i|\mathbf{x}_i)\begin{cases}\Pr[y=0\,|\,\mathbf{x},\beta_1]=1-\dfrac{e^{x'\beta_1}}{1+e^{x'\beta_1}} \\ \Pr[y=1\,|\,\mathbf{x},\beta_1]f(y|y>0,\mathbf{x},\beta_2) \\ \qquad\qquad =\left(\dfrac{e^{x'\beta_1}}{\left(1+e^{x'\beta_1}\right)\left(1-e^{-x'\beta_2}\right)}\right)\left(\dfrac{e^{-x'\beta_2}\left(x'\beta_2\right)^y}{y!}\right)\end{cases} \quad (4)$$

The Poisson Latent Class Model

The NB regression model presented previously can be interpreted as a continuous mixture model. The Latent Class Poisson model (LCM) is an alternative approach that uses instead a discrete mixture representation of unobserved heterogeneity.

The previous HM approach assumes a clear dichotomy between the population of users and nonusers. On the contrary, in the Latent Class formulation the factor that splits the population is the intensity of use of medical services. In the case of a two-point finite mixture model, the dichotomy is made between "high" and "low" users. (Cameron and Triverdi, 1998)

This model assumes that the observations are drawn from a finite discrete mixture of Poisson distributions. These distributions differ in the intercept and in the coefficients of the explanatory variables in the regression component of the model (Wedel et al., 1993). Thus, it accounts for heterogeneity in the coefficients of the Poisson regression models instead of assuming a probability distribution of the mean event rate over the sample done in NB regression model (Wedel et al., 1993).

These models work quite well for count data studies of health care utilization. See for example Deb and Triverdi (1997, 2002), Lourenço and Ferreira (2005), Bago d'Uva (2005), Sarma and Simpson (2006).

We assume that the density of y is a linear combination of S different Poisson densities, with $s = 1,2,\ldots,S$. Thus an S-component finite Poisson[1] mixture density function is given by:

$$f(y_i|\mathbf{x}_i,\mu,\pi) = \sum_{s=1}^{S} \pi_s \left(\frac{e^{-\mu_s} \mu_s^{y_i}}{y_i!} \right) \tag{5}$$

$$\text{with } 0 \leq \pi_s \leq 1 \text{ and } \sum_{s=1}^{S} \pi_s = 1$$

and where $S = 2$ is the unconditional probability that an individual belongs to class s and is known as a point-mass.

Due to the empirical and statistical evidence of two types of users of medical care[2], we will estimate this model with two components. In the case where $S = 2$, the density function is given by:

$$f(y_i|\mathbf{x}_i,\mu,\pi) = \pi \left(\frac{e^{-\mu_1} \mu_1^{y_i}}{y_i!} \right) + (1-\pi) \left(\frac{e^{-\mu_2} \mu_2^{y_i}}{y_i!} \right) \tag{6}$$

with $0 \leq \pi \leq 1$.

1 As in the Hurdle Model, the choice of Poisson in this case is made to consider only one source of heterogeneity.

2 Deb and Triverdi (2002) used only two classes after previous analysis that indicated overparameterization of the model with three classes.

4. Data

Our data source is the National Health Surveys realized in Portugal by the Portuguese Ministry of Health in collaboration with *Instituto Nacional de Saúde Dr. Ricardo Jorge* between October 1998 and September 1999. The size of the sample is 48,607 individuals belonging to 21,808 households. The survey was conducted by direct interviews with the intervention of Portuguese National Statistic Institute (INE), and the sample was distributed in order to ensure an adequate geographical dispersion.[3]

The dependent variable of our study is the number of visits to a doctor in the three months before the interview.

We restrict our analysis to adult individuals, with age higher or equal to 18[4] without any type of long-term incapacity, i.e. not permanently on bed or chair and with mobility not limited to their own house. We have also excluded those individuals that have not reported the variables in which we are interested. We ended up with 36,259 individuals in the analysis after imposing these restrictions.

The independent variables are organized in four groups: demographic variables, socioeconomic variables, health status variables and health insurance variables. Table 1 presents the description of the variables used in the study and Table 2 contains the summary statistics of those variables.

5. Model Estimation and Results

5.1 Model Estimation and Specification Tests

The Poisson, Negative Binomial and Hurdle models were estimated using Stata 9.0 and the Latent class model was estimated using GLIMMIX 3.0.

The first three models were estimated by maximum likelihood method. The likelihood function is the product of the individual densities conditioned on the regressors and the estimator maximizes the log-likelihood function (Cameron and Triverdi, 2005).

The LCM was estimated with Expectation Maximization (EM) algorithm, an iterative computation of maximum likelihood that successively improves upon some sets of starting values of the parameters, and permits simultaneous

3 In "Methodological Note of National Health Care Survey 1998–1999".

4 We are interested to model the behavior of adult individuals and thus this restriction on the age of the population used does not introduce any type of bias in the results.

Table 1: Description of the Variables

Variable	Description
Utilization Variable	
utiliz	Number of visits to a doctor in the last three months
Demographic Variables	
age	Age in years
agesquared	Age in years squared (divided by 100)
male	If the person is male = 1, if female = 0
married	If the person is married = 1, if not = 0
educ	Number of years of education
region1	If the person lives in *Norte* = 1, if not = 0
region2	If the person lives in *Centro* = 1, if not = 0
region3	If the person lives in *Lisboa e Vale do Tejo* = 1, if not = 0
region4	If the person lives in *Alentejo* = 1, if not = 0
Socioeconomic Variables	
unemp	If the person was unemployed in the last two weeks = 1, if not = 0
familyinc1	If total family net income is less than 219€ = 1, if not = 0
familyinc2	If total family net income is between 219€ and 314€ = 1, if not = 0
familyinc3	If total family net income is between 315€ and 422€ = 1, if not = 0
familyinc4	If total family net income is between 423€ and 546€ = 1, if not = 0
familyinc5	If total family net income is between 547€ and 678€ = 1, if not = 0
familyinc6	If total family net income is between 679€ and 814€ = 1, if not = 0
familyinc7	If total family net income is between 815€ and 990€ = 1, if not = 0
familyinc8	If total family net income is between 991€ and 1235€ = 1, if not = 0
familyinc9	If total family net income is between 1236€ and 1681€ = 1, if not = 0
Health Status Variables	
avphysical	Number of days in the last week that the person realized some physical activity
milk	Number of days in the last week that the person drank milk
avcigarr	Average number of packs of cigarrettes smoked per day
alcohol	If the person drank alcoholic drinks more than 1 time per week in the last year = 1, if not = 0
diabetes	If the person has diabetes = 1, if not = 0
asthma	If the person has asthma = 1, if not = 0
chronbronq	If the person has chronical bronquitis = 1, if not = 0
allergy	If the person has any allergy = 1, if not = 0
hypertension	If the person has hyper tension = 1, if not = 0
back	If the person has some problem on the back = 1, if not = 0
Insurance Variables	
priv_ins	If the person uses private insurance = 1, if not = 0
pub_subsist	If the person uses a public health subsystem = 1, if not = 0
priv_subsist	If the person uses a private health subsystem = 1, if not = 0

Table 2: Summary Statistics of the Variables

Variable	Mean	Std.Dev.	Min	Max
Utilization Variable				
utiliz	1.3396	2.0862	0	30
Demographic Variables				
age	48.695	1.8448	18	98
agesquared	2711.5	1.8583	324	9604
male	0.4709	0.4992	0	1
married	0.6726	0.4693	0	1
educ	5.8333	4.4543	0	24
region1	0.2958	0.4564	0	1
region2	0.1943	0.3957	0	1
region3	0.2699	0.4439	0	1
region4	0.1253	0.3311	0	1
Socioeconomic Variables				
unemp	0.4757	0.4994	0	1
familyinc1	0.0820	0.2744	0	1
familyinc2	0.1120	0.3154	0	1
familyinc3	0.1236	0.3292	0	1
familyinc4	0.1225	0.3278	0	1
familyinc5	0.1182	0.3229	0	1
familyinc6	0.1088	0.3114	0	1
familyinc7	0.0926	0.2899	0	1
familyinc8	0.0985	0.2979	0	1
familyinc9	0.0666	0.2494	0	1
Health Status Variables				
avphysical	0.2080	0.8565	0	7
milk	4.6855	3.1609	0	7
avcigarr	0.1814	9.0391	0	5
alcohol	0.3963	0.4891	0	1
diabetes	0.0633	0.2435	0	1
asthma	0.0613	0.2398	0	1
chronbronq	0.0326	0.1777	0	1
allergy	0.1483	0.3555	0	1
hypertension	0.2076	0.4056	0	1
back	0.4802	0.4996	0	1
Insurance Variables				
priv_ins	0.0487	0.2153	0	1
pub_subsist	0.1181	0.3228	0	1
priv_subsist	0.0150	0.1217	0	1

estimation of all model parameters (Wedel and Kamakura, 1999). Its name is due to the fact that iterations of the algorithm consist of an expectation step followed by a maximization step (Dempster et al., 1977).

Models can be compared and evaluated in two levels. First, model selection criteria may be used to choose between the models. Next, a goodness-of-fit criterion can be used to evaluate whether the chosen model offers a good fit to the data (Cameron and Triverdi, 1998).

Table 3: Model Specification Tests

Model	AIC	BIC	CAIC
Poisson	125684.8	125965.3	125998.3
NB	110903.9*	111192.9*	111226.9*
HM	117450.3	118011.2	118077.2
LCM	111764.4	112325.3	112391.3

* Model preferred by all criteria

The Poisson Model is nested in HM and LCM, but HM and LCM are non-nested. This implies that we will have to use adequate methods of evaluation. We will make use of three information criteria[5] to compare the different specifications: the Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC) and the Consistent Akaike Information Criterion (CAIC).[6] The model with smallest value is preferred.

Observing the calculations presented in Table 3 for all criteria, we conclude that the preferred model is the NB. The HM and LCM approaches may be over-parameterized which penalizes the values obtained for the criteria.

Deb and Triverdi (2002) contrasted the LCM with HM applied to health care demand. They found a strong evidence in favor of LCM as compared to

5 These are non-nested model discrimination criteria within the likelihood framework (Cameron and Triverdi, 2005).

6 $AIC = -\ln L + K$, $BIC = -2\ln L + K\ln(N)$, $CAIC = -2\ln L + K[1 + \ln(N)]$ where L is the maximized log likelihood value of the model, K is the number of parameters and N is the number of observations.

HM for counts of utilization, supported by both in-sample and cross-validation model selection tests.

Within the Poisson specifications we select the LCM as preferred. We shall evaluate it in terms of goodness-of-fit. We have computed a pseudo R-squared measure of goodness-of-fit proposed by Cameron and Windmeijer (1997) and given by:

$$R^2_{DEV} = 1 - \frac{\sum_{i=0}^{N}\left[y_i \ln\left(\frac{y_i}{\mu}\right) - (y_i - \mu)\right]}{\sum_{i=0}^{N}\left[y_i \ln\left(\frac{y_i}{\bar{y}}\right)\right]} \tag{7}$$

where $\bar{y}$ is the simple average of the sample and $\hat{\mu}$ is the estimated mean parameter.

The pseudo R-squared shown in (7) is based on the decomposition of the total deviance (a generalization of the total sum of squares used in linear regressions). On the numerator of the quotient there is the deviance in the fitted model and on the denominator is the deviance in the intercept-only model (Cameron and Triverdi, 1998).

The value obtained with LCM is 0,193 and measures the reduction in the deviance due to the inclusion of regressors.

5.2 Discussion of the Results

The discussion of the results will be based mainly in the last estimations that were done, the Hurdle model and the Latent class model. The results obtained with Poisson and Negative Binomial specifications work as benchmarks for comparisons. All the results are presented in Table 4 and the marginal effects are calculated in Table 5.

Our analysis will be organized by groups of variables.

Table 4: Estimation Results by Type of Specification of the Model

	Poisson		Negative Binomial		Hurdle Model				Latent Class Poisson Model			
					1st Stage		2nd Stage		Low Users		High Users	
Variable	Coeff.	(*t*-value)	Coeff.	(*t*-value)	Coeff.	(*t*-value)	Coeff.	(*t*-value)	Coeff.	(*t*-value)	Coeff.	(*t*-value)
Demographic Variables												
age	0.02	7.73	0.02	7.98	0.03	6.26	0.01	5.19	0.04	16.00	0.00	1.25
agesquared	−0.02	(−6.71)	−0.02	(−6.53)	−0.01	(−3.34)	−0.01	(−6.56)	−0.03	(−12.50)	0.00	(−1.89)
male	−0.05	(−2.53)	−0.09	(−5.64)	−0.28	(−10.59)	0.07	5.12	−0.11	(−7.50)	0.11	7.95
married	0.09	4.29	0.11	6.10	0.15	5.14	0.04	2.56	0.09	5.81	0.07	4.74
educ	0.01	1.96	0.01	2.52	0.02	5.29	−0.01	(−2.66)	0.01	5.90	0.00	0.24
region1	0.22	6.88	0.22	8.68	0.43	10.75	0.05	2.51	0.33	13.91	0.04	1.85
region2	0.33	10.27	0.35	12.95	0.56	12.95	0.16	7.46	0.45	18.89	0.25	11.56
region3	0.20	6.47	0.21	8.30	0.38	9.42	0.06	2.83	0.31	12.85	0.08	3.84
region4	0.12	3.38	0.10	3.54	0.24	5.25	0.01	0.46	0.22	8.12	0.00	0.03
Socioeconomic Variables												
unemp	0.39	20.03	0.40	24.25	0.35	13.20	0.35	26.71	0.34	23.54	0.41	29.16
familyinc1	0.06	1.18	0.06	1.37	−0.01	(−0.19)	0.07	2.33	0.03	0.95	−0.08	(−2.57)
familyinc2	−0.01	(−0.24)	0.00	(−0.06)	−0.07	(−1.11)	0.01	0.38	−0.04	(−1.13)	−0.09	(−2.93)
familyinc3	0.07	1.58	0.08	2.22	0.05	0.83	0.08	2.63	0.08	2.60	−0.02	(−0.60)
familyinc4	−0.01	(−0.26)	−0.01	(−0.25)	−0.03	(−0.53)	0.00	0.12	−0.01	(−0.20)	−0.17	(−5.79)
familyinc5	0.00	(−0.10)	0.01	0.22	0.00	0.04	−0.01	(−0.22)	−0.01	(−0.46)	−0.13	(−4.30)
familyinc6	0.03	0.66	0.03	0.90	0.01	0.21	0.04	1.17	0.00	0.14	−0.06	(−2.10)
familyinc7	0.04	0.85	0.04	1.19	0.02	0.41	0.04	1.40	0.04	1.16	−0.04	(−1.32)
familyinc8	0.04	0.95	0.05	1.28	0.08	1.36	0.02	0.51	0.03	0.99	0.01	0.24

Table 4 (continued)

	Poisson		Negative Binomial		Hurdle Model				Latent Class Poisson Model			
					1st Stage		2nd Stage		Low Users		High Users	
Variable	Coeff.	(*t*-value)	Coeff.	(*t*-value)	Coeff.	(*t*-value)	Coeff.	(*t*-value)	Coeff.	(*t*-value)	Coeff.	(*t*-value)
familyinc9	0.06	1.37	0.06	1.63	0.02	0.36	0.08	2.42	0.03	0.93	−0.01	(−0.18)
Health Status Variables												
avphysical	−0.01	(−0.59)	0.00	(−0.29)	0.04	3.14	−0.03	(−4.39)	0.02	2.44	−0.03	(−3.71)
milk	0.01	4.94	0.01	5.67	0.03	8.03	0.00	1.25	0.02	10.40	0.01	3.54
avcigarr	−0.01	(−0.36)	0.01	0.53	−0.06	(−2.19)	0.06	4.20	−0.09	(−5.14)	0.03	2.53
alcohol	−0.24	(−12.20)	−0.24	(−14.16)	−0.35	(−12.92)	−0.14	(−10.83)	−0.28	(−19.72)	−0.16	(−11.58)
diabetes	0.31	12.11	0.36	13.65	0.90	15.16	0.17	10.29	0.36	20.62	0.27	14.32
asthma	0.18	6.44	0.20	7.37	0.38	7.30	0.11	6.12	0.19	9.99	0.17	7.94
chronbron	0.17	4.91	0.19	5.18	0.37	5.15	0.10	4.46	0.20	8.09	0.07	2.13
allergy	0.24	11.44	0.26	13.87	0.40	11.75	0.16	11.76	0.25	17.11	0.25	16.72
hypertension	0.26	14.40	0.29	16.87	0.71	21.72	0.09	7.47	0.33	25.05	0.17	11.58
back	0.33	18.13	0.34	22.10	0.48	19.13	0.21	16.82	0.38	27.65	0.23	16.75
Insurance Variabes												
priv_ins	0.09	2.37	0.10	2.82	0.16	2.94	0.03	1.16	0.10	3.33	0.07	2.10
pub_subsist	0.02	0.87	0.04	1.57	−0.01	(−0.30)	0.05	2.62	0.00	0.23	0.07	3.97
priv_subsist	0.01	0.09	0.01	0.16	0.19	1.97	−0.10	(−1.94)	0.10	2.01	−0.20	(−3.52)
constant	−1.21	(−12.13)	−1.17	(−14.86)	−1.82	(−14.64)	0.04	0.63	−2.16	(−30.07)	1.07	17.42
ln L	−62809.41		−55417.97		−22395.22		−36263.96		−55816.18			
R-Squared	0.08		0.04		0.10		0.04		0.19			
α			0.91									
π									0.91		0.09	

Table 5: Marginal Effects by Type of Specification of the Model

Variable	Poisson	NB	HM		LCM	
			1st St.	2nd St.	Low Us.	High Us.
Demographic Variables						
age	2%	2%	3%	1%	4%	NS
agesquared	–2%	–2%	0%	–1%	–3%	NS
male	–5%	–9%	–24%	7%	–10%	12%
married	10%	12%	16%	4%	9%	7%
educ	1%	1%	2%	–1%	1%	NS
region1	24%	25%	54%	6%	40%	NS
region2	40%	42%	74%	17%	58%	29%
region3	22%	24%	47%	6%	36%	9%
region4	13%	11%	28%	NS	24%	NS
Socioeconomic Variables						
unemp	48%	49%	43%	43%	41%	51%
familyinc1	NS	NS	NS	8%	NS	–7%
familyinc2	NS	NS	NS	NS	NS	–8%
familyinc3	NS	NS	NS	8%	8%	NS
familyinc4	NS	NS	NS	NS	NS	–16%
familyinc5	NS	NS	NS	NS	NS	–12%
familyinc6	NS	NS	NS	NS	NS	–6%
familyinc7	NS	NS	NS	NS	NS	NS
familyinc8	NS	NS	NS	NS	NS	NS
familyinc9	NS	NS	NS	8%	NS	NS
Health Status Variables						
avphysical	NS	NS	4%	–3%	2%	–3%
milk	1%	1%	3%	NS	2%	1%
avcigarr	NS	NS	–6%	7%	–9%	4%
alcohol	–21%	–21%	–30%	–13%	–25%	–15%
diabetes	37%	43%	145%	18%	44%	32%
asthma	20%	22%	47%	12%	21%	19%
chronbron	18%	21%	45%	11%	22%	8%
allergy	27%	30%	49%	17%	28%	28%
hypertension	30%	34%	103%	10%	39%	19%
back	40%	41%	61%	24%	46%	26%
Insurance Variabes						
priv_ins	10%	10%	17%	NS	11%	7%
pub_subsist	NS	NS	NS	5%	NS	8%
priv_subsist	NS	NS	20%	NS	10%	–18%

Observation: NS stands for nonsignificant effect at 5% level

5.2.1 Demographic Variables

Age has a quadratic relationship with health care utilization, a result that is consistent with the general applications of count data models (Windmeijer and Santos Silva, 1997; Pohlmeier and Ulrich, 1995; Cameron et al., 1988; Lourenço and Ferreira, 2005). However, when we consider the LCM[7], the significance disappears for the frequent users' class, i.e. if the person is a high user of health care, his decision does not depend on his age.

Men tend to consume less health care services than women. Once again this result is consistent with previous health care demand studies (Windmeijer and Santos Silva, 1997; Pohlmeier and Ulrich, 1995; Cameron et al., 1988; Deb and Triverdi, 2002). The results with the basic count models show clearly that result, but when we consider the HM and the LCM the results become different.

When we separate the contact decision from the frequency decision we observe that in the first, men tend to do it less than women and in the second tend to do it more. A possible explanation is that men wait longer before seeking health services and therefore they have more conditions that require long term treatment after the first contact (Deb and Triverdi, 1997).

With LCM, the result obtained outlines that in the group of low users men tend to consume less, however, if we analyze the high users group the result is the opposite.

Essentially, on the one hand, men tend to more readily substitute home for market medical care than women but, on the other hand, men expect to suffer larger health losses because of life-style choices (Sindelar, 1982).

The marital status seems to influence the demand for health care in Portugal. The results, in the related literature, usually present this variable as insignificant (Deb and Triverdi, 1997) or significant but low (Windmeijer and Santos Silva, 1997).

In our sample, being married increases the demand for health services in all specifications considered. Considering HM and LCM we observe that it increases both stages/classes but with low magnitude in the second stage and the high users class.

The number of years of education seems to have a negligible effect in health services utilization in Portugal. Even if the coefficients associated with this

7 It is important to refer, in this first allusion to the LCM on the results, that the estimated unconditional probability of low users and high users are approximately 91% and 9%, respectively.

variable are almost always positive and significant they present low magnitudes. The same results have been obtained by Deb and Triverdi (2002) and Pohlmeier and Ulrich (1995).

People with higher education will be able to improve health more efficiently, generating fewer visits to general practitioners. Although, this higher education may be correlated with medical knowledge and as a consequence people with more education years may go to a specialist more often. In our case, the dependent variable aggregates both visits to general practitioners and to specialists and the two effects may cancel.

The last demographic variable considered is the region of inhabitance of the individual. In accordance with NUTS II[8] classification, Portugal has seven regions. The survey does not include the two autonomous regions, *Açores* and *Madeira* and, as a consequence, we use four dummies assuming, without explicit justification, the reference region to be *Algarve*. The other regions are *Norte*, *Centro*, *Lisboa e Vale do Tejo* and *Alentejo*.

We observe that the use of medical services is statistically different across regions. However, we can conclude that the region influences more the contact decision than the choice of visits' frequency to a doctor and seems to influence more low users than high users of health care services.

5.2.2 Socioeconomic Variables

In this group we have only considered unemployment status and net family income, both of them dummies described in Table 1.

Being unemployed increases significantly the utilization of health care in Portugal (approximately 45%), independently of the specification considered. This variable is commonly seen as an income variable, and therefore researchers do not give much attention to it and prefer to explicitly include income variables. In fact, in the few studies where this variable is included it is not significant (Pohlmeier and Ulrich, 1995; Windmeijer and Santos Silva, 1997).

Unemployment status captures not only income but also time effects. A person which is unemployed has more time to spend going to a doctor, but since he is unemployed he probably has less income to spend in health care services.

As we will observe next, net family income is not significant in almost all specifications and thus we should justify the value of the coefficients for unemployment with the dominance of disposable time effect.

8 Official Territorial Nomenclature for Statistical analysis in Portugal.

The high magnitude of the effect of unemployment status suggests that there are other factors influencing their utilization of health services. Being unemployed may be related to some short-term incapacity or disease that can strongly contribute to the increase of visits to doctors. Furthermore, in Portugal the unemployed people have not to pay the usual fee to access health care which can motivate excess consumption.[9]

In the literature, the effect of net family income is either insignificant (Cameron et al., 1988) or has a slightly nonlinear relationship (Windmeijer and Santos Silva, 1997) with number of doctor visits. Pohlmeier and Ulrich (1995) obtain that the number of visits to a general practitioner decrease with income and visits to a specialist increase with income.

For the Portuguese case, this variable has no effect for both the basic specifications and HM. This outcome may be the result of the specification of our dependent variable (number of visits to a doctor) which does not distinguish between types of doctors.

When considering the results obtained with the LCM we obtain that the low users do not modify their number of visits with variations of net income. Observing the behavior of the high users we obtain a kind of nonlinear relationship between income and health care utilization: the use of health care services tends to decrease with income for lower income individuals and to increase with income to higher income classes.

5.2.3 Health Status Variables

In this group of variables we have included some personal habits, like practicing physical activity, drinking milk, smoking and drinking alcohol[10] and dummies referred to chronic conditions like diabetes, asthma, chronic bronchitis or allergy.

The results obtained for the different chronic conditions are in line with all studies (Cameron et al., 1988; Pohlmeier and Ulrich, 1995; Deb and Trivedi, 1997, 2002; Lourenço and Ferreira, 2005) and are empirically reasonable. A person with any chronic disease needs additional health care and as a result tends to increase the visits to a doctor.

9 Called *taxas moderadoras* in the Portuguese Health System.

10 A variable that could have been used in this group is the self-reported health status. However, it was not used because of endogeneity problems: as the health status is self evaluated in the final of the period of survey it is logically influenced by the number of visits to a doctor in that period. See Windmeijer and Santos Silva (1997).

The dummies related to these diseases are all statistically significant and positively related to the explained variable. An interesting remark is the difference in the magnitude of the effects between low and high users; what we obtained is that one person that is a high user tends to increase less the demand of health care services due to chronic diseases than a low user.

The results obtained for personal habits clearly show that with this kind of models we are not estimating "production" of health but utilization of health services. For example, smoking is a personal habit that is associated with worse health status. This fact can motivate utilization of health services in the sense that the person will need more health services to compensate the damage caused by the smoke. As a matter of fact, this personal habit may decrease the demand for health services: it is usually associated with people that do not care much about health and so only in a last situation will go to a doctor.

The effect of the consumption of alcohol shows this second effect. In all specifications it decreases the utilization, more in the contact decision than in frequency, and more in low users than in high users.

When we analyze the effect of practicing physical activity or smoking we observe the two effects. In the contact decision and for low users, smoking will have a negative effect and practicing weekly exercise will have positive effect. In the analysis of the frequency and the determinants for high users we obtain the opposite effect.

5.2.4. Insurance Variables

In this group we have included dummies for the use of private insurance, public subsystems and private subsystems[11], using as the reference class the use of Portuguese National Health System (SNS).

All Portuguese citizens have universal and unrestricted access to SNS, in the sense that the contributions from their own taxes cannot be canceled if one citizen wants to use private insurance instead of the public system. As a result, the dummies included refer to use and not to possession.

These variables try to capture the response of demand to prices (in terms of ease of access and properly in financial terms). As the three insurance options are more favorable than the SNS, mainly in terms of access, we expect the coefficients associated with them to be positive.

11 Subsystems are mechanisms of insurance based on individual professional activity. In the private subsystems class is included only SAMS and in the public subsystems are included ADSE, SSMJ, ADMA, ADFA, ADME, SAD/GNR and SAD/PSP.

Both private insurance and subsystems ensure additional cover and are widely used by approximately 25% of the Portuguese population. They may constitute an object of study about excess of consumption associated with moral hazard problems (Barros, 2005).

The results obtained illustrate this problem. People that use both private insurance and subsystems will tend to increase the health care services utilization, more in the first case than in the second case.

Even though we find these results, we have to be careful in the analysis. It is arguable that the public sector (SNS) does not give a sufficient response to the real needs of the population and as a result, this additional consumption originated by subsystems and private insurance is not totally a result of moral hazard problems (Barros, 2005).

6. Conclusions

The main objective of this work was to understand the determinants of health care utilization in Portugal. This objective was successfully achieved through the use of specific methods that account for the characteristics of our dependent variable.

Within the Poisson regression based models, the Latent Class approach applied to health care demand is seen to dominate the basic approach and the usual specification of HM providing the most accurate estimations and achieving in the best way the purpose of this paper.

Demographic, socioeconomic, health status and individual insurance characteristics are seen as a set of factors that clearly determines the demand for health care services.

For the Portuguese case, demographic factors like age, gender and marital status seem to influence health care utilization more in low users than in high users. Being unemployed strongly increases utilization and income effects are only seen for high users. Chronic conditions increase demand for health services and personal habits like smoking or practising physical activity have mixed effects. Private insurance and the use of public or private subsystems are seen to increase the utilization.

We think that the policy makers in Portugal recognize that the results obtained are useful in the design and reformulation of health care systems to ensure efficiency and equity in the provision of health care services to the population.

Behind this formulation, there are some techniques that could improve future research and obtain even more accurate estimations. Formulations that take into

account endogenous covariates like in Windmeijer and Santos Silva (1997) or interdependence between demands for health insurance and health care (Cameron et al., 1988) are examples of these techniques.

In terms of model specification, a recent approach is obtained by mixing the features of the two main models used in this work, the Hurdle Model and the Latent Class Model. Bago d'Uva (2005) explores a two-stage process of demand within a latent class framework.

References

Bago d'Uva, T. (2005), "Latent Class Models for Utilization of Health Care", *Health Economics*, 14, pp. 873–892.

Barros, P. P. (2005), "Economia da Saúde: Conceitos e Comportamentos", *Almedina*.

Cameron, A. C., P. K. Trivedi, F. Milne, and J. Piggot (1988), "A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia", *Review of Economic Studies*, 55, pp. 85–106.

Cameron, A. C., and P. K. Trivedi (1998), *Regression Analysis of Count Data*, Econometric Society Monographs, vol. 30. New York, Cambridge University Press.

Cameron, A. C., and P. K. Trivedi (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press.

Deb, P., and P. K. Triverdi (1997), "Demand for Medical Care by the Elderly: A Finite Mixture Approach", *Journal of Applied Econometrics*, 12, pp. 313–336.

Deb, P., and P. K. Triverdi (2002), "The Structure of Demand for Health Care: Latent Class versus Two-Part Models", *Journal of Health Economics*, 21, pp. 601–625.

Dempster, A., N. Laird and D. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, pp. 1–38.

Gerdtham, U. G. (1997), "Equity in Health Care Utilization: Further Tests Based on Hurdle Model and Swedish Micro Data", *Health Economics*, 6.

Gourieroux, C., A. Monfort and A. Trognon (1984), "Pseudo Maximum Likelihood Methods: Theory", *Econometrica*, 52, pp. 681–700.

Grogger, J. T., and R. T. Carson (1991), "Models for Truncated Counts", *Journal of Applied Econometrics*, 6, pp. 225–238.

Grossman, M. (1972), "On the Concept of Health Capital and the Demand for Health", *Journal of Political Economics*, 80, pp. 223–235.

Hausman, J., B. H. Hall and Z. Griliches (1984), "Econometric Models for Count Data with an Application to the Patents-R&D Relationship", *Econometrica*, 52, pp. 909–938.

Lourenço, O. D., and P. L. Ferreira (2005), "Utilization of Public Health Centers in Portugal: Effect of Time Costs and Other Determinants. Finite Mixture Models applied to Truncated Samples", *Health Economics*, 14, pp. 939–953.

Pohlmeier, W., and V. Ulrich (1995), "An Econometric Model of the Two-Part Decision Making Process in the Demand for Health Care", *Journal of Human Resources*, 30, pp. 339–361.

Sarma, S., and W. Simpson (2006), "A Microeconometric Analysis of Canadian Health Care Utilization", *Health Economics*, 15, pp. 219–239.

Sindelar, J. L. (1982), "Differential Use of Medical Care by Sex", *The Journal of Political Economy*, 90, pp. 1003–1019.

Verbeek, M. (2004), *A Guide to Modern Econometrics,* 2nd Edition, John Wiley & Sons, Ltd.

Wedel, M., W. S. Desarbo, J. R. Bult and V. Ramaswamy (1993), "A Latent Class Poisson Regression Model for Heterogeneous Count Data", *Journal of Applied Econometrics*, 8, pp. 397–411.

Wedel, M., and W. A. Kamakura (1999), *Market Segmentation: Conceptual and Methodological Foundations*, Dordrecht, Kluwer, 2nd edition.

Windmeijer, F. A., and J. M. C. Santos Silva (1997), "Endogeneity in Count Data Models: An Application to Demand for Health Care", *Journal of Applied Econometrics*, 12, pp. 281–294.

SUMMARY

This work aims to identify the determinants of health care utilization in Portugal and to quantify their effects. A clear idea about this may help the policy debate on the delivery and organization of health services. Toward this end, we developed an approach based on count data models. We found that demographic factors influence health care demand; unemployment strongly increases utilization, and income effects are only observed for frequent users. Furthermore, chronic conditions increase demand for health services, and personal habits have mixed effects. Both private insurance and the use of subsystems increase the utilization.