FIFTEENTH EUROPEAN
WORKSHOP ON
ECONOMETRICS AND
HEALTH ECONOMICS

Econometric analysis of healthcare utilization. An alternative hurdle specification using latent class models

Óscar Domingos Lourenço, Pedro Lopes Ferreira and Pedro Pita Barros

6-9 September 2006

ECONOMETRIC ANALYSIS OF HEALTH CARE UTILIZATION
AN ALTERNATIVE HURDLE USING LATENT CLASS MODELS

Óscar Domingos LOURENÇO[1]; Pedro Lopes FERREIRA[2]; Pedro Pita BARROS[3]

This version: **August 2006**

Comments and suggestions are welcome

Abstract:

The analysis of medical care utilization has usually been made using econometric models, and two types of specifications (one-part *vs* two-part/hurdle models) have disputed the label for the best model to describe the data. The choice of one empirical model over another has usually been essentially an empirical question, being in this domain that this paper contributes to the literature.

Until 1997, when latent class models were introduced to study the subject, hurdle models, when compared to simple one-part models, like Poisson or negative binomial, generally arises as the preferred specification, however, when compared to latent class models, its statistical adequacy to fit the data does not emerge as the better one.

It should, however, be noted that LCM models have usually been compared with one particular hurdle formulation, the one that uses a binary model for the first part and truncated-at-zero negative binomial model for the positive observations. In general the use of the negative binomial is advocated because the specification must include the unobserved factors present in this stage of the model, nevertheless accounting for unobserved heterogeneity in a model for the positives is not a linear task and some pitfalls may come up during specification process. In this paper we argue that the popular hurdle specification accounts incorrectly for the unobserved heterogeneity, causing a misspecification in the model for the positives.

Accordingly, in this paper we propose a new specification for the hurdle model which we believe to be the correct. The innovative hurdle specification suggested is different from the popular hurdle only in the model for the second stage, where we propose a correct specification based on latent class models, gaining statistical flexibility and departing from strong distributional assumptions in the modelling process.

After comparing the new hurdle with other competing count data models using information criteria statistics, Vuong and GoF tests, we have found that our new hurdle specification outperforms all other competing models, including the traditional hurdle and Latent class models.

Keywords: Health care utilization, hurdle models, latent class models, unobserved heterogeneity.

1) Corresponding author: Centre of Studies and Research in Health of the University of Coimbra. Avenida Dias da Silva, nº 165, 3004-512, Coimbra. e-mail: osl@fe.uc.pt

2) Centre of Studies and Research in Health of the University of Coimbra:

3) Universidade Nova de Lisboa

1. Introduction

Over the course of the last thirty years, or so, two major classes of econometric specifications have tended to dominate the empirical literature on medical care utilization[1]; we are referring to one-part models and to two-part models. The debate over the merits of each approach to model health care utilization has been intense and interesting.

One-part models can be considered as specifications based on Grossman's Human Capital model (Grossman, 1972, Wagstaff, 1986). As is well known, in Grossman's framework the individual is taken as the primary decision maker, fully controlling the choices regarding medical care. Basically, one-part regression models are regression models, with a linear or non-linear reduced form equation, where a dependent variable, which represents medical care utilization, are explained as a function of a set of medical care determinants. Examples of this approach is provided by Cameron *et al.* (1988) and Vera-Hernandez (1999), who used insights provided by Grossman's model to develop a '*theoretical framework*' in which frame individual behaviour. Subsequently, the '*theoretical model*' is used as a justification to the empirical regression model. Conversely, two-part models belong to the other class of specifications widely use to explain medical care utilization. Early applications of this model were justified only by statistical reasons. The model was initially proposed to deal with the characteristics of the medical care consumption indicators, namely the high number of individuals reporting as non-users of health care (Duan et al., 1983, Manning et al., 1987). Later on, Pohlmeier and Ulrich (1995) framed this class of models in the principal-agent set-up. Under this framework, the patient is no longer the unique and sovereign decision maker, as he transfers to the doctor the responsibility about the amount of medical services to consume. The empirical counterpart of the principal-agent model is the two-part model, which assumes a two part decision structure, with different decision agents in each part of the process. The first decision, the contact to a physician, is controlled by individual, hence Grossman's type approach is likely to be relevant (Wagstaff, 1986). After the contact decision, the choices necessary at the second stage are taken essentially by the physician, possibly including patient preferences, therefore, Grossman type models are likely to be less relevant in this stage. The two components of this dual process might diverge in the respective economic determinants and can provide different evidence to policy making.

[1] Early studies on medical care demand have relied also on other statistical methods other than econometric specifications, for example, analysis of variance and analysis of covariance. A review of early empirical methods for medical care utilization study can be found on DUAN, N., MANNING, W. G., MORRIS, C. N. & NEWHOUSE, J. P. (1983) A comparison of alternative models for the demand for medical care. *Journal of business and Statistics,* 1, 115-126.

In summary, both types of empirical specifications have a theoretical support, however, the lack of consensus about the appropriate framework, Grossman like models or principal-agent models, to represent individual behaviour concerning medical care decisions has been the norm. Therefore, the choice of one econometric specification over another has usually been essentially an empirical question being in this domain that this paper makes a contribution to the literature. To pursue this goal we use data taken from the Portuguese National Health Survey (NHS, 1999) to estimate and tests statistical hypothesis on a wide variety of regression models specified in the spirit of one-part model and on the spirit of two part models. Our indicator of health care utilization is the total number of doctor visits in a period of three months.

Some authors advocate the utilization of two-part models, also referred to as hurdle models in count data, because they are the empirical counterpart of the principal agent set-up, which, they believe, represent well the actual decision process (Jimenez-Martin et al., 2002, Pohlmeier and Ulrich, 1995). In fact, we also consider that hurdle models are more appealing specifications to explain medical care utilization because it reflects more accurately the decisional structure regarding health care choices. Moreover, relative to simpler models, the hurdle can be more enlightening as it allows separating and quantifying the determinants of medical care concerning the decision to see the physician as well as the frequency of visits, what can be significant for health policy making (Pohlmeier and Ulrich, 1995 113).

Until 1997, when Deb and Trivedi introduced latent class models in the study of this theme, the hurdle framework was viewed as the appropriate empirical methodology to explain the usage of medical care services, essentially because it had a theoretical support, and also due to its statistical properties, which seem appropriate to deal with the characteristics of medical care data. In fact, when compared to simple one-part models, like Poisson or negative binomial, the hurdle model generally arises as the preferred specification (Gerdtham, 1997, Deb and Trivedi, 1997, Van Ourti, 2004, Deb and Trivedi, 2002).

Deb and Trivedi (1997, , 2002) did some criticisms to the hurdle specification and proposed an alternative econometric framework to study medical care utilization. The authors advocated that latent class models (henceforth referred to as LCM) present statistical properties that makes the model more appealing to study utilization data, and in fact, they provided evidence favouring the LCM specification over the hurdle model. Other authors, also compared the statistical performance of LCM with hurdle and found similar results (Gerdtham and Trivedi, 2001, Sarma and Simpson, 2006). Therefore, the empirical evidence seems to make the case that LCM, making use of a combination of several statistical processes in the same model gaining in this way additional statistical flexibility, offer always a better framework to analyse health care utilization data.

A question now arises: how to deal with this apparent contradiction between theory — we believe that the decision making process regarding health care choices should be framed in the principal-agent set-up — and empirical findings showing that one-step models, namely the LCM, provide better fit relative to the hurdle model? To begin with, there is evidence, although scarce, pointing toward the statistical gains of the hurdle specification over LCM. For instance, Jiménez-Martín et al. (2002), using data from the European Community Household Panel estimated hurdle and LCM models to evaluate the determinants of individual utilization of medical care using several utilization measures, namely, the number of visits to a GP and the number of visits to a specialized doctor. They found mixed evidence concerning the model that better describes the data. LCM models were found to be more suitable than hurdle models when the dependent variable is the number of visits to a General Practitioner, while the opposite is reported for visits to specialist physicians. Winkelmann (2004) have also found that hurdle specification performs better than LCM models.

We note that LCM models have usually been compared with one particular hurdle formulation, the one that uses a binary model for the first part and truncated-at-zero Poisson or negative binomial model for the positive observations. In general the use of the negative binomial is advocated to deal with the positives because the specification must include the unobserved factors present in this stage of the model (Pohlmeier and Ulrich, 1995). However, accounting for the presence of unobserved heterogeneity for the positives — in a truncated-at-zero sample — is not a linear task, and some pitfalls may come up during the specification process (Santos-Silva, 2003, Lourenço and Ferreira, 2005). Therefore, in the hurdle model, in part of specifying a model for the second stage, the way one account for the unobserved heterogeneity is not irrelevant and it may affect model performance. In fact, Santos-Silva (2003) argues that in the popular hurdle model, the one based on the negative binomial probability function, the unobserved heterogeneity is incorrectly modelled, thus, causing the hurdle model to be misspecified. This potential misspecification can be one explanation for the poor statistical performance of the popular hurdle model when compared to the LCM.

Accordingly, in this paper we propose a new specification for the hurdle model which we believe to be the correct. The hurdle model that we are suggesting in this paper is different from the popular hurdle only in the formulation of the model for the second stage. In this stage, we suggest a specification that uses latent class models, therefore, in this way we account for the presence of unobserved heterogeneity in a flexible way and not depending on strong distributional assumptions.

Therefore, the contribution of this paper is twofold: On the one hand we suggest an alternative hurdle formulation, namely, a new way to account for the unobserved heterogeneity in the specification of the model for the positives. Moreover, we compare the statistical performance of our alternative hurdle with the performance of popular hurdle models as well as LCM models, using information criteria statistics, Vuong and GoF tests. These comparisons may allow us to reconsider the relevance or irrelevance of the principal-agent models as theoretical frameworks framing individual behaviour. On the other hand, we assess the stability of the effect of various covariates, e.g. income, health insurance generosity and rural status, across the various models under evaluation, using average marginal effects statistics. The results of these comparisons may be relevant to evaluate if models specified under different assumptions about the unobserved heterogeneity are stable, or, on the contrary, change from model to model.

The paper is organized as follows: Section 2 aims at presenting some aspects of the main empirical specifications that have been used to model medical care utilization. We will give emphasis to mixture models. Section 3 proposes a new hurdle formulation, combining insights from hurdle and latent class models. The next section, 4, presents the dataset used, the dependent as well as the independent variables selected for this empirical application. Section 5 presents and discusses some results.

2. Empirical specifications to analyse health care utilization: Emphasis on mixture models

In the analysis of count data the Poisson regression model (PRM) is usually indicated as the reference model. However the model is usually unsatisfactory in fitting real data, mainly because real data are '*overdispersed*' and present a proportion of zeros (*excess zeros*) inconsistent with the PRM . Mullahy (1997) points out that *overdispersion* and *excess zeros* are consequences of unobserved heterogeneity, which must be included in the model, otherwise the estimates will lose efficiency, inducing biases on variances and, consequently, on testing procedures (Cameron and Trivedi, 1996, Gourieroux and Visser, 1997). These considerations motivate the utilization of specifications that include unobserved heterogeneity.

Mixture models: general definition

One well known approach to account for the presence of unobserved heterogeneity in health care utilization is to assume that the data generating process (dgp) is a mixture model. Two types of mixture models can be considered: continuous and discrete mixtures.

Lindsay and Lesperance (1995) define a continuous mixture as a probabilistic model where the conditional density of y_i is defined as

$$f(y_i \mid x_i, \beta, \gamma) = \int_{\upsilon_i} f(y_i \mid x_i; \upsilon_i, \beta) * h(\upsilon_i \mid x_i, \gamma) d\upsilon_i \qquad [1]$$

where $h(\upsilon_i \mid x_i, \gamma)$ is the mixing distribution.

When the unobserved heterogeneity is assumed to be represented by a discrete random variable, with an unknown number *(P)* of support points, the conditional probability function of y_i is given by

$$f(y_i \mid x_i, \beta) = \sum_{j=1}^{P} \pi_j f_j\left(y_i \mid x_i, \upsilon_i^j, \beta_j\right) \qquad [2]$$

This formulation assumes that $f(y_i \mid x_i, \beta)$ is a convex linear combination of the component distributions $f_j\left(y_i \mid x_i, \upsilon_i^j, \beta_j\right)$ $j = 1 \ldots P$, in proportions $\pi_1, \cdots, \pi_P$, such that $\sum_{j=1}^{P} \pi_j = 1,\ \pi_j \geq 0$. Particular instances of these two approaches have been used to model healthcare utilization. The most well known are the negative binomial model and latent class models.

The Negative Binomial Model

The negative binomial distribution (henceforth NB) can be derived in a number of ways however, perhaps most popular is obtained as a continuous mixture model allowing the mean parameter of the PRM to vary randomly across the population according to a gamma distribution (Cameron and Trivedi, 1998).

Following Deb and Trivedi (2002), the NB probability function of y_i, conditional on x_i, can be written as,

$$f(y_i \mid x_i) = \frac{\Gamma(y_i + \eta_i)}{y_i! \Gamma(\eta_i)} \eta_i^{\eta_i} (\lambda_i + \eta_i)^{-(\eta_i + y_i)} \lambda_i^{y_i} \quad y_i = 0, 1, 2 \ldots \qquad [3]$$

where $\Gamma(.)$ is the gamma function, $\lambda_i = \exp\left(x_i^{'}\beta\right)$ and $\eta_i = \left(\frac{1}{\alpha}\right)\lambda_i^k$. The two most applied versions of the NB are obtained setting $k = 1$ (NB1), or $k = 0$ (NB2).

The NB family of densities is considered as the most general and flexible discrete distribution, nevertheless its utilization in applied work may present some weaknesses. First, it is a fully parametric model relying on explicit assumptions about the distribution of unobserved heterogeneity, and economic theory that gives insight about the unknown functional form for the distribution is often lacking (Wedel et al., 1993); second, in the NB the zeros and positives are assumed to share the same dgp, therefore, if the impact of a covariate differs across the support of the dependent variable, the NB model does not capture that different impact (Deb and Trivedi, 2002, Winkelmann, 2004). Hence, empirical health economists have suggested the formulation of more general count data models.

Latent class models

A more recent approach to address the problem of unobserved heterogeneity uses latent class models (LCM). A Latent Class Model (LCM) specification arises when the random variable that represents the unobserved heterogeneity, υ_i, is assumed to be discrete with P support points. In this formulation it is implicit the assumption that the population consists of P homogeneous, relative to the unobserved factors, latent subpopulations. Latent class models were first applied to the study of health care data by Deb and Trivedi (1997) and since then, a number of other empirical applications have appeared in the literature (Deb and Trivedi, 2002, Deb and Trivedi, 1997, Deb, 2001 #58, Bago d'Uva, 2006, Bago d'Uva, 2005, Lourenço and Ferreira, 2005, Jimenez-Martin et al., 2002, Atella et al., 2004, Gerdtham and Trivedi, 2001).

Under the LCM framework, the conditional probability function of y_i is given by yhe probability function presented in 2.

The advantages of using Latent Class Models over the use of continuous mixture models has been emphasized by a number of authors therefore it is worthless to reproduce them here (Deb and Trivedi, 2002, Cameron and Trivedi, 1998, Wedel et al., 1993, Heckman and Singer, 1984). However, Jimenez Martin el al. (2002) while agreeing that latent class presents good statistical properties, also criticize its application because it is driven only by statistical reasons, not being the empirical counterpart of an economic model.

One can argue that all models presented so far, PRM, NB and the standard LCM are specified in the spirit of one-part models. In this models the impact of individual and doctor inputs to the decision making process regarding the number of visits, are entangled, therefore, non-separable. Even in the LCM, which assumes that two stochastic processes govern health care choices, individual and doctor contributions to the choice are mixed in each stochastic process. In this situation it is difficult to know who, doctor or patient, contributes more to make decisions regarding the number of visits. Therefore, the evolvement from the PRM to the NB and, from the later to the standard LCM has been motivated for statistical reasons, implicitly neglecting the economic motivation behind model specification. It has been widely acknowledged that in the context of healthcare choices, due to the asymmetry of information between patient and physician, the decision process involve two stages, with different key decision makers in each step (Zweifel, 1981). Accordingly, it has been argued that the specification of empirical models should recognize this two stage decisional structure.

The Hurdle regression model

Cragg (1971) proposed an econometric specification suited to analyse consumption decisions that can be considered as being made in two stages. Later, Mullahy (1986) suggested a model with a similar decision structure adapted to a count variable.

In the case of health care utilization this type of two-stage decision process still to be pertinent because the key decision makers in each stage are different. In the first stage the decision to seek medical care is individual based, while in the second stage decisions are mainly doctor based, possibly including patient preferences (Pohlmeier and Ulrich, 1995, Santos-Silva and Windmeijer, 2001).

For a general formulation of the hurdle model, let y_i, denote the count variable and $x_i' = [x_{i1}, x_{i2}, \cdots, x_{ik}]$ a $(1xk)$ vector of covariates. Assume also that $f_1(.)$ and $f_2(.)$ are discrete probability functions, where $f_1(.)$ governs the first part of the model and a truncated-at-zero version of $f_2(.)$ governs the process after the hurdle has been crossed. Then, the probability function of the hurdle model is given by,

$$f(y_i \mid x_i;\beta_1,\beta_2) = \begin{cases} f_1(0 \mid x_i;\beta_1) & if \quad y_i = 0 \\ [1 - f_1(0 \mid x_i;\beta_1)] * f_2(y_i \mid x_i;\beta_2, y_i > 0) & if \quad y_i = 1, 2, \ldots, \end{cases} \qquad [4]$$

Poisson and NB probability functions are common choices for $f_1(.)$, being also possible to specify the first part as binary models like probit and logit. For $f_2(.)$ the most popular alternatives are Poisson and NB (Winkelmann, 2004). The main reason driving the utilization of the NB probability function in the second stage, instead of the Poisson, is that the modelling in this stage needs to account for the presence of unobserved heterogeneity, that is to be likely present because household micro-data hardly measures supply side influences regarding medical care decisions about the duration of the treatment (Pohlmeier and Ulrich, 1995).

From the above we have learned that hurdle specifications has a two-fold appeal to explain health care utilization. On one hand it is an empirical model whose structure is suggested by a 'theoretical' model, and on the other hand it has a good statistical performance in explaining health care utilization (Cameron and Trivedi, 1998, Pohlmeier and Ulrich, 1995). Despite this double advantage of the hurdle specification, it has been subjected to some critical observations.

The popular hurdle specification: some comments

The popular hurdle specification has been subjected to some critical remarks, which can be broadly classified into two groups: In the first group we include data related problems, while in the second group we incorporate specification issues objections.

The data related problems were fully addressed by Pohlmeier and Ulrich (1995). In short, they arise because the hurdle model assumes that during the period of analysis individuals go through one illness spell only. This assumption is verified when the data regarding medical consultations is measured per illness episode and not during a fixed time period as occurs in cross-section datasets. If the single illness spell is violated then the second stage parameters are not identified (Santos-Silva and Windmeijer, 2001). Regarding this difficulty, we follow Pohlmeier and Ulrich (1995) and Gerdtham (1997) and assume that the occurrence of multiple illness spells is a rare event, therefore, all parameters of the model will be identified. Note that in our application the observation period used to capture doctor visits is short (3 months), meaning that the occurrence of multiple illness spell are less likely to occur. In addition, roughly 70% of the individuals had less than one visit, thus we believe that this provides enough evidence to support the assumption of a single illness spell during the observation period.

The second group of unfavourable comments are related to specification issues, focusing primarily on the difficulties to account for the unobserved heterogeneity in the specification of the second part of the hurdle. As was mentioned previously, in applied work it has been common to specify the second stage of the hurdle as a truncated-at-zero Poisson or NB. Gurmu (1997) object to the utilization of truncated-at-zero NB on the grounds that the model rests on the explicit assumption that unobserved heterogeneity is Gamma distributed, what can be considered arbitrary especially without any prior information about the true distribution of the unobserved heterogeneity. In the event of a misspecification of the unobserved heterogeneity, then the estimation process would lead to inconsistent estimates. In an attempt to respond to some of these criticisms, Gurmu (1997) proposed a semi-parametric mixture hurdle model, thus not requiring prior knowledge about the distribution of unobserved heterogeneity, however, even Gurmu's alternative hurdle model may misspecify the part for the strictly positive counts. This is a much more subtle form of misspecification that may arise when modelling the positives in hurdle models. This issue is closely related to the specification and estimation of models for truncated counts with unobserved heterogeneity.

Models for truncated counts: short notes

As is well known, the application of standard count data models to truncated samples leads to inconsistent parameter estimates, thus, suitable modification of standard count data models have to be made to make valid inference (Cameron and Trivedi, 2005, Grogger and Carson, 1991, Gurmu and Trivedi, 1992, Santos-Silva, 2003).

Let $f(y_i \mid x_i)$ represent the density function of the i^{th} person in the actual population, then, the probability function of y_i in the sample is given by,

$$f(y_i \mid x_i, y_i > 0) = f_s(y_i \mid x_i) = \frac{f(y_i \mid x_i)}{P(y_i > 0 \mid x_i)} \quad y_i = 1, 2, \ldots \qquad [5]$$

where $f(y_i \mid x_i)$is the probability function in the actual population (Grogger and Carson, 1991, Brannas and Rosenqvist, 1994, Gurmu and Trivedi, 1992). Given the likely presence of unobserved factors affecting medical care utilization, when modelling truncated data the researcher must explicitly account for its presence. As pointed out in a previous section, one natural way of accounting for the presence of unobserved heterogeneity is through the use of mixture models, therefore, mixture models for truncated counts should be specified (Santos-Silva, 2003, Cameron and Trivedi, 1998, Grogger and Carson, 1991, Brannas and Rosenqvist, 1994). However the specification of mixture models in truncated samples is not as linear as in the standard case thus careful analysis is required.

Specification of mixture models for truncated samples

Santos-Silva (2003) studied the impact of endogenous sampling, of which truncation is a specific instance (Cameron and Trivedi, 2005), in the distribution of the unobserved heterogeneity. The author highlighted that in truncated samples the investigator may account for the unobservables in two different ways, and the choice of the correct one can be an important issue.

As was shown in the previous section, the suitable probability model to analyse truncated-at-zero data is given by expression 5. Consider now that the researcher intends to include unobserved heterogeneity in the specification. Accordingly, the probabilistic model in the actual (overall) population, $f(y_i \mid x_i)$, if specified as an LCM, is expressed as,

$$f(y_i \mid x_i, \beta) = \sum_{j=1}^{P} \pi_j f_j\left(y_i \mid x_i, \beta_j, \upsilon_i^j\right) \qquad [6]$$

Therefore, it is immediate that,

$$P(y_i > 0 \mid x_i, \beta) = \sum_{j=1}^{P} \pi_j P\left(y_i > 0 \mid x_i, \beta_j, \upsilon_i^j\right) \qquad [7]$$

Plugging equations 6 and 7 into equation 5, under the LCM framework, the density in the sample, can be written as,

$$f_s(y_i \mid x_i, \beta) = \frac{\sum_{j=1}^{P} \pi_j f_j\left(y_i \mid x_i, \upsilon_i^j, \beta_j\right)}{\sum_{j=1}^{P} \pi_j P\left(y_i > 0 \mid x_i, \upsilon_i^j, \beta_j\right)} \qquad [8]$$

The above expression can also be expressed as

$$f_s(y_i \mid x_i, \boldsymbol{\beta}) = \sum_{j=1}^{P} \left(\frac{f_j\left(y_i \mid x_i, \upsilon_i^j, \boldsymbol{\beta}_j\right)}{1 - \left(\sum_{j=1}^{P} \pi_j P\left(y_i = 0 \mid x_i, \boldsymbol{\beta}_j, \upsilon_i^j\right)\right)} \right) \pi_j \qquad [9]$$

Working out equation 9, it can be re-expressed as 12, presented below,

$$f_s(y_i \mid x_i, \boldsymbol{\beta}) = \sum_{j=1}^{P} \left(\frac{f_j\left(y_i \mid x_i, \upsilon_i^j, \boldsymbol{\beta}_j\right) * \left[1 - P\left(y_i = 0 \mid x_i, \boldsymbol{\beta}_j, \upsilon_i^j\right)\right]}{\left[1 - \left(\sum_{j=1}^{P} \pi_j P\left(y_i = 0 \mid x_i, \boldsymbol{\beta}_j, \upsilon_i^j\right)\right)\right] * \left[1 - P\left(y_i = 0 \mid x_i, \boldsymbol{\beta}_j, \upsilon_i^j\right)\right]} \right) \pi_j \qquad [10]$$

$$= \sum_{j=1}^{P} \left(\frac{f_j\left(y_i \mid x_i, \upsilon_i^j, \boldsymbol{\beta}_j\right)}{1 - P\left(y_i = 0 \mid x_i, \boldsymbol{\beta}_j, \upsilon_i^j\right)} \right) \hat{\pi}_j \qquad [11]$$

$$= \sum_{j=1}^{P} f_j^s\left(y_i \mid x_i, \upsilon_i^j, \boldsymbol{\beta}_j\right) \hat{\pi}_j \qquad [12]$$

In 12, $f_j^s\left(y_i \mid x_i, \upsilon_i^j, \beta_j\right)$ is the probability function of the j component distribution in the sample,

$$f_j^s\left(y_i \mid x_i, v_i^j, \beta_j\right) = \frac{f_j\left(y_i \mid x_i, \upsilon_i^j, \beta_j\right)}{1 - P\left(y_i = 0 \mid x_i, \upsilon_i^j, \beta_j\right)} \qquad [13]$$

and the mixing probabilities, in the truncated population are now given by,

$$\hat{\pi}_j = \frac{1 - P\left(y_i = 0 \mid x_i, \upsilon_i^j; \boldsymbol{\beta}_j\right)}{1 - \left(\sum_{j=1}^{P} \pi_j P\left(y_i = 0 \mid x_i, \boldsymbol{\beta}_j, \upsilon_i^j\right)\right)} \pi_j \qquad [14]$$

Summing up, to account for the presence of unobservable factors in truncated datasets the density of the count in the sample, $f_s\left(y_i \mid x_i\right)$, can be written in two different ways

$$f_s(y_i \mid x_i, \boldsymbol{\beta}) = \sum_{j=1}^{P} \left(\frac{f_j\left(y_i \mid x_i, \upsilon_i^j, \boldsymbol{\beta}_j\right)}{1 - \left(\sum_{j=1}^{P} \pi_j P\left(y_i = 0 \mid x_i, \boldsymbol{\beta}_j, \upsilon_i^j\right)\right)} \right) \pi_j \qquad [15]$$

and

$$f_s(y_i \mid x_i,\beta)=\sum_{j=1}^{P} f_j^s\left(y_i \mid x_i,\upsilon_i^j,\beta_j\right)\hat{\pi}_j \qquad [16]$$

where $f_j^s\left(y_i \mid x_i,\upsilon_i^j,\beta_j\right)$ and $\hat{\pi}_j$ are defined above.

Equation 15 assumes that the unobserved factors belong to the individuals present in the actual population, thus this specification assumes that the researcher are making assumptions about the distribution of the unobserved heterogeneity in the overall population. On the other hand, equation 16 assumes that the unobserved factors belong to the individuals present in the truncated population, hence, making assumptions about the distribution of the unobserved heterogeneity in this population. These two possibilities to specify the distribution of the unobservables, therefore, the distribution of the positive counts, leads the analyst to a cross-road regarding the choice of the proper specification to model the data. Is it indifferent to choose between 15 and 16? Santos-Silva (2003) point out that it is not indifferent, indicating also that care is needed in deciding which model is the most suitable to fulfil the aims of the analysis.

Specification $f_z(y_i \mid x_i,\beta)=\sum_{j=1}^{P}\left(\frac{f_j\left(y_i \mid x_i,\upsilon_i^j,\beta_j\right)}{1-\left(\sum_{j=1}^{P}\pi_j P\left(y_i=0 \mid x_i,\beta_j,\upsilon_i^j\right)\right)}\right)\pi_j$ is the proper formulation when the population of interest is the actual population, while the researcher should use $f_s(y_i \mid x_i,\beta)=\sum_{j=1}^{P} f_j^s\left(y_i \mid x_i,\upsilon_i^j,\beta_j\right)\hat{\pi}_j$ if the empirical analysis aims at analysing the population induced by the sampling scheme, in this case, the truncated population (Santos-Silva, 2003).

There is an intuition behind this result. On the one hand, when the study aims at analysing the actual population, the researcher assumes that the unobserved factors are in the overall population, thus, the unobserved factors aggregate in the overall population to generate the latent classes. Hence it makes sense first to specify a mixture model in the actual population and only after that truncate the mixture, resulting equation 15. On the contrary, when the target of the study is the truncated population, then the analyst assumes that the unobserved factors are in this population. In this event the unobserved factors aggregate in the population of positives to form latent classes (of users), hence, one should at first to specify a truncated distribution to represent each latent class of users, and only after that mixture the truncated distributions that represent each latent class of users, resulting Equation 16.

3. An alternative hurdle formulation using latent class models

Given the reasoning presented in the previous section, it is reasonable to ask: which mixture specification, given by 15 or 16, is more suitable to use when one intends to specify the second part of the hurdle model? In the case of hurdle models, it is natural to estimate the model for the positive counts by making assumptions concerning the distribution of the unobserved heterogeneity in the truncated population (Santos-Silva, 2003), thus, the LCM for the positives should be specified according to 16, that is,

$$f_s(y_i \mid x_i, \beta) = \sum_{j=1}^{P} f_j^s\left(y_i \mid x_i, \upsilon_i^j, \beta_j\right)\hat{\pi}_j = \sum_{j=1}^{P}\left(\frac{f_j\left(y_i \mid x_i, \upsilon_i^j, \beta_j\right)}{1-P\left(y_i = 0 \mid x_i, \beta_j, \upsilon_i^j\right)}\right)\hat{\pi}_j \qquad [17]$$

However the popular hurdle formulation specifies the second stage according to 15, meaning that the researcher are making assumptions regarding the unobserved heterogeneity in the actual population, and consequently, according to Santos-Silva's, are making inference to the actual population and not the truncated population as it should be in the case of hurdle models.

From this discussion we can conclude that in hurdle contexts the inclusion of the unobserved heterogeneity in the model for the positives should be done following clear specification hypothesis. The assumed dgp for the positives must be specified using 16, implying that models using specification 15 are, by definition, misspecifed, therefore inappropriate in hurdle contexts. Consequently, one can argue that the popular hurdle specifications based on the NB distribution, when the use of this density is justified as a manner to account for the presence of unobserved heterogeneity in the second part of the model, are misspecified.

Mullahy (1986), Pohlmeier (1995), Gerdtham (1997), Jimenez-Martin (2002) and Grootendorst (2002) are only some examples of applications of hurdle models specified in such, supposedly, inadequate way. Even the semi-parametric hurdle model proposed by Gurmu (1997) use the continuous counterpart of density 15 to model the positives.

One exception of a hurdle specification that departs from this popular specification is the hurdle formulation proposed by Winkelmann (2004). The author specifies the second stage of the model according to the continuous counterpart of 16, thus correctly pointing towards the truncated population. Winkelmann's proposes a probit for the first part and a truncated Poisson-log-normal model for the strictly positive observations. For the second part the model initially specifies a truncated Poisson, and next assumes that the unobservables, present in the truncated population, are normally distributed. Regarding the statistical performance of this hurdle, Winkelmann concludes that it offers a substantial improvement over all other models[2]. This finding can be considered evidence in favour of Santos-Silva thesis that, in the case of hurdle models, the hypothesis regarding the distribution of the unobservables should be made in the truncated population. We note that Winkelmann's specification is fully parametric, consequently may be somewhat arbitrary. The author do not offer any justification to support the assumption that the unobserved heterogeneity in the truncated population follows a normal density. Hence, we consider that a specification less dependent on strong distributional assumptions is fully desirable.

Bago d'Uva (2006) pointed out that the models to analyse medical care utilization are not constrained to be only hurdle or only LCM being possible to combine the features of both formulations. Although she mentions that combining hurdle and LCM are restricted to be applied in panel data contexts, otherwise, one would face identification problems, this is only true in her formulation.

In fact, the new hurdle that we propose in this paper combines features of both formulations, hurdle and LCM, and can be estimated with cross-section data. It merges the original hurdle formulation as suggested by the principal-agent set-up with LCM specifications to deal with the positives. Basically, we merely propose a modification in the approach to treat the individuals with a positive number of visits. The suggestion is to use the LCM framework to specify a density for the strictly positive utilization, providing in this manner statistical flexibility to the second stage of the hurdle. Note that our approach respond to at least one criticism that Deb and Trivedi (2002) made to the popular hurdle. The authors expressed a preference for the hurdle over LCM because the consequences of a misspecification of the dgp will be smaller in the case of LCM, as it can serve as a better approximation to any true, but unknown, probability density. In our model, because we are specifying a hurdle based on LCM, the consequences of a misspecification will be similar.

For a formulation of the model, let, y_i , denote the dependent variable and $x_i' = [x_{i1}, x_{i2}, \cdots, x_{ik}]$ a $(1xk)$ vector of covariates. Furthermore, assume that both $f_0(.)$ and $f_j(\cdot)$, $j = 1 \ldots P$ are discrete probability functions. Assume that $f_0(.)$ governs the hurdle part and a truncated-at-zero LCM governs the process after the hurdle has been crossed. Under these conditions, the probability function y_i is given by,

[2] The competing models estimated by Winkelmann were: Poisson, Negative Binomial, Poisson log-normal, Hurdle negative binomial, Probit-Poisson-log-normal, two components finite mixture negative Binomial, and a multi-episode Poisson logarithmic

$$f\left(y_i \mid x_i;\beta_0,\beta_j\right)=\begin{cases} f_0\left(0 \mid x_i;\beta_0\right) & if \quad y_i=0 \\ \left[1-f_0\left(0 \mid x_i;\beta_0\right)\right]*\left[\sum_{j=1}^{P}\left(\frac{f_j\left(y_i \mid x_i,\beta_j,u_i^j\right)}{1-P\left(y_i=0 \mid x_i,\beta_j,u_i^j\right)}\right)\hat{\pi}_j\right] & if \quad y_i=1,2,... \end{cases} \quad [18]$$

In the first step a binary model is estimated, while in the second step, constraining our sample to the individuals with a positive number of doctor visits, we estimate a model that assumes that the density in the truncated sample is given by a latent class model specified according to 16.

The construction of the likelihood function for this model does not present any relevant difficulties.

In our view, the main advantage of a hurdle model with the second part specified as an LCM is that the new model continues to be the empirical counterpart of an economic model, the principal-agent model, and at the same time unobserved heterogeneity is modelled through a semi-parametric approach, therefore, moving away from strong, and always somewhat arbitrary, distributional assumptions. In addition, in our model the positives are analyzed under the assumption that the unobserved heterogeneity exists in the population of health care users.

4. Data and variables

All results presented in this article are based on cross-section data taken from the National Health Survey, 1999 version (henceforth referred to as NHS_99). The survey is a representative sample of the Portuguese population and collected data from 48.606 individuals. It provides a wide range of information, at an individual level, about socioeconomic and demographic variables, life styles, health status indicators and medical services utilization (Ministério da Saúde - Instituto Nacional de Saúde, 1999).

After dropping the 4.8% of individuals reporting to hold a private health insurance in addition to the observations with missing values on any of the interest variables, the final sample comprise 42.501 observations. The elimination of observations due to the presence of missing values may raise sample selection issues. They may occur whenever one estimate models using a sub-sample and the unobservable characteristics influencing inclusion in the sub-sample, in our case influencing non-response, are correlated with the unobservable factors that influence the dependent variable (Vella, 1998). If deletion is non-random then standard procedures applied to the final sample would result in incorrect inference regarding the impact of the observables on doctor visits (Wooldridge, 2002, Vella, 1998).

In our application 'income' is the variable that the individuals most lack to respond (about 6%). However, only a small share (10%) of those who had not filled the income question did it intentionally, with remaining 90% declaring not knowing the household income. In this situation the existence of correlated unobserved factors influencing both the decision to respond and the number of doctor visits seems unlikely. However, to test whether this reduction in the sample is random we performed a statistical test suggested by Wooldridge (2002) (*procedure 17.2, page 568*). Due to space constraints we omit the details about the test, namely, the instrumental variables used (details will be sent upon request). Concerning the result of the test, the interest coefficient (the 2SLS parameter on $\hat{\lambda}$) is -0.769 (*se = 0.53, t = -1.43*) showing no evidence of sample selection bias.

The dependent variable - VISITS - is the total number of visits to physicians in a 3 months period. The empirical distribution of the dependent variable is given in the Table 1,

Insert Table 1 about here

The maximum number is 30 visits, the average is 1.29 (se = 2.06), showing that the variance is almost four fold the mean, being a sign of 'overdispersion', which was confirmed by formal tests of overdispersion.

As covariates we have selected those that have been found to influence medical care utilization in similar studies. The covariates were clustered into four groups, encompassing socioeconomic and demographic variables, health status indicators, a supply side determinant, and, finally, a group of variables capturing each individual's health insurance status.

Insert Table 2 about here

The intuition to use these covariates is given by several authors. Therefore we skip a detailed description about each variable, namely, why to include it in the regression and the channels through which they, supposedly, impact health care utilization. Nevertheless it is worth to explain how some variables were created.

Education is measured as the total number of years in school. In the case of individuals aged less than 14, education is measured as the maximum education among the adults in household.

Income, as is common in surveys like the one we use, is captured through a categorical ordinal variable measuring disposable net household monthly income. In this paper, in place of use dummy variables to include income, we opted for computing the monthly equivalent disposable income. To create it we have used the modified OECD scale.

The *rural_area* variable was created by augmenting the NHS_99 with data gathered from the National Bureau of Statistics, who classified each '*Freguesia*'[3] in predominantly urban, medially urban and predominantly rural. After including this information in our dataset we end up with a new variable, which classifies each individual's place of residence as predominantly rural, predominantly urban and medially urban. We merged the categories predominantly urban medially urban status, creating in this way two dummy variables, '*rural_area*' and '*urban_area*'.

Regarding health status, one usual way to measure it is using self-assessed health (SAH). Indeed, the NHS_99 includes SAH, however we decided not to use it because a large number of individuals have failed to respond to it. The inclusion of SAH would lead to the elimination of 36% of observations, dropping to 27.044. One may argue that 27.044 observations provide enough degrees of freedom to estimate with sufficient precision the model parameters, what in fact is true, however we suspect that the loss of these individuals may cause sample selection bias. It can be argued that individuals in worse (*unmeasured*) health status are more reluctant to self-assess their health, or because they are unable to do it or because they do not want express their opinion about it. This suggests the existence of unobserved factors, unobserved health status, that influence both, the non-response and the number of visits, thus causing sample selection bias. Once again, to test whether the elimination of these individuals is random we performed the statistical test suggested by Wooldridge (2002), mentioned above. The result of the test show clear evidence of sample selection bias. The relevant coefficient is -0.525 (*se = 0.117, t = -4.48*). This result provides a first indication that inclusion of SAH would cause the working sample to depart from a random sample, and a sound econometric analysis should have this into account if one want to include SAH in the analysis. Our strategy to avoid the problem was to leave out of the analysis the SAH variables. This approach of excluding SAH can have a cost and also lead to results potentially misleading results (Cameron and Trivedi, 2005), nevertheless, we believe that in our application this is less likely to occur because we have a large array of variables to capture health status measuring it sufficiently well. Therefore, we assume that leaving out of the analysis SAH indicators does not cause any econometric problems.

The next variable, '*Phy_1000_residents*', was created by adding external data to the NHS_99. A first step to create this variable was the assignment of each individual area of residence to a territorial region referred to as '*Nut III*'[4]. Afterwards, using data from the National Bureau of

[3] For some purposes the Portuguese territory is divided in three hierarchic administrative divisions: In the first level is the '*Distrito*', containing '*Concelho*' (municipality) and finally, each '*Concelho*' contains a number of '*Freguesias*'. In 1999 Portugal was divided in approximately 4000 '*Freguesias*'

[4] An alternative administrative division of Portugal is at the level of what is known as '*NUT*', where *NUT I* is Portugal mainland, NUT II represent the five health regions (North, Centre, Lisbon and Tagus valley, Alentejo and Algarve), and *NUT III* further divides the territory in more 28 territorial units

Statistics about the total number of physicians and the total population in each Nut III was possible to compute the total number of physicians per 1000 residents at the level of NUT III.

Finally, the last group of variables considered as determinants of medical care utilization is the health insurance variables. Portugal provides health insurance with two main types of insurance schemes; first, the coverage provided by the statutory National Health Service (NHS), covering about 75% of the population; second, the health insurance supplied by various public and private insurance funds whose membership is based on professional or occupational category, referred to as Health Subsystems (HSS). Among the HSS the fund that cover all civil servants, usually referred to as ADSE, is different from the remaining mainly because the scale of operation. Because membership to these funds comes with the profession or occupation, it easy to argue that the variables representing insurance are exogenous in our models. Summarizing, in terms of health insurance, we identify three types of access groups; 1) individuals covered only by the NHS, ('NHS-only'); 2) those individuals covered by the ADSE fund and 3) the individuals who benefits from a health insurance contract provided by a HSS other than the ADSE (referred to as OHSS).

Sample statistics of the independent variables considered in the analysis are presented in Table 3.

Insert here Table 3

5. Results and discussion

Stata 9.0 was used to estimate all models and to perform all numerical computations presented throughout this paper. To account for the possibility of model misspecification the variance-covariance matrix was computed using the robust sandwich estimator (Cameron and Trivedi, 1998, White, 1982). Conversely to the simple negative binomial and hurdle models the estimation of models including LCM specifications may be challenging and time consuming. We opted for estimating those models by direct optimization of the likelihood function, despite the existence of other feasible methods, namely, the EM algorithm. The challenges associated to this estimation method are that the likelihood function of such models may have multiple local maxima, hence one cannot exclude the possibility of convergence to local solutions (McLachlan and Peel, 2000). In this paper we have guarded against this possibility estimating repeatedly each model using a number of different initial solutions. Simpler models were used to generate the initial solutions. We did not observe any relevant convergence problems.

In this article all models that include an LCM specification assumes the existence of only 2 latent classes. Wedel *et al.* (1993) mentioned that the empirical evidence has shown that a small number of latent classes provide enough flexibility to reproduce the data accurately. Moreover, all, at least the one that we know, empirical applications of this methodology to health care data have reported that two latent classes provide sufficient flexibility to explain medical care counts quite well (Deb and Holmes, 2000, Deb and Trivedi, 2002, Deb and Trivedi, 1997, Jimenez-Martin et al., 2002, Atella et al., 2004, Lourenço and Ferreira, 2005). In addition models with $P > 2$ will become rapidly overparametrized, thus, difficult, if not impossible, to estimate. Actually, we did some experiments with $P = 3$ and experienced severe difficulties in the estimation process. Regarding the choice of the density for each latent class we have chosen the NB probability function. In this paper we have estimated the most general LCM specification allowing that the distributions governing the latent classes vary in all parameters, intercept, slopes and the dispersion parameter.

In what follows, this section is organized into three subsections. In the first subsection we address the question of determining if the new hurdle specification is, from a statistical point of view, preferred relative to the competing specifications. In the second subsection, we compare the segmentation of the population into latent classes generated by the popular LCM and the new LCM based hurdle. Finally, in the last subsection we estimate the impact of selected covariates in the mean function of several models estimated to assess the stability of some potentially relevant health policy indicators across models, making also some interpretation work of the most relevant results in terms of health economics conclusions.

Does the new hurdle present better fit?

This subsection addresses the subject of determining if the new hurdle specification is, from a statistical point of view, preferred relative to the competing specifications estimated in this paper.

Table 4 presents the acronyms along with the description of all competing models.

Insert here Table 4

We have used likelihood ratio tests (LR) to choose among nested specifications, while to discriminate among non-nested specifications we have relied on Vuong tests (Vuong, 1989, Winkelmann, 2003), information criteria [*BIC (Bayesian Information criteria) and CAIC (Consistent Akaike Information criteria)*] (Sin and White, 1996, Deb and Trivedi, 1997, Deb and Trivedi, 2002) and GoF tests (Leamer, 1986, Cameron and Trivedi, 2005, Andrews, 1988, Deb and Trivedi, 2002).

Table 5 reports the LR tests results. Some of these tests are made in the boundary of the parameter space, consequently, the rejection region must be adjusted to make correct decisions (Cameron and Trivedi, 1998).

Insert here Table 5

Both versions of the NB model were rejected in favour of the popular NB based hurdle (*tests LR1 and LR4*), which were also rejected in favour of the standard LCM models (*tests LR2 and LR5*).

This result shows that single index of the NBi family impose constraints not verified by the data. Therefore, more general count data models are necessary to adequately describe the data. Still in Table 5, it shows that the models H_NB1 and H_NB2 are rejected when compared, respectively, to HLCM_NB1 and HLCM_NB2 (*tests LR 3 and LR 6*), suggesting that our new LCM based hurdle model outperforms the popular NB based hurdle. This seems to indicate that, conversely to the common practice of specify a simple truncated-at-zero negative binomial model in the second stage of the hurdle, the second part of the model is better described by a 2-component LCM model. An alternative interpretation of this result is that the popular NB based hurdle accounts incorrectly for the unmeasured factors, being necessary more sophisticated ways to deal with the presence of individual unobserved effects.

Regarding the performance of models LCM_NBi *vs* HLCM_NBi, because they are non-nested we compare them using Vuong tests, information criteria, and GoF tests.

Before analysing the results of the Vuong tests, we briefly present some details about its implementation. Consider the statistical decision of choosing between two non-nested models, f_{β_1} and g_{β_2}. Voung's hypothesis can be formulated as follows,

$$\left[H_0 : f_{\beta_1} \text{ and } g_{\beta_2} \text{ are equivalent models}\right] \; vs \; \left[H_1 : \left(f_{\beta_1} \text{ is better than } g_{\beta_2}\right) \; or \; \left(g_{\beta_2} \text{ is better than } f_{\beta_1}\right)\right]$$

When the models are strictly nested Voung proposed the utilization of the following test statistic,

$$V = \frac{\sum_{i=1}^{n}\left[\ln\left(f\left(y_i \mid x_i, \hat{\beta}_1\right)\right) - \ln\left(g\left(y_i \mid x_i, \hat{\beta}_2\right)\right)\right]}{\sqrt{\sum_{i=1}^{n}\left[\ln\left(\frac{f\left(y_i \mid x_i, \hat{\beta}_1\right)}{g\left(y_i \mid x_i, \hat{\beta}_2\right)}\right)\right]^2 - \frac{1}{n}\left[\sum_{i=1}^{n}\ln\left(\frac{f\left(y_i \mid x_i, \hat{\beta}_1\right)}{g\left(y_i \mid x_i, \hat{\beta}_2\right)}\right)\right]^2}} \qquad [19]$$

Voung showed that under H_0 the test statistic V follows a standard normal distribution, therefore, at a significant level α, the decision rule presented by Vuong (1989) is as follows: 1) If $V > Z_\alpha$ then one rejects the null hypothesis of equivalent models in favour of f_{β_1} being better than g_{β_2}, 2) If $V < -Z_\alpha$ then one rejects the null hypothesis in favour of g_{β_2} being better than f_{β_1} and finally, 3) If $-Z_\alpha \leq \text{V} \leq Z_\alpha$ it is not possible to discriminate among the competing models.

Table 6 reports the results of Voung tests, providing clear evidence about the better performance of our new hurdle formulation. The Voung test statistic for the comparison of the non-nested models LCM_NBi against HLCM_NBi (*i=1,2*) clearly rejects, at significance level of 1%, the two LCM_NBi models (*tests V2 and V5*).

Insert Table 6 about here

In addition, the test to contrast model HLCM_NB1 against HLCM_NB2 rejects the new hurdle with NB1 as baseline density, evidencing that the new hurdle with NB2 as baseline distribution is the preferred model to fit our data.

Table 7 presents values of BIC and CAIC for each model estimated. On the basis of this criterion the model with lower values for both statistics is preferred (Deb and Trivedi, 2002).

Insert here Table 7

The figures in the table show that NBi models perform poorly relative to all other econometric specifications, as is evidenced by the large BIC and CAIC values, which is entirely in line with the LR tests conclusions presented earlier. Comparing the models LCM_NBi with H_NBi in the spirit of Deb and Trivedi (2000) the results offer mixed evidence regarding which model performs better. The LCM_NB1 model performs better than H_NB1, the opposite result occurs when the NB2 is the base model. Voung tests results reported in Table 6 support this view. However, what is more relevant from Table 7 is that both BIC and CAIC present lower values for the new hurdle showing that it performs better than all competing specifications. Moreover, overall, the model with lower values for BIC and CAIC is model HLCM_NB2. This finding is a further input to add to Voung tests conclusion supporting the superiority of HLCM_NB2 specification over all alternative specifications.

In addition to all these tests, we have also performed some goodness of fit tests to verify the robustness of the previous findings. Due to space constraints we omit its results (will be sent upon request), however it is worth to mention that their conclusions are completely in line with the conclusions presented so far.

In summary, all model selection tests converge into the same conclusion: the model HLCM_NB2 is the specification that performs better in fitting our data. Our global results are in accordance with Winkelmann's (2004) conclusions that hurdle specifications that deviate from the popular hurdle negative binomial model can outperform the familiar LCM specification. In line with Winkelmann, our findings also suggest that the evidence that has been reported favouring latent class models over the popular hurdle specification (e.g. (Deb and Holmes, 2000, Deb and Trivedi, 2002, Deb and Trivedi, 1997, Jimenez-Martin et al., 2002, Sarma and Simpson, 2006) can be interpreted as evidence against that particular specification of the hurdle but not against the general hurdle framework. Therefore, this evidence, in some way gives new life to the principal agent model as a feasible economic framework in which to base the econometric specifications to analyse medical care utilization, and it may renew the discussion about the better econometric model to explain medical care utilization models.

The majority of empirical work that have been reporting the better statistical performance of LCM specifications over the hurdle framework have in common the fact that their dependent variable (number of doctor visits) are measured during one year period. This is a long period that may cause the violation of the single illness spell. This probable violation of the single illness spell assumption along with the misspecification of the model for the positives in the hurdle may explain the reported superiority of the LCM over the hurdle. When the data regarding the number of doctor consultations are collected during a shorter time period (for example, 3 months), decreasing the probability of violating the single illness spell assumption, the hurdle structure may again emerge as a suitable model to explain individual behaviour. In this data paradigm, if the second part of the hurdle correctly specifies the unobservables then the hurdle structure outperforms the LCM framework. This is precisely what our and Winkelmann's (2004) results have showed. In both empirical applications doctor visits are measured during a three months period, and the second part of the hurdle are correctly specified. The results of both applications presented similar conclusions regarding the preference for the hurdle structure. Therefore, in our view, the performance of one specification over another, being dependent on the manner one specifies the model for the positives may also depend on the characteristics of the data, namely, on the survey design. Clearly, to confirm this thesis one would have to apply the new hurdle formulation to medical care data gathered during the period of one year, and verify whether the new specification still to outperform the popular LCM model.

Unobserved heterogeneity and latent class characteristics

The results reported in the previous subsection have shown that the new hurdle specification, presents better statistical performance when compared to all competing specifications. In addition the results also show that the traditional LCM with NB1 as baseline distributions are, among the traditional count data models, the model that better represents the data[5]. Both specifications use the LCM framework to develop the model to analyse the data, however under different hypothesis regarding the unobserved heterogeneity. The pure LCM is formulated under the assumption that the unobserved heterogeneity is present in the overall population under study, in contrast, the LCM part of the new hurdle is specified under the hypothesis that the unobserved heterogeneity is present in the population of health care users. Therefore, it would be interesting to analyse if the latent classes of users generated by the two LCM based models are similar or, on the contrary, are different.

[5] Despite Voung tests contrasting some LCM versions with some versions of the popular hurdle specifications lead to inconclusive results.

Table 8 presents the expected number of visits and respective standard errors as estimated by the models LCM_NB1 and HLCM_NB2.

Insert here Table 8

It is worth to quickly mention how the figures presented in Table 8 were computed.

For the LCM_NB1 we estimated the fitted mean for the overall population (column 1) by taking the sample averages of the estimates of the LCM mean function given by,

$$E(y_i \mid x_i) = \pi * \exp\left(x_i'\beta_1\right) + (1-\pi) * \exp\left(x_i'\beta_1\right) \qquad [20]$$

To compute the fitted mean for each latent class, we have borrowed the procedure proposed by Bago d'Uva (2006). First, one has to assign each individual to a latent classe, which can easily be done after calculating the posterior probability that individual *i* belongs to the latent class *j*, which is given by (Deb and Trivedi, 2002)

$$P\left(I_i \in C_j \mid Y_i = y_i, x_i\right) = \frac{\pi_j * f_j\left(y_i \mid x_i, \boldsymbol{\beta}_j\right)}{\sum_{j=1}^{2} \pi_j * f_j\left(y_i \mid x_i, \boldsymbol{\beta}_j\right)} \quad j = 1,2 \qquad [21]$$

After computing the posterior probabilities, each individual was assigned to a latent class according to the highest posterior probability. Then, conditional on the latent class, we have estimated the distribution of visits, according to,

$$E\left(y_i \mid x_i, \boldsymbol{\beta}_j\right) = \lambda_j = \exp\left(x_i'\boldsymbol{\beta}_j\right) \quad j = 1,2 \qquad [22]$$

Finally, the fitted values for each class (columns 2 and 3) were obtained by taking the sample average for each class, selecting, obviously, only the individuals of that class.

The model HLCM_NB2 respects the original idea of a hurdle model assuming that the population is, *a priori*, segmented in the population of health care users and non-users. For the health care users ($y_i > 0$) one can compute the fitted mean for each latent class (columns 6 and 7) in the users population. Similar to what was done previously to the LCM_NB1 model, we have computed the posterior probability that individual *i* belongs to the latent class *j*, however, because we are working in the users population, formula 21 has to be slightly modified,

$$P\left(I_i \in C_j \mid Y_i = y_i, x_i, y_i > 0\right) = \frac{\pi_j * \dfrac{f_j\left(y_i \mid x_i, \boldsymbol{\beta}_j\right)}{\left(1 - P\left(y_i = 0 \mid x_i, \boldsymbol{\beta}_j\right)\right)}}{\sum_{j=1}^{2}\left[\pi_j * \dfrac{f_j\left(y_i \mid x_i, \boldsymbol{\beta}_j\right)}{\left(1 - P\left(y_i = 0 \mid x_i, \boldsymbol{\beta}_j\right)\right)}\right]} \quad j = 1,2 \qquad [23]$$

After computing these posterior probabilities in the users population, once again, each individual was assigned to a latent class according to the highest posterior probability. Then, conditional of the latent class, we have estimated the distribution of visits, according to

$$E\left(y_i \mid x_i, \beta_j, y_i > 0\right) = \frac{\exp\left(x_i'\beta_j\right)}{P\left(y_i > 0 \mid x_i, \beta_j\right)} = \frac{\lambda_{ij}}{1 - P\left(y_i = 0 \mid x_i, \beta_j\right)} \quad j = 1,2 \qquad [24]$$

Finally, the fitted values for each class were obtained by taking the sample average for each class.

On the other hand, the fitted mean in the whole users population, is given by

$$E(y_i \mid x_i, y_i > 0) = \pi * \frac{\exp\left(x_i'\beta_1\right)}{1 - P(y_i = 0 \mid x_i, \beta_1)} + (1 - \pi) * \frac{\exp\left(x_i'\beta_1\right)}{1 - P(y_i = 0 \mid x_i, \beta_2)} \qquad [25]$$

We note that given our specification for the LCM in the users population, the π_s reflect the proportions of individuals belonging to each latent class in the users population and not on the full population as would be the case if we have used specification 15.

Finally, to estimate the fitted mean for the overall population under study (users+non-users), we have used the expression,

$$E(y_i \mid x_i) = \left[1 - P(y_i = 0 \mid x_i, \beta_0)\right] * \left[\pi * \frac{\lambda_1}{1 - P(y_i = 0 \mid x_i, \beta_1)} + (1 - \pi) * \frac{\lambda_2}{1 - P(y_i = 0 \mid x_i, \beta_2)}\right] \qquad [26]$$

In every case the standard error was computed using the Delta method (Oehlert, 1992, Wooldridge, 2002). After computing the standard error, the construction of the confidence intervals is straightforward (Cameron and Trivedi, 1998).

As is well known, standard LCM models usually permit to categorize the overall population into two latent classes. One class is formed by low intensity users of health care, while the people that form the other class are classified as high intensity users (Deb and Holmes, 2000, Deb and Trivedi, 2002, Deb and Trivedi, 1997, Sarma and Simpson, 2006, Gerdtham and Trivedi, 2001, Winkelmann, 2004). The LCM_NB1 results obtained in our application are fully in line with these findings, as they also suggest that the overall population can be segmented into a population of high intensity users, about 14% of the sample, visiting the doctor, in average, 4.3 (se = 0.454) times in each quarter, while the remaining 86% of the sample are grouped in the low intensity users class, visiting the doctor 1.13 (se = 0.015).

Before analyse the classes generated by HLCM_NB2 note that the LCM part of this model deals with the unobserved heterogeneity in the health care users population, which might be different from the unobservables in the overall population. The possible differences in the sources of unobserved heterogeneity assumed in each model may cause the emergence of dissimilar latent classes.

The model HLCM_NB2 suggests that the users population can be segmented into two clusters. The first cluster *(column 6)* who comprise about 18% of the users population have, on average, 4.27 (se = 0.406) visits to a physician, while the remaining 82% of individuals (*U_Class 2*) seek care about 2.03 (se = 0.014) times. Note that the model predicts that expected number of visits for the user population is 2.38 (se = 0.014).

Comparing the latent classes generated by the two models, we can conclude that the high users class formed by each model presents a similar expected utilization, about 4.3 visits, despite the difference in the estimated proportion of individuals, 14% as estimated by the LCM_NB1 and 18% as estimated by the HLCM_NB2. This seems to indicate that the high users class suggested by both models overlap in a large extent. However, concerning the low users class, it show differences in the expected utilization as well as in the percentage of individuals. The LCM_NB1 model predicts an average utilization of 1.13 visits, while the other model predicts an expected utilization of about 2 visits. Note that the respective 99% confidence intervals do not overlap what suggests statistically significant differences in the estimates. The different characteristics of the classes generated by the two models, especially the differences in the low users class, show that, in fact, the sources of unobserved heterogeneity may vary across populations (full population *vs* users population).

Among the health care users population, one can clearly label one class as high intensity users and the other as low intensity users. However, more important than this classification into low and high users, it is its classification as, respectively, '*healthy*' and '*ill*' clusters of individuals, as has been frequent to interpret the latent classes suggested by the general LCM (Deb and Trivedi, 1997, Deb and Trivedi, 2002). Is this split of the population into '*healthy*' *vs* '*ill*' individuals still valid in the health care users population? We believe that the classification is still valid, even more valid, due to the role of the physician in the choices regarding the number of visits among the health care users. The argument goes as follows:

Under the hurdle framework, the initial decision to seek care is entirely dependent on the individual, however, after the contact decision, the physician's role in deciding about the number of visits increases. Therefore, it can be argued that, in the health care users population, it is mainly the doctor who decides, based on characteristics of the individuals, health status included (most of them unobserved to the researcher), who is visits the doctor regularly (high user) and who visits less regularly (low user). That is, it can be argued that it is the physician who makes the larger contribution to the allocation of individuals to each latent class. It is widely recognized that doctors, compared to patients, when make decisions on behalf of the patient tend to base the decisions more on the health status of the individual, and less on other factors, like socio-demographic and '*patient preference*' factors. Therefore in the presence of an *ill* individual, the doctor, based on observed and unobserved health status, will advice a high number of visits. On the contrary, in the presence of a *healthy* individual, will advice a low number of visits. In this way, the high users group will comprise the '*ill*' individuals, while the low users group will comprise the '*healthy*' individuals, perhaps the ones that seek mostly preventive care and have that may go through sporadic episodes of illness.

If the former reasoning is accepted then we expected that, in model HLCM_NB2 the observed health status variables to have a similar effect in explaining the number of visits in each latent class. For example, consider the effect, conditional on a latent class, of suffering from back pain in the mean number of visits. When the decision is mainly doctor based, he assesses the patient health care needs in function of the health status of the individual, part of it unobserved to the researcher, being unlikely that other unobserved (for the researcher) non-related health status factors contribute much to the decision. Therefore we expect that the impact of back pain in the number of visits to be similar across latent class. On the contrary, in the full population (model LCM_NB1), when the individual has an increased role on the decision process, the observed health status (back-pain) may interact differently with the unobserved factors in each latent class, causing different individual decisions in different classes. In order to assess to what extent the impact of health status covariates are different across latent classes, we have tested the equality of health status variables slopes across the latent classes. The results are presented in Table 9.

Insert Table 9 about here

The results are what we expected. In the HLCM_NB2 model the health status group of variables have similar roles in explaining the average number of visits in both latent classes (*p = 0.021*), conversely, in the LCM_NB1 model the health status variables show a different, statistically significant, impact in each latent class (*p = 0.000*).

Do different models provide different evidence?

An important health policy theme in Portugal is that of access to health care, where access should dependent only on medical need rather than on socio-economic or demographic variables, like, for instance, area of residence, income, insurance status and so on. Thus, the empirical test to examine if those in equal need have equal access to health care could be relevant for the Portuguese health authorities. However, as was clearly asserted by Deb and Holmes (2000), the estimates of medical care utilization are dependent on the empirical specification used to examine the data, thus, if health care utilization models do not reflect properly the behavioural structures then estimates will not reflect real use, and suggested policies may have unexpected consequences. In order to shed some light on the extent that policy relevant measures depends on the form of the empirical specification used to analyse health care utilization, and at the same time to determine if access to health care in Portugal is dependent only medical need, in this subsection we estimate the effect of '*income*', '*education*', area of residence — '*rural area*' — and insurance coverage — '*NHS-only*' — on the mean function for a number of models estimated. Table 10 reports the results of the analysis.

Insert Table 10 about here

Beginning with the analysis of the estimates of the effect of the covariates in the overall population (*columns 1, 2, 3, 5 and 8*), we note that, for each covariate considered, all models present estimates statistically significant, and with the same sign. Despite some variations in the estimates, the 99% all confidence intervals generally overlap in a large extent, meaning that after accounting for the standard deviation, the estimates of the average marginal effect resulting from different models are not statistically different. This means that if the goal of the research is to verify how the mean utilization varies in response to a change in covariates then it seems that the extra effort to estimate more sophisticated models do not uncover any new relevant results, and a simple single index model like Poisson or Negative Binomial will be sufficient. However, more sophisticated models generally permit to enhance the analysis.

Comparing now the effect of the covariates on the frequency decisions (columns 4 and 9) estimated by the two hurdle versions presented in Table 10, once again the estimates of the two models are very similar, both in sign, intensity and statistical significance. A quick look to the table also shows that the impact of the covariates in the expected number of visits for the high users class of both models (column 6 and 10) although presenting somewhat large differences in the intensity of the estimate, they have in common the lack of statistical significance, meaning that those factors are not important in explaining medical care utilization. This similarities regarding the impact of the covariates is not surprising, especially after having concluded above that the high users class created by the two models are similar in the predicted utilization. In Table 10 the main differences arise in the estimates of the AME for the low users class generated by the different models (columns 7 and 11). For each covariate, the AME effect is statistically significant in the low users class generated by the LCM_NB1 model, while they loose the statistical significance in the low users class generated by the HLCM_NB2 model.

In summary, calculations based on all models generally show only small differences, possibly without statistical significance, on AME estimates. In our view, when the analyst objective is to study the mean function, the advantage of using structural models, like hurdle and LCM, is that they enhance the analysis, allowing extracting more information from the data.

However, most of the times, among the analyst aims are the study of the effect of covariates in other policy relevant measures other than the mean. For instance the goal can be to study the probability of having at least one visit, or the probability that use exceeds a given value. Perhaps, this could reveal that different models generate different conclusions, however, we will leave this analysis for future work.

Table 11 reports the parameter estimates and standard errors for models LCM_NB1 and HLCM_NB2.

Insert Table 11 about here

A quick scan through the results of the new hurdle model, reported in the three right most columns of Table 11, show that the signs of the probability[6] of being a user are consistent with the results found in similar studies. Moreover, almost all covariates considered in the analysis present statistical significance, meaning that they influence the probability of visit a doctor. Generally lower health status is associated to a higher probability of making a visit to the doctor and being only covered by the NHS lowers that probability. In addition, the demographic variables also present the anticipated signs. Among this class of variables, we highlight that income increases the probability of visiting a doctor, while living in a rural area is a factor that decreases that probability. These results suggest that access to health care dependent not only on health status but also on income (more income increases the probability of use), living place (those who reside in rural areas are less likely to have a visit to the doctor) and on the type of insurance coverage (those who are only covered by the NHS are less probable to see the doctor).

Regarding the effect of covariates in the number of visits after crossing the hurdle, they are presented in the two right most columns of Table 11. The columns labelled '*High users|user*' present the parameters for the high intensity users class, while the column labelled '*Low users|user*' presents the estimates for the low intensity users class. After the contact decision, the results show that in the high users class only some of the health status variables presents statistical significance. This is a significant result as it shows that, once in the system, those in similar need for medical care receive a similar quantity.

Regarding the impact of the covariates in the low users class, although more variables have gained statistical significance [female (+), (north, centre, LVT, alentejo (+) and not_work (+)], the result that we found more relevant is that income, insurance generosity (NHS-only) and living place (Rural_Area) still without statistical significance, meaning that they do not exert any relevant impact on the expected number of visits to the doctor. This result has also been shown in Table 10 (columns 10 and 11). Like in the high users class, the results for the low users class also indicate that after the contact decision, the frequency of utilization does not depend on income, neither on the place of residence or on the generosity of the insurance status.

Therefore, one can conclude that in the Portuguese health care system, the effect income, place of residence and insurance generosity related inequities are in the contact decision, and not on the frequency of utilization, conditional on the individual had jump the first hurdle of deciding to see the doctor. Note that this conclusion illustrates the importance of using more sophisticated models to analyse medical care utilization. Using, for instance, the simple NB1 model, one would have concluded that income, rural area and NHS-only covariates played a significant role in explaining

[6] Notice that we have specified a NB model for the first part of the hurdle, therefore, the figures presented in the '*first part*' should not be interpreted as measuring the intensity of the impact on the probability of being a user. However, the sign and the direction of the effect are coincident.

medical care utilization. However this global result hidden crucial information, namely, that the bottle neck (relative to the three variables in analysis) is on the probability of initiating the process of care, and not on the frequency of utilization, after crossing the bottle neck.

Also note that the use of LCM_NB1, would have lead to the same incomplete conclusion regarding the effect of income, living place and insurance generosity. Column 7 of Table 10 show that in the low users population, income, rural area and NHS-only covariates exert an effect on the mean utilization, meaning that in this low uses population non- medical care need related variables play a role in the frequency of utilization. However, the model HLCM_NB2, the one that we believe that better represents the data, once more, shows that after contact decision, the covariates income, NHS-only and rural area do not influence utilization the low users (the class generated by the model HLCM_NB2).

6. References

- ANDREWS, D. W. K. (1988) Chi-Square Diagnostic-Tests for Econometric-Models - Introduction and Applications. *Journal of Econometrics*, 37, 135-156.
- ATELLA, V., BRINDISI, F., DEB, P. & ROSATI, F. C. (2004) Determinants of access to physician services in Italy: a latent class seemingly unrelated probit approach. *Health Economics*, 13, 657-668.
- BAGO D'UVA, T. (2005) Latent class models for use of primary care: evidence from a British panel. *Health Economics*, 14, 873-892.
- BAGO D'UVA, T. (2006) Latent class models for utilisation of health care. *Health Economics*, 15, 329-343.
- BRANNAS, K. & ROSENQVIST, G. (1994) Semiparametric Estimation of Heterogeneous Count Data Models. *European Journal of Operational Research*, 76, 247-258.
- CAMERON, A. C. & TRIVEDI, P. K. (1996) Count data models for financial data. IN MADDALA, G. S. & RAO, C. R. (Eds.) *Handbook of Statistics*.
- CAMERON, A. C. & TRIVEDI, P. K. (1998) *Regression analysis of count data*, Cambridge, UK ; New York, NY, USA, Cambridge University Press.
- CAMERON, A. C. & TRIVEDI, P. K. (2005) *Microeconometrics : methods and applications*, New York, NY, Cambridge University Press.
- CAMERON, A. C., TRIVEDI, P. K., MILNE, F. & PIGGOTT, J. (1988) A Microeconometric Model of the Demand for Health-Care and Health-Insurance in Australia. *Review of Economic Studies*, 55, 85-106.
- CRAGG, J. G. (1971) Some Statistical Models for Limited Dependent Variables with Application to Demand for Durable Goods. *Econometrica*, 39, 829-&.
- DEB, P. (2001) A discrete random effects probit model with application to the demand for preventive care. *Health Economics*, 10, 371-383.
- DEB, P. & HOLMES, A. M. (2000) Estimates of use and costs of behavioural health care: A comparison of standard and finite mixture models. *Health Economics*, 9, 475-489.
- DEB, P. & TRIVEDI, P. K. (1997) Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 12, 313-336.
- DEB, P. & TRIVEDI, P. K. (2002) The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics*, 21, 601-625.
- DUAN, N., MANNING, W. G., MORRIS, C. N. & NEWHOUSE, J. P. (1983) A comparison of alternative models for the demand for medical care. *Journal of business and Statistics*, 1, 115-126.
- GERDTHAM, U. G. (1997) Equity in health care utilization: Further tests based on hurdle models and Swedish micro data. *Health Economics*, 6, 303-319.

- GERDTHAM, U. G. & TRIVEDI, P. K. (2001) Equity in Swedish health care reconsidered: New results based on the finite mixture model. *Health Economics.*
- GOURIEROUX, C. & VISSER, M. (1997) A count data model with unobserved heterogeneity. *Journal of Econometrics,* 79, 247-268.
- GROGGER, J. T. & CARSON, R. T. (1991) Models for Truncated Counts. *Journal of Applied Econometrics.*
- GROOTENDORST, P. V. (2002) A Comparison of Alternative Models of Prescription Drug Utilization. IN ANDREW M. JONES, O. O. D. (Ed.) *Econometric Analysis of Health Data.*
- GROSSMAN, M. (1972) Concept of Health Capital and Demand for Health. *Journal of Political Economy,* 80, 223-225.
- GURMU, S. (1997) Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization. *Journal of Applied Econometrics,* 12, 225-242.
- GURMU, S. & TRIVEDI, P. K. (1992) Overdispersion Tests for Truncated Poisson Regression-Models. *Journal of Econometrics.*
- HECKMAN, J. & SINGER, B. (1984) A Method for Minimizing the Impact of Distributional Assumptions in Econometric-Models for Duration Data. *Econometrica.*
- INSTITUTO NACIONAL DE ESTATÍSTICA. (1999) *Indicadores Urbanos do Continente,* Lisboa, Instituto Nacional de Estatística.
- JIMENEZ-MARTIN, S., LABEAGA, J. M. & MARTINEZ-GRANADO, M. (2002) Latent class versus two-part models in the demand for physician services across the European Union. *Health Economics,* 11, 301-321.
- JIMÉNEZ-MARTÍN, S., LEBEAGA, J. M. & MARTÍNEZ-GRANADO, M. (2002) Latent Class versus Two-Part Models in the Demand for Physician Services Across the European Union. IN ANDREW M. JONES, O. O. D. (Ed.) *Econometric Analysis of Health Data.*
- LEAMER, E. (1986) Model Choice and Specification Analysis. *Handbook of Econometrics,* 1, 285-330.
- LINDSAY, B. G. & LESPERANCE, M. L. (1995) A Review of Semiparametric Mixture-Models. *Journal of Statistical Planning and Inference,* 47, 29-39.
- LOURENÇO, Ó. D. & FERREIRA, P. L. (2005) Utilization of public health centres in Portugal: effect of time costs and other determinants. Finite mixture models applied to truncated samples. *Health Economics,* 14, 939-953.
- MANNING, W. G., NEWHOUSE, J. P., DUAN, N., KEELER, E. B., LEIBOWITZ, A. & MARQUIS, M. S. (1987) Health-Insurance and the Demand for Medical-Care - Evidence from a Randomized Experiment. *American Economic Review,* 77, 251-277.
- MCLACHLAN, G. J. & PEEL, D. (2000) *Finite mixture models,* New York, Wiley.
- MINISTÉRIO DA SAÚDE - INSTITUTO NACIONAL DE SAÚDE (1999) *INS 1998/1999. Continente. Dados Gerais.*, INSA - Instituto Nacional de Saúde.
- MULLAHY, J. (1986) Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics,* 33, 341-365.
- MULLAHY, J. (1997) Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics.*
- OEHLERT, G. W. (1992) A Note on the Delta Method. *American Statistician,* 46, 27-29.
- POHLMEIER, W. & ULRICH, V. (1995) An Econometric-Model of the 2-Part Decision-Making Process in the Demand for Health-Care. *Journal of Human Resources.*
- PORTUGAL (2002) Decreto-Lei n.o 244/2002. N.o 255 — 5 de Novembro de 2002 ed., MINISTÉRIO DAS CIDADES, ORDENAMENTO DO TERRITÓRIO E AMBIENTE.
- SANTOS-SILVA, J. M. C. (2003) A note on the estimation of mixture models under endogenous sampling. *Econometrics Journal,* 36, 46-52.
- SANTOS-SILVA, J. M. C. & WINDMEIJER, F. (2001) Two-part multiple spell models for health care demand. *Journal of Econometrics.*
- SARMA, S. & SIMPSON, W. (2006) A microeconometric analysis of Canadian health care utilization. *Health Economics,* 15, 219-239.
- SIN, C. Y. & WHITE, H. (1996) Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics.*
- VAN OURTI, T. (2004) Measuring horizontal inequity in Belgian health care using a Gaussian random effects two part count data model. *Health Economics,* 13, 705-724.
- VELLA, F. (1998) Estimating models with sample selection bias: A survey. *Journal of Human Resources,* 33, 127-169.

- VERA-HERNANDEZ, A. M. (1999) Duplicate coverage and demand for health care. The case of Catalonia. *Health Economics,* 8, 579-598.
- VUONG, Q. H. (1989) Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica,* 57, 307-333.
- WAGSTAFF, A. (1986) The Demand for Health - Some New Empirical-Evidence. *Journal of Health Economics,* 5, 195-233.
- WEDEL, M., DESARBO, W. S., BULT, J. R. & RAMASWAMY, V. (1993) A Latent Class Poisson Regression-Model for Heterogeneous Count Data. *Journal of Applied Econometrics,* 8, 397-411.
- WHITE, H. (1982) Maximum-Likelihood Estimation of Mis-Specified Models. *Econometrica,* 50, 1-25.
- WINKELMANN, R. (2003) *Econometric analysis of count data,* Berlin ; New York, Springer.
- WINKELMANN, R. (2004) Health care reform and the number of doctor visits - An econometric analysis. *Journal of Applied Econometrics,* 19, 455-472.
- WOOLDRIDGE, J. M. (2002) *Econometric analysis of cross section and panel data,* Cambridge, Mass., MIT Press.
- ZWEIFEL, P. (1981) Supplier-Induced Demand in a Model of Physician Behavior. IN VAN DER GAAG, J. & PERLMAN, M. (Eds.) *Health, Economics, and Health Economics.* North-Holland.

7. Appendix A - Tables

Table 1 - Empirical distribution of the total number of physician visits (VISITS)

Visits	Relative Frequency
0	46.01%
1	23.75%
2	12.40%
3	9.19%
4	3.25%
5	1.92%
6	1.61%
7	0.40%
8	0.32%
9	0.11%
10	0.40%
≥ 11	0.57%
Mean	1.29
Variance	4.24
Variance/mean	3.29
n	*42.501*

Table 2 - Variable Definitions

Variable name	Variable Definition
Visits	Total number of doctor visits during a 3 months period
Socioeconomic	
Age [/10]	Age, in years, divided by 10
sqAge	Square of age[/10]
Female	= 1 if the individual is female
Married	= 1 if the individual is married
Education	Number of years of schooling. In the case of child, the education of most educated adult living in the household
Not_work	= 1 if the individual did not work in the two weeks previous to the application of the survey
Retired	= 1 if the individual is retired
Unemployed	= 1 if the individual is unemployed
(log) Income	Logarithm of equivalised monthly real income in hundreds of Euros
North	= 1 if the individual lives in the north region
Centre	= 1 if the individual lives in the centre region

LTV	= 1 if the individual lives in the Lisbon and Tagus Valley region
Alentejo	= 1 if the individual lives in the Alentejo region
Rural_Area	= 1 if the individual lives in a rural area
Health Status	
Diabetes	= 1 if the individual suffers from diabetes
Insulin	= 1 if the individual is insulin dependent
EBP	= 1 if the individual suffers from elevated blood pressure
Asthma	= 1 if the individual suffers from asthma
Bronchitis	= 1 if the individual suffers from bronchitis
Allergy	= 1 if the individual suffers from an allergy
Back pain	= 1 if the individual suffers from back pain
sick_long_Run	= 1 if the individual suffers from an illness for more than 3 months
sick_Short_run	= 1 if the individual reports had been sick in the previous two weeks
Limited	= 1 if the individual has some sort of physical handicap that impedes him to execute certain physical daily activities
Stress	= 1 if the individual is took sleeping pills in the last two weeks
NeverSmoked	= 1 if the individual never smoked during her/his lifetime
not physical activity	= 1 if the individual's daily activities do not require physical activity
mild_exercise	= 1 if the individual engages in mild sports activities at least four hours a week.
Supply side	
Phy_1000_residents	Total number of licensed physicians per 1000 inhabitants [Nut III regional level]
Insurance Status	
NHS-only	= 1 if the individual is covered only through the NHS
ADSE	= 1 if the individual is covered by the ADSE insurance scheme
HOSS	= 1 if the individual is covered by a HSS, other than ADSE (reference class)
Seasonality	
Winter	= 1 if the period of observation was in the Winter
Spring	= 1 if the period of observation was in the Spring
Summer	= 1 if the period of observation was in the Summer

Table 3 - Summary statistics for covariates (N = 42.501)

	Mean	**S.d**	**Max**	**Min**
Age [/10]	4.240	2.331	0	10.3
Female	0.527	0.499	0	1
Married	0.540	0.498	0	1
Education	5.380	4.300	0	24
not_work	0.589	0.492	0	1
Retired	0.202	0.401	0	1
Unemployed	0.030	0.171	0	1
Income[/100]	3.656	2.718	0.231	24.939
North	0.315	0.464	0	1
Centre	0.200	0.400	0	1
LVT	0.246	0.431	0	1
Alentejo	0.119	0.324	0	1
Rural_Area	0.170	0.376	0	1
Diabetes	0.056	0.231	0	1
Insulin	0.006 [/diabetes =0.11]	0.079	0	1
EBP	0.178	0.383	0	1
Asthma	0.062	0.241	0	1
Bronchitis	0.030	0.170	0	1
Allergy	0.144	0.351	0	1
Back pain	0.407	0.491	0	1
sick_long_Run	0.009	0.096	0	1
sick_Short_run	0.344	0.475	0	1
Limited	0.045	0.207	0	1
Stress	0.113	0.317	0	1
NeverSmoked	0.629	0.483	0	1
not physical activity	0.609	0.488	0	1
mild_exercise	0.149	0.356	0	1
Phy_1000_Inhabitants	2.774	2.220	0.579	9.152
NHS-only	0.848	0.359	0	1

ADSE	0.095	0.293	0	1
Winter	0.248	0.432	0	1
Spring	0.252	0.434	0	1
Summer	0.244	0.429	0	1

Table 4 – Description of competing models estimated

Model Acronym	**Model description**	**Number of parameters to estimate**	**Log_L NegBin1**	**Log_L NegBin2**
NBi, *i = 1,2*	Negative binomial regression model	*36*	- 61,837	- 62,200
H_NBi, *i=1,2*	Popular Hurdle specification: both hurdle and positive part are specified using a NB2 probability function. In the hurdle the dispersion parameter was set to one.	71	- 61,470	- 61,275
LCM_NBi, *i=1,2*	Two latent class model with NBi as component distributions.	73	- 61,170	- 61,389
HLCM_NBi, *i=1,2*	New Hurdle model: first part based on the NB distribution. The second part is specified as a 2-LCM with NBi as baseline distribution. The dispersion parameter was set to one in the first stage.	108	- 60,938	- 60,740

Table 5 - Likelihood ratio tests results

Test number	Null	Alternative	LR statistic	Df	Result
LR1	NB1	H_NB1	734.08	35	HNB1
LR 2♦	NB1	LCM_NB1	1332.7	37	LCMNB1
LR 3♦	H_NB1	HLCM_NB1	1064.0	37	HLCMNB1
LR 4	NB2	H_NB2	1851.1	35	HNB2
LR 5♦	NB2	LCM_NB2	1622.3	37	LCMNB2
LR 6♦	H_NB2	HLCM_NB2	1068.6	37	HLCMNB2

♦ - Tests made in the boundary of the parameter space.

Table 6 – Results of Vuong tests to discriminate among non-nested alternatives

Test	Model 1 F_{β_1}	Model 2: G_{β_2}	Vuong Statistic	Result (Preferred model)
V1	LCM_NB1	LCM_NB2	12.1	LCM_NB1**
V2	LCM_NB1	HLCM_NB1	-10.0	HLCM_NB1**
V3	LCM_NB1	HLCM_NB2	-15.6	HLCM_NB2**
V4	LCM_NB2	HLCM_NB1	-13.3	HLCM_NB1**
V5	LCM_NB2	HLCM_NB2	-19.1	HLCM_NB2**
V6	HLCM_NB1	HLCM_NB2	-12.0	HLCM_NB2**
V7	LCM_NB1	H_NB1	6.3	LCM_NB1**
V8	LCM_NB2	H_NB2	-2.2	Inconclusive
V9	LCM_NB1	H_NB2	2.1	Inconclusive
V10	LCM_NB2	H_NB1	1.6	Inconclusive
V11	NB1	NB2	6.7	NB1**
V12	H_NB1	H_NB2	-6.1	H_NB2

** $p < 0.01$

Table 7 - Information Criteria results

Model	**BIC**	**CAIC**
NB1	124,057.9	124,093.9
H_NB1	123,696.8	123,767.8
LCM_NB1	123,119.5	123,192.5
HLCM_NB1	123,027.1	123,135.1

NB2	124,785.0	124,821.0
H_NB2	123,306.9	123,377.9
LCM_NB2	123,557.0	123,630.0
HLCM_NB2 ♣	122,632.7	122,740.7

♣ - Model preferred by BIC and CAIC

Table 8 – Model fitted means, standard errors (in parenthesis) and confidence intervals

	Model LCM_NB1			**Model HLCM_NB2**			
					Users population		
	Overall Population	*High users 13.8%*	*Low users 86.2%*	*Overall Population*	*Overall Population/Users*	*High users/Users 17.8%*	*Low users/Users 82.2%*
Estimated mean	1.29 (0.010)	4.29 (0.454)	1.13 (0.015)	1.29 (0.009)	2.38 (0.014)	4.27 (0.406)	2.03 (0.014)
99% CI	[1.26 – 1.32]	[3.12 – 5.46]	[1.09 – 1.17]	[1.27 – 1.31]	[2.34 – 2.42]	[3.22 – 5.32]	[1.99 – 2.07]
Column nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**

Table 9 - Results of statistical tests to test the equality of parameters across latent classes. Applied for the models LCM_NB1 and HLCM_NB2

	DOF	Chi(DOF) value **LCM_NB1 (p)**	Chi(DOF) value **HLCM_NB2 (p)**
All slopes	34	196.7 (0.000)	106.8 (0.000)
Health status	14	104.3 (0.000)	26.6 (0.021)
Socioeconomic	15	31.1 (0.008)	42.2 (0.0002)
Insurance	2	0.37 (0.832)	0.15 (0.929)

Table 10 - Average marginal effect of selected variables for different models. Impact on the mean function

Model	**Poisson**	**NB1**	**HNB2**		**LCM_NB1**			**HLCM_NB2**			
									Health care users		
Covariate	*Overall pop.*	*Overall pop.*	*Overall pop.*	*Users pop.*	*Overall pop.*	*High users*	*Low users*	*Overall pop.*	*Users pop.*	*High users*	*Low users*
Income (elasticity)	0.036 (0.013)*	0.045 (0.010)*	0.053 (0.013)*	*-0.001* *(0.010)*	0.048 (0.012)*	0.069 (0.059)	0.042 (0.014)*	0.053 (0.013)*	*-0.002* *(0.01)*	0.007 (0.03)	-0.005 (0.01)
Education	0.008 (0.003)*	0.011 (0.003)*	0.007 (0.003)*	*-0.005* *(0.005)*	0.011 (0.003)*	0.037 (0.054)	0.010 (0.003)*	0.007 (0.003)*	*-0.004* *(0.005)*	-0.002 (0.027)	-0.005 (0.004)
Rural_Area	-0.103 (0.024)*	-0.107 (0.020)*	-0.107 (0.024)*	*-0.07* *(0.038)*	-0.106 (0.025)*	-0.259 (0.379)	-0.101 (0.024)*	-0.107 (0.024)*	*-0.07* *(0.038)*	-0.214 (0.209)	-0.04 (0.03)
NHS-only	-0.147 (0.045)*	-0.143 (0.035)*	-0.152 (0.045)*	*-0.103* *(0.070)*	-0.156 (0.04)*	-0.658 (0.638)	-0.124 (0.038)*	-0.152 (0.043)	*-0.101* *(0.070)*	-0.316 (0.368)	-0.06 (0.052)
Column nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**

Notes:

1. The income-elasticity reported in columns 1, 2, 6 and 7, is a parameter that comes out directly from estimation results. This is the case because income is included in logs in the linear index of the mean function, given by $E(y_i \mid x_i) = \exp\left(x_i'\beta\right)$.
2. All other figures reported in the previous table represent Average Marginal effects (AME), that is, the sample average of individual marginal effects (Cameron and Trivedi, 2005). The individual marginal effects are the derivative of the respective mean function, when one is evaluating the effect of a continuous covariate, or the individual discrete changes from 1 to 0, when one is evaluating the effect of a dummy variable. In the case of columns 6, 7, 10 and 11, the average effects were computed using only the individuals that were allocated to each class.

3. Concerning the popular hurdle model, with NB2 as baseline distribution in both stages: one can identify two mean functions; the mean for the users population, given by $E(y_i \mid x_i, y_i > 0) = \frac{\exp(x_i'\beta_2)}{1 - P_2(y_i = 0 \mid x_i)}$ [column 4], and the mean for the overall population, $E(y_i \mid x_i) = [1 - P_1(y_i = 0 \mid x_i, \beta_1)] * \frac{\exp(x_i'\beta_2)}{1 - P_2(y_i = 0 \mid x_i, \beta_2)}$ [column 3].
4. Regarding the model LCM_ NB1, the mean for the overall population [column 5] is given by equation 20. In what concerns the expressions of the mean functions for columns 8, 9, 10 and 11, see, respectively, equations 26, 25 and 24.
5. The
6. Standard errors, on parenthesis, were computed using the delta method, implemented in Stata. Details will be sent upon request.
7. * p < 0.001

Table 11 – LCM_NB1 and HLCM_NB1 estimation results

N = 42.501	**Hurdle Latent class model: NB2 as parent distribution**		
	First part	***Second part***	
		High users/user (17.8%)	***Low users/user (82.2%)***
Constant	-0.905 (0.088)	0.743 (0.257)	-0.596 (0.092)
Socioeconomic			
Age [/10]	-0.319 (0.031)	0.062 (0.061)	0.055 (0.029)
sqAge	0.035 (0.003)	-0.014 (0.007)	-0.005 (0.003)
female	0.322 (0.024)	-0.132 (0.058)	0.101 (0.024)
Married	0.314 (0.029)	-0.026 (0.062)	0.115 (0.025)
Education	0.022 (0.004)	-0.001 (0.009)	-0.004 (0.004)
Retired	0.305 (0.044)	0.229 (0.086)	0.045 (0.03)
Unemployed	-0.072 (0.067)	0.173 (0.144)	0.001 (0.066)
(log) Income	0.119 (0.02)	0.01 (0.042)	-0.01 (0.018)
North	0.265 (0.037)	-0.104 (0.089)	0.104 (0.039)
Centre	0.42 (0.04)	0.082 (0.093)	0.175 (0.039)
LVT	0.274 (0.04)	-0.129 (0.09)	0.128 (0.04)
Alentejo	0.25 (0.044)	-0.137 (0.1)	0.111 (0.044)
Rural_Area	-0.166 (0.031)	-0.068 (0.068)	-0.034 (0.027)
not_work	0.153 (0.031)	0.185 (0.069)	0.206 (0.03)
Health Status			
Diabetes	0.842 (0.064)	0.052 (0.097)	0.153 (0.031)
insulin	0.882 (0.238)	0.632 (0.182)	0.193 (0.078)
HBP	0.699 (0.035)	0.034 (0.062)	0.126 (0.023)
Asthma	0.052 (0.097)	0.052 (0.097)	0.052 (0.097)
Bronchitis	0.24 (0.072)	-0.067 (0.109)	0.128 (0.041)
Allergy	0.372 (0.033)	0.143 (0.061)	0.104 (0.024)
Backp	0.316 (0.026)	0.048 (0.059)	0.176 (0.024)
sick_long_Run	1.542 (0.17)	0.675 (0.157)	0.629 (0.053)
sick_Short_run	1.002 (0.025)	0.452 (0.051)	0.466 (0.022)
Limited	-0.126 (0.067)	0.439 (0.096)	0.091 (0.041)
Stress	0.875 (0.044)	0.285 (0.063)	0.245 (0.024)
NeverSmoked	-0.106 (0.025)	-0.161 (0.063)	-0.103 (0.025)
not physical activity	0.026 (0.036)	0.156 (0.085)	0.039 (0.036)
mild_exercise	-0.087 (0.042)	0.083 (0.102)	0.018 (0.043)
Supply side			
Phy_1000_Inhabitants	0.018 (0.005)	0.014 (0.012)	0.005 (0.005)
Insurance Status			
NHS-only	-0.22 (0.048)	-0.096 (0.107)	-0.05 (0.042)
ADSE	-0.205 (0.057)	-0.083 (0.127)	-0.053 (0.052)
Seasonality			
Winter	0.071 (0.03)	-0.102 (0.067)	0.009 (0.028)
Spring	0.185 (0.03)	0.024 (0.069)	0.082 (0.027)
Summer	0.066 (0.03)	-0.139 (0.069)	0.089 (0.028)
	---	1.242 (0.193)	0.158 (0.035)

Note: Robust standard errors in parenthesis

Funding

We are grateful for funding and support for the workshop from:

- The European Commission
- University of Macedonia
- University of York

MARIE CURIE ACTIONS