HEDG Health, Econometrics and Data Group
THE UNIVERSITY of York

HEDG Working Paper 05/01

Latent class models for utilisation of health care

Teresa Bago d'Uva

June 2005

ISSN 1751-1976

york.ac.uk/res/herc/hedgwp

Latent Class Models for utilisation of health care

Teresa Bago d'Uva*

Centre for Health Economics, University of York

July 8, 2005

Abstract

This paper explores different approaches to econometric modelling of count measures of health care utilisation, with an emphasis on latent class models. A new model is proposed that combines the features of the two most common approaches: the hurdle model and the finite mixture Negative Binomial. Additionally, the panel structure of the data is taken into account. The proposed model is shown to perform better than the existing models for a particular application with data from the RAND Health Insurance Experiment.

Keywords: count data, finite mixture models, hurdle model, panel data.

∘1. INTRODUCTION

In the recent literature, there are various examples of empirical modelling of count measures of health care. These measures include visits to different types of physi-

*Centre for Health Economics, Alcuin A Block, University of York, York, YO10 5DD, United Kingdom. Tel: 0044 1904 321411. Fax: 0044 1904 321402. E-mail: tbd100@york.ac.uk.

Funding from Fundação para a Ciência e Tecnologia is gratefully acknowledged (PhD grant: SFRH/BD/10551/2002). Data from the RAND Health Insurance Experiment were kindly provided by Partha Deb.

cians, hospital stays and number of medicines used. The count data models that have been most commonly used are the hurdle model (Pohlmeier and Urlich, 1995; Grootendorst, 1995, Hakkinen et al, 1996, Gerdtham, 1997) and the finite mixture Negative Binomial, FMNB (Deb and Trivedi, 1997, 2002; Gerdtham and Trivedi, 2000; Deb and Holmes, 2000). The hurdle model considers that the count measure of health care utilisation is a result of two different decision processes and implies a distinction between users and non-users. Deb and Trivedi argue that the FMNB is a better approach to model a count measure of health care use. It is argued that the distinction low versus high users is more adequate than the one between users and non-users of health care. This view is supported by the empirical applications of Deb and Trivedi (1997, 2002), Deb and Holmes (2000) and Gerdtham and Trivedi (2000). However, Jimenez et al (2002) compare the performance of the two models for visits to different types of doctor in 12 EU countries and find that the hurdle model performs better in some cases. More recently, Winkelmann (2004) showed that the FMNB also outperforms the traditional hurdle model in his application but a hurdle model with different distributional assumptions outperforms the FMNB.

This paper attempts to contribute to the debate by exploring different approaches to econometric modelling of count measures of health care. In particular, I explore the possibility of combining features of the FMNB and hurdle, in a single model: the finite mixture hurdle (FMH). This model draws on the FMNB model in that the individuals are considered to belong to latent classes. The innovation is that, similarly to the hurdle model, it allows for a two stage decision process, given the latent class. It is also shown how the finite mixture hurdle model can be specified as a panel data model (FMH-Pan). The individual heterogeneity is accounted for by considering the panel structure of the data in the formulation of the finite mixture. The FMH-Pan, nests a finite mixture negative binomial for panel data, which is itself an interesting alternative to the cross-section models previously used for health care

utilisation, when panel data are available.

The contribution of this paper is two-fold. First, it proposes a new approach to model health care utilisation that combines the features of the hurdle model and the finite mixture model. Second, within the finite mixture Negative Binomial framework, it suggests a way to account for unobserved individual heterogeneity, when panel data are available.

This paper is organised as follows: Section 2 introduces some count data models that are used for health care utilisation; Section 3 provides a proposal of combination of FMNB and hurdle models; Section 4 describes the data; Section 5 deals with some relevant identification issues; Section 6 defines finite mixture models for panel data; the results are given in Section 7; finally, Section 8 concludes.

2. ECONOMETRIC MODELS FOR THE NUMBER OF VISITS TO THE DOCTOR

Hurdle model

The hurdle model considers that the count measure of health care utilisation is the result of two different decision processes. The first part specifies the decision to seek care, and the second part models the positive values of the variable for the individuals who receive some care. This can be interpreted as a principal-agent type model, where the physician (the agent) determines utilisation on behalf of the patient (the principal) once initial contact is made. Thus, it is assumed that the decision to seek care is taken by the individual, while the level of care depends also on supply factors.

It has been defended in recent literature of health care utilisation that the two part hurdle model is a better starting point than the NB class (for example, Pohlmeier and Urlich, 1995, Grootendorst, 1995, Gerdtham, 1997). Another motivation for the

hurdle model is the high proportion of zeros that remains even after allowing for overdispersion.

The hurdle model for count data was proposed by Mullahy (1986). The participation decision and the positive counts are determined by two different processes $P_1(.)$ and $P_2(.)$. The log-likelihood for the hurdle model is given by:

$$\begin{aligned} LogL &= \sum_{y=0} \log\left[1 - P_1\left(y > 0|x\right)\right] \\ &\quad + \sum_{y>0} \left\{\log\left[P_1\left(y > 0|x\right)\right] + \log\left[P_2\left(y|y > 0, x\right)\right]\right\} \\ &= \sum_{y=0} \log\left[1 - P_1\left(y > 0|x\right)\right] + \sum_{y>0} \log\left[P_1\left(y > 0|x\right)\right] \\ &\quad + \sum_{y>0} \log\left[P_2\left(y|y > 0, x\right)\right] \\ &= LogL_1 + LogL_2. \end{aligned} \tag{1}$$

The two parts of the hurdle model can be estimated separately. For the participation decision, a binary model is defined. The second decision concerns the quantity of health care consumed, given it is positive, and is modelled by a truncated at zero count data model. The distribution P_1 is usually logit, probit, Poisson or NB, while the most common choices for P_2 are the Poisson and the NB.

In Mullahy (1986), the underlying distribution for both stages is the Poisson. Pohlmeier and Ulrich (1995) argue that it is necessary to account for remaining unobserved heterogeneity, since "supply side effects are rarely well captured in household data at the micro level". Thus, these authors use the NB1 distribution for both stages of the model, instead of the Poisson. They note this specification allows for explicit testing of distributional assumptions (for example, against Poisson) and the equality of the two parts of the decisionmaking process (thus, assessing the importance of considering that the number of physician visits is determined by two different processes). The NB (with $k = 1, 2$) is the most used distribution in empirical applications of the

hurdle model to the utilisation of health care. In particular, when the hurdle model is defined using the NB, the probability of zero visits and the distribution of utilisation conditional on some use are given by:

$$g\left(0|x_i\right) = P\left(y_i = 0|x_i, \beta_1\right) = \left(\lambda_{1i}^{1-k} + 1\right)^{-\lambda_{1i}^{k}} \tag{2}$$

$$g\left(y_i|x_i, y_i > 0\right) = \frac{\Gamma\left(y_i + \frac{\lambda_{2i}^{k}}{\alpha}\right)\left(\alpha\lambda_{2i}^{1-k} + 1\right)^{-\frac{\lambda_{2i}^{k}}{\alpha}}\left(1 + \frac{\lambda_{2i}^{k-1}}{\alpha}\right)^{-y_i}}{\Gamma\left(\frac{\lambda_{2i}^{k}}{\alpha}\right)\Gamma\left(y_i + 1\right)\left[1 - \left(\alpha\lambda_{2i}^{1-k} + 1\right)^{-\frac{\lambda_{2i}^{k}}{\alpha}}\right]}, \tag{3}$$

where $\lambda_{1i} = \exp\left(x_i'\beta_1\right)$ and $\lambda_{2i} = \exp\left(x_i'\beta_2\right)$.

Gurmu (1997) criticises the use of the NB distribution in hurdle models, in that it is based on the arbitrary assumption that the unobserved heterogeneity is gamma distributed. This author notes that, in the hurdle framework, a misspecification of the distribution of the unobserved heterogeneity leads to inconsistent estimates. In order to avoid this misspecification, Gurmu (1997) develops a semi-parametric hurdle model that does not require knowledge of the distribution of unobserved heterogeneity.

Winkelmann (2004) also moves away from the usual choice of the NB distribution but with a parametric approach. Winkelmann develops a hurdle model with different distributional assumptions, the probit-Poisson-log-normal. The probability of having at least one visit to the doctor is determined by a probit model. The positive part of the distribution is defined as a truncated Poisson-log-normal :

$$y_i|y_i > 0 \sim \text{truncated Poisson}\left(\lambda_i\right) \tag{4}$$

where $\lambda_i = \exp\left(x_i'\beta + u_i\right)$ and u_i follows a normal distribution with variance σ^2. Therefore, the probability of having y_i visits, given that y_i is positive, is given by:

$$g\left(y_i|x_i, y_i > 0\right) = \int_{-\infty}^{\infty} \frac{\exp\left(-\lambda_i\left(u_i\right)\right)\lambda_i\left(u_i\right)^{y_i}}{\left[1 - \exp\left(-\lambda_i\left(u_i\right)\right)y_i!\right]}\phi\left(u_i\right)du_i \tag{5}$$

Under the assumption of independence, the two parts of the model can be estimated

separately.[1] The likelihood of the second part of the model can be evaluated using Gauss-Hermite integration. Apart from the distribution used (log-normal instead of gamma), there is another difference concerning the way this model and the usual NB hurdle account for the unobserved heterogeneity. The NB hurdle considers a truncated NB for the second part, that is, a truncated version of a Poisson mixture. The model proposed by Winkelmann uses a mixture of a truncated Poisson distribution.

Finite Mixture Models (Latent Class)

Deb and Trivedi (1997) propose the use of finite mixture models as an alternative to the hurdle models in the empirical modelling of health care utilisation. In a more recent paper (2002) these authors point out that "a more tenable distinction for typical cross-sectional data may be between an 'infrequent user' and a 'frequent user' of medical care, the difference being determined by health status, attitudes to health risk, and choice of lifestyle". They argue that this is a better framework than the hurdle model, that distinguishes between users and non-users of care.

In a finite mixture model the population is assumed to be divided in C distinct populations in proportions $\pi_1, \ldots, \pi_C$, where $\sum_{j=1}^{C} \pi_j = 1$, $0 < \pi_j < 1$ $(j = 1, \ldots, C)$. The C-point finite mixture model is given by:

$$f\left(y_i|\Theta\right) = \sum_{j=1}^{C} \pi_j f_j\left(y_i|\theta_j\right), \tag{6}$$

where the mixing probabilities π_j are estimated along with all the other parameters, denoted Θ. Also, $\pi_C = 1 - \sum_{j=1}^{C-1} \pi_j$.

The component distributions in a C-point finite mixture negative binomial model, FMNB, are defined as:

$$f_j\left(y_i\right) = \frac{\Gamma\left(y_i + \psi_{j,i}\right)}{\Gamma\left(\psi_{j,i}\right)\Gamma\left(y_i + 1\right)} \left(\frac{\psi_{j,i}}{\lambda_{j,i} + \psi_{j,i}}\right)^{\psi_{j,i}} \left(\frac{\lambda_{j,i}}{\lambda_{j,i} + \psi_{j,i}}\right)^{y_i} \tag{7}$$

[1] Winkelmann also considered the case where the two parts are allowed to be correlated. However, no evidence of correlation was found.

where $j = 1, \ldots, C$ are the latent classes, $\lambda_{j,i} = \exp\left(x_i'\beta_j\right)$ and $\psi_{j,i} = (1/a_j)\,\lambda_{j,i}^k$.

Plugging the expression for $\psi_{j,i}$ in $f_j\left(y_i\right)$ leads to:

$$f_j\left(y_i|x_i\right) = \frac{\Gamma\left(y_i + \frac{\lambda_{ji}^k}{\alpha_j}\right)}{\Gamma\left(\frac{\lambda_{ji}^k}{\alpha_j}\right)\Gamma\left(y_i+1\right)}\left(\alpha_j\lambda_{ji}^{1-k}+1\right)^{-\frac{\lambda_{ji}^k}{\alpha_j}}\left(1+\frac{\lambda_{ji}^{k-1}}{\alpha_j}\right)^{-y_i}. \tag{8}$$

In the most general specification, all the elements of the vectors β_j as well as the overdispersion parameters α_j are allowed to vary across the latent classes. However, more parsimonious specifications can arise from restrictions on the components of β_j. For example, the slope parameters can be restricted to be equal across all latent classes. In this case, the differences between classes are given only by the intercept differences and the overdispersion parameters.

Deb and Trivedi (1997) point out a number of advantages of the finite mixture approach. It provides a natural representation since each latent class can be seen as a "type" of individual, while still accommodating heterogeneity within each class. It can also be seen as a discrete approximation of an underlying continuous mixing distribution, that does not need to be specified. Furthermore, the number of points of support needed for the finite mixture model in empirical applications is low, usually two or three.

In the finite mixture (latent class) formulation of unobserved heterogeneity, the latent classes are assumed to be based on the individual latent long-term health status, which may not be well captured by proxy variables such as self-perceived health status and chronic health conditions (Cameron and Trivedi, 1998). The two-point finite mixture model suggests the dichotomy between the 'healthy' and the 'ill' groups, whose demands for health care are characterised by, respectively, low mean and low variance and high mean and high variance.

Jimenez-Martin et al (2002) agree with the advantages of the finite mixture model described above but also note some disadvantages. Namely, while the hurdle model

is a natural extension of an economic model (the principal-agent model), the finite mixture model is driven by statistical reasoning.

3. AN EXTENDED MODEL: FINITE MIXTURE OF HURDLE MODEL

Unlike the hurdle model, the latent class model does not allow explicitly for a two-stage decision process. In the hurdle model, the probability of being a non-user of health care is modelled separately and the effects of the relevant factors are allowed to differ from the effects on the total number of visits, given that this is positive. Deb and Trivedi (1997, 2002) argue that the difference between low users and high users is more relevant than the difference between users and non-users. Therefore, the latent class model represents the unobserved heterogeneity in a finite number of latent classes. Each of these classes can be seen as representing a "type" of users.

Recent literature provide comparisons between the performance of the hurdle model and the latent class model in estimating health care use. In the empirical applications in Deb and Trivedi (1997, 2002) it is found that a two-point mixture of NB is sufficient to explain health care counts very well and outperforms the NB hurdle. Deb and Holmes (2000) also present evidence that the finite mixture model outperforms the hurdle model. However, Jimenez et al (2002) present evidence that, in some cases, the hurdle model can provide better results than the finite mixture model. They compare the hurdle and the finite mixture specifications for visits to specialists and GPs in 12 EU countries. It is found that the finite mixture model performs better for the visits to GPs while the hurdle model is preferred for specialists.

This paper investigates to what extent it is possible to combine features of both models in a single model. Drawing on the latent class model, it is proposed that the unobserved individual heterogeneity is represented by a finite number of classes. Each observation is assumed to be a random draw from a population in which there

are C distinct classes in proportions $\pi_1, ..., \pi_C$. These classes are heterogeneous as to the distribution of the number of visits. The novelty of the proposed model is that it introduces the two part feature of the hurdle model within a latent class framework. For each class, the hypothesis that the decision concerning the number of visits is taken in two steps is not discarded. This model allows for a two-stage decision process for each individual, given his class. Conditional on the latent class, the two decisions are independent.

Similarly to what was presented for the FMNB and for the hurdle model, the distribution used to define the finite mixture hurdle model (FMH) is the Negative Binomial. It can be argued that the model should not consider two different sources of heterogeneity (the overdispersion parameter in the NB and the finite mixture) and, therefore, the Poisson distribution should be used. However, since the aim is to compare the performance of the FMH with the models previously used in the literature, the same underlying distribution is used here. The mixture density in the FMH is given by[2]:

$$g\left(y|x\right) = \sum_{j=1}^{C} \pi_j f_j\left(y|\beta_{j1}, \beta_{j2}\right), \tag{9}$$

where $\sum_{j=1}^{C} \pi_j = 1$. For each component j, the probability of zero visits and the probability of observing y visits, given that y is positive, are given by the following expressions:

$$f_j\left(0|x\right) = P\left(y=0|x,\beta_{j1}\right) = \left(\lambda_{j1}^{1-k}+1\right)^{-\lambda_{j1}} \tag{10}$$

$$f_j\left(y|x,y>0\right) = \frac{\Gamma\left(y+\frac{\lambda_{j2}^k}{\alpha_j}\right)\left(\alpha_j\lambda_{j2}^{1-k}+1\right)^{-\frac{\lambda_{j2}^k}{\alpha_j}}\left(1+\frac{\lambda_{j2}^{k-1}}{\alpha_j}\right)^{-y}}{\Gamma\left(\frac{\lambda_{j2}^k}{\alpha_j}\right)\Gamma\left(y+1\right)\left[1-\left(\alpha_j\lambda_{j2}^{1-k}+1\right)^{-\frac{\lambda_{j2}^k}{\alpha_j}}\right]}, \tag{11}$$

where $\lambda_{j1} = \exp\left(x\beta_{j1}\right)$, $\lambda_{j2} = \exp\left(x\beta_{j2}\right)$ and $j = 1, ..., C$.

[2]The subscripts i are omitted here for simplicity.

Similarly to the hurdle model, the fact that β_{j1} can be different from β_{j2}, reflects the possibility that the zeros and the positives are determined by two different processes. On the other hand, in line with the finite mixture model, having $\left[\beta_{j1}\beta_{j2}\right] \neq [\beta_{l1}\beta_{l2}]$ when $j \neq l$ reflects the differences between the latent classes. The same set of regressors is considered in both parts of the model and for all the classes. As usual in the hurdle model, all the effects are allowed to differ in both parts of the model. As to the variation between classes, it can be assumed that all the slopes are the same, varying only the constant terms, $\beta_{j10}$and β_{j20}, and the overdispersion parameters α_j. Alternatively, these restrictions might not be imposed, allowing for all the parameters to differ. The FMH also accommodates a mixture of sub-populations for which health care use is determined by a NB (the two decision processes are indistinguishable) and sub-populations for which utilisation is determined by a hurdle model. This is obtained by setting $\beta_{j1} = \beta_{j2}$, for some classes j.

Other possible restricted versions of the FMH are obtained when it is assumed that the classes differ only regarding one of the parts of the model, assuming that the parameters of the other part are the same across classes. Consider that the latent classes differ in the positive part of the distribution but not in the binary part. In other words, the individuals are homogeneous as to the probability of seeking medical care, whereas the individual unobserved heterogeneity in the distribution of the conditional number of visits is accounted for by the finite mixture. This restricted model is close to the model proposed by Winkelmann (2004) when the two parts are considered to be uncorrelated.[3] Consider further that the truncated Poisson distribution is used for the conditional part of the model instead of the truncated Negative Binomial. This restricted FMH is composed by a binary part (without mixture) and a conditional part which consists of a finite mixture of truncated Poisson distributions. The two

[3]This is the preferred model in Winkelmann (2003), since there was no evidence of correlation between the binary step and the conditional distribution.

parts can be estimated separately. Subject to the same choice of binary distribution, the first part corresponds to the first part of the hurdle model in Winkelmann (2004). The conditional part corresponds to a discrete approximation of the conditional part considered by Winkelmann (2004).

4. DATA

This paper uses data from the Rand Health Insurance Experiment. The choice is based on the advantages of these data to model utilisation of health care. Since individuals were randomised into health insurance plans, the characteristics of the plan the individual is in are exogenous. This overcomes one of the major difficulties in modelling health care use, since the features of health insurance coverage are among the most important determinants of the demand for health care. Another advantage of these data is the reliability of health care utilisation variables. One of the drawbacks of using survey data to model the number of visits to the doctor is the fact that the reference period is quite long (usually a year) and recall errors are likely to occur. The dataset used is the same that was used in Deb and Trivedi (2002) to assess the performances of the FMNB and the hurdle model. This dataset was kindly provided by Partha Deb. The features of the dataset that are most important for the application presented are pointed out below. A full description of the variables used can be found in Deb and Trivedi (2002).

The individuals were assigned to insurance plans for a period of 3 or 5 years. The sample consists of 20186 observations (each observation is an individual in a given year).

The utilisation variable that is used here is the total number of outpatients visits during a year (there are 28.3% of zero observations and the maximum is 147). The covariates are in line with the factors usually considered to determine demand for health care. The socio-economic characteristics include the logarithm of family in-

come, $LINC$, and the number of years of education of household head, $EDUCDEC$. The health status variables considered are: a dummy for the existence of a physical limitation, $PHYSLIM$; an index of chronic diseases, $DISEA$; and three dummies indicating whether the individual considers his health to be good, fair or poor (the category excellent is omitted). As to the insurance plan, the variables included are the logarithm of the coinsurance rate plus 1, LC[4]; a dummy for the existence of deductible in the plan, IDP; a participation-incentive payment function, LPI; and a maximum expenditure function, $FMDE$.

5. SOME IDENTIFICATION ISSUES

The estimation of the FMH encountered some problems that cast doubt on the identifiability of the model. The problems occurred in the mixture of the binary part. As noted by McLachlan and Peel (2000), in order to estimate the parameters in a mixture distribution, the mixture should be identifiable. These authors define a mixture as identifiable if two sets of parameters which do not agree after permutation do not yield the same mixture distribution. Teicher (1960) proved the identifiability of finite mixtures of Poisson distributions without covariates. For the class of Poisson regression mixtures, Wang et al (1996) show that a sufficient condition for identifiability is that the matrix of regressors is of full rank. The concern here is the identifiability of mixture of two part models but this case is not covered by the literature. Nevertheless, there are some theoretical results for a related case that might be relevant, the mixture of binomial models, which includes as special case the mixture of binary models (when $m = 1$). The interest in this case lies in the problems encountered in the estimation of the mixture of hurdles in the first (binary) part of the model. The identifiability conditions involve a bound on the number of latent classes, which is usually a function of m. Teicher (1963) and Wang (1994) give the following condition

[4]The insurance plans had coinsurance rates equal to 0, 25, 50 or 95.

for identifiability of the mixture with C classes:

$$C \leq \frac{m+1}{2}. \tag{12}$$

Condition (12) implies that, for the binary case, no mixture model is identified, since $C \leq 1$.

Another relevant example of latent class models that has been considered in the literature is the case of multiple binary outcomes. Consider that the vector of binary variables $y = \left[\begin{array}{cccc} y_1 & y_2 & \ldots & y_T \end{array} \right]$ has the following distribution:

$$g(y) = \sum_{j=1}^{C} \pi_j \prod_{t=1}^{T} \theta_{jt}^{y_t} (1-\theta_{jt})^{1-y_t}.$$

This case can also represent a binary response observed over a panel of T periods. This framework implies the assumption that the individuals belong to the same latent class throughout the observed years. In other words, the latent class structure represents a discrete distribution of the time-invariant unobservables at the individual level. A necessary condition for identification of this model is presented, for example, in Garrett (2000). The maximum number of classes that can be identified is:

$$C < \frac{2^T+1}{T+1}. \tag{13}$$

When $T = 1$, this model reduces to the binary model. In line with the conditions given by Teicher (1963) and Wang (1964), the condition imposes the upper bound of one in the number of classes that can be identified in the binary model. Repeated observations of the binary variable assure identification of the mixture model, since, for $T > 1$, the upper bound is equal or greater than 2.

In an attempt to shed some more light on the identification issues for mixtures of binary models, some simulations were made, covering models with cross-section data and panel data. Consider the Panel mixture logit, defined in the following way:

$$g(y|x) = \pi \times \prod_{t=1}^{3} f_1(y_t|x_t) + (1-\pi) \times \prod_{t=1}^{3} f_2(y_t|x_t) \tag{14}$$

where $y = [y_t]$ and $x_t = [x_t]$ and, for $t = 1, 2, 3$ and $j = 1, 2$:

$$f_j\left(y_t|x_t\right) = \left(\frac{1}{1+\exp\left(x_t'\beta_1\right)}\right)^{y_t=0}\left(\frac{\exp\left(x_t'\beta_1\right)}{1+\exp\left(x_t'\beta_1\right)}\right)^{y_t=1}. \tag{15}$$

Within this framework, the repeated individual observations are assumed to belong to the same latent class.

Consider also the cross-section Mixed logit:

$$g\left(y_t|x_t\right) = \pi \times f_1\left(y_t|x_t\right) + (1-\pi) \times f_2\left(y_t|x_t\right)$$

where $f_j\left(y_t|x_t\right)$ is as defined in (15).

The vector of regressors x_t is composed by a constant term, x_0, a continuous variable, x_1, that follows the standard normal distribution and a dummy variable, x_2 The size of the samples drawn in the simulations is 30000 observations, which correspond to 10000 individuals over 3 years. For each sample, three models are estimated and compared:

- Mixed logit with pooled sample**.**
- Mixed logit with cross-section year 3.
- Mixed panel logit with full sample.

The third model is the true model used in the simulation. The second model is also the true model for a cross-section of observations. This model corresponds to a situation when there is only data for a single year or the remaining years are ignored. As to the first model, it ignores the panel structure of the data or, equivalently, the information that each individual belongs to the same latent class across the panel is not accounted for. Two cases concerning the variation in the parameters across classes are considered: first, the constant term is allowed to vary and the slopes are the same for both classes; second, constant and slopes are allowed to vary between classes. For each case, three samples of the true model Panel mixture logit were drawn. The performance of the three estimated models was then compared.

The performance of the panel mixture logit was always good, providing precise estimates that are close to the true parameters. Moreover, the fit always improves substantially when compared to the one component logit. As to the estimations of the cross-section mixed logit, in some cases, the estimated parameters were very different from the true parameters and their precision was generally poor. Another signal of non-identifiability of the mixture is the fact that it does not provide a better fit than the one component model (even when the parameter estimates are satisfactory). This happens in some of the examples with the cross-section mixed logit. In short, these simulations show that the estimation of binary models with cross-section or pooled data can give unsatisfactory results.

While, in the simulations, the performance of the binary mixture model with panel data was very good, showing no problems of identifiability, it should not be forgotten that these problems can arise with real data if these do not provide enough information to estimate all the parameters of the mixture precisely. In order to assess the behaviour of this model with the RHIE data, this was estimated using a binary variable that equals 1 if the number of visits to the doctor is positive and 0, otherwise. The panel mixed logit was estimated for a balanced panel of 5325 individuals over 3 years. The coefficients of the model are generally well determined and the model provides a considerable better fit than the one component logit.

6. FINITE MIXTURE OF HURDLE MODEL WITH PANEL DATA

Problems encountered in the estimation of the FMH with pooled data, the identification issues covered in section 5, and the fact that dataset used in this paper is a panel of individuals motivate a reformulation of the FMH as a panel data model. A further motivation is the previous use of the binary mixture model for panel data (or with multiple binary responses in a cross-section) in empirical modelling of health care utilisation. Atella et al (2004) model the probability of visiting three types of

physician jointly. The individuals are assumed to belong to one of C latent classes. Within each latent class, the decision to visit each physician type follows a probit distribution. An example of a binary mixture model with panel data is the discrete random effects probit. Deb (2001) uses a latent class model where only the constant varies across classes. The discrete random effects probit is a discrete approximation of the distribution of the unobserved family effects in the random effects probit.

The FMH presented in section 3 is reformulated in order to account for the panel structure of the data in the same way as the Panel mixture logit in (14). This is the same way that the individual effects are accounted for in Atella et al (2004). The finite mixture represents individual unobserved time-invariant heterogeneity. For each class, the number of visits to the doctor in a given year is determined by a hurdle model. The model is estimated for an unbalanced panel of 5908 individuals over T_i years, $T_i = 1, ..., 5$. The finite mixture hurdle for panel data, FMH-Pan, is defined by the joint distribution of the number of visits across the panel:

$$g\left(y_i|x_i\right) = \sum_{j=1}^{C} \pi_j \prod_{t=1}^{T_i} f_j\left(y_{t,i}|x_{t,i}\right) \tag{16}$$

where $y_i = [y_{t,i}]$ and $x_t = [x_{t,i}]$ and, for $t = 1, ..., T_i$ and $j = 1, ..., C$:

$$f_j\left(0|x_{i,t}\right) = P\left(y_{i,t} = 0|x_{i,t}, \beta_{j1}\right) = \left(\lambda_{j1,i,t}^{1-k} + 1\right)^{-\lambda_{j1,i,t}} \tag{17}$$

$$f_j\left(y_{i,t}|x_{i,t}, y_{i,t} > 0\right) = \frac{\Gamma\left(y + \frac{\lambda_{j2,i,t}^{k}}{\alpha_j}\right)\left(\alpha_j \lambda_{j2,i,t}^{1-k} + 1\right)^{-\frac{\lambda_{j2,i,t}^{k}}{\alpha_j}}\left(1 + \frac{\lambda_{j2,i,t}^{k-1}}{\alpha_j}\right)^{-y_{i,t}}}{\Gamma\left(\frac{\lambda_{j2,i,t}^{k}}{\alpha_j}\right)\Gamma\left(y_{i,t}+1\right)\left[1 - \left(\alpha_j \lambda_{j2,i,t}^{1-k} + 1\right)^{-\frac{\lambda_{j2,i,t}^{k}}{\alpha_j}}\right]}. \tag{18}$$

where $\lambda_{j1,i,t} = \exp\left(x_{i,t}\beta_{j1}\right)$, $\lambda_{j2,i,t} = \exp\left(x_{i,t}\beta_{j2}\right)$.

Similarly to the FMH defined by equations (9) to (10), having $\beta_{j1} \neq \beta_{j2}, j = 1, ..., C$, allows for the zeros and the positives to be determined by two different processes. The latent class framework is reflected by the inequalities $\left[\beta_{j1}\beta_{j2}\right] \neq$

$[\beta_{l1}\beta_{l2}]$ for $j \neq l$. The same types of restrictions that were mentioned for the FMH in Section 3 can be considered for the FMH-Pan.

The specification of the finite mixture hurdle as a panel data model has some attractive features. Firstly, it is a hurdle model for panel data. The previous applications of the hurdle model used, in general, cross-section data. A notable exception is Van Ourti (2004), that considers a hurdle model for panel data with a common random effect in both parts. However, Van Ourti (2004) does not consider individual heterogeneity in the effects of the covariates. The FMH-Pan accounts for unobserved individual heterogeneity in the constants and the slopes by considering directly the panel structure of the data in the formulation of the mixture. Secondly, it allows the identification of the mixture, which was problematic in the pooled FMH. Thirdly, it disentangles the differences across latent classes in differences on the probability of visiting a doctor and differences on the conditional positive number of visits. Finally, the FMH-Pan accounts for the correlation between the two parts (as the models in Winkelmann, 2004, and Van Ourti, 2004, do). The two stages of the decision process are independent for each class of users but, if one does not condition on the class, the hurdle step is correlated with the distribution of the positive number of visits. Thus, the FMH-Pan does not impose independence between the two stages, since it allows for common unobservables to affect both.

On the other hand, some disadvantages of the FMH-Pan should be noted, such as the requirement for panel data. Additionally, a rich dataset is needed to allow to identify all the parameters of the model, that is much less parsimonious than the FMNB and the hurdle. It is possible to consider more restricted versions of the FMH-Pan like the examples given in Section 3.

Another interesting feature of the FMH-Pan is the fact that it nests a FMNB for panel data, FMNB-Pan. The FMNB-Pan is obtained from the FMH-Pan by imposing $\beta_{j1} = \beta_{j2}$, $j = 1, ..., C$. In this case, for each latent class, the zeros and the positives

are determined by the same distribution. The FMNB-Pan is defined by:

$$g\left(y_i|x_i\right) = \sum_{j=1}^{C} \pi_j \prod_{t=1}^{T_i} f_j\left(y_{t,i}|x_{t,i}\right) \tag{19}$$

where $y_i = [y_{t,i}]$ and $x_i = [x_{t,i}]$ and, for $t = 1, ..., T_i$ and $j = 1, ..., C$:

$$f_j\left(y_{i,t}|x_{i,t}, y_{i,t} > 0\right) = \frac{\Gamma\left(y + \frac{\lambda_{j1,i,t}^k}{\alpha_j}\right)}{\Gamma\left(\frac{\lambda_{j1,i,t}^k}{\alpha_j}\right)\Gamma\left(y_{i,t}+1\right)} \left(\alpha_j \lambda_{j1,i,t}^{1-k} + 1\right)^{-\frac{\lambda_{j1,i,t}^k}{\alpha_j}} \left(1 + \frac{\lambda_{j1,i,t}^{k-1}}{\alpha_j}\right)^{-y_{i,t}}. \tag{20}$$

The FMNB-Pan model can be tested against the more flexible FMH-Pan using a log-likelihood ratio test of equality of the parameters of the two parts. This test is a way of assessing the importance of allowing for two different decision processes. The FMNB-Pan differs from FMNB (Deb and Trivedi, 2002) in that it accounts for the panel structure of the data, which was not explored in that paper to identify individual heterogeneity. Comparison of the non-nested models FMNB and FMNB-Pan shows the relevance of accounting for the panel structure in the latent class framework.

7. RESULTS

This section presents the estimation results for the pooled models, Hurdle and FMNB (pooled sample of 20186 observations), and for the panel data models, FMNB-Pan and FMH-Pan (unbalanced panel of 5908 individuals over 1 to 5 years). For all the models, the parent distribution is the NB1. Thus, the models are defined by the expressions given in Sections 2 and 6, taking $k = 1$. For the mixture models, I considered $C = 2$, which corresponds to having two latent classes in the population. This is in line with the FMNB model estimated in Deb and Trivedi (2002, these authors used only two classes after preliminary analysis indicated overparameterisation of the models with three classes). The two latent classes can be seen as low users and high users, or healthy and ill.

All the estimations were done by maximum likelihood using TSP 4.5. For the hurdle model with one component, the Newton method is used. The Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm is implemented in TSP as an option of the maximum likelihood estimation. This is the method used to estimate the mixture models presented here. Alternative methods could be used for the mixture models, such as the EM algorithm (Dempster et al, 1977) and its extensions (stochastic EM and simulated annealing EM). The description and discussion of these methods can be found, for example, in Celeux et al (1995) and McLachlan and Krishnan (1997).

Due to the possibility of convergence to local maxima in mixture models, the estimation should be repeated using different sets of starting values for the parameters being estimated. These starting values can be obtained as combinations of the estimates of the one component version of the model. Another option is to use as starting values the estimates of a mixture model with a restricted specification for the component densities (for example, use the results of the Poisson mixture as starting values for the NB mixture).

The comparison of the estimated models makes use of several criteria. For the non-nested models, the most common information criteria are used: AIC $= -2 \ln L + 2K$ and BIC $= -2 \ln L + K \ln N$, where $\ln L$ is the maximised likelihood, K is the number of estimated parameters and N is the sample size. The following pairs of non-nested models are compared according to AIC and BIC: Hurdle, FMH-Pan; Hurdle, FMNB; Hurdle, FMNB-Pan; and FMNB, FMNB-Pan.

The FMNB and the FMNB-Pan are further compared in greater detail in order to assess the importance of accounting for the panel structure of the data in the estimation of the finite mixture of NB. Analysis of the estimates of both models gives an insight to the extent to which the estimated covariate effects and their precision change. Additionally, I used Vuong's (1989) test for model selection among non-

nested, which test statistic is given by:

$$V = \frac{\sum_{i=1}^{N} \left(\log g\left(y_i|x_i\right) - \log f_P\left(y_i|x_i\right) \right)}{\sqrt{\sum_{i=1}^{N} \left(\log g\left(y_i|x_i\right) - \log f_P\left(y_i|x_i\right) \right)^2}}.$$

$g\left(y_i|x_i\right)$ is the joint density of $y_i = [y_{t,i}]$ in the FMNB-Pan, defined by equations (19) and (20). The logarithm of the joint density of y_i in the FMNB is given by $\log f_P\left(y_i|x_i\right) = \sum_{t=1}^{T_i} \log f\left(y_{t,i}|x_i\right)$, where $f\left(y_{t,i}|x_i\right)$ is defined by (6) and (8).[5] Under the null-hypothesis that the FMNB and the FMNB-Pan are equivalent, the test statistic follows a standard normal distribution.

Since the FMH-Pan nests the FMNB-Pan, the null hypothesis that the "nesting" restrictions hold is tested using a log-likelihood ratio test. These models are further compared as to the estimated proportion of positive number of visits in the sample, expected number of visits as well as effect of price of health care on number of visits.

7.1 Pooled models - Hurdle and FMNB[6]

Hurdle

The estimation results for the hurdle model are given in Table 1:

Table 1 about here

In general, the results are as expected, indicating higher demand for people in poorer health and higher demand for women, which is observed consistently in both parts of the model. There are some differences in the slopes of the two parts of the model. For example, the health factors have greater effects on the positive part of the distribution. As to the characteristics of insurance plan, the results show that the

[5] The pooled model assumes independence of the individual observations across time, therefore, the joint log-likelihood of the individual observations is the summation of the log-likelihoods of those observations.

[6] Only a brief presentation of the results is included here since these were covered by Deb and Trivedi (2002).

coinsurance rate and the existence of deductible have a significant negative impact on both parts of the model but this is more significant on the probability of seeking medical care. The maximum expenditure function has a positive and significant effect on the probability of visiting a doctor while its estimated effect on the conditional number of visits is not significantly negative.

Finite mixture Negative Binomial - FMNB

Table 2 presents the results for the finite mixture NB:

Table 2 about here

All the estimates have the expected signs. The estimated class proportions are 79% for the low users and 21% for the high users. The two classes are easily distinguished by the overdispersion parameters and the constant terms. Moreover, the slopes are allowed to differ. The slopes of the two classes are jointly different and these differences occur especially for the slopes of the demographic variables and the health variables.

Table 3 presents the log-likelihood, AIC and BIC of the Hurdle and the FMNB.

Table 3 about here

As was seen in Deb and Trivedi (2002), the three criteria favour the mixture NB against the hurdle model.

7.2 Panel data models - FMNB-Pan and FMH-Pan

Finite mixture Negative Binomial for panel data - FMNB-Pan

Table 4 about here

As Table 4 shows, in the FMNB-Pan, the probability of belonging to the class of high users is 0.25. The results for the FMNB-Pan can be compared to the results for the FMNB with pooled data, Table 2. It is interesting to compare the models as to the estimates obtained, precision of the estimates and adjustment. In general,

the parameters are more precisely estimated in the panel data model, especially in the case of the class proportions, overdispersion parameters and coefficients of class of high users. As to the estimated coefficients, all of them have the same sign in both models, except for the case of the effect of $CHILD$ for high users. In the pooled model, this coefficient is negative, although not significant. In the panel model, the effect of that dummy becomes significantly positive. LPI and AGE are not significant in the class of high users in the pooled model but become significant when the panel structure is accounted for. The negative effect of the family size is more significant in the pooled model, for both classes of users.

Apart from being more precisely estimated, the estimated overdispersion parameters are smaller in the panel model. A possible explanation for this is that the individual heterogeneity is more accurately taken into account in the panel model, where the finite mixture represents the individual time-invariant heterogeneity. The self-assessed health variables are less significant in panel model. Similarly to what was said for the overdispersion parameters, this might mean that the panel model accounts for unobserved individual heterogeneity (that comprises unobserved health status) in a better way, thus diminishing the relevance of observed self-assessed health. The panel model fits the data better than the pooled model, despite having the same number of parameters. Additionally, the Vuong test statistic for the comparison of the FMNB-Pan and FMNB is $V = 17.6$, which clearly rejects the pooled specification in favour of the panel specification.

There are significant differences between the two classes of users obtained with the FMNB-Pan. The differences in the constant terms as well as overdispersion parameters are easily noted. In order to assess to what extent the two classes respond differently to the covariates considered, tests of equality of all the slopes as well as of slopes of some groups of variables were performed. The definition of the groups was borrowed from Deb and Trivedi (2002): insurance variables, health variables

and demographics (this group includes the remaining covariates). Table 5 shows the results of the log-likelihood ratio tests of equality of slopes:

Table 5 about here

When considered jointly, the responses of the two classes of users to the covariates are significantly different. The same occurs for the groups of insurance variables and demographics. However, in what concerns the effect of health on the number of visits, the difference between the classes is only significant at a 10% level.

The expected number of visits to the doctor as well as the proportion of positive observations, as estimated by the FMNB-Pan, are shown in Table 6. These were calculated for each class and for the overall sample using the following procedure. First, the posterior probabilities for each individual were calculated using the usual expressions:

$$P[y_i \in \text{class } j] = \frac{\pi_j \prod_{t=1}^{T_i} f_j\left(y_{t,i}|x_{t,i}\right)}{\sum_{j=1}^{2} \pi_j \prod_{t=1}^{T_i} f_j\left(y_{t,i}|x_{t,i}\right)}, j = 1, 2.$$

The class of each individual i, $i = 1, ..., 5908$, was chosen according to the highest posterior probability. Then, conditional on the class, $E[y_{i,t}|x_{i,t}]$ and $P[y_{i,t} > 0|x_{i,t}]$ were obtained for individuals i, $i = 1, ..., 5908$, in year t, $t = 1, ...T_i$, using the estimated distribution of visits for the respective class. In order to obtain the fitted values for the sample, and for each class, simple averages were taken over all the respective observations.

Table 6 about here

The two classes differ significantly in terms of the expected number of visits. The results also show that the estimated probability of visiting the doctor at least once is significantly higher for the class of high users.

Finite mixture hurdle for panel data - FMH-Pan

Table 7 about here

Table 7 shows the estimates of the parameters of the model FMH-Pan, with a

class of high users with probability 0.28, and a class of low users with probability 0.72. The results are in line with what is commonly found in empirical models of utilisation of health care. Consistently across classes and in both parts of the models, there is a positive effect of income, being female and in poor health. All the coefficients of the variables denoting coinsurance and existence of deductible are negative. The parameters in the model seem to be well identified, except for the coefficient of $HLTHP$ in the first part of the model for high users. This can be explained by the small proportion of observations for which $HLTHP = 1$ (0.015), making it difficult to identify four parameters for this dummy variable. The log-likelihood ratio tests of joint significance of groups of parameters support that, in general, the estimates are well determined. Table 8 presents the log-likelihood ratio test statistic for these tests. The unrestricted model is the FMNH-Pan in table 7 and, for each cell, the restricted model corresponds to setting the respective parameters to 0.

Table 8 about here

As to the latent class structure of the model, significant differences are found between the high users and the low users. The FMH-Pan disentangles the differences between classes in differences in the probability and differences in the conditional number of visits. Firstly, the comparison of the constant terms across classes for both parts as well as the overdispersion parameters in the positive part, shows that the two classes differ significantly in both parts of the decision process. Moreover, differences can also occur in the responses to the covariates. In order to better assess the differences between the two classes of individuals, a set of tests were performed. The p-values of tests of equality of parameters across classes are given in Table 9.

Table 9 about here

There are significant differences in the effects of the covariates on the probability of seeking health care. This is observed when all the slopes are considered jointly as well as by sub-groups. As to the conditional number of visits, the two classes differ

significantly in the constant term and in the overdispersion parameter. However, there is not enough evidence to conclude that the effects of the covariates on the conditional positive number of visits to the doctor vary across classes.

The main feature introduced with the finite mixture hurdle model in the latent class framework is the fact that it allows for the two different decisions to be driven by different distributions. It is now possible to evaluate the extent to which the parameters of the two decision processes can be considered different, for each latent class. Once again, this is done using log-likelihood ratio tests of equality of parameters. Table 10 summarises.

Table 10 about here

For high users, the effects of insurance variables on the probability of zero visits are not significantly different from the effects on the number of visits, given this number is positive. Within the higher users, the remaining effects vary significantly depending on the stage of the decision process. For low users, all the sub-groups of covariates considered impact differently the probability of use and the conditional use.

Considering figures of tables 8 to 10, it is possible to assess the relative importance of some factors to determine the zeros and the positives (and, similarly, the relative importance between classes). For example, health status is more relevant on the conditional number of visits than in the probability of having at least one. Regarding the effect of health on the probability of use, this is more significant for low users than for high users.

Table 11 presents the estimated proportion of positives, expected conditional positive number of visits and expected value of visits.

Table 11 about here

The two classes have significantly different probabilities of going to the doctor as well different expected number of visits, given this is positive. Consequently, the expected number of visits is multiplied by almost 5 when passing from the low class

to the high class.

Table 12 allows a better understanding of the effect of insurance on utilisation of health care. The expected number of visits as well as the probability of use and the expected positive number of visits, were calculated for each individual, conditional on the individual covariates and setting the coinsurance rate equal 0, 25 and 95.

Table 12 about here

The effect of price of health care, measured by the coinsurance rate, is consistently negative both on the probability of seeking medical care and on the positive number of visits. The combined effect on the expected number of visits is greater for the class of low users. The difference between the response to price of high and low users is driven mainly by the difference in the probability of seeking medical care.

7.3 Model selection

Table 13 shows the log-likelihood, AIC and BIC for the estimated models.

Table 13 about here

The information criteria AIC and BIC are used in comparisons of non-nested models. The hurdle model is outperformed by all the alternative models. In particular, the fact that the FMH-Pan outperforms the Hurdle is evidence that there is a mixture within the hurdle framework (since the information criteria are the preferred criteria to assess the existence of finite mixture). A better performance of the FMNB over the hurdle had been already shown by Deb and Trivedi (2002) for these data. However, the FMNB is now outperformed by its panel version, FMNB-Pan. The latter is also preferred to the former according to the Vuong test, presented in 7.2.

The information criteria are not the preferred criteria to compare the FMNB-Pan and the FMH-Pan, since the 2 models are nested. Nevertheless, these are worth looking at. The FMH-Pan outperforms the FMNB-Pan even when it is penalised for the number of parameters in the way the AIC does. However, according to the BIC, that

penalises more heavily the number of parameters, the FMNB-Pan would be preferred. This criterion highlights the lack of parsimony of the FMH-Pan. More parsimonious versions of the model can be obtained by imposing restrictions of equality on groups of parameters that are not significantly different (such as the parameters of the insurance variables on both parts of the model for the class of high users, or the slopes of all the variables for both classes in the second part of the model). Both of these restricted versions of the FMH-Pan outperform the FMNB-Pan for both information criteria. However, the preferred criteria to compare the FMH-Pan and the FMNB-Pan is the hypothesis test for the "nesting" restrictions (36 restrictions of equality of the coefficients of the 18 covariates in both parts, for each class). The log-likelihood ratio test statistic of the FMH-Pan against the FMNB-Pan equals 294.8 which clearly leads to reject the null hypothesis that the two parts of the model are determined by the same distribution. The two models can further be compared in terms of the expected number of visits and the expected proportion of positives. The results in tables 6 and 11 show that the estimated probability of going to the doctor according to each model is very close the sample proportion 0.72, being only slightly closer for the FMH-Pan. As to the expected number of visits, the estimate for this model is again slightly closer to the sample average than the FMNB-Pan. The two models are further compared as to the fitted values for different coinsurance rates in table 14.

Table 14 about here

The relative increase in the expected number of visits is higher when the price goes from 0 to 25 than from 25 to 95. The FMH-Pan provides better estimates of the expected number of visits overall and for each of the categories of coinsurance. In terms of relative changes, both models provide an estimate that is very close the sample average when the coinsurance rate goes from 25 to 95. As to the change when the coinsurance rate goes from 0 to 25, the estimate given by the hurdle model is closer to the sample value.

In general, the panel models perform better than the pooled models, showing the relevance of accounting for the panel structure of the data within the latent class framework. Among the panel models, the hypothesis that the two stages of the model can be aggregated in a NB for each class of users is not accepted, which gives preference to the FMH-Pan over the FMNB-Pan. The lack of parsimony of the hurdle mixture was, however, highlighted by the BIC making it reasonable to consider more parsimonious versions of the FMH-Pan.

8. CONCLUSIONS

This paper explores different approaches to econometric modelling of count measures of health care utilisation. A new approach is proposed that draws on the hurdle model (Pohlmeier and Urlich, 1995) and the finite mixture model (Deb and Trivedi, 1997, 2002). The latent class framework is maintained, whilst not discarding the hypothesis that the decision regarding utilisation of health care is made in two stages.

The number of outpatient visits in a year is modelled using data from the Rand Health Insurance Experiment. The dataset is an unbalanced panel of individuals which motivates the specification of the finite mixture model as a panel data model. The finite mixture represents the individual time-invariant unobserved heterogeneity.

The finite mixture hurdle identifies two classes of users with very different levels of health care utilisation and different responses to the covariates considered (health status, features of insurance plan, demographics). The hurdle feature of the model disentangles the differences across latent classes in the probability of visiting a doctor and the conditional positive number of visits. Additionally, the latent class specification of the hurdle model does not rule out a correlation between the two parts, as is usually done with the standard hurdle model (Winkelmann, 2004, is an exception)

Similarly to what was found with the FMNB, the effect of price is greater for low users. The FMH-Pan allows to further understand that this difference is driven

mainly by the difference in the probability to visit a doctor. Additionally, it is found that the effect of health status is more relevant on the conditional number of visits than on the probability of having at least one. Regarding the effect of health on the probability of use, this is more significant for low users than for high users.

The FMH-Pan accommodates more parsimonious versions that can be of interest. For example, considering the slopes to be the same across classes, while the constant is allowed to vary, one obtains a random effects hurdle model with a flexible distribution of the individual effects. The FMH-Pan also nests a finite mixture model for panel data in which the number of visits for each class is determined by a Negative Binomial distribution. The FMNB-Pan differs from FMNB (Deb and Trivedi, 2002) in that it accounts for the panel structure of the data. Comparison of the results of the non-nested models FMNB and FMNB-Pan shows the relevance of accounting for the panel structure in the latent class framework. While in this application the FMNB-Pan is rejected in favour of the more flexible FMH-Pan, and does not allow for a two-stage decision process for each class, it might actually be preferred in some cases. The panel version of FMNB is itself an interesting alternative to the cross-section FMNB, when panel data are available.

This study has some drawbacks regarding the empirical application and the specification of the econometric model. These drawbacks point towards some possibilities for future research. Despite the advantages of the RHIE data for empirical modelling of health care utilisation, the interest in the results obtained is somewhat limited, since the experiment was conducted between 1974 and 1982. Applications to more recent non-experimental data, such as the ECHP will be of interest. More careful analysis of partial effects of relevant variables is important in an application which aims at producing policy relevant results. Additionally, within the econometric framework, there is still scope for further work. On the one hand, all the models are based on the Negative Binomial distribution, which has been shown to be limited by Winkelmann

(2004). The proposed model could, therefore, be adapted in order to relax the distributional assumptions. On the other hand, the hurdle specification is questionable. The main goal of this paper is to account for the two-stage process of demand for health care within a latent class framework. However, the traditional hurdle model imposes the assumption that there is only one sickness spell throughout the observed period. In order to overcome this limitation, the finite mixture two-part model can consider, for each latent class, that the number of visits is determined by the multiple spell model of Santos Silva and Windmeijer (2001), instead of the standard single spell hurdle model.

References

Atella, V., Brindisi, F., Deb, P. and Rosati, F.C., 2004, Determinants of access to physician services in Italy: a latent class seemingly unrelated probit approach, forthcoming in Health Economics.

Cameron, A.C. and P.K. Trivedi, 1998, Regression analysis of count data, Cambridge University Press, Cambridge.

Celeux, G., Chaveau, D. and Diebolt, J., 1995, On stochastic versions of the EM algorithm, Working paper 2514, Institut National de Recherche en Informatique et en Automatique.

Deb, P., 2001, A discrete random effects probit model with application to the demand for preventive care, Health Economics, 10, 5.

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, 6.

Deb and Trivedi, 1997, Demand for medical care by the elderly: a finite mixture approach, Journal of Applied Econometrics, 12.

Deb, P. and Trivedi, C., 2002, The structure of demand for health care: latent class versus two-part models, Journal of Health Economics, 21.

Deb and Trivedi, 1997, Demand for medical care by the elderly: a finite mixture approach, Journal of Applied Econometrics, 12.

Dempster, A.P., Laird, N.M., and Rubin, P.B., 1977, Maximum-likelihood estimation from incomplete data via the EM algorithm, Journal of the Royal Statistical society B, 39.

Follmann, D.A., and Lambert, D., 1991, Identifiability of finite mixture of logistic regression models, Journal of Statistical Planning and Inference, 27.

Garret, E., 2000, Graphic diagnostic tools for standard latent class and latent class regression model assessment with an application in describing depression and validating diagnostic criteria for depression, Ph.D. dissertation, The John Hopkins University, Baltimore, Maryland.

Gerdtham, 1997, Equity in health care utilization: further tests based on hurdle models and swedish micro data, Health Economics, 6.

Gerdtham and Trivedi, 2000, Equity in swedish health care reconsidered: new results based on the finite mixture model, Working Paper Series in Economics and Finance, Stockholm School of Economics.

Grootendorst, 1995, A comparison of alternative models of prescription drug utilization, Journal of Health Economics, 4.

Gurmu, S., 1997, Semi-Parametric Estimation of Hurdle Regression Models with an Application to Medicaid Utilizations, Journal of Applied Econometrics, 12, 225-242.

Jimenez-Martin, S., J.M. Labeaga and M. Martinez-Granado, 2002, Latent class versus two-part models in the demand for physician services across the european union, Health Economics, 11.

McLachlan, G.J., and Peel, D., 2000, Finite Mixture Models, Wiley, New York.

McLachlan, G.J., and Krishnan, T., 1997, The EM algorithm and extensions, Wiley, New York.

Mullahy, J., 1986, Specification and testing in some modified count data models,

Journal of Econometrics, 33.

Pohlmeier and Urlich, 1995, An econometric model of the two-part decision making process in the demand for health care, The Journal of Human Resources 30.

Santos Silva and Windmeijer, 2001,Two-Part Multiple Spell Models for Health Care Demand, Journal of Econometrics, 104.

Teicher, H., 1960, On the mixture of distributions, Annals of Mathematical Statistics, 31.

Van Ourti, T., 2004, Measuring horizontal inequity in Belgian health care using a Gaussian random effects two part count data model, forthcoming in: Health Economics.

Vuong, Q.H., 1989, Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica, 57, 307-333.

Wang, P., Puterman, M.L., Cockburn, I. and Le, N.D., 1996, Mixed Poisson regression models with covariates dependent rates, Biometrics, 52.

Wang, P., 1994, Mixed regression models for discrete data, Ph.D. dissertation, University of British Columbia, Vancouver.

Winkelmann, R., 2004, Health Care Reform and the Number of Doctor Visits - An Econometric Analysis, forthcoming in: Journal of Applied Econometrics.

Table 1: Hurdle estimation results for number of outpatient visits

N = 20186 LogL = -45351.2	Part 1 - P[Y>0]		Part 2 - P[Y|Y>0]	
	Estimate	Std. Error	Estimate	Std. Error
Constant	-0.456	0.101	-0.027	0.243
LC	-0.217	0.014	-0.085	0.034
IDP	-0.566	0.028	-0.534	0.059
LPI	0.029	0.004	0.010	0.008
FMDE	0.041	0.008	-0.031	0.018
LINC	0.061	0.010	0.053	0.027
LFAM	-0.039	0.020	-0.222	0.039
AGE	0.036	0.010	0.040	0.019
FEMALE	0.432	0.025	0.390	0.054
CHILD	0.335	0.039	0.162	0.092
FCHILD	-0.442	0.039	-0.346	0.107
BLACK	-0.791	0.030	-0.835	0.117
EDUCDEC	0.030	0.004	0.022	0.007
PHYSLM	0.176	0.033	0.385	0.050
DISEA	0.019	0.002	0.023	0.002
HLTHG	-0.005	0.021	0.230	0.049
HLTHF	0.069	0.040	0.458	0.068
HLTHP	0.358	0.089	0.704	0.095
α			6.745	0.157

Table 2: FMNB results for number of outpatient visits

N = 20186 LogL = -45036.5	Low users		High users	
	Estimate	Std. Error	Estimate	Std. Error
Constant	-0.238	0.127	1.194	0.256
LC	-0.204	0.016	-0.072	0.033
IDP	-0.542	0.028	-0.407	0.068
LPI	0.021	0.004	0.019	0.011
FMDE	0.039	0.008	-0.028	0.018
LINC	0.061	0.014	0.063	0.027
LFAM	-0.056	0.021	-0.229	0.051
AGE	0.027	0.010	0.019	0.023
FEMALE	0.394	0.027	0.275	0.061
CHILD	0.332	0.041	-0.008	0.104
FCHILD	-0.360	0.043	-0.357	0.110
BLACK	-0.771	0.042	-0.680	0.092
EDUCDEC	0.025	0.004	0.021	0.009
PHYSLM	0.182	0.029	0.447	0.062
DISEA	0.018	0.001	0.028	0.003
HLTHG	0.023	0.022	0.089	0.055
HLTHF	0.182	0.040	0.176	0.088
HLTHP	0.494	0.066	0.341	0.134
α	1.707	0.079	10.110	0.524
π	0.793	0.017		

Table 3: Selection between Hurdle and FMNB

Model	LogL	AIC	BIC	K
Hurdle	-45351.2	90776	91069	37
FMNB	-45036.5	90151	90460	39

Table 4: FMNB-Pan results for number of outpatient visits

N = 5908 LogL = -43762.9	Low users		High users	
	Estimate	Std. Error	Estimate	Std. Error
Constant	-0.764	0.139	1.132	0.137
LC	-0.249	0.017	-0.120	0.020
IDP	-0.580	0.029	-0.322	0.045
LPI	0.027	0.005	0.020	0.007
FMDE	0.051	0.009	0.013	0.011
LINC	0.075	0.015	0.052	0.013
LFAM	-0.036	0.022	-0.067	0.030
AGE	0.027	0.011	0.038	0.014
FEMALE	0.454	0.027	0.261	0.036
CHILD	0.355	0.040	0.268	0.061
FCHILD	-0.464	0.041	-0.273	0.063
BLACK	-0.958	0.044	-0.483	0.046
EDUCDEC	0.035	0.004	0.019	0.005
PHYSLM	0.248	0.031	0.211	0.040
DISEA	0.021	0.001	0.020	0.002
HLTHG	0.014	0.022	0.027	0.034
HLTHF	0.072	0.045	0.090	0.053
HLTHP	0.443	0.079	0.185	0.085
α	1.190	0.045	5.991	0.180
π	0.751	0.010		

Table 5: Loglikelihood ratio tests of equality of parameters across classes (p-values)

	dof	FMNB-Pan
All slopes	17	0.000
Insurance	4	0.000
Demographics	8	0.000
Health status	5	0.075

Table 6: Expected probability of having one visit and expected number of visits - FMNB-Pan

	P[Y>0]	E[Y]
Sample	0.72	3.55
Estimates		
Overall	0.70	3.37
Low users	0.64	1.85
High users	0.91	8.81

Table 7: FMH-Pan results for number of outpatient visits

N = 5908 LogL = -43615.5	Low users				High users			
	Part 1 - P[Y>0]		Part 2 - P[Y|Y>0]		Part 1 - P[Y>0]		Part 2 - P[Y|Y>0]	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Constant	-0.999	0.170	-0.592	0.275	0.416	0.246	0.918	0.218
LC	-0.293	0.021	-0.161	0.035	-0.128	0.035	-0.135	0.030
IDP	-0.640	0.037	-0.498	0.049	-0.400	0.101	-0.385	0.062
LPI	0.034	0.006	0.021	0.007	0.026	0.014	0.023	0.009
FMDE	0.069	0.011	0.005	0.017	0.004	0.020	0.009	0.016
LINC	0.067	0.018	0.090	0.032	0.073	0.020	0.052	0.023
LFAM	-0.008	0.028	-0.095	0.035	0.004	0.056	-0.146	0.043
AGE	0.037	0.014	0.018	0.017	0.051	0.030	0.035	0.019
FEMALE	0.524	0.035	0.343	0.045	0.346	0.073	0.289	0.052
CHILD	0.431	0.052	0.154	0.070	0.488	0.128	0.210	0.086
FCHILD	-0.520	0.052	-0.319	0.076	-0.458	0.126	-0.219	0.090
BLACK	-1.064	0.050	-0.672	0.099	-0.436	0.075	-0.764	0.091
EDUCDEC	0.045	0.005	0.022	0.006	0.016	0.011	0.026	0.007
PHYSLM	0.228	0.044	0.319	0.046	-0.107	0.089	0.327	0.051
DISEA	0.020	0.002	0.021	0.002	0.010	0.005	0.026	0.003
HLTHG	0.008	0.028	0.062	0.037	0.011	0.068	0.086	0.049
HLTHF	-0.013	0.059	0.217	0.063	0.288	0.121	0.134	0.074
HLTHP	0.455	0.125	0.452	0.104	0.224	0.230	0.234	0.106
α			1.217	0.069			6.881	0.232
π	0.722	0.011						

Table 8: Loglikelihood ratio test statistics of significance of groups of parameters

		P[Y>0]		P[Y¦Y>0]	
	dof	Low users	High users	Low users	High users
Insurance	4	569	38	226	96
Demographics	8	672	85	215	209
Health status	5	189	15	269	208

Table 9: Loglikelihood ratio tests of equality of parameters across classes (p-values) - FMH-Pan

	dof	P[Y>0]	P[Y|Y>0]
All but C and alfa	17	0.000	0.170
Insurance	4	0.001	0.136
Demographics	8	0.000	0.625
Health status	5	0.000	0.193

Table 10: Loglikelihood ratio tests of equality of parameters of binary and positive part (p-values)

	dof	Low users	High users
Insurance	4	0.000	0.995
Demographics	8	0.000	0.003
Health status	5	0.013	0.000

Table 11: Expected probability of having one visit and expected number of visits - FMH-Pan

	P[Y>0]	E[Y|Y>0]	E[Y]
Sample	0.72	4.94	3.55
Estimates			
Overall	0.71	4.77	3.39
Low users	0.63	2.85	1.80
High users	0.95	8.58	8.16

Table 12: Effect of coinsurance rate on expected number of visits, by components - FMH-Pan

	P[Y>0|p]			E[Y|Y>0,p]			E[Y,p]		
	overall	Low users	High users	overall	Low users	High users	overall	Low users	High users
p=0	0.85	0.81	0.98	5.92	3.45	10.84	4.70	2.73	10.60
p=25	0.63	0.53	0.93	4.30	2.56	7.74	2.86	1.38	7.27
p=95	0.54	0.42	0.90	3.87	2.34	6.90	2.32	1.00	6.29
Relative changes									
0 to 25	-0.26	-0.34	-0.05	-0.27	-0.26	-0.29	-0.39	-0.49	-0.31
25 to 95	-0.15	-0.22	-0.03	-0.10	-0.09	-0.11	-0.19	-0.28	-0.14

Table 13: LogL and information criteria

	LogL	AIC	BIC	k
Hurdle	-45351.2	90776	91069	37
FMNB	-45036.5	90151	90460	39
FMNB-Pan	-43762.9	87604	87864	39
FMH-Pan	-43615.5	87381	87882	75

Table 14: Expected number of visits by coinsurance rate - FMNB-Pan and FMH-Pan

	Sample	FMNB-Pan	FMH-Pan
E[Y]	3.55	3.37	3.39
E[Y|p=0]	4.55	4.99	4.70
E[Y|p=25]	3.33	2.80	2.86
E[Y|p=95]	2.69	2.25	2.32
Relative changes			
0 to 25	-0.27	-0.44	-0.39
25 to 95	-0.19	-0.20	-0.19