Two-Part Multiple Spell Models for Health Care Demand

João M. C. Santos Silva
Frank Windmeijer

THE INSTITUTE FOR FISCAL STUDIES
WP 99/2

Two-Part Multiple Spell Models for Health Care Demand

João M. C. Santos Silva*

ISEG/Universidade Técnica de Lisboa, R. do Quelhas 6, 1200 Lisboa, jmcss@iseg.utl.pt

Frank Windmeijer

Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, f.windmeijer@ifs.org.uk

Abstract

The demand for certain types of health care services depends on decisions of both the individual and the health care provider. This paper studies the conditions under which it is possible to separately identify the parameters driving the two decision processes using only count data on the total demand. It is found that the frequently used hurdle models may not be adequate to describe this type of demand, especially when the assumption of a single illness spell per observation period is violated. A test for the single spell hypothesis is developed and alternative modelling strategies are suggested, including one that allows for correlated unobserved heterogeneity. The results of the paper are illustrated with an application.

JEL Classification: C12, C13, C25, I10

Key words: Count data; Generalised method of moments; Hurdle models; Non-parametric maximum likelihood estimator; Stopped-sum distributions.

*Corresponding author.

1. INTRODUCTION

It has become generally accepted in the literature on the demand for health care that the demand for certain types of these services depends on two different decision processes (Stoddart and Barer, 1981; Pohlmeier and Ulrich, 1995). In the first stage, the individual decides whether or not she needs medical attention. In the second stage, the health care providers together with the individual decide on the intensity of the treatment. Because this kind of demand is often measured as a count variable, for example the number of visits to a doctor in a given period, the more recent applied literature on this field has used hurdle models of the type proposed by Mullahy (1986) to account for the nature of the data, while trying to separately estimate the parameters of the two decision processes (Pohlmeier and Ulrich, 1995; and Gurmu, 1997).[1]

This paper studies the conditions under which it is possible to separately identify the parameters driving the two decision processes using only data on the total demand, and allowing for multiple illness spells. It is argued that the hurdle models used in the literature are in general not adequate to describe this two stage decision process, especially when the basic assumption of one illness spell is violated. A test for the validity of the single spell hypothesis is proposed which is based on robust estimation by the Generalised Method of Moments (GMM) (Hansen, 1982).

Alternative modelling strategies are proposed for the cases in which the hurdle model is rejected. Specifically, it is suggested that the parameters of the two processes can be jointly estimated by GMM, using only the specification of the conditional

[1] Two-part models have been extensivly used in other areas of applied health economics. The Rand Health Insurance Experiment team (Newhouse et al., 1993) used a two-part type model to analyse health care *expenditures*, with a binary model for positive versus zero expenditures and a linear model for the positive expenditures. They also analyse expenditures at the level of illness episodes. For a recent reappraisal of related two-stage models see Mullahy (1998). Deb and Trivedi (1999) consider as an alternative to the two-part model a latent class model, where the distinction is not so much between a user or a nonuser, but between the healthy and the ill.

mean of the total demand. Identification of the parameters in this double index model will depend on the functional forms specified for the conditional means of the two processes, and on the exclusion restrictions imposed. Alternatively, maximum likelihood procedures can be employed by using the framework of generalised, or stopped-sum, distributions where the distribution of the illness spells is generalised by the distribution of the referral process.

Although attention is focused on the demand for health care services, the methods developed here are equally adequate to study other cases in which the count variable is determined by two different processes. An example is the analysis of worker absenteeism or counts of workdays lost due to sickness. In this case the total number of lost working days depends both on the number of spells, and on the number of days in each spell. Indeed, Johansson and Brännäs (1998) use the framework of generalised distributions to motivate that their data on work absence is overdispersed. This type of framework is also used by Winkelmann (1998) to study under-reporting. In this case, the number of reported occurrences is obtained as a sum of binary variables that equal one if the event is reported and zero otherwise.

One of the maintained assumptions for the adequacy of the generalised distributions to describe the two-part decision process is conditional independence of the two underlying processes. To allow for correlation, we discuss and apply a non-parametric maximum likelihood estimator, where correlated unobserved individual heterogeneity is introduced into these processes. The bivariate distribution of the heterogeneity is approximated by a finite number of support points with corresponding probability masses (see for example Heckman and Singer, 1984, and Moon and Stotsky, 1993).

The paper is organised as follows. Section 2 sets out the problem and presents the framework of stopped sum distributions. Section 3 presents the hurdle model as a special case of this class of distributions. Section 4 discusses the estimation by maximum likelihood and GMM of the two-stage model, and presents appropriate specification tests. Section 5 considers the effects of departures from the assumptions used to derive the proposed estimators and introduces the non-parametric maximum likelihood estimator for generalised distributions with correlated unobserved heterogeneity. Section 6 contains an empirical illustration and section 7 concludes the paper.

2. TWO-PART DECISION PROCESS AND STOPPED-SUM DISTRIBUTIONS

Following Stoddart and Barer (1981), let an illness spell be defined as a set of medical services received continuously by an individual in response to a particular request. Therefore, the total amount of services that an individual receives in a given period depends on two different decision mechanisms. The process is started by an individual decision to seek medical care. In the second stage, the health care provider has an important role in deciding how many times, if any, the individual should return to complete the treatment.

Taking the number of visits to a doctor as an example, this definition of an illness spell implies that the total number of visits resulting from S spells of illness can be expressed as

$$V = \sum_{j=1}^{S} R_j^*, \tag{1}$$

where V is the total number of visits in S spells and R_j^* is the number of visits in the j-th spell.[2] Noting that in S spells of illness, S visits result from individual decisions and the rest are determined jointly by the individual and health care provider, (1) can be expressed as $V = S + \sum_{j=1}^{S} R_j$, where $R_j = R_j^* - 1$ is the number of referrals in the j-th spell.

Suppose that, given a set of covariates x, the researcher is interested in estimating the conditional expectation of S and R_j^*. If data on S and R_j^* are available, estimation of the parameters of interest is a simple regression exercise. However, in practice, analysis of data for health care demand is often complicated by the fact that the

[2]Notice that S can be zero, in which case V is also zero. However, the summation in (1) has to start at $j = 1$ since R_j^* is only defined for $j > 0$.

researcher only has information on the total number of visits in a given period, rather than on the number of spells and on the number of visits in each spell. Assuming that the conditional expectation of R_j^* does not depend on j and under the hypothesis that S and R_j^* are conditionally independent, (1) implies that V has a stopped-sum distribution in which the distribution of the number of spells is generalized with the distribution of the number of visits in each spell. Under these assumptions, a simple application of the law of iterated expectations shows that

$$\mathrm{E}(V|x,\beta,\gamma) = \mathrm{E}_S[S\,\mathrm{E}(R^*|S,x,\gamma)] = \mathrm{E}(S|x,\beta)\,\mathrm{E}(R^*|x,\gamma)\,,$$

where $\mathrm{E}(R^*|x,\gamma) = \mathrm{E}\Big(R_j^*|x,\gamma\Big)$, $j = 1,\ldots,S$.

Johnson, Kotz and Kemp (1992) describe many examples of stopped-sum distributions for this type of data. A distribution that is often applied to count data is the negative binomial. This distribution is popular as it allows for the often observed overdispersion in the data. The negative binomial distribution can be motivated as a stopped-sum distribution, and we will consider this particular distribution in more detail.

When S is Poisson(λ) and R_j^* is logarithmic(θ) distributed, V follows a negative binomial distribution.[3] The logarithmic distribution is defined by

$$\mathrm{P}[R^* = r] = \frac{a\theta^r}{r} \quad ; \quad r = 1,2,\ldots$$

where $a = -\left[\ln(1-\theta)\right]^{-1}$ and $0 < \theta < 1$, with moments

$$\begin{aligned} \mathrm{E}[R^*] &= \frac{a\theta}{1-\theta}, \\ \mathrm{Var}[R^*] &= \frac{a\theta\,(1-a\theta)}{(1-\theta)^2} = \mathrm{E}[R^*]\left\{(1-\theta)^{-1} - \mathrm{E}[R^*]\right\}. \end{aligned}$$

The negative binomial distribution of V is then given by

$$\mathrm{P}[V = v] = \binom{a\lambda + v - 1}{a\lambda - 1}\theta^v\,(1-\theta)^{a\lambda}$$

[3]Notice that the logarithmic distribution is an attractive model for the number of visits in each spell since it does not assume independence between the occurrence of events, and has been successfully used to model phenomena where dependence is likely to be present (see Johnson, Kotz and Kemp, 1992).

with moments

$$\begin{aligned}\mathrm{E}[V] &= a\lambda\frac{\theta}{1-\theta}, \\ \mathrm{Var}[V] &= a\lambda\frac{\theta}{(1-\theta)^2} = \frac{1}{1-\theta}\,\mathrm{E}[V]\,.\end{aligned}$$

The variance of V is proportional to $\mathrm{E}[V]$, as in the Negbin1 model described in Cameron and Trivedi (1986), where θ is assumed to be a constant. However, more interestingly, the parameter θ can be specified as a function of explanatory variables and parameters determining the referral process. A convenient specification is $\theta = \frac{\exp(x'\gamma)}{1+\exp(x'\gamma)}$, which, together with $\lambda = \exp(x'\beta)$, leads to

$$\mathrm{P}[V=v|x,\beta,\gamma] = \frac{\Gamma\left(v+\frac{\exp(x'\beta)}{\ln[1+\exp(x'\gamma)]}\right)\exp\left[-\exp(x'\beta)\right]}{\Gamma(v+1)\,\Gamma\left(\frac{\exp(x'\beta)}{\ln[1+\exp(x'\gamma)]}\right)\left(1+\exp(-x'\gamma)\right)^{v}}, \quad (2)$$

$$\mathrm{E}[V|x,\beta,\gamma] = \frac{\exp(x'\beta+x'\gamma)}{\ln\left[1+\exp(x'\gamma)\right]}, \quad (3)$$

$$\mathrm{Var}[V|x,\beta,\gamma] = \left[1+\exp(x'\gamma)\right]\mathrm{E}[V|x,\beta,\gamma]\,. \quad (4)$$

Again, the variance function is of the Negbin1 type, with the overdispersion parameter depending on the covariates. Moreover, the conditional mean function (3) has a different parametrization in this model. We call (2) the Negbin$_X$ model.

3. HURDLE MODELS

In order to account for the two different decision processes determining the demand for health care, hurdle models of the kind introduced by Mullahy (1986) were used by Pohlmeier and Ulrich (1995) and Gurmu (1997). A related model was also used by Grootendorst (1995) but with a different motivation. The hurdle model is characterized by the assumption that zeros and positive counts are generated by processes with different sets of parameters. That is, the hurdle model combines a binary model for the probability of observing a zero or a positive realization of the count variable with a model for the positive realizations.

This type of model can be successfully used to describe health care demand data, and if certain conditions are met, its parameters can have a structural interpretation. In fact, a hurdle model can be viewed as a stopped-sum distribution obtained by generalizing a Bernoulli distribution with a generalizer that has support on the positive integers.[4] Therefore, if the individuals in the sample have at most one illness spell, the parameters of the two parts of the hurdle model can be identified as the parameters of the two stages in the decision process underlying the demand for medical services. Of course, the assumption that individuals can have at most one spell of illness during the observation period can be very restrictive, and its validity will depend critically on the length of the observation period. However, in general there is no *a priori* information to justify the imposition of such restriction.[5]

Another important point is that the hurdle models are generally constructed assuming that the generalizing distribution has the form of a truncated distribution (see Mullahy, 1986). Besides being too restrictive, the use of a truncated distribution to model R_j^*, the number of visits *in each spell*, might result in some confusion because the distribution of R_j^* is not truncated by excluding from the sample individuals with zero visits. In fact, R_j^* is defined only for individuals having an illness spell, and these individuals are all retained in the sample when only positive counts are considered. Therefore, R_j^* should be modelled using a distribution with support on the positive integers, but not necessarily having the form of a truncated distribution.

It is important to notice that, even if V does not follow a generalized Bernoulli distribution, the standard hurdle model allows the identification of the first stage parameters. Let $d = \min\{V, 1\} = \min\{S, 1\}$ and note that $\mathrm{P}(S = 0|x, \beta) = \mathrm{P}(V = 0|x, \beta)$. Then, the individual contribution to the log-likelihood function of a hurdle model has

[4] The probability generating function for a hurdle model can be written as $G(z) = 1 - \mathrm{P}(V > 0|\beta, x) + \mathrm{P}(V > 0|\beta, x)\, G_1(z)$, where $G_1(z)$ is the probability generating function of the second stage, that is $\mathrm{P}(V = v|\gamma, x, V > 0) = \frac{1}{v!}\frac{\partial^v G_1(z)}{\partial z^v}\Big|_{z=0}$, and $G_1(0) = 0$.

[5] A referee pointed out to us that there are cases in which it is natural to assume that there is at most one spell of illness. An example would be the analysis of the demand for health care generated by self-immunising diseases like varicella.

the well known form

$$\ell = (1-d)\ln\left[\mathrm{P}(S=0|x,\beta)\right] + d\ln\left[\mathrm{P}(S>0|x,\beta)\right] + d\ln\left[\mathrm{P}(V=v|x,\gamma^*,V>0)\right], \quad (5)$$

where v is a positive integer and γ^* is a vector of parameters. This log-likelihood can be decomposed into two parametrically independent functions which can be estimated separately. Therefore, whether or not one assumes the hypothesis of a single spell, the hurdle model allows the identification of the parameters determining S. Moreover, if $S \in \{0,1\}$, $\mathrm{P}(V=v|\gamma^*,x,V>0) = \mathrm{P}(R_1^* = v|\gamma,x)$ and the model also permits the identification of the parameters governing the distribution of the number of visits in each spell. That is, only under the hypothesis of a single spell can the hurdle model be viewed as a stopped-sum distribution and only in this case $\gamma^* = \gamma$. Otherwise, $\mathrm{P}(V=v|x,\gamma^*,V>0)$ will depend on the parameters of both decision stages, and γ^* will be a function of β and γ. Of course, even in this case, the hurdle model may be adequate if the objective of the researcher is not the identification of β and γ, but just the description of the distribution of V.

4. ESTIMATION AND TESTING OF THE TWO-PART MODEL

4.1. Estimation of the Two-Part Model

At least some of the parameters of the conditional distribution of V can be estimated under very mild assumptions. Recalling that $d = \min\{V,1\}$,

$$\mathrm{E}(d|x) = 1 - \mathrm{P}(S=0|x,\beta), \quad (6)$$

and assuming that $\mathrm{P}(S=0|x,\beta)$ depends on x only through the index $x'\beta$, the first stage parameters β can be semi-parametrically estimated up to a normalizing constant using, for example, the estimator proposed by Klein and Spady (1993). Unfortunately, it is not possible to estimate the parameters of the distribution of R_j^* under such mild assumptions.

In order to proceed, assume that in the observation period only complete illness spells are observed, and that S and R_j^* are conditionally independent. If the researcher is prepared to specify the distributions of S and R_j^*, the fully efficient maximum likelihood estimator can be used to estimate both β and γ. As described in section 2, if it is assumed that S follows a Poisson distribution, that R_j^* has a logarithmic distribution, and that S and R_j^* are conditionally independent, then V follows a negative binomial distribution.

The problem with the maximum likelihood approach is that there is no pseudo-likelihood type result for the stopped-sum distributions. As with the hurdle specifications, misspecification of the stopped-sum distributions can lead to inconsistent parameter estimates. Naturally, the adequacy of the adopted distributional assumptions can be tested and, in case no evidence is found against them, this is probably the estimation method to be preferred.

If interest is focused only on the conditional expectation $\mathrm{E}(V|x)$, it is easy to see that the parameters from this regression can be consistently estimated under mild assumptions on the conditional distributions of S and R_j^*. As mentioned before, assuming that the conditional expectation of R_j^* does not depend on j and under the hypothesis that S and R_j^* are conditionally independent, $\mathrm{E}(V|x,\beta,\gamma) = \mathrm{E}_S[S\,\mathrm{E}(R^*|S,x,\gamma)] = \mathrm{E}(S|x,\beta)\,\mathrm{E}(R^*|x,\gamma)$, where $\mathrm{E}(R^*|x,\gamma) = \mathrm{E}\Big(R_j^*|x,\gamma\Big)$, $j = 1,\ldots,S$. Therefore, as long as the conditional expectations of S and R_j^* are both correctly specified, a moment condition of the form

$$\mathrm{E}[V - \mathrm{E}(S|x,\beta)\,\mathrm{E}(R^*|x,\gamma)] = 0 \tag{7}$$

can be used to estimate β and γ by GMM, without knowing the form of the conditional distribution of V. The major problem with this approach is that separate identification of the parameters of these two conditional expectations might be difficult. In fact, identification in a double index model like (7) will depend critically on the functional forms of $\mathrm{E}(S|x,\beta)$ and $\mathrm{E}(R^*|x,\gamma)$, and on the imposition of (exclusion) restrictions on β and γ. Therefore, except in special cases, this approach will often be inadequate in applied work.

Alternatively, identification can be achieved by using additional assumptions. For example, if in the observation period an individual has at most one illness spell, this information can be used to construct additional moment conditions that will help to separately identify the parameters of $\mathrm{E}(S|x,\beta)$ and $\mathrm{E}(R^*|x,\gamma)$. Additional moment conditions can also be obtained if a model for $\mathrm{P}(S=0|x,\beta)$ is specified assuming a given distribution for S. Naturally, in any of these cases, consistent estimation of β and γ will depend on the validity of the additional information being used.

4.2. Specification tests for the Two-Part Model

Specification tests for the two-part model can be constructed using the fact that some parameters can be identified in different ways. This section describes two types of specification tests based on this idea. First, a specification test for parametric models based on a stopped-sum distribution is presented. Essentially, this test compares the maintained specification with its hurdle counterpart. Second, a specification test is suggested for the cases in which the model assumes the single spell hypothesis.

4.2.1. A test for the parametric model

As pointed out above, the first stage parameters can always be estimated from (6), without using any information on the specification of the second stage. However, if there are individuals with more than one illness spell, the same parameters can be estimated jointly with those for the second stage, in the model for the positive counts based on an appropriate stopped-sum distribution. Therefore, a hurdle model constructed from the stopped-sum distribution provides two independent sets of estimates for the first stage parameters. If it can be accepted that the two sets of estimates obtained for the first stage parameters are the same, there is strong evidence in favour of the adopted specification. For example, in the case of the Negbin_X model the first stage of the hurdle specification is:

$$\mathrm{P}(V>0)=1-\exp\left(-\exp\left(x'\beta\right)\right). \tag{8}$$

The model for positive counts is

$$\mathrm{P}(V=v|x,V>0)=\frac{\Gamma\left(v+\frac{\exp(x'\beta^*)}{\ln[1+\exp(x'\gamma^*)]}\right)\exp\left[-\exp\left(x'\beta^*\right)\right]\left(1+\exp\left(-x'\gamma^*\right)\right)^{-v}}{\Gamma\left(v+1\right)\Gamma\left(\frac{\exp(x'\beta^*)}{\ln[1+\exp(x'\gamma^*)]}\right)\left(1-\exp\left(-\exp\left(x'\beta^*\right)\right)\right)},\tag{9}$$

and a test for the null hypothesis $H_0:\beta=\beta^*$ is a general test for correct specification.

4.2.2. A test for the single spell hypothesis

It was noted before that the hurdle models used in the studies of demand for health care assume that in the observation period an individual has at most one illness spell (see Pohlmeier and Ulrich, 1995). However, this is actually a testable assumption and, in case it holds, it might be used to construct additional moment conditions that will help to separately identify the parameters of $\mathrm{E}(S|x,\beta)$ and $\mathrm{E}(R^*|x,\gamma)$.

If each individual has at most one illness spell, S is observed since in this case $S=d=\min\{V,1\}$. Therefore, the following moment condition holds

$$\mathrm{E}[d-\mathrm{E}(S|x,\beta)]=0.\tag{10}$$

An additional moment condition can be obtained considering only the positive observations of V. The conditional expectation of the positive observations of V can be expressed as

$$\mathrm{E}(V|V>0,x,\beta,\gamma)=\mathrm{E}(S|S>0,x,\beta)\ \mathrm{E}(R^*|x,\gamma)\,.$$

In case the sample contains only individuals for which $S\leq 1$, S is identically 1 for the positive observations of V and so

$$\mathrm{E}(V|V>0,S\leq 1,x,\gamma)=\mathrm{E}(R^*|x,\gamma)\,.$$

That is, under these assumptions, it is possible to observe R^* for those individuals with $V>0$. Therefore, under the assumptions that there is at most one illness spell and that S and R^* are conditionally independent, it is possible to estimate $\mathrm{E}(R^*|x,\gamma)$ based on the moment condition

$$\mathrm{E}\{[V-\ \mathrm{E}(R^*|x,\gamma)]\,d\}=0,\tag{11}$$

where, as before, $d = \min\{V, 1\}$.

Using (10) and (11) it is possible to estimate β and γ by GMM, without the need to fully specify the distributions of S and R^*. A test for the overidentifying restrictions defined by (7) can then be used to test the hypothesis that the number of illness spells does not exceed one, conditional on the independence of S and R^*, and on the correct specification of their first moments. This test can be conveniently performed using the conditional moments tests described by Newey (1985) and Tauchen (1985).[6]

5. INCOMPLETE SPELLS AND DEPENDENCE

So far, it has been assumed that the data in the sample comes from S complete illness spells. However, applied studies commonly use data on the total number of visits in a given period, rather than on the total number of visits in S complete spells of illness. Therefore, as noted by Pohlmeier and Ulrich (1995), the sample will probably include individuals with incomplete illness spells because part of the recorded visits may result from an illness spell that started before the observation period. Moreover, some individuals may continue to visit the doctor after the observation period as a consequence of an illness spell started during the observation period. Therefore, in a sample, the distribution of R_j^* is a mixture of distributions with left, right and no truncation.

It is important to analyse the consequences of this on the models for the demand for health care. The first obvious consequence is that S has to be interpreted as the number of illness spells the individual suffers during the observation period. Therefore S is larger than or equal to both the number of complete spells and the spells started in that period, either of which would probably be more interesting to model.

On the other hand, the possibility of truncated spells implies that the number of visits in an illness spell registered during the observation period cannot be seen as 1 plus the number of referrals. Therefore, R_j^* should be viewed as the number of visits from the j-th spell that occurred during the observation period, which may be

[6] See Davidson and MacKinnon (1993, pp. 571-578) for a clear exposition of this type of tests.

smaller than the total number of visits in the j-th illness spell. Another complication that arises due to the existence of incomplete illness spells is that the conditional expectation of R_j^* may depend on j as the first and last illness spells are more likely to be truncated. Again, the hypothesis of a single illness spell is relevant here since if $S \leq 1$ what is being estimated is the conditional expectation of R_1^*, not of R_j^*. Of course, the distribution of R_1^* is still a mixture of distributions with different forms of truncation.

The other assumption that has been maintained throughout the paper is that R_j^* and S are conditionally independent. As noted in section 4, identification of the first stage parameters does not depend on this hypothesis and therefore they can be consistently estimated whether or not this assumption holds. On the other hand, the conditional independence between S and R_j^* is critical for the identification of the second stage parameters using the ML and GMM estimators studied in section 4. Whether the conditional independence assumption is valid or not is an empirical matter that should be checked in every application. Although they are not specifically directed towards this form of misspecification, the general specification tests proposed in section 4 can be used to assess the validity of this assumption in practice.

If the researcher is able to specify the distributions of S and R_j^*, conditional both on the regressors and on unobservable individual effects, then it is possible to introduce correlation in the underlying latent processes determining the two stages of the decision process by allowing the unobservables to be correlated.[7] A model of this sort has been developed by Winkelmann (1998).[8] In this case, given the regressors and unobservables, V has a stopped-sum distribution, but conditioning only on the observable individual characteristics the distribution will not be of the stopped-sum

[7]Unobserved heterogeneity is often an argument for specifying models that accomodate overdispersion. Note that the specification for the total number of visits as in (2) allows for individual over- (or under-) dispersion. Including unobserved heterogeneity in the illness spell process now allows for this process to be overdispersed. Note that the logaritmic distribution for the number of visits is overdispersed when $\theta > 1 - e^{-1}$ and underdipersed when $\theta < 1 - e^{-1}$.

[8]See also Yen and Jones (1996).

type. Naturally, one can make additional assumptions on the form of the distribution of the unobservables and estimate by maximum likelihood, integrating out the jointly distributed unobservables. Alternatively, one can approximate the bivariate distribution by a non-parametric finite mixing distribution, following Heckman and Singer (1984) and Moon and Stotsky (1993). In the next section we will present estimation and test results for the Negbin$_X$ model on the basis of this estimation procedure. For this model the unobserved individual heterogeneity for observations $i = 1, ..., N$ can be introduced in the Poisson spell mean parameter λ_i and visits per spell parameter θ_i as

$$
\begin{aligned}
\lambda_{iu} &= \exp\left(x_i'\beta + u_i\right), \\
\theta_{iw} &= \frac{\exp\left(x_i'\gamma + w_i\right)}{1 + \exp\left(x_i'\gamma + w_i\right)},
\end{aligned}
$$

where the (u_i, w_i) are iid bivariate random variables. The unconditional maximum likelihood estimator is obtained by summing over a finite number of support points u_j and w_q, $j = 1, .., J$, $q = 1, ..., Q$, with corresponding masses $\pi_{u_j w_q}$. The individual likelihood contribution is then given by

$$
\begin{aligned}
L_i\left(\beta, \gamma, u_j, w_q, \pi_{u_j w_q}\right) &= \sum_{j=1}^{J}\sum_{q=1}^{Q} L_{ijq}\left(\beta, \gamma, u_j, w_q\right)\pi_{u_j w_q} \\
L_{ijq}\left(\beta, \gamma, u_j, w_q\right) &= \frac{\Gamma\left(v_i + \frac{\exp\left(x_i'\beta + u_j\right)}{\ln\left[1+\exp\left(x_i'\gamma + w_q\right)\right]}\right)\exp\left[-\exp\left(x_i'\beta + u_j\right)\right]}{\Gamma\left(v_i + 1\right)\Gamma\left(\frac{\exp\left(x_i'\beta + u_j\right)}{\ln\left[1+\exp\left(x_i'\gamma + w_q\right)\right]}\right)\left(1 + \exp\left(-x_i'\gamma - w_q\right)\right)^{v_i}}
\end{aligned}
$$

and the non-parametric maximum likelihood estimator (NPMLE) can be obtained by employing an EM algorithm (see for example Aitkin, 1999). The estimated support points and probability masses can be used to assess whether the two processes are correlated. A test for independence is a likelihood ratio test comparing the likelihood of the unrestricted model with the likelihood of a model imposing independence via the restrictions $\pi_{u_j w_q} = \pi_{u_j}\pi_{w_q}$, $j = 1, .., J$, $q = 1, ..., Q$, where π_{u_j} and π_{w_q} are the marginal probabilities of u_j and w_q respectively.

6. AN EMPIRICAL ILLUSTRATION

In this section, the data studied in the paper by Pohlmeier and Ulrich (1995) are used to illustrate the application of the proposed methodology. These data consists of 5096 observations for employed individuals, and it is taken from the 1985 wave of the German Socioeconomic Panel (SOEP). Pohlmeier and Ulrich (1995) studied the demand for health care services as measured by the number of visits to a general practitioner in the last quarter, and by the number of visits to a specialist (except gynaecology or paediatrics) in the same period. The variables used in the analysis are described in Table A.1 in the Appendix. For more detailed information on the sample and on the variables used see Pohlmeier and Ulrich (1995).

The models favoured by Pohlmeier and Ulrich (1995) are hurdle models based on the Negbin1 distribution.[9] Specifically, the second stage is modelled using a Negbin1 truncated at zero and the binary model for the occurrence of an illness is based on a censored at 1 Negbin1. Because the Negbin1 is a Poisson stopped-sum model, ignoring the information from the counts greater than 1 destroys all the information about the generalizing distribution, allowing only the identification of the parameters from the parent distribution. Therefore, the binary model for S specified by Pohlmeier and Ulrich (1995) cannot be distinguished from a model based on the Poisson distribution.

6.1. The single spell assumption

As argued before, hurdle models are based on the testable assumption that individuals have at most one illness spell during the observation period. Pohlmeier and Ulrich (1995) notice that the hurdle model is inadequate in the presence of multiple illness spells. However, they argue that in the present example this is unlikely to be problematic as more than 75% of the individuals in the sample have at most one visit

[9] Besides presenting the general hurdle models for the number of visits to general practitioners and to specialists, Pohlmeier and Ulrich (1995) also present restricted versions where some covariates are excluded from the second stage model.

to the doctor. In order to check if this assumption holds in this particular example, the specification test proposed in section 4 was performed.

The hurdle model preferred by Pohlmeier and Ulrich (1995) specifies

$$\mathrm{E}(S|x,\beta) = 1 - \exp\left(-\exp(x'\beta)\right),$$

$$\mathrm{E}(R^*|x,\gamma) = \frac{\exp(x'\gamma)}{1 - \exp\left(-\exp(x'\gamma)\right)}.$$

Therefore, assuming that the single spell hypothesis holds, the parameters of the first and second stages can be consistently estimated by GMM solving the sample analogs of the following moment conditions[10]

$$\mathrm{E}\{[d - 1 + \exp\left(-\exp(x'\beta)\right)]\, x\} = 0 \qquad (12)$$

$$\mathrm{E}\left\{\left[V - \frac{\exp(x'\gamma)}{1 - \exp\left(-\exp(x'\gamma)\right)}\right] xd\right\} = 0, \qquad (13)$$

where $d = \min\{V, 1\}$. Given this specification of the hurdle model, the validity of the single spell hypothesis can be tested by checking that the sample analogs of

$$\mathrm{E}\left\{\left[V - \exp(x'\gamma)\frac{1 - \exp\left(-\exp(x'\beta)\right)}{1 - \exp\left(-\exp(x'\gamma)\right)}\right] x\right\} = 0$$

hold when evaluated at the estimates obtained from (12) and (13). This test was performed using a conditional moments test as described in Newey (1985).[11] As the results in Table 1 show, the single spell hypothesis cannot be rejected at the 5% level in the case of visits to general practitioners, but is clearly rejected in the model for visits to specialists.

As explained in section 4, this test for overidentifying restrictions can be interpreted as a test of the hypothesis that the number of illness spells does not exceed one, conditional on the correct specification of the first moments of S and R_j^* and on the

[10] Note that these moment conditions are independent of the overdispersion parameter. Pohlmeier and Ulrich (1995) present an estimate for the overdispersion parameter in the first stage, but that is an unfortunate typographical error.

[11] All computations in this section were performed using TSP 4.3, Hall (1996), apart from the non-parametric maximum likelihood procedure, which was programmed in Gauss (1994).

Table 1:
Results of the test for the single spell hypothesis

	Test statistic	p-value*
General practitioners	33.49	0.055
Specialists	59.18	0.000

* Null distribution is χ^2_{22}.

conditional independence of these variates. Therefore, the rejection of the null in the case of the visits to specialists clearly indicates some sort of misspecification, although not necessarily the existence of multiple illness spells.[12]

6.2. The Negbin$_X$ model

It was pointed out before that the Negbin1 is based on a stopped-sum distribution and therefore can be used to model the two-part decision process, without the need to impose the single spell restriction. Therefore, it is interesting to see if this specification can be used to obtain a better model for the number of visits to specialists.

Maximum Likelihood estimation results for the visits to specialists model using the Negbin$_X$ specification described by (2) are reported in Table 2. Before analysing in detail the results obtained, it is necessary to check if this specification is adequate. Since the main question about this particular specification of the Negbin concerns its ability to adequately model the high number of zero counts, the model was tested against its hurdle counterpart, as suggested in section 4. In this case, the alternative is a hurdle model in which the first stage is a censored Poisson, (8), and the second stage is obtained by truncating at zero the Negbin$_X$ specification, (9). The score test statistic for

[12] In Germany, an individual can choose to visit a specialist, without a referral from a general practitioner. Means (standard deviations) of general practitioner and specialist visits are 1.33 (3.24) and 1.24 (3.37) respectively.

Table 2: Negbin_X MLE results for visits to specialists

$N = 5096$	Spells		Referrals	
$Logl = -6278.43$	$\hat{\beta}$	Std. err.	$\hat{\gamma}$	Std. err.
Constant	−0.9575	0.3772	1.2565	0.8172
FEMALE	0.7269	0.0504	−0.1515	0.1073
SINGLE	−0.2413	0.0823	−0.1750	0.1864
AGE	−0.3698	0.1739	0.1990	0.3595
AGE^2	0.3962	0.2084	−0.2593	0.4256
INCOME	0.1451	0.0750	−0.1656	0.1862
CHRONIC COMPLAINTS	0.7594	0.0561	0.2832	0.1093
PRIVATE INSURANCE	0.1887	0.0902	−0.4285	0.1821
EDUCATION	0.0223	0.0082	−0.0455	0.0175
HEAVY LABOUR	−0.0095	0.0674	−0.1799	0.1376
STRESS	0.0516	0.0573	0.0842	0.1170
VARIETY ON JOB	0.0871	0.0537	0.0851	0.1075
SELF DETERMINED	0.0668	0.0544	−0.0652	0.1068
CONTROL	0.0142	0.0679	0.2230	0.1401
POP-0/5	−0.5569	0.0926	−0.0650	0.1915
POP-5/20	−0.3777	0.0673	0.0598	0.1397
POP-20/100	−0.2075	0.0636	−0.0516	0.1301
PHYSICIANS DENSITY	0.8951	0.4649	−0.2517	0.9576
UNEMPLOYMENT	−0.0079	0.0204	−0.0503	0.0418
HOSPITALIZED	0.1640	0.0907	0.2610	0.1668
SICK LEAVE	0.3524	0.0674	0.5859	0.1239
DISABILITY	0.2296	0.0911	0.0628	0.1778
Specification test*	21.00		p-value = 0.52	

* This is a score test for $H_0 : \beta = \beta^*$ in the hurdle model as specified in (8) and (9).

the hypothesis that the first stage parameters are the same in both specifications has a value of 21.00, to which corresponds a p-value of 0.52. Therefore, the null is not rejected.

The results in Table 2 can be summarized as follows. Females have more illness spells than males but gender does not have a statistically significant impact on the number of visits per spell. Being single decreases the expected number of illness spells, but has no effect on the number of referrals. Age affects the occurrence of spells in a quadratic way but it is not possible to identify clearly its effect on the second stage. Income, physicians density, unemployment in the previous year, being more than seven days hospitalized in the previous year, and job characteristics do not have significant individual impacts on either stage. On the other hand, suffering from chronic complaints and having had more than 14 days of sick leave in the previous year significantly increases both the expected number of illness spells and the number of visits in each spell, while having a degree of disability greater than 20% increases the number of illness spells, but not the number of visits in each spell. Both having private insurance and having more years of schooling significantly increase the expected number of first contacts, having the opposite effect on the number of visits in each spell. Finally, although these variables are not individually significant for the second stage, the model suggests that the expected number of first contacts increases with size of the place of residence.

6.3. The NEGBIN$_X$ model with unobserved heterogeneity

Tables 3a and 3b present the results of the non-parametric maximum likelihood estimator, using the bivariate mass-point distribution to approximate the joint distribution of individual unobserved heterogeneity in the two processes, as discussed in section 5. The number of support points were found by increasing J and Q one at a time until there were no further changes in the log likelihood function. Two points of support for both processes were identified. Table 3a gives the parameter estimates for β and γ, whereas Table 3b displays the results for the estimated support points and associated probability masses.

The estimation results for the points of support and associated probabilities clearly indicate that the two processes u_i and w_i are uncorrelated. The estimated bivariate probabilities are almost all equal to the product of the marginal probabilities. A model that imposes independence identifies the same number of support points, and the likelihood ratio test statistic, comparing the log likelihood of the unrestricted model with that of a model that imposes independence restrictions, is equal to 0.015.[13] Therefore, conditional independence is clearly not rejected. Although this is the case, some estimated parameter values do change due to the inclusion of unobserved heterogeneity. The estimates of the parameters of the illness spells, β, are generally very similar to the ones obtained without allowing for unobserved heterogeneity. The main differences are the coefficients for female, income and chronic complaints that are all larger (positive) than before. For the referral process parameters, γ, the female coefficient becomes more negative and the effect of chronic complaints is now insignificant. The effect of physician density is very poorly determined. Given the nature of the phenomenon being modelled and the covariates available to explain it, it is not surprising to find that many coefficients, especially for the referral process, are not significantly different from zero. The imposition of exclusion restrictions such that only the female, private insurance, education and sick leave variables are included in the model for referrals are not rejected by the data. Independence of the unobserved heterogeneity processes is again not rejected in the restricted model.

Overall the estimates obtained with the Negbin$_X$ model, whether allowing for correlated unobserved heterogeneity or not, are in line with what was expected and are not much different from those produced by the Negbin1 hurdle model, as given by Pohlmeier and Ulrich (1995). However because the structure of the models is very

[13] Results for the estimation imposing independence are not presented as they are simillar to those obtained with the unrestricted Negbin$_X$ NPMLE.

Table 3a: Negbin_X NPMLE results for visits to specialists, controlling for unobserved heterogeneity

$N = 5096$	Spells		Referrals	
$Logl = -6258.36$	$\hat{\beta}$	Std. err.	$\hat{\gamma}$	Std. err.
Female	0.8835	0.0654	−0.5316	0.1671
Single	−0.2673	0.0961	−0.0900	0.2547
Age	−0.3578	0.2019	0.2202	0.5060
Age^2	0.3639	0.2424	−0.2457	0.6000
Income	0.2510	0.0914	−0.2630	0.3189
Chronic complaints	0.9537	0.0788	−0.0655	0.1772
Private insurance	0.1725	0.1050	−0.4184	0.2694
Education	0.0250	0.0096	−0.0705	0.0251
Heavy labour	0.0080	0.0818	−0.1432	0.1971
Stress	0.0544	0.0686	0.0995	0.1671
Variety on job	0.1084	0.0641	0.0575	0.1504
Self determined	0.0839	0.0647	−0.1247	0.1505
Control	0.0207	0.0810	0.2211	0.2002
Pop-0/5	−0.6828	0.1079	0.0578	0.2580
Pop-5/20	−0.4314	0.0802	0.1210	0.1964
Pop-20/100	−0.2392	0.0762	0.0655	0.1891
Physicians density	1.2528	0.5803	−1.2626	1.4553
Unemployment	−0.0070	0.0286	−0.0496	0.0629
Hospitalized	0.1982	0.1167	0.2060	0.2433
Sick leave	0.4685	0.0870	0.5308	0.1778
Disability	0.2951	0.1191	−0.1755	0.2874

Table 3b. NPMLE results for points of support and probabilities

	$\widehat{u}_1$: −2.0674 (0.5282)	$\widehat{u}_2$: −0.3677 (0.4770)	
$\widehat{w}_1$: 0.2415 (1.1940)	$\widehat{\pi}_{u_1w_1}$: 0.3807 (0.1688)	$\widehat{\pi}_{u_2w_1}$: 0.2469 (0.0559)	$\widehat{\pi}_{w_1}$: 0.6278 (0.1912)
$\widehat{w}_2$: 2.1929 (1.1520)	$\widehat{\pi}_{u_1w_2}$: 0.2142 (0.2226)	$\widehat{\pi}_{u_2w_2}$: 0.1582 (0.0653)	$\widehat{\pi}_{w_2}$: 0.3722 (0.1912)
	$\widehat{\pi}_{u_1}$: 0.5952 (0.1023)	$\widehat{\pi}_{u_2}$: 0.4048 (0.1023)	1
L. R. Test for Independence: 0.0150 (dof = 1, p-value = 0.9025)			

Standard errors in brackets.

different, the Negbin1 hurdle and Negbin$_X$ models have different performances in terms of goodness of fit. In fact, using the Schwarz (1978) criterion,[14] the Negbin1 hurdle model is outperformed by both Negbin$_X$ models. The values for the Schwarz criterion for the three models are −6466, −6467 and −6491, for the Negbin$_X$, Negbin$_X$ with unobserved heterogeneity and Negbin1 hurdle, respectively. Furthermore, imposing independence in the heterogeneous Negbin$_X$ the value of this criterion rises to −6463. Therefore, the heterogeneous Negbin$_X$ imposing independence is the preferred model according to this criterion.

The difference in performance of the Negbin1 hurdle and Negbin$_X$ is further illustrated in Table 4, which displays the true and predicted frequencies for the three models. Starting with the total number of visits, the Negbin1 hurdle is again outperformed by the Negbin$_X$ models, especially by the Negbin$_X$ with unobserved heterogeneity. However, the main difference between these models is in terms of the predicted frequencies of the number of spells and visits in each spell. Because it imposes that each individual has at most one illness spell during the observation period, the hurdle model attributes relatively low probabilities to spells with few visits.

[14] The Schwarz (1978) criterion is calculated as $Logl - 0.5\ln(N)k$, where k is the total number of parameters estimated.

Table 4. True and predicted frequencies[15]

		0	1	2	3	>3
Data	V	0.678	0.120	0.061	0.044	0.097
Negbin1 hurdle	$\hat{V}$	0.679	0.138	0.055	0.023	0.105
	$\hat{S}$	0.679	0.321	0	0	0
	$\widehat{R^*}$	—	0.479	0.178	0.060	0.283
Negbin$_X$	$\hat{V}$	0.681	0.115	0.060	0.037	0.107
	$\hat{S}$	0.681	0.240	0.060	0.014	0.005
	$\widehat{R^*}$	—	0.485	0.194	0.104	0.217
Heterogeneous Negbin$_X$	$\hat{V}$	0.679	0.122	0.062	0.037	0.099
	$\hat{S}$	0.679	0.202	0.067	0.027	0.026
	$\widehat{R^*}$	—	0.602	0.179	0.078	0.141

On the other hand, the Negbin$_X$ models predict that 8 to 10% of the individuals have more than 1 illness spell during the observation period, and attribute higher probabilities to spells with few visits.

7. CONCLUSIONS

Total demand for health care depends both on the individual needs and on the decisions of the health care providers. Recently these two stages of the decision process have explicitly been incorporated in empirical models for the demand of health care services. Ideally, these studies should use data on the number of illness spells and on the number of visits in each spell. However, very often the demand for

[15] Predicted frequencies for the Negbin1 hurdle and Negbin$_X$ models are obtained by summing the probability of each occurrence for all individuals in the sample, using the appropriate model. Predicted frequencies for the heterogeneous Negbin$_X$ are obtained using the method described in Aitkin (1996).

health care services is simply measured by the number of times an individual received medical care during a given period. Under certain conditions, this two-part decision process implies that total demand follows a stopped-sum distribution.

Hurdle models have frequently been used to describe the demand for this type of services. Although hurdle models can be interpreted as models based on stopped-sum distributions, they are far too restrictive to be generally adequate. In particular, these models assume that the individuals have at most one illness spell during the observation period. Here, it has been shown that this assumption can easily be tested and that, in case it is accepted, it can be used to help identifying the parameters of interest under mild assumptions on the distribution of the number of spells and referrals.

In case the single spell hypothesis cannot be accepted, hurdle models only permit the identification of the first stage parameters. However, it may still be possible to separately identify the two sets of parameters using alternative assumptions on the distribution of the data. In particular, if the interest is focused on the conditional expectations of both the number of illness spells and the demand in each spell, a generalized method of moments estimator can be used to estimate the conditional mean parameters using only data on the total demand for health care services, as measured by the number of times these services are used.

The merit of this approach is that it requires only a minimal set of assumptions and is therefore robust. On the other hand, because only the conditional expectation is specified, it is not possible to make inference concerning the probabilities of interesting events. Furthermore, the gain in robustness has a price in terms of efficiency and, unless the data set is rich enough, this approach is bound to be marred by identification problems.

In view of this, it is convenient to use some further information in the estimation of the parameters of interest. One possibility is to specify the distribution of the number of spells and to use this information to obtain a set of overidentifying restrictions that can be used to help in the estimation of the parameters of interest.

Naturally, if the researcher is willing to fully specify the stopped-sum distribution for the total demand in the observation period, maximum likelihood can be used to estimate the two sets of parameters of interest. In the empirical illustration presented in section 6, a particular Negbin specification was used to model the data on the number of visits to specialists studied by Pohlmeier and Ulrich (1995). Furthermore, a non-parametric maximum likelihood estimator that allows for correlated unobserved heterogeneity was also described and implemented.

ACKNOWLEDGMENTS

We are indebted to four anonymous referees, Murray Aitkin, Kurt Brännäs, Paul Grootendorst, Andrew Jones, Rainer Winkelmann and participants at the Sixth European Workshop on Econometrics and Health Economics for helpful comments. The data used in the application in section 6 was kindly made available by Winfried Pohlmeier and are from the public use version of the German Socio-Economic Panel Study. We thank the Deutsches Institut für Wirtschaftsforschung for authorizing its use. This work was partially carried out while both authors were at the Department of Economics of the University College London with the financial support of the Human Capital and Mobility program of the European Commission, which is gratefully acknowledged. This study is further part of the program of research of the ESRC Centre for the Microeconomic Analysis of Fiscal Policy at IFS. Frank Windmeijer gratefully acknowledges the financial support of the ESRC. João Santos Silva is thankful to the Fundação para a Ciência e a Tecnologia, programme Praxis XXI, for partially sponsoring his research. The usual disclaimer applies.

REFERENCES

Aitkin M. (1996) 'Empirical Bayes Shrinkage using Posterior Random Effect Means from Nonparametric Maximum Likelihood Estimation in General Random Effects Models', in Forcina, A., G.M. Marchetti, R. Hatzinger and G. Galmacci (eds.), *Proceedings of the 11th International Workshop on Statistical Modelling*, Graphos, Citta di Castello.

Aitkin, M. (1999) 'A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models', *Biometrics*, 5, 117-128.

Cameron, A. C. and P. K. Trivedi (1986) 'Econometric Models based on Count Data: Comparisons and Applications of some Estimators and Tests', *Journal of Applied Econometrics,* 1, 29-53.

Davidson, R. and J. G. MacKinnon (1993) *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

Deb, P. and P. K. Trivedi (1999) 'Which Tale of Two Parts Describes Demand for Health Care - Users and Nonusers or the Healthy and Ill?, paper presented at the 8th European Workshop on Econometrics and Health Economics, Catania.

Grootendorst, P. V. (1995) 'A Comparison of Alternative Models of Prescription Drug Utilization', *Journal of Health Economics*, 4, 183-98.

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

Hall, B.H. (1996) *Time Series Processor Version 4.3 User's Guide.* TSP International, Stanford.

Hansen, L. P. (1982) 'Large Sample Properties of Generalized Method of Moments Estimators', *Econometrica*, 50, 1029-1054.

Heckman, J. J. and B. Singer (1984) 'A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data, *Econometrica*, 52, 271-320.

Gauss (1994), Aptech Systems, Maple Valey.

Johansson, P. and K. Brännäs (1998) 'A Household Model for Work Absence', *Applied Economics*, 30, 1493-1503.

Johnson, N. L., S. Kotz and A.W. Kemp (1992) *Univariate Discrete Distributions*, 2nd Ed., John Wiley & Sons, Inc., New York.

Klein, R. W. and R. H. Spady (1993) 'An Efficient Semiparametric Estimator for Binary Response Models', *Econometrica*, 61, 387-421.

Moon, C.-G. and J. G. Stotsky (1993) 'The Effect of Rent Control on Housing Quality Change: A Longitudinal Analysis', *Journal of Political Economy*, 101, 1114-1148.

Mullahy, J. (1986) 'Specification and Testing in Some Modified Count Data Models', *Journal of Econometrics*, 33, 341-365.

Mullahy, J. (1998) 'Much Ado About Two: Reconsidering Retransformation and the Two-Part Model in Health Econometrics', *Journal of Health Economics*, 17, 247-281.

Newey, W. (1985) 'Maximum Likelihood Specification Testing and Conditional Moments Tests', *Econometrica*, 53, 1047-1070.

Newhouse, J. P. and the Insurance Experiment Group (1993), *Free for All? Lessons from the RAND Health Insurance Experiment*, Harvard University Press, Cambridge, Massachusetts.

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

Schwarz, G. (1978) 'Estimating the Dimension of a Model', *Annals of Statistics*, 6, 461-464.

Stoddart, G. L. and M. L. Barer (1981) 'Analyses of the Demand and Utilization Through Episodes of Medical Care', *in* van der Gaag, J. and M. Perlman, (eds.), *Health, Economics and Health Economics*, Amsterdam, North-Holland.

Tauchen, G. (1985) 'Diagnostic Testing and Evaluation of Maximum Likelihood Models', *Journal of Econometrics*, 30, 415-443.

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

Winkelmann, R. (1998) 'Count Data Models with Selectivity', *Econometric Reviews*, 17, 339-359.

Yen, S.T. and A.M. Jones (1996) 'Individual Cigarette Consumption and Addiction: A Flexible Limited Dependent Variable Approach', *Health Economics*, 5, 105-117.

APPENDIX

Table A.1: Description of Variables

FEMALE	1 if female
SINGLE	1 if single
AGE	age in years
INCOME	net monthly household income
CHRONIC COMPLAINTS	1 if has chronic complaints for at least one year
PRIVATE INSURANCE	1 if had private medical insurance in the previous year
EDUCATION	number of years in education after the age of sixteen
HEAVY LABOUR	1 if has a job in which physically heavy labour is required
STRESS	1 if has a job with high level of stress
VARIETY ON JOB	1 if job offers a lot of variety
SELF DETERMINED	1 if has a job where the individual can plan and carry out job tasks
CONTROL	1 if has a job where work performance is strictly controlled
POP-0/5	1 if place of residence has less than 5000 inhabitants
POP-5/20	1 if place of residence has between 5000 and 20.000 inhabitants
POP-20/100	1 if place of residence has between 20.000 and 100.000 residents
PHYSICIANS DENSITY	number of physicians per 100.000 inhabitants in the place of residence
UNEMPLOYMENT	number of months of unemployment in the previous year
HOSPITALIZED	1 if was more than seven days hospitalized in the previous year
SICK LEAVE	1 if missed more than fourteen work days due to illness in the previous year
DISABILITY	1 if the degree of disability is greater than 20%