Estimating Health Care Demand with Ordinal Endogenous Variable

Shamzaeffa Samsudin

Department of Economics, University of East Anglia, Norwich NR4 7TJ
December 2007

Abstract

Understanding factors that affect health care demand has been subject of social and policy interest for many years. It requires continuous assessment as those factors are constantly changing over time. This paper discusses several count data techniques in estimating demand for health care and by using data from British General Household Survey 2004 (GHS 2004), self-perceived general health state is found to be one of the significant determinants of health care demand. Nonetheless, care needs to be exercised in using this explanatory variable, since there is a strong possibility that it is simultaneously determined by other variables appearing within the model, and therefore endogenous. While most previous studies that have addressed this issue use a binary health index, health state in this study has three possible ordered outcomes: not good; fairly good; good. Hence, the model has been extended to take into account the endogeneity of this ordinal variable

1 Background

Demand for health care is a derived demand which mainly depends on one's health status. Beside unobserved health status, demand for health care depends on other variables namely wage rate, price of medical service, age, level of education and other environmental variables. Wagstaff (1986) estimates both reduced and structural demand function for health care by considering health capital as a latent variable in the latter. The structural demand function for health care, M_i, takes the form

$$\ln M_{it} = \beta_0 + \ln H^*_{it} + \beta_1 \ln w_{it} - \beta_2 \ln P^m_{it} + \beta_3 A_{it} + \beta_4 X_{it} + \beta_5 E_{it} + u_{it} \quad (1.1)$$

where H^*_i is the latent variable of health state, w_i, P^m_i, A_i, X_i, E_i, represent wage rate, price of medical care, age, environmental variables and education variables respectively. This form of health care demand treats the individual as the main decision maker in determining the amount of health care used. This approach is later being used as a foundation and extension by other studies (Mocan, Tekin, & Zax, 2004; Pohlmeier & Ulrich, 1995). With a different approach, Cameron, Trivedi, Frank & Pigott (1988) have developed a microeconometric model to understand not only factors that affect demand of healthcare but also demand of insurance in Australia under uncertainty[1]. By using 1977-1978 Australian Health Survey (AHS) of 40,650 individuals in the analysis, health status is seen to be

[1] The theoretical model of their work is based on two-period utility function.

statistically influenced by the utilization of health care services while income plays an important role in determining the insurance choice. Their main findings suggest that income is important in determining insurance choice while health status has large impact on health care utilization rather than insurance choice. In a different setting of a health system, Windmeijer and Santos Silva (1997) have estimated the demand for health care, measured by the number of the visits to the General Practitioner by using data from British Health and Lifestyle Survey 1991-1992 (HALS2). Unlike Cameron *et al*. (1988), there is no insurance variable in the model since every individual in the UK is entitled to free visits to the GP under the National Health Service (NHS). Many studies on health care demand suggest that variables representing health condition are the key factors determining the demand for health care (Deb & Trivedi, 2002; Gurmu, 1997; Pohlmeier & Ulrich, 1995; Sarma & Simpson, 2006).

One of the challenges one may encounter in modelling health care demand is endogeneity of the regressors. For instance, regressors in health care demand equation may be simultaneously determined by another set of variables. If these variables are correlated with the error term, least square estimation leads to inconsistent estimated coefficients. Alternative estimation techniques can be used to overcome the problem such as Generalised Method of Moment (GMM) or two-stage specification. Early studies that have discussed the treatment of endogenous variables include Amemiya (1974) who estimated a simultaneous equation model that is nonlinear in both variables and parameter and Kalejian (1971) who estimate equation that are linear in parameter but nonlinear in the endogenous variable by using two-stage least square. Windmeijer and Santos Silva (1997) estimate the demand for doctor visits assuming that self-reported health index is a potential endogenous variable in the estimated count model. Poisson pseudo-likelihood (PL) and GMM are utilised in a model with binary endogenous variable, while PL two- stage is used in a model with unobserved endogenous variable.

With a different motivation, which is to understand cigarette smoking behaviour, Mullahy (1997), on the other hand, suggests the usage of a nonlinear instrumental variable in estimating models with endogeneity problem. While previous studies like Windmeijer *et al*. (1997), Cameron *et al*. (1988) and Terza (1998) study models with binary endogenous variables, this paper tries to generalise those models by incorporating ordinal endogenous variables in order to estimate the determinants of four types of health care demand namely doctor consultations (*consultation*), practise nurse visits (*nurse*), outpatient visits (*outpatient*) and being inpatient (*inpatient*). This paper consists of five main sections. After an

introduction in Section 1, Section 2 discusses some statistical models in modelling health care demand. Section 3 describes the methods and the data used in the empirical analysis while Section 4 and Section Five present the results and the conclusion.

2 Statistical Models for Modelling Demand For Health Care

There are numbers of approaches taken in modelling demand for health care by considering the nature of the data used. This paper outlined some models that are regularly used in the literatures starting with logit and probit models and count data models which consist of Poisson, negative binomial and hurdle model. The discussions are frequently referred to Cameron and Trivedi (2006; , 2005) and Winkelmann and Zimmermann (1995).

2.1 Logit and Probit Model

Logit or probit models are frequently used in the first part of a two-part model (*see* Sarma & Simpson, 2006). It includes binary responses such as whether the respondents have visited their General Practitioners (GPs) or whether they have consumed any other health services within specific time period, for instance, in the last two weeks before the interview. The first part of the model is said to represent a contact decision with health care while the second part shows the degree of the demand which involve truncated count data models. This type of model, also known as hurdle model, is discussed in Section 2.5. Besides health care demand, logit and probit models have also been used in modelling demand for health which represented by demand for good health state (Windmeijer *et al*,1997).

1) Logit Model

Consider the dependent variable for each observation, y_i , takes value 1 with probability P_i and 0 with probability $1\text{-}P_i$. Y_i follows the Bernoulli probability distribution where

$$E(y_i|x_i) = \beta_1 + \beta_2 x_i = P_i \ , \qquad i=1,2\ldots..N \tag{2.1}$$

Suppose y_i^* is an unobserved latent health variable with the index function model and x_i be the vector of exogenous variables that determine health.

$$y_i^* = x_i'\beta + u_i \tag{2.2}$$

We only can observe y_i that is linked to y_i^* by

$$y_i = \begin{cases} 1 \, if \; y_i^* > 0 \\ 0 \, if \; y_i^* \le 0 \end{cases} \tag{2.3}$$

The cumulative distribution function (cdf) of the logistic is

$$\Pr(y_i = 1) = P_i = \frac{\exp(x_i'\beta)}{[1+\exp(x_i'\beta)]} \tag{2.4}$$

The log-likelihood function of this model is:

$$\begin{aligned} \log L &= \sum_{i=1}^{n} \left[y_i \log P_i + (1-y_i)\ln(1-P_i) \right] \\ &= \sum_{i=1}^{n} \left[y_i (x_i'\beta) - \log(1+\exp(x_i'\beta)) \right] \end{aligned} \tag{2.5}$$

In order to obtain unknown parameter, β, we differentiate Equation (2.5) with respect to β and set it equal to 0. Since the equations obtained are nonlinear, there are no unique solutions for β which requires us to use non-linear estimation. Unknown parameter can also be obtained by using statistical software, eg. STATA by using *logit* command and later be substituted into Equation (2.4) in order to estimate $P_{i.}$

2) Probit Model

By referring to the same model in Equation (2.2), we now assume that error term u_i has a standard normal distribution; $u_i \sim N(0,1)$. The probability of observing y_i, giving the value of x_i is

$$\Pr\left(y_i = 1 | x_i\right) = \Phi\left(x_i'\beta\right) \tag{2.6}$$

where $\Phi(.)$ is the cdf of the standard normal. The log-likelihood function for probit is

$$\log L = \sum_{i=1}^{n} \left[(1-y_i) \log\left\{1 - \Phi\left(x_i'\beta\right)\right\} + y_i \log \Phi\left(x_i'\beta\right) \right] \tag{2.7}$$

2.2 Ordered logit and probit

Ordered logit and probit have been used in analysing the discrete outcomes which have more than two possible responses. These models have been used in various applications including in health economics (Arendt, 2005; Fu, Liu, & Christensen, 2004; Lindeboom & van Doorslaer, 2004; Ngalula, Urassa, Mwaluko, Isingo, & Ties Boerma, 2002). Unlike the binary index in the previous logit and probit models, the outcomes are given in ordinal values (*see* Daykin & Moffatt, 2002; Jones, 2000). For instance, self-perceived health status index can take more than two values like excellent, good, fairly good and poor.

Consider again the latent health model which takes the form as below:

$$y_i^* = x_i'\beta + u_i \tag{2.8}$$

Since y_i^* is unobserved, we observe $y_i = j$ if $\kappa_{j-1} < y_i^* \leq \kappa_j$, j=1,……$m$

The parameters $\kappa_1, \ldots \kappa_{j-1}$ are the threshold parameters or the cut-points. The probability of observing an observation in interval *J* in the logit model is

$$\Pr(y_i = j) = P_{ij} = \Pr(\kappa_{j-1} < \sum x_i'\beta + u_i < \kappa_j) \tag{2.9}$$

while in ordered probit, when the error term is assumed to be normally distributed; $u_i \sim N(0,1)$ the probability becomes

$$\Pr(y_i = j) = \Phi(\kappa_j - \sum x_i'\beta) - \Phi(\kappa_{j-1} - \sum x_i'\beta) \tag{2.10}$$

where Φ is standard normal distribution function. The log-likelihood for each individual, *i*, is

$$LogL = \sum_i \log\left[\Phi\left(\kappa_j - x_i'\beta\right) - \Phi\left(\kappa_{j-1} - x_i'\beta\right)\right] \tag{2.11}$$

The threshold parameters $\kappa_1, \ldots \kappa_{j-1}$ and β can be obtained by maximising Equation (2.11). The coefficients are constant across categories of the dependent variable, y_{ij}, because of the assumption of a single linear index in the model (Jones, Rice, d'Uva, & Balia, 2007).

2.3 Poisson

The Poisson model is a basic model for count data which assumes that (1) The probability that each event occurs is the same for each unit measure (e.g time, space); (2) The number of events in a specific unit of measure is independent of the number of events occur in other units and (3) The conditional mean and variance are equal due to (1) and (2). The Poisson distribution can be used to model the probability of specified occurrences that happen within a unit of time. The number of occurrences in one unit of time is independent of the number of the previous occurrences. Poisson distribution can be used to estimate non-negative independent variable, for example number of visits to hospital in a specific time frame. The number of episodes is distributed with probability mass function:

$$\Pr(Y = y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0,1,2..., \tag{2.12}$$

with properties

$$\mathrm{E}(Y) = \mu \qquad \text{and} \qquad \mathrm{V}(Y) = \mu \tag{2.13}$$

From Equation (2.12) and (2.13), the mean μ can be associated with regressors *x*, which also shows the equality of mean and variance (equidispersion). For each observation *i*,

$$\mathrm{E}(y_i|x_i) = \mu_i = \exp(x_i'\beta) = \mathrm{V}(y_i|x_i), \qquad i=1, 2\ldots N \tag{2.14}$$

By considering Equation (2.12) and (2.14) together with the assumption of independent occurrences of $(y_i \mid x_i)$, the maximum likelihood estimator is used with the log-likelihood function

$$\ln L(\beta) = \sum_{i=1}^{N} \{y_i x_i'\beta - \exp(x_i'\beta) - \ln y_i!\} \tag{2.15}$$

The standard estimator for this model is the maximum likelihood estimator . By assuming of *K* linearly independent covariates (regressors), the first order conditions of *K* nonlinear equations is given by the Poisson MLE, $\hat{\beta}$. The first order conditions

$$\sum_{i=1}^{N} (y_i - \exp(x_i'\beta))x_i = 0 \tag{2.16}$$

The summation of the residuals $y_i - \exp(x_i'\beta)$ in Equation (2.16) is equal to 0, provided if the regressors include a constant term. The fact that the log-likelihood function is globally concave; the equations are solved by Newton-Raphson iterative technique. Besides Poisson model that have been estimated by maximum likelihood, there are many Poisson related models that can be estimated after correct specification of the mean and variance (*see* Cameron & Trivedi, 1998; 59-70). Another Poisson related model is Zero Inflated Poisson (ZIP). This model takes into account the distribution with excess zeros and tries to correct it. These excess zeros are believed to be generated by different data generating process (DGP). A number of studies in health economics that have already utilised excess zeros model suggest that the model fit the data better than the ordinary model (Affleck, 2006; Liming Xiang, 2007; Rose, Rose, Martin, Wannemuehler, & Plikaytis, 2006).

2.4 Negative Binomial

Count data may turn out to be overdispersed because of unobserved heterogeneity, a different reason for occurrence of the same consequent events, or the number of the events are dependent on the number of events occur in the previous units. In these cases, a negative binomial model is used as an alternative of the Poisson model. Negative binomial is a more flexible model that assumes the variance has a multiple or quadratic function of the mean.

Suppose random variable *y* has a Poisson distribution with parameter λ. If we now specified λ as a gamma distributed random variable with parameters μ, α $(\mu, \alpha > 0)$, the

distribution becomes negative binomial which every observation has a distinct λ because of unobserved heterogeneity. With α represents the scalar parameter, the density is given by

$$f(y|\mu,\alpha)=\frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha-1}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}, \quad \alpha\geq 0, \quad y=0,1,2,.... \tag{2.17}$$

where

$$\mathrm{E}(y|\mu,\alpha)=\mu \qquad \mathrm{V}(y|\mu,\alpha)=\mu+\alpha\mu^{k}, \qquad \mu,\alpha>0 \tag{2.18}$$

Since $\mu,\alpha>0$, the variance therefore exceed the mean. There are two variance functions depending on *k*. If we set *k*=1, the variance becomes proportional to the mean (NB1 model) while by setting *k*=2, the variance becomes a quadratic function of the mean (NB2) model. Setting α =0 in Equation (2.17), simplifies the model to the Poisson.

2.5 Hurdle Model (Two-part model)

Hurdle model treats participation or contact decision and the frequency of visit as a different process that are explained by separate DGPs. Suppose, the first process has the probability $P_1(.)$ and the second process $P_2(.)$. The zeros are explained by the first process while the positive counts are determined by the second process. The probability of these two distinct processes are given by

$$\Pr(y=0)=f_1(0)$$

$$\Pr(y=j|\,j>0)=\frac{\Pr(y=j)}{\Pr y>0}=\frac{f_2(j)}{1-f_1(0)}, \quad j>0 \tag{2.19}$$

The log-likelihood is

$$LogL=\sum_{y=0}\log\left[1-P_1(y>0|x\right]+\sum_{y>0}\left\{\log\left[P_1(y>0|x\right]+\log\left[P_2(y|x,y>0\right]\right\}$$

$$=\left\{\sum_{y=0}\log\left[1-P_1(y>0|x\right]+\sum_{y>0}\log\left[P_1(y>0|x\right]\right\}+\left\{\sum_{y>0}\log\left[P_2(y|x,y>0\right]\right\}=LogL_1+LogL_2 \tag{2.20}$$

To understand health care demand in the light of hurdle model, we have to divide decision process into two processes. First, is when the individual decide to demand health care in a certain period of time and the second process is when the health care provider, after the first contact, determines the next visit. These processes, however, assume that the demands are determined by a single spell of illness. Studies that utilise hurdle model include Pohlmeier & Ulrich (1995), Jiménez-Martín, Labeaga, & Martínez-Granado (2002), Mocan,

Tekin, & Zax (2004) and Sarma and Simpson (2006). Pohlmier and Ulrich (1995) employ a negative binomial distributed hurdle model to explain the demand for health care. They suggest that a two-part model is essential because of different decision processes whereby the initial visit to the physician is determined by the individual while the frequency is decided by the physician. The theoretical model used is largely based on Grossmann (1972), Muurinen (1982) and Wagstaff (1986). Martín *et al.* (2002) compare the two-part models with the latent class models in order to estimate demand for physician services of twelve countries in European Union. By using two tests which are known as the *Akaike Information Criterion (AIC)* and the *Bayesian Information Criterion (BIC)*, it was found that the two part models are more favoured for specialist demand framework while latent class models are better in explaining demand for GPs.

By using the data from Canadian National Population and Health Survey, Sarma and Simpson (2006) have also compared several demand frameworks that include the negative binomial, zero-inflated negative binomial (ZINB), hurdle and latent class. Both AIC and BIC criterions suggest that the hurdle model are less preferred compared to the latent class model for modelling doctors and GPs' visits while ZINB model is found to be more appropriate in modelling specialist visits and number of nights spent as inpatient. By utilising the data from RAND Health Insurance Experiment, the AIC and BIC in Deb and Trivedi (2002), also favour the latent-class model over the two-part model.

3 Methods and Data

3.1 Data

For this empirical analysis data from British General Household Survey 2004 (GHS2004) are used. GHS is a national survey on various topics concerning private households which have been carried out annually. It has two types of questionnaires namely (1) household questionnaire which is completed by the Household Reference Person (HRF) and (2) individual questionnaire which is completed by a household member aged 16 and over. In 2004/2005 survey, it covers 8,700 households which consists 20,421 individuals. All adults aged 16 or over were interviewed while proxies were used to answer for children. Sampling process involves two stage sampling technique. The first stage, known as Primary Sampling Units (PSU) were based on postcode sectors while the second stage or Secondary Sampling Units (SSU) were addresses within those sectors. Some variables had been recoded

from their original values in order to suit the assumptions of some econometric models[2]. The summary statistics are described in Table 3.1 and Table 3.2.

Table 3.1 Summary Statistics of Dependent Variables

Variables	Definitions	Mean	Std. Dev	Min	Max
consultation	No. of consultations to doctor in the past 2 weeks.	0.196	0.532	0	9
nurse	No. visits to practise nurse in the past 2 weeks.	0.0818	0.336	0	7
outpatient	No. of outpatient visits in the last 3 months.	0.293	1.26	0	42
inpatient	No. of separate days as inpatient in the past one year apart from maternity stays.	0.0907327	0.3953832	0	6

Table 3.2 Summary Statistics of Explanatory Variables

Variables[3]	Definitions	Mean	Std. dev	Min	Max
GH	Self-perceived health state Not Good …… 0 Fairly good ……...1 Good ……2	1.534089	.688828	0	2
age	Age measured by absolute numbers	38.8696	22.93802	0	99
age2	Square of age	2036.972	1921.563	0	9801
male	Male	.4853337	.4997971	0	1
edulevel					
edulevel_1	Child (reference group)				
edulevel_2	No qualification	.160948	.3674938	0	1
edulevel_3	Other qualification	.3678222	.4822269	0	1
edulevel_4	Higher qualification	.2209056	.4148692	0	1
ndyscutd1	Number of activities cutdown because of illness during the last 2 weeks, including Saturday and Sunday	1.091262	3.392061	0	14

3.2 Empirical Specifications

The initial empirical model for health care demand is specified as below:

$$HC_{ij} = \exp(x_i'\beta_j) + u_i, \qquad i = 1...N \tag{3.1}$$

where HC_{ij} is the demand for health care type j and x_i are exogenous variables that determine HC_{ij}. As no assumption has been made to distinguish different decision process between contact decision and density of visits, this model is estimated by both Poisson and

[2] For example, number of demand in *consultation*, *nurse*, *outpatient* and *inpatient* had been recoded such that it takes into account the respondents with 'No' answers which represent zero utilisation.

[3] Initially, income and lifestyle variables such as drinking and smoking behaviour have been included in the models to see whether they have a direct impact on demand for health care. However, they seem not to be statistically significant that I suspect those variable might have some influence in determining health status.

negative binomial instead of hurdle or zero inflated models. Besides, it is difficult to identify from the survey whether demand for a particular health care is from the same episode of illness which makes us unable to differentiate the type of process. While developing the demand model, one of the regressors, GH_i, is suspected to be correlated with the error term, u_{ij}, which suggest that $E(u_{ij}|x_i) \neq 0$. This is because any short term illness might influence demand for health care which in turn affect the value one put on his or her general health. As such, GH_i is instrumented by a set of covariates z_i so that $E(u_{ij}|z_i) = 0$. In health care demand models, true health capital is unobserved, so self-perceived of general health state is chosen in this paper to represent the level of health. Ordinal values are assigned for this health state where each respondent chose to rate his or hers as 'Not good', Fairly good' or 'Good' (*see* Table 3.3 for frequency distribution of *GH*). The model takes form as

$$HC_{ij} = \exp(\alpha GH_i + x_i'\beta_j) + u_{ij}, \qquad j = 1,....J \quad \text{and} \quad i = 1...N \tag{3.2}$$

$$GH_i^* = \varphi HC_{ij} + z_i'\gamma + w_i \quad , \qquad i = 1...N \tag{3.3}$$

GH_i^* is unobserved health capital, z_i are exogenous variables that determine health; consists of age, gender, education level, drinking and smoking status, income, accommodation type and number of long standing illness as described in Table 3.4. Since GH_i^* is unobserved, we observe $GH_i = s$ if $\kappa_{s-1} < GH_i^* \leq \kappa_s$, s=1, 2……*m*. As pointed out by Windmeijer & Santos-Silva (1997), the model is only *coherent* when the system is triangular, that is either α=0 or φ=0, such that $\sum_{s=1}^{3} P_{is} = \sum_{s=1}^{3} \Pr(GH_i = s) = 1$ (*see* Blundell & Smith, 1994; Gourieroux, Laffont, & Monfort, 1980). In this case, we assume φ=0, which means that the short term health care consumptions, HC_{ij}, do not directly affect GH_i^*, which represents long term health capital. The model is estimated in stages where GH_i is estimated by ordered probit which assumes a single linear index, such that the coefficients do not change across categories (Jones *et al.*, 2007; 40). Since the system is recursive (triangular) the first stage requires the ordinal health index, GH_i to be regressed on z_i. The predicted value of GH_i is substituted into Equation (3.2) which then be estimated by the negative binomial. Instrumental variables z_i, are summarised in Table 3.4.

Table 3.3 Frequency Distribution of Self-Perceived Health State

Self-perceived Health State (GH)	Frequency	Percent	Cumulative
Not Good	2,161	11.28	11.28
Fairly Good	4,603	24.03	35.31
Good	12,392	64.69	100.00
Total	**19,156**	**100.00**	

Table 3.4 Summary Statistics of Instruments used in Health Equation

Variables	Definitions	Mean	Std. Dev	Min	Max
age	age measured by absolute numbers	38.8696	22.93802	0	99
age2	Square of age	2036.972	1921.563	0	9801
male	Male	.4853337	.4997971	0	1
edulevel					
edulevel_1	Child (reference group)				
edulevel_2	No qualification	.160948	.3674938	0	1
edulevel_3	Other qualification	.3678222	.4822269	0	1
edulevel_4	Higher qualification	.2209056	.4148692	0	1
loginc	Log of households' weekly income in GBP	5.868545	1.105867	0	10.30
drink1	Drinking status (Ever drink at all) children being recoded as 0.	0.7324323	0.4427017	0	1
cignow4	Smoke once in a month	0.1869613	.38989	0	1
acctyp3	Accommodation type				
acctyp3_1	House				
acctyp3_2	Flat	.1295725	.3358407	0	1
acctyp3_3	Caravan	.001518	.0389335	0	1
nillness	Number of Longstanding Illness	.4792196	.876951	0	6

To test for endogeneity, a linear form of Equation (3.2) is considered[4]. This step permits us to use modelling technique by using instrumental variable (IV) approach with GMM option (IV-GMM). This approach could provide us statistics for underidentification test overidentification test (Sargan statistics) and endogeneity test.

4 Results

Results from IV-GMM are presented in Table 4.1. All types of demand pass both underidentification and overidentification tests except for *outpatient* which shows some symptom of correlation between chosen instruments and the error term. Overidentification

[4] Cameron *et a.l* (1988, p 103) suggest that results from comparing IV and OLS model for Hausman test can be generalized for count model as the outcome are similar.

test is used to identify whether the selected instruments for GH_i are not correlated with the error term, u_{ij}[5]. Thus we could conclude that instruments used, except for *outpatient* are orthogonal to the disturbance process which implies $E(u_{ij}|z_i) = 0$. Endogeneity test for all j also rejects the null hypothesis of exogeneity. By this it means that the results statistically suggest that GH_i should be treated as endogenous in the model. However, by looking at the nature of the data, two-stage estimation that consists of ordered probit in the first stage and count model in the second stage is considered[6]. Equation (3.3) is estimated by using ordered probit and predicted value for GH_i is substituted into health care demand equation; equation (3.2). Results are presented in Table 4. 2.

Table 4.1 IV-GMM Estimates

Variable	*consultation*	*Nurse*	*outpatient*	*inpatient*
GH	-0.2024***	-0.0863***	-0.5888***	-0.1804***
	(0.0153)	(0.0099)	(0.0359)	(0.0109)
age	-0.008***	-0.0027**	-0.0385	0.0034***
	(0.0014)	(0.0012)	(0.0033)	(0.0010)
age2	0.00007***	0.00004***	.00004	0.00003***
	(0.000017)	(0.000013)	(0.0000398)	(0.000012)
male	-0.0571***	-.02638***	0.0009	-0.0108*
	(0.0080)	(0.0056)	(0.0188)	(0.0057)
edulevel_2	0.1194***	-1.1143***	-0.0384	0.01583
	(0.0240)	(0.0370)	(0.0563)	(0.0172)
edulevel_3	0.1281***	-1.1039***	0.0107	0.0215
	(0.0216)	(0.0366)	(0.0506)	(0.0154)
edulevel_4	0.1616***	-1.1041***	0.1045*	0.0347*
	(0.0232)	(0.0372)	(0.0543)	(0.0165)
ndyscutd1	0.0342***	.0041***	0.0324***	0.0092***
	(0.0017)	(0.0011)	(0.0040)	(0.00124)
constant	0.5687***	1.3474***	1.313***	0.4037***
	(0.0311)	(0.0377)	(0.0730)	(0.0223)
Underidentification Test (*p-value*)	0.0000	0.0000	0.0000	0.0000
Overidentification test (*p-value*)	0.5580	0.1156	0.0108	0.1446
Endogeneity test (*p-value*)	0.0001	0.0001	0.0000	0.0000
N	14999	11300	14995	14995

* Figures in parenthesis are t-statistics. The symbols ***, ** and * denote 1, 5 and 10% level of significance, respectively
** *edulevel_1* is the reference variable

The fact that the results in Table 4.1 is from a linear model, the results of two-stage negative binomial model are used for discussion. Even so, it is worth to note that the directions of all coefficients are consistent between these two models. As suspected, the

[5] Null hypothesis for overidentification test is the instruments are not correlated with the error term.
[6] Both LR and Wald overdispersion tests suggest of strong rejection of Poisson model at 1% critical value. Thus negative binomial model is used for estimation in the second stage.

predicted value of self-perceived health state index, $GHhat_i$, has a very strong negative effect in all equations. It suggests that healthier people are less likely to demand health care than people in poor health state. While age has a concave relationship with maximum point in Gurmu (1997) and Windmeijer & Santos Silva (1997), this paper suggest the opposite findings which is consistent with Pohlmier and Ulrich (1995) and Cameron *et al.* (1988). Nonetheless, these results cannot be directly compared as contradiction occurs might be due to variation in type of demand investigated or utilisation of different dataset which focus on specific age group. Age is significant in all equations except for *nurse* with 'U-shaped' relationship. Minimum number for doctor consultations, outpatient visits and inpatient are at age 51, 40 and 44 respectively. *Male* also plays an important role in *consultation, nurse* and *inpatient* which signify that males demand less health care than females. All dummies for education show a significant impact in all equations at least at 10% confident level except for *edulevel_2* in *outpatient*. As for *consultation, outpatient* and *inpatient*, education levels have negative effect when compared to the reference group but have opposite roles in *nurse*. Representing the short-term health status, *ndyscutd1*, also has a strong positive relationship with demand intensity.

Table 4.2 Two-Stage Negative Binomial Estimates

Variable	***consultation***	***nurse***	***Outpatient***	***Inpatient***
GHhat	-0.3432***	-0.3515***	-0.6901***	-0.5987***
	(0.0289)	(0.0439)	(0.040)	(0.0473)
age	-0.0547***	-0.0179	-0.0289***	-0.0529***
	(0.0079)	0.0153)	(0.0095)	(0.0131)
age2	0.0005***	0.0003**	0.0004***	0.0005***
	(0.00009)	(0.0002)	(0.00011)	(0.00015)
male	-0.3196***	-0.4418***	0.0254	-0.1257**
	(0.0438)	(0.0789)	(0.0524)	(0.0730)
edulevel_2	0.8327***	-3.424***	0.1478	0.5101**
	(0.1337)	(0.2932)	(0.1604)	(0.2246)
edulevel_3	0.8476***	-3.3060***	0.2490*	0.4809**
	(0.1232)	(0.2828)	(0.1457)	(0.2089)
edulevel_4	1.0255***	-3.4039***	0.5014***	0.6018***
	(0.1321)	(0.2948)	(0.1568)	(0.2255)
ndyscutd1	0.1113***	0.0640***	0.1020***	0.0978***
	(0.0049)	(0.0083)	(0.0070)	(0.0082)
constant	-1.4847***	0.35636***	-1.5597***	-2.4232***
	(0.0715)	(0.1834)	(0.0901)	(0.1195)
Log *L*	-7250.02	-2726.69	-8195.76	-3734.76
N	15003	11302	14999	14999

* Figures in parenthesis are t-statistics. The symbols ***, ** and * denote 1, 5 and 10% level of significance, respectively
** *edulevel_1* is the reference variable

5 Conclusion

This paper deals with problem of ordinal endogenous variable in count data models of health care demand. An ordinal endogenous variable contains more information of the respondents' self-perceived general health state, which is believed, could provide more precise results than the binary endogenous variable. Endogeneity test on all equations suggest that self-perceived health state, GH_i is endogenous. Thus, two-stage estimation has been utilised which consist of ordered probit on health equation in the first stage and the negative binomial, with predicted value of GH_i ($GHhat_i$) on health care demand equations in the second stage. As expected, health related variables, $GHhat_i$ and *ndyscutd1* are statistically significant in determining all types of health care demand. Results also suggest that age, gender and education levels also play some roles in determining demand, though in mixed directions which depending on the type of health care.

References

Affleck, D. L. R. (2006). Poisson mixture models for regression analysisof stand-level mortality. *Canadian Journal of Forest Research, 36*(11), 2994-3006.

Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. *Journal of Econometrics, 2*(2), 105-110.

Arendt, J. N. (2005). Does education cause better health? A panel data analysis using school reforms for identification. *Economics of Education Review, 24*(2), 149-160.

Blundell, R., & Smith, R. J. (1994). Coherency and estimation in simultaneous models with censored or qualitative dependent variables. *Journal of Econometrics, 64*(1-2), 355-373.

Cameron, A. C., & K.Trivedi, P. (2006). *Regression Analysis of Count Data* (Vol. New York): Cambridge University Press.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Application*. New York: Cambridge University Press.

Cameron, A. C., Trivedi, P. K., Frank, M., & Pigott, J. (1988). A Microeconomic Model of the Demand for Health Care and Health Insurance in Australia. *The Review of Economics Studies, 55*(1), 85-106.

Daykin, A. R., & Moffatt, P. G. (2002). Analyzing Ordered Response: A Review of the Ordered Probit Model. *Understanding Statistics, 1*(3), 156-166.

Deb, P., & Trivedi, P. K. (2002). The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics, 21*(4), 601-625.

Fu, A. Z., Liu, G. G., & Christensen, D. B. (2004). Inappropriate Medication Use and Health Outcomes in the Elderly. *Journal of the American Geriatrics Society, 52*(11), 1934-1939.

Gourieroux, C., Laffont, J. J., & Monfort, A. (1980). Coherency Conditions in Simultaneous Linear Equation Models with Endogenous Switching Regimes. *Econometrica, 48*(3), 675-695.

Gurmu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization. *Journal of Applied Econometrics, 12*(3), 225-242.

Jiménez-Martín, S., Labeaga, J. M., & Martínez-Granado, M. (2002). Latent class versus two-part models in the demand for physician services across the European Union. *Health Economics, 11*(4), 301-321.

Jones, A. M. (2000). Health Econometrics. In A. J. Culyer & J. P. Newhouse (Eds.), *Handbook of Health Economics* (Vol. 1A). London: North-Holland Elsevier.

Jones, A. M., Rice, N., d'Uva, T. B., & Balia, S. (2007). *Applied Health Economics*. London: Routledge.

Kelejian, H. H. (1971). Two-Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variable. *Journal of the American Statistical Association, 66*(334), 2.

Liming Xiang, A. H. L. K. K. W. Y. G. J. M. (2007). A score test for overdispersion in zero-inflated poisson mixed regression model. *Statistics in Medicine, 26*(7), 1608-1622.

Lindeboom, M., & van Doorslaer, E. (2004). Cut-point shift and index shift in self-reported health. *Journal of Health Economics, 23*(6), 1083-1099.
Mocan, H. N., Tekin, E., & Zax, J. S. (2004). The demand for medical care in urban China. *World Development, 32*(2), 289-304.
Mullahy, J. (1997). Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior. *The Review of Economics and Statistics, 79*(4), 586-593.
Ngalula, J., Urassa, M., Mwaluko, G., Isingo, R., & Ties Boerma, J. (2002). Health service use and household expenditure during terminal illness due to AIDS in rural Tanzania. *Tropical Medicine & International Health, 7*(10), 873-877.
Pohlmeier, W., & Ulrich, V. (1995). An Econometric-Model of the Two Part Decision-Making Process in the Demand for Health-Care. *Journal of Human Resources, 30*(2), 339-361.
Rose, C. E., Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data. *Journal of Biopharmaceutical Statistics, 16*(4).
Sarma, S., & Simpson, W. (2006). A microecenometric analysis of Canadian health care utilization. *Health Economics, 15*(3), 219-239.
Terza, J. V. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics, 84*(1), 129-154.
Wagstaff, A. (1986). The Demand for Health - Some New Empirical-Evidence. *Journal of Health Economics, 5*(3), 195-233.
Windmeijer, F. A. G., & Santos-Silva, J. M. C. (1997). Endogeneity in Count Data Models: An Aplication To Deman For Health Care. *Journal of Applied Econometrics, 12*(3), 281-294.
Winkelmann, R., & Zimmermann, K. F. (1995). Recent Developments in Count Data Modelling: Theory and Application. *Journal of Economic Surveys, 9*(1), 1-24.