Windmeijer, Santos Silva:

Estimation of count data models with endogenous regressors; an application to demand for health care

Sonderforschungsbereich 386, Paper 20 (1996)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Estimation of count data models with endogenous regressors; an application to demand for health care*

F.A.G. Windmeijer
University College London

J.M.C. Santos Silva
ISEG/Universidade Técnica de Lisboa

Abstract

The generalized method of moments (GMM) estimation technique is discussed for count data models with endogenous regressors. Count data models can be specified with additive or multiplicative errors. It is shown that, in general, a set of instruments is not orthogonal to both error types. Simultaneous equations with a dependent count variable often do not have a reduced form which is a simple function of the instruments. However, a simultaneous model with a count and a binary variable can only be logically consistent when the system is recursive. The GMM estimator is used in the estimation of a model explaining the number of visits to doctors, with as a possible endogenous regressor a self-reported binary health index. Further, a model is estimated, in stages, that includes latent health instead of the binary health index.

Address for correspondence: Dept. of Economics, University College London, Gower Street, London WC1E 6BT

*__Acknowledgements:__ We are grateful to Richard Blundell, Rachel Griffith, Costas Meghir and two anonymous referees for helpful comments, and to Susan Harkness and Steve Machin for providing us with the health data used in this study. This research has been carried out at University College London with the financial support of the Human Capital and Mobility program of the European Commission, which is gratefully acknowledged. We thank the department of economics at UCL for their hospitality and support. The usual disclaimer applies.

1. Introduction

In many economic applications the variable of interest is a count process, for example the number of visits to the doctor, job applications or patent applications. The modelling and estimation of such inherently nonlinear processes are well established nowadays, with early references Gourieroux, Monfort and Trognon (1984), Cameron and Trivedi (1986) and McCullagh and Nelder (1983, 1989). Recent developments are surveyed by Gurmu and Trivedi (1994) and Winkelmann and Zimmermann (1995).

In microeconomic applications, explanatory variables are often simultaneously determined with the dependent variable, resulting in coefficient estimates that are inconsistent when obtained by standard methods. Techniques for dealing with simultaneity in count data regression models are largely underdeveloped when compared to the continuous data case. In a recent paper, Mullahy (1996) discusses instrumental variables, or generalized method of moments, and two stage estimation methods for Poisson regression models in a specification where unobserved heterogeneity is correlated with the regressors. Testing for exogeneity has been discussed earlier by Grogger (1990), who proposed use of the Hausman test after estimation of the model by non-linear instrumental variables.

In this paper we discuss the generalized method of moments (GMM) estimator for count data models with endogenous regressors, utilizing first-order moment conditions only. Given a set of exogenous instruments for the endogenous variables, the GMM estimation method provides consistent estimates for the parameters.

Count data regression models can be specified with additive or multiplicative errors. These specifications are in principle observationally equivalent when only the first order conditional mean is specified. Differences arise, however, when the choice of instruments in the two specifications under endogeneity is considered, as the same set of instruments is, in general, not orthogonal to both type of errors.

For nonlinear models, the reduced form for endogenous regressors often has a very complicated structure. This is also the case for the reduced form of simultaneous equation models that include a dependent count variable. A specification we analyse and use for the modelling of demand for health care, is that of a binary variable which is simultaneously determined with the count variable. Such a system is coherent in the sense of Blundell and Smith (1994), only when it is recursive. For this model, an alternative to GMM estimation, which amounts to instrumenting the binary variable, is the estimation in two stages of a model which is specified in terms of a latent continuous variable. This latent variable determines the binary outcome by means of a threshold transformation.

The model specifications and estimators are applied to a model for visits to

or by a doctor (general practitioner) in the last month before the interview, utilizing data from the British Health and Lifestyle Survey 1991-1992. Models that explain the number of visits to the doctor have also been analysed by Cameron et al. (1988) and Pohlmeier and Ulrich (1995) for Australian and German data respectively. Factors such as income and education are likely to have a direct effect on demand for medical care, but are also important determinants of health, which in turn affects demand. In order to estimate the direct effects of income and education, a measure of health has to be included in the model. The measure we consider is a self-reported health index that is likely to have measurement error which is correlated with the number of visits to the doctor. We therefore instrument the binary health index by estimated reduced form probabilities. Finally, a model is estimated which includes underlying latent health instead of the binary health index.

The outline of the paper is as follows. In section 2 the GMM estimation technique is presented. Section 3 discusses simultaneous models and section 4 presents the estimation results for the demand for doctors. Section 5 concludes.

2. Generalized Method of Moments Estimation

Before discussing the estimation of count data models with endogenous regressors by generalized method of moments, we first present the standard Poisson model for count data.

Let y_i, $i = 1, \ldots, N$, denote the dependent count variable, which is independently Poisson distributed, with the conditional mean specified as

$$E\left(y_i|x_i\right) = \mu_i = \exp(x_i'\beta), \tag{2.1}$$

where x_i is a k-vector of explanatory variables and β is a k-vector of parameters. The maximum likelihood estimator, denoted $\widehat{\beta}_{ML}$, solves the first-order condition $X'\left(y-\mu\right) = 0$, where X is an $N \times k$ matrix, and y and μ are N-vectors, and $\sqrt{N}\left(\widehat{\beta}_{ML} - \beta\right)$ has a limiting normal distribution with mean zero and variance the limit of $\left(\frac{1}{N}X'MX\right)^{-1}$, where $M = diag\left(\mu_i\right)$. In practice, the standard errors are often biased due to the presence of over- or underdispersion. Correct standard errors in these cases are computed from the estimated variance of the Poisson pseudo-likelihood estimator $\widehat{\beta}_{PL}$ $(= \widehat{\beta}_{ML})$, (Gourieroux, Monfort and Trognon (1984)):

$$\widehat{Var}\left(\widehat{\beta}_{PL}\right) = \left(X'\widehat{M}X\right)^{-1}\left(\sum_i \left(y_i - \widehat{\mu}_i\right)^2 x_i x_i'\right)\left(X'\widehat{M}X\right)^{-1},$$

where $\widehat{M} = diag\left(\widehat{\mu}_i\right)$, and $\widehat{\mu}_i = \exp\left(x_i'\widehat{\beta}_{PL}\right) = \exp\left(x_i'\widehat{\beta}_{ML}\right)$.

The conditional mean specification (2.1) implicitly defines a regression model

$$y_i = \mu_i + u_i = \exp(x_i'\beta) + u_i,$$

with $E(u_i|x_i) = 0$. The generalized method of moments (GMM) estimator (see Hansen (1982), Ogaki (1993)), based on this moment condition only, minimizes

$$(y-\mu)' X W_N^{-1} X' (y-\mu), \tag{2.2}$$

where W_N is a weight matrix. As the minimum of (2.2) is obtained at $X'(y-\mu) = 0$, the GMM estimator for β will be the same as the Poisson maximum likelihood estimator. The efficient GMM estimator is obtained for $W_N = Var(X'(y-\mu)) = X'\Omega X$. When the variance equals the mean, Ω is equal to M and the variance of the GMM estimator is equal to the variance of the Poisson maximum likelihood estimator. The variance of the GMM estimator is equivalent to the pseudo-likelihood variance when $(X'\Omega X)^{-1}$ is estimated by $\sum_i (y_i - \widehat{\mu}_i)^2 x_i x_i'$.

When some elements of x_i are endogenous, implying that

$$E(u_i|x_i) \neq 0, \tag{2.3}$$

the Poisson ML estimator will be inconsistent. If there are instruments z_i available such that

$$E(u_i|z_i) = 0,$$

then the consistent non-linear instrumental variables (NLIV) estimator (see Amemiya (1985)) is given by minimisation of

$$(y-\mu)'Z(Z'Z)^{-1}Z'(y-\mu), \tag{2.4}$$

and is a one step GMM estimator.[1] The efficient two step GMM estimator, given the instruments, is found by minimisation of

$$(y-\mu)'Z(\widetilde{Z'\Omega Z})^{-1}Z'(y-\mu),$$

where

$$\widetilde{Z'\Omega Z} = \sum_{i=1}^{N} (y_i - \widetilde{\mu}_i)^2 z_i z_i',$$

is an estimate of the (asymptotic) variance of $Z'(y-\mu)$, with $\widetilde{\mu}_i = \exp(x_i'\widetilde{\beta})$, and $\widetilde{\beta}$ is the first round estimate of β. Denote the two step GMM estimator

[1] The NLIV estimator clearly does not take into account the heteroscedasticity of u. A 'Poisson' type first round estimator is obtained by minimizing $(y-\mu)'Z(Z'MZ)^{-1}Z'(y-\mu)$, which should be iterated till convergence.

by $\widehat{\beta}_{GMM2}$. Under standard regularity assumptions, the limiting distribution of $\sqrt{N}(\widehat{\beta}_{GMM2} - \beta)$ is normal with mean zero and variance the limit of

$$\left(\frac{1}{N}\left(X'MZ\right)\left(Z'\Omega Z\right)^{-1}\left(Z'MX\right)\right)^{-1}.$$

Optimal instruments, which minimize the variance of the two step GMM estimator, are given by (see, for example, Davidson and MacKinnon (1993))

$$Z^* = E\left(\Omega^{-1}D|Z\right)$$

where D is the matrix of derivatives $\partial(y-\mu)/\partial\beta$, which is equal to $-MX$. The optimal instruments therefore are

$$Z^* = E\left(\Omega^{-1}MX|Z\right).$$

When $Z = X$, i.e. there is no endogeneity, the optimal instruments for the Poisson model are given by X. For $Z \neq X$, it is in general impossible to get consistent estimates of Z^*. A reasonable working hypothesis may be to specify $\Omega = M$, which leads to the use of the instruments $\widehat{X} = E\left(X|Z\right)$. A test for overidentifying restrictions can be obtained by augmenting the instruments $\widehat{X}$ by the elements of Z different from and not collinear with $\widehat{X}$, and using the standard test for overidentifying restrictions for the GMM estimator that utilizes the augmented instrument matrix.

2.1. Additive vs Multiplicative Errors

A multiplicative model is specified as

$$y_i = \exp\left(x_i'\beta + \tau_i\right) = \exp\left(x_i'\beta\right)v_i = \mu_i v_i, \tag{2.5}$$

which is motivated by treating the unobservables τ_i and observables x_i symmetrically. In principle, the multiplicative and additive models are observationally equivalent when only the first order conditional mean is specified (see Wooldridge (1992)). Differences arise, however, when the choice of instruments in the two specifications under endogeneity is considered.

Endogeneity occurs in (2.5) when $E(v_i|x_i) \neq 1$. Let z_i be a set of instruments which satisfy $E(v_i|z_i) = 1$, then an instrumental variable estimator can be based on the residual $v_i - 1$, which is equal to $(y_i - \mu_i)/\mu_i = u_i/\mu_i$. Under endogeneity the same set of instruments can, in general, not be orthogonal to u_i and u_i/μ_i at the same time, i.e. when $E\left(u_i|z_i\right) = 0$, it follows in general that $E\left(\frac{u_i}{\mu_i}|z_i\right) \neq 0$ due to the correlation between u_i and μ_i.

Which transformation should be used in the GMM estimation when there is endogeneity present is an empirical issue which can be tested by the standard test for overidentifying restrictions in cases where there are more instruments than endogenous variables. The two step GMM estimator in the multiplicative model minimizes the objective function

$$(y-\mu)'M^{-1}Z(Z'\widetilde{\Omega^*}Z)^{-1}Z'M^{-1}(y-\mu),$$

where $Z'\Omega^*Z$ is the asymptotic variance of $ZM^{-1}(y-\mu)$. When $Z=X$ in the Poisson model, with $\Omega^*=M^{-1}$, this becomes

$$(y-\mu)'M^{-1}X(X'M^{-1}X)^{-1}X'M^{-1}(y-\mu), \tag{2.6}$$

which is equivalent to a heteroscedasticity corrected objective function. Clearly, (2.6) will not yield Poisson ML estimates. However, the optimal instruments for the multiplicative model are given by

$$Z^*=E\left(\Omega^{*-1}WX|Z\right),$$

where $W=diag(y_i/\mu_i)$. When $Z=X$ and $\Omega^*=M^{-1}$, this reduces to $Z^*=MX$, and the estimator minimizing (2.6) with MX as instruments is the same as the Poisson ML estimator. Equivalently, if Z are valid instruments in the additive model, then MZ are valid instruments in the multiplicative model, giving the same estimation results for the two model specifications.

Mullahy (1996) proposed use of the transformed residual u_i/μ_i in a model with unobserved heterogeneity which is correlated with (some of) the regressors. His model is specified as

$$y_i=\exp\left(x_i'\beta\right)\eta_i+\varepsilon_i=\mu_i\eta_i+\varepsilon_i, \tag{2.7}$$

where η_i is the unobserved heterogeneity term and

$$E\left(\varepsilon_i|x_i\right)=0 \quad ; \quad E\left(\eta_i|x_i\right)\neq 1.$$

Clearly, model (2.7) is observationally equivalent to the multiplicative model (2.5) with endogeneity.

3. Simultaneous Equations

Specifying a simultaneous equations model where one dependent variable is a count, and which has a reduced form with a simple structure, is virtually impossible. Consider for example a two equation model with a count variable and a

continuous variable that are simultaneously determined. The system of equations can be specified as

$$\begin{aligned} y_1 &= \exp(\alpha y_2 + x_1'\beta) + u_1 \\ y_2 &= \gamma y_1 + x_2'\delta + u_2, \end{aligned}$$

and the reduced form for y_2 has to be derived from

$$y_2 = \gamma \exp(\alpha y_2 + x_1'\beta) + x_2'\delta + \gamma u_1 + u_2,$$

which does not reduce to a simple equation for y_2.

A model specification we are particularly interested in, given the application in the next section, is that of a simultaneous model with a count and a binary variable as endogenous regressors. The possible model specifications are

$$\begin{aligned} y_1 &= \exp(\alpha y_2^* + x_1'\beta) + u_1^* \\ y_2^* &= \gamma y_1 + x_2'\delta + u_2 \\ Cov(u_1^*, u_2) &= \Sigma^*, \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} y_1 &= \exp(\alpha y_2 + x_1'\beta) + u_1 \\ y_2^* &= \gamma y_1 + x_2'\delta + u_2 \\ Cov(u_1, u_2) &= \Sigma, \end{aligned} \tag{3.2}$$

where in both cases y_2^* is and unobserved latent variable. The binary variable y_2 is related to y_2^* by

$$\begin{aligned} y_2 &= 1 \quad if \quad y_2^* > 0 \\ y_2 &= 0 \quad otherwise. \end{aligned}$$

As discussed above, for model (3.1) the reduced form of y_2^* is a complicated function of the exogenous variables. For model (3.2), where the binary variable enters the mean function of the count variable directly, the logical consistency, or coherency, is an issue. Following Maddala (1983, p.118) and Blundell and Smith (1994) it is easily established that model (3.2) is only coherent when $\gamma = 0$ or $\alpha = 0$ as,

$$\begin{aligned} P(y_2 &= 1) + P(y_2 = 0) \\ &= F(\gamma \exp(\alpha + x_1'\beta) + x_2'\delta) + (1 - F(\gamma \exp(x_1'\beta) + x_2'\delta)) \\ &= 1 \quad if \quad \gamma = 0 \quad or \quad \alpha = 0. \end{aligned}$$

This result means that if a binary variable is included in the exponential mean function of the count variable, and the two variables are simultaneously determined, this simultaneity arises via the correlation of u_1 and u_2.

For model (3.1), which is specified in the latent variable only, the parameters cannot be estimated by GMM as y_2^* is not observed. The model may, however, be estimated by substituting some reduced form of y_2^* into the mean function of y_1. For example, if $\gamma = 0$, (3.1) is written as

$$y_1 = \exp\left(\alpha x_2'\delta + x_1'\beta\right)\exp\left(\alpha u_2\right) + u_1, \tag{3.3}$$

and the parameters of this model can be estimated consistently if u_2 is independent of x_2 and x_1. One way to estimate $\sigma_{u_2}\alpha$ is to estimate the model in stages. The first stage is to estimate δ/σ_{u_2} by logit or probit. The second stage is to substitute the estimator for δ/σ_{u_2} into (3.3) and to estimate the model by, for example, Poisson pseudo-likelihood. The standard errors of such an estimator have to be adjusted to take account of the estimation in stages.

For model (3.2), however, estimation in stages does not give consistent estimates of the parameters. If the conditional mean of y_2 is specified as

$$E\left(y_2|x_{i2}\right) = F\left(x_{i2}'\delta\right),$$

with F a CDF, then estimation in stages of the equation

$$y_1 = \exp\left(\gamma F\left(x_{i2}'\delta\right) + x_{i1}'\beta\right)\exp\left(\gamma w_2\right) + u_2$$

leads to biased results due to the fact that the moments of $w_2 = y_2 - F\left(x_{i2}'\beta\right)$ are dependent on x_{i2}, as $E\left(w_{i2}^2|x_{i2}\right) = F\left(x_{i2}'\delta\right)\left(1 - F\left(x_{i2}'\delta\right)\right)$, and so $E\left(\exp\left(w_2\right)\right)$ is a function of the parameters and regressors.

A consistent estimator for (α, β) in model (3.2) is the GMM estimator, and a natural choice of instrument for y_2 is $F\left(x_{i2}'\widehat{\delta}\right)$, where $\widehat{\delta}$ is the logit or probit estimator of δ.

4. Demand for Health Care

In this section we estimate models for the demand for health care in terms of the number of visits to the doctor. A self-reported health index is included as a regressor, which is likely to be endogenous.

The data are taken from the British Health and Lifestyle Survey 1991-1992 (HALS2). This survey is a follow-up from a previous survey in 1984-1985. In the first survey the people interviewed were 18 years and older, and so the minimum age of the people interviewed in the second survey is 25. Of the original sample,

59% were interviewed in the follow-up. The drop out rates between the two samples, due to death, refusal and non-tracing, affect the sampling distribution, and the HALS2 survey cannot be considered to be a representative sample of the adult population of 25 years and older (see HALS2 User Guide). For example, a higher proportion of interviews was achieved in non-manual than in manual groups. However, the age/sex distribution of the HALS2 and the 1991 census data compare reasonably well. The total number of respondents is 5352, but due to missing information in some of the variables, especially the income variable, the sample size for the estimation of the models is reduced to 4814.

The dependent count variable is the number of visits to or by a doctor (general practitioner) in the month before the interview. Variables that are included in the demand equation are sex, age, marital status, education, employment status, income, short term health status, and the self-reported general health index. The binary health index, denoted H_i, is defined as

$$\begin{aligned} H_i &= 1 \quad \text{if health is fair-poor} \\ H_i &= 0 \quad \text{if health is excellent-good.} \end{aligned}$$

Descriptions and summary statistics of the variables are given in Table 2 in the Appendix.

In column (1) of Table 1 we first present Poisson pseudo-likelihood (PL) results for the model which includes the binary general health variable. The coefficient of H_i is 0.38 and is significant with a t-value of 5.78. As mentioned before, it is likely that the self-reported health index has measurement error that is correlated with the number of visits to the doctor, as people who have recently visited a doctor may under-report their general health. In order to test whether the health index is endogenous, we specify the model as

$$\begin{aligned} y_i &= \exp(\alpha H_i + x_i'\beta) + u_i \qquad (4.1) \\ H_i^* &= z_i'\delta + w_i \qquad (4.2) \\ H_i &= 1 \ \ if \ \ H_i^* \geq 0; \ \ H_i = 0 \ \ otherwise. \end{aligned}$$

The z_i are the instruments, which are variables that explain health, but are not likely to determine demand for doctors, other than via health. These variables are alcohol consumption and smoking behaviour, socio-economic status, regional variables, housing variables, work status and long term health indicators. A listing of the instrumental variables is given in the Appendix. Further, z_i also contains x_i.

The equation for health, (4.2), is the reduced form for H_i^*. As argued in section 3, the structural equation of H_i^* cannot be a function of the number of visits to

the doctor for the system to be coherent. This does not seem an unreasonable specification for this particular application. H_i^* can be interpreted as long term health, whereas the number of visits to the doctor is in the last two weeks before interview. If there are any positive idiosyncratic shocks to *short* term health this will be reflected by a higher than average number of visits to the doctor in that period. It may also lead to a perception of a person's own long term health being worse than they would normally have reported. This means that the idiosyncratic shocks u and w in equations (4.1) and (4.2) are correlated, while the number of visits to the doctor does not have a direct effect on long term health per se.

After logit estimation of the reduced form (4.2), a RESET-type test for misspecification and tests for heteroscedasticity did not indicate misspecification. Further, powers (of the few non-dummy variables) and cross products that were considered most likely to contribute to the explanation of health, were not jointly significant. Therefore, the linear reduced form specification for H_i^* does not seem inappropriate.

INSERT TABLE 1 HERE

In column (2) of Table 1, the results of the GMM estimator are presented, using as instruments $\left(F\left(z_i'\widehat{\delta}\right), z_i'\right)$, with F the logistic CDF. The value of the parameter of the health index is now equal to 0.24, a smaller value than before which is expected given the presumed positive correlation of H_i with the number of visits to the doctor. The difference, however, is not statistically significant as indicated by the Hausman test for endogeneity, comparing the estimated coefficient of H_i and its variance in column (2) with those in column (1). This result is due to the large estimated standard error of the coefficient in the GMM model. When the Hausman test is computed comparing all coefficients, the statistic is significant. This result, however, seems to arise from the fact that some standard errors are very close in both models. For example, when the constant is not considered in the test, the Hausman test statistic is insignificant. Further, some standard errors are smaller in the GMM model than in the PL model, giving an indefinite variance matrix for the difference in the parameters. Following an idea in Browning and Meghir (1991), we split the sample randomly into two equal subsamples and compared the estimation results for the PL model based on one subsample with the estimation results of the GMM model for the other subsample. The resulting Hausman test statistics were not significant.

Although the GMM estimator does not give conclusive evidence of the endogeneity of the self-reported health index, due to its relative imprecision, a comparison of the results of the additive specification to those of the multiplicative model (of which the full set of results are not presented here) gives some indication of endogeneity. First of all, we found that the test for overidentifying restrictions

was rejected in the multiplicative model with a p-value of 0.00035, compared to a p-value of 0.0607 in the additive model. From the results of section 2.1, this would indicate that z and u are uncorrelated, but that z and u/μ are correlated, which could occur when u and μ are correlated. Further, the pseudo-likelihood estimated coefficient of H_i in the multiplicative model was 0.4279, whereas the GMM estimated coefficient was *higher* and equal to 0.4975. This result seems implausible given the positive correlation between visits to the doctor and self reported health. The two results together indicate that the z_i seem proper instruments for the additive model, whereas they are not for the multiplicative model.

Next, we proceed by estimating, in stages, the model that includes latent health H_i^* instead of H_i, which is perhaps the most appropriate way of modelling health. The estimation results are presented in column (3) of Table 1. The first-stage parameter vector δ is estimated by the logit model, using the same reduced form (4.2), and the second stage is estimated by Poisson pseudo-likelihood. The coefficient of H^* is equal to 0.11 and is significant. The results are quite similar to those in columns (1) and (2), but overall, the latent health specification adjusts the parameters more for the health effect. The results indicate that males visit the doctor less often than females. Marital status does not have an effect, and the age structure for demand is quadratic with a peak at around 57 years. This seems quite a low age, but this result is primarily driven by the conditioning on short term health (Hlimit), which is positively correlated with age. Higher educated people visit the doctor less frequently than lower educated people, with the exception of the highest educated. Being unemployed does increase the demand, but this effect is not significant. The effect of income is slightly nonlinear, with the higher and lowest income groups having lower demand. The short-term health indicators are the most important in determining the demand for doctors in the last month before the interview.

5. Conclusions

This paper has examined the estimation of count data models with endogenous regressors, using the generalized method of moments (GMM) estimation technique. It has been shown that for model specifications with an additive or a multiplicative error term, the same set of instruments will not, in general, be orthogonal to both error types.

The GMM estimation technique has been used to estimate a model for the explanation of the number of visits to the doctor by individuals. In the model specification, a self-reported health index was likely to be simultaneously determined. For nonlinear models it is often difficult to obtain the reduced form of the endogenous regressor as a simple function of the instruments. However, a

simultaneous model of a count and a binary variable can only be logically consistent when the system is recursive. In the demand for doctors model, the binary health index was instrumented by the estimated probabilities of a reduced form logit model, which resulted in a smaller estimated coefficient when compared to the estimation by Poisson pseudo-likelihood. This difference was, however, not significant.

The model was finally specified in terms of continuous latent health, instead of the binary health index. This specification was estimated by a two stage estimation procedure, using the same reduced form logit estimates to form predictions of health.

A. Formula for Adjusted Standard Errors in Two-Stage Estimation Procedure

Consider the model

$$\begin{aligned} y_i &= \exp(\alpha \widehat{H_i^*} + x_i'\beta) + u_i \\ \widehat{H_i^*} &= z_i'\widehat{\delta}, \end{aligned}$$

where $\widehat{\delta}$ is the logit estimator. The asymptotic variance of the Poisson ML estimator $\widehat{\theta}$, with $\theta = (\alpha, \beta')'$ follows from the asymptotic identity (see Maddala (1983, p 243))

$$\widehat{\theta} - \theta = -\left[E\left(\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right)\right]^{-1}\left[\frac{\partial \log L}{\partial\theta} + E\left(\frac{\partial^2 \log L}{\partial\theta\partial\delta'}\right)\left(\widehat{\delta} - \delta\right)\right],$$

If we then let the variance of the count model be general, the asymptotic variance of $\widehat{\theta}$ is given by

$$\begin{aligned} &(X^{*\prime}MX^*)^{-1}\left[X^{*\prime}\Omega X^* + \alpha^2 X^{*\prime}MZ\left(Z'PZ\right)^{-1}Z'MX^*\right. \\ &\left.-\alpha X^{*\prime}MZ\left(Z'PZ\right)^{-1}Z'WX^* - \alpha X^{*\prime}WZ\left(Z'PZ\right)^{-1}Z'MX^*\right](X^{*\prime}MX^*)^{-1} \end{aligned}$$

where

$$X^* = [Z\delta\ X]$$

$$M = diag(\mu_i) \quad ; \quad P = diag\left(F_i\left(1 - F_i\right)\right)$$

$$\Omega = diag\left(E\left(y_i - \mu_i\right)^2\right) \quad ; \quad W = diag\left(E\left(y_i - \mu_i\right)\left(H_i - F_i\right)\right),$$

and which is estimated by substituting the estimates $\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\delta}$ for α, β, and δ, and by replacing Ω and W by $diag\left(\left(y_i - \widehat{\mu}_i\right)^2\right)$ and $diag\left(\left(y_i - \widehat{\mu}_i\right)\left(H_i - \widehat{F}_i\right)\right)$ respectively.

B. Summary Statistics and Instruments

Summary statistics of the dependent and explanatory variables are given in Table 2.

INSERT TABLE 2 HERE

The instruments used in the GMM estimation are the explanatory variables plus:

Workstatus	- fulltime, parttime, permanently sick or disabled, retired, full time education, home
Social Class	- 7 classes
Accomodation	- house, bungalow, other
Region	- 11 regions
Alcohol	- nondrinking, units taken last week (and squared) of beer, wine and spirits
Smoking	- ever smoked, smoke now
Long term health	- long-standing illness, disability or infirmity
Other	- using oral contraceptive

References

[1] Amemiya T (1985) Advanced Econometrics, Basil Blackwell, Oxford

[2] Blundell R, Smith RJ (1994) Coherency and estimation in simultaneous models with censored or qualitative dependent variables. Journal of Econometrics 64 : 355-373

[3] Browning M, Meghir C (1991) The effects of male and female labor supply on commodity demands. Econometrica 59 : 925-951

[4] Cameron AC, Trivedi PK (1986) Econometric models based on count data: comparisons and applications of some estimators and tests. Journal of Applied Econometrics 1 : 29-53

[5] Cameron AC, Trivedi PK, Milne F, Piggott J (1988) A microeconometric model of the demand for health care and health insurance in Australia. Review of Economic Studies 55 : 85-106

[6] Cameron AC, Windmeijer FAG (1995) R-squared measures for count data regression models with applications to health care utilization. Journal of Business & Economic Statistics, forthcoming

[7] Davidson R, MacKinnon JG (1993) Estimation and Inference in Econometrics, Oxford University Press

[8] Hansen LP (1982) Large sample properties of generalized method of moments estimators. Econometrica 50 : 1029-1054

[9] Health and Lifestyle Survey: Seven Year Follow-Up, 1991-1992 (HALS2), User Guide, ESRC Data Archive, University of Essex, Colchester

[10] Gourieroux C, Monfort A, and Trognon A (1984) Pseudo maximum likelihood methods: applications to Poisson models. Econometrica 52 : 701-720

[11] Grogger J (1990) A simple test for exogeneity in probit, logit and Poisson regression models. Economics Letters 33 : 329-322

[12] Gurmu S, Trivedi PK (1994) Recent developments in models of event counts: a survey. Thomas Jefferson Centre Discussion Paper No. 261, University of Virginia

[13] Maddala GS (1983) Limited-Dependent and Qualitative Variables in Econometrics, Econometric Society Monographs 3, Cambridge University Press

[14] McCullagh P, Nelder JA (1983, 1989) Generalized Linear Models, First and Second Editions, Chapman and Hall, London

[15] Mullahy J (1996), Instrumental variables estimation of Poisson regression models, applications to models of cigarette smoking behavior. Review of Economics and Statistics, forthcoming

[16] Ogaki M (1993) Generalized method of moments: econometric applications. In: GS Maddala, CR Rao, HD Vinod (eds) Econometrics, Handbook of Statistics 11, North-Holland, Amsterdam

[17] 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 : 339-361

[18] Winkelmann R, Zimmermann KF (1995) Recent developments in count data modelling: theory and application. Journal of Economic Surveys 9 : 1-24

[19] Wooldridge J (1992) Some alternatives to the Box-Cox regression model. International Economic Review 33: 935-955

Table 1. Estimation Results

	(1)		(2)		(3)	
Model	PL		GMM		PL two stage	
Var	b	se	b	se	b	se
Const	-2.2752	*0.3297*	-2.5864	*0.3298*	-1.9693	*0.3707*
Male	-0.2748	*0.0590*	-0.2583	*0.0620*	-0.2953	*0.0628*
Single	-0.0104	*0.1040*	-0.0082	*0.1064*	-0.0388	*0.1077*
Age	0.0343	*0.0124*	0.0469	*0.0125*	0.0324	*0.0129*
Age^2	-0.0298	*0.0114*	-0.0418	*0.0112*	-0.0277	*0.0118*
Edu2	-0.2336	*0.1518*	-0.1984	*0.1487*	-0.1864	*0.1564*
Edu3	-0.2146	*0.0853*	-0.2206	*0.0887*	-0.1923	*0.0883*
Edu4	-0.2630	*0.1200*	-0.2974	*0.1250*	-0.2217	*0.1228*
Edu5	0.1477	*0.1008*	0.0759	*0.1056*	0.1718	*0.1057*
Unem	0.1416	*0.1596*	0.0939	*0.1685*	0.1559	*0.1587*
Inc2	0.1089	*0.0644*	0.0859	*0.0646*	0.1254	*0.0649*
Inc3	-0.1894	*0.1080*	-0.2534	*0.1148*	-0.1470	*0.1128*
Inc4	-0.5484	*0.1920*	-0.6924	*0.2137*	-0.5112	*0.1937*
Tempsick	0.5898	*0.1423*	0.6363	*0.1401*	0.5565	*0.1501*
Pregnant	0.9187	*0.2025*	0.8793	*0.2202*	1.0346	*0.2351*
Hlimit2	0.8191	*0.0700*	0.8917	*0.0874*	0.7629	*0.0918*
Hlimit3	1.1313	*0.0895*	1.2176	*0.1232*	1.0462	*0.1281*
Hlimit4	1.5032	*0.1053*	1.6230	*0.1497*	1.3621	*0.1715*
H	0.3832	*0.0662*	0.2448	*0.2030*		
H^*					0.1112	*0.0412*
R^2	0.1910				0.1841	

	Test	DoF	p-value
Hausman endogeneity test	0.5195	1	*0.4711*
Overidentification test	47.61	34	*0.0607*

Notes to Table: The sample size is 4814. The dependent variable is "Visits". PL is the Poisson pseudo-likelihood estimator. GMM is iterated till convergence and is based on the instrument set $\left(F\left(z_i'\delta\right), z_i'\right)$, with F the logistic CDF. The R^2 measure is based on deviance residuals (see Cameron and Windmeijer (1995)). The Hausman test for endogeneity is based on the coefficients of H_i. The test for overidentifying restrictions is the standard GMM χ^2-test. H_i^* is predicted by $z_i'\widehat{\delta}$, with $\widehat{\delta}$ the logit estimator. The standard errors for the two stage estimator are corrected for the estimation in stages (see Appendix A).

Table 2. Summary Statistics

Var	Description	Mean	St dev	Min	Max
Visits	# visits to/by a GP in the month before interview	0.4026	0.8059	0	10
Male	sex = male	0.4348	0.4958	0	1
Age	Age	51.32	15.87	25	96
Age^2	Age×Age/100	28.86	17.22	6.25	92.16
Single	Single	0.0956	0.2940	0	1
Unem	Unemployed	0.0270	0.1621	0	1
Education					
Edu1	CSE Grades 1-5	0.1664	0.3724	0	1
Edu2	GCE 'A' level	0.0357	0.1856	0	1
Edu3	ONC/OND/HNC/HND	0.1388	0.3457	0	1
Edu4	Teacher/Nurse	0.0538	0.2256	0	1
Edu5	Professional/Degree	0.1248	0.3306	0	1
Edu6	Other	0.0372	0.1893	0	1
After tax weekly personal income (£)					
Inc1	(.,100)	0.4508	0.4976	0	1
Inc2	[100,250)	0.3637	0.4811	0	1
Inc3	[250,400)	0.1359	0.3427	0	1
Inc4	[400,.)	0.0496	0.2172	0	1
Short term health					
Pregnant	Pregnant at time of interview	0.0058	0.0761	0	1
Tempsick	Out of work as temporarily sick	0.0042	0.0643	0	1
Hlimit1	Activities in last month not limited by health	0.7154	0.4513	0	1
Hlimit2	a little limited	0.1751	0.3801	0	1
Hlimit3	quite a lot	0.0669	0.2499	0	1
Hlimit4	a great deal	0.0426	0.2019	0	1
Long term health					
H	self-reported health fair/poor	0.2528	0.4347	0	1