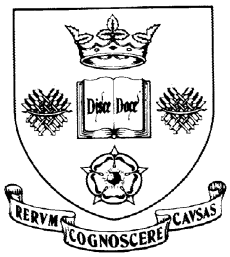
Sheffield Economic Research Paper Series

SERP Number: 2012026

ISSN 1749-8368



Sarah Brown
Mark N. Harris
Jennifer Roberts
Karl Taylor

Modelling Primary Health Care Use: A Panel Zero Inflated Interval Regression Approach

October 2012

Department of Economics
University of Sheffield
9 Mappin Street
Sheffield
S1 4DT
United Kingdom
www.shef.ac.uk/economics

Modelling Primary Health Care Use: A Panel Zero Inflated Interval Regression Approach

Sarah Brown[a], Mark N. Harris[b], Jennifer Roberts[a] and Karl Taylor[a]

[a]Department of Economics, University of Sheffield, United Kingdom
[b]Department of Econometrics and Quantitative Modelling, Curtin University, Australia

Abstract: We introduce the (panel) zero-inflated interval regression (ZIIR) model, to investigate GP visits using individual-level data from the British Household Panel Survey. The ZIIR is particularly suitable for this application as it jointly estimates the probability of visiting the GP and then, conditional on visiting, the frequency of visits (defined by given numerical intervals in the data). The results show that different socio-economic factors influence the probability of visiting the GP and the frequency of visits.

Key Words: GP Visits; Panel Data; Zero-Inflated Interval Regression

JEL Classification: I10; C24; C25

Acknowledgements: We are grateful to Arne Risa Hole and Anita Ratcliffe for excellent comments. We are also grateful to the Data Archive, University of Essex, for supplying the *British Household Panel Surveys*, 1991 to 2008. Funding from the Australian Research Council is kindly acknowledged. The normal disclaimer applies.

October 2012

I. Introduction and Background

Understanding primary health care utilisation is important for policy-makers. To achieve an efficient and equitable healthcare allocation, general practitioner (GP) services should be used in accordance with need. Evidence suggests that other factors, such as socioeconomic status, also influence GP visits. A substantial amount of empirical research has explored GP visits focusing on explaining the number of visits made within a specified time period, typically characterised by a significant proportion of zero observations and a small number of observations indicating frequent visits. As such, count data techniques have been popular in the existing literature. A particular focus relates to whether 'zero' observations reflect non-participants (individuals who never visit a GP) or individuals who are potential, or infrequent, participants (they do visit their GP, but not during the study period). Zero-inflated count models distinguish between these two sources of zeros, treating the cluster at zero as a mixture of these two processes (for example, Freund *et al.*, 1999, Wang, 2003, and Gurmu and Elder, 2008).

In a similar vein, here we introduce the zero-inflated interval regression (ZIIR) model as our data is in the form of grouped counts. This new model is particularly appropriate here, but clearly could be used in a wide range of applications. Common approaches to modelling grouped count data include ordered probit (OP)[1] and zero-inflated ordered probit (ZIOP) models.[2] A ZIOP approach loses information, as would a standard OP, compared to an interval regression (IR) approach with known boundary points. In an IR-based approach, it is possible to estimate the scale of the dependent variable: the latent process underlying the "amount of consumption" has direct quantitative meaning. In a ZIOP, we can discuss partial effects of variables on the probabilities of outcomes (low, medium, high *etc.*), whereas, with

[1] As suggested by Cameron and Trivedi (2005).
[2] A grouped count data model with excess zeros has also been considered by Moffatt and Peters (2000).

the ZIIR, we can estimate partial effects on the expected number of GP visits, thus providing more accurate information to policy-makers.

II. The Zero-Inflated Interval Regression Model

We analyse grouped count data on the frequency of GP visits in an OP-type set-up. As with the ZIOP model of Harris and Zhao (2007), we define an observable random variable *y* that assumes the discrete ordered values of 0, 1,..., *J*, where unlike the former, here these outcomes have direct quantitative meaning. Unlike the OP approach, in the IR-case, due to the known grouping structure, the boundary parameters are fixed (at μ = 1, 3, 6 and 11, see below). As with the ZIOP model, the proposed ZIIR model involves two latent equations: a probit selection equation and an IR one. As with double-hurdle models (Jones, 1989), to observe non-zero "consumption", individuals must overcome two hurdles: whether to participate, and, conditional on participation, how much to "consume".

Let *r* denote a binary variable indicating the split between Regime 0 (*r* = 0 for non-participants) and Regime 1 (*r* = 1 for participants). Although unobservable, *r* is related to a latent variable r^* via the mapping *r* = 1 for $r^* > 0$ and *r* = 0 for $r^* \leq 0$. r^* represents the propensity for participation and is related to a set of explanatory variables $(\mathbf{X}_r)$ with unknown weights β_r, and a standard-normally distributed error term, ε_r:

$$r^* = \mathbf{X}_r'\beta_r + \varepsilon \, . \tag{1}$$

Conditional on *r* = 1, consumption levels under Regime 1 for "participants" are represented by a discrete variable $\tilde{y}$ $(\tilde{y} = 0,1,\dots,J)$ generated by an IR model via a second latent variable $\tilde{y}^*$

$$\tilde{y}^* = \mathbf{X}_y'\beta_y + v, \tag{2}$$

with explanatory variables $(\mathbf{X}_y)$ with unknown weights β_y and a normally distributed error term v, with the standard mapping of:

$$\tilde{y} = \begin{cases} 0 \text{ if } \tilde{y}^* \leq \mu_0, \\ j \text{ if } \mu_{j-1} < \tilde{y}^* \leq \mu_j \text{ , } (j = 1, \ldots, J-1) \\ J \text{ if } \mu_{j-1} \leq \tilde{y}^*, \end{cases} \quad (3)$$

Thus the major difference between the ZIIR and the ZIOP, is that in the former the μ are known and therefore that the scale of *y* can now be identified, σ_v. Neither $\tilde{y}$ nor *r* are directly observed. The observability criterion for observed *y* is

$$y = r \times \tilde{y}. \quad (4)$$

An observed $y = 0$ outcome can arise from two sources: $r = 0$ (the individual is a non-participant); $r = 1$ (the individual is a participant) and jointly that $r = 1$ and $\tilde{y} = 0$ (the individual is a zero-consumption participant). To observe positive *y*, the individual is a participant ($r = 1$) and $\tilde{y}^* > 0$. As the unobservables ε and *v* relate to the same individual, they are likely to be related with covariance $\sigma_{\varepsilon v} = \rho_{\varepsilon v}\sigma_v$. So, on the assumption of joint normality we have:

$$\Pr(y = 0|\mathbf{X}) = [1 - \Phi(\mathbf{X}_r'\beta_r)] + \Phi_2\left(\mathbf{X}_r'\beta_r, \left[\mu_0 - \mathbf{X}_y'\beta_y\right]/\sigma_v; -\rho_{\epsilon v}\right) \quad (5)$$

and

$$\Pr(y = j|\mathbf{X}) = \Phi_2\left(\mathbf{X}_r'\beta_r, \left[\mu_j - \mathbf{X}_y'\beta_y\right]/\sigma_v; -\rho_{\epsilon v}\right) - \Phi_2\left(\mathbf{X}_r'\beta_r, \left[\mu_{j-1} - \mathbf{X}_y'\beta_y\right]/\sigma_v; -\rho_{\epsilon v}\right),$$

$(j = 1, J-1)$

$$\Pr(y = J|\mathbf{X}) = \Phi_2\left(\mathbf{X}_r'\beta_r, \left[\mathbf{X}_y'\beta_y - \mu_{J-1}\right]/\sigma_v; -\rho_{\epsilon v}\right)$$

where $\Phi_2(.,.;\rho)$ represents the standardised bivariate normal distribution, with correlation coefficient, ρ. Thus a zero observation is explicitly allowed to come from one of two sources, and this can account for the observed "excess" build-up of such zeros.

As a further extension, we condition on individual unobserved heterogeneity by including unobserved effects in equations (1) and (2), which are assumed to be normally-distributed with mean zero and covariance matrix $\sum$

$$\sum = \begin{pmatrix} \sigma_r^2 & \sigma_{ry} \\ \sigma_{ry} & \sigma_y^2 \end{pmatrix} \quad (6)$$

This further innovation complicates estimation meaning that each unit's *it* likelihood contributions are no longer independent; and the likelihood for each *i* is the product over T_i. These unobserved effects need to be integrated out of the likelihood function; here undertaken via simulation techniques using Halton sequences of length 50.[3] The simulated log-likelihood function is

$$L_S(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} P_{it} \quad (7)$$

where P_{it} corresponds to the probability of the chosen outcome by individual *i* in period *t* as given by the appropriate element of equation (5). In usual IR, expected values (EVs) are simply given by $\mathbf{X}_y'\beta_y$, thus *ex post* here we consider overall expected values as

$$E(y|\varepsilon) = P(y=0)0 + P(y>0)E(y|y>0) \quad (8)$$
$$= \Phi(\mathbf{X}_r'\beta_r)\left(\mathbf{X}_y'\beta_y + \rho_{\varepsilon v}\sigma_v IMR[\mathbf{X}_r'\beta_r]\right)$$

where $IMR[.]$ is the Inverse Mills Ratio evaluated at its argument; and then $E(y|\varepsilon, y>0)$ is $E(y|\varepsilon)/\Phi(\mathbf{X}_r'\beta_r)$.[4]

III. Data

We use the British Household Panel Survey (BHPS), a survey conducted by the *Institute for Social and Economic Researc*h, 1991 to 2008. We analyse an unbalanced panel of data comprising 51,713 observations focusing on males in England only.[5] Individuals were asked, over the last 12 months, '*approximately how many times have you talked to or visited a GP or family doctor about your own health?*' The possible responses were: none (33%); one or two (38%); three to five (17%); six to ten (7%); or more than ten (5%).

[3] The results were essentially unchanged for a larger number of draws.
[4] These are evaluated at the expected values of both observed and unobserved heterogeneity.
[5] We focus on England only as health system policies have evolved differentially across the different countries of the United Kingdom.

In the probit selection equation, we follow the existing literature and include controls for: aged 18-30 (omitted category), 31-45, 46-60, 61-75 and over 75; married/cohabiting; non-white; highest educational qualification; owner occupier; household size; children in the household aged 0-2, 3-4, 5-11, 12-15 and 16-18; employed/self-employed (omitted category), unemployed and out of the labour force; real household annual gross income[6]; region[7]; urban area; registered disabled; smoker; and self-assessed health (SAH) status, excellent, good, fair and poor (omitted category).[8] For identification, we include two additional variables in the probit part: whether the individual has had dental or eyesight checks in the previous year. With the exception of the dental and eyesight checks, we include the same set of explanatory variables in the IR-part of the model as well as additional controls for: number of hours spent caring for an adult in the household; caring for someone outside the household; use of a car; and weekly hours spent on housework.

IV. Results

Table 1 presents the marginal effects (MEs) associated with EVs of: (i) the unconditional number of GP visits, and; (ii) the number of GP visits conditional on visiting the GP. The final column shows the MEs associated with the probability of non-participation. The overall expected value predicts just over 2 visits to the GP over the last 12 months, with the expected value conditional on participation being slightly higher. In general, the ancillary parameters are strongly statistically significant. In terms of the EVs, the influence of SAH has a large monotonic negative effect on the number of GP visits, *i.e.*, those in worse health visit GPs more frequently. For both types of EVs, smokers visit the GP less frequently than non-smokers. The number of GP visits increases (decreases) monotonically with age (educational

[6] Deflated to 1991 prices.
[7] We control for the eleven standard regions of England.
[8] To allow for the potential endogeneity of SAH, we follow Terza et al. (2008)'s two stage residual inclusion, where the first stage residuals from modelling SAH (as a consistently estimated dynamic random effects OP model) are included as additional regressors in the second stage along with the observed value of SAH.

attainment) relative to those aged 18-30 (those with no education). The role of household size increases the unconditional EV but, once conditioned on visiting the GP, has a negative effect. Out of the additional controls in the IR-part, those men who have the use of a car visit their GP more frequently, with both EVs being of similar magnitude.

Focusing on the MEs associated with the probability of non-participation, older men are more likely not to visit the GP and this effect is monotonically increasing in age. Whilst having children under the age of five has no influence on the EVs, having dependents in this age range is associated with a higher probability of non-participation. For example, those men with children aged between 0-2 years old are 3.67 percentage points (pp) more likely to non-participate. There are clear effects of labour market status on the propensity to visit the GP. Whilst the unemployed are more likely to non-participate in comparison to those individuals who are employed or self-employed, around a 7pp higher probability, the converse is evident for those not in the labour market, approximately a 3.4pp lower probability. Whilst men in excellent/good/fair health visit the GP infrequently compared to those in poor health, such individuals are less likely to non-participate: for excellent health around a 13pp lower probability.[9] There are positive income effects, where a one percent increase in annual income is associated with a 2.68pp higher probability of non-participation. The two identifying variables in the participation (probit) part of the model, indicators for dental and eyesight checks, are both statistically significant and exert negative effects on the probability of non-participation, perhaps signifying that such individuals generally are more likely to engage with health care.[10]

[9] The marginal effects for the first stage residuals are positive and statistically significant throughout, indicating that self-assessed health is an endogenous variable thereby endorsing our two stage residual inclusion approach.
[10] We have also explored specifications with Mundlak fixed effects by including individual level mean variables for all time varying control variables.

V. Conclusion

We have proposed a ZIIR model for instances where there are groupings of data with a build-up of observations at "zero", and applied this to a problem of grouped counts of GP visits. The findings indicate that socio-economic factors have different influences across the two parts of the model, which should be of interest to policy makers concerned with healthcare allocation. Furthermore, it is apparent that the new model is widely applicable to areas where the outcome of interest is in the form of grouped counts.

References

Cameron, C. and Trivedi, P. (2005). *Microeconometrics*, Cambridge University Press.

Freund, D. A., Knieser, T. J. and LoSasso, A. T. (1999). Dealing with the Common Econometric Problems of Count Data with Excess Zeros, Endogenous Treatment Effects, and Attrition Bias. *Economics Letters*, 62, 7-12.

Gurmu, S. and Elder, J. (2008). A Bivariate Zero-Inflated Count Data Regression Model with Unrestricted Correlation. *Economics Letters*, 100, 245-248.

Harris, M. N. and Zhao, X. (2007). A Zero-Inflated Ordered Probit Model, with an Application to Modelling Tobacco Consumption. *Journal of Econometrics*, 141, 1073-1099.

Jones, A. (1989). A Double-Hurdle Model of Cigarette Consumption. *Journal of Applied Econometrics*, 141(2), 1073-1099.

Moffatt, P. and Peters, S. (2000). Grouped Zero-Inflated Count Data Models of Coital Frequency. *Journal of Population Economics*, 13, 205-220.

Terza, J. V., Basu, A. and Rathouz, P. J. (2008). Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling. *Journal of Health Economics*, 27(3), 531-43.

Wang, P. (2003). A Bivariate Zero-Inflated Negative Binomial Regression Model for Count Data with Excess Zeros. *Economics Letters*, 78, 373-378.

TABLE 1: Determinants of the frequency of GP visits and the probability of non-participation

	EXPECTED VALUES[#]				PROBABILITY OF	
	UNCONDITIONAL		CONDITIONAL		NON-PARTICIPATION	
	M.E.s	S.E.s	M.E.s	S.E.s	M.E.s	S.E.s
Intercept	7.7930*	0.1901	9.0750*	0.1552	0.4129*	0.0354
Aged 31-45	0.0868	0.0462	0.0186	0.0512	-0.0315*	0.0068
Aged 46-60	0.1024	0.0543	0.1343*	0.0591	0.0119	0.0065
Aged 61-75	0.5590*	0.0657	0.6209*	0.0699	0.0165*	0.0076
Aged 76+	0.5226*	0.0789	0.6290*	0.0830	0.0366*	0.0088
Married	0.0883*	0.0399	-0.0397	0.0404	-0.0577*	0.0074
Non white	0.5363*	0.1250	0.4201*	0.1298	-0.0610*	0.0164
Degree	0.1868*	0.0745	0.1704*	0.0789	-0.0107	0.0070
A level	0.2418*	0.0551	0.2073*	0.0578	-0.0197*	0.0055
O level	0.1909*	0.0557	0.1302*	0.0585	-0.0302*	0.0060
Own home	-0.0876*	0.0388	-0.1076*	0.0410	-0.0071	0.0038
Household size	0.1472*	0.0245	-0.0515*	0.0201	-0.0898*	0.0076
Children aged 0-2	0.0087	0.0587	0.0929	0.0625	0.0367*	0.0144
Children aged 3-4	-0.0149	0.0609	0.0543	0.0635	0.0306	0.0166
Children aged 5-11	0.0035	0.0498	0.0708	0.0508	0.0294*	0.0130
Children aged 12-15	-0.0589	0.0463	-0.0125	0.0464	0.0214	0.0134
Children aged 16-18	-0.1293	0.0819	-0.0689	0.0829	0.0289	0.0221
Unemployed	0.3264*	0.0679	0.4999*	0.0723	0.0697*	0.0089
Out of the labour market	0.7008*	0.0433	0.6542*	0.0466	-0.0337*	0.0065
Health excellent	-7.9520*	0.1035	-8.6040*	0.0732	-0.1339*	0.0116
Health good	-7.4260*	0.1057	-8.1140*	0.0705	-0.1601*	0.0132
Health fair	-4.3950*	0.0637	-4.7100*	0.0507	-0.0544*	0.0067
Generalised health residuals	1.3490*	0.0314	1.4860*	0.0280	0.0342*	0.0036
Registered disabled	0.1970*	0.0504	0.2778*	0.0522	0.0316*	0.0064
Smoker	-0.1705*	0.0377	-0.1501*	0.0396	0.0121*	0.0041
Live in urban area	0.2384*	0.0416	0.1534*	0.0437	-0.0418*	0.0053
Log income	-0.1597*	0.0196	-0.1054*	0.0209	0.0268*	0.0033
Dental check	0.1649*	0.0192	0.0089	0.0109	-0.0714*	0.0070
Sight check	0.1873*	0.0215	0.0102	0.0125	-0.0811*	0.0076
Number hours caring	0.0689	0.0468	0.0719	0.0489		
Care outside household	0.0057	0.0113	0.0059	0.0118		
Has use of a car	0.1441*	0.0393	0.1504*	0.0409		
Weekly hours housework	-0.1164*	0.0164	-0.1215*	0.0191		
Log likelihood	-67,746.83					
Expected value	2.099 (0.0303)					
Conditional expected value	2.200 (0.0338)					
AIC (BIC)	135,572.66 (136,351.08)					
IR sigma	2.4280 (0.0071)					
Covariance – OP (se)	3.3930 (0.0630)					
Covariance – probit (se)	1.7810 (0.1171)					
Covariance $\sigma_{\varepsilon v}$ (se)	0.1677 (0.0409)					
Correlation $\rho_{\varepsilon v}$ (se)	-0.0272 (0.0333)					
OBSERVATIONS	51,713					

Notes: (i) [#] The marginal effects relate to the actual number of trips; (ii) * denotes statistical significance at the 5 or 1 percent level.