
Article:
Brown, S., Duncan, A., Harris, M.N. et al. (2 more authors) (2015) A Zero Inflated Regression Model for Grouped Data. Oxford Bulletin of Economics and Statistics, 77 (6). pp. 822-831. ISSN 1468-0084

https://doi.org/10.1111/obes.12086

A Zero Inflated Regression Model for Grouped Data[1]

SARAH BROWN[†], ALAN DUNCAN[‡], MARK N. HARRIS[§], JENNIFER ROBERTS[*], and KARL TAYLOR[*]

†Department of Economics, University of Sheffield, 9 Mappin Street, Sheffield. S1 4DT. UK (e-mail: sarah.brown@sheffield.ac.uk)

‡Bankwest Curtin Economics Centre, Curtin University, Perth, Australia

§School of Economics and Finance, Curtin University, Perth, Australia

*Department of Economics, University of Sheffield, Sheffield. UK

Abstract: We introduce the (panel) zero-inflated interval regression (ZIIR) model, which is ideally suited when data are in the form of groups, and there is an 'excess' of zero observations. We apply our new modelling framework to the analysis of visits to the general practitioner (GP) using individual-level data from the British Household Panel Survey. The ZIIR model simultaneously estimates the probability of visiting the GP and 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.

JEL Classification numbers: C33; C35; I12;

Keywords: GP Visits; Grouped Data; Interval Regression; Zero Inflated

November 2014

[1] Funding from the Australian Research Council is kindly acknowledged.

I. Introduction and Background

In this paper, we introduce the zero-inflated interval regression (ZIIR) model which is ideally suited when the variable of interest is grouped in some way and there is an excess of zero observations. The standard approach to modelling grouped data is the interval regression approach (see, for example, Greene and Hensher, 2010), which is based on the ordered probit model but with known boundary parameters. A key advantage of this approach is that it is now possible to identify the scale of the dependent variable (in contrast to the ordered probit approach). However, there are circumstances in which outcomes at the extensive margin may be driven by different processes than those that dictate positive outcomes. Grouped dependent data that exhibit a build-up of 'excess' zeros is one likely manifestation of such a situation. It is therefore necessary to introduce a more flexible parametric specification into the standard interval regression to accommodate such divergent processes in order to avoid the potential for biased and inconsistent estimates. In such a case we propose generalising the interval regression framework along the lines suggested by Harris and Zhao (2007) for ordered dependent variables.

Grouped data are commonly found in surveys where, for example, individuals are asked to provide their responses within particular ranges. This occurs across a wide spectrum of areas such as income bands, the number of general practitioner (GP) or hospital visits, and drug, alcohol and cigarette consumption. And in many of these cases, moreover, there is a strong possibility for the presence of excess zeros. In order to illustrate our modelling framework, we apply the ZIIR approach to the modelling of grouped data on visits to the GP.

II. The Zero-Inflated Interval Regression Model

As with the Zero Inflated Ordered Probit (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 individual level outcomes have direct quantitative meaning.

Unlike the ordered probit approach, in the interval regression case, due to the known grouping structure, the boundary parameters are fixed (at μ = 1, 3, 6 and 11, in the example presented in Section III below). As with the ZIOP model, the proposed ZIIR model involves two latent equations: a binary probit equation and an interval regression (or an ordered probit one, in the former). 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, generally defined) 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:

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

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 interval regression 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:

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

Thus the major difference between the ZIIR and the ZIOP models, 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 $\tilde{y}^* > \mu_0$. As the unobservables ε and v relate to the same individual, we will assume that $\mathrm{E}(\upsilon|\varepsilon) = \rho_{\varepsilon\upsilon}\sigma_\upsilon\varepsilon$, that is they are related with covariance $\sigma_{\varepsilon v} = \rho_{\varepsilon v}\sigma_v$. For ease of notation, let the combined set of explanatory variables $\mathbf{X}$ represent the union of $\mathbf{X}_r$ and $\mathbf{X}_y$. Then, on the assumption of joint normality, we have that:

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

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, when panel data are available, we can condition on individual unobserved heterogeneity by including (the usual additive and time-invariant) unobserved effects in equations (1) and (2), call these α_{ir} and α_{iy} respectively, 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.[2] Collecting all parameters of the model together in θ, the simulated log-likelihood function is

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

where $P_{it,m}$ corresponds to the probability of the chosen outcome by individual i in period t as given by the appropriate element of equation (5). Note that in order to integrate the unobserved effects out of the likelihood function, the $P_{it,m}$ probabilities are a function of the m^{th} draw $(m = 1, \ldots, M = 50)$ of the jointly normally distributed variates with mean zero and covariance matrix Σ. That is, these probabilities are a function of both of $\alpha_{ir,m}$ and $\alpha_{iy,m}$, where these respectively enter into $P_{it,m}$ via equations (1) and (2).

It is useful to summarise here explicitly how this approach, and therefore likelihood function, differs from that of Harris and Zhao (2007), upon which the current approach is based. Firstly, as we have panel data, unlike Harris and Zhao (2007), we can readily condition on the (likely) unobserved heterogeneity in both equations $(\alpha_{ir}, \alpha_{iy})$, and their correlation. These accordingly need to be integrated out of the likelihood function, and hence the need for simulated maximum likelihood (clearly this would not be required if only cross-sectional data were available). Also, due to the cardinal nature of the dependent variable here, the scale of this is meaningful, such that we can now estimate σ_v whilst fixing the boundary parameters at their known values (which are parameters to be estimated in Harris and Zhao, 2007).

In the usual interval regression, expected values (EVs) are simply given by $\mathbf{X}_y'\beta_y$. For our zero-inflated interval regression we consider two sets of expected values to be of interest,

[2] The results discussed in Section III were essentially unchanged for a larger number of draws.

each of which is analogous to measures typically reported for censored or incidentally truncated regressions. First, the conditional expectation $E\left(\tilde{y}^* r \middle| \mathbf{X}, y>0\right)$ provides a measure of the expected value of $\tilde{y}^*$ for positive observed y only. This requires both that $r=1$ (which characterises the individual as a 'participant') and that $y>0$ (which denotes positive 'consumption'). Noting that $p_y = \Pr(y>0|\mathbf{X})$ is the complement of equation (5), and defining $r_h = (1-\rho_{\varepsilon v}^2)^{-1/2}$, we have that

$$\begin{aligned}
E\left(\tilde{y}^* r \middle| \mathbf{X}, y>0\right) &= E\left(\tilde{y}^* \middle| \mathbf{X}, r=1, y>0\right) \\
&= E\left(\tilde{y}^* \middle| \mathbf{X}, r^*>0, \tilde{y}^*>\mu_0\right) \\
&= E\left(\tilde{y}^* \middle| \varepsilon > -\mathbf{X}_r'\beta_r, 0, v>\mu_0 - \mathbf{X}_y'\beta_y\right) \\
&= \mathbf{X}_y'\beta_y + E\left(v \middle| \varepsilon > -\mathbf{X}_r'\beta_r, 0, v>\mu_0 - \mathbf{X}_y'\beta_y\right) \\
&= \mathbf{X}_y'\beta_y + \rho_{\varepsilon v}\sigma_v \phi\left(-\mathbf{X}_r'\beta_r\right)\Phi\left(r_h[\mathbf{X}_y'\beta_y - \mu_0]/\sigma_v - r_h\rho_{\varepsilon v}\mathbf{X}_r'\beta_r\right)/p_y \\
&\quad + \sigma_v \phi\left(\left[\mu_0 - \mathbf{X}_y'\beta_y\right]/\sigma_v\right)\Phi\left(r_h\mathbf{X}_r'\beta_r + r_h\rho_{\varepsilon v}[\mu_0 - \mathbf{X}_y'\beta_y]/\sigma_v\right)/p_y .
\end{aligned} \tag{8}$$

The second relevant expected value is an unconditional expectation of $\tilde{y}^*$ that takes account of the censoring of observed y at zero. Defining $c_y = \mathbf{1}(y>0)$ where $\mathbf{1}(.)$ is the indicator function, the censored expected value $E\left(c_y \tilde{y}^* \middle| \mathbf{X}\right)$ can be derived as

$$\begin{aligned}
E\left(c_y\tilde{y}^* \middle| \mathbf{X}\right) &= \Pr(y=0|\mathbf{X})\,E\left(c_y\tilde{y}^* \middle| \mathbf{X}, y=0\right) + \Pr(y>0|\mathbf{X})\,E\left(c_y\tilde{y}^* \middle| \mathbf{X}, y>0\right) \\
&= p_y E\left(c_y\tilde{y}^* \middle| \mathbf{X}, y>0\right) \equiv p_y E\left(\tilde{y}^* r \middle| \mathbf{X}, y>0\right).
\end{aligned} \tag{9}$$

This is equivalent to equation (8) scaled by the (bivariate) probability $p_y = \Pr(y>0|\mathbf{X})$.[3] The importance of both expected values (8) and (9) is that each inherits the scale of the underlying

[3] Both expressions are most easily evaluated at the expected values of both observed and unobserved heterogeneity components.

measure of interest, $\tilde{y}^*$, rather than the grouped index y. This provides us with meaningful interpretations when calculating the marginal effects corresponding to each expected value.

As with any such multi-equation model specification, issues of identification naturally arise for the ZIIR model. Akin to the related Heckman model for mixed discrete-continuous choice, the ZIIR may be estimated under circumstances where $\mathbf{X}_r \equiv \mathbf{X}_y$ due to the nonlinearities in each component of the model specification.[4] However, nonparametric identification requires at least one exclusion restriction to be imposed on the parameters of the 'participation' equation (see, for example, Wooldridge, 2010).

III. Application

A substantial amount of empirical research has explored GP visits focusing on explaining the number of visits made by individuals 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). Although in our illustration we have data on grouped counts, as Cameron and Trivedi (2005), p.682, point out 'count data can be modelled by discrete choice model methods, possibly after some grouping of counts ... a sequential model that recognises the ordering of the data should be used .... o(O)ne such model is an ordered model.'[5]

Data

[4] Although this is not so if the two equations are not treated as independent.
[5] A grouped count data model with excess zeros has been considered Moffatt and Peters (2000).

To illustrate the ZIIR model we utilise data drawn from the British Household Panel Survey (BHPS), a survey conducted by the Institute for Social and Economic Research. Specifically, we analyse an unbalanced panel of 51,713 observations from the BHPS on the number of GP visits made by males in England between 1991 and 2008.[6] Individuals were asked, over the last 12 months, 'how many times have you talked to or visited a GP or family doctor about your own health?'. Possible responses to this question (with unconditional averages in parentheses) were: none (33%); one or two (38%); three to five (17%); six to ten (7%); or more than ten (5%). These answers serve to fix the boundary parameters as discussed in Section II above.

In the probit equation, we follow existing literature and include controls for: aged 18-30 (the 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 (the omitted category), unemployed and out of the labour force; real household annual gross income;[7] region;[8] urban area; registered disabled; smoker; and self-assessed health (SAH) status, excellent, good, fair and poor (the omitted category). For identification, we include two additional variables in the probit component: whether the individual has had dental or eyesight checks in the previous year. Our justification is that the initial participation decision is influenced by the individual's general attitudes towards health related behaviours which are reflected in their propensity to undergo regular elective health screening.[9]

[6] As is common in the health economics literature, we split our analysis by gender. For brevity, here we focus on males. Additionally, we focus on England only as health system policies have evolved differentially across the different countries of the United Kingdom.

[7] Deflated to 1991 prices.

[8] We control for the eleven standard regions of England.

[9] Note that, although the primary rationale for introducing a probit equation into the interval regression is to build a more flexible specification to deal with excess zeros, the parameters of the participation equation are likely to be of interest in their own right in capturing potential drivers of visits to the GP. We are grateful to an anonymous referee for highlighting this important point.

With the exception of the dental and eyesight checks, we include the same set of explanatory variables in the interval regression part of the model as well as additional controls for the number of hours spent caring for an adult in the household, whether or not they care for someone outside the household, whether they have use of a car, and their weekly hours spent on housework. The assumption here is that the frequency of visits is determined by the availability of time and ease of travel to the GP.

Results

Table 1 presents the marginal effects associated with the expected values of: (i) the unconditional number of GP visits, and; (ii) the number of GP visits conditional on visiting the GP, where the marginal effects relate to the actual number of GP visits.[10] The final column shows the marginal effects associated with the probability of non-participation. The overall expected value predicts 2.3 visits to the GP over the last 12 months, with the expected value conditional on participation being higher at almost 3.5 visits.

Turning first to the marginal effects associated with the probability of non-participation, we see a clear age effect on participation, with older men far less likely to visit their GP than younger age cohorts: moreover, this effect is steeply increasing in age. Married men and non-white men are both around 5 percentage points less likely to be non-participants. Men with formal high school qualifications (either at O-level or A-level) are less likely to visit their GP than those with other educational qualifications. There are also clear differences in participation by labour market status. The unemployed are significantly less likely to participate in comparison to those individuals who are employed or self-employed (by around 6 percentage points on average). The converse is evident for those not in the labour market, who have a 3 percentage point higher probability of participation than the employed

[10] Expressions for both expected values are derived in equations (8) and (9), with each set of marginal effects evaluated at the means of observed and unobserved heterogeneity.

reference category. Whilst men in excellent/good/fair health visit the GP less frequently than those in poor health, they are also more likely to participate (for excellent health, with around an 11 percentage point higher probability).[11] Positive income effects are evident in the probability of non-participation, with a one per cent increase in annual income reducing the chance of visiting a GP by around 2.3 percentage points. Smokers are around 1 percentage point more likely to be a non-participant. 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. These findings perhaps signify that such individuals are generally more likely to engage with health care professionals.[12,13]

Turning now to the first two columns in Table 1, we look at the influence of the explanatory variables on both the unconditional and conditional frequency of GP visits. We have seen above that older men are more likely to be non-participants. However, they also have a higher expected number of visits - the oldest age group have 0.43 more visits on average per year compared to the youngest age group in the unconditional expectation (0.31 more visits in the conditional expectation). Similarly education exerts a positive effect on the frequency of visits. The role of household size and being married increases the unconditional expected value but, once conditioned on visiting the GP, household size has a negative effect and marital status is insignificant. Being unemployed or out of the labour market are both associated with a higher expected number of visits.

[11] 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 ordered probit model) are included as additional regressors in the second stage along with the observed value of SAH. 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.

[12] We have also explored specifications with Mundlak fixed effects by including individual level mean variables for all time varying control variables.

[13] Note that although dental and eyesight checks are only included in the binary probit equation for GP visits, all variables in the model have a direct and/or an indirect effect on the expected values as can be seen from equations (8) and (9).

For both types of expected values, smokers visit the GP less frequently than non-smokers. Out of the additional controls in the interval regression part, those men who have the use of a car, and thus can travel more easily, visit their GP more frequently and those who do more housework, and thus have less free time, visit less frequently. Both findings are in line with our justification for including these variables.[14] Quantitatively the most important determinant of the number of GP visits is, somewhat unsurprisingly, SAH, which has a monotonic effect; someone with excellent health has, on average, almost seven fewer visits (in the unconditional expectation) that someone in poor health.

In order to compare our results to a more 'naïve' estimator with no flexibility at the extensive margin, we report the results of a standard interval regression in Table 2; a number of key differences emerge. The standard approach suggests that men aged 31-45 visit the GP less than younger men, whereas our model reveals that this age group are in fact less likely to be a non-participant with no significant effect on frequency of visits. Household size and living in an urban area have no effect in the standard model; in our model they both impact negatively on the probability of being a non-participant, and have significant effects on both the unconditional and conditional frequency of GP visits. Finally, weekly hours of housework have a positive effect in the standard interval model and a negative effect in the extended framework.

IV. 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 data on GP visits. The findings from this flexible statistical framework indicate that socio-economic factors have different influences across the two parts of the model, which potentially provides accurate

[14] Caring responsibilities are not statistically significant but these behaviours are not very prevalent in our sample with only 8% of men providing care for another adult.

information to policy-makers concerned with healthcare allocation. Furthermore, this new model is widely applicable to areas where the outcome of interest is grouped.

References

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

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, Vol. 62, pp. 7-12.

Greene, W. and Hensher, D. (2010). Modelling Ordered Choices: A Primer, Cambridge University Press, Cambridge, UK.

Gurmu, S. and Elder, J. (2008). 'A Bivariate Zero-Inflated Count Data Regression Model with Unrestricted Correlation', Economics Letters, Vol. 100, pp. 245-8.

Harris, M. N. and Zhao, X. (2007). 'A Zero-Inflated Ordered Probit Model, with an Application to Modelling Tobacco Consumption', Journal of Econometrics, Vol. 141, pp. 1073-99.

Jones, A. (1989). 'A Double-Hurdle Model of Cigarette Consumption', Journal of Applied Econometrics, Vol. 141(2), pp. 1073-99.

Moffatt, P. and Peters, S. (2000). 'Grouped Zero-Inflated Count Data Models of Coital Frequency', Journal of Population Economics, Vol. 13, pp. 205-20.

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, Vol. 27(3), pp. 531-43.

Wang, P. (2003). 'A Bivariate Zero-Inflated Negative Binomial Regression Model for Count Data with Excess Zeros', Economics Letters, Vol. 78, pp. 373-8.

Wooldridge, J. (2010) Econometric Analysis of Cross Section and Panel Data, 2e, MIT Press, Cambridge, MA.

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	6.5540*	0.1589	4.4500*	0.0857	0.3455*	0.0328
Aged 31-45	0.0667	0.0387	0.0018	0.0245	-0.0265*	0.0060
Aged 46-60	0.0865	0.0461	0.0664*	0.0288	0.0098	0.0056
Aged 61-75	0.4718*	0.0553	0.3040*	0.0334	0.0136*	0.0066
Aged 76+	0.4389*	0.0669	0.3088*	0.0400	0.0306*	0.0077
Married	0.0867*	0.0344	-0.0183	0.0199	-0.0487*	0.0067
Non white	0.4599*	0.1081	0.2030*	0.0642	-0.0512*	0.0144
Degree	0.1596*	0.0638	0.0830*	0.0389	-0.0091	0.0061
A level	0.2075*	0.0472	0.1008*	0.0284	-0.0167*	0.0048
O level	0.1654*	0.0478	0.0627*	0.0288	-0.0255*	0.0053
Own home	-0.0710*	0.0331	-0.0522*	0.0201	-0.0065*	0.0033
Household size	0.1486*	0.0223	-0.0217*	0.0097	-0.0768*	0.0072
Children aged 0-2	0.0039	0.0497	0.0469	0.0301	0.0308*	0.0126
Children aged 3-4	-0.0156	0.0532	0.0277	0.0317	0.0256	0.0144
Children aged 5-11	0.0002	0.0431	0.0358	0.0248	0.0246*	0.0114
Children aged 12-15	-0.0521	0.0424	-0.0054	0.0251	0.0180	0.0117
Children aged 16-18	-0.1122	0.0709	-0.0325	0.0409	0.0242	0.0194
Unemployed	0.2694*	0.0575	0.2470*	0.0352	0.0584*	0.0080
Out of the labour market	0.5970*	0.0370	0.3183*	0.0229	-0.0286*	0.0058
Health excellent	-6.7130*	0.0749	-4.2080*	0.0380	-0.1124*	0.0106
Health good	-6.2650*	0.0776	-3.9700*	0.0364	-0.1345*	0.0123
Health fair	-3.7110*	0.0478	-2.3030*	0.0257	-0.0459*	0.0061
Generalised health residuals	1.1360*	0.0249	0.7267*	0.0133	0.0290*	0.0033
Registered disabled	0.1635*	0.0429	0.1368*	0.0262	0.0264*	0.0056
Smoker	-0.1446*	0.0322	-0.0727*	0.0195	0.0099*	0.0035
Live in urban area	0.2084*	0.0358	0.0740*	0.0215	-0.0356*	0.0048
Log income	-0.1375*	0.0170	-0.0500*	0.0102	0.0227*	0.0030
Dental check	0.1474*	0.0177	0.0017	0.0499	-0.0602*	0.0064
Sight check	0.1670*	0.0197	0.0019	0.0056	-0.0682*	0.0070
Number hours caring	0.0053	0.0097	0.0032	0.0059		
Care outside household	0.0584	0.0400	0.0352	0.0239		
Has use of a car	0.1169*	0.0333	0.0704*	0.0200		
Weekly hours housework	-0.0940*	0.0114	-0.0566*	0.0091		
Log likelihood	-67,531.84					
Expected value	2.323 (0.0254)					
Conditional expected value	3.461 (0.0161)					
AIC (BIC)	135,142.69 (135,921.11)					
IR sigma	2.4280 (0.0071)					
Covariance – OP (se)	3.3950 (0.0631)					
Covariance – probit (se)	1.7940 (0.1208)					
Covariance $\sigma_{\varepsilon v}$ (se)	0.1612 (0.0426)					
Correlation $\rho_{\varepsilon v}$ (se)	-0.0118 (0.0342)					
OBSERVATIONS	51,713					

Notes: * denotes statistical significance at the 5 or 1 percent level; regional dummy variables not reported.

TABLE 2

Determinants of the frequency of GP visits – Interval regression

	M.E.s	S.E.s
Intercept	8.7220*	0.1269
Aged 31-45	-0.1385*	0.0492
Aged 46-60	0.1012	0.0522
Aged 61-75	0.5958*	0.0648
Aged 76+	0.5569*	0.0785
Married	0.2583*	0.0387
Non white	0.4655*	0.0821
Degree	0.1413*	0.0577
A level	0.2320*	0.0434
O level	0.1288*	0.0418
Own home	-0.2223*	0.0372
Household size	0.0041	0.0190
Children aged 0-2	0.0480	0.0685
Children aged 3-4	0.0212	0.0703
Children aged 5-11	0.0464	0.0547
Children aged 12-15	0.0089	0.0568
Children aged 16-18	-0.0929	0.0911
Unemployed	0.4209*	0.0695
Out of the labour market	0.6857*	0.0481
Health excellent	-8.3070*	0.0632
Health good	-7.8990*	0.0638
Health fair	-4.5600*	0.0524
Registered disabled	0.2314*	0.0636
Smoker	-0.3099*	0.0352
Live in urban area	0.0337	0.0360
Log income	0.1098*	0.0211
Number hours caring	-0.0094	0.0121
Care outside household	0.1142*	0.0515
Has use of a car	0.2238*	0.0398
Weekly hours housework	0.0920*	0.0139
Log likelihood	-68,596.21	
AIC (BIC)	137,231.42 (137,615.71)	
OBSERVATIONS	51,713	

Notes: * denotes statistical significance at the 5 or 1 percent level.