A Bayesian Two-stage Regression Model for Zero-inflated Longitudinal Outcomes

Technical Report No. ASU/2016/5
Dated: 17 October, 2016

Prajamitra Bhuyan
Applied Statistics Unit,
Indian Statistical Institute,
Kolkata
bhuyan.prajamitra@gmail.com

Jayabrata Biswas
Interdisciplinary Statistical Research Unit,
Indian Statistical Institute,
Kolkata

Pulak Ghosh
Department of Quantitative Methods and Information Sciences,
Indian Institute of Management,
Bangalore

Kiranmoy Das
Interdisciplinary Statistical Research Unit,
Indian Statistical Institute,
Kolkata

A Bayesian two-stage regression model for zero-inflated longitudinal outcomes

Prajamitra Bhuyan [1], Jayabrata Biswas [2], Pulak Ghosh [3], and Kiranmoy Das [2]

[1] Applied Statistics Unit, Indian Statistical Institute, Kolkata, India

[2] Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, India

[3] Department of Quantitative Methods and Information Sciences, Indian Institute of Management, Bangalore, India

Address for correspondence: Kiranmoy Das, Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, India-700108.

E-mail: kmd@isical.ac.in.

Phone: +91-8017354328.

Fax: .

Abstract: Two-stage regression methods are quite popular for analysing the simultaneous equations models which are used often in economics and social sciences. However, the methods are challenging when response and/or the endogenous covariate(s) contain excess zeros. We propose a Bayesian approach for the joint modelling of a zero-inflated longitudinal continuous response and a zero-inflated count endogenous covariate. We define latent continuous variables for handling the excess zeros and develop a Gibbs sampler for the simultaneous estimation of the model parameters for both the models. We consider partially varying coefficients regression models containing covariates with time-varying and time invariant effects on the response. We analyse data from the health and retirement study for modelling the out-of-pocket medical cost with the number of hospital visits as the endogenous covariate. Simulation studies are performed for assessing the usefulness of the proposed method compared to its competitors. The proposed method will be useful in biomedical research, economic research and studies related to the social sciences.

Key words: Gibbs sampling, Latent Variable, Longitudinal data, Two-stage regression, Zero-inflated outcomes.

1 Introduction

Two-stage regression techniques, typically based on the two-stage least squares, are commonly used for analysing the structural equations where the errors of the dependent variables are correlated with the independent variables. In the simultaneous equations models, we analyse the relationship between two sets of "somewhat different" variables, for example, the exogenous and the endogenous variables in an econometric model. Scott and Holt (1982) studied the effect of a two-stage sampling on the ordinary least squares method. Angrist and Imbens (1995) used two-stage regression technique for estimating the average causal effect of the variable treatments. Holliday et al. (1998) proposed similar two-stage regression model for engine mapping experiment. Khan and Lewbel (2006) used two-stage regression method for semi-parametric truncated regression models. For jointly modelling the longitudinal prostate specific antigen and the recurrence of prostate cancer, Ye et al. (2008) proposed a two-stage estimation method in a semi-parametric framework. In economics and social science literature two-stage regression methods are used quite often. Simar and Wilson (2011) used two-stage estimation method for data envelopment analysis. Dong and Lewbel (2015) proposed models for analysing binary response with endogenous regressors. Two-stage least squares estimation of spatial models with endogenous regressor and instruments has been proposed by Liu and Lee (2011).

In this paper, we consider two-stage regression model for longitudinal data where the response and the endogenous predictor contain a weighty proportion of zeros. In recent years, methods based on the mixture models have been used for analysing longitudinal data with “excess zeros”. Zero Inflated Poisson (ZIP) model was proposed by Lambert (1992) for handling the excess zeros in count data. Hall (2000) developed zero-inflated Binomial models for upper bounded count data. Tobit model (Amemiya 1984), Hurdle model (Mullahy 1986), two-part model (Duan et al. 1983), finite mixture model (Aitkin and Rubin 1985) are the commonly used models for analysing zero-inflated longitudinal outcomes. In the context of joint modelling, Ghosh and Tu (2008) developed models for handling excess zeros in longitudinal outcomes. Min and Agresti (2002) gives a nice tutorial on various methods for handling zero-inflated data in different contexts. Although there is a vast literature on it, handling the excess zeros in the context of longitudinal outcomes is still an issue, specifically when one or more endogenous covariates contain excess zeros.

We develop a Bayesian approach of analysing the continuous longitudinal zero-inflated response with the covariates having time-varying effects and time-invariant effects on the response. We consider a count endogenous covariate containing excess zeros. We define continuous latent variables for the zero-inflated variables and develop two-stage regression method on the latent variables. The time-varying effects of the covariates are modelled non-parametrically using the orthogonal Legendre Polynomials (LP). Suitable spline functions can also be used but we prefer LP basis for the ease of computation. A simple Gibbs sampler is developed following Albert and Chib (1993) where in each iteration, we sample the model parameters as well as the latent variables. Our method is simple because it is based on Gibbs sampler, and it is fast since we estimate the parameters for both the models simultaneously in an automated manner.

Our work is motivated by a sample dataset from the Health and Retirement Study (HRS) conducted by the University of Michigan for an aging population. The study started in 1992 and ended in 2012 with the subjects (over the age of 50 years) who were either retired or about to retire in 1992. These people were followed for 20 years with a follow-up frequency of 2 years. The financial status, disease status, health insurance status, total medical cost, number of hospital visits, body mass index (BMI) etc. were measured for the subjects over time with the goal of deciphering the challenges of aging in terms of the health condition and medical cost. The world population is aging and it is expected that the growth of the older people will have a significant impact on the world economy. In the United States, life expectancy has increased from 49 years to 78 years in the last century (Arias, 2006). In China it has changed from 40 years to 76 years in the last few decades. Similar pictures are observed even for the third world countries.

Our goal is to model the out-of-pocket medical expenditure (OOPME) which includes all the medical costs excluding those which are paid through or reimbursed by the health insurance. Obviously, OOPME depends on the financial status, disease status, number of hospital visits and the insurance status of the subjects over the years. However, there is an inherent endogeneity since the number of hospital visits also depends on the financial status and the disease status of the subjects. An aged individual from an affluent family visits hospital even when the condition is not that serious. On the other hand, an aged person with poor financial status is forced to visit hospital under critical health condition. We analyse the dataset in a two-stage regression framework where in the first stage we model OOPME in terms of the other predictors, and in the second stage we model the number of hospital visits using the relevant predictors. In each stage, we have some covariates with time-varying effects on the response and the others having time-invariant effects. Thus, our model at each stage becomes a partially varying coefficients linear model (Senturk et al. 2013).

The rest of the paper is organized as the following. In Section 2, we discuss our proposed model and the Bayesian estimation method in detail. The non-parametric approach of modelling the time-varying effects of the covariates using the Legendre Polynomials is also discussed in this section. We analyse the HRS data and report the inference in Section 3. Simulation studies are performed for assessing the effectiveness of the proposed approach compared to the other methods and the results are summarized in Section 4. Some concluding remarks and limitations of the proposed method are given in Section 5.

2 Proposed Model and Methods

In the following presentation, we consider a continuous response measured over T different evenly spaced time points from n subjects. We consider a set of covariates, some of which possibly have time-varying effects on the response. The response for the i-th subject at the t-th time point, which we denote by $Y_i(t)$, can thus be modelled as the following:

$$Y_i(t) = \sum_{j=1}^{J} \beta_j(t) X_{ji}(t) + \sum_{j'=1}^{J'} \gamma_{j'} Z_{j'i}(t) + u_i + e_i(t), \tag{2.1}$$

where we consider J covariates with time-varying effects on the response, and J' denotes the number of covariates with time-invariant effects on the response. Subject-specific random effects u_is capture the longitudinal dependence and are assumed to be iid N(0,σ_u^2). The residuals $e_i(t)$s are assumed to be iid $N(0, \sigma_e^2)$. Without loss of generality, let us assume that the count covariate Z_1 is endogenous and thus $cov(Z_1, e) \neq 0$. Hence, we consider the following regression model for $Z_{1i}(t)$ with some additional covariates:

$$g(\mu_i(t)) = \sum_{k=1}^{K} \theta_k(t) W_{ki}(t) + \sum_{k'=1}^{K'} \delta_{k'} S_{k'i}(t) + v_i, \tag{2.2}$$

where g denotes the appropriate link function and $\mu_i(t) = E\left(Z_{1i}(t) | W_i(t), S_i(t)\right)$. The random effects v_is are assumed to be iid $N(0, \sigma_v^2)$. We note that the sets X and W; and Z and S can be identical, although in our application, they are not identical.

In the usual two-stage regression problem, one first fits the model given in equation (2.2) and then fits the model given in equation (2.1) with $Z_{1i}(t)$ replaced by the estimated $\mu_i(t)$. However, when both the response and the endogenous predictor are zero-inflated, the models in equation (2.1) and (2.2) will give very poor fit. Such data are not so uncommon, specially in biomedical studies and social sciences. We consider Tobit models for modelling the excess zeros using latent variables as the following:

$$Y_i(t) = \begin{cases} Y_i^*(t), & \text{for } Y_i^*(t) > 0; \\ 0, & \text{for } Y_i^*(t) \leq 0, \end{cases}$$

where $Y_i^*(t)$ is a latent random variable. The endogenous covariate $Z_1(t)$ is a count variable taking values $0, 1, \ldots, C$, and we define a continuous (normal) latent random variable $Z_1^*(t)$ as the following:

$Z_{1i}(t) = 0, \quad \text{if } Z_{1i}^*(t) < B_0 = 0,$

and $Z_{1i}(t) = c, \quad \text{if } B_{c-1} < Z_{1i}^*(t) < B_c$, for $c > 0$, where we consider the bin boundaries $B_1, \ldots, B_C$ unknown.

We then rewrite the two-stage regression models given in equations (2.1) and (2.2) in terms of the latent random variables as the following:

$$Y_i^*(t) = \sum_{j=1}^{J} \beta_j(t) X_{ji}(t) + \gamma_1 Z_{1i}^*(t) + \sum_{j'=2}^{J'} \gamma_{j'} Z_{j'i}(t) + u_i + e_i(t), \tag{2.3}$$

and

$$Z_{1i}^*(t) = \sum_{k=1}^{K} \theta_k(t) W_{ki}(t) + \sum_{k'=1}^{K'} \delta_{k'} S_{k'i}(t) + v_i + \epsilon_i(t), \tag{2.4}$$

where the residuals $\epsilon_i(t)$ are iid $\mathrm{N}(0, \sigma_\epsilon^2)$.

For the above models, we can use the usual two-stage regression but this requires values of the latent variables at each step. We propose a Bayesian estimation method for simultaneous estimation of the model parameters using Gibbs sampling. We note that for more than one zero-inflated covariates, one can simply extend the above model and define a latent covariate for each of the zero-inflated covariate.

2.1 Modelling the time-varying coefficients using basis functions

The time-varying coefficients $\beta_j(t)$ and $\theta_k(t)$ in equations (2.3) and (2.4) respectively are to be modelled appropriately for meaningful inference on the effects of the respective covariates. Note that parametric modelling of these coefficients is less appealing since the effects are usually not known in advance. We consider non-parametric approach of modelling the time-varying coefficients using the basis functions. There is a rich literature in statistics on such modelling using Fourier basis, wavelets, B-splines, P-splines etc. Here, we consider the orthogonal Legendre Polynomials (LP) as our basis functions. These Polynomials have already been proven as powerful tool by several authors for non-parametric regression (Das et al. 2011, Cui et al. 2008, Meyer 2000, Huskova and Sen 1985).

The general form of a Legendre Polynomial of order r is given by the following sum

$$P_r(x) = \sum_{l=0}^{L} (-1)^l \frac{(2r-2l)!}{2^r l!(r-l)!(r-2l)!} x^{r-2l}, \tag{2.5}$$

where $L{=}\frac{r}{2}$ or $\frac{r-1}{2}$, whichever is an integer. These polynomials are defined over [-1, 1] and are orthogonal to each other in this interval in the sense that $\int_{-1}^{1} P_r(x)P_s(x)dx = 0$, for $r \neq s$. First few LPs are as the following:
$P_0(x) = 1$, $P_1(x) = x$, $P_2(x) = \frac{1}{2}(3x^2 - 1)$, $P_3(x) = \frac{1}{2}(5x^3 - 3x)$. We take these polynomials as the basis functions.

In our context, we denote the r-th order Legendre Polynomial (LP) at time t by $P_r(t)$. We transform the original time points t to the adjusted time points t' for fitting the orthogonal LP range [-1, 1]. This is done by defining $t^{'} = -1 + 2(\frac{t-t_{min}}{t_{max}-t_{min}})$, t_{min} and t_{max} are the smallest and the highest time points respectively. Let $P^{(r)}(t) =$

$[P_0(t), P_1(t), \ldots, P_r(t)]^T$ denote the family of the first $r+1$ basis functions and express the functions $\beta_j(t)$ and $\theta_k(t)$ as some linear combinations of these basis functions:

$$\beta_j(t) = \mathbf{a}_j^T P^{(r_1)}(t); \;\; \theta_k(t) = \mathbf{b}_k^T P^{(r_2)}(t), \tag{2.6}$$

where $\mathbf{a}_j = (a_{j0}, a_{j1}, \ldots, a_{jr_1})^T$, and $\mathbf{b}_k = (b_{k0}, b_{k1}, \ldots, b_{kr_2})^T$ are called base vectors. The optimal orders r_1 and r_2 are chosen by the information criteria, e.g. AIC/BIC etc. By using the LPs as the basis function, we avoid the computational complexities (e.g., knot selection, knot location) related to the other basis functions.

2.2 Bayesian Estimation using Gibbs Sampler

We employ a Bayesian approach of estimating the model parameters for the equations (2.3) and (2.4) using the Gibbs sampler. Let Θ_1 and Θ_2 denote the set of all the model parameters for the equations (2.3) and (2.4) respectively. Thus, we have $\Theta_1 = [\mathbf{a}, \boldsymbol{\gamma}, \sigma_u^2, \sigma_e^2]$; and $\Theta_2 = [\mathbf{b}, \boldsymbol{\delta}, \sigma_v^2, \sigma_\epsilon^2]$, where the bold symbols denote the vector of the respective coefficients. We note that for both the models, we have the unknown parameters as well as the unknown latent variables which we need to sample from the posterior distribution.

Considering the independent priors, the joint posterior density for the model in the equation (2.4) can be expressed as the following:

$$\begin{aligned} \pi(\mathbf{B}, \Theta_2, Z_1^*|Z_1) \propto \pi(\mathbf{b}, \boldsymbol{\delta}, \sigma_v^2, \sigma_\epsilon^2) \times \prod_{i=1}^{n} \int \prod_{t=1}^{T} f(Z_{1i}^*(t)|v_i) \\ \times \left\{ \sum_{c=0}^{C} I(Z_{1i}(t) = c) I(B_{c-1} < Z_{1i}^*(t) < B_c) \right\} f(v_i) dv_i, \end{aligned} \tag{2.7}$$

where the joint prior $\pi(\mathbf{b}, \boldsymbol{\delta}, \sigma_v^2, \sigma_\epsilon^2)$ is the product of the individual prior densities. Following the tradition of the Bayesian regression, we consider normal priors for $\mathbf{b}$ and

$\boldsymbol{\delta}$, inverse gamma priors for σ_v^2 and σ_ϵ^2, for conjugacy. Here in addition, we consider a diffuse prior for $\mathbf{B}$.

The posterior distributions of Θ_2 conditional on $\mathbf{B}, Z_1^*$ and Z_1 is a routine work and hence we skip the details on it. For $\mathbf{b}$, and $\boldsymbol{\delta}$, the full conditionals are normal and for σ_v^2, and σ_ϵ^2, they are inverse gamma. The latent variables $Z_{1i}^*(t)$ are sampled from the following conditional density:
$(Z_{1i}^*(t)|v_i, \Theta_2, \mathbf{B}, Z_{1i}(t) = c)$ distributed as truncated $N\left(\sum\limits_{k=1}^{K} \theta_k(t)W_{ki}(t) + \sum\limits_{k'=1}^{K'} \delta_{k'}S_{k'i}(t) + v_i, \sigma_\epsilon^2\right)$, truncated at the left by B_{c-1} and at the right by B_c.

Finally, the full conditional density of B_c given all the other parameters can be written as the following:
$\prod\limits_{i=1}^{n}\prod\limits_{t=1}^{T}\left[I(Z_{1i}(t) = c)I(B_{c-1} < Z_{1i}^*(t) < B_c) + I(Z_{1i}(t) = c+1)I(B_c < Z_{1i}^*(t) < B_{c+1})\right].$

This conditional density is a uniform density on $[max\{max\{Z_{1i}^*(t) : Z_{1i}(t) = c\}, B_{c-1}\}, min\{min\{Z_{1i}^*(t) : Z_{1i}(t) = c+1\}, B_{c+1}\}]$.

Thus, our Gibbs sampler runs as the following. We start with some initial estimates of Θ_2 and $\mathbf{B}$ and sample Z_1^* and B using the above full conditionals. Then we use the sampled Z_1^* values and B to get the updated estimates of Θ_2 based on the equation (2.4).

The joint posterior density for the model in equation (2.3) can be written as the following:

$$\begin{aligned}\pi(\Theta_1, Y^*|Y) \propto \pi(\mathbf{a}, \boldsymbol{\gamma}, \sigma_u^2, \sigma_e^2) \times \prod_{i=1}^{n}\int\prod_{t=1}^{T} f(Y_i^*(t)|u_i) \\ \times \{1(Y_i^*(t) > 0)1(Y_i(t) > 0) + 1(Y_i^*(t) \leq 0)1(Y_i(t) = 0)\} f(u_i)du_i,\end{aligned} \tag{2.8}$$

where the joint prior is again the product of the individual prior densities. Normal priors are taken for $\mathbf{a}$ and $\boldsymbol{\gamma}$, and inverse gamma priors are taken on σ_u^2 and σ_e^2.

We again employ a Gibbs sampler for the parameter estimation. Note that given the model parameters Θ_1, the random effects u_i, the observed $Y_i(t)$, and the sampled $Z_{1i}^*(t)$ from the other model, we sample the latent variable $Y_i^*(t)$ as the following:

$$\begin{cases} Y_i(t); & \text{for } Y_i(t) > 0; \\ N\left(\sum_{j=1}^{J} \beta_j(t) X_{ji}(t) + \gamma_1 Z_{1i}^*(t) + \sum_{j'=2}^{J'} \gamma_{j'} Z_{ji}(t) + u_i, \sigma_e^2\right); & \text{right truncated by 0, for } Y_i(t) = 0. \end{cases} \tag{2.9}$$

Once the $Y_i^*(t)$ values are sampled, it is a routine work to estimate the parameters Θ_1 in equation (2.3) using the standard Bayesian regression technique. Note that here we run the two-stage regression simultaneously; in each iteration, we use the sampled $Z_{1i}^*(t)$ values for sampling $Y_i^*(t)$ values. Since we are using Gibbs sampler, the algorithm becomes very fast and computationally easy. This is an advantage over the traditional two-stage regression modelling, where one has to fit two models separately and estimate the parameter in each stage.

The starting values of the model parameters are taken from the ordinary least squares (OLS) estimates for the usual two-stage regression approach. We run MCMC for 65,000 iterations and discard the first 5,000 iterations as “burn-in”. We also thin the chains by saving every 10-th iteration. The convergence of the chains are monitored graphically.

3 HRS Data Analysis

We analyse the data from the health and retirement study conducted by the University of Michigan on an aging population. The study started in 1992 and ended in 2012 with a follow-up frequency of every two years, which we refer as waves. The study started with the subjects within the age group of [52-62], i.e. the subjects who are either retired or about to retire. The subjects were measured for disease status, financial status, number of hospital visits, total medical expenditures, out of pocket medical cost etc. The goal of the study was to decipher the challenges in aging and the effects of the disease status and the financial status on the out-of-pocket medical expenditure (OOPME). In the health economics, OOPME is considered as an important measure of financial risk (Mukherji et al. 2015).

We analyse a part of the data from this study. Our current dataset contain 630 subjects who are measured for all the 10 waves. We consider OOPME as our dependent variable which depends on the financial status, number of hospital visits, body mass index, health insurance status, and disease status. However, there is an obvious endogeneity since the number of hospital visits clearly depends on the financial status, and the disease status etc. Thus, a two-stage regression model is meaningful in this context. However, note that we have excess zeros for the number of hospital visits and OOPME. If the health condition is good at some wave then there is no hospital visit, and even for the non-zero hospital visits the OOPME can be zero if all the medical costs are reimbursed through health insurance. For our data, we have nearly 12% zeros for the number of hospital visits and nearly 15% zeros for the OOPME. The disease status includes the binary outcomes (yes/no) for 8 diseases (blood pressure, diabetes, cancer, lung problem, heart problem, stroke, arthritis and psychological disorders) and the health insurance category includes the binary outcomes (yes/no) for three different health insurances (employment insurance, government insurance, and other private insurance). The financial status includes the total family income (annual), value of the total assets, and total value of debt.

3.1 Model and Priors

Let $Y_i(t)$ denote the OOPME for the i-th subject at the t-th wave, $(t = 1, 2, \ldots, 10)$ and we treat this variable as response. Since this is a zero-inflated variable, we define a latent random variable $Y_i^*(t)$ as discussed in Section 2. Also let $Z_{1i}(t)$ be the number (count) of the hospital visits for the i-th subject at the t-th wave and we define latent random variable $Z_{1i}^*(t)$ similar to Section 2. For our data, the maximum value of Z_1 is 5.

We fit models given in equations (2.3) and (2.4) for OOPME and the number of hospital visits. For the latent $Y_i^*(t)$ we consider three financial variables and three (binary) insurance variables as the covariates with time-varying effects. The disease status and BMI are considered as the covariates with time-invariant effects on the response. Thus for modeling $Y_i^*(t)$, we take J=6 and J'=9. The residuals $e_i(t)$ are assumed to be iid $N(0, \sigma_e^2)$. However, for modelling $Z_{1i}^*(t)$ we consider the financial variables as the covariates with time-varying effects on response and the disease status as the covariates with time-invariant effects on the response. Thus, we take K=3 and K'=8, and assume that $\epsilon_i(t)$ are iid $N(0, \sigma_\epsilon^2)$.

The time-varying coefficients for both the models are modelled using LPs as discussed in Section 2.1. For each $\mathbf{a}_j$, and $\mathbf{b}_k$, we consider multivariate normal priors with mean=0, and covariance matrix=$100I$. For the $\gamma_{j'}$ and $\delta_{k'}$, we consider normal priors with mean=0 and variance=20; and for the variance parameters $\sigma_u^2, \sigma_v^2, \sigma_e^2$ and σ_ϵ^2, we take inverse gamma (2,4.5) priors. As mentioned in Section 2.2, we take diffuse prior for $\mathbf{B}$ for the computational ease. We perform a sensitivity analysis to investigate the effects of the prior parameter values on the final inference and noticed that the results are not sensitive to these specific choices of the prior parameters (results not shown).

The model parameters are estimated by the Gibbs sampler discussed in Section 2.2. For assessing the convergence of the Markov chains, we compute the multivariate potential scale reduction factor as proposed by Brooks and Gelman (1998). For our data analysis, the computed scale reduction factors are below 1.1 indicating the convergence.

3.2 Results

In Table 1, we summarize the parameter estimates and the corresponding 95% Bayesian credible intervals for the predictors with time-invariant effects on the response used in modelling the OOPME and the number of hospital visits. Note that the binary disease status for cancer, stroke and heart problem are significant for both the models, since the corresponding parameter estimates are different from zero and the corresponding 95% CIs do not contain zero. However, blood pressure, arthritis and lung problem are significant for the number of hospital visits but not significant for OOPME. Diabetes and psychological disorders are not significant for both the cases. This reflects that the aged patients who suffer from cancer, stroke or heart problems often get hospitalized and the insurance possibly do not cover much for such diseases resulting a higher OOPME. People (aged) suffering from hypertension (blood pressure), arthritis, and/or lung problem also visit hospital often but the cost related to those are mostly covered by the health insurance. BMI, diabetes, and psychological disorder do not play much significant role in either the hospital visit or OOPME.

In Figure 1, we show the time-varying coefficients for the 6 covariates used in modeling

OOPME. We note that the total family income has an increasing effect on OOPME over 10 waves. This means the people with higher family income will have higher OOPME over time. This is quite intuitive since for the older people insurance do not cover much and the people with sound family income can easily pay the excess cost out of their own pockets. Total value of assets has initially an increasing trend till wave 5 and then decreases slowly. This reflects that the people with higher value of assets bear the extra cost (which is not covered by the insurance) nearly upto age 72 and then slowly become reluctant to spend out of their own pockets. The total debt more or less has a constant effect over the waves with a very slow decreasing trend at the end. Employment insurance shows a constant trend till wave 4 and then shows a decreasing trend. The government insurance shows a sharp increasing trend after wave 3. This dictates that the "medicare" health insurance given by the US government to the people older than 65 years has a higher effect on the OOPME. The trend for the other private insurance is decreasing reflecting that the private insurances are not much effective in terms of the coverage for the older people.

In Figure 2, we show the similar plots for the number of hospital visits. Note that the total family income shows an increasing trend reflecting that the older people with higher family income will visit hospital quite often. The total debt has a decreasing trend dictating that the people with higher debt visit hospital less often over time. However, the effect of the total value of assets is more or less constant and hence can be treated as a covariate with fixed effect on the number of hospital visit.

4 Simulation Studies

We investigate the operating characteristics of the proposed approach through simulation studies. We consider a longitudinal response Y measured on n=100 subjects at T=5 evenly spaced time points and consider four covariates X_1, X_2, X_3, and X_4, where X_2 is the endogenous count covariate. For $Y_i(t)$, the response from the i-th subject at the t-th time point, we consider the following distribution:
$Y_i(t) \sim 0.15\delta_0 + 0.85N(\mu_i(t), \sigma_e^2)$, where δ_0 is a point mass at zero, and for $\mu_i(t)$, we consider the following model:

$$\mu_i(t) = \beta_0 + \beta_1(t)X_{1i}(t) + \gamma_1 X_{2i}(t) + \gamma_2 X_{3i}(t) + u_i. \tag{4.1}$$

For the count covariate X_2, we consider the following zero-inflated Poisson distribution:
$X_{2i}(t) \sim 0.10\delta_0 + 0.90$ Poisson (λ_{it}), where for the mean parameter λ_{it} we consider the following model:

$$log(\lambda_{it}) = \theta_0 + \theta_1(t)X_{1i}(t) + \delta_1 X_{3i}(t) + \delta_2 X_{4i}(t) + v_i. \tag{4.2}$$

For the simulation purpose, we take θ_0=1.5; $\delta_1 = 2.2$; $\delta_2 = 3.5$; and generate v_i from $N(0, 5)$. We take $\theta_1(t) = \theta_{10} + \theta_{11}t + \theta_{12}t^2$, with $\theta_{10} = 0.6, \theta_{11} = 1.4, \theta_{12} = 2$. Generate X_1 from 5 variate normal density with mean = 0 and covariance matrix

$$\Sigma_1 = \begin{bmatrix} 2.0 & 1.4 & 1.7 & 2.0 & 2.1 \\ & 4.0 & 2.9 & 3.4 & 3.6 \\ & & 6.0 & 2.8 & 2.9 \\ & & & 8.0 & 5.9 \\ & & & & 9.0 \end{bmatrix}.$$

Similarly, generate X_3 from a 5 variate normal density with mean=$[1.2, 2.3, 4, 3.8, 5]^T$ and covariance matrix

$$\Sigma_1 = \begin{bmatrix} 3.0 & 1.6 & 2.1 & 3.8 & 1.4 \\ & 6.0 & 1.9 & 4.2 & 2.9 \\ & & 9.0 & 3.6 & 1.5 \\ & & & 4.0 & 3.9 \\ & & & & 2.0 \end{bmatrix}.$$

And we take iid $X_{4i}(t)$ from Gamma (2,5) density.

For generating $Y_i(t)$, we take β_0= 2.5, $\gamma_1 = 1.7, \gamma_2 = 2.6$. The random effects u_i are generated from N(0,3). We take $\beta_1(t) = \beta_{10} + \beta_{11}t$, with $\beta_{10} = 1.3, \beta_{11} = 2.5$.

Once we generate the data, we fit two competing models to the data. The first model is the traditional two-part model for zero-inflated response, and the endogenous covariate. For the response, define $p_i(t)$ as the probability of a zero response and model this as $\text{logit}(p_i(t)) = \kappa_0 + \kappa_1(t)X_{1i}(t) + \kappa_2 X_{2i}(t) + \kappa_3 X_{3i}(t) + c_i$. Then we model the non-zero responses using the the model given in equation (4.1).

For the zero-inflated count covariate, similarly, define $\pi_i(t)$ as the probability of a zero value and model $\text{logit}(\pi_i(t))$ using the covariates and the subject-specific random effects. Finally, model the non-zero counts as a Poisson $(\lambda_i^*(t))$, and model $\log(\lambda_i^*(t))$ similar to equation (4.2). We refer this as Model I.

Then we fit our proposed approach of simultaneously modelling the response and the endogenous predictor using latent random variables. We refer this as Model II. Model parameters are estimated for both the approaches using the Bayesian MCMC methods.

We consider 100 replicates of the dataset generated above and for each dataset fit both the models. We compute three important model selection criteria, e.g. (i) mean squared error, (ii) log pseudo-marginal likelihood, and (iii) deviance information criterion for both the models under consideration.

For each replicated dataset, once we fit the models, we compute the mean squared error (MSE). Based on 100 replicates we obtain 100 MSE values for both the models and compute the average MSE (AMSE) for the model selection. A smaller AMSE value indicates a better fit.

Secondly, we use the conditional predictive ordinate (CPO) (Gelfand et al. 1992) defined as $CPO_i = P(Y_i|Y_{-i}) = E_{\boldsymbol{\theta}}\left[P(Y_i|\boldsymbol{\theta}, Y_{-i}\right]$, where Y_{-i} denotes the data excluding Y_i and $\boldsymbol{\theta}$ denotes the set of all model parameters. We compute the log pseudo-marginal likelihood (LPML)=$\sum\limits_{i=1}^{n} log\widehat{CPO}_i$. A higher LPML value indicates a better fit.

Finally we compute the conditional deviance information criteria (DIC) proposed in Celeux et al. (2006). This DIC is based on the conditional likelihood $l(\mathbf{Y}|\boldsymbol{\gamma})$ and is given by
DIC=$-4\mathbf{E}[logl(\mathbf{Y}|\boldsymbol{\gamma})]+2logl(\mathbf{Y}|\hat{\boldsymbol{\gamma}})$, where $\boldsymbol{\gamma}$ denotes the vector of random effects and $\hat{\boldsymbol{\gamma}}$ is the corresponding estimate (posterior mean). A smaller value indicates a better fit.

Results are summarized in Table 2. We note that AMSE values are quite similar for both the models. Model II gives a higher value of LPML compared to Model I. For DIC, Model I gives a higher value compared to Model II. This indicates, in general, Model II is giving a better fit than Model I in the simulated data.

In Table 3, we show the average bias and average width of the 95% Bayesian credible intervals of the model parameters given in equations (4.1) and (4.2) based on 100 replications. We also provide the corresponding coverage probabilities. We note that for most of the parameters, the average bias is smaller for Model II. Also Model II provides shorter CIs with satisfactory coverage probabilities. This illustrates the better practical usefulness of our proposed approach compared to a traditional two-part model in two-stage regression.

5 Discussion

Two-stage regression technique is quite popular in the presence of one or more endogenous covariates. However, the problem is challenging for zero-inflated longitudinal response and zero-inflated endogenous covariates. We have developed a general Bayesian approach of handling such data and our method does not require a continuous response and/or continuous covariate. The proposed method is computationally simple since it is based on the Gibbs sampler. All our computations are performed in R.

In this paper, we have only one endogenous zero-inflated covariate which is a count variable. But our method is general and even in the presence of two or more endogenous covariates our method can be used directly. The advantage of the proposed approach is the continuity of the latent random variables which can be assumed as Normal and hence the full conditionals become simple.

For longitudinal data, we often have missingness in our data. Although we have not explicitly considered missing data in our current analysis, but if the missingness is ignorable (i.e. missing at random) then a simple data augmentation technique has to be added to our proposed approach. For non-ignorable missingness, one needs more complex techniques discussed in detail in Daniels and Hogan (2008). We leave this issue as a possible future work.

References

Aitkin, M. and Rubin, D. (1985) Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B.* 47, 67-75.

Albert, J. and Chib, S. (1993) Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88, 669–79.

Amemiya, T. (1984) Tobit models: A survey. *Journal of Econometrics*, 24, 3-61.

Angrist, J.D. and Imbens, G.W. (1995) Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity. *Journal of the American Statistical Association*, 90, 431-42.

Arias, E. (2006) United States Life Tables, 2006. *National Vital Statistics Reports*, 58(21).

Cui, Y., Zhu, J., and Wu, R. (2006) Functional mapping for genetic control of programmed cell death. *Physiol. Genomics*, 25, 458-69.

Daniels, M.J. and Hogan, J.W. (2008) Missing data in longitudinal studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Chapman and Hall (CRC Press).

Das, K., Li, J., Wang, Z., Tong, C., Fu, G., Li, Y., Xu, M., Ahn, K., Mauger, D., Li, R. and Wu, R.L. (2011) A dynamic model for genome-wide association studies. *Human Genetics*, 129, 629-39.

Dong, Y. and Lewbel, A. (2015) A simple estimator for binary choice models with endogenous regressors. *Econometric Reviews*, 34, 82-105.

Duan, N., Manning, W., Morris, C. and Newhouse, J.P. (1983) A comparison of alternative models for the demand for medical care (Corr: V2 P413). *Journal of Business and Economic Statistics*, 1, 115-126.

Ghosh, P. and Tu, W. (2008) Assessing sexual attitudes and behaviors of young women: a joint model with nonlinear time eects, time varying covariates, and dropouts. *Journal of the American Statistical Association*, 103, 1496–1507.

Hall, D. (2000) Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics*. 56, 1030-39.

Holliday, T., Lawrance, A.J. and Davis, T.P. (1998) Engine-Mapping Experiments: A Two-Stage Regression Approach. *Technometrics*, 40, 120-26.

Huskova, M. and Sen, P.K. (1985) On sequentially adaptive asymptotically efficient rank statistics. *Sequential Analysis*, 4, 125–51.

Khan, S. and Lewbel, A. (2007) Weighted and two-stage least squares estimation of semi-parametric truncated regression models. *Econometric Theory*, 23, 309-47.

Lambert, D. (1992) Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 34, 1-14.

Liu, X. and Lee, L.F. (2013) Two Stage Least Squares Estimation of Spatial Autoregressive Models with Endogenous Regressors and Many Instruments. *Econometric Reviews*, 32, 734-53.

Meyer, K. (2000) Random regressions to model phenotypic variation in monthly weights of australian beef cows. *Livestock Production Sci.*, 65, 19-38.

Min, Y. and Agresti, A. (2002) Modeling nonnegative data with clumping at zero: A survey. *Journal of Iranian Statistical Society*, 1, 7–33.

Mukherji, A. , Roychoudhury, S., Ghosh, P. and Brown, S. (2015) Estimating Health Demand for an Aging Population: A Flexible and Robust Bayesian Joint Model. *Journal of Applied Econometrics*, doi: 10.1002/jae.2463.

Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 34165.

Simar, L. and Wilson, P.W. (2011) Two-stage DEA: caveat emptor. *Journal of Productivity Analysis*, 36, 205-18.

Scott, A.J. and Hold, D. (1982) The Effect of Two-Stage Sampling on Ordinary Least Squares Methods. *Journal of the American Statistical Association*, 77, 848-54.

Ye, W., Lin, X. and Taylor, J.M. (2008) Semiparametric Modeling of Longitudinal Measurements and Time-to-Event DataA Two-Stage Regression Calibration Approach. *Biometrics*, 64, 123846.

Table 1: Parameter estimates and 95% Bayesian CI for the covariates with time invariant effects on OOPME and the number of hospital visits for the HRS data.

	OOPME		No. of hospital visits	
Predictor	Coefficient Estimate	95% C.I.	Coefficient Estimate	95% C.I.
Blood pressure	0.025	(-2.62,1.13)	1.49	(0.63,3.54)
Diabetes	0.13	(-1.04,1.77)	0.28	(-1.18,0.96)
Cancer	4.92	(1.28, 6.54)	5.29	(3.01,8.66)
Lung problem	0.36	(-1.23, 1.46)	3.62	(1.24,5.58)
Heart problem	3.32	(1.32,6.08)	3.75	(1.28, 5.01)
Stroke	9.53	(7.36, 12.70)	7.38	(5.07,10.42)
Arthritis	-0.32	(-3.38,1.06)	2.34	(0.69,4.52)
Psychological problem	0.03	(-1.09,0.94)	-0.05	(-1.48,0.63)
BMI	0.68	(-2.87,1.49)	-	-

Table 2: Goodness of fit (GOF) measures for Model I and Model II in the simulation study.

GOF Measure	Model I	Model II
AMSE	13.85	14.11
LPML	-234.62	-184.25
DIC	38.54	25.16

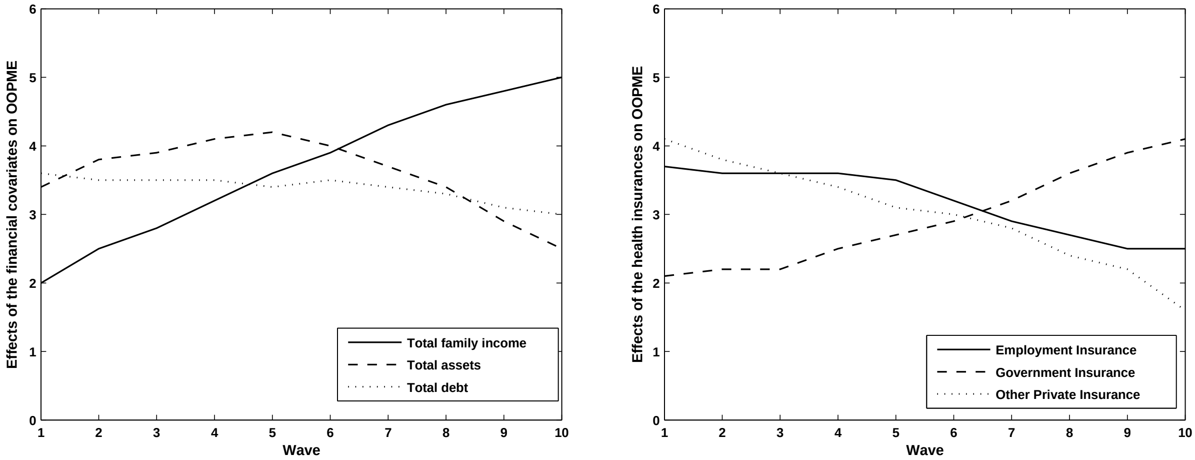

Figure 1: Time varying effects of the financial covariates and the health insurances on the out-of-pocket medical expenditure.

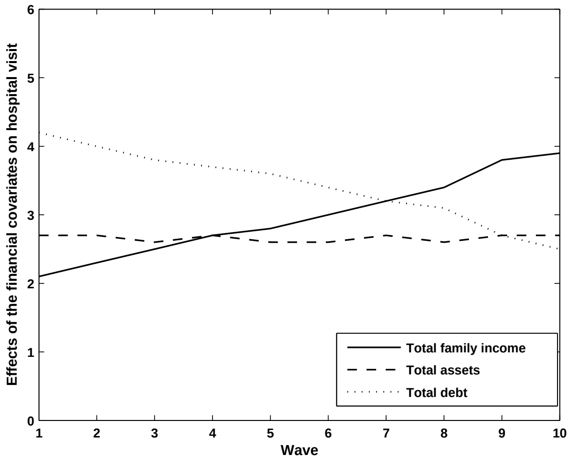

Figure 2: Time-varying effects of financial covariates on the number of hospital visits.

Table 3: Estimated bias, width of the 95% CI and the coverage probabilities for Model 1 and Model II in the simulation study. CI width stands for the width of the 95% Credible intervals and Cov.P stands for the corresponding coverage probability.

	Model I		Model II	
Parameter	Bias	CI width(Cov.P)	Bias	CI width(Cov.P)
β_0	0.23	1.23(0.96)	0.12	1.02(0.94)
β_{10}	0.21	1.45(0.96)	0.13	0.86(0.95)
β_{11}	0.18	1.58(0.95)	0.14	0.99(0.95)
θ_0	0.16	1.02(0.95)	0.12	0.76(0.94)
θ_{10}	0.24	0.97(0.96)	0.16	0.68(0.95)
θ_{11}	0.14	1.11(0.95)	0.14	1.06(0.94)
γ_1	0.17	0.98(0.95)	0.18	0.89(0.94)
γ_2	0.16	2.21(0.96)	0.10	1.14(0.94)
δ_1	0.10	2.03(0.95)	0.11	1.33(0.95)
δ_2	0.15	1.31(0.95)	0.09	0.94(0.94)