MCMC Methods for Truncated Poisson Regression of Insurance Claims Data

David Pitt
Australian National University, School of Finance and Applied Statistics
Faculty of Economics and Commerce
Canberra, Australia
David.Pitt@anu.edu.au

Terry O' Neill
Australian National University, School of Finance and Applied Statistics
Faculty of Economics and Commerce
Canberra, Australia
Terry.Oneill@anu.edu.au

Abstract

Reinsurers often receive truncated data of insurance claims. One common form of truncation arises due to claims only being reported to the reinsurer if they exceed a certain level. This paper aims to fit a regression model to the truncated data with the aim of estimating the untruncated sample size in addition to suitable regression coefficients for certain covariates in the model. Hence the reinsurer will be able to get an estimate of the total number of claims and the significance of potential explanatory variables for claim size. A Poisson regression model with logarithmic link is proposed for the analysis. A Bayesian framework is utilised whereby the estimation of parameters becomes quite complicated potentially involving multi-dimensional integrals. Markov Chain Monte Carlo methods, such as the Gibbs Sampler and the Metropolis Hastings Algorithm are used to simplify the process of estimating the parameters.

1. Introduction

Truncated data arises when the range of possible responses has been restricted in some way. Methods for the regression modeling of such data have been proposed by Grogger and Carson (1991) and Shaw (1988). O' Neill and Barry (1995) recently proposed a truncated model for grouped binary data.

This paper aims to examine the regression analysis of data from a truncated Poisson distribution. The Poisson distribution arises from grouping together claims into bins of width 5000. The first bin, corresponding to the first 5000 dollars of claims is assumed to be unknown and regression is performed on the remaining untruncated data. An estimate of both the mean of the total sample size

and also the sensitivity of the amount of data in each claim size bin will be provided.

2. Bayesian Model

The results will be developed using a Poisson model. The bins numbered from zero upward will be dentoed Y_i and assumed to follow a Poisson distribution

$$P(Y_i = y) = \frac{e^{-m_i} m_i^{y_i}}{y_i!}; y = 0,1,2,... \qquad (1)$$

where μ_i is the mean of the process calculated using a logarithmic link such that

$$g(m_i) = \log m_i = \beta_0 + \beta_1(Time) \qquad (2)$$

The sample contains n untruncated observations and is assumed to have come from a sample containing N observations where N is unknown.

The likelihood of the complete sample is given by

$$[Y, X, n \mid \beta, n] = \binom{N}{n} \prod_{i=1}^{n} [Y_i \mid X_i, \beta][X_i] \prod_{i=n+1}^{N} [Y_i = 0 \mid X_i, \beta][X_i] \qquad (3)$$

We next calculate the marginal distribution of the untruncated sample size,n, and the observed data. Under the assumption that [N|λ] is Poisson(λ) we get

$$[Y^*, X^*, n \mid \beta, \lambda] = \prod_{i=1}^{n} [Y_i \mid X_i, \beta][X_i] \lambda^n e^{-\lambda P(\beta)} \qquad (4)$$

where P(β) is the unconditional probability of observing a unit randomly chosen from [X].

We next specify prior distributions for each of β, η, λ and determine the posterior density conditional on all other variables as

$$[\beta, \lambda, \eta \mid \bullet] \propto \prod_{i=1}^{n} [Y_i \mid X_i, \beta] \left(\prod_{i=1}^{n} \eta_i \right)^{\frac{2}{n}} \lambda^n \frac{e^{-\lambda \eta' p}}{\eta' p} \qquad (5)$$

where p is the vector of probabilities of non-truncation and q is its complement.

From this joint posterior density, marginal posteriors are found for each of the parameters. Note the

introduction of dummy variables t and j. These are designed to produce marginal posteriors from which simulation can be readily performed.

The resulting marginal posteriors are given below

$$[t \mid \bullet] \propto (\lambda + t)^{j \bullet} e^{-t} \qquad (6)$$

$$[\lambda \mid \bullet] \propto (\lambda + t)^{j \bullet} e^{-\lambda}\, \lambda^{n} \qquad (7)$$

$$[\beta \mid \bullet] \propto \prod_{i=1}^{n} [Y_i \mid X_i, \beta] \prod_{i=1}^{n} q_i^{j_i} \qquad (8)$$

$$[\theta \mid \bullet] \propto \prod_{i=1}^{n} \theta_i^{\alpha/n + j_i} \qquad (9)$$

$$[j \mid \bullet] \propto \prod_{i=1}^{n} \frac{(\theta_i q_i)^{j_i}}{j_i!} (\lambda + t)^{j \bullet} \qquad (10)$$

Equations (6), (7), (9) and (10) can readily be simulated from using mixture distributions. Equation (8) though is not easily recognised and so the Metropolis Hastings Algorithm is employed to simulate from this density. The Metropolis Hastings Algorithm simulates from the given marginal posterior distribution (8) by forming a Markov Chain which converges to a stationary distribution. It is necessary to specify a transition matrix in order to form the Markov Chain.

The technique of Gibbs sampling is then used to estimate the above parameters. This method essentially involves simulating from each marginal distribution and then repeating this process many times. The resulting simulated values from the marginal distributions will tend towards simulated values from the joint posterior density (5). The Gibbs Sampler is then iterated and the mode of the simulated outputs from the runs of the Gibbs method are used as the parameter estimates.

The technique has been applied to some Workers' Compensation claims for an Australian general insurer. The results of the analysis are shown below.

Estimated value of β_0 is 1.44 and the estimated value of β_1 is –0.012. An estimate of λ was found to be 142.

The results are subject to variation depending on the number of iterations both of the Gibbs Sampler and of the Metropolis Hastings Algorithm. Raftery and Lewis have proposed a method for determining the convergence of the Markov Chain to a stationary distribution, with a given probability.

References

Amemiya, T. (1984). Tobit models: a review. *J.Econometrics*, 4, 3-61

Devroye, L. (1986) Non-uniform random variate generation. Springer-Verlag, New York

McCullagh, P. & Nelder, J. (1989). *Generalised Linear Models*. Chapman & Hall, New York.

O' Neill, T.J. and Barry, S.C. (1995) Truncated logistic regression. *Biometrics*, 51,533 – 541

Tanner M.A. (1993) *Tools for Statistical Science*. Springer-Verlag, New York.

Raftery, A. & Lewis, S. (1992) Comment: One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statistical Science*, 7, 493-497.

Resume

Cet article présente une régression tronqué pour des données d' assurance en utilisant les méthodes de Monte Carlo Markov Chain (MCMC) et le Gibbs Sampler. La distribution marginale postériore est trouvé pour chaque variable. Le Gibbs Sampler peut etre utilisé pour simuler de la distribution posteriore. L' algorithme de Metropolis et Hastings est utilise pour simuler les coefficients de la regression.