Int. J. Contemp. Math. Sci., Vol. 2, 2007, no. 13, 639 - 648

A Bayesian Mixture Model with Application to Typhoon Rainfall Predictions in Taipei, Taiwan[1]

Tsai-Hung Fan

Graduate Institute of Statistics
National Central University
Jhongli, Taiwan 320

Yun-Huan Lee

Department of Applied Statistics and Information Science
Ming Chuan University, Taoyuan, Taiwan, 333

Abstract

In linear regression, it is typically assumed that the response variables are normally distributed. In practice, however, quite often the response variables are not only positive but with many zero measurements. In this article, we use Bayesian approach to analyze the data in which the distribution of the response variable is considered to be a mixture of a continuous distribution and a point mass at zero. Gibbs sampling algorithm is conducted to draw Bayesian inference and predictions and the results are quite accurate from the simulation study. The proposed model is also employed to make rainfall predictions for the typhoon data in Taipei, Taiwan.

Keywords: mixture distribution, point mass, probit model, Bayesian prediction

1 Introduction

In linear regression, it is typically assumed that the response variables are normally distributed. When the observations are positive, log-normal regression

[1]This work was supported by the MOE Program for Promoting Academic Excellent of Universities and the Central Weather Bureau in Taiwan under the grant numbers 91-H-FA07-1-4 and MOTC-CWB-93-2M-06, respectively.

analysis is considered usually. In practice, the data may not only be positive but with many zero measurements, such as manufacturing defects, typhoon rainfalls etc. More often, these observations are influenced by some other covariates and we are more interested in future predictions. How to find an appropriate regression model is thus needed.

If the response variable is discrete, for example the defect counts collected from manufacturing process, commonly used models such as Poisson or geometric regression distributions often underestimate the zero-defect probability. Zero-inflated Poisson (ZIP) regression models (Lambert 1992, Gupta 1996) are more suitable to model this type of data. Ghosh and Mukhopadhyay et al. (2002) present Bayesian analysis of the ZIP regression models. This approach overcomes most of the limitations in classical inference procedures. Albert and Chib (1993) use a Bayesian approach to analyze binary data by introducing a latent variable for each observation through the probit model. In this paper, we consider zero-inflated model with continuous response variables in which the data are mostly positive real measurements, but with many zero observations. Traditional linear regression models are not appropriate since the observations are obviously not normally distributed but with a positive probability at zero. Logistic regression models are not suitable either for it wastes the information of those positive data. We will consider a mixture distribution of a regression model with positive responses and a degenerate distribution at zero. That is, given the explanatory vectors, $\mathbf{X}_i = (X_{1i}, X_{2i}, \ldots, X_{ki})^t$, $i = 1, \ldots, n$, the response variables are

$$Y_i = \begin{cases} 0 & \text{with probability } p_i. \\ Z_i & \text{with probability } 1 - p_i, \end{cases} \tag{1}$$

where $Z_i > 0$ and $0 < p_i < 1$. Conditionally on $\mathbf{X}_i$, Y_i are assumed to be independent observations for $i = 1, \ldots, n$. It is furthermore assumed that Z_i is of truncated normal distribution with mean $\mathbf{X}_i^t\underline{\beta}$ and variance σ^2, where $\underline{\beta} = (\beta_1, \ldots, \beta_k)^t$ and σ^2 are unknown parameters. Typically, one may assume $p_i = H(\mathbf{X}_i^t\underline{\beta})$, where H is a known cumulative distribution function linking the probabilities with the linear structure $\mathbf{X}_i^t\underline{\beta}$. A reference can be seen in McCullagh and Nelder (1989). In this paper, we consider the probit model, with $p_i = \Phi(\mathbf{X}_i^t\underline{\beta}/\sigma)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution since it allows for a convenient sampling scheme in computation, as originally developed by Albert and Chib (1993). We will conduct a Bayesian analysis with respect to the noninformative priors on the unknown parameters $(\underline{\beta}, \sigma^2)$ based on the given observations. The key idea is to introduce independent latent variables into the problem when the observations

are zeros. These principal observations, combined with the Gibbs sampler (Geman and Geman (1984)), allow us to simulate the posterior distribution of $\underline{\beta}$ and σ^2 as well as the predictive distribution of Y. We will illustrate the Bayesian analysis via the Gibbs sampling algorithm for the model in Section 2. In Section 3 we give some simulation results to confirm our study and apply this procedure to typhoon rainfalls data in Taipei, Taiwan and Section 4 concludes our study.

2 Bayesian analysis for mixture regressions

Given the data $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, from the mixture model (1), we are not only interested in estimating p_i, $i = 1, \ldots, n$ and $\underline{\beta}$, but also in predicting the future response value Y^* when the covariate is given at $\mathbf{X} = \mathbf{X}^*$, that is independent of the previous ones. A Bayesian analysis will be considered. First, we introduce a binary variable T_i for each observation such that $T_i = 1$ if $Y_i = 0$, and $T_i = 0$ if $Y_i > 0$. Thus $P(T_i = 1) = p_i = \Phi(\mathbf{X}_i^t \underline{\beta}/\sigma)$, for $i = 1, \ldots, n$. Moreover, for each $i = 1, \ldots, n$, if $Y_i = 0$, a latent variable Z_i^0 is embedded, where $Z_i^0 < 0$ follows the truncated normal distribution with mean $\mathbf{X}_i^t \underline{\beta}$, and variance σ^2 and independent with each other; otherwise, let $Z_i = Y_i$, if $Y_i > 0$. Without loss of generality, suppose that $Y_i = 0$, for $1 \le i \le r < n$; and $Y_i > 0$, for $i = r+1, \ldots, n$. Denote $\mathbf{Z}^0 = (Z_1^0, \ldots, Z_r^0)$ and let $\pi(\underline{\beta}, \sigma^2)$ be the joint prior density of $\underline{\beta}$ and σ^2. Then given $(\mathbf{X}, \mathbf{Y}, \mathbf{T}) = \{(\mathbf{X}_1, Y_1, T_1), \ldots, (\mathbf{X}_n, Y_n, T_n)\}$, the posterior density of $(\underline{\beta}, \sigma^2, \mathbf{Z}^0)$ is

$$\pi(\underline{\beta}, \sigma^2, \mathbf{Z}^0 | \mathbf{X}, \mathbf{Y}, \mathbf{T}) \propto \pi(\underline{\beta}, \sigma^2) \prod_{i=1}^{r} \phi(Z_i^0; \mathbf{X}_i^t \underline{\beta}, \sigma^2) I(Z_i^0 < 0) \prod_{i=r+1}^{n} \phi(Z_i; \mathbf{X}_i^t \underline{\beta}, \sigma^2), \tag{2}$$

where $\phi(x; \mu, \sigma^2)$ is the density of the $N(\mu, \sigma^2)$ distribution, and $I(\cdot)$ is the indicator variable.

If the usual noninformative prior, namely $\pi(\underline{\beta}, \sigma^2) \propto 1/\sigma^2$, is considered, then, for given data $(\mathbf{X}, \mathbf{Y}, \mathbf{T})$, the conditional posterior distribution of $\underline{\beta}$ given $\mathbf{Z}^0$ and σ^2 is

$$\underline{\beta} | \mathbf{Z}^0, \sigma^2, \mathbf{X}, \mathbf{Y}, \mathbf{T} \sim N\left(\underline{\hat{\beta}}_{LSE}, \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}\right), \tag{3}$$

where $\underline{\hat{\beta}}_{LSE} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Z}$, the usual least square estimator based on the complete data $\mathbf{Z} = (\mathbf{Z}^0, Z_{r+1}, \ldots, Z_n)$. Similarly, given $\underline{\beta}$ and σ^2, for each $i = 1, \ldots, r$, Z_i^0 has conditional posterior distribution

$$Z_i^0 | \underline{\beta}, \sigma^2, \{Z_j^0, j \ne i\}, \mathbf{X}, \mathbf{Y}, \mathbf{T} \sim N(\mathbf{X}_i^t \underline{\beta}, \sigma^2) I(Z_i^0 < 0);$$

and the conditional posterior of σ^2 given $\mathbf{Z}$ and $\underline{\beta}$ is

$$\sigma^2|\mathbf{Z}^0, \underline{\beta}, \mathbf{X}, \mathbf{Y}, \mathbf{T} \sim \text{Inverse Gamma } \left(n/2; (\mathbf{Z} - \mathbf{X}\underline{\beta})^t(\mathbf{Z} - \mathbf{X}\underline{\beta})/2\right). \quad (4)$$

Therefore, we can simulate an approximate posterior sample of $(\underline{\beta}, \sigma^2)$ via the Gibbs sampling (Casella and George (1992)) by iteratively generating random samples from the above conditional distributions. Let $(\underline{\beta}_k, \sigma_k^2)$, $k = 1, \ldots, N$, be the resulting (approximate) posterior sample of $(\underline{\beta}, \sigma^2)$, then $\hat{\underline{\beta}} = \sum_{k=1}^N \underline{\beta}_k/N$ and $\hat{\sigma^2} = \sum_{k=1}^N \sigma_k^2/N$ can be used as the (approximate) Bayesian estimates of $\underline{\beta}$ and σ^2, respectively. Furthermore, for each $i = 1, \ldots, n$, the resulting $p_i^k = \Phi(\mathbf{X}_i^t \underline{\beta}_k/\sigma_k)$, $k = 1, \ldots, N$, also form an approximate posterior sample of p_i and its sample mean, $\hat{p}_i = \sum_{k=1}^N p_i^k/N$, indeed is the corresponding Bayesian estimate of p_i.

Note that given the observations $\mathbf{X}, \mathbf{Y}$, the predictive density of the response Y^* at $\mathbf{X} = \mathbf{X}^*$ is

$$f(y|\mathbf{X}^*, \mathbf{X}, \mathbf{Y}) = \int f(y|\mathbf{X}^*, \underline{\beta}, \sigma^2, p(\beta, \sigma^2))\pi(\underline{\beta}, \sigma^2|\mathbf{X}, \mathbf{Y}) d\underline{\beta} d\sigma^2, \quad (5)$$

where $f(y|\mathbf{X}^*, \underline{\beta}, \sigma^2, p)$ is the pdf of the mixture distribution defined by (1), and $\pi(\underline{\beta}, \sigma^2|\mathbf{X}, \mathbf{Y})$ is the posterior of $\underline{\beta}$ and σ^2 . Thus, the response Y^* at $X = \mathbf{X}^*$ can be estimated via the predictive expectation of Y^* with respect to (5), that can be approximated by the sample mean of a sample drawn from the posterior predictive distribution of $Y^\star$ from a Bayesian perspective. The following algorithm gives the prediction of Y^* at $\mathbf{X} = \mathbf{X}^*$.

Step 1): For each posterior sample point $\underline{\beta}_k, \sigma_k^2$, $k = 1, \ldots, N$, generated from the above algorithm, compute

$$p^{*k} = \Phi(\mathbf{X}^{*t} \underline{\beta}_k/\sigma_k).$$

Step 2): Generate a Bernoulli random variate with p^{*k} as the success probability for each $k = 1, \ldots, N$, say T_k^*. Let $Y_k^* = 0$, if $T_k^* = 1$; otherwise, generate $Z^* > 0$ from $N(\mathbf{X}^{*t}\underline{\beta}_k, \sigma_k^2)$ and take $Y_k^* = Z^*$.

Step 3): An estimate of Y^* from a Bayesian perspective is the sample mean

$$\hat{Y}^* = \frac{1}{N} \sum_{k=1}^N Y_k^*, \quad (6)$$

with associated posterior variance $\sum_{k=1}^N (Y_k^* - \hat{Y}^*)^2/(N-1)$.

Here, we only consider the noninformative prior distribution on the unknown parameters $(\underline{\beta}, \sigma^2)$. The algorithm is also valid when conjugate prior

on $(\underline{\beta}, \sigma^2)$ is considered. For example, if $\underline{\beta}$ is of k-variate normal distribution with mean vector $\underline{\mu}$ and variance-covariance matrix Σ and σ^2 is of inverse gamma $(\nu_0/2, \tau_0/2)$ prior, then, under independence assumption of $\underline{\beta}$ and σ^2, the conditional posterior of $\underline{\beta}$ in (3) should be replaced by $N(D^{-1}(\sigma^{-2}\mathbf{X}^t\mathbf{Z} + \Sigma^{-1}\underline{\mu}), D^{-1})$, where $D = \sigma^{-2}\mathbf{X}^t\mathbf{X}+\Sigma^{-1}$, and that of σ^2 in (4) must be replaced by inverse gamma $((n + \nu_0)/2, [(\mathbf{Z} - \mathbf{X}\underline{\beta})^t(\mathbf{Z} - \mathbf{X}\underline{\beta}) + \tau^0]/2)$ in the sampling procedures.

3 Simulation and application

We first present a simulation study for the proposed analysis. The data sets were generated through model (1) with $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (1.1301, 2.5620, 1.2015, 1.5231, 3.0023)$ and $\sigma^2 = 1$, and the independent variables were all generated from $U(0, 3)$ distribution for sample sizes $n = 100$ and 500, respectively. In each case, the Gibbs sampling procedure by iteratively generating pseudo random variates from the conditional distributions described in Section 2 was conducted. The first 1000 iterations were discarded then one sample point in every 50 iterations was taken to form an approximate posterior sample of θ of size 1000. Tables 1 and 2, for $n = 100$ and 500, respectively, list the resulting averages of the Bayesian estimates, posterior standard deviations, and the estimated frequentist's mean square errors (MSE) of the parameters based on 1000 simulation runs. The last three columns are the (averaged) end points of the 95% credible intervals as well as the corresponding coverage probabilities. For each data set, we had also made 500 predictions in which the dependent variables were still simulated from $U(0, 3)$ and the results of the 500 predictions based on the same data sets are given in Table 3. The first two columns are the MSE's and the sample correlation coefficients (corr) between the averaged true and predicted values, obtained by (6), and the last column gives the coverage probabilities of the predictive intervals from the 1000 simulation runs. Figures 1 and 2 also show the 500 predictive intervals for $n = 100$ and 500, respectively, from only one data set in each case. It shows that our proposed method is quite accurate from the simulation results in parameters estimation as well as in prediction.

Table 1: Simulation results for the parameters estimation for $n = 100$.

Parameter	True	Estimate	S.E.	MSE	Lower	Upper	C.P.
β_1	1.1301	1.1322	0.1219	0.0131	0.8916	1.3708	0.957
β_2	2.5620	2.5673	0.1221	0.0121	2.3256	2.8065	0.965
β_3	1.2015	1.2015	0.1220	0.0101	0.9602	1.4406	0.981
β_4	1.5231	1.5192	0.1219	0.0097	1.2785	1.7583	0.986
β_5	3.0023	2.9997	0.1221	0.0117	2.7589	3.2385	0.964
σ^2	1	1.1866	0.3178	0.0658	0.8023	2.0076	0.949

Table 2: Simulation results for the parameters estimation for $n = 500$.

Parameter	True	Estimate	S.E.	MSE	Lower	Upper	C.P.
β_1	1.1301	1.1314	0.0478	0.0022	1.0373	1.2247	0.958
β_2	2.5620	2.5609	0.0477	0.0021	2.4671	2.6541	0.958
β_3	1.2015	1.1991	0.0478	0.0023	1.1049	1.2924	0.946
β_4	1.5231	1.5239	0.0477	0.0023	1.4299	1.6171	0.946
β_5	3.0023	3.0033	0.0477	0.0026	2.9093	3.0963	0.936
σ^2	1	1.0296	0.0762	0.0049	0.8983	1.1963	0.963

Next, We demonstrate a typhoon rainfall prediction model for Taipei, Taiwan based on the proposed mixture model. The observational data, collected and recorded hourly by the Central Weather Bureau of Taiwan, include the surface wind, the pressure, and the rainfalls in Taipei, the center maximum wind speed, the typhoon moving speed, direction, the distance between Taipei and the typhoon center, and the Julian day number function (Neumann 1992) of each typhoon that reached the area across between the east longitudes 120° and 125° and the north latitudes 20° and 28° from 1961 to 1994. The response

Table 3: Simulation results for 500 predictions.

n	MSE	corr	C. P.
100	1.0545	0.9678	0.9722
500	1.0126	0.9688	0.9541

variable is the rainfall after $\triangle t$ hours, and all the other eight variables mentioned above are the predictors. Yeh, Fan and Lee (2001) used linear regression models to the rainfall prediction based on the same data set. The models were constructed separately in each one by one degree sub-domain according to the latitudinal and longitudinal position of typhoons to remove the dependence. When a typhoon approached Taiwan, say within the area across between the east longitudes 120° and 125°, and the north latitudes 22° and 25°, it carried heavy rainfall and the usual linear regression models can be employed satisfactory. However, when the typhoon was far away from Taipei, it usually brought no rainfall at all. Indeed when a typhoon's center was in 122° and 125° east longitudes across from 20° to 22° or from 25° to 28° north latitudes, at least 50% of the data had zero rainfalls . Obviously, usual linear regression models are not appropriate in these areas, hence we will only focus on these areas here.

Within each sub-domain, the mixture regression model was constructed based on all $(\mathbf{X}, Y_{\triangle t})$, where $\mathbf{X}$ represents the observations of the above predictor variables and $Y_{\triangle t}$ represents the rainfall after $\triangle t$ hours. Similar Gibbs sampler procedure was performed in each sub-domain. The mean square errors by comparing each observation with its predicted value in all sub-domains considered are given in Table 4 for $\triangle t = 1$, 3 and 6 hours predictions along with their sample correlation coefficients (corr), respectively. It shows that the proposed mixture regression models resulted in smaller mean square errors, but high correlation coefficients, compared with those predicted by the linear regression models.

Table 4: The MSE(corr) for the rainfall predictions in Taipei.

	1 Hour	3 Hours	6 Hours
Mixture Model	17.87 (0.46)	16.47 (0.50)	12.20 (0.57)
Simple Regression	20.69 (0.40)	16.96 (0.46)	12.74 (0.53)

4 Conclusion and discussions

We develop a Bayesian mixture regression model where the data of the non-negative response variable contain many zero measurements. We also applied the model to make typhoon rainfall predictions in Taipei, Taiwan. Due to

the complexity of the typhoon data, the results might not be extremely satisfactory, but it was much more improved than the existing method. We have also tried to use logistic regression in the first stage to pre-estimate the possibility of positive measurement; and if high probability at the first stage was predicted, the second stage prediction using regression model with non-zero measurements only will be proceeded. However, the mixture regression approach seemed to perform better again. We believe valuable prior information can yield better results and the computational algorithm will be no harder with conjugate structure.

References

Albert, J. H., and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88, 669-679.

Casella, G., and George, E. I. (1992). "An Introduction to Gibbs Sampling." *American Statistician*, 46, 167-174.

Geman, S., and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." *IEEE Trans. Pattern Anal, Mach. Intell.*, 6, 721-740.

Ghosh, S. K., Mukhopadhyay, P., Lu, J. C., and Chen, D. (2002). Bayesian Analysis of Zero-Inflated Regression Models. *Technique Report.*

Gupta, P. L., Gupta, R. C., and Tripathi, R. C. (1996). "Analysis of Zero-Adjusted Count Data." *Computational Statistics and Data Analysis*, 23, 207-218.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd Edition. Chapman and Hall. London.

Neumann, C. J. (1992). "An alternate to the HURRAN tropical cyclone forecast system." *NOAA Technical Memorandum NWS SR-62*, 25.

Yeh, T. C., Fan, T. H. and Lee, Y. H. (2000). "Typhoon Rainfall Regression Predictions over Taiwan Area (I) The Linear Regression Model for

Predicting Rainfalls at Taipei (In Chinese)." *Atmospheric Sciences,* 29, 77-96.

Received: October 23, 2006

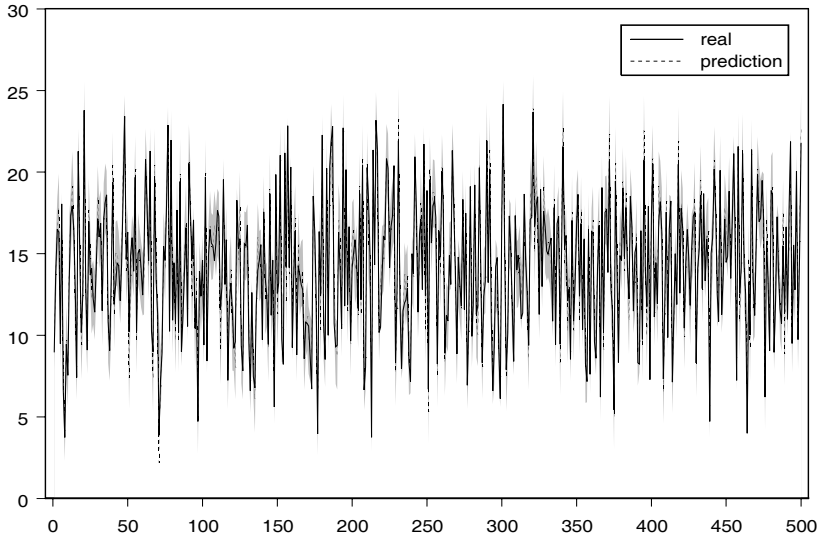

Figure 1: Simulation of 500 predictive intervals for $n = 100$.

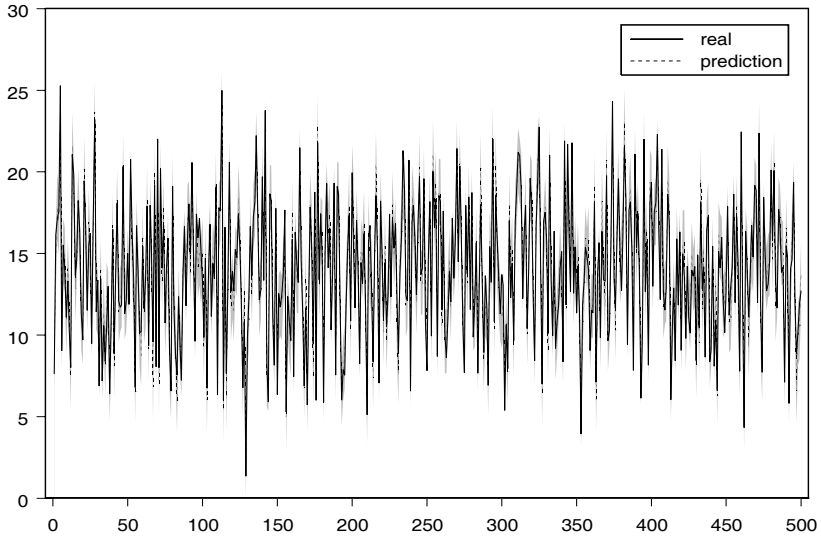

Figure 2: Simulation of 500 predictive intervals for $n = 500$.