OUTLIER DETECTION USING GENERALIZED LINEAR MODEL IN MALAYSIAN BREAST CANCER DATA

MARDZIAH BT NAWAMA
BALANCE OF PAYMENTS DIVISION
DEPARTMENT OF STATISTICS, MALAYSIA
29 SEPTEMBER 2017

LITERATURE REVIEW

- Outlier refers to an observation with abnormal properties compared to the others such as being surprisingly far from the main data set or having large residual. Current outlier detection methods, for example, distance based approach, looks for those data points which are far away from their neighbours, etc..

- Fisher (1993) summarized ways in which outliers can occur, due to mis-recording, sampling from a second population, or vagaries of sampling resulting in the occasional isolated values.

- In the Bayesian set up, Freeman (1980) defined an outlier as 'any observation that has not been generated by the mechanism that generated the majority of the observation in the data set. In other word, the detection of outliers in this framework is reduced to the problem of estimating the parameters of the distribution of the contaminated observations (Bayarri and Morales (2003)).

MODIFIED GLM WITH OUTLIER

- Let $\mathrm{N} = \{1, \ldots, N\}$ be a finite population with known N . For each unit $i \in$ N, we have the real valued response variable y_i and known $p \times 1$ vector of explanatory variables $\mathbf{x}_i$ where $\mathbf{x}_i^{'} = \begin{pmatrix} x_{i1} & \cdots & x_{ip} \end{pmatrix}$

- Assume that a random sample of size n is obtained with a number of suspected outliers. Let v^k be the set of all outlying observations, where k denotes the number of outliers. We consider the models with/without outliers based on GLM such that

$$p(y_i \mid \theta_i, \phi_i, \delta) = \begin{cases} \exp\{\phi_i(y_i\theta_i) + c_i(\theta_i, \phi_i) + d_i(y_i)\} & i \in \mathrm{N} - v^k \\ \exp\left\{\dfrac{\phi_i}{\delta}(y_i\theta_i) + c_i(\delta, \theta_i, \phi_i) + d_i(y_i)\right\} & i \in v^k \end{cases} \qquad (1)$$

where θ_i is a location parameter, ϕ_i and δ are scale parameters and $c(.)$ $d(.)$ are known functions. The parameters θ_i are modelled through a specific link function $h(.)$given by

$$h(\theta_i) = \mathbf{x}_i^{'}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, N$$

where $\boldsymbol{\beta}' = \begin{pmatrix} \beta_1 & \ldots & \beta_p \end{pmatrix}$ and the error components ε_i's are independently and normally distributed. Consequently, we can write $h(\theta_i) \mid \sigma^2 \sim N\left(\mathrm{x}_i^{'}\boldsymbol{\beta}, \sigma^2\right)$. Commonly, we usually assume that the model have the same mean for all observations but we expect to see higher variance for outlying observations, that is, when $\delta > 1$.

LOCAL BREAST CANCER DATA

- Breast cancer is the commonest cancer in Malaysia with the incidence rate for females of 47.4 per 100,000 women. Only recently the breast cancer specific survival information in Malaysia is available through the National Cancer Registry program under the purview of the Ministry of Health Malaysia. One of the established Breast Cancer Centre in the country is situated at the University of Malaya Medical Centre (UMMC) Kuala Lumpur.

- The UMMC is a 900 bed tertiary public hospital located in urban Kuala Lumpur. Prospective cohort studies of women with breast cancer treated in the UMMC are considered. The cohort comprises of patients who are diagnosed from year 1998 to 2002 and are followed up until March 2006. Patients underwent surgery and adjuvant chemotherapy under the care of general surgery and then followed by radiotherapy in UMMC.

- The information collected from the patients consists of race, age, date of diagnosis, and pathological characteristics of tumour. In addition, the survival times and status of patients are recorded at the end of the study. The mortality information is confirmed by referring to the record in the National Registry of Births and Deaths. The data set has been used in several other papers including Taib *et al.* (2008, 2011).

- For our case, we consider the size of tumour as independent variable and survival time to death in months from first diagnosis of the disease as dependent variable . Only patients who registered in 2000 with age 60 years old and above are considered. The scatter plot of the data with the exponential fitted curve is given in Figure 1. It can be seen that the data appear to follow the exponential distribution with one extreme observation is a candidate to be an outlier. It is of interest to identify the outlier using the modified model (1) in the Bayesian framework.

THE MODEL – PRIOR DISTRIBUTION

- The appropriate model to study the exponential relationship between y and x in the present data set corresponding to the general model (1) is given by

$$f(y_i \mid \theta_i, \phi, \delta) = \begin{cases} \phi\theta_i \exp(-\phi\theta_i y_i) & \text{for } i \notin v^k \\ \frac{\phi\theta_i}{\delta} \exp\left(-\frac{\phi\theta_i}{\delta} y_i\right) & \text{for } i \in v^k \end{cases}, \tag{2}$$

with the link function $\log\theta_i = \beta(x_i - \bar{x}) + \varepsilon_i$, where $\bar{x}$ is the mean of the sample. We intend to detect outlying observation using the hierarchical Bayesian approach. Hence, we consider the following hierarchical prior distributions of the parameters:

$$\left.\begin{array}{l} \sigma^{-2} \mid a_\sigma, b_\sigma \sim \text{gamma}\left(\frac{a_\sigma}{2}, \frac{b_\sigma}{2}\right), \qquad \beta \mid a_\beta, b_\beta \sim \text{gamma}\left(\frac{a_\beta}{2}, \frac{b_\beta}{2}\right), \\ \delta \sim \text{Uniform}(1, \delta_{\max}), \qquad \phi \mid a_\phi, b_\phi \sim \text{gamma}\left(\frac{a_\phi}{2}, \frac{b_\phi}{2}\right), \\ \log\theta_i \mid \sigma^2, \beta \sim N(x_i^* \beta, \sigma^2) \quad \text{where } x_i^* = (x_i - \bar{x}) \text{ and } p(v^k \mid k) = \binom{N}{k}^{-1} \end{array}\right\} \tag{3}$$

THE MODEL – POSTERIOR DISTRIBUTION

- Now, the joint likelihood function is given by

$$L\left(\mathbf{y}\mid\boldsymbol{\theta},\phi,\delta,v^{k}\right)=\prod_{i\notin v^{k}}\phi\theta_{i}\exp\left(-\phi\theta_{i}y_{i}\right)\times\prod_{i\in v^{k}}\frac{\phi\theta_{i}}{\delta}\exp\left(-\frac{\phi\theta_{i}}{\delta}y_{i}\right) \tag{4}$$

- Correspondingly, from the result obtained in equations (3) and (4), the full joint posterior distribution for the parameters $\boldsymbol{\theta},\sigma^{2},\beta,\phi,\delta,v^{k}$ is given by

$$p\left(\boldsymbol{\theta},\phi,\delta,\beta,\sigma^{2},v^{k}\mid\mathrm{y}\right)\propto\prod_{i\notin v^{k}}\phi\theta_{i}\exp\left(-\phi\theta_{i}y_{i}\right)\times\prod_{i\in v^{k}}\frac{\phi\theta_{i}}{\delta}\exp\left(-\frac{\phi\theta_{i}}{\delta}y_{i}\right)$$

$$\times\prod_{i=1}^{n}\frac{1}{\theta_{i}\left(2\pi\sigma^{2}\right)^{1/2}}\exp\left\{-\frac{1}{2}\left(\frac{\log\theta_{i}-x_{i}^{*}\beta}{\sigma}\right)^{2}\times\left(\sigma^{-2}\right)^{\left(\frac{a_{\sigma}}{2}-1\right)}\exp\left(-\frac{b_{\sigma}}{2\sigma^{2}}\right)\right\}$$

$$\times\beta^{\frac{a_{\beta}}{2}-1}\exp\left(-\beta\frac{b_{\beta}}{2}\right)\times\phi^{\frac{a_{\phi}}{2}-1}\exp\left(-\phi\frac{b_{\phi}}{2}\right)\times\frac{1}{\delta_{\max}-1}\times\left(\frac{N!}{(N-k)!k!}\right)$$

- It is clear that the full joint posterior distribution is intractable. Hence, we employ the MCMC sampling method, in particular, using Gibbs sampler with MH algorithm in the outlier detection procedure.

SAMPLING METHODS OF THE PARAMETERS

- The sampling methods for each of the parameters $\boldsymbol{\theta}, \sigma^2, \beta, \phi, \delta, v^k$ are given below:
- a) $\sigma^{-2} \sim inv\text{gamma}\left(\frac{n}{2} + \frac{a_\sigma}{2}, \frac{b_\sigma}{2} + \frac{1}{2}\sum_{i=1}^{n}\left(\log\theta_i - x_i^*\beta\right)^2\right)$
- b) $\beta \sim \text{gamma}\left(\frac{a_\beta}{2}, \frac{b_\beta}{2}\right)$
- c) δ given $\boldsymbol{\theta}, \phi, \sigma^2, \beta, v^k$ ~ uniform $(1, \delta_{\max})$
- d) $\phi \mid \sigma^2, \beta, \delta, \boldsymbol{\theta}, v^k \sim \text{gamma}\left(n + \frac{a_\phi}{2}, \frac{b_\phi}{2} + \sum_{i=1}^{n}\theta_i y_i\left(1 - \omega_i + \frac{\omega_i}{\delta}\right)\right)$
- e) $\theta_i \sim \text{lognormal}\left(x_i^*\beta, \sigma^2\right)$
- f) v^k

RESULTS

- The sampling methods for each of the parameters $\sigma^2, \beta, \phi, v^k$ are given below:

Figure 1: Plot of patients' survival time versus size of tumour

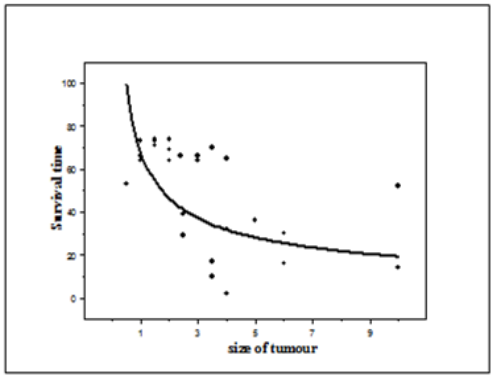


Figure 2: Histogram of the marginal posteriors for σ^2

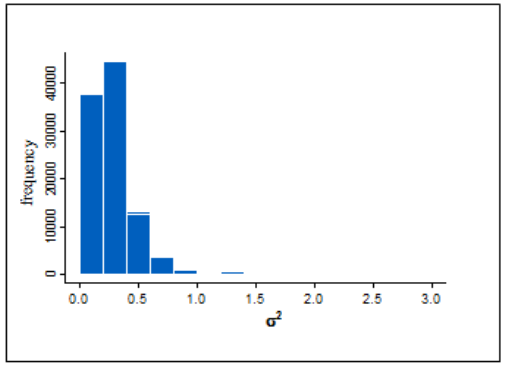


Figure 3: Histogram of the marginal posteriors for β

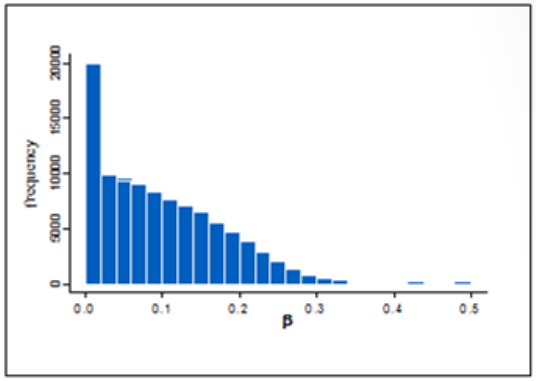


Figure 4: Histogram of the marginal posteriors for ϕ

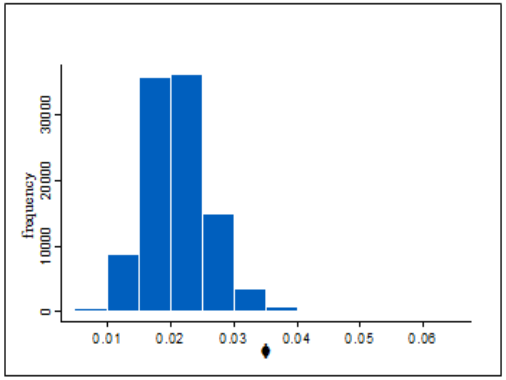
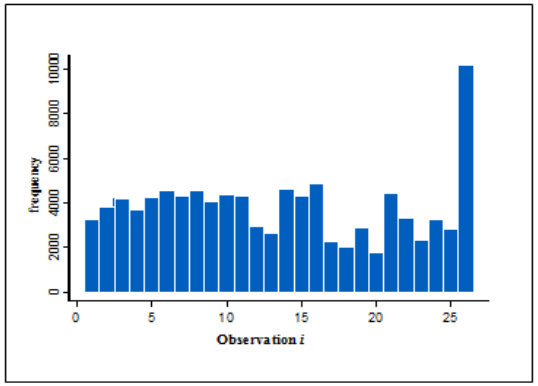


Figure 5: Probability for an observation being outlier in Breast Cancer data



DISCUSSIONS

- We use 100,000 iterations, with a burn-in of 50,000 iterations. It can be seen that the shape of the histograms of the three parameters resemble that of the gamma distributions. Given that there is one outlier, we identify observation 26, which has the highest probability (close to 0.20), as an outlier. This probability is distinctly higher if compared to the other probabilities corresponding to the other patients in the data.

- In survival data, many authors have tried to give specific meaning to the outlier due to the special features of the data. Collet (2003) referred an outlier in survival as an individual who has extremely long survival time, but the values of the explanatory variables suggested the individual should have died earlier, and vice versa. Therneau *et al.* (1990) and Nardi and Schemper (1999) associated outlier to individuals who "died too soon" or "lived too long", while Maller and Zhou (1994) identified outlier as individual who is already "immune" or "cured" after being released from prison.

- Using these definitions, patient 26 fits into the definition of outliers such that the patient's survival time is rather long even though the size of tumour for this patient is amongst the largest in the data set. Such identification enables the breast cancer specialists to monitor the background of such patients in finding the insight on factors that contribute to the improved survival life times for patients with similar prognosis.

CONCLUSIONS

- It is considered the problem of detecting outlier using Bayesian approach in generalized linear model.

- It is shown that with the choice of prior distribution for the parameters, we can obtain the information from samples generated using MCMC sampling, in particular using the Gibbs sampler with MH algorithm.

- When applied to the local breast cancer data, observation 26 who has a large size of tumour but with long survival time which is 52 months from diagnosed time, is identified as an outlier.