MODEL SELECTION AND CLAIM FREQUENCY FOR WORKERS' COMPENSATION INSURANCE

BY

JISHENG CUI, DAVID PITT AND GUOQI QIAN

ABSTRACT

We consider a set of workers' compensation insurance claim data where the aggregate number of losses (claims) reported to insurers are classified by year of occurrence of the event causing loss, the US state in which the loss event occurred and the occupation class of the insured workers to which the loss count relates. An exposure measure, equal to the total payroll of observed workers in each three-way classification, is also included in the dataset. Data are analysed across ten different states, 24 different occupation classes and seven separate observation years. A multiple linear regression model, with only predictors for main effects, could be estimated in $2^{23+9+1+1} = 2^{34}$ ways, theoretically more than 17 billion different possible models! In addition, one might expect that the number of claims recorded in each year in the same state and relating to the same occupation class, are positively correlated. Different modelling assumptions as to the nature of this correlation should also be considered. On the other hand it may reasonably be assumed that the number of losses reported from different states and from different occupation classes are independent. Our data can therefore be modelled using the statistical techniques applicable to panel data and we work with generalised estimating equations (GEE) in the paper. For model selection, Pan (2001) suggested the use of an alternative to the AIC, namely the quasi-likelihood under independence model criterion (QIC), for model comparison. This paper develops and applies a Gibbs sampling algorithm for efficiently locating, out of the more than 17 billion possible models that could be considered for the analysis, that model with the optimal (least) QIC value. The technique is illustrated using both a simulation study and using workers' compensation insurance claim data.

KEYWORDS

Model selection; QIC; Longitudinal study; Workers' compensation insurance; Gibbs sampler.

Astin Bulletin 40(2), 779-796. doi: 10.2143/AST.40.2.2061136

1. Introduction

For the past fifteen years actuaries in general insurance have routinely used generalised linear models (GLMs) as a statistical modelling tool. The outputs from such analyses have been used to inform their premium and reserving calculations. The most authoritative text on GLMs is McCullagh and Nelder (1989). An early paper on the application of GLMs in general insurance was Brockman and Wright (1992). They considered the analysis of motor vehicle insurance claim data and highlighted the potential for simpler and possibly better-informed assessment of risk premia using GLMs. Haberman and Renshaw (1996) provided a very informative summary of the uses that actuaries have made of GLMs which include the determination of outstanding claims provisions in general insurance, the description of renewal rates for insurance policies, the description of observed mortality rates and the prediction of future mortality rates for insured lives.

When using a GLM we assume that the dependent variable of the regression, which could be for example claim frequencies, claim severities or insurance policy renewal rates, is a collection of mutually independent random variables. The dependent variable is assumed to come from the linear exponential family of distributions. Associated with each of the observed values of the dependent variable is a set of regressor variables. The mean of the dependent variable, conditional on the values taken by the observed regressor variables, is set equal to some monotonically increasing transformation of a linear predictor formed from the regressor variables. Akaike's Information Criterion (AIC), Akaike (1974), has commonly been used as a method for selecting between a collection of GLMs that have been fitted to a given dataset.

In workers' compensation insurance, where workers are paid compensation for workplace related injuries, individual payments are made to workers over time. These payments may replace lost income, they may cover medical expenses or the costs of rehabilitation or retraining. Such repeated observations on individuals are common not only in insurance but also in many studies in the social sciences and in econometrics where such data are known as panel data. In the workers' compensation insurance example above, a panel refers to an individual on which many observations may be made. It is useful to model the payments to policyholders under such insurance arrangements using regression type analysis. Ordinary GLM analysis, as described in the previous paragraph, is not appropriate for panel data, because of the correlation we would anticipate between the size of payments made to the same individual over time. Another example of panel data in the area of workers' compensation insurance, which we shall explore in this paper, relates to aggregate level claim frequency analysis. Observations from the same occupation class and state across different years are assuemd to be correlated. The analysis is said to run at an aggregate level because data are grouped by state, occupation class and year. In the context of analysing correlated dependent variable observations, Frees et al. (1999) and Antonio and Berlant (2007) discuss the use of subject

specific (random effects) generalised linear mixed models (GLMMs). In this paper we consider the use of population-averaged (PA) or so-called marginal models, Liang and Zeger (1986), for incorporating within-subject correlations between the observations from the same panel.

Pan (2001) argues that the AIC measure is not suitable for selecting between population-averaged, generalised estimating equation type models which are the models employed in this paper. He proposes a new measure, the quasi-likelihood under independence model criterion (QIC). Cui and Qian (2007) discuss the application of this model selection criterion to biomedical data. In this paper we develop a methodology, using the Gibbs sampler, for determining quickly the model with the lowest (best) value of QIC. The Gibbs sampler has been used in the past for model selection using the AIC with both ARIMA models, Qian and Zhao (2007), and logistic regression models, Qian and Field (2000).

The paper is organised as follows. Section 2 describes the workers' compensation insurance claim frequency dataset used in the paper. Section 3 is devoted to methodology. The GEE population-averaged models along with the QIC selection criterion and the Gibbs sampler algorithm for locating an optimal model are described. Section 4 assesses the ability of the Gibbs sampler algorithm to select an optimal model using a simulation study. Section 5 provides the results of our analysis of the workers' compensation insurance dataset using the methodology described in Section 3. Section 6 concludes the paper.

2. Data

As mentioned in the introduction, the dataset analysed in Section 5, which provides the motivation for the methodological development presented in Section 3 of this paper, are given in Klugman (1992) as Data Set 4. Antonio and Beirlant (2007) also discuss these data in Section 2 of their paper. We do not repeat the exploratory data analysis that they conducted on this dataset here but instead we emphasise the main characteristics of the dataset which make it suitable for illustrating our methodology.

The dataset includes the number of observed losses on workers' compensation insurance policies split by year, occupation class and state. Our analysis works from data on 24 different occupation classes and ten different states. Data are recorded annually over seven years. For each observation in the dataset, a measure of exposure, equal to the total payroll of all insured individuals in that particular classification of state, occupation class and year is also recorded. The first ten records of the dataset are shown in Table 1. For example, the payroll total in year 2 for those in the first state and the first occupation group was \$33.779 times ten million (payroll figures have been divided by 10^7 for ease of readability in Table 1). The observed number of losses in this category was 4. This dataset has been analysed in a Bayesian framework by Klugman (1992) and later by Scollnik (1996) and Makov et al. (1996).

TABLE 1

WORKERS' COMPENSATION CLAIM FREQUENCY DATA. FIRST 10 RECORDS SHOWN.

Year	State	Occupation Class	Payroll	Count of Losses
1	1	1	32.322	1
2	1	1	33.779	4
3	1	1	43.578	3
4	1	1	46.686	5
5	1	1	34.713	1
6	1	1	32.857	3
7	1	1	36.600	4
1	1	2	45.995	3
2	1	2	37.888	1
3	1	2	34.581	0

Antonio and Beirlant (2007) present a scatterplot matrix of the observed number of losses by year of occurrence. The matrix shows very clear positive correlation between the number of losses reported in year i and in year j for $i, j \in \{1, 2, ..., 7\}$. In addition they include a selection of the number of losses recorded plotted against the year. The positive correlation between observed number of losses and year of occurrence for losses occurring in the same state and for the same occupation class is evident.

3. METHODOLOGY

3.1. GLMs and the Extension to Population-Averaged GEE Models

Population-averaged Generalised Estimating Equation (PA-GEE) models, introduced by Liang and Zeger (1986), are an extension of GLMs. They address the issue of panel or subject level correlation. In the context of the workers' compensation claim frequency data considered in this paper, we allow for dependence between the observed claim frequency values across the seven years of data for claims recorded in the same state and the same occupation class.

In this section, we provide a quick review of the procedure for estimating the coefficients in a GLM. We then describe the PA-GEE models and cast them as a natural extension of GLMs. McCullagh and Nelder (1989) describe the maximum likelihood procedure which is applied to the estimation of regression model coefficients in the GLM setting. In the GLM regression model framework, the dependent variables are assumed to be drawn from the linear exponential family of probability distributions. These variables are assumed

to be mutually independent. The probability density (or mass) function for a variable Y from the linear exponential family is

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right), \tag{1}$$

where θ is the natural parameter, ϕ is the dispersion parameter and $a(.)$, $b(.)$ and $c(.)$ are functions that take particular forms depending on the particular distribution being considered. For example, for the Poisson distribution with mean λ, it can be shown that $\theta = \ln\lambda$, $b(\theta) = \lambda$, $\phi = a(\phi) = 1$ and $c(y;\phi) = -\ln y!$. It can also be shown that the natural parameter θ is related to the mean of Y, specifically $E[Y] = b'(\theta)$. Similarly the dispersion parameter ϕ is related to the variance of Y through $Var(Y) = b''(\theta)\,a(\phi)$. Since the second derivative term $b''(\theta)$ is connected to both the mean and variance of Y, we have the notation $V(\mu) = b''(\theta)$.

For a set of n independent observations from the linear exponential family of distributions, the log-likelihood function, where we consider the observed values of Y as given and the parameters as unknown is

$$\ln L(\theta_i,\phi \mid y_1, \ldots, y_n) = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right). \tag{2}$$

Note that a subscript i has been added to each of the observed values of the variables Y and also to the natural parameter θ. Note also that the dispersion parameter is not indexed and is estimated only once for the entire model rather than separately for each individual data record. Heller et al (2007) consider joint modelling of the natural and dispersion parameters using motor vehicle insurance data. The natural parameters θ_i are related to the means of the dependent regression variable μ_i. In the GLM we restrict the mean parameter to be of the form

$$g(\mu_i) = \eta_i = \sum_{j=1}^{p}\beta_j x_{ji}, \tag{3}$$

where the β_j are regression coefficients and the x_{ji} are the observed independent regression variables for the ith observation. Note that very often x_{i1} would be 1, as the models very often include an intercept.

The maximum likelihood estimation procedure involves equating the vector of first order partial derivatives (where derivatives are taken with respect to each of the regression model coefficients) of the log-likelihood function to the zero vector and solving for the resultant coefficient estimates. The equations formed in this process are known as the score equations and take the general form

$$\left(\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{n}\frac{y_i - \mu_i}{a(\phi)\,V(\mu_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i x_{ji}\right)_{j=1,\ldots,p} = [0]_{p\times 1}. \tag{4}$$

Suppose now that instead of having n independent observations from the variables Y_i, we have n panels of data where the ith panel contains n_i data points. Dependent variable observations are denoted $\{y_{it}\}_{i=1,\ldots,n;\ t=1,\ldots,n_i}$. Suppose that all variables, including those in the same panel, are mutually independent. The score equations are then

$$\left(\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{n}\sum_{t=1}^{n_i} \frac{y_{it}-\mu_{it}}{a(\phi)\,V(\mu_{it})}\left(\frac{\partial \mu}{\partial \eta}\right)_{it} x_{jit}\right)_{j=1,\ldots,p} = [0]_{p\times 1}, \tag{5}$$

where μ_{it} is the mean of the tth observation from the ith panel, and x_{jit} is the jth covariate value for the tth observation from the ith panel. Following Hardin and Hilbe (2003) these score equations can be written in matrix form as

$$\left(\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{n} x_{ji}^{T} D\left(\frac{\partial \mu}{\partial \eta}\right)\left[V(\mu_i)\right]^{-1}\left(\frac{y_i-\mu_i}{a(\phi)}\right)\right)_{j=1,\ldots,p} = [0]_{p\times 1}, \tag{6}$$

where $D(\,)$ represents an $n_i \times n_i$ diagonal matrix. Note that $V(\mu_i)$ is also an $n_i \times n_i$ diagonal matrix which can be written as

$$V(\mu_i) = \left[D\big(V(\mu_{it})\big)^{\frac{1}{2}} I_{n_i\times n_i} D\big(V(\mu_{it})\big)^{\frac{1}{2}}\right]. \tag{7}$$

The transition from GLM to PA-GEE model replaces the identity matrix in the above decomposition with a correlation matrix for the observations within a panel. We therefore have for the PA-GEE model that

$$V(\mu_i) = \left[D\big(V(\mu_{it})\big)^{\frac{1}{2}} R(\rho)_{n_i\times n_i} D\big(V(\mu_{it})\big)^{\frac{1}{2}}\right]. \tag{8}$$

We note that the resulting equations (6) need not correspond to series of first order partial derivatives of a log-likelihood function. The specification of $R(\rho)$ need not match up with a specific likelihood function. In addition, it is possible to express the relationship between the mean and the variance of the dependent variable in a way that does not coincide with that for a standard linear exponential family probability density or mass function.

Different forms of the correlation matrix, $R(\rho)$ have been proposed. In this paper we will consider independent, exchangeable and autoregressive correlation of order 1 (AR(1)). The independent correlation structure has $R(\rho)$ equal to the identity matrix, the exchangeable correlation structure has

$$\mathbf{R}(\rho) = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix},$$

while the AR(1) correlation structure has

$$\mathbf{R}(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n_i-1} \\ \rho & 1 & \rho & \cdots & \rho^{n_i-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n_i-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \cdots & 1 \end{pmatrix}.$$

The estimation of a PA-GEE model involves solving the system of equations (6). The model outputs include parameter estimates for each regression coefficient and an estimated correlation matrix $R(\rho)$.

3.2. QIC and PA-GEE Model Selection

Model selection for GLMs is frequently performed by selecting the model with the lowest value of Akaike's Information Criterion (AIC). The AIC is calculated as

$$AIC = -2\ln L + 2p, \tag{9}$$

where $\ln L$ denotes the log-likelihood for the fitted model and p is the number of parameters included in this model.

The AIC model selection criterion is only valid for comparing models that are estimated based on the maximisation of a log-likelihood function. As described in Section 3.1, the PA-GEE models, which are estimated based on a modified quasi-likelihood function, do not necessarily use a log-likelihood function. Pan (2001) modified the AIC so that it could be employed for selecting amongst PA-GEE models. His proposed model selection criterion is called the quasi-likelihood under independence model criterion (QIC) and is calculated as

$$QIC = -2Q(\hat{\mu}; I) + 2\text{trace}\,(\hat{\Omega}_I \widehat{V}_R), \tag{10}$$

where I represents the independence correlation structure under which the quasi-likelihood $Q(.)$ is calculated. The negative Hessian $\hat{\Omega}_I$ is estimated from the quasi-likelihood under the independence correlation structure also. The robust regression coefficients $\hat{\beta}$ and their associated variance matrix $\widehat{V}_R$ are estimated from the estimating equations (6) and (8) with the specified working correlation matrix R. Therefore, in order to calculate the QIC value we need to run the GEE model twice, once under the independence correlation structure and then again under the specified working correlation structure.

The QIC value can be used to select both the best correlation structure and the best fitting GEE model. A correlation structure with the smallest QIC value is usually chosen as the best correlation structure. A mean response model with the smallest QIC value can be considered as the best fitting model.

3.3. Gibbs Sampler and Model Selection Involving Many Candidate Models

In this section we describe how the Gibbs sampler, Geman and Geman (1984), can be used to select the optimal PA-GEE model based on the QIC model selection criterion. We describe the methodology for doing this by making reference to the workers' compensation insurance claim frequency dataset described in Section 2.

Suppose $y_i = (y_{i1}, y_{i2}, \ldots, y_{i7})^T$ is a vector containing the observed numbers of claims in a particular state and a particular occupation class over the seven years of observation. Given that we have data relating to ten states and 24 occupation classes, we therefore have 240 separate observation vectors y_i. Denote the mean vector for the response variables $E[Y_i] = \mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{i7})^t$. Denote the variance-covariance matrix for the response variables as $Var(Y_i) = \Sigma_i$. Let $g(.)$ denote a link function.

Associated with each element of each response variable vector, we have a vector of independent predictor variables. One variable relates to year, and one variable relates to payroll. Since state is a categorical variable having 10 states, we need 9 indicator variables to encode it. Also 23 indicator variables are needed to encode the occupation class since it is a categorical variable with 24 occupation classes observed. In total there are 34 predictor variables relating to panel i and year t which are stored in a column vector x_{it}. The predictor variables relating to panel i are collected into a design matrix $X_i = (x_{i1}, x_{i2}, \cdots, x_{i7})^t$. Note X_i does not include the model intercept column of 1's.

The full model for the number of claims occurring is $g(\mu_i) = \beta_0 1_7 + X_i\beta$ where β_0 is the intercept coefficient, 1_7 is a column vector of seven 1's, and β is a column coefficient vector of length 34. The variance-covariance matrix Σ_i is modelled using one of the structures from Section 3.1. Clearly, the full model can be reduced most of the time because the response variable may not be significantly dependent on some of the predictors. This has induced the issue of model selection.

A sub model, denoted M_α, is a model that includes the intercept plus a subset of the 34 available predictor variables in the linear predictor. Let α be a subset of $\{1, 2, \ldots, 34\}$. So, α represents the columns of the design matrix X_i that are included in the sub model under consideration. The predictor for the mean of the dependent variable under this sub model is written

$$M_\alpha : g(\mu_i) = \beta_0 1_7 + X_{i\alpha}\beta_\alpha, \tag{11}$$

where $X_{i\alpha}$ consists of those columns of X_i indexed by α and β_α is similarly defined.

To help in the description of the Gibbs sampler, we define an alternative method for representing a sub model, M_α. Define a $1 \times K$ ($K = 34$ for the dataset here) vector $\gamma_\alpha = (\gamma_{\alpha,1}, \ldots, \gamma_{\alpha,34})$, where $\gamma_{\alpha,k} = 1$ if $k \in a$ and $\gamma_{\alpha,k} = 0$ if $k \notin \alpha$, $k = 1, \ldots, 34$. The domain of γ_α is $\{0, 1\}^{34}$, which represents all possible sub models. This model space has 2^{34} points!

Suppose we use QIC to perform model selection. The best model, denoted γ^* would be

$$\gamma^* \equiv \gamma_{\alpha^*} = \arg \min_{\gamma \in \{0,1\}^{34}} QIC(\gamma). \tag{12}$$

That is, we choose as our optimal model the one which has the minimum value of the QIC model selection criterion amongst all 2^{34} models in our model space.

In order to speed up the process of locating this model with the minimum QIC compared with the time that would be required to do an exhaustive search, we apply the method of Gibbs sampling. Define

$$P_{QIC,\lambda}(\gamma) = \Delta^{-1} e^{-\lambda QIC(\gamma)}, \quad \gamma \in \{0,1\}^{34}, \tag{13}$$

where $\Delta = \sum_{\gamma \in \{0,1\}^{34}} e^{-\lambda QIC(\gamma)}$ and $\lambda > 0$ is a tuning parameter. It is clear that $P_{QIC,\lambda}(\gamma)$ is a multivariate discrete probability mass function defined over the space $\{0,1\}^{34}$, and that

$$\gamma^* = \arg \min_{\gamma \in \{0,1\}^{34}} QIC(\gamma) = \arg \max_{\gamma \in \{0,1\}^{34}} P_{QIC,\lambda}(\gamma). \tag{14}$$

Suppose now that we were able to generate a sample of γ values from $\{0,1\}^{34}$ based on the probability function $P_{QIC,\lambda}(\gamma)$. We know that γ^* has the highest probability of appearing in this sample. We would therefore expect γ^* to be generated early (after a suitable burn-in period) and most frequently in the sample. Therefore, it is with probability converging to 1 in the limit that the optimal model $\hat{\gamma}$ found from the generated sample of γ is the global optimal model γ^*. In other words, $\hat{\gamma}$ would be a very good (i.e. consistent) estimator of γ^* in practice. This implies that, in finite sample situations, $\hat{\gamma}$ is either γ^* itself, which often is the case if $\mathrm{QIC}(\gamma^*)$ is noticeably smaller than QIC values of other sub-models, or captures all significant variables in γ^* if the function $\mathrm{QIC}(\gamma)$ is fairly level. The optimal properties of $\hat{\gamma}$ in related model selection problems have been discussed in Qian (1999), Qian and Field (2000) and Qian and Zhao (2007).

As mentioned previously, the method of generating a sample $(\gamma_1, \gamma_2, \ldots, \gamma_Q)$ from $\{0,1\}^{34}$ involves the Gibbs sampler. The simulation algorithm adopted is given below:

- Start from an initial value of γ, say $\gamma_0 = (1, \ldots, 1) \in \{0,1\}^K$.
- Suppose $\gamma_0, \gamma_1, \ldots, \gamma_q$ have already been generated, with their QIC values also having been calculated. Write $\gamma_q = (\gamma_{q,1}, \ldots, \gamma_{q,K})$ and $\gamma_{q,l:m} = (\gamma_{q,l}, \gamma_{q,l+1}, \ldots, \gamma_{q,m})$ with $1 \le l \le m \le K$. Note that $\gamma_{q,a:b} = \emptyset$ if $a > b$.

- For $k = 1, \ldots, K$ do the following loop

  Calculate $\mathrm{QIC}(\gamma_{q,1:(k-1)}, 1-\gamma_{q,k}, \gamma_{q,(k+1):K})$ where $(\gamma_{q,1:(k-1)}, 1-\gamma_{q,k}, \gamma_{q,(k+1):K})$ is the same 0-1 indicator vector as γ_q except that its kth component equals 1 if $\gamma_{q,k} = 0$ and 0 if $\gamma_{q,k} = 1$.

  Calculate the conditional probability

$$\xi_k = P_{\mathrm{QIC},\lambda}\left(\gamma_{q,k} = 1 \mid \gamma_{q,1:(k-1)}, \gamma_{q,(k+1):K}\right)$$

$$= \frac{\exp(-\lambda\,\mathrm{QIC}(\gamma_{q,1:(k-1)}, 1, \gamma_{q,(k+1):K}))}{\exp(-\lambda\,\mathrm{QIC}(\gamma_q)) + \exp(-\lambda\,\mathrm{QIC}(\gamma_{q,1:(k-1)}, 1-\gamma_{q,k}, \gamma_{q,(k+1):K}))}$$

  and generate a random number, denoted as $\gamma'_{q,k}$, from the Bernoulli distribution with ξ_k as the probability of 'success'

  Update $\gamma_q = (\gamma_{q,1:(k-1)}, \gamma'_{q,k}, \gamma_{q,(k+1):K})$ and $\mathrm{QIC}(\gamma_q)$.

  At the end of each loop, update $\gamma_{q+1} = \gamma_q$.

- Continue this procedure until $\gamma_1, \ldots, \gamma_Q$ are generated.
- The generated $\gamma_1, \ldots, \gamma_Q$ constitute a Markov chain which, when Q is sufficiently large, will become stationary with the equilibrium distribution being $P_{\mathrm{QIC},\lambda}$. The generated Markov chain, after a burn-in period, can serve as a random sample from $P_{\mathrm{QIC},\lambda}$.

The next two sections are devoted to an illustration of this methodology using, in turn, simulated data and workers' compensation claim frequency data.

4. Simulation Study

In this section we report the results of a simulation study designed to assess the ability of the Gibbs sampler algorithm described in the previous section to locate an optimal model. We simulate data according to a PA-GEE model from the Poisson family with independent observations within the panels. The model linear predictor is of the form

$$\log(E(\text{number of losses})) = \beta_0 + \sum_{i=1}^{23} \beta_i x_i + \sum_{i=1}^{9} \gamma_i z_i, \tag{15}$$

where the β_i parameters relate to the intercept and the regression coefficients for occupations while the γ_i parameters relate to the regression coefficients for the states. The assumed values of these parameters are shown in Figure 1 where the year variable is given first followed by the state variables and finally the occupation variables. The simulated dataset is the same size and has the same values for the predictor variables as in the claim frequency dataset described in Section 2.

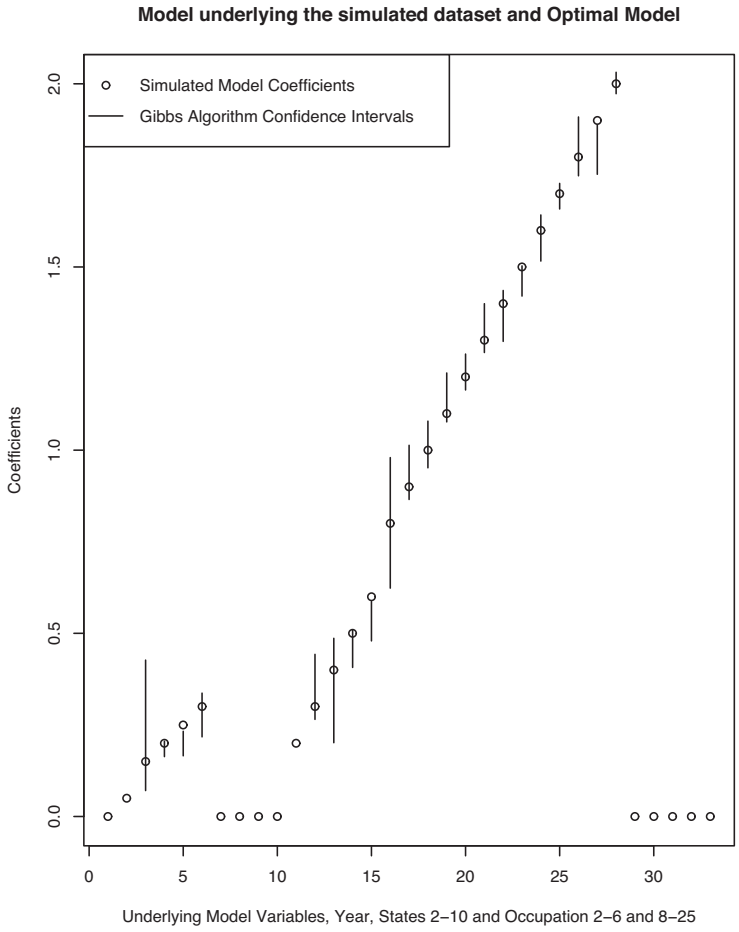

FIGURE 1: Simulation Study results.

Using this simulated set of data, we apply the algorithm described in Section 3.3 to determine the optimal PA-GEE model using the QIC as our indicator of model quality. We apply the Gibbs sampling algorithm assuming, in turn, independence, exchangeable and AR(1) correlation within panels.

Applying our Gibbs sampler model selection algorithm, assuming independence of observations within panels we find that the optimal model has a QIC of 1349.9. In fact the top five models, based on our Gibbs algorithm, have QIC values all less than 1350.5. Figure 1 shows the model coefficients (with a circle) used to simulate the claims data along with 95% confidence intervals for each regression coefficient in the optimal model (the model with minimum QIC value from the application of our Gibbs sampler algorithm). These confidence intervals are shown as vertical bars. These confidence intervals were

determined from the large sample distribution of the maximum likelihood estimators for the regression coefficients in the PA-GEE model. Analysis of Figure 1 shows that the QIC Gibbs sampler algorithm has chosen as optimal a model which estimates the regression coefficients to be very close to those used in the model from which the data were simulated. The only small potential anomolies relate to the impact of being in occupation group 2, where the model underlying the simulation had a regression coefficient equal to 0.2 while this term failed to appear in the model estimated as optimal by our model selection algorithm. In relation to this, we note that the impact of being in Occupation Group 2 was included in the three out of the five most optimal models identified by our model selection algorithm. Also evident from Figure 1 is that the impact of being in State 2 is not included in the model identified as optimal by the model selection algorithm but the model underlying the simulation of the data employed here had a regression coefficient for State 2 of 0.05. We note from Figure 1 that the 95% confidence intervals for the impacts of being in the various occupation groups and states go very close to overlapping (or do indeed overlap) the underlying regression coefficient employed in the simulation. We also note that with the large number of confidence intervals being compared here, it is appropriate to compute wider confidence intervals, using for example the method of Bonferroni, see Kutner et al (2005). Such wider intervals would cover the underlying model coefficients used to create the simulated dataset.

It is also important to note that while a model has been used to simulate the claim data from, it is indeed possible, particularly with a moderately sized dataset being employed here, that the simulated data may give a different message to that of the model employed to generate it. So while the investigation summarised in Figure 1 is useful, it can be complemented by determining the proportion of models chosen by our QIC algorithm, in the 1200 models considered after burn in (see Section 5 for discussion of this). Figure 2 is a plot of the proportion of Gibbs sampler models which include each of the possible regression coefficients against the actual value of the regression coefficient assumed in the model used to generate the simulated dataset. We see that the Gibbs algorithm performs very well in that it includes statistically significant regression coefficient variables for the parameters that indeed generated the data in the vast majority of cases.

Our simulated dataset was created on the basis of independence between observations within a panel. Nevertheless, it is of interest to employ our QIC Gibbs sampler algorithm to determine the optimal model under each of an exchangeable and AR(1) assumed correlation between observations within a panel. For the exchangeable correlation for observations in a panel, we note that the correlation estimated in our optimal model is $\rho = -0.0127$ which is very close to zero. Similarly for the AR(1) correlation for observations in a panel, the lag 1 autocorrelation estimated for our optimal model is $\rho = 0.0231$. This is very close to zero and is in agreement with the model underlying our simulated dataset.

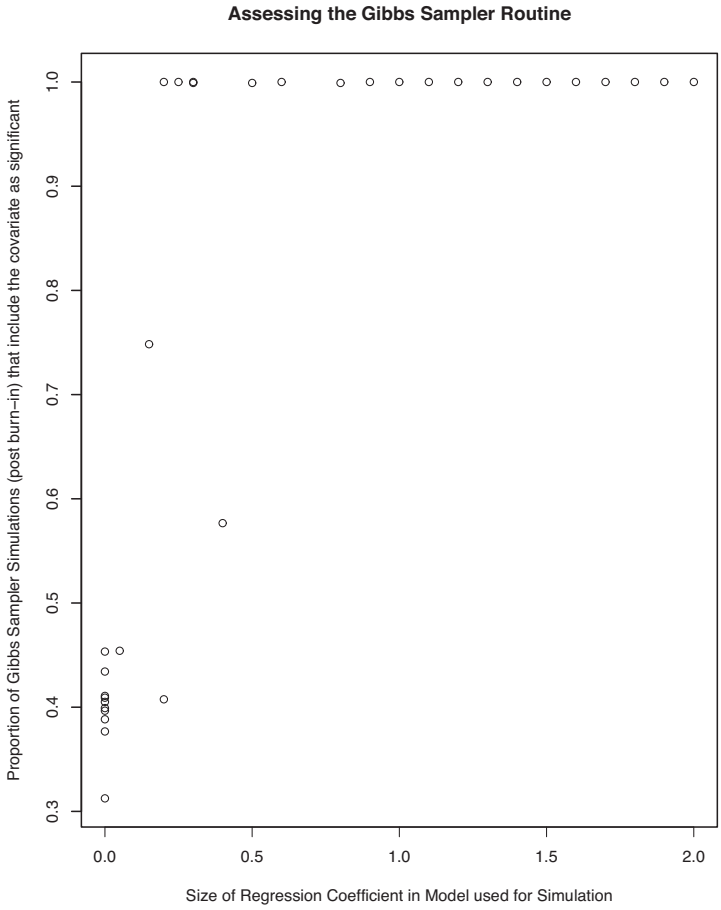

FIGURE 2: Ability of the Gibbs Sampler Algorithm to select models in agreement with the model used for the simulation

5. CLAIM FREQUENCY ANALYSIS

In this section we report the results of applying the Gibbs Sampler QIC model selection methodology from Section 3 to the US workers' compensation insurance dataset described in Section 2. We propose a PA-GEE model from the Poisson family, of the form

$$\log(E(\text{Number of Claims})) = \log(\text{payroll}) + \beta_0 + \sum_{j=1}^{33} \beta_j X_j, \qquad (16)$$

where X_1 relates to the observation year, and, noting that occupation class 7 has zero exposure and so is not included in the coding, we have

$$X_j = \begin{cases} 1, & \text{if state } j,\ j = 2, 3, ..., 10 \\ 0, & \text{otherwise} \end{cases}$$

$$X_j = \begin{cases} 1, & \text{if occ}(j-9) = 1,\ j = 11, 12, ..., 15 \\ 0, & \text{otherwise} \end{cases}$$

$$X_j = \begin{cases} 1, & \text{if occ}(j-8) = 1,\ j = 16, 17, ..., 33 \\ 0, & \text{otherwise} \end{cases}$$

Note that the state variable has 10 categories, thus one would need at most 9 indicator variables in order for a sufficient coding of it. Similarly, at most 23 indicator variables would be needed to sufficiently encode the occupation class variable. Certainly, there are many ways to define the indicator variables required. For example, different indicator variables would result if choosing different reference categories. Also the number of indicator variables required can be reduced if some categories of a categorical variable are merged. Clearly, which categories are used as the reference ones and which categories can be merged are issues that can be addressed in a model selection procedure. However, we will ignore these issues in this paper in order to convey the idea of Gibbs sampling selection in a concise way. Nevertheless, a more complicated model selection framework may be designed to allow these issues to be adequately addressed where the Gibbs sampling technique can still be applied.

We assess the suitability of independence, exchangeable and AR(1) assumed correlation for observations within panels. For the purposes of describing the process of implementing the methodology from Section 3, we assume that we are working with the AR(1) correlation within panels. Beginning with the full model, containing all of the 32 indicator variables for the categorical variables state and occupation class, we estimate the parameters of PA-GEE model with AR(1) within panel correlation and predictor of the form (16). We then apply the algorithm given at the end of Section 3.3 to draw a sample of size 5000 from $P_{\text{QIC},\lambda}(\gamma)$. Before we determine the model from our sample of 5000 which has the lowest QIC value, and is therefore optimal in our simulation, we assess the convergence of the Gibbs sampler. When we apply the Gibbs sampler to the problem of simulating values from the multivariate discrete probability mass function $P_{\text{QIC},\lambda}(\gamma)$, the initial values generated in our chain of simulated values will not be genuine simulated values from this distribution. It is known that the Gibbs sampler simulated values only approach the target distribution, in this case $P_{\text{QIC},\lambda}(\gamma)$, after a suitable number of iterations of the algorithm have been performed. It is therefore common practice, particularly in the application of Bayesian statistics where the Gibbs sampler is routinely adopted, to discard the values generated from the Gibbs sampler from the early iterations of the algorithm. This period of iterations, where the outputted values are discarded, is known in the literature as the "burn-in period". We will work

with a burn-in period of 3000 iterations, leaving us with, potentially (if the algorithm has indeed converged) 2000 simulations from $P_{\mathrm{QIC},\lambda}(\gamma)$.

One method for assessing convergence, discussed in Qian and Field (2000), is to apply the well-known chi-squared test of association. The idea behind the application of this test to the assessment of the convergence of the Gibbs sampler algorithm is simple. We take our 2000 output values from our Gibbs sampler, which relate to after the burn-in period, and divide them up into 4 bins. The first bin contains outputs 3001 to 3500, the second bin contains outputs 3501 to 4000, the third bin contains outputs 4001 to 4500 and the fourth and final bin contains outputs 4501 to 5000. For each bin, we count the number of QIC output values which fall into each of four intervals. These four intervals are determined by considering the minimum and maximum generated QIC values in our sample of 2000 values and dividing this range up into four bins of equal expected frequency. We then have a complete 4×4 contingency table for our QIC values. We use the chi-squared test of association to determine whether or not there is a relationship between position in the chain, as recorded by the four bins of observations and the relative magnitudes of the QIC values, as recorded by the bins used to count the number of outputs in each range of the QIC values. This methodology was applied to the AR(1) model, 2000 QIC simulations and the results are shown in Table 2. Note that in Table 2, Group A contains QIC outputs in the first bin, that is outputs from 3001 to 3500 in the chain, and Groups B, C and D contain the QIC outputs in the second, third and fourth bins described above. The range of QIC values, recorded in each bin, are self explanatory from the table. Observed counts are recorded in each cell along with expected counts (under the hypothesis of no association between position in chain and magnitude of simulated model QIC values) are recorded in parentheses in each cell. From this table, the observed chi-squared test statistic value is 12.6 which is below the critical value, at the 5% significance level, of 16.9. The *p*-value for the test of no association vs an association between position in the chain and simulated model QIC is 0.181.

Other diagnostic checks, routinely performed in Bayesian applications of the Gibbs sampler, such as I-Charts and the Gelman-Rubin test of convergence were also conducted for this chain and found to give satisfactory results. We therefore have no reason to believe that our Markov Chain of simulated

TABLE 2

CHI-SQUARED TEST FOR INDEPENDENCE BETWEEN POSITION IN CHAIN AND QIC VALUE

	4055.87-4059.33	4059.33-4063.41	4063.41-4068.79	Greater than 4068.79
Group A	115 (125)	126 (125)	135 (125)	124 (125)
Group B	101 (125)	136 (125)	133 (125)	130 (125)
Group C	121 (125)	124 (125)	131 (125)	124 (125)
Group D	139 (125)	110 (125)	178 (125)	118 (125)

values from $P_{\mathrm{QIC},\lambda}(\gamma)$ do not correspond to simulations from this distribution. We therefore proceed to the task of finding the optimal model.

The optimal model is identified as that model with the lowest QIC value in the post burn-in Gibbs sampler model simulations. This model also appeared the most frequently in the post burn-in QIC simulations. The optimal estimated model is of the form (16) with parameter estimates shown in Table 3 below. The model intercept is −3.53. The regression coefficients presented in the table below are all statistically signficantly different from zero. We note that there is a strong effect due to state, with quite wide variation in the coefficient estimates. There is some grouping in the values of the occupation indicator variables, although again we see marked differences in the predicted number of claims by occupation class. Given the relatively low level of estimated correlation between observations within the panels, it is of interest to compare the coefficients obtained from our selected model with those obtained by estimating a simple GLM where independence between observations within panels and between panels is assumed. The coefficient estimates from a GLM from the Poisson family with log link function and log exposure as an offset variable are included in the table below in parentheses in each cell. We note that the simple GLM coefficient estimates are broadly similar to those obtained from the GEE based model but do differ in some instances by a material amount. We have also provided, in Figure 3, a histogram of the differences in fitted values for the independence GLM and the GEE based model that includes AR1 autocorrelation within the panels. The benefit of including the correlation within the panels is evident here with some marked differences evident between the predicted number of losses under the two modelling approaches.

TABLE 3

FINAL MODEL COEFFICIENT ESTIMATES AND COEFFICIENTS FOR ASSOCIATED GLM

Predictor	Coefficient	Predictor	Coefficient	Predictor	Coefficient
		Occupation 2	–	Occupation 14	–
Year	–	Occupation 3	0.077 (0.05)	Occupation 15	0.56 (0.58)
State 2	0.82 (0.85)	Occupation 4	−0.31 (−0.43)	Occupation 16	–
State 3	1.32 (1.44)	Occupation 5	0.44 (0.48)	Occupation 17	0.16 (0.23)
State 4	0.43 (0.50)	Occupation 6	1.09 (1.14)	Occupation 18	−0.093 (−0.08)
State 5	−0.03 (0.03)	Occupation 8	–	Occupation 19	−0.35 (−0.15)
State 6	0.82 (0.73)	Occupation 9	–	Occupation 20	0.12 (0.13)
State 7	–	Occupation 10	−0.56 (−0.32)	Occupation 21	−0.12 (−0.12)
State 8	1.32 (1.30)	Occupation 11	0.15 (0.27)	Occupation 22	0.32 (0.19)
State 9	0.02 (0.04)	Occupation 12	0.058 (0.02)	Occupation 23	−0.66 (−0.63)
State 10	1.07 (1.08)	Occupation 13	0.21 (0.23)	Occupation 24	0.32 (0.27)

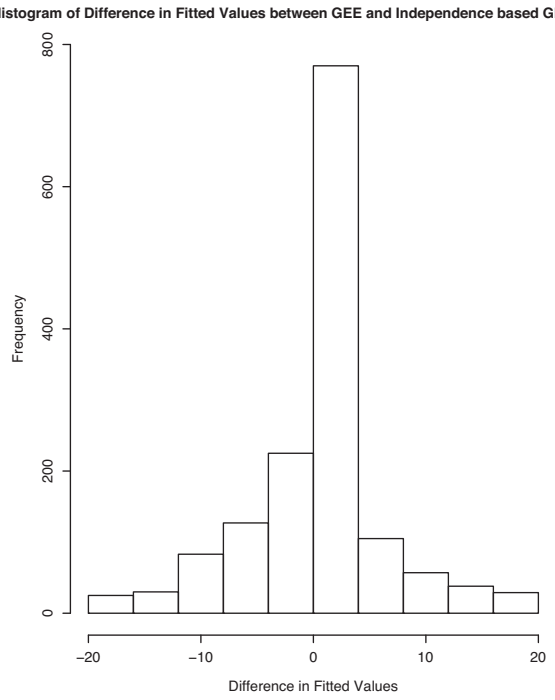

FIGURE 3: Histogram of Difference in Fitted Values between GEE and Independence Based GLM.

6. CONCLUSION

This paper has given a new algorithm for determining optimal statistical models using the QIC model selection criterion. We used both simulated data and a workers' compensation claim dataset to illustrate this algorithm. This algorithm works well with different assumptions about the correlation structure of the observations with the panels. The methodology described in this paper has the advantage of quickly determining an optimal model when the space of potential models is extremely large. The algorithm uses the Gibbs sampler and has, for the purposes of illustrations in this paper, been programmed using the statistical language R. The algorithm could equally be written in most other statistical software packages. The methodology from this paper can be adapted to other model selection problems of interest to actuaries including GLM selection, graduation model selection and state space models which are beginning to see application in actuarial practice.

ACKNOWLEDGEMENT

The authors would like to thank the two anonymous referees for their helpful comments and suggestions which certainly improved the quality of the paper.

References

Akaike, H. (1974) A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716-723.

Antonio, K. and Beirlant, J. (2007) Actuarial Statistics with generalized linear mixed models, *Insurance: Mathematics and Economics*, **40**, 58-76.

Brockman, M.J. and Wright, T.S. (1992) Statistical Motor Rating: Making Effective Use of your Data, *Journal of the Institute of Actuaries*, **119**, 457-543.

Cui, J. & Qian, G. (2007) Selection of Working Correlation Structure and Best Model in GEE Analyses of Longitudinal Data, *Communications in Statistics – Simulation and Computation*, **36(5)**, 987-996.

Frees, E.W., Young, V.R. and Luo, Y. (1999) Case studies using panel data models, *North American Actuarial Journal*, **5(4)**, 24-42.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. Pattn. Anal. Mach. Intel.*, **6**, 721-741.

Haberman, S. and Renshaw, A.E. (1996) Generalized linear models and actuarial science, *The Statistician*, **45**, 407-436.

Hardin, J.W. and Hilbe, J.M. (2003) *Generalized estimating equations.* Chapman and Hall.

Heller, G.Z., Stasinopoulos D.M., Rigby, R.A. and De Jong, P. (2007) Mean and dispersion modelling for policy claims costs, *Scandinavian Actuarial Journal*, **4**, 281-292.

Klugman, S.A. (1992) *Bayesian Statistics in Actuarial Science.* Kluwer.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005) *Applied Linear Statistical Models.* 5th ed. McGraw-Hill.

Liang, K. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models, *Biometrika*, **73(1)**, 13-22.

Makov, U., Smith, A.F.M. and Liu, Y.H. (1996) Bayesian methods in actuarial science, *The Statistician*, **45(4)**, 503-515.

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

Pan, W. (2001) Akaike's information criterion in generalized estimating equations, *Biometrics*, **57**, 120-125.

Qian, G. (1999) Computations and analysis in robust regression model selection using stochastic complexity, *Computational Statistics*, **14**, 293-314.

Qian, G. and Field, C. (2000) Using MCMC for Logistic Regression Model Selection Involving Large Number of Candidate Models, *Monte Carlo and Quasi-Monte Carlo Methods*, K.T. Fang, F.J. Hickernell & H. Niederrerter (eds), Springer, 460-474.

Qian, G. and Zhao, X. (2007) On time series model selection involving many candidate ARMA models, *Computational Statistics & Data Analysis*, **51(1)**, 6180-6196.

R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Scollnik, D.P.M. (1996) An Introduction to Markov Chain Monte Carlo methods and their actuarial applications, *Proceedings of the Casualty Actuarial Society*, **LXXXIII**, 114-165.

David Pitt
Department of Actuarial Studies
Macquarie University
NSW 2109
Australia
E-Mail: david.pitt@mq.edu.au