The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2632-279X.htm

Received 21 August 2019
Accepted 29 August 2019

MCMC and GLMs for estimating regression parameters

Evidence from non-life Egyptian insurance sector

Mahmoud ELsayed and Amr Soliman
Department of Mathematics and Insurance, Faculty of Commerce, Cairo University, Giza, Egypt

Abstract

Purpose – The purpose of this study is to estimate the linear regression parameters using two alternative techniques. First technique is to apply the generalized linear model (GLM) and the second technique is the Markov Chain Monte Carlo (MCMC) method.

Design/methodology/approach – In this paper, the authors adopted the incurred claims of Egyptian non-life insurance market as a dependent variable during a 10-year period. MCMC uses Gibbs sampling to generate a sample from a posterior distribution of a linear regression to estimate the parameters of interest. However, the authors used the R package to estimate the parameters of the linear regression using the above techniques.

Findings – These procedures will guide the decision-maker for estimating the reserve and set proper investment strategy.

Originality/value – In this paper, the authors will estimate the parameters of a linear regression model using MCMC method via R package. Furthermore, MCMC uses Gibbs sampling to generate a sample from a posterior distribution of a linear regression to estimate parameters to predict future claims. In the same line, these procedures will guide the decision-maker for estimating the reserve and set proper investment strategy.

Keywords Insurance, Regression, Sampling, MCMC, GLM

Paper type Research paper

1. Introduction

Modeling of random events is one of the most vital research aspects in insurance and actuarial sciences. In insurance particularly, modeling and predicting the amount of claims has an extremely importance to both insurers and academics. In addition, Bayesian approach is one of the best statistical methods that estimate outstanding claims. In Bayesian modeling we should distinguish between the observable quantities and the unknown parameters that can be treated as random variables. Moreover, Bayesian approach provides a technique that combine prior information from the given data to estimate posterior distribution. In the same vein, posterior distribution can be used to describe the model

© Mahmoud ELsayed and Amr Soliman. Published in *Journal of Humanities and Applied Social Sciences*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

JEL classification – C22, C53, G22

Journal of Humanities and Applied Social Sciences
Vol. 2 No. 1, 2020
pp. 46-55
Emerald Publishing Limited
2632-279X
DOI 10.1108/JHASS-08-2019-0028

parameters via mean, median, percentiles, point estimate and credible intervals. Markov Chain Monte Carlo (MCMC) simulation may follow Bayesian statistics to estimate parameters that is impossible to be estimated by maximum likelihood estimate (MLE) or other statistical methods. MCMC is a technique that is used for sampling probability mass functions or density functions. Furthermore, MCMC does not require optimization algorithm such as MLE and generalized methods of moments, but it provides small sample inference of parameters. MCMC has been improved to fit nonlinear regression models; this approach fills the gap in literature of non-life insurance market. In addition, MCMC uses Gibbs sampling to generate a sample from a posterior distribution of a linear regression to estimate the linear regression parameters. In the same line, generalized linear models (GLMs) can be used for non-identical residuals and nonlinear functions and it uses a transformation to increase straighten of the regression, GLMs is considers as an extension to ordinary least square method when the variances are not equal (i.e. heteroscedastic models). The aim of this paper is to estimate the linear regression parameters using MCMC and GLM methods for incurred claims of the non-life Egyptian insurance market. Data adopted in this research consist of ten year incurred claims from 2007/2008 to 2016/2017 of 22 non-life Egyptian insurance companies.

Table I and Figure 1 describe the amount of incurred claims during the period from 2007/2008 to 2016/2017 for non-life Egyptian insurance market, we can notice that there is an increase in claims but there is a drop in the period between 2010/2011 to 2013/2014 The due to the Egyptian revolution.

The remainder of this paper will be as follows: Section 2 presents the literature review; Section 3 gives the methodology; Section 4 presents the models; Section 5 gives the empirical study; Section 6 concludes the paper.

Year	Incurred claims
2007/2008	2,40,9495
2008/2009	2,08,0829
2009/2010	2,72,8816
2010/2011	2,65,2673
2011/2012	2,21,7705
2012/2013	2,47,0999
2013/2014	2,21,5532
2014/2015	2,74,7947
2015/2016	2,56,1639
2016/2017	3,31,2489

Source: Financial Regulatory Authority (FRA)

Table I. Amount of incurred claims for non-life Egyptian insurance market in thousands EGP

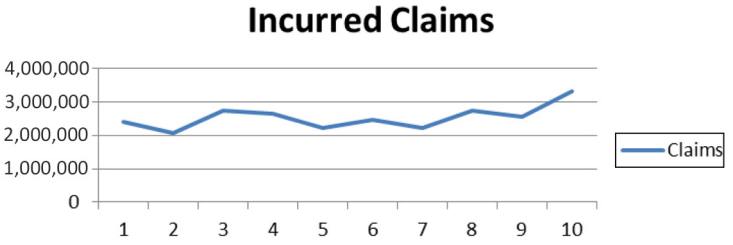

Figure 1. Amount of Incurred claims for non-life Egyptian insurance market

2. Literature review

Alba (2008) used Munich Chain Ladder (MCL) to optimize paid and incurred claims via MCMC method using WinBUGS software. This paper suggested many modifications to the MCL method. Moreover, he presented a Bayesian approach to the MCL.

Jackie (2007) compared many stochastic reserving methods such as MCMC and GLMs by considering the structure of the model, the assumption and estimation. This paper applied these methods on claims to estimate the outstanding claims and risk margin for each individual accident and aggregate risk margin.

Pang *et al.* (2007) emphasized on modeling loss distributions for insurance claims, by considering Pareto distribution to calculate the probability of extreme claims. They used Bayesian and MCMC techniques to estimate Pareto parameters.

Scollnik (2001, 2004) reviewed several actuarial models that consider Bayesian method. Afterwards, he implemented the MCMC simulations for Bayesian estimation BUGS (Bayesian inference Using Gibbs Sampling) of reserves via several programming languages (e.g. WinBUGS).

Peremans *et al.* (2017) focused on claim reserving using GLMs on chain ladder based on past claims, also used an alternative technique to obtain inference by using bootstrapping. In addition, he estimated a distribution of risk measures using several bootstrap procedures.

Boj and Costa (2017) estimated the parameters of loss distribution and predicted the error using GLMs to the claim amounts of a chain ladder method. Furthermore, they used a parametric family to estimate error distribution. In addition, they assumed a Poisson distribution with logarithmic link function as a deterministic chain ladder method.

Verdonck *et al.* (2009) illustrated how to forecast claim reserves using two methods. Firstly, robust chain ladder method that observes outliers. Secondly, robust GLMs that estimate the claim reserve as if the data has no outliers. They concluded that the robust chain ladder method is showing a better performance than robust GLMs.

Carrato and Visintin (2019) introduced a new approach which is machine learning techniques in actuarial sciences that has more accuracy in prediction than traditional techniques. They focused on the elements of machine learning rather than traditional forecasting techniques to predict property and casualty loss reserving.

Ravenzwaaij *et al.* (2018) introduced the MCMC methods as a technique that estimate the posterior distributions and provides the benefits and limitation of sampling using MCMC.

Luoma *et al.* (2008) applied the Bayesian approach and regression method to evaluate the American-Style option, also used MCMC method to estimate the model and parameter errors. Moreover, they concluded that the proper choice of the model is a vital issue in risk management.

Hogg and Foreman (2018) used MCMC to estimate the density function of the posterior distribution, fitting models to data and probabilistic inferences. In this paper, they illustrated the MCMC method and parameter estimation, they concluded that this method provides the best estimate.

Zhang (2017) used Apache Spark across a cluster of computers to estimate distribution using Bayesian approach. In addition, he used Bayesian hierarchical Tweedie model to big data of insurance claims as a predictive model.

Yu (2015) adopted a statistical model for health insurance claims, to predict future claims. In this paper, he used generalized exponential growth model (GEGM) and estimated the parameters of the model based on MCMC.

Lim (2011) applied the MCMC method to solve Bayesian method, estimate parameters and prediction of reserves. He also concluded that MCMC method is much better than classical methods (e.g. chain ladder and Bayesian over-dispersed Poisson model).

3. Data and methodology

Data adopted in this research consists of a 10-year time series of incurred claims for 22 non-life Egyptian insurance market, these data reported in FRA since 2007/2008 to 2016/2017. In this paper, we will estimate the parameters of a linear regression model using MCMC method via R package. Furthermore, MCMC uses Gibbs sampling to generate a sample from a posterior distribution of a linear regression to estimate parameters to predict future claims. In the same line, these procedures will guide the decision-maker for estimating the reserve and set proper investment strategy.

4. Models

4.1 Linear regression

Consider the linear regression model:

$$y = b_0 + b_1 x + \varepsilon$$

where y is the dependent variable, x the independent variable, b_0 and b_1 are the parameters of the model and ε is the white noise $\varepsilon \sim N(0, \sigma^2)$.

4.2 Generalized linear model

The GLM is formed with two ingredients: link function and variance function. The link function relates the means of the observations to the predictors (linearization), while the variance function relates the means to variances Lindsey (1997).

The link function can be expressed by:

$$g(\mu_i) = \zeta_i$$

and the variance function is defined by:

$$var(Y_i) = \varnothing \; V(\mu)$$

where the dispersion parameter $\varnothing$ is a constant.

4.3 Bayesian statistics

Bayesian analysis emphasis in estimation of posterior distribution depending on prior distribution and the likelihood function of the parameters. In addition, normalize the final posterior distribution:

$$P(\theta|x) = \frac{P(\theta) \times P(x|\theta)}{P(x)}$$

$$P(\theta|x) \propto P(\theta) \times P(x|\theta)$$

$$Posterior \propto Prior \times Likelihood$$

4.4 Markov chain

According to Andrieu *et al.* (2003) let X_t be the value of a certain random variable at time t and possible values of X represents the state space. A stochastic process is considered as a Markov stochastic process if the state space depends only on the current state:

$$P\left[X_{t+1} = s_j | X_0 \;=\; s_k, \ldots, X_t = s_i\right] = P\left[X_{t+1} = s_j |\; X_t = s_i\right]$$

That means to predict future value of the process we need only the current state, the probability of such event is called transition probability $P(i,j) = P(i \rightarrow j)$:

$$P(i,j) = P(i \rightarrow j) = \;\; P\left[X_{t+1} = s_j |\; X_t = s_i\right]$$

4.5 Monet Carlo

According to Andrieu *et al.* (2003), assume that $h(x)$ is a complex function and we need to find the integration of $h(x)$:

$$\int_a^b h(x)dx$$

and $h(x)$ can be expressed as a function $f(x)$ multiplied by a probability $p(x)$ then:

$$\int_a^b h(x)dx = \int_a^b f(x) \cdot p(x)dx = \;\; E_{p(x)}\left[f(x)\right]$$

this function is the expected value of $f(x)$ over a density function $p(x)$. If we draw a large number of observations x_i, $i = 1, 2, ., n$ of a random variable with density function $p(x)$ then:

$$\int_a^b h(x)dx = \;\; E_{p(x)}\left[f(x)\right] \cong \;\; \frac{1}{n}\sum_{i=1}^{n} f(x)$$

4.6 Markov chain Monte Carlo

According to Bayesian statistics, MCMC method is an iterative sampling technique that allows sampling through $P(\theta \,|\, x)$. MCMC is an effective approach that generates samples from posterior distributions $P(\theta \,|\, x)$. Moreover, the target density is the posterior density $\pi(\theta) = P(\theta \,|\, x)$ and MCMC can be implemented when posterior cannot be formed properly:

Suppose we seek an expectation $\mu = E_\pi\left[g(\theta)\right] = \int g(\theta) \times P(\theta|x)d\theta$.

As illustrated by Monte Carlo method above.

5. Empirical results

5.1 Descriptive statistics

In this section, we will illustrate the statistical characteristics of amount of incurred claims as shown in Table II.

From Table II and Figure 2 we can notice that the data is positively skewed, also the data is Platykurtic under the normal curve.

5.2 Generalized linear model

Table III presents the results of GLM applied for incurred claims to estimate the linear regression parameters b_0 and b_1. Moreover, Akaike Information Criterion (AIC) found to be 285.36 that represents the model selection that estimates the quality of the model. In addition, Figure 3 visualizes the residuals of the model and shows the Skewness of the data.

5.3 Markov chain Monte Carlo

In this section, we implement the MCMC method in R package to estimate the parameters of the linear regression; we performed 1000 iterations on MCMC that means we applied the

Measure	Incurred claims
Mean	2,53,9812
Standard Deviation	3,54,429.1
Minimum	2,08,0829
First Quartile	2,26,5652
Median	2,51,6319
Third Quartile	2,70,9780
Maximum	3,31,2489
Range	1,23,1660
Skewness	0.7
Kurtosis	−0.3
Standard Error	112080.3

Source: Authors' calculations

Table II.
Descriptive analysis of incurred claims

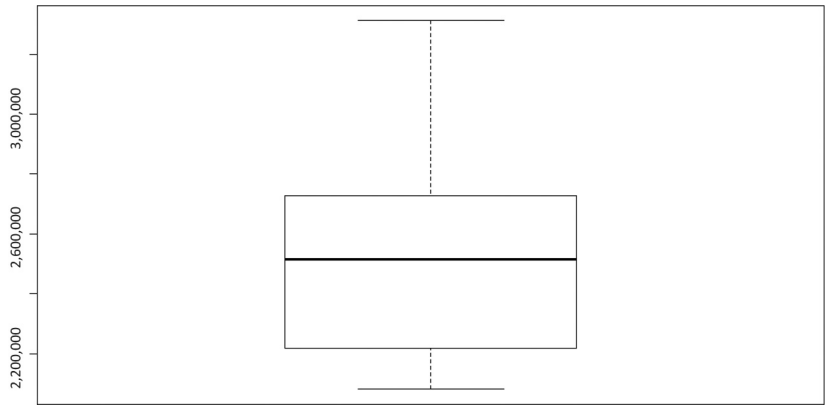

Figure 2.
Box-plot of incurred claims

Markov chain 1000 times. Table IV summarizes the results from the R package of the mean, standard deviation and the standard error of the parameters b_0, b_1 and σ^2. Moreover, Figures 4, 6 and 8 present the trace (time series plot) of the parameters, while Figures 5, 7 and 9 presents the density of the parameters.

Parameter	Value	Standard error	*t*-value	*p*-value
b_0	−1,25,83,2191	6,97,93,362	−1.803	0.109
b_1	6,3819	3,4697	1.839	0.103
Degrees of freedom	9			
AIC	285.36			

Table III.
GLM Results

Source: Authors' Calculations

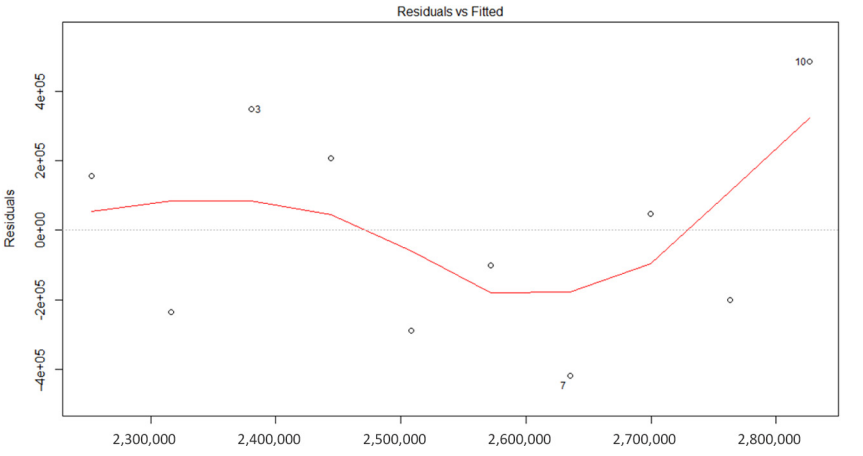

Figure 3.
GLM residuals

Parameter	Mean	SD	Standard error
b_0	−1.240e + 08	7.511e + 07	2.096e + 06
b_1	6.292e + 04	3.734e + 04	1.042e + 03
σ^2	1.269e + 11	8.748e + 10	3.622e + 09

Table IV.
MCMC Results

Source: Authors' Calculations

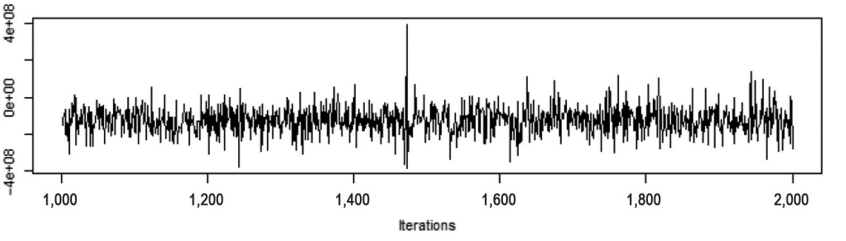

Figure 4.
Trace of b_0

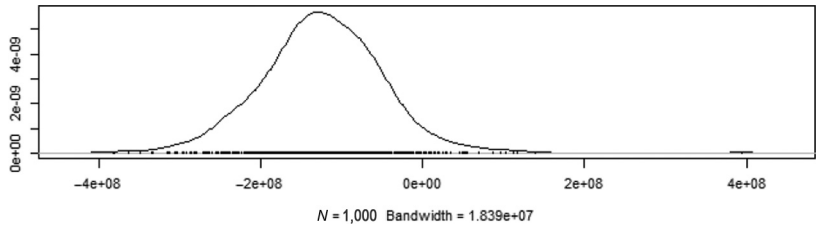

Figure 5.
Density of b_0

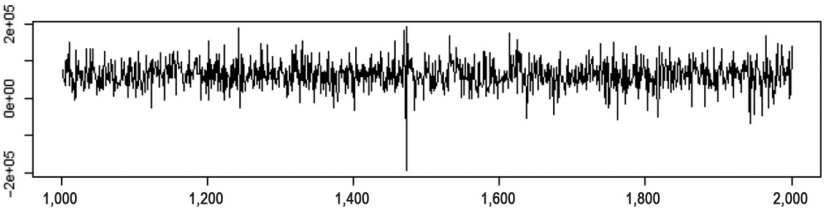

Figure 6.
Trace of b_1

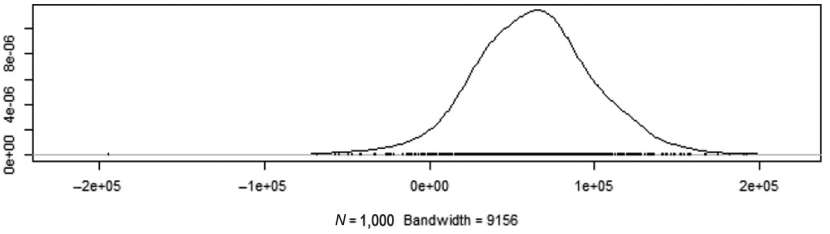

Figure 7.
Density of b_1

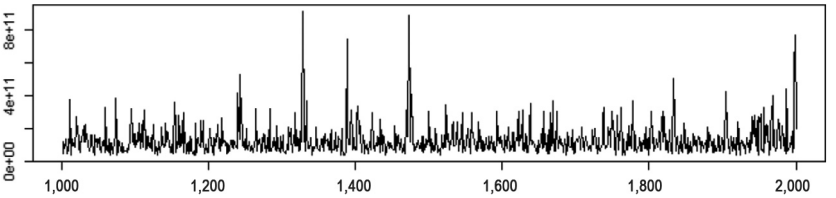

Figure 8.
Trace of σ^2

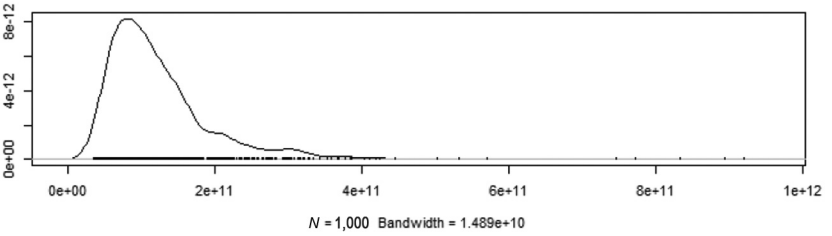

Figure 9.
Density of σ^2

6. Conclusion

MCMC simulation may follow Bayesian statistics to estimate parameters. In addition, MCMC uses Gibbs sampling to generate a sample from a posterior distribution of a linear regression. MCMC is a technique implemented to estimate the linear regression parameters b_0, b_1 and σ^2. In this paper, we adopted the incurred claims of non-life Egyptian insurance industry reported in Financial Regulatory Authority (FRA) during 10-year period, from 2007/2008 to 2016/2017 as an explanatory variable. We applied GLM to estimate the regression parameters and we performed 1000 iterations (i.e. 1000 Markov Chains) on MCMC to estimate the linear regression parameters on R package, MCMC performs many iterations of chains for sampling to estimate regression parameters that yield more information to reach the true values of the parameters. Moreover, these procedures will guide the decision maker for estimating the reserve and set proper investment strategy.

References

Alba, E. (2008), "The Munich Chain-Ladder method: a Bayesian approach", *Institute of Insurance and Pension Research*, University of Waterloo, Ontario, pp. 1-25.

Andrieu, C., Freitas, N., Doucet, A. and Jordan, M. (2003), "An introduction to mcmc for machine learning", *Machine Learning*, Vol. 50 Nos 1/2.

Boj, E. and Costa, T. (2017), "Provisions for claims outstanding, incurred but not reported, with generalized linear models: prediction error formulated according to calendar year", *Cuadernos de Gestión*, Vol. 17 No. 2, pp. 157-174.

Carrato, A. and Visintin, M. (2019), "From the chain ladder to individual claims reserving using machine learning techniques", *ASTIN Colloquium*, Vol. 1, pp. 1-19.

Hogg, D. and Foreman, D. (2018), "Data analysis recipes: using Markov Chain Monte Carlo", *The Astrophysical Journal Supplement Series*, Vol. 236 No. 1, pp. 1-18.

Jackie, L. (2007), "Comparison of stochastic reserving methods", *Australian Actuarial Journal*, Vol. 12 No. 4, pp. 489-569.

Lim, K. (2011), "Bayesian analysis of claim run-off triangles", The Australian National University, Thesis for Bachelor of Actuarial Studies.

Lindsey, J. (1997), *Applying Generalized Linear Models*, Springer, New York, NY.

Luoma, A., Puustelli, A. and Koskinen, L. (2008), *Bayesian Analysis of Participating Life Insurance Contracts with American-Style Options*, Insurance Supervisory Authority, pp. 1-16.

Pang, W., Hou, S., Troutt, D., Yu, W. and Li, K. (2007), "A Markov Chain Monte Carlo approach to estimate the risks of extremely large insurance claims", *International Journal of Business and Economics*, Vol. 6 No. 3, pp. 225-236.

Peremans, K., Segaert, P., Van Aelst, S. and Verdonck, T. (2017), "Robust bootstrap procedures for the Chain-Ladder method", *Scandinavian Actuarial Journal*, Vol. 10, pp. 870-897.

Ravenzwaaij, D., Pete, C. and Scott, B. (2018), "A simple introduction to Markov chain Monte-Carlo sampling", *Psychonomic Bulletin and Review*, Vol. 25, pp. 143-154.

Scollnik, D. (2001), "Actuarial modeling with MCMC and BUGS", *North American Actuarial Journal*, Vol. 5 No. 2, pp. 96-125.

Scollnik, D. (2004), "Bayesian reserving models inspired by Chain-Ladder methods and implemented using WinBUGS", available at: www.soa.org/library/proceedings/arch/2004/arch04v38n2_3.pdf

Verdonck, T., Wouwe, M. and Dhaene, J. (2009), "A robustification of the Chain-Ladder method", *North American Actuarial Journal*, Vol. 13 No. 2, pp. 280-298.

Yu, G. (2015), "Hierarchical Bayesian modeling of health insurance claims", Master thesis, The University of Melbourne. Faculty of Science. The Department of Statistics and Actuarial Science.

Zhang, Y. (2017), "Bayesian analysis of big data in insurance predictive modeling using distributed computing", *Astin Bulletin*, Vol. 47 No. 3, pp. 943-961.

Corresponding author
Amr Soliman can be contacted at: s.amr@foc.cu.edu.eg

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**