The hierarchical generalized linear model and the bootstrap estimator of the error of prediction of loss reserves in a non-life insurance company

Alicja Wolny-Dominiak

December 14, 2016

Abstract

This paper presents the hierarchical generalized linear model (HGLM) for loss reserving in a non-life insurance company. Because in this case the error of prediction is expressed by a complex analytical formula, the error bootstrap estimator is proposed instead. Moreover, the bootstrap procedure is used to obtain full information about the error by applying quantiles of the absolute prediction error. The full R code is available on the Github https://github.com/woali/BootErrorLossReserveHGLM.

Keywords. Loss Reserves; Hierarchical Generalized Linear Model; Error of prediction; MSE; Parametric Bootstrap

JEL: G22, C21, C53

1 Introduction

This analysis concerns a non-life insurance company. The largest item on an insurer’s balance sheet are technical provisions. Any variations in their values have a great impact on the insurer’s financial strength. A large part of the provisions is the reserve for incurred but not reported (IBNR) claims or simply – the loss reserve, which is crucial to the insurer’s solvency. The total loss reserve is generally determined by statistical methods based on both deterministic techniques and stochastic models and is a sum of outstanding claims liabilities.

A wide variety of the total loss reserve prediction tools are presented in literature in the group of stochastic methods. For years the most popular method with insurers was the chain-leader technique (e.g. (Mack [1994], (Mack and Venter [2000]), (Verrall [2000])). Parallel methods based on the regression model have been developed (e.g. (Kremer [1982], (Mack [1991])). In (Renshaw and Verrall 1998) it is shown that the chain-leader method is equivalent to the well-known generalized linear model (GLM) with fixed effects assuming over-dispersed Poisson (ODP) distribution for incremental claims. This paper puts forward a mixed
model for loss reserving as an extension of the GLM in which data for different homogenous groups are assumed to be dependent in some way. The model belongs to the class of HGLMs and contains both fixed effects as well as random effects. Model’s parameters are estimated in the frequentist approach using extended likelihood [Lee and Nelder 1996], [Lee and Nelder 2001], [Lee and Nelder 2003]. As the total loss reserve is actually a random variable, its value has to be predicted. In order to measure the prediction accuracy, the bootstrap root mean squared error of prediction (RMSEP) estimator is proposed.

2 Data – single loss triangle and total loss reserve

In loss reserving, data have a specific form referred to as the loss triangle. Consider the random variable \( Y_{ij} \) with \( y_{ij} \) realizations \( i, j = 0, \ldots, n \) being the incremental quantity for insurance claims that occurred in year \( i \) (origin year) and were reported to the insurer after \( j \) years (development year). It could be the value of claims as well as their number. The matrix \( [Y_{ij}]_{(n+1) \times (n+1)} \) represents random variables in the loss triangle for a single LOB as shown in Table 1 below.

|   | 0   | ... | n - 1 | n  |
|---|-----|-----|-------|----|
| 0 | \( Y_{00} \) | \( \cdots \) | \( Y_{0n-1} \) | \( Y_{0n} \) |
| 1 | \( Y_{10} \) | \( \cdots \) | \( Y_{1n-1} \) |
| ... | \( \cdots \) | \( \cdots \) |
| n | \( Y_{n0} \) |

The realizations \( y_{ij} \) for \( i + j \leq n \) are observed data while \( y_{ij} \) for \( i + j > n \) represent the future unobserved data. The goal in loss reserving is to predict the random variable \( R \) being the total loss reserve defined as the sum of unknown future claims:

\[
R = \sum_{i, j; i+j > n} Y_{ij}
\]

(1)

In order to assess the uncertainty of predictor \( \hat{R} \), usually

\[
RMSEP(\hat{R}) = \sqrt{E[(\hat{R} - R)^2]}
\]

(2)

is adopted.

3 The model and the prediction

Let us assume \( Y_{ij} \sim ODP(\mu_{ij}, \phi) \) as independent and identically distributed (i.i.d) with \( E[Y_{ij}] = \mu_{ij} \) and \( Var(Y_{ij}) = \phi \mu_{ij} \) and the unknown dispersion parameter \( \phi \). In the GLM the mean has a linear form \( \log(\mu_{ij}) = x_{ijk} \beta \), where
\[ \beta = (c, u_1, \ldots, u_n, \beta_1, \ldots, \beta_n)' \] is the vector of fixed effects, \( x'_{ijk} \) is the row of the design matrix \( X \) and \( u_0 = \beta_0 = 1 \). The dispersion parameter \( \phi \) is the same for all elements of the loss triangle. It is easy to extend this model to the general class of models assuming that incremental claims follow the three-parametric Tweedie \( T(\mu, \phi, p) \) (Jorgensen, 1987) with dispersion parameter \( \phi \) and power \( p \in (0, \infty) \setminus (0, 1) \). The GLM is then of the form:

\[ E[Y_{ij}] = \mu_{ij} = \exp(x'_{ij}\beta), \ Var(Y_{ij}) = \phi \mu_{ij}^p, \]  

(3)

In order to draw an inference about the total loss reserve \( R \), MLE estimators of parameters \( (\hat{c}, \hat{u}_1, \ldots, \hat{u}_n, \hat{\beta}_1, \ldots, \hat{\beta}_n, \hat{\phi})' \) can be plugged into Formula (1) giving the predictor:

\[ \hat{R}_{GLM} = \sum_{i,j;i+j>n} \hat{Y}_{ij} = \sum_{i,j;i+j>n} \exp(\hat{c} + \hat{u}_i + \hat{\beta}_j). \]  

(4)

The total loss reserve prediction accuracy is typically measured using the \( RMSEP \). According to (England and Verrall, 1999), the predictor defined in Formula (4) can be expressed as:

\[ RMSEP(\hat{R}_{GLM}) = \sum_{i,j;i+j>n} \phi_{ij} \mu_{ij} + \sum_{i,j;i+j>n} \mu_{ij}^2 Var(x'_{ij}\beta) + \sum_{i,j;i+j>n} \mu_{ij} Var(z'_{ij}\beta). \]  

(5)

The fundamental assumption in the GLM is the independence between random incremental claims \( Y_{ij}, i,j = 0, \ldots, n \) taken from a single loss triangle. However, this assumption may be inadequate, as indicated by the fact that origin years reflect the same process of loss development. This means that there is a dependence between claims from the same development year \( j \) but from different origin years. The dependence may be taken into account by applying a mixed model with fixed and random effects in place of the GLM with fixed effects only.

Let us assume that effects \( u_0, \ldots, u_n \) occurring in model (3) are independent realizations of random variable \( U \) following Tweedie distributions \( u_i \sim T(\psi_{iu}, \phi_u) \) with \( E[u_i] = \psi_{iu} \) and \( Var(u_i) = \phi_u \psi_{iu}^p \). This means that incremental claims from the loss triangle are now conditionally independent, which can be described in the following way: \( Y_{ij} \) of different origin years are independent, but due to the random effects, for any origin year \( i \), the random variables \( Y_{ij} \) for the different development years \( j \) are dependent (cf. (Gigante et al., 2013b), pp. 383, (a1)–(4)). This approach is examined in (Gigante et al., 2013a, 2013b), where the hierarchical GLM (HGLM) is described. The authors assume a conditional ODP distribution for incremental claims \( Y_{ij} | u_i \). Like for the GLM, this assumption is extended to any \( p \) such that \( Y_{ij} | u_i \sim T(\mu_{ij}, \phi, p) \), which gives the following general form of the HGLM:

\[ E[Y_{ij} | u_i] = \mu_{ij}(u_i) = \exp(x'_{ij}\beta + z'_{ij}v) \]  

\[ Var(Y_{ij} | u_i) = \phi \mu_{ij}^p(u_i), \]  

(6)
where $\beta = (c, \beta_1, \ldots, \beta_n)'$, $v = (\log(u_1), \ldots, \log(u_n))'$ and $z_{ijk}$ is the row of the design matrix $Z$. Under the model defined in (6) and using the conditional independence assumption, the total loss reserve is defined as the conditional expected value:

$$R_{HGLM} = \sum_{i,j:i+j>n} E[Y_{ij}|u_i] = \sum_{i,j:i+j>n} \exp(x_{ij}'\beta + z_{ij}'v) \quad (7)$$

The estimation of the HGLM parameters is more complex than in the case of the GLM due to the presence of random effects. Following (Gigante et al., 2013a), the likelihood-based approach is applied using the h-likelihood function. The idea is to treat the vector of random effects $v$ as the vector of fixed effects and transform the HGLM into an augmented GLM. Details concerning the h-likelihood estimation are presented in two fundamental papers (Lee and Nelder, 1996), (Lee and Nelder, 2001). This numerical analysis uses the R implementation of the algorithm taken from the hglm package (Ronnegard et al., 2010).

After estimators $\hat{\beta}$ and predictors $\hat{v}$ are found, the total loss reserve predictor is of the form:

$$\hat{R}_{HGLM} = \sum_{i,j:i+j>n} \exp(\hat{c} + \hat{\beta}_j)\hat{u}_i \quad (8)$$

A similar mixed model-based approach is described in (Antonio et al., 2006), but only for a lognormal distribution. The authors adopted the generalized linear mixed model (GLMM) and the REML as the method of estimation of the model parameters.

### 4 Bootstrap MSEP

In (Gigante et al., 2013b) the estimator of $RMSEP(\hat{R})$ is derived analytically in the case of a single loss triangle. Unfortunately, the formula describing the estimator has a rather complex structure. An alternative may be to use the parametric bootstrap technique, like for the $RMSEP(\hat{R}_{GLM})$ proposed in (England and Verrall, 1999). This paper proposes another procedure.

Let us denote the realization of $Y_{ij}$ as $y_{ij}$. As the HGLM is commonly an extension of the GLM, Pearson residuals $r_{ij} = \frac{\hat{y}_{ij} - y_{ij}}{\sqrt{\hat{\mu}_{ij}}}$ are bootstrapped instead of the data from the loss triangles. A single $b$th simulation, $b = 1, \ldots, B$ in the full bootstrap procedure has the following steps:

- take a sample $r_{ij}^{*b}$ with replacement from Pearson residuals $r_p$ based on data from the loss triangle $i + j \leq n$
- calculate new data $y_{ij}^{*b} = r_{ij}^{*b} \sqrt{\hat{\mu}_{ij}} + \hat{\mu}_{ij}, i + j \leq n$
- estimate the vector of parameters $(\beta^{*b}, v^{*b}, \phi^{*b}, \phi^{*b}_u)'$ based on new data $y_{ijk}^{*b}$
- calculate predicted values $\hat{y}_{ij}^{*b} = \exp(\hat{c} + \hat{\beta}_j)\hat{u}_i$ for $i + j > n$
• generate values $y_{ij}^{*b}$ for $i + j > n$ from Tweedie distribution with the dispersion parameter $\phi^{*b}$

After $B$ iterations of the bootstrap procedure the estimator of $RMSEP(\hat{R}_{HGLM})$ is of the following form:

$$\text{RMSE}(\hat{R}) = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left( \sum_{i,j;i+j>n} \hat{y}_{ij}^{*b} - \sum_{i,j;i+j>n} y_{ij}^{*b} \right)^2}$$  \hspace{1cm} (9)

Let us notice that the commonly used $RMSEP$ gives information only about the average error of prediction. If full distribution of the prediction error is of interest, quantiles of the absolute prediction error can be used (see (Zadro, 2013), where the $p$th quantile is defined as $Q_p(\hat{R}) = Q_p(|\hat{R} - R|)$. Quantiles reflect the relation between the magnitude of the error and the probability of its realization. In order to estimate $Q_p(\hat{R})$, the bootstrap procedure described above can be applied. In the last step, Formula (9) is replaced by the formula presented below:

$$\hat{Q}_p(\hat{R}) = Q_p(\sum_{i,j;i+j>n} \hat{y}_{ijk}^{*b} - \sum_{i,j;i+j>n} y_{ijk}^{*b})$$  \hspace{1cm} (10)

As a result, the bootstrap estimator of the $Q_p(\hat{R})$ error is obtained.

5 Numerical example

The subject of this analysis is the single loss triangle taken from (Wüthrich and Merz, 2008), pp. 33, like in (Gigante et al., 2013a). In order to demonstrate the HGLM for loss reserving, it is assumed that incremental claims conditionally follow Tweedie distribution $Y_{ij|u_i} \sim T(\mu_{ij}, \phi_u)$ with power $p = 1$ and random effects follow Tweedie distribution $u_i \sim T(\psi_u, \phi_u)$ with power $p = 2$. This model is equivalent to the ODP-Gamma HGLM. Moreover, dispersion parameters $\phi$, $\phi_u$ are assumed as constant for both incremental claims and random effects, respectively. The vector of parameters $$(c, \beta_1, \ldots, \beta_9, u_0, \ldots, u_9, \phi, \phi_u)^\prime$$ is estimated using the h-likelihood function. Plugging the obtained values into Formula (9), the total loss reserve and the loss reserves for origin years are obtained. In order to find the estimator of $RMSEP$, the bootstrap procedure was implemented. The number of simulations was $n = 1000$. The results are presented in Table 2.

It can be seen that the HGLM gives higher loss reserve values compared to the GLM but the errors are generally lower. Naturally, errors rise for subsequent origin years in both cases, which results directly from the fact that there are fewer and fewer observations in the loss triangle. The two models are not compared directly because in the case of the HGLM it is the conditional value of reserves that is determined.
Table 2: Total loss reserve and bootstrap estimator $\hat{RMSE}$

| Origin year $i$ | $\hat{R}_{HGLM}$ | $\hat{RMSEP}_{HGLM}$ | $\hat{R}_{GLM}$ | $\hat{RMSEP}_{GLM}$ |
|-----------------|------------------|----------------------|-----------------|---------------------|
| 1               | 15 239           | 19 763               | 15 125          | 19 620              |
| 2               | 26 415           | 24 220               | 26 257          | 24 508              |
| 3               | 35 264           | 27 182               | 34 538          | 27 613              |
| 4               | 86 572           | 38 331               | 85 302          | 40 690              |
| 5               | 157 840          | 49 339               | 156 494         | 54 876              |
| 6               | 290 991          | 66 545               | 286 121         | 70 027              |
| 7               | 465 024          | 81 939               | 449 161         | 86 568              |
| 8               | 1 078 769        | 130 374              | 1 043 237       | 132 874             |
| 9               | 3 988 971        | 281 813              | 3 950 809       | 315 437             |
| Total           | 6 145 085        | 373 914              | 6 047 044       | 403 506             |

$\phi = 14 739$  \quad $\phi = 14 714$

$\phi_u = 0.0054$

Figure 1: Measures of accuracy for origin years and the total loss reserve
The bootstrap procedure makes it possible to obtain not only the RMSEP but also the quantiles of the absolute error of prediction $Q_p$. Figure 1 presents quantiles $Q_{0.5}$, $Q_{0.75}$, $Q_{0.9}$ and $Q_{0.95}$ for origin years and for the total loss reserve. It can be seen that in earlier origin years the RMSEP value is more or less equal to quantiles of the order of 0.75, whereas for the total loss reserve the value is close to quantiles of the order of 0.9.

6 Conclusions

GLMs are popular statistical techniques in actuarial practice, especially in ratemaking but also in loss reserving. However, the independence assumption needed in GLMs is generally violated in many cases. There are three basic advantages of the mixed HGLM application. Firstly, by introducing random effects into the model, the dependencies between development years are taken into account. Secondly, by imposing certain constraints on random effects, it is possible to take account of external information which does not come from the sample directly, like in (Gigante et al., 2013a), and which has an impact on the total loss reserve value. Thirdly, distributions can be set flexibly within the Tweedie family, e.g. the Gamma, the inverse-Gaussian or the compound Poisson distribution. The downside, however, is the complex form of the error prediction. Therefore, it is proposed herein that the error should be determined by means of the bootstrap technique. Although this solution is not perfect, its important advantage is that full information on the absolute error distribution can be obtained easily using quantiles.

References

K. Antonio, J. Beirlant, T. Hoedemakers, and R. Verlaak. Lognormal mixed models for reported claims reserves. *North American Actuarial Journal*, 10 (1):30–48, 2006.

P. England and R. Verrall. Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: mathematics and economics*, 25(3):281–293, 1999.

P. Gigante, L. Picech, and L. Sigalotti. Claims reserving in the hierarchical generalized linear model framework. *Insurance: Mathematics and Economics*, 52(2):381–390, 2013a.

P. Gigante, L. Picech, and L. Sigalotti. Prediction error for credible claims reserves: an h-likelihood approach. *European Actuarial Journal*, 3(2):453–470, 2013b.

B. Jorgensen. Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 127–162, 1987.
E. Kremer. Ibnr-claims and the two-way model of anova. *Scandinavian Actuarial Journal*, 1982(1):47–55, 1982.

Y. Lee and J. Nelder. Extended-reml estimators. *Journal of Applied Statistics*, 30(8):845–856, 2003.

Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 619–678, 1996.

Y. Lee and J. A. Nelder. Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4):987–1006, 2001.

T. Mack. A simple parametric model for rating automobile insurance or estimating ibnr claims reserves. *Astin Bulletin*, 21(1):93–109, 1991.

T. Mack. Which stochastic model is underlying the chain ladder method? *Insurance: mathematics and economics*, 15(2):133–138, 1994.

T. Mack and G. Venter. A comparison of stochastic models that reproduce chain ladder reserve estimates. *Insurance: Mathematics and Economics*, 26(1):101–107, 2000.

A. E. Renshaw and R. J. Verrall. A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, 4(04):903–923, 1998.

L. Ronnegard, X. Shen, and M. Alam. hglm: A package for fitting hierarchical generalized linear models. *The R Journal*, 2(2):20–28, 2010.

R. J. Verrall. An investigation into stochastic claims reserving models and the chain-ladder technique. *Insurance: Mathematics and Economics*, 26(1):91–99, 2000.

M. V. Wüthrich and M. Merz. *Stochastic claims reserving methods in insurance*, volume 435. John Wiley & Sons, 2008.

T. Žďálo. On parametric bootstrap and alternatives of mse. In *Proceedings of the 31st International Scientific Conference Mathematical Method in Economics, (Hana Vojáčková ed.)*, College of Polytechnics Jihlava, Jihlava, pages 1081–1086, 2013.