CLAIMS RESERVING WITHIN THE PANEL DATA FRAMEWORK

MICHAL GERTHOFER, PAVEL ZIMMERMANN, CLAUDIA FEDORČÁKOVÁ, RADEK BUDKA

University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability,
W. Churchill Sq. 4, Prague, Czech Republic
e-mail: michal.gerthofer@gmail.com, zimmerp@vse.cz, xfedc00@vse.cz, radekbudka@gmail.com

Abstract

Reserving in non-life insurance is a key factor for the financial position of a company. The text introduces the basic actuarial notation, terminology and methods. In the presented text the issue of dependency between response variables within the subjects in the generalized linear models framework is investigated. The main part is focused on panel data framework, especially Generalized Estimating Equations (GEE), and their application to claims reserving. The aim is to show the advantages, disadvantages and limitations of this approach on real dataset. Significant focus is set on model selection and residual diagnostics used for this purpose, which as well can help to justify the assumption of used models.

Key words: *stochastic claims reserving, panel data, generalized estimating equations.*

1. Introduction

The claims reserving problem or the run-off problem, has been one of the most important issues handled in general insurance for many years. The proper determination of the reserve amount has a key impact on the financial position of an insurance company, especially in the non-life insurance. Many various methods have been developed for this purpose beginning with deterministic approaches such as an original chain-ladder. Later, stochastic models have been proposed.

However, various approaches and models lead to different results. Every method is based on different ideas and uses different principles. Therefore, the most important question in the practical application of these methods is model selection and the adequacy of chosen models.

Stochastic methods for modelling of claims development became popular among practitioners in order to use more information about the data and consequently, to get deeper detail about the estimate as well. The major part of these approaches requires independency of the incremental claims. In practice, this assumption often does not hold and the methods can provide misleading results. This article deals with the stochastic models based on generalized linear models, especially Generalized Estimating Equations (GEE), which are able to handle the aforementioned dependency.

The purpose of the paper will be the construction of suitable models for claims reserving framework, then a description of model selection on real dataset. The structure of this article is as follows. In the first section, claims reserving problem in non-life insurance and its standard notation are introduced. The incremental or cumulative claims are understood as random variables ordered in the development triangles. At the end of the section, basic reserving methods are introduced.

The aim of the second section is to present the GEE models with all technicalities, data structure and advantages. For the purpose of the simplification of GEE models, testing of coefficients is discussed as well.

In the last section, the focus lies on the application of the GEE models. Specification of the models, which satisfy the theory associated with the actuarial practice, is discussed. In order to choose the most adequate model, useful residual properties are presented. Moreover, the reasons for simpler models, if it is possible, are listed.

Finally, the real-life data application of the GEE models will be performed. Consequently, model selection based mainly on residual diagnostic and ability to reduce number of parameters is performed using `R` software.

2. Introduction to Reserving Theory

This section deals with claims reserving, which is the main problem in non-life insurance. Models for life insurance are rather different due to the structure of products, nature of claims, risk drivers, term of contracts etc. These are the reasons for the separation of life and non-life insurance.

Non-life insurance offers financial coverage against various types of random occurrences in case that well-specified event happens. The value, which the insurer is obligated to pay as coverage, is called claim amount or the loss amount.

According to the type of claim non-life insurance is split into several Lines of Business (LoB), e.g., motor/car insurance, property insurance, liability insurance, accident insurance, etc. Number and types of LoBs vary through different insurance companies.

Reserving in non-life insurance needs a special approach because of a time-lag between claims occurrence and claims reporting to the insurer, which is called reporting delay. It can also take several years until the process is finally closed after the claim is reported. It is also possible that an already closed claim will need to be reopened because of new facts.

Due to the mentioned time-lag, the claim cannot be settled right after its accident day and so-called claims reserves have to be created. These reserves should represent all future claims arising from policies currently in force and policies written in the past. This amount of money should be held by the insurance company with the aim to meet its future liabilities.

There are two main types of reserves. The first one is reserves for claims that have been reported but have not been settled yet, so called RBNS (Reported But Not Settled). The second one is reserves for claims that have occurred but have not been reported, so called IBNR (Incurred But Not Reported). The last mentioned often contains reserves for not enough reported incurred claims IBNeR (Incurred But Not enough Reported).

It is worth mentioning that claim costs are often impacted by inflation. The main effect of inflation is not related on the salary or price but on the specifications of a particular LoB. For example, in the motor hull LoB, it is driven by the complexity of car repairing techniques and in LoB accident insurance, it is driven by improvements in medical care or in medicine. The impact of inflation develops through accident years as well as development years.

2.1 Reserving Terminology and Notation

Reserving approaches are based on history of claims. In order to capture all this information in standardized form, so-called claims development triangle is used, see Table 1. Let Y_{it} stand for all the claim amounts in development year t with accident year i. We refer to Y_{it} as incremental

claims in accident year i made in the accounting year $i+t$. Current year n corresponds to the most recent accident year as well as the most recent development year. The history of claims is placed in right-angled isosceles triangle $\{Y_{it}\}$, where $i = 1,\dots,n$ and $t = 1,\dots,n+1-i$. It is worth to note that we use following notation for indices, $\mathbf{Y}_{it} \equiv \mathbf{Y}_{i,t}$ and $\mathbf{Y}_{itj} \equiv \mathbf{Y}_{i,t,j}$.

Table 1: Run-off triangle for incremental claim amounts Y_{it}

Accident year	Development year t						
i	1	2	$\cdots$	t	$\cdots$	$n-1$	n
1	Y_{11}	Y_{12}	$\cdots$	Y_{1t}	$\cdots$	Y_{1n-1}	Y_{1n}
2	Y_{21}	Y_{22}	$\cdots$	Y_{2t}	$\cdots$	Y_{2n-1}	
$\vdots$	$\vdots$	$\cdots$	$\ddots$	$\cdots$	$\cdots$		
i	Y_{i1}	$\cdots$	$\cdots$	Y_{it}			
$\vdots$	$\vdots$	$\cdots$	$\cdots$				
n	Y_{n1}						

Source: the authors.

Let us denote a random variables C_{it}, cumulative payments or cumulative claims, in origin year i after t development years, e.i. $C_{it} = \sum_{k=1}^{t} Y_{ik}$. Observations of C_{it} for $i+t-1 \leq n$ form a cumulative run-off triangle. The task is to estimate the ultimate claims amount C_{in} and consequently, on calculating reserves for all accident years $i = 2,\dots,n$ as follows

$$R_i^{(n)} = C_{in} - C_{in+1-i}. \tag{1}$$

This article deals only with reserves defined in (1) and does not assume any tail factor.

2.2 Basic Reserving Methods

Early methods for distributing risk were practiced by Chinese and Babylonian traders as long ago as the 3rd and 2nd millennia BC, respectively. Modern insurance began in Europe where it became far more sophisticated and specialized. Insurance as we know it today is dated to 1667 and was founded after the Great Fire of London in 1666.

Claims reserving has significantly developed relatively recently. The first deterministic reserving model, original chain-ladder, was developed by Fisher and Lange (1973). Later, more complex approaches were needed, so a random part was added to the existing models, which resulted in stochastic models. The basis for these models was founded by Mack in 1993 and was built on assumption of proportionality of columns in a run-off triangle. These stochastic approaches are presented and described in Wuthrich and Merz (2008), in England and Verrall (2002) or in Huang and Wu (2012). Almost all of these proposed stochastic models require independence of incremental claims Y_{it}, which in practice does not often hold. Due to this, models such as GEE were introduced because of their ability to cope with dependencies within subjects.

3. Generalized Estimating Equations

In this section we present another approach called Generalized Estimating Equations (GEE) which is able to cope with correlated data within subjects (accident years). The main idea

behind GEE is to generalize and extend the usual likelihood equations from GLM (Generalized Linear model) by including the covariance matrix of the vector $\mathbf{Y}$. The biggest advantage of this model is that we do not need to specify the whole distribution of the response. Only the mean structure, the mean-variance relationship and specification of the covariance structure need to be defined. The second and the third conditions are similar to GLM, see definition below.

3.1 Definition of GEE Model

- We assume unbalanced design with independence between particular accident years $\mathbf{Y}_i$, $i = 1, \ldots, N$.
- Denote expected value of response $\mu_{it} \equiv \mathsf{E}(Y_{it})$, which depends on covariates, $\mathbf{X}_{it}$ as follows

$$g(\mu_{it}) = \eta_{it} = \mathbf{X}_{it}^{\top}\beta,$$

  where $\beta = (\beta_1, \ldots, \beta_p)^{\top}$ is the vector of regression parameters, $g(\cdot)$ is the link function and together with the linear predictor η_{it} fully specify the mean structure μ_{it}.
- It is also assumed for the variance of Y_{it} that

$$Var(Y_{it}) = \varphi V(\mu_{it}), \tag{2}$$

  where $V(\cdot)$ is a known variance function and $\varphi > 0$ is a scale or dispersion parameter, that can be known or may need to be estimated. It is worth to mention that we could consider that the (unknown) distribution belongs to the exponential family distributions.
- Furthermore, correlation between components of $\mathbf{Y}_i$ is represented by a working correlation matrix $\mathbf{C}_i \equiv \mathbf{C}_i(\alpha)$, where α is for our purpose scalar unknown parameter, however in general it could be assumed as vector.

The name "working" comes from the fact that the structure of $\mathbf{C}_i$ does not need to be correctly specified and asymptotic properties of estimate still hold. The corresponding working covariance matrix for i-th accident year can be constructed as the product of standard deviations and working correlation matrix

$$\mathbf{V}_i = \varphi \mathbf{A}_i^{1/2} \mathbf{C}_i(\alpha) \mathbf{A}_i^{1/2},$$

where $\mathbf{A}_i$ is diagonal matrix with $V(\mu_{it})$ along the diagonal.

3.2 Estimation of Parameters

In GEE the estimator of β minimizes the objective function (weighted least squares)

$$\sum_{i=1}^{N} (\mathbf{Y}_i - \mu_i)^{\top} \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mu_i), \tag{3}$$

where $\mathbf{V}_i$ is treated as known, $\mathbf{Y}_i \equiv (Y_{i1}, \ldots, Y_{in-i+1})^{\top}$ represents the particular row in run-off triangle and $\mu_i \equiv (\mu_{i1}, \ldots, \mu_{in-i+1})^{\top}$ is vector with elements given by

$$\mu_{it} = g^{-1}(\mathbf{X}_{it}^{\top}\beta).$$

Consequently, it can be shown that if a minimum of the function (3) exists, it must solve the generalized estimating equation

$$\mathbf{u}(\beta) = \sum_{i=1}^{N} \mathbf{D}_i^{\top} \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mu_i),$$

where $\mathbf{D}_i = \partial\mu_i/\partial\beta \equiv \{\partial\mu_{it}/\partial\beta_k\}_{t,k=1}^{n-i+1,p}$ and $\mathbf{u}(\beta)$ is the so-called quasi-vector. The estimate of β solves equation $\mathbf{u}(\widehat{\beta}) = 0$. Usually, parameters φ and α from $\mathbf{V}_i$ are unknown, so they can be estimated by moment estimates. We use the iterative algorithm. In one step, we estimate β and in the next step, we use this result for re-estimating $(\widehat{\varphi}, \widehat{\alpha})$ until the required precision is obtained.

The most important properties of the estimate $\widehat{\beta}$ are consistency, efficiency and asymptotic normality (for $N \to \infty$) with mean equal to β, variance equal to $Cov(\widehat{\beta})$, which hold even when the working correlation matrix $\mathbf{C}_i(\alpha)$ is misspecified.

It is shown in Fitzmaurice et al. (2004, Chapter 11.3), that for large samples the variance of $\widehat{\beta}$ can be expressed as follows

$$Cov(\widehat{\beta}) = \mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1},$$

where

$$\mathbf{B} = \sum_{i=1}^{N} \mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{D}_i, \qquad \mathbf{M} = \sum_{i=1}^{N} \mathbf{D}_i^\top \mathbf{V}_i^{-1} Cov(\mathbf{Y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i.$$

φ, α, β in $\mathbf{M}$ and $\mathbf{B}$ can be replaced by their estimates. Moreover $Cov(\mathbf{Y}_i)$ can be also replaced by $(\mathbf{Y}_i - \widehat{\mu}_i)(\mathbf{Y}_i - \widehat{\mu}_i)^\top$. Consequently, the estimate for variance of $\widehat{\beta}$, known as the empirically corrected variance estimates for $\widehat{\beta}$ or so-called sandwich estimator, is given by

$$\widehat{Cov}(\widehat{\beta}) = \\ = \left(\sum_{i=1}^{N} \widehat{\mathbf{D}}_i^\top \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i \right)^{-1} \left\{ \sum_{i=1}^{N} \widehat{\mathbf{D}}_i^\top \widehat{\mathbf{V}}_i^{-1} (\mathbf{Y}_i - \widehat{\mu}_i)(\mathbf{Y}_i - \widehat{\mu}_i)^\top \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i \right\} \left(\sum_{i=1}^{N} \widehat{\mathbf{D}}_i^\top \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i \right)^{-1}. \quad (4)$$

The expression from (4) is consistent estimator of $Cov(\widehat{\beta})$. Other properties of this estimate can be found in Liang and Zeger (1986), in Fitzmaurice et al. (2004, Chapter 11.3) or in Zeger et al. (1988).

3.3 Correlation Structure

Despite the fact that asymptotic normality of $\widehat{\beta}$ holds even when the correlation matrix is misspecified, a more precise choice of this matrix to the true one leads to more efficient estimates of β. Parameter α from $\mathbf{C}_i(\alpha)$ is assumed to be the same for all individuals and should be estimated in cases where it is unknown. Moreover, we have to specify how the correlation matrix should look like. There are several common choices for $\mathbf{C}_i(\alpha) = \{c_{jk}\}_{j,k=1}^{n-i+1,n-i+1}$.

- The first and simplest one is uncorrelated (or independent) structure

$$c_{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k. \end{cases}$$

- The next one is exchangeable structure

$$c_{jk} = \begin{cases} 1 & \text{if } j = k \\ \alpha & \text{if } j \neq k. \end{cases}$$

- The last one is an $AR(1)$ correlation structure

$$c_{jk} = \alpha^{|j-k|}.$$

For more details about the working correlation structure see in Hin and Wang (2009).

In order to determine a suitable correlation structure and variance function, Pearson residuals must be defined as follows

$$r_{it} = \frac{Y_{it} - \widehat{\mu}_{it}}{\sqrt{V(\widehat{\mu}_{it})}}, \tag{5}$$

where $V(\cdot)$ is still deterministic function defined in 2.

Now we describe a general strategy to estimate parametrized correlations by the method of moments. Firstly, an estimate of β under working independence must be calculated. Furthermore, Pearson residuals based on this model are expressed. Consequently, if the mean structure is correct, the following should hold

$$\begin{aligned} \mathsf{E}\, r_{it} &\approx 0 \\ Var(r_{it}) &\approx \varphi \\ \mathsf{E}\, r_{it} r_{ik} &\approx \varphi \{\mathbf{C}_i\}_{tk}, \quad i = 1,\ldots,N, \quad t \neq k \in \{1,\ldots,n-i+1\}. \end{aligned} \tag{6}$$

Next, moment estimate of α is calculated using these Pearson residuals, e.g., estimate of α for independent correlation structure is given by

$$\widehat{\alpha} = \frac{1}{\widehat{\varphi}} \frac{1}{\sum_{i=1}^{N}(n-i+1) - N - p} \sum_{i=1}^{N} \sum_{t=1}^{n-i} r_{it} r_{it+1},$$

where p is a number of parameters in vector β, $\widehat{\alpha}$ is in this case a one dimensional estimate and $\widehat{\varphi}$ is a moment estimate of φ given by

$$\widehat{\varphi} = \frac{1}{\sum_{i=1}^{N}(n-i+1) - p} \sum_{i=1}^{N} \sum_{t=1}^{n-i+1} r_{it}^2.$$

3.4 Testing Coefficients

This section deals with testing hypotheses for coefficients of vector β. We assume that coefficient vector β consists of two sub vectors γ, δ with r and l components respectively, for which $p = r + l$ holds. Thus, the vector of coefficients can be expressed as $\beta^\top = (\gamma^\top, \delta^\top)$. The main aim of this section is testing of hypothesis

$$H_0 : \gamma = \gamma_0,$$

where γ_0 is hypothesized value of γ.

There are several approaches for constructing of test statistics for hypothesis tests, e.g., likelihood ratio test, Wald test or score test. The first one mentioned cannot be applied to GEE directly because there is no likelihood underlying the model. However, this test can be used with modified assumptions, where the likelihood ratio is calculated under associated independence model. If zero hypothesis holds, then given test statistic has χ^2 distribution with r degrees of freedom.

The second mentioned test is calculated after model estimation. The test statistics are typically calculated without adjusting the degrees of freedom and use the sandwich estimate or model based estimate of $Cov(\widehat{\beta})$ from (4). The generalized Wald test statistic with sandwich estimate of variance is given by

$$W = n(\widehat{\gamma} - \gamma_0)^\top \widehat{Cov}(\widehat{\beta})^{-1} (\widehat{\gamma} - \gamma_0). \tag{7}$$

The test statistics is also assumed to follow χ^2 distribution with r degrees of freedom.

As we mentioned in the previous part dealing with the estimation of $Cov(\widehat{\beta})$, the sandwich estimate is not always the best choice. If there are more covariates than subjects or panels, it can cause the sandwich estimate of variance to be singular. Due to this fact, an alternative to the generalized Wald test is the working Wald test, where the sandwich estimate of covariance from (7) is replaced by model based estimate of covariance matrix. To use this approach, it must be assumed that working correlation matrix $\mathbf{C}_i$ describes the true correlation structure of the data. Likelihood ratio test and score test are described in Hardin and Hilbe (2003, Chapter 4.5).

Model selection can also be done using modified information criteria, but it should be used only in cases, where there is no other way to choose between models. Several information criteria for GEE models are presented in Hudecová and Pešta (2013).

4. Practical Application of GEE Models

In this section, the theory of GEE from Section 3 is applied on claims reserving presented in Section 2. It is worth to note that advantages of these models seem to be a suitable solution for the problem which is possible dependence among the incremental claims within accident year i.

4.1 Specification of the Models suitable for Reserving Problem

Before the application of the models to the data we have to define which models will be used, i.e., link function, variance function, mean structure as well as correlation structure.

4.1.1 Link Function

Commonly used link functions are the log-link, $g(\cdot) = \log(\cdot)$, or identity link function, $g(\cdot) = (\cdot)$. By using the log-link, the expected value of the response is assumed to be positive. Thus, it has some constrains on run-off triangle as well. In insurance practice, log-link is preferred due to its interpretation, so we will use this link function as well.

4.1.2 Linear Predictor

The choice of the linear predictor is a bit limited due to the interpretation and structure of claims data. Firstly, basic linear predictor, which uses $2(n-1)+1$ unknown parameters, is given by

$$\eta_{it} = \gamma + \alpha_i + \beta_t, \tag{8}$$

where $\alpha_1 = \beta_1 = 0$ and α_i represents effect of accident year i, β_t effect of development year t. This can also be expressed into vector notation

$$\eta_{it} = \mathbf{X}_{it}^{\top}\beta,$$

where $\beta = (\gamma, \alpha_2, \ldots, \alpha_n, \beta_2, \ldots, \beta_n)^{\top}$ and vector $\mathbf{X}_{it}$ is defined using dummy variables as follows $(1, d_{2i}, \ldots, d_{ni}, d_{2t}, \ldots, d_{nt})^{\top}$. Dummy variables are defined $d_{nm} = 1$ for $m = n$ and zero otherwise.

There are several other more complex and sophisticated linear predictors, e.g., Hoerl curve with with $3(n-1)+1$ parameters. However, we should realize that we have only $n(n+1)/2$ observations and model with $3(n-1)+1$ parameters, which is not very useful for our purpose.

The model from (8) is used in the practical part, even though it still has a lot parameters. It is important to have model with small number of parameters relatively to the number of observations because such models lead to better interpretations, the estimates become more precise and efficient, which are very important and practical properties.

Due to this, testing coefficient of particular accident or development year could be made in order to obtain suitable model with appropriate number of coefficients. In order to do this, the Wald test described in Subsection 3.4 can be used.

4.1.3 Variance Function

It is possible to define various functions but, in order to avoid confusion in the amount of models used in the practical part, only three basic variance functions are assumed

$$V(\mu_{it}) = \begin{cases} 1, \\ \mu_{it}, \\ \mu_{it}^2. \end{cases}$$

4.1.4 Correlation Structure

It is worth to mention that there exist way more complex correlation structures but only correlation structures introduced in Subsection 3.3 are used in our practical analysis. Reasons for our choice are the data structure of claims and the number of parameters that need to be estimated in a working correlation matrix. In chosen structures one or no parameter needs to be estimated which is in line with our needs for the low number of the unknown parameters.

The appropriateness of a model with given correlation structure is assessed after fitting the model, according to properties of Pearson residuals (6), as well as the whole residual diagnostic. In the practical part, we will use plot of fitted values with respect to the observed values to see how well the model will fit the data. Next, QQ plot, scatter plot, histogram of residuals and plot of classical residuals will be listed as well.

As we already mentioned at the end of Subsection3.4, there exist information criteria for comparison of the models. However, according this, only one, number we are not able to assess whether after fitting the data, assumptions of the model hold. This could be done by detailed residual diagnostic which is the main aim of the following text. It is necessary to note that each model and approach has different assumptions, therefore it is not always possible to compare more models. We should be aware of the fact that using model on data where assumptions do not hold, might cause that the results of this model could be misleading.

4.2 Application on Real Data

For GEE, nine models are analysed. They differ by the choice of variance function $V(\mu_{it})$ which can be equal to 1, μ_{it} or μ_{it}^2. Furthermore, the model is defined by correlation structure which, in our case, can be independent, exchangeable or $AR(1)$.

For the purpose of clarity in our tables and figures, abbreviations for GEE models are introduced. First letters stands for correlation structure, i.e., AR for $AR(1)$, IND for independent and EX for exchangeable correlation structure. Letters behind the underscore denote variance function, i.e., 1 for $V(\mu_{it}) = 1$, L for $V(\mu_{it}) = \mu_{it}$ and Q for $V(\mu_{it}) = \mu_{it}^2$. For example, GEE model with exchangeable correlation structure and variance function equal to 1 is labeled EX_1.

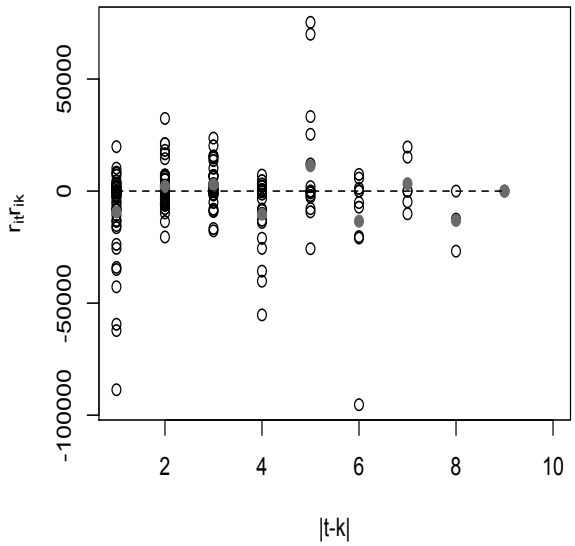
4.2.1 Dataset

The strengths of GEE models are shown on dataset Millers Mut Ins. Assoc. from Workers' compensation where all incremental claims in the upper triangle are positive as it is needed for application of GEE models with log-link. This data comes from the National Association of Insurance Commissioners (NAIC) database, which can be found in Meyers and Shi (2011). The volumes of claims from the database are reported in thousand USD. Chosen dataset contains the lower triangle as well, which is further used for calculation of real reserves and retrospective residual diagnostic.

4.2.2 Residual Diagnostic

According to the residual diagnostic of all models, GEE models fit the data very well. Especially models with exchangeable and $AR(1)$ correlation structure, with the same variance function equal to 1 have quite good results in diagnostic figures. In order to determine which correlation structure is more appropriate, we focus on the third property of Pearson residuals from (6). So we plotted these residuals from the model with independent correlation structure and variance function equal to 1 in Figure 1, where grey points symbolize arithmetic means for given values.

Figure 1: Products of Pearson residuals with respect to their distance within accident year based on the upper triangle



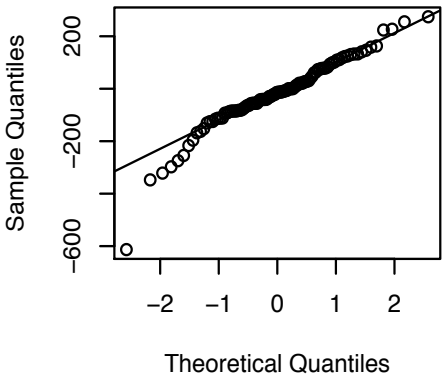
Source: the authors.

Based on this figure, we are not able to choose one of these two models due to a lot of outliers which influence these arithmetic means and an insufficient amount of data. Finally, we decide for model with exchangeable correlation structure because, as it will be shown later, some coefficients are not statistically significant and simpler linear predictor can be used. Unlike the model with $AR(1)$ correlation structure where all coefficients are statistically significant.

It is worth to state that we did not choose the model with independent correlation structure and variance function equal 1 due to the worse residual diagnostic, especially QQ plot see Figure 2. Note that for the variance function equal to 1 we could consider that the response has Gaussian distribution, therefore the points in the QQ plot should be placed on the drawn

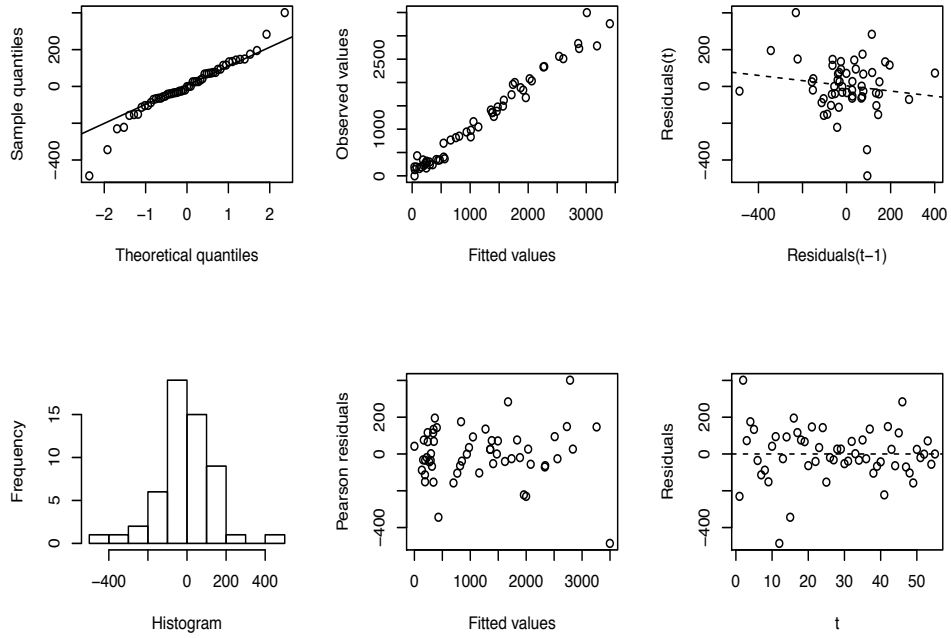
line. Other variance functions were not considered because of the decreasing pattern in Pearson residuals with respect to the fitted values.

Figure 2: QQ plot generated using upper triangle for model *IND*_1



Source: the authors.

Figure 3: Residual diagnostic generated using upper triangle for model *EX*_1



Source: the authors.

Figure 3 shows residual diagnostic for chosen model with exchangeable correlation structure, where the first diagnostic is a QQ plot, where almost all points are placed on the line except some outliers. Then, a plot of observed values with respect to fitted values is listed, where all points lay almost on a diagonal line, as expected. Next, a scatter plot of residuals indicates a very small correlation between residuals in development year t and $t-1$. The histogram is a bit skewed but close to Gaussian one. In the plot of Pearson residuals with respect to fitted values, no pattern is visible which is a very important indicator for the right choice of variance function as well. The last plot of classical residuals looks like we expected due to their symmetrical

placement around the zero. So, according to all above mentioned, the model with exchangeable correlation structure and variance function equal to one seems to be the most suitable.

Further, information from the lower triangle is used for the purpose of reserve calculation and retrospective residual diagnostic. Residual diagnostic based on whole rectangle, upper and lower triangle, was even better than the diagnostic from just upper triangle. To sum up, it confirms the suitability of the chosen model.

4.2.3 Claims Reserving

The predicted values are quite precise, except for the last accident year, where the biggest gap occurs in the second development year. This fact causes the predicted reserve to be quite higher than the real one.

According to the Table 2, our final model is the second worst in accuracy of prediction from all GEE models. However, GEE model with $AR(1)$ correlation structure and variance function equal to one, which also has similar reasonable residual diagnostic has the third best prediction of total reserve. In this case, the Mack-chain-ladder model provides reasonable prediction as well.

These results do not imply that our model is wrong. The difference is caused by outlier in the last accident year. It is necessary to remind that assumptions for our final model hold, however assumptions of other models do not have to hold as it was pointed out in residual analysis, thus results from other models could provide misleading results.

Table 2: Real and predicted reserves in thousand USD (shown values are rounded to integer values)

Reserves					
Real	Mack	GEE models	Predictions	GEE models	Predictions
9259	11064	*AR_Q*	10817	*IND_L*	11064
		AR_L	11084	*IND*_1	11194
		*AR*_1	10953	*EX_Q*	10656
		IND_Q	10656	*EX_L*	10994
				*EX*_1	11171

Source: the authors.

4.2.4 Coefficient Tests

The chosen software provides the generalized Wald test with sandwich estimate of variance as mentioned in Subsection 3.4. According to our results, coefficients for accident years 4 and 6 are not statistically significant on a given level equal to 5 %. This fact is quite reasonable because the values of these coefficients are very close to zero.

It is not suitable to put all coefficients for these two accident years equal to 0 based on the current results. It would be possible to use the test of joint hypotheses, but in this analysis, we apply a backward elimination approach using P-values as a criterion. Hence, the coefficient for accident year 6 with the highest P-value is excluded from the model and then again whole residual diagnostic is made in order to ensure that the model still fits the data well. In this case, the residual diagnostic does not change significantly.

Next, a test of coefficients is computed on the simpler model without effect for accident year 6. According to the results of this test, coefficient for accident year 4 has the highest P-value (0.949), so this factor is excluded from the model as well. After this, a residual diagnostic of model without effects for accident years 4 and 6 was generated. This model still provides comparable results with the original model.

It is worth to note that the precision of reserve prediction gets worse, from 11172 in original model to 13316 in model without two accident year effects. It is worth mentioning that after excluding only the first mentioned coefficient, prediction was even worse (13642) than after excluding both of them.

5. Conclusion

The GEE approach was presented in order to demonstrate its ability to cope with the within-subject correlation using the working correlation matrix. Our attention has been paid to estimates of unknown regression parameters and unknown working correlation matrix as well as to their properties. Consequently, instruments such as Pearson residual and tests for coefficients used for model selection were introduced.

One of the goals has been to choose the models, which are proper for the insurance data. Focus has been put to the interpretation of the model and the number of unknown regression parameters that must be estimated. Due to this, log-link function has been chosen and testing of the coefficients as well as choice of the appropriate correlation structures has been made.

Finally, the application of the proposed models on real data has been carried out for the purpose of their analysis and comparison of their performance. All computations have been performed in R software. The GEE models applied on real dataset has quite good residual diagnostic as well as prediction of reserves. We were not able to choose the best model out of the two final models according to residual diagnostic, therefore properties of Pearson residual were used for the correlation analysis, however, it did not work due to the insufficient number of observations. Nevertheless, the final model was selected according to the possibility of reduction of the coefficient. Very important fact is that independent models were not even considered between final ones. This fact implies that incremental data in our dataset are not independent.

The reserves estimate by the final model was not the best, however this does not imply that this model is wrong. The difference is caused by the development of the last accident year which could be categorized as an outlier. Nevertheless, assumptions of this model hold therefore it is the most suitable for this purpose. Further, we improved the efficiency of the estimation by reduction of the parameters using Wald test.

Further, this approach could be used for modelling claims reserving in micro models which means that we would be able to model claims on each policy. It demands more granular data, however, we could use more complex models which could lead to more precise results. Moreover using bootstrap predictive distribution could be obtained.

Acknowledgements

The support of the grant scheme IGA 70/2016 is gladly acknowledged.

References

[1] ENGLAND, P. D., VERRALL, R. J. 2002. Stochastic claims reserving in general insurance. In British Actuarial Journal, 2002, vol. 8, iss. 3, pp. 443-518.

[2] FISHER, W. H., LANGE, J. T. 1973. Loss reserve testing : A report year approach. In Poceedings of the Casualty Actuarial Society, 1973, vol. 60, pp. 189-207.

[3] FITZMAURICE, G. M., LAIRD, N. M., WARE, J. H. 2004. Applied longitudinal analysis. Boston : Wiley, 2004.

[4] HARDIN, J. W., HILBE, J. 2003. Generalized estimating equations. Boca Raton : Chapman and Hall, 2003.

[5] HIN, L. Y., WANG, Y. G., 2009. Working-correlation-structure identification in generalized estimating equations. In Statistics in Medicine, 2009, vol. 28, pp. 642-658.

[6] HUANG, J., WU, X. 2012. Stochastic claims reserving in general insurance : Models and methodologies. China International Conference on Insurance and Risk Management. Qingdao, 2012, pp. 694-724.

[7] HUDECOVÁ, Š., PEŠTA, M. 2013. Modeling dependencies in claims reserving with GEE. In Insurance : Mathematics and Economics, 2013, vol. 53, pp. 786-794.

[8] LIANG, K., ZEGER, S. L. 1986. Longitudinal data analysis using generalized linear models. In Biometrika, 1986, vol. 73, iss. 1, pp. 13-22.

[9] MEYERS, G.G., SHI, P. 2011. Loss reserving data pulled from NAIC Schedule P. [cit. 17-04-2016] http://www.casact.org/research/index.cfm?fa=loss_reserves_data.

[10] WUTHRICH, M. V., MERZ, M. 2008. Stochastic claims reserving methods in insurance. Hoboken : Wiley. 2008.

[11] ZEGER, S., LIANG, K., ALBERT, P. 1988. Models for longitudinal data : A generalized estimating equation approach. In Biometrics, 1988, vol. 44, iss. 4, pp. 1049-1060.