FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Tereza Smolárová

Tweedie models for pricing and reserving

Department of Probability and Mathematical Statistics

Supervisor of the master thesis:	RNDr. Michal Pešta, Ph.D.
Study programme:	Mathematics
Study branch:	Financial and Insurance Mathematics

Prague 2017

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In Prague, 12th May 2017 Tereza Smolárová

Title: Tweedie models for pricing and reserving

Author: Tereza Smolárová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Michal Pešta, Ph.D., Department of Probability and Mathematical Statistics

Abstract:

This presented thesis deals with applications of Tweedie compound Poisson model in non-life insurance pricing and claims reserving. Tweedie models are exponential dispersion models with power mean-variance relationships and compound Poisson distribution is a particular Tweedie model. The interest in Tweedie compound Poisson model is motivated by its applications to generalized linear models (GLMs) and generalized estimation equations (GEE). The purpose of this thesis is to construct pricing and claims reserving models in which the response variables follow Tweedie compound Poisson model. Theoretical approaches are applied on the real datasets.

Keywords: Tweedie models, non-life insurance, pricing, reserving, generalized linear models

I would like to thank my supervisor, RNDr. Michal Pešta, Ph.D., for his time, guidance and valuable advice.

I am also grateful to my family and friends for their support during my studies.

Contents

List of Figures

List of Tables

Introduction

A non-life insurance policy is a contract between a policyholder and an insurer, e.g. an insurance company. In such contract, the insurer undertakes to compensate the policyholder for certain unpredictable losses during a time period, usually one year, in exchange for a fee called *premium*. A claim is an event reported by the policyholder, for which they demand economic compensation.

A non-life insurance policy may cover property damages, motor-car accidents, personal injuries resulting from an accident etc. In effect, any insurance that is not life insurance is classified as non-life insurance.

Tweedie models are a class of exponential dispersion models with power mean-variance relationships. This thesis focuses on compound Poisson distribution which is a particular Tweedie model. This distribution is a mixed distribution with mass at zero and continuous on the positive real numbers. The approach based on this distribution provides a highly efficient method of analysing insurance claims data, especially claim totals. The reason for concern in Tweedie compound Poisson model is its applications to the generalized linear models (GLMs) and the generalized estimation equations (GEE).

The purpose of this thesis is to construct pricing and claims reserving models in which the response variables follow Tweedie compound Poisson model.

The advantage of all Tweedie regression models is that they allow a simultaneous modelling of both the occurrence and the amount of claim. In addition, Tweedie GEE models also allow multi-subject modelling of claim, whilst taking into account the dependent nature of claims data. The difficulty in using GEE models is that there is no software available that fits the GEE with Tweedie compound Poisson model, therefore it is necessary to write a code so this type of modelling could be implemented. All computations in this thesis will be performed in the R statistical software.

This thesis is structured as follows. In the first chapter, exponential dispersion models as well as Tweedie models are introduced. These distributions represent a key component of GLMs theory. The second chapter presents all the important components of generalized linear models framework. The generalized estimating equations are discussed as a possible extension of GLMs to handle correlated data. The third chapter defines compound Poisson distribution. By choosing a suitable parametrization, it will be shown that the distribution corresponds to a Tweedie model with specific values of power index parameter. The aim of the fourth and the fifth chapters is to show how Tweedie compound Poisson model can be used in the context of non-life insurance pricing and claims reserving.

1. Exponential Dispersion Models

Exponential dispersion models (EDMs) represent a family of distributions that has found place in various fields. They have an important role in statistical data analysis as the response distributions for generalized linear models. Jørgensen (1997) undertook a detailed study of their properties and applications.

Many well-known distributions, such as normal, Poisson, gamma or inverse Gaussian are members of the exponential dispersion model family.

1.1 Definition and properties

Let a random variable Y follow an exponential dispersion model. The probability density function or the probability mass function of Y can be written in the following form

$$f_Y(y;\theta,\phi/w) = a(y;\phi/w)\exp\left\{\frac{w}{\phi}\big(y\theta-\kappa(\theta)\big)\right\}, \qquad y\in S\subseteq\mathbb{R}, \tag{1.1}$$

where $a(\cdot;\cdot)$ and $\kappa(\cdot)$ are real functions, $\theta\in\mathbb{R}$ is a *canonical parameter*, $\phi>0$ is a *dispersion parameter* and $w>0$ is a known prior *weight*. The function $\kappa(\cdot)$ is assumed to be twice continuously differentiable with an invertible first derivative. The function $a(\cdot;\cdot)$ only acts as a normalizing function and it cannot always be written in closed form.

The moment generating function $M_Y(t)$ of Y exists, is finite over the whole range and equals to

$$M_Y(t) = \mathsf{E}\exp\{tY\} = \exp\left\{\frac{w}{\phi}\left(\kappa(\theta+\frac{t\phi}{w})-\kappa(\theta)\right)\right\}. \tag{1.2}$$

Note that since $\kappa(\theta)$ is twice continuously differentiable then $M_Y(t)$ is twice differentiable at $t=0$. Thus, we use the property of moment generating function to obtain the following relations for mean and variance

$$\begin{aligned}\mathsf{E}\,Y &= \mu = \kappa'(\theta),\\ \operatorname{var}Y &= \frac{\phi}{w}\kappa''(\theta),\end{aligned}$$

where prime denotes differentiation with respect to θ.

An EDM is uniquely characterized by its *variance function* $V(\cdot)$ defined as

$$V(\mu) = \kappa''(\theta) = \kappa''\big((\kappa')^{-1}(\mu)\big),$$

which describes the mean-variance relationship of the distribution when the dispersion parameter is held constant. Then the variance of Y can be written as

$$\operatorname{var}Y = \frac{\phi}{w}V(\mu).$$

The specification of the parameters and the functions from (1.1) as well as the variance functions of the forenamed distributions are listed in McCullagh and Nelder (1989, Section 2.2.2).

Any EDM can be parametrized also in terms of its mean instead of its canonical parameter. We use the notation $Y \sim \mathrm{ED}(\mu, \phi/w)$ to indicate that a random variable Y follows an EDM with parameters μ and ϕ.

Following EDM's property is relevant in practice as we see later on. Assume that $Y_1, \ldots, Y_n$ are independent and

$$Y_i \sim \mathrm{ED}(\mu, \phi/w_i) \quad \text{for } i = 1, \ldots, n.$$

Then we have

$$Y = \frac{1}{w_+} \sum_{i=1}^{n} w_i Y_i \sim \mathrm{ED}(\mu, \phi/w_+), \tag{1.3}$$

where $w_+ = w_1 + \ldots + w_n$. The formula (1.3) shows that w-weighted average Y follows the same EDM but with the weight w_+. Thus, we say that EDMs are *reproductive.*

1.2 Tweedie Models

Of our special interest within exponential dispersion models is a class of distributions with the variance functions of the form

$$V(\mu) = \mu^p, \tag{1.4}$$

for some p. Following Jørgensen (1997), these EDMs are called *Tweedie models* after Maurice C. K. Tweedie, a British statistician and medical physicist, who presented the first thorough study of them in 1984. Dunn and Smyth (2005) gave a survey of published applications showing that Tweedie models have been used in various fields including actuarial studies, survival analysis, ecology and meteorology. Notation $Y \sim \mathrm{ED}_p(\mu, \phi/w)$ indicates that the random variable Y is distributed as a Tweedie model with the mean μ, dispersion ϕ and the *power index parameter* $p \in \mathbb{R}$.

Jørgensen (1997, Section 4.1.1) demonstrated that Tweedie models are the only EDMs which are closed with respect to scale transformations, i.e. scale invariant. Thus, if $Y \sim \mathrm{ED}_p(\mu, \phi/w)$ then $cY \sim \mathrm{ED}_p(c\mu, c^{2-p}\phi/w)$ for any positive constant c. This fundamental property makes them obvious candidates for modelling data when the unit of measurement is arbitrary. For example, the result of an actuarial analysis should not depend on the currency used.

The class of Tweedie models includes discrete, continuous as well as mixed distributions. The value of the power index parameter p determines the distribution. The normal ($p = 0$), Poisson ($p = 1$ and $\phi = 1$), gamma ($p = 2$) and inverse Gaussian ($p = 3$) distributions are the special cases. Although the other distributions might be less well-known, Jørgensen (1997, Proposition 4.2) showed that Tweedie models exist for all values of p outside the interval $(0, 1)$.

Tweedie models with $1 < p < 2$ are Poisson mixtures of gamma distributions. These so called *compound Poisson distributions* are mixed distributions supported on non-negative real numbers with positive probability of taking value zero. The presence of the discrete mass at zero makes these distributions suitable for many applications where observations are often zero but sometimes are positive. In non-life insurance, they can be applied in pricing and also in claims reserving. They have also been known as *compound gamma* and *Poisson-gamma distributions.*

For $p > 2$, the distributions are continuous with support on positive real numbers and have similar shape to the gamma distribution but more rightly skewed. Together with gamma distribution, they are often suggested as distributions for the claim severity. Negative values of p give continuous distributions on the whole real axis, but no application in insurance has been proposed yet.

The canonical parameter θ and the mean μ can be found for a Tweedie model by equating $\kappa''(\theta) = V(\mu) = \mu^p$. Hence

$$\mu^p = \frac{\partial^2\kappa}{\partial\theta^2} = \frac{\partial}{\partial\theta}\Big(\frac{\partial\kappa}{\partial\theta}\Big) = \frac{\partial\mu}{\partial\theta}.$$

Taking the reciprocals of both sides, integrating with respect to μ and setting the arbitrary constant of integration to zero gives expressions for canonical parameter θ, for specific values of the power index parameter p. Thus

$$\theta = \begin{cases} \dfrac{\mu^{1-p}}{1-p} & \text{for } p \neq 1, \\ \log\mu & \text{for } p = 1, \end{cases} \tag{1.5}$$

with inverse

$$\mu = \begin{cases} \big((1-p)\theta\big)^{\frac{1}{1-p}} & \text{for } p \neq 1, \\ e^\theta & \text{for } p = 1. \end{cases} \tag{1.6}$$

Let $\kappa_p(\theta)$ denote the function $\kappa(\theta)$ for a Tweedie model. We use the above expression for μ to find function $\kappa_p(\theta)$. Integrating both sides of the equation $\kappa_p'(\theta) = \mu$ with respect to θ and setting the arbitrary constant of integration to zero gives

$$\kappa_p(\theta) = \begin{cases} \dfrac{1}{2-p}\big((1-p)\theta\big)^{\frac{2-p}{1-p}} & \text{for } p \neq 1, 2, \\ e^\theta & \text{for } p = 1, \\ -\log(-\theta) & \text{for } p = 2. \end{cases} \tag{1.7}$$

By using (1.5) and (1.6), we can express $\kappa_p(\theta)$ as a function of mean μ

$$\kappa_p(\theta) = \begin{cases} \dfrac{\mu^{2-p}}{2-p} & \text{for } p \neq 2, \\ \log\mu & \text{for } p = 2. \end{cases} \tag{1.8}$$

The moment generating function of a Tweedie model is obtained by inserting the expression for function $\kappa_p(\theta)$ (1.7) into (1.2). Hence

$$\begin{aligned} M_Y(t) &= \exp\left\{\frac{w}{\phi}\Big(\kappa_p(\theta + \frac{t\phi}{w}) - \kappa_p(\theta)\Big)\right\} \\ &= \exp\left\{\frac{w}{\phi}\kappa_p(\theta)\Big(\Big(1 + \frac{t\phi}{w\theta}\Big)^{\frac{2-p}{1-p}} - 1\Big)\right\}. \end{aligned}$$

We can express $M_Y(t)$ as a function of mean μ by using the identities (1.8) and (1.5)

$$M_Y(t) = \exp\left\{\frac{w}{\phi}\frac{\mu^{2-p}}{(2-p)}\left(\left(1 - \frac{t\phi}{w}\frac{(p-1)}{\mu^{1-p}}\right)^{\frac{2-p}{1-p}} - 1\right)\right\} \quad \text{for } p \neq 1, 2. \tag{1.9}$$

The above expressions show that the moment generating function of a Tweedie model has a simple analytic form. On the other hand, according to Dunn and Smyth (2005), apart from the four special cases, none of the Tweedie models has density functions which can be written in closed form. Hence, the Tweedie densities can be expressed in general form as

$$f_Y(y; \mu, \phi/w, p) = a(y; \phi/w, p) \exp\left\{\frac{w}{\phi}\left(y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right)\right\} \quad \text{for } p \neq 1, 2,$$

where the function $a(y; \phi/w, p)$ needs to be evaluated numerically. Dunn and Smyth (2001) proposed numerical methods for their computation.

Plots of some Tweedie densities for various values of p are given in Figure 1.1. In all cases, the mean and the variance are fixed at unity. Note that the distribution approaches the Poisson as $p \to 1$ and the gamma as $p \to 2$.

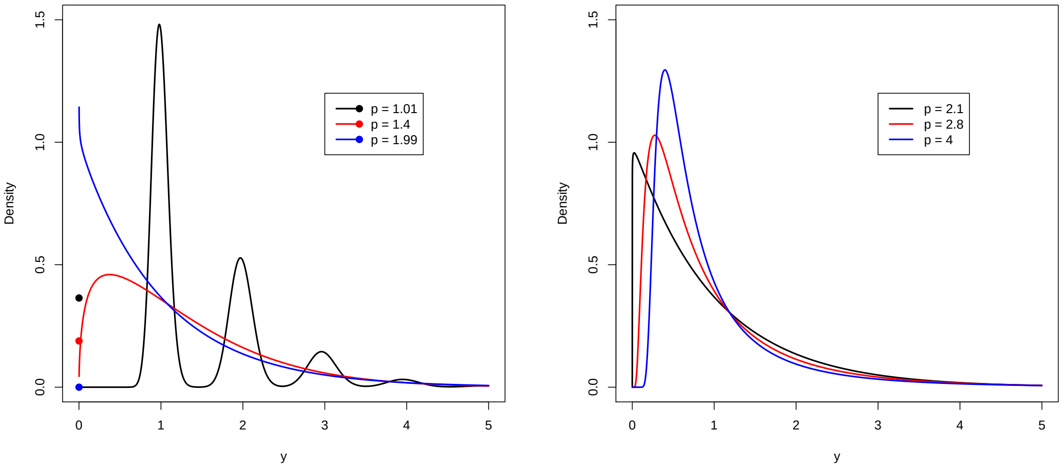

Figure 1.1: Tweedie densities for various values of p and other parameters being set to one. The solid circles represent discrete probability of $Y = 0$.

2. Generalized Linear Models and their extension

Generalized linear models (GLMs) and their extensions are becoming the premier statistical analysis methods for insurance data. Haberman and Renshaw (1996) demonstrated the applications of these models to a wide range of actuarial problems, such as mortality, lapses, premium setting and claims reserving in non-life insurance. This thesis focuses on the last two mentioned.

2.1 Classical Linear Model

The classical linear models are considered as the special case of GLMs. The primary interest of the classical linear regression analysis is to model the marginal expectation of a response variable given the explanatory variables.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ denote an $(n \times 1)$ random vector called a *response vector.* Then the expected value of vector $\boldsymbol{Y}$ is denoted by a vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$.

2.1.1 Definition of Classical Linear Model

Based on McCullagh and Nelder (1989, Section 2.2), to simplify the transition to generalized linear models, a classical linear model is defined as follows:

1. The random variables $Y_1, \ldots, Y_n$ are independent and the distribution of Y_i depends on the explanatory variables $\mathbf{x}_i$ through a $(k \times 1)$ vector of unknown regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$.

2. The response variable Y_i is normally distributed with mean μ_i and dispersion parameter ϕ, i.e.
$$Y_i \sim \mathcal{N}(\mu_i, \phi/w_i),$$
where w_i is a known prior weight.

3. A linear combination of explanatory variables is considered
$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \sum_{j=1}^{k} x_{ij}\beta_j,$$
where η_i is called a *linear predictor* and $\mathbf{x}_i^\top$ is a row vector of an $(n \times k)$ known matrix $\boldsymbol{X}$ called a *regression* or a *design matrix.* We assume that the design matrix $\boldsymbol{X}$ has full column rank and $k < n$.

4. The expected value μ_i is related to the linear predictor through the identity
$$\mu_i = \eta_i.$$

2.2 Generalized Linear Models

Generalized linear models represent a rich class of statistical models, which generalizes the classical linear models in two different directions. Firstly, the response variable Y may have a distribution other than normal distribution, i.e it can follow any distribution that belongs to the class of exponential dispersion models. Secondly, some monotone transformation of the mean is a linear function of the explanatory variables.

2.2.1 Definition of Generalized Linear Model

We would like to describe the dependence of $\mu_i = \mathsf{E}\, Y_i$ on the explanatory variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^\top$ by a regression model that is more general than the linear model.

A generalized linear model is defined as follows:

1. The random variables $Y_1, \ldots, Y_n$ are independent and the distribution of Y_i depends on the explanatory variables $\mathbf{x}_i$ through regression parameters $\boldsymbol{\beta}$.

2. The response variable Y_i follows an EDM with mean μ_i and dispersion parameter ϕ, i.e.
$$Y_i \sim \mathrm{ED}(\mu_i, \phi/w_i),$$
where w_i is a known prior weight.

3. A linear combination of explanatory variables is considered
$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \sum_{j=1}^{k} x_{ij}\beta_j,$$
where η_i is called a *linear predictor.*

4. There exists a known strictly monotone and twice continuously differentiable *link function* $g(\cdot)$ such that
$$g(\mu_i) = \eta_i.$$

Each of the EDMs has a natural link function, a so-called *canonical link function.* This link function relates the canonical parameter θ_i directly to the linear predictor η_i,
$$\eta_i = g(\mu_i) = \theta_i.$$
Nevertheless, the link function can be chosen independently of the response's distribution and the choice depends on the character of data. There are several commonly used link functions, e.g. identity, logarithmic or reciprocal function. In non-life insurance pricing and claims reserving the logarithmic link function is by far the most common one, since a multiplicative model is often reasonable.

Summing up the previous conclusions, the parameters are related to each other as follows:

- $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$;
- $\mu_i = \kappa'(\theta_i), \theta_i = (\kappa')^{-1}(\mu_i)$;
- $\eta_i = g(\mu_i), \mu_i = g^{-1}(\eta_i)$;
- $\eta_i = g\big(\kappa'(\theta_i)\big), \theta_i = (\kappa')^{-1}\big(g^{-1}(\eta_i)\big)$.

2.2.2 Parameter estimation

Let the definition of GLM hold. A vector of independent observations $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ is a realization of response vector $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$. Maximum likelihood method is used to estimate the vector of regression parameters $\boldsymbol{\beta}$.

The log-likelihood function for parameters $(\boldsymbol{\beta}, \phi)$ has the following form

$$\ell(\boldsymbol{\beta}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \theta_i, \phi/w_i) = \sum_{i=1}^{n} \log a(y_i; \phi/w_i) + \frac{1}{\phi}\sum_{i=1}^{n} w_i\big(y_i\theta_i - \kappa(\theta_i)\big), \tag{2.1}$$

where θ_i is connected to $\boldsymbol{\beta}$ through

$$\theta_i = (\kappa')^{-1}\big(g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})\big).$$

The partial derivative of log-likelihood function (2.1) with respect to β_j is then

$$\frac{\partial \ell}{\partial \beta_j} \overset{\substack{\text{chain}\\ \text{rule}}}{=} \frac{\partial \ell}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{V(\mu_i)g'(\mu_i)} x_{ij}.$$

By setting all these k partial derivatives equal to zero and multiplying by ϕ, which does not have any effect of maximization, we get the maximum likelihood equations

$$\sum_{i=1}^{n} w(\mu_i)g'(\mu_i)(y_i - \mu_i)x_{ij} = 0, \qquad j = 1, \dots, k, \tag{2.2}$$

where $\mu_i = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$ and $w(\cdot)$ is a *weight function* of form

$$w(\mu_i) = \frac{w_i}{V(\mu_i)\big(g'(\mu_i)\big)^2} > 0.$$

We may rewrite (2.2) into a matrix form. Let $\boldsymbol{W} = diag\{w(\mu_1), \dots, w(\mu_n)\}$ and $\boldsymbol{G} = diag\{g'(\mu_1), \dots, g'(\mu_n)\}$ be an $(n \times n)$ diagonal matrices. Then a maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ solves the equation

$$\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{0}. \tag{2.3}$$

By adding the term $(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})\boldsymbol{\beta}$ to both sides of equation (2.3) and rearranging it under the regularity assumption of matrix $(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})$ we obtain

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}(\boldsymbol{X}^\top \widehat{\boldsymbol{W}}\widehat{\boldsymbol{z}}), \tag{2.4}$$

where the vector $\widehat{\boldsymbol{z}} = (\widehat{z_1}, \dots, \widehat{z_n})^\top$ with components $\widehat{z_i} = g(\widehat{\mu_i}) + g'(\widehat{\mu_i})(y_i - \widehat{\mu_i})$ is called an *adjusted response*.

We can not calculate $\widehat{\boldsymbol{\beta}}$ directly from (2.4) because the matrix $\widehat{\boldsymbol{W}}$ and the vector $\widehat{\boldsymbol{z}}$ depend on $\widehat{\boldsymbol{\mu}}$ and hence on $\widehat{\boldsymbol{\beta}}$. We use *iterative weighted least squares* (IWLS) algorithm to obtain the maximum likelihood estimates:

1. Take initial values $\widehat{\mu_i}^{(0)} = y_i$ and set l to zero.
2. Compute matrix of weights $\widehat{\boldsymbol{W}}^{(l)} = diag\{w(\widehat{\mu_1}^{(l)}), \dots, w(\widehat{\mu_n}^{(l)})\}$ and components of adjusted response vector $\widehat{z_i}^{(l)} = g(\widehat{\mu_i}^{(l)}) + g'(\widehat{\mu_i}^{(l)})(y_i - \widehat{\mu_i}^{(l)})$.
3. Compute $\widehat{\boldsymbol{\beta}}^{(l+1)}$ by following iterative formula
$$\widehat{\boldsymbol{\beta}}^{(l+1)} = (\boldsymbol{X}^\top \widehat{\boldsymbol{W}}^{(l)} \boldsymbol{X})^{-1} (\boldsymbol{X}^\top \widehat{\boldsymbol{W}}^{(l)} \widehat{\boldsymbol{z}}^{(l)}).$$
4. Calculate $\widehat{\mu_i}^{(l+1)} = g^{-1}(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^{(l+1)})$ and increase l by one.
5. Repeat steps 2–4 until the convergence criterion $\|\widehat{\boldsymbol{\beta}}^{(l+1)} - \widehat{\boldsymbol{\beta}}^{(l)}\| < \varepsilon$ is fulfilled. Note that ε is a pre-specified tolerance parameter.

The dispersion parameter ϕ is also often unknown. One of the most commonly used estimators for ϕ in generalized linear models is the *Pearson estimator* defined as
$$\widehat{\phi} = \frac{1}{n-k} \sum_{i=1}^{n} w_i \frac{(y_i - \widehat{\mu_i})^2}{V(\widehat{\mu_i})},$$
where k is the number of regression parameters and $V(\widehat{\mu_i})$ is estimated variance function for the considered distribution.

2.3 Generalized Estimating Equations

While the GLM extends the classical linear model by having a more general family of distributions and relation between the mean and the linear predictor, the independence assumption is retained. Generalized estimating equations (GEE) were first proposed by Liang and Zeger (1986) as a method for estimating the regression parameters in case the independence assumption is violated.

Suppose there are K subjects with n_i measurements available for the i^{th} subject. Let $\boldsymbol{Y}_i = (Y_{i,1}, \dots, Y_{i,n_i})^\top$ denote a response vector and $\boldsymbol{y}_i = (y_{i,1}, \dots, y_{i,n_i})^\top$ represents a vector of response realizations for subject i. Further, $\boldsymbol{X}_i$ is an $(n_i \times k)$ design matrix for subject i and an expected value of $\boldsymbol{Y}_i$ is denoted by a vector $\boldsymbol{\mu}_i = (\mu_{i,1}, \dots, \mu_{i,n_i})^\top$.

2.3.1 Definition of GEE Model

We would like to describe the dependence of $\mu_{i,j} = \mathsf{E}\, Y_{i,j}$ on the explanatory variables $\mathbf{x}_{i,j}$ by a regression model which is able to deal with correlated data within the subjects.

A GEE model is defined as follows:

1. The random vectors $\boldsymbol{Y}_i, \dots, \boldsymbol{Y}_K$ are independent, however, the components of $\boldsymbol{Y}_i$ are allowed to be correlated.

2. A linear combination of explanatory variables is considered
$$\eta_{i,j} = \mathbf{x}_{i,j}^\top \boldsymbol{\beta},$$
where $\eta_{i,j}$ is called a *linear predictor*.

3. The expected value $\mu_{i,j}$ satisfies the identity
$$g(\mu_{i,j}) = \eta_{i,j},$$
where $g(\cdot)$ is a known strictly monotone and twice continuously differentiable link function. Linear predictor $\eta_{i,j}$ together with the link function $g(\cdot)$ fully specify the *mean structure*.

4. The variance of response $Y_{i,j}$ can be expressed as a function of the mean
$$\operatorname{var} Y_{i,j} = \frac{\phi}{w_{i,j}} V(\mu_{i,j}),$$
where $V(\cdot)$ is a known variance function and $w_{i,j} > 0$ is a known prior weight. The dispersion parameter $\phi > 0$ may or may not be known.

5. The correlation between components of $\boldsymbol{Y}_i$ is represented by an $(n_i \times n_i)$ *working correlation matrix* $\boldsymbol{R}_i \equiv \boldsymbol{R}_i(\boldsymbol{\vartheta})$, where $\boldsymbol{\vartheta}$ is an $(s \times 1)$ vector of unknown parameters, which is the same for all the subjects.

An important aspect of GEE framework is that we have to specify a form of the working correlation matrix $\boldsymbol{R}_i$. The name "working" comes from the fact that we still obtain a consistent and asymptotically normal estimate of $\boldsymbol{\beta}$ even if the correlation structure of $\boldsymbol{R}_i$ is misspecified though a closer choice to the true correlation structure leads to a more efficient estimate.

Some of the most commonly used structures of working correlation matrix $\boldsymbol{R}_i(\boldsymbol{\vartheta}) = \{r_{j,l}\}_{j,l=1}^{n_i,n_i}$ are presented below. Their detailed descriptions and other possibilities can be found in Hardin and Hilbe (2003, Section 3.2.1).

- The simplest choice is an *uncorrelated* (or *independent*) structure
$$r_{j,l} = \begin{cases} 1 & \text{if } j = l, \\ 0 & \text{if } j \neq l. \end{cases}$$

- The simple extension of previous structure is an *exchangeable* (or *equal*) correlation structure
$$r_{j,l} = \begin{cases} 1 & \text{if } j = l, \\ \vartheta & \text{if } j \neq l. \end{cases}$$

- Another common choice is a first-order *autoregressive AR(1)* correlation structure
$$r_{j,l} = \begin{cases} 1 & \text{if } j = l, \\ \vartheta^{|j-l|} & \text{if } j \neq l. \end{cases}$$

- The most general of the correlation structures presented is an *unstructured* correlation structure
$$r_{j,l} = \begin{cases} 1 & \text{if } j = l, \\ \vartheta_{jl} & \text{if } j \neq l. \end{cases}$$

Consequently, a *working covariance matrix* for ith subject is defined as

$$\boldsymbol{V}_i = \phi \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\vartheta}) \boldsymbol{A}_i^{1/2},$$

where $\boldsymbol{A}_i = diag\{V(\mu_{i,1})/w_{i,1}, \ldots, V(\mu_{i,n_i})/w_{i,n_i}\}$ is an $(n_i \times n_i)$ diagonal matrix. This working covariance matrix will be equal to true covariance matrix $\mathsf{cov}\boldsymbol{Y}_i$ if $\boldsymbol{R}_i(\boldsymbol{\vartheta})$ is indeed the true correlation matrix for the response vector $\boldsymbol{Y}_i$.

Note that within GEE framework, it is not necessary to specify the whole distribution of the response variable. Only the mean structure, the mean-variance relationship and the form of the correlation structure need to be defined.

2.3.2 Parameter estimation

Let the definition of GEE model hold. For now, we assume that estimates $(\widehat{\phi}, \widehat{\boldsymbol{\vartheta}})$ of parameters $(\phi, \boldsymbol{\vartheta})$ are given. The estimator $\widehat{\boldsymbol{\beta}}$ is defined as the solution of the generalized estimating equations

$$\sum_{i=1}^{K} \boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0},$$

where $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta} \equiv \{\partial \mu_{i,j} / \partial \beta_l\}_{j,l=1}^{n_i,k}$ is an $(n_i \times k)$ matrix of partial derivatives.

To compute the estimates of unknown parameters, we iterate between a modified Fisher scoring method for $\boldsymbol{\beta}$ and moment estimation of $(\phi, \boldsymbol{\vartheta})$:

1. Start with independent structure, i.e. $\boldsymbol{R}_i = \boldsymbol{I}_{n_i}$. Then matrix $\boldsymbol{V}_i = \phi \boldsymbol{A}_i$ is diagonal and initial $\boldsymbol{\beta}^{(0)}$ is estimated as if the observations were independent using GLM estimation techniques presented in Section 2.2.2.

2. Calculate Pearson residuals
$$r_{i,j}^{(P)} = \sqrt{w_{i,j}} \frac{(y_{i,j} - \widehat{\mu}_{i,j})}{\sqrt{V(\widehat{\mu}_{i,j})}}.$$

3. Compute moment estimates of ϕ and $\boldsymbol{\vartheta}$. The moment estimator of ϕ is defined as
$$\widehat{\phi} = \frac{1}{n-k} \sum_{i=1}^{K} \sum_{j=1}^{n_i} w_{i,j} \frac{(y_{i,j} - \widehat{\mu}_{i,j})^2}{V(\widehat{\mu}_{i,j})} = \frac{1}{n-k} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(r_{i,j}^{(P)}\right)^2,$$
where $n = \sum_{i=1}^{K} n_i$ is the total number of observations across all subjects and k is the number of regression parameters.
However, the moment estimator of $\boldsymbol{\vartheta}$ varies depending on the chosen correlation structure. Specific estimators are given in Hardin and Hilbe (2003, Section 3.2.1).
The working correlation matrix $\boldsymbol{R}_i$ can now be determined using the $\boldsymbol{\vartheta}$ value calculated and the assumed correlation structure.

4. Calculate estimate of the working covariance matrix $\boldsymbol{V}_i$

$$\widehat{\boldsymbol{V}_i} = \widehat{\phi}\widehat{\boldsymbol{A}}_i^{1/2}\boldsymbol{R}_i(\widehat{\boldsymbol{\vartheta}})\widehat{\boldsymbol{A}}_i^{1/2}.$$

5. Compute $\widehat{\boldsymbol{\beta}}^{(l+1)}$ by following iterative formula

$$\widehat{\boldsymbol{\beta}}^{(l+1)} = \widehat{\boldsymbol{\beta}}^{(l)} + \left[\sum_{i=1}^{K}\widehat{\boldsymbol{D}_i}^{\top}\widehat{\boldsymbol{V}_i}^{-1}\widehat{\boldsymbol{D}_i}\right]^{-1}\left[\sum_{i=1}^{K}\widehat{\boldsymbol{D}_i}^{\top}\widehat{\boldsymbol{V}_i}^{-1}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)\right].$$

6. Repeat steps 2–5 until convergence criterion is fulfilled.

3. Tweedie compound Poisson model

The compound Poisson distributions have been applied in various fields in which continuous data with exact zeros regularly arise. The presence of the discrete mass at zero makes them suitable for modelling aggregate distributions. For example, in actuarial applications the total claim amount usually has a continuous distribution on positive values, except of being zero when the claim does not occur. Our interest in these distributions is motivated by their applications to GLMs and GEE.

The main goal of this chapter is to demonstrate that the compound Poisson distribution can be re-parametrized to be a Tweedie model with $p \in (1, 2)$.

3.1 Compound Poisson model

Let $N, Z_1, Z_2, \ldots$ denote a sequence of random variables such that the following assumptions hold:

1. N follows a Poisson distribution with a parameter $\lambda w > 0$, where $w > 0$ is a known volume measure called the *exposure*. This distribution is denoted by $Po(\lambda w)$.
2. $Z_1, Z_2, \ldots$ are independent and follow a gamma distribution with a shape parameter $\alpha > 0$ and a rate parameter $\tau > 0$ denoted by $\Gamma(\alpha, \tau)$.
3. N and $(Z_1, Z_2, \ldots)$ are independent.

We define a random variable

$$Z = \mathbb{1}_{\{N>0\}} \times \sum_{i=1}^{N} Z_i,$$

where $\mathbb{1}_{\{\}}$ denotes an indicator function. Under these assumptions Z has a *compound Poisson distribution* denoted by $CPG(\lambda w, \alpha, \tau)$.

We are more interested in the ratio between Z and the exposure w, i.e. a random variable Y defined as

$$Y = \frac{Z}{w},$$

than in the random variable Z itself.

According to Jørgensen (1997, Chapter 4), the moment generating function $M_Y(t)$ of Y equals to

$$M_Y(t) = M_Z(t/w) = \exp\left\{\lambda w\left(\left(1 - \frac{t}{\tau w}\right)^{-\alpha} - 1\right)\right\} \quad \text{for } \tau w > t. \tag{3.1}$$

Hence $Y \sim CPG(\lambda w, \alpha, \tau w)$.

The distribution of Y can be re-parametrized in such a way that it takes form of a Tweedie model. To show that compound Poisson distribution corresponds to a Tweedie model with $1 < p < 2$, we simply compare the moment generating functions of both distributions.

Let $Y \sim \mathrm{ED}_p(\mu, \phi/w)$. Then the moment generating function of Y equals to (1.9) and it has the same form as the moment generating function (3.1). By matching term by term, we see that the Tweedie model is hence the compound Poisson distribution with the following parameter mapping

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \quad \alpha = \frac{2-p}{p-1}, \quad \tau = \frac{\mu^{1-p}}{\phi(p-1)}. \tag{3.2}$$

The requirement of gamma shape parameter α to be positive means that the representation of Y as a Poisson mixture of gamma random variables holds only for $p \in (1, 2)$, i.e $\alpha > 0$ implies $p \in (1, 2)$. Also, note that $\lambda > 0$ and $\tau > 0$ imply $\mu > 0$ and $\phi > 0$. The relations between parameters (3.2) provide a convenient mechanism of transferring from one parametrization to another and are exploited for evaluating the Tweedie density function.

A Tweedie model with $p \in (1, 2)$ is known as *Tweedie compound Poisson model.*

3.2 The joint distribution of Y and N

Following Jørgensen (1997)[Section 4.2.4], the joint density function of Y and N is given by the formula

$$f_{Y,N}(y, n) = f_{Y|N}(y \mid n) \times f_N(n) \quad \text{for } y > 0 \text{ and } n > 0,$$

in which the conditional distribution of Y given $N = n$ is $\Gamma(n\alpha, \tau w)$ for $n > 0$ and $N \sim Po(\lambda w)$. This gives the following joint density function

$$\begin{aligned} f_{Y,N}(y, n; \lambda, \alpha, \tau) &= \frac{(\tau w)^{n\alpha}}{\Gamma(n\alpha)} y^{n\alpha-1} \exp\{-\tau w y\} \times \frac{(\lambda w)^n}{n!} \exp\{-\lambda w\} \\ &= \frac{1}{n!\Gamma(n\alpha)y} \big(\lambda \tau^\alpha y^\alpha w^{\alpha+1}\big)^n \exp\Big\{w\big(-\tau y - \lambda\big)\Big\}. \end{aligned} \tag{3.3}$$

Now we re-parametrize the joint density (3.3) using the relations in (3.2). Hence, we obtain

$$f_{Y,N}(y, n; \mu, \phi/w, p) = \frac{1}{n!\Gamma(n\alpha)y} \Big(\frac{(w/\phi)^{\alpha+1} y^\alpha}{(p-1)^\alpha (2-p)} \Big)^n \exp\Big\{ \frac{w}{\phi} \Big(y \frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p} \Big) \Big\} \tag{3.4}$$

for $p \in (1, 2)$. Note that the term $\mu^{2-p}(\mu^{1-p})^\alpha$ is equal to 1. For the sake of simplicity, the parameter α remains expressed in other.

The joint distribution of Y and N also has a positive probability in zero

$$\mathsf{P}[Y = 0, N = 0] = \mathsf{P}[N = 0] = \exp\Big\{ -\frac{w}{\phi} \frac{\mu^{2-p}}{2-p} \Big\}.$$

We may also derive the marginal density of Y. For $y > 0$ the distribution is continuous and density of Y can be obtained from the joint density (3.4) by summing over positive values of N. Thus

$$\begin{aligned} f_Y(y;\mu,\phi/w,p) &= \sum_{n=1}^{\infty} f_{Y,N}(y,n;\mu,\phi/w,p) \\ &= \sum_{n=1}^{\infty} \frac{1}{n!\Gamma(n\alpha)y}\Big(\frac{(w/\phi)^{\alpha+1}y^{\alpha}}{(p-1)^{\alpha}(2-p)}\Big)^n \exp\Big\{\frac{w}{\phi}\Big(y\frac{\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p}\Big)\Big\}. \end{aligned}$$

Furthermore, the marginal distribution of Y has a positive probability in zero,

$$\mathsf{P}[Y=0] = \mathsf{P}[N=0] = \exp\Big\{-\frac{w}{\phi}\frac{\mu^{2-p}}{2-p}\Big\}.$$

Combining the previous results together gives the following form of probability density function of the Tweedie compound Poisson model

$$\begin{aligned} f_Y(y;\mu,\phi/w,p) &= a(y;\phi/w,p)\exp\Big\{\frac{w}{\phi}\Big(y\frac{\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p}\Big)\Big\} && (3.5) \\ &= a(y;\phi/w,p)\exp\Big\{\frac{w}{\phi}\big(y\theta-\kappa_p(\theta)\big)\Big\} && (3.6) \end{aligned}$$

for $y \geq 0$ and $p \in (1,2)$ where

$$a(y;\phi/w,p) = \begin{cases} \displaystyle\sum_{n=1}^{\infty} \frac{1}{n!\Gamma(n\alpha)y}\Big(\frac{(w/\phi)^{\alpha+1}y^{\alpha}}{(p-1)^{\alpha}(2-p)}\Big)^n & \text{for } y>0, \\ 1 & \text{for } y=0. \end{cases}$$

Note that the forms of marginal density (3.6) and (3.5) confirm that the compound Poisson distribution is an exponential dispersion model and, even more precisely, it is a Tweedie model with $p \in (1,2)$.

3.3 GLM and Tweedie compound Poisson model

Being EDM, Tweedie compound Poisson model fits the generalized linear model framework discussed in Section 2.2. Therefore, we can assume that the response variable Y follows a Tweedie compound Poisson model with mean μ, dispersion parameter ϕ and its variance function is of the form (1.4) for any $p \in (1,2)$. The exposure w is used as a weight in the GLMs.

Within R software, the libraries `statmod` and `tweedie` are needed to create regression models for pricing and claims reserving in Sections 4.6 and 5.6. To fit a GLM with a Tweedie compound Poisson model (Tweedie GLM) the generalized linear model family function `tweedie()` is used.

3.3.1 Parameter Estimation

In order to fit Tweedie GLM, only the aggregated variable Y is used without using any knowledge of the frequency. Thus, some information provided by the data is lost. As stated in Quijano Xacur and Garrido (2015), this issue can be solved by maximizing the joint log-likelihood of (Y, N) instead of the log-likelihood of Y. The joint log-likelihood approach includes the frequency so that information given by N is not lost.

We consider m independent pairs of observations of the form $(y_1, n_1), \ldots,$ (y_m, n_m), such that the i^{th} pair (y_i, n_i) is a realization of random vector (Y_i, N_i) with joint density (3.4). The joint log-likelihood function for the parameters $(\boldsymbol{\beta}, \phi, p)$ has the following form

$$\begin{aligned} \ell(\boldsymbol{\beta}, \phi, p; \boldsymbol{y}, \boldsymbol{n}) &= \sum_{i=1}^{m} \log f_{Y_i,N_i}(y_i, n_i; \mu_i, \phi/w_i, p) \\ &= \sum_{i=1}^{m} \Bigg[n_i \log\left(\frac{(w_i/\phi)^{\alpha+1} y_i^{\alpha}}{(p-1)^{\alpha}(2-p)} \right) - \log\big(n_i! \Gamma(n_i \alpha) y_i\big) \\ &\qquad + \frac{w_i}{\phi}\left(y_i \frac{\mu_i^{1-p}}{1-p} - \frac{\mu_i^{2-p}}{2-p} \right) \Bigg], \end{aligned} \tag{3.7}$$

where $\mu_i = g^{-1}(\mathbf{x}_i^{\top}\boldsymbol{\beta})$.

Case: p known

In case the true power index parameter $p \in (1, 2)$ is known, the estimates of unknown parameters may be found just from the marginal densities of Y_i (3.6). Vector of regression parameters $\boldsymbol{\beta}$ and dispersion parameter ϕ are estimated by the GLMs estimation techniques presented in Section 2.2.2.

Also when the logarithmic link function is used, the general maximum likelihood equations (2.2) simplify to

$$\sum_{i=1}^{m} \frac{w_i}{\mu_i^p \, \mu_i^{-2}} \mu_i^{-1}(y_i - \mu_i) x_{ij} = \sum_{i=1}^{m} w_i \frac{y_i - \mu_i}{\mu_i^{p-1}} x_{ij} = 0 \quad \text{for } j = 1, \ldots, k,$$

where μ_i is connected to $\boldsymbol{\beta}$ through

$$\mu_i = \exp\left\{ \sum_{j=1}^{k} x_{ij}\beta_j \right\} \quad \text{for } i = 1, \ldots, m.$$

Case: p unknown

Now we examine the case when the power index parameter p is unknown. Based on Dunn and Smyth (2005), we use the profile likelihood approach to estimate the power index and dispersion parameters. Note that we can profile out the mean parameters as they are obtained for a given value of power index parameter.

At first, we need to specify a grid of possible values P from the interval $(1, 2)$ that p can take. Further, we exploit the fact that the estimation of $\boldsymbol{\beta}$ depends only on p, i.e. the dispersion parameter ϕ has no influence on its estimation.

Given a fixed $p \in P$, maximum likelihood estimate $\widetilde{\boldsymbol{\beta}}_p$ is calculated as in the previous part. Then conditional on this p and the corresponding $\widetilde{\boldsymbol{\beta}}_p$, the maximum likelihood estimator $\widetilde{\phi}_p$ solves the equation

$$\frac{\partial \ell(\widetilde{\boldsymbol{\beta}}_p, \phi, p)}{\partial \phi} = -\frac{(\alpha+1)}{\phi}\sum_{i=1}^{m} n_i - \frac{1}{\phi^2}\sum_{i=1}^{m} w_i\left(y_i \frac{\widetilde{\mu}_i^{1-p}}{1-p} - \frac{\widetilde{\mu}_i^{2-p}}{2-p}\right) = 0. \tag{3.8}$$

Hence the maximum likelihood estimator of ϕ is given by

$$\widetilde{\phi}_p = \frac{-\sum_{i=1}^{m} w_i\left(y_i \frac{\widetilde{\mu}_i^{1-p}}{1-p} - \frac{\widetilde{\mu}_i^{2-p}}{2-p}\right)}{(\alpha+1)\sum_{i=1}^{m} n_i}. \tag{3.9}$$

In such way, we have determined the corresponding $(\widetilde{\boldsymbol{\beta}}_p, \widetilde{\phi}_p)$ for each fixed $p \in P$. Then the maximum likelihood estimate of p is obtained by maximizing the profile log-likelihood with respect to the grid of pre-specified values

$$\widehat{p} = \underset{p \in P}{\mathsf{arg\,max}}\{\ell(\widetilde{\boldsymbol{\beta}}_p, \widetilde{\phi}_p, p)\},$$

where

$$\begin{aligned}
\ell(\widetilde{\boldsymbol{\beta}}_p, \widetilde{\phi}_p, p) &= \sum_{i=1}^{m}\left[n_i \log(w_i/\widetilde{\phi}_p)^{(\alpha+1)} + n_i \log\left(\frac{y_i^{\alpha}}{(p-1)^{\alpha}(2-p)}\right)\right. \\
&\qquad \left. - \log\big(n_i!\Gamma(n_i\alpha)y_i\big) - n_i(\alpha+1)\right] \\
&= \sum_{i=1}^{m}\left[(\alpha+1)n_i\big(\log(w_i/\widetilde{\phi}_p) - 1\big) - \log\big(n_i!\Gamma(n_i\alpha)y_i\big)\right. \\
&\qquad \left. + n_i \log\left(\frac{y_i^{\alpha}}{(p-1)^{\alpha}(2-p)}\right)\right]
\end{aligned}$$

is the profile log-likelihood evaluated at $\widetilde{\boldsymbol{\beta}}_p, \widetilde{\phi}_p$ and p, using the Tweedie compound Poisson density evaluation methods provided by Dunn and Smyth (2001). The corresponding estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ are obtained according to the previous part by using $\hat{p}$. Note that often in practice, p is assumed to be known.

In R, the maximum likelihood estimate of power index parameter p is found by the `tweedie.profile` function.

3.4 GEE and Tweedie compound Poisson model

One of the components which has to be specified within the GEE framework is the variance function. Assuming the Tweedie compound Poisson model, the variance function has the following form

$$V(\mu_{i,j}) = \mu_{i,j}^{p} \qquad \text{for all } i, j \text{ and } p \in (1,2).$$

3.4.1 Parameter Estimation

Following the iterative algorithm in Section 2.3.2, the GLMs estimation technique is used to estimate the initial $\boldsymbol{\beta}$. Therefore, the power index parameter p and initial regression parameters $\boldsymbol{\beta}$ are estimated according to the previous section.

Also, some terms in the iterative algorithm can be specified more precisely for all $i = 1, \ldots, K$:

- Pearson residuals and matrix $\boldsymbol{A}_i$, as the form of variance function is known
$$V(\boldsymbol{\mu}_i) = \boldsymbol{\mu}_i^p \quad \text{for any } p \in (1, 2).$$
- Matrix
$$\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \overset{\substack{\text{chain}\\ \text{rule}}}{=} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} = \frac{1}{g'(\boldsymbol{\mu}_i)} \times \boldsymbol{X}_i$$
has the following form when the logarithm link function is used
$$\boldsymbol{D}_i = \boldsymbol{\mu}_i \times \boldsymbol{X}_i.$$
Note that this is not a matrix multiplication.

In R, the GEE estimates cannot be obtained by `gee` function from `gee` library because the function `gee` does not allow using the Tweedie compound Poisson model. Therefore, it was necessary to write a code so this type of modelling could be implemented. The GEE parameter estimation for Tweedie generalized estimating equation model (Tweedie GEE) is based on the code from Swan (2006).

4. Pricing using Tweedie compound Poisson model

The main goal of this chapter is to show how Tweedie compound Poisson model can be used in non-life insurance pricing. The chapter is structured as follows.

The first section introduces some important definitions concerning pricing in non-life insurance. A brief overview of well-known methods used for setting the premium is given in the consecutive section. Section 4.3 recalls the definition of Tweedie compound Poisson model. The definition is supplemented by the interpretation of random variables within context of modelling a pure premium. Following the distributional assumption, Sections 4.4 and 4.5 present the applications of GLMs and GEE for pure premium estimation. The general frameworks from Chapter 2 are modified to fit the collected insurance policies data. A suitable form of linear predictor and a type of link function are suggested. Moreover, within the GEE framework, the appropriate choice of correlation structure is discussed. The last section illustrates an application and a performance of proposed Tweedie GLM and GEE models on a real dataset. The results of the individual models are summarized and consequently used for their comparison.

4.1 Pricing terminology

One of the most important problems in every insurance company is to set the premiums for the policyholders. Following Ohlsson and Johansson (2010, Chapter 1), we introduce some key insurance terms that are used within non-life insurance pricing context.

A *tariff* is a formula by which the premium for any policyholder can be computed. The underlying work the actuaries perform to obtain the tariff is known as a *tariff analysis*. It is mainly based on the insurance company's historical data on policies and claims.

The premium charged to the policyholders can be broken down into a so-called *pure premium* and other components such as claims handling costs, administration costs, salaries, profit margin etc. On top of this, the insurance company may adjust the price for individual customers for various reasons, e.g. extra high margins for young motorcycle drivers who represent a larger risk to the insurance company. In this thesis, we will only consider the part of premium which is directly connected to the losses.

The *duration of a policy* is the amount of time a policy is in force and it is usually measured in years. The *pure premium* is the total claim amount divided by the duration, i.e. the average cost per time unit. The pure premium can be expressed as a product of *claim frequency* (the average number of claims per time unit) and *claim severity* (the average cost per claim).

The pure premium varies between policies and can be estimated based on several variables called *rating factors*. The rating factors are usually related to the properties of policyholders or insured objects, e.g. gender, age, type of car etc. Each combination of values of these variables defines a homogeneous risk group called a *tariff class*. Insurance is justified by the law of large numbers and for this reason it is desirable to have tariff classes of a sufficient volume. Therefore, the continuous rating factors are often categorized into intervals and all values within an interval are treated as identical. We follow such practice in this thesis. Generally, the insurance portfolio is divided into tariff classes and all policyholders within the same tariff class are charged the same premium.

4.2 Overview of pricing methods

Determining the right level of premium is not an easy task. In order to be solvent, an insurance company needs to charge enough money in premiums to be able to face its liabilities. On the contrary, when the policyholders think that they are being overcharged they probably leave. Also a ”wrong” rating system may attract bad risks. A number of different methods for setting appropriate premiums has been invented.

Bailey-Simon method and consecutive more robust Bailey-Jung method also known as the *method of total marginal sums* are considered to be simple tariffcation methods because they are not directly based on a stochastic model. In the 1990's, generalized linear models were introduced by British actuaries as a tool for tariff analysis and since then they have become the standard approach in many insurance companies. Traditional GLM tariff analysis is based on two separate GLM models for claim frequency and claim severity. Then the estimates of pure premiums are found by simply multiplying the results.

Direct modelling of pure premium is problematic since a typical pure premium distribution will consist of a large spike at zero (where policies have not had claims) and then a wide range of amounts (where policies have had claims). Jørgensen and Souza (1994) used a Tweedie GLM to model the pure premium directly. Due to its ability to simultaneously model the zeros and the continuous positive outcomes a Tweedie GLM has been a widely used method in tariff analysis.

4.3 Tweedie compound Poisson model

Following Jørgensen and Souza (1994), to model pure premium using the Tweedie compound Poisson model the following assumptions for all insurance policies $i = 1, \dots, n$ should hold:

1. The number of reported claims N_i on policy i is Poisson distributed with a parameter $\lambda_i w_i > 0$.
2. The individual claim amounts $Z_i^{(l)}$ are independent gamma distributed with a shape parameter $\alpha > 0$ and a policy-specific parameter $\tau_i > 0$.
3. Random variables N_i and $Z_i^{(l)}$ are independent for all l.

Then the total claim amount Z_i and the pure premium Y_i for the ith policy over a given time period are respectively defined as

$$Z_i = \mathbb{1}_{\{N_i > 0\}} \times \sum_{l=1}^{N_i} Z_i^{(l)} \quad \text{and} \quad Y_i = \frac{Z_i}{w_i},$$

where w_i denotes the duration of ith policy.

According to Section 3.1, Y_i follows a Tweedie compound Poisson model with parameters μ_i and ϕ, i.e.

$$Y_i \sim \mathrm{ED}_p(\mu_i, \phi/w_i) \quad \text{for } 1 < p < 2.$$

The mean and the variance of Y_i are then given by

$$\begin{aligned} \mathsf{E}\, Y_i &= \mu_i, \\ \operatorname{var} Y_i &= \frac{\phi}{w_i}\mu_i^p. \end{aligned}$$

4.4 GLM framework in pricing

We assume that all Y_i are independent and $Y_i \sim \mathrm{ED}_p(\mu_i, \phi/w_i)$ for $p \in (1,2)$. Furthermore, we consider M rating factors, each one divided into categories, where m_l denotes the number of categories for rating factor l and i_l denotes the category of rating factor l.

For the policy i with rating factors pertaining to the categories $i_1, \ldots, i_M$, based on Ohlsson and Johansson (2010, Chapter 1), we assume a multiplicative structure for μ_i

$$\mu_i \equiv \mu_{i_1,\ldots,i_M} = \gamma_0 \prod_{l=1}^{M} \gamma_{l,i_l},$$

where the parameter γ_0 is called a *base value* and the other parameters $\{\gamma_{l,i_l}, i_l = 1, \ldots, m_l\}$ are known as the *price relativities* for rating factor l. We specify a *base tariff class* $\{i_1 = b_1, \ldots, i_M = b_M\}$ and set $\{\gamma_{l,b_l} = 1, l = 1, \ldots, M\}$, so that the estimation problem is well-defined. The base value γ_0 can now be interpreted as the mean in the base tariff class and the price relativities measure the relative difference in relation to the base class.

It is now straightforward to choose the logarithmic link function. Then we have

$$\log \mu_i = \log \mu_{i_1,\ldots,i_M} = \eta_i = \log \gamma_0 + \sum_{l=1}^{M} \log \gamma_{l,i_l}. \tag{4.1}$$

The linear predictor η_i can be expressed in vector notation

$$\eta_i = \boldsymbol{x}_i \boldsymbol{\beta}$$

where a design vector

$$\boldsymbol{x}_i \equiv \boldsymbol{x}_{i(i_1,\ldots,i_M)} = (1, \delta_{i(1,1)}, \ldots, \delta_{i(1,m_1)}, \ldots, \delta_{i(M,1)}, \ldots, \delta_{i(M,m_M)})$$

is for all l and i_l defined by

$$\delta_{i(l,i_l)} = \begin{cases} 1 & \text{if } \log \gamma_{l,i_l} \text{ is included in } \log \mu_i, \\ 0 & \text{ortherwise.} \end{cases}$$

The vector of corresponding unknown regression parameters has the following form

$$\boldsymbol{\beta} = (\log \gamma_0, \log \gamma_{1,1}, \ldots, \log \gamma_{1,m_1}, \ldots, \log \gamma_{M,1}, \ldots, \log \gamma_{M,m_M})^\top,$$

where $\log \gamma_{l,b_l} = 0$ for $l = 1, \ldots, M$. Then the number of unknown parameters equals to

$$k = \sum_{l=1}^{M} m_l - M + 1.$$

The parameter $\boldsymbol{\beta}$ is estimated by the maximum likelihood method, see Section 3.3.1.

4.5 GEE framework in pricing

Within GLMs framework, we assume that the average costs per time unit for different policies are independent variables, i.e. we assume that Y_i are independent for all $i \in \{1, \ldots, n\}$. If this assumption is not fulfilled, then the GLMs may provide incorrect estimates of pure premiums.

The portfolio of insurance policies can be divided into tariff classes. A tariff class consists of policies which have the same values of all rating factors. Therefore, it is natural to suspect that the observations within the same tariff class might be dependent. Hence, we can handle the insurance portfolio as a special type of clustered data in which the tariff classes represent the subjects.

The GEE approach enables modelling these dependencies via a working correlation matrix. Some of the most commonly used correlation structures were presented in Section 2.3.1. The exchangeable correlation structure might be the most suitable choice since the observations are simply clustered based on the rating factors and time ordering within the tariff class lacks sense. However if the number of clusters is small, then the independent structure could be a good choice as well. The advantage of GEE approach is that the estimates of pure premiums are still valid even if the correlation structure is not chosen correctly.

Furthermore, in the proposed Tweedie GEE model for the pure premium, the additive rating mean structure (4.1) is assumed.

4.6 Practical application

In the following part of this chapter, we illustrate a practical application of presented pricing theory concerning Tweedie GLM and GEE models on a real dataset. The main aim of the following analyses is to determine how the pure premium varies with the number of rating factors.

Note that there is no software available that fits a Tweedie GEE model. The parameter estimation for this model is based on the code from Swan (2006). A selected R code is attached in Appendix A.

4.6.1 Dataset

We consider authentic insurance data from the former Swedish insurance company WASA which concern partial casco insurance for motorcycles during years 1994-1998 studied by Ohlsson and Johansson (2010, Section 2.4). Partial casco covers theft and some other causes of loss, like fire. The analysed dataset consists of 64 548 insurance policies and for each policy the information about the following variables (in Swedish acronyms) is available:

agarald - The owners age.

kon - The owners gender.

zon - Geographic zone numbered from 1 to 7 in a standard classification of all Swedish parishes.

mcklass - MC class, a classification by the so called EV ratio, defined as (Engine power in kW $\cdot$ 100) / (Vehicle weight in kg + 75), rounded to the nearest lower integer. The 75 kg represent the average driver weight. The EV ratios are divided into seven categories.

fordald - Vehicle age.

duration - The number of policy years.

antskad - The number of claims.

skadkost - The claim cost in SEK.

Note that we assume that the owners are also the drivers.

It is reasonable to have a closer look at selected dataset before we create any model for the pure premium. Swedish Transport Agency learner's permit can be issued at the age of 16. Therefore, we omit all drivers younger than 16 from the original dataset. We also assume that drivers at the age 90 and more are not capable of driving so we eliminate them as well. Furthermore, we exclude zero durations from the analysis. The adjusted dataset consists of 62 435 insurance policies over the years 1994-1998 in which 61 769 (98.9 %) policies had no claim. Thus, the Tweedie approach of modelling zeros and positive observations together appears to be perfectly adequate here.

For each policy, the data contain the exact age of the owners and the vehicles. Following the approach mentioned in Section 4.1, we divide the values of these variables into classes. We choose to have just five age classes and three vehicle age classes. Our decision was based on age requirements for different types of motorcycles and various studies provided by Swedish Transport Agency. Obviously, a large number of alternative groupings is possible. We consider all categorical variables as rating factors and their detailed description is given in Table 4.1. Variable *duration* is used as weight.

The drivers can be divided into 1 072 tariff classes on the basis of rating factors. Note that $2 \times 3 \times 5 \times 7 \times 7 = 1\ 470$ is a total number of tariff classes but we do not have the observations for every tariff class. For each rating factor, it is customary to choose the class with the highest duration as the base class. Thus, the base tariff class is $(\mathrm{M}, 3, 3, 4, 3)$ and it corresponds to middle age male drivers living in small towns and countrysides who own older motorcycles with middle MC class.

4.6.2 Application of GLM to pricing

An appropriate value of the power index parameter p needs to be found first in order to fit a Tweedie GLM. We use the profile log-likelihood function to estimate p, see Section 3.3.1. The maximum likelihood estimate is $\hat{p} = 1.5673$. Figure 4.1

Rating factor	Class	Class description
Owners gender	M	Male
	K	Female
Vehicle age	1	0 − 3 years
	2	4 − 10 years
	3	11+ years
Owners age	1	16 − 21 years
	2	22 − 35 years
	3	36 − 50 years
	4	51 − 65 years
	5	66+ years
Geographic zone	1	Central and semi-central parts of Sweden's three largest cities
	2	Suburbs and middle-sized towns
	3	Lesser towns, except those in 5 or 7
	4	Small towns and countrysides, except 5 − 7
	5	Northern towns
	6	Northern countrysides
	7	Gotland (Sweden's largest island)
MC class	1	EV ratio − 5
	2	EV ratio 6 − 8
	3	EV ratio 9 − 12
	4	EV ratio 13 − 15
	5	EV ratio 16 − 19
	6	EV ratio 20 − 24
	7	EV ratio 25+

Table 4.1: Rating factors in motorcycle insurance.

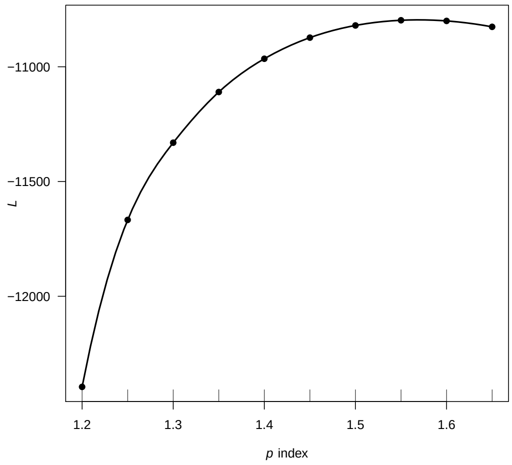

Figure 4.1: Profile log-likelihood plot for p.

shows the profile log-likelihood plot where the points represent the profile log-likelihood computed at each value in the pre-specified grid P.

Now we can fit the model in terms of equation (4.1) using all five rating factors. The estimate of dispersion parameter ϕ is equal to 2 454. The estimated regression coefficients along with the estimated price relativities are listed in Table 5.3.

We want to investigate whether all rating factors are important and hence should be included in the tariff. We test the significance of each of rating factors in turn against the full model. For such sub-model testing, we use a likelihood ratio test with a significance level of 5 %. This test belongs among the basic tests within the GLM framework and its detailed description is presented e.g. in Ohlsson and Johansson (2010, Section 3.1). Based on the individual p-values, all five rating factors are significant and therefore we keep the initial full Tweedie GLM model. Note that we do not consider any interactions between the rating factors in the analyses since the interactions produce a model which is too complex for interpretation and practical use in setting insurance tariffs.

Model diagnostics

The numerous diagnostic plots can be drawn to check whether the model adequately fits the data. A general tool used in diagnostic analysis are residuals. The most commonly used residuals for GLMs are the Pearson residuals.

Before we look at the various diagnostic plots, it is wise to recall two important facts about the analysed dataset. Firstly, there are 986 out of 1072 tariff classes with more than one observation available and secondly 98.9 % of policies had no claim during 1994-1998. These circumstances will have a major impact on the following plots visualisation.

One of the most important components of a GLM is that the correct distribution is chosen for the response variable. The Pearson residuals were used to asses how well the distribution fits the data. They have an approximate normal distribution $\mathcal{N}(0, \phi)$, provided that the correct distribution is used. To check whether the Tweedie compound Poisson model is suitable for the analysed data, a Q-Q plot and a histogram of Pearson residuals are produced. Both graphs are shown in Figure 4.2. The larger values which do deviate from normality line indicate that the model does not fit extreme values very well. These graphs suggest that the Tweedie GLM model with the estimated power index parameter $\hat{p} = 1.5673$ and the estimated dispersion parameter $\hat{\phi} = 2\ 454$ is not appropriate.

Residual plots are shown in Figure 4.3. The first graph illustrates the Pearson residuals. The residuals should be symmetrically located around zero and any pattern observed indicates problems with the fitted model. However, we see that the magnitude of the positive residuals is much larger than the magnitude of the negative ones. This is caused by the fact that the model does not predict extreme values very accurately. The Pearson residuals versus the linear predictor are displayed in the second graph. If a model accurately represents the data, all points should be uniformly spread. As the plot does not show the uniformity, Tweedie GLM model provides a poor fit for this data.

Another diagnostic plot that demonstrates the overall fit of the model is given in Figure 4.4. It illustrates the observed values with respect to the fitted values. It is important to realize that we have one fitted value for the observations that

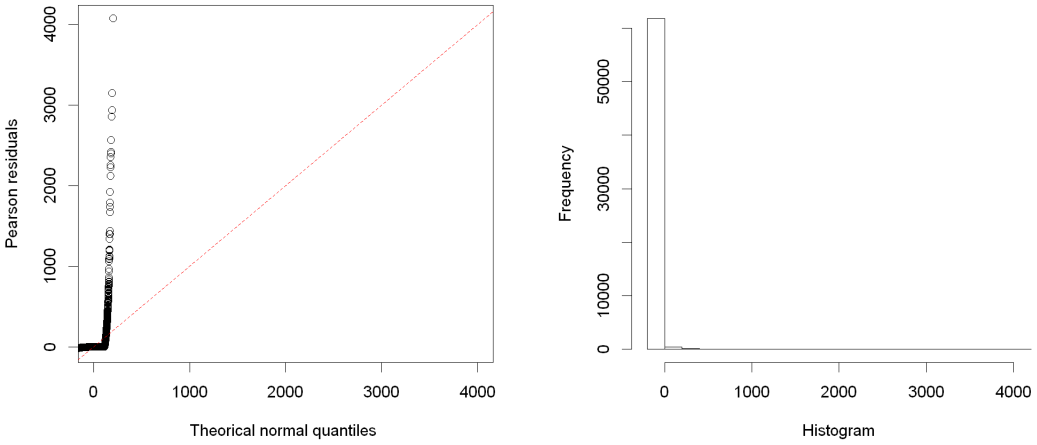

Figure 4.2: Q-Q plot and histogram of Pearson residuals.

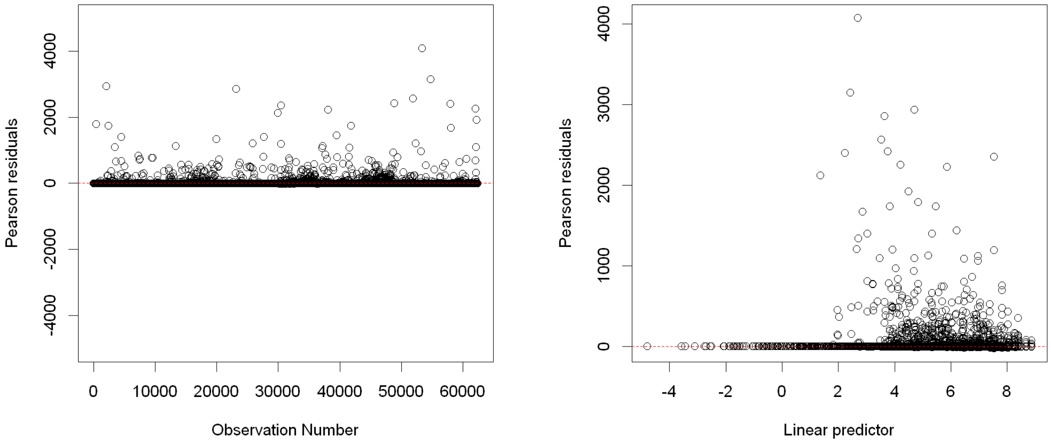

Figure 4.3: Pearson residuals plots for Tweedie GLM model.

belong to the same tariff class. We see that for the large observed values the corresponding fitted ones are very small. This is caused by the large number of zeros in the data.

4.6.3 Application of GEE to pricing

The results obtained from the application of Tweedie GEE model with exchengeable correlation structure to the same dataset are summarized in this part. The reasons for choosing this structure are explained in Section 4.5.

A corresponding Tweedie GLM is fitted in the initial stage of the Tweedie GEE estimation process. Therefore, the estimate of the power index parameter p is the same for both models. The maximum likelihood estimate is $\hat{p} = 1.5673$

A Tweedie GEE model with exchangeable correlation structure is fitted in the terms of equation (4.1). The estimate of parameter ϑ for this structure equals to

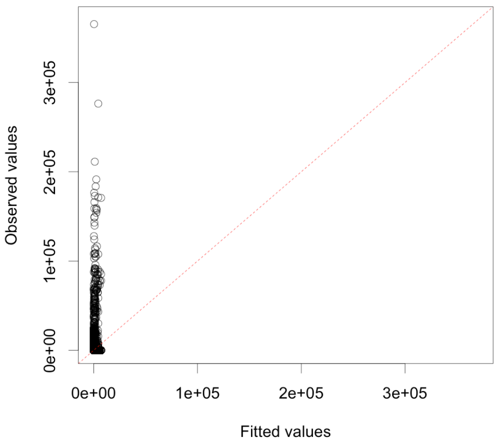

Figure 4.4: Observed vs. fitted values of Tweedie GLM model.

0.2927. The value of $\hat{\vartheta}$ indicates that there exists a significant positive correlation between the average costs per annum over the years 1994-1998 within the same tariff class. The estimate of dispersion parameter ϕ is $\hat{\phi} = 18\ 249$. The estimated regression parameters together with the estimated price relativities are given in Table 4.2.

Model diagnostics

The residual plots and graph showing the fitted values plotted against the observed are very similar to the Figures 4.3 and 4.4. Therefore, the same justifications also apply for the Tweedie GEE model.

4.6.4 Comparison of results

We fitted two different models for the pure premium to the same dataset, namely the Tweedie GLM model and the Tweedie GEE model with exchangeable correlation structure. Both models were examined in detail to determine if any of these models are adequate for pure premium.

Based on the diagnostic checks performed, Tweedie GLM model does not provide a good representation of data and the distributional assumptions are not satisfied. Tweedie GEE model does not perform satisfactory either. Both models provide a poor fit to the analysed data. Overall, the models can be considered as the inadequate models for the expected pure premium. But again, we have to keep in mind the structure of underlying dataset. Almost all of the observations are zeros. Therefore, we can not really decide whether these models are completely useless.

Even thought, the models did not perform well, we still interpret and compare the estimates gained from the fitted models. We only interpret the estimated price relativities for Tweedie GLM. The interpretation for Tweedie GEE is the same. The pure premium is estimated to be 127 SEK per year in the base tariff class, i.e. $(\mathrm{M}, 3, 3, 4, 3)$ which corresponds to middle age male drivers living in small towns and countrysides who own older motorcycles with middle MC class.

To get a prediction of pure premium for other classes, the base value should be multiplied by the price relativities given in the Table 4.2 for each rating factor category. For example, the pure premium for a young woman living in the largest Swedish city having a new motorcycle with MC class 7 is estimated to be:

$$\hat{\mu} = 127.4342 \times 0.4484 \times 14.8513 \times 1.2156 \times 1.6206 \times 0.2895 = 484$$

SEK per year. By comparing the estimates of the price relativities between two considered models, some significant differences can be observed. These differences are probably cause by the detected dependence within the tariff classes.

	Tweedie GLM		Tweedie GEE	
Variable	Estimate	Relativity	Estimate	Relativity
Base tariff class	4.8476	127.4342	7.1724	1302.9680
Female	−0.8021	0.4484	−0.6384	0.5281
Vehicle age (1)	2.6981	14.8513	2.6088	13.5827
Vehicle age (2)	1.3559	3.8802	1.0294	2.7994
Owners age (1)	0.1952	1.2156	−0.3493	0.7052
Owners age (2)	0.3627	1.4372	0.1763	1.1928
Owners age (4)	−1.0582	0.3471	−1.4098	0.2442
Owners age (5)	−2.7435	0.0643	−5.7652	0.0031
Geographic zone (1)	0.4828	1.6206	−0.3658	0.6936
Geographic zone (2)	0.2827	1.3267	−0.8995	0.4068
Geographic zone (3)	−0.6084	0.5442	−1.0697	0.3431
Geographic zone (5)	−2.3514	0.0952	−3.2088	0.0404
Geographic zone (6)	−1.8002	0.1653	−3.5688	0.0282
Geographic zone (7)	−5.5416	0.0039	−7.2148	0.0007
MC class (1)	−1.3516	0.2588	−2.0772	0.1253
MC class (2)	−0.1318	0.8765	−0.5797	0.5601
MC class (4)	−1.0081	0.3649	−1.7279	0.1777
MC class (5)	−0.1335	0.8750	−0.8376	0.4327
MC class (6)	0.4827	1.6205	−0.7560	0.4695
MC class (7)	−1.2397	0.2895	−4.2099	0.0148

Table 4.2: Estimated regression parameters and price relativities for Tweedie GLM and GEE models.

5. Claims reserving using Tweedie compound Poisson model

The previous chapter described the application of Tweedie compound Poisson model in non-life insurance pricing. The main goal of this chapter is to show how Tweedie compound Poisson model can be used in claims reserving. The structure of this chapter is identical to the previous one.

The classical claims reserving terminology and notation are introduced in the first section. The following section gives a brief overview of popular reserving methods. Section 5.3 recalls the definition of Tweedie compound Poisson model. The definition is supplemented by the interpretation of random variables within claims reserving context. Sections 5.4 and 5.5 present the application of GLMs and GEE for estimation of claims reserves. The general frameworks from Chapter 2 are modified to fit the data structure of claims. A suitable form of linear predictor and a type of link function are proposed. Moreover, within the GEE framework, the appropriate choice of correlation structure is discussed. Last section illustrates an application of proposed Tweedie GLM and GEE models on a real dataset. It focuses especially on performance of these models and also on their comparison.

5.1 Claims reserving terminology and notation

Claims reserving is another classical problem in non-life insurance. The main issue is that a typical non-life insurance claim cannot be settled immediately at its occurrence. This is often caused by reporting and settlement delays or by re-opening of an already closed claim due to unexpected new developments. The history of a typical non-life insurance claim can be illustrated as in Figure 5.1 taken from Wüthrich and Merz (2008, Section 1).

As a consequence, the insurance company needs to build *claims reserves*, so that it is able to fulfil future payments arising from claims that have occurred in the past and are only settled in the future. Setting an appropriate amount of claims reserves is called *claims reserving* and is one of the most fundamental actuarial tasks in every insurance company.

There are two main categories of claims reserves. The first one is a reserve on

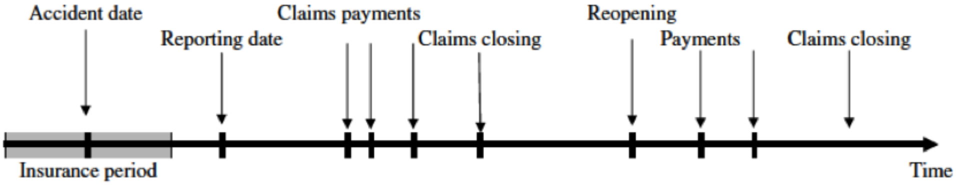

Figure 5.1: Typical time line of a non-life insurance claim.

Accident	Development year j				
year i	1	2	$\dots$	$n-1$	n
1	$Y_{1,1}$	$Y_{1,2}$	$\dots$	$Y_{1,n-1}$	$Y_{1,n}$
2	$Y_{2,1}$	$Y_{2,2}$	$\dots$	$Y_{2,n-1}$	$Y_{2,n}$
			$\ddots$		
$\vdots$	$\vdots$	$\vdots$	$Y_{i,n+1-i}$	$\vdots$	$\vdots$
				$\ddots$	
$n-1$	$Y_{n-1,1}$	$Y_{n-1,2}$			
n	$Y_{n,1}$	$Y_{n,2}$	$\dots$		$Y_{n,n}$

Table 5.1: Run-off triangle for normalized incremental claims payments $Y_{i,j}$.

claims that have incurred but have not been reported, called *IBNR reserve*, and the second is a reserve on claims that have been reported but have not yet been settled, called *RBNS reserve*.

The historical data used in claims reserving are typically structured in a so-called *claims development triangle* or a *run-off triangle*, see Table 5.1. In this triangle, the row i denotes the accident year and the column j indicates the delay, also assumed to be measured in years. Each diagonal corresponds to one calendar year. The current year is n and it represents the most recent accident as well as development year.

Let a random variable $Z_{i,j}$ denote all payments done in the accounting year $i+j$ for claims occurred in the year i. We refer to $Z_{i,j}$ as the *incremental payments* for claims with origin in the accident year i during development period j. The *normalized incremental claims payments* are then given by

$$Y_{i,j} = \frac{Z_{i,j}}{w_i},$$

where w_i denotes a known exposure for the accident year i. Note that there exist several different possibilities from which we can choose the appropriate w_i, e.g. total number of policies in the i^{th} accident year.

At time n, the following payment information is available

$$D_n^U = \{Y_{i,j}; i+j \leq n+1 \text{ and } i,j = 1,\dots,n\},$$

which corresponds to an *upper triangle* in Table 5.1 and the outstanding payments, i.e. a *lower triangle*

$$D_n^L = \{Y_{i,j}; i+j > n+1 \text{ and } i,j = 1,\dots,n\}$$

needs to be predicted in order to estimate the outstanding claims reserves

$$R_i = \sum_{j=n+2-i}^{n} Y_{i,j} = \frac{1}{w_i} \sum_{j=n+2-i}^{n} Z_{i,j} \quad \text{for } i = 2,\dots,n. \tag{5.1}$$

The total claims reserve at time n is then given by

$$R = \sum_{i=2}^{n} R_i. \tag{5.2}$$

5.2 Overview of claims reserving methods

There is a wide range of different methods and models used for setting appropriate claims reserves, but only practical experience will tell which one should be applied in any particular situation.

The chain-ladder and Bornhuetter-Ferguson methods belong to the deterministic claims reserving methods. Their simplicity and accurate results still make them the most commonly used techniques in practice.

In the last 30 years, there has been an increasing interest in stochastic models underlying forenamed methods. The practicality of these models lies in them being able to provide more information which may be useful in the claims reserving process and in the overall management of insurance company.

The basics of stochastic models were introduced by Mack in 1993. Mack distributional-free chain-ladder model is probably the most famous stochastic claims reserving model. It is straightforward from a stochastic point of view and it is easy to apply. Renshaw in 1995 was the first who implemented the standard generalized linear model techniques in claims reserving context. Since then these models have played an important role therein. Among popular claims reserving GLMs belong the overdispersed Poisson, gamma and Tweedie compound Poisson models. An overview on stochastic reserving models used in non-life insurance can be found in e.g. England and Verrall (2002) or Wüthrich and Merz (2008).

Almost all of the proposed approaches require the independence of incremental claims amounts in different years. However, this assumption does not often hold in practice and we need methods which enable to model the possible dependencies among the incremental claims payments. Thus, the claims reserving GEE models were presented.

5.3 Tweedie compound Poisson model

Following Wüthrich (2003, Section 2), to estimate claims reserves using the Tweedie compound Poisson model the following assumptions for all $i, j \in \{1, \ldots, n\}$ should hold:

1. The number of claims payments $N_{i,j}$ is Poisson distributed with a parameter $\lambda_{i,j} w_i > 0$.
2. The individual claims payments $Z_{i,j}^{(l)}$ are independent gamma distributed with a common shape parameter $\alpha > 0$ and a cell-specific parameter $\tau_{i,j} > 0$.
3. Random variables $N_{i,j}$ and $Z_{i,j}^{(l)}$ are independent for all l.

Then the incremental claims payments $Z_{i,j}$ and the normalized incremental claims payments $Y_{i,j}$ in cell (i, j) are defined respectively by

$$Z_{i,j} = \mathbb{1}_{\{N_{i,j}>0\}} \times \sum_{l=1}^{N_{i,j}} Z_{i,j}^{(l)} \quad \text{and} \quad Y_{i,j} = \frac{Z_{i,j}}{w_i}. \tag{5.3}$$

According to Section 3.1, normalized incremental claims payments $Y_{i,j}$ follows a Tweedie compound Poisson model with parameters $\mu_{i,j}$ and ϕ, i.e

$$Y_{i,j} \sim \mathrm{ED}_p(\mu_{i,j}, \phi/w_i) \quad \text{for } 1 < p < 2.$$

The mean and the variance of $Y_{i,j}$ are then given by

$$\begin{aligned} \mathsf{E}\, Y_{i,j} &= \mu_{i,j}, \\ \operatorname{var} Y_{i,j} &= \frac{\phi}{w_i} \mu_{i,j}^p. \end{aligned}$$

5.4 GLM framework in claims reserving

We assume that all $Y_{i,j}$ are independent and $Y_{i,j} \sim \mathrm{ED}_p(\mu_{i,j}, \phi/w_i)$ for $p \in (1,2)$. As stated in Wüthrich (2003, Section 4), an additional structure for the means of $Y_{i,j}$ needs to be specified. We assume a multiplicative structure

$$\mu_{i,j} = \gamma_i \nu_j, \tag{5.4}$$

for all i, j with the constrain $\gamma_1 = 1$, so that the estimation problem is well-defined. Parameter γ_i can be interpreted as the expected ultimate claim in accident year i and ν_j is the proportion paid in development year j.

The multiplicative structure (5.4) reduces the number of unknown parameters from $n \cdot n$ to $2n-1$ and defines exactly how the information from the upper triangle is transferred to the lower triangle.

It is now straightforward to choose the logarithmic link function. Then we have

$$\log \mu_{i,j} = \eta_{i,j} = \log \gamma_i + \log \nu_j. \tag{5.5}$$

The linear predictor $\eta_{i,j}$ can be expressed in vector notation

$$\eta_{i,j} = \boldsymbol{x}_{i,j} \boldsymbol{\beta},$$

where a design vector

$$\boldsymbol{x}_{i,j} = (0, \delta_{2,i}, \ldots, \delta_{n,i}, \delta_{1,j}, \ldots, \delta_{n,j})$$

is defined by Kronecker's deltas $\delta_{i,j}$. The vector of corresponding unknown regression parameters has the following form

$$\boldsymbol{\beta} = (0, \log \gamma_2, \ldots, \log \gamma_n, \log \nu_1, \ldots, \log \nu_n)^\top.$$

The parameter $\boldsymbol{\beta}$ is estimated by the maximum likelihood method, see Section 3.3.1. Having the estimates of regression parameters, the predicted normalized incremental payments for the lower triangle are obtained as follows

$$\hat{Y}_{i,j} = \hat{\mu}_{i,j} = \exp(\hat{\eta}_{i,j}) = \hat{\gamma}_i \hat{\nu}_j.$$

Accident year i and overall reserve estimates can then be found by summing these values according to (5.1) and (5.2) respectively.

5.5 GEE framework in claims reserving

Within GLMs framework, we assume that normalized claims payments in different accident and development years are independent variables, i.e. we assume that

$Y_{i,j}$ are independent for all $i, j \in \{1, \ldots, n\}$. If this assumption is not fulfilled, then the GLMs may provide incorrect estimates of the claims reserves.

The claims development triangle consists of observations which are ordered in time. Therefore, according to Hudecová and Pešta (2013), it is natural to suspect that the observations are correlated. The most common approach is to assume that observations within the same accident year are dependent and observations of different accident years are independent. Hence we can handle the claims development triangle as a special type of panel data in which accident years represent the subjects.

The GEE approach enables modelling these dependencies via a working correlation matrix. In Section 2.3.1 some of the most commonly used correlation structures were presented. The AR(1) correlation structure might be the most suitable choice since the claims payments within an accident year are ordered in time and it is natural that correlation between two observations weakens with their time distance. However, when the observations are strongly dependent regardless on their time distances, the exchangeable correlation structure could also be a good choice. The advantage of GEE approach is that the estimates of claims reserves are still valid even if the correlation structure is not chosen correctly.

Furthermore, in all the proposed Tweedie GEE models for claims reserving, the additive accident-development year mean structure (5.5) is assumed.

5.6 Practical application

In the following part of this chapter, we illustrate a practical application of presented claims reserving theory concerning Tweedie GLM and GEE models on a real dataset. The main aim of following analyses is to predict future claims payments and calculate claims reserves.

As it was mentioned several times, the difficulty in using Tweedie GEE model is that there is no function available to compute the parameters directly. Thus, the estimation for this model is based on code from Swan (2006). A selected R code is attached in Appendix B.

5.6.1 Dataset

We consider data from workers' compensation line of business of Lumber insurance company published by Meyers and Shi (2011). The selected dataset contains observations of incremental paid losses in thousands of dollars from accident years 1988-1997 with ten years development lag. The upper triangle, as well as the lower triangle, is included in the data and both triangles are given in Table 5.2. We use the upper triangle to construct the loss reserving models and the lower triangle to calculate the real reserves. These values will be compared with their estimates as a part of model diagnostics.

The visualization of upper triangle is given in Figure 5.2. The graphs show that the data are regular with a significant increase of volume over last accident years. The most of payments are done in the second development year and since then a strong decreasing trend in payments is visible.

Following the notation from Section 5.1, n is equal to ten and the incremental claims payments in Table 5.2 are a realization of random variable $Y_{i,j}$. Note that the exposure is not available in the data, i.e. $w_i = 1$ for all i and hence $Y_{i,j} = Z_{i,j}$ for all i and j.

5.6.2 Application of GLM to claims reserving

To fit a Tweedie GLM, an appropriate value of the power index parameter p needs to be found first. We use the profile log-likelihood function to estimate p, see Section 3.3.1. Note that information about the number of loss payments is not provided in the analysed dataset. Therefore, a simplification in which all n's are set to one has to be used. The maximum likelihood estimate is $\hat{p} = 1.3286$. Figure 5.3 shows the profile log-likelihood plot where the points represent the profile log-likelihood computed at each value in the pre-specified grid P.

Now we can fit the model in terms of equation (5.5). The estimate of the dispersion parameter ϕ is equal to 3.8128 and the estimates of regression parameters are listed in Table 5.3. Note that the values of estimated regression parameters fully coincide with our description of the data illustrated in Figure 5.2. The estimated parameters of accident years reflect the changes in volume over the years. As we expected, the parameter $\widehat{\log \nu_2}$ has the highest value because most payments are done in the second development year.

Model diagnostics

After a model has been fit, it is always wise to check how well the model fits the data. The various diagnostic plots can help us to check the appropriateness of the assumed distribution of the response variable, detect any isolated and systematic discrepancies and assess the predictive ability of the selected model.

One of the most important components of the GLM is that the correct distribution is chosen for the response variable. The Q-Q plot and the histogram of the Pearson residuals are produced to check whether the Tweedie compound Poisson model is adequate for the data. Both graphs are shown in Figure 5.4. The Q-Q plot suggests that the Tweedie GLM is appropriate as all residuals lie close to the line indicating normality. The shape of the histogram also confirms its suitability. Thus, the Tweedie model with the estimated power index parameter $\hat{p} = 1.3286$ and the estimated dispersion parameter $\hat{\phi} = 3.8128$ fits the claims payments appropriately.

The Pearson residuals are displayed in the first graph in Figure 5.5. The residuals are uniformly scattered around zero and no outliers are present. The second graph shows the Pearson residuals plotted against the linear predictor. There is no pattern observed which indicates the right form of the variance function.

The observed values versus fitted values are illustrated in Figure 5.6. The plot shows that nearly all points lie very close to the diagonal line which implies that Tweedie GLM fits the data well.

Figure 5.7 shows the comparison of real claims payments with the fitted and predicted ones of Tweedie GLM model separately for each accident year. True claims payments are illustrated by a black line. A red dashed line distinguishes the fitted values from the predicted values of Tweedie GLM model which are denoted by a green line. Based on these graphs, we see that the fitted values do not significantly differ from the true values besides the early development years

Accident	Development year j									
year i	1	2	3	4	5	6	7	8	9	10
1988	3 270	4 335	2 936	1 675	931	666	273	303	185	96
1989	1 936	3 626	1 963	985	635	242	118	41	85	34
1990	1 187	1 791	1 020	519	412	141	117	43	31	27
1991	1 257	1 771	965	544	159	138	101	−34	18	39
1992	2 592	3 387	2 056	841	694	289	180	180	141	135
1993	3 853	6 343	2 824	1 885	951	505	455	703	67	187
1994	4 727	6 421	3 106	1 903	879	728	543	499	418	300
1995	5 586	6 712	2 974	1 868	1 578	915	519	417	478	234
1996	8 110	8 190	4 130	2 466	1 506	808	434	368	328	55
1997	7 226	7 884	4 569	2 856	1 796	1 024	875	749	466	321

Table 5.2: Development triangle of incremental claims payments (in USD thousands).

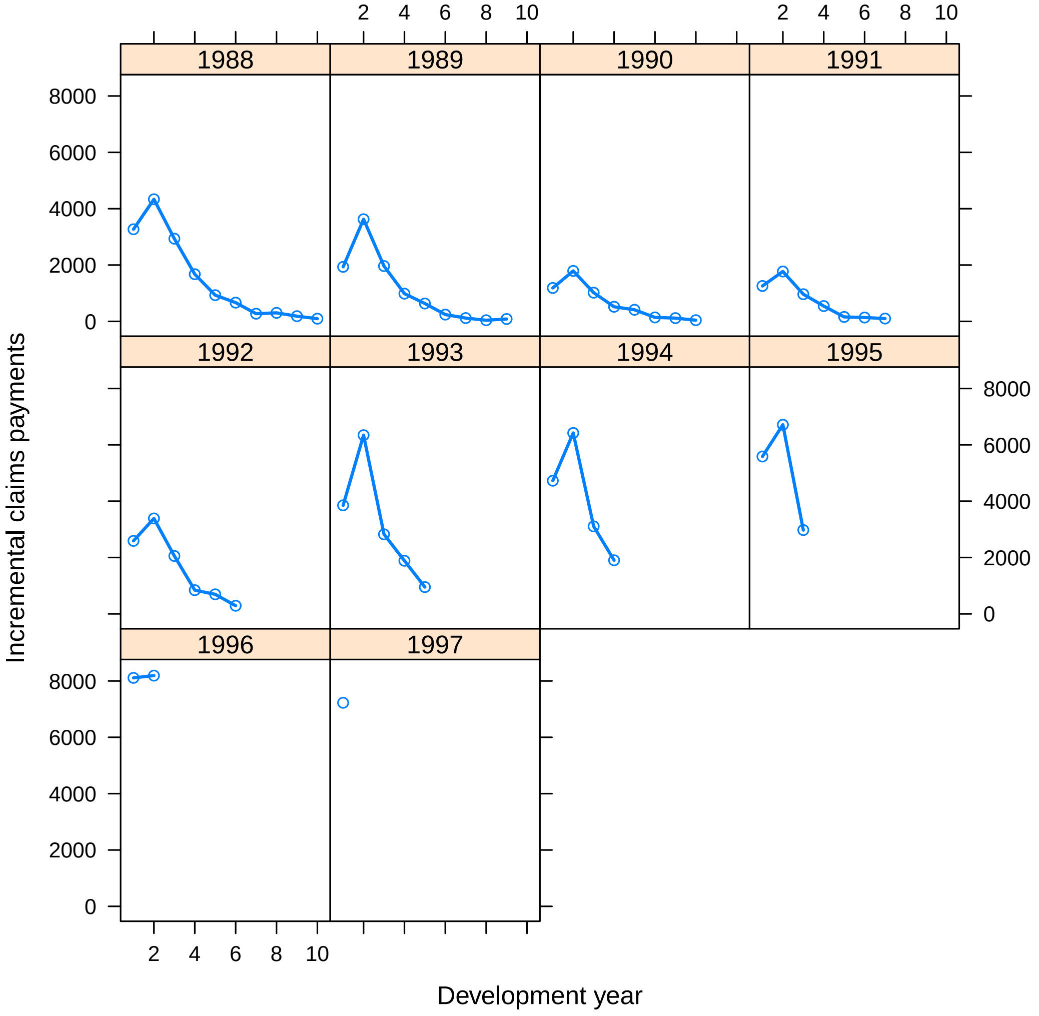

Figure 5.2: Development of incremental claims payments for each accident year.

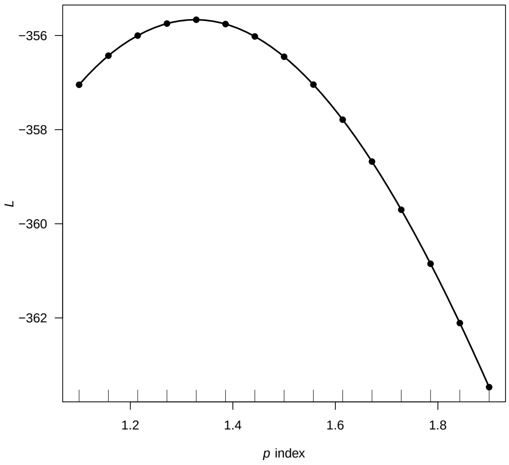

Figure 5.3: The profile log-likelihood plot for p.

s	1	2	3	4	5	6	7	8	9	10
$\widehat{\log\gamma_s}$	0	-0.4776	-1.0480	-1.1127	-0.3904	0.1169	0.2063	0.2752	0.6203	0.6477
$\widehat{\log\nu_s}$	8.2377	8.5256	7.9072	7.3535	6.8260	6.1816	5.5906	5.2225	5.1049	4.5643

Table 5.3: Parameter estimates for Tweedie GLM model.

in 1988 and the first development year in 1996. The predicted values are also quite accurate, except for the years 1996 and 1997 where large gaps are visible in the second development year. This causes the predicted reserves in these years to be higher than the true ones.

5.6.3 Application of GEE to claims reserving

The Pearson residuals of Tweedie GLM model indicate a dependence between incremental claims payments within the same accident year. The Pearson correlation coefficient of the first and the second development year is equal to -0.4274 implying that the application of GEE might be appropriate here.

The results obtained from the application of two Tweedie GEE models to the same dataset are summarized in this part. The AR(1) and exchangeable correlation structures are considered respectively. The reasons for choosing these structures are explained in Section 5.5.

The estimate of the power index parameter p is the same as for Tweedie GLM model. Thus the maximum likelihood estimate is $\hat{p} = 1.3286$.

1. AR(1) correlation structure

First, we fit Tweedie GEE model with AR(1) correlation structure in the terms of equation (5.5).The estimate of parameter ϑ for AR(1) correlation structure equals to -0.0024. The value of $\hat{\vartheta}$ is very close to zero and it suggests that incremental claims payments of the same accident year might be independent.

The estimated regression parameters are given in Table 5.4 and the estimate of dispersion parameter ϕ is $\hat{\phi} = 3.8129$. The detected possible independence be-

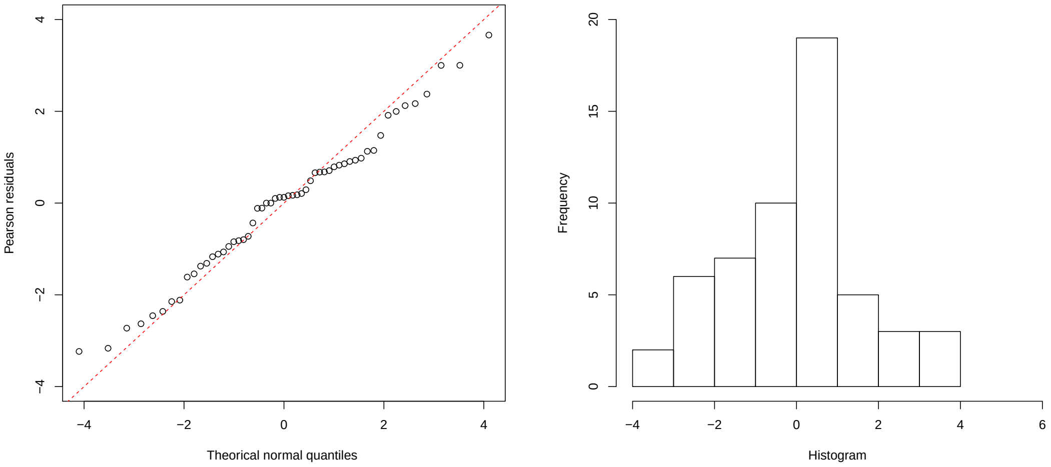

Figure 5.4: Q-Q plot and histogram of Pearson residuals.

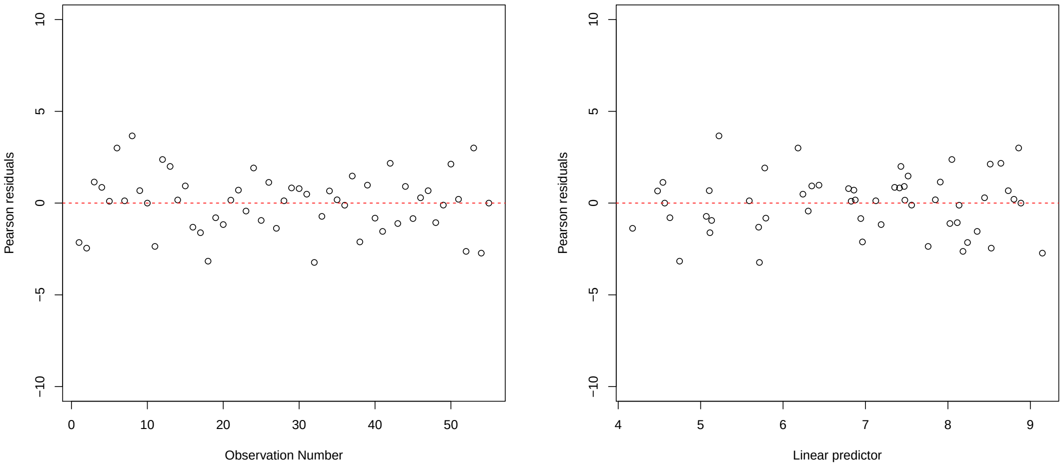

Figure 5.5: Pearson residuals plots for Tweedie GLM model.

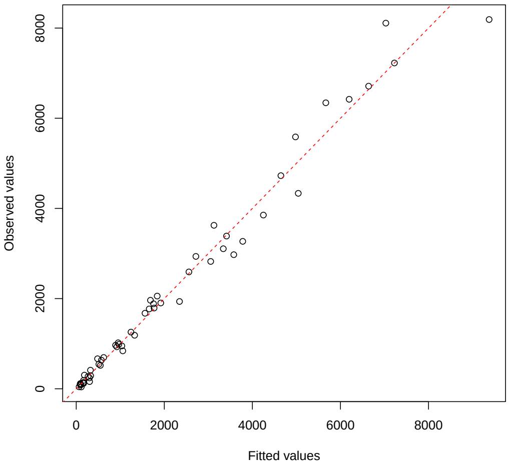

Figure 5.6: Observed vs. fitted values of Tweedie GLM model.

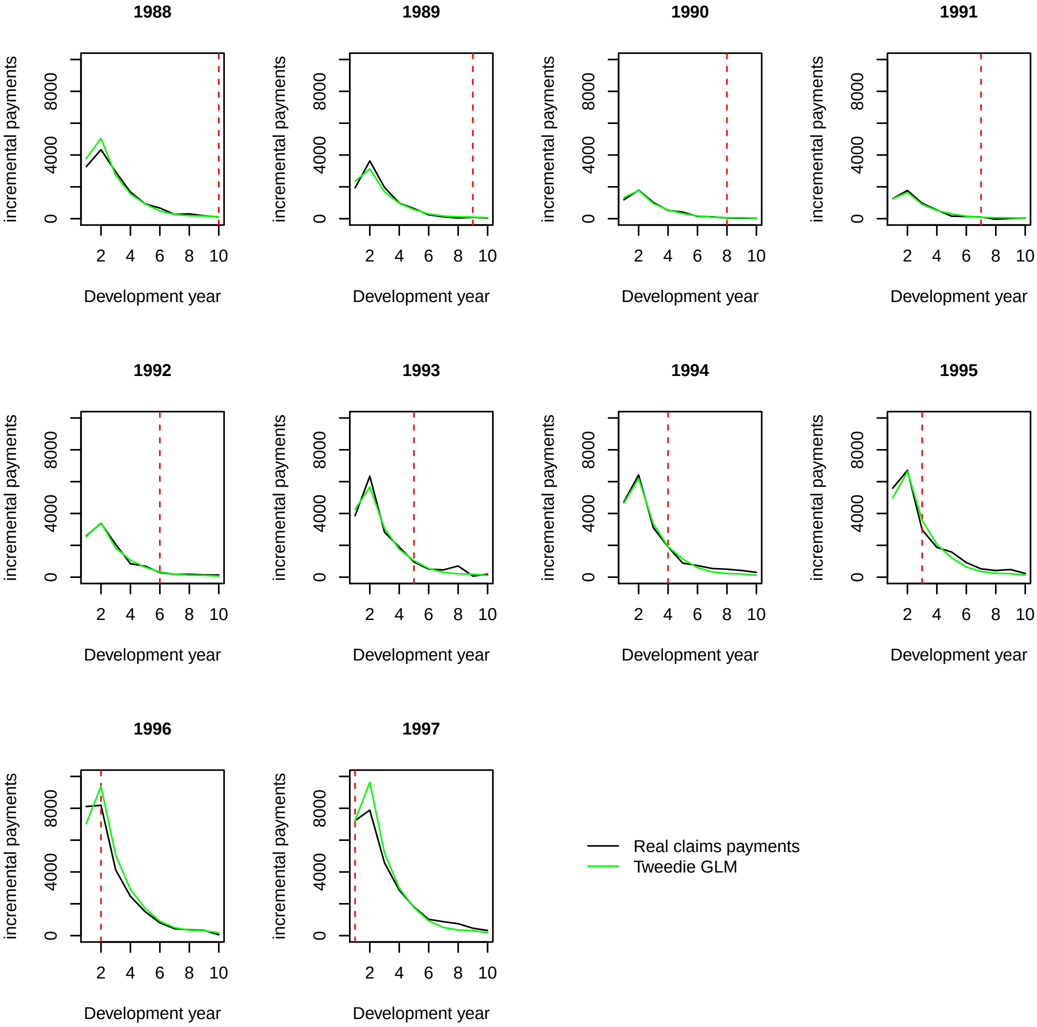

Figure 5.7: Fitted and predicted claims payments vs. real claims payments.

s	1	2	3	4	5	6	7	8	9	10
$\widehat{\log \gamma_s}$	0	-0.4776	-1.0481	-1.1129	-0.3905	0.1168	0.2063	0.2753	0.6202	0.6474
$\widehat{\log \nu_s}$	8.2377	8.5257	7.9074	7.3538	6.8261	6.1816	5.5907	5.2224	5.1049	4.5635

Table 5.4: Parameter estimates for Tweedie GEE AR(1) model.

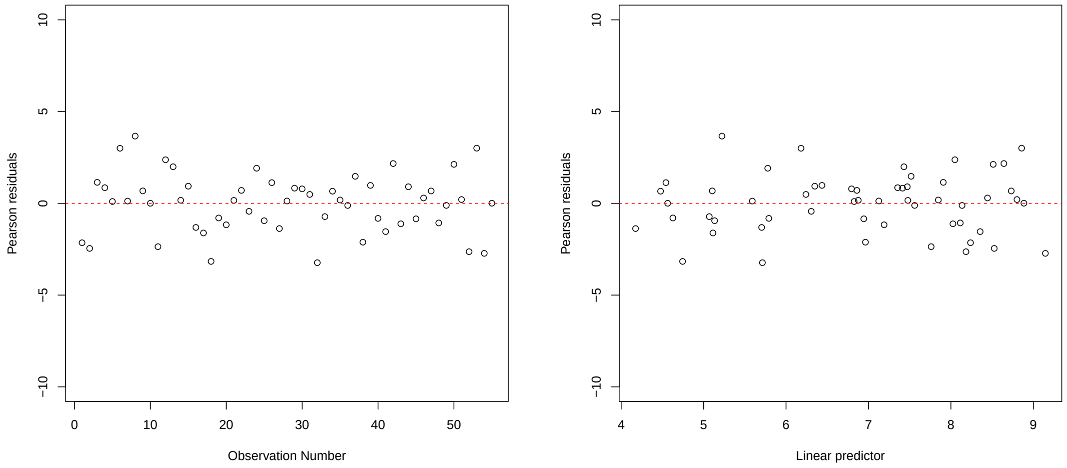

Figure 5.8: Pearson residuals plots for Tweedie GEE AR(1) model.

tween incremental payments causes the parameter estimates to be almost identical to the Tweedie GLM model estimates.

Model diagnostics

Residual diagnostics from this model are shown in Figure 5.8. The first graph illustrates the Pearson residuals. There is no pattern observed as the residuals are symmetrically located around zero. The Pearson residuals versus the linear predictor are displayed in the second graph. The plot suggests that the form of variance function is appropriate since no pattern is visible and the points are uniformly spread. The observed values with respect to the fitted ones are shown in Figure 5.9. The plot confirms a good fit of Tweedie GEE AR(1) model as almost all points are placed along the diagonal line. Moreover, in all the presented plots, there are no outliers. Note that the graphs are nearly indistinguishable from the Figures 5.5 and 5.6.

2. Exchangeable correlation structure

Next, Tweedie GEE model with exchangeable correlation structure is fitted in the terms of equation (5.5). For this structure, the estimate of parameter ϑ equals to -0.0121. The value of $\hat{\vartheta}$ indicates that there might be some very small negative correlation between the incremental payments within the same accident year. The estimate of dispersion parameter ϕ is $\hat{\phi} = 3.8194$ and the estimated parameters are listed in Table 5.5. Comparing them with Tweedie GLM estimates 5.3, some differences can be observed. These differences are more visible from the estimated

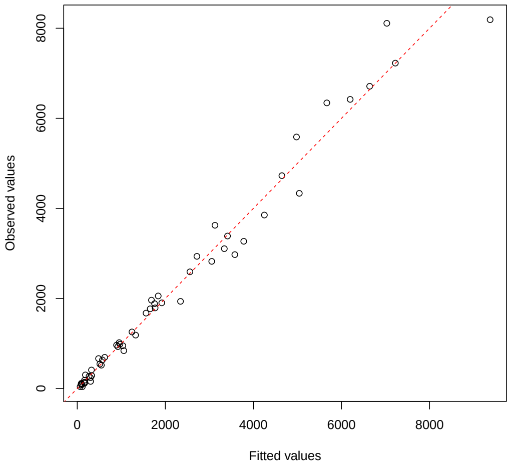

Figure 5.9: Observed vs. fitted values of Tweedie GEE AR(1) model.

s	1	2	3	4	5	6	7	8	9	10
$\widehat{\log \gamma_s}$	0	-0.4778	-1.0484	-1.1131	-0.3906	0.1169	0.2064	0.2752	0.6204	0.6478
$\widehat{\log \nu_s}$	8.2371	8.5250	7.9064	7.3526	6.8249	6.1801	5.5887	5.2203	5.1028	4.5619

Table 5.5: Parameter estimates for Tweedie GEE exchangeable model.

claims reserves given in the following section. The character of estimated values is similar to the parameters estimated in Tweedie GLM model.

Model diagnostics

The same diagnostic plots are produced for Tweedie GEE exchangeable model. The Pearson residuals and the Pearson residuals plotted against the linear predictor are shown in Figure 5.10. The model has analogical residual diagnostics as the previous model. Figure 5.11 illustrates the fitted values with respect to their observed values. The plot indicates a good fit of model as almost all points lie close to the diagonal. It is noticeable that all graphs are again nearly identical to the ones from Tweedie GLM and Tweedie AR(1) models.

5.6.4 Comparison of results

We fitted three different claims reserving models to the same dataset, namely Tweedie GLM model and Tweedie GEE models with AR(1) and exchangeable correlation structures.

Moreover, using the information provided in the lower triangle, we calculated the true values for claims reserves separately for each accident year, as well as in total for all years together. Note that the negative value in the year 1991 was replaced by zero. In 1997, these amounts should have been held by Lumber insurance company in order to meet future payments arising from claims within workers' compensation line of business. We use them as a benchmark for prediction. The estimated claims reserves for all of the considered models together

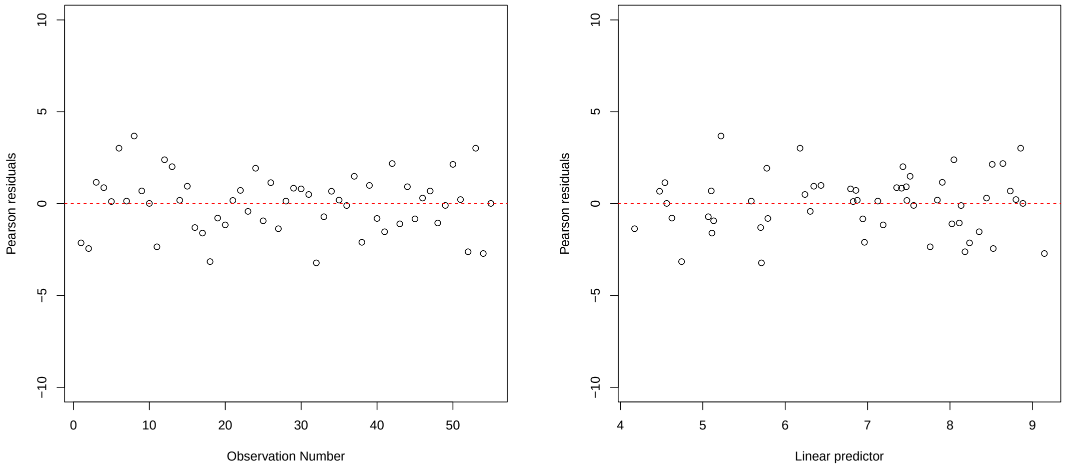

Figure 5.10: Pearson residuals plots for Tweedie GEE exchangeable model.

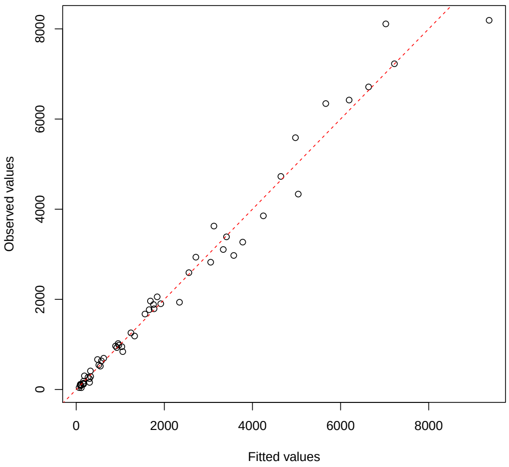

Figure 5.11: Observed vs. fitted values of Tweedie GEE exchangeable model.

	Accident year i									
Model	1989	1990	1991	1992	1993	1994	1995	1996	1997	Total
True	34	58	57	636	1 917	3 367	6 009	10 095	20 540	42 679
GLM	60	91	147	483	1 346	2 605	4 847	11 897	21 863	43 340
GEE AR(1)	59	91	147	483	1 346	2 605	4 848	11 897	21 860	43 337
GEE Exchangeable	59	91	146	482	1 344	2 601	4 841	11 885	21 847	43 297

Table 5.6: True and predicted claims reserves from Tweedie compound Poisson models (in USD thousands).

with their true values are listed in Table 5.6.

Looking at the Table 5.6, we notice that all the models underestimate or overestimate the claims reserves for a single accident years but in total the estimated reserves are higher than the true ones. The Tweedie GEE exchangeable model provides the closest prediction of total claims reserves. Overall, there are very small differences between the predicted reserves. We see that Tweedie GEE model with AR(1) correlation structure provides almost identical reserves estimates as the Tweedie GLM model. This is probably the result of the negligible correlation detected in Tweedie GEE AR(1) model. The predicted reserves obtained from the Tweedie GEE model with exchangeable correlation structure and Tweedie GLM model are very similar.

We have compared the models in terms of their precision of predictions in numbers. Now we can look at their various diagnostic plots. Based on the results from the separate model diagnostics, we can conclude that all of the proposed models performed well and therefore all of them could be considered adequate for estimating claims reserves.

Even though the Tweedie GEE models were more accurate in prediction of claims reserves, we have chosen the Tweedie GLM as our final model. According to the diagnostics checks performed, this model provides a good representation of the data and the distributional assumptions are satisfied. Moreover, the Tweedie GLM is simpler to implement in practice.

Note that the final choice among these models could be further based on some other model diagnostics, e.g. the information criteria or the mean square error of prediction, see Hudecová and Pešta (2013).

Conclusion

Of our main interest was Compound Poisson distribution. We demonstrated that this distribution corresponds to the Tweedie model with $p \in (1, 2)$. Being EDM, Tweedie compound Poisson model fits the GLM and GEE frameworks and its parameters can be estimated by standard GLM and GEE methods, except for the estimation of power index parameter, where the profile likelihood approach is used.

The aim of the presented thesis was to illustrate the applications of Tweedie compound Poisson models in the non-life insurance pricing and claims reserving. The Tweedie compound Poisson model has been found suitable for modelling pure premiums as well as normalized incremental claims. The general frameworks were modified and two different real insurance datasets were used to illustrate such applications. The obtained results were discussed, summarized and compared.

The difficulty in using Tweedie GEE models is that there is no software available that fits Tweedie GEE, therefore it was necessary to write a code so this type of modelling could be implemented. In this thesis, we developed the required R code necessary to compute the parameters of Tweedie GEE models with AR(1) and exchangeable correlation structures. The selected R code for non-life insurance pricing and claims reserving is included in the corresponding appendices.

Bibliography

Dunn, P. and G. Smyth (2001). Tweedie Family Densities: Methods of Evaluation. *Proceedings of the 16th International Workshop on Statistical Modelling, Odense, Denmark.*

Dunn, P. and G. Smyth (2005). Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing 15*, 267–280.

England, P. and R. Verrall (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal 8*(3), 443–544.

Haberman, S. and A. Renshaw (1996). Generalized linear models and actuarial science. *Journal of the Royal Statistical Society 45*(4), 407–436.

Hardin, J. and J. Hilbe (2003). *Generalized Estimating Equations.* Boca Raton: Chapman & Hall/CRC.

Hudecová, Š. and M. Pešta (2013). Modeling Dependencies in Claims Reserving with GEE. *Insurance: Mathematics and Economics 53*, 786–794.

Jørgensen, B. (1997). *The Theory of Dispersion Models.* 1st ed. London: Chapman & Hall.

Jørgensen, B. and M. C. P. D. Souza (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal 1*, 69–93.

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

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

Meyers, G. and P. Shi (2011). Loss Reserving Data Pulled From NAIC Schedule P. `http://www.casact.org/research/index.cfm?fa=loss_reserves_data`. Accessed: 25.1.2017.

Ohlsson, E. and B. Johansson (2010). *Non-Life Insurance Pricing with Generalized Linear Models.* Verlag Berlin Heidelberg: Springer.

Quijano Xacur, O. A. and J. Garrido (2015). Generalised linear models for aggregate claims: to Tweedie or not ? *European Actuarial Journal 5*, 181–202.

Swan, T. (2006). Generalized estimating equations when the response variable has a Tweedie distribution: An application for multi-site rainfall modelling. Master's thesis, The University of Southern Queensland, Toowoomba, QLD.

Wüthrich, M. (2003). Claims reserving using Tweedie's compound Poisson model. *ASTIN Bulletin 33*(2), 331–346.

Wüthrich, M. and M. Merz (2008). *Stochastic claims reserving methods in insurance.* Chichester: Wiley.

A. Source code: Pricing

Appendix A provides a selected R code used for performing the analyses in Section 4.6.

A.1 Exchangeable correlation structure

```
N=length(MCdataU$kon)
K=length(pocty$V1)
n<-pocty$V1

# Estimation of parameter p.
power <- tweedie.profile(skadkost ~ kon+fordald_f+agarald_f+
zon+mcklass, p.vec=seq(1.2, 1.65, length=10), do.plot=TRUE,
data = MCdataU, weights = duration,do.ci=F,method="interpolation",
control=list( maxit=800))
(p=power$p.max)

# Tweedie compound Poisson GEE initial model.
summary(glm_model0<- glm(skadkost ~ kon+fordald_f
+agarald_f+zon+mcklass, family = tweedie(var.power=p,
link.power=0), data=MCdataU,weights = duration,x=TRUE))

beta=glm_model0$coefficients
X_model=glm_model0$x
k=glm_model0$rank
phi = power$phi.max
fits=glm_model0$fitted.values

# Attaching fitted values and residuals to dataset.
MCdataU<-cbind(fits,MCdataU)
p.residuals=sqrt(MCdataU$duration)*
(MCdataU$skadkost-MCdataU$fits)/sqrt(MCdataU$fits^p)
MCdataU<-cbind(p.residuals,MCdataU)
head(MCdataU)

dev=sum(tweedie.dev(MCdataU$skadkost,MCdataU$fits,p))
devold=100*dev
epsilon =1e-5

## Dividing into groups.
groups<- list()
groups<-dlply(MCdataU, .(kon,fordald2,agarald2,zon,mcklass))

### Repeat steps 2-5 until the convergence criterion
is not fulfilled.
```

```
while (abs(dev - devold)/(0.1 + abs(dev)) > epsilon) {

### Step 2 - Calculate Pearson residuals for the model.
MCdataU$p.residuals=sqrt(MCdataU$duration)*
(MCdataU$skadkost-MCdataU$fits)/sqrt(MCdataU$fits^p)
groups<-dlply(MCdataU, .(kon,fordald_f,agarald_f,zon,mcklass))

### Step 3 - Calculate estimates of phi and vartheta
phi<-sum(MCdataU$p.residuals^2)/(N-k)

vartheta=NULL
pom_sum=NULL
pom_sum1=NULL
pom_sum2=NULL

for (j in 1:K){
if (n[j]>1){
## First part of the sum.
for (u in 1:n[j]){
for (v in 1:n[j]){
pom_sum[v]=groups[[j]][u,1]*groups[[j]][v,1]
}
pom_sum1[u]=sum(pom_sum)
pom_sum=NULL
}
##Second part of the sum.
        for(u in 1:n[j]){
          pom_sum2[u]=groups[[j]][u,1]^2
         }

  vartheta[j]= (sum(pom_sum1)-sum(pom_sum2))/(n[j]*(n[j]-1))
   pom_sum1=NULL
   pom_sum2=NULL
}  else { vartheta[j]=0}
}

vartheta=sum(vartheta)/phi

## Calculate R for Exch. correlation structure
R<- list()
for (i in 1:K){
R[[i]]=matrix(vartheta,nrow=n[i],ncol=n[i])
for (j in 1:n[i]){
R[[i]][j,j]=1
   }
}

### Step 4: Calculate A.
```

```
A<- list()
for(i in 1:K){
if (n[i]>1){
A[[i]]=diag(groups[[i]][,2]/groups[[i]][,8])^(p/2)
} else
{A[[i]]=matrix((groups[[i]][,2]/groups[[i]][,8])^(p/2),
nrow=1,ncol=1)}
}

### Step 5: Update beta.
## Calculate D.
D<- list()

D[[1]]=matrix(nrow=n[1],ncol=k)
for(j in (1:k)){
 D[[1]][,j]=groups[[1]][,2]*X_model[,j][1:n[1]]
     }

for (i in 2:K){
D[[i]] = matrix(nrow=n[i],ncol=k)
fitted=NULL
fitted<-matrix(groups[[i]][,2],nrow=n[i],ncol=1)
for(j in (1:k)){
     D[[i]][,j]=fitted*X_model[,j][(sum(n[1:(i-1)])+1):
     (sum(n[1:i]))]
     }
}

## Calculate V.
V<- list()
for (i in 1:K){
V[[i]] = phi*(A[[i]]  %*% R[[i]]  %*% A[[i]])
}

## Calculate inverse of V.
invV<- list()
for (i in 1:K){
invV[[i]]=solve(V[[i]])
}

## Calculate C.
C<- list()
for (i in 1:K){
C[[i]]=t(D[[i]]) %*% invV[[i]] %*% D[[i]]
}

## Calculate B.
B<- list()
```

```
for (i in 1:K){
pom=NULL
pom<-matrix(groups[[i]][,10]-groups[[i]][,2],nrow=n[i],ncol=1)
B[[i]]=t(D[[i]]) %*% invV[[i]] %*% pom
}

## Get final matrices B and C.
C_final<-Reduce('+', C)
B_final<-Reduce('+', B)

### Fit the new values..
beta=beta + (solve(C_final) %*% B_final)

MCdataU$fits=exp(X_model %*% beta)
groups<-dlply(MCdataU, .(kon,fordald_f,agarald_f,zon,mcklass))
devold<-dev
dev=sum(tweedie.dev(MCdataU$skadkost,MCdataU$fits,p))
} # End of the while cycle.
```

B. Source code: Claims reserving

Appendix B provides a selected R code used for performing the analyses in Section 5.6.

B.1 AR(1) correlation structure

```
# Estimation of parameter p.
power=tweedie.profile(inc_pdloss ~  devy + ay -1 ,data=data_inc,
p.vec=seq(1.1,1.9,length=15),
do.plot=T,do.smooth=T,do.ci=F,method="interpolation")
(p=power$p.max)

# Tweedie compound Poisson GEE initial model.
glm_model0 = glm(inc_pdloss ~ devy + ay -1 ,
family = tweedie(var.power=p, link.power=0), data=data_inc,x=TRUE)
summary(glm_model0)

n=length(data_inc$ay)
beta=glm_model0$coefficients
k=glm_model0$rank
phi = power$phi.max
fits=glm_model0$fitted.values

# Set the variables to be used in the convergence criteria.
dev=sum(tweedie.dev(data_inc$inc_pdloss,fits,p))
devold=100*dev
epsilon =1e-8

### Repeat steps 2-5 until the convergence criterion
is not fulfilled.

while (abs(dev - devold)/(abs(dev)+0.1) > epsilon) {

### Step 2 - Calculate Pearson residuals for the model.
p.residuals=(data_inc$inc_pdloss-fits)/sqrt(fits^p)

### Step 3 - Calculate estimates of phi and vartheta
phi<-sum(p.residuals^2)/(n-k)

## AR(1) correlation structure.
vartheta=NULL
pom_sum=NULL

for (j in 1:(n-1)){
pom_sum[j]=p.residuals[j]*p.residuals[j+1]
}
vartheta=(sum(pom_sum))/(phi*(n-1)*(n-k))
print(paste0("vartheta=",vartheta))

## Calculate R.
index<-seq(0,n-1,by=1)
longindex<-c(seq(n-1,1,by=-1),index)
i=0

R=matrix(nrow=n,ncol=n)
while(i<n){
     R[i+1,]=vartheta^longindex[(n-i):(2*n-i-1)]
      i=i+1
}

### Step 4: Calculate A.
A=diag(fits)^(p/2)

### Step 5: Update beta.
## Model matrix X.
X_model=glm_model0$x

## Calculate D.
D = matrix(nrow=n,ncol=k)
for(i in (1:k)){
D[,i]=fits*X_model[,i]
}

V = phi*(A  %*% R  %*% A)
svdV=svd(V)

C= t(D) %*% svdV$v%*%diag(1/svdV$d)%*%t(svdV$u) %*% D
B= t(D) %*% svdV$v%*%diag(1/svdV$d)%*%t(svdV$u) %*%
(data_inc$inc_pdloss-fits)

## Fit the new values.
svdC=svd(C)
beta=beta+ (svdC$v%*%diag(1/svdC$d)%*%t(svdC$u) %*% B)

fits<-exp(t(beta) %*% t(X_model))
fits<-as.vector(fits)
devold<-dev
dev<-sum(tweedie.dev(data_inc$inc_pdloss,fits,p))
} # End of the while cycle.
```

B.2 Exchangeable correlation structure

```
## Step 3 - Calculate estimates of phi and vartheta.
```

```
phi<-sum(p.residuals^2)/(n-k)

vartheta=NULL
pom_sum=NULL
pom_sum1=NULL
pom_sum2=NULL

for(u in 1:n){
   for(v in 1:n){
pom_sum[v]=p.residuals[u]*p.residuals[v]
    }
  pom_sum1[u]=sum(pom_sum)
  pom_sum=NULL
}

for(j in 1:n){
pom_sum2[j]=(p.residuals[j])^2
}

vartheta= ((sum(pom_sum1)-sum(pom_sum2))/(phi*n*(n-1)))

R = matrix(vartheta,nrow=n,ncol=n)
for (j in 1:n){
R[j,j]=1
}
```