VRIJE
UNIVERSITEIT
BRUSSEL

Graduation thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Mathematics

EXPLORATION OF VARIOUS MACHINE LEARNING TECHNIQUES IN THE CONTEXT OF NON-LIFE INSURANCE

Othman El Hammouchi

June 2023

Promotors: prof. dr. Robin Van Oirbeek prof. dr. Tim Verdonck

Sciences and Bioengineering Sciences

.
.

VRIJE
UNIVERSITEIT
BRUSSEL

Proefschrift ingediend met het oog op het behalen van de graad van Master of Science in de Wiskunde

VERKENNING VAN VERSCHILLENDE MACHINE LEARNING-TECHNIEKEN IN DE CONTEXT VAN NON-LIFE INSURANCE

Othman El Hammouchi

Juni 2023

Promotors: prof. dr. Robin Van Oirbeek prof. dr. Tim Verdonck

Wetenschappen en Bio-ingenieurswetenschappen

Abstract

Your abstract would go here.

Contents

List of Symbols

The next list describes several symbols that will be later used within the body of the document

C_{ij} Cumulative claim amount

Chapter 1

Introduction

Chapter 2

Pattern break detection

2.1 Introduction

The most defining characteristic of the insurance industry is the inverted nature of its production cycle. In manufacturing, commerce, transport, etc., payment is usually received only upon delivery of goods or services. By contrast, insurance products are purchased long before the adverse events which they protect against have occured, if they ever do. Insurers therefore face the challenge of forecasting the amount and variability of funds needed to settle outstanding contracts, a process known as *claims reserving*. In this the reserving actuary relies historical data which is most often presented in the form of a *loss* or *run-off triangle* $\mathcal{D}_I$, which consists either of cumulative or incremental amounts of some actuarial variable (payments, number of claims, etc.), respectively denoted by C_{ij} and X_{ij}. Here $1 \leq i \leq I$ denotes the *cohort*, *origin year* or *accident year* and $1 \leq j \leq J$ the *development year*, so that

$$\mathcal{D}_I = \{C_{ij} \mid 1 \leq j \leq J, i+j \leq I+1\} \quad \text{or} \quad \mathcal{D}_I = \{X_{ij} \mid 1 \leq j \leq J, i+j \leq I+1\} \ .$$

To simplify the formulas, we assume throughout this exposition that $I = J$. Embedding $\mathcal{D}_I$ into a matrix on and above the anti-diagonal, the actuary then seeks to predict the *total outstanding loss liabilities*

$$R = \sum_{i=2}^{I} (C_{i,I} - C_{i,I+1-i})$$

by forecasting the values in the lower triangle $\mathcal{D}_I^{\complement}$. A special difficulty arising in the actuarial context is the relatively small number of observations which is usually available.

One of the most frequently used loss reserving techniques in practice is the so-called *chain ladder* (CL), which predicts the cumulative claim in development year j by multiplying the previous year's amount by a so-called *age-to-age factor*, *link ratio* or *development factor*. It was originally conceived as a purely computational algorithm, but there have since been various attempts to frame it in terms of a stochastic model. The central assumption it makes is that the pattern observed in earlier cohorts is applicable to later ones. In one sense, this is of course perfectly reasonable: all models ultimately use the past as a guide to the future. The dearth of data typically available to the actuary makes it challenging to verify its validity, as it limits the efficacy of classical statistical techniques. In particular, this makes it difficult to detect structural breaks in the claims development pattern.

Our aim in this chapter is to investigate whether it is possible to use bootstrap simulations to remedy this problem. Specifically, we start from two widely-used chain ladder models (Mack's

model and the Poisson GLM) and simulate run-off triangles which perfectly follow their assumptions. We then perturb the triangles in a myriad of ways and calculate a bootstrap reserve for the resulting data, allowing us to investigate how the simulated reserve is impacted by deviations from the model assumptions.

The chapter is divided as follows. The next section introduces the bootstrap method and explains how it can be applied to regression models, which will

2.2 The bootstrap method

When using a statistical model to describe a dataset in terms of a reduced number of parameters, we are not only interested in producing point estimates of these parameters, but also in quantifying their *uncertainty*. In classical statistics, the usual approach to achieve this is to start from the model assumptions and derive from them analytically the sampling distribution of the estimators. In most cases (the Gaussian distribution being a notable exception) this leads to intractible calculations, so that one is either forced to rely on approximations and asymptotic results, or make unrealistic simplifying assumptions. Moreover, estimates obtained in this way often heavily depend on their underlying assumptions, which can potentially lead to gross errors if these are violated.

The bootstrap method aims to remedy this problem by using numerical simulations to compute estimates of model uncertainty. At its core, it is premised on the idea that the empirical distribution of the sample forms a good proxy for that the population distribution. Consequently, we can approximate sampling from the population by *resampling our data*, which, to the uncareful observer, can give the impression that we're 'magically' producing new information, using our single sample to 'pull ourselves up by our own bootstraps', which is where the procedure derives its name from. Let's see how this can be done concretely for a simple estimation problem.

2.2.1 Bootstrapping an estimator

Let $X_1, \dots, X_n$ be an i.i.d. sample drawn from a distribution F, and consider an estimator $\widehat{h(F)} = g(X_1, \dots, X_n)$ of some quantity $h(F)$ whose uncertainty we wish to estimate, using e.g. the variance of the sampling distribution. Depending on the assumptions we are willing to make, we can choose between two broad approaches: *parameteric* methods and *nonparametric* ones.

In the nonparametric bootstrap, we use the data directly, drawing with replacement to simulate new samples $X_1^{(b)}, \dots, X_n^{(b)}$. In other words, we approximate F using the *empirical cumulative distribution function*

$$\widehat{F}_n(x) \coloneqq \sum_{k=1}^{n} I_{\{X_k \leq x\}} \,,$$

which we use to generate new data. We then compute the statistic of interest on these pseudo-samples, yielding pseudo-observations $g^{(b)} = g(X_1^{(b)}, \dots, X_n^{(b)})$ which approximate the sampling distribution of $\widehat{h(F)}$. Writing B for the total number of bootstrap samples, we can esimate the variance of $\widehat{h(F)}$ by

$$\frac{1}{B-1} \sum_{b=1}^{B} (g^{(b)} - \bar{g})^2 \,,$$

with $\bar{g} = \frac{1}{B} \sum_{b=1}^{B} g^{(b)}$. Provided $F \approx \widehat{F}_n$ holds with sufficient accuracy, this will yield a reasonable approximation to $\mathrm{Var}(\widehat{h(F)})$.

By contrast, in the parametric bootstrap, we first fit a model using the data, and then simulate samples from this with the help of a random number generator. As usual, the parametric approach offers the advantage of efficiency if its assumptions are met, at the risk of increased error when they are violated. If we assume that F belongs to some family $\{F_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta\}$, then we can use the sample $X_1, \dots, X_n$ to produce an estimate $\widehat{\boldsymbol{\theta}}$ of the parameter. Plugging this in then gives us $F_{\widehat{\boldsymbol{\theta}}}$, from which we can simulate $X_1^{(b)}, \dots, X_n^{(b)}$ and $g^{(b)}$ as before. An estimate of the sampling variance is likewise obtained the same manner.

Even though the previous example involves the calculation a single statistic, it is clear that the bootstrap produces a complete *simulated distribution* of the estimator, which can be used for any arbitrary form of inference. This shows the tremendous potential of bootstrap as a tool for statistical analysis, which explains its rise in popularity with the advent of powerful personal computers capable of carrying out the requisite calculations.

2.2.2 Bootstrapping a regression model

Altough we have introduced the bootstrap in the context of a classical one-sample estimation problem, the same principles can be applied to data structures of arbitrary complexity, so long as we have a model for the probabilistic mechanism generating the observations (see [1, Chapter 8] for a general exposition of this methodology). In particular, bootstrap methods for regression models are well-established in the literature. We now turn our attention to these, as they will form the foundation for developing bootstrap methods for claims triangles.

Consider a set of covariates $X_1, \dots, X_p$ and a response variable Y whose relationship we model by a parametrised mapping $f(X_1, \dots, X_p; \boldsymbol{\beta})$. Given a sample of pairs $(\mathbf{x_1}, Y_1), \dots, (\mathbf{x_n}, Y_N)$ and a choice of loss function, we can fit this model to obtain an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. For new values $x_1^+, \dots, x_N^+$ of the regressors, we can then predict the response Y^+ as $f(x_1^+, \dots, x_N^+; \widehat{\boldsymbol{\beta}})$. It is worth emphasising these as two distinct operations, which correspond to different bootstrap procedures (see [2, Sections 6.3.3 and 7.2.4]). *Estimation* seeks to *identify* the value of a quantity which is *fixed but unknown*; *prediction* aims to *forecast* the value of a *random variable*.

Under the least squares criterion, for example, we know that the optimal predictor for Y is the conditional expectation $\mathbb{E}[Y \mid X = x]$. This is an ordinary function which returns a real number for any $x \in \mathbb{R}^p$, and which can therefore be estimated from a sample. Such an estimate will contain some error, which we have to take into account when doing inference. If we additionally want to measure the error we make when predicting Y using $\mathbb{E}[Y \mid X = x]$, we also have to incorporate the intrinsic randomness of the response variable. Prediction is therefore a two-phase procedure involving an intermediate estimation step.

Let's illustrate this in the case of the all-familiar linear regression model, which is given by

$$Y_i = \mathbf{x_i}^T \boldsymbol{\beta} + \varepsilon_i \,, \qquad i \in 1, \dots, n \tag{2.1}$$

with $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma$ and $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$. Considering the nonparametric bootstrap first, we need to identify a fundamental unit of resampling such that the resulting variables are interchangeable. One option would be to fit the model and use it to compute some kind of suitably scaled residuals, which we will then resample (see [2, Algorithm 6.1]). This approach is sometimes referred to as *semiparametric*, because it only uses the specification of certain aspects of the data distribution in terms of some parameters, but does not assume a specific form for it. If we then choose the standardised residuals

$$r_i = \frac{Y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}}{\widehat{\sigma}\sqrt{1 - h_{ii}}} \,,$$

for example, we obtain from these the bootstrap samples $r_1^{(b)}, \dots, r_n^{(b)}$, which in turn yield bootstrap responses

$$Y_i^{(b)} \coloneqq \mathbf{x}^T \widehat{\boldsymbol{\beta}} + \widehat{\sigma}\sqrt{1-h_{ii}} r_i^{(b)} .$$

By refitting the model to this new data, we finally obtain simulated regression parameters $\widehat{\boldsymbol{\beta}}^{(b)}$. If we also want to bring in the variability of the response, we can resample the residuals a second time and the result to the regression line in order to simulate the fluctuations around it.

An alternative approach, which is fully nonparametric, is to resample the pairs $(\mathbf{x_i}, Y_i)$ themselves (see [1, Section 9.5], [2, Algorithm 6.2]), which corresponds to approximating the multivariate distribution of $(X_1, \dots, X_n, Y)$ by the empirical distribution of our data. This has the significant benefit of parsimony, making no other assumption beside the i.i.d.-ness of the sample. The model is then fitted to the bootstrap samples $(\mathbf{x_1}^{(b)}, Y_1^{(b)}), \dots, (\mathbf{x_n}^{(b)}, Y_n^{(b)})$ to produce pseudo-realisations $\widehat{\boldsymbol{\beta}}^{(b)}$ of the regression parameter estimator. One drawback of this approach is that it does not have a good mechanism for simulating the process error, so that we need to borrow this part either from one of the other methods.

For the parametric case, we have to make an additional assumption about the distribution of ϵ, the classical choice being the normal distribution. We then begin to fit (2.1), which gives us estimates $\widehat{\beta}$ for the regression parameters. With the help of a random number generator, we then produce bootstrap responses $Y_1^{(b)}, \dots, Y_n^{(b)}$ by drawing from the estimated distribution $\mathcal{N}(\mathbf{x}^T\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$, and fit the model to this new data to obtain bootstrap samples $\widehat{\boldsymbol{\beta}}^{(b)}$ of the regression parameters.

With this introduction complete, we now move on to discussing our first model.

2.3 Mack's model

In his seminal paper [3], Mack proposed the following model for cumulative claims triangles, which remains among the most influencial in actuarial reserving.

Model 1 (Mack Chain Ladder).

(i) There exist development factors $f_1, \dots, f_{I-1}$ such that

$$\mathbb{E}[C_{ij} \parallel C_{i,j-1}, \dots, C_{i1}] = \mathbb{E}[C_{ij} \parallel C_{i,j-1}] = f_{j-1} C_{i,j-1} \tag{2.2}$$

for $1 \le i \le I$.

(ii) There exist parameters $\sigma_1, \dots, \sigma_{I-1}$ such that

$$\mathrm{Var}[C_{ij} \parallel C_{i,j-1}, \dots, C_{i1}] = \mathrm{Var}[C_{ij} \parallel C_{i,j-1}] = \sigma_{j-1}^2 C_{i,j-1} ,$$

for $1 \le i \le I$.

(iii) Cumulative claims processes $(C_{ij})_j, (C_{i'j})_j$ are independent for $i \neq i'$.

The development factors are estimated by

$$\widehat{f}_j(\mathcal{D}_I) = \widehat{f}_j(C_{1j}, \dots, C_{I-j,j}, \dots, C_{1,j+1}, \dots, C_{I-j,j+1}) \coloneqq \frac{\sum_{i=1}^{I-j} C_{i,j+1}}{\sum_{i=1}^{I-j} C_{i,j}} . \tag{2.3}$$

If we define the *single* or *individual* development factors as

$$F_{i,j+1} \coloneqq \frac{C_{i,j+1}}{C_{ij}} ,$$

then $\widehat{f}_j$ can be obtained as the weighted average

$$\widehat{f}_j = \frac{\sum_{i=1}^{I-j} C_{ij} F_{i,j}}{\sum_{i=1}^{I-j} C_{ij}} .$$

The σ_j are estimated by

$$\widehat{\sigma}_j \coloneqq \frac{1}{I-j} \sum_{i=1}^{I-j} C_{ij} \left(F_{i,j+1} - \widehat{f}_j \right)^2$$

for $j < I-1$. This formula does not work for $j = I-1$, as we only have a single pair of observations in the last two columns of the triangle. To remedy this, Mack proposed a simple extrapolation from the previous development years, leading to the estimate

$$\widehat{\sigma}^2_{I-1} = \min \left\{ \frac{\widehat{\sigma}^4_{I-2}}{\widehat{\sigma}^2_{I-3}}, \widehat{\sigma}^2_{I-2}, \widehat{\sigma}^2_{I-3} \right\}$$

and this appears to be the most widely adopted solution in the literature.

Under the assumptions of Model 1, it can be shown (see [4, pp. 17 sqq.]) that $\widehat{f}_j$ and $\widehat{\sigma}_j$ are (conditionally) unbiased, and moreover that the $\widehat{f}_j$ are uncorrelated. Predicted ultimate claim amounts C_{iI} are obtained by substituting the estimates for the unknown development factors f_j in the conditional expectation. In other words, we predict the ultimate loss using the conditional mean $\mathbb{E}[C_{iI} \parallel C_{i,I+1-i}]$, and estimate the latter by plugging in $\widehat{f}_j$, yielding

$$\widehat{C}_{iI} \coloneqq \widehat{\mathbb{E}}[C_{i,I} \parallel C_{i,I-i}] = C_{i,I+1-i} \prod_{j=I-i}^{I-1} \widehat{f}_j .$$

From this, we then finally obtain the reserve predictor

$$\widehat{R} = g(\mathcal{D}_I) \coloneqq \sum_{i=2}^{I} (\widehat{C}_{i,I} - C_{i,I+1-i}) . \tag{2.4}$$

Model 1 is often referred to as "distribution-free" because it only makes assumptions about the first two moments of the claims triangle variables. Indeed, we will show that the Mack CL can be viewed as a series of linear regressions through the origin (i.e. without intercept term), hence these are same assumptions as for the Gauss-Markov theorem, i.e. the minimal ones[1] required to guarantee optimality. Introduce, for any development year $j \in \{1, \dots, I-1\}$, the notation

$$\mathbf{c_j} \coloneqq \begin{bmatrix} C_{1,j} \\ \vdots \\ C_{I-j,j} \end{bmatrix} ,$$

then the first two assumptions of Model 1 can be equivalently stated as

$$\mathbf{c_{j+1}} = f_j \mathbf{c_j} + \boldsymbol{\varepsilon} ,$$

with $\boldsymbol{\varepsilon}$ a random vector satisfying

$$\mathbb{E}[\boldsymbol{\varepsilon} \parallel C_{1,j}, \dots, C_{i,I-j}] = \mathbf{0} \qquad \mathrm{Var}(\boldsymbol{\varepsilon} \parallel C_{1,j}, \dots, C_{i,I-j}) = \sigma_j^2 \begin{bmatrix} C_{1j} & & \\ & \ddots & \\ & & C_{I-j,j} \end{bmatrix} .$$

[1] If we want to be completely precise, the third assumption is slightly stronger than needed, as Gauss-Markov only requires the errors to be uncorrelated.

Consequently, it follows (see [5, Proposition 1.7]) that the weighted least squares method with weights matrix

$$\mathbf{W} = \frac{1}{\sigma_j^2} \begin{bmatrix} 1/C_{1j} & & \\ & \ddots & \\ & & 1/C_{I-j,j} \end{bmatrix},$$

leads to an estimator for f_j which has minimal variance in the class of linear unbiased estimators. This estimator is given by

$$\widehat{f}_j^{\mathrm{WLS}} = (\mathbf{c_j}^T \mathbf{W} \mathbf{c_j})^{-1} \mathbf{c_j}^T \mathbf{W} = \frac{\sum_{i=1}^{I-j} C_{i,j+1}}{\sum_{i=1}^{I-j} C_{i,j}},$$

which is the same expression as (2.3).

2.3.1 A challenging simulation

Owing to its recursive nature, Mack's model does not readily lend itself to application of the theory from Section 2.2. The actuarial literature on bootstrap methods is not very helpful in this regard either, as it has mostly tended to focus on generalised linear models—even papers like [6] which address the Mack CL do so by reframing it in this way. As will become clear shortly, this passes over some subtleties related to the particular structure of Mack's model, and we will therefore take a different approach. In particular, our starting point will be the problem of deriving a closed-form estimate of the so-called conditional *mean square error of prediction* (MSEP) for the Mack predictor. While this might appear at first glance to be unrelated to the bootstrap, we will see that it furnishes us with the necessary theoretical framework to understand the special issues involved in resampling a recursive model.

The MSEP is a measure for the total uncertainty associated with a given predictive model. It is defined as the Euclidean distance between the predictor and the response in the underlying filtered probability space, i.e.

$$\operatorname*{MSEP}_{R \,|\, \mathcal{D}_I}(\widehat{R}) \coloneqq \mathbb{E}\Big[(\widehat{R} - R)^2 \;\Big\|\; \mathcal{D}_I\Big]$$

for our special case of predicting the reserve. The MSEP admits a decomposition, similar to the familiar bias-variance decomposition from classical statistics into so-called *parameter* or *estimation error* and *process error*:

$$\begin{aligned}
\mathbb{E}\Big[(\widehat{R} - R)^2 \;\Big\|\; \mathcal{D}_I\Big] &= \mathbb{E}\big[(R - \mathbb{E}[R \parallel \mathcal{D}_I])^2 \parallel \mathcal{D}_I\big] + \mathbb{E}\Big[(\mathbb{E}[R \parallel \mathcal{D}_I] - \widehat{R})^2 \;\Big\|\; \mathcal{D}_I\Big] \\
&\quad - 2\mathbb{E}\Big[(R - \mathbb{E}[R \parallel \mathcal{D}_I])(\mathbb{E}[R \parallel \mathcal{D}_I] - \widehat{R}) \;\Big\|\; \mathcal{D}_I\Big] \\
&= \operatorname{Var}(R \,\|\, \mathcal{D}_I) + (\mathbb{E}[R \parallel \mathcal{D}_I] - \widehat{R})^2 \\
&\quad - 2(\mathbb{E}[R \parallel \mathcal{D}_I] - \widehat{R})(\mathbb{E}[R - \mathbb{E}[R \parallel \mathcal{D}_I] \parallel \mathcal{D}_I]) \\
&= \underbrace{\operatorname{Var}(R \,\|\, \mathcal{D}_I)}_{\text{process error}} + \underbrace{(\mathbb{E}[R \parallel \mathcal{D}_I] - \widehat{R})^2}_{\text{estimation error}},
\end{aligned}$$

corresponding to the two stages of bootstrapping a predictor which we discussed in Section 2.2.2. Consider now, for any accident year $i \in \{1, \dots, I\}$, the MSEP for the associated ultimate

$$\operatorname*{MSEP}_{C_{iI} \,|\, \mathcal{D}_I}(\widehat{C}_{iI}) = (\mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \widehat{C}_{iI})^2 + \operatorname{Var}(C_{iI} \parallel \mathcal{D}_I),$$

and suppose we are interested in obtaining a closed-form estimator for it. Such an expression can be derived relatively straightforwardly for the process error from the assumptions of Model 1 in the following way. We begin by applying the law of total variance in conjunction with (2.2) to obtain

$$\begin{aligned}
\mathrm{Var}(C_{iI} \parallel \mathcal{D}_I) &= \mathrm{Var}(C_{iI} \parallel C_{i,I+1-i}) \\
&= \mathbb{E}[\mathrm{Var}(C_{iI} \parallel C_{i,I-1}) \parallel C_{i,I+1-i}] + \mathrm{Var}(\mathbb{E}[C_{iI} \parallel C_{i,I-1}] \parallel C_{i,I+1-i}) \\
&= \sigma_{I-1}^2 \mathbb{E}[C_{i,I-1} \parallel C_{i,I+1-i}] + f_{I-1}^2 \mathrm{Var}(C_{i,I-1} \parallel C_{i,I+1-i}) \\
&= \sigma_{I-1}^2 C_{i,I+1-i} \prod_{j=I+1-i}^{I-2} f_j + f_{I-1}^2 \mathrm{Var}(C_{i,I-1} \parallel C_{i,I+1-i}) \,,
\end{aligned}$$

which is a linear recurrence equation of the form

$$x_n = a_{n-1} x_{n-1} + g_{n-1}$$

with $x_n = \mathrm{Var}(C_{in} \parallel C_{i,I+1-i})$ and

$$g_{n-1} = \sigma_{n-1}^2 C_{i,I+1-i} \prod_{j=I+1-i}^{n-1} f_j \,, \qquad a_{n-1} = f_{n-1}^2 \,.$$

The general solution is given by

$$x_n = \left(\prod_{j=n_0}^{n-1} a_j \right) \left(x_{n_0} + \sum_{k=n_0}^{n-1} \frac{g_k}{\prod_{l=n_0}^{k} a_l} \right)$$

where n_0 denotes the first index of the sequence x_n, in our case $I+1-i$. Using the initial condition $x_{I+1-i} = \mathrm{Var}(C_{i,I+1-i} \parallel C_{i,I+1-i}) = 0$, we finally obtain

$$\begin{aligned}
\mathrm{Var}(C_{iI} \parallel \mathcal{D}_I) &= \left(\prod_{j=I+1-i}^{I-1} f_j^2 \right) \left(\sum_{k=I+1-i}^{I-1} \frac{\sigma_k^2 C_{i,I+1-i} \prod_{j=I+1-i}^{k-1} f_j}{\prod_{j=I+1-i}^{k} f_j^2} \right) \\
&= \left(\prod_{j=I+1-i}^{I-1} f_j^2 \right) C_{i,I+1-i}^2 \left(\sum_{k=I+1-i}^{I-1} \frac{\sigma_k^2 / f_k^2}{\prod_{j=I+1-i}^{k-1} f_j C_{i,I+1-i}} \right) \\
&= \mathbb{E}[C_{iI} \parallel C_{I+1-i}]^2 \sum_{k=I+1-i}^{I-1} \frac{\sigma_k^2 / f_k^2}{\mathbb{E}[C_{ik} \parallel C_{i,I+1-i}]} \,,
\end{aligned}$$

which we can estimate by plugging in $\widehat{f}_j$ and $\widehat{\sigma}_j$ for f_j and σ_j, respectively.

For the parameter error, if we use the definitions from the previous section to rewrite it as

$$\begin{aligned}
(\mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \widehat{C}_{iI})^2 &= C_{i,I+1-i}^2 \left(\prod_{j=I+1-i}^{I-1} f_j - \prod_{j=I+1-i}^{I-1} \widehat{f}_j \right)^2 \\
&= C_{i,I+1-i}^2 \left(\prod_{j=I+1-i}^{I-1} f_j^2 + \prod_{j=I+1-i}^{I-1} \widehat{f}_j^2 - 2 \prod_{j=I+1-i}^{I-1} f_j \widehat{f}_j \right) \,,
\end{aligned} \tag{2.5}$$

it becomes clear that things are more complicated than with process error. Indeed, we cannot simply subtitute the $\widehat{f}_j$ for the unknown parameters in this expression as that would cause it to vanish, yielding an estimate which will generally not be accurate. This problem was recognised by Mack himself in [7], and is caused by the fact that the claims triangle observations are used for both estimation and forecasting (see [8, Section 2] for a more general discussion). His suggested solution was to apply some kind of conditional averaging to the $\widehat{f}_j$. Ideally, one would like to condition on all available observations in $\mathcal{D}_I$, but the $\mathcal{D}_I$-measurability of the $\widehat{f}_j$ would then bring us right back where we started. We must therefore use a smaller set in order to allow $\widehat{f}_{I+1-i}, \dots, \widehat{f}_{I-1}$ to fluctuate around $f_{I+1-i}, \dots, f_{I-1}$. This corresponds to asking which other values $\widehat{f}_j$ could have taken, given that we fix a certain subset of the data—in other words, it's a resampling scheme on the parameter estimates. Thus, one can obtain an estimate of the parameter error by specifying a mechanism for generating new realisations of $\widehat{f}_j$ (see [9], [4, pp. 44 sqq.]), with different mechanisms yielding different estimates. While the literature uses this mostly as a theoretical device to facilitate analytical calculations, there is no reason why it could not be employed as a basis for bootstrapping. In the remainder of this section, we outline two approaches for estimating (2.5) and indicate the corresponding bootstrap methods.

Denote the subset of observations in $\mathcal{D}_I$ up to and including development year k by

$$\mathcal{B}_k \coloneqq \{C_{ij} \in \mathcal{D}_I \mid j \leq k\} \,,$$

and write

$$\mathcal{D}^O_{I,k} \coloneqq \{C_{ij} \in \mathcal{D}_I \mid j > I + 1 - k\}$$

for its complement. One option would then be to take the conditional expectation of $\widehat{f}_j$ with respect to $\mathcal{B}_{I+1-i}$, leading to the estimate

$$\begin{aligned}
\mathbb{E}[(\mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \widehat{C}_{iI})^2 \parallel \mathcal{B}_{I+1-i}] &= C^2_{i,I+1-i}\mathbb{E}\left[\left(\prod_{j=I+1-i}^{I-1} f_j - \prod_{j=I+1-i}^{I-1} \widehat{f}_j\right)^2 \,\middle\|\, \mathcal{B}_{I+1-i}\right] \\
&= C^2_{i,I+1-i}\left(\mathbb{E}\left[\prod_{j=I+1-i}^{I-1} \widehat{f}_j^2 \,\middle\|\, \mathcal{B}_{I+1-i}\right] - \prod_{j=I+1-i}^{I-1} f_j^2\right) \,,
\end{aligned}$$

where we used the fact that the $\widehat{f}_j$ are uncorrelated. This corresponds to averaging over the distribution of $\mathcal{D}^O_{I,k}$, or, expressed in terms of resampling, to generating new observations in the upper right triangle. Borrowing the nomenclature from [9], we call this the *unconditional approach*. Alternatively, we could average each $\widehat{f}_j$ only over the observations after j. This is equivalent to fixing the denominator $\sum_{i=1}^{I-j} C_{ij}$ in the development factor estimator (2.3) and allowing the numerator $\sum_{i=1}^{I-j} C_{i,j+1}$ to vary. Formally, it corresponds to taking the expectation with respect to the probability measure defined on $\mathcal{D}^O_{I,i}$ by

$$\mathbb{P}^*_{\mathcal{D}_I}(\{dz_{ij}\}_{i+j\leq I+1}) \coloneqq \prod_{j=1}^{I-1}\prod_{i=1}^{I-j} \mathbb{P}_{C_{i,j+1}}(dz_{i,j+1} \mid C_{ij} = c_{ij}) \,,$$

yielding the estimate

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^*_{\mathcal{D}_I}}\left[\left(\mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \widehat{C}_{iI}\right)^2\right] &= C_{iI}^2\, \mathbb{E}_{\mathbb{P}^*_{\mathcal{D}_I}}\left[\left(\prod_{j=I+1-i}^{I-1} f_j - \prod_{j=I+1-i}^{I-1} \widehat{f}_j\right)^2\right] \\ &= C_{i,I+1-i}^2 \left(\mathbb{E}_{\mathbb{P}^*_{\mathcal{D}_I}}\left[\prod_{j=I+1-i}^{I-1} \widehat{f}_j^2\right] - \prod_{j=I+1-i}^{I-1} f_j^2\right) \\ &= C_{i,I+1-i}^2 \left(\prod_{j=I+1-i}^{I-1} \mathbb{E}\left[\widehat{f}_j^2 \,\middle\|\, \mathcal{B}_j\right] - \prod_{j=I+1-i}^{I-1} f_j^2\right).\end{aligned}$$

We refer to this as the *conditional approach*, and it corresponds to a scheme in which only the observations from the next period are resampled to produce a new realisation of the parameter estimate for the current period.

There has been some controversy about which of these approaches should be preferred, leading to the vigorous discussion found in [9]–[12]. As we will see in Section 2.3.3, difference between the results which they produce is negligeable, and so the question is mainly of theoretical interest. Nevertheless, based on the previous exposition, it seems reasonable to prefer whichever method produces resampled parameter estimates approximating the original $\widehat{f}_j$ most closely. In particular, we note that these posses the following property, the proof of which can be found in [10].

Theorem 1. *The squares of two successive development factor estimates in the Mack chain ladder are negatively correlated:*

$$\operatorname{Cov}(\widehat{f}_j, \widehat{f}_{j-1}) < 0\,.$$

In the conditional approach, the resampled parameter estimates are independent by construction, and so they cannot incorporate this covariance structure. In light of this, it would appear that the unconditional scheme has slightly better theoretical properties. As the empirical difference between the two is minimal, however, the conditional version is a reasonable approximation to fall back on if necessary. In the next section, we will see how both approaches give rise to a variety of different bootstrap methods.

2.3.2 Bootstrap methodology

Using the taxonomy from Section 2.2.2, we now consider in turn the application of the semiparameteric, nonparametric and parametric type to Model 1. For the semiparametric bootstrap, the crucial step is to find a suitable definition for the residuals which ensures that they are interchangeable. Here the view we outlined in Section 2.3 of the Mack CL as a series of weighted linear regressions will be useful, because it gives us access to the results of classical regression theory. The distribution-free nature of the model does lead to one additional complication, however: it means that we cannot make precise statements about the distribution of the errors. This can be resolved in one of two ways. A first option would be to extrapolate from homogeneity of the first two moments to homogeneity of the distributions. In that case, the *raw residuals*

$$e_i \coloneqq C_{i,j+1} - \widehat{C}_{i,j+1} = C_{i,j+1} - \widehat{f}_j C_{ij}$$

are not an option, as these suffer from heteroscedasticity,

$$\operatorname{Var}(e_i \parallel C_{ij}) = \sigma_j^2 \left(C_{ij} - \frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}\right).$$

Hence we could address this by dividing out this variance, i.e. we consider the errors

$$\varepsilon_{i,j+1} \coloneqq \frac{(F_{i,j+1} - f_j)\sqrt{C_{i,j}}}{\sigma_j \sqrt{1 - \frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}}},$$

which satisfy $\mathbb{E}[\varepsilon_{i,j+1}] = 0$ and $\mathrm{Var}(\varepsilon_{i,j+1}) = 1$. Provided the sampling variability of $\widehat{f}_j$ and $\widehat{\sigma}_j$ is not too bad (which is not obvious given the small sample sizes we're usually dealing with), the same should hold approximately for the corresponding residuals

$$r_{i,j+1} \coloneqq \frac{(F_{i,j+1} - \widehat{f}_j)\sqrt{C_{i,j}}}{\widehat{\sigma}_j \sqrt{1 - \frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}}},$$

obtained by substituting these estimators. We emphasize that such additional assumptions should not be made lightly: it is perfectly possible for the error distribution to exhibit heterogeneity in other ways than through its mean and variance (see [1, p. 114] for an example where the *percentiles* vary with the value of the regressor).

Alternatively, we could choose to augment our model with some distributional assumptions, allowing us to make more precise statements about errors and residuals. In particular, we will consider the autoregressive Gaussian time series model

$$C_{i,j+1} = f_j C_{ij} + \sigma_j \sqrt{C_{ij}}\, \varepsilon_{i,j}, \qquad \varepsilon \sim \mathcal{N}(0,1)\,, \tag{2.6}$$

which can easily be seen to be compatible with Model 1. Recallling that the hat matrix corresponding to the regression of column $j+1$ on column j is given by

$$\begin{aligned}\mathbf{H} &= \mathbf{c_j}(\mathbf{c_j}^T \mathbf{W} \mathbf{c_j})^{-1} \mathbf{c_j}^T \mathbf{W} \\ &= \frac{1}{\sum_{i=1}^{I-j} C_{ij}} \begin{bmatrix} C_{1j} & \dots & C_{1j} \\ \vdots & \ddots & \vdots \\ C_{I-j,j} & \dots & C_{I-j,j} \end{bmatrix},\end{aligned}$$

we know, for example, that the *externally studentised residuals*

$$r_{i,j+1} \coloneqq \frac{e_i}{\widehat{\sigma}_{j(i)}\sqrt{1 - \mathbf{H}_{ii}}} \sqrt{\mathbf{W}_{ii}} = \frac{(F_{i,j+1} - \widehat{f}_j)\sqrt{C_{i,j}}}{\widehat{\sigma}_{j(i)} \sqrt{1 - \frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}}},$$

with $\widehat{\sigma}_{j(i)}$ denoting the leave-i-out estimator of σ_j, follow a t_{I-j-1} distribution. Another option are the *standardised* or *internally studentised* residuals

$$r_{i,j+1} \coloneqq \frac{(F_{i,j+1} - \widehat{f}_j)\sqrt{C_{i,j}}}{\widehat{\sigma}_j \sqrt{1 - \frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}}}$$

which also share the same distribution, albeit a more complicated one (see [13, pp. 267 sqq.]). Note that all residuals can be written in the form $\delta(C_{ij}, \hat{C}_{ij})$ for an appropriate δ.

For any particular choice of residuals, the algorithm then proceeds in the following way. First, we resample the residuals to obtain pseudo-realisations $r_{ij}^{(b)}$. Under the conditional approach, bootstrapped claims triangle variables are then obtained by inverting the appropriate formula, which can be stated in general as solving

$$\delta$$

Algorithm 1 Semiparametric resampling for Mack CL

Input: Cumulative claims triangle $\mathcal{D}_I$, required number of bootstrap samples B, parameter CONDITIONAL specifying the resampling approach
for $b \leftarrow 1, B$ **do**
for $j \leftarrow 1, I-1$ **do**
$r_{i,j+1} \leftarrow \delta(C_{i,j+1}, \hat{C}_{i,j+1})$
$\widehat{f}_j^{(b)} \leftarrow \sum_{i=1}^{I-j} C_{i,j+1}^{(b)} / \sum_{i=1}^{I-j} C_{ij}^{(b)}$
for $i \leftarrow 1, I-j$ **do**
$F_{i,j+1}^{(b)} \leftarrow C_{i,j+1}^{(b)} / C_{ij}^{(b)}$
end for
$\widehat{\sigma}_j^{(b)} \leftarrow \dfrac{1}{I-j} \displaystyle\sum_{i=1}^{I-j} C_{ij}^{(b)} \left(F_{i,j+1}^{(b)} - \widehat{f}_j^{(b)} \right)^2$
end for
end for
return $\{(\widehat{f}_j^{(b)}, \widehat{\sigma}_j^{(b)}) \mid b = 1, \ldots, B\}$

The fully nonparametric bootstrap requires us to resample the pairs $(C_{ij}, C_{i,j+1})$ in the regression of the. We then refit the model to obtain bootstrapped development factor and dispersion parameter estimates.

The only additional remark we must make here, is that it is not possible to apply the unconditional approach to this type of bootstrap =¿ YES WE CAN!!!!

Algorithm 2 Pairs resampling for Mack CL (conditional)

Input: Cumulative claims triangle $\mathcal{D}_I$, required number of bootstrap samples B
for $b \leftarrow 1, B$ **do**
for $j \leftarrow 1, I-1$ **do**
$\{(C_{i,j}^{(b)}, C_{i,j+1}^{(b)}) \mid i = 1, \ldots, I-j\} \leftarrow$ RESAMPLE$(\{(C_{i,j}, C_{i,j+1}) \mid i = 1, \ldots, I-j\})$
$\widehat{f}_j^{(b)} \leftarrow \sum_{i=1}^{I-j} C_{i,j+1}^{(b)} / \sum_{i=1}^{I-j} C_{ij}^{(b)}$
for $i \leftarrow 1, I-j$ **do**
$F_{i,j+1}^{(b)} \leftarrow C_{i,j+1}^{(b)} / C_{ij}^{(b)}$
end for
$\widehat{\sigma}_j^{(b)} \leftarrow \dfrac{1}{I-j} \displaystyle\sum_{i=1}^{I-j} C_{ij}^{(b)} \left(F_{i,j+1}^{(b)} - \widehat{f}_j^{(b)} \right)^2$
end for
end for
return $\{(\widehat{f}_j^{(b)}, \widehat{\sigma}_j^{(b)}) \mid b = 1, \ldots, B\}$

Finally, we discuss how to incorporate the process error in our bootstrap, which will be done in the same way for all the different types.

$$\widehat{\sigma}_j^{(b)} = \frac{1}{I-j} \sum_{i=1}^{I-j} C_{ij}^{(b)} \left(F_{ij}^{(b)} - \widehat{f}_j^{(b)} \right)^2$$

We then use these to simulate lower pseudo-triangles $(\mathcal{D}_I^{\mathsf{c}})^{(b)} = (C_{ij}^{(b)})$, which in turn yield simulated reserve samples

$$R^{(b)} \coloneqq \sum_{i=2}^{I} (C_{iI}^{(b)} - C_{i,I+1-i}) \,.$$

Notice that this allows us to bypass (2.4) completely, altough we still expect that

$$\frac{1}{N} \sum_{k=1}^{N} R_k^{(b)} \approx \widehat{R} \,.$$

In view of (2.6), the most evident way to generate $(\mathcal{D}_I^{\mathsf{c}})^{(b)}$ is to start from the antidiagonal in $\mathcal{D}_I$ and successively sample

$$C_{i,j+1}^{(b)} \sim \mathcal{N}(\widehat{f}_j^{(b)} \, C_{ij}, \widehat{\sigma}_j^{(b)} \, C_{ij}) \,. \tag{2.7}$$

major drawback, however: it makes it possible to draw negative samples for the cumulative loss amounts. One way to remedy this would obviously be to simply discard a simulated triangle as soon as it contains a negative value.

An alternative is to follow the suggestion given in [6, p. 238] and subtitute in place of (2.7) a gamma distribution with the same mean and variance. If we write $C_{ij} \sim \Gamma(\alpha, \beta)$, this means that α, β must satisfy

$$\frac{\alpha}{\beta} = f_{j-1} C_{i,j-1} \quad \text{and} \quad \frac{\alpha}{\beta^2} = \sigma_{j-1}^2 C_{i,j-1} \,,$$

from which it follows that

$$\alpha = \frac{f_{j-1}^2 C_{i,j-1}}{\sigma_{j-1}^2} \quad \text{and} \quad \beta = \frac{f_{j-1}}{\sigma_{j-1}^2} \,.$$

This ensures that Mack's assumptions are still valid while avoiding nonsensical outcomes.

2.3.3 Numerical implementation and results

We now move on to the second model which we consider in this chapter.

2.4 Overdispersed Poisson model

The (overdispersed) Poisson model (ODP), proposed by Renshaw and Verrall in [14], belongs to the family of so-called *generalised linear models* (GLM). In contrast to the Mack CL, it describes the incremental claims X_{ij}. As the concept of overdispersion is explained in Section 2.4.1, we only state the assumptions of the ordinary variant at this point.

Model 2 (Poisson GLM).

1. *The incremental claims are independent from each other.*
2. *There exist parameters* c, $a_1, \ldots, a_I$ *and* $b_1, \ldots, b_I$ *such that*

$$\log(\mathbb{E}[X_{ij}]) = c + a_i + b_j \,, \tag{2.8}$$

   with $a_1 = b_1 = 0$.

3. *The incremental claims follow a Poisson distribution with* $\mu_{ij} = \mathbb{E}[X_{ij}]$*:*

$$X_{ij} \sim \text{Pois}(e^{c+a_i+b_j}) \,.$$

The condition $a_1 = b_1 = 0$ is necessary to obtain an identifiable model. Without it, any set of parameters $c, a_1, \ldots, a_I, b_1, \ldots, b_I$ satisfying the assumptions would yield an infinite number of alternatives $c + a_0 + b_0, a_1 - a_0, \ldots, a_I - a_0, b_1 - b_0, \ldots, b_I - b_0$ for $a_0, b_0 \in \mathbb{R}$. We can therefore see that we have two superfluous degrees of freedom, which we can get rid of by imposing two conditions on the parameters.

By defining $\xi_i \coloneqq e^{c+a_i}$ and $\gamma_j \coloneqq e^{b_j}$, we can obtain a different parametrisation of the model with a multiplicative structure for the mean,

$$\mathbb{E}[X_{ij}] = \xi_i \gamma_j \,,$$

which is often preferred to the previous one for reasons of interpretability. Indeed, it is clear that the multiplicative form has one fewer degree of freedom than the linear one, and if we remove it by imposing the constraint

$$\sum_{j=1}^{I} \gamma_j = 1$$

then we can view the ξ_i as expected ultimate claim amounts, and the γ_j as the expected development pattern.

As mentioned in the introduction, stochastic claims reserving models have to reproduce the chain ladder point predictions in order to be acceptable to practitioners. While less obvious than for the Mack CL, it can be shown that the Poisson model also satisfies this requirement (see [4, Lemma 2.16]).

In the next section, we will give a general overview.

2.4.1 Generalised linear models

GLMs were first conceived by Nelder and Wedderburn in [15] as a way of unifying the many disparate generalisations of linear regression with Gaussian errors which were then in existence. These sought to extend the classical model by allowing the use of different functional forms for the conditional mean and different distributions for the response, thus making it suited to modelling counts data (Poisson regression) or the probability of binary events (logistic regression), among others. For a set of covariates $X_1, \ldots, X_p$ and a response variable Y, a GLM consists of three parts:

1. The *random component*, a distribution for response Y belonging to the so-called *exponential dispersion model* family (EDM), which consists of all probability distributions whose density (with respect either to the Lebesgue or counting measure) has the form

$$p(y \mid \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\} , \tag{2.9}$$

   where a, b and c are known functions, and b is at least twice differentiable. We call θ the *canonical parameter* of the distribution and ϕ the *dispersion parameter*.

2. The *systematic component*, a predictor $\eta \coloneqq \mathbf{x}^T \boldsymbol{\beta}$ which is a linear function of the covariates.

3. A monotonic differentiable link function $g : \mathbb{R} \to \mathbb{R}$ giving the relation between the conditional expectation and the linear predictor,

$$\mu \coloneqq \mathbb{E}[Y \parallel X_1, \ldots, X_p] = g^{-1}(\eta) \,.$$

The Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ can be seen to belong to the EDM family by rewriting its density as

$$\begin{aligned}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\} &= \exp\left\{-\frac{y^2}{2\sigma^2}+\frac{y\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}-\log(\sqrt{2\pi}\sigma)\right\}\\ &= \exp\left\{\frac{y\mu-\mu^2/2}{\sigma^2}-\frac{y^2}{2\sigma^2}-\log(\sqrt{2\pi}\sigma)\right\},\end{aligned}$$

which is of the form (2.9) with $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $\phi = \sigma^2$, $a(\phi) = \phi$ and $c(y, \sigma) = -\frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)$. Thus, the familiar normal linear model can be obtained from the GLM framework with response distribution $\mathcal{N}(\mu, \sigma^2)$ and identity link $g(\mu) = \mu$.

The EDM family has a number of properties which greatly facilitate the computations involved in estimation. Recall from likelihood theory that $l(\theta \mid y, \phi) \coloneqq \log p(y \mid \theta, \phi)$ satisfies

$$\mathbb{E}\left[\frac{\partial l(\theta \mid Y)}{\partial\theta}\right] = 0\,, \qquad \mathrm{Var}\left(\frac{\partial l(\theta \mid Y)}{\partial\theta}\right) = -\mathbb{E}\left[\frac{\partial^2 l(\theta \mid Y)}{\theta^2}\right], \tag{2.10}$$

where $\frac{\partial l(\theta|Y)}{\partial\theta}$ is known as the *score function*. Using (2.9), we then find that

$$\mathbb{E}\left[\frac{Y-b'(\theta)}{a(\phi)}\right] = 0\,, \qquad \mathrm{Var}\left(\frac{Y-b'(\theta)}{a(\phi)}\right) = -\mathbb{E}\left[\frac{-b''(\theta)}{a(\phi)}\right],$$

from which we obtain the elegant relations

$$\mu = b'(\theta)\,, \qquad \mathrm{Var}(Y) = a(\phi)b''(\theta)\,.$$

Observe that this implies that $\frac{d\mu}{d\theta} = b''(\theta) > 0$ (because the variance is always positive), which means that $\theta \mapsto \mu(\theta)$ is one-to-one and therefore invertible. In particular, we can always write the likelihood as function of the mean. The function $V(\mu) \coloneqq b''((b')^{-1}(\mu))$ is called the *variance function* and determines how the scale of the response varies as a function of its mean.

Special care has to be taken with the parameter ϕ, as it occupies a rather awkward position in GLM theory. The trouble is that we want to incorporate two-parameter distributions, such as the normal and gamma distribution, into the GLM framework which can fundamentally only handle a single parameter gracefully (the more flexible framework of *vector GLMs* is an attempt to remedy this; see [16, Chapter 2] for a general discussion). The dispersion is therefore regelated to the role of nuisance parameter and subjected to severe (and often unrealistic) constraints. Basically, we ϕ to be the constant as a function of the covariates, but this would preclude certain special cases such as binomial regression with a different number of trials for each observation in the sample. To take this into account, we allow the function a in the denominator of (2.9) to vary across different sample responses as $a_i(\phi) = \phi/w_i$, where w_i is a known weight. Not a very elegant solution, perhaps, but one which is foisted upon us by the limitations of the theory. The parameter ϕ itself is then considered as known, and estimated outside of the GLM framework, most commonly using the Pearson statistic

$$\widehat{\phi} \coloneqq \frac{1}{n-p}\sum_{i=1}^{N}\frac{(Y_i-\widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}\,.$$

Given a sample $(\mathbf{x_1}, Y_1), \ldots, (\mathbf{x_N}, Y_N)$, the standard way to fit a GLM is by means of maximum likelihood estimation (MLE). The joint log-likelihood of the sample is given by

$$l(\boldsymbol{\beta} \mid \mathbf{y}, \phi) = \sum_{i=1}^{N}\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\,,$$

which we must differentiate with respect to β_j to obtain the likelihood equations. An application of the chain rule gives us

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta} \mid \mathbf{y}, \phi)}{\partial \beta_j} &= \sum_{i=1}^{N} \frac{\partial l(\boldsymbol{\beta} \mid y_i, \phi)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_{i=1}^{N} \frac{y_i - b_i'(\theta)}{a_i(\phi)} \frac{1}{b_i''(\theta_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\ &= \sum_{i=1}^{N} \frac{y_i - \mu_i}{\operatorname{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}\,, \end{aligned} \tag{2.11}$$

and setting this equal to 0 yields a system of p (usually nonlinear) equations. It is generally impossible to solve these analytically, and so we must resort to numerical methods. In particular, we use a modified version of the Newton-Raphson algorithm known as *Fisher scoring*, which replaces the negative Hessian of the log-likelihood, called the *observed information*, by its expectation

$$\mathcal{I}_{jk} \coloneqq \mathbb{E}\left[-\frac{\partial^2 l(\boldsymbol{\beta} \mid \mathbf{y}, \phi)}{\partial \beta_j \partial \beta_k} \right],$$

which is known as the *Fisher information matrix*. Thus, starting from an initial guess $\boldsymbol{\beta}^{(0)}$ for the parameters, we compute a sequence of approximations, where each new value obtained from the previous one via

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \mathcal{I}(\boldsymbol{\beta}^{(k)})^{-1} \nabla l(\boldsymbol{\beta} \mid \mathbf{y}, \phi)\,. \tag{2.12}$$

Similarly to (2.10), it can be shown that

$$\mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\beta} \mid \mathbf{y}, \phi)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right] = -\operatorname{Var}\big(\nabla l(\boldsymbol{\beta} \mid \mathbf{y}, \phi) \nabla l(\boldsymbol{\beta} \mid \mathbf{y}, \phi)^T \big) < 0\,,$$

from which we also see that the log-likelihood is concave, and will therefore have a global maximum. Using the fact that the Y_i are independent, so that $\mathbb{E}[(Y_i - \mu_i)(Y_l - \mu_l)] = 0$ for $i \neq l$, we then obtain

$$\begin{aligned} I_{jk} &= \mathbb{E}\left[\left(\sum_{i=1}^{N} \frac{Y_i - \mu_i}{\operatorname{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \right) \left(\sum_{l=1}^{N} \frac{Y_l - \mu_l}{\operatorname{Var}(Y_l)} \frac{\partial \mu_l}{\partial \eta_l} x_{lk} \right) \right] \\ &= \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \frac{\mathbb{E}\big[(Y_i - \mu_i)^2\big]}{\operatorname{Var}(Y_i)^2} x_{ij} x_{ik} \\ &= \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \frac{x_{ij} x_{ik}}{\operatorname{Var}(Y_i)} \\ &= \mathbf{x}_{\boldsymbol{j}}^T \mathbf{W} \mathbf{x}_{\boldsymbol{k}} \end{aligned} \tag{2.13}$$

where $\mathbf{W}$ is a diagonal matrix with

$$\mathbf{W}_{ii} = \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \frac{1}{\operatorname{Var}(Y_i)}\,. \tag{2.14}$$

Hence we have $\mathcal{I} = \mathbf{X}^T \mathbf{W} \mathbf{X}$ and we see from (2.11) that

$$\nabla l(\boldsymbol{\beta} \mid \mathbf{y}, \phi) = \mathbf{X}^T \mathbf{W} \tilde{\mathbf{z}} \tag{2.15}$$

with $\tilde{\mathbf{z}}_i = (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$. Multiplying both sides of 2.12 by $\mathcal{I}(\boldsymbol{\beta}^{(k)})$ and using (2.13) to (2.15), we finally obtain

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}^{(k+1)} = \mathbf{X}^T\mathbf{W}\mathbf{z}$$

with $\mathbf{z} = \mathbf{X}\boldsymbol{\beta}^{(k)} + \tilde{\mathbf{z}}$ and all quantities evaluated at the current estimate $\boldsymbol{\beta}^{(k)}$ of the parameter vector. In other words, the Fisher scoring is equivalent to a series of weighted least squares problems, where the new parameter estimates are obtained by regressing the vector $\mathbf{z}$ on the original covariates $\mathbf{x_1}, \ldots, \mathbf{x_N}$ using weight matrix $\mathbf{W}$, and $\mathbf{z}$ and $\mathbf{W}$ are determined by the current estimate $\boldsymbol{\beta}^{(k)}$—hence why the algorithm is called *iteratively reweighted least squares* (IRWLS).

This procedure can be specialised to the particular case of Model 2 in the following way. First, in order to obtain the matrix-vector form used above, we must flatten the tabular response (using, for example, the colexicographical ordering $(i, j) \mapsto jI + i$, i.e. column-major order). If we define the parameter vector

$$\boldsymbol{\beta} \coloneqq \begin{bmatrix} c & a_2 & \cdots & a_I & b_2 & \cdots & b_I \end{bmatrix}^T ,$$

then (2.8) can be rewritten as

$$\log(\mu_{ij}) = \mu + a_i + b_j = (\mathbf{e_1} + \mathbf{e_i} + \mathbf{e_{I+j-1}})^T\boldsymbol{\beta}$$

where $\mathbf{e_k}$ denotes the kth standard basis vector in $\mathbb{R}^{(2I-1)}$. Hence we see that the covariates are binary vectors of length $2I - 1$, with the position of the nonzero entries determined by the indices of the observation in the triangle, forming the rows of a very sparse design matrix. As the Poisson model uses the log link, we have $\mu_{ij} = e^{\eta_{ij}}$ and

$$\frac{\partial \mu_{ij}}{\partial \eta_{ij}} = e^{\eta_{ij}} ,$$

from which, using (2.14), we finally obtain

$$\mathbf{W}_{ii} = \frac{1}{e^{\eta_{ij}}}(e^{\eta_{ij}})^2 = e^{\eta_{ij}} ,$$

giving us all the required components of the IRWLS algorithm.

We have assumed up to this point that a GLM requires us to specify an exact distribution for the response variable. In many practical situations, however, this is either infeasible or leads to unrealistic models. An example which is particularly common with count data is a phenomenon known as *overdispersion*, where the variability of the data is greater than would be suggested by e.g. the Poisson or binomial distribution. Recall that the variance of a Pois(λ) distribution is λ, and that of a $B(n, p)$ distribution is $np(1-p)$; in both cases, it is fully determined by the mean, and we have no degree of freedom with which to adjust it in order to obtain a better fit to the data, as would be the case with the normal distribution, for example.

To remedy this, an extension can be made to the GLM framework, which only relies on the specification of a relation between mean and variance. Recall from above that the MLE works by setting the score equal to 0. If we write the likelihood in terms of $\boldsymbol{\mu}$, this will have components

$$\frac{\partial l(\boldsymbol{\mu} \mid \mathbf{y}, \phi)}{\partial \mu_j} = \frac{y_j - \mu_j}{\phi V(\mu_j)} , \tag{2.16}$$

and is therefore completely determined by $V(\cdot)$. Suppose now, conversely, that we start from $V(\cdot)$. We could then define functions

$$Q_i(\mu \mid y_i, \phi) \coloneqq \int_{y_i}^{\mu} \frac{y_i - u}{\phi V(u)}\, du\,,$$

and estimate $\boldsymbol{\mu}$ (and therefore $\boldsymbol{\beta}$) by minimising

$$Q(\boldsymbol{\mu} \mid \mathbf{y}, \phi) \coloneqq \sum_{i=1}^{N} Q_i(\mu_i \mid y_i, \phi)\,.$$

It must be stressed that Q has no probabilistic significance: it does not, in general, correspond to the log-likelihood of any distribution. Rather, it functions a device to obtain estimates of the desired parameters, fulfilling in this a similar role to that of the log-likelihood, which is why we refer to it as a *quasi-likelihood function*[2]. It is usual to identify quasi-likelihood models derived from specific distributions (i.e. using the corresponding variance function) by prefixing 'quasi' to the name of said distribution, e.g. quasi-Poisson or quasi-binomial. The derived quasi-model yields the same parameter estimates as the classical GLM if the data follow the original distribution.

We are now finally in a position to describe the overdispersed variant of the Poisson GLM.

Model 3 (Overdispersed quasi-Poisson).

1. *The incremental claims are independent from each other.*
2. *There exist parameters* c, $a_1, \dots, a_I$ *and* $b_1, \dots, b_I$ *such that*

$$\log(\mu_{ij}) = c + a_i + b_j\,,$$

   with $\mu_{ij} \coloneqq \mathbb{E}[X_{ij}]$ *and* $a_1 = b_1 = 0$.
3. *There exists a parameter* ϕ *such that*

$$\operatorname{Var}(X_{ij}) = \phi \mu_{ij}\,.$$

When the data consists entirely of positive integers (e.g. a triangle of claims counts), it follows from the previous remark that this model yields the same predictions as the chain ladder. More generally, the CL results will be reproduced as long as the additional condition

$$\sum_{i=1}^{I} X_{ij} \geq 0$$

is satisfied for $j \in \{1, \dots, I\}$ (see [14, Section 2]). The quasi-Poisson is therefore robust to the presence of a limited number of negative claim amounts, which is sometimes observed in practice. Moreover, it lifts the unrealistic restriction that the response values must be integers, and gives us a way of accounting for overdispersion, which is a feature of many claims triangles. The absence of a likelihood also poses some difficulties, however, notably in the area of inference and diagnostics.

[2]Strictly speaking, it would be more correct to call Q a quasi-*log*-likelihood, but the current nomenclature has been widely adopted and the literature seems to have resigned itself to it.

2.4.2 Bootstrap methodology

Developing a bootstrap procedure for the Poisson and quasi-Poisson models is in some respects easier than for the Mack CL. The absence of a recursive structure makes it more straightforward to reason about resampling. Furthermore, bootstrap methods for claims triangle GLMs have seen more discussion in the literature (see e.g. [17] and [6]), and so we can draw upon this material for our exposition. As with Mack's model, we shall take Section 2.2 as starting point. Recall that we made a distinction there between nonparametric, semiparametric and parametric approaches to bootstrapping, depending on the assumptions which they make. We will consider each of these in turn, and see how they can be applied to the model under consideration.

For the nonparameteric bootstrap, the essential step is to find a satisfactory definition for the residuals such that they are i.i.d. Things are more complicated here than for Mack's model, as there generally exists no natural seperation of the response into mean and additive error for a non-Gaussian response (this problem was recognised early on in the literature on GLM bootstrapping, see [18]). Consequently, a multitude of different residual types are available. We will consider three of these in particular.

The *Pearson residuals*

$$r_{ij} \coloneqq \frac{X_{ij} - \widehat{\mu}_{ij}}{\sqrt{V(\widehat{\mu}_{ij})}}\,, \tag{2.17}$$

attempt to deal with the inherent heteroscedasticity of the GLM response by dividing out the component of the variance which is specific to each observation. In this, they resemble the standardised residuals in the context of weighted linear regression. Extending this analogy further, we can adjust (2.17) for the leverage of the observation, i.e.

$$\tilde{r}_{ij} \coloneqq \frac{X_{ij} - \widehat{\mu}_{ij}}{\sqrt{V(\widehat{\mu}_{ij})(1-h_{ij})}}\,,$$

where h_{ij} is the appropriate diagonal element in the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$$

corresponding to the final iteration of the IRWLS algorithm.

Another kind of residuals are based on a goodness-of-fit measure for GLMs known as the *deviance*. It can be derived from (2.16) by noticing that the mean parametrisation of the log-likelihood is maximised at $\boldsymbol{\mu} = \mathbf{y}$, so that the quantity

$$D(\mathbf{y}, \boldsymbol{\mu}) \coloneqq \sum_{i=1}^{N} d(y_i, \mu_i) \coloneqq 2\sum_{i=1}^{N}(l(y_i \mid y_i) - l(\widehat{\mu}_i \mid y_i))$$

expresses the departure of our model from a perfect fit. The functions $D(\mathbf{y}, \boldsymbol{\mu})$ and $d(y_i, \mu_i)$ are called the *total* and *unit deviance*, respectively. The *deviance residuals* are then defined as

$$r_{ij} \coloneqq \operatorname{sign}(x_{ij} - \mu_{ij})\sqrt{d(x_{ij}, \mu_{ij})}\,.$$

Finally, we consider a third type known as *quantile residuals* (see [19] for a general discussion), which are most easily explained for continuous response distributions. In that case, an elementary fact from probability theory states that

$$F(Y \mid \mu, \phi) \sim U(0,1)\,,$$

and it should therefore follow that the empirical distribution of the transformed sample

$$r_i \coloneqq F(Y_i \mid \widehat{\mu}_i, \widehat{\phi})$$

is approximately uniform, provided the sampling variability of $\widehat{\boldsymbol{\mu}}$ and $\widehat{\phi}$ is not too severe. If $F(\cdot \mid \boldsymbol{\mu}, \phi)$ is discrete, the definition is amended as follows: for every observation y_i, set

$$a_i \coloneqq \lim_{y \uparrow y_i} F(y \mid \widehat{\mu}_i, \widehat{\phi}), \qquad b_i \coloneqq F(y_i \mid \widehat{\mu}_i, \widehat{\phi}),$$

and define the r_i as mutually independent random variables which are uniformly distributed on $(a_i, b_i]$.

2.4.3 Numerical implementation and results

Conclusion

Bibliography

[1] B. Efron and R. Tibshirani, *An introduction to the bootstrap*. Boca Raton, Fla Chapman & Hall/Crc, 1998, ISBN: 9780412042317.

[2] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*. Cambridge University Press, 1997.

[3] T. Mack, "Distribution-free calculation of the standard error of chain ladder reserve estimates," *Astin Bulletin*, vol. 23, no. 2, 1993.

[4] M. V. Wüthrich and M. Merz, *Stochastic Claims Reserving Methods in Insurance*. John Wiley & Sons, 2008.

[5] F. Hayashi, *Econometrics*. Princeton, Nj ; Oxford: Princeton University Press, 2000, ISBN: 9780691010182.

[6] P. D. England and R. J. Verrall, "Predictive distributions of outstanding liabilities in general insurance," *Annals of Actuarial Science*, vol. 1, no. 1, 2006.

[7] T. Mack, "Measuring the variability of chain ladder reserve estimates," 1999.

[8] M. Lindholm, F. Lindskog, and F. Wahl, "Estimation of conditional mean squared error of prediction for claims reserving," *Annals of Actuarial Science*, vol. 14, no. 1, 93–128, 2020. DOI: 10.1017/S174849951900006X.

[9] M. V. Wüthrich, M. Buchwalder, H. Bühlmann, and M. Merz, "The mean square error of prediction in the chain ladder reserving method (mack and murphy revisited)," *Astin Bulletin*, vol. 36, no. 2, 2006.

[10] T. Mack, G. Quarg, and C. Braun, "The mean square error of prediction in the chain ladder reserving method – a comment," *ASTIN Bulletin: The Journal of the IAA*, vol. 36, no. 2, 543–552, 2006. DOI: 10.1017/S051503610001463X.

[11] A. Gisler, "The estimation error in the chain-ladder reserving method: A bayesian approach," *ASTIN Bulletin: The Journal of the IAA*, vol. 36, no. 2, 554–565, 2006. DOI: 10.1017/S0515036100014653.

[12] G. G. Venter, "Discussion of the mean square error of prediction in the chain ladder reserving method," *ASTIN Bulletin: The Journal of the IAA*, vol. 36, no. 2, 566–571, 2006. DOI: 10.1017/S0515036100014665.

[13] G. A. F. Seber and A. J. Lee, *Linear regression analysis*. Wiley-Interscience, 2003, ISBN: 9780471415404.

[14] A. Renshaw and R. Verrall, "A stochastic model underlying the chain-ladder technique," *British Actuarial Journal*, vol. 4, no. 4, 903–923, 1998. DOI: 10.1017/S1357321700000222.

[15] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972, ISSN: 00359238.

[16] T. W. Yee, *Vector Generalized Linear and Additive Models*. Springer, 2015, ISBN: 9781493928187.

[17] P. J. R. Pinheiro, J. M. A. e Silva, and M. de Lourdes Centeno, "Bootstrap methodology in claim reserving," *The Journal of Risk and Insurance*, vol. 70, no. 4, pp. 701–714, 2003, ISSN: 00224367, 15396975.

[18] L. H. Moulton and S. L. Zeger, "Bootstrapping generalized linear models," *Computational Statistics & Data Analysis*, vol. 11, no. 1, pp. 53–63, 1991, ISSN: 0167-9473. DOI: `https://doi.org/10.1016/0167-9473(91)90052-4`.

[19] P. K. Dunn and G. K. Smyth, "Randomized quantile residuals," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 236–244, 1996, ISSN: 10618600.