VRIJE
UNIVERSITEIT
BRUSSEL

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

SENSITIVITY ANALYSIS OF STOCHASTIC RESERVING MODELS USING BOOTSTRAP SIMULATIONS

Othman El Hammouchi

June 2023

Promotors: 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

GEVOELIGHEIDSANALYSE VAN STOCHASTISCHE SCHADERESERVERINGSMOD-ELLEN AAN DE HAND VAN BOOTSTRAP-SIMULATIES

Othman El Hammouchi

Juni 2023

Promotors: 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

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

$$\mathcal{D}_I = \{C_{ij} \mid 1 \leq j \leq J, i+j \leq I+1\} \quad \text{or} \quad \mathcal{D}_I = \{X_{ij} \mid 1 \leq j \leq J, i+j \leq I+1\} \ . \tag{1.1}$$

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*

$$R = \sum_{i=2}^{I} (C_{i,I} - C_{i,I+1-i}) \tag{1.2}$$

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

C_{11}	C_{12}	C_{13}	C_{14}	C_{15}
C_{21}	C_{22}	C_{23}	C_{24}	
C_{31}	C_{32}	C_{33}		
C_{41}	C_{42}			
C_{51}				

(a) Cumulative

X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_{21}	X_{22}	X_{23}	X_{24}	
X_{31}	X_{32}	X_{33}		
X_{41}	X_{42}			
X_{51}				

(b) Incremental

Table 1.1: General notation for a 5 by 5 claims triangle

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 has since been framed as a stochastic model in a variety of ways. 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, however, as it limits the efficacy of classical statistical techniques. In particular, it makes it difficult to detect structural breaks in the claims development pattern.

To illustrate this point, consider Table 1.2, which contains the dataset of cumulative payments for a motor insurance account from the UK given in [1] (this will serve as the running example throughout this text). It consists of a 7 by 7 claims triangle with a total of 28 observations. Figure 1.1 shows the diagnostic plot of standardised residuals against fitted value for the Mack chain ladder model, which will be discussed in Chapter 3. The details are not important at this point; all that matters at present is that it should be symmetric around the x-axis and exhibit no structural patterns if the model gives a good fit.

j i	1	2	3	4	5	6	7
2007	3511	6726	8992	10704	11763	12350	12690
2008	4001	7703	9981	11161	12117	12746	
2009	4355	8287	10233	11755	12993		
2010	4295	7750	9773	11093			
2011	4150	7897	10217				
2012	5102	9650					
2013	6283						

Table 1.2: UK Motor claims triangle from Christofides [1]

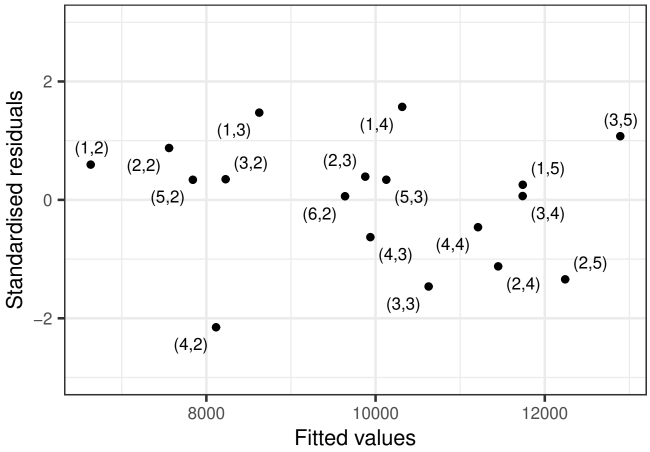
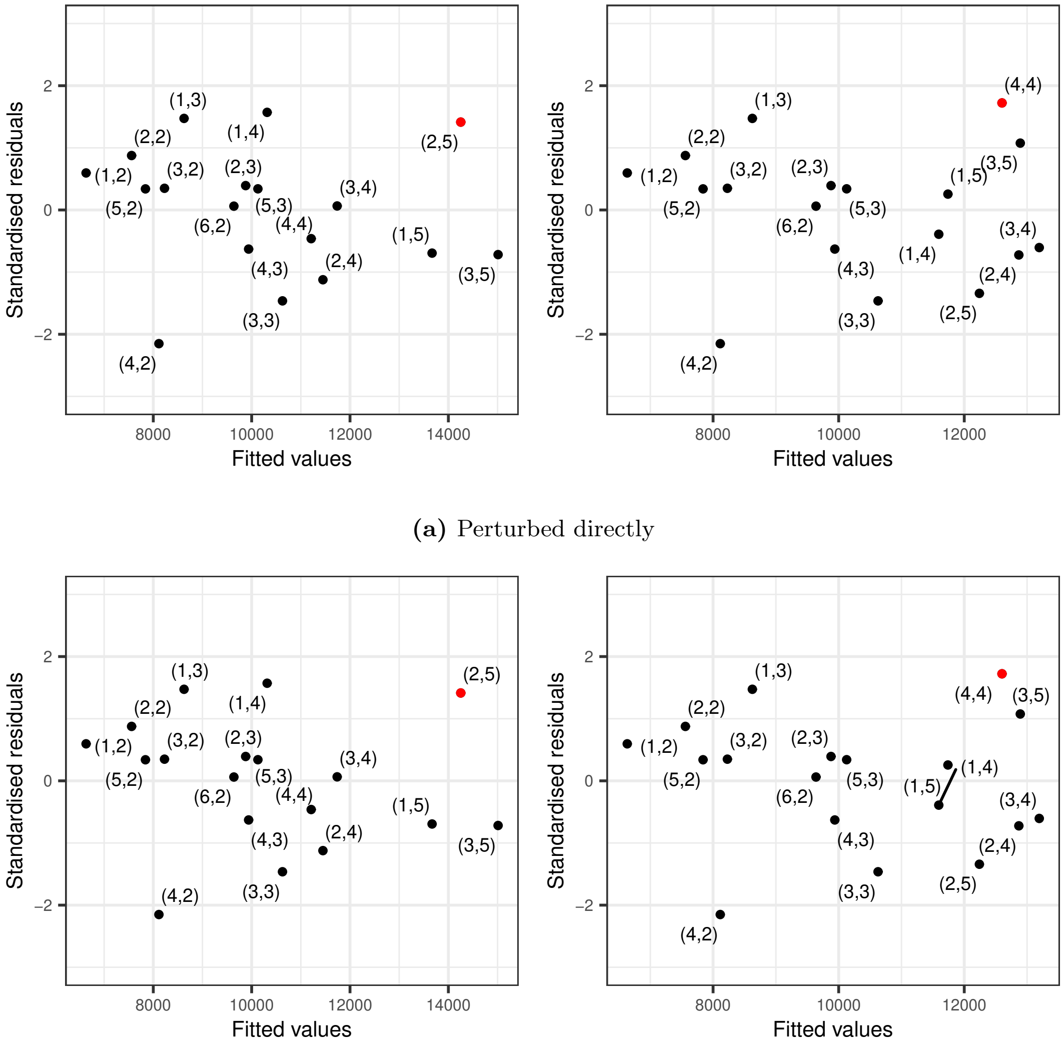


Figure 1.1: Diagnostic plot for original triangle

The same diagnostic plots are shown in Figure 1.2 for the case where the single points $(2, 5)$ and $(4, 4)$ of this triangle has been perturbed by a factor 1.5, either by direct multiplication or through simulation from the underlying model. The residual corresponding to the pathological observation has been highlighted in red. As these examples demonstrate, it is not always feasible to identify deviations from the model assumptions by examining such plots, even for the trained eye.



(a) Perturbed directly

(b) Perturbed according to model

Figure 1.2: Diagnostic plots for perturbed triangles

Our aim in this chapter is to investigate whether it is possible to use bootstrap simulations to remedy this problem. The chapter is divided as follows. The next section introduces the bootstrap and explains how it can be used for inference on regression models, which is how we will frame the reserving methods under consideration. In the subsequent two sections, we will then apply this theory to two of the most widespread stochastic claims reserving models, the Mack chain ladder and the (overdispersed) Poisson GLM, in order to study whether it is possible to identify pattern breaks by using the bootstrap. Specifically, we simulate claims triangles which perfectly follow the model assumptions, perturb these, and generate a bootstrap reserve from the resulting dataset. This will allow us to investigate how the simulated reserve is impacted by deviations from the model assumptions.

Chapter 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.

2.1 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*

$$\widehat{F}_n(x) := \sum_{k=1}^{n} I_{\{X_k \leq x\}} \, , \tag{2.1}$$

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

$$\frac{1}{B-1}\sum_{b=1}^{B}(g^{(b)}-\bar{g})^2\,, \tag{2.2}$$

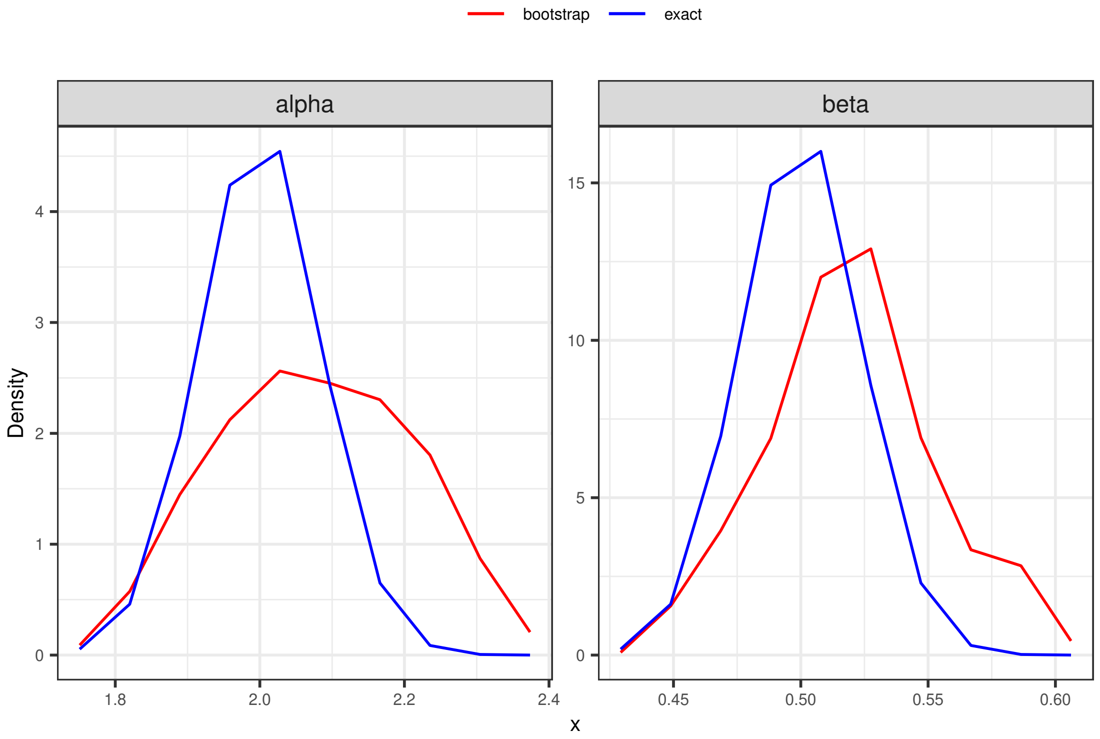
with $\bar{g}=\frac{1}{B}\sum_{b=1}^{B}g^{(b)}$. Provided $F\approx\widehat{F}_n$ holds with sufficient accuracy, this will yield a reasonable approximation to $\mathrm{Var}(\widehat{h(F)})$.

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 $X_1,\ldots,X_n$ 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 $X_1^{(b)},\ldots,X_n^{(b)}$ and $g^{(b)}$ as before. An estimate of the sampling variance is likewise obtained the same manner.

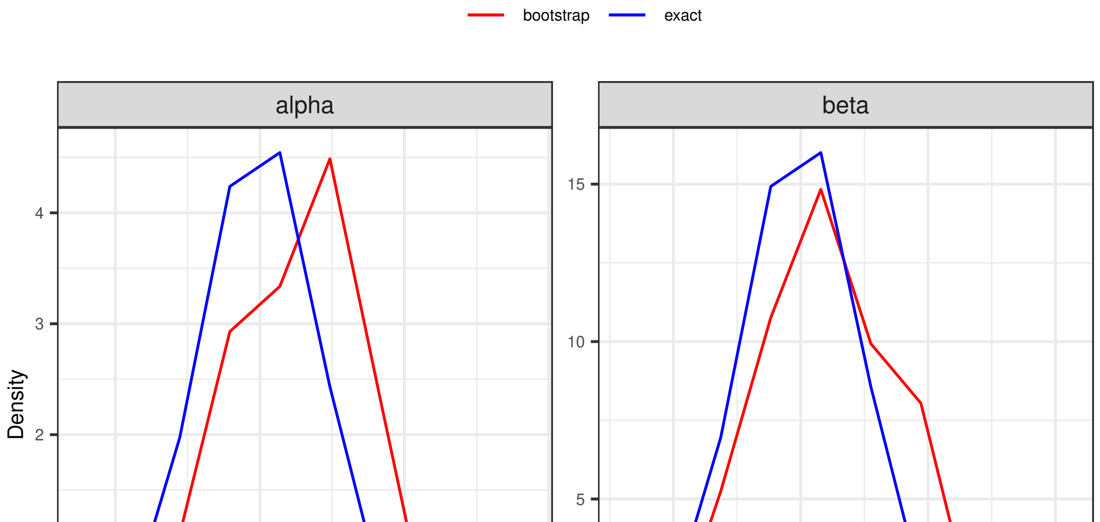
Although we have thus far only used it for the calculation of a single statistic, it is clear that the bootstrap produces a complete *simulated distribution* of the estimator, which can be used for any arbitrary form of inference. This shows its tremendous potential as a tool for statistical analysis, which explains the rise in popularity of these methods with the advent of powerful personal computers capable of carrying out the requisite calculations. Let us give an example to illustrate this. Figure 2.1 compares, for a $\Gamma(\alpha,\beta)$-distribution with $\alpha=2$ and $\beta=0.5$, the bootstrap distributions of the maximum likelihood estimators to their exact counterparts. We use a simulated sample of size $n=1000$ and compute . The analytic distributions are based on the well-known fact from likelihood theory that the asymptotic distribution of the MLE is given by the multivariate normal distribution $\mathcal{N}(\boldsymbol{\theta},I(\boldsymbol{\theta})^{-1})$, where $\boldsymbol{\theta}$ is the parameter vector and $I(\boldsymbol{\theta})$ the Fisher information matrix. For the gamma distribution, the latter is given by

$$I(\alpha,\beta)=n\begin{pmatrix}\psi'(\alpha) & -1/\beta\\ -1/\beta & \alpha/\beta^2\end{pmatrix} \tag{2.3}$$

where $\psi(x):=\frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$ is the so-called digamma function. It is clear that



(a) Parametric



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 [2, 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 $X_1, \ldots, X_p$ and a response variable Y whose relationship we model by a parametrised mapping $f(X_1, \ldots, X_p; \boldsymbol{\beta})$. Given a sample of pairs $(\mathbf{x_1}, Y_1), \ldots, (\mathbf{x_n}, Y_N)$ and a choice of loss function, we can fit this model to obtain an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. For new values $x_1^+, \ldots, x_N^+$ of the regressors, we can then predict the response Y^+ as $f(x_1^+, \ldots, x_N^+; \widehat{\boldsymbol{\beta}})$. It is worth emphasising these as two distinct operations, which correspond to different bootstrap procedures (see [3, 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 or *process variance* of the response variable. Prediction is therefore a two-stage procedure involving an intermediate estimation step.

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

$$Y_i = \mathbf{x}_{\boldsymbol{i}}^T \boldsymbol{\beta} + \varepsilon_i\,, \qquad i \in 1, \ldots, n \tag{2.4}$$

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 use some suitably standardised residuals which we obtain by fitting the model (see [3, 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. Choosing, for example, the residuals

$$r_i := \frac{Y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}}{\widehat{\sigma}\sqrt{1 - h_{ii}}}\,, \tag{2.5}$$

we resample these for B times to obtain pseudo-residuals $r_1^{(b)}, \ldots, r_n^{(b)}$, which in turn yield pseudo-responses

$$Y_i^{(b)} := \mathbf{x}^T \widehat{\boldsymbol{\beta}} + \widehat{\sigma}\sqrt{1 - h_{ii}} r_i^{(b)}\,. \tag{2.6}$$

By refitting the model to this new pseudo-data, we then obtain bootstrapped regression parameter estimates $\widehat{\boldsymbol{\beta}}^{(b)}$ for $1, \ldots, B$.

An alternative approach, which is fully nonparametric, is to resample the pairs $(\mathbf{x_i}, Y_i)$ themselves (see [2, Section 9.5], [3, Algorithm 6.2]), which corresponds to approximating the multivariate distribution of $(X_1, \ldots, X_n, Y)$ by the empirical distribution of the 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 $(\mathbf{x_1}^{(b)}, Y_1^{(b)}), \ldots, (\mathbf{x_n}^{(b)}, Y_n^{(b)})$ to produce pseudo-realisations $\widehat{\boldsymbol{\beta}}^{(b)}$ of the regression parameter estimator.

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 eq. (2.4), 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.

2.3 Process variance and predictive distributions

If we want to do predictive inference on a regression model using the bootstrap, we additionally have to take into account the inherent variability of the response. Considering once again the example of the normal linear model, suppose we are interested in predicting the response Y_+ at new value $\mathbf{x}_+$ of the regressors. One way to quantify the accuracy of our forecast is to consider the *prediction error*

$$\delta := Y_+ - \widehat{Y}_+ \tag{2.7}$$

(see [3, Algorithm 6.4]). The second term in this expression can be bootstrapped using any one of the methods described in the previous section, as these yield replicates $\widehat{\boldsymbol{\beta}}^{(b)}$ of the parameter estimator and hence of the predictor $\widehat{Y}_+^{(b)} = \mathbf{x}_+{}^T\widehat{\boldsymbol{\beta}}^{(b)}$. It then remains for us to produce bootstrap simulations of the response Y_+ itself. In the semiparametric approach, this can be achieved by resampling the residuals a second time to obtain pseudo-realisations $r^{(s)}$ for $s = 1, \dots, S$, and adding these (after correctly scaling them) to the value of the regression line at x_+. The resulting pseudo-responses $Y_+^{(s)}$ mimick the random fluctuations of Y_+, allowing us to approximate δ by the bootstrapped prediction errors

$$\widehat{\delta}^{(b,s)} := Y_+^{(s)} - \widehat{Y}_+^{(b)} = (\widehat{\mathbf{x}}_+^T\widehat{\boldsymbol{\beta}} + r^{(s)}) - \widehat{\mathbf{x}}_+^T\widehat{\boldsymbol{\beta}}^{(b)}\,. \tag{2.8}$$

A similar procedure can be outlined for the parametric bootstrap by appropriately adjusting the method of generating bootstrap response replicates. The method cannot be applied to the pairs bootstrap, however, as it lacks a mechanism for simulating new response values by themselves; we must therefore borrow this part from one of the alternative approaches, forsaking part of its parsimony and robustness properties in the process. We can then use Equation (2.8) to do all kinds of inference, e.g. obtaining prediction intervals or computing the mean squared error of prediction.

While the prediction error offers an avenue of incorporating process variance into bootstrap procedures for predictive inference, it is not the most fitting one for our purposes. Recall from Chapter 1 that we want to study how the reserve is impacted by violation of the assumptions of certain actuarial models, which can be framed in terms of regression, as we shall see in Chapters 3 and 4. In other words, we would like to simulate the distribution of the response itself. In Equation (2.8), this was done by generating fluctuations around $\widehat{\boldsymbol{\beta}}$ through resampling (or in the parametric case, with the help of a random number generator). The trouble with this approach is that it fails to account for the error in our parameter estimates, leading to an underestimation of the prediction uncertainty. Although we will endeavour in this exposition to remain agnostic with respect to philosophical questions about the interpretation of probability, it will be necessary in this case, for reasons which will become apparent shortly, to borrow a concept from Bayesian school of statistics in order to address this problem.

Recall that the Bayesian point of view is premised on the idea that the parameters $\boldsymbol{\theta}$ governing a statistical model $p(y \mid \boldsymbol{\theta}, x_1, \dots, x_p)$ are themselves random variables. These are assumed to follow a so-called *prior distribution* $p(\theta)$, which is a probabilistic expression of the beliefs we have about them before observing any data[1]. When presented with a sample $D = \{(\mathbf{x_1}, Y_1), \ldots, (\mathbf{x_{\boldsymbol{n}}}, Y_n)\}$, we are then led to update our beliefs, and using the formula

$$p(\boldsymbol{\theta} \mid D) \propto p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})\,, \tag{2.9}$$

which is known as *Bayes' rule*, we obtain the *posterior distribution* $p(\boldsymbol{\theta} \mid D)$ expressing the likelihood of different values of $\boldsymbol{\theta}$ given our observations. For any value of the parameters, the likelihood of the response at a new input, conditional on this value and the sample D, is now given by $p(y_+ \mid \boldsymbol{\theta}, D)$. By marginalising over the posterior distribution, we then incorporate all possible values of $\boldsymbol{\theta}$ in proportion to their likelihood under the data, resulting in the *posterior predictive distribution.*

$$p(y_+ \mid D) = \int p(y_+ \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}\,. \tag{2.10}$$

of Y_+ given D; this distribution incorporates both the intrinsic variability of the response as well as our uncertainty regarding the parameters (or, in the classical view, their estimates). Moreover, it can be very easily integrated into the bootstrap framework. Indeed, we can follow the same steps as when simulating the prediction errror, but instead of Equation (2.8), we compute pseudo-realisations

$$Y_+^{(b,s)} := \mathbf{x}_+^T \widehat{\boldsymbol{\beta}}^{(b)} + \widehat{\sigma}\sqrt{1 - h_{ii}}\, r^{(s)}\,. \tag{2.11}$$

where h_{ii} is the leverage corresponding to the.

Equation (2.11) also explains why it is difficult to fit this approach within a classical frequentist paradigm, as it is unclear in that case which theoretical quantity these bootstrap replicates would be approximating. We hasten to add that much work has been done to remedy this, leading, among other things, to the concept of a *confidence distribution* (see, for example, [4], [5]). It is true, however, that there has been a multiplicity of disparate frequentist versions of the predictive distribution, as remarked for instance by Dickson, Tedesco, and Zehnwirth [6], lending credence to the idea that the notion fits more naturally into the Bayesian framework.

[1] Strictly speaking, this is only one possible paradigm within Bayesian probability, known as subjective Bayes. Objective Bayes dispenses with the subjective aspect of the method through the use of so-called *uninformative priors*. Since this distinction is not important for our exposition, however, we will not pursue it further.

Chapter 3

Mack's model

3.1 Introduction

In his seminal paper [7], 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

$$\mathbb{E}[C_{ij} \parallel C_{i,j-1}, \dots, C_{i1}] = \mathbb{E}[C_{ij} \parallel C_{i,j-1}] = f_{j-1} C_{i,j-1} \tag{3.1}$$

for $1 \leq i \leq I$.

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

$$\operatorname{Var}[C_{ij} \parallel C_{i,j-1}, \dots, C_{i1}] = \operatorname{Var}[C_{ij} \parallel C_{i,j-1}] = \sigma_{j-1}^2 C_{i,j-1}\,, \tag{3.2}$$

for $1 \leq i \leq I$.

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

The development factors are estimated by

$$\widehat{f}_j(\mathcal{D}_I) = \widehat{f}_j(C_{1j}, \dots, C_{I-j,j}, \dots, C_{1,j+1}, \dots, C_{I-j,j+1}) \coloneqq \frac{\sum_{i=1}^{I-j} C_{i,j+1}}{\sum_{i=1}^{I-j} C_{i,j}}\,. \tag{3.3}$$

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

$$F_{i,j+1} \coloneqq \frac{C_{i,j+1}}{C_{ij}}\,, \tag{3.4}$$

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

$$\widehat{f}_j = \frac{\sum_{i=1}^{I-j} C_{ij} F_{i,j}}{\sum_{i=1}^{I-j} C_{ij}}\,. \tag{3.5}$$

The σ_j are estimated by

$$\widehat{\sigma}_j \coloneqq \frac{1}{I-j} \sum_{i=1}^{I-j} C_{ij} \left(F_{i,j+1} - \widehat{f}_j\right)^2 \tag{3.6}$$

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

$$\widehat{\sigma}_{I-1}^2 = \min\left\{\frac{\widehat{\sigma}_{I-2}^4}{\widehat{\sigma}_{I-3}^2}, \widehat{\sigma}_{I-2}^2, \widehat{\sigma}_{I-3}^2\right\} \tag{3.7}$$

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

Under the assumptions of Model 1, it can be shown (see [8, 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

$$\widehat{C}_{iI} := \widehat{\mathbb{E}}[C_{i,I} \parallel C_{i,I-i}] = C_{i,I+1-i} \prod_{j=I-i}^{I-1} \widehat{f}_j \,. \tag{3.8}$$

From this, we then finally obtain the reserve predictor

$$\widehat{R} = g(\mathcal{D}_I) := \sum_{i=2}^{I} (\widehat{C}_{i,I} - C_{i,I+1-i}) \,. \tag{3.9}$$

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 $j \in \{1, \ldots, I-1\}$, the notation

$$\mathbf{c_j} := \begin{bmatrix} C_{1,j} \\ \vdots \\ C_{I-j,j} \end{bmatrix}, \tag{3.10}$$

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

$$\mathbf{c_{j+1}} = f_j \mathbf{c_j} + \boldsymbol{\varepsilon} \,, \tag{3.11}$$

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

$$\mathbb{E}[\boldsymbol{\varepsilon} \parallel C_{1,j}, \ldots, C_{i,I-j}] = \mathbf{0} \qquad \mathrm{Var}(\boldsymbol{\varepsilon} \parallel C_{1,j}, \ldots, C_{i,I-j}) = \sigma_j^2 \begin{bmatrix} C_{1j} & & \\ & \ddots & \\ & & C_{I-j,j} \end{bmatrix}. \tag{3.12}$$

Consequently, it follows (see [9, Proposition 1.7]) that the weighted least squares method with weights matrix

$$\mathbf{W} = \begin{bmatrix} 1/C_{1j} & & \\ & \ddots & \\ & & 1/C_{I-j,j} \end{bmatrix}, \tag{3.13}$$

[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.

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

$$\widehat{f}_j^{\text{WLS}} = (\mathbf{c}_{\boldsymbol{j}}^T \mathbf{W} \mathbf{c}_{\boldsymbol{j}})^{-1} \mathbf{c}_{\boldsymbol{j}}^T \mathbf{W} = \frac{\sum_{i=1}^{I-j} C_{i,j+1}}{\sum_{i=1}^{I-j} C_{i,j}} , \tag{3.14}$$

which is the same expression as eq. (3.3).

3.2 A challenging simulation

Owing to its recursive nature, Mack's model does not readily lend itself to application of the theory from Chapter 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 [10] 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.

$$\underset{R \,|\, \mathcal{D}_I}{\text{MSEP}}(\widehat{R}) \coloneqq \mathbb{E}\Big[(\widehat{R} - R)^2 \;\Big\|\; \mathcal{D}_I\Big] \tag{3.15}$$

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*:

$$\begin{aligned}
\mathbb{E}\Big[(\widehat{R} - R)^2 \;\Big\|\; \mathcal{D}_I\Big] &= \mathbb{E}\big[(R - \mathbb{E}[R \parallel \mathcal{D}_I])^2 \parallel \mathcal{D}_I\big] + \mathbb{E}\Big[(\mathbb{E}[R \parallel \mathcal{D}_I] - \widehat{R})^2 \;\Big\|\; \mathcal{D}_I\Big] \\
&\qquad - 2\mathbb{E}\Big[(R - \mathbb{E}[R \parallel \mathcal{D}_I])(\mathbb{E}[R \parallel \mathcal{D}_I] - \widehat{R}) \;\Big\|\; \mathcal{D}_I\Big] && (3.16) \\
&= \text{Var}(R \,\|\, \mathcal{D}_I) + (\mathbb{E}[R \parallel \mathcal{D}_I] - \widehat{R})^2 \\
&\qquad - 2(\mathbb{E}[R \parallel \mathcal{D}_I] - \widehat{R})(\mathbb{E}[R - \mathbb{E}[R \parallel \mathcal{D}_I] \parallel \mathcal{D}_I]) && (3.17) \\
&= \underbrace{\text{Var}(R \,\|\, \mathcal{D}_I)}_{\text{process error}} + \underbrace{(\mathbb{E}[R \parallel \mathcal{D}_I] - \widehat{R})^2}_{\text{estimation error}} , && (3.18)
\end{aligned}$$

corresponding to the two stages of bootstrapping a predictor which we discussed in Section 2.2. Consider now, for any accident year $i \in \{1, \dots, I\}$, the MSEP for the associated ultimate

$$\underset{C_{iI} \,|\, \mathcal{D}_I}{\text{MSEP}}(\widehat{C}_{iI}) = (\mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \widehat{C}_{iI})^2 + \text{Var}(C_{iI} \parallel \mathcal{D}_I) , \tag{3.19}$$

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 eq. (3.1) to obtain

$$\begin{aligned}
\mathrm{Var}(C_{iI} \parallel \mathcal{D}_I) &= \mathrm{Var}(C_{iI} \parallel C_{i,I+1-i}) && (3.20)\\
&= \mathbb{E}[\mathrm{Var}(C_{iI} \parallel C_{i,I-1}) \parallel C_{i,I+1-i}] + \mathrm{Var}(\mathbb{E}[C_{iI} \parallel C_{i,I-1}] \parallel C_{i,I+1-i}) && (3.21)\\
&= \sigma_{I-1}^2 \mathbb{E}[C_{i,I-1} \parallel C_{i,I+1-i}] + f_{I-1}^2 \mathrm{Var}(C_{i,I-1} \parallel C_{i,I+1-i}) && (3.22)\\
&= \sigma_{I-1}^2 C_{i,I+1-i} \prod_{j=I+1-i}^{I-2} f_j + f_{I-1}^2 \mathrm{Var}(C_{i,I-1} \parallel C_{i,I+1-i}) \,, && (3.23)
\end{aligned}$$

which is a linear recurrence equation of the form

$$x_n = a_{n-1} x_{n-1} + g_{n-1} \tag{3.24}$$

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

$$g_{n-1} = \sigma_{n-1}^2 C_{i,I+1-i} \prod_{j=I+1-i}^{n-1} f_j \,, \qquad a_{n-1} = f_{n-1}^2 \,. \tag{3.25}$$

The general solution is given by

$$x_n = \left(\prod_{j=n_0}^{n-1} a_j \right) \left(x_{n_0} + \sum_{k=n_0}^{n-1} \frac{g_k}{\prod_{l=n_0}^{k} a_l} \right) \tag{3.26}$$

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

$$\begin{aligned}
\mathrm{Var}(C_{iI} \parallel \mathcal{D}_I) &= \left(\prod_{j=I+1-i}^{I-1} f_j^2 \right) \left(\sum_{k=I+1-i}^{I-1} \frac{\sigma_k^2 C_{i,I+1-i} \prod_{j=I+1-i}^{k-1} f_j}{\prod_{j=I+1-i}^{k} f_j^2} \right) && (3.27)\\
&= \left(\prod_{j=I+1-i}^{I-1} f_j^2 \right) C_{i,I+1-i}^2 \left(\sum_{k=I+1-i}^{I-1} \frac{\sigma_k^2 / f_k^2}{\prod_{j=I+1-i}^{k-1} f_j C_{i,I+1-i}} \right) && (3.28)\\
&= \mathbb{E}[C_{iI} \parallel C_{I+1-i}]^2 \sum_{k=I+1-i}^{I-1} \frac{\sigma_k^2 / f_k^2}{\mathbb{E}[C_{ik} \parallel C_{i,I+1-i}]} \,, && (3.29)
\end{aligned}$$

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

$$\begin{aligned}
(\mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \widehat{C}_{iI})^2 &= C_{i,I+1-i}^2 \left(\prod_{j=I+1-i}^{I-1} f_j - \prod_{j=I+1-i}^{I-1} \widehat{f}_j \right)^2 && (3.30)\\
&= C_{i,I+1-i}^2 \left(\prod_{j=I+1-i}^{I-1} f_j^2 + \prod_{j=I+1-i}^{I-1} \widehat{f}_j^2 - 2 \prod_{j=I+1-i}^{I-1} f_j \widehat{f}_j \right) \,, && (3.31)
\end{aligned}$$

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 [11], and is caused by the fact that the claims triangle observations are used for both estimation and forecasting (see [12, 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 [13], [8, pp. 44 sqq.]), with different mechanisms yielding different estimates. The literature uses this mostly as a theoretical device to facilitate analytical calculations; for the specific approach developed by Mack, it leads to the estimator

$$\widehat{\mathrm{MSEP}}(\widehat{R}_i) := \widehat{C}_{iI} \sum_{j=I+1-i}^{I-1} \frac{\widehat{\sigma}_j^2}{\widehat{f}_j} \left(\frac{1}{\widehat{C}_{ij}} + \frac{1}{\sum_{i=1}^{I-j} C_{ij}} \right) \tag{3.32}$$

(see [11, p. 11]). In this case, however, the theory happens to fit in perfectly with the resampling framework, and we can therefore employ it as a basis for bootstrap procedures. In the remainder of this section, we outline two approaches for estimating eq. (3.30) and indicate the corresponding resampling methods.

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

$$\mathcal{B}_k := \{C_{ij} \in \mathcal{D}_I \mid j \leq k\} \,, \tag{3.33}$$

and write

$$\mathcal{D}_{I,k}^O := \{C_{ij} \in \mathcal{D}_I \mid j > I + 1 - k\} \tag{3.34}$$

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

$$\mathbb{E}[(\mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \widehat{C}_{iI})^2 \parallel \mathcal{B}_{I+1-i}] = C_{i,I+1-i}^2 \mathbb{E}\left[\left(\prod_{j=I+1-i}^{I-1} f_j - \prod_{j=I+1-i}^{I-1} \widehat{f}_j \right)^2 \middle\| \mathcal{B}_{I+1-i} \right] \tag{3.35}$$

$$= C_{i,I+1-i}^2 \left(\mathbb{E}\left[\prod_{j=I+1-i}^{I-1} \widehat{f}_j^2 \middle\| \mathcal{B}_{I+1-i} \right] - \prod_{j=I+1-i}^{I-1} f_j^2 \right) , \tag{3.36}$$

where we used the fact that the $\widehat{f}_j$ are uncorrelated. This corresponds to averaging over the distribution of $\mathcal{D}_{I,k}^O$, or, expressed in terms of resampling, to generating new observations in the upper right triangle. Borrowing the nomenclature from [13], 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 eq. (3.3) 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 $\mathcal{D}_{I,i}^O$ by

$$\mathbb{P}_{\mathcal{D}_I}^*(\{dz_{ij}\}_{i+j \leq I+1}) := \prod_{j=1}^{I-1} \prod_{i=1}^{I-j} \mathbb{P}_{C_{i,j+1}}(dz_{i,j+1} \mid C_{ij} = c_{ij}) \,, \tag{3.37}$$

yielding the estimate

$$\mathbb{E}_{\mathbb{P}^*_{\mathcal{D}_I}}\left[(\mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \widehat{C}_{iI})^2\right] = C_{iI}^2\, \mathbb{E}_{\mathbb{P}^*_{\mathcal{D}_I}}\left[\left(\prod_{j=I+1-i}^{I-1} f_j - \prod_{j=I+1-i}^{I-1} \widehat{f}_j\right)^2\right] \tag{3.38}$$

$$= C_{i,I+1-i}^2 \left(\mathbb{E}_{\mathbb{P}^*_{\mathcal{D}_I}}\left[\prod_{j=I+1-i}^{I-1} \widehat{f}_j^2\right] - \prod_{j=I+1-i}^{I-1} f_j^2\right) \tag{3.39}$$

$$= C_{i,I+1-i}^2 \left(\prod_{j=I+1-i}^{I-1} \mathbb{E}\left[\widehat{f}_j^2 \,\middle\|\, \mathcal{B}_j\right] - \prod_{j=I+1-i}^{I-1} f_j^2\right) . \tag{3.40}$$

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 [13]–[16]. As we will see in Section 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 [14].

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

$$\operatorname{Cov}(\widehat{f}_j, \widehat{f}_{j-1}) < 0 . \tag{3.41}$$

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 when needed. In the next section, we will see how both approaches give rise to a variety of different bootstrap methods.

3.3 Bootstrap methodology

In Section 2.2, we introduced a taxonomy for the different kinds of bootstrap, distinguishing between the semiparameteric, nonparametric and parametric type. We now consider how each of these can be applied to Model 1. For comparison, Table 3.1 shows the results of applying Model 1 with the estimator Equation (3.32) to the dataset from Table 1.2.

For the semiparametric bootstrap, the crucial step is to find a suitable definition for the residuals which ensures that they are interchangeable. The distribution-free nature of the model makes this difficult, however, as it limits the statements we can make about the errors to the first two moments. We can resolve this in one of two ways. The first option would be to extrapolate from homogeneity of the first two moments to homogeneity of the distributions. In that case, the *raw residuals*

$$e_{i,j+1} \coloneqq C_{i,j+1} - \widehat{C}_{i,j+1} = C_{i,j+1} - \widehat{f}_j C_{ij} \tag{3.42}$$

$i\,/\,j$	$\widehat{f}_j^{\text{CL}}$	$\widehat{\sigma}_i^{\text{CL}}$	$\widehat{R}_i^{\text{CL}}$	$\widehat{\text{MSEP}}(\widehat{R}_j)$
2	1.89	2.83	350.9	3.62
3	1.28	3.34	1037.54	22.9
4	1.15	2.98	2044.86	141.98
5	1.1	1.07	3663.4	426.7
6	1.05	0.16	7162.15	692.39
7	1.03	0.02	14396.92	900.58

Table 3.1: Mack CL results for UK Motor triangle

are not an option, as these suffer from heteroscedasticity,

$$\text{Var}(e_{i,j+1} \parallel C_{ij}) = \sigma_j^2 \left(C_{ij} - \frac{C_{ij}^2}{\sum_{i=1}^{I-j} C_{ij}} \right) . \tag{3.43}$$

We can address this by dividing out this variance, i.e. we consider the errors

$$\varepsilon_{i,j+1} := \frac{C_{i,j+1} - f_j C_{ij}}{\sigma_j \sqrt{C_{ij}} \sqrt{1 - \frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}}} , \tag{3.44}$$

which satisfy $\mathbb{E}[\varepsilon_{i,j+1} \parallel C_{ij}] = 0$ and $\text{Var}(\varepsilon_{i,j+1} \parallel C_{ij}) = 1$. Provided the sampling variability of the $\widehat{f}_j$ 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

$$r_{i,j+1} := \frac{C_{i,j+1} - \widehat{f}_j C_{ij}}{\widehat{\sigma}_j \sqrt{C_{ij}} \sqrt{1 - \frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}}} , \tag{3.45}$$

obtained by substituting these estimators. Note that the factor $\sqrt{1 - \frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}}$ in the denominator corresponds to the leverage adjustment, as can be seen by computing the hat matrix:

$$\mathbf{H} = \mathbf{c_j}(\mathbf{c_j}^T \mathbf{W} \mathbf{c_j})^{-1} \mathbf{c_j}^T \mathbf{W} \tag{3.46}$$

$$= \frac{1}{\sum_{i=1}^{I-j} C_{ij}} \begin{bmatrix} C_{1j} & \dots & C_{1j} \\ \vdots & \ddots & \vdots \\ C_{I-j,j} & \dots & C_{I-j,j} \end{bmatrix} . \tag{3.47}$$

It's worth emphasising, however, that the above extrapolation should not be made lightly, as it is perfectly possible for the error distribution to exhibit heterogeneity in other ways than through its mean and variance (see [2, p. 114] for an example where the *percentiles* vary with the value of the regressor). In light of this, an alternative approach would be to augment our model with some explicit distributional assumptions, which is more transparent and allows us to make precise statements about errors and residuals. One such augmentation that has been studied in the literature (see [8, p. 49]) is the autoregressive Gaussian time series model

$$C_{i,j+1} = f_j C_{ij} + \sigma_j \sqrt{C_{ij}}\, \varepsilon_{i,j+1}, \qquad \varepsilon \sim \mathcal{N}(0,1) , \tag{3.48}$$

which can easily be seen to be compatible with Model 1. Because the Mack CL can be viewed as a series of weighted linear regressions, as we say in Section 3.1, this has the benefit of making available to us the results of classical regression theory. We know, for example, that the *externally studentised residuals*

$$r_{i,j+1} := \frac{e_{i,j+1}}{\widehat{\sigma}_{j(i)}\sqrt{1-\mathbf{H}_{ii}}}\sqrt{\mathbf{W}_{ii}} = \frac{C_{i,j+1} - \widehat{f}_j C_{ij}}{\widehat{\sigma}_{j(i)}\sqrt{C_{ij}}\sqrt{1-\frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}}}\,, \tag{3.49}$$

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

$$r_{i,j+1} := \frac{C_{i,j+1} - \widehat{f}_j C_{ij}}{\widehat{\sigma}_j\sqrt{C_{ij}}\sqrt{1-\frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}}} \tag{3.50}$$

which also share the same distribution, albeit a more complicated one (see [17, pp. 267 sqq.]).

The Gaussian model has a major shortcoming: it makes it possible to have a negative realisation in the next step of the time series, in which case all future observations from that point on are undefined, because of the square root factor appearing in the variance. This is not merely a theoretical problem: we have sometimes observed this phenomenon in our numerical implementation, where the resampling produces negative pseudo-realisations of certain claim amounts, particularly when the model has been severely perturbed. One obvious way of avoiding it would be to simply discard the current bootstrap iteration as soon as a negative value is produces. This has the drawback of requiring more computational power, sometimes beyond the realm of what is reasonable. Moreover, we have seen cases in which the probability of having no negative replicates was so vanishingly small as to cause the program to get stuck indefinitely. We must therefore develop an alternative approach, if only to have a fail-safe method to fall back upon in case of problems with Equation (3.48).

Fundamentally, we must find of guaranteeing

$$\varepsilon_{i,j+1} > -\frac{f_j C_{ij}}{\sigma_j}\,. \tag{3.51}$$

This can be achieved by choosing an alternative error distribution in eq. (3.48), one whose support is bounded from below. For example, we could use a shifted lognormal distribution, i.e.

$$\log(\varepsilon_{i,j+1} - \alpha_{ij}) \sim \mathcal{N}(\beta_{ij}, \gamma_{ij}^2) \tag{3.52}$$

for certain parameters α_{ij}, β_{ij} and γ_{ij}. This has support $(\alpha_{ij}, +\infty)$, so that the requirement eq. (3.51) is satisfied with $\alpha_{ij} = -\frac{f_j C_{ij}}{\sigma_j}$. It then remains for us to determine β_{ij} and γ_{ij} such that the assumptions of Model 1 are satisfied, meaning

$$\mathbb{E}[\varepsilon_{i,j+1} \parallel C_{ij}] = \exp\left(\beta_{ij} + \frac{\gamma_{ij}^2}{2}\right) + \alpha_{ij} = 0, \qquad \operatorname{Var}(\varepsilon_{i,j+1} \parallel C_{ij}) = (\exp\gamma_{ij}^2 - 1)\exp(\gamma_{ij}^2 + 2\beta_{ij}) = 1\,. \tag{3.53}$$

Solving these equations, we obtain

$$\gamma_{ij} = \sqrt{\log\left(1 - \frac{1}{\alpha_{ij}^2}\right)}, \qquad \beta_{ij} = \log(-\alpha_{ij}) - \frac{\gamma_{ij}}{2}\,. \tag{3.54}$$

Provided the sampling variability of $\widehat{f}_j$ and $\widehat{\sigma}_j$ is not too severe, this means that the residuals

$$r_{i,j+1} := \frac{\log\left(C_{i,j+1} - \widehat{f}_j C_{ij} + \widehat{f}_j C_{ij}/\widehat{\sigma}_j\right) - \widehat{\beta}_{ij}}{\widehat{\gamma}_{ij}} \tag{3.55}$$

$$= \frac{\log\left(C_{i,j+1} - \widehat{f}_j C_{ij} + \widehat{f}_j C_{ij}/\widehat{\sigma}_j\right) - \log(\widehat{f}_j\sqrt{C_{ij}}/\widehat{\sigma}_j) + \frac{1}{2}\left(\sqrt{\log\left(1 + \widehat{\sigma}_j/\widehat{f}_j C_{ij}\right)}\right)}{\sqrt{\log\left(1 + \widehat{\sigma}_j/\widehat{f}_j C_{ij}\right)}} \tag{3.56}$$

are approximately $\mathcal{N}(0,1)$-distributed.

After selecting a particular type of residual, the next step is to fit the model and compute the residuals from it. We then resample these to generate bootstrap residuals $r_{ij}^{(b)}$, from which a bootstrap triangle is obtained by inverting the appropriate residuals formula. In the conditional approach, the inversion is based on the original triangle, whereas the unconditional version uses the previously generated bootstrap observations. Finally, the model is refitted to the generated triangle to obtain bootstrap development factor and dispersion parameter estimators $\widehat{\boldsymbol{f}}$ and $\widehat{\boldsymbol{\sigma}}$. The entire procedure is outlined in Algorithms 1 and 2 for the case of standardised residuals, and the results for the example data from Table 1.2 are given in Table 3.2.

Algorithm 1 Conditional semiparametric bootstrap for Mack CL

```
Input: Cumulative claims triangle $\mathcal{D}_I$, required number of bootstrap samples $B$
  $(\{r_{ij} \mid i+j \le I+1\}, \widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\sigma}}) \leftarrow \textsc{fit}(\mathcal{D}_I)$
  for $b \leftarrow 1, B$ do
      $\{r_{ij}^{(b)} \mid i+j \le I+1\} \leftarrow \textsc{resample}(\{r_{ij} \mid i+j \le I+1\})$
      for $j \leftarrow 1, I-1$ do
          for $i \leftarrow 1, I-j$ do
              $C_{i,j+1}^{(b)} \leftarrow \widehat{f}_j C_{ij} + \widehat{\sigma}_j \sqrt{C_{ij}} \sqrt{1 - \frac{C_{ij}}{\sum_{i=1}^{I-j} C_{ij}}} r_{i,j+1}$
              $F_{i,j+1}^{(b)} \leftarrow C_{i,j+1}^{(b)}/C_{ij}$
          end for
          $\widehat{f}_j^{(b)} \leftarrow \sum_{i=1}^{I-j} C_{i,j+1}^{(b)} / \sum_{i=1}^{I-j} C_{ij}$
          if $j < I-1$ then
              $\widehat{\sigma}_j^{(b)} \leftarrow \frac{1}{I-j-1} \sum_{i=1}^{I-j} C_{ij} \left(F_{i,j+1}^{(b)} - \widehat{f}_j^{(b)}\right)^2$
          else
              $\widehat{\sigma}_{I-1}^{(b)} \leftarrow \sqrt{\min\left\{\frac{(\widehat{\sigma}_{I-2}^{(b)})^4}{(\widehat{\sigma}_{I-3}^{(b)})^2}, (\widehat{\sigma}_{I-2}^{(b)})^2, (\widehat{\sigma}_{I-3}^{(b)})^2\right\}}$
          end if
      end for
  end for
  return $\{(\widehat{\boldsymbol{f}}^{(b)}, \widehat{\boldsymbol{\sigma}}^{(b)}) \mid b = 1, \ldots, B\}$
```

Algorithm 2 Unconditional semiparametric bootstrap for Mack CL

```
Input: Cumulative claims triangle $\mathcal{D}_I$, required number of bootstrap samples $B$
  $(\{r_{ij} \mid i + j \leq I + 1\}, \widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\sigma}}) \leftarrow \textsc{fit}(\mathcal{D}_I)$
  for $b \leftarrow 1, B$ do
      for $i \leftarrow 1, I$ do
          $C_{i1}^{(b)} \leftarrow C_{i1}$
      end for
      for $j \leftarrow 1, I - 1$ do
          for $i \leftarrow 1, I - j$ do
              $C_{i,j+1}^{(b)} \leftarrow \widehat{f}_j C_{ij}^{(b)} + \widehat{\sigma}_j \sqrt{C_{ij}^{(b)}} \sqrt{1 - \frac{C_{ij}^{(b)}}{\sum_{i=1}^{I-j} C_{ij}^{(b)}}} r_{i,j+1}$
              $F_{i,j+1}^{(b)} \leftarrow C_{i,j+1}^{(b)} / C_{ij}^{(b)}$
          end for
          $\widehat{f}_j^{(b)} \leftarrow \sum_{i=1}^{I-j} C_{i,j+1}^{(b)} / \sum_{i=1}^{I-j} C_{ij}^{(b)}$
          if $j < I - 1$ then
              $\widehat{\sigma}_j^{(b)} \leftarrow \frac{1}{I-j-1} \sum_{i=1}^{I-j} C_{ij}^{(b)} \left( F_{i,j+1}^{(b)} - \widehat{f}_j^{(b)} \right)^2$
          else
              $\widehat{\sigma}_{I-1}^{(b)} \leftarrow \sqrt{\min \left\{ \frac{(\widehat{\sigma}_{I-2}^{(b)})^4}{(\widehat{\sigma}_{I-3}^{(b)})^2}, (\widehat{\sigma}_{I-2}^{(b)})^2, (\widehat{\sigma}_{I-3}^{(b)})^2 \right\}}$
          end if
      end for
  end for
  return $\{(\widehat{\boldsymbol{f}}^{(b)}, \widehat{\boldsymbol{\sigma}}^{(b)}) \mid b = 1, \ldots, B\}$
```

j	$\widehat{f}_j^B$	$\widehat{\sigma}_j^B$	$\widehat{R}_j^B$	$\widehat{\mathrm{MSEP}}(\widehat{R}_j)$
2	1.89	2.45	335.1	35.26
3	1.27	4.18	979.83	158.99
4	1.14	5.29	2055.74	533
5	1.1	4.36	3641.79	919.08
6	1.05	1.07	6988.07	1023.69
7	1.03	0.29	14232.38	1328.99

(a) Conditional

j	$\widehat{f}_j^B$	$\widehat{\sigma}_j^B$	$\widehat{R}_j^B$	$\widehat{\mathrm{MSEP}}(\widehat{R}_j)$
2	1.89	2.3	350.9	3.29
3	1.28	3.03	1032.43	11.41
4	1.14	2.24	2040.81	98.51
5	1.1	0.84	3617.58	305.97
6	1.05	0.06	7101.28	582.25
7	1.03	0.01	14337.49	600.69

(b) Unconditional

Table 3.2: Semiparametric bootstrap results

Next, we consider the fully nonparametric bootstrap, in which we resample the pairs $(C_{ij}, C_{i,j+1})$ at every development year index j. For this procedure, we have no choice regarding the resampling scheme which is used: the only possibility is conditional resampling. To see why this is the case, consider what it would mean to implement unconditional resampling. If the resampled pairs for the first two columns are denoted by

$$\{(C^*_{11}, C^*_{12}), \ldots, (C^*_{I-j,1}, C^*_{I-j,2})\}\,,$$

this would mean using the generated C^*_{i2} as the regressor column in the second step. However, as the pairs are fundamental i.i.d. unit for this method, we have to ensure that these remain paired to the same response from the third column. In effect, this means that we are forced to permute the rows of the triangle at every stage. But this creates a problem: the last point in the second column does not have a successor in the triangle, and we therefore become stuck in the second step if we had previously drawn it, as illustrated in Figure 3.1. Hence the only option is to carry out the resampling of each column pairs independently of the others, using the original data, and compute bootstrapped development factor and dispersion parameter estimates from these. The entire procedure is outlined in Algorithm 3, and the results for the UK Motor dataset are given in Table 3.3.

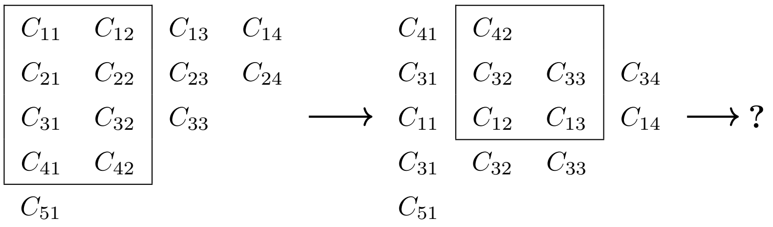

C_{11} C_{12} C_{13} C_{14}
C_{21} C_{22} C_{23} C_{24}
C_{31} C_{32} C_{33}
C_{41} C_{42}
C_{51}
C_{41} C_{42}
C_{31} C_{32} C_{33} C_{34}
C_{11} C_{12} C_{13} C_{14}
C_{31} C_{32} C_{33}
C_{51}
?

Figure 3.1: Failure of unconditional pairs resampling

Algorithm 3 Pairs bootstrap 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
          $\{(C^{(b)}_{i,j}, C^{(b)}_{i,j+1}) \mid i = 1, \ldots, I-j\} \leftarrow$ RESAMPLE($\{(C_{i,j}, C_{i,j+1}) \mid i = 1, \ldots, I-j\}$)
          $\widehat{f}^{(b)}_j \leftarrow \sum_{i=1}^{I-j} C^{(b)}_{i,j+1} / \sum_{i=1}^{I-j} C^{(b)}_{ij}$
          for $i \leftarrow 1, I-j$ do
              $F^{(b)}_{i,j+1} \leftarrow C^{(b)}_{i,j+1} / C^{(b)}_{ij}$
          end for
          $\widehat{\sigma}^{(b)}_j \leftarrow \frac{1}{I-j} \sum_{i=1}^{I-j} C^{(b)}_{ij} \left(F^{(b)}_{i,j+1} - \widehat{f}^{(b)}_j\right)^2$
      end for
  end for
  return $\{(\widehat{\boldsymbol{f}}^{(b)}, \widehat{\boldsymbol{\sigma}}^{(b)}) \mid b = 1, \ldots, B\}$
```

Finally, we discuss the parametric bootstrap, in which we simulate directly from the fitted model. Based on Equation (3.48), we might be tempted to simply substitute $\mathcal{N}(0,1)$-distributed

j	$\widehat{f}_j^B$	$\widehat{\sigma}_j^B$	$\widehat{R}_j^B$	$\widehat{\text{MSEP}}(\widehat{R}_j)$
2	1.88	10.22	350.9	2.65
3	1.28	7.97	1041.46	22.93
4	1.16	7.36	2036.93	138.29
5	1.1	0.68	3789.38	431.8
6	1.05	0.01	7345.81	677.97
7	1.03	0	14492.71	958.93

Table 3.3: Pairs bootstrap results

draws from a random number generator for the residuals in Algorithms 1 and 5. The problem with this approach, is that it doesn't extend easily to other distributions, because it is not possible, in general, to write these as the sum of a mean and a scaled error term. A better idea is therefore to generate bootstrap responses directly from the fitted distribution. Depending on whether the conditional or the unconditional scheme is used, this will either depend on the original triangle observations or the ones generated at the previous step. The Mack CL is then refitted to the bootstrapped triangle in order to obtain $\widehat{\boldsymbol{f}}^{(b)}$ and $\widehat{\boldsymbol{\sigma}}^{(b)}$.

Algorithm 4 Conditional parametric bootstrap for Mack CL

```
Input: Cumulative claims triangle $\mathcal{D}_I$, required number of bootstrap samples $B$
  $(\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\sigma}}) \leftarrow \textsc{fit}(\mathcal{D}_I)$
  for $b \leftarrow 1, B$ do
      for $j \leftarrow 1, I-1$ do
          for $i \leftarrow 1, I-j$ do
              $C_{i,j+1}^{(b)} \leftarrow \textsc{sample}(\mathcal{N}(\widehat{f}_j C_{ij}, \widehat{\sigma}_j^2 C_{ij}))$
              $F_{i,j+1}^{(b)} \leftarrow C_{i,j+1}^{(b)} / C_{ij}$
          end for
          $\widehat{f}_j^{(b)} \leftarrow \sum_{i=1}^{I-j} C_{i,j+1}^{(b)} / \sum_{i=1}^{I-j} C_{ij}$
          if $j < I-1$ then
              $\widehat{\sigma}_j^{(b)} \leftarrow \frac{1}{I-j-1} \sum_{i=1}^{I-j} C_{ij} \left( F_{i,j+1}^{(b)} - \widehat{f}_j^{(b)} \right)^2$
          else
              $\widehat{\sigma}_{I-1}^{(b)} \leftarrow \sqrt{\min\left\{ \frac{(\widehat{\sigma}_{I-2}^{(b)})^4}{(\widehat{\sigma}_{I-3}^{(b)})^2}, (\widehat{\sigma}_{I-2}^{(b)})^2, (\widehat{\sigma}_{I-3}^{(b)})^2 \right\}}$
          end if
      end for
  end for
  return $\{(\widehat{\boldsymbol{f}}^{(b)}, \widehat{\boldsymbol{\sigma}}^{(b)}) \mid b = 1, \ldots, B\}$
```

Algorithm 5 Unconditional parametric bootstrap for Mack CL

```
Input: Cumulative claims triangle $\mathcal{D}_I$, required number of bootstrap samples $B$
  $(\{r_{ij} \mid i+j \le I+1\}, \widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\sigma}}) \leftarrow \textsc{fit}(\mathcal{D}_I)$
  for $b \leftarrow 1, B$ do
      for $i \leftarrow 1, I$ do
          $C_{i1}^{(b)} \leftarrow C_{i1}$
      end for
      for $j \leftarrow 1, I-1$ do
          for $i \leftarrow 1, I-j$ do
              $C_{i,j+1}^{(b)} \leftarrow \textsc{sample}(\mathcal{N}(\widehat{f}_j C_{ij}^{(b)}, \widehat{\sigma}_j^2 C_{ij}^{(b)}))$
              $F_{i,j+1}^{(b)} \leftarrow C_{i,j+1}^{(b)} / C_{ij}^{(b)}$
          end for
          $\widehat{f}_j^{(b)} \leftarrow \sum_{i=1}^{I-j} C_{i,j+1}^{(b)} / \sum_{i=1}^{I-j} C_{ij}^{(b)}$
          if $j < I-1$ then
              $\widehat{\sigma}_j^{(b)} \leftarrow \frac{1}{I-j-1} \sum_{i=1}^{I-j} C_{ij}^{(b)} \left(F_{i,j+1}^{(b)} - \widehat{f}_j^{(b)}\right)^2$
          else
              $\widehat{\sigma}_{I-1}^{(b)} \leftarrow \sqrt{\min\left\{\frac{(\widehat{\sigma}_{I-2}^{(b)})^4}{(\widehat{\sigma}_{I-3}^{(b)})^2}, (\widehat{\sigma}_{I-2}^{(b)})^2, (\widehat{\sigma}_{I-3}^{(b)})^2\right\}}$
          end if
      end for
  end for
  return $\{(\widehat{\boldsymbol{f}}^{(b)}, \widehat{\boldsymbol{\sigma}}^{(b)}) \mid b = 1, \ldots, B\}$
```

j	$\widehat{f}_j^B$	$\widehat{\sigma}_j^B$	$\widehat{R}_j^B$	$\widehat{\text{MSEP}}(\widehat{R}_j)$
2	1.89	2.71	335.74	120.77
3	1.27	4.71	926.98	275.7
4	1.15	5.25	2004.1	350.68
5	1.1	3.24	3696.87	933.33
6	1.04	1.35	7096.49	1159.61
7	1.03	0.75	14301.84	1367.22

(a) Conditional

j	$\widehat{f}_j^B$	$\widehat{\sigma}_j^B$	$\widehat{R}_j^B$	$\widehat{\text{MSEP}}(\widehat{R}_j)$
2	1.89	3.1	350.99	9.34
3	1.27	3	1034.62	32.31
4	1.15	2.34	2066.53	122.15
5	1.1	1.05	3689.86	322.74
6	1.05	0.16	7049	540.6
7	1.03	0.05	14289.03	739.27

(b) Unconditional

Table 3.4: Parametric bootstrap results

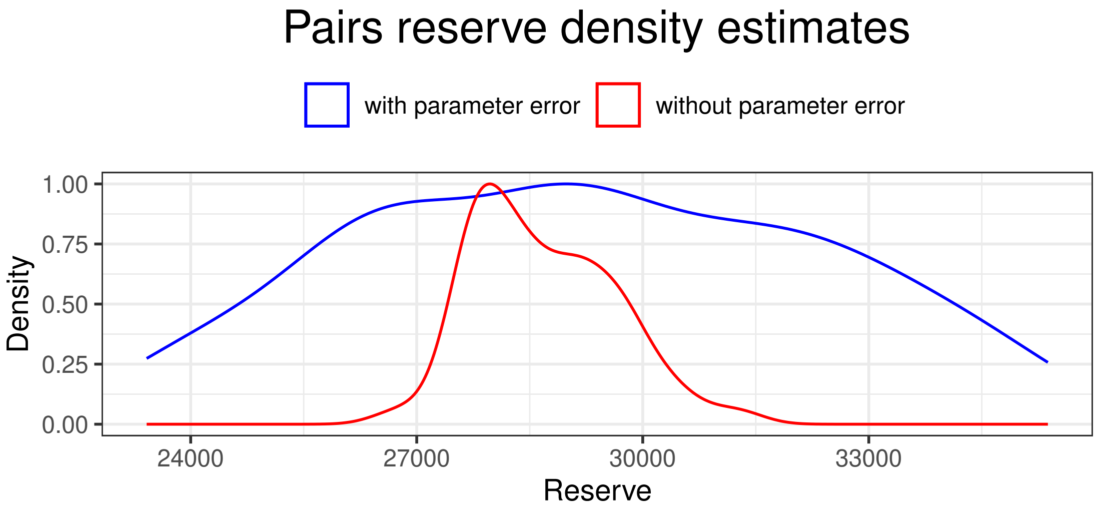


Figure 3.2: Comparison of the simulated distribution of the reserve with and without parameter error

3.4 Incorporating the process error

We end this chapter by discussing how the process error can be incorporated into these bootstrap procedures. As described in Section 2.3, our aim is to obtain a predictive distribution of the reserve which incorporates both parameter and process error. Following the procedure outlined there, we can achieve this by simulating the lower triangle $\mathcal{D}_I^\mathsf{c} = \{C_{ij} \mid i + j > I + 1\}$, giving us pseudo-realisations C_{ij} of the relevant claim amounts. This, in turn, yields bootstrap replicates

$$R^{(b)} := \sum_{i=2}^{I} (C_{iI}^{(b)} - C_{i,I+1-i}) \tag{3.57}$$

for the reserve. In view of (3.48), one way of achieving this would be to start from the antidiagonal of $\mathcal{D}_I$ and successively sample

$$C_{i,j+1}^{(b)} \sim \mathcal{N}(\widehat{f}_j^{(b)}\, C_{ij}^{(b)}, \widehat{\sigma}_j^{(b)}\, C_{ij})\,. \tag{3.58}$$

As with the parameter error in Section 3.3, however, we are faced with the problem of possibly drawing negative samples. Again, we could solve this by simply discard a simulated triangle as soon as it contains a negative value, but this can lead to computationally infeasible situations.

we can follow the suggestion given in [10, p. 238] and subtitute in place of (3.58) 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

$$\frac{\alpha}{\beta} = f_{j-1} C_{i,j-1} \quad \text{and} \quad \frac{\alpha}{\beta^2} = \sigma_{j-1}^2 C_{i,j-1}\,, \tag{3.59}$$

from which it follows that

$$\alpha = \frac{f_{j-1}^2 C_{i,j-1}}{\sigma_{j-1}^2} \quad \text{and} \quad \beta = \frac{f_{j-1}}{\sigma_{j-1}^2}\,. \tag{3.60}$$

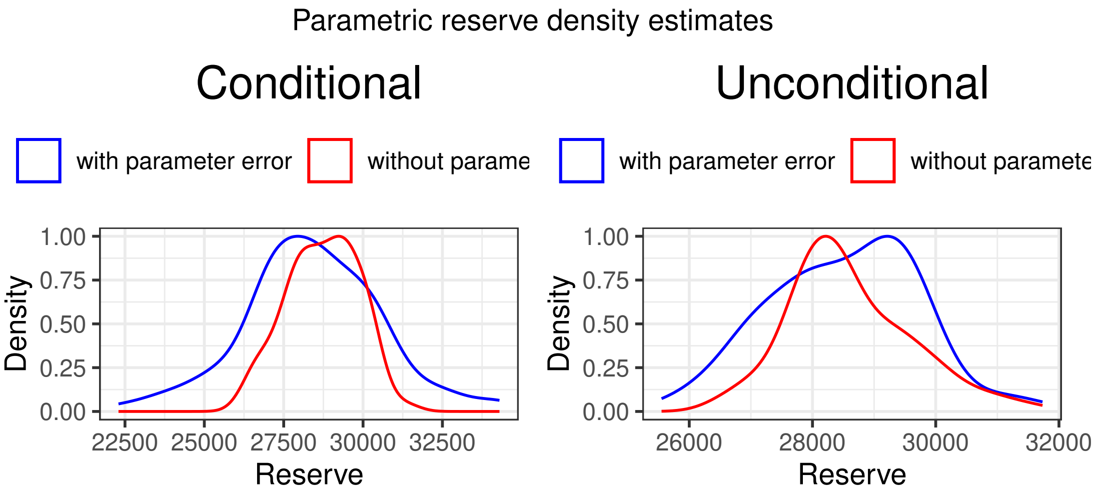


Figure 3.3: Comparison of the simulated distribution of the reserve with and without parameter error

Chapter 4

Poisson GLM

The (overdispersed) Poisson model (ODP), proposed by Renshaw and Verrall in [18], 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 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, $a_1, \dots, a_I$ and $b_1, \dots, b_I$ such that*

$$\log(\mathbb{E}[X_{ij}]) = c + a_i + b_j \,, \tag{4.1}$$

with $a_1 = b_1 = 0$.

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

$$X_{ij} \sim \mathrm{Pois}(e^{c+a_i+b_j})\,. \tag{4.2}$$

The condition $a_1 = b_1 = 0$ is necessary to obtain an identifiable model. Without it, any set of parameters $c, a_1, \dots, a_I, b_1, \dots, b_I$ satisfying the assumptions would yield an infinite number of alternatives $c + a_0 + b_0, a_1 - a_0, \dots, a_I - a_0, b_1 - b_0, \dots, b_I - b_0$ 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 $\xi_i := e^{c+a_i}$ and $\gamma_j := e^{b_j}$, we can obtain a different parametrisation of the model with a multiplicative structure for the mean,

$$\mathbb{E}[X_{ij}] = \xi_i \gamma_j \,, \tag{4.3}$$

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

$$\sum_{j=1}^{I} \gamma_j = 1 \tag{4.4}$$

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 [8, Lemma 2.16]).

4.1 Generalised linear models

GLMs were first conceived by Nelder and Wedderburn in [19] 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 $X_1, \dots, X_p$ 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

$$p(y \mid \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}, \tag{4.5}$$

   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 $\eta := \mathbf{x}^T \boldsymbol{\beta}$ 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,

$$\mu := \mathbb{E}[Y \parallel X_1, \dots, X_p] = g^{-1}(\eta). \tag{4.6}$$

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

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{y-\mu}{\sigma} \right)^2 \right\} = \exp\left\{ -\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma) \right\} \tag{4.7}$$

$$= \exp\left\{ \frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma) \right\}, \tag{4.8}$$

which is of the form eq. (4.5) 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 $l(\theta \mid y, \phi) := \log p(y \mid \theta, \phi)$ satisfies

$$\mathbb{E}\left[\frac{\partial l(\theta \mid Y)}{\partial \theta} \right] = 0, \qquad \mathrm{Var}\left(\frac{\partial l(\theta \mid Y)}{\partial \theta} \right) = -\mathbb{E}\left[\frac{\partial^2 l(\theta \mid Y)}{\theta^2} \right], \tag{4.9}$$

where $\frac{\partial l(\theta|Y)}{\partial\theta}$ is known as the *score function.* Using eq. (4.5), we then find that

$$\mathbb{E}\left[\frac{Y-b'(\theta)}{a(\phi)}\right]=0\,,\qquad \mathrm{Var}\left(\frac{Y-b'(\theta)}{a(\phi)}\right)=-\mathbb{E}\left[\frac{-b''(\theta)}{a(\phi)}\right],\tag{4.10}$$

from which we obtain the elegant relations

$$\mu=b'(\theta)\,,\qquad \mathrm{Var}(Y)=a(\phi)b''(\theta)\,.\tag{4.11}$$

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 $V(\mu):=b''((b')^{-1}(\mu))$ 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 [20, 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 eq. (4.5) 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

$$\widehat{\phi}:=\frac{1}{n-p}\sum_{i=1}^{N}\frac{(Y_i-\widehat{\mu_i})^2}{V(\widehat{\mu_i})}\,.\tag{4.12}$$

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

$$l(\boldsymbol{\beta}\mid\mathbf{y},\phi)=\sum_{i=1}^{N}\frac{y_i\theta_i-b(\theta_i)}{a_i(\phi)}+c(y_i,\phi)\,,\tag{4.13}$$

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

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta}\mid\mathbf{y},\phi)}{\partial\beta_j}&=\sum_{i=1}^{N}\frac{\partial l(\boldsymbol{\beta}\mid y_i,\phi)}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j} &(4.14)\\ &=\sum_{i=1}^{N}\frac{y_i-b_i'(\theta)}{a_i(\phi)}\frac{1}{b_i''(\theta_i)}\frac{\partial\mu_i}{\partial\eta_i}x_{ij} &(4.15)\\ &=\sum_{i=1}^{N}\frac{y_i-\mu_i}{\mathrm{Var}(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}x_{ij}\,, &(4.16)\end{aligned}$$

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

$$\mathcal{I}_{jk} \coloneqq \mathbb{E}\left[-\frac{\partial^2 l(\boldsymbol{\beta} \mid \mathbf{y}, \phi)}{\partial\beta_j \partial\beta_k}\right], \tag{4.17}$$

which is known as the *Fisher information matrix.* Thus, starting from an initial guess $\widehat{\boldsymbol{\beta}}^{(0)}$ for the parameters, we compute a successive approximations via

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} + \mathcal{I}(\widehat{\boldsymbol{\beta}}^{(k)})^{-1} \nabla l(\widehat{\boldsymbol{\beta}}^{(k)} \mid \mathbf{y}, \phi)\,. \tag{4.18}$$

Similarly to eq. (4.9), it can be shown that

$$\mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\beta} \mid \mathbf{y}, \phi)}{\partial\boldsymbol{\beta}\, \partial\boldsymbol{\beta}^T}\right] = -\mathrm{Var}\big(\nabla l(\boldsymbol{\beta} \mid \mathbf{y}, \phi) \nabla l(\boldsymbol{\beta} \mid \mathbf{y}, \phi)^T\big) < 0\,, \tag{4.19}$$

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

$$\begin{aligned}
I_{jk} &= \mathbb{E}\left[\left(\sum_{i=1}^{N} \frac{Y_i - \mu_i}{\mathrm{Var}(Y_i)} \frac{\partial\mu_i}{\partial\eta_i} x_{ij}\right)\left(\sum_{l=1}^{N} \frac{Y_l - \mu_l}{\mathrm{Var}(Y_l)} \frac{\partial\mu_l}{\partial\eta_l} x_{lk}\right)\right] &(4.20)\\
&= \sum_{i=1}^{N} \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \frac{\mathbb{E}\big[(Y_i - \mu_i)^2\big]}{\mathrm{Var}(Y_i)^2} x_{ij} x_{ik} &(4.21)\\
&= \sum_{i=1}^{N} \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \frac{x_{ij} x_{ik}}{\mathrm{Var}(Y_i)} &(4.22)\\
&= \mathbf{x}_{\boldsymbol{j}}^T \mathbf{W} \mathbf{x}_{\boldsymbol{k}} &(4.23)
\end{aligned}$$

where $\mathbf{W}^{(k)}$ is a diagonal matrix with

$$\mathbf{W}_{ii}^{(k)} = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2_{\widehat{\boldsymbol{\beta}}^{(k)}}\,. \tag{4.24}$$

Hence we have $\mathcal{I}(\widehat{\boldsymbol{\beta}}^{(k)}) = \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X}$ and we see from eq. (4.16) that

$$\nabla l(\boldsymbol{\beta} \mid \mathbf{y}, \phi) = \mathbf{X}^T \mathbf{W} \tilde{\mathbf{z}} \tag{4.25}$$

with $\tilde{\mathbf{z}}_i = (y_i - \mu_i)\left(\frac{\partial\eta_i}{\partial\mu_i}\right)$. Multiplying both sides of 4.18 by $\mathcal{I}(\boldsymbol{\beta}^{(k)})$ and using eqs. (4.23) to (4.25), we finally obtain

$$\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X} \widehat{\boldsymbol{\beta}}^{(k+1)} = \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{z} \tag{4.26}$$

with $\mathbf{z} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)} + \tilde{\mathbf{z}}$ 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 $\mathbf{x_1}, \ldots, \mathbf{x_{\boldsymbol{N}}}$ 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

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix}}_{\mathbf{X}^T} \underbrace{\begin{bmatrix} e^{\widehat{c}^{(k)}} & & & \\ & e^{\widehat{c}^{(k)}+\widehat{a}_1^{(k)}} & & \\ & & \ddots & \\ & & & e^{\widehat{c}^{(k)}+\widehat{a}_I^{(k)}+\widehat{b}_I^{(k)}} \end{bmatrix}}_{\mathbf{W}^{(k)}} \underbrace{\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} c \\ a_2^{(b)} \\ \vdots \\ a_I^{(b)} \\ b_2^{(b)} \\ \vdots \\ b_I^{(b)} \end{bmatrix}}_{\widehat{\boldsymbol{\beta}}^{(b)}}$$

$$= \underbrace{\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix}}_{\mathbf{X}^T} \underbrace{\begin{bmatrix} e^{c} & & & \\ & e^{c+a_1} & & \\ & & \ddots & \\ & & & e^{c+a_I+b_I} \end{bmatrix}}_{\mathbf{W}^{(k)}} \mathbf{z}$$

Figure 4.1: IRWLS equation for Poisson GLM in matrix form

(using, for example, the colexicographical ordering $(i, j) \mapsto jI + i$, i.e. column-major order). If we define the parameter vector

$$\boldsymbol{\beta} := \begin{bmatrix} c & a_2 & \cdots & a_I & b_2 & \cdots & b_I \end{bmatrix}^T , \tag{4.27}$$

then eq. (4.1) can be rewritten as

$$\log(\mu_{ij}) = c + a_i + b_j = (\mathbf{e_1} + \mathbf{e_{\boldsymbol{i}}} + \mathbf{e_{\boldsymbol{I+j-1}}})^T \boldsymbol{\beta} \tag{4.28}$$

where $\mathbf{e}_{\boldsymbol{k}}$ 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

$$\frac{\partial \mu_{ij}}{\partial \eta_{ij}} = e^{\eta_{ij}} , \tag{4.29}$$

from which, using eq. (4.24), we finally obtain

$$\mathbf{W}_{ii} = \frac{1}{e^{\eta_{ij}}} (e^{\eta_{ij}})^2 = e^{\eta_{ij}} , \tag{4.30}$$

giving us all the components of the IRWLS algorithm. Figure 4.1 shows in matrix form.

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 $\text{Pois}(\lambda)$ 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

$$\frac{\partial l(\boldsymbol{\mu} \mid \mathbf{y}, \phi)}{\partial \mu_j} = \frac{y_j - \mu_j}{\phi V(\mu_j)}, \tag{4.31}$$

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

$$Q_i(\mu \mid y_i, \phi) := \int_{y_i}^{\mu} \frac{y_i - u}{\phi V(u)}\, du, \tag{4.32}$$

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

$$Q(\boldsymbol{\mu} \mid \mathbf{y}, \phi) := \sum_{i=1}^{N} Q_i(\mu_i \mid y_i, \phi). \tag{4.33}$$

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*[1]. 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, $a_1, \ldots, a_I$ and $b_1, \ldots, b_I$ such that*

$$\log(\mu_{ij}) = c + a_i + b_j, \tag{4.34}$$

with $\mu_{ij} := \mathbb{E}[X_{ij}]$ and $a_1 = b_1 = 0$.

3. *There exists a parameter ϕ such that*

$$\text{Var}(X_{ij}) = \phi \mu_{ij}. \tag{4.35}$$

[1] It would actually 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.

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

$$\sum_{i=1}^{I} X_{ij} \geq 0 \tag{4.36}$$

is satisfied for $j \in \{1, \ldots, I\}$ (see [18, 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.

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. [21] and [10]), and so we can draw upon this material for our exposition.

As with Mack's model, we shall take Section 2.2 as our starting point. We distinguished there between nonparametric, semiparametric and parametric approaches to bootstrapping. Of these, the nonparametric bootstrap, which involves resampling predictor-response pairs from the original data, cannot be applied to the Poisson model. To understand why, recall from Equation (4.26) that the IRWLS algorithm fits a linear model at each step of the iteration in order to obtain a new estimate of the parameters, which requires the matrix on the left-hand side to be invertible. Hence, it follows that the design matrix must have full rank. We also saw in Equation (4.28) that the rows of $\mathbf{X}$ are binary vectors indicating the origin and development year (row and column in the claims triangle) to which an observation belongs. Because of the structure of the triangle, every observation corresponds

This leaves us with only the semiparameteric and parametric variants to consider.

For the semiparameteric 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 [22]). Consequently, a multitude of different residual types are available. We will consider three of these in particular.

The *Pearson residuals*

$$r_{ij} := \frac{X_{ij} - \widehat{\mu}_{ij}}{\sqrt{V(\widehat{\mu}_{ij})}}\,, \tag{4.37}$$

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 eq. (4.37) for the leverage of the observation, i.e.

$$\tilde{r}_{ij} := \frac{X_{ij} - \widehat{\mu}_{ij}}{\sqrt{V(\widehat{\mu}_{ij})(1 - h_{ij})}}\,, \tag{4.38}$$

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

$$\mathbf{H} = \mathbf{X}(\mathbf{XWX})^{-1}\mathbf{X}^T\mathbf{Wz} \tag{4.39}$$

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 eq. (4.31) by noticing that the mean parametrisation of the log-likelihood is maximised at $\boldsymbol{\mu} = \mathbf{y}$, so that the quantity

$$D(\mathbf{y}, \boldsymbol{\mu}) := \sum_{i=1}^{N} d(y_i, \mu_i) := 2\sum_{i=1}^{N}(l(y_i \mid y_i) - l(\widehat{\mu}_i \mid y_i)) \tag{4.40}$$

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

$$r_{ij} := \operatorname{sign}(x_{ij} - \mu_{ij})\sqrt{d(x_{ij}, \mu_{ij})}\,. \tag{4.41}$$

Finally, we consider a third type known as *quantile residuals* (see [23] for a general discussion), which are most easily explained for continuous response distributions. In that case, an elementary fact from probability theory states that

$$F(Y \mid \mu, \phi) \sim U(0,1)\,, \tag{4.42}$$

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

$$r_i := F(Y_i \mid \widehat{\mu}_i, \widehat{\phi}) \tag{4.43}$$

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

$$a_i := \lim_{y \uparrow y_i} F(y \mid \widehat{\mu}_i, \widehat{\phi}), \qquad b_i := F(y_i \mid \widehat{\mu}_i, \widehat{\phi})\,, \tag{4.44}$$

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

Chapter 5

Numerical implementation

Chapter 6

Results

Conclusion

Bibliography

[1] S. Christofides, "Section d5. regression models based on log-incremental payments," in *Claims Reserving Manual*, vol. 2, Institute of Actuaries, 1997.

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

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

[4] O. E. Barndorff-Nielsen and D. R. Cox, "Prediction and asymptotics," *Bernoulli*, vol. 2, no. 4, pp. 319–340, 1996, ISSN: 13507265.

[5] J. F. Lawless and M. Fredette, "Frequentist prediction intervals and predictive distributions," *Biometrika*, vol. 92, no. 3, pp. 529–542, 2005, ISSN: 00063444.

[6] D. C. M. Dickson, L. M. Tedesco, and B. Zehnwirth, "Predictive aggregate claims distributions," *The Journal of Risk and Insurance*, vol. 65, no. 4, pp. 689–709, 1998, ISSN: 00224367, 15396975. (visited on 04/18/2023).

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

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

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

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

[11] T. Mack, "Section d6. measuring the variability of chain ladder reserve estimates," in *Claims Reserving Manual*, vol. 2, Institute of Actuaries, 1997.

[12] 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, pp. 93–128, 2020. DOI: 10.1017/S174849951900006X.

[13] 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.

[14] 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, pp. 543–552, 2006. DOI: 10.1017/S051503610001463X.

[15] 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, pp. 554–565, 2006. DOI: 10.1017/S0515036100014653.

[16] 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, pp. 566–571, 2006. DOI: 10.1017/S0515036100014665.

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

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

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.