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

Chapter 1

Introduction

The defining characteristic of the insurance industry is the inverted nature of its production cycle. In manufacturing, commerce, transportation, etc. payment is usually received upon delivery of goods or services. By contrast, insurance products are paid up-front, long before the adverse events they provide protection against occur, if at all. Insurance contracts exchange fixed payments today (premia) for contingent claims in the future, and the insurer must set aside a sufficient portion of the former in order to cover the latter. The *claims reserving problem* involves forecasting the funds which will be needed to settle outstanding contracts as well as their uncertainty. This poses clear challenges from a risk-management perspective, especially in long-tailed lines of business, and makes careful and rigorous statistical modelling indispensible to the actuary. Failure to do so not only invites potentially burdensome regulatory scrutiny, but can also lead to solvency problems.

Chapter 2

Pattern break detection

2.1 Introduction

The chain ladder method ranks among the most frequently applied loss reserving techniques in insurance. Originally conceived as a purely computational algorithm, various models have since been proposed to cast it in a stochastic framework. Regardless of which one the reserving actuary chooses to employ, the central assumption he is obliged to make is that the development pattern observed in earlier cohorts is applicable to later ones. While this requirement seems eminently reasonable - all models ultimately rely on the past serving as a sufficiently reliable guide to the future - it turns out to be difficult to verify quantitatively in practice.

A special difficulty which arises in the actuarial context is the relatively small quantity of data which is typically available. This dearth of observations sharply constrains the efficacy of classical statistical tests. Historical data 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* or *accident year* and $1 \leq j \leq J$ the *development year*, so that

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

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

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

by forecasting the values in the lower triangle $\mathcal{D}_I^c$.

Our aim in this chapter is to use bootstrap simulation methods to investigate whether it is possible to detect structural breaks in the claims development pattern. We do this by examining the sensitivity of widely-used actuarial models to the deviations from their assumptions.

2.2 The bootstrap method

In its simplest form, the bootstrap is a statistical technique which employs simulation to compute estimates for the uncertainty of a given model. When applied correctly, it is an enormously

powerful method which can free us from the dilemma of either having to navigate convoluted analytical expressions which are most likely intractible, or being forced to make debilitating oversimplifications in order to make the calculations feasible. Fundamentally, bootstrapping is premised on the idea that the distribution of the observed samples serves as a good proxy for the population, and that we may therefore approximate i.i.d. sampling from the latter by resampling the former. Depending on their assumptions, we can broadly distinguish[1] between two classes of bootstrapping: *parametric* and *non-parametric*. The following example will illustrate both.

Suppose we are given a sample $X_1, \dots, X_n$ of independent and identical random variables drawn from a distribution F and wish to estimate a quantity $h(F)$ which is a function of it. Assume moreover that we have already fixed an estimator $\widehat{h(F)} = g(X_1, \dots, X_n)$. How can we quantify the uncertainty of our result? The non-parametric bootstrap does this by approximating F with the *empirical cumulative distribution function* (ECDF) defined by

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

We then use $\hat{F}$ to simulate new samples $X_1^{(b)}, \dots, X_n^{(b)}$, which simply corresponds to drawing with replacement from $X_1, \dots X_n^*$. For each of these B bootstrap samples, we can then compute $g^{(b)} = g(X_1^{(b)}, \dots, X_n^{(b)})$, which we think of as an approximate sample from the true distribution of $\widehat{h(F)}$. The uncertainty of our estimate could then for instance be quantified by

$$\frac{1}{B-1} \sum_{b=1}^{B} (g^{(b)} - \widehat{h(F)})^2$$

The parametric bootstrap follows the same logic, but makes the additional assumption that F has a known form which is fully determined by some parameters $\boldsymbol{\theta}$. We can then compute an estimate $\hat{\boldsymbol{\theta}}(X_1, \dots, X_n)$ from the available data and approximate $F_{\boldsymbol{\theta}} \approx F_{\hat{\boldsymbol{\theta}}}$. A random number generator can then be used to produce the simulated samples $X_1^{(b)}, \dots, X_n^{(b)}$ and $g^{(b)}$ needed to perform inference.

Uses the fitted model to generate new samples.

2.3 Mack's model

We begin by considering the distribution-free model due to Mack[2], which applies to the cumulative claims triangle $\mathcal{D}_I = (C_{ij})$, and makes the following assumptions:

Model 1 (Mack Chain Ladder).

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

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

for $1 \leq i \leq I$.

(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 $1 \leq i \leq I$.

[1] **davison**.
[2] **mack**.

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

Assumption (**??**) states that the conditional expectation of the cumulative claim amount in any period depends only on that of the previous period, and that this dependence is moreover linear. Assumption (**??**) makes a similar assertion about the conditional variance, and assumption (**??**) states that the rows of our data triangle represent randomly drawn sample paths.

Estimates for the ultimate claim amounts C_{iI} are obtained by substituting the chain ladder estimates $\hat{f}_j$ for the unknown development factors f_j in the expression for the conditional expectation, yielding

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

from which we obtain the reserve estimate

$$\hat{R} = \sum_{i=1}^{I} (\hat{C}_{i,I} - C_{i,I+1-i}) \,.$$

Observe that this is, in fact, a two-step process: the ultimate is estimated by the conditional mean, which we estimate in turn by plugging in $\hat{f}_j$.

There exist a number of approaches to defining a bootstrap procedure for Mack's model. Following the literature[3], we will describe them in the context of gauging the uncertainty of our chain ladder estimates, which provides a suitable backdrop to describe the different choices involved. As with any model, the error made by Mack's chain ladder can be quantified using the (conditional) *mean square error of prediction*

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

which, in line with previous remarks, we can divide into *estimation error* and *process error* as follows:

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

It is not difficult to see the analogy between this expression and the familiar bias-variance decomposition from classical statistics. Using the standard rules of conditional probability, we can

[3] **wuthrich:mse**.

then express the MSEP in terms of the individual accident years:

$$\begin{aligned}
\mathbb{E}\Big[(\hat{R}-R)^2 \;\Big\|\; \mathcal{D}_I\Big] &= \sum_{i=1}^{I} \operatorname{Var}(C_{iI} - C_{i,I+1-i} \,\|\, \mathcal{D}_I) \\
&\quad + \left(\sum_{i=2}^{I} \mathbb{E}\left[(\hat{C}_{iI} - C_{iI}) \,\middle\|\, \mathcal{D}_I \right] - \sum_{i=2}^{I} (\hat{C}_{iI} - C_{i,I+1-i}) \right)^2 \\
&= \sum_{i=2}^{I} \operatorname{Var}(C_{iI} \,\|\, \mathcal{D}_I) + \left(\sum_{i=2}^{I} \mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \hat{C}_{iI} \right)^2 \\
&= \sum_{i=2}^{I} \underbrace{\Big(\operatorname{Var}(C_{iI} \,\|\, \mathcal{D}_I) + (\mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \hat{C}_{iI}) \Big)}_{=: \operatorname{MSEP}_{C_{iI} \mid \mathcal{D}_I}(\hat{C}_{iI})} \\
&\quad + \sum_{\substack{2 \le i,j \le I \\ i \neq j}} (\mathbb{E}[C_{iI} \parallel \mathcal{D}_I] - \hat{C}_{iI})(\mathbb{E}[C_{jI} \parallel \mathcal{D}_I] - \hat{C}_{jI}) \,.
\end{aligned}$$

The problem of esitmating the total MSEP has thus been reduced to estimation of the MSEP for a single accident along with the additional cross terms. As the latter is not important the

2.3.1 Bootstrapping the parameter error

Using the definitions from the previous section, we can write the process error for a single accident year $i \in \{1, dots, I\}$ as

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

Notice that this expression contains the unknown development factors f_j and so cannot be computed directly. If we tried to estimate it by subtituting the $\hat{f}_j$'s for them, however, the result would be a constant 0, which is clearly not accurate. We must therefore find a way to express the variability of $\hat{f}_j$ around f_j. Until now, we have worked under the assumption that $\mathcal{D}_I$ is held fixed, which is problematic because, conditional on $\mathcal{D}_I$, $\hat{f}_j$ is a scalar. Relaxing this assumption therefore means reducing the dataset we condition on.

Several ways of achieving this have been proposed in the literature, see for example [**wuthrich:stoch**] and [**mack:var**]. We will focus on two of them in particular, the so-called *conditional* and *unconditional* approaches. They can be most easily understood if we think of conditioning as providing a 'recipe' for a particular kind of resampling, e.g. $\mathbb{E}[X \parallel Y]$ corresponds to sampling observations from X for a fixed value of Y and taking their average. Unsurprisingly, this point of view will allow us to transition easily from analytic formulae to a bootstrapping procedure.

If we define the subset of observation up to (and including) development year k as

$$\mathcal{B}_k := \{ C_{ij} \in \mathcal{D}_I \,|\, j \le k \} \,,$$

then the unconditional approach fixes $\mathcal{B}_{I+1-i}I$ in accident year i, resampling the observations in

$$\mathcal{D}_{I,i}^{O} \coloneqq \{C_{ij} \in \mathcal{D}_I \mid j > I+1-i\} \,.$$

completely, which leads to the estimate

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

where we used the fact that the $\hat{f}_j$'s are uncorrelated. If we upgrade Assumptions **??** to the stronger autoregressive Gaussian time series model

$$C_{i,j+1} = f_j C_{i,j} + \sigma_j \sqrt{C_{i,j}} \varepsilon_{i,j} \sim \mathcal{N}(f_j C_{i,j}, \sigma_j^2 C_{i,j}) \,,$$

then we can view **??** as resampling $\mathcal{D}_{I,i}^{O}$ completely, computing $\prod_{j=I+1-i}^{I-1} \hat{f}_j^2$ and taking the average.

By contrast, the conditional approach only allows us to vary a given factor $\hat{f}_j$ over the observations after j, so that every point in $\mathcal{D}_{I,i}^{O}$ (except the upper right corner) will be conditioned on for at least one $\hat{f}_j$ when resampling. Phrased in terms of (**??**), we keep the original C_{ij} fixed at every step of the time series and only resample the next value $C_{i,j+1}$. Denoting the probability measure induced by this process on $\mathcal{D}_{I,i}^{O}$ by $\mathbb{P}_{\mathcal{D}_I}^*$, our estimate for the parameter error in the conditional approach then becomes

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

We will implement both approaches in **??**.

2.3.2 Bootstrapping Mack's model

Although we introduced the bootstrap in **??** as a device for gauging the uncertainty of estimates, its scope is in reality far wider than this. The output of the bootstrap is in fact a complete *simulated distribution* which can be used for all kinds of inference, e.g. estimating higher moments, computing confidence intervals. Recall as well that we distinguished between the parametric and non-parametric bootstrap, which differ in the method used for resampling. In this section, we will see how both can be applied to Mack's model.

Our first concern is to identify a set of variables which are i.i.d. In the context of bootstrapping statistical models, these are invariably taken to be the residuals. Defining the individual development factors $F_{i,j+1} \coloneqq \frac{C_{i,j+1}}{C_{ij}}$

(**??**)

$$\hat{\varepsilon}_{ij} \coloneqq \frac{(F_{i,j} - \hat{f}_j)\sqrt{C_{i,j}}}{\sigma_j} \,.$$

Note that the true errors are standard normally distributed

$$\varepsilon_{ij} \coloneqq \frac{(F_{i,j} - f_j)\sqrt{C_{i,j}}}{\sigma_j} \sim \mathcal{N}(0,1) \,,$$

which suggests the following parametric bootstrap procedure:

Algorithm 1 Normal parametric bootstrap, Mack's model

Loss triangle $\mathcal{D}_I = (C_{ij})_{1 \le i,j \le I}$, number of iterations N

Simulated reserve sample $\hat{R}_1, \dots, \hat{R}_N$

procedure RESAMPLE
  for $i \leftarrow 1, I-1$ **do**
    $F_{ij} \leftarrow \dfrac{C_{i,j+1}}{C_{ij}}$
    $\hat{f}_j \leftarrow \dfrac{\sum_{i=1}^{I+1-j} C_{i,j+1}}{\sum_{i=1}^{I+1-j} C_{ij}}$
    $\hat{\sigma}_j \leftarrow \dfrac{1}{I-j} \displaystyle\sum_{i=1}^{I-j} C_{ij} \left(F_{ij} - \hat{f}_j\right)^2$
  end for
  $\boldsymbol{\varepsilon}^* \leftarrow$ RNORM($\mu = 0$, $\sigma = 1$, $n = (I + I^2)/2$)

The difficulty with the normal model is that it makes it possible to draw negative samples during the bootstrap simulation.

To remedy this, we follow the suggestion given in [**england:dist**] and additionally consider use a different distribution which still respects Mack's assumptions.

If we take for instance $C_{ij} \sim \Gamma(\alpha, \beta)$, then we must choose α, β to satisfy

$$\begin{cases} \frac{\alpha}{\beta} = f_{j-1} C_{i,j-1} \\ \frac{\alpha}{\beta^2} = \sigma_{j-1}^2 C_{i,j-1} \,, \end{cases}$$

giving the values

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

for the distribution parameters.

2.3.3 Numerical results

Conclusion