From the Chain Ladder to Individual Claims Reserving using Machine Learning techniques

Alessandro Carrato[1], Michele Visintin[2]

[1]alessandro.carrato@gmail.com

[2]michele.visintin@hotmail.com

March 25, 2019
v1.0 (ASTIN Colloquium 2019 Version)

Abstract

Recent years have seen the emergence of a lot of research on the application of machine learning techniques in actuarial science. In particular, there has been a noticeable amount of papers regarding machine learning applied to P&C Loss Reserving. Overall, there is a growing understanding that machine learning techniques provide better prediction accuracy of the outstanding liabilities compared to traditional methods. Nevertheless, the greater accuracy is offset by a difficult interpretations of results. This makes them in line of principle not suitable in a increasing regulated world, as it is the insurance business. Our objective is to show how we can introduce elements of machine learning into the traditional actuarial reserving methods in a gradual way. We strive to achieve a balance between predictive power and interpretability by introducing step-by-step new machine learning elements, with the possibility to simply start from the legacy paid/incurred datasets underlying the loss claim triangles without introducing any cumbersome data requirement or significant IT investment.

Keywords: claims reserving, Mack's CL model, individual claims reserving, machine learning, non-life insurance

Contents

Preface

Our starting point is the most classic of the actuarial reserving models, i.e. the Chain Ladder model. By means of interpretation of the Chain Ladder model as a constrained linear regression, in section 1 we show how we can apply an aggregated model to individual data. In section 2 we move a step further, removing some constraints and making use of clustering techniques to aggregate claims in homogeneous groups such that for each of them we can accept the validity of the chain-ladder assumptions. Subsequently, in section 3, we describe a model we developed and calibrated on real data that makes jointly use of paid and incurred data; in particular, we show how the clustering of the paid-incurred trajectories lets us obtain groups of very homogeneous data for which we can use a linear regression model. We conclude the paper by discussing potential extensions which make use of more powerful, tree-based regression techniques.

1 The Chain Ladder model as linear regression on individual loss data

This section shows how the Chain Ladder model can be reinterpreted as a constrained linear regression. This also satisfies the underlying assumptions of the Mack model [1]. Then, it is proved that a linear regression which yields the same estimates can be applied to individual loss data. Finally, suggestions to generalize the linear models are discussed; this topic is presented more in-depth in the following sections.

1.1 Recalling the Mack model

Let t be the evaluation date and let $C_{i,j},\ i = 0, 1, ..., t;\ j = 0, 1, ..., t$ be the cumulative payments, i.e. the total amount of payments generated by claims of accident year i up to time $i + j$; the following assumptions are adopted:

1. $(C_{0,0}, ..., C_{0,t}), ..., (C_{t,0}, ...C_{t,t})$ stochastically independent;
2. $(C_{i,0}, ..., C_{i,t})$ Markov chain for each accident year i;
3. There exist $f_0, ..., f_{t-1}$ strictly positive such that:

$$E(C_{i,j}|C_{i,0}, ..., C_{i,j-1}) = E(C_{i,j}|C_{i,j-1}) = C_{i,j-1}f_{j-1}; \tag{1.1}$$

4. There exist $\sigma_0^2, ..., \sigma_{t-1}^2$ strictly positive such that:

$$V(C_{i,j}|C_{i,0}, ..., C_{i,j-1}) = V(C_{i,j}|C_{i,j-1}) = C_{i,j-1}\sigma_{j-1}^2. \tag{1.2}$$

Mack proposes the following estimator for the ultimate cost:

$$\tilde{C}_{i,t} = C_{i,t-i} \prod_{j=t-i}^{t-1} \tilde{f}_{j-1}. \tag{1.3}$$

One can prove that the following are unbiased and uncorrelated estimators of $f_0, ..., f_{t-1}$:

$$\tilde{f}_{j-1} = \frac{\sum_{i=0}^{t-j} C_{i,j}}{\sum_{i=0}^{t-j} C_{i,j-1}}, \tag{1.4}$$

which, as Mack proved, are the same as the Chain Ladder model. Thus, the estimator of the reserve is:

$$\tilde{R}_i = \tilde{C}_{i,t} - C_{i,t-i}. \tag{1.5}$$

Mack then proposes the following estimators for the parameters $\sigma_0^2, ...\sigma_{t-2}^2$:

$$\tilde{\sigma}_{j-1}^2 = \frac{1}{t-j} \sum_{i=0}^{t-j} C_{i,j-1}(\frac{C_{i,j}}{C_{i,j-1}} - \tilde{f}_{j-1})^2, \tag{1.6}$$

where an estimate for σ_{t-1}^2 may be obtained by means of extrapolation.

1.2 Some notes on linear regression

Let consider this zero-intercept simple linear regression model:

$$Y_i = \beta X_i + \varepsilon_i,$$

where the following hold true:

1. $E(\varepsilon_i|X_i) = 0$;
2. $(X_i, Y_i), (X_j, Y_j)$ stochastically independent if $i \neq j$; and
3. $V(\varepsilon_i|X_i) = \alpha_i\sigma^2$, with $\alpha_i > 0, \sigma^2 > 0$.

Consistently with assumption 3), weights $\omega_i = \frac{1}{\alpha_i}$ are assigned and β is estimated by means of weighted OLS, i.e. the following quantity is minimized:

$$\sum_{i=1}^{n} \omega_i(y_i - \beta x_i)^2,$$

and therefore the estimate is:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \omega_i x_i y_i}{\sum_{i=1}^{n} \omega_i x_i^2}. \tag{1.7}$$

One can prove that the estimator

$$\tilde{\beta} = \frac{\sum_{i=1}^{n} \omega_i X_i Y_i}{\sum_{i=1}^{n} \omega_i X_i^2} \tag{1.8}$$

is unbiased and consistent.

1.3 The linear regression underlying the Mack model

In this section it is shown that the Mack model can be interpreted as a constrained linear regression.

Lemma 1.1. *Let us consider* $C_{i,j}$*:*

$$C_{i,j} = f_{j-1} C_{i,j-1} + \varepsilon_{i,j-1} \tag{1.9}$$

and assume the following:

I. $(C_{0,0}, ..., C_{0,t}), ..., (C_{t,0}, ...C_{t,t})$ *stochastically independent;*

II. $(C_{i.0}, ..., C_{i,t})$ *Markov chain for each* i*;*

III. $E(\varepsilon_{i,j-1} | C_{i,0}, ..., C_{i,j-1}) = 0$*;*

IV. $V(\varepsilon_{i,j-1} | C_{i,0}, ..., C_{i,j-1}) = C_{i,j-1} \sigma_{j-1}^2$*;*

then the assumptions above are equivalent to the the assumptions of the Mack model.

Proof. We start by proving that the Mack model assumptions imply the linear regression assumptions:

- I and II are the same 1) and 2) of the Mack model;
- Assumption 3) implies $C_{i,j} = f_{j-1} C_{i,j-1} + \varepsilon_{i,j-1}$ and III; and
- Assumptions 2), 3) and 4) imply IV: indeed $V(\varepsilon_{i,j-1} | C_{i,0}, ..., C_{i,j-1}) =^1$ $= V(C_{i,j} - f_{j-1} C_{i,j-1} | C_{i,j-1}) =^2 V(C_{i,j} | C_{i,j-1}) = \sigma_{j-1}^2 C_{i,j-1}$.

[1] 2) implies $V(\varepsilon_{i,j-1} | C_{i,0}, ..., C_{i,j-1}) = V(C_{i,j} - f_{j-1} C_{i,j-1} | C_{i,0}, ..., C_{i,j-1}) = V(C_{i,j} - f_{j-1} C_{i,j-1} | C_{i,j-1}) = V(\varepsilon_{i,j-1} | C_{i,j-1})$.

[2] In general, $V(Y - X|X) = V(Y|X) + V(X|X) - 2cov(X, Y|X)$, with $V(X|X) = 0$ and $cov(X, Y|X) = E(XY|X) - E(X|X)E(Y|X) = XE(Y|X) - XE(Y|X) = 0$

We now prove that the linear regression assumptions imply the Mack model assumptions:

- 1) and 2) are the same as I and II of the linear regression model;
- Assumption 3) trivially descends from $C_{i,j} = f_{j-1}C_{i,j-1} + \varepsilon_{i,j-1}$ and III; and
- $V(C_{i,j}|C_{i,j-1}) = V(f_{j-1}C_{i,j-1} + \varepsilon_{i,j-1}|C_{i,j-1}) = V(\varepsilon_{i,j-1}|C_{i,j-1}) =$
  $= \sigma^2_{j-1}C_{i,j-1}$

□

The estimates of $f_0, ..., f_{t-1}$ can be obtained from the known cumulative payments, i.e. those such that $i + j \leq t$; accordingly to 4), weights $\omega_i = \frac{1}{C_{i,j-1}}$ are assigned. By means of (1.8), the following estimators are derived:

$$\tilde{f}_{j-1} = \frac{\sum_{i=0}^{t-j} C_{i,j}}{\sum_{i=0}^{t-j} C_{i,j-1}} \tag{1.10}$$

These estimators are the same that Mack proposes and therefore give the same reserve estimates as the Chain Ladder deterministic model.
To estimate $\sigma^2_0, ..., \sigma^2_{t-2}$, the usual unbiased estimator based on the squared residuals is adopted. In this context, because weights have been introduce, the weighted residuals $\sqrt{\omega_i}(\hat{y}_i - y_i)$ are used. Thus, the estimates are:

$$\hat{\sigma}^2_{j-1} = \frac{1}{t-j}\sum_{i=0}^{t-j}\frac{1}{c_{i,j-1}}(c_{i,j} - \hat{f}_{j-1}c_{i,j-1})^2 = \frac{1}{t-j}\sum_{i=0}^{t-j} c_{i,j-1}(\frac{c_{i,j}}{c_{i,j-1}} - \hat{f}_{j-1})^2. \tag{1.11}$$

The estimator is equivalent to the estimator Mack proposes:

$$\tilde{\sigma}^2_{j-1} = \frac{1}{t-j}\sum_{i=0}^{t-j} C_{i,j-1}(\frac{C_{i,j}}{C_{i,j-1}} - \tilde{f}_{j-1})^2. \tag{1.12}$$

As in the Mack model, the estimate for σ^2_{t-1} cannot be obtained directly within the model, and an extrapolation may be considered.

1.3.1 The regression with the intercept

The interpretation of the Mack model, or Chain Ladder, as a linear regression suggests immediately how to generalize the model. As a very first step, an intercept can be added to the linear model:

$$C_{i,j} = \beta_{j-1} + f_{j-1}C_{i,j-1} + \varepsilon_{i,j-1}. \tag{1.13}$$

The model can be further expanded by introducing other regressors, or even by changing the structure of the regression function. As a very general model, one can assume:

$$C_{i,j} = f(X_{i,j-1}) + \varepsilon_{i,j-1}, \tag{1.14}$$

where $X_{i,j-1}$ is a vector of regressors (which can include, for example, all the past payments, information about the case reserves and related transformations) and f is the regression function, which can be estimated parametrically (linear regression, GLMs,...) or nonparametrically (tree based models).
In the next sections, it is showed how we can use this regression framework on individual loss data.

Remark. *The introduction of an intercept in the model can also be used to test the validity of the Chain Ladder assumptions by performing the following statistical test:*

$$H_0 : \ \beta_{j-1} = 0 \quad vs \quad H_1 : \ \beta_{j-1} \neq 0 \tag{1.15}$$

The test on the intercept is a known technique to assess the goodness of fit of the Chain Ladder model. To this extent, it is worth to mention Halliwell [13], who presents the topic as Chain Ladder "bias".

1.4 Regression on individual loss data

In this section the individual claims reserving framework is introduced. Instead of considering the claims on an aggregated basis, a model on individual claims payments is proposed. It is shown that this model yields the same estimates of the Chain Ladder (or Mack) model.
Let n_i be the number of claims of accident year i. We denote $C^h_{i,j},\ h = 1, ..., n_i$ the cumulative payment up to time $i + j$ of the $h-$th claim of accident year i, where:

$$C_{i,j} = \sum_{h=1}^{n_i} C^h_{i,j}. \tag{1.16}$$

As in the aggregate model, the cumulative payments $C^h_{i,j}$ where $i + j \leq t$, are known at evaluation date t; the others are the object of our forecasting.
If we consider the linear model:

$$C^h_{i,j} = f_{j-1} C^h_{i,j-1} + \varepsilon^h_{i,j-1},$$

we can prove that this model provides the same estimates of the aggregated one.
As before, weights $\omega^h_i = \frac{1}{C^h_{i,j-1}}$ are assigned and the parameters are estimated

accordingly to the aggregate model (as before, only claims of accident year $i \leq t-j$ are considered).

Remark. *Please note that for the initial development periods, $C^h_{i,j-1}$ may be nil and therefore ω^h_i could fail to exist. Therefore, when implementing the reserving algorithm, we can either ignore the claims where $C^h_{i,j-1} = 0$ or set $C^h_{i,j-1}$ to a relatively small number (eg. $10^(-5)$). As the second option is more pragmatic, to the extent of the paper we will consider only claims with cumulative non-nil.*

By means of (1.8) and (1.16), and assuming $\omega^h_i > 0$ the following estimators are derived:

$$\tilde{f}_{j-1} = \frac{\sum_{i=0}^{t-j}\sum_{h=1}^{n_i} C^h_{i,j}C^h_{i,j-1}\frac{1}{C^h_{i,j-1}}}{\sum_{i=0}^{t-j}\sum_{h=1}^{n_i}(C^h{}_{i,j-1})^2\frac{1}{C^h_{i,j-1}}} = \frac{\sum_{i=0}^{t-j} C_{i,j}}{\sum_{i=0}^{t-j} C_{i,j-1}} \tag{1.17}$$

which yield the same estimates of the Mack model.

Remark. *Also in this case, an intercept can be introduced. As it will be shown in the next sections, this is particularly useful in presence of null cumulative payments.*

1.5 Criticisms and remarks

1.5.1 Dishomogeneous claims

Mack model assumptions may be unacceptable if the claims considered are not homogeneous. In the actuarial common practice, claims are usually aggregated on a balance-sheet logic, which can lead to groups of very dishomogeneous claims. In particular, one have to pay attention to the separation of large claims from attritional ones. In the next section, it is showed how homogeneous claims can be aggregated using clustering methods.

1.5.2 Tails

In the model, it has been assumed that a claim cannot generate cash-flows after calendar year $t+i$. In practice, that may not be true. Thus, it could be necessary to extrapolate the cumulative payments $C^h_{i,j},\ j > t$ so that a prediction for the ultimate cost $C^h_{i,+\infty}$ can be obtained. In section 4, a method to cope with the tails on an individual basis is described.

2 Linear regression and clustering on individual loss data

In the previous section it has been remarked that the Chain-Ladder (or Mack) model may not be suitable if the claims are dishomogeneous. To solve this problem, in this section it is proposed to cluster the individual claims and to use this information to generalize the linear regression of the previous section. Matters regarding clustering procedures are, for the moment, postponed: it is assumed that, using the information available for each claim at calendar year $i+j-1$ (to estimate f_{j-1}, j is fixed and i depends on the claim considered), it is possible to identify m clusters such that two claims belonging to the same cluster can be considered similar and two claims belonging to different clusters can be considered different. A linear model is assumed for each group of claims:

$$C_{i,j}^{k_h} = \beta_{j-1}^{k} + f_{j-1}^{k} C_{i,j-1}^{k_h} + \varepsilon_{i,j-1}^{k_h}, \quad k = 1, ..., m; \; h = 1, ..., n_k, \tag{2.1}$$

where k indicates the cluster, k_h indicates the $h-$th claim which belongs to the $k-$th cluster in calendar year $i + j$.

Remark. *The idea of extending the regression underlying the Mack model was already proposed by Barnett and Zehnwirth [9]; in this section we show how to apply the same idea on individual loss data.*

2.1 Model assumptions and estimates

The following assumptions are considered:

1. $E(\varepsilon_{i,j-1}^{k_h} | C_{i,j-1}^{k_h}) = 0$;
2. Vectors of cumulative payments related to different claims are stochastically independent; and
3. $V(\varepsilon_{i,j-1}^{h_k} | C_{i,j-1}^{h}) = C_{i,j-1}^{h} \sigma_{j-1}^{2}$.

Weights $\omega_i^h = \frac{1}{C_{i,j-1}^h}$ are assigned and the estimates of the development factors are obtained by means of the weighted least squares method.

2.2 Forecast

2.2.1 One-step forecast

We consider a fixed development year $j = 1, ..., t$ and all the claims of accident year i such that $i + j - 1 = t$. The cumulative payments up to time $i + j - 1$ are

known; the first step is to forecast $C_{i,j}^{k_h}$.
We observe that, for a claim belonging to cluster k, the following equation holds true:

$$E(C_{i,j}^{k_h}|C_{i,j-1}^{k_h}) = \beta_{j-1}^{k} + f_{j-1}^{k} C_{i,j-1}^{k_h}. \tag{2.2}$$

It is therefore natural to consider as the predicted value of $C_{i,j}^{k_h}$ the following:

$$\hat{C}_{i,j}^{k_h} = \hat{\beta}_{j-1}^{k} + \hat{f}_{j-1}^{k} C_{i,j-1}^{k_h}. \tag{2.3}$$

Remark. *If the intercept is discarded, a model conceptually similar to the Chain-Ladder (or Mack) is obtained; however, in this model, the development factor may be different across different groups (eg. the development factors related to attritional claims may be different to the ones related to large claims). We assert that by clustering claims in homogeneous groups we can obtain estimates which are more reliable than the ones obtained with the aggregate model.*
Furthermore, we observe that, if a number of clusters equal to one is chosen (i.e. there is no clustering) and if intercept is discarded, the model yields the same estimates of the Chain-Ladder (or Mack) model.

2.2.2 Forecasting the ultimate cost

In the previous section, the estimates $\hat{C}_{i,j}^{k_h}$ have been obtained; this section shows how it is possible to estimate the ultimate cost $C_{i,t}^{h}$ and the reserves.Therefore, the estimate $\hat{C}_{i,j}^{k_h}$ is now considered as a known feature[3] of the claim in calendar year $i+j$; by means of that and of the other features, the claim is classified in one of the m clusters, which have already determined, of the model related to factors $f_j^1, ..., f_j^m$ (which have been estimated by means of the claims such that $i+j \leq t$). Supposing that the claim belong to cluster k in calendar year $i+j$, the following estimate is obtained:

$$\hat{C}_{i,j+1}^{k_h} = \hat{\beta}_0^j + \hat{f}_j^{k_j} \hat{C}_{i,j}^{h}. \tag{2.4}$$

Considering a single claim, we denote with k_j the cluster it belongs to in calendar year $i+j$; by means of iterative application of (2.4) for all development years j, the estimate of the ultimate cost is obtained:

$$\hat{C}_{i,t}^{h} = C_{i,t-i}^{h} \prod_{j=t-i}^{t-1} \hat{f}_j^{k_j} + \sum_{j=t-i}^{t-1} \hat{\beta}_0^j. \tag{2.5}$$

[3]Formally, if $\mathscr{F}_{i+j}$ is the sigma-algebra representing the information available in $i+j$, we say that the random variable $C_{i,j}^{h}$ is $\mathscr{F}_{i+j}$-measurable

Thus, we obtain the estimate of the individual reserve:

$$\hat{R}_i^h = \hat{C}_{i,t}^h - C_{i,t-i}^h, \tag{2.6}$$

and therefore the estimate of the total reserve:

$$\sum_{i=0}^{t} \sum_{h=1}^{n_i} \hat{R}_i^h. \tag{2.7}$$

2.3 Claims clustering

In this section it is shown how to find the clusters which have been used in the previous sections. Let us consider a fixed development period $j = 1, ..., t$ and the claims of accident years i such that $i + j \leq t$. In calendar year t, every claim is characterized by a $n-$dimensional vector of features, which we denote with $X_{h,i,j}$ (features of the $h-$th claim of accident year i in calendar year $i+j$). Some of these features may be static (i.e they do not change over time), while some others may be dynamic. We remark that the reason we consider different clusters for each development year is due to the dynamic features.
We cite as examples of static variables the line of business of the related policy, the affected guarantee, the number of people involved etc.; as examples of dynamic features, we cite the payment sequence, the case reserves sequence, the payments related to medical and legal expenses, the numbers of lawyers allocated to the claim etc. (for a formal and complete presentation of the available features, we refer to Wüthrich [3]). We remark that if we make use of dynamic features, we need to predict them as well, just like the cumulative payments (we refer againt to Wüthrich [3] for the details).
To cluster the claims, either the most common clustering techniques (e.g. k-means) or more sophisticated ones (gaussian mixtures) can be deployed. A great deal of attention must be paid to the choice of the number of clusters: it may be made either by means of common techniques found in literature (e.g., Hastie et al. [7]), or by choosing a priori a number of clusters consistent with one's objective (for example, if one simply would like to separate attritional claims from large ones,the number of clusters can be set equal to two). In the next section, a method to select the best number of clusters based on the predictive power of the model will be described.

2.3.1 Claims clustering with paid and case reserves data

As a basic setup, we recall the paid/case reserve approach suggested by Mack [10]. Let us fix development year j, which means that the objective is to estimate

$f_{j-1}^1, ..., f_{j-1}^m$. We consider all the claims of accident years i such that $i+j \leq t$ and we denote with $B_{i,j}^h$ the case reserve at year $i+j$ of the $h-$th claim of accident year i. For each claim we are considering, there is a vector

$$(C_{i,1}^h, ..., C_{i,j}^h, B_{i,1}^h, ..., B_{i,j}^h) \equiv P_{i,j}^h \tag{2.8}$$

called the paid-reserve trajectory. We can identify m clusters using, for example, the k-means algorithm.

Remark. *We note that, to remain consistent with the Mack markovian assumption, we can cluster using only the last paid-incurred information: indeed, if we use all the trajectory, we are implicitly assuming that the distribution of the future cumulative payments depend not only on the last one, but also on the previous ones.*

2.4 General considerations

2.4.1 Compatibility with legacy actuarial datasets

We remark that, if the features in the clustering phase consist only of paid and incurred information, this model requires no more data than those that actuaries already have. That makes the implementation of the model very easy in practice, because it does not require any change in how the data are collected and stored.

2.4.2 Why clustering?

One could ask why we should cluster the claims and then proceed to the regression when in fact it is possible to use the individual claims features directly as regressors. This choice is consistent with the aim of proceeding step by step: in this way a model which is conceptually similar to the Chain-Ladder is obtained and which is, as a matter of fact, a generalization of it.

2.4.3 Generalizing the regression

The next step, therefore, is to generalize the regression model. First of all, some individual claims features may be used as regressors (and as a result, avoiding or reducing the clustering phase). Then, more accurate regression models could be deployed. For example, one may identify large and attritional claims and then assume a Pareto distribution for the first ones and a gamma glm for the second ones. As a further step, "black-box" models for predicting the future payments can be introduced. This topic is further discussed in section 5.

3 A joint paid-incurred model on individual loss data

3.1 Introduction

In practice, especially for long tailed lines of business, actuaries do not rely only on paid data, but also on incurred one. Even though incurred data is used in the clustering phase, it is still possible that the predicted ultimate cost of a claim with a large case reserve for a claim with a nil last cumulative paid is small. For this reason, it may be appropriate to consider a model also for incurred data. We observe that the model we described in the previous section can be applied on incurred data as well. This section shows it is possible to jointly model paid and incurred data to obtain a single ultimate cost for each claim.
As before, a model on paid data is introduced:

$$C_{i,j}^{k_h} = \beta_{j-1}^{k} + f_{j-1}^{k} C_{i,j-1}^{k_h} + \varepsilon_{i,j-1}^{k_h} \tag{3.1}$$

In addition to this model, the associated incurred model is considered:

$$I_{i,j}^{k_h} = \beta_{j-1}^{k} + f_{j-1}^{k} I_{i,j-1}^{k_h} + \varepsilon_{i,j-1}^{k_h} \tag{3.2}$$

We remark that the clusters are the same for both the models. Furthermore, we specify that in this model only paid and incurred data are used (further regressors are introduced in section 4).
First of all, an easy procedure to decide whether to introduce an intercept into the model is presented. Then, it is described how to choose an appropriate number of clusters, so that to maximise the joint predictive power of the models. At the conclusion of the section an heuristic (but accurate) method to combine the ultimate costs obtained with the two models is discussed.

3.2 Model selection

3.2.1 Selecting whether to include the intercept

Let us, for the moment[4], fix a number of cluster equals to m for development period j. Furthermore, let us focus on the paid model (the method is the same for the incurred model).

[4]As it is showed in the next section, the procedure for selecting the optimal number of clusters is iterative, and at each step a decision about whether to include the intercept or not is taken.

To decide whether to include the intercept, we fit the model with intercept and the model without intercept. The two AIC are then compared and the model with the lowest AIC value is chosen. For linear regression models, the AIC has the following expression:

$$AIC = 2p + n\log(\hat{\sigma^2}), \tag{3.3}$$

where p is the number of parameters, n the number of observations used to fit the model and $\hat{\sigma^2}$ is the residuals squared sum divided by the number of observations. In this case, p is equal to 1 or 2 depending on whether the model has the intercept and n is the number of elements in the cluster.
The same procedure is then repeated on the incurred model.

Remark. *With this method, for a fixed development period, we can include the intercept in one model and not in the other. Our experience suggests that typically the models on incurred data do not require the introduction of an intercept. On the other side, models on paid data usually require the intercept for clusters including large claims (where often the Chain Ladder assumptions do not hold true), while it is often not included for clusters consisting of mainly attritional claims (where often the Chain Ladder assumptions hold true).*

3.2.2 Choosing the optimal number of clusters

For each cluster a decision about whether to include the intercept in the associated paid and incurred models has been taken. It is now necessary to choose the optimal number of clusters, so that the model has strong predictive power for both paid and incurred. At this point, it is possible to compute the total AIC of the model with m clusters.
We denote with $p_h^P \in \{1, 2\}$ the number of parameters of the paid model associated with the $h-$th cluster, and with $p_h^I \in \{1, 2\}$ the number of parameters of the incurred model associated with the $h-$th cluster. In this way, the number of parameters of the model with m cluster is $\sum_{h=1}^{m}(p_h^P + p_h^I)$. We can now compute the AIC for this model as stated in (3.3), where RSS is the sum of the RSSs of each model and n is the total number of claims which are used to fit the model.

Remark. *In this way, the two models are considered equivalent, ie. they have the same importance. In fact, this may not be desirable. Indeed, it is natural to trust more in incurred data for the recent accident years, where the ratio paid/incurred is still low, and to trust more paid data for older accident years. Following this logic, weights can be introduced*

$$\alpha_j = \frac{C_{j,t-j}}{I_{j,t-j}}, \tag{3.4}$$

together with a new target variabile

$$Y_{i,j}^{k_h} = \alpha_j C_{i,j}^{k_h} + (1 - \alpha_j) I_{i,j}^{k_h}. \tag{3.5}$$

The new RSS can be heuristically considered as the weighted average of the RSSs of the paid and incurred models, where the effective number of parameters is

$$\alpha_j p_h^P + (1 - \alpha_j) p_h^I. \tag{3.6}$$

This weighting is also consistent to what is proposed in the next section to combine the results of the two models together.

At this point, the AIC of the model with m clusters has been computed. The same procedure can be repeated for m in $1, ..., maxclusters$, where the last is a parameter that must be set in advance[5]. We then select the model with the lowest AIC.

Remark. *Instead of using the AIC, machine learning procedures usually require the use of cross validation. Our recommendation of using AIC is only driven by cost/benefit analysis, ie. we are currently observing that AIC selects very similar models to cross validation using a fraction of the computational power (and runtime). Furthermore, Stone [11] showed that cross-validation and AIC are asymptotically equivalent.*

3.3 Combining the two models

The prediction of the future payments and incurred data is the same as has been presented in the previous section. Thus, an ultimate cumulative paid amount with the paid models and an ultimate incurred amount with the incurred models are obtained. Therefore, we need to combine the two results into a unique, selected ultimate cost. One heuristic way to do this is to consider a weighted average of the two predictions. The underlying idea is that we trust more in incurred data for the recent accident years, when the ratio paid/incurred is still low, and vice versa. So, for a claim of accident year i, we select the ultimate cost defined as follows:

$$\alpha_i \hat{C^h}_{i,t} + (1 - \alpha_i) \hat{I^h}_{i,t}, \tag{3.7}$$

[5]From a theoretical point of view, this parameter should be set as high as possible. However, the computational cost of adding clusters can be very high, and our experience suggests that the optimal number of clusters is, typically, not very high.

where α_i is defined as in 3.4. We remark that 3.7 is consistent with 3.5.
Another possibility is to consider weights specific for each claim, for example:

$$\alpha_{i,j}^h = \frac{C_{i,t-i}^h}{I_{i,t-i}^h} \tag{3.8}$$

3.4 Tails

In section 1 it has been remarked the need to implement a model to extrapolate future payments beyond development period t. Therefore, we will introduce a model which applies to paid data only. This model yields a new ultimate cumulative payment, which then is averaged with the ultimate incurred obtained as described above.
For each claim, the individual development factors are defined as follows:

$$f_{i,j-1}^h = \frac{C_{i,j}^h}{C_{j-1}^h} \tag{3.9}$$

(to streamline the notation, we do not specify if the quantities are observed or estimated; furthermore, these ratios are not defined when $C_{j-1}^h = 0$ and $C_{i,j}^h > 0$ and are equal to 1 when both the quantities are zero).
These factors can be extrapolated with the same procedures used for the one related to the cumulative payments. We propose to fit a linear regression such that:

$$\log(f_{i,j}^h - 1) = \beta_{i,j}^h + \gamma_{i,j}^h j + \varepsilon_{i,j}^h. \tag{3.10}$$

Having fixed a tail projection period equals to m, we can predict $C_{i,t+m}^h$ as follows:

$$\hat{C}_{i,t+m}^h = \hat{C}_{i,t}^h [\prod_{k=t+1}^{m} (e^{\hat{\beta_{i,j}^h} + \hat{\gamma_{i,j}^h} k} + 1)]. \tag{3.11}$$

To avoid obtaining unrealistic results, some extreme development factors may be discarded from the fitting process.

4 Model extensions

In this section some potential extensions of the model are proposed. These can come into two forms: addition of more regressors (or features) and change of the regression model structure.

Remark. *In this section clusters are not used anymore: their introduction was due to the use of a constrained linear regression. By dropping every constrain and moving away from linear regressions, clustering is not needed anymore.*

In general, a structure of this kind can be assumed:

$$C_{i,j}^h = f(X_{i,j-1}^h) + \varepsilon_{i,j-1}^h \tag{4.1}$$

with the assumption:

$$E(\varepsilon_{i,j-1}^h | X_{i,j-1}^h) = 0 \tag{4.2}$$

where $X_{i,j-1}^h$ is a set of regressors related to the $h-$th claim of accident year i known and known in calendar year $i + j - 1$.
As in the previous section, an associated model on incurred data can be assumed. However, the need to consider two models together should decrease with the introduction of more information: if the features related to a claim are really significant to predict its ultimate cost, a paid-only model is sufficient to obtain good predictive power.

4.1 Making use of more features

In previous sections, only paid and incurred data were used. However, insurance companies collect far more information which can be relevant to predict the ultimate cost a of a claim.
As stated in section 2.3, some features can be static, whilst other ones can be dynamic. The introduction of static features into the model is straightforward, and it has been proved to be of great importance in improving the predictive power of the model, see Wüthrich [2] and Taylor [12]. The introduction of dynamic features, instead, requires their modeling (in a way similar to how paid and incurred data were modelled in the previous sections). This can lead to very computational demanding models and even to a reduction of the predictive power of the model, especially if the features are not so relevant to predict the ultimate cost.

4.2 Refining the regression structure

In the previous sections, the regression function f has been modeled as a linear function of the regressors. However, this choice is quite restrictive and may yield bad predictive results.
Given the high number of potential features which can be used, it may make sense to use some tree-based model (random forest, gradient boosting) to estimate the regression function f. Indeed, they operate an implicit selection of the relevant features; furthermore, compared to neural networks, they are also easy to fine tune.

References

[1] T. Mack (1993). *Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates*, Astin Bullettin 23, 213-221

[2] M. Wüthrich (2018). *Neural Networks Applied to Chain-Ladder Reserving*

[3] M. Wüthrich (2016). *Machine Learning in Individual Claims Reserving*, Swiss Finance Institute, Research Paper Series, N. 16-67

[4] M. Wüthrich, C. Buser (2017). *Data Analytics for Non-Life Insurance Pricing*

[5] Astin Working Group (2017). *Individual Claim Development with Machine Learning*

[6] Astin Working Group (2018). *Machine Learning and Traditional Methods Synergy in Non-Life Reserving*

[7] T. Hastie, R. Tibshirani, J. Friedman (2009). *The Elements of Statistical Learning*, Springer, Second Edition

[8] M. Wüthrich, M. Merz (2008) *Stochastic Claims Reserving Methods in Insurance*, Wiley

[9] G. Barnett, B. Zehnwirth (2000). *Best Estimates for Reserves*

[10] T. Mack (2002). *Schadenversicherungsmathematik*,
Verlag Versicherungswirtschaft, 2. Auflage

[11] M. Stone (1977). *An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion* , Journal of the Royal Statistical Society, Vol. 39, N. 1, pp 44-47

[12] G. Taylor, G. McGuire, J. Sullivan (2008). *Individual Claim Loss Reserving Conditioned by Case Estimates*

[13] L. J. Halliwell (2007). *Chain-Ladder Bias: Its Reason and Meaning* Variance 1:2, pp. 214-247.