Self-assembling insurance claim models using regularized regression and machine learning

Gráinne McGuire

Taylor Fry
Level 22, 45 Clarence Street
Sydney NSW 2000
AUSTRALIA

Greg Taylor

School of Risk and Actuarial Studies
University of New South Wales
UNSW Sydney, NSW 2052
AUSTRALIA

Hugh Miller

Taylor Fry
Level 22, 45 Clarence Street
Sydney NSW 2000
AUSTRALIA

August 2018

Abstract. The lasso is applied in an attempt to automate the loss reserving problem. The regression form contained within the lasso is a GLM, and so that the model has all the versatility of that type of model, but the model selection is automated and the parameter coefficients for selected terms will not be the same.

There are two applications presented, one to synthetic data in conventional triangular form, and another to real data.

The secret of success in such an endeavour is the selection of the set of candidate basis functions for representation of the data set. Cross-validation is used for model selection.

The lasso performs well in modelling, identifying known features in the synthetic data, and tracking them accurately. This is despite complexity in those features that would challenge, and possibly defeat, most loss reserving alternatives. In the case of real data, the lasso also succeeds in tracking features of the data that analysis of the data set over many years has rendered virtually known.

A later section of the paper discusses the prediction error associated with a lasso-based loss reserve. It is seen that the procedure can be readily adapted to the estimation of parameter and process error, but can also estimate one component of model error. To the authors knowledge, no other loss reserving model in the literature does so.

Keywords: bootstrap, cross-validation, GLM, feature selection, lasso, loss reserving, machine learning, regularized regression.

1. Introduction

Many claim data sets are modelled, and estimates of loss reserve produced, by means of simple statistical structures. The chain ladder model, and its simple derivatives, such as Bornhuetter-Ferguson, Cape Cod, etc. (Taylor, 2000; Wüthrich & Merz, 2008) may be singled out for special mention in this respect.

Other data sets are modelled by means of more complex statistical structures. For example, Taylor & McGuire (2016) describe in detail the application of Generalized Linear Models (**GLMs**) to claims data. This approach is especially suitable for data sets that contain features such that the chain ladder model is inapplicable.

More recently, interest has been growing in Machine Learning (Harej, Gächter & Jamal, 2017; Jamal et al., 2018). This category of model includes the Artificial Neural Net (**ANN**), which has been studied in earlier literature (Mulquiney, 2006), and shown to be well adapted to data sets with complex features, such as those modelled with GLMs.

The drawback of a GLM is that its fitting to a complex data set requires a skilled resource, and is time-consuming. Many diagnostics will require programming, much time will be absorbed by their review, and many iterations of the model will be required.

For a data set such as that described in Section 4.3.1, a total of 15-25 hours may be required, even assuming that all diagnostics have already been programmed, and excluding documentation. The cost of this will vary from place to place, but a reasonable estimate of the consultancy fee in a developed country might be US$5,000-10,000.

The time and cost might be cut if an ANN could be applied. However, the ANN model will not be transparent. It will provide only limited revelation of the claim processes and mechanics (i.e. superimposed inflation (**SI**)) generating the data set. Although the ANN might produce a good fit to past data, the absence of this information on claim processes might render their extrapolation to the future difficult. This can induce considerable uncertainty in any estimate of loss reserve.

A GLM, on the other hand, is a structured and transparent model that rectifies these shortcomings, but, as mentioned above, at a cost.

Regularized regression, specifically the **lasso**, provides a compromise between the GLM and the ANN, with the potential to achieve the best of both worlds. The purpose here is to establish a framework in which the lasso may be automated for application to claim modelling in such a way as to cut the modelling cost a small fraction of that required by a GLM, but with an output equivalent to that of a GLM, and with all of the latter's transparency.

In order to achieve these objectives, the model is to be **self-assembling** in the sense that, on submission of a data set to the lasso-based software, the model will assemble itself in a form that may be read, validated, understood and extrapolated exactly as would a GLM.

The lasso has found its way into recent actuarial literature, though not always in application to loss reserving. Gao & Meng (2018) apply a **Bayesian lasso** to loss reserving triangles, using B-splines as basis functions for development (column-wise) pattern. This enables them to model development patterns that vary over accident year.

Venter (2018) also uses a Bayesian lasso for loss reserving, and Li, O'Hare & Vahid (2017) and Venter & Şahin (2018) apply it to age-period-cohort models. This is the terminology used in mortality studies, but, as those authors point out, precisely the same concept arises in insurance claim modelling, where the translation is accident year-development year-payment year.

Section 3 describes the construction of a lasso model aimed at these objectives, after Section 2 deals with notational and other preliminaries. Section 4 illustrates numerically the application of the model to both synthetic and real data, and Section 5 discusses the prediction error associated with this type of forecasting. Section 6 examines a couple of possible areas of further investigation, and Section 7 closes with some concluding remarks.

2. Framework and notation

Many simple claim models use the conventional **data triangle**, in which some random variable of interest Y is labelled by **accident period** $i = 1,2, \ldots, I$ and **development period** $j = 1,2, \ldots, I - i + 1$. In this set-up, the combination (i,j) is referred to as a **cell** of the triangle, and the quantity Y observed in this cell denoted Y_{ij}. The durations of the accident and development periods will be assumed equal, but need not be years. As a matter of notation, $E[Y_{ij}] = \mu_{ij}, Var[Y_{ij}] = \sigma_{ij}^2$. A realization of Y_{ij} will be denoted y_{ij}.

The real data set used in this paper consists of individual claim histories. These are capable of representation in the above triangular form but, for most purposes, this is unnecessary. However, some aggregation of claims occurs, as explained in Section 4.3.1, simply in order to reduce computation. The quantity of interest throughout will be individual claim size at claim finalization, converted to constant dollars on the basis of wage inflation. A claim often involves multiple payments on various dates. The indexation takes account of the different dates.

Claim size for individual claim k will be denoted $Y_{[k]}$. A vector $v_{[k]}$ of labelling quantities will also be associated with claim k. These quantities may include $i, j, t = i + j - 1 =$ calendar period in cell (i, j), and others.

One non-routine quantity of interest later is **operational time (OT)** at the finalization of a claim. OT was introduced to the loss reserving literature by Reid (1978), and is discussed by Taylor (2000) and Taylor & McGuire (2016).

Let the OT for claim k be denoted $\tau_{[k]}$, defined as follows. Assume that claim k belongs to accident period $i_{[k]}$, and that $\hat{N}_{i_{[k]}}$ is an estimator of the number of claims incurred in this accident period. Let $F_{i_{[k]}:[k]}$ denote the number of claims from the accident period finalized up to and including claim k. Then $\tau_{[k]} = F_{i_{[k]}:[k]} / \hat{N}_{i_{[k]}}$. In other words, $\tau_{[k]}$ is the proportion of claims from the same accident period as claim k that are finalized up to and including claim k.

There will be frequent reference to **open-ended ramp functions**. These are single-knot linear splines with zero gradient in the left-hand segment and unit gradient in the right-hand segment. Let $R_K(x)$ denote the open-ended ramp function with knot at K. then

$$R_K(x) = max(0, x - K) \tag{2.1}$$

For a given condition c, define the **indicator function** $I(c) = 1$ when c is true, and $I(c) = 0$ when c is false.

3. Regularized regression

3.1. In general

Consider an ordinary least squares regression model

$$y = X\beta + \varepsilon \tag{3.1}$$

where $y = (y_1, \dots, y_n)^T$ is the response vector, $\beta = (\beta_1, \dots, \beta_p)^T$ the parameter vector, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ an error vector subject to $\varepsilon \sim N(0, I)$, and X the $n \times p$ design matrix. The parameter vector β is estimated by that $\hat{\beta}$ which minimizes the loss function $L(y; \hat{\beta}) =$

$\left(y - X\hat{\beta}\right)^2$. It will be convenient to express this as $L(y;\hat{\beta}) = \left(\|y - X\hat{\beta}\|_2\right)^2$, where, for vector x with components x_m, $\|x\|_q$ denotes the L_q-norm $(p \geq 1)$

$$\|x\|_q = \left(\sum_m |x_m|^q\right)^{1/q}$$

Regularized regression includes in the loss function a penalty for large regression coefficients (components of $\hat{\beta}$. The regularized loss function is

$$L(y;\hat{\beta}) = \left(\|y - X\hat{\beta}\|_2\right)^2 + \lambda\left(\|\hat{\beta}\|_q\right)^q \tag{3.2}$$

for some selected q, where $\lambda > 0$ is a tuning constant.

The generalization of (3.1) to a GLM assumes the form

$$y = h^{-1}(X\beta) + \varepsilon \tag{3.3}$$

where h is the (possibly non-linear) link function, h^{-1} operates component-wise on the vector $X\beta$, and ε may now be non-normal, but still with $E[\varepsilon] = 0$ and independence between observations still assumed. The parameter vector β is estimated by the maximum likelihood estimator $\hat{\beta}$ which minimizes the loss function (negative log-likelihood)

$$L(y;\hat{\beta}) = -\sum_{m=1}^{n} l(y_m;\hat{\beta}) \tag{3.4}$$

where the summand is the log-likelihood of observation y_i when $\beta = \hat{\beta}$.

The regularized form of (3.4) is similar to (3.2):

$$L(y;\hat{\beta}) = -\sum_{m=1}^{n} l(y_m;\hat{\beta}) + \lambda\left(\|\hat{\beta}\|_q\right)^q \tag{3.5}$$

The effect of variation of λ is as follows. As $\lambda \to 0$, (3.5) tends to revert to the unregularized form of GLM. As $\lambda \to \infty$, the penalty for any non-zero coefficient increases without limit, forcing the model towards one with a single coefficient, in which case the model consists of just an intercept term and all predictions are equal to $\bar{y}$, the mean response.

3.2. The lasso

3.2.1. Loss function

There are two recognizable special cases of (3.2):

(a) $\boldsymbol{q = 0}$. Here the second member of (3.2) reduces to just λ multiplied by the number of nonzero terms in the model. This is a computationally intractable problem, although it is noteworthy that setting $\lambda = 1$ produces a metric equivalent to the Akaike Information Criterion.

(b) $\boldsymbol{q = 2}$. In this case, (3.2) is recognized as the **Ridge regression** (Bibby & Toutenburg, 1977) loss function. All included covariates will be nonzero in this approach.

A further case of interest arises when $q = 1$. This is the **Least Absolute Shrinkage and Selection Operator (lasso)** (Tibshirani, 1996), which is the modelling device used throughout this paper. Its loss function (3.5) may be written explicitly as

$$L(y;\hat{\beta}) = -\sum_{m=1}^{n} l(y_m;\hat{\beta}) + \lambda\|\hat{\beta}\|_1 = -\sum_{m=1}^{n} l(y_m;\hat{\beta}) + \lambda\sum_{r=1}^{p}|\hat{\beta}_r| \tag{3.6}$$

The appearance of absolute values of coefficients in the loss function will generate many corner solutions when that function is minimized. Hence, the covariates included in the model will be a subset of those included in the regression design, and the lasso acts (as its full name suggests) as both a selector of covariates and as a shrinker of coefficient size. Efficient algorithms such as Least Angle Regression (Efron et al, 2004), which generates a path of solutions across all λ, make the lasso computationally feasible.

The strength of the shrinkage of covariate set can be controlled with the tuning parameter λ, as discussed at the end of Section 3.1, with the number of covariates decreasing as λ increases.

An obvious generalization of loss function (3.6) is

$$L(y;\hat{\beta}) = -\sum_{m=1}^{n} l(y_m;\hat{\beta}) + \sum_{r=1}^{p} \lambda_r|\hat{\beta}_r| \tag{3.7}$$

where the $\lambda_r (> 0)$ may differ with r. This generalization will be discussed further in Sections 3.2.2 and 4.3.3.

3.2.2. Lasso interpretations

The lasso has been presented in Section 3.2.1 rather in heuristic terms. However, it is useful for some purposes (see e.g. Section 4.3.3) to consider specific models for which the lasso is the optimal estimator in some sense. There are two main interpretations.

Bounded coefficients. Consider the optimization problem:

$$\hat{\beta} = \underset{\sum_{r=1}^{p}|\hat{\beta}_r| \le B}{\arg\min} \sum_{m=1}^{n} l(y_m;\hat{\beta})$$

i.e. maximum likelihood subject to the bound $\sum_{r=1}^{p}|\hat{\beta}_r| \le B$.

This problem is soluble by the method of Lagrange multipliers, then (3.6) is the Lagrangian, and λ the Lagrange multiplier. So the solution is as for the lasso.

It is straightforward to show by a parallel argument that, if the problem is modified to the following:

$$\hat{\beta} = \underset{|\hat{\beta}_r| \le B_r, r=1,\dots,p}{\arg\min} \sum_{m=1}^{n} l(y_m;\hat{\beta})$$

then the solution is as for the lasso with loss function (3.7).

Bayesian interpretation. Suppose that each parameter $\beta_r, r = 1, \dots, p$ is subject to a Bayesian prior that follows a centred Laplace distribution with density

$$\pi(\beta_r) = (2s)^{-1} \exp(-|\beta_r|/s) \tag{3.8}$$

where s is a scale parameter. In fact, $E[\beta_r] = 0, Var[\beta_r] = 2s^2$. All priors are supposed stochastically independent.

The posterior negative log-density of β is

$$-\sum_{m=1}^{n} l(y_m; \beta) + s^{-1} \sum_{r=1}^{p} |\beta_r| \tag{3.9}$$

up to a normalizing coefficient, and this is the same as (3.6) with $\lambda = 1/s = (½Var[\beta_r])^{-½}$. The lasso, in minimizing this function, maximizes the posterior density of β, so that $\hat{\beta}$ is the **maximum a posteriori (MAP)** estimator of β.

It is straightforward to show by a parallel argument that, if the prior (3.8) is replaced by

$$\pi(\beta_r) = (2s_r)^{-1} \exp(-|\beta_r|/s_r) \tag{3.10}$$

then the posterior negative log-density of β becomes

$$-\sum_{m=1}^{n} l(y_m; \beta) + \sum_{r=1}^{p} s_r^{-1} |\beta_r| \tag{3.11}$$

and MAP estimation lasso is equivalent to the lasso with loss function (3.7) and $\lambda_r = 1/s_r$.

3.2.3. Covariate standardization

Standardization of covariates, and hence associated coefficients, is often moot in regression analysis. In the case of regularized regression, however, it is usually desirable. Without it, the true values of some coefficients will be large and some small, but the penalty applied in (3.6) will be the same for all cases.

This would favour the exclusion of covariates whose coefficients large by their nature. And it would lead to the undesirable situation in which the regularized model would change as a consequence of mere re-scaling of covariates.

It is assumed that all covariates in a regression take numerical values with physical meaning, or are categorical. A categorical variate is converted into a collection of Boolean variates for the purpose of the regression.

Covariates are also classified as of three types:

- fundamental;
- derived;
- interaction.

Fundamental covariates do not depend on others (e.g. accident period), while derived variates are functions of fundamental variates (e.g. $max(0, accident\ year - 2003)$). It will be assumed that any derived covariate depends on just one fundamental covariate, and the latter will be called its **parent**.

Three types of standardization were tested for the present investigation. All took the form

$\bar{\bar{x}}_{ms} = x_{ms}/f_s$

where

- $\bar{\bar{x}}_{ms}$ is the standardized form of x_{ms}, the (m, s) element of the unstandardized design matrix X, i.e. value of the s-th covariate associated with the m-th observation; and
- f_s is a scaling constant for given s.

Further, let $\bar{x}_{ms}$ is the centred version of x_{ms}, i.e.

$$\bar{x}_{ms} = x_{ms} - \eta_s, \eta_s = \sum_{\ell=1}^{n} x_{\ell s}/n$$

and let $\bar{X}$ denote the design matrix after centring of covariates and excluding any column relating to an intercept coefficient.

Note that standardization here consists of scaling but not centring. The three versions of scaling constant assumed the following forms (nomenclature in the first two cases from Sardy, 2008):

$\boldsymbol{\rho}$ **− scaling**: $f_s = \left[\sum_{\ell=1}^{n} \bar{x}_{\ell s}{}^2/n\right]^{½}$.

$\boldsymbol{\Sigma}$ **− scaling**: $f_s = d_s^{-½}$, where d_s is the s-th diagonal element of $\Sigma = (\bar{X}^T\bar{X})^{-1}$.

$\boldsymbol{\eta}$ **− scaling**: $f_s = \eta_s$.

For all Boolean covariates, the scaling factors f_s are as set out here. Similarly, for non-Boolean fundamental covariates. A non-Boolean derived covariate takes on the same scaling factors as its parent.

Of the three forms of standardization, that generated by $\rho -$ scaling appeared most effective in terms of goodness-of-fit, and is the only one pursued further here.

3.2.4. Cross-validation

The parameter λ in (3.6) is free, and its value must be selected with suitable compromise between goodness-of-fit (small λ) and parsimony of parameters (large λ). A common approach to this selection is ***k*-fold cross-validation** (Hastie, Tibshirani & Friedman, 2009). This procedure is as follows:

(1) Select a maximum value of λ to be considered
(2) Run the lasso on all the data to generate a sequence of λ values to be considered
(3) Randomly partition the data set into k (roughly) equal subsets.
(4) Select one of these subsets as the validation set, and aggregate the remaining $k - 1$ subsets to form the training set.
(5) Fit the model to the training set.
(6) Compute a goodness-of-fit statistic for this model. In the present exercise, this is the **Poisson deviance** $\sum_m (y_m \ln \hat{y}_m - \hat{y}_m)$, where m runs over all observations y_m in the validation set, and $\hat{y}_m$ is the value fitted by the model to the m-th observation. Note that this is an out-of-sample comparison.

(7) Repeat steps (5) and (6) $k-1$ times, in each case using a different one of the k subsets as the validation set.
(8) Average the k goodness-of -fit statistics thus obtained and produce an average error and standard deviation over the folds. This is referred to as the **CV error**.
(9) Select a new (lower) value of λ, and repeat steps (3) to (8). Continue until the selected λ attains its chosen minimum value.

Note that the results of this cross-validation process are random due to the partitioning of the data into k subsets. This randomness may be reduced by running the cross-validation many times and averaging the error curves; however, this has not been carried out here.

At this point, one has obtained a collection of ordered pairs $(\lambda, \text{CV error})$, and can plot CV error against λ. The curve will usually take a vaguely U-shaped form, with high values λ resulting in poor fit, and low values of λ resulting in over-fit. The optimal value of λ may be taken as $\lambda = \lambda_{min}$, that which minimizes CV error. The model based on this value of λ will be referred to as the **min CV model**. For the purpose of the present investigation, $k = 8$.

An alternative choice of optimal λ, intended to recognize the sampling error in the CV error, proceeds as follows. In step (8) of the above procedure, the standard error of the goodness-of-fit statistic replications is also calculated. Let the value obtained from the model defined by $\lambda = \lambda_{min}$ be denoted se_{min}. The optimal value of λ is then taken as $\lambda = \lambda_{se}(> \lambda_{min})$, the least λ for which CV error $< \lambda_{min} + se_{min}$. Evidently, the requirement $\lambda_{se} > \lambda_{min}$ implies that the model with $\lambda = \lambda_{se}$ will be more parsimonious than the min CV model.

This alternative was tested for the data sets discussed here, but the resulting models appeared overly parsimonious, tending to model inefficiently features that were known to exist in the data. As a compromise, other intermediate models were tested, in which the optimal λ was selected as the least for which CV error $< \lambda_{min} + \gamma \times se_{min}$ for various $0 < \gamma < 1$, e.g. $\gamma = ½, ⅓, ¼$. Again, however, such models appeared to confer no advantage over the min CV model. Henceforth, only min CV models will be considered.

3.2.5. Other error diagnostics

There are a couple of other metrics of fitting error of interest. These are based on a model trained on the entire data set. These are:

- The **training error**, defined as the mean squared error, $\sum_m (y_m - \hat{y}_m)^2 / \hat{y}_m$ where the $\hat{y}_m$ are taken from the model trained on the entire data set, and the summation runs over all m in that data set; and
- The Akaike information criterion **(AIC)**.

The first of these is a goodness-of-fit test only, with no regard to the predictive power of the model. The second includes an over-parameterization penalty, and so endeavours to provide an indication of predictive power.

In the application of the lasso, although the CV error is taken as the most efficient criterion for selection of the tuning parameter λ, training error and AIC are routinely computed and displayed in results obtained from synthetic data (Section 4.2.2).

4. Application of the lasso to claim modelling

4.1. Preamble on covariate selection

Section 4.2 studies the application of the lasso to a number of synthetic data sets, while Section 4.3 studies its application to a well understood real data set. But, before that, some general comment on covariate selection is pertinent.

A typical claim data set will cover two independent time dimensions, accident and development periods ("row and column effects"), and a third dimension, payment period ("diagonal effects"), which is constructed from the other two. A necessary decision in any modelling of the data set will be the number of dimensions to be included within covariates.

It will usually be essential to include at least two. Moreover, some data sets exhibit all three effects: accident, development and payment. Nonetheless, inclusion of a third dimension should always be approached with caution. It clearly introduces collinearity between covariates, and therefore the potential for an ill-behaved model and poor prediction.

Further, the number of potential terms in a lasso linear predictor can become very large in models that include all possible covariates plus interactions, with obvious implications for computational load.

Even when the model fits well to data, the inclusion of all three effects may lead to mis-allocation between them. For example, data containing SI as a payment period effect, but no accident period effect, may return a model in which SI is allocated partly to each. The ramifications of this for forecasts require careful consideration. More on that shortly.

In the meantime, one may note that collateral information, external to the model itself, might exist with a bearing on the covariates to be included in the model. If, for example, there existed such information, indicating a high likelihood of payment quarter effects but low likelihood of accident quarter effects, then payment quarter might be included as a model covariate and accident quarter excluded.

This preliminary selection of covariates is referred to as **feature selection**. It may not be essential in some cases, but is a means of exploiting collateral information to reduce computational load.

Section 4.2 approaches the synthetic data set as if blind to any features. It therefore employs a **saturated lasso**, in which accident quarter **(AQ)**, development quarter **(DQ)** and payment quarter **(PQ)** effects are all included in the models, together with all possible interactions of a defined type.

On the other hand, the real-world example of Section 4.3 comes accompanied by a legislative background, involving major changes to the conditions affecting claim sizes, with these changes taking effect from a particular accident date. Any potential AQ effects are muted by the fact that this is a model of average claim sizes, and so AQ effects due to changing exposure are absent.

There are no other known major AQ effects, so feature selection has been employed, with AQ excluded from main effects and included in only a few very specific interactions related to the legislative change. PQ effects are included in full.

As already mentioned, decisions on feature selection carry ramifications for forecasts. Intuition may be derived here from the operation of the separation method (Taylor, 1977). Consider a claim triangle constructed to contain only the column and diagonal effects contained in the separation method and model the resulting triangle with a GLM containing only these effects. Call this "Model $GLM1$". Suppose the diagonal effect is an increase of 5% from each diagonal to the next. The model will produce an estimate of this.

Now model the same triangle with the same GLM, augmented to contain a row effect. Call this "Model $GLM2$". This model may mis-allocate the diagonal effect, partitioning it between rows and diagonals, say 3% per diagonal and 1.94% per row ($1.03 \times 1.94 = 1.05$).

Models $GLM1$ and $GLM2$ are distinctly different. However, their forecasts will be identical if the $GLM1$ forecast assumes 5% inflation per future diagonal, and the $GLM2$ forecast assumes 1.94%.

The point of this discussion is that it hints at the following proposition.

Proposition. Consider a data set containing DQ and PQ effects but no AQ effect. Let $M1$ denote a model that contains explicit DQ and PQ effects but no AQ effect, and let $M2$ denote a model that is identical except that it also contains explicit AQ effects. Then, in broad terms, $M1$ and $M2$ will generate similar forecasts of future claim experience if each extrapolates future PQ effects at a rate representative of that estimated for the past by the relevant model. ■

The saturated lasso of Section 4.2 corresponds to model $M2$, and so forecasts from that model incorporate future PQ effects at rates consistent with those estimated for the past, despite that fact that mis-allocation might have caused their under-statement.

The lasso of Section 4.3 largely excludes AQ effects, and so corresponds to model $M1$. There is little scope for mis-allocation of PQ effects, and the model estimates of these are genuine. In accordance with the proposition, they may be extrapolated to the future at face value. Alternatively, the allowance for future inflation may be reduced to zero, giving estimates in the currency values of the valuation date (closing date of last observed diagonal). This latter course is the one followed in the forecasts of Section 4.3.

4.2. Synthetic data

4.2.1. Data sets

The lasso was first applied to several synthetic data sets. The advantage of this is that the features of the data sets are known, and one is able to check the extent to which the lasso identifies and reproduces them. The data sets can be manufactured to include specific features identifiability of which is marked by differing degrees of difficulty.

Four synthetic data sets were constructed and analysed. Each consisted of a 40×40 quarterly triangle of incremental paid claims Y_{ij}. In each case, it is assumed that Y_{ij} is subject to a log normal distribution with mean μ_{ij} and variance σ_{ij}^2:

$$Y_{ij} \sim logN\left(ln\,\mu_{ij} - ½\tau_{ij}^2, \tau_{ij}^2\right) \qquad (4.1)$$

where $\sigma_{ij}^2/\mu_{ij}^2 = exp\,\tau_{ij}^2 - 1$

It is assumed, other than where specifically noted, that

$$ln\,\mu_{ij} = \alpha_i + \beta_j + \gamma_t \qquad (4.2)$$

for parameters $\alpha_i, \beta_j, \gamma_t$, so that (4.1) amounts to a GLM with AQ (row), DQ (column) and PQ (diagonal) effects.

The cell variances are set as

$$\sigma_{ij}^2 = C\mu_{ij} \qquad (4.3)$$

for constant $C > 0$, selected so that $\sigma_{ij}^2 = \left(0.1\mu_{ij}\right)^2$ for $(i,j) = (1{,}16)$, i.e. $C = 0.01\mu_{1,16}$. This is considered a not unrealistic variance structure, and (4.3) accords with the variance assumption of the Mack chain ladder model (Mack, 1993).

Data set 1: Cross-classified model chain ladder model

In this model, $\gamma_t = 0$, and so (4.2) reduces to

$$ln\,\mu_{ij} = \alpha_i + \beta_j \qquad (4.4)$$

and the model includes accident and development quarter, but not calendar quarter, effects. The mean and variance structure are the same as for the **cross-classified chain ladder models** (Taylor, 2011).

The numerical values of α_i, β_j are:

$$\alpha_i = ln\,100{,}000 + 0.1R_1(i) + 0.1R_{15}(i) - 0.2R_{20}(i) - 0.05R_{30}(i) \qquad (4.5)$$

$$\beta_j = (a-1)ln\,j - bj \text{ with } a/b = 16, a/b^2 = 48 \qquad (4.6)$$

Here the AQ effects exhibit an upward trend initially, with an increase in gradient at AQ 15, and then a downward trend from AQ 20. The DQ effects follow a Hoerl curve with delay to payment subject to a mean of 16 quarters, and a standard deviation of $\sqrt{48}$ quarters.

Data set 2: Addition of payment quarter effect

The model is as for data set 1 except that term γ_t of (4.2) is now manifest, specifically

$$\gamma_t = 0.0075\left(R_1(t) - R_{12}(t)\right) + f(t)$$

with

$$\Delta f(t) = 0.001\left(R_{12}(t) - R_{24}(t)\right) + 0.002R_{32}(t)$$

where Δ is the backward difference operator $\Delta\gamma_t = \gamma_t - \gamma_{t-1}$.

This represents a PQ effect that increases at the rate of 0.0075 per quarter from zero at PQ 1 to 0.0825 at PQ 12. The rate of increase then increases linearly from 0.0010 at PQ 13 to 0.0120 per quarter at PQ 24; the PQ effect is then flat up to PQ 32, after which its rate of change increases linearly to 0.0160 per quarter up to PQ 40. This represents SI at a continuous rate of 0.75% per quarter for the first 12 payment quarters, at a steadily increasing rate for the next 12 quarters, at a constant rate over the next 8 quarters, and finally at a steadily increasing rate over the last 8 quarters.

Since t is linearly related to i and j, there is potential for multi-collinearity when predictors related to all three are included in the linear response, as in (4.2) (see e.g. Kuang, Nielsen & Nielsen (2008), Venter & Şahin 2018, Zehnwirth, 1994). Selection of predictors that avoid this is often difficult in such models. The lasso offers an automated procedure for predictor selection, although it obviously does not always guarantee a correct parametric description of multicollinearity.

Data set 3: Addition of a simple interaction

The model is as for data set 2 except that an additional term is added to the linear predictor (4.2), thus:

$$ln\,\mu_{ij} = \alpha_i + \beta_j + \gamma_t + 0.3H_{17}(i)H_{21}(j)\beta_j$$

where $H_k(x)$ is the Heaviside function

$$\begin{aligned} H_k(x) &= 0, x < k \\ &= 1, x \geq k \end{aligned}$$

This data set is the same as data set 2, except that all DQ effects β_j are increased to $1.3\beta_j$ wherever AQ exceeds 16 **and** DQ exceeds 20. It may be verified that the interaction, while a large step change for affected cells, affects only 10 cells (those in the sub-triangle with vertices at (17,21), (17,24) and (20,21)) out of the 820 constituting the 40×40 triangle, so this is a subtle effect within the data set, and its identification by a model can be expected to be difficult.

Data set 4: Addition of more complex interactions

The model is as for data set 2 except that the diagonal effect there is now modified:

$$ln\,\mu_{ij} = \alpha_i + \beta_j + \frac{[40 - j]}{39}\gamma_t \tag{4.7}$$

where γ_t is as in data set 2. This means that the strength of the diagonal effect is greatest at $j = 1$, where it is the same as for data set 2. The strength of the effect declines linearly with j, until it vanishes at $j = 40$. This may be interpreted as SI that is large in respect of early claim settlements, and small in respect of late settlements (as well as varying over PQ).

4.2.2. Results

A data set amounts to a surface of observed values, as functions of a number of explanatory variables. In the present case, it is a surface of quarterly paid claim amounts as a function i, j, t. *A priori*, the function may assume any shape.

It will be assumed that the observations can be adequately approximated by a vector space with a finite basis of functions $X_r(i, j, t), r = 1, \dots, p$ (the **basis functions** of the space). Let observations y_{ij} be arranged in a vector (the ordering is unimportant), and hence denoted y_m. Then observation y_m will be approximated by some linear combination of these basis functions:

$$y_m \cong \sum_{r=1}^{p} \beta_r X_r(i_m, j_m, t_m), m = 1, \dots, n \tag{4.8}$$

where (i, j, t)=(i_m, j_m, t_m) for observation y_m.

An alternative representation of this is

$$y = X\beta + \varepsilon \tag{4.9}$$

where $y, \hat{\beta}$ are vectors with components y_m, β_r respectively, X is the matrix with elements $X_r(i_m, j_m, t_m)$, and the approximation errors in (4.8) are represented by components of the vector ε, assumed to be stochastic subject to a defined distribution.

This is the same as (3.1), and so the whole discussion of Section 3 may be applied to the estimation of the surface in question here. The lasso may applied, yielding a parameter estimate vector $\hat{\beta}$ and a vector of **fitted values**

$$\hat{y} = X\hat{\beta}$$

Any application of this structure requires a selection of the set of basis functions. Basis functions for the present investigation were chosen as:

- $R_K(i), R_K(j), R_K(t), K = 1,2, \dots, 39$ for main effects; and
- $H_k(i)H_\ell(j), H_k(i)H_g(t), H_g(t)H_\ell(j), k, g, \ell = 2,3, \dots, 40$ for interactions.

This produces a total of 117 main effects basis functions and 4,563 interaction basis functions. The values of K, k, g, ℓ were selected to remove redundant terms (for example, $H_k(i) = 1$ everywhere for $k = 1$), though these would be automatically eliminated by a lasso model in any case.

The main effects basis functions generate the **vector space of all linear splines** in i, j, t respectively, with knots at integer values in their domains. The interaction basis functions relate to 2-way interactions in each of which two of the three row, column and diagonal effects undergo a **step change**.

The lasso is applied thus to the four data sets described in Section 4.2.1, using a Poisson distribution assumption. This choice of distribution was essentially dictated by the lasso software used. See Section 6.1 for further discussion.

In the case of synthetic data, there is another goodness-of-fit metric of interest. This is referred to subsequently as the **test error**, and is calculated by the same formula as in step (6) of the CV error (see Section 3.2.4) except that the "observations" y_m are taken from the lower triangle consisting of future diagonals ($t > 40$). These observations are generated as described in Section 4.2.1 with $\gamma_t = \gamma_{40}, \hat{\gamma}_t = \hat{\gamma}_{40}$ for $t > 40$, i.e. nil SI in the future, neither actual nor forecast in the underlying data generative model.

The test error cannot be computed in the case of real data, since the true underlying model is unknown and the actual experience unobserved. It is worthy of note that, even if the actual experience were available, the underlying process generating it might have undergone structural change so that the computed test error would be contaminated with model error.

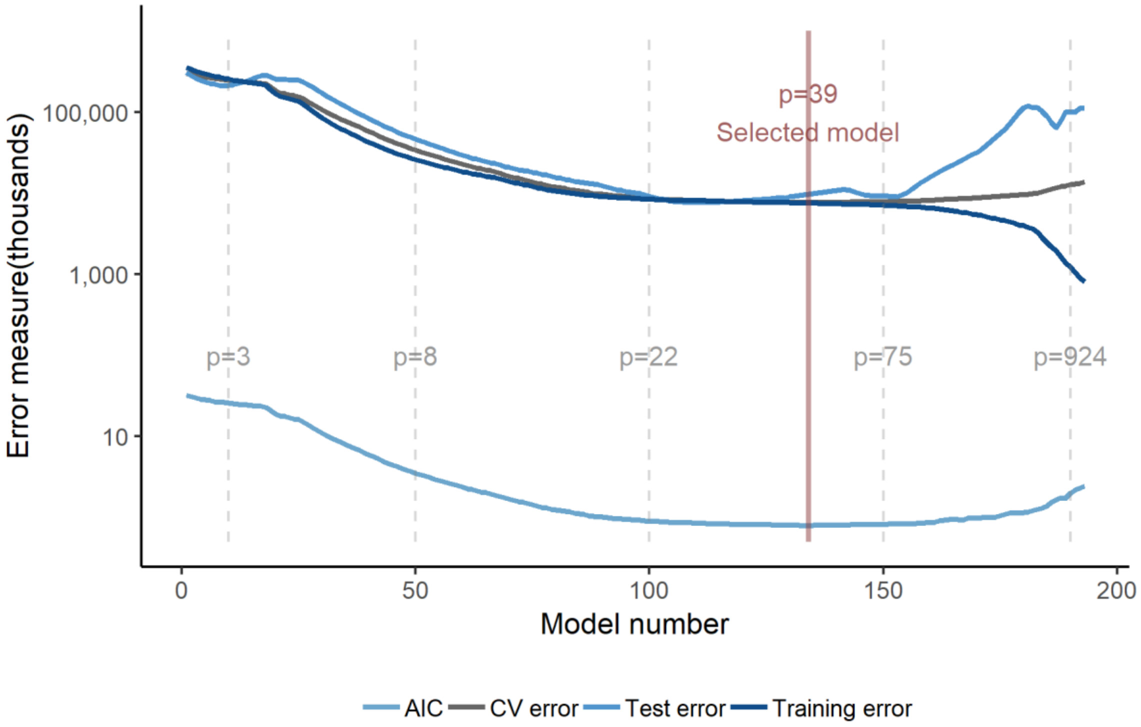
The results of lasso regression for each of the four data sets follow.

Data set 1: Cross-classified model chain ladder model

Figure 4-1 displays the CV error, training error, test error and AIC for a sequence of models with various values of λ, decreasing from left to right. Also shown is the number of parameters in the model, steadily increasing as λ decreases.

The CV error initially decreases with decreasing λ, but then increases. It attains a minimum in a model involving 39 parameters. The test error would choose a similar model, and the AIC a less parameterized model. The training error, as pointed out in Section 3.2.5 is a mere goodness-of-fit statistic, and so declines monotonically as the parameterization of the model increases.

Figure 4-1 Model selection for data set 1



Note: $p = \#$ indicates the number of model parameters p at the corresponding model number (10, 50, 100, 150, 190), while the selected model corresponds to number 134.

Figure 4-2 and Figure 4-3 illustrate the accuracy with which the model tracks the known AQ and DQ effects included in the data, both in training data (the past, indicated by the grey area in the plots) and the test data (the future, white background). The AQ effect, for example, measures the variation of the response variate, and the lasso fit to it, across the range of AQs.

In general, this fit will depend on DQ (though not in the case of data set 1), and hence the labelling in Figure 4-2 as “AQ effect where DQ=20”.

Figure 4-2 Model AQ effect tracking for data set 1 where DQ=20

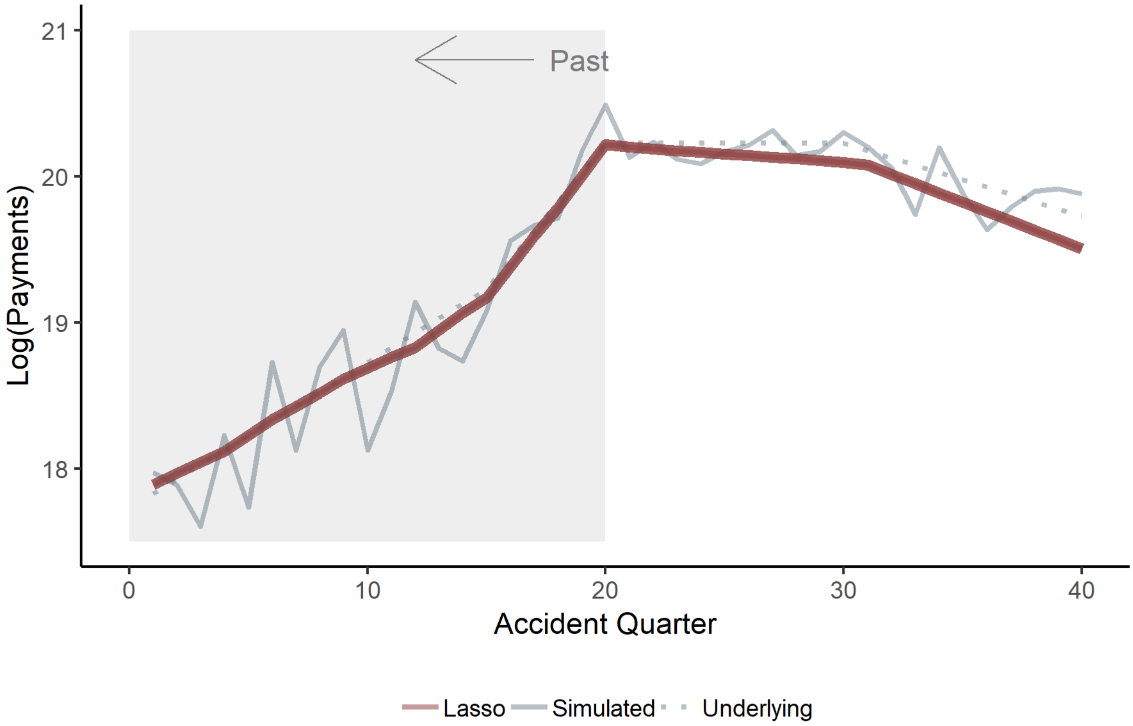
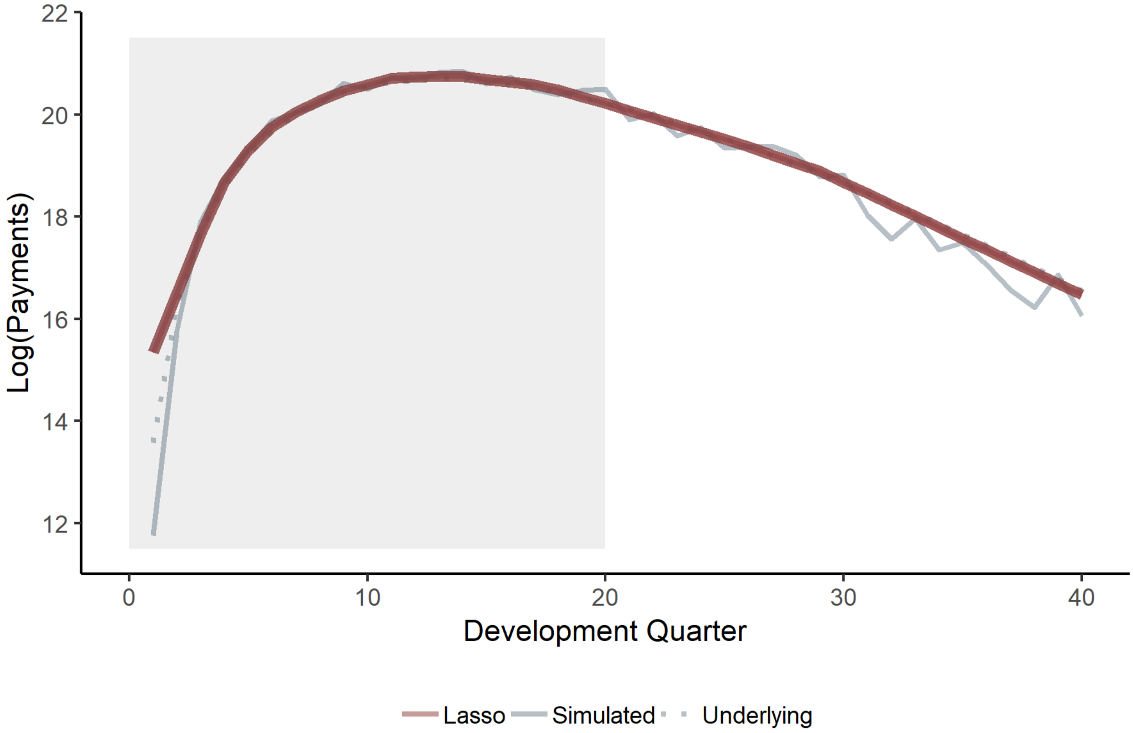


Figure 4-3 Model DQ effect tracking for data set 1 where AQ=20



The values of $y_{i,20}, i = 1, \dots, 40$ in the particular instance of synthetic data are labelled as "Simulated" and represented by the solid grey line. The plotted values of "Underlying" and "Lasso" are $\mu_{i,20}, \hat{y}_{i,20}, i = 1, \dots, 40$. It is seen that the fitted values track actuals closely. Similar remarks apply to Figure 4-3.

Data set 2: Addition of payment quarter effect

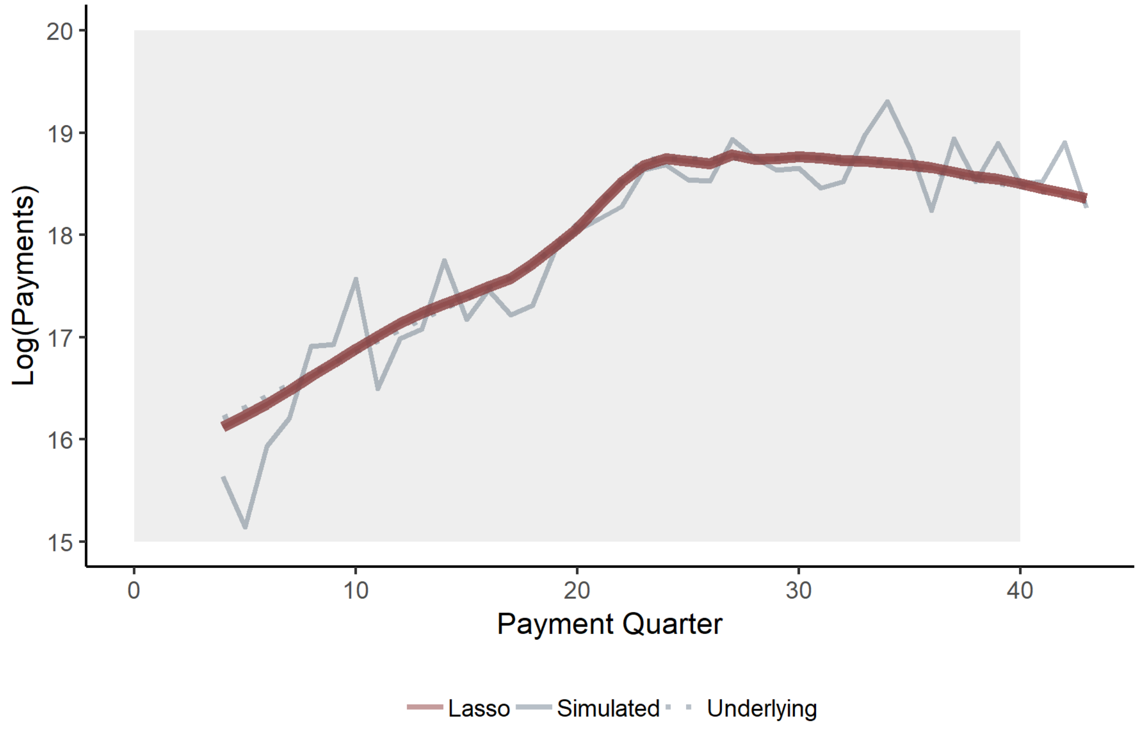
In this case, the model involves 59 parameters. Due to space limitations, the plots of AQ and DQ effects are not reproduced here. The model's tracking of these effects is, however, accurate.

Figure 4-4 and Figure 4-5 track PQ effects at DQ=4, 14 respectively. The PQ effects differ in the two cases because PQ t at DQ j corresponds to AQ $t - j + 1$ and, for given t, the AQ effects differ at $t - 4$ and $t - 14$. Note that, by (4.2), the true values for PQ t at DQ j (j fixed) are given by

$$ln\, \mu_{t-j+1,j} = \left(\alpha_{t-j+1} + \gamma_t\right) + \beta_j, t = 1, \dots, 40$$

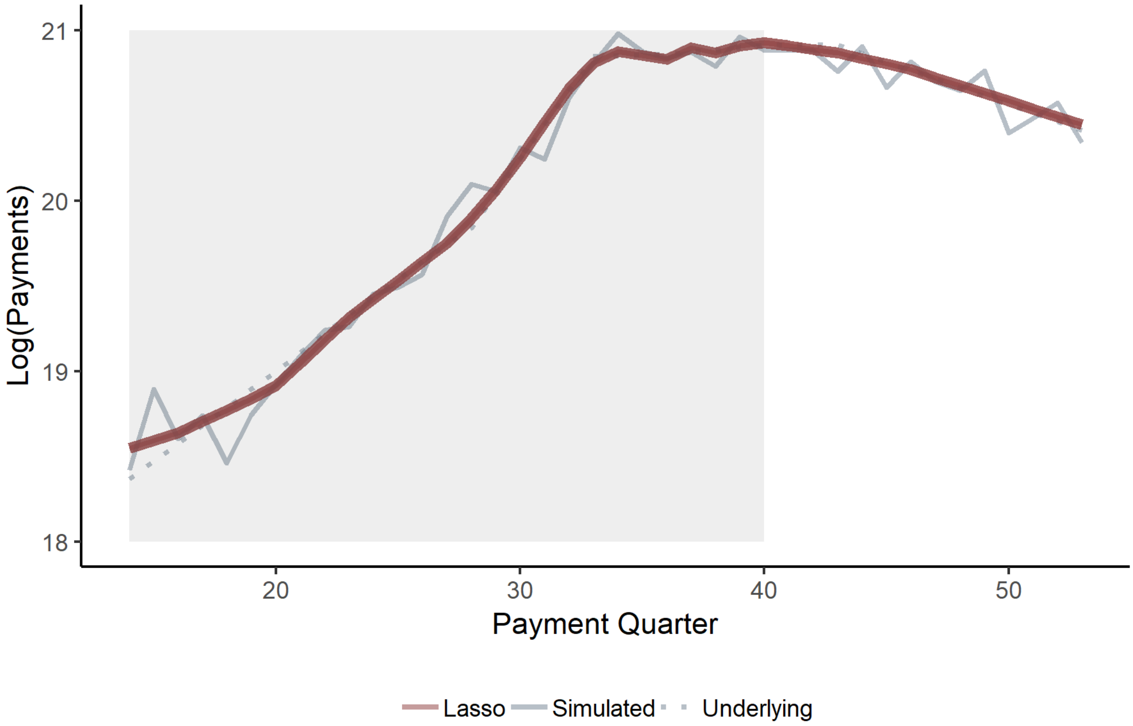
The value β_j is constant as t varies, but the bracketed member illustrates how the plotted "PQ effect" is actually a mixture of AQ and PQ effects, and that the plot will vary with the value selected for j.

Figure 4-4 Model PQ effect tracking at DQ=4 for data set 2



The lasso fit of PQ effects is accurate; the "Underlying" and "Lasso" trajectories in the plots are almost indistinguishable except at low payment quarters.

Figure 4-5 Model PQ effect tracking at DQ=14 for data set 2



Data set 3: Addition of a simple interaction

The model involves 77 parameters. It was remarked in Section 4.2.1 that the interaction added to the model generating this data set affects only 10 cells out of 820, and so modelling of the interaction was expected to be difficult. However, Figure 4-6 shows the lasso to be remarkably effective in tracking the sudden increase in paid amounts at DQ 21. On the other hand, the lasso model somewhat overstates the tail of the DQ effect.

Similarly, Figure 4-7 illustrates the sudden increase in paid amounts at AQ 17 in the AQ effect at DQ 24. AQs 17 to 20 are the only ones for which observations on the increase exist. The lasso recognizes the increase accurately for AQs 17 to 19, but extrapolates a lesser increase to subsequent AQs.

This seems excusable. The data triangle contains no experience of the increase for those later AQs, and in fact provides no particular basis for assuming that those AQs would be subject to the same increase.

Figure 4-6 Model DQ effect tracking at AQ=20 for data set 3

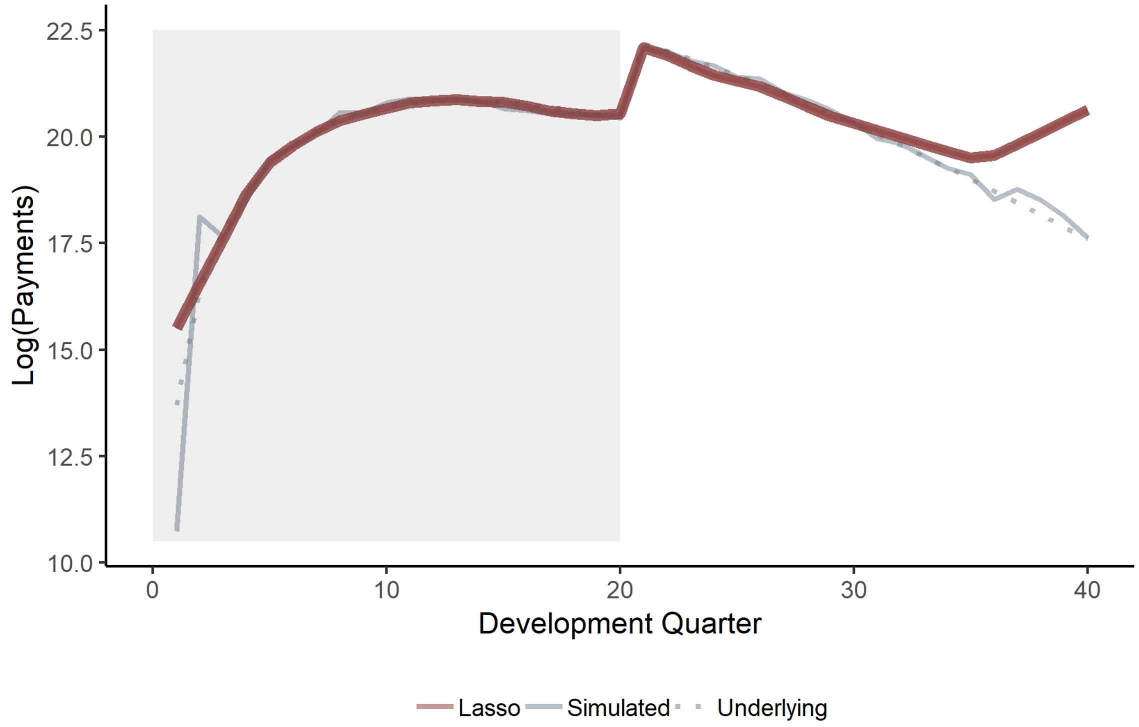
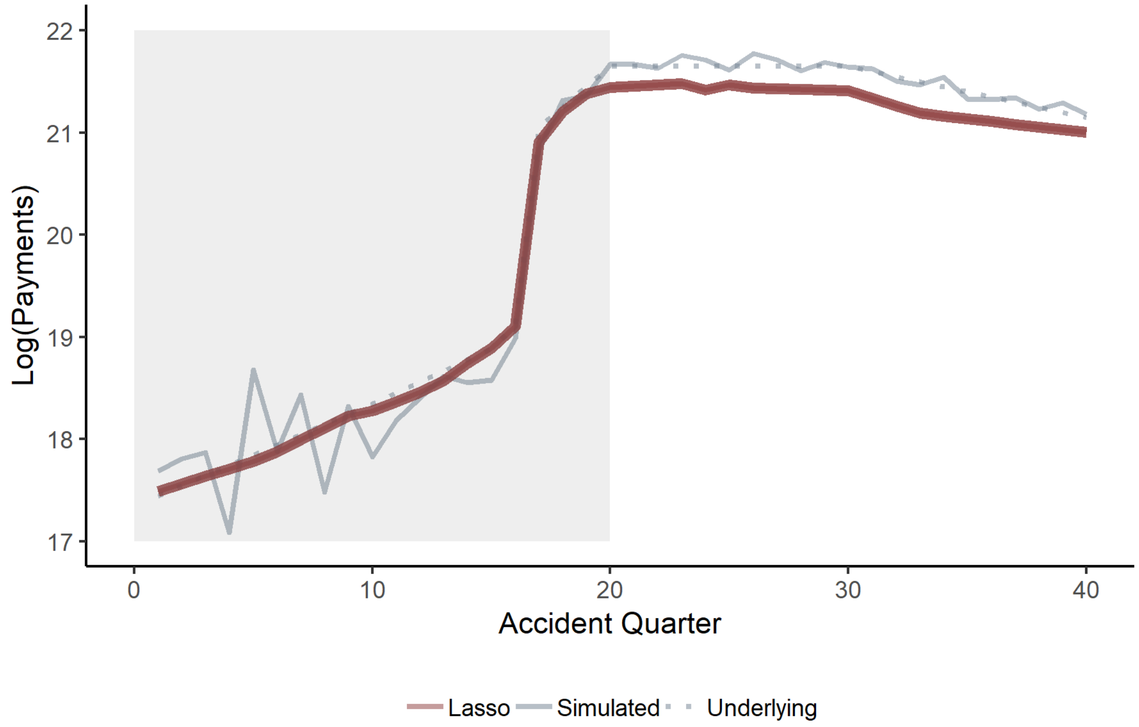


Figure 4-7 Model AQ effect tracking at DQ=24 for data set 3



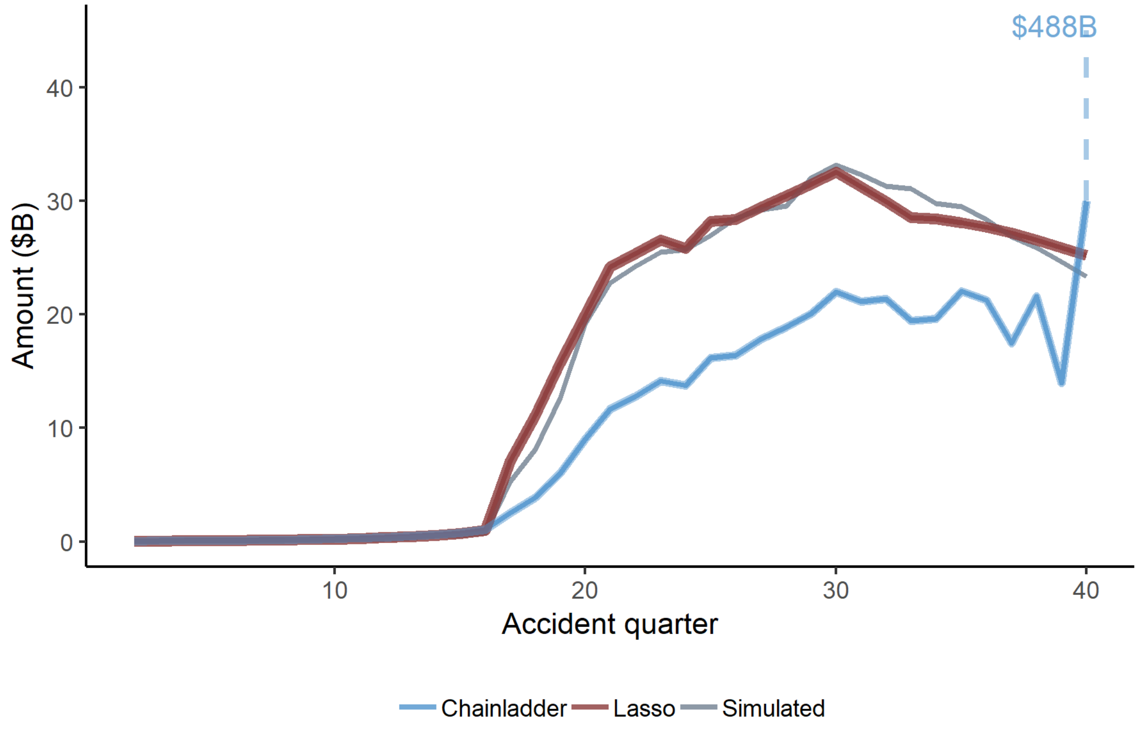
The chain ladder is based on an assumption that the payment delay pattern is the same for all accident periods. It can therefore lead to poor modelling, and erroneous forecasting, in the case of data sets contain abrupt changes in delay pattern. This is illustrated by Figure 4-8, which plots loss reserves by accident quarter, as given by:

- The lasso, where future payments are forecast as the $\hat{y}_{ij}, j > 41 - i$;
- The chain ladder, with age-to-age factors estimated from the experience of the last 8 PQs;
- Actual future payments (simulated) $y_{ij}, j > 41 - i$.

The chain ladder's age-to-age factors are influenced by the elevated experience at DQs 21 to 24, but only slightly. As a result, the loss reserve for each AQ after 16 is seriously under-estimated. As explained in connection with Figure 4-7, the lasso does recognize the increase for AQs 17 to 19, and most AQs from 17 onward are **not** under-estimated.

The fact that this occurs despite the lasso's under-estimation of accident quarter effect apparent in Figure 4-7 was investigated further, and the under-estimation found to be compensated elsewhere in the model. Specifically, SI was over-estimated in the later payment quarters of experience, and then extrapolated into the future. This is an example of the identifiability problem when AQ, DQ and PQ effects are all included in the basis functions.

Figure 4-8 Estimated loss reserves for data set 3



Note: the chainladder result for accident quarter 40 exceeds the scale of the plot.

Data set 4: Addition of more complex interactions

The model involves 53 parameters. The main feature of this data set was SI that varied over both PQ and DQ. Therefore, Figure 4-9 and Figure 4-10 plot PQ effects for DQ = 5, 15 respectively. These plots correspond to Figure 4-4 and Figure 4-5 for data set 2 and, as there, the plotted “PQ effect” is actually a mixture of AQ and PQ effects. In any event, the lasso appears to track the true data trends closely.

Figure 4-9 Model PQ effect tracking at DQ=5 for data set 4

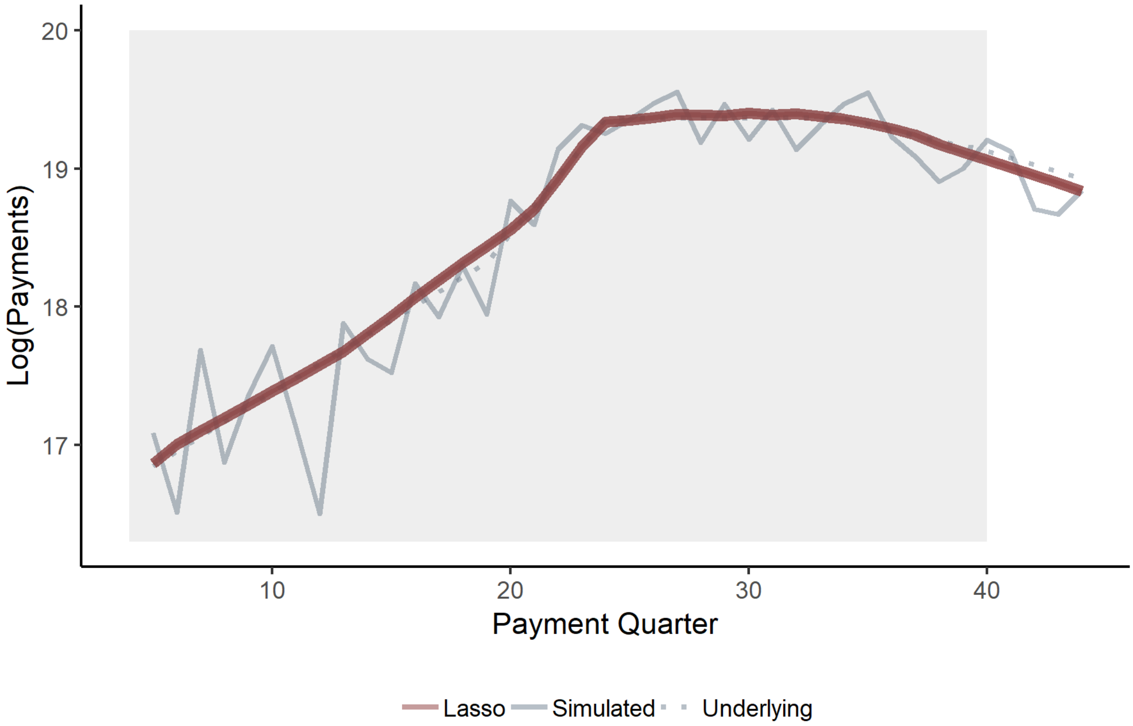
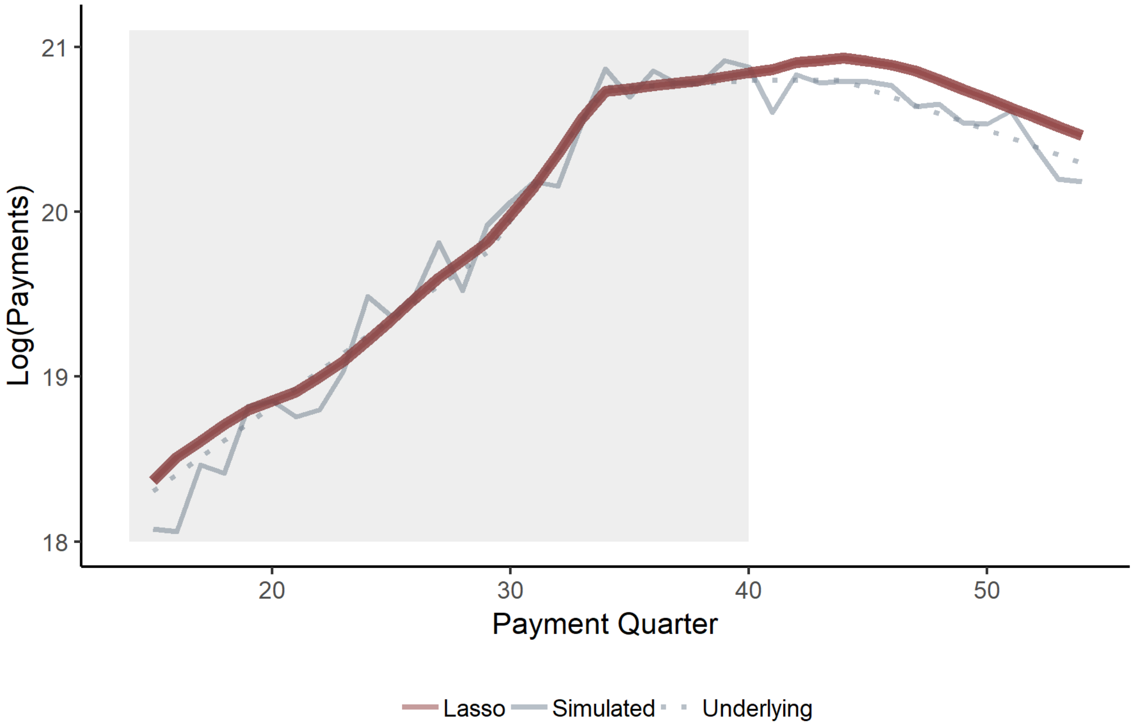


Figure 4-10 Model PQ effect tracking at DQ=15 for data set 4



Of course, the ultimate purpose of the claim modelling here is forecast of a loss reserve. For this reason, Figure 4-11 (top) plots the lasso loss reserve estimates, separately by AQ, compared with the true expected values. The figure also includes error bars for each reserve. These have been obtained as follows:

- The simulation of the data set was replicated 500 times, generating a sample of 500 data sets.
- The lasso has been applied to each replication, in each case with λ set equal to the λ min value corresponding to the originally chosen model (strictly speaking, a path of decreasing values of λ values was used, each corresponding to a different model, from which the model corresponding to the original model's λ min value was selected, since it is often faster to fit a whole path in *glmnet* than compute a single fit).

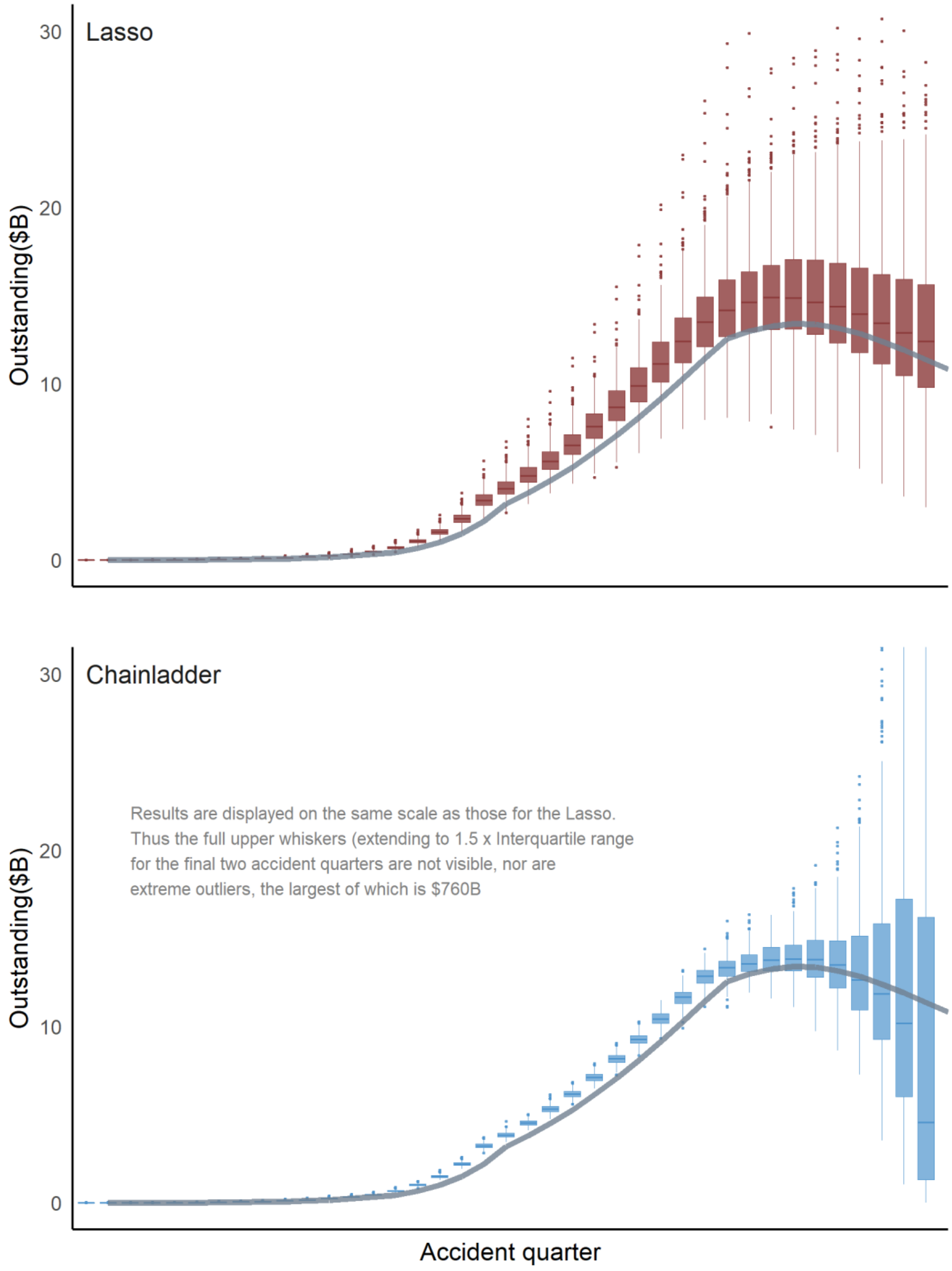
Some experimentation was conducted with an alternative (and considerably more time-consuming) approach in which the value of λ was optimized within each of 100 replications, and little change in results was found.

Figure 4-11 (bottom) gives the corresponding plot for the chain ladder on the same scale.

It is seen that:

- Both models exhibit some upward bias, except that, for the most recent AQs, the chain ladder under-estimates considerably.
- As is often the case, the chain ladder exhibits large prediction error in connection with the more recent AQs, whereas the lasso produces much tighter forecasts.

Figure 4-11 Lasso and chain ladder estimates of loss reserves for data set 4 compared to the underlying mean



Results are displayed on the same scale as those for the Lasso. Thus the full upper whiskers (extending to 1.5 x Interquartile range for the final two accident quarters are not visible, nor are extreme outliers, the largest of which is $760B

It may be possible to shrink the parameter set further in all four of the above examples. The basis functions selected at the start of this sub-section for the modelling of column effects are open-ended ramp functions. Some number of these are required to fit the kind of curvature observed in Figure 4-3.

The profile displayed there resembles a Hoerl curve, for which

$$\beta_j = A + B \ln j + Cj$$

Thus, the addition of the functions $\ln j$ and j to the set of basis functions might lead to a more economical model. The ramp basis functions could be retained to accommodate deviations from the Hoerl curve.

4.3. Real data

4.3.1. Data set

As explained at the start of Section 4.2.1, synthetic data form a useful test of a model's ability to detect and reproduce known (albeit complex) features. In the case of real data, on the other hand, the correct model is unknown and so validation of the fitted model is less sure. However, the real data set may have the advantage of challenging the model with subtle, unknown features that may not have been contemplated in the construction of synthetic data.

The real data set selected for use here is drawn from a privately underwritten, but publicly regulated, scheme of Auto Bodily Injury Liability in one of the Australian states. It consists of the claim history from the scheme commencement on 1 September 1994 to 31 December 2014. This totals 82 quarters if the initial month September 1994 is counted as 1994Q3.

The data set comprises a unit record claim file, containing only **finalized claims**, of which there are about 139,000. Each record includes *inter alia* the following fields in respect of the claim to which it relates:

- Date of accident;
- Date of finalization;
- Injury severity score at finalization;
- Legal representation status at finalization;
- For each individual payment:
  - Date of claim payment;
  - Amount of claim payment.

The injury severity score is the **Motor Accident Injury Severity (MAIS)**, which assumes values 1 to 5 for injuries, 6 for a fatality, and 9 in the case where generally there is insufficient information to determine a score. Values 1 to 5 ascend with increased severity. Severity 5 would usually involve paraplegia, quadriplegia, or serious brain injury. Claim sizes vary very considerably according to MAIS.

Legal representation simply notes whether or not the claimant is represented. The great majority of claimants with MAIS 2 to 5 are legally represented, but MAIS 1 includes a substantial proportion of minor injuries for which legal representation has not been obtained. For the purpose of the present investigation, a new covariate *maislegal* has been created, defined as:

maislegal = 0, if MAIS = 1 and the claimant is unrepresented;

= 1, if MAIS = 1 and the claimant is represented;
= MAIS, otherwise.

Claim payments have been summarized into a (finalized) claim size. Each claim payment has been adjusted by the State wage inflation index from the date of payment to 31 December 2014, and all adjusted payments then summed. This produces a claim size expressed in 31 December 2014 dollars.

Most claims involve a sequence of payments, but there is usually a dominant one, and it is usually not greatly separated in time from the finalization date. Hence, for the purpose of the analysis, all payments in respect of a specific claim are regarded as made on the date of finalization (albeit correctly indexed for inflation). This has some implications for the measurement of SI.

Finally, the data have been aggregated into cells labelled by maislegal, accident and development quarter (and, by implication, payment quarter). The number of finalizations in each cell has also been recorded for use as a weight in the analysis.

This step was not essential in principle, but was aimed purely at reducing the computational load so that a greater number of basis functions could be considered. With an additional load, modelling at the individual claim level would be possible.

Two of the authors have worked on the data set continually over a collective period of more than 15 years, conducting quarterly analyses. There are a number of data complexities but, with this experience, the data set is believed to be well understood.

Complexities include the following:

(a) Claim processes undergo change from time to time. In consequence, there have been occasional material changes in the rate of claim settlement.
(b) SI experience has been typical of schemes of this type, with periods of rapid increase in claim sizes, punctuated by periods of quiescence.
(c) It is apparent that SI does not affect all claim sizes equally. The largest claims are largely unaffected by it. These are claims involving the provision of income and medical support for life, and there is little scope for dispute over the extent of liability. The opposite is true of less serious claims. These often involve soft tissue injury for which objective determination of severity is difficult. In the absence of change to the rules of assessment, the trend in claim sizes is usually upward.
(d) The scheme was subject to major legislative change from December 2002. This was the **Civil Liability Act (CLA)**. The changes affected claims with accident dates after that date, and had the effect of elimination of many of the smallest claims, and reducing the cost other relatively small claims. This caused a radical change in the distribution of claim sizes.
(e) As a result of the elimination of a material proportion of claims from the scheme, claim management resources were released to process and settle the remaining claims earlier than had been the case in AQs prior to the CLA.
(f) The reduction in claim sizes resulting from the CLA was gradually eroded in AQs subsequent to 2003Q1, causing further changes in the distribution of claim sizes.

All of these changes are incompatible with the chain ladder model, or indeed with any model that assumes the same payment delay pattern for all accident periods. Some of the changes (specifically, (a) and (d)-(f)) are row effects, whereas some are diagonal effects ((b) and (c)). This creates a challenge for many models.

In view of (a), **operational time (OT)**, as defined in Section 2, is a much more useful metric of time than real development time, such as development quarters. The OT $\tau_{[k]}$ is computed for each claim k according to its AQ, and attached to the relevant claim record. In the aggregated data, the operational time for in an (accident, development) quarter cell within a particular maislegal is taken as the average of the individual operational times in that cell.

4.3.2. Results

The data were aggregated by quarter, and the lasso was used to model average claim size where

- $s =$ maislegal;
- $i =$ AQ;
- $j =$ DQ;
- $t =$ PQ.
- $\tau_{i,j,s} =$ OT in AQ i and DQ j for maislegal s;
- $y_{i,j,s} =$ average claim size (response variate) in AQ i and DQ j for maislegal s;
- $n_{i,j,s} =$ number of claim finalizations (weight variate) in AQ i and DQ j for maislegal s.

All variates here other than DQ, the response and the number of claim finalizations are covariates.

The basis functions for main effects in the lasso were chosen as:

- $I(s = m), m = 0,1,2, \dots ,6, 9$
- $R_\ell(\tau_{i,j,s}), \ell = 0, 0.05, 0.10, \dots ,0.95, 1$
- $R_T(t), t = 1,2, \dots ,82$

where the calendar quarters 1994Q3,…,2014Q4 are labelled 1,…,82 for convenience.

All observations were assumed Poisson distributed, as discussed later in Section 6.1.

The basis functions for interactions were chosen as:

- $I(s = m)R_T(t), m = 0, \dots ,6,9; T = 1,2, \dots ,82;$
- $I(s = m)R_\ell(\tau_{i,j,s}), m = 0, \dots ,6,9; \ell = 0,0.05,0.10, \dots ,1;$
- $I(s = m)H_h(i)H_\ell(\tau_{i,j,s}), m = 0, \dots ,6,9; h = 1,2, \dots ,82; \ell = 0,0.05,0.10, \dots ,1;$
- $I(s = m)H_g(t)H_\ell(\tau_{i,j,s}), m = 0, \dots ,6,9; g = 1,2, \dots ,82; \ell = 0,0.05,0.10, \dots ,1.$

There are 109 main effects basis functions and 26,728 interaction basis functions. Regression problems of this size are computation-intensive. The analyses of the present sub-section were performed using a PC with 16Gb RAM, 2 cores, and a 2.9GHz chip using the Microsoft R Open 3.5.0 release. With these resources, a single regression occupied 5 hours.

It should be pointed out that this regression alone is only part of the entire program of loss reserving. For example, the computation of OT requires an estimate of the ultimate numbers of claims incurred in each AQ (see Section 2), so the estimation of IBNR counts is required as a preliminary exercise.

The loss reserve estimate depends on future finalizations in respect of past AQs. These must be assigned to future PQs if SI is to be accounted for correctly, so a forecast of future finalization counts by AQ and PQ forms another preliminary exercise. Moreover, although the lasso model will have provided estimates of past rates of SI, future rates are exogenous and will require forecasts.

The loss reserve consists of the ultimate cost of future finalizations (for past AQs), less the indexed amount already paid in respect of claims open at the valuation date (which will be finalized in future). There may be other subsidiary issues to be addressed. For example, finalized claims may sometimes re-open.

Thus, the loss reserving model of this example consists of multiple sub-models. The present sub-section deals with just one of those.

The lasso model contained 94 non-zero coefficients including the intercept. Although this is a moderately large number of coefficients, it should be recognised as covering the 8 distinct *maislegal* categories. These were all modelled separately in the consulting exercise described in Section 4.3.1. In this sense, the 94 parameters might be thought of as about 12 per *maislegal* model.

In the case of analysis of real data, it is not possible to compare actual and fitted effects as in Figure 4-2 to Figure 4-7, Figure 4-9 and Figure 4-10, since the actual effects are unknown. The following model validation diagnostics therefore compare actual and fitted total cost of finalization.

For example, Figure 4-12 compares, for each value of *maislegal* and for each development quarter, the total (inflation indexed) cost of finalizations to 31 December 2014 with the total of the fitted costs of all the claims involved. Note that these values, and all in subsequent figures, have been scaled by a constant value for confidentiality, and, in some cases, the scaled values are shown on a log scale for readability.

Attention is given in the following discussion to *maislegal* 1, since it is by far the largest category, accounting for 62.5% of all finalizations in the data set.

Figure 4-12 Model fit to real data by development quarter

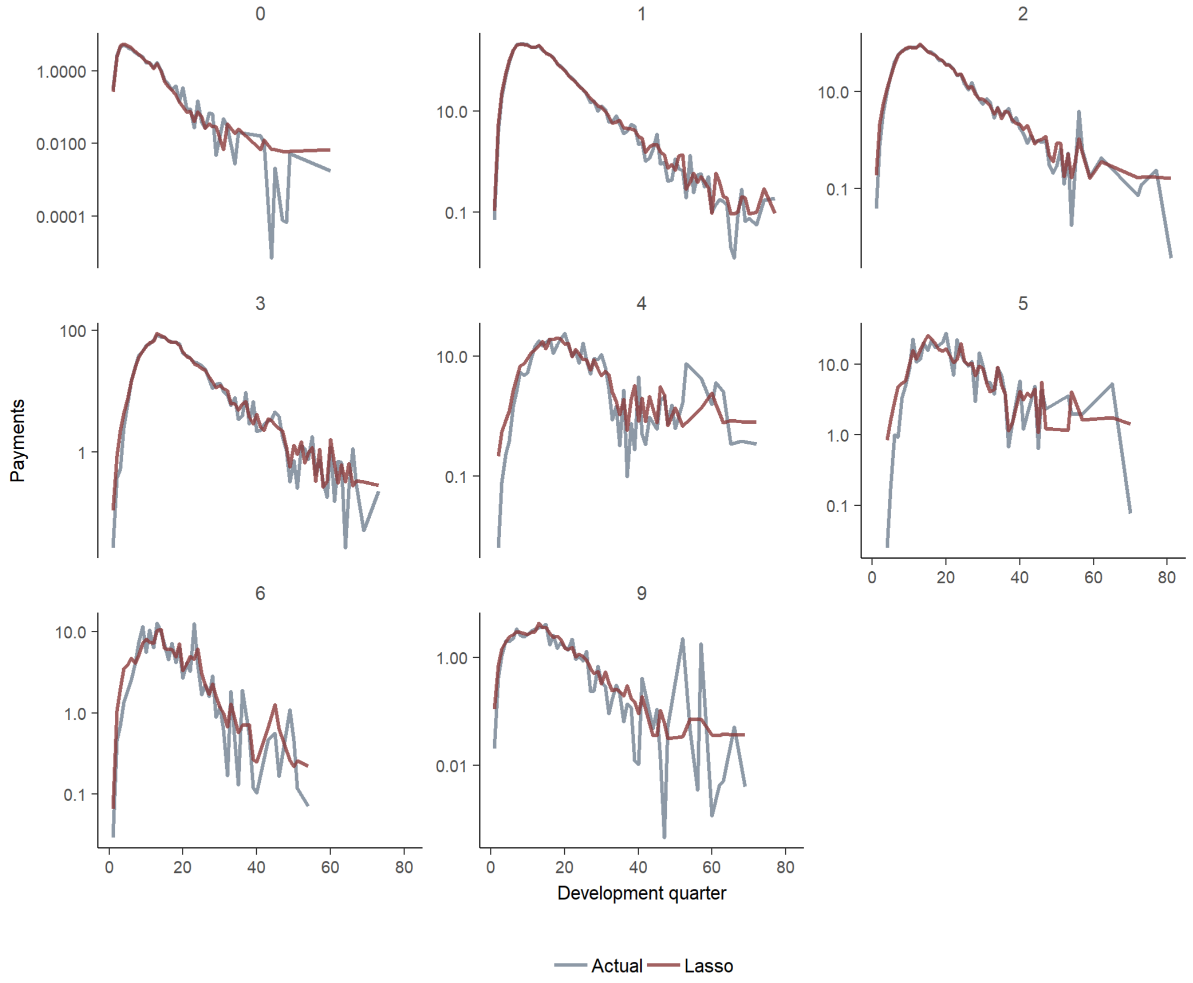


Figure 4-13 gives the corresponding plot by payment quarter.

Figure 4-13 Model fit to real data by payment quarter

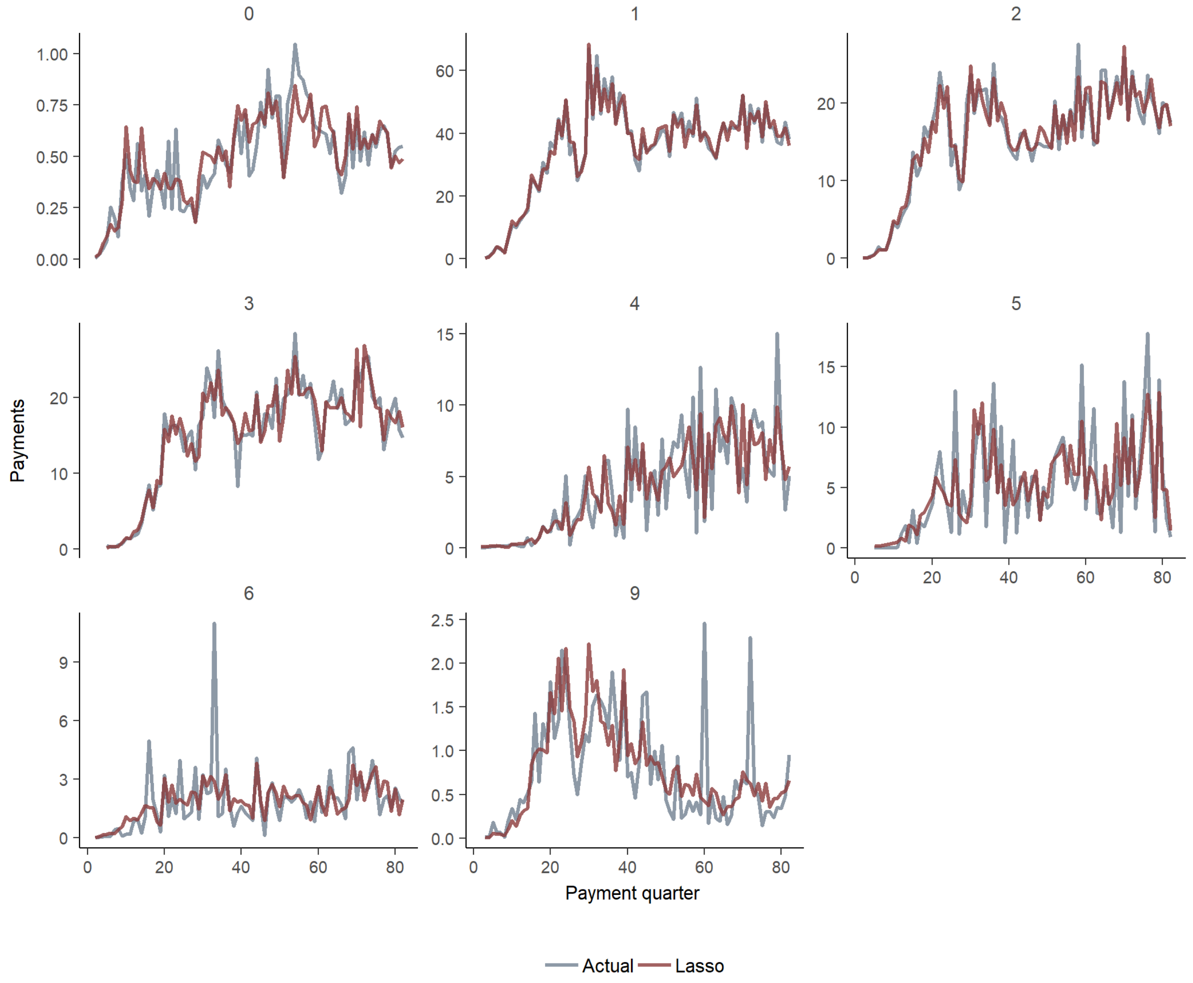


Figure 4-12 and Figure 4-13 illustrate a relatively faithful reproduction of the data by the model, at least in aggregate over development and payment quarters respectively. Similarly the models tracks the data reasonably well by accident quarter, though there are two notable exceptions for *maislegal*=1 as may be seen in Figure 4-14 where there are differences over AQs 2 to 8 (1995Q1 to 1996Q4) and 35 to 40 (2003Q2 to 2004Q3).

An anomaly in the second of these periods is not altogether surprising. Note the mention in points (d) to (f) of Section 4.3.1 of legislative change from late 2002, known to have caused abrupt changes in the cost of claim finalizations. Figure 4-14 (left) suggests that the model has smeared the abruptness over a year or two.

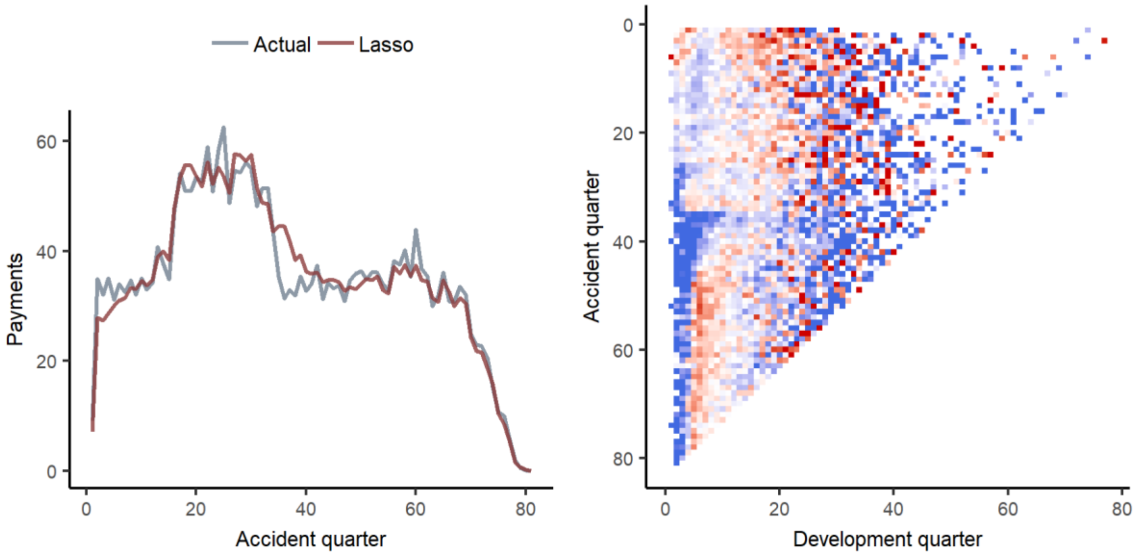
Figure 4-14 (right) sheds additional light on this matter. It is a 2-dimensional heat map of the ratio A/F, where A, F denote actual and fitted total finalization costs at the cell level. Several features are prominent:

(a) An abrupt change in fit is indeed apparent at AQ 34 or 35. For 3 or 4 years after this, the model over-states claim costs in the early DQs.
(b) In later AQs, this over-statement persists at a more moderate level in the first development year or so, but then tends to under-estimate in the second development

year. This is probably explained by the changing payment pattern mentioned in point (e) of Section 4.3.1.

(c) In the early AQs mentioned above, the model exhibits the reverse tendency, with over-estimation in the first DQs, and under-estimation in the next few.

Figure 4-14 Model fit to real data for maislegal=1



4.3.3. Adjustment for special circumstances

The legislative change discussed in Sections 4.3.1 and 4.3.2 appears somewhat troublesome in the abruptness of its effects, and is worthy of further discussion. It is accommodated in the model by interactions of main effects with the step function $H_k(i)$ for k in the vicinity of 34. These interactions could have been anticipated.

Recall that the lasso applied in Section 4.3.2 used loss function (3.6) in which the same penalty λ is applied to each covariate. Recall also the Bayesian interpretation of the lasso in Section 3.2.2 in which this λ relates to the reciprocal dispersion of the prior on each covariate's regression coefficient.

Such a penalty treats the $H_k(i)$ interactions as no more likely to be non-null in the vicinity of $k = 34$ then anywhere else, in contradiction of strong expectations to the contrary. It would be possible to recognize the virtual inevitability of these interactions by using the alternative loss function (3.7) with λ_r for these interactions assuming a lesser value than for other λ_r.

The objective stated in Section 1 was to automate the modelling of claims experience, and the contemplation of λ_r varying with r seems to run counter to this objective. On the other hand, it would appear perverse to deliberately overlook the effects of a known material change to the claim environment. For this reason, we are broadly in favour of removing the penalty term corresponding to parameters that would track known discontinuities in the data.

We would also take this a step further and recommend the consideration of customized parameters (i.e. beyond those included in the set of features) that capture specific experience. These customised parameters would be included with a penalty of 0.

In the case of the anomaly relating to the legislative change here, we have conducted a small side experiment by including some customized variables based on the following considerations:

- The legislative change introduced a discontinuity at AQ 35 for *maislegal* 1. This suggests including a Boolean covariate for AQ 35 and higher.
- The effect differed in the first year compared to later years. This suggests including some specific modelling for AQ 35-38.
- The legislative change had the effect of removing lower value claims but leaving more serious claims unaffected. This translates to an increase in claim sizes at lower operational times but this increase gradually wears off as operational time increases. This suggests using variables that contain a reverse operational time spline – the spline is non-zero at low operational times but reduces to 0 for higher operational time terms.

To select exact terms to be included in the model, a GLM was fitted with over-dispersed Poisson error and using the same covariates as selected by the lasso illustrated above. Additional terms were then added to deal with the special features just identified. A heat map similar to that in Figure 4-14 was used to guide covariate selection.

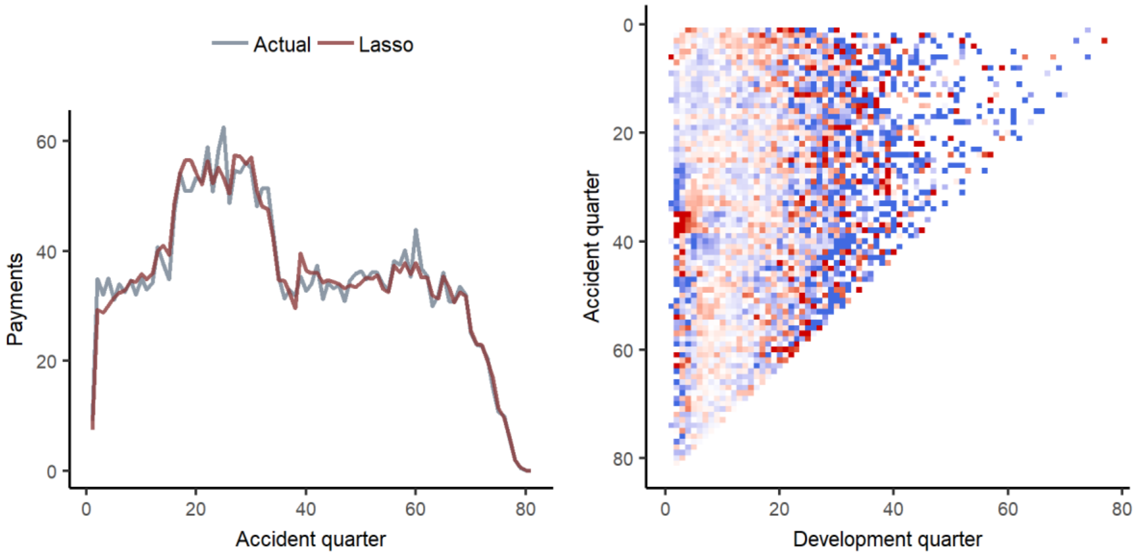
The following custom terms were selected for inclusion in the lasso model after a short period of experimentation:

- $I(s=1)H_{35}(i)$;
- $I(s=1)\big(H_{35}(i)-H_{39}(i)\big)$;
- $I(s=1)H_{35}(i)(1-\tau)$
- $I(s=1)H_{35}(i)\ \max(0, 0.05-\tau)$
- $I(s=1)\big(H_{35}(i)-H_{39}(i)\big)\ \max(0, 0.4-\tau)$
- $I(s=1)\big(H_{35}(i)-H_{47}(i)\big)\ \max(0, 0.2-\tau)$

The lasso model was then fitted in the usual way, except that the 6 variates above were included without penalty, along with the other basis functions described in Section 4.3.1.

After refitting the lasso, the tracking of model and experience for *maislegal* 1 was re-examined. This has improved, as is apparent from Figure 4-15.

Figure 4-15 Customized model fit to real data for maislegal=1



4.3.4. Comparison of lasso and custom-built GLM predictions

Projections from model from the previous section (i.e. the lasso model together with the custom modifications for the known legislative effect) were compared with those from a custom built GLM of the individual claim size data. This model was manually constructed by those intimately familiar with the data, and has previously been discussed by Taylor & Sullivan (2016). As discussed in Section 1, building such a model consumes a significant amount of a skilled resource.

Both models are projected forward excluding SI. Comparisons of the loss reserve by accident quarter (on a log scale) are presented in Figure 4-16. Again payments have been scaled.

Overall the comparison is reasonable. The results for *maislegal* 1 are very similar, while other *maislegal* groups are generally comparable. Note that the differences in *maislegal* 9 are magnified due to the scale of that plot. Furthermore, this class contains small numbers of claims.

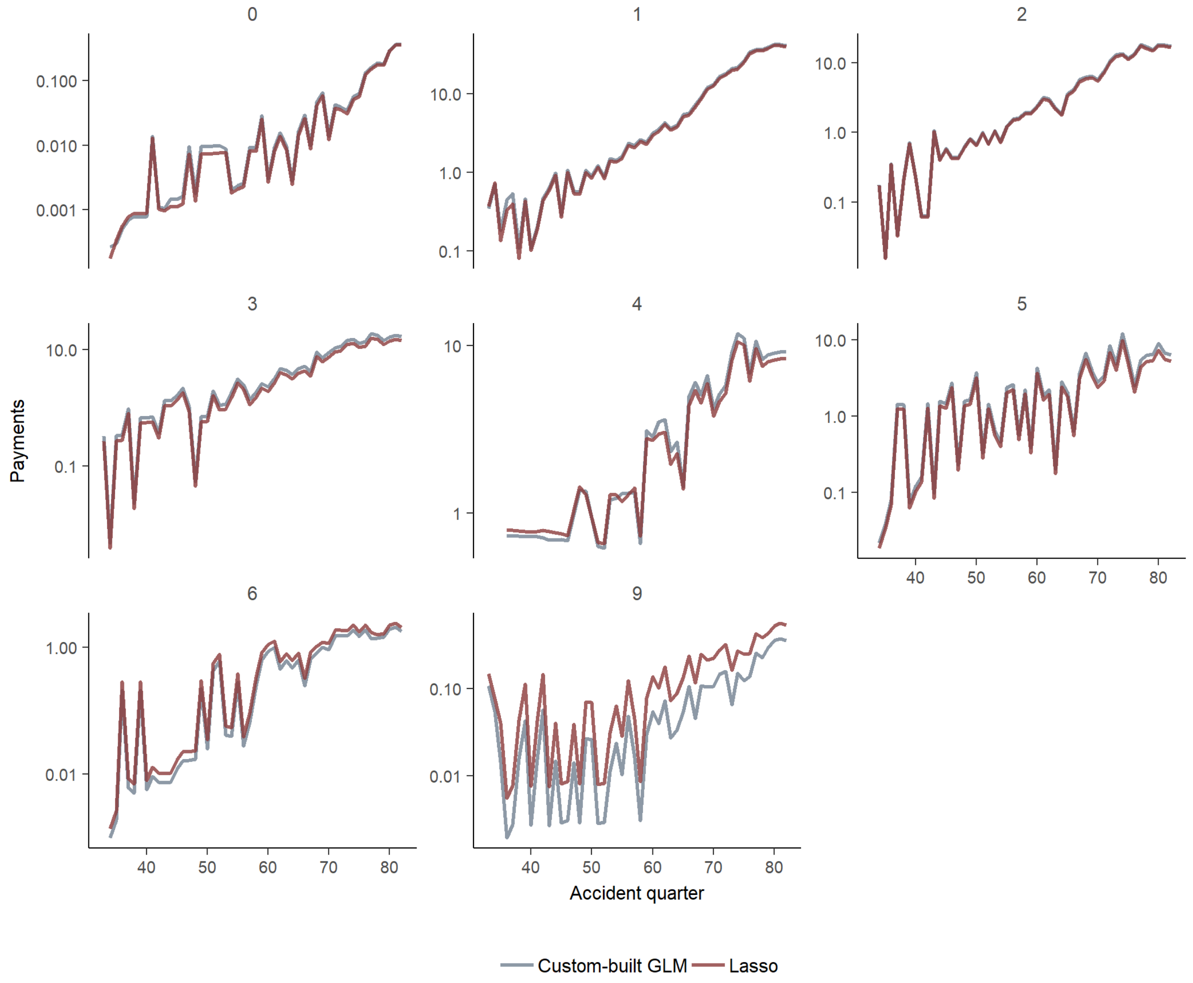
Evidently, the lasso, with a small amount of customization, has produced numerical results very close to those derived from many hours of a skilled consultant's time. The lasso might therefore be considered as effective as the consultant in the estimation of quantities such as average claim size, ultimate claim cost, run-off schedules, etc.

However, the lasso model, in common with other machine learning models, is subject to the **interpretability problem**. This manifests itself in a model that fits well to data, forecasts efficiently, yet is a rather abstract combination of basis functions. A physical interpretation will often be possible, but may be achieved only on by means of some detailed analysis.

While it is true that the lasso produces a more interpretable model than an ANN, for example, it is still the case that modelled effects may be a combination of a number of spline and interaction terms. Additional work would be required to interpret these and translate them to specific data features, which in turn could form the basis for extrapolation to a loss reserve.

The consultant's modelling, on the other hand, will be more targeted at specific data features, and consequently less abstract and more easily interpretable.

Figure 4-16 Comparison of projections under the lasso model and the custom-built GLM



5. Prediction error

Every forecast needs to be accompanied by some information on its dispersion, possibly standard error, but preferably the entire predictive distribution. This can be achieved in relation to a loss reserve generated from a lasso model by a bootstrap of that model.

This approach is well documented in the literature, and can be readily bolted onto the forecasting procedure outlined in Section 4. In view of this lack of novelty, the bootstrap is not followed up numerically in the present paper. However, the following three sub-sections add a few comments on each.

To consider prediction error, let Z be the future random variable to be forecast, and let $E[Z] = \mu$. Let $\hat{Z}$ be the forecast from the thinned lasso model. The prediction error is

$$\hat{Z} - Z = \left(E[\hat{Z}] - \mu\right) + \left(\hat{Z} - E[\hat{Z}]\right) + (\mu - Z) \tag{5.1}$$

The first of the three members on the right is the error in the long-run model estimate (i.e. the average over many replications of the model estimate), and is the effect of the difference between the selected model and reality. It is called **model error**. The second member is the component of prediction error arising from the limited size of the data set, and the resulting sampling error in parameters estimates. It is called **parameter error**. The final member reflects the assumption that the predictand is drawn from a stochastic process, and represents the noise in that process. It is called **process error**.

It is usual to assume that all three components of (5.1) are stochastically independent, so that the **mean square error of prediction (MSEP)** of forecast $\hat{Z}$ is

$$MSEP[\hat{Z}] = E\left[\left(\hat{Z} - Z\right)^2\right] = E\left[\left(E[\hat{Z}] - \mu\right)^2\right] + E\left[\left(\hat{Z} - E[\hat{Z}]\right)^2\right] + E[(Z - \mu)^2]$$

$$= MSE_{model}[\hat{Z}] + MSE_{parameter}[\hat{Z}] + MSE_{process}[\hat{Z}] \tag{5.2}$$

where the last three members are the mean square errors of model, parameter and process errors.

5.1. Bootstrap of a GLM

Bootstrap of a lasso is an extension of bootstrapping a GLM, and so some aspects of the latter are first recalled as background to the former.

The application of the bootstrap to a loss reserving GLM is discussed in some detail in Taylor & McGuire (2016). It is noted there that various forms of bootstrap are available, most notably (nomenclature varies somewhat from place to place):

- **Parametric bootstrap.** Bootstrap replications are generated by re-sampling of model parameters from a selected distribution (usually normal) with first and second moments as estimated by the GLM.
- **Non-parametric bootstrap.** A bootstrap replication is generated by re-sampling of the data set, and the fitting of the GLM form to it.

In these bootstrap procedures, the GLM is taken as the correct model, and so model error is not considered. However, an estimate of $MSE_{parameter}[\hat{Z}]$ may be obtained from replications of $\hat{Z}$, and $MSE_{process}[\hat{Z}]$ may be estimated by simulation of the noise.

5.2. Bootstrap of lasso

In an analogous process, a non-parametric bootstrap may be applied to the entire lasso. The replications then generate a collection of different data sets, and hence possibly different models (i.e. different subsets of selected covariates). Aggregation of the first two members on the right side of (5.1) yields

$$\hat{Z} - Z = \left(\hat{Z} - \mu\right) + (\mu - Z) \tag{5.3}$$

Now, if it is reasonable to assume that the lasso is unbiased, then the average $\bar{Z}$ of $\hat{Z}$ over all replicates (and therefore over different models) will approximate μ, and an approximate version of (5.3) is

$$\hat{Z} - Z = (\hat{Z} - \bar{Z}) + (\mu - Z) \tag{5.4}$$

Then replicates of $(\hat{Z} - \bar{Z})$ can be used to estimate $MSE_{model}[\hat{Z}] + MSE_{parameter}[\hat{Z}]$. An estimate of $MSE_{process}[\hat{Z}]$ may be obtained in the usual way.

The computational load from a bootstrap of the entire lasso would vary depending on the number of basis functions considered and the size of the data set. For example, a single run of the *glmnet* procedure for the synthetic data took 6-10 seconds while the full cross-validation run required approximately 3 minutes. By contrast, the run-times for the real data example were approximately 20 minutes and 4.5 hours, mainly due to the significantly greater number of basis functions. Therefore, bootstrapping using a single *glmnet* procedure, as described in relation to Figure 4-11, with cross-validation omitted, will produce acceptable computation times in many, though not all, cases.

Note also that bootstrapping a lasso, as described above, would not enable partition of $MSE_{model}[\hat{Z}] + MSE_{parameter}[\hat{Z}]$ into its two components. This would require replications within replications, a form of iterated bootstrap (Hall, 2013). A single replication, involving a fixed GLM, would be expanded to multiple replications of re-sampled parameter estimates with that model held fixed. This would provide an estimate of $MSE_{parameter}$ for that single model. Evidently, this would be even more computer-intensive, though the replications to estimate parameter error would run more quickly due to the constrained set of basis functions.

Unfortunately, the above estimate of $MSE_{model}[\hat{Z}]$ would probably fall short of its true value in two respects.

First, the universe of models from which any lasso model is selected is only the vector space spanned by the chosen basis functions. This may be smaller than the space of all conceivable models.

Second, resampling of the data set can generate only pseudo-data sets that are broadly consistent with the trends and features of the original data set. For example, it is unlikely to generate a data set consistent with rates of SI that are uniformly double those underlying the original data set. Yet such a regime of high SI might occur in future.

Thus, the model MSE obtained from a bootstrap lasso will include only models of the future that broadly resemble those of the past. This sort of distinction is discussed in some detail in Risk Margins Taskforce (2008), who decompose model MSE into (their nomenclature):

- **External systemic error**, which includes model error induced by features that are possible in future data but not present in the data set; and
- **Internal systemic error**, which accounts for only error in the selection of a model consistent with the data set.

No model based only on claim data is likely to include allowance for external systemic error, so the most that one can expect of the lasso is inclusion of internal systemic error.

6. Further development

6.1. Model thinning

The lasso is implemented here by means of the R procedure *glmnet*. This provides a limited choice of error distributions. The most appropriate available for analysis of claim experience is Poisson, and this has been used for initial lasso modelling. The result is that a min CV model often includes some number of very small coefficients, seemingly with very little influence on the model.

The statistical significance of these coefficients was tested as follows. In the case of each data set considered here, a GLM was fitted to the full data set, including only those covariates included in the min CV lasso model and assuming, in common with the lasso, a Poisson error distribution. The resulting regression coefficients do not equal those of the lasso, because the GLM is unpenalized. The majority are found to be significant.

However, the conclusion on significance fails to recognize the effect of the very restrictive Poisson assumption that the mean of observations equals the variance. In practice, this can estimate an unrealistically low variance. The standard errors associated with estimated regression coefficients reflect this variance, and can also be unrealistically low.

This creates a very low hurdle for significance of the coefficients, and consequent over-statement of that significance. In order to overcome this shortcoming, an experimental GLM was fitted, again including only those covariates included in the min CV lasso model, but now assuming more realistic error distributions, either over-dispersed Poisson or gamma.

The major effect was that the standard errors of coefficients were considerably expanded relative to those of the Poisson GLM, significance was reduced, and a much more economical model resulted. This model is referred to as a **thinned model**.

In common with other Bayesian estimators that shrink towards a prior mean, lasso estimates of regression coefficients are biased. The thinned model, consisting of an unpenalized GLM, would mitigate this bias.

The thinned model is preferable when it does not degrade fit or forecast performance. However, this was not always the case. It usually performed almost as well as the unthinned model in fitting to observations. In forecasting, it usually performed well on the real data set, but poorly on the synthetic, though even this was not consistently the case.

Model thinning might be a useful area of further investigation, and it might be useful in the case of more highly supervised modelling. However, in view of the uncertainty as to its performance, the thinned model is not recommended as a self-assembling model at this stage. A possible future project might consist of revisiting this question.

6.2. Bayesian lasso

The version of the lasso used in this paper is non-Bayesian. A Bayesian version is also available.

A Bayesian interpretation of the lasso model (where the penalty term is interpreted as a Laplacian prior on the parameters; see equation (3.8)) permits the usage of standard Bayesian machinery to infer distributional properties for the parameters as well as the observation error, which allows a broader estimate of uncertainty.

One common approach is to interpret the Laplacian prior as a mixture of Gaussians which means that the β_j and their corresponding variance parameters can be estimated as a hierarchical model; see for instance Park and Casella (2008). This had the added advantage of normal distributions throughout, which is a convenient assumption for conjugacy and Gibbs sampling.

The joint sampling of the distribution then allows estimation of the full posterior distribution of other statistics, such as the reserve implied by a set of β estimates. This might be valuable for the estimation of variability and quantiles of reserve, including allowance for internal systemic error. As in the case of the bootstrap (Section 5.1), such estimates would include no allowance for external systemic error.

The Bayesian approach also presents options for the estimation of the penalty term λ. There are plug-in estimates possible using empirical Bayes approaches, or a low-information prior can be applied to give a posterior distribution for λ, jointly with the other parameters. Again, see Park and Casella (2008).

7. Conclusions

The objective of this paper is to identify an automated system of claims experience modelling that will track complex data sets sufficiently well without supervision. The lasso, with judiciously chosen basis functions as covariates, seems to achieve this. This is subject to the issue of feature selection, discussed in Section 4.1, and which might sometimes require a brief preliminary investigation.

A routine procedure has been developed, and once the basis functions have been specified, no parameter input nor supervision is required. The model will self-assemble the model that is optimal according to the lasso criterion.

This procedure has been applied to both synthetic and real data. The synthetic data contain known complexities, and so form a control against which to assess any model. The lasso-based procedure appears to identify these features and estimate that describe them reasonably accurately.

The real data set relates to Auto Bodily Injury claims, and so is comparatively long tailed. Eccentricities of the data cannot be known with certainty. However, as the authors have more than 15 years' experience with the data set, a number of features are known with reasonable confidence. Of course, real data might also contain other unknown subtleties.

The data set contains a number of features that are awkward for traditional claim modelling ("traditional" here means those that stop short of a GLM framework). They include changes in all three of row, column and diagonal effects. Even within a GLM, considerable time and effort is required to explore these features thoroughly, and account for them satisfactorily in a model.

But the lasso-based procedure appears to identify them and model them relatively accurately. This reduces a claims modelling assignment from a duration of possibly several days of senior analyst time to a few hours of junior analyst time. It was found that the lasso model closely reproduces the forecasts of a manually and expensively custom built GLM. However, a custom built GLM is likely to provide a less abstract representation of the data, be more interpretable, and yield insight into the dynamics of the claim process with less analysis than would be required by the lasso for the same outcome.

One weakness that emerged was failure of the model to recognize instantaneous material changes such as might result from legislation with a drop-dead date for changes.

However, where the occurrence (though perhaps not the effectiveness nor efficiency) of such changes can be anticipated, as would be the case for legislation, the model is capable of modification in such a way as to enhance its recognition of the changes. This aspect was tested in the case of real data, and found satisfactory.

The proposed procedure is applicable to data at any level of granularity, from traditional aggregation to individual transaction level. The required set of basis functions may vary from one application to another, and time spent in careful choice of these will probably repay itself.

Estimation of prediction error may be performed by one of several bolt-on procedures that are already well known. Of particular note is the fact that one of them, non-parametric bootstrapping of the entire lasso, will provide an estimate of prediction error that includes at least a part of model error.

To the authors' knowledge, no other documented loss reserving procedure does this, though one might conjecture that some future machine learning methods that contemplate a universe of alternative model forms (e.g. neural nets), will be able to do so.

It must be said that the lasso-based procedure discussed here can be computer-intensive. While a fit to an aggregate triangle data can be relatively fast, a single model fit (with cross-validation) to a large unit record dataset was not possible on a relatively heavy-duty PC..

A few cautionary comments. First, although the model can accurately estimate past effects, such as variations in SI and variations in claim payment delays due to changing rates of claim settlement (and others), it is not an oracle. It cannot pronounce on future rates of SI and claim settlement. These must be inserted "by hand" into any forecasts.

The unsupervised procedure suggested here amounts to a form of machine learning. One would be well advised, in handing control of one's destiny to a robot, to maintain strict surveillance to ensure adequate performance of the robot.

Although an unsupervised procedure is proposed, it would be advisable to supplement it with strong back-end supervision. This would consist of a number of comparisons of the model with the data to which it is fitted. Examples (by no means exhaustive) appear in Sections 4.2.2 and 4.3.2, and include actual-to-fitted heat maps, plots of actual and fitted response against major individual covariates (AQ, DQ, etc.), and extraction of specific model effects.

Acknowledgement

The authors acknowledge the valuable assistance of Josephine Ngan of the University of New South Wales in the modelling documented in this paper.

The research reported here received financial supported under Australian Research Council's Linkage Projects funding scheme (project number LP130100723).

References

Bibby J & Toutenburg H (1977). **Prediction and improved estimation in linear models**. John Wiley & Sons, Chichester UK.

Efron, B, Hastie, T, Johnstone, I, & Tibshirani, R (2004). Least angle regression. **The Annals of statistics**, *32*(2), 407-499.

Frees J W, Meyers G & Derrig R G (eds.). (2016). **Predictive Modeling Applications in Actuarial Science, Volume II.** Cambridge University Press, New York NY, USA.

Gao G & Meng S (2018). Stochastic claims reserving via a Bayesian spline model with random loss ratio effects. **ASTIN Bulletin**, 48 (1), 55–88.

Hall, P (2013). **The bootstrap and Edgeworth expansion.** Springer Science & Business Media.

Harej B, Gächter R & Jamal S (2017). Individual claim development with machine learning. Report of the ASTIN Working Party of the International Actuarial Association. Available at http://www.actuaries.org/ASTIN/Documents/ASTIN_ICDML_WP_Report_final.pdf.

Hastie T, Tibshirani R & Friedman J. (2009). **The Elements of Statistical Learning: Data Mining, Inference and Prediction**. Springer, New York USA.

Jamal S, Canto S, Fernwood R, Giancaterino C, Hiabu M, Invernizzi L, Korzhynska T, Martin Z & Shen H (2018). Machine learning & traditional methods synergy in non-life reserving. Report of the ASTIN Working Party of the International Actuarial Association. Available at https://www.actuaries.org/IAA/Documents/ASTIN/ASTIN_MLTMS%20Report_SJAMAL.pdf.

Kuang D, Nielsen B & Nielsen J P (2008). Identification of the age-period-cohort model and the extended chain-ladder model. **Biometrika**, 95, 979–986.

Li H, O'Hare C & Vahid F (2017). A flexible functional form approach to mortality modeling: Do we need additional cohort dummies? **Journal of Forecasting**, 36, 357–367.

Mack T (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. **ASTIN Bulletin**, 23, 213–225.

Mulquiney P (2006). Artificial neural networks in insurance loss reserving. **9th Joint Conference on Information Sciences 2006 – Proceedings**. Atlantis Press. http://www.atlantis-press.com/php/download_paper?id=67.

Park T & Casella G (2008). The Bayesian Lasso. **Journal of the American Statistical Association**, 103 (482), 681-686.

Reid D H (1978). Claim reserves in general insurance. **Journal of the Institute of Actuaries**, 105, 211-296.

Risk Margins Taskforce (2008). A framework for assessing risk. Presented to the **Institute of Actuaries of Australia 16th General Insurance Seminar**, 9-12 November 2008, Coolum, Australia.

Sardy S (2008). On the practice of rescaling covariates. **International Statistical Review**, 76(2), 285-297.

Taylor G (2000). **Loss reserving: an actuarial perspective**. Kluwer Academic Publishers, Dordrecht, Netherlands.

Taylor G (2011). Maximum likelihood and estimation efficiency of the chain ladder. **ASTIN Bulletin**, 41(1), 131-155.

Taylor G C (1977). Separation of inflation and other effects form the distribution of non-life insurance claim delays. **Astin Bulletin,** 9 (1977), 219-30.

Taylor G & McGuire G (2016). Stochastic loss reserving using Generalized Linear Models. **CAS Monograph Series, number 3**. Monograph commissioned by the Casualty Actuarial Society, Arlington VA, USA.

Taylor G & Sullivan J (2016). GLMs as predictive claim models. Chapter 3 of Frees, Meyers & Derrig (2016).

Tibshirani R (1996). Regression Shrinkage and Selection via the lasso. **Journal of the Royal Statistical Society. Series B (methodological)**, 58 (1), 267–88

Venter G G (2018). Loss reserving using estimation methods designed for error reduction. At https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3096178.

Venter G G & Şahin Ş (2018). Parsimonious parameterization of age-period-cohort models by Bayesian shrinkage. **Astin Bulletin**, 48 (1), 89–110.

Wüthrich M V & Merz M (2008). **Stochastic claim reserving methods in insurance**. John Wiley & Sons, Ltd, Chichester, UK.

Zehnwirth B (1994). Probabilistic development factor models with applications to loss reserve variability, prediction intervals, and risk based capital. **CAS Forum, Spring 1994, 2**, 447-605.