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

Gráinne McGuire
Taylor Fry

Greg Taylor
University of New South Wales
Joint work with

Hugh Miller, Taylor Fry
Josephine Ngan, University of New South Wales
Outline of presentation

• Motivation

• Regularized regression and the LASSO

• Case studies
  – Synthetic data
  – Real data

• Discussion

• Conclusions
Outline of presentation

• **Motivation**

• Regularized regression and the LASSO

• Case studies
  – Synthetic data
  – Real data

• Discussion

• Conclusions
Motivation

• We consider the modelling of claim data sets containing complex features
  – Where chain ladder and the like are inadequate (examples later)

• When such features are present, they may be modelled by means of a Generalized Linear Model (GLM)

• But construction of this type of model requires many hours (perhaps a week) of a highly skilled analyst
  – Time-consuming
  – Expensive

• Objective is to consider more automated modelling that produces a similar GLM but at much less time and expense
Outline of presentation

• Motivation

• **Regularized regression and the LASSO**

• Case studies
  – Synthetic data
  – Real data

• Discussion

• Conclusions
Regularized regression and the LASSO

• Consider general GLM structure
  \[ y = h^{-1}(X\beta) + \varepsilon \]

• Regularized regression loss function becomes
  \[ L = -2\ell(y; X, \hat{\beta}) + \lambda \|\hat{\beta}\|_p \]

  – Penalty included for more coefficients and larger coefficients, so tends to force parameters toward zero
    • \( \lambda \to 0 \): model approaches conventional GLM
    • \( \lambda \to \infty \): all parameter estimates approach zero
    • Intermediate values of \( \lambda \) control the complexity of the model (number of non-zero parameters)

  – Special case: \( p = 1 \), **Least Absolute Square Shrinkage Operator (LASSO)**
    \[ L = -2\ell(y; X, \hat{\beta}) + \lambda \sum_j |\hat{\beta}_j| \]
Outline of presentation

• Motivation
• Regularized regression and the LASSO
• **Case studies**
  – **Synthetic data**
  – Real data
• Discussion
• Conclusions
Synthetic data sets: construction

- Purpose of synthetic data sets is to introduce known trends and features, and then check the accuracy with which the lasso is able to detect them.
- 4 data sets with different underlying model structures considered:
  - In increasing order of stress to the model.
- Notation:
  - $k$ = accident quarter ($= 1, 2, \ldots, 40$)
  - $j$ = development quarter ($= 1, 2, \ldots, 40$)
  - $t = k + j − 1 = payment quarter$
  - $Y_{kj} = incremental paid losses in (k, j) cell$
  - $\mu_{kj} = E[Y_{kj}], \sigma^2_{kj} = Var[Y_{kj}]$
  - Assumed that $ln \mu_{kj} = \alpha_k + \beta_j + \gamma_t$ (generalized chain ladder)
Synthetic data sets: features

\[ \ln \mu_{kj} = \alpha_k + \beta_j + \gamma_t \]

- **Data set 1**: \( \beta_j \) follows Hoerl curve as function of \( j \), \( \gamma_t=0 \) (no payment year effect), \( \alpha_k \) as in diagram
- **Data set 2**: \( \alpha_k, \beta_j \) as for data set 1, \( \gamma_t \) as in diagram
- **Data set 3**: \( \alpha_k, \beta_j \) as for data sets 1&2, \( \gamma_t \) as for data set 2, AQ-DQ interaction (35% increase) as in diagram
- **Data set 4**: \( \ln \mu_{kj} = \alpha_k + \beta_j + \theta_j \gamma_t \), \( \alpha_k, \beta_j \) as for data sets 1-3, \( \gamma_t \) as for data sets 2&3, \( \theta_j \) as in diagram

06 September 2018
Model formulation, selection and performance measurement

• Model formulation
  – Regressors consist of set of basis functions that form a vector space:
    • All single-knot linear spline functions of \( k, j, t \)
    • All 2-way interactions of Heaviside functions of \( k, j, t \)

• Model selection
  – For each \( \lambda \), calculate 8-fold cross-validation error
  – Select model with minimum CV
  – Forecast with extrapolation of any PQ trend (to be discussed later)

• Model performance
  – AIC
  – Training error \[ \text{sum of (actual-fitted)2/fitted values for training data set} \]
  – Test error \[ \text{sum of (actual-fitted)2/fitted values for test data set} \] (N.B. unobservable for real data)
Synthetic data set 1: results

\[ \lambda \text{ decreasing} \]
Synthetic data set 2: results

At DQ 4

At DQ 14
Synthetic data set 3: results

At AQ 20

At DQ 24

Loss reserves
Synthetic data set 4: results

At DQ 5

At DQ 15

Loss reserves: lasso v true expectation

06 September 2018
Outline of presentation

• Motivation
• Regularized regression and the LASSO
• Case studies
  – Synthetic data
  – Real data
• Discussion
• Conclusions
Real data: nature of data set

• Motor Bodily injury (moderately long tail)
• (Almost) all claims from one Australian state
  – AQ 1994M9 to 2014M12
  – About 139,000 claims
  – Cost of individual claim finalizations, adjusted to 31 December 2014 $

  • Each claim tagged with:
    – Injury severity score ("maislegal") 1 to 6 and 9
    – Legal representation: maislegal set to 0 for unrepresented severity 1 claims
    – Its operational time (OT), proportion of AQ’s ultimate number of claims finalized up to and including it
Real data: known data features

• Collectively, presenters have worked continually with data set for about 17 years

• The Civil Liability Act affected AYs ≥ 2003
  – Eliminated many small claims
  – Reduced the size of some other small to medium claims

• There have been periods of material change in the rate of claim settlement

• There is clear evidence of superimposed inflation (SI)
  – This has been irregular, sometimes heavy, sometime non-existent
  – SI has tended to be heavy for smallest claims, and non-existent for largest claims
Real data: lasso model

• Lasso applied to the data set summarized into quarterly cells
  – This summary is not theoretically essential but reduces computing time

• Basis functions:
  – Indicator function for severity score (maislegal)
  – All single knot linear splines for OT, PQ
  – All 2-way interactions of maislegal*(OT or PQ spline)
  – All 3-way interactions maislegal*(AQ*OT or PQ*OT Heaviside)

• Forecasts do NOT extrapolate any PQ trend

• Model contains 94 terms
  – Average of about 12 per injury severity

• By comparison, the custom-built consultant’s GLM included 70 terms
Real data: model fit by DQ

Payments have been scaled.
Real data: model fit by PQ
Real data: model fit by AQ (injury severity 1)
Real data: known data features

• Failure of fit results from data features that were known in advance
  – Legislative change affecting AQ $\geq$ 35

• Perverse to ignore it in model formulation

• Introduce a few simple interactions between injury severity, AQ, OT without penalty
  – Brief side investigation required to formulate these

• Model fit considerably improved
Real data: Human vs Machine

- Same data set modelled with GLMs for many years as part of consulting assignment
  - Separate GLM for each injury severity
  - Many hours of skilled consultant’s time
- Loss reserves from two sources very similar
  - Note that severity 9 is a small and cheap category
- **BUT** consultant’s analysis
  - More targeted
  - Less abstract
  - Conveys greater understanding of claim process
Outline of presentation

• Motivation
• Regularized regression and the LASSO
• Case studies
  – Synthetic data
  – Real data
• Discussion
• Conclusions
Discussion: feature selection

• How many covariates out of AQ, DQ, PQ should be included?
  – Usually at least 2
  – But 3 will generate collinearity
    • Enlarges model dimension
    • May cause mis-allocation of model features between among dimensions
    • So caution before introducing 3
• Make use of feature selection where features are known/strongly suspected

• Implications for forecasting
• Forecasts depend on future PQ effects
  – Should these be extrapolated?
  – How will forecasts be affected by mis-allocation?
• Proposition. Consider data set containing DQ and PQ effects but no AQ effect. Let $M_1$ denote model containing explicit DQ, PQ effects but no AQ effect. Let $M_2$ denote identical model except that also contains explicit AQ effects. Then, in broad terms, $M_1$ and $M_2$ 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.
Discussion: interpretability

• Most machine learning models subject to the **interpretability problem**
  – Model is an abstract representation of the data
  – May not carry an obvious interpretation of model’s physical features
  – Physical interpretation usually possible, but requires some analysis for visualization
Discussion: miscellaneous matters

- **Prediction error**
  - Bootstrap can be bolted onto lasso
  - Preference for non-parametric bootstrap
  - Computer-intensive if min CV chosen separately for each replication
    - Lasso for real data
      - 20 minutes without CV
      - 4½ hours with CV
  - Bootstrap will include at least part of internal model error, but not external model error

- **Model thinning**
  - Most appropriate distribution provided by lasso software *glmnet* is Poisson
  - Low significance hurdle
  - Reduce number of parameters by applying GLM with gamma error and same covariates as lasso
  - Model performance sometimes degraded, sometimes not

- **Bayesian lasso**
  - Lasso can be given a Bayesian interpretation
    - Laplacian prior with $\lambda$ as dispersion parameter
    - Software (Stan) then selects $\lambda$ according to defined performance criterion
Outline of presentation

• Motivation
• Regularized regression and the LASSO
• Case studies
  – Synthetic data
  – Real data
• Discussion
• Conclusions
Conclusions (1)

• Objective was to develop an automated scheme of claim experience modelling

• Routine procedure developed
  – Specify basis functions and performance criteria
  – Then model self-assembles without supervision

• Tested against both synthetic and real data, with reasonable success
  – Lasso succeeds in modelling simultaneous row, column and diagonal features that are awkward for traditional claim modelling approaches

• Procedure is applicable to data of any level of granularity
Conclusions (2)

• Some changes of unusual types may be difficult for an unsupervised model to recognize
  – If these are foreseeable, a small amount of supervision might be added with minimal loss of automation

• Standard bootstrapping can be bolted on for the measurement of prediction error
  – Uniquely, this can be formulated so as to incorporate part of model error (internal systemic error) within estimated prediction error

• As with any form of unsupervised learning, strong back-end supervision is recommended
The views expressed in this [publication/presentation] are those of invited contributors and not necessarily those of the IFoA. The IFoA do not endorse any of the views stated, nor any claims or representations made in this [publication/presentation] and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this [publication/presentation].

The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this [publication/presentation] be reproduced without the written permission of the IFoA [or authors, in the case of non-IFoA research].