Institute of Actuaries of Australia

Combining GLM and data-mining techniques for modelling accident compensation data

Prepared by Peter Mulquiney

Presented to the Institute of Actuaries of Australia
XIth Accident Compensation Seminar 1-4 April 2007
Grand Hyatt Melbourne, Australia

The Institute of Actuaries of Australia
Level 7 Challis House 4 Martin Place
Sydney NSW Australia 2000
Telephone: +61 2 9233 3466 Facsimile: +61 2 9233 3446
Email: actuaries@actuaries.asn.au Website: www.actuaries.asn.au

Abstract

The purpose of this paper is to illustrate the potential advantages of combining GLM and data-mining techniques for modelling accident compensation claims data.

A GLM framework often provides an effective way of modelling claims when a model structure has been determined. However, the process of arriving at the structure by iterative trial models can be laborious and fallible. Artificial Neural Networks (ANN) can provide a useful tool for seeking out influential components of the required structure.

The benefits of combining GLM and ANN are illustrated with a case study using CTP data.

Keywords: Data-mining, GLM, neural networks, CTP.

1. Introduction

Accident compensation data often exhibit features which complicate loss reserving and premium rate setting. Features often observed in the data of accident compensation schemes include:

- Speeding up or slowing down of claim payments;
- Abrupt changes in payment patterns due to, for example, changes in claim management practices;
- Abrupt changes in trends, say due to legislative changes;
- Changes in the profile of claims;
- Seasonality; and
- Other changes which emerge as superimposed inflation.

One method of dealing with these features is through the statistical modelling technique Generalised Linear Modelling ("GLM"). GLM have proven useful for modelling accident compensation data because the structure of GLM can be chosen to represent features such as those listed above.

The structure of a GLM is usually chosen in an *ad hoc* manner, using an iterative series of trial models. This process can be laborious and can be fallible and a more rigorous approach is desirable.

Another modeling methodology that may prove useful for seeking out influential components of the required structure is Artificial Neural Networks (ANN). ANN have some advantages over GLM in seeking out data structures because:

- They have a flexible architecture that can fit almost any data structure; and
- Model fitting is largely automated so there is no need for an *ad hoc* structure identification approach.

ANN are just one of many data-mining techniques which have the advantages listed above. These techniques include CART (Brieman et al, 1984), MARS (Friedman, 1991) and MART (Friedman, 2001).

In the following paper, I consider how ANN may assist in formulating the structure of a GLM. Specifically the paper:

- Examines the general form of a model of claims data;
- Examines the specific case of a GLM to represent the data;
- Considers how the GLM structure is chosen;
- Introduces and discusses ANN; and
- Considers how these may assist in formulating a GLM

I would also like to acknowledge that this paper draws heavily on discussions with my colleague Dr Greg Taylor.

2. General form of claims data model

In general, claims data can be modeled with a function of the following form:

$Y_i = f(\boldsymbol{X}_i; \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_i$, where [1]

- Y_i = some observation on claims experience;
- $\boldsymbol{\beta}$ = vector of parameters that apply to all observations;
- $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ = vector of p attributes (covariates) of the *i*-th observation; and
- $\boldsymbol{\varepsilon}_i$ = vector of centred stochastic error terms.

For example, Y_i, could be equal to Y_{ad}, the paid losses in cell (a,d) of a paid loss triangle where a = accident period and d = development period. Another example could be that Y_i is the cost of the i-th completed claim.

A simple example of a claims data model is the paid loss chain ladder model. In this model:

$$\mathrm{E}[Y_{ad}] = \beta_d \sum_{r=1}^{d-1} Y_{ar} \qquad [2]$$

where β_d are the chain ladder factors.

3. Using GLMs to model claims data

3.1 Form of GLM model

Given a vector of inputs $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, the GLM has a regression function of the form

$f(\boldsymbol{X}_i; \boldsymbol{\beta}) = g^{-1}(\eta_i)$ [3]

where $\eta_i = \beta_0 + \sum_{r=1}^{p} \beta_r x_{ir}$

with β_i being unknown parameters and the variables X_{ir} being:

- direct quantitative inputs such as accident quarter, quarter of finalisation, etc.
- transformations of quantitative inputs such as x_{ir}^2, x_{ir}^3, $x_{ir}^{1/2}$, log (x_{ir}), and $(x_{ir} - c)_+$. The last function in the list is known as a linear spline and the "+" subscript means that the function is zero when $x_{ir} - c$ is negative.
- numeric coding of the levels of a categorical input. For example, for a two level categorical input such as sex we could create $x_{i1} = I$(sex = male) and $x_{i2} = I$ (sex = female). Here I(.) is the indicator function which is 1 when the statement within the parentheses is true and 0 when not. Using this coding, the effect of sex is modelled by two sex-dependent constants.

- Interactions between input variables such as $x_{i3} = x_{i2}. x_{i1}$.

The function $g(.)$ is known as the link function and for many insurance applications, the log function is used for the link function. η is often referred to as the linear predictor.

As indicated by Equation [3], the GLM regression function has a large amount of flexibility. The link function, input transformations, and interaction terms allow one to construct regression functions for quantities which are complicated and non-linear functions of their inputs. This flexibility is one reason for the widespread use of GLMs in actuarial applications.

However, determining the appropriate input transformations and interactions to include in a GLM model can be difficult to do in practice. This is an area where the skill of the model builder can play a large part in determining how well the regression function will model the data.

As an illustration of the types of model that could be constructed with a GLM consider the following simple GLM model:

- Y_i = cost of i-th completed claim;
- $Y_i \sim Gamma$. In other words the centred stochastic error terms in Eqn [1] have a gamma distribution;
- $E[Y_i] = \exp[\alpha_i + \beta_i t_i]$ with
  - α_i = accident period to which i-th claim belongs
  - t_i = operational time at completion of i-th claim = proportion of claims from the accident period a_i completed before i-th claim.

More generally we could model Y_i as:

$$E[Y_i] = \exp[\textit{function of operational time}]. \quad [4]$$

Further, legislative changes may mean that our model could be improved with a model of the form;

$$E[Y_i] = \exp[\textit{function of operational time} + \textit{function of accident period (legislative change)}] \quad [5]$$

Taking this approach further, one may be able to construct a model with the form:

$$E[Y_i] = \exp[\textit{function of operational time} + \textit{function of accident period (legislative change)} + \textit{function of completion period (superimposed inflation)} + \textit{joint function (interaction) of operational time \& accident period (change in payment pattern attributable to legislative change}))] \quad [6]$$

Models of this type may be very detailed and may include a number of features such as:

- Operational time effects (payment pattern effects)
- Seasonality
- Creeping changes in payment pattern
- Abrupt changes in payment pattern

- Accident period effects (legislative change)
- Completion quarter effects (superimposed inflation)
- Variations in superimposed inflation over time
- Variations of superimposed inflation with operational time.

3.2 Choosing GLM structure

A case study of a Motor Bodily Injury (CTP) insurance data set in one state of Australia is presented as an example of choosing an appropriate GLM structure to model a data set. The payments for Motor Bodily Injury are usually dominated by a single lump sum near the date of claim completion (finalisation). Hence a common approach to such payment types is to:

- Model the expected number of claim finalisations to be made at future dates; and
- Model the expected size of completed claims at each future finalisation date.

In the following paper attention is restricted to the model of expected claim sizes, however the general conclusions apply equally to the model of claim finalisations.

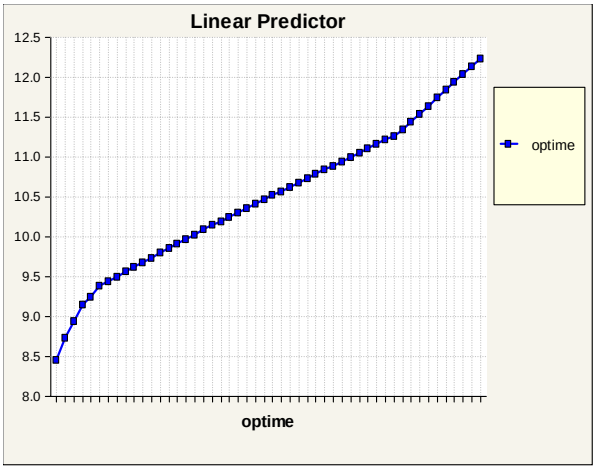
The data set consists of a claim file with approximately 60,000 claims. For each claim various items were recorded, including the date of injury, date of notification, and histories of paid losses, case estimates and finalised/unfinalised status including dates of change of status.

The GLM structure was chosen in a largely *ad hoc* manner by using:

- Trial and error regressions; and
- By looking at a number of statistical diagnostics such as residual plots, AIC values, etc.

For example, in the first instance a model with just an operational time effect was fitted to the data (Figure 1).

Figure 1 Individual claim regression estimate of trend in average claim size by operational time.

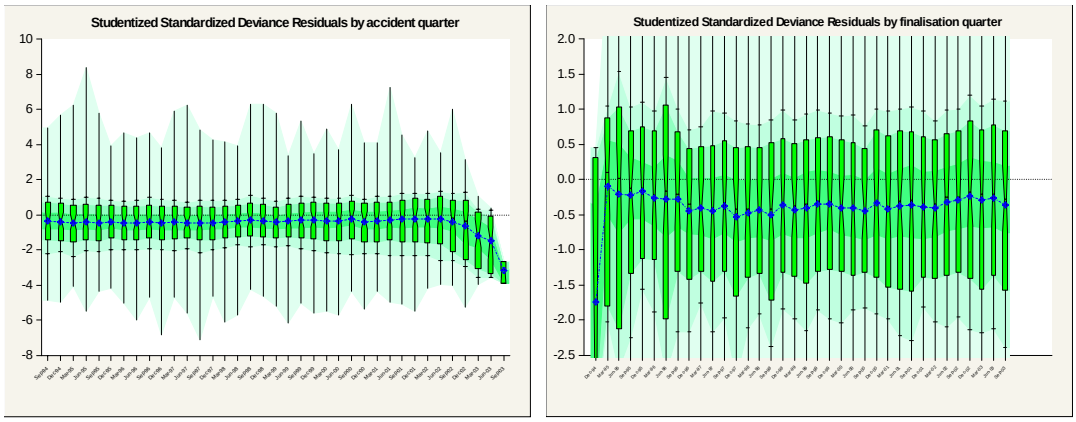


Residuals were then examined as a function of the other predictors that were not included in the model. Examination of these residuals revealed that there appeared to be unmodelled trends by:

- Accident quarter (Figure 2); and
- Completion (finalisation) quarter (Figure 2).

For example, there is a clear trend in the extreme right of the accident quarter residual plot (seen in the blue line).

Figure 2 Deviance residuals as a function of accident quarter and finalisation quarter



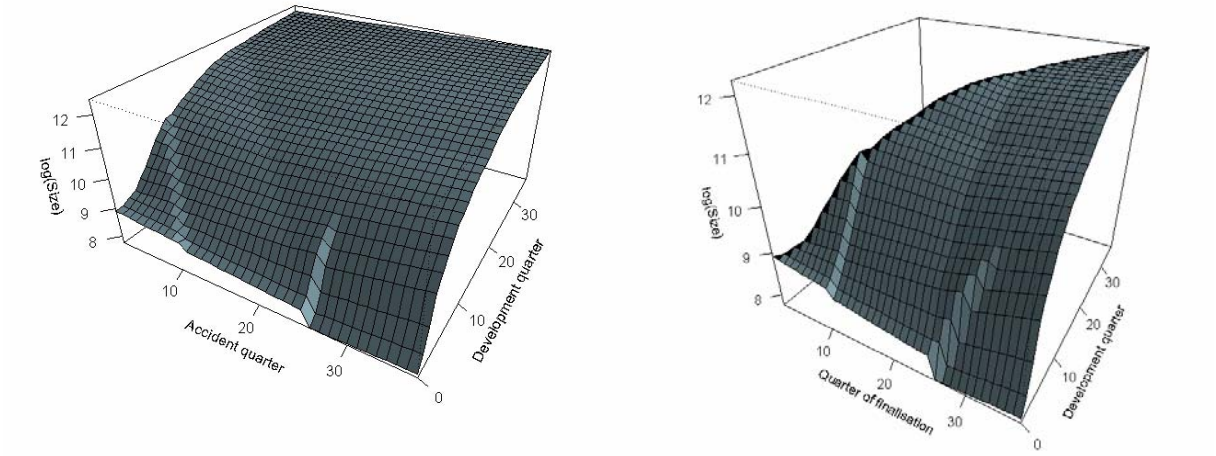
Terms which were functions of accident quarter and finalisation quarter were then introduced into the model and the process of trial and error model refinement continued.

By the end of the structure selection process the model had a complex structure and included terms for:

- Operational time - the average size of finalised claims increases with operational time;
- Seasonality;
- Claim frequency – a decrease induces increased claim sizes;
- Accident quarter. This feature resulted from legislative changes that came into effect in September 2000. This legislation placed limitations on the payment of plaintiff costs and effectively eliminated a certain proportion of smaller claims in the system in all subsequent accident quarters;
- Change in operational time effect with change in Scheme rules; and
- Superimposed inflation (payment quarter trends) varying with operational time. This brings out the feature that smaller and larger finalised claims are subject to different rates of superimposed inflation.

More details of the model fitting process are found in Taylor and McGuire (2004). Two-dimensional plots of the linear predictor of the GLM model are shown in Figure 3.

Figure 3 Plot of the linear predictor of the GLM model. To smooth these plots I have assumed that the rates of finalisation in each accident quarter are equivalent and I have ignored the effect of seasonality.



In general, a more rigorous approach to structure selection seems desirable compared to the one presented above. Artificial Neural Networks may be an alternative approach to seeking out influential components of the required structure. They are introduced in the following Section.

4. Artificial Neural Networks (ANN)

4.1 Form of ANN Model

In the previous section, we saw that the basic approach of modelling with GLM was to match the structure of the regression function to the data in an iterative trial manner. The approach of modeling with ANN is different. Instead of matching the model to the data, the ANN regression function is given an initial structure that is so flexible it can model almost anything. Careful fitting is then used to constrain the function so that it will only describe the underlying features of the data.

Starting with a vector of p inputs $\boldsymbol{X} = (x_1, x_2, \ldots, x_p)$, we can construct a neural network regression function as follows. First we create M linear combinations of inputs

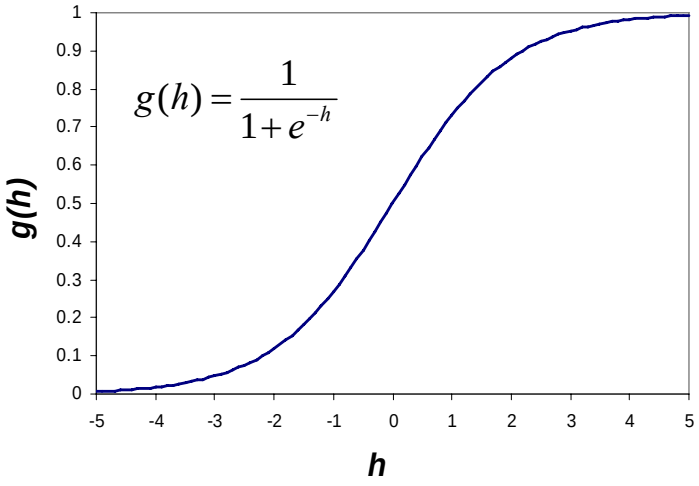
$$h_m = \sum_{i=1}^{p} w_{mi} x_i \qquad [7]$$

The actual value that we choose for M will be determined in the tuning/fitting process. These M linear combinations are then passed through a layer of activation functions $g(h_m)$ (Figure 4) to produce the outputs Z_m

$$Z_m = g(h_m) = g(\sum_{i=1}^{p} w_{mi} x_i) \qquad [8]$$

These first steps correspond to the middle (or hidden) layer of the neural network (Figure 5).

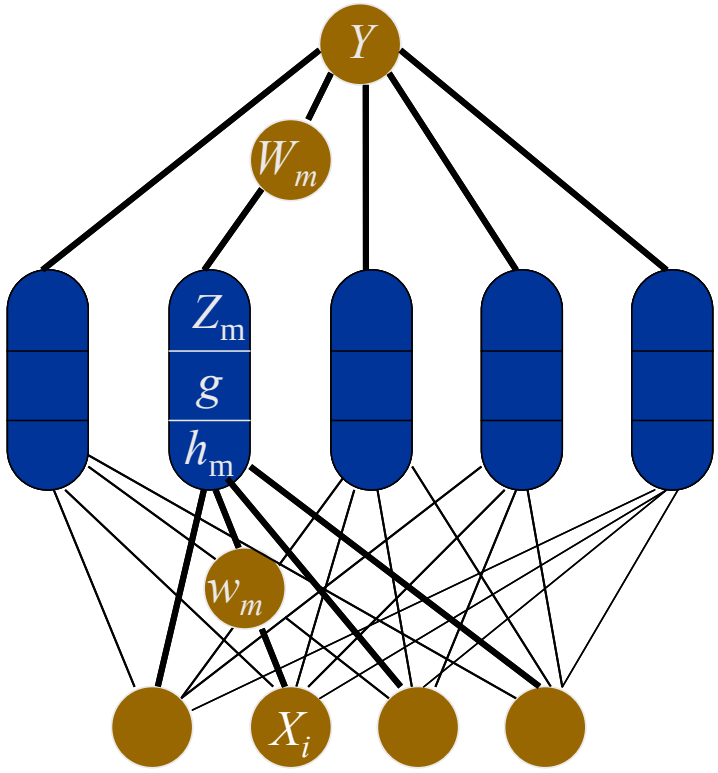
Figure 4 A sigmoidal activation function. A sigmoidal curve is usually chosen as it introduces non-linearity into the regression function while keeping responses bounded.



The regression function is then taken to be a linear combination of the outputs from the hidden layer.

$$f(X) = \sum_m W_m Z_m = \sum_m W_m g(\sum_{i=1}^{p} w_{mi} x_i) \qquad [9]$$

Figure 5 The structure of ANN. This neural network has a single hidden layer with 5 hidden units (M = 5). Figure adapted from Gershenfeld (1999).



The parameters of this regression model are the weights. In their simplest form these regression functions will have $(p+1) \times M$ parameters. Typically there are many more parameters in a neural network regression function compared to a GLM regression function.

As might be expected, this structure produces a regression function that is very flexible. Indeed, it has been shown that a neural network regression function with a single hidden layer and enough hidden units can describe any continuous function to any desired degree of accuracy. Further, if you introduce a second hidden layer, it can be shown that the neural network can describe any function with a finite number of discontinuities.

Weights are usually determined by minimising the least-squares error:

$$Err = \frac{1}{2} \sum_{i=1}^{N} (y_i - f(x_i))^2 \qquad [10]$$

Overfitting is prevented in neural networks by adding a penalty function to the sum of squares error function which becomes larger as the regression function becomes less smooth. The penalty function is typically defined by

$$Err + \lambda(\sum_m W_m^{\ 2} + \sum_m \sum_p w_{mp}^{\ 2}) \qquad [11]$$

where the W_m and w_{mp} are the weight parameters from the neural network regression function (Equation [9]). It is seen that the weight decay parameter, λ, controls the magnitude of the penalty. So by choosing a larger λ, the fitted weights are forced to be smaller and the regression function to be smoother.

λ is typically determined by cross-validation. For cross-validation, the data is randomly divided into a training data set and a test data set. We then fit a number of neural network models to the training data using a number of values of λ. The sum of squares in the test data set is then determined for each of the models. The λ value that minimises the sum of squares in the test set, is the λ value that is chosen. The following references provide greater detail on ANN theory: Bishop, 1995; Hastie et al. 2001 and Ripley,1996.

4.2 Choosing ANN structure

The fitting of an ANN model is illustrated in this subsection using the same data as was used for the previous GLM model. A random subset of 70% of the data was assigned to be the training data set, while the remaining 30% formed the test data set. The tuning parameters were determined using cross-validation and the final neural network consisted of a single hidden layer with 20 units and a weight decay, λ, of 0.05.

The predictive accuracy of the ANN on the test data set compared favourably to the GLM for two different measures (Table 1). In addition, it took significantly less time to fit the ANN compared to the GLM model. The ANN algorithm was largely automated while fitting the GLM required significant input from the model builder.

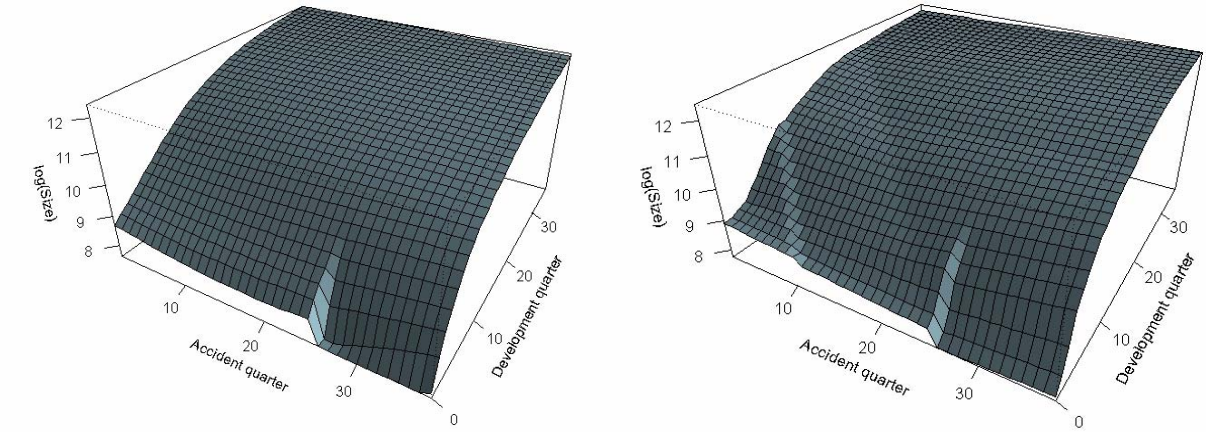
Table 1 Test errors for the ANN and GLM models

Model	Average Sum of squares	Average Absolute Error
GLM	$99,965[2]	$33,777
ANN	$99,843[2]	$33,559

A variety of 1 dimensional residual plots showed that there appeared to be no systematic bias in the model fits across the predictors. The quality of the residual plots was similar between the ANN and the GLM model.

Because the ANN has a complex algebraic structure, it is usually necessary to use graphical means to understand the features of the model. One example is shown in Figure 6 where a plot of the log of the claim size for the ANN model is compared to that produced by the GLM model.

Figure 6: Plot of log(size) for the ANN model (left) and the previous GLM model (right). Smoothing as for Figure 3.



One-dimensional graphical plots can also be useful for visualising data features. For example Figure 7 shows a plot of historical and future superimposed inflation from the ANN. In this figure superimposed inflation is defined to be the gradient of the trend in finalised claim size as a function of payment quarter, with all other predictors held constant.

Figure 7 Historical and projected (from finalisation quarter 38) superimposed inflation for the ANN model as a function of finalisation quarter and development quarter. Development quarter was: red line, 10; green line, 20; yellow line, 30; blue line, 40. All other predictors were constant.

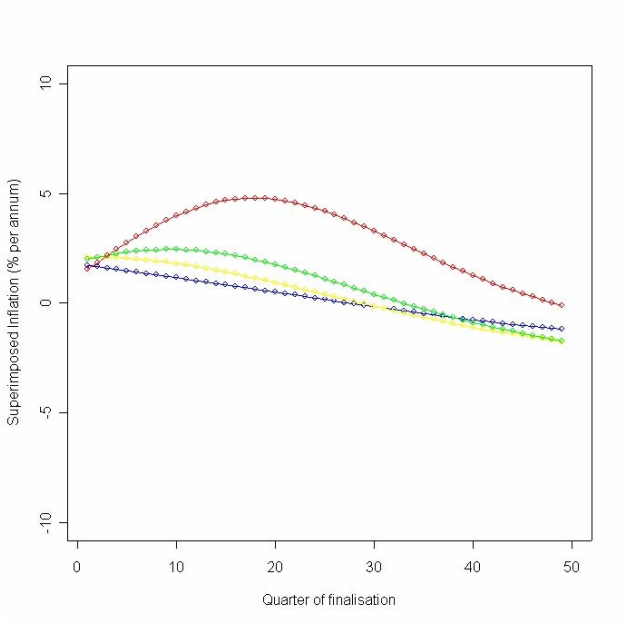


Figure 7 illustrates how the ANN has searched out the general form of past superimposed inflation. Future superimposed inflation has then been determined by simple extrapolation. By simple extrapolation I mean that the fitted ANN function has been used to project values of superimposed inflation for finalisation quarters outside the range of finalisation quarters that were used for model fitting. The main features to note from this figure are:

- ANN have been useful in searching out the general from of past superimposed inflation; and
- Simple extrapolation of the ANN has produced forecasts of negative superimposed inflation.

Forecast negative SI may be undesirable and one would need to consider the expected claims environment in the future to determine an appropriate SI forecast. In this example, simple extrapolation was used for the forecast. However because the ANN is fitted over the range of the input values in the data set, the complex nature of the ANN function means that there is little control of its behaviour outside the input data ranges when simple extrapolation is used.

It is however possible to project future claim size with the ANN without any finalisation quarter (superimposed inflation) effects. Doing this removes the problem of extrapolation in the finalisation quarter dimension. This can be done by keeping the finalisation quarter constant (and equal to the most recent historical finalisation quarter) for all records in the data set used to project future values of claim size.

The projections without superimposed inflation can then be supplemented by a separate model of expected future superimposed inflation.

5. Using ANN to choose GLM structure

In the previous Sections of this paper an illustration was presented of how both GLM and ANN could be used to model important features in insurance claim data. However, there are some reasons why it may be preferable to use GLM over ANN.

In particular GLM models tend to be simpler and more transparent than ANN models. This is illustrated by the fact that the ANN model presented above contained 181 parameters compared to the 13 parameters in the GLM. In addition, the linear predictor of the GLM has identifiable components introduced to model specific features of the claim data. This is of particular advantage when it comes to decisions about how future claim experience should be projected.

However one issue with GLM is that the process of arriving at the structure by iterative trial models can be laborious and fallible.

This suggests that we may get the best out of ANN and GLM if we use them in combination in the following manner:

- Use **ANN** as an automated tool for seeking out trends in data:
  - Apply ANN to the data set
  - Study trends in the fitted model against a range of predictors or pairs of predictors using graphical means (Figures 6 and 7 provide some simple examples).

- Use this knowledge to choose the functional forms to include in the **GLM** model.

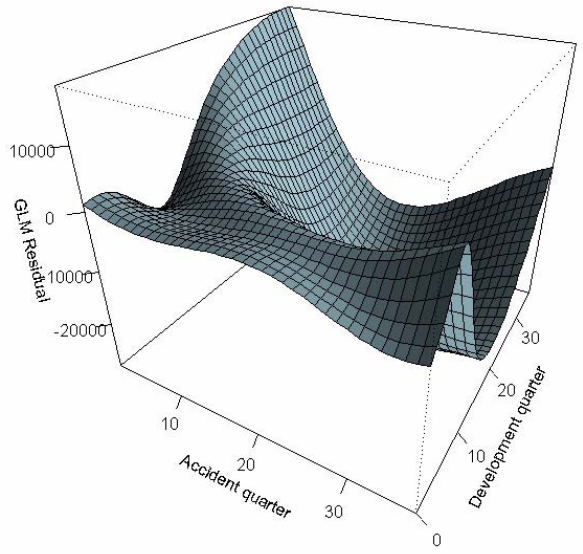
I note that it would be possible to perform the same type of analysis using other data-mining techniques such as CART and MARS. In some cases the more transparent functional forms behind the CART and MARS models may aid in identifying important features of the data. However in other cases the greater flexibility of the ANN architecture may be better at picking out the features of the data.

In all cases ,however, I acknowledge that while the data-mining techniques allows one to visualise the gross features of the data, there may still be some difficulty in translating this knowledge into an appropriate GLM model. In many cases the final model may be no better than could be constructed using the usual *ad hoc* approach to GLM fitting.

However an ultimate test of the GLM would be to apply ANN to the residuals from the GLM fit in an attempt to seek out any remaining structure. If the GLM fit is good there should be no remaining structure in the residuals. Of course, the parameters of the ANN would need to be determined using cross-validation.

In the following diagram (Figure 8) an ANN was fitted to residuals from a main effects GLM model. A main effects model is one in which no interaction terms have been included.

Figure 8 Neural Network fit to the residuals from the main effects GLM model



The figure indicates that the chosen GLM structure may:

- Over-estimate the more recent experience at the mid-ages of claim
- Under-estimate it at the older ages

6. Conclusions

GLM provide a powerful and flexible family of models for claims data. However, complex GLM structures may be required for adequate representation of the data. The identification of these may be difficult and the identification procedures are likely to be *ad hoc*.

ANN provide an alternative form of non-linear regression to GLM. However, the use of ANN to model claims data is likely to involve its own shortcomings if the ANN are left to stand on their own. These shortcomings relate, in particular, to their reduced transparency in comparison to GLM. ANN may, however, provide considerable assistance if used in parallel with GLM to identify GLM structure.

7. References

Bishop, C,1995, *Neural networks for pattern recognition*, Clarendon Press, Oxford.

Breiman L, Friedman J, Olshen R, and Stone C, 1984, *Classification and Regression TreesI,* Chapman & Hall.

Friedman J, 1991, *Multivariate adaptive regression splines (with discussion),*Annals of Statistics 19, 1-141.

Friedman, J, 2001, *Greedy function approximation: the gradient boosting machine*, Annals of Statistics 39, 1189-1232

Hastie, T, Tibshirani, R & Friedman, J, 2001, *The elements of statistical learning*, Springer-Verlag.

Gershenfeld, N, 1999, *The nature of mathematical modeling*, Cambridge University Press.

Ripley, B,1996, Pattern *recognition and neural networks*, Cambridge University Press.

Taylor, G & McGuire, G, 2004 *Loss reserving with GLMs: A Case Study*, Casualty Actuarial Society 2004 Discussion Paper Program.