British Actuarial Journal, Vol. 17, part 3, pp. 491–509. © Institute and Faculty of Actuaries 2012
doi:10.1017/S1357321712000190
First published online 16 October 2012

Redefining the deviance objective for generalised linear models

A. C. Lovick* and P. K. W. Lee

[Presented to the Institute and Faculty of Actuaries, London: 28 March 2011; Norwich: 6 June 2011]

Abstract

This paper defines the 'Case Deleted' Deviance - a new objective function for evaluating Generalised Linear Models, and applies this to a number of practical examples in the pricing of general insurance. The paper details practical approximations to enable the efficient calculation of the objective, and derives modifications to the standard Generalised Linear Modelling algorithm to allow the derivation of scaled parameters from this measure to reduce potential over fitting to historical data. These scaled parameters improve the predictiveness of the model when applied to previously unseen data points, the most likely being related to future business written. The potential for over fitting has increased due to number of factors now used, particularly in pricing personal lines business and the advent of price comparison sites which has increased the penalties of mis-estimation. New material in this paper has been included in a UK patent application No. 1020091.3.

Keywords

Generalised Linear Modelling; General Insurance Pricing; Parameter Uncertainty; Case Deletion; Deviance; Non-Linear Modelling; Demand Modelling; Price Comparison Site Pricing; Winner's Curse.

1. Introduction

1.1 Current Statistical Techniques

1.1.1 Current modelling techniques use the generalised linear modelling framework to estimate parameters for a given model structure based upon calculating the minimum deviance (maximum likelihood) estimates for the parameters from a given dataset.

1.1.2 First the structure of the model to be produced (both appropriate link function and distribution) is established using an understanding of the data, and then considering residual plots and the results of the Tweedie distribution test (Box-Cox transformation).

1.1.3 The significance of parameter estimates can then be judged according to the standard errors calculated from the information matrix, and various statistical tests for example the Chi Squared and F-Tests can be used to compare two competing models.

*Correspondence to: Tony Lovick, Towers Watson, Saddlers Court, 64–74 East Street, Epsom, Surrey, KT17 1HB. Tel: +44 (0)1372 751060; E-mail: Tony.Lovick@TowersWatson.com

1.1.4 A range of other statistics such as the Akaike Information Criteria 'AIC', and Bayesian Information Criteria 'BIC' can also be considered.

1.1.5 These statistical approaches were originally utilised in the context of relatively few factors and levels and relatively few interactions. The range of factors, number of levels within factors and the number of interactions has increased significantly in UK personal lines insurance as insurers have sought competitive advantage and more recently to prevent anti-selection on price comparison sites ('Winner's Curse').

1.2 Short-comings of Statistical Techniques

1.2.1 As discussed, over time the size of modelling datasets has increased (datasets up to 100m rows are becoming more common), and this has highlighted the differences between academic methods designed for a few thousand rows and actual insurance specific models deployed to determine prices.

1.2.2 In particular the Degrees of Freedom is defined by the number of rows of data – number of parameters (unaliased). This becomes effectively constant where the dataset is large, and the parameter list rarely exceeding 1000.

1.2.3 The Deviance for a model decreases as new parameters are added. Hence the ratio of residual deviance to number of degrees of freedom always improves when the degrees of freedom is effectively constant. This causes Chi Squared tests on nested models and F-Tests to accept parameters which would be rejected from a business perspective as spurious and over-parameterised.

1.2.4 For example Figure 1 shows a completely random factor which when added to the dataset, proves to be a significant factor using traditional methods. Both the parameter values are accepted by the standard errors, and the Chi-Square test proves significant. If a 95% significance test is used, then this would be expected to happen once in twenty times.

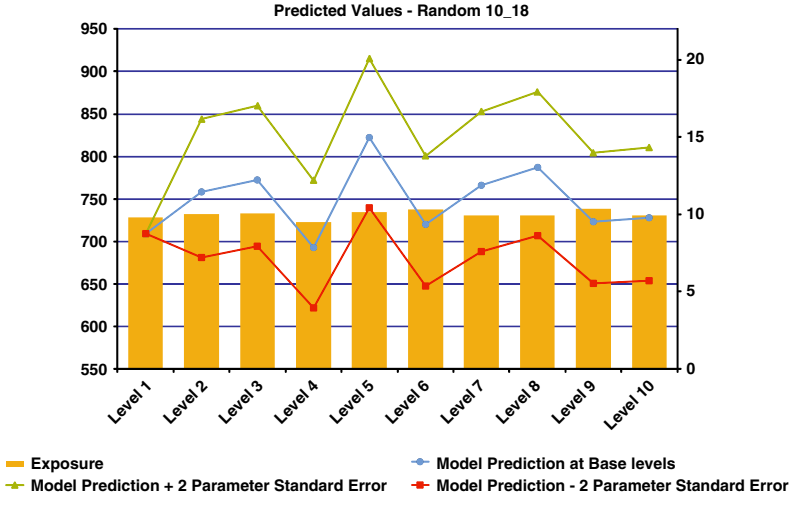

Figure 1. A Random Factor – the 18th attempt

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

1.2.5 The likelihood of over-fitting clearly increases the more parameters that are added to the model, be it factors, levels or interactions.

1.3 Current Business Techniques

1.3.1 There is wide recognition that modelling is not a pure science and better results can be achieved using domain knowledge (by applying some 'art'). The statistical techniques are usually supported by checking the models against business understanding of the factors, their usual significance and trends from past time periods and other datasets.

1.3.2 Time consistency testing is used to ensure that a factor shape is both consistently present for given time periods, and to establish if there is trend for the shape to strengthen, weaken or change shape over time. This is essential if the chosen values of the parameters are to be predictive for a future time period, which is normally the business objective.

1.3.3 To mitigate the problems outlined in 1.2, extensive use is also often made of hold-out sample data. This is where a model is built on a sample of the data, say 80% (modelling or training data) and then the performance is judged by comparing results scored against the remaining 20% (hold-out sample).

1.4 Price Comparison Website Developments, Efficient Market, Winners Curse

1.4.1 In recent years the rise of price comparison websites, particularly in the UK motor market, has created a near perfect market for consumers. Which, coupled with the fact that many view motor insurance as a commodity product, has resulted in observed new business elasticities ranging in magnitude from 10 to 100.

1.4.2 The estimates from a pricing model are best estimates in the statistical sense and hence are subject to uncertainty. In these circumstances the Winner's Curse operates as a powerful anti-selection effect which imposes a heavy penalty where the uncertainty randomly results in an estimate which is below the true value.

1.4.3 In this business context insurers have responded by increasing the range of factors, levels within factors and number of interactions as they have tried to minimise the level of anti-selection. But in doing so there is an increased likelihood of over-fitting which does present a real business dilemma. When presented with a new factor to implement which makes sense from a business viewpoint and is significant, the view will, more often than not, be to introduce the factor. In fact it is very likely that when one systematically reviews the inclusion of each term in a sophisticated model that a business sense argument can be made for each and every one, but it is likely when taking together there will be an element of over-fitting.

1.4.4 In addition to the parameter estimates, the modelling process makes available results which reveal the uncertainty attached to these expressed as a Variance/Covariance matrix. Also the Hat matrix which displays the influence that each data point has had on its corresponding estimate.

1.4.5 There are a number of elements which influence how this uncertainty varies from model to model, and by risk within the model. Two elements of this uncertainty will be tackled by this paper; the remainder can only be properly treated in another paper.

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

1.4.6 The first is the tendency for over-parameterised models to replicate noise within the data which will not be repeated in future observations. This noise is one source of estimate uncertainty.

1.4.7 The second is the tendency for models to be used over a heterogeneous domain. Some areas of the domain are well populated and hence estimates are subject to less uncertainty. The fringes of the domain which tend to be sparsely populated with observations resulting in greater levels of uncertainty. Extrapolation to future time periods is a special case of this which is necessary for the deployment of most predictive models.

1.4.8 The aim of this paper is to consider if the 'best estimates' produced by a model can be adjusted to make them more predictive.

1.4.9 More formally the aim is to temper the model outputs by penalising uncertain parameters to the extent that they are only rewarded for improving the likelihood (reducing Deviance) of the estimates as measured against hold-out sample data.

1.5 Nirvana Method of Scaling Back Parameters

1.5.1 A true mechanism for scaling back parameters will also help in another way. Simply allowing a model to become over-parameterised as it is developed, and reporting parameter errors which state they are not significant is not enough. With this method we can go further, and scale back the poor parameters effectively neutralising them from the model. Pruning processes can then operate to remove them altogether. This will allow the user to focus upon finding potential factors in the knowledge that unsuccessful attempts will not damage the output.

1.5.2 A company may choose to build the model on top of a market rates model, so that rather than scaling back towards the mean, parameters are scaled back towards market rates instead. Therefore a company would only differ from market structures where it had sufficient data to confirm a significant difference in experience.

2. Case Deletion

2.1 Elimination of Outliers by Case Deletion

2.1.1 Using a measure of residuals such as the Cook's Statistic, Outlier points can be excluded from the model based on their undue influence on the parameter estimates.

2.1.2 This technique is supported by leading statistical packages, but for datasets of the scale currently in use, deleting outliers is an onerous and unproductive task.

2.2 Effect of Case Deletion

2.2.1 In essence each data point acts to pull the model towards itself, and the exclusion of that point and refitting the parameters will result in a new set of parameter values and hence new "Case Deleted" Estimate for that data point. By definition that estimate will lie further from the observed data point than the estimate produced by the full model (see figure 2).

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

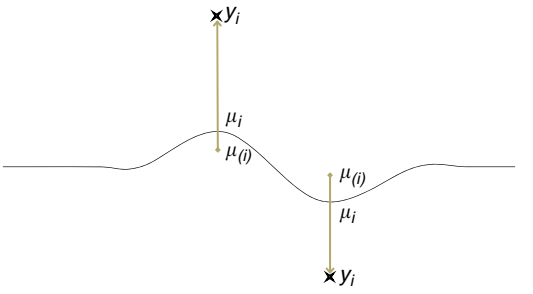

Figure 2. Standard and 'Case Deleted' estimates

2.3 "Case Deleted" Deviance

2.3.1 The Standard Deviance is a measure of the distance from the observed values to the estimate. For an identity link, Normal distribution error structured model, this equates to the sum of squared distances of each observed to estimate. For more general model structures the parameters are found by maximising the likelihood function, and the deviance is defined from this. However it can still be usefully thought of as a form if distance measure.

2.3.2 In the extreme case where the model contains a parameter for every data point, the estimates and the observed values will be equal and the deviance will have a minimum value. The model here is replicating both the Pattern in the data and the Noise.

2.3.3 Taking the set of 'Case Deleted' Estimates, one for each data point, provides a means to calculate a new 'Case Deleted' Deviance. This is in effect the limiting case of calculating the estimate for a hold-out sample of one row against a model based on 'n-1' rows, as the new estimate is not influenced by the observed value itself. The deviance is then calculated in the normal way from the estimates, and summing over the dataset.

2.3.4 Because the 'Case Deleted' Deviance is calculated from estimates which are independent of the observed values it represents the pattern but without the noise related to the observed data point in question. An extreme model will still include noise generated by the other data points, but provided the data points are independent these should average to zero.

2.3.5 A number of practical tests have been conducted comparing the Standard and 'Case Deleted' Deviances. From these it is helpful to define some terms. Let SD_1, SD_2 be the Standard Deviances from a base model and an adjusted model. If the adjusted model is created by adding parameters to the base model, then we know that $SD_1 > SD_2$. Similarly take CDD_1, CDD_2 to be the 'Case Deleted' equivalents. Interestingly it is possible for CDD_2 to be larger than CDD_1 in circumstances where the extra parameters are adding more noise to the model than pattern.

2.3.6 To assist with the investigation of these intuitive concepts of 'pattern' and "noise", we will now propose a more formal definition. $Pattern_{1,2} = CDD_1 - CDD_2$ and $Noise_{1,2} = SD_1 - SD_2 - Pattern_{1,2}$.

This allows us to consider the value of these measures and compare them to existing tests.

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

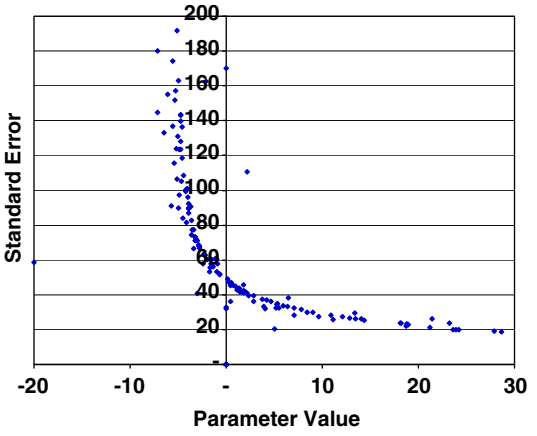

Figure 3. Correlation between the 'Value' Measure and Standard Errors

2.4 Correlation with Standard Errors

2.4.1 The first example involved a Log-Poisson model with an Accident Damage Frequency dataset containing 1m rows, with around 200 parameters covering a range of factors.

2.4.2 For each parameter in the model a new sub-model was created with that single parameter deleted. Then the Noise and Pattern measures were calculated between the full model and the sub-model.

2.4.3 When testing a parameter with a normal distribution error structure, a 95% significance test would reject parameters where the standard error exceeds 50% of the parameter value itself. For other error structures the 50% ratio is also used as an acceptance threshold.

Pleasingly the two tests showed a strong correlation. Defining $Value_{1,2} = Pattern_{1,2} - 5 * Noise_{1,2}$ (x-axis) shows a positive value when the Standard Error % (y-axis) is less than 50% and negative above, as Figure 3 demonstrates. While 5 appears a sensible value to choose in this example, work remains to investigate whether this is a stable value, or changes for other model structures and datasets.

3. Applications to Model Comparison

3.1 Random Factor Example

3.1.1 Using the same random factor example mentioned in 1.2.4

3.1.2 Deviance values calculate as $CDD_1 = 10242$, $CDD_2 = 10249$ $SD_1 = 14631$, $SD_2 = 14605$ and this gives $Pattern_{1,2} = CDD_1 - CDD_2 = -7$ and $Noise_{1,2} = SD_1 - SD_2 - Pattern_{1,2} = 33$ which would reject the random factor as detrimental $Value_{1,2} = Pattern_{1,2} - 5 * Noise_{1,2} = -172$.

3.1.3 Hence the Case Deleted Deviance rejects the random factor where traditional tests would accept it.

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

3.2 Optimal Knot Position Application for Factor Splines

3.2.1 This example is a case study using the same dataset as 2.4.1, and using the 'Value' measure from 2.4.3 above.

3.2.2 For a number of factors Policyholder Age, Vehicle Group, Rating Area, NCD, Convictions, Number of Years Licence Held, a number of different spline functions were compared. In all cases a standard cubic, with x, x^2 and x^3 terms, is fitted to the data. Then in addition, one or many corrections to the standard cubic were added corresponding to the different knot positions.

3.2.3 The first line shows how the 'Value' measure varies as the knot position for the spline is varied. The second line shows the effect when the first knot position is fixed and the position of a second knot is changed and so forth. The charts below show the knot position on the x-axis, and the 'Value' measure on the y-axis. The best position for the knot was selected and then the process repeated adding another knot. Continuing until an extra knot reduces 'Value'.

3.2.4 This is not an efficient process since the smooth results obtained indicate the maximum 'Value' position could be obtained with fewer steps. However the result graphs are more complete if every position is calculated for the result charts below.

3.2.5 Once an extra knot has been added, this method did not recheck that the existing ones should remain in their current positions. An efficient implementation would first derive the number of knots required, then find their approximate positions, and finally jiggle them to find the global optimum.

3.2.6 Although the process of calculating the additional 'Noise' was performed at each step, this turned out to be quite stable, hence the knot position could be estimated from the unadjusted deviance alone. The 'Noise' adjustment only being needed to define the absolute 'Value' of adding an extra knot.

3.2.7 The method suggests two knots at ages 17 and 49, but rejects a third one at 53 (see figure 4).

3.2.8 Standard Error values would also accept these two knots and reject the third, as would F-Tests.

3.2.9 Vehicle Group

3.2.9.1 This factor suggests two knots at 5 and 19 (see figure 5).

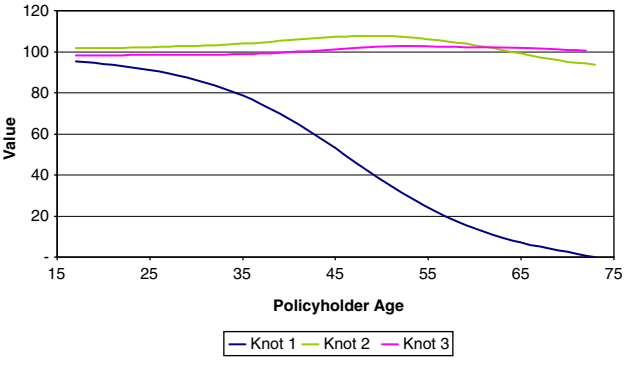

Figure 4. The 'Value' Measure as the Number and Position of Knots are varied by Policyholder Age

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

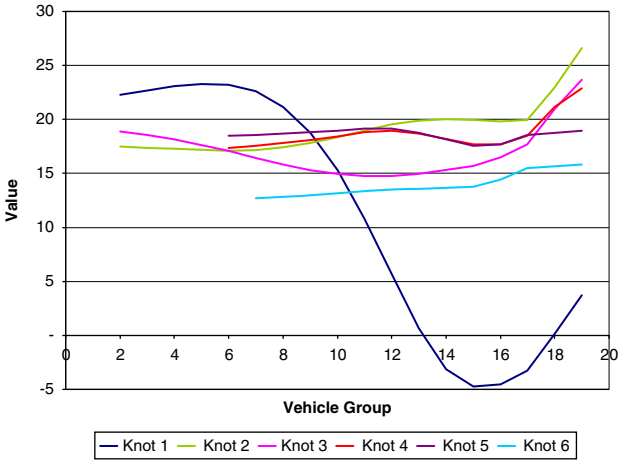

Figure 5. The 'Value' Measure as the Number and Position of Knots are varied by Vehicle Group

3.2.9.2 The Standard Error test is confusing here. Testing each of the knot definitions built up by the procedure, ultimately it would accept knot positions 5 and 2, but only from a five knot spline.

Spline Description	Accepted knots	Rejected Knots
(5, 19)	5	19
(5, 19, 2)		5, 19, 2
(5, 19, 2, 18)		5, 19, 2, 18
(5, 19, 2, 18, 6)	5, 2	19, 18, 6

3.2.9.3 The F-Test considers the splines (5) similar to (5, 19), (5, 19, 2) and (5, 19, 2, 14), but claims the spline (5, 19, 2, 14, 6) is different from (5, 19, 2, 14) yet similar to (5, 19, 2).

3.2.9.4 Hence the 'Value' measure appears useful as a global absolute statistic. The standard error only describes the certainty of an individual parameter, and becomes difficult when the SE values of several parameters vary from one model to the next. The F-Test only describes if two models are significantly different, not if one is better than the other. Whilst assuming the simplest model from two similar models, and the complex model from two different models is standard practice, this does not ensure that the ultimate model is the most predictive.

3.2.10 Rating Area

3.2.10.1 This factor gives two knots at 1 and 29. In agreement with SE and F-Tests (see figure 6).

3.2.11 NCD

3.2.11.1 This factor gives one knot at 4. In agreement with SE and F-Tests (see figure 7).

3.2.12 Conclusion

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

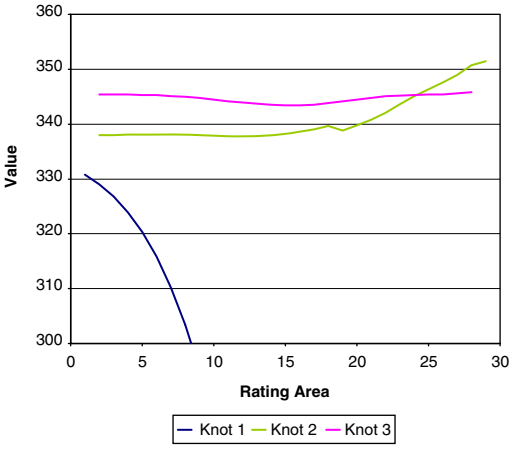

Figure 6. The 'Value' Measure as the Number and Position of Knots are varied by Rating Area

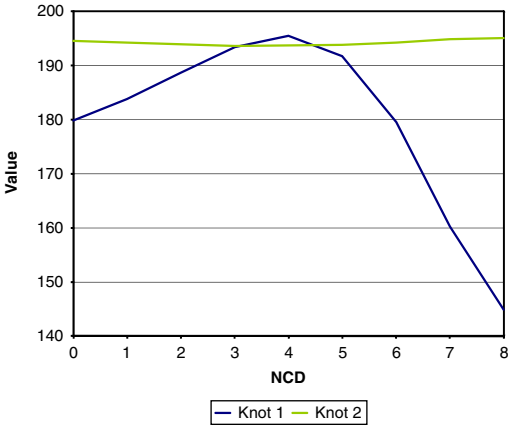

Figure 7. The 'Value' Measure as the Number and Position of Knots are varied by NCD

3.2.12.1 The "Value" measure has therefore proved consistent with existing test results where small changes are being made from one model to the next, lending to its credibility for use as a single global measure by which to compare the results of any two models.

4. Calculation of 'Case Deleted' Estimates

4.1 Formula for Case Deleted Parameters

4.1.1 McCullagh & Nelder (1989) suggest two methods for calculating 'Case Deleted' parameters.

4.1.2 McCullagh & Nelder (1989:396) discusses the idea of Case Deletion in the standard sense, as a means to identify whether to exclude individual outlier points from an analysis. They talk about the impact on the model fit of removing the point. Also that this is slow if the model needs to be refitted, and suggest that a first step approximation is used. For our purposes even if a single

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

iteration was accurate enough a set of 'Case Deleted' Parameters is still required for every data point which as noted in Berry would be impractically slow.

4.1.3 On p406 they quote a result from Atkinson (1987) for the linear case $\hat{\beta}_{(j)}-\hat{\beta}_j = \frac{-(X^T WX)^{-1}x_i(y_i-\mu_i)}{(1-h_i)}$ where the Hat diagonal is defined as $h_i = diag_i\left(W^{1/2}X(X^T WX)^{-1}X^T W^{1/2}\right)$ and suggest a modification for the generalised linear case $\hat{\beta}_{(j)}-\hat{\beta}_j = \frac{-(X^T WX)^{-1}x_i(z_i-\eta_i)}{(1-h_i)}$ where $z_i = g(y_i)$.

4.1.4 For the linear case this can be used to generate the estimate directly $\eta_{(i)} = \eta_i - \left(\frac{h_i}{1-h_i}\right)\frac{(y_i-\mu_i)}{W_i}$ with the $h_i = \sum_{jk} X_{ij}C_{jk}X_{ik}W_i$.

4.2 Formula for Generalised Linear "Case Deleted" Estimates

4.2.1 This same formula was also tested in the generalised linear case of a Log-Poisson model and found to be 99.8% accurate, albeit with a slight bias. Figure 8 shows $-\left(\frac{h_i}{1-h_i}\right)\frac{(y_i-\mu_i)}{W_i}$ on the x-axis and $\left(\eta_{(i)}-\eta_i\right)/\left(-\left(\frac{h_i}{1-h_i}\right)\frac{(y_i-\mu_i)}{W_i}\right)$ on the y-axis where the actual $\eta_{(i)}$ have been calculated with a full model fit per data point.

4.2.2 Likewise the second formula $\hat{\beta}_{(j)}-\hat{\beta}_j = \frac{-(X^T WX)^{-1}x_i(z_i-\eta_i)}{(1-h_i)}$ was also tested and rejected.

4.2.3 The formula from 4.1.4 has also been checked on a Logit-Binomial model, giving the results illustrated in Figure 9.

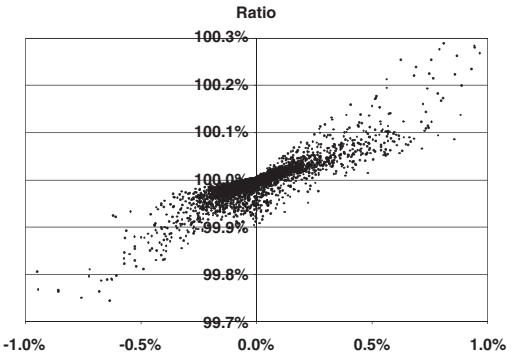

Figure 8. 'Case Deleted' Estimate approximation for Log-Poisson

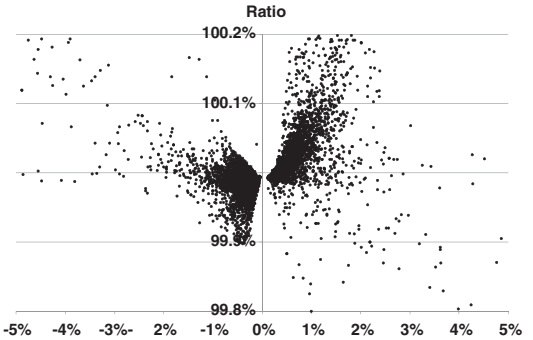

Figure 9. 'Case Deleted' Estimate approximation for Logit-Binomial

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

4.2.4 Armed with this new method we now have the ability to generate $\eta_{(i)}$ directly from a single model fit.

4.3 Bayesian Understanding of the 'Case Deleted' Estimates

4.3.1 The Hat matrix provides the influence of each data point on the parameters. The total of each row adding to one, and hence can be thought of as a credibility in a Bayesian context.

4.3.2 For a linear model the estimate will be formed as follows: $\eta_i = \sum_p h_p y_p$. This can be rearranged as follows $\eta_i = h_i y_i + \sum_{p \neq i} h_p y_p$ then observing that $\eta_{(i)}$ is the equivalent developed from one less data point $\eta_{(i)} = \frac{\sum_{p \neq i} h_p y_p}{\sum_{p \neq i} h_p} = \frac{\sum_{p \neq i} h_p y_p}{(1-h_i)}$, $\eta_i = h_i y_i + (1-h_i)\eta_{(i)}$ so $\eta_{(i)} = \frac{\eta_i - h_i y_i}{1-h_i} = \eta_i - \left(\frac{h_i}{1-h_i}\right)(y_i - \eta_i)$

4.3.3 For the Generalised Linear Model a first order approximation would be $\eta_{(i)} = \eta_i - \left(\frac{h_i}{1-h_i}\right)\frac{\partial \eta_i}{\partial \mu_i}(y_i - \mu_i) = \eta_i - \left(\frac{h_i}{1-h_i}\right) g'(\mu_i)(y_i - \mu_i)$.

For Log-Poisson and Logit-Binomial models it can be shown that $g'(\mu_i)V(\mu_i) = 1$, giving 4.1.4 $\eta_{(i)} = \eta_i - \left(\frac{h_i}{1-h_i}\right)\frac{(y_i - \mu_i)}{W_i}$ for unit weights.

4.3.4 We undertook a numerical checking of a Log-Gamma model as for this structure $W_i g'(\mu_i) = \frac{\omega_i}{\varphi \mu_i}$. For this model we found that $\eta_{(i)} = \eta_i - \left(\frac{h_i}{1-h_i}\right) g'(\mu_i)(y_i - \mu_i)$ and hence reject 4.1.4.

5. Noise Reduced Parameters

5.1 Desire for an Amended Set of Parameters

5.1.1 The realisation that the 'Case Deleted' Estimate $\mu_{(i)}$ is a useful noise independent measure, and easily calculated, led to a couple of initial attempts to use it directly to influence the model output estimates.

5.2 First and Second Attempts, Mean Adjustors

5.2.1 The first thought was that the noise in the model output could be reduced by artificially offsetting each data point to remove an equivalent amount, $y_i^* = y_i + \mu_{(i)} - \mu_i$. These can then be refitted to obtain a new set of estimates, μ_i^*.

5.2.2 The second attempt applied a second tier model to the 'Case Deleted' Estimates from the first $y_i^* = \mu_{(i)}$ to try to produce some new estimates μ_i^* with less noise.

5.2.3 Neither of these produces results which are significantly different from the original estimates. This can be understood by reflecting on the way that GLM models select their parameters by placing them at the 'mean' position of the sub-domain for each parameter. Hence the data has a symmetry about this mean, and the noise $\mu_i - \mu_{(i)}$ reflects this too. So both methods above represent symmetrical adjustments to the data which have little effect on the new estimates.

5.2.4 Consider the example illustrated in figure 10. Here we have a well populated domain with data points on the left defining a value of μ_i shown as the lower green dashed line. Then a new parameter based solely upon two data points y_1,y_2 is considered, this will move the ordinary estimates to the mid-point of the two points shown as μ_i^* the red dashed line. With this parameter

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

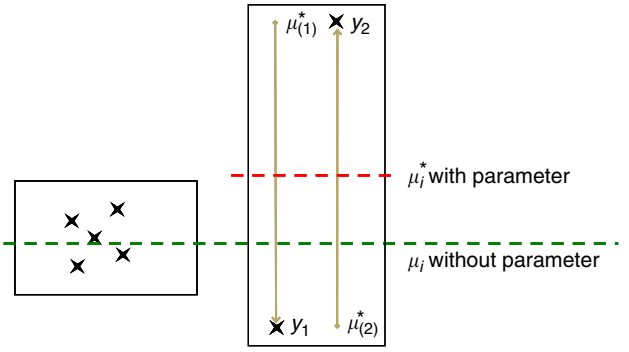

Figure 10. 'Case Deleted' Deviance – Increasing with Model Complexity

included the Case Deleted model for y_1 will produce $\mu^*_{(1)} = y_2$ and similarly the Case Deleted model for y_2 will produce $\mu^*_{(2)} = y_1$.

5.2.5 The Deviances calculated for this parameter will show $SD(y_i, \mu_i) > SD^*\left(y_i, \mu^*_i\right)$ but here the "Case Deleted" Deviance will be substantially worse $CDD^*\left(y_i, \mu^*_{(i)}\right) >> SD(y_i, \mu_i)$. The symmetry of the adjustments can be seen easily here, and hence despite the failure of the extra parameter to add value, we can see why its value remains unchanged.

5.3 The Need for a Variance Penalty Function to Drive the Adjustor

5.3.1 Looking again at the formulation of the 'Case Deleted' Estimates $\mu_{(i)}$, notice that they involve terms representing the mean μ_i and through the Hat diagonal h_i the variance. Instead therefore we need to develop a penalty function to reward the model for good mean values and penalise by increasing variance.

5.3.2 However we cannot simply replace μ_i with $\mu_{(i)}$ in the likelihood and refit, since the extra deviance introduced already possesses the symmetry above, and hence there is little impact on the parameters values by the method.

5.3.3 Now let's focus instead on a more direct penalty function. Take the results of the free fit μ_i with corresponding $\mu_{(i)}$. Now consider that the variance introduced by a parameter, as expressed by the Variance/Covariance matrix, will be scaled if the parameter itself is artificially scaled. Specifically the impact on the covariances will allow the model to rebalance in the presence of correlated parameters.

5.3.4 The Variance/Covariance matrix itself will adjust simply according to the normal result for scaled variances. $Var(\lambda Y_i) = \lambda^2 Var(Y_i)$. In this case the elements of the Variance/Covariance matrix need to be replaced with

$$C^*_{jk} = \begin{cases} Var\left(\lambda_j \beta_j\right) = \lambda_j^2 Var\left(\beta_j\right) & ,j = k \\ Cov\left(\lambda_j \beta_j, \lambda_k \beta_k\right) = \lambda_j \lambda_k Cov\left(\beta_j, \beta_k\right) & ,j \neq k \end{cases} \text{ where } \lambda_j = \frac{\beta^*_j}{\beta_j}$$

5.3.5 From this a scaled version of the Hat diagonal can be calculated. $h^*_i = \sum_{jk} X_{ij} C^*_{jk} X_{ik} W_i =$

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

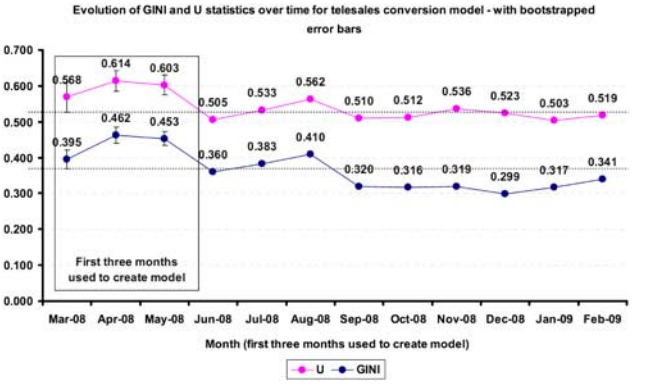

Figure 11. Source: Model Validation Working Party

$\sum_{jk} \frac{X_{ij}\beta_j^* C_{jk} X_{ik} \beta_k^* W_i}{\beta_j \beta_k}$ which produces new Linear Predictors $\eta_{(i)}^* = \eta_i^* - \left(\frac{b_i^*}{1-b_i^*}\right) g'(\mu_i)(y_i - \mu_i)$ and "Case Deleted" Estimates $\mu_{(i)}^* = g^{-1}\left(\eta_{(i)}^*\right)$.

5.4 Idea of a Model Depreciation Index

5.4.1 To draw an analogy, the value of a model is like that of a used car. The instant it rolls off the forecourt it loses a chunk of its predictive power simply by virtue of the fact that it is now being used on new data rather than measured in a circular fashion against the data used to define it.

5.4.2 As time passes, the value decreases further, as was illustrated by the Model Validation working party last year. Figure 11 is an extract from page 10 of their report (Berry *et al.*, 2009).

5.4.3 The Noise Reduction technique provides an indication of that initial depreciation, by reference to the scale factors which have been derived.

5.4.4 Without applying the scale factors, deploying the full model, would result in a worse model than the scaled one.

6. Calculation Of Noise Reduced Model

6.1 Specification of Penalty Function and Two Tier Modelling Process

6.1.1 First obtain the results of the normal Generalised Linear Model fit, as outlined in A.1.10. Next calculate the 'Case Deleted' Linear Predictors $\eta_{(i)} = \eta_i - \left(\frac{h_i}{1-h_i}\right) g'(\mu_i)(y_i - \mu_i)$, and Estimates $\mu_{(i)} = g^{-1}\left(\eta_{(i)}\right)$. Now using the superscript * to denote new parameters and estimates β_j^*, η_i^*, μ_i^* which we will estimate from the new penalty function.

6.1.2 The Hat diagonal h_i is a measure of the influence attaching to the data point y_1 with $(1-h_i)$ the influence of the remaining points. This includes the effect of the Variance of parameter β_j, $Var(\beta_j)$ and the Covariance of this with the other parameters $Cov(\beta_j, \beta_k)$. Now suppose that β_j is scaled back to a value β_j^*, this will reduce the variance to $Var\left(\beta_j^*\right) = \left(\frac{\beta_j^*}{\beta_j}\right)^2 Var\left(\beta_j\right)$ and the Covariances to $C_{jk}^* = Cov\left(\beta_j^*, \beta_k^*\right) = \left(\frac{\beta_j^*}{\beta_j}\frac{\beta_k^*}{\beta_k}\right) Cov\left(\beta_j, \beta_k\right)$. These are not the same as the variance results that would occur from a model which had generated these parameter values directly. Using these values we can scale back the 'Case Deleted' Estimates that would apply to the new parameters.

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

$\eta^*_{(i)} = \eta^*_i - \left(\frac{h^*_i}{1-h^*_i}\right) g'(\mu_i)(y_i - \mu_i)$ where $h^*_i = \sum_{jk} \frac{X_{ij}\beta^*_j C_{jk} X_{ik} \beta^*_k W_i}{\beta_j \beta_k}$. The superscript * quantities are then derived using a similar method to original parameters, described in Appendix A, with non-linear adjustments (Murphy, 2005) to allow for the more complex definition of $\eta_{(i)}$.

7. Worked Example

7.1 Log Poisson Frequency Model

7.1.1 This example is taken from a Motor Third Party Bodily Injury example dataset. This model has a large sample size of 500,000 with around 30,000 responses.

7.1.2 A full complexity model was built upon the data, using 31 factors with 54 parameters, of which 8 were interactions.

7.1.3 Figure 12 shows the relationship between the Standard Error (x-axis) reported by the GLM and the Scale Factor (y-axis) recommended by the Noise Reduction technique.

7.1.4 A few parameters were retained beyond the normal acceptance threshold, to show the fall-off between higher errors and the scale factor.

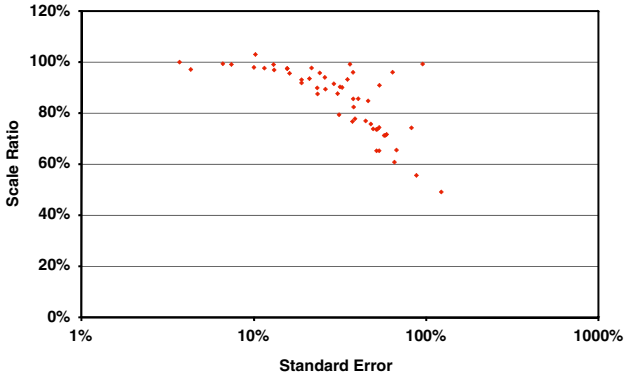

Figure 12. Scatter Chart of GLM Standard Error and Noise Reduced Scale Factor for a Log-Poisson Frequency Model

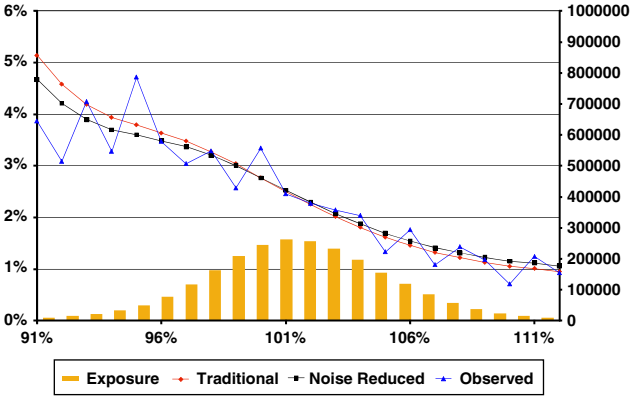

Figure 13. Out of Sample Model Comparison Chart for a Log-Poisson Frequency Model

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

7.1.5 Figure 13 shows the ratio of the two models (x-axis), and the average observed response and model prediction values (y-axis), plus the exposure as bars (2nd y-axis). Models here have been fitted on the training dataset, and then rescored against the hold-out dataset, the chart then measures their value against observed data from the hold-out dataset.

7.1.6 The models show varying predictions with a ratio substantially between +/− 5%. The noise reduced model produces predictions which are scaled towards the mean, which temper the predictions made by the GLM at the extremes of the distribution.

7.1.7 Using a simple business model with a price comparison website level of elasticity fixed at 10 shows a profit margin improvement in this example of 0.57% at constant volumes.

7.2 Log Gamma Severity Model

7.2.1 This example is taken from a Motor Accidental Damage Severity example dataset. To contrast with the previous frequency model, a sample size of 12,000 was used with an average response of 1,450.

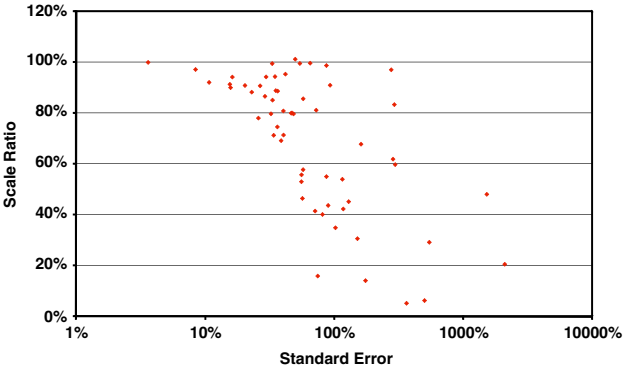

Figure 14. Scatter Chart of GLM Standard Error and Noise Reduced Scale Factor for a Log-Gamma Severity

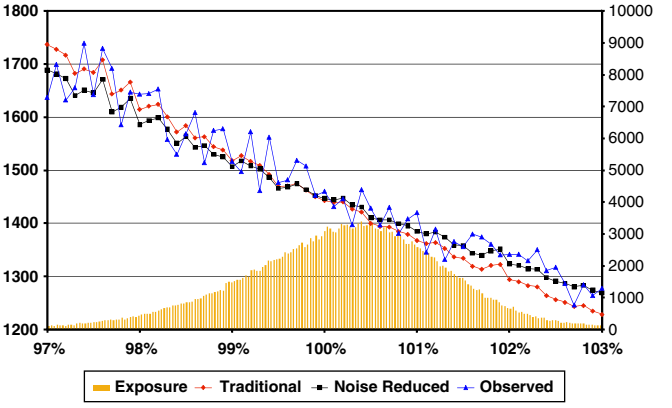

Figure 15. Out of Sample Model Comparison Chart for a Log-Gamma Severity Model

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

7.2.2 A full complexity model was built upon the data, using 18 factors with 59 parameters, of which 17 were interactions (see figures 14 and 15).

7.2.3 Using a simple business model with a price comparison website level of elasticity fixed at 10 shows a profit margin improvement in this example of 0.69% at constant volumes.

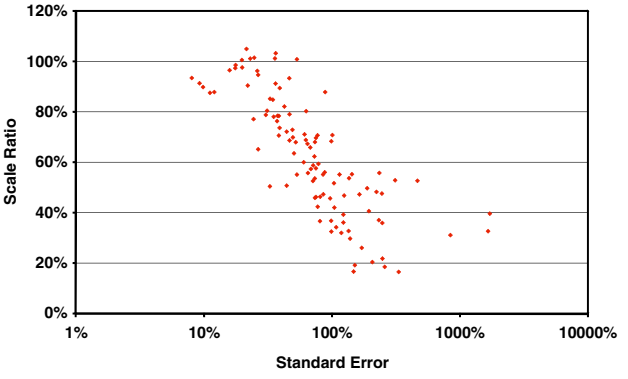

Figure 16. Scatter Chart of GLM Standard Error and Noise Reduced Scale Factor for a Logit-Binomial Propensity Model

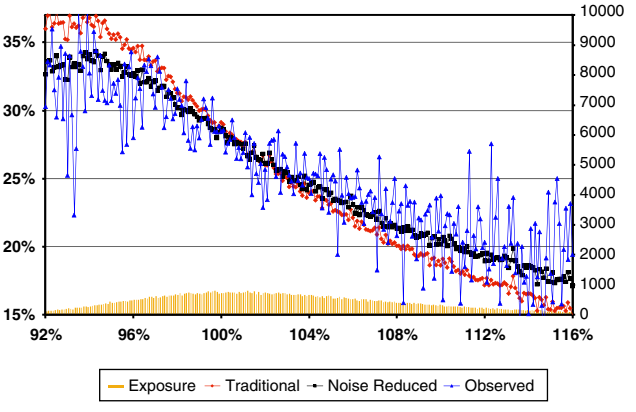

Figure 17. Out of Sample Model Comparison Chart for a Logit-Binomial Propensity Model

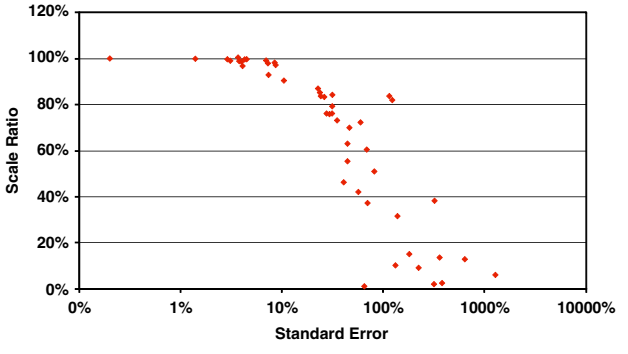

Figure 18. Scatter Chart of GLM Standard Error and Noise Reduced Scale Factor for a Poor Model

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

7.3 Logit Binomial Proportion of Collisions with Bodily Injury Model

7.3.1 This example is a propensity model built on a Motor dataset using collision as the exposure measure, and proportion of Bodily Injuries on the claim as the response. Such an approach is sometimes used to increase the patterns detected in sparse Bodily Injury data. The sample size was 22,000.

7.3.2 The model is using 19 factors with 108 parameters with no interactions (see figures 16 and 17).

7.3.3 Using a simple business model with a price comparison website level of elasticity fixed at 10 shows a large profit margin improvement in this example of 3.4% at constant volumes.

7.4 Poor Model

7.4.1 In this example a particularly poor set of parameters were retained to find out how effectively the technique was at removing ones that are not significant. The chart below shows that scale factors quite close to zero are achieved. The resultant model however was still very poor, as the technique does nothing to add significant factors which are missing from the original model (see figure 18).

8. Conclusions

8.1 Summary of results

8.1.1 Larger datasets and increased competition spurred by price comparison websites have created a natural environment for the development of models at the limit of complexity. In this situation over-fitting is an increasing problem that cannot simply be avoided by the rejection of borderline parameters, as this would risk the penalty of anti-selection.

8.1.2 The Case Deleted Deviance measure discussed in this paper provides an objective method to compare two models that is consistent with the standard error when the two models only differ by a single parameter. It can also be used to compare more complex differences between nested models, and also where the two models are not nested. This is an improvement on the current deviance measure.

8.1.3 The Noise Reduction technique then makes use of this measure to produce scaled back parameters for a model. This is a novel alternative to the standard approach of accepting or rejecting a parameter completely. It also represents a mechanism to select parameters which will be most predictive on a hold-out dataset, rather than simply optimising the parameter values based on the training dataset.

8.1.4 Given the real business risks of anti-selection from using a model which is too simple, another alternative using this technique would be to deliberately over-fit the model including some borderline parameters, and then use this method to scale them back. Further research could be conducted to investigate the relative benefits of this new process compared to the standard method.

8.1.5 This technique clearly will not directly solve the issue of identifying patterns not in the data due to insufficient exposure in the 'corners' of risk segments. Currently this issue is dealt with by underwriting overlays either through additional loadings or acceptance/referral criteria and this process will still need to be applied after the "pure" modelling process.

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press

8.1.6 Beyond the demonstration of this technique to GLM modelling, this concept can be applied to any model forms where parameters are derived from a dataset. Classification techniques for postcode smoothing and vehicle grouping may benefit from this, as could other methods such as decision trees and neural networks.

Acknowledgements

The authors would like to thank the following people for their helpful comments and input during the production of this paper: Mark Sinclair-McGarvie, Duncan Anderson, Catherine Scullion, Sami Abdel-Gadir, Michael Korner, Karl Murphy, Colin Towers, John Berry.

References

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E. & Thandi, N. (2007). *A Practitioner's Guide to Generalized Linear Models (Third Edition)*, CAS Study Note.

Atkinson, A.C. (1987). *Plots, Transformations and Regression*. Oxford University Press, ISBN 978-0-198-53371-9.

Berry, J. (Chair), Hemming, G., Matov, G. & Morris, O. (2009). Report of the Model Validation and Monitoring Personal Lines Pricing Working Party. *Available at:* http://www.actuaries.org.uk/research-and-resources/documents/report-model-validation-and-monitoring-personal-lines-pricing-worki

Dobson, A.J. (2001). *An introduction to generalized linear models 2nd Ed*. Chapman & Hall, ISBN 978-1-58488-165-8.

English, A. (2000–9) EMB Emblem User's Guide, EMB Software Ltd, 2000–9.

Hocking, R.R. (1996). *Methods and applications of Linear Models*. John Wiley & Sons. Inc, ISBN 978-0-471-59282-2.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models 2nd Ed*. Chapman & Hall, ISBN 978-0-41231-760-5.

Murphy, K., Lee, P. & Brockman, M. (2005). Generalized Nonlinear Models: Applications to Auto. COSIS: Predictive Modelling.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (2002). *Numerical Recipes in C++: the art of scientific computing 2nd Ed*. Cambridge University Press, ISBN 978-0-521-75033-4.

Additional reference

Smyth, G.K. & Jørgensen, B. (2002). Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling. *ASTIN Bulletin*, **Vol. 32**, No. 1.

Appendix A

Generalised Linear Models

A.1 *Derivation and Notation*

A.1.1 The following derivation is drawn from Anderson *et al*. (2007) and Dobson (2001) and although well known is included so that the non-linear variant can be derived using the same notation in the main body of the text. Another useful overview can be found in Hocking (1996).

A.1.2 Let Y_i be a series of random variables belonging to the exponential family of distributions, expressed in canonical form with natural parameter θ_i by the pdf. $f(y_i,\theta_i) = \exp\left(\frac{\omega_i}{\varphi}(y_i\theta_i - a(\theta_i)) + b(y_i,\varphi)\right)$ where ω_i is a constant related to Y_i representing the weight which is commonly the exposure for insurance applications, and φ is the scale parameter

A.1.3 Given $\int f(y_i,\theta_i)\,dy_i = 1$ we have $\int \frac{\partial}{\partial\theta_i} f(y_i,\theta_i) = 0 = \int \frac{\omega_i}{\varphi}(y_i - a'(\theta_i)) f(y_i,\theta_i)$ and $\int \frac{\partial^2}{\partial\theta_i^2} f(y_i,\theta_i) = 0 = \int \left[\frac{\omega_i}{\varphi}(-a''(\theta_i)) + \left(\frac{\omega_i}{\varphi}(y_i - a'(\theta_i))\right)^2\right] f(y_i,\theta_i)$.

A.1.4 The first of these gives $E[Y_i] = a'(\theta_i)$ and substituting this into the second gives $a''(\theta_i) = \frac{\omega_i}{\varphi} E\left[(Y_i - E[Y_i])^2\right] = \frac{\omega_i}{\varphi} Var[Y_i]$ we define $\mu_i = E[Y_i] = a'(\theta_i)$ and $V(\mu_i) = a''(\theta_i) = a''\left(a'^{-1}(\mu_i)\right) = \frac{\varphi}{\omega_i} Var[Y_i]$.

A.1.5 Let the log likelihood function be denoted by $l(y_i,\theta_i) = \sum_i \frac{\omega_i}{\varphi}(y_i\theta_i - a(\theta_i)) + b(y_i,\varphi)$

A.1.6 Further define the linear predictor and the link function for the model $\eta_i = g(\mu_i)$, where the linear predictor is a linear combination of the parameters $\eta_i = \sum_j X_{ij}\beta_j$.

A.1.7 First we define the score statistic $U_j = \frac{\partial l}{\partial \beta_j}$ and obtain the result by deriving each of the following terms in order: $U_j = \sum_i \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}\frac{\partial l_i}{\partial \theta_i} = \frac{\omega_i}{\varphi}(y_i - a'(\theta_i)) = \frac{\omega_i}{\varphi}(y_i - \mu_i)$, $\frac{\partial \mu_i}{\partial \theta_i} = a''(\theta_i) = V(\mu_i)$, $\frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i)$ $\frac{\partial \eta_i}{\partial \beta_j} = X_{ij}$. Giving $U_j = \frac{\partial l}{\partial \beta_j} = \sum_i \frac{\omega_i X_{ij}(y_i - \mu_i)}{\varphi g'(\mu_i) V(\mu_i)} = \sum_i W_i g'(\mu_i) X_{ij}(y_i - \mu_i)$. Where $W_i = \frac{\omega_i}{\varphi (g'(\mu_i))^2 V(\mu_i)}$ for reasons that will become clearer below. Note also that $E[U_j] = E\left[\sum_i W_i g'(\mu_i) X_{ij}(y_i - \mu_i)\right] = \sum_i W_i g'(\mu_i) X_{ij}(E[Y_i] - \mu_i) = 0$.

A.1.8 Next, Dobson (2001) derives an approximation by first defining the information matrix $J_{jk} = Cov(U_j, U_k)$, and using $E[U_j] = 0$. $J_{jk} = E\left[(U_j - E[U_j])(U_k - E[U_k])\right] = E[U_j U_k] = \sum_i \left(\frac{\omega_i}{\varphi g'(\mu_i) V(\mu_i)}\right)^2 X_{ij} X_{ik} E\left[(Y_i - \mu_i)^2\right]$. $J_{jk} = \sum_i \left(\frac{\omega_i}{\varphi g'(\mu_i) V(\mu_i)}\right)^2 X_{ij} X_{ik} Var[Y_i] = \sum_i \frac{\omega_i}{\varphi (g'(\mu_i))^2 V(\mu_i)} X_{ij} X_{ik} = \sum_i X_{ij} W_i X_{ik}$.

A.1.9 To solve for the parameters in the general case we use an extension of the Newton Raphson formula (Press, 2002) ${}^{m+1}\beta_j = {}^{m}\beta_j - \sum_k \left({}^{m}U'_{jk}\right)^{-1} {}^{m}U_k$ to find the root of $\sum_j U_j = 0$, $U'_{jk} = \frac{\partial U_j}{\partial \beta_k} = \sum_i \frac{\partial (W_i g'(\mu_i))}{\partial \beta_k} X_{ij}(y_i - \mu_i) + W_i g'(\mu_i) X_{ij}\left(\frac{\partial (y_i - \mu_i)}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k}\right)$.

At the stationary point we are seeking $\sum_i \omega_i X_{ij}(y_i - \mu_i)$ will be close to zero.

For the structures noted in 4.3.3 this will be exactly zero, and $g'(\mu_i) V(\mu_i) = 1$, giving $\frac{\partial (W_i g'(\mu_i))}{\partial \beta_k} = 0$. Hence the first term is normally ignored. $U'_{jk} = \frac{\partial U_j}{\partial \beta_k} = \sum_i -W_i g'(\mu_i) X_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k}\right) = \sum_i -X_{ij} W_i X_{ik} = -J_{jk}$.

https://doi.org/10.1017/S1357321712000190 Published online by Cambridge University Press