Report of the Model Validation and Monitoring in Personal Lines Pricing Working Party

John Berry (Chair)
Gary Hemming
Georgy Matov
Owen Morris

October 2009

The members of the working party would like to thank David Brown, Stephen Jones, David Pandit and Emma Potter for their contributions.

Report of the Model Validation and Monitoring in Personal Lines Pricing Working Party

Contents

1. Introduction

1.1 Motivation

Reserving and capital modelling have received significant attention from GIRO working parties in recent years. Arguably, less formal consideration has been given to pricing issues at an industry level, particularly given the relatively large number of actuaries working within personal lines pricing. The working party felt that this is partly the product of commercial sensitivity surrounding rating methodologies used by competing insurers. The working party decided that the validation and monitoring of technical models is, however, an area where there is a shared interest in research. Importantly, this research can be carried out without the need to discuss areas of particular commercial sensitivity.

Of the two areas of model validation and model monitoring, we felt that model monitoring is an area that is less well covered by the available literature. Two recent trends in the UK personal lines market have added to the importance of monitoring technical models. The first is the widespread implementation of price optimisation, which has led to sophisticated insurers maintaining many more models than would have been the case five or ten years ago. Secondly, the rise of price comparison websites ("aggregators") and associated increase in customer price elasticities has placed renewed emphasis on the accurate estimation of risk premiums. To address the area of model monitoring we have reviewed the literature and where possible tried to bring methods from other fields into an insurance context. Our desire was to illustrate practical methods and techniques using real insurance data in order to help company actuaries.

The working party was keen to consider problems such as:

- How can we detect material differences between modelled and actual results as quickly as possible in order to prevent financial loss?

- Can we use model monitoring techniques to migrate away from calendar based model refresh/rebuild cycles so that we can focus our resources on models which are broken?

- Can we identify processes that are simple to carry out and which produce output that can be understood by the full range of pricing stakeholders?

- Can we create a monitoring process which would improve the risk management of the pricing function?

This report touches briefly on model validation techniques in Appendix A, sharing some interesting techniques and insight gained during the research. However, given the breadth of the topic we took the tough decision not to investigate model validation techniques in as much detail as model monitoring. It was felt that there is already a broad coverage of model validation within existing statistical literature. The working party agreed however that individual companies should seek to define a model validation "recipe" for all technical models. In defining this recipe, companies should consider whether the validation process should be carried out by someone other than the person responsible for building the model.

1.2 Scope

The techniques discussed and examples presented in this report are all based on generalised linear models. We would expect that many of findings would be similar for other types of models but have not tested this.

Much of our work has focussed on what might be called global model performance diagnostics. By this we mean that we have investigated techniques and considered metrics which attempt to describe, at a high level, whether a model is working or not. For reasons of commercial sensitivity we have not produced, or investigated in detail, charts showing experience plotted against individual rating factors. The focus on overall predictive power rather than performance in individual segments has enabled us to concentrate on diagnostics which could feasibly be produced regularly and included within technical model dashboards for provision to stakeholders in the pricing process. The report demonstrates how several diagnostics can be used to monitor technical models. However, it was not our intention to produce a definitive guide to diagnostic tools.

In the examples which follow, we have based our investigations on data which are fully run off. In the conversion data example this is not a major issue. Depending on the precise definition of conversion and the workings of the de-duping process for multiple quotations, there may be a delay of up to a few weeks for conversion data to be fully run off. In the claims modelling context the issue of whether the data is run off or not is more important. This will be discussed in more detail later in the report.

1.3 Data used in the worked examples

The working party would like to thank RBS Insurance Services Ltd for making claims data available to the working party and Zurich Financial Services Ltd for providing data on customer conversion.

We make it clear that the models built in the process of carrying out this research are not used by or endorsed by RBS Insurance Services Ltd or Zurich Financial Services Ltd.

A disguise was applied to some of the charts shown in the example model dashboards at the end of the report.

2. Statistical background for worked examples

If a detailed understanding of the statistics underlying the worked examples is not immediately required, then this section can be read after the section containing the worked examples.

Here we describe four tools for assessing the performance of a predictive model over time. Throughout this section, when explaining the construction and calculation of the statistics, we only refer to conversion rate models, but the techniques are also valid for the assessment of retention, cancellation and cross-sell propensity models.

Ideally, the model performance statistics should be easy to calculate and understand, and their calculation should be capable of being automated. The statistics we describe below were calculated using relatively simple Visual Basic code within EMB Emblem. We are confident that it is also a simple process to calculate these statistics in other common modelling packages.

The statistics are as follows:

(1) The construction of the "cumulative gains curve" – a diagram that summarises the performance or "gain" a model shows relative to a random model (i.e. a model with zero predictive power)

(2) The Gini coefficient – closely related to the area under the gains curve

(3) The construction of the "receiver operating characteristic curve" (ROC curve) – a similar diagram to the cumulative gains curve

(4) The Mann Whitney *U* statistic – closely related to the area under the ROC curve

We describe the construction of each of these statistics below and explain how they are related.

2.1 The Cumulative Gains Curve

Gains curves are popular tools with marketing professionals as they are a direct and visual means to assess the performance of a predictive model. We describe below how a gains curve is created in the context of a conversion rate modelling exercise.

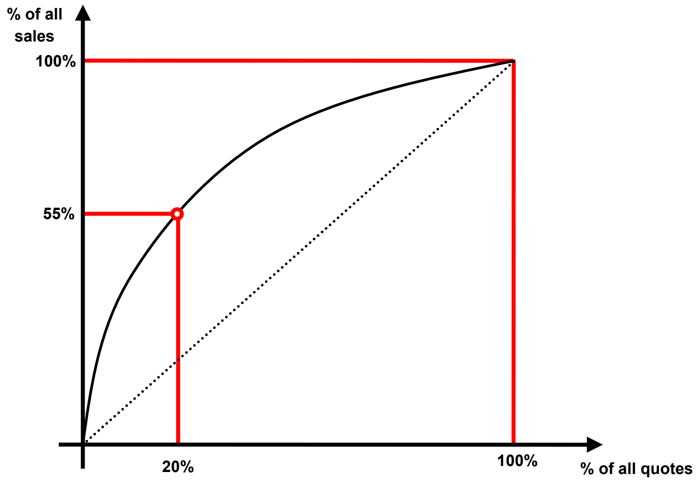
The gains curve is created as follows:

(1) "Score" (calculate the predicted conversion rate using the chosen conversion rate model) each data point, and rank the scores from the highest to the lowest predicted conversion rate (*n* observations in total, *i*=1,…,*n*, with the *n*th observation having the lowest predicted conversion rate)

(2) Loop through the ranked dataset counting the cumulative number of **actual** sales (C_i) for each ranked data point

(3) Plot the percentage of **all** observations on the x-axis against the percentage of actual sales on the y-axis (plot $x = i/n$ versus $y = C_i/C_n$)

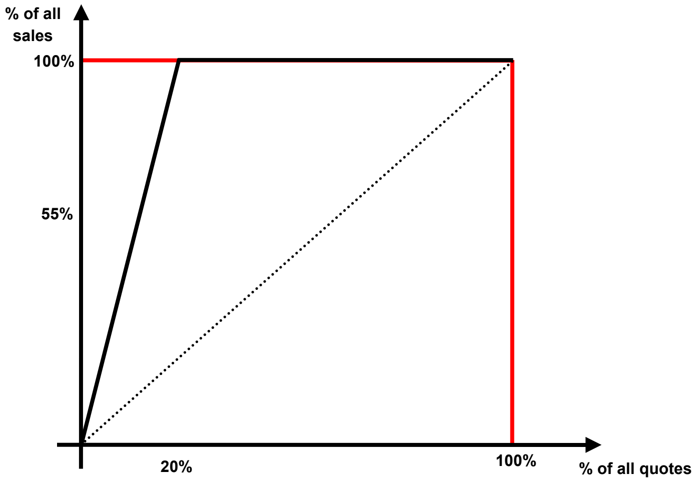
The result of this procedure is a curve that joins the origin to the point (100%, 100%);



In this example, the highlighted point on the curve has the following meaning: the highest 20% of the conversion rate scores accounted for 55% of the actual sales.

The straight dotted line joining the origin to the point (100%, 100%) is often drawn on the gains curve to represent the performance of a model that assigns scores at random – in this case the highest 20% of the conversion rate scores should only account for 20% of the actual sales.

Intuitively, the greater the area between the gains curve itself and the dotted line, the greater the predictive power of the model. In the diagram below we show the gains curve for a "perfect" model. In this case the conversion rate is 20%, and so 100% of sales are contained in the first 20% quotes when ranked by the predicted conversion rate.



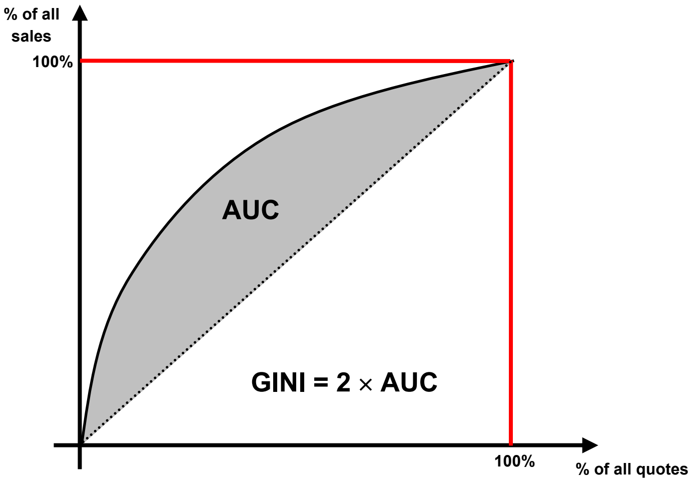
The diagram above shows the "perfect" gains curve in the case of a Binomial dataset. For a frequency model, where the possible responses are non-negative integers (corresponding to the number of claims reported on the policy, generally modelled using a Poisson distribution), the "perfect" gains curve would be represented by a series of connected straight line segments.

Gains curves can be used at the model validation stage to compare alternative models. Often a "hold-out" set of data, not used in the original model construction is used for this purpose.

2.2 The Gini coefficient

The Gini coefficient (or Gini "index") is probably most commonly known as a measure of the inequality of income and is used by economists to help measure the distribution of a society's wealth. However, it is also a useful statistic to help measure the performance of a predictive model.

The simplest definition of the Gini coefficient is in terms of the area under the gains curve ("AUC") – it is twice the area between the gains curve and the line joining the origin to the point (100%, 100%).



We have seen slightly different definitions of the Gini coefficient, but they are all variations on the definition above.

In most sensible situations the Gini coefficient (G) is a number greater than zero and less than one. In fact, the Gini coefficient, as defined here, cannot exceed one minus the actual conversion rate (C) for the dataset. This is because, for the "perfect" model, the area under the curve is a triangle with a base of (1-C) and a height of 1. Hence, in this situation, the area under the triangle is (1-C)/2 and the Gini coefficient is 1-C.

The Gini coefficient, as described in this paper, is calculated from a sample and so is a sample statistic. Hence we should try to estimate its sampling distribution, or at least its standard error. Unfortunately this is not a simple matter analytically, and so we have resorted to bootstrap techniques to estimate the error around the point estimate of this statistic. This is discussed in section 2.5.

2.3 The Receiver Operating Characteristic Curve (ROC curve)

ROC curves were first used to help analyse the performance of radio receiving equipment, where radio users could alter the operation of the equipment to trade-off *sensitivity* (the correct identification of a positive signal) for *specificity* (the correct identification of a negative signal). In this way, the sensitivity and specificity are related to the concepts of Type I and Type II errors used in statistical analysis.

In the context of a conversion rate modelling exercise, the ROC curve is derived as follows:

(1) "Score" (calculate the predicted conversion rate using the chosen conversion rate model) each data point

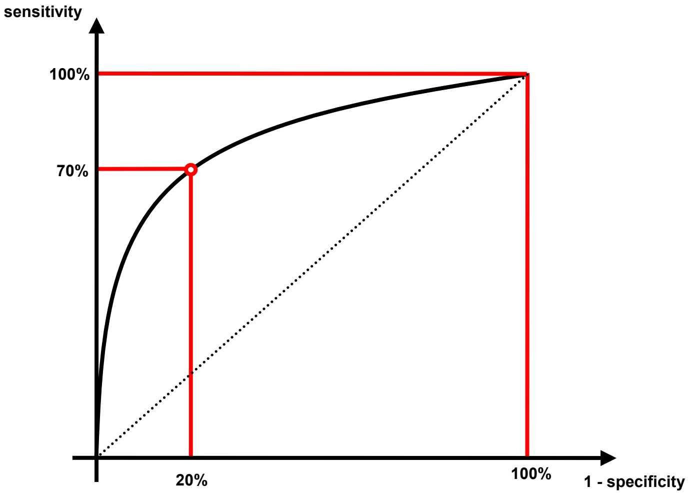
(2) Separate the data into two groups – the **actual** sold and unsold quotations.

(3) Calculate the model *sensitivity* as the percentage of actual sales where the predicted value is greater than the value of an arbitrary "cut-off", c ($0 \le c \le 1$) (c is also called the "discrimination threshold")

(4) Calculate the model *specificity* as the percentage of the actual unsold quotations where the predicted value is less than the value c

(5) Plot $x = 1 - \textit{specificity}$ against $y = \textit{sensitivity}$ for all values of c

Like the cumulative gains chart, the result of this procedure is a curve that joins the origin to the point (100%, 100%);



In this example, suppose that the value of c corresponding to the highlighted point shown is 0.3. In this case the interpretation of the highlighted point is as follows:

(1) "70% of cases where the predicted conversion rate was **greater** than 0.3 were, in fact, actual sales".

(2) "20% of cases where the predicted conversion rate was **less** than 0.3 were, in fact, actual sales".

At first the ROC curve can appear less intuitive than the cumulative gains curve. However, it is very useful when assessing which customers should be prioritised for an outbound telesales operation or in situations where a cut-off must be made, for example when setting a minimum credit score for accepting a consumer loan application.

2.4 The Mann-Whitney *U* statistic

The Mann-Whitney *U* statistic (also called the Mann-Whitney-Wilcoxon statistic) is a non-parametric statistic used to assess whether two independent samples of observations come from the same distribution. It is well covered in many statistics textbooks so we do not discuss it in detail in this report.

Here it is used to compare the predicted conversion rates for the **sold** and **unsold** quotations.

The value of the statistic is calculated as follows:

(1) "Score" each data point (regardless of whether the quote is converted or not), and rank the scores from the highest to the lowest predicted conversion rate

(2) Sum the ranks for the **sold** policies only (R_1)

(3) The value of the *U* statistic is given by:

$$U = \frac{1}{n_1 n_2} \times \left(R_1 - \frac{n_1(n_1+1)}{2} \right)$$

where

n_1 = the number of sales

n_2 = the number of unsold quotations

R_1 = the sum of the ranks for the sold policies only

Often, the value of *U* is defined without the $1/n_1n_2$ term, but we have included it here since it "normalises" the value of *U* so that its maximum value is 1.

The *U* statistic has the following properties

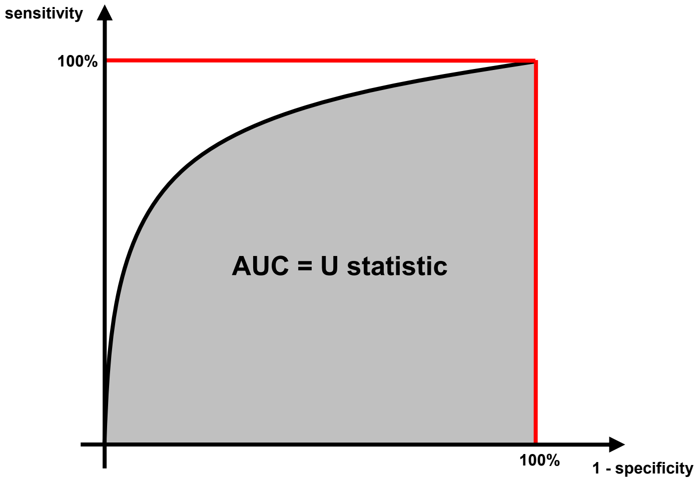
(1) *U* is equal to the probability that a randomly chosen **sold** quotation will have a higher predicted value than a randomly chosen **unsold** quotation.

(2) *U* is equal to the area underneath the ROC curve (see the diagram below).

(3) *U* is equal to related to the Gini coefficient as follows:

$$2U - 1 = \frac{G}{1-C}$$

where *C* = the actual conversion rate for the dataset



In the same way that the Gini coefficient is related to the area under a cumulative gains chart, the Mann-Whitney *U* statistic is related to the area under the ROC curve, although the *U* statistic is equal to the total area under the curve, not just the segment between the curve and the line joining the origin to the point (100%, 100%).

For a “random” model, the ROC curve is a straight line joining the origin to the point (100%, 100%), as for the cumulative gains curve. In this case the value of *U* would be ½.

For this reason, we have slightly modified the definition of the *U* statistic as follows:

$$U^{'} = 2U - 1$$

This simple transformation means that the value of *U* for a “random” model is zero, while for the “perfect” model, the value of *U* is 1. In addition, the relationship between U’ and G is given by

$$U^{'} = \frac{G}{1-C}$$

2.5 Using bootstrapping to estimate confidence intervals

Ideally, we should try to estimate the uncertainty surrounding our estimates of the Gini coefficient and the *U* statistic, so that we can assess whether deterioration over time in the value of either statistic is statistically significant. Unfortunately, we have found no easy method to calculate the confidence interval and so we have used bootstrapping techniques to provide the estimates.

In the context of a logistic model, the non-parametric bootstrap procedure is applied as follows:

(1) Resample the original dataset used to construct the model, *with replacement*, to construct a new dataset with exactly the same number of observations.

(2) Refit the statistical model on the new dataset to obtain new estimates of the GLM parameters. This step does not involve a review of parameter selection and smoothing, and the structure of the model is unchanged – rather it is only the parameter values that are re-estimated. Hence this step requires no human input and can be automated.

(3) Calculate the value of the Gini and *U* statistics for the resampled data and refitted model

(4) Repeat steps 1 to 3 many times (ideally at least 1,000 times), and save the values of the Gini and *U* statistics for each repetition.

(5) Calculate the 5th and 95th percentiles of the bootstrapped estimates of the Gini and *U* statistics to provide an estimate of the statistics’ confidence intervals.

The bootstrap procedure is relatively easy to understand and implement, but it is computationally intensive. For the aggregator dataset referred to in this section it took around 12 hours to run the bootstrap procedure 1,000 times, on a typical modern desktop computer. However, the procedure requires little human input and can be left to run overnight, so the estimation of the confidence intervals does not pose any significant difficulties.

In practice, we believe that the confidence intervals calculated by this method should only be used as a guide to assess model performance over time. The confidence intervals should not used as an inflexible decision criterion upon which to trigger a model refitting exercise, but rather they should be considered in the context of several alternative assessments of model performance.

3. Worked examples

In this section we present some results based on the analysis of a UK direct telesales motor conversion dataset and a UK aggregator motor conversion dataset. These datasets included all of the relevant rating factor information along with estimates of the competitive position for each quote.

We tried to replicate a fairly typical situation that might be encountered in an insurance company, where the conversion rate model, with the price elasticity component, was built based on a sufficiently large number of historical observations to provide a robust model. This model would then be employed in subsequent months for the following activities:

(1) To monitor the actual versus the expected conversion rate,

(2) As an input to a price optimisation process, or

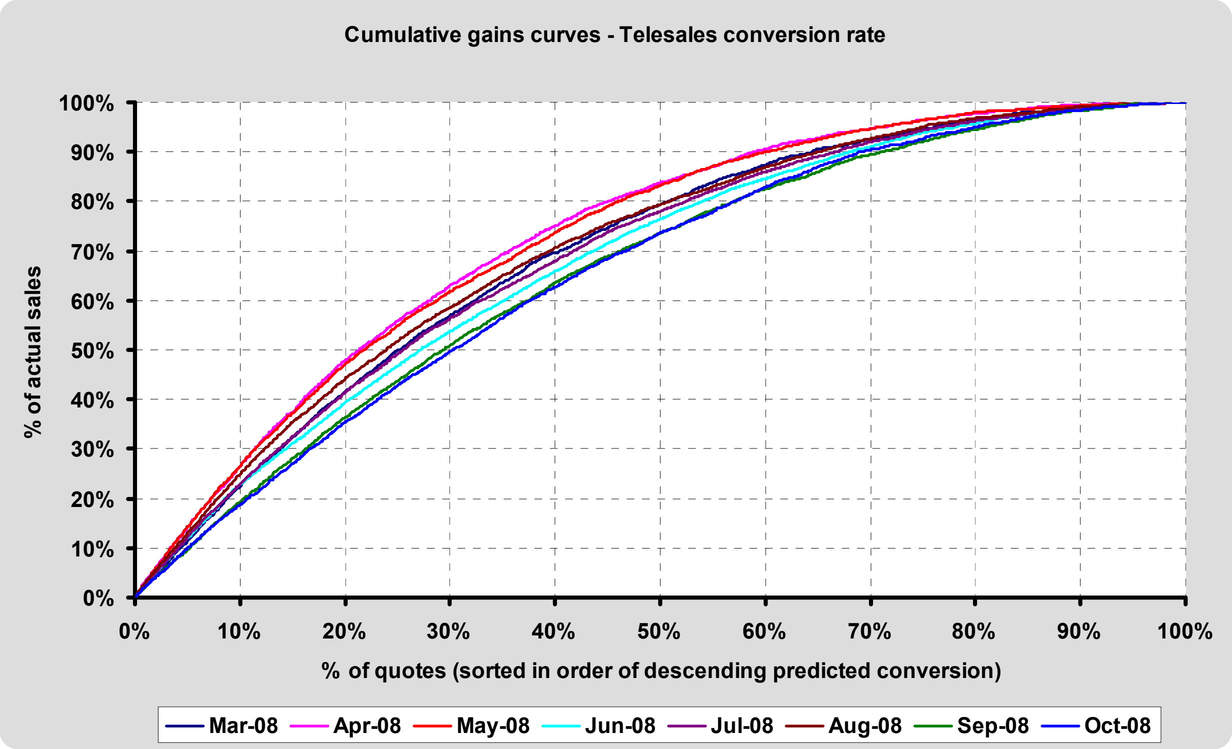
(3) As an input to a marketing promotion analysis (for example, to help identify which customers are most likely to respond to a particular direct marketing campaign and then go on to make a purchase)

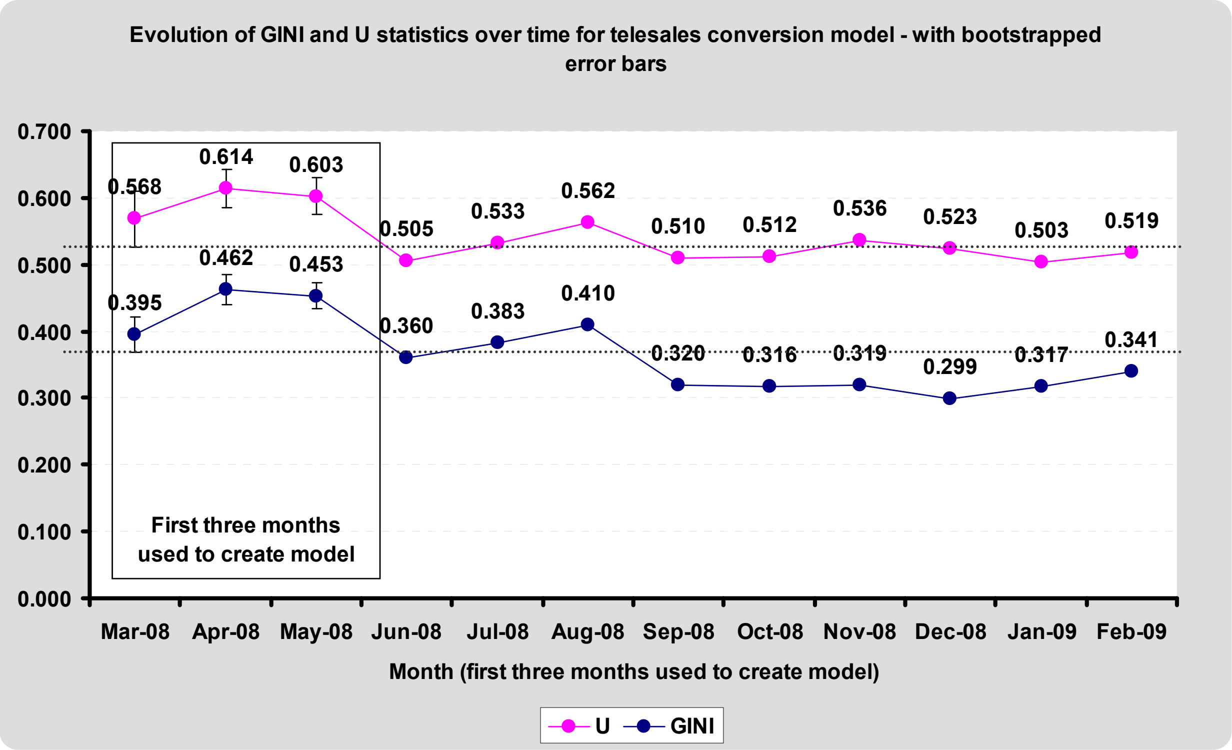
We have deliberately not focussed on the price elasticity component of the models in order to avoid straying into an area of commercial sensitivity. In practice, we expect that companies will track price elasticities on a segmental basis over time as part of the model monitoring process.

3.1 Direct motor telesales conversion – worked example

Quotation data from March 2008 to May 2008 were used to construct a robust model of telesales conversion, complete with relevant interactions and factor smoothing. The dataset contained around 35,000 observations with a typical motor telesales conversion rate in the range of 10% to 40% depending on the segment – we have not been more specific here in order to respect the confidentiality of the insurer that supplied the data.

For the three months used in the construction of the model, and also for each subsequent month in 2008 we evaluated the Gains curve, Gini coefficient and *U* statistic. The results are shown below.





For clarity we have only shown the months from March 2008 through to October 2008 in the cumulative gains chart.

The highest two curves on the gains curve chart correspond to the months of April 2008 and May 2008, which were two of the months used to create the original conversion rate model. It is unsurprising that the months used to create the original model should show the highest gains curves. These two months also have the highest values of the Gini coefficient and *U* statistic over the period shown, which is to be expected given the close relationship between these statistics and the cumulative gains curve.

Interestingly, the first month in the dataset (March 2008) has lower values of the Gini and *U* statistics, and correspondingly the gains curve for this month is rather lower than for April and May. The March quote profile was significantly different to the April and May quote profiles in that it contained a significantly different mix of quotes by marketing media type and we believe that this drove the lower value of the Gini and *U*.

We have used the bootstrap procedure described in section 2.5 to derive the 5th and 95th percentiles of the estimated Gini and *U* statistics for the months used to create the conversion rate model and we have included these error bars on the second chart. As a rough guide to the level of model deterioration over time we have drawn a horizontal line across the chart coincident with the lowest level of the error bars.

Unfortunately, the very first month after the modelled period, June 2009 shows both the Gini and the *U* statistics dropping below the horizontal line. However, an issue was identified with the rating of quotations in June meaning that the experience for this month was not typical and therefore it is unsurprising that the model performance deteriorated. This issue was corrected in early July, and the Gini and *U* statistics rise back above the horizontal line for July and August.

Both the Gini and the *U* statistic fall in September, with the Gini coefficient showing a particularly large decrease. Both statistics also fall below the value of the lower error bar for March. In an automated model evaluation environment this could provide a trigger that the model performance has deteriorated significantly and that a refitting or remodelling exercise should be considered. Interestingly, March and September are months where significant volumes of new vehicles are registered in the UK. A more likely reason for the observed behaviour is however that a significant change in the marketing activity occurred for this account in September, and therefore it is not surprising that the model performance deteriorated. The change in marketing activity remained in place for the remainder of the model period, and the Gini and *U* statistics remained relatively level from September 2008 through to February 2009.

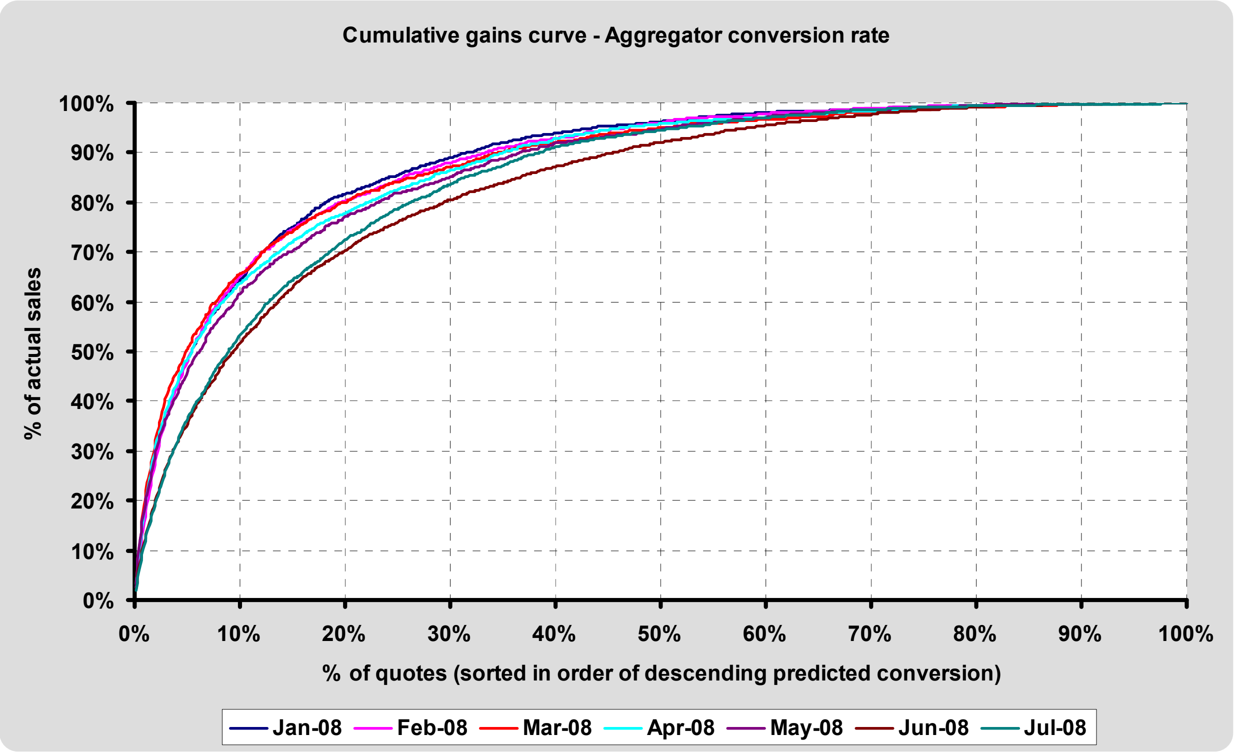
There is an explanation for the larger fall in the value of the Gini coefficient in September than the fall in the value of the *U* statistic. As a result of the change in marketing activity, the number of quotations fell in September, but the conversion rate rose significantly. As mentioned in section 2.2, the maximum possible value of the Gini coefficient, as defined here, is equal to 1 minus the observed conversion rate. Hence it is reasonable to conclude that when the conversion rate increases (in this case due to a change in the mix of quotations) that the observed value of the Gini coefficient is likely to decrease.

This observation highlights a potential problem with the Gini coefficient when used as a tool to identify deterioration in model performance. It is quite possible that when the underlying model has not deteriorated in any way, the Gini coefficient can significantly decrease due to an increase in the observed conversion rate, caused merely by a shift in the quotation profile. However, in practice, large movements in conversion rate for a well established channel are unusual and since the actual versus predicted conversion rate is almost certain to be part of any diagnostic suite, this is not a significant problem. However, we believe that the *U* statistic does not suffer from this problem and so is probably more suitable for use in an automated process for assessing model deterioration over time.

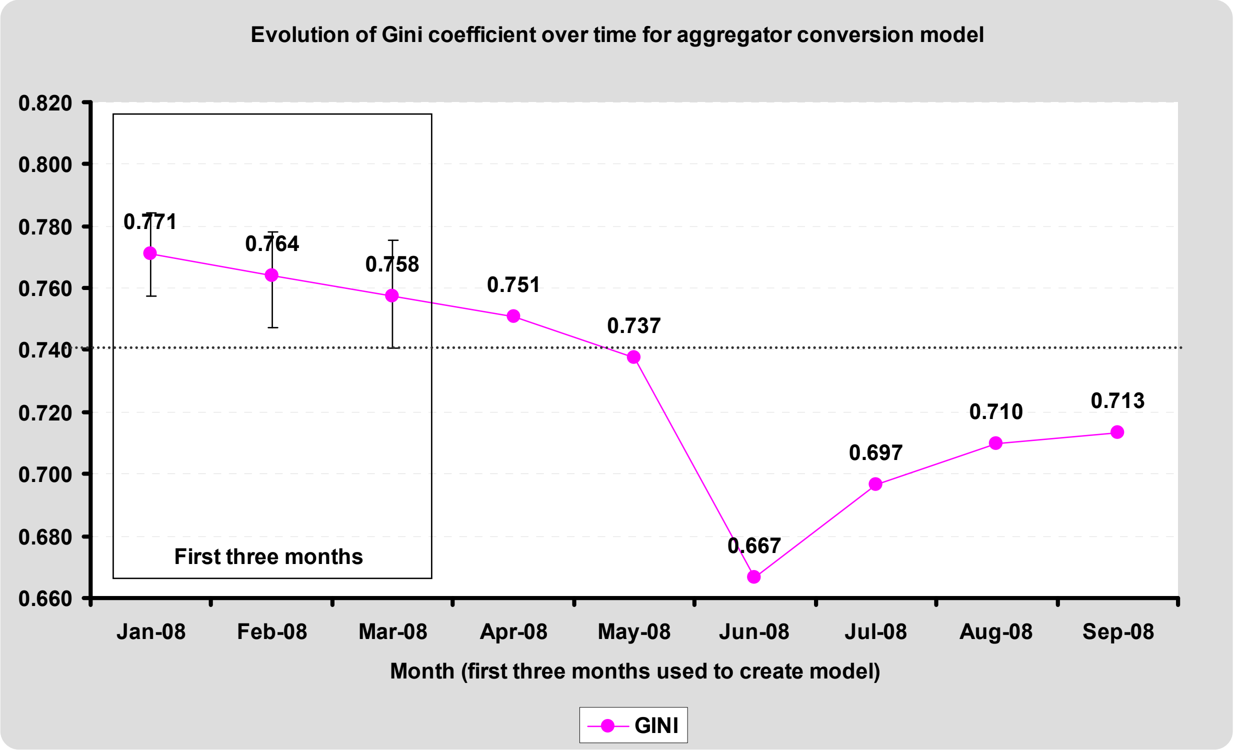
3.2 Direct motor aggregator conversion – worked example

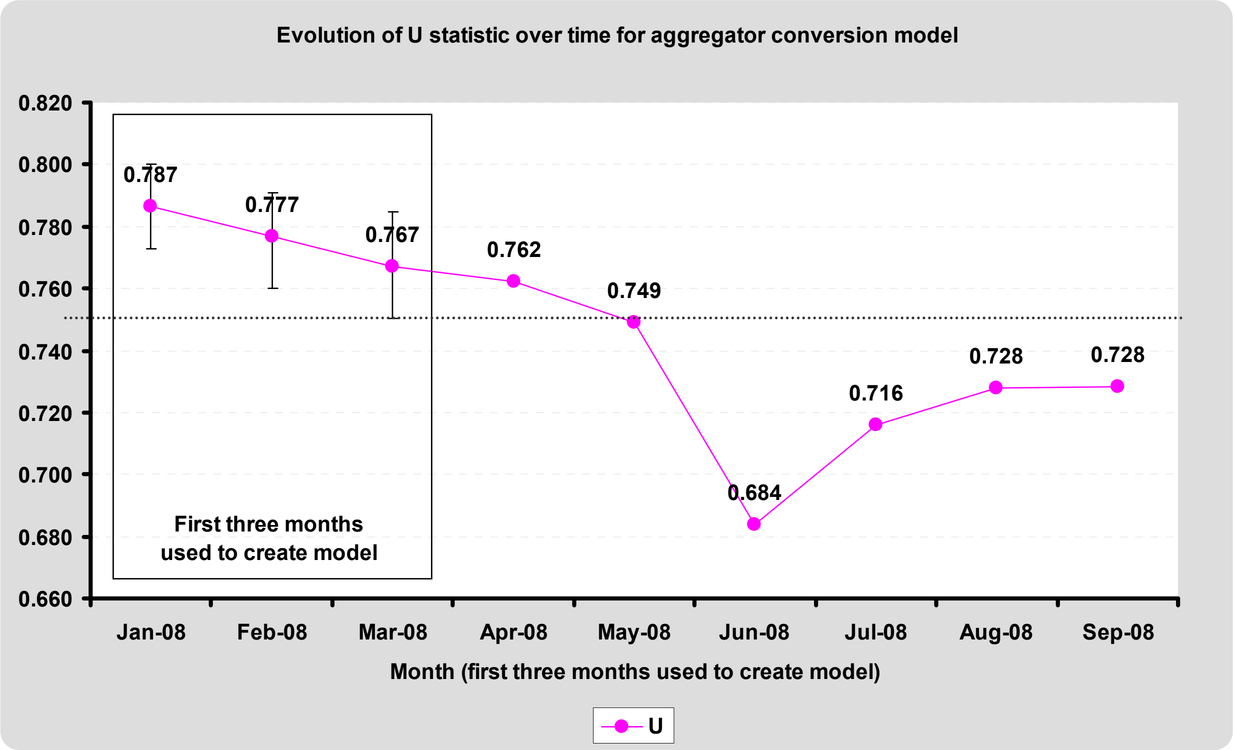
Quotation data from January to March 2008 inclusive were used to construct a robust model of aggregator conversion. The dataset contained around 340,000 observations and had the typically low conversion rates expected for UK aggregator sourced business.

For the three months used in the construction of the model, and also for each subsequent month in 2008 we evaluated the Gains curve, Gini coefficient and *U* statistic. The results are shown below.



For clarity we have only shown the months from January 2008 through to July 2008 in the cumulative gains chart.





We first note that the gains curves and the larger values of the Gini and *U* statistics indicate that the aggregator conversion model is more effective at identifying quotations with relatively high conversion rates than the telesales model. For the aggregator dataset, the competitive position was a very powerful predictive factor, while in the telesales dataset it was only weakly predictive of conversion rate. We believe that this is the primary factor driving the superior performance of the aggregator conversion model, based on the assessments of the Gini and *U* statistics.

We also note that the values of the Gini coefficient and *U* statistic are very similar for this dataset. This is sensible since the two statistics only differ by the factor 1-*C*, where *C* is the observed conversion rate, and for the aggregator channel conversion rates tend to be very low (typically no more than a few percent).

The model performance does appear to deteriorate slightly in the first two months (April and May) following the modelled period, but clearly there is a significant deterioration in performance in the month of June. The subsequent months recover slightly but remain well below the horizontal line drawn to show the lowest error bar for the three modelled months. The cumulative gains curves confirm that the model is significantly less effective at identifying policies with the highest conversion rates in June and July.

A closer examination of the actual versus predicted conversion rates by factor for June revealed several factors where the original model was no longer adequately explaining the observed variation in conversion rate. In section 3.3 we have analysed how the model performance improves following a refitting and remodelling exercise.

We have no clear-cut explanation for the sudden deterioration in performance and why the June performance in particular was so poor – however, we believe that a major insurance group placed several of its brands onto some of the price comparison sites at around this time, triggering a round of rate reviews across the market.

An insurer employing some of the diagnostic tools we have presented in this paper might reasonably expect to refit its conversion rate model following the poor June performance. In the next section we present an analysis of such an exercise.

3.3 Results obtained by refitting the aggregator model using more recent data

We choose to examine the June deterioration further by conducting a remodelling exercise, using data from June 2008 to August 2008. We chose two approaches:

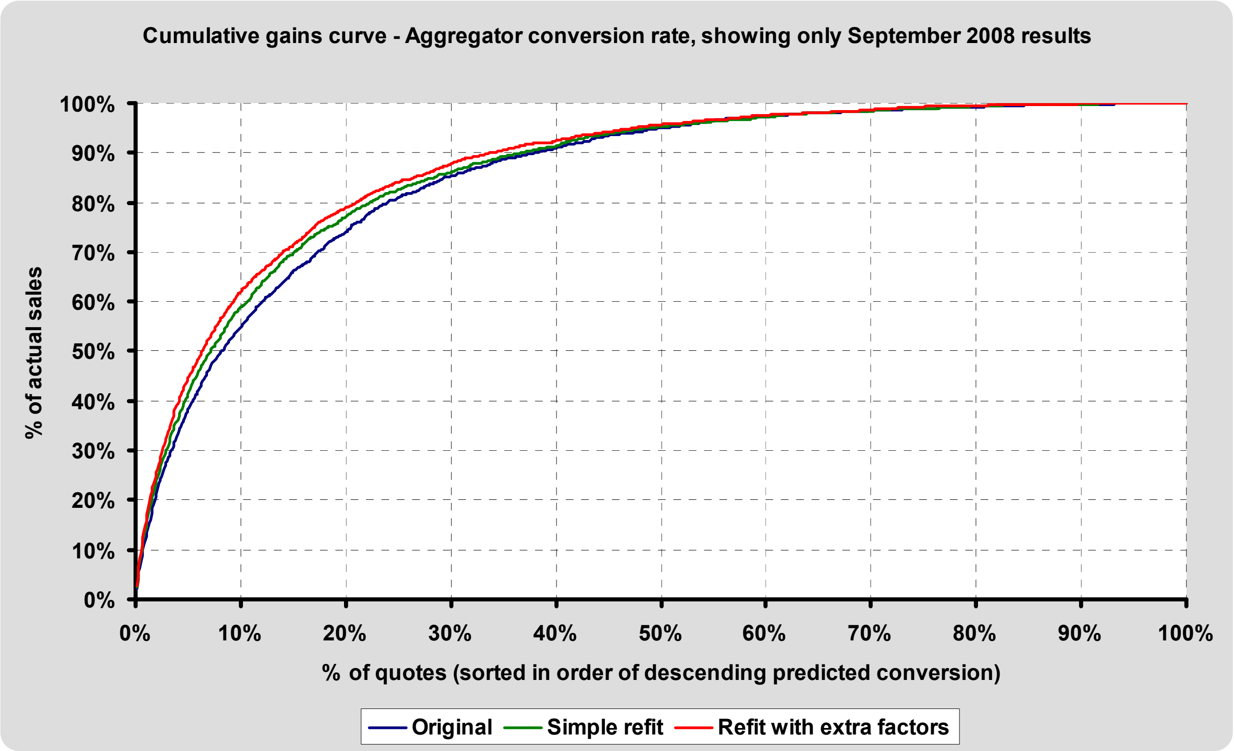
(1) A simple parameter refit, with no change to the existing model factors, interactions or smoothing, and;

(2) A more thorough review of all factors, including those not originally present in the model, to assess whether existing factors were still statistically significant and whether other factors should be included in the model

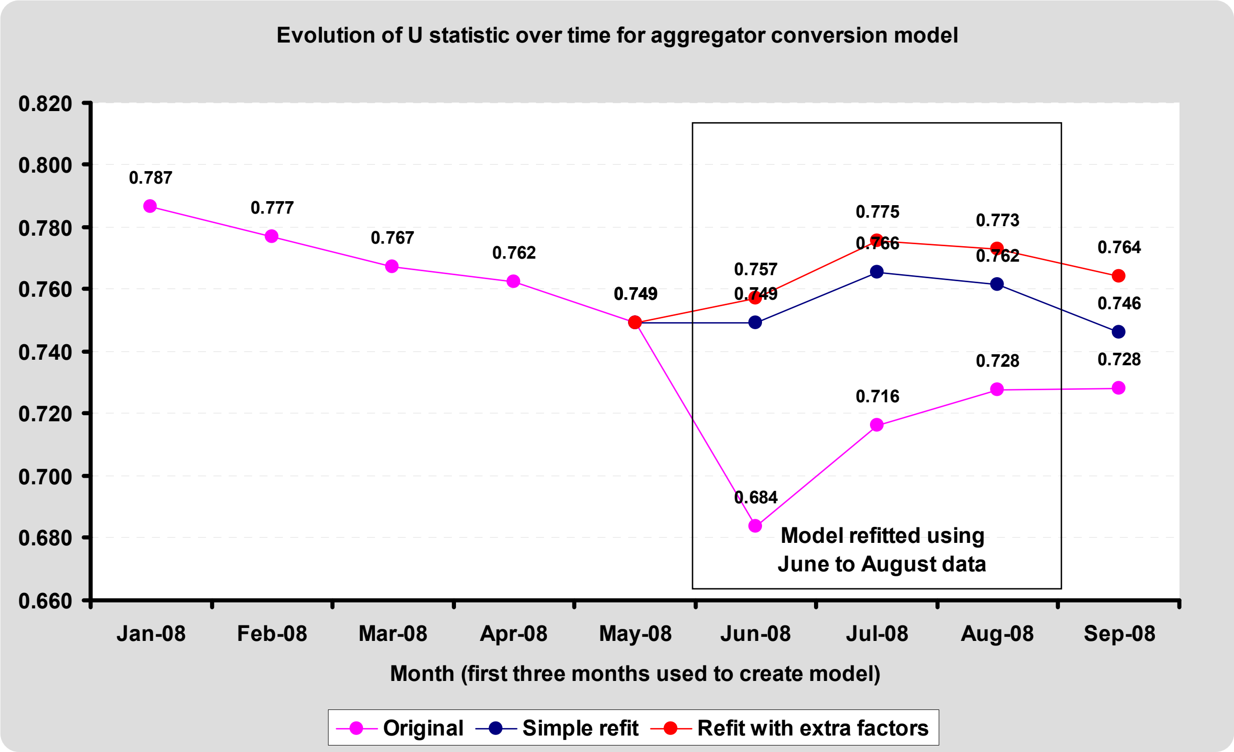
We created the conversion rate model using June 2008 to August 2008 quotation data and tested the results on the final month in the dataset, September 2008.

The simple parameter refit improved the model performance considerably. However, the more thorough review clearly indicated that two additional factors should be included in the model based on the June to August data. These factors were not judged to be significant in the January to March data.

The results of the remodelling exercise are shown below. The chart shows the cumulative gains curves for the month of September 2008, for the original model, the simple parameter refit model and finally the thoroughly reviewed model.



The following chart shows the values of the *U* statistic over time for the three models.



The cumulative gains curves demonstrate the improved model performance. To further illustrate the improvement in September, some key statistics from these curves are reproduced below:

	% of sales in highest 5% of model predictions	% of sales in highest 10% of model predictions	% of sales in highest 25% of model predictions
Original model	38.2%	54.8%	81.0%
Simple refit model	41.3%	58.9%	82.5%
Refitted model with extra factors	44.5%	62.1%	84.0%

In this example, the refitted model with extra factors delivers significantly better performance than the original model, or the simple refitted model for the first non-modelled month (September 2008). During the course of our analysis as part of this working party, we have concluded that it is rarely appropriate to simply refit a previous model, based on more up to date information, and to use this model unchecked in a business application. The human review of the model is still essential since it is very likely that additional factors may prove to be significant, or that previous factor smoothing is no longer appropriate. However, the time spent on this review need not be onerous since the majority of the components of the original model will still prove to be valid.

3.4 Final comments on the assessment of demand model performance

Only two datasets have been analysed in preparing the results for this section, and so we cannot put forward definitive recommendations for the frequency of model assessment in a typical insurer. However, depending on the application, it seems reasonable that modellers should consider remodelling conversion rate datasets at least once every three months to avoid models becoming out of date. This is particularly true in the case of aggregator conversion rate modelling, where we found that model performance can deteriorate rapidly over a very short period of time.

We originally expected a telesales conversion rate model to have a longer “lifetime” than an aggregator conversion rate model. This is possibly true for a telesales account with a stable marketing profile, but we found that for this dataset the model is degraded by changes in the marketing, and so modellers should also consider remodelling telesales datasets at least once every three months.

We have not analysed a retention dataset, but we believe a retention model should have a longer lifetime than a conversion rate model – generally, over time, retention rates are more stable than conversion rates, and retention rates are less exposed to rapid changes in the competitive position in the marketplace. However, we haven’t analysed any suitable data and so can’t propose a typical remodelling frequency for retention data.

In the following sections we move on to look at the assessment of claims model performance over time. Claims analysis differs considerably from the demand analysis considered up to this point and introduces additional complexity, particularly in terms of claims run off, seasonality in claims experience and the choice of accident or underwriting analysis periods. These complexities are discussed in the following sections.

We found that the Gini and *U* statistics described above were not suitable for monitoring claims models, and an alternative statistic (based on the deviance of the models) is used to assess the performance of the claims models over time.

3.5 Additional considerations when monitoring claims models

The following sections assume that claims modelling uses GLMs (see appendix for references to relevant literature), typically with frequency and severity GLMs for each claim type. It is further assumed that a log link function is used to impose a multiplicative structure.

There is additional complexity in monitoring models of claims experience. This complexity arises from:

(1) the time it takes for claims to develop sufficiently for modelling, which is typically much longer than the time taken for conversion or retention data to develop

(2) the fact that different types of monitoring analyses require models to be based on either an accident period basis or an underwriting period basis

(3) the fact that seasonality has a large effect on claims models

(4) the need to project trends in claim frequency and inflation rates for claim severities

The first point means that there is a delay before model performance can be assessed for a particular exposure period. This delay will vary according to type of claim. The impact of this is that for some claim types it may take a long time to assess whether a model is broken or not. A method for assessing the predictive performance of a pricing model on undeveloped data is discussed in section 3.9.

The second point, concerning analyses performed on accident year vs. underwriting year, is important because different types of analysis require different approaches. We expect that regular model monitoring will normally take place on an accident period basis, perhaps quarterly. If the model fails to accurately predict the claims experience of a recent accident period then the solution in some cases will be to update the model using the most recent data.

Monitoring the model on an underwriting period basis can be important as a means of checking that the model has proved robust to events such as the following:

a) a change in cover or rating (such as the implementation of a revised risk premium model or a new rating factor)

b) a change in marketing or distribution strategy

c) a change in the marketplace, perhaps caused by competitor actions

The benefit of monitoring on an underwriting period basis in these cases is to provide warnings in advance of results being fully earned. If, following such an event, the model fails to accurately predict the claims experience for business written during or after the event then a warning can be issued to management. If the event was the introduction of a new rating factor then potentially the warning may be to suggest that the new rating basis may be flawed. The use of an underwriting period basis means that discrepancies can be spotted earlier than would be the case under an accident period approach.

The worked examples in this report are based on an accident period analysis. We felt that analyses based on accident periods should form the basis of regular model review processes within companies and that analyses based on underwriting periods would normally be used for ad-hoc reviews following particular events.

The third point means that any attempt to monitor model performance (over any time period less than a full year) needs to make allowance for the sometimes complex effect of seasonality. Pricing models for annual policies typically do not include allowance for seasonality. As such, in order to monitor models more frequently than annually, it is essential to model seasonality by means of a variable describing the time of year. This variable will be included in the model as a main effect and will also have interactions with variables whose effects vary by time of year. For most medium and large UK motor and home accounts, it is possible to capture most of the seasonality through the use of an accident quarter variable and its interactions. This introduces the requirement of having policy episodes split by accident quarter. This quadruples the number of records in the underlying data and in the frequency modelling dataset. However, we found that even the most complex of models still fitted quickly.

There is however an additional concern with converting policy episodes to be quarterly rather than annual. The concern is that the use of quarterly policy episodes might violate the assumption made by the variance function of the GLM. Specifically, the GLM assumes that the variance is a function of the mean. If, for example, the variance function does not correctly describe the different variances in Q4/Q1 and in Q2/Q3, then the model could be invalid unless a varying scale parameter is introduced. The working party noted this issue but also commented that there are tests that can be used to identify whether this problem exists or not. We felt that the most likely approach to regular model monitoring in the short term is to introduce a second model. The first model is the model used in business as usual pricing. It is built on annual policy episodes and has no requirement to capture seasonality. The second model is a copy of the first model except for the fact that the second model has been fitted to quarterly episodes and explicitly models seasonality. Aside from model monitoring, there are other advantages to having the second model. One benefit is that the second model can be used for accurate forecasting of quarterly earned claims experience, allowing for complex effects of seasonality. Another benefit is that understanding complex aspects of seasonality, particularly for motor business, adds much insight to the pricing process.

The final point about projecting the base rate raises an important issue. It is helpful to think of each generalised linear model as having two components. The first component is the *structure* of the model, meaning the choice of variables and interactions included within the model and all parameter estimates except for the overall mean parameter (or base rate). The second component is the mechanism by which the *base rate is projected* to allow for trends in frequency and claim severity.

The key point to note is that if the model has failed, then it isn’t always obvious which of the two components is to blame. It would be a mistake to jump to the conclusion that the structure of the pricing model is wrong simply because the model has underestimated the cost of claims in a recent period. The real problem could be that the estimate for claims inflation was too small. The worked example which follows discusses the methods for assessing whether the structure of the model is broken. We decided that an investigation of the estimation and projection of inflation and trends in frequency was out of scope. If, following investigation, it turns out that the model structure is still reliable then we can look elsewhere for the source of the problem. For companies which have separate statistical modelling and street pricing teams the results of the analysis will help to identify which team “owns” the problem.

It’s worth issuing a warning here. If a company is using a pricing model which is more simplistic than those of its competitors then it is possible for anti-selection to occur in a manner which doesn’t manifest itself as a failure of the structure of the model. For example, suppose that every company in the market except one starts to use a new rating factor which is uncorrelated with other rating factors. Then the company which didn't use the new rating factor will experience anti-selection through a jump in overall frequencies and/or severities, whilst the structure of that company's pricing models might still appear to be correct. This illustrates the important point that unexpectedly high inflation can be a sign that the model is inadequate, even if the structure of the model seems to be fine. If a company is using a weak or outdated approach to claims modelling, then it has a much higher chance of experiencing trends in frequency and severity which cannot be explained by changes to the structure of the model. Having an accurate claims model is a prerequisite to understanding trends in frequency and claims inflation.

From this point onwards the report concentrates on the investigation into whether the structure of the model is correct or not.

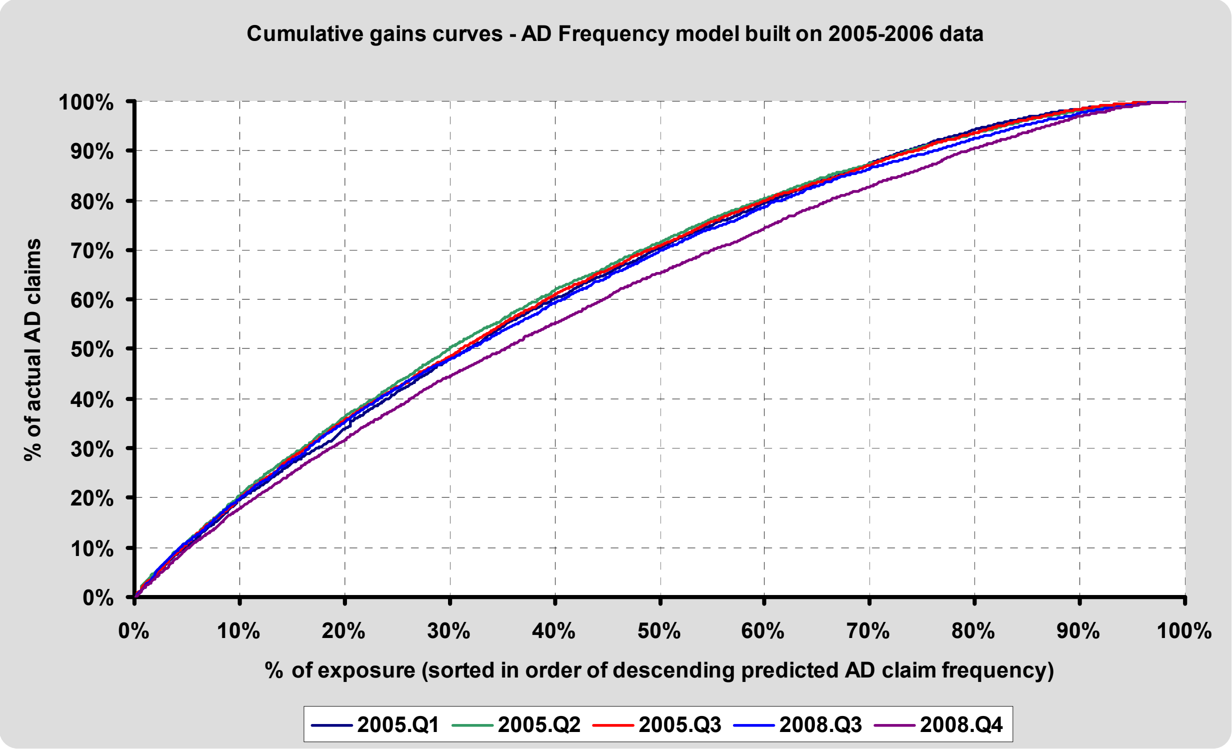
3.6 Claims models – worked example

The example which follows is based on recent UK private motor claims data provided by RBS Insurance Services Ltd. The analyses are based on accidental damage (AD) claims (claims and part-claims relating to the repair or replacement of the insured's own vehicle). The period of data is 2005 to 2008 inclusive. The as-at-date of the data is such that there is sufficient time for the claims to be considered to be close to fully developed.

Exposure episodes were split by accident quarter, in order to allow models to capture seasonality. GLMs were built of claim frequency and claim severity using Poisson and Gamma response distributions. These models were built and validated by the working party using the 2005 and 2006 data. The 2007 and 2008 data was held back in order to test different model monitoring methodologies. We deliberately built models which could loosely be described as being of an average technical standard for the UK motor market. The volume of data available is consistent with that of a large UK motor insurer. Within the models, seasonality is captured through use of an accident quarter variable and its interactions with other variables. In practice, as noted above, it may be the case that the model built using quarterly exposure episodes exists to monitor a model built using annual exposure episodes.

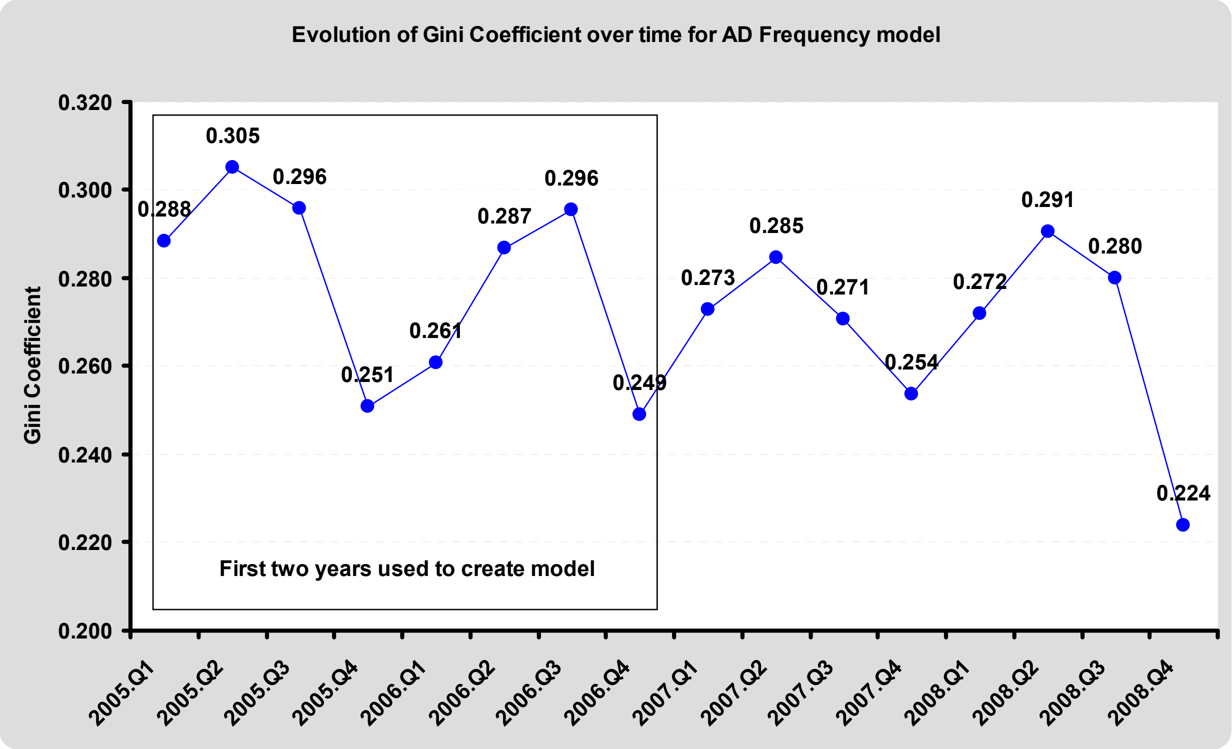
As with the demand data example above, the intention is to focus on simple ways to monitor the performance of the model over time.

Firstly, for consistency with the earlier analyses, the Gini coefficient will be examined as a measure for monitoring the performance of the accidental damage frequency GLM. The following chart tracks the gains curves on a quarterly basis for a model built on the 2005 and 2006 data.



For clarity, only first 3 and last 2 quarters are shown. The intention is to demonstrate the shallowness of the gains curves.

The associated Gini coefficient can also be plotted:



The first chart shows gains curves which are far shallower than those produced for the conversion datasets. This in part reflects the greater randomness of the claims process compared with the conversion process. Closer inspection of the charts revealed that there is a gradual decay of the gains curves throughout 2007 and 2008, with 2007Q4 and 2008Q4 behaving particularly badly. It could be that part of the reason for the poor performance in 2008Q4 is that the data is not sufficiently developed.

The chart of Gini coefficients plotted over time highlights a couple of issues:

1) the effect of quarterly seasonality dominates the chart, even though the model captures seasonality through the use of an accident quarter variable and its interactions

2) although it could be argued that there is a gradual reduction in performance, the degradation is not obvious until 2008Q4, the final quarter in the dataset

We did not expect the Gini coefficient to be a suitable metric for monitoring the performance of claims models but its inclusion here helps to highlight issues which also affect other metrics. However, the gains curves are useful because the gradual reduction in steepness indicates a weakening model. A change in the shape of the gains curve could be used as an indication that there has been a change in the underlying claims process.

The Mann-Whitney *U* statistic suffers from some of the same issues as the Gini coefficient and so is not discussed in detail here.

3.7 Desirable properties of a model monitoring metric for claims models

Before testing alternative metrics for monitoring performance we noted some desirable properties of such metrics:

- the value of the metric should not be biased by moderate changes in volume of business
- the value of the metric should not be unduly influenced by seasonality
- ideally the metric could be used as the basis for decisions concerning whether to refresh or totally rebuild the model
- ideally the metric will give an indication of the financial loss in the event that the model is not performing well
- the metric should be relatively easy to understand and explain
- the metric should be robust to changes in average frequencies or severities

As noted earlier, it is useful to consider the pricing GLM in two components. The first component is the structure of the GLM. The second component is the projection of overall average frequencies and severities. As such, if we want to understand whether the structure of the GLM is still correct, we don’t want to use a metric which is totally influenced by whether the average observed value matches the average expected value.

It is assumed that one of the first stages of any model monitoring exercise is to plot the average observed frequencies or severities and the averages predicted by the models, for each accident period. Following this analysis, if we wish to investigate the performance of the structure of the GLM then it is very helpful to scale the individual model predictions (normally using multiplicative scaling for consistency with the use of a log link function) such that the average model prediction equals the average observed response for each historical accident period included in the investigation. If this is carried out then metrics will not be influenced by the fact that the observed average experience doesn't match the predicted average experience. For metrics such as the Gini coefficient or the Mann-Whitney *U* statistic this is not an issue because the metrics rely on only the rank ordering of the observations and the rank ordering of the model predictions. For other diagnostics however, this scaling avoids the diagnostic being dominated by the overall average being incorrectly predicted.

3.8 Alternative model monitoring metrics

Any model monitoring statistic should seek to identify how well the model explains the variation present in the observed data.

In linear regression, the proportion of the observed variation which is explained by the model can be a useful metric. It is known as the "R squared" or R^2 of the model. There are various extensions to this which are designed for use with binary response data and with GLMs. One of the attractions of such a measure is the ease with which it can be communicated. R^2 for a linear regression model is defined as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

where y_i is the observed response for the *i*th observation, $\hat{y}_i$ is the predicted response for the *i*th observation and $\bar{y}$ is the mean observed response.

In order to adapt this for consideration in the claims modelling context, a couple of adjustments are required. Firstly, we wish to perform the scaling exercise outlined in section 3.7 above, so that the average predicted response equals the average observed response for each accident quarter. Secondly, we need to allow for the weights defined by the exposure period of the frequency model. If we do this then we end up with the following metric, which can be calculated for each accident quarter:

$$1 - \frac{\sum w_i(y_i - a\hat{y}_i)^2}{\sum w_i(y_i - \bar{y})^2}$$

where $a = \frac{\sum y_i}{\sum \hat{y}_i}$ is the scale factor, w_i are the weights of the model and other notation is as before.

To avoid cluttering the report, charts of this statistic and a detailed discussion are omitted on the basis that the working party agreed that it would be better to proceed with a statistic more closely related to the theory of GLMs. For the record, the results using this statistic were broadly similar to those which are about to be presented but lacked precision in some areas.

The deviance of a GLM is a measure of how well it fits the data. A small deviance implies a better fit. It also happens to be the quantity which is minimised in the GLM fitting algorithm. The proposed metric could be called a "deviance based R^2" and is defined as:

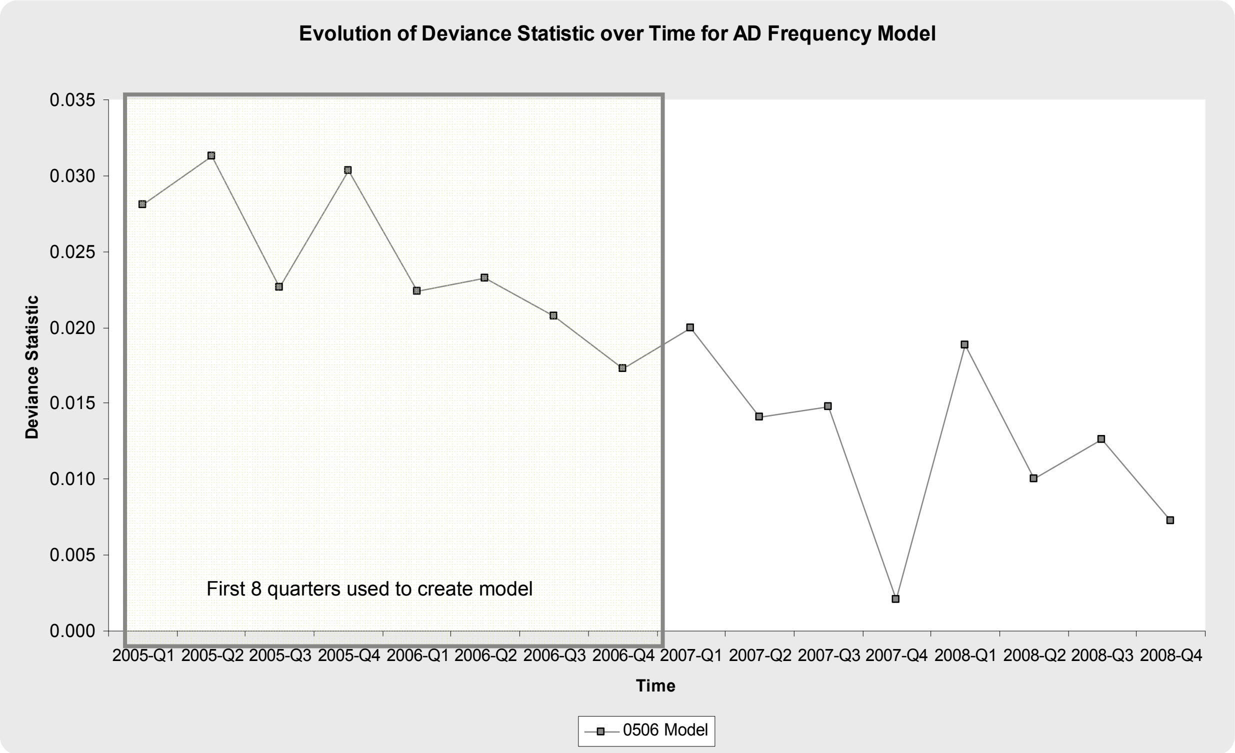
$$1 - \frac{deviance_{\text{model}}}{deviance_{\text{null}}}$$

where $deviance_{\text{model}}$ is the deviance of the selected model and $deviance_{\text{null}}$ is the deviance of a model containing just a mean parameter. In both cases the predicted values have been scaled as described in section 3.7, such that the overall predicted mean response matches the mean observed response for each accident quarter.

The statistic can still loosely be explained as the proportion of variation that is explained by the model, so a large value of the metric means that the model is performing well and a small value means that the model is performing badly.

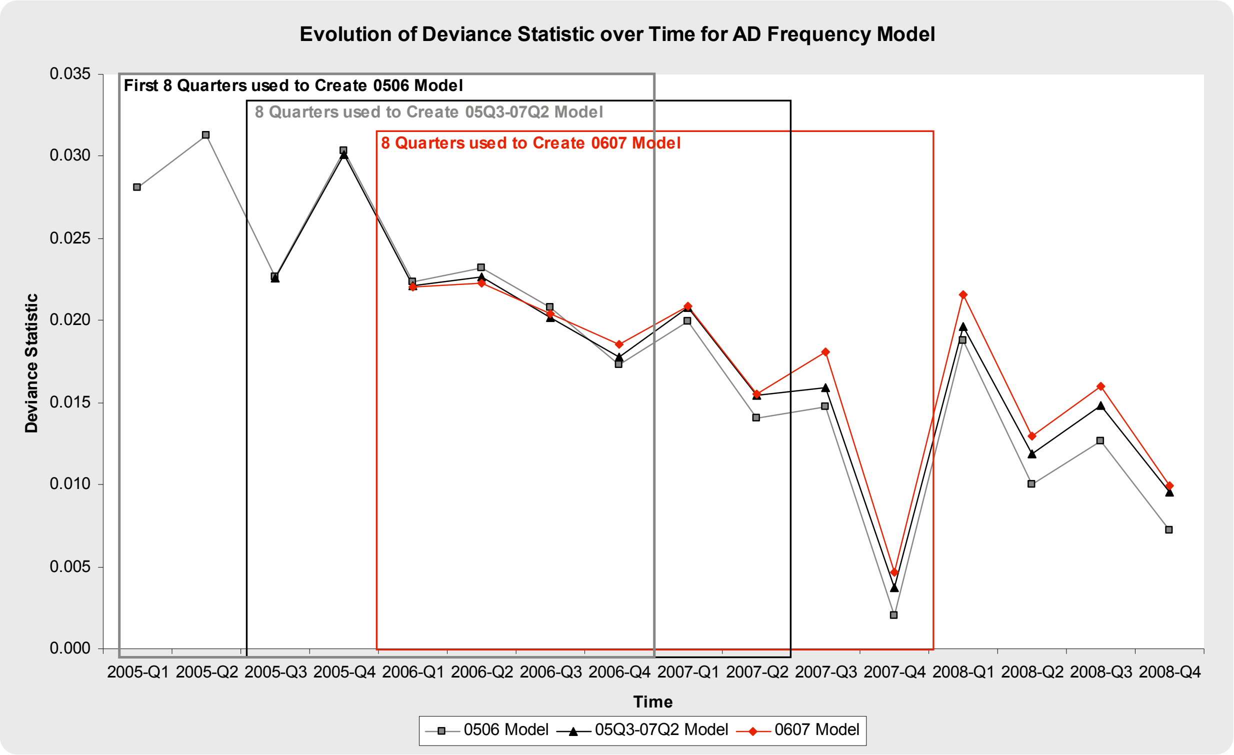
As an aside, it's worth noting that this statistic should not be used to compare the performance of two models with different numbers of parameters on an "in-sample" dataset because there is no adjustment for the number of degrees of freedom.

The following chart shows a plot of the statistic for the AD frequency data:



The reason for the model degradation over the period of modelling (2005-2006) could be found by investigating the underlying claim severity distributions over the period in question. A change had occurred, possibly a change in cover, terms, target market or claims strategy which led to an evolution of the claim severity distribution over the period. This pointed to a change in the type of claims which would therefore have an impact on the claim frequency model. This highlights the fact that histogram plots of both claim frequency and severity over successive accident quarters are important for establishing whether there has been an underlying shift in the claims experience. If this data isn't available then it will be harder to understand why the statistical diagnostics suggest that the model has performed poorly. Similarly, it is important to monitor the underlying mix of "event types" which give rise to accidental damage claims. For example, a shift towards more accidental damage claims being associated with third party property damage will have an effect on the performance of the accidental damage model.

The results using the deviance based R^2 suggest that model performance continues to degrade over 2007 and 2008. In particular, the model performance in 2007Q4 appears to be particularly poor. The following chart includes results from models which have been refreshed based on more recent data. The first new model is based on data from 2005Q3 to 2007Q2. The second new model is based on data from 2006Q1 to 2007Q4.



The chart shows that refitting the model to more recent data does not prevent the overall slide in model performance in quarters leading to 2007Q2. Re-fitting the model to more recent data does however improve model performance in the future. This can be seen by noting that for 2007Q3 onwards, the model performance of the model parameterised on 2006 and 2007 data (plotted in red) performs better than the models built on older data.

The extent to which this statistic suffers from seasonality is unclear. Three of the four Q4 periods have relatively low values of the statistic. If the statistic does suffer from mild problems to do with seasonality then this could simply be a fair reflection of the risk environment being different during the winter period.

Referring back to the other ideal requirements, this metric is reasonably invariant to volumes although sampling error will introduce noise on smaller claim severity datasets. The metric is fairly easy to explain as the proportion of the variation which is explained by the model. There may be a tendency for the metric to drift slightly upwards when observed average claim frequencies drift upwards and vice versa. The metric however gives neither an estimate of the financial impact of the model being out of date, nor a concrete methodology for deciding whether to refresh a model or to completely rebuild it. Bootstrapping could be used to provide a confidence interval for the statistic. However, an easier approach would be to carry out a calibration exercise based on historical data in order to set thresholds for triggering remodelling exercises. Such thresholds could be defined as percentage reductions in the values of diagnostic statistics relative to the values observed on the modelling dataset.

We next set out to investigate whether it's possible to produce a metric which highlights the cost of a model being out of date. It was thought that this would work better as a second metric, to be considered in addition to a more statistical measure of whether the model is performing or not.

This area wasn't considered in the context of the conversion dataset, on the basis that different companies will put their conversion models to different uses. Some companies will use conversion models which directly drive the price offered to the customer. Other companies will have conversion models for assessing different pricing strategies in a more offline sense. So the extent to which a poor conversion model causes loss of profit depends very much on way that the conversion model is used in the pricing process.

Similarly, in the context of claims modelling, different companies will have different approaches to using their GLMs. Some companies will use their GLMs directly (or with very slight moderation) in their rating engines and other companies will use their GLMs merely to guide the people who are setting the rates. In some situations discounting by brokers or agents will mean that the GLMs play a limited role in determining the price that the customer pays. However, following the rise of aggregators (price comparison websites) in the UK market and the resulting increase in customer price elasticities there has been increased recognition of the benefits of using GLMs in the pricing basis.

The working party made a couple of attempts at deriving metrics to assess the financial cost of a claims GLM being out of date. Whilst neither was explored fully, it's worth documenting the approaches in case this area receives attention in the future.

Both attempts used the following assumptions. Firstly let's make the simplifying assumption that the GLMs predicting claim frequency and severity are housed in the rating engine and are used without modification or moderation in the premium calculation formula. If a few more assumptions are made then it's possible to derive an estimate of the cost of the GLM being out of date. The assumptions aren't particularly realistic but the intention here is to trigger thoughts about how this could be tailored to a particular company. The focus here will be limited to the AD frequency model but could be extended to cover all GLMs.

The first attempt assumed that there is a model which is in force and that it is possible to quickly re-estimate the parameters of this model based on the most recent experience (allowing an appropriate time for claim development). Let's call this second model the updated model. The updated model is quickly inspected to check that its parameter estimates and implied factor smoothings are still reasonable. Then, on an individual risk level, the difference in fitted (predicted) values for claim frequency and severity are calculated between the two models. This enables the difference in predicted risk premium cost between the two models to be known. Assuming a simple rating basis, such as a cost plus approach where gross premiums are broadly x% greater than risk premiums, it's possible to calculate the difference in gross premium implied by the two models. Next, if customer price elasticities are assumed, it's possible to assess the impact on volumes of using the in force model rather than the updated model. Then, if it is assumed that the updated model has a more accurate estimate of risk premium than the in force model, it's possible to calculate the impact on profitability of charging a premium based on the in force model. By summing these volume and profitability metrics across all risks in a representative profile, some measure of the cost (both in terms of volumes and in terms of profitability) of using the in force model compared to the updated model can be known.

The results from such an approach are shown on the example claims model dashboard (chart 4) contained in Appendix C.

The second attempt at deriving a metric describing the financial cost of using an old model relied on three datasets, rather than two. This time the approach was to compare the in force model with an updated model but to perform the comparison on a hold out sample (the third dataset). Suppose the scope is limited to testing the cost of using an out of date AD (own damage) frequency model. One approach considered was to use the hold out sample to see what improvement the updated model gave over the in force model. In order to create a metric with units of currency, we calculated the absolute deviations in predicted frequency between the in force model and the updated model for each risk in the sample. The deviations were assigned to be positive if the updated model gave a more accurate prediction and negative if the in force model gave a more accurate prediction, where accuracy was evaluated with reference to the observed number of claims from each risk in the hold out sample. The next stage was to scale each such deviation in predicted frequencies by an assumed average claim severity. The scaled deviations then represent the improvement in AD partial risk premium estimation. By summing these scaled deviations over a representative profile, a metric was calculated which describes the overall improvement in AD partial risk premium estimation, with units of currency.

We tested this approach on real AD frequency data. The updated model was based on data which was six months more recent than the data used for the in force model. Both models were based on datasets containing two years of exposure data. The annualised improvement in partial risk premium estimation, scaled to an assumed typical portfolio size of one million vehicles, was only £40,000. Obviously this figure can’t directly be interpreted as an amount which would be added directly to underwriting profit if the updated model had been used. The aim was simply to try to find a way of quantifying the financial lift available to insurers by taking advantage of the most recent analytical data. A more thorough study into this area is required before conclusions can be drawn.

We did not carry out any further investigations in this area but noted that it is an area in which a future working party could undertake research aimed at demonstrating the value added by actuaries and statisticians by using superior and up-to-date technical models. The relevance of such an analysis to individual companies will of course depend on the degree of detachment between the technical models and the price actually paid by the customer.

3.9 Dealing with data which isn’t run-off

A common problem is that stakeholders want to know if the model is still valid and working over a period where the claims have yet to fully develop.

An approach to solving this problem is to obtain historical claims data at the same stage in its development as the cohort in question. Copies of the in-force GLM(s) corresponding to the combination(s) of claim types and frequency/severity of interest can then be parameterised using the historical undeveloped data. This will probably require variables and interactions to be dropped from the models to reflect the fact that the data is likely to be more ‘noisy’ than would be the case when it is fully run-off. The performance of this model on more recent data can then be assessed using the methods described above. Of course it may also be useful to perform detailed segmental comparisons of predicted and observed experience which are not discussed in this report. The results of any analysis could be affected by any changes in the claim settlement pattern.

3.10 Monitoring changes in mix of business

We briefly considered monitoring mix of business and how this relates to monitoring technical pricing models. In summary, it was thought that mix should always be monitored regardless of the level of monitoring of technical models. Indications that the mix has changed will serve as a useful input into the model monitoring process. The extent to which the change in mix degrades the model performance depends on the quality and granularity of the model.

If a more sophisticated method for monitoring mix is required then a multivariate visualisation technique, such as the star plot, could be used to track the centroid of an evolving mix over time. We briefly experimented with tracking centroids over time and found that interesting and useful results could be obtained even if the analysis is limited to the five most important rating variables. Whilst measures of change in mix are not included in the dashboards in Appendix C, they could be a useful addition.

4. How would model monitoring work in practice?

4.1 Introduction

The four questions that we attempted to address in this report are:

1. How can we detect material differences between modelled and actual results as quickly as possible in order to prevent financial loss?

2. Can we use model monitoring techniques to migrate away from calendar based model refresh/rebuild cycles so that we focus our resources on models which are broken?

3. Can we identify processes that are straightforward enough to be used by a wide audience and have analogies with well-established model comparison methods?

4. Can we create a monitoring process which would improve the risk management of the pricing function?

The aim of this section and the next section on recommendations is to answer these questions.

We felt that existing monitoring and review processes could typically be enhanced by the following processes:

- The use of dashboards created regularly as part of a semi-automated process. The content of the dashboards would be defined by experienced technical practitioners but ideally they would not require specialist resources to be updated regularly.

- A regular review of internal and external events against a list of key “trigger events”. The working party defined a trigger event to be anything which should lead to a consideration of whether the technical model(s) are still valid. Ideally dashboards should contain lists of trigger events which have occurred since the last update of each model. See Appendix B for a list of example trigger events.

The remainder of this section concentrates on the creation of dashboards for monitoring technical models.

4.2 What might model monitoring dashboards look like?

Before discussing how detailed dashboards can be produced for individual technical models it is worth considering what a “master dashboard” might look like. This dashboard will collect and summarise the most important information about each model and will act as a high-level overview of the technical models currently in use. The main challenge in the production of this dashboard is to present a metric which describes the overall performance of each model and is capable of being interpreted by all pricing stakeholders – technical and non-technical. Detailed dashboards for individual technical models will be considered later.

Detailed considerations when designing the master dashboard include:

- Ideally a single metric could be applied to all models to describe their overall performance but this may not be possible given the variety of purposes served by different models.
- Significance of the model - this could be a simple assessment of the importance of the model without requiring any calculations. For example, in motor, it could be decided that a model used to predict windscreen claims is of lesser importance than a model used to predict bodily injury claims. The importance of models could be described by a simple score out of five. Alternatively, the importance of a model could be derived as the result of a scenario testing investigation into the cost of the model being broken.
- The dashboard should contain the date of the last model refit/refresh exercise.
- The dashboard should note the period of data used to fit the model and the as-at-date where relevant.
- Lists of trigger events (defined above) which have occurred since the last modelling exercise can be included for each model on this master dashboard. This enables senior stakeholders to make recommendations on the need for model updates.
- If the plan is to have detailed dashboards for pairs of peril risk premium models (i.e. frequency and severity), then it may also be useful for the master dashboard to have a metric describing the overall performance of all risk premium models. A simple way to do this might be to show a lift curve comparing predicted total claims costs with actual total claims costs for recent periods.

Appendix C contains example model monitoring dashboards which were constructed using real data. These dashboards contain a number of statistical diagnostics and charts. The aim of these is to prove the concept of model monitoring dashboards by illustrating the types of metrics that could be used.

Descriptions of the key diagnostics shown on these example dashboards are provided elsewhere. In this section we focus on one or two practical issues.

We felt that it is possible to summarise the performance of a conversion or retention model using fairly simple metrics. The chart showing the Mann Whitney *U* statistic plotted over time, along with bootstrap confidence intervals was found to be useful. These charts have the label “4” on the example conversion model dashboard. The chart showing the “deviance based R^{2}“ statistic shows that the use of this alternative statistic would lead to similar conclusions. The decaying gains curves were also found to be a useful diagnostic in this setting.

The working party agreed that it is much harder to find simple metrics in the claims context. This is perhaps partly because of the greater financial importance (in most cases) of claims models compared with demand models. We also felt that greater complexity in the claims process gave rise to the need for a variety of diagnostics on model performance. The conclusion was that no single metric is capable of telling the complete story about the performance of a claims GLM.

In order to properly understand the performance of claims models it is necessary to have a solid understanding of how the claims process is evolving. One example of particular relevance to the UK motor market is the escalation of policy excesses on comprehensive business, partly driven by intense price competition on aggregators. This of course has various impacts on the claims process which will affect the performance of technical models. One such impact is the reduction in small own damage claims and consequent increase the proportion of claims which are associated with a third party claim cost. In the experience of the members of the working party, this type of background information is essential in order to understand the trends shown by statistical monitoring diagnostics. As such, the example dashboard relating to AD (own damage) claims has a number of diagnostic charts along the bottom which are designed to provide some of this background information. Several of the charts look at the evolution of the claim severity distribution. This is important because a change in the severity distribution is a strong indication that there has been a change in the process which gives rise to claims. This in turn advises the owners of the models that they should tread carefully. It is worth noting that a shift in the severity distribution is a sign that the frequency *and* severity models could need review.

In order to protect the intellectual property of the companies which provided the data, the charts shown on the dashboards have been disguised. In a normal company setting it will be easier to see the links between the simple diagnostics (shown along the bottom of the AD dashboard) and the more complex monitoring statistics.

To conclude the discussion of dashboards for claims models, the working party felt that it was important to monitor simple diagnostics such as the claim severity distribution in addition to using a global performance metric such as the “deviance based R^2”.

4.3 How can we better focus our resources on models which are broken?

Leading European personal lines insurers may have at least 15 claims models and 20 demand models in each territory, perhaps with different suites of models for different brands. These are usually maintained by means of a well-embedded model review process based on calendar cycles. The problem with this approach is that resources may not be used in the most efficient way. For example, some models may be replaced before they are broken and others may have been broken for a long time before they are replaced.

By creating the dashboards described above, companies can migrate towards a model refresh system which is more intelligent than simply relying on calendar cycles. By regularly updating the dashboards using the most recent data and using agreed traffic lights to describe model performance, management and technical resources can quickly see which models are in greatest need of attention. If it is necessary to define a minimum model refresh frequency for the purposes of risk management, the process of setting up the dashboards using historical data can give the answer. Referring to the conversion dashboard in Appendix C it’s easy to get a feel for how often the aggregator and telesales conversion models needed to be refreshed.

Another question when allocating resources is “Is it possible to know in advance whether a full rebuild of the model is required or if a simple refit to the latest data will suffice?”. Although it may seem like a good idea to have rules based indicators to decide when a model requires a simple refit as opposed to a complete rebuild, we concluded that it isn't possible to define such rules. In the example of the aggregator conversion model, the performance in June 2008 was found to be very poor. Only after a reasonably detailed investigation of the June data was it possible to say that the solution was to add two variables to the model. In other circumstances a full rebuild may have been required but this would never be known until a more detailed investigation of the data had been performed. We noted that the first stage of any investigation triggered by the dashboard should be to identify whether a full rebuild is required. It is only after this investigation is carried out that the extent of required resources can be known.

We felt that the greatest value from adequate monitoring came from the speed and cost advantages of swiftly identifying that a model is broken. There will always remain a requirement for technical resources to address the complex problem of how to repair a degraded model. The model monitoring metrics however can act as a strong foundation for debate between senior pricing management, model users, model builders and the wider business.

4.4 Can we produce output that can be understood by all pricing stakeholders?

Stakeholders in the pricing process have a diverse range of skills. We felt that it was vital that dashboards should be designed with the ease of use of all such users in mind in order to:

1. Enable senior stakeholders to quickly assess and comment on model performance
2. Promote a wider understanding of how technical models are used in the business
3. Promote a wider understanding of the benefit of good technical models and the cost of poor technical models
4. Promote an understanding of the types of internal and external events which are likely to break technical models
5. Promote feedback to the technical pricing team of information from other areas which is likely to be of importance in the pricing process

These considerations led to the use of gains curves in the monitoring of conversion models. Many pricing stakeholders will already be familiar with gains curves as they are frequently used by marketing departments, for example when deciding on the extent of direct marketing campaigns. Use of metrics which are already understood will allow broader communication of results and/or combining of any monitoring procedures across departments.

Where stakeholders only need a high level view it may suffice to provide them with the master dashboard discussed above, rather than the dashboards for individual models. Where the detailed dashboards are provided (as in Appendix C), it is helpful to have labels explaining whether large values or small values of individual metrics indicate good performance.

Similar considerations should apply when communicating the results of a model validation exercise.

4.5 How can model monitoring improve pricing risk management?

It should already be clear that a formal approach to model monitoring can improve risk management by delivering an early warning when models are broken.

The working party briefly discussed the benefits from model monitoring in the following areas:

- Governance – insurance risk is the key risk category that general insurance companies have to manage on a daily basis. Sound governance of insurance risk requires regular monitoring of a wide range of key risk indicators with reference to the enterprise risk appetite. Model monitoring dashboards could form an integral part of the underwriting and pricing suite of monitors that feed into the wider key risk indicators dashboard.

- Audit trails – the use of model monitoring dashboards will aid communication with internal and external audit processes. The dashboards will demonstrate to internal auditors that there are sound systems and controls in place to monitor the technical models which feed into the pricing decision making process. It is also possible that a best practice enterprise risk management framework under Solvency II (in the EU) might require increased visibility and tractability of the modelling process which leads to pricing decisions, possibly including stringent external disclosure requirements.

- Definition of roles – the use of a formal model monitoring process can help to separate the responsibilities for model building and model monitoring. Where a clear separation of duties is possible, such segregation of responsibilities ensures sound pricing governance and a timely response when models have degraded over time.

5. Recommendations

We believe that model monitoring is an important part of the pricing process and can add substantial value. Some of the leading personal lines insurers in the UK are currently in the process of testing techniques and processes similar to those described in this report. We made the following notes to try to help pricing managers who wish to set up a regular monitoring process for their technical models.

Firstly, the extent of the possible benefits from a model monitoring process will depend on the size of the company. Large companies which have teams of individuals working on the modelling and price setting processes stand to gain more than smaller companies. It is likely that smaller companies will rely on fewer models and in smaller companies the communication of model performance across teams will be less of a challenge. In particularly small companies it may not be possible to separate the role of model building from that of model monitoring. Secondly, in order to obtain the maximum benefit from the monitoring, it is essential that it is used as an input to the risk management process as well as an input to the pricing decision making process.

When defining the scope for a model monitoring process, it is probably sensible to opt for a phased implementation, possibly running a pilot project dealing only with demand models (e.g.) in the first instance. The need for a master dashboard as well as detailed individual dashboards should be discussed with stakeholders. The use of bootstrap techniques (or other) to understand the uncertainty in performance metrics can slow things down and may not be required on day one. Likewise, the metrics which seek to define the financial cost of models being out of date may wisely be left out of scope for an initial implementation. A key consideration in the claims setting is the need to produce models which explicitly capture seasonality and the possibility of splitting policy episodes by accident quarter in the underlying data. All other diagnostics shown on the example dashboards are easy to obtain using standard statistical and modelling software.

Once the scope has been agreed, the first stage of any implementation is to test the chosen metrics on historical data using the historical models which were in force at the time. This calibration process will include the definition of thresholds for deciding whether a model is in need of review. The definition of such thresholds should be done with input from all pricing stakeholders so that they feel confident to use the chosen metrics and decision rules in the business as usual pricing process. The working party has deliberately avoided defining decision rules based on threshold values of particular metrics because we feel that individual companies should own this decision. It is likely that different companies will use different metrics and different thresholds to define model performance indicators, such as the traffic light system labelled “Current Performance” on the example model dashboards in Appendix C. The most important thing is that pricing stakeholders feel at ease with the metrics in use and that the robustness of the metrics has been tested on historical data. An output of this calibration process will be an agreed frequency for the production of the dashboards, which may differ for different models.

Once a regular technical model monitoring process has been in place its effectiveness should be reviewed. One of the success criteria should be whether all pricing stakeholders feel that they are aware of the current performance of the technical models. Another success criterion should be whether the process has been integrated into the pricing and underwriting risk management framework.

Appendix A: Model validation

A.1 Introduction

Model validation is the use of a range of processes and techniques to investigate whether a model is predictive in the market environment in which it will operate.

The working party noted that there is lots of information about model validation on the internet and in the statistical literature. Recent presentations to the Casualty Actuarial Society by James Guszcza and Christopher Monsour (see References) are a good starting point for material specific to the insurance industry. What follows is a brief overview which highlights a few areas of particular interest to those involved in personal lines pricing. The overview focuses heavily on the use of hold out samples as a means of validating models.

A.2 Why should models be validated?

Without checking the performance of a model on a hold out sample, models which seem to fit data reasonably well could be very poor at predicting future experience.

Confusion frequently exists over the need for model validation. Some modelling techniques explicitly require the use of hold out samples. Other modelling techniques do not obviously require the use of hold out samples to validate the model. For example, GLMs are supported by statistical diagnostics based on maximum likelihood theory. So in the case of GLMs there would seem to be less need to use hold out samples because the user can rely on the statistical theory.

However, various features of the data or model search methodologies can to some extent undermine the statistical theory:

- The consideration of large numbers of variables increases the chance of Type I error (error whereby a variable is erroneously included in the model)

- Automated (or semi-automated) search processes effectively reduce the number of degrees of freedom

- Heteroskedasticity (or invalid assumptions about the variance function in the context of GLMs) can invalidate the assumptions of factor inclusion tests

- The use of spatial smoothing or vehicle classification techniques, whilst adding substantial predictive power to models, reduce the number of degrees of freedom in a manner which cannot be calculated

The reduction in the number of degrees of freedom makes it more likely that the modeller will include insignificant factors and therefore end up with an over-parameterised model.

Another benefit of model validation is that it can be done as part of an internal peer-review process and thereby enhance the risk management of the pricing process. Some model validation techniques may even reduce the amount of time spent reviewing a model. For example, for models which have a low business impact it may be sufficient to assess the performance of the model on a hold out sample, rather than undertaking a painstaking review of all of the modelling steps.

With these points in mind, we believe that all models should be validated.

A.3 A few warnings

In order to carry out a fair model validation process, the model selection process must not have used the hold out sample in any way. In practice this is usually difficult if not impossible because it is always necessary to reconcile the full dataset (including the hold out sample) to other sources before the modelling starts. It is also common for the practitioner to be aware of trends which have been identified through the business as usual monitoring processes. To the extent that this is the case care should be taken when treating the hold out sample as being independent of the rest of the data.

In our experience the methods described below, when used in isolation, are often limited in their power to identify weak models. This leads to a few rules of thumb:

- Even if a model passes a particular validation test it doesn't mean to say that the model is robust in a general sense
- If a model passes many validation tests but fails one then the model is probably still invalid
- A robust model validation recipe will use several approaches and several metrics for assessing the model's performance

It's also worth noting that where models are used to drive the price offer made to the customer, the real quality of the model remains unknown until after it is implemented in the rates. The act of implementing the model will possibly lead to a change in mix of business, meaning that validation on historical data is of limited usefulness. In such cases it could be considered that the roles of model validation and model monitoring become blurred, with the final stage of model validation and the first stage of monitoring becoming the same activity.

A.4 Hold out samples

Firstly it helps to introduce some terminology:

- The *training set* is the dataset used to build the model
- The *validation set* is a dataset which was not used in the model building which can be used to assess the performance of the model on unseen data
- The *test set* is a third dataset which is required if the model is adapted based on its performance on the validation set. This dataset gives a final check that the model is up to scratch.

If GLMs are being used it is possible that only the training and validation sets are required. The test set becomes necessary if the validation set is used to influence model selection. For the remainder of this report it is assumed that only a training set and a validation set are required. The validation set is used as the hold out sample.

A hard decision is how to choose the type and size of the validation set.

The most common method in the statistical literature is to take a sample of observations at random. Judgement is required in the selection of the proportion of observations to set aside. The validation set needs to be large enough that the predictiveness of the model can be properly assessed. The obvious problem is that a large hold out sample significantly reduces the amount of data available for building the model in the first place.

We did not find any definitive advice in the literature on suitable sizes for the validation set. All suggestions were within the 10-30% bracket. We feel that 20% might be a reasonable figure for a first attempt.

The members of the working party think that it can be helpful for the validation set to be of the same size as datasets used in any regular reports. For example, if the problem is to validate a renewal price elasticity model then it's helpful to use a dataset comparable to monthly renewal volumes. Any variations in experience can then be compared with deviations seen in monthly business as usual monitoring.

The other common way to define the validation set is to withhold data corresponding to the most recent time period. In a claims modelling context, this might involve setting aside the data from the most recent accident period. For price elasticity modelling this may mean setting aside data from the most recent month(s). We believe that this is a more powerful test of the model. Since the primary use of the model will be to predict future experience, it makes sense to validate the model against recent unseen experience. In particular, this approach is essential for models which are used to predict weather related events. The key here is to test the model on events which weren't used in the model parameterisation process.

A.5 "Biased" hold out samples

Although not based on statistical theory, the following method has been suggested by experienced practitioners. It is assumed that this method would supplement a more conventional model validation process.

It could be argued that models are most often made obsolete by changes in mix of business. Scenario testing the robustness of a model to changes in mix is possible through the use of biased validation sets. To give a fictitious example, a company may currently write business through two price comparison sites ("aggregators") and plan to increase volume by joining a third aggregator. Suppose that this company is relying on a price elasticity model to predict customer level conversion rates as a function of price. In this case a useful approach might be to remove one of the aggregators for use as a validation set. The model is then built on the training set and applied to the validation set. The performance of the model on the validation set may either give comfort that the model is robust to this type of change in mix or may identify problems which might exist if the existing model were to be used to price business written through the third aggregator.

The aim of performing these experiments is to judge the likely robustness of the model to an accidental or deliberate change in mix. Performing these experiments allows the modeller to better understand the strengths and weaknesses of models and to communicate these to stakeholders.

A.6 Cross validation

Instead of partitioning the data into one training set and one validation set, cross validation partitions the data into multiple training and validation sets. So instead of selecting one random hold out sample, the idea is to select multiple hold out samples. The performance of the model is then assessed by looking at the results accumulated over each of the validation sets. This then gives an indication of how well the model will predict the response on independent data (future data in the context of pricing).

Two common methods of cross validation are “leave one out cross validation” and “k-fold cross validation”.

Leave one out cross validation

Under this approach training sets always include all observations except for one randomly selected observation. This observation then forms the validation set. So there are as many validation sets as there are observations. This leads to computational challenge of fitting as many models as there are observations. This isn’t currently practicable for personal lines rating where large datasets can contain many millions of observations.

k-fold cross validation

A less extreme approach is to randomly partition the data into k training sets and k validation sets such that each observation is in exactly one validation set and is in k-1 of the training sets. Typically k might take the value 5 or 10. The process is as before. k models are fitted to the training sets and the performance of each model is then assessed on the validation sets. Since the process only requires that k models are fitted, this is far less computationally intensive than leave one out cross validation.

Challenges with cross validation

In order to respect the principles of cross validation, it is assumed that each of the models is built entirely independently of the corresponding validation set and of every other model. This is quite an onerous requirement, except where automated model selection and fitting tools are being used. So this is an onerous requirement in the context of claims modelling, where detailed GLMs are widely regarded as being the best practice modelling framework. In this context the GLM is likely to have been built on the full dataset (or the full dataset excluding an out of time validation set). It would be far too time consuming for the practitioner to build k models from scratch. The additional need for the k models to be built independently makes the task almost impossible.

There is however a pragmatic approach which yields useful results even if it breaks the principles of cross validation. The approach involves taking the model fitted to the whole dataset and re-fitting the model (with the same structure) to each of the k training sets. So the model search and smoothing of the model is performed once but this model is then fitted to each of the k training sets in turn. The performance of the models is then assessed.

When this approach is combined with the use of gains curves or other appropriate cost metrics, it can be a powerful method of detecting an over-parameterised model, as discussed in an example later in this section.

There is one further caveat which needs to be applied to this pragmatic approach. The approach assumes that the parameterisation of the model is ‘stable’. By stable it is meant that there are no factor smoothings which rely on, for example, the use of high order polynomials which whilst giving an appropriate smoothing on the full dataset might give wild and spurious trends if only a handful of observations were to be removed from the dataset. Experienced modellers should be able to avoid this.

A.7 Predictive accuracy and loss functions

A simple way to assess the performance of a model on a validation set is to define a loss function. A loss function quantifies the inaccuracy of the model by comparing the model's predictions with the observed response values.

Common loss functions include the model deviance (in a GLM environment) or the sum of the squared differences between observed values and predicted values.

Other metrics can be used to assess the predictive accuracy of the model. For example, Gini coefficients can be useful when predicting conversion rates. In logistic models, McFadden's pseudo R squared and Efron's pseudo R squared might be used.

In an ideal world the loss function would be based on the financial cost to the company of using inaccurate predictions. In practice, this is possible for sophisticated insurers which have technical models stored in a software package which can be used for scenario testing and projecting estimates of profitability. However, it may not be feasible to set up this environment at the model building stage.

Having said this, it may be possible to construct a loss function which is loosely related to the financial cost of using inaccurate estimates.

A.8 Commonly used charts

Gains curves

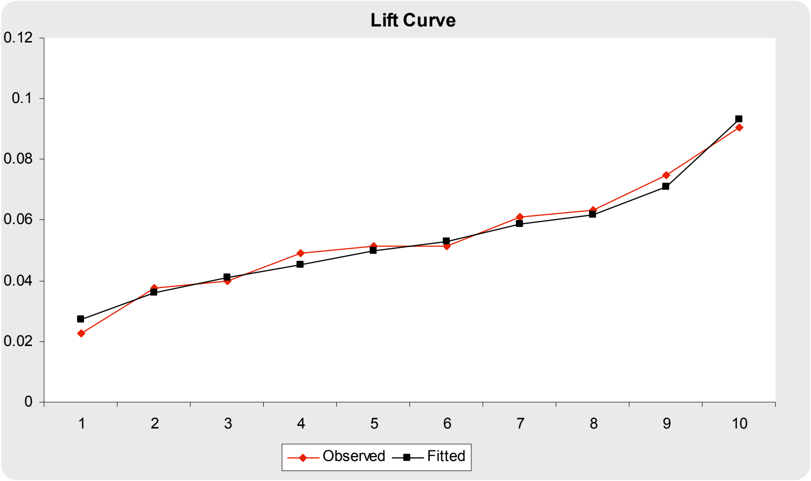
Gains curves were discussed earlier in the report.

Lift curves

Another way to present the results of a model validation is to use a lift curve.

A lift curve requires the observations to be sorted by increasing fitted value. The observations are then grouped by increasing fitted value. This is either done using a customised banding or simply by grouping the fitted values by percentile or decile.

Plotted on the chart are the average actual and average predicted response rates for each group of observations.

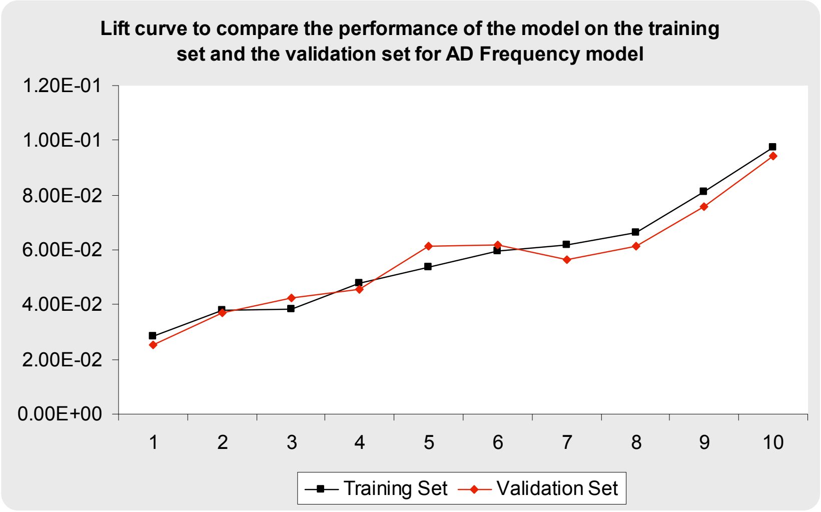


The ideal situation would be to have two entirely coincident lines. This never happens in practice, but the chart can sometimes be used to spot strange behaviour which may lead to further investigations.

In personal lines, for models of claim frequency or severity, the higher deciles correspond to observations which have the greater prediction error. For binomial models, the greatest prediction error will be for fitted values of about 0.5. So high predicted conversion rates will typically be subject to the greatest uncertainty in a conversion rate model (where the overall average conversion rate is assumed to be below 0.5).

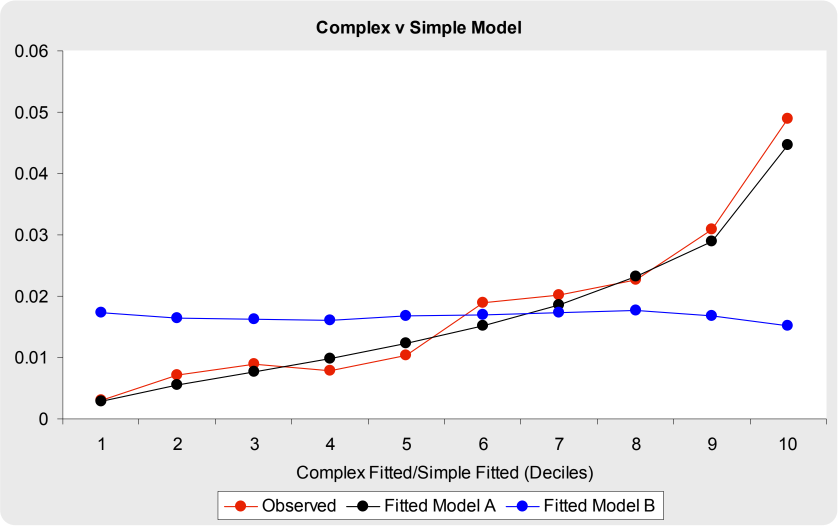
Lift curves can be a good way to present results to non-technical audiences as they are simple to understand.

The lift curve can also be used to compare the performance of the model on the training set and the validation set, as shown in the following chart:



The extent of any overfitting can be seen by the lower gradient for the validation set.

Variations on the standard lift curve can also be used to compare the predictive power of two competing models on a validation set. In this case the observations are sorted by increasing values of (model A prediction – model B prediction). Average predicted values for each model is plotted along with the average observed response.

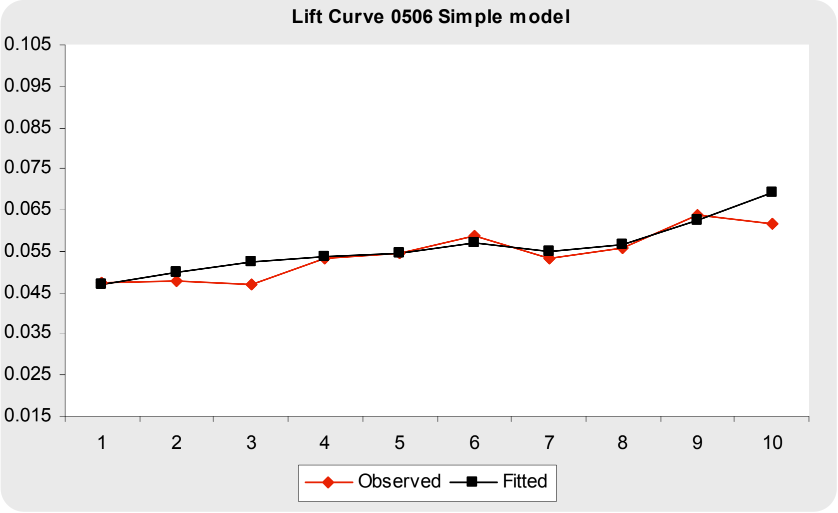


This lift curve chart clearly demonstrates that model A has significantly better predictive power than model B.

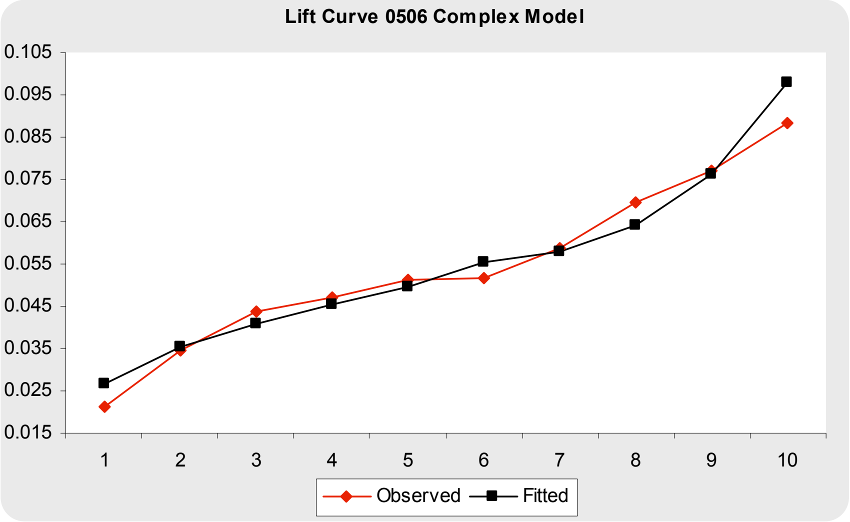
Drawbacks of gains curves and lift curves

Gains curves and lift curves can give misleading results when considered in isolation.

For example, the following lift curve is based on a risk model.



The lift curve doesn't reveal anything untoward. However, it was constructed using a risk model which only contains one variable. The following lift curve has been produced from a far more sophisticated model.



A similar story could be constructed using gains curves. The conclusion is that lift curves and gains curves may not be very useful if they are only used to look at the results of one model. Modellers should be wary when presented with a lift curve if they don't have a well understood benchmark to use as a comparison.

A.9 Detailed investigations

In addition to the methods described so far, modellers will wish to perform detailed analyses of validation performance against all rating factors and underwriting segment definitions. Standard one and two way analyses are commonly used in this context.

It's important to remember that the validation process doesn't need to be limited to variables which could be included in the models. As an example, for various reasons it's unlikely that a conversion model would use expected profit at customer level as an explanatory variable. However, assessing the performance of the model against such a variable can be very useful. Variables which may not be considered for inclusion in models but which are useful for validation are:

(1) Expected profit

(2) Competitiveness

(3) Expected conversion or retention rate

(4) Customer price elasticity

(5) Marketing segmentation

(6) Other external data

A.10 Should we refit to the full dataset once the validation has been performed?

Different opinions exist on this. The assumption is made in this section that the experience in the validation set is well predicted by the model and therefore that the model passes the validation test.

Some would argue that the final model (the model to be implemented) should be parameterised on all of the available data including any validation set. If an out-of-time sample is used the argument would be that it is important that the model captures the most recent experience.

Some however argue that reintroducing the validation data for the final parameterisation runs against the principles of model validation. They argue that the model parameterised on the combined data would then need another validation set to prove that it is predictive.

A pragmatic solution may be to include the validation set for final estimation of parameter values before the model is implemented. In the context of GLMs, this would mean taking the structure of the model fitted to the training data and updating the parameter estimates based on the combined training and validation sets. No modifications would be made to the structure of the model provided that the model still satisfied any commercial constraints (e.g. the need for smoothness).

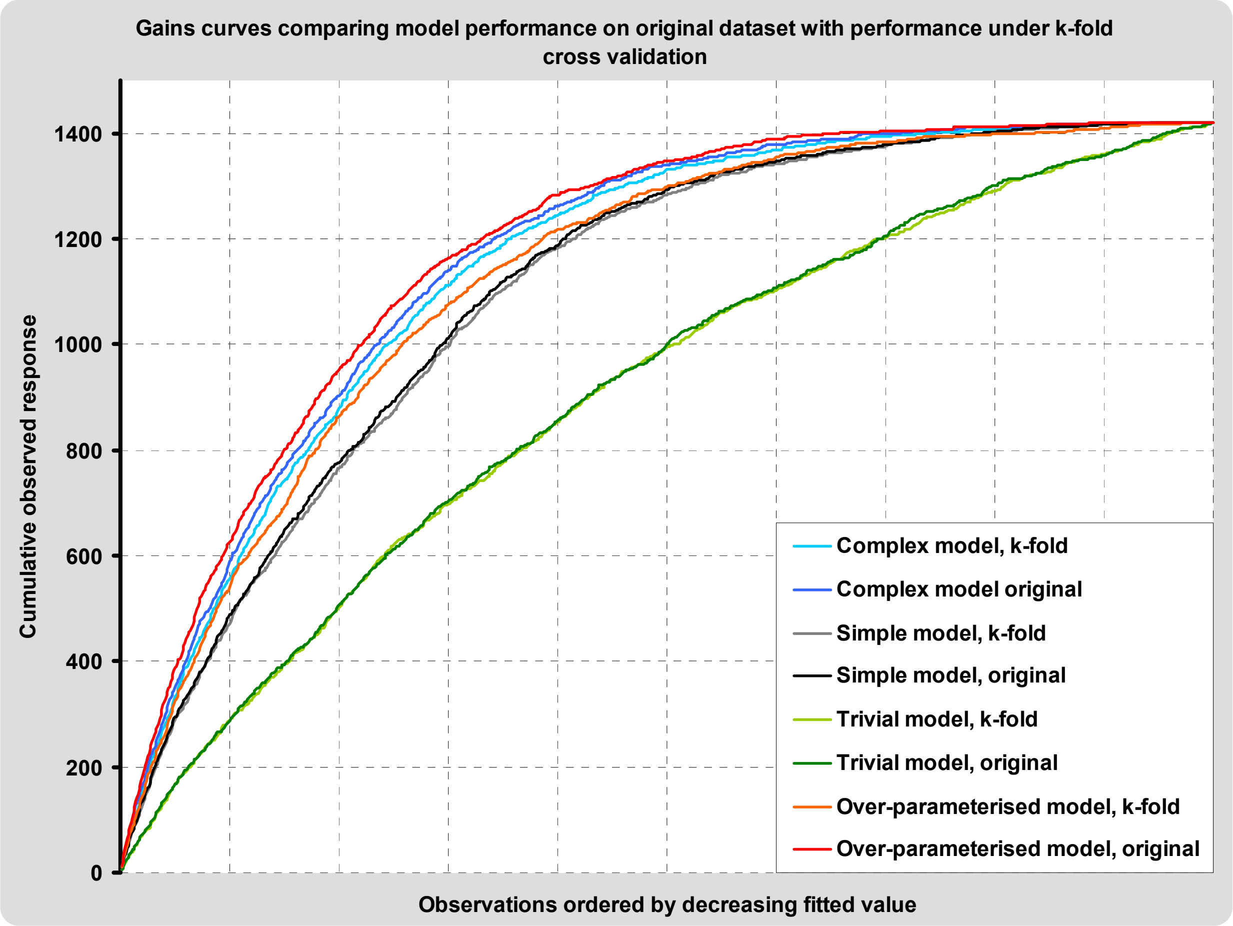
In certain overseas territories regulators may object to the use of models which have been fitted to only part of an insurer's historical experience. Under these circumstances the final model would need to be fitted to all available data.

Validation for fine tuning variable selection

Frequently the modelling process involves decisions over whether to include a variable or interaction which is of marginal significance. Less frequently a revised modelling process is proposed. In either case, validation techniques can give solutions. The comparison of global model fit statistics calculated on a hold out sample for two different modelling approaches can be an objective way of settling any argument.

We tried k-fold cross validation as a way to compare models with different levels of complexity. As an example, suppose that a company is familiar with gains curves. The gains curve for a particular dataset and model can be plotted, along with a plot of the gains curve constructed by considering the performance on all of the validation sets in a k-fold cross validation. If the gains curve substantially changes for the out of sample data then this is an indication that the model may be over parameterised.

The following chart shows the results of this process carried out for models with four different levels of complexity. There are two gains curves for each of four different models. The most interesting comparison is that of the modelled labelled "over-parameterised" with the model labelled "complex". The over-parameterised model beats the complex model on the in-sample data (gains curves for in-sample data are labelled "original") but is inferior on the out of sample data (gains curves for out of sample data are labelled "k-fold" on this chart).



If this technique is used then theoretically the selected model should again be validated on data unseen thus far.

Appendix B: Trigger events

The working party defined a trigger event to be any event, internal or external, which should lead to a consideration of whether technical model(s) are still valid.

The following list provides examples of such events, some of which have occurred in recent years in the markets in which the members of the working party are familiar. A similar list could be considered for use as part of a company's formal risk management procedures. It may be necessary to have a different list of such events for different types of model. In any case, we felt that for each model, it would be useful to list the relevant trigger events which have occurred since the last modelling exercise on a master technical model dashboard. This enables senior management to have a better overview of the models which are in force and to be able to identify models which are underperforming.

The working party does not consider the following list to be exhaustive in any way.

Internal

- Changes in product or underwriting, such as changes to:
    - cover
    - product features, including add-ons
    - underwriting terms and conditions e.g. compulsory or voluntary excess
    - monthly payment schedules/patterns
    - use of external data (e.g. fraud databases) at point of sale
    - level of NCD validation
    - use of CUE (Claims and Underwriting Exchange)
    - underwriting footprint (expansion or contraction)
- Changes in pricing, such as:
    - changes to explicit pricing constraints
    - revisions to models or rates through the regular rate review process
    - improved capability of sales systems (e.g. more tables available, introduction of real-time external databases) or availability of new predictive factors through other means
- Changes in distribution, such as:
    - an expansion or contraction of distribution (e.g. use of telesales, new affinity/scheme, joining a panel or aggregator)
    - a material shift in target volumes by source

  - revised approaches to up-sell and cross-sell
  - changes to commission agreements
  - changes in the availability of MI and other data from distributors
- Changes in administration and claims processing procedures, such as:
  - changes in case estimation
  - changes in claim development patterns
  - changes to the number of nil claims
  - outsourcing the handling of particular types of claims
  - changes to repair network
  - installation of new administration systems
  - the use of overseas call centres
- Changes in marketing, such as:
  - major changes in marketing spend
  - rebranding
  - the use of different media
  - new campaigns
  - large campaigns
  - marketing messages and channel discounts, e.g. (10% off for web customers)
- Significant changes in reinsurance cover purchased
- Acquisition or disposal of portfolio(s) of policies

External

- Changes in legislation, for example:
  - 4th and 5th EU directives (introducing minimum TPL cover across Europe)
  - EU gender directive
  - UK Ogden tables and Courts Act 2003
  - Knock-for-knock agreement in Italy

- Changes in regulation, for example:
    - increased emphasis on TCF in the UK (unlikely to be a direct cause for model review)
    - Italian regulator introduced a “filing of rates” requirement in 2003
- Changes in economic environment, for example a recession associated with:
    - cheaper cover in demand, e.g. increased demand for higher excesses and lower cover
    - rating factor abuse (mileage, overnight location, occupation, etc.)
    - exaggerated claims
    - more fraudulent claims
    - less frequent use of cars
    - higher theft claim frequency
- Changes in shopping behaviour, such as:
    - changes caused by new distribution channels, e.g. aggregators increasing buyers’ price elasticity
    - increased propensity for renewing customers to “switch” to the new business rate
- Changes in claims behaviour, such as:
    - changes in frequencies caused by external factors, e.g. hire purchase agreements
    - changes in severities caused by external factors, e.g. higher PI and/or PD claims inflation (e.g. major increase in the impact of accident management and credit hire firms)
- Competitor actions, such as:
    - changes in competitors’ target markets and/or rating structures, e.g. the use of price optimisation
    - significant change in marketing spend or share of voice by a major competitor
    - technological changes, e.g. PAYD
    - new market entrant(s) with aggressive goals and strategies
- Any other external events driving changes to the mix and volumes of:
    - quotes
    - new business sales
    - renewals

Appendix C: Example model monitoring dashboards

These are designed to be printed on A3 paper. The data has been disguised where necessary for reasons of commercial sensitivity.

Aggregator Conversion

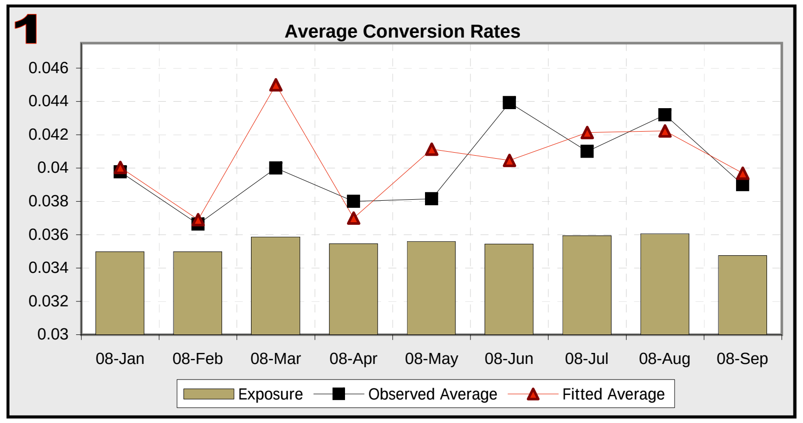
Model Importance

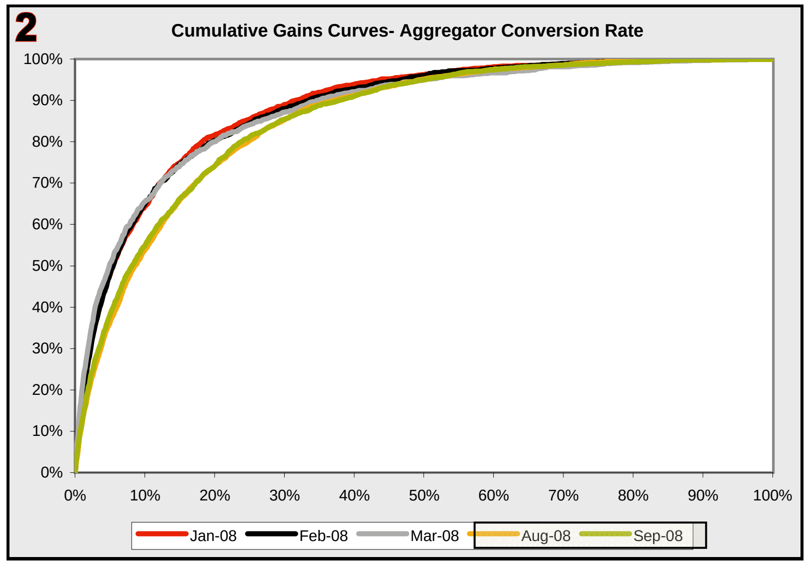
JAN 08 - MAR 08	Data Used in Model
APRIL 2008	Date Model Last Fit

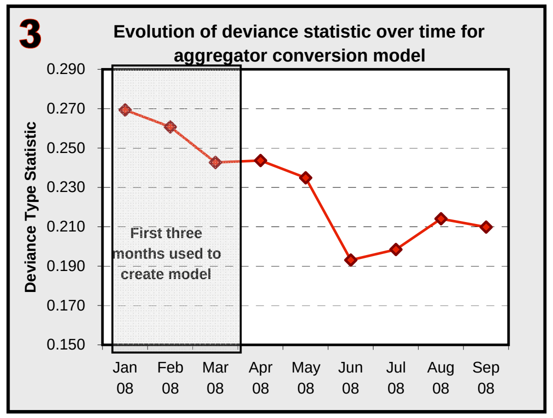
☒ 🟧 🟨 🟩 Current Performance

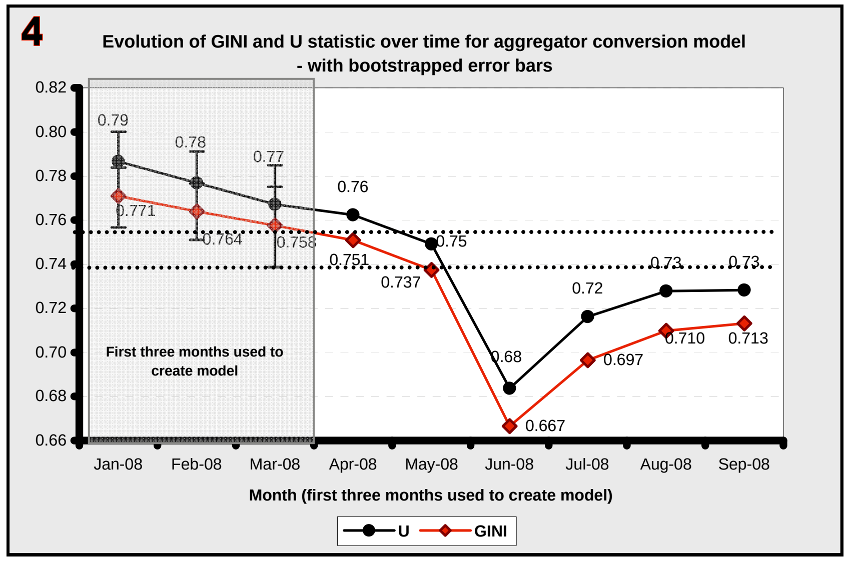
Model performance degraded significantly in June 08. Refit required.

Unique quotes (Year to Date):	1,056,528
Sales (Year to Date):	42,853
Average conversion rate:	0.0406


1 Average Conversion Rates


2 Cumulative Gains Curves- Aggregator Conversion Rate


3 Evolution of deviance statistic over time for aggregator conversion model


4 Evolution of GINI and U statistic over time for aggregator conversion model - with bootstrapped error bars

5

Elasticity Chart goes here

Telesales Conversion

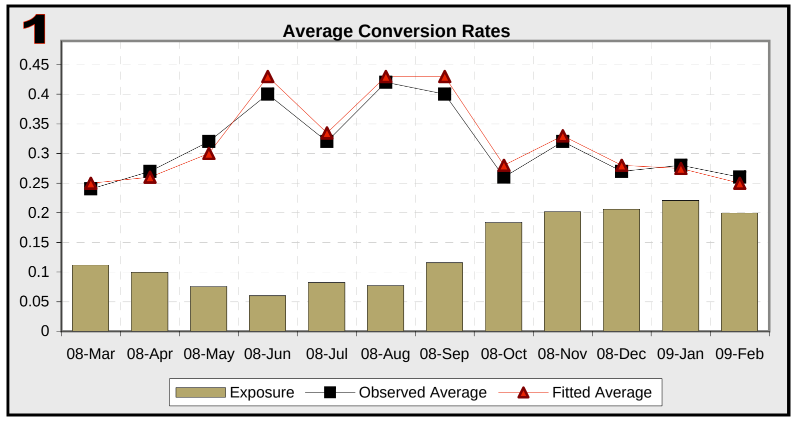
Model Importance

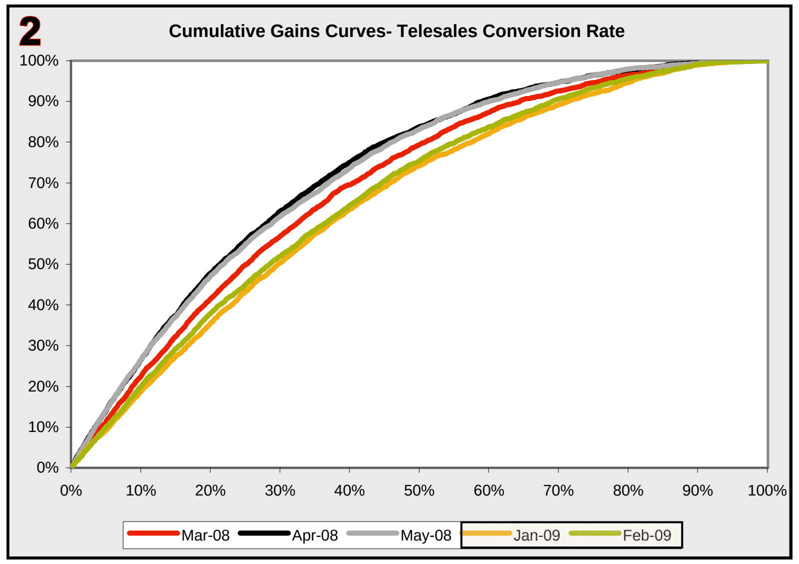
MAR 08 - MAY 08	Data Used in Model
JUNE 2008	Date Model Last Fit

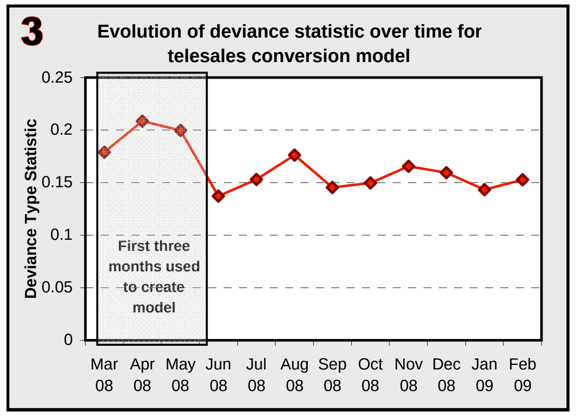
☒ 🟧 🟨 🟩 Current Performance

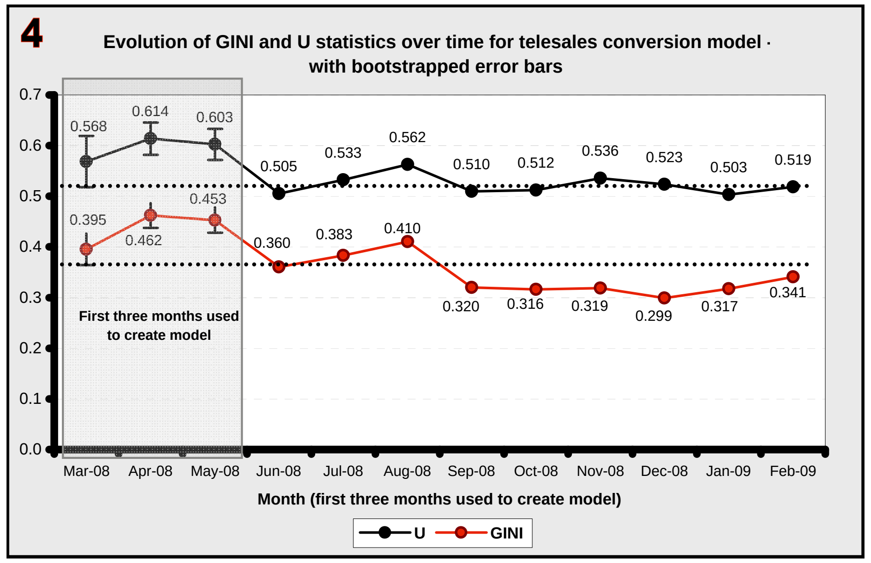
Model degradation has been gradual. Refit required.

Unique quotes (Year to Date):	111,595
Sales (Year to Date):	34,078
Average conversion rate:	0.3054


1 Average Conversion Rates


2 Cumulative Gains Curves- Telesales Conversion Rate


3 Evolution of deviance statistic over time for telesales conversion model


4 Evolution of GINI and U statistics over time for telesales conversion model - with bootstrapped error bars

5

Elasticity Chart goes here

This dashboard would normally also contain:
(1) a table/chart of monthly marketing spend
(2) a table describing rating changes and dates
(3) a table describing market events and dates
(4) a table describing product changes (including changes to upsell/add-on strategy)

Note that the dates vary between the conversion models - this is just for illustration - time periods would normally be consistent

1 Average conversion rates by month along with fitted average (modelled) conversion rates
2 Gains curves by month, only showing months used in model build and most recent months
3 Calculated as 1 - deviance(model) / deviance(null model)
4 See main body of report for explanation
5 Chart tracking modelled elasticity by month, with confidence intervals

FIRM LOGO

AD FREQUENCY

Model Importance

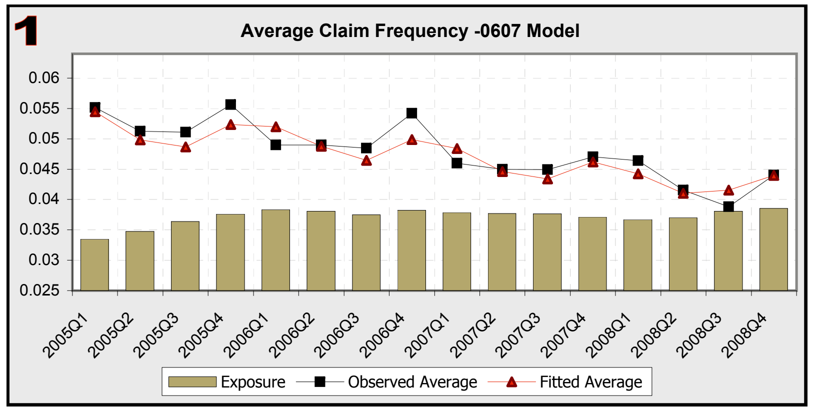
JAN 06 - DEC 07 Data Used in Model

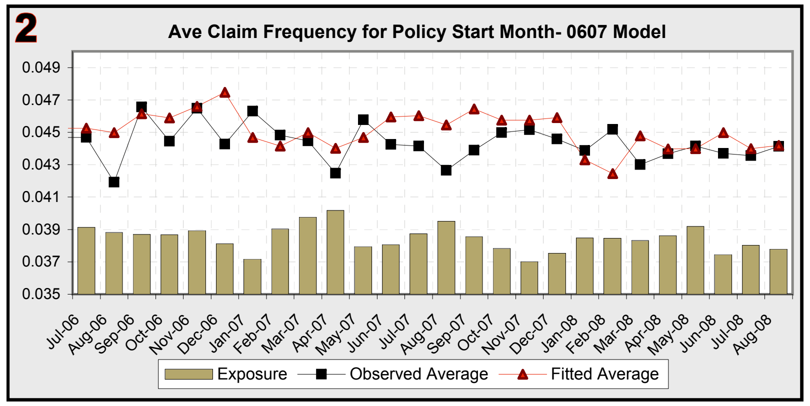
JUNE 2008 Date Model Last Fit

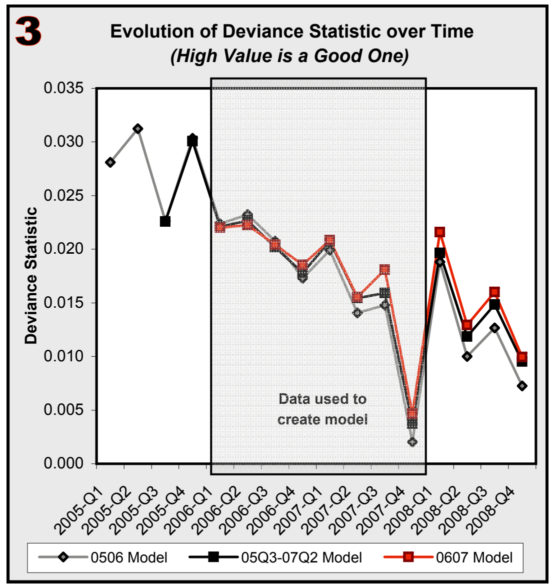
Current Performance

Model performance has improved in 2008 relative to 07Q4. 08Q4 is badly predicted by the model.

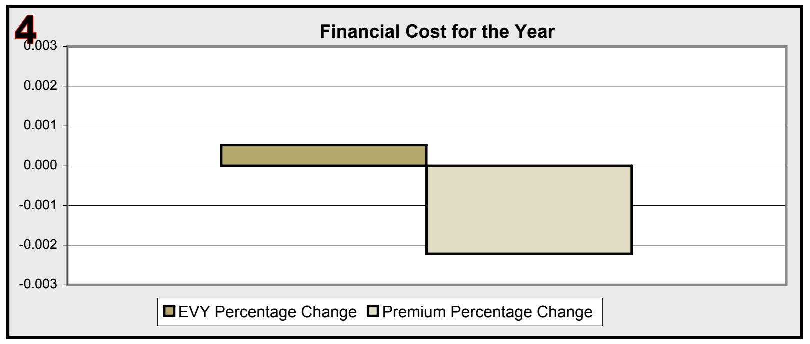
Claim Freq YTD	0.04512
Claim Freq (YTD last year)	0.04893


1 Average Claim Frequency -0607 Model


2 Ave Claim Frequency for Policy Start Month- 0607 Model


3 Evolution of Deviance Statistic over Time

(High Value is a Good One)


4 Financial Cost for the Year

AD SEVERITY

Model Importance

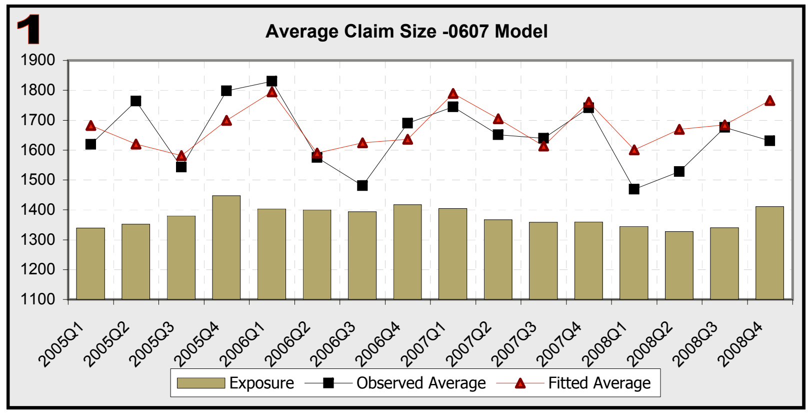
JAN 06 - DEC 07 Data Used in Model

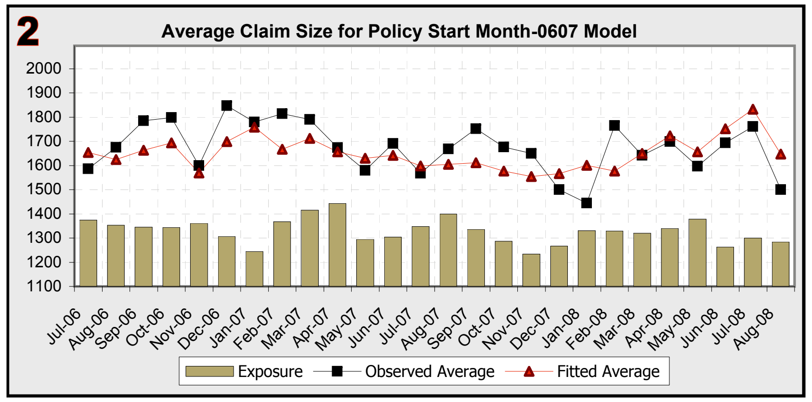
JUNE 2008 Date Model Last Fit

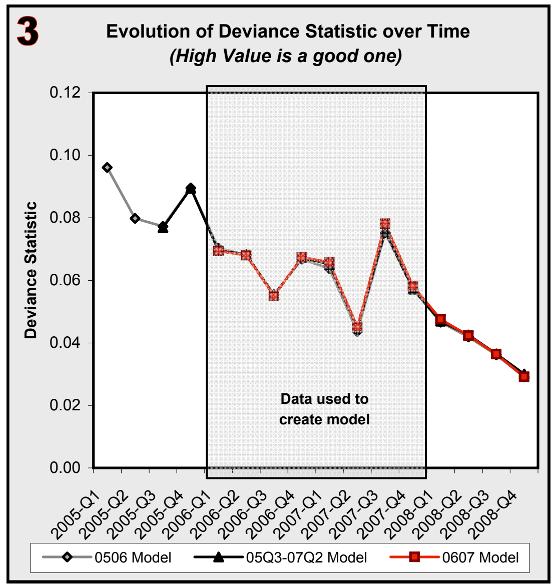
Current Performance

Gradual degradation in the model.

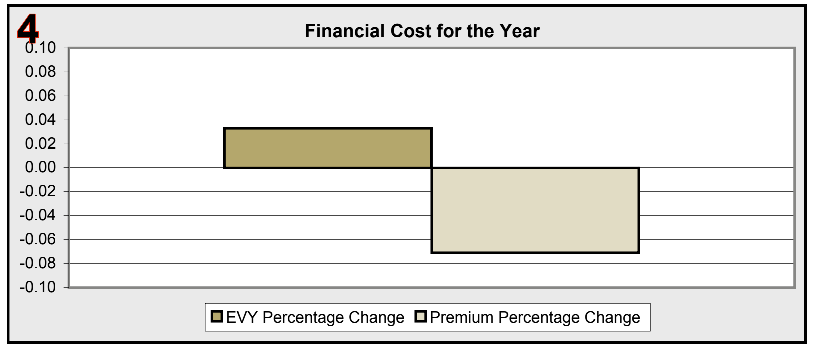
Claim Sev YTD	£1603
Claim Sev (YTD last year)	£1673

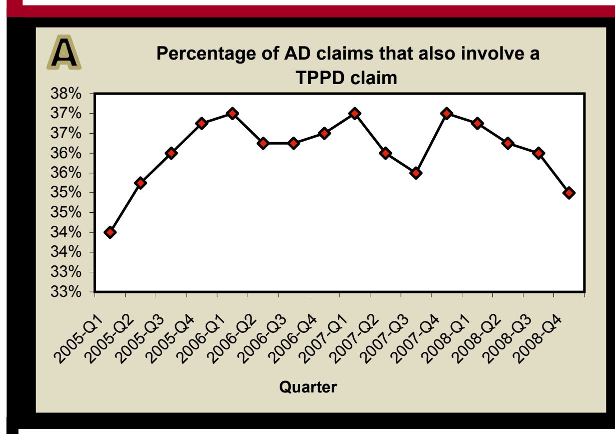

1 Average Claim Size -0607 Model

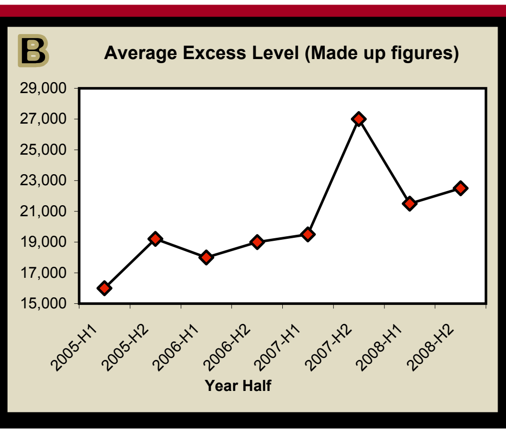

2 Average Claim Size for Policy Start Month-0607 Model

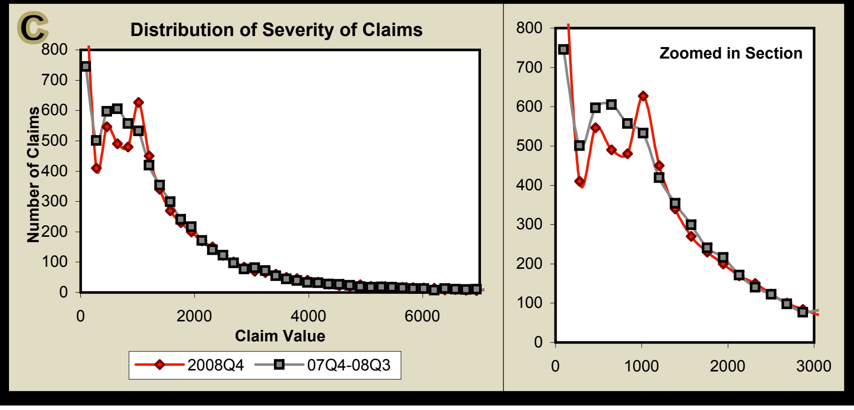

3 Evolution of Deviance Statistic over Time

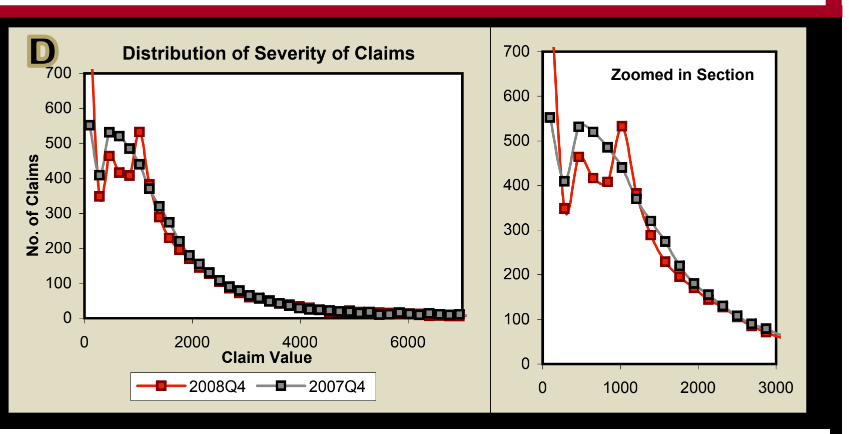
(High Value is a good one)


4 Financial Cost for the Year


A Percentage of AD claims that also involve a TPPD claim


B Average Excess Level (Made up figures)


C Distribution of Severity of Claims


D Distribution of Severity of Claims

A Dashboard should include some information on how mix of underlying claim types/events is evolving.

B Trends in excess will have a major effect on model performance

C Shifts in the severity distribution are a warning sign that both frequency and severity models could need updating. Chart shows data for latest quarter vs. rolling year.

D Same as chart C except this shows latest quarter vs. same quarter one year ago.

1 Average claim frequencies/severities, showing observed average and modelled (fitted) average trends

2 Same as 1 except uses underwriting month rather than accident quarter

3 Calculated as 1 - deviance(model)/deviance(null model)

4 See main body of report for explanation

FIRM LOGO

Appendix D: References

Duncan Anderson et al, *Report of the General Insurance Premium Rating Issues Working Party ("GRIP")* (GIRO, 2007, http://www.actuaries.org.uk/__data/assets/pdf_file/0008/20051/grip_report_jan07.pdf)

Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher and Neeza Thandi, *A Practitioner's Guide to Generalized Linear Models* (http://www.casact.org/library/studynotes/anderson9.pdf)

MJ Brockman and TS Wright, *Statistical Motor Rating: Making Effective Use of Your Data* (JIA 119III, http://www.actuaries.org.uk/files/pdf/library/JIA-119/0457-0543.pdf)

Richard Brookes, *"Individual" modeling –what, when and a little bit of how* (Institute of Actuaries of Australia, 2006, http://www.actuaries.asn.au/IAA/upload/Public/1520_Brookes_PPT.pdf)

A.C. Davison and D.V. Hinkley, *Bootstrap Methods and their Application* (Cambridge University Press, 1997)

James Guszcza, *Model Validation and Bootstrapping* (CAS Predictive Modeling Seminar, 2004, http://www.casact.org/education/specsem/f2004/handouts/guszcza1.ppt)

James Guszcza, *The Basics of Model Validation* (CAS Predictive Modeling Seminar, 2005, http://www.casact.org/education/specsem/f2005/handouts/validation.ppt)

James Guszcza, *Predictive Modelling for Commercial Insurance* (General Insurance Pricing Seminar, 2008, http://www.actuaries.org.uk/__data/assets/pdf_file/0010/133786/Guszcza.pdf)

David Hindley et al, *Report of the Effective Management Information Working Party* (GIRO, 1994, http://www.actuaries.org.uk/__data/assets/pdf_file/0010/26965/0001-0072.pdf)

David Isaacs and Chris Hope, *Actuarial Control Cycle in Pricing – Using Data Mining Techniques to Enhance Monitoring* (Institute of Actuaries of Australia, 2007, http://www.actuaries.asn.au/IAA/upload/public/GIPS%20paper%20David%20Isaacs%20&%20Chris%20Hope.doc.pdf)

Christopher Monsour, *Model Validation Techniques* (CAS Ratemaking and Product Management Seminar, 2009, http://www.casact.org/education/rpm/2009/handouts/monsour.pdf)

Christopher Monsour, *Validating Models* (CAS Special Interest Seminar on Predictive Modeling, 2006, http://www.casact.org/education/specsem/f2006/handouts/monsour.pdf)

Karl Murphy, Michael Brockman and Peter Lee, *Using Generalized Linear Models to Build Dynamic Pricing Systems* (Casualty Actuarial Society Forum, 2000 Winter, http://www.casact.org/pubs/forum/00wforum/00wf107.pdf)