Multimodel Inference for Reserving

Luke Nichols
Disclaimer

The material in this report is copyright of Luke Nichols.

The views and opinions expressed in this report are solely that of the author’s and do not reflect the views and opinions of the Australian Prudential Regulation Authority.

Any errors in this report are the responsibility of the author. The material in this report is copyright.

Other than for any use permitted under the Copyright Act 1968, all other rights are reserved and permission should be sought through the author prior to any reproduction.
Multimodel Inference for Reserving

Luke Nichols

Brian Gray Memorial Scholarship Report

December 2009

Abstract

Actuaries tend to choose a model for their loss process and project the future conditional on that model, when in fact there may be a number of reasonable models for the process. Model sensitivity is a topic widely considered by actuaries, and indeed forms part of APRA's GPS 310 and the IAAust's PS 300. Although model sensitivity is an important issue, measuring the degree to which a predictive distribution is affected by model choice is only one component in properly accounting for model selection uncertainty.

This paper describes how sensitivity and model adequacy combine into a single framework for inference involving multiple plausible models which can account for the contribution of model selection uncertainty to the predictive distribution. The ‘multimodel inference’ framework can be derived from the Bayesian, information-theoretic or frequentist paradigms, with a similar structure in each case.

The multimodel inference framework is then applied in four reserving examples to produce a model averaged predictive distribution in each case. The 18 models fitted include popular reserving models and an autoregressive model in the payment year direction. The data analysis illustrates that model averaging causes both the variance and the heaviness of the tail to be impacted. While the resulting uncertainties are still lower bounds, they are able to include much more of the underlying predictive uncertainty than the usual approach.

I would like to express my gratitude to my supervisor, Dr Glen Barnett, whose insight and enthusiasm in discussing ideas have definitely enhanced the quality of this research. I would also like to thank Mr Philip Bell for his invaluable advice and criticism of this paper and the Australian Prudential Regulatory Authority for financial support from the Brian Gray Scholarship.
## Contents

1 Introduction .................................................................................................................. 1

2 Literature Review .......................................................................................................... 1
   2.1 Regulatory Framework .............................................................................................. 3
      2.1.1 Outstanding Claims Liabilities ........................................................................... 3
      2.1.2 Prudential Standard GPS 310 (APRA) .............................................................. 3
      2.1.3 Professional Standard PS 300 (IAAust) ............................................................ 4
      2.1.4 Summary ............................................................................................................ 4

3 Model Selection ............................................................................................................ 4
   3.1 Akaike Information Criterion .................................................................................. 5
   3.2 Bayesian Information Criterion ............................................................................... 7

4 Components of Model Error .......................................................................................... 8
   4.1 Model selection uncertainty ..................................................................................... 8
   4.2 Model specification uncertainty ................................................................................. 9
   4.3 Model sensitivity ....................................................................................................... 9
   4.4 Model uncertainty and actuarial judgement .............................................................. 10
   4.5 The problem with sensitivity analysis ...................................................................... 11

5 Multimodel Inference .................................................................................................... 11
   5.1 Model Averaging ..................................................................................................... 12
   5.2 The Bayesian Framework for Multimodel Inference .............................................. 13
      5.2.1 Posterior model probabilities ........................................................................... 14
   5.3 The Information Theoretic Framework for Multimodel Inference ......................... 15
      5.3.1 Akaike weights ................................................................................................. 16
   5.4 The Frequentist Framework for Multimodel Inference .......................................... 17
   5.5 Collinearity or similar models ................................................................................. 18

6 Application to reserving ................................................................................................ 18
   6.1 Data Sets ................................................................................................................ 19
   6.2 Models .................................................................................................................... 19
   6.3 Model fit and diagnostics ......................................................................................... 21

7 Multimodel Inference Data Analysis ........................................................................... 22
   7.1 Mortgage Data ....................................................................................................... 22
   7.2 GenIns Data .......................................................................................................... 24
   7.3 ABC Data .............................................................................................................. 25
   7.4 CTP Data .............................................................................................................. 26
   7.5 Comments on data analysis ................................................................................. 29

8 Conclusion ..................................................................................................................... 29

References ......................................................................................................................... 31
1 Introduction

Model selection is generally acknowledged as fundamental to good inference and prediction; however uncertainty about the selected model has rarely been fully incorporated into measures of precision. For example, an actuary usually has a number of sensible models for a loss process, and will carefully select the one “best” model from the model set to project the future, assuming that the chosen model is correct. These predictions may be presented with measures of ‘precision’ that ignore the impact of model selection even when it is a potentially important component of variability.

This paper focuses on model risk and how this risk can be addressed. Model risk may be broken down into two main components: model selection risk and model specification risk. Model selection risk is the risk of choosing the wrong model within the set of candidate models; for example, choosing an incorrect set of predictors in a regression model. Model specification risk, on the other hand, is the risk that the best model is not included in the set of candidate models. This risk may include the risk that the model changes through time, the risk that unused variables are useful, and the risk that the best model is not even considered in the set of possible models. Model specification risk also includes the risk of choosing a model class which is not best; for example, using a linear model for a distinctly non-linear process. Some model specification risk can be transferred to model selection risk by using flexible models or a wider model class.

One way that model selection risk can be formally accounted for is by using multimodel inference. Multimodel inference provides a framework for inferences to be made that are robust to model selection effects, based on several plausible, competing models. The contribution of each model depends on the extent to which the model is supported by the data. This framework allows model selection risk to be accounted for, while having a consistent structure for the frequentist, information theoretic or Bayesian view.

The major interest of this research is the application of multimodel inference to reserving: setting provisions for a general insurance company’s outstanding claims liabilities. Four examples are discussed which give evidence for the use of multimodel inference, especially in the case of multiple plausible models.

2 Literature Review

There is a substantial amount of literature on the subject of model selection, but until recently, it has mostly been focused on how to select the single best model (Barron, Rissanen and Yu, 1998; Berger and Pericchi, 1996 and 1998; Myrvold and Harper, 2002; Rao et al., 2001). Traditionally, hypothesis testing methodology has been used for model selection; however there
is little theoretical justification for this approach (Akaike, 1974; Burnham and Anderson, 2002). Groundbreaking papers by Akaike (1974) and Schwarz (1978) developed separate “mathematical formulae of the principle of parsimony in model building” (Schwarz, 1978); introducing the “minimum information theoretical criterion,” later referred to as Akaike’s information criterion (AIC), and the Bayesian information criterion (BIC) respectively. This changed the way model selection was done by eliminating much of the subjective judgement previously required and outlining a simple, objective measure to choose the best model. While the AIC and BIC led the way in model selection, seemingly little thought or research was spent on understanding and addressing model selection risk until recently. Chatfield (1995) notes: “it is indeed strange that we often admit model uncertainty by searching for a best model but then ignoring this uncertainty by making inferences and predictions as if certain that the best fitting model is actually true.”

In recent times, model uncertainty has been an active area of research, promoted strongly by several excellent papers: Chatfield (1995), Draper et al. (1987) and Draper (1995) (also see Hodges, 1987). Chatfield (1995) places emphasis on the whole model building process, including model specification, model estimation, model validation and meta-analysis. His widened viewpoint leads to analysing the sources of model biases and uncertainties which arise as a consequence of the entire model building process and discussing ways to circumvent or reduce the biases and uncertainties, including using Bayesian model averaging. Chatfield points out a number of issues which arise as a result of formulating and fitting a model using the same data: 1) least squares theory is not valid; 2) parameter estimates and residual variance may be biased after model selection; and 3) analysts tend to be too optimistic about the model fit. 4) Prediction intervals are usually too narrow. Draper et al. (1987) use the example of forecasting oil prices in the 1980s is used to highlight the types of uncertainty that were not accounted for by the forecasters. Building on the idea of an unconditional forecast obtained by “integrating out the model” introduced by Draper (1987), Draper et al. (1987) propose new techniques to implement a weighted average prediction taking into account the usefulness of each model using a ‘prior’ and ‘likelihood’ of each model. They describe this weighting method as a logical extension of sensitivity analysis. Following this, Draper (1995) makes four main arguments: 1) model uncertainty needs to be dealt with; 2) the number of possible models must be kept small, otherwise computation time is an issue; 3) the possible models should be weighted by probability; and 4) the Bayesian methodology works well to take account of model uncertainty. Buckland et al. (1997) argue that “model selection uncertainty should be fully incorporated into statistical inference whenever estimation is sensitive to model choice and that choice is made with reference to the data.”
Only a few papers address this in the context of actuarial work. A recent paper by Peters, Shevchenko and Wüthrich (2009) considers model uncertainty in general insurance claims reserving for a single family of reserving models: Tweedie’s compound Poisson models (Tweedie, 1984).

2.1 Regulatory Framework
Insurance companies sell promises to make a future payment contingent on future events. In exchange for this possible future payment, customers pay premiums now. Clearly for the system to work, insurers must hold enough capital to be reasonably certain that they can fulfil the promises they have made to customers. The Australian Prudential Regulation Authority (APRA), as the prudential regulator of insurance companies, requires general insurance companies to set aside reserves for outstanding claims and unearned premiums. Outstanding Claims Liabilities (reserves) are described as “claims incurred prior to the calculation data, whether or not they have been reported to the insurer” (GPS 310, 72), while Premiums liabilities are future claims arising from post-valuation date events which are insured under existing, unexpired policies (GPS 310, 73). Unearned premiums calculations often involve only simple adjustments to outstanding claims, and are not considered further in this paper.

2.1.1 Outstanding Claims Liabilities
The outstanding claims liabilities of an insurance company at any given time are uncertain. Part of the actuary’s role is to estimate the expected outstanding claims liabilities and to advise on liability provisions. The estimation of liabilities is a probabilistic problem; we are trying to forecast the distribution of outstanding claims. The forecast distribution - which incorporates uncertainties in addition to process uncertainty (e.g. parameter uncertainty) - is called a predictive distribution. Clearly, the predictive distribution is dependent on the model, and different models can not be expected to yield the same predictive distributions or reserves.

The outstanding claims liability directly impacts the minimum capital requirement, through the Insurance Risk Capital Charge in the standard method (GPS 115, 2008), and the 0.5% probability of default requirement in the internal model method (GPS 113, 2009). The impact of even a 5% deviation of the outstanding claims liability can have a large effect on the capital requirement. Therefore it is very important for the insurer that the outstanding claims liability is estimated accurately, allowing for all material uncertainties. Also, the method of calculation must be specific; otherwise companies may be able to manipulate their capital requirement.

2.1.2 Prudential Standard GPS 310 (APRA)
APRA’s GPS 310 (2008) states that the value of insurance liabilities should be the unbiased central estimate plus a risk margin reflecting the uncertainty of the central estimate (e.g. model
selection uncertainty). The risk margin required in Australia is such that, with respect to the predictive distribution of outstanding claims, the outstanding claims reserve is set at the greater of the 75th percentile and the mean plus half the standard deviation. In many cases, both of these requirements are difficult to quantify using standard techniques. Such estimation will require “both the exercise of judgement and technical analysis” (GPS 310, p.26). Judgement is involved in determining the impact on the risk margin of “the robustness of the valuation method, the reliability and volume of available data, past experience of the insurer and the characteristics of the class of business” (GPS 310, p.25), while the technical analysis should involve “comprehensive actuarial analysis and modelling techniques, subject to materiality” (GPS 310, p.27) in order to understand and communicate the key drivers of uncertainty and its consequences. The Insurance Liability Valuation Report should include ‘... an indication of the uncertainty' in the central estimates, including statistics such as the standard deviation; the sensitivity analyses undertaken; …’ (GPS 31, p.30)

2.1.3 Professional Standard PS 300 (IAAust)
In addition to following APRA’s GPS 310, actuaries must have proper regard for the IAAust’s Code of Professional Conduct and relevant professional standards, such as PS 300 (2008). PS 300 takes a more technical focus on the principles and directions which must be followed in General Insurance liabilities valuations. As with GPS 310, allowing for uncertainty is a key focus, and the actuary is expected to consider and document material risks and uncertainty and their implications (PS 300, p.20). The actuary must provide a quantitative indication of uncertainty which “can be achieved by examining scenario analyses, sensitivity analyses and/or statistics such as the estimated standard deviation of any assumed probability distribution of claim cost outcomes” (PS 300, p.18, p.20).

2.1.4 Summary
The prudential and professional standards clearly intend actuaries to make allowance for all material risks, including model selection risk. Both PS 300 and GPS 310 require the use of sensitivity analyses and scenario testing. Although sensitivity analysis is useful, this approach can cause considerable over or under-estimation of actual predictive uncertainty; it is only one component in allowing for model selection uncertainty.

3 Model Selection
Actuaries are regularly faced with the fundamental problem of model selection. Given some data, there are often a large number of reasonable models. There will be models that fit the data well overall or fit sections of the data and others which provide insight into certain aspects of
the data. Model selection is deciding which model of a set of models is preferable, with the available given data. It involves choosing a model with optimal complexity in order to make the most valuable inferences and/or predictions.

The process of arriving at a set of candidate models must firstly include much careful thought about the actual problem, analysis theory and the data (see Burnham and Anderson, 2002; Chatfield, 1995; Leamer, 1978). For example, models under consideration must be reasonable, and model complexity must be related to the information available. Unfortunately, in practice, emphasis is often on the number crunching without regard to the actual problem or the reason and assumptions behind the models.

The choice of model is intrinsically linked with the choice of model selection criterion. Thus model selection criteria should select the best model through an objective and repeatable process, based on either the statistical concept of parsimony or shrinkage (smoothing) methods (some criteria are parsimonious as a consequence of their derivation). Parsimony is the concept that a model with fewer parameters is preferred over a model with more parameters of similar fit, while shrinkage methods downplay variables lacking evidence from the data – they are effectively a continuous model selection approach. Parameter estimates and predictions should be accurate and unbiased, and the model should approximate reality sufficiently well for the model’s purposes. A model with too many parameters produces parameter estimates with large variances since there is not enough data relative to parameters. On the other hand, a model with too few parameters may be biased and not realistic enough to make useful inferences or predictions (e.g. fitting a line to a process with distinct curvature). Model selection criteria should be well-justified with a deep level of theoretical support within a general statistical inference framework (e.g. Bayesian) be easy to use, and work well for realistic sample sizes (Pötscher, 1991). Finally, for each model under consideration, the criteria should be computable from the data, and condense to a single number from which the models can be ranked. In multimodel inference, these numbers are used to calculate model weights which allow an assessment of model uncertainty.

3.1 **Akaike Information Criterion**

Akaike (1973) proposed the use of Kullback-Leibler (K-L) information as a fundamental basis for model selection, considering the loss of information due to using the specified model to approximate reality. The K-L information (Kullback and Leibler, 1951) measures the loss of information by using model $g$ to approximate reality $f$. It can be thought of as the “distance between two models.” The K-L distance has strong links to Boltzmann’s entropy:
K-L distance = \( I(f, g) = E_f (-\text{Boltzmann’s entropy}) \)

\[
= \int \log \left( \frac{f(x)}{g(x)} \right) \cdot f(x) \, dx
\]

(Boltzmann, 1877 in Akaike, 1985; Burnham and Anderson, 2002).

Thus minimising K-L information is the same as maximising expected entropy. After some simple manipulation, the K-L information may be written in an easily interpretable form:

\[
I(f, g) = \int f(x) \log (f(x)) \, dx - \int f(x) \log \left( \frac{g(x)}{f(x)} \right) \, dx
\]

\[
= K - E_f \left[ \log \left( \frac{g(x)}{f(x)} \right) \right]
\]

where K is a constant.

Without knowing K we don’t know the absolute value of the K-L distance, but relative differences of the K-L information between models \( g_1 \) and \( g_2 \) can be calculated (subtraction causes K to disappear) and will estimate (parameters are generally unknown) which model loses the least information, and thus which is a better approximation to reality. Through the 1970s, Akaike (1973, 1974, and 1978) progressively introduced his entropy maximisation based model selection criterion. The “minimum information theoretical criterion” (Akaike, 1973), later known as the Akaike information criterion (AIC), approximates

\[
\hat{AIC} = -2 \log L(\hat{\theta}; x, M_i) + 2k
\]

The first term \(-2 \log L(\hat{\theta}; x, M_i)\) measures the lack of fit of the model and will be smaller for better fitting models; while the second term \(2k\) enforces parsimony by penalising the number of parameters in the model by increasing AIC by two per parameter. The model with the lowest AIC should be chosen as the best model for inference.

Sugiura (1978) proposes the AIC\( C \) to correct bias in the AIC where there is a small sample size relative to the number of parameters that are fitted.

\[
AIC_C = -2 \log L(\hat{\theta}; x, M_i) + 2k + \frac{2k(k+1)}{n-k-1}
\]

Hurvich and Tsai (1989) find the AIC\( C \) superior to AIC in time series applications. AIC\( C \) selected the correct model more often, and when incorrect, the model was generally much more parsimonious than that selected by AIC. Burnham and Anderson (2004) recommend that the AIC\( C \) be employed regardless of sample size. Other improvements and generalisations of the AIC have been derived, such as the AIC\( T \) (or TIC) of Takeuchi (1976), which generalises the AIC’s penalty term (see Rao et al., 2001). This paper uses the AIC\( C \) instead of the AIC.

Many researchers have tested and commented on the usefulness of the AIC. The derivation based on Kullback-Leibler distance doesn’t assume full truth is in the candidate model set. The AIC is a criterion for the best model fit, rather than one which estimates the true model order. It tends to overestimate the true model order, and thus it may select an over-fitted model, though Burnham and Anderson (2002) point out that the AIC is unlikely to select a seriously overfit model. Still, several other authors have questioned whether the \( 2k \) penalty is enough to prevent
an overly complex model from being chosen (Rissanen, 1976; Bozdogan, 2000). This must be considered when using the AIC for predicting future insurance claims.

### 3.2 Bayesian Information Criterion

The Bayesian information criterion (BIC) (Schwarz, 1978) is of a similar form to AIC, but the assumptions underlying the derivation differ. Under a particular set of assumptions, the BIC is an asymptotic approximation to \(-2\) times the log of the posterior model probability (see Kass and Raftery, 1995).

\[
\hat{\text{BIC}}_i = -2 \log \hat{L}(\hat{\theta}; x, M_i) + k \log(n)
\]

As with the AIC, the first term \(-2 \log \hat{L}(\hat{\theta}; x, M_i)\) measures the lack of fit of the model, being smaller for better fitting models; while the second term \(k \log(n)\) enforces parsimony by penalising the number of parameters in the model by increasing BIC by a factor depending on the number of observations. For \(n \geq 8\), this penalty will be higher than that of the AIC, and it increases with sample size. Hence in practical situations (reasonable size \(n\)), the BIC is likely to choose a simpler model than the AIC.

![Figure 4.1](image)

**Figure 4.1**

AIC and BIC penalty crossover points. The drawn lines are where the penalty terms cross over, hence to the left of the lines the AIC penalty is heavier, while to the right of the lines, the BIC penalty is heavier.

Like the AIC, much research has been conducted on the BIC. The main assumption difference between the AIC and BIC is that the BIC assumes that a true model exists (and generated the data) and is in the model set. Then the BIC estimates the dimension of this true model, and many researches have shown that it works well in this situation. The strong penalty on complexity often results in an under-fit model, leading to biased parameter estimates and overestimated precision estimates (Burnham and Anderson, 2002). The divergence in penalty terms for large samples means that the BIC and AIC usually result in very different models being chosen, and this is still a major problem in model selection. Although AIC and BIC are the most popular model selection criteria, there are many other possible criteria. Hurvich and Tsai (1998) provide a comparison and overview of many popular selection criteria.
4 Components of Model Error

Statistical analysis broadly involves three stages: model building (which includes model specification, model fitting, model validation and the combination of data from multiple sources), model selection and forecasting. Each of these tasks introduces model biases and uncertainties, which should not be considered separately. Ad hoc adjustments to predictions and measures of precision are sometimes made to reflect such biases and uncertainties, but ideally they should be allowed for within the statistical analysis framework.

Making inferences as if a model is true when it has been selected from the same data to be used for model estimation is ‘well known’ to be ‘logically unsound and practically misleading’ (Zhang, 1992). However, the impact of formulating and then fitting a model to the same data is not well understood (Chatfield, 1995; Burnham and Anderson, 2002), although it often leads to an overfit model, because residual variation is interpreted as model structure. Chatfield (1995) labels the resulting biases model selection biases. As the models under consideration grow larger, the bias is very relevant. Two problems expressed by Zucchini (2000) are 1) With many models in the candidate set, it is very likely that the ‘best’ model is chosen because of ‘lucky’ random variation rather than merit; and 2) The selected model generally fits the data too well, leading to underestimates of measures of variance including predictive variance.

Throwing everything in the pot adds to the first bias. Although it can not be eliminated completely, it may be reduced through careful specification of the models under consideration and a priori model formulation (Anderson and Burnham, 2002). The second problem can be fully allowed for within the multimodel inference framework. The ramifications of using the same data set for model selection and inference has been an area of much research for many years. Burnham and Anderson (1998, 2002 and 2004), Chatfield (1995) and Leamer (1978) provide a detailed discussion of the issues arising from the common practice of data dredging: analysing the data and fitting a model by searching the data for trends and “significance.”

4.1 Model selection uncertainty

Often there is considerable uncertainty in choosing a certain model as the best approximating model. If another independent set of data were available, maybe a different model would be chosen. This is one component of model selection uncertainty. Another component is that certain approximating models may be good to describe different parts of reality. Consider an estimator \( \hat{\theta} \). The variance of \( \hat{\theta} \) should have two parts: 1) \( \text{var}(\hat{\theta} | M) \) (sampling variance given the model); and 2) \( \text{var}(\hat{\theta}) - \text{var}(\hat{\theta} | M) \) (variance due to uncertainty about the model). In other
words, measures of unconditional precision are needed, rather than the usual measures which are conditional on the model.

Model selection uncertainty can have a large effect where the single “best” model is only marginally better than some alternatives, and these alternatives give different predictions or inferences. If this occurs, it should at least be reported. Since the statistical analysis undertaken depends on model selection, inferences should incorporate model selection uncertainty; that is, the variance component due to model uncertainty should be included in estimates of precision. Where model selection uncertainty is not allowed for, estimated sampling variances and covariances are too low, hence confidence interval coverage is below the nominal level (Burnham and Anderson, 2002). Such an underestimation of the variance within the general insurance context may lead to under-reserving and under-pricing, which could potentially place an insurer in financial difficulty in the future. Clearly model selection uncertainty should be taken into account within the estimates. Burnham and Anderson (2002, 2004), Chatfield (1995), Draper (1995) and Hurvich and Tsai (1990) provide good descriptions of model selection uncertainty.

4.2 Model specification uncertainty

Model specification uncertainty refers to uncertainty about the form of models within the candidate set. There are a number of risks contributing to model specification uncertainty, such as parameter evolution risk; the risk of hidden variables and the risk that past patterns do not continue into the future. However, the main component is whether the best approximating model is in the set of possible models. Since model specification uncertainty contributes risk to inferences and predictions, it is prudent to allow for specification uncertainty as much as possible. This can be done within the multimodel inference framework.

In order to reduce specification uncertainty, careful thought is needed about the problem in choosing the candidate models. By considering the likely ways that the candidate models could be wrong and expanding the model class, some specification uncertainty can be “transferred” to model selection uncertainty, which is fully accounted for within the multimodel inference framework. Note that this should be done thoughtfully, since widening the model class contributes to selection bias.

4.3 Model sensitivity

Model sensitivity is a topic widely considered by actuaries and indeed forms part of the APRA’s GPS 310 and the IAAust PS 300 (see Section 2.1). Both GPS 310 and PS 300 recommend the use of sensitivity analyses and scenario testing as an indication of uncertainty in the central estimates of the value of insurance liabilities. In actuarial work, a sensitivity analysis
can be defined in a number of ways. It can refer to the process of running a model using different assumptions to check which assumptions are most relevant to the model (Bellis, Shepherd and Lyon, 2003). More generally, a sensitivity analysis is the process of running projections or making inferences on each plausible model given the data. Actuarial analysis generally involves choosing a model, and then making predictions and inferences based on the model. In sensitivity analysis, each model in turn is assumed to be the selected model, and projections are made conditionally on that model. The idea is to give the actuary some idea about how much the final reserve varies among models.

Some regulators prescribe a set of adverse “what if” scenarios to assess the capital adequacy of a company. In this case, the analysis is called scenario testing instead of a sensitivity analysis. While a sensitivity analysis tends to focus more on results from plausible models, scenario testing generally refers to analysis done on models which project higher future claims. For the purposes of this paper, differentiation between sensitivity analysis and scenario testing is not important, thus “sensitivity analysis” will be used to refer to any such testing or analysis.

4.4 Model uncertainty and actuarial judgement

Bardis, Majidi and Murphy (2009) present a chain ladder model which introduces more statistical theory in order to better understand the approach and appropriateness of the technique. They point out that actuaries tend to select link ratios using judgement rather than following the chain ladder method, since trends in claims triangles can often be recognised. Such flexibility is certainly practical and logical, however even when the actuary is experienced, their judgement can be called into question. This is indeed the problem posed by subjective model selection techniques. Thus the choice of link ratio can be viewed as an example of model selection.

A recent study by Blumsohn and Laufer (2009) asked 52 actuaries to select loss development factors for the same claims triangle. The ratios selected varied greatly – the resulting reserves ranged from $10.7m to $60.2m with a coefficient of variation of 30% and this ignores the tail factor! Blumsohn and Laufer (2009) comment that ‘In a world which many employers, regulators, auditors, and investors seem to think actuaries can make loss picks that are within 10% of the “truth,” this should be sobering.’ Given that ratio selection is a type of model selection, this highlights the need for an objective, statistical justification for the model selected. Clearly, relying entirely on actuarial judgement to make such material decisions can be problematic. This brings into question whether current regulation is enough – certainly the use of methods that minimise this subjectivity would be very desirable.
4.5 The problem with sensitivity analysis

Sensitivity analysis explicitly tries different models, in order to give a crude allowance for model uncertainty. Although sensitivity analysis does give a measure of the impact of a particular scenario, it does not give any indication of the likelihood of the event occurring – sensitivity testing cannot even rank events from most likely to least likely. This makes it difficult to quantify variance across models and confidence in the chosen model (or equivalently, the base valuation).

In other words, sensitivity analysis does not formally consider a material and uncertain component of risk assessment: the relevance of the model. The probability that the model is correct should reflect the credibility and plausibility of the model. The most obvious candidate to include in the weighting is how well the model fits the data. By incorporating fit to the data along with actuarial judgement, the resulting model probability provides a well defined weight of evidence for each model.

5 Multimodel Inference

Multimodel inference is inference based on all plausible models. It has been described as “an extension of sensitivity analysis, in which predictions and uncertainty assessments from some or all of the separate model-scenario combinations examined in the sensitivity analysis are combined quantitatively” (Draper et al., 1987). A number of researchers have found that multimodel inference is better than model selection in many situations (e.g. Raftery, Madigan and Hoeting, 1997; Burnham and Anderson, 2002; Yang, 2003). One of the reasons for this is that the multimodel inference framework can take account of model risks, such as model selection risk and model specification risk. The formal framework has existed for a number of years, and is used in many non-actuarial fields.

Consider an example, in which we have 7 models: $M_1, \ldots, M_7$; this is analogous to a sensitivity analysis type problem. Under each model, the expected predictions given the data can be calculated. The following graph shows these $E(y_f | y, M_i)$ for each model.

![Graph showing expected predictions for seven models](image)

Figure 5.1: Expected predictions given the data for the seven models under consideration. The lines of equal height reflect the probabilities of the models being viewed as equal.

But obviously these models are not equally good at predicting future claims. What is missing are weights according to the credibility of the model.
Sensitivity analysis ignores the probability of event occurrence. It focuses on the outcome, rather than the risk, leaving the rest to the actuary’s judgement. For example, an event which could cause an insurer a loss of $1 trillion with probability 0.00000001 would generally be regarded as less risky than a $1 billion event with probability 0.1. The outcome alone cannot determine the risk, rather a better measure of risk should be of the form: risk = outcome \times probability of occurrence.

The widespread use of sensitivity analysis among actuaries is designed to quantify risks such as model risk, yet the relevance of results is left entirely up to actuarial judgement. The problem with actuarial judgement is that it is subjective: for example, different actuaries under the same conditions could potentially choose very different liability reserves. The inconsistent nature of actuarial judgement is obvious and material in the Blumsohn and Laufer (2009) study described in Section 4.4. Clearly a more robust framework is needed, which takes proper account of the contribution of model selection to predictive uncertainty.

5.1 Model Averaging

The information in the data can often be described well by several models and the data does not support the selection of only one. In this case, some form of multimodel inference (e.g. model averaging) should be applied (Burnham and Anderson, 2002). There are two forms of model averaging: averaging the estimate of an unknown quantity (which is common across models); and averaging the predictive distributions of the models.

Consider an unknown quantity of interest \( Q \) which is common across the candidate models. \( Q \) may be a parameter \( \theta \) or a future observation \( Y \). Under each model, \( Q \) can be estimated given some data. An overall average estimate of \( Q \) can then be calculated by weighting each estimate according to the plausibility of the model which generated it, \( p_i \) \( E(Q|x) = \sum_{i=1}^{K} E(Q|M_i, x) p_i \). The variance of the model average estimate is \( \text{var}(Q|x) = \sum_{i=1}^{K} [\text{var}(Q|M_i, x) + (E(Q|M_i, x))^2] p_i - E(Q|x)^2 \) where \( E(Q|M_i, x) \) and \( \text{var}(Q|M_i, x) \) are the mean and variance of the parameter \( Q \) under model \( i \), given the data. The estimate of \( Q \) will not be conditional on one single model, thus model selection bias is minimised. Furthermore, the corresponding unconditional variance fully incorporates model selection uncertainty and can partly incorporate model specification.
uncertainty. The equations above assume the Bayesian methodology, in which all unknown quantities can be thought of as random variables. Model averaging has a similar form under the information theoretic and frequentist methodologies (replace $Q$ by $\hat{Q}$).

This simple argument can easily be extended. Instead of considering the mean of $Y$ under each model, consider the entire distribution. Under each model, the height of the density function at each point can be weighted, leading to the “unconditional” predictive distribution of $Y$:

$$f(Y \mid x) = \sum_{i=1}^{K} f(Y \mid M_i, x) p_i.$$

The resulting distribution is not dependent on any one model, thus model selection uncertainty and to some extent specification uncertainty are included in the measures of precision. Model selection bias is also reduced (Zucchini, 2000). It is this model averaged distribution from which predictions should be made, and variability of predictions measured.

### 5.2 The Bayesian Framework for Multimodel Inference

There are several ways to obtain the posterior model probabilities ($p_i$) within the Bayesian framework: 1) the direct calculation of Bayes factor integrals; 2) MCMC to estimate the posterior distribution; or 3) the BIC (or other approach) approximation to the Bayes factor.

Under the Bayesian methodology, all parameters can be thought of as random variables. Thus once a joint distribution of parameters and data can be written down, unwanted parameters can be integrated out to provide the marginal distribution of the quantity of interest, be it an unknown parameter or a forecast future observation. For example, in model selection, the quantity of interest is the probability of obtaining the data under each model. Bayes factors (Kass and Raftery, 1995; Wasserman, 2000) represent the strength of evidence for a certain model compared to another model. Kass and Raftery (1995) provide some guidelines for interpretation of Bayes factor values. The Bayes factor for model $i$ versus model $j$ is:

$$B_{ij} = \frac{g(x \mid M_i)}{g(x \mid M_j)} = \frac{\int f(x \mid \theta_i, M_i) g(\theta_i \mid M_i) d\theta_i}{\int f(x \mid \theta_j, M_j) g(\theta_j \mid M_j) d\theta_j},$$

where $f(x \mid \theta_i, M_i)$ is the pdf of the data under model $i$ (or Likelihood function), and $g(\theta_i \mid M_i)$ is the prior on the parameters of model $i$. Clearly, setting up priors on every parameter of every model is not feasible, so an automatic, objective method of model selection is needed. If one wishes to use prior information, it may be reasonable to set priors on the top few models, and use an objective method for the rest of the model set.

The integrals on the right are in most cases intractable (an exception being exponential family with conjugate priors), so numerical methods or an approximation must be used. Evans and Swartz (1995) review numerical methods to solve such integrals. Markov chain Monte Carlo
(MCMC) can also be used to solve the integrals. One simple approach to MCMC is the Metropolis-Hastings algorithm (Hastings, 1970).

Numerical methods and techniques such as MCMC usually require much effort to implement successfully. However, the BIC (Section 4.2) provides a simple approximation which is automatic, as it does not require priors to be specified. By Laplace’s method, the quantity \( S = \frac{\text{BIC}_i - \text{BIC}_j}{2} \) may be used as an approximation to the log of the Bayes factor (see Kass, Tierney and Kedane, 1990; and Kass and Raftery, 1995). The BIC approach is “intuitively reasonable” (Kass and Wasserman, 1995) and “should provide a reasonable indication of the evidence,” as long as the degrees of freedom is small relative to sample size (Kass and Raftery, 1995). It is also well defined for non-nested models. Furthermore, the BIC can be derived from frequentist principles (Burnham and Anderson, 2002).

5.2.1 Posterior model probabilities

The Bayesian framework allows posterior probabilities of all models under consideration to be calculated. Posterior model probabilities represent a weight of evidence for each model. By Bayes’ theorem, the posterior probability of the \( i^{th} \) model is

\[
P(M_i | x) = \frac{P(x | M_i) P(M_i)}{\sum_{j=1}^{R} P(x | M_j) P(M_j)}
\]

Posterior model probabilities also follow directly from Bayes factors. Let \( M_0, M_1, ..., M_K \) be models under consideration, let \( \alpha_i \) be the prior odds of \( M_i \) to \( M_0 \), and let \( B_{ij}, i = 1, ..., K \), be the Bayes factors for \( M_i \) versus \( M_0 \). Then the posterior odds of each model are \( \alpha_i B_{ij} \). Dividing through by the sum of all the odds leads to a probability and rescales so that the sum of posterior model probabilities is 1. That is, \( P(M_i | x) = \alpha_i B_{i0} \sqrt{\sum_{j=1}^{K} \alpha_j B_{j0}} \). This equation can be used to find the posterior model probabilities from calculating Bayes factors by direct calculation or numerical methods. The data analysis in this paper uses the BIC approximation. Let \( \Delta_i = \text{BIC}_i - \text{BIC}_0 \) denote the BIC difference, and let \( p_i \) denote the posterior model probability, then using the quantity \( S = -\left( \frac{\text{BIC}_i - \text{BIC}_j}{2} \right) \) to approximate the log of the Bayes factor, the posterior model probability may be written as:

\[
p_i = \alpha_i e^{\frac{\Delta_i}{2}} \left( \sum_{j=1}^{K} \alpha_j e^{\frac{\Delta_j}{2}} \right)^{-\frac{1}{2}}
\]

Notice that the BIC method is not a full Bayesian approach – it doesn’t allow priors to be set up on any parameters in the models. The use of model priors is also optional, and it is often
assumed that models are equally likely a priori. Hence the prior probabilities $P(M_i)$ are commonly assigned equal probabilities $1/K$, so that the $\alpha$’s disappear in the above equation. Sometimes we do have a priori information, and this should ideally be reflected in the model priors, but this is usually very difficult to do. Madigan, Gavrin and Raftery (1995) discuss the eliciting of prior information. Furthermore, collinear or similar models should have an adjusted prior weight to ensure the models are weighted correctly overall (See Section 7). So, posterior model probabilities take the following form, by using the BIC to approximate the Bayes factor.

$$p_i = \frac{\Delta_i e^{-\frac{\Delta_i}{2}} P(M_i)}{\sum_{j=1}^K e^{-\frac{\Delta_j}{2}} P(M_j)}$$

One of the criticisms of Bayesianism is the “subjective” nature of the prior, and whether this method is sensitive to the choice of prior is a major consideration. However, a number of papers investigating the Bayesian multimodel inference framework found that the model average is not overly sensitive to prior judgement and produces a more accurate predictive distribution than the one best model (Raftery, Madigan and Hoeting, 1997). Madigan and Raftery (1994) measure predictive performance by splitting the data into sets and find that a model average (using exact Bayes factors) produces better predictive performance than the one “best” model, as measured by the logarithmic scoring rule of Good (1952). Model averages based on Markov Chain Monte Carlo (MCMC) methods also have better performance than one “best” model.

5.3 The Information Theoretic Framework for Multimodel Inference

The purpose of statistical experiments and observations is to obtain information (Kullback, 1997). Information theory deals with the inferences that can be made from such experiments. In other words, information theory attempts to quantify the information contained in the observations. Kullback (1997) provides a good overview of information theory in statistics.

In model selection, we are concerned with how well each model describes reality. In other words, we want a model which expresses the information available in the best way. Kullback and Leibler (1951) introduce the Kullback-Leibler distance (see Section 3.1), which can be understood to represent the amount of information about reality $f$ conveyed by an approximating model $g$. After some simple manipulation, it was shown that the above equation may be written in the form $I(f, g) - K = -E_f \left[ \log \left( g(x|\theta) \right) \right]$ where $K$ is an unknown constant.

The AIC is derived as an approximation to $-E_f \left[ \log \left( g(x|\theta) \right) \right]$, which has no meaning in itself because of the unknown constant. However, the AIC differences do have meaning, describing how much better one model is than another. So, define the AIC differences
as \( \Delta_i = \text{AIC}_i - \text{AIC}_0 \). The AIC differences should be used to rank the models from best to worst, with the smallest \( \Delta_i \) being the best. Burnham and Anderson (2002, p.70) give some rough rules of thumb to link the value of the AIC differences with the evidence for model \( i \) given the data:

<table>
<thead>
<tr>
<th>( \Delta_i )</th>
<th>Level of Empirical Support of Model ( i ) vs Model 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 2</td>
<td>Substantial</td>
</tr>
<tr>
<td>4 – 7</td>
<td>Considerably less</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>Essentially none</td>
</tr>
</tbody>
</table>

### 5.3.1 Akaike weights

The previous section describes how differences between AIC values can be used to rank models from best to worst. Akaike weights take this one step further, by transforming the AIC differences into a quantity which can be interpreted as the probability of a model given the data (Burnham and Anderson, 2002). The Akaike weight of model \( i \) is

\[
 w_i = \frac{e^{-\frac{\Delta_i}{2}}}{\sum_{j=1}^{K} e^{-\frac{\Delta_j}{2}}}
\]

(Burnham and Anderson, 2002). Akaike (1978) first recommended an \( e^{(-1/2) \text{AIC}} \) type factor to weight models in an “average of the models.” Since AIC = \(-2 \log L + 2k\), the likelihood of a model given the data may be computed as: \( L_i = L(M_i | x) \propto e^{-\frac{\Delta_i}{2}} \) and Akaike weights can be expressed in terms of the likelihood function: \( w_i = L_i e^{-\frac{q_i}{2}} \sum_{j=1}^{K} L_j e^{-\frac{q_j}{2}} \) where \( q_i = 2k \) is the penalty term for model \( i \). This form of the Akaike weights shows more clearly why the weights make sense. The numerator is the penalised likelihood of model \( i \), while the denominator is the sum of the penalised likelihoods of all the other models, so that the weights sum to one, and can be interpreted as the probability that a given model is correct. Furthermore, two models with the same AIC will receive the same weight, regardless of whether they have the same penalty. A further justification is the interpretation of the AIC penalty (divided by -2) as a bias adjustment term for the expected log likelihood (Buckland et al., 1997). The AIC can also be derived in a Bayesian context and an Akaike weight is a valid posterior model probability (Burnham and Anderson, 2002). To get Akaike weights from Bayesian posterior model probabilities, the following model prior can be used \( P(M_i) = e^{\left(\frac{\Delta_{\text{BIC}}}{2}\right)} \cdot e^{\left(\frac{\Delta_{\text{AIC}}}{2}\right)} \) (Burnham and Anderson, 2002), although it would seem unlikely that a Bayesian would choose this prior.
Finally, Akaike weights can also be generalised, allowing manual a priori weights $\tau_i$ to be selected, giving the same form for model weights as the Bayesian posterior model probabilities.

$$w_i = \frac{\tau_i e^{\frac{\Delta_i}{2}}}{\sum_{j=1}^{K} \tau_j e^{\frac{\Delta_j}{2}}}$$

Analogous to the Bayes factor, information theorists have the “evidence ratio” which represents the strength of evidence for a certain model compared to another fitted model, in a Kullback-Leibler information sense. The evidence ratio for model $i$ versus model $j$ is just the ratio of Akaike weights $w_i / w_j$. Of particular interest in model selection is the evidence ratio for the best model versus the $i^{th}$ best model.

### 5.4 The Frequentist Framework for Multimodel Inference

Like Bayesians and information theorists, frequentists acknowledge that the model selection process introduces uncertainties into parameter estimates, and the usual methods of precision underestimate the true variability. Several plausible models may fit the data well, but give rise to non-overlapping confidence intervals. Buckland, Burnham and Augustin (1997) explore frequentist philosophies to incorporate model selection uncertainty into inference, including estimating model weights by a bootstrap method (see also Buchholz, Holländer and Sauerbrei, 2008) and by using approximation methods similar to the Bayesian and Information Theoretic approaches. In its simplest form, the bootstrap method involves drawing samples from the data set then selecting the best model based on each sample. The proportion of resamples which choose model $i$ as the “best” model approximates the model weight. For the approximation method, Buckland et al. (1997) base model weights on an information criterion of the form $I = -2 \log L + q$ although they do not recommend a particular criterion. The value of $q$ is a penalty used for enforcing parsimony; for the AIC, $q = 2k$, while for the BIC, $q = k \log n$. The model with the smallest $I$ is the best model in some sense. Such a criterion leads to model weights of the form $w_i = \frac{\frac{I_i}{2}}{\sum_{j=1}^{K} \frac{I_j}{2}}$ (Buckland et al., 1997).

Note that this is equivalent to the model weights under the Bayesian and Information Theoretic paradigms, even using AIC or BIC to choose model weights.

Frequentist approaches choose model weights on a reasonable basis, which may include the use of AIC or BIC, but evaluate the model average and performance of the model-averaged estimators, in a frequentist fashion. The investigations conducted by Buckland et al. (1997) and Buchholz et al. (2008) into the performance of model averaging compared to selecting a single
best model found that where model selection uncertainty is ignored, confidence interval coverage is “well below the nominal level.” Buckland et al. (1997) found that using AICc to calculate weights, coverage achieved the nominal coverage of 95% to a good approximation. Buchholz et al. (2008) also found that model averaging coverage was very good when using a bootstrap method of weight selection. For examples of frequentist model averaging, see Buckland et al. (1997), Buchholz et al. (2008) and Schomaker et al. (2009).

5.5 Collinearity or similar models
An issue arises when we have a number of similar models (or copies) included in the set of candidate models in the analysis; model weights are distorted, and predictions are biased. The effect can be substantial if several copies of a bad model are included, as it will cause the overall weighting of that type of model to be too high, stealing weight from the other models. A Bayesian interpretation of the similar models issue is that when similar models are included, it is effectively assigning a high prior weighting to that type of model. Thus it makes sense to adjust the prior weights of the similar models, so that they are on equal footing with the other models under consideration. Where there are $k$ similar models, a prior probability of $1/K$ should be a sufficient approximation to the effect of the similar models, as long as $k$ is small. As the models become less similar, more weight could be assigned to each. A more formal allowance may be possible using some measure of discrepancy to assign the prior weights.

6 Application to reserving
General insurance companies must set aside reserves in order to pay future claims arising from current or past policies. A loss triangle contains all observed losses for a particular product, split by accident year and development year. The accident year refers to the period of insurance coverage (usually a year), while the development year refers to the time of payment following the period of coverage. Calendar year payments are also useful, referring to the year of payment of the loss. For certain types of insurance (e.g. liability insurance) it can be many years before all claims related to a certain accident year are paid out, thus models based on the observed data must be used to predict future losses.

If parameter values were known, the model could predict a future observation from the distribution $Y_f \sim N(\mu, \sigma^2)$. In practice, however, the parameter values are not known, and parameter estimates based on past data are used. Thus predictions come from a distribution such as $Y_f \sim N(\hat{\mu}, \hat{\sigma}^2)$. Since $\hat{\mu}$ is not fixed but random, the usual measures of variance underestimate the average variation between $Y_f$ and $\hat{Y}_f$. The extra variability from parameter uncertainty should be included in the variance of predictions. The Australian Prudential
Regulation Authority’s requirements for the calculation of the risk margin only mention uncertainty about the mean, not uncertainty about the variance (e.g. GPS 310, 2008, paragraph 77). Indeed, the popular lognormal model results in a log-t predictive distribution when both mean and variance uncertainty are taken into account (Dickson, Tedesco and Zehnwirth, 1998), and none of the moments of the log-t distribution exist. This implies an infinite reserve, which is neither practical for the insurer nor comforting for APRA as the prudential regulator. Therefore, calculations of predictive variance in reserving generally ignore the uncertainty surrounding the variance parameter - it is estimated and then assumed to be fixed.

Traditionally, the chain ladder method has been used to estimate future insurance claims (see England and Verrall, 2002 for a description). Now, the chain ladder is a distribution free model, hence the likelihood cannot be written down. This problem can be partly overcome by using the ratio model given in Section 6.2, which replicates chain ladder forecasts. However, the full likelihood is not able to be calculated for the ratio model, as there are no fits for the first development year. Since only the conditional likelihood is available, the chain ladder can not be directly compared with other models. Some rough comparisons can be made by: 1) calculating other model’s likelihoods without the first year’s data; then producing model weights; 2) an ad hoc scaling the chain ladder likelihood by a factor such as $\frac{X}{X_m}$ to make a rough comparison; and 3) examining the contributions of the first development year to the log likelihood in other models, and adding the average contribution to the chain ladder conditional log likelihood. It should be noted that all these ideas ignore the fact that chain ladder is not being penalised for parameters which attempt to explain the first development year, while other models are.

### 6.1 Data Sets

Four loss triangles were used to compare the results of multimodel inference with model selection:

- Mortgage (Sanders, 1990 in Mack, 1993, Table 4)
- GenIns (Taylor and Ashe, 1983)
- ABC (Barnett and Zehnwirth, 2000, Table 3.5)
- CTP (Taylor, 1981, p. 97)

### 6.2 Models

In order to study a useful application of multimodel inference, a number of models are needed. The models used in this example include standard regression models, chain ladder, Hoerl curves, natural cubic splines and generalised linear models (England and Verrall, 2002; Nichols, 2009). In addition to these standard models, I introduce two new models, an autoregressive time series model, fitting an AR(1) to the calendar year residuals (on the log scale) (Nichols, 2009), and a nonparametric regression model which smooths in the development year direction (Nichols, 2009), using the local mean (Nadaraya-Watson) estimator.
A smoothing model will probably work better to describe the accident year effect, and local linear regression smoothing (Wand and Jones, 1995, Chapter 5), which fits a straight line locally, will perform better than the local mean estimator.

Let \( p_{ij} \) be the paid loss, \( C_{ij} \) be the cumulative paid loss and \( \mu_{ij} \) be the actual mean paid loss in accident year \( i \), development year \( j, i = 1, ..., m; j = 1, ..., n \). The models used are as follows:

1) Normal model in development year
\[
p_{ij} = \mu_j + \epsilon_{ij}\text{ where } \epsilon_{ij} \sim N(0, \sigma_j^2)
\]

2) Normal model in development year with a calendar year trend
\[
p_{ij} = \mu_j + (i + j - 1) \tau + \epsilon_{ij}\text{ where } \epsilon_{ij} \sim N(0, \sigma_j^2)
\]

3) Lognormal model in development year
\[
\log(p_{ij}) = \mu_j + \epsilon_{ij}\text{ where } \epsilon_{ij} \sim N(0, \sigma_j^2)
\]

4) Lognormal model in development year with a calendar year trend
\[
\log(p_{ij}) = \mu_j + (i + j - 1) \tau + \epsilon_{ij}\text{ where } \epsilon_{ij} \sim N(0, \sigma_j^2)
\]

5) Chain Ladder (Ratio model)
\[
C_{ij} = \lambda_{j-1} C_{i,j-1} + \epsilon_{ij}\text{ where } \epsilon_{ij} \sim N(0, \sigma_j^2 C_{i,j-1})
\]

6) Hoerl Curve with the same level in each accident year (Hoerl Curve 1)
\[
\log(p_{ij}) = \alpha + \beta \log(j) + \gamma j + \epsilon_{ij}\text{ where } \epsilon_{ij} \sim N(0, \sigma_j^2)
\]

7) Hoerl Curve with a different level in each accident year (Hoerl Curve 2)
\[
\log(p_{ij}) = \alpha_i + \beta \log(j) + \gamma j + \epsilon_{ij}\text{ where } \epsilon_{ij} \sim N(0, \sigma_j^2)
\]

8) Four Natural Cubic Splines on the log paid losses with internal knots at \((2, 4); (2, 5); (3, 5)\) and \((2, 4, 5)\) and boundary knots all at \((1, 7)\).

9) Overdispersed Poisson GLM in accident year and development year
\[
\log(\mu_{ij}) = \alpha_i + \beta_j
\]

10) Gamma GLM in development year
\[
\log(\mu_{ij}) = \beta_j
\]

11) Gamma GLM in development year and calendar year
\[
\log(\mu_{ij}) = \beta_j + (i + j - 1) \tau
\]

12) Gamma GLM in accident year and development year
\[
\log(\mu_{ij}) = \alpha_i + \beta_j
\]

13) Splines-type GLM in accident year and development year
\[
\log(\mu_{ij}) = \alpha + \beta I_1 + \gamma I_2 + \tau I_3
\]
where \( I_1 = (j - 2) \text{ if } j = 3, ..., n; \quad I_2 = j^2 \text{ if } j = 1, 2; \quad I_3 = j^3 \text{ if } j = 1, 2 \)

14) AR(1) in calendar year
\[
\log(p_{ij}) = \mu_j + \beta \left( \log(p_{i,j-1}) - \mu_{i-1,j-1} \right) + \epsilon_{ij}\text{ where } \epsilon_{ij} \sim N(0, \sigma_j^2)
\]

15) Nonparametric regression in development year by the kernel density approach
\[
p_{ij} = m(j) + \epsilon_{ij}\text{ with } \epsilon_{ij} \sim N(0, \sigma_j^2)
\]
The model fitting and calculations required for multimodel inference are done in R (R Development Core Team, 2009). The chain ladder model was fitted using the package ‘ChainLadder’ (Gesmann, 2009) and the nonparametric regression model was fitted using the package ‘sm’ (Bowman and Azzalini, 2007).

6.3 Model fit and diagnostics

Some useful diagnostics when fitting models to claims triangles include residual plots against development year, accident year and calendar year as well as residual plots against fitted values and any other potential predictors. The residual plots allow us to better answer the questions: Is the model for the mean reasonable? Are there effects that it misses? Does the model describe the variance accurately? It is well known that a well-fit model may not be good for prediction, thus testing the model fit does not test the quality of predictions. The process of validation tests of the suitability of the model for prediction. It removes the most recent few years of data, refits the model and predicts the removed data (this is similar to one-step-ahead prediction errors in time series analysis). Validation can also be conducted when new data arrives each year, to monitor the usefulness of the model over time. A suitable standardised validation residual is (Barnett, 2007, p. 20):

\[
\frac{(\text{observed} - \text{predicted})}{\text{predicted standard error}}
\]

The quality of predictions from multimodel inference can be tested and compared to that of the single best model. Since a model averaged distribution describes the data more accurately than a single best model, this type of testing should illustrate that the model averaged predictive distribution produces more accurate predictions than the single “best” model. Validation is not considered any further in this paper.

The sensitivity of the outstanding claims liability estimate to changes in the data should also be considered in the modelling process. Ye’s (1998) concept of Generalised Degrees of Freedom (GDF) is a measurement of the sensitivity of fitted values to changes in the observed value:

\[
GDF = \sum \frac{\partial y_{ij}}{\partial y_{ij}}
\]

Venter and Tampubolon (2008) and Tampubolon (2007) apply this concept to reserving, considering instead the derivative of the reserve with respect to each data point. Their “robust analysis” is able to detect influential data points, whose random components may have a large impact on the reserve.
7 Multimodel Inference Data Analysis

7.1 Mortgage Data

The Mortgage data is an interesting data set, as there is considerable model selection uncertainty, with vastly different predictions between the top two individual models. Figure 7.1 shows the model averaged distribution for the Mortgage data.

![Figure 7.1: This plot shows the model average of the predictive distributions (i.e. allows for uncertainty in the mean parameters, but not in the variance parameters) based on the AIC weights (top left) and BIC weights (top right). The predictive distributions of the top two models are included.](image)

In this example, a very large component of the variance is model selection uncertainty. If the task was model selection, the AIC, AIC$_C$ and BIC would all choose the Hoerl Curve 2 model. However this model leads to a reserve which is likely to be much too low. If the insurer sets their outstanding claims liability reserve at the 75$^{th}$ percentile of the predictive distribution, the insurer will have under-reserved by $7-9m!

The model average is influenced in a major way by the *Hoerl Curve 2* model, which has a much higher weight than other models, and a much lower mean reserve. The following figure shows the diagnostic plots for the fitted *Hoerl curve 2* model.

<table>
<thead>
<tr>
<th>Model Average</th>
<th>AIC$_C$</th>
<th>AIC$_C$ w</th>
<th>BICw</th>
<th>Mean Res</th>
<th>SD</th>
<th>75%</th>
<th>95%</th>
<th>99.5%</th>
<th>99.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Average AIC$_C$</td>
<td>10910</td>
<td>6015</td>
<td>15640</td>
<td>21688</td>
<td>28553</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Average BIC</td>
<td>12838</td>
<td>6423</td>
<td>17780</td>
<td>23317</td>
<td>31015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splines (2,5)</td>
<td>0.000</td>
<td>0.006</td>
<td>0.039</td>
<td>18793</td>
<td>3716</td>
<td>21040</td>
<td>25443</td>
<td>30531</td>
<td></td>
</tr>
<tr>
<td>Splines (3,5)</td>
<td>0.000</td>
<td>0.006</td>
<td>0.038</td>
<td>19129</td>
<td>3859</td>
<td>21455</td>
<td>26043</td>
<td>31364</td>
<td></td>
</tr>
<tr>
<td>Hoerl Curve 1</td>
<td>0.000</td>
<td>0.015</td>
<td>0.190</td>
<td>17700</td>
<td>4173</td>
<td>20154</td>
<td>25257</td>
<td>31363</td>
<td></td>
</tr>
<tr>
<td>Hoerl Curve 2</td>
<td>0.844</td>
<td>0.701</td>
<td>0.511</td>
<td>8012</td>
<td>4214</td>
<td>9896</td>
<td>15986</td>
<td>25325</td>
<td></td>
</tr>
<tr>
<td>AR(1) in CY</td>
<td>0.083</td>
<td>0.256</td>
<td>0.124</td>
<td>17701</td>
<td>3528</td>
<td>19831</td>
<td>24017</td>
<td>28861</td>
<td></td>
</tr>
</tbody>
</table>
Figure 7.2: Plots of the standardised residuals by calendar year, accident year, development year and fitted values for the Hoerl curve 2 model.

There appears to be strong calendar year effects. There is also an outlier, which is from accident year 1, development period 1. The other plots are quite noisy and it is difficult to see any hidden trends. One of the contributing factors to the low reserve for this model is the very low level of claims in the final two accident years. Since the model fits a parameter to each accident year, this leads to low predicted claims for all future developments in these accident years. As discussed in Section 6.3, the sensitivity of the total reserve to changes in the data should be examined as part of the diagnostic checks.

Figure 7.3: Residuals vs Calendar Year for the AR(1) in PY. The other plots do not show any trends.
The AR(1) model does not fit well. Ignoring the first residual, there is a strong downward trend to calendar year 6, then a strong upward trend.

### 7.2 GenIns Data

As with the Mortgage data, the GenIns data has a number of plausible models with different mean reserve, and hence model selection uncertainty is again a material consideration. The following plot shows the model averaged predictive distribution and the individual model predictive distributions.

![Model Averaged Predictive Distribution for GenIns Data](image)

**Figure 7.4**
Left: Model averaged predictive distribution (i.e. allowing for uncertainty surrounding mean parameters, but not variance parameters) for the GenIns data, based on the AIC\(_c\) weights.
Below: Mean reserve, SD, percentiles and model weights for the top weighted models

<table>
<thead>
<tr>
<th>Model</th>
<th>AICw</th>
<th>AIC(_c)w</th>
<th>BICw</th>
<th>Mean Res</th>
<th>SD</th>
<th>75%</th>
<th>95%</th>
<th>99.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Average AIC(_c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Average BIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gamma GLM (DY+CY)</td>
<td>0.119</td>
<td>0.281</td>
<td>0.088</td>
<td>19440</td>
<td>2437</td>
<td>21084</td>
<td>23449</td>
<td>25718</td>
</tr>
<tr>
<td>AR(1) in CY</td>
<td>0.157</td>
<td>0.268</td>
<td>0.016</td>
<td>16878</td>
<td>1749</td>
<td>18000</td>
<td>19898</td>
<td>21907</td>
</tr>
<tr>
<td>Gamma GLM (DY)</td>
<td>0.080</td>
<td>0.189</td>
<td>0.022</td>
<td>16676</td>
<td>1311</td>
<td>17539</td>
<td>18893</td>
<td>20257</td>
</tr>
<tr>
<td>Splines GLM</td>
<td>0.004</td>
<td>0.114</td>
<td>0.461</td>
<td>16830</td>
<td>1157</td>
<td>17610</td>
<td>18733</td>
<td>19810</td>
</tr>
</tbody>
</table>

When model averaging is done based on AIC\(_c\)-weights, it results in a mixture distribution which is very different to any single model. For example, the mean of the single best model (*Gamma GLM (DY+CY)*) is $2m greater than that of the model average, and $2.4m greater than the 75\(^{th}\) percentile of the model average. Most models’ predictive distributions are centred around $16.7m, while the best model is at $19.5m, so that the model mixture is very spread out. The much larger spread means there is a great deal of model selection uncertainty in this example. Furthermore, the BIC weights are quite different to the AIC\(_c\) weights, so the choice of model selection criterion is an also important consideration.
### 7.3 ABC Data

The ABC data is interesting because the Chain Ladder model fits quite well, so it gives an example of the issues of evaluating its likelihood (Section 6).

**Left: Figure 7.5** The model averaged predictive distribution (i.e. allowing for uncertainty surrounding mean parameters, but not variance parameters) for the ABC data set. The distributions based on AIC and AIC<sub>c</sub> are not materially different.

**Right: Figure 7.6** The model averaged distribution where posterior model probabilities are calculated from log-likelihoods which ignore the first development year. This plot is based on the BIC weight, but plots based on the AIC and AIC<sub>c</sub> are almost identical.

Since the mean reserves are similar, the main element of model selection uncertainty in this case is the difference in the standard deviations under each model.

Founding weights on log likelihoods which omit the first development year leads to posterior weights of over 0.99 (by AIC, AIC<sub>c</sub> and BIC) for the Gamma GLM (AY+DY), while the chain ladder posterior probabilities decrease to below 0.01. The increase in posterior is due to the fact that the log-likelihood under the Gamma GLM decreases by around 123 points when omitting the fits from the first development year, while the Chain Ladder model only decreases by 114 points. Other models under consideration have their likelihoods decreasing by more than 140 points, when the first development year is omitted. This suggests that the scaling up of the chain ladder is not sufficient or reasonable in this case. Unless the scaling factor leads to a similar contribution for the first development year compared to other models, the scaling will not work. Thus the results from Figure 7.5 are invalid, as the model weight for the chain ladder is likely to be much lower such as in Figure 7.7.

<table>
<thead>
<tr>
<th>Model Average BIC</th>
<th>AICw</th>
<th>AIC&lt;sub&gt;c&lt;/sub&gt;w</th>
<th>BICw</th>
<th>Mean Res</th>
<th>SD</th>
<th>75%</th>
<th>95%</th>
<th>99.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Average BIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chain Ladder</td>
<td>0.000</td>
<td>0.002</td>
<td>0.001</td>
<td>5278</td>
<td>153</td>
<td>5381</td>
<td>5529</td>
<td>5671</td>
</tr>
<tr>
<td>Gamma GLM (AY+DY)</td>
<td>0.993</td>
<td>0.991</td>
<td>0.993</td>
<td>5238</td>
<td>249</td>
<td>5406</td>
<td>5648</td>
<td>5880</td>
</tr>
</tbody>
</table>

25
Figure 7.7: Model averaging results for the ABC data, based on the BIC, omitting the first development year fits in the likelihood.

Figure 7.6 and Figure 7.7 illustrate the situation of one dominant model with most of the weight. When this is the case, the model averaging gives a distribution which is nearly identical to the model with high weight, as it should.

7.4 CTP Data

The CTP data analysis results show the classic case of when multimodel inference should be used (when based on BIC weights).

![Model Average for CTP data based on BIC weights](image)

Figure 7.8: Top Left: model averaged predictive distribution (i.e. allowing for uncertainty surrounding mean parameters, but not variance parameters) based on the BIC for the CTP data. Top Right: model averaged predictive distribution for the CTP data based on AIC. Bottom: This table shows the weights for the top few individual models, along with the mean reserve, standard deviation of the reserve and percentiles of the predictive distribution for these models and the model averages based on the AIC and the BIC.

<table>
<thead>
<tr>
<th>Model Average</th>
<th>AICw</th>
<th>AICcw</th>
<th>BICw</th>
<th>Tot Res</th>
<th>Tot SD</th>
<th>75%</th>
<th>95%</th>
<th>99.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Average AIC</td>
<td>15249</td>
<td>2504</td>
<td>16799</td>
<td>19679</td>
<td>22902</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Average BIC</td>
<td>13182</td>
<td>1960</td>
<td>14055</td>
<td>16982</td>
<td>20797</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splines (4,2)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.020</td>
<td>12652</td>
<td>1276</td>
<td>13472</td>
<td>14853</td>
<td>16311</td>
</tr>
<tr>
<td>Hoerl Curve 1</td>
<td>0.000</td>
<td>0.004</td>
<td>0.751</td>
<td>12627</td>
<td>1318</td>
<td>13473</td>
<td>14905</td>
<td>16422</td>
</tr>
<tr>
<td>Hoerl Curve 2</td>
<td>1.000</td>
<td>0.994</td>
<td>0.210</td>
<td>15264</td>
<td>2501</td>
<td>16811</td>
<td>19687</td>
<td>22909</td>
</tr>
</tbody>
</table>

The model averaged predictive distributions are quite different depending on whether AIC or BIC weights are used. The AIC weights give the Hoerl curve 2 model a weight close to 1. Thus the model average distribution is very similar to the predictive distribution under the single “best” model. However, if BIC weights are used, the Hoerl curve 2 model is penalised much more (see Figure 4.1), since it has $m + 3$ parameters, and the Hoerl curve 1 model
receives most of the weight (it has only 4 parameters). This leads to a model averaged predictive distribution which is lower than that produced by AICc weights.

The plot of the Model Average predictive distribution based on the BIC weights has a heavier upper tail than the single “best” model. Therefore, upper percentiles (e.g. 75% or 99.5%) of the single best model underestimate the actual percentile implied by the data. This is the classic case of where multimodel inference should be used.

Since the different methods result in very different distributions, diagnostic plots of the top BIC-ranked models are examined (Hoerl curve 1 and Hoerl curve 2). Figure 7.9 shows the residual plots for the Hoerl curve 1 model, while Figure 7.10 shows the residual plots for the Hoerl curve 2 model. Note that the data has not been adjusted for exposure or inflation.

Figure 7.9: Standardised residuals against calendar year, accident year, development year and fitted values for the Hoerl curve 1 model.

For the first few calendar years, the residual plot against calendar year shows a slight pattern, but this appears to stabilise after payment year 7. The residuals against accident year show a wave type effect, which is not being modelled in the simple Hoerl curve. The residuals against development year and fitted values plots appear fine. There is a large standardised residual of 7.65 in accident year 4, development year 8.
Figure 7.10: Standardised residuals against calendar year, accident year, development year and fitted values for the Hoerl curve 2 model.

Similarly to the simpler Hoerl curve model, the standardised residuals vs calendar year shows a small trend for the first 7 payment years, but seems to stabilise after that. The accident year trend observed in the Figure 7.9 is now allowed for, with the residuals vs accident year showing no trend, the residuals vs development year and residuals vs fitted values do not show any patterns. There are two large standardised residuals in this model fit, one in AY1, DY2 (2.83) and one in AY4, DY8 (5.51) (compared to one of 7.65 from the simpler Hoerl curve model).

For predictive purposes, the Hoerl curve 2 model does fit the accident year effect better than the simpler model, but whether this is worth the extra parameter uncertainty is debatable. After all, if the Hoerl curve 1 model is used, and the accident year trend continues in a wave fashion, the under-predictions of future claims in accident years 2 to 5 will be offset by over-predictions in accident years 6 to 8. Also, there is not much information available about accident years past 8. Based on all these points, in my opinion the simpler Hoerl curve model should predict the future well with lower parameter uncertainty, thus the BIC weights are better to use in this case.
7.5 Comments on data analysis
The examples shown in this paper used either AIC$_C$ or BIC, as AIC tended to choose models with too many parameters. The choice between AIC$_C$ and BIC depends on the situation. For forecasting, a simpler model is preferred, and BIC does assign more weight to simpler models. However, the fit of the models should be checked, as sometimes the top BIC-weighted models may not describe the data sufficiently for forecasting. AIC$_C$ performed consistently in the examples, generally assigning weights between the AIC weights and BIC weights.

Kass and Raftery (1995) state that model uncertainty should be accounted for, and thus choosing a single best model is only reasonable if one model is clearly better than the rest. If this is not the case, the predictions are biased, since the selected model has been given too much weight. The gains from model averaging depend on the situation. When one model is clearly better than the rest, the model selection uncertainty may be small, so the gain from model averaging is small (e.g. Figure 7.5). However when there are a number of plausible models, model selection uncertainty can be large (e.g. Figures 7.1 and 7.4). In this case, model averaging can provide a much better description of the data than the single best model.

8 Conclusion
General insurance actuaries face a difficult task when calculating the outstanding claims liability and recommending provisions. The estimation of the 75th percentile of the predictive distribution and the mean plus half the standard deviation are in many cases difficult to calculate directly, hence judgement must be utilised along with technical analysis (GPS 310, 2008, p. 26). However, the focus on sensitivity analyses to give actuaries an “indication of the uncertainty in the central estimates” (GPS 310, 2008 p. 30; PS 300, 2008, p. 18) may leave too much to the subjective judgement of the actuary, since it provides a number of predictions under different models, but no indication of how plausible each model is.

This paper argues that instead of such a crude allowance for model uncertainty, a formal framework should be used. Recent advancements in model selection and multimodel inference, as well as computing power, have made it possible to incorporate model selection uncertainty. Using multimodel inference leads to a more accurate description of the observed data; a model averaged predictive distribution that will be better calibrated to predict the future than the single “best” model. At the simplest level, model uncertainties can be allowed for using a simple model weighting method based on AIC, AIC$_C$, BIC or other criterion. This is the approach that is used in this paper. More complex methods have also been described but not implemented, including bootstrap methods, MCMC methods and full Bayesian approaches. Although these
other methods tend to perform a little better than a simple weighting approach, they require much more effort to implement.

The data analysis in Section 7 looks at multimodel inference for the reserving problem, based on a wide range of plausible models. It illustrates that unless one model is clearly better than the other models, a model averaged predictive distribution describes the data better than any single model, and is thus more useful for predicting the future. This is because where there is evidence for more than one model, considerable uncertainty about the model can exist, especially where the models have material differences in predictions.

Multimodel inference takes into account the plausibility of the models, and can be thought of as a logical extension of sensitivity analysis (Draper et al., 1987). This does not replace actuarial judgement, it formalises it. For example, where the actuary has prior beliefs about which model will be better or expects the future to be different to past or current data, this can be incorporated by placing priors on the models. Priors can also be used where there are multiple similar models, to prevent this from having an excessive influence on the overall predictive distribution. This framework allows for increased consistency across actuaries, with differences being more readily describable; they can be explicitly traced back to the actuaries’ judgments, or the models included in the candidate sets. Furthermore, model averaging minimises the risk of forecasting using a poor model as it is not a winner-takes-all method like model selection.

Further research could be valuable in the following areas

- Specification of priors and the application of a full-Bayesian approach for the model averaged predictive distribution
- Implementing MCMC methods to approximate the full-Bayesian solution for averaging over all models in reserving examples (e.g. Peters, Shevchenko and Wüthrich, 2009), and comparing this to the model averages from using simple weighting methods.
- Assessing the predictive performance of model averaged distributions in reserving. This could be measured by standardised residuals from one-step-ahead predictions or some other method.
- A more formal allowance for similar models using a measure of discrepancy to assign the prior weights.
- Exploring the new models introduced in this paper; and in particular fitting a time series in payment year which allows for accident year effects as well as development year effects and fitting a local linear regression smoothing model in the accident year direction.
References


31


32


Yang, Y. (2003). "Regression with Multiple Candidate Models: Selecting or Mixing?" Statistica Sinica 13, 783-809.

