Institute of Actuaries of Australia

A STATISTICAL BASIS FOR CLAIMS EXPERIENCE MONITORING

Prepared by Greg Taylor

Presented to the Institute of Actuaries of Australia
17th General Insurance Seminar
7 – 10 November 2010
Gold Coast

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

Summary

By claims experience monitoring is meant the systematic comparison of the forecasts from a claims model with claims experience as it emerges subsequently. In the event that the stochastic properties of the forecasts are known, the comparison can be represented as a collection of probabilistic statements. This is stochastic monitoring.

The paper defines this process rigorously in terms of statistical hypothesis testing. If the model is a regression model (which is the case for most stochastic claims models), then the natural form of hypothesis test is a number of likelihood ratio tests, one for each parameter in the valuation model. Such testing is shown to be very easily implemented by means of GLM software.

This tests the formal structure of the claims model and is referred to as **micro-testing**. There may be other quantities (e.g. amount of claim payments in a defined interval) that require testing for practical reasons. This sort of testing is referred to as **macro-testing**, and its formulation is also discussed.

Keywords: claims experience monitoring, GLM, hypothesis testing, macro-testing, micro-testing, primary targets, stochastic monitoring.

Claims experience monitoring

1. Introduction

1.1 Background

The purpose of the present paper is to discuss an area of common insurance/actuarial practice whose foundations have been greatly neglected. This is the practice of claims experience monitoring, in which the claims experience of a defined period is compared with a set of model-based forecasts. In more formal language, the issue is as follows.

Suppose a model of claims experience has been formulated and calibrated. Suppose further that it is a **predictive model** in the sense that it is capable of generating forecasts of future claims experience.

In the periods subsequent to the formulation of the model, further claims experience will accumulate and one will be interested in testing whether or not that experience is consistent with the model. The natural test consists of a comparison of the post-model experience with model forecasts.

This sort of comparison of claims experience with model forecasts will be referred to generically as **claims experience monitoring**, or just **monitoring**. It is a form of *post hoc* **model validation** in which the new claims experience provides out-of-sample data.

Commonly this situation will arise in the context of loss reserving. The out-of-sample data will accumulate progressively during the inter-valuation period, and will provide the basis for advance warning of any valuation model failures. Usually, the earlier such failures are detected, the less the balance sheet shock arising from their correction.

Henceforth, for brevity, the post-model experience will be referred to as simply the **experience**. Consider a specific observation Y within the experience. Suppose that the model has produced a forecast $\hat{Y}$ of this quantity. Monitoring (in respect of this quantity) will consist of making some sort of comparison of Y with $\hat{Y}$.

There will be many possible choices of the subject observation Y, and hence many possible comparisons. Natural questions to ask about these comparisons are:

- (Q1) How should the observations Y featured in the monitoring be selected from the available experience?
- (Q2) What form should the comparison between Y and $\hat{Y}$ take, e.g. $Y - \hat{Y}$, $Y/\hat{Y}$, etc?
- (Q3) What criteria should be adopted for deciding whether or not the experience is consistent with the model?

Claims experience monitoring is widely practiced by insurance companies, especially in relation to pricing and valuation models. However, there is almost no body of theory to guide it.

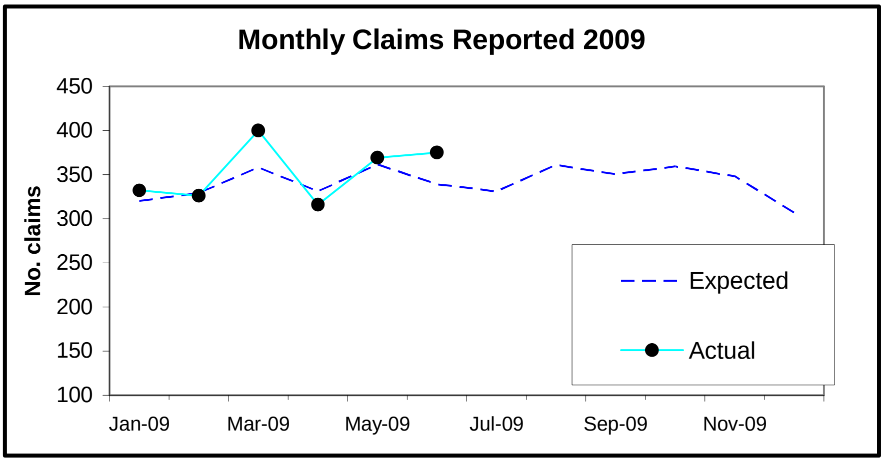
An exception to this is the paper by Berry, Hemming, Matov & Morris (2009), which introduces a number of statistics for assessment of a model in the light of subsequent data. However, this paper is largely directed toward pricing models and their concomitants, such as models of sales conversion, policy retention, cancellation and cross-sell.

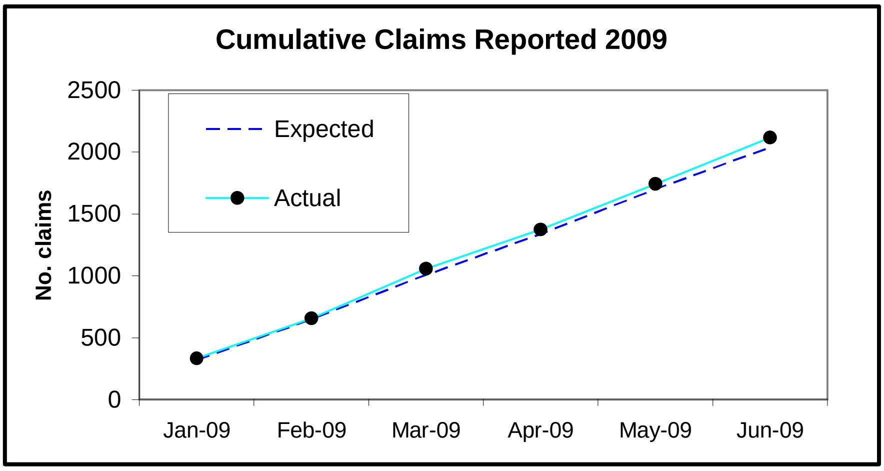
The absence of an integrated body of theory for claims experience monitoring creates a risk that it will be misdirected or incomplete.

1.2 Form and coverage of monitoring

Common forms of monitoring encountered in practice consist of selecting Y to be claim payments and/or numbers of claims reported, dissected by accident period. Figure 1.1 provides an example. These are clearly important quantities but are there others that are relevant to validation of the valuation model?

Figure 1.1 – Graphic representation of claims experience monitoring





Some claims experience models are relatively simple with straightforward (if not necessarily few) parameters (e.g. chain ladder), whereas other models are more complex, containing parameters whose meaning is more subtle (e.g. payments per claim incurred with a Hoerl curve payment pattern).

In general, it will be reasonable to assert that the claims experience monitoring will achieve full coverage of the subject model's features only if it includes at least one comparison table for each parameter in the valuation model. Section 3 will provide a formal justification of this assertion.

1.3 Stochastic monitoring

Claims experience monitoring commonly consists, in practice, of forming the ratios $Y/\hat{Y}$ and considering whether the differences between them and 100% are acceptable. Usually, a ratio of say 102% would be considered acceptable, while 150% might not. But how about 110%? Is this sufficiently different from 100% to call the valuation model seriously into question?

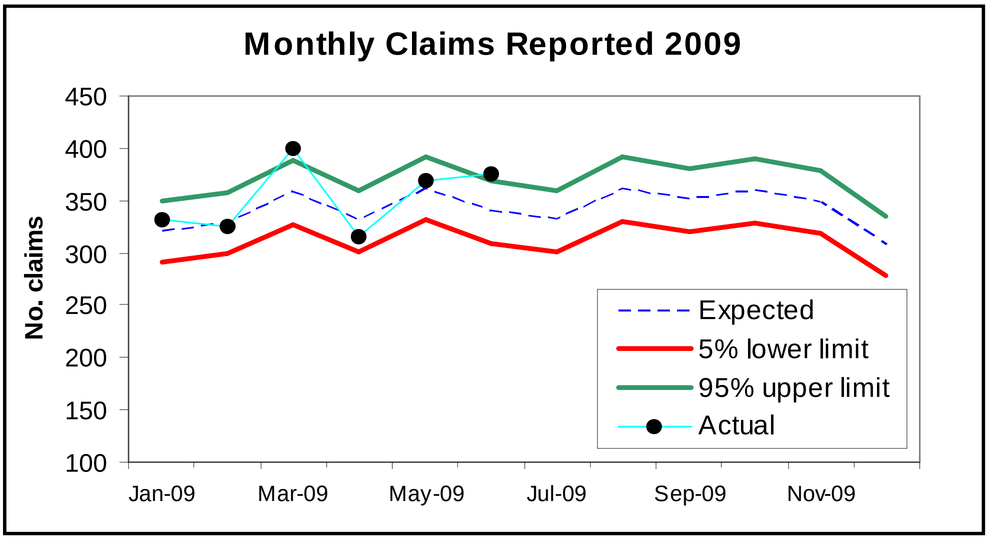
Clearly, the acceptability of such a result would depend on a number of matters, such as:

- What level of sampling error might be expected in the observation Y?
- How precise were the model forecasts $\hat{Y}$?

These are questions about the stochastic properties of the model forecasts. The question inherent in any monitoring is as follows: if $\hat{Y}$ is forecast from model $\mathcal{M}$ [the model under test], what is the probability that $|Y/\hat{Y} - 100\%| > \Delta\%$, where $\Delta\%$ is the observed difference?

If monitoring can be formulated in this way, then it is possible to accompany each observation $Y/\hat{Y}$ with a $100p\%$ confidence interval (l,u) with $l<100\%$, $u>100\%$. Equivalently, each observation Y is accompanied by $100p\%$ confidence interval $(l\hat{Y},u\hat{Y})$. Any observation falls outside this interval with probability $(100-p)\%$. Figure 1.2 extends Figure 1.1 by the addition of a 90% confidence interval.

Figure 1.2 – Graphic representation of stochastic claims experience monitoring



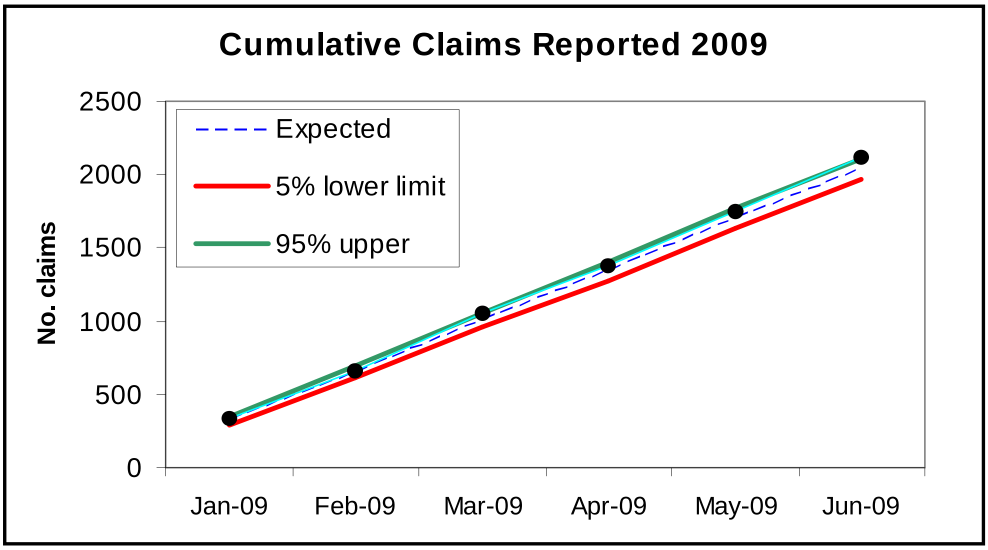


Figure 1.2 sets a 90% confidence level, according to which the model of claim number forecasts is invalidated, or at least called into question, at the 10% significance level. This form of monitoring will be referred to here as **stochastic monitoring**, to be distinguished from the **deterministic monitoring** illustrated in Figure 1.1.

It is apparent from the mere language in which stochastic monitoring has been described that it is an exercise in **hypothesis testing**. The model $\mathcal{M}$ is the null hypothesis, Y (or $Y - \hat{Y}$ or $Y/\hat{Y}$) is the test statistic, and (100-p)% the significance level.

It is apparent that the feasibility of stochastic monitoring depends on the existence of a stochastic valuation model.

1.4 Definition of stochastic monitoring

Against this background, it is possible to give a formal definition of a system of stochastic monitoring that provides full coverage of the valuation model's parameters. However, it is first necessary to define a stochastic model.

Let $X = (X_1, X_2, \ldots, X_n)$ denote a random vector of claims observations. Let $F(x;\theta)$ denote the distribution function of x, dependent on a parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$. This structure will be called a **stochastic claims model**.

Note that, within this model, $E[Y] = g(\theta)$ for some function $g(.)$. A **deterministic claims model** contains no specification of F, but only of $g(.)$. Now to the definition of a monitoring system.

Consider a stochastic claims model $\mathcal{M}$ of claims experience dependent on parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$. Suppose that $\mathcal{M}$ has been calibrated with an estimate $\hat{\theta}$ of θ. The calibrated model generates a forecast of the joint d.f. $G(y; \hat{\theta})$ of a random vector $Y = (Y_1, Y_2, \ldots, Y_r)$ with mean $\hat{Y}(\hat{\theta}) = (\hat{Y}_1(\hat{\theta}), \hat{Y}_2(\hat{\theta}), \ldots, \hat{Y}_r(\hat{\theta}))$. The d.f. G provides the capacity for testing the null hypothesis $\theta = \hat{\theta}$ on the basis of an observation on Y, and such a test will be called a **stochastic claims experience monitoring system**.

Suppose further that $r \geq q$, and that each Y_i relates to θ_i, $i=1,2,\ldots,q$ in a sense yet to be defined. Then the system will be said to provide **full coverage** (of the valuation model's parameters). The quantities $Y_1, Y_2, \ldots, Y_q$ will be called **test forecasts**. There may be other forecasts $Y_{q+1}, Y_{q+2}, \ldots, Y_r$, but they are not required for the testing of the hypothesis $\theta = \hat{\theta}$. Selection of the test forecasts will be discussed in Section 3.1.

A **deterministic monitoring system** is a lesser structure in which $\mathcal{M}$ generates only point estimate forecast of $\hat{Y}$, with no associated distributional information.

Note that the definition of full coverage is as yet incomplete due to the vagueness of the statement that "each Y_i relates to θ_i". The intuitive objective is that each test statistic Y_i be some statistic from which an efficient test of the hypothesis $\theta_i = \hat{\theta}_i$ can be constructed. Section 3 will consider the precise meaning of the relation between Y_i and θ_i.

The procedures established there, referred to as micro-testing, will test the fine structure of the model. Section 4 will discuss macro-testing, the testing of the extent to which post-model experience is consistent with its main purposes.

2. Heuristic selection of test forecasts

In many simple models the parameters may have simple physical meanings, and there may be obvious choices of test forecasts that are reasonable. Consider the following example.

Example 2.1 (stochastic form of chain ladder). Let N_{kj} denote the number of claims notified in development period j of accident period k for k=1,2,…,,K and j=1,2,…,K-k+1 (a data triangle), and define

$X_{kj} = \sum_{i=1}^{j} N_{ki}$ = cumulative row sum

Suppose that the X_{kj} satisfy the following assumptions.

(CL1) Accident periods are stochastically independent, ie $N_{k_1 j_1}, N_{k_2 j_2}$ are stochastically independent if $k_1 \neq k_2$.

(CL2) For each k = 1, 2, …, K, the N_{kj} (j varying) form a Markov chain.

(CL3) $N_{k,j+1} \sim Poisson\left(X_{kj} f_j\right)$ for parameters $f_j > 0, j = 1, 2, ..., K-1$.

Note that, by (CL3),

$$E\left[N_{k,j+1}\right] = Var\left[N_{k,j+1}\right] = X_{kj} f_j \tag{2.1}$$

Assumptions (CL1) and (CL2), together with (2.1), describe a special case of the Mack chain ladder model (Mack, 1993). That model was distribution free and so the model described by (CL1) – (CL3) is the Mack chain ladder equipped with the Poisson cell distributions.

Mack demonstrated a certain justification for the chain ladder estimates of the parameters (age-to-age factors) f_j, ie the estimates

$$\hat{f}_j = \sum_{k=1}^{K-j} X_{k,j+1} / \sum_{k=1}^{K-j} X_{kj} \tag{2.2}$$

Taylor (2010) showed that these estimates were also maximum likelihood and minimum variance unbiased.

The model then generates one-step-ahead forecasts (note that observations on the last diagonal of the data triangle are $N_{k,K-k+1}, k = 1, 2, ..., K$)

$$\hat{N}_{k,K-k+2} = X_{k,K-k+1}\left(\hat{f}_{K-k+1} - 1\right) \tag{2.3}$$

It seems reasonable, or even “obvious”, that claims experience monitoring compare $N_{k,K-k+2}$ with $\hat{N}_{k,K-k+2}$ for k = 1, 2, …, K and $\sum_{k=1}^{K} N_{k,K-k+2}$ (diagonal sum) with $\sum_{k=1}^{K} \hat{N}_{k,K-k+2}$. □

Example 2.2 (stochastic form of payments per claim incurred model)

Let Y_{kj} denote inflation-corrected claim payments in development period j of accident period k for $k = 1, 2, \ldots, K$ and $j = 1, 2, \ldots K - k + 1$. Let N_k denote the number of claims incurred in accident period k, supposed known with certainty. Define the payments per claim incurred ("PPCI") in cell (k,j) as

$$P_{kj} = Y_{kj} / N_k \tag{2.4}$$

PPCI models are discussed in Section 4.2 of Taylor (2000).

Suppose the Y_{kj} satisfy the following assumptions.

(PPCI1) The Y_{kj} are all stochastically independent.

(PPCI2) Y_{kj} ~ Gamma, with

(a) $E\left[Y_{kj}\right] = N_k \pi_j$.

(b) $Var\left[Y_{kj}\right] = N_k \sigma_j^2$

for parameters $\pi_j, \sigma_j^2 > 0$.

By (2.4) and (PPCI2a),

$$E\left[P_{kj}\right] = \pi_j \tag{2.5}$$

and so π_j is recognised as the expected PPCI for development period j.

Further

$$Var\left[P_{kj}\right] = \sigma_j^2 / N_k$$

and so, if π_j is to be estimated by a weighted average of the P_{kj} (j fixed), then the minimum variance estimator will be

$$\begin{aligned} \hat{\pi}_j &= \sum_{k=1}^{K-j+1} N_k P_{kj} \Big/ \sum_{k=1}^{K-j+1} N_k \\ &= \sum_{k=1}^{K-j+1} Y_{kj} \Big/ \sum_{k=1}^{K-j+1} N_k \end{aligned} \tag{2.6}$$

which is the estimator often encountered in practice.

Forecasts of future observations are given by

$$\hat{Y}_{kj} = N_k \hat{\pi}_j \tag{2.7}$$

In parallel with Example 2.1, the "obvious" form of monitoring consists of a comparison of $Y_{k,K-k+2}$ with $\hat{Y}_{k,K-k+2}$ and $\sum_{k=1}^{K-k+1} Y_{k,K-k+2}$ with $\sum_{k=1}^{K-k+1} \hat{Y}_{k,K-k+2}$. □

The discussion in Examples 2.1 and 2.2 was made particularly simple by the single-parameter nature of the forecasts (2.3) and (2.7). Consider now an example in which individual cells depend on multiple parameters.

Example 2.3 (Hoerl curve PPCI). Let Y_{kj}, P_{kj} be defined as in Example 2.2. Suppose the Y_{kj} satisfy (PPCI1) and (PPCI2) except that the π_j , instead of being free parameters, take the form:

$$\pi_j = A\left(j-\tfrac{1}{2}\right)^b \exp -c\left(j-\tfrac{1}{2}\right) \tag{2.8}$$

where A, $c > 0$ and b are now the free parameters.

This model is reminiscent of De Jong & Zehnwirth (1983) and Wright (1990). The parametric form (2.8), as a function of j, is referred to as a **Hoerl curve**.

For this model, one does not estimate π_j directly from the data, as in (2.6), but

$$\hat{\pi}_j = \hat{A}\left(j-\tfrac{1}{2}\right)^{\hat{b}} \exp -\hat{c}\left(j-\tfrac{1}{2}\right) \tag{2.9}$$

where $\hat{A},\hat{b},\hat{c}$ are estimates of A, b, c.

Forecasts of future observations are still given by (2.7). Note, however, that $\hat{Y}_{kj}$ now depends on the multiple parameter estimates $\hat{A},\hat{b},\hat{c}$. One may still carry out the same comparisons between Y and $\hat{Y}$ terms as in Example 2.2, but it is not obvious how this tests the parameter estimates $\hat{A},\hat{b},\hat{c}$.

It is possible, however, to construct heuristic tests of these parameters by considering their physical interpretations. It is evident from (2.8) that the Hoerl curve is a discretised gamma distribution (not to be confused with the gamma distribution of Y_{kj} in (PPCI2)) with multiplier A.

Thus

(i) the Y_{kj} are proportional A;

(ii) if (2.8) is regarded as describing the distributions of expected PPCI with respect to j, and if (2.8) is approximated by its continuous (in j) form, then the mean and variance of j are $(b + 1)/c$ and $(b + 1)/c^2$ respectively.

It follows from (i) that the comparison of $\sum_{k=1}^{K-k+1} Y_{k,K-k+2}$ with $\sum_{k=1}^{K-k+1} \hat{Y}_{k,K-k+2}$ (just as in Example 2.2) provides a test of $\hat{A}$.

From (ii),

$$b+1=m^2/s^2 \tag{2.10}$$

$$c=m/s^2 \tag{2.11}$$

where m and s^2 are the mean and variance of j respectively.

Therefore form estimators of m and s^2 as follows. Define P_j to be the same estimator as in (2.6), ie an estimator of π_j ignoring the latter's parametric dependency (2.8). Then define estimators $\hat{m}$ and $\hat{s}^2$ that are the empirical mean and variance of j with respect to the distribution of the P_j. Specifically,

$$\hat{m}=\sum_{j=1}^{J} jP_j \Big/ \sum_{j=1}^{J} P_j \tag{2.12}$$

$$\hat{s}^2=\sum_{j=1}^{J} (j-\hat{m})^2 P_j \Big/ \sum_{j=1}^{J} P_j \tag{2.13}$$

Then (2.10) and (2.11) imply estimates

$$\hat{b}=\hat{m}^2/\hat{s}^2-1 \tag{2.14}$$

$$\hat{c}=\hat{m}/\hat{s}^2 \tag{2.15}$$

These estimates would be used as follows for monitoring. Suppose the valuation model is based on the data set $\{Y_{kj} : k = 1, \ldots, K; j = 1, \ldots, K - k + 1\}$, ie a triangle with K diagonals. Estimates of b and c will have been made on the basis of analysis of this data set.

Consider now the addition of a $(K + 1)$-th diagonal so that the data set becomes $\{Y_{kj} : k = 1, \ldots, K + 1; j = 1, \ldots, K - k + 2\}$. Form estimates $\hat{b},\hat{c}$ according to (2.10) – (2.15) but with K replaced by $K + 1$ in the calculation of P_j by means of (2.6). [Note that, for comparability of $\hat{b},\hat{c}$ with b, c, J should **not** be changed to $J + 1$ in (2.12) and (2.13)].

The resulting estimates $\hat{b},\hat{c}$, based on the triangle of dimension $K + 1$ may be compared with the valuation parameters b, c, derived from the triangle of dimension K.

□

While example 2.3 indicates how to construct a monitoring system with full coverage in the sense defined in Section 1.4, it can be seen to differ from Examples 2.1 and 2.2 in one major respect. In Example 2.3, two valuation parameters, b and c, were monitored by means of quantities $\hat{b}$ and $\hat{c}$ that were constructed from the entire data set, not from the increment in that data set since the valuation under test (the $(K + 1)$-th diagonal).

This situation arose from the coupling of all parameters in each observation and, as a result, the inability to isolate subsets of observations that relate to just individual parameters.

The procedure applied in Example 2.3 can be generalised to any other example in which the model specifies a parametric form as a function of development period *j* and this form has finite moments. Generally, if a model of the sort under test contains *q* parameters, then *q* moments (possibly including the 0-th) may be calculated, as in (2.12) and (2.13), and equated to parametric expressions for those amounts, as in (2.14) and (2.15). This approach is evidently a version of the **method of moments**.

Example 2.4 (payments per claim finalised model). Let Y_{kj} be claim payments as in Examples 2.2 and 2.3. Let F_{kj} denote the number of claim finalisations in the (k, j) cell. Instead of (2.4), define the payments per claim finalised ("PPCF")

$$P_{kj} = Y_{kj} / F_{kj} \tag{2.16}$$

PPCF models are discussed in Section 4.3 of Taylor (2000).

Suppose the Y_{kj} satisfy the following assumptions.

(PPCF1) The $Y_{kj} \mid F_{kj}$ are stochastically independent.

(PPCF2) $Y_{kj} \mid F_{kj} \sim$ Gamma, with

(a) $E\left[Y_{kj} \mid F_{kj}\right] = F_{kj}\pi_j$

(b) $Var\left[Y_{kj} \mid F_{kj}\right] = F_{kj}\sigma_j^2$

(c) $\pi_j = A - B\left(j - \frac{1}{2}\right)^{-\alpha}$ for parameters *A*, *B* and $\alpha > 0,\ \sigma_j^2 > 0$.

This example is more difficult. The function $\pi_j = A - B\left(j - \frac{1}{2}\right)^{-\alpha}$ does not even converge to zero for large *j* and so none of its moments will be finite and the method of moments cannot be applied.

While it would be possible to find functions of the π_j, other than moments, that can be expressed in terms of the parameters, this is not seen as rewarding here and the example is not pursued further. □

The main impression left by Examples 2.1 – 2.4 is of their *ad hoc* nature. In all but the simplest models there is difficulty in identifying functions of the data that relate to individual parameters and so lead to straightforward tests of those parameters. The next section will propose a rigorous basis for the construction of the necessary test statistics.

3. Monitoring as hypothesis testing

3.1 Form of test

3.1.1 Testing the claims model as a whole

Section 1.4 defines monitoring as a test of the hypothesis $\theta = \hat{\theta}$. It is natural therefore to take advantage of the statistical theory of hypothesis testing. The formal statement of the test is as follows.

Test of hypothesis. Assuming that $Y \sim G(y;\theta)$

H_0 (null hypothesis): $\theta = \hat{\theta}$

H_1 (alternative hypothesis): $\theta \neq \hat{\theta}$ □

This will be straightforward if $G(y;\theta)$ is known, which requires specification of the functional relation between observations *Y* and the parameter vector θ, as well as the stochastic error structure. For all practical purposes, this amounts to the formulation of a (possibly non-linear) regression model.

It will be assumed here that $G(y;\theta)$ may be expressed as a **generalised linear model** (GLM) (Nelder & Wedderburn, 1972), ie the d.f. *G* is a member of the exponential dispersion family such that

$$E[Y] = h^{-1}(A\beta) \tag{3.1}$$

where *h* is the link function, *A* the design matrix and, for the sake of conventional notation, the parameter vector has been denoted by β instead of θ.

The null and alternative hypotheses are now $H_0 : \beta = \hat{\beta}$ and $H_1 : \beta \neq \hat{\beta}$ respectively. It will be convenient to rewrite these yet again as $H_0 : \delta\beta = 0, H_1 : \delta\beta \neq 0$ where $\delta\beta = \beta - \hat{\beta}$.

The formulation of the alternative hypothesis in this manner does not account for model error. The existence of any model error would therefore be revealed only indirectly by the proposed hypothesis testing as inconsistencies between the specified model and subsequent observations emerged. Some specific forms of model error are considered further in Section 4.2.

The testing of H_0 is now subject to the well known theory of GLM hypothesis testing (McCullagh & Nelder, 1989, particularly pp. 471-478). This amounts to a likelihood ratio test of H_0 against H_1, carried out as follows.

Consider the GLM formulated above, ie

$$Y \sim G(\mu,\phi)$$

for ϕ a scale parameter and $\mu = E[Y]$ subject to (3.1), ie

$$\mu = h^{-1}\left(A\hat{\beta} + A\delta\beta\right) \tag{3.2}$$

Here $\hat{\beta}$ will have been fixed at the original calibration of the model. Let $\delta\tilde{\beta}$ denote the maximum likelihood estimate (MLE) of $\delta\beta$ within this model, that is on the basis of post-calibration experience $Y = y$.

Let $D(y;\delta\beta)$ denote the deviance associated with this model, ie

$$D(y;\delta\beta) = 2\ell^*(y) - 2\ell(y;\delta\beta) \tag{3.3}$$

where $\ell(y;\delta\beta)$ is the log-likelihood of y for given $\delta\beta$ and $\ell^*(y)$ is the log-likelihood for the saturated model in which the fitted value corresponding to y is y itself.

Then the likelihood ratio test statistic for H_0 against H_1 is

$$T = D(y;0) - D\left(y;\delta\tilde{\beta}\right) \tag{3.4}$$

If the dimension of Y is large, then approximately

$$T \sim \chi_q^2 \tag{3.5}$$

where q is the number of parameters in the model. A large value of T indicates that $Y = y$ is inconsistent with the model. Statistical significance of such inconsistency may be evaluated by means of (3.5).

Note that the data set for the test represented by (3.4) and (3.5) is specifically just **post-calibration experience**, not the full data set that includes pre-calibration experience also.

3.1.2 Testing subsets of the claims model's parameter set

The procedure outlined in Section 3.1.1 tests the claims model as a whole against subsequent data. It will also be of interest to test individual parameters, or possibly larger subsets of the parameter vector. This will be particularly so if the model as a whole fails its test.

Let $S = \{i_1, i_2, ..., i_k\}$ denote a subset of $\{1, 2, ..., q\}$. Let β_S denote the sub-vector of β consisting of the i_1-th, …, i_k-th components of β, and let $\hat{\beta}_S, \delta\beta_S$ have similar meanings. Consider the null and alternative hypotheses $H_0 : \delta\beta_S = 0, H_1 : \delta\beta_S \neq 0$.

Define $\delta\tilde{\beta}^{(S)}$ to be the vector with j-th component

$$\delta\tilde{\beta}_j^{(S)} = 0 \text{ for } j \in S$$
$$= \text{ MLE of } \delta\beta_j \text{ for } j \notin S \qquad (3.6)$$

In other words, $\delta\tilde{\beta}^{(S)}$ is the MLE of $\delta\beta$ when the components $\delta\tilde{\beta}_j^{(S)}, j \in S$ are held to zero.

The likelihood ratio test statistic for H_0 against H_1, replacing (3.4), is now

$$T_S = D\left(y;\delta\tilde{\beta}^{(S)}\right) - D\left(y;\delta\tilde{\beta}\right) \qquad (3.7)$$

The large sample asymptotic result corresponding to (3.5) is

$$T_S \sim \chi_k^2 \qquad (3.8)$$

An individual parameter may be tested by setting S to be a singleton. Thus, if $S = \{i\}$, then (asymptotically)

$$T_{\{i\}} \sim \chi_1^2 \qquad (3.9)$$

with $T_{\{i\}}$ given by (3.7).

Some care is required in the interpretation of significance tests based on statistics such as (3.8) and (3.9). For example, if all model parameters are tested individually by means of (3.9) at the 100*p*% significance level, the expected number of parameter increments $\delta\tilde{\beta}$ statistically different from zero will be *pq*. Testing whether **any** of these increments is statistically different from zero is a much stricter than a significance test of the entire model based on (3.5).

3.2 Numerical implementation

The use of GLM software (SAS GENMOD, R, etc) renders the hypothesis setting set out in Section 3.1 extremely simple. The theory set out there may be simply skipped.

3.2.1 Testing the claims model as a whole

Consider the test (3.4) – (3.5) established in Section 3.1.1 and return to the interpretation of the null and alternative hypotheses $H_0 : \beta = \hat{\beta},\ H_1 : \beta = \hat{\beta} + \delta\beta,\ \delta\beta \neq 0$. The deviance associated with H_0, the term $D(y;0)$ in (3.4), is obtained by setting the **offset** vector in the GLM (see McCullagh & Nelder, 1989) equal to $\hat{\beta}$ and carrying out no further fit.

Likewise, the deviance associated with H_1, the term $D\left(y;\delta\tilde{\beta}\right)$ in (3.4), is obtained by setting the same offset and then re-fitting all covariates in the model on the basis of **post-calibration experience**. The statistic T in (3.4) is then simply the amount of the decrease in deviance that arises from this re-fit.

In summary, the test procedure consists of the following steps:

1. Set the offset vector to $\hat{\beta}$, to obtain deviance $D(y;0)$.
2. Re-fit all covariates to obtain deviance $D\left(y;\delta\tilde{\beta}\right)$.

Then the test statistic T in (3.4) is the amount by which the deviance decreased by virtue of Step 2.

Some GLM software packages automatically output the change in deviance at the fit of any model. This covers the case of Step 2 above.

3.2.2 Testing subsets of the claims model's parameter set

The argument here is quite parallel to that given in Section 3.2.1 but applied to (3.7) instead of (3.4). The procedure consists of the following steps:

1. Set the offset vector to $\hat{\beta}$, as before.
2. Re-fit all covariates other than $i_1, i_2, ..., i_k$ on the basis of **post-calibration experience**. This gives deviance $D\left(y;\delta\tilde{\beta}^{(S)}\right)$.
3. Re-fit all covariates, to obtain deviance $D\left(y;\delta\tilde{\beta}\right)$.

Then the test statistic T_S in (3.7) is the amount by which the deviance decreased by virtue of Step 3.

Consider the testing of a single parameter, as in (3.9). In this case, it will usually be unnecessary to carry out the above procedure explicitly. Most GLM software, in re-fitting all covariates (Step 2 in Section 3.2.1) will produce a χ_1^2 statistic for each. The χ_1^2 statistic associated with covariate i is precisely $T_{\{i\}}$ in (3.9).

Thus, in carrying out the test of the whole valuation model, as in Section 3.2.1, using GLM software, one usually obtains a test of each individual parameter in addition.

3.3 Full coverage

Sections 3.1 and 3.2 clarify the meaning of the concept of full coverage introduced in Section 1.4. The test forecasts $Y_1, ..., Y_q$ defined there are seen to be $T_{\{1\}}, T_{\{2\}}, ..., T_{\{q\}}$. This set of test forecasts tests the consistency of post-calibration data with each model parameter $\beta_1, ..., \beta_q$.

4. Other components of monitoring

4.1 Macro-control

The test forecasts defined in Section 3 test the micro-structure of the predictive model. However, it will also be desirable to test the accuracy with which the model is forecasting its primary targets.

If, for example, the model is a valuation model, its primary target is the quantum of incurred but unpaid liabilities, so it is desirable to test the accuracy of the prediction of that quantity. Otherwise, there would be a possibility of the model performing well in the fine detail of its forecasts but failing at the macro-level.

The performance of monitoring at this level requires the identification of the model's primary targets. These will not be internal to the model, as were the parameters under test in Section 3. Rather they will be determined by the business purpose to which the model is put.

Once the primary target forecasts have been determined, each can be tested as in the following sub-sections.

4.1.1 General form of macro-testing

Let the primary target quantities estimated by the model at time t be denoted L_{it}, i = 1,2, etc. Let $\hat{L}_{it|s}$ denote an estimate of L_{it} made on the basis of data up to time s. For example, if the model is a liability valuation model, then $\hat{L}_{it|t}$ will be the valuation estimate of liabilities at time t and $\hat{L}_{it|s}$ for $s > t$ will be a hindsight estimate of the same quantity on the basis of data to time s.

In the notation of Section 3, $\hat{L}_{it|t}$ will be a function of $\hat{\beta}$, so write $\hat{L}_{it|t}\left(\hat{\beta}\right)$.

Now consider the situation at time t + 1. Data from the time interval (t, t + 1] has now been realised, and the estimate of parameter vector β has changed from $\hat{\beta}$ to $\hat{\beta}+\delta\tilde{\beta}$. Thus there is a new estimate $\hat{L}_{it|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right)$ of L_{it}. Not only has the estimate of β changed but also any forecast of data from the interval (t, t + 1] contained in $\hat{L}_{it|t}\left(\hat{\beta}\right)$ will have been replaced by observations in $\hat{L}_{it|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right)$, ie $\hat{L}_{it|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right)$ is a hindsight estimate of L_{it} taking account of data up to t + 1.

Define

$$\Delta_{it} = \hat{L}_{it|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right) - \hat{L}_{it|t}\left(\hat{\beta}\right) \tag{4.1}$$

which is the shift in estimate of L_{it} by virtue of data from (t, t + 1].

Now, from the viewpoint of time t, $\hat{L}_{it}\left(\hat{\beta}\right)$ is known but $\hat{L}_{it|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right)$ is a random variable. Thus Δ_{it} is also a random variable.

The hypotheses for testing $\hat{L}_{it|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right)$ are as follows.

$$H_0: \quad E\left[\Delta_{it} \mid \mathcal{F}_t\right] = 0$$

$H_1: \quad E\left[\Delta_{it} \mid \mathcal{F}_t\right] \neq 0$

where $\mathcal{F}_t$ denotes data up to time *t*.

Now the distribution of $\hat{L}_{it|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right) \mid \mathcal{F}_t$, and hence of $\Delta_{it} \mid \mathcal{F}_t$, is estimable from the model at time *t*. Let *F*(.) denote the estimated d.f.

Then, for any given $\delta > 0$,

$$\text{Prob}\left[\mid \Delta_{it} \mid > \delta \mid \mathcal{F}_t\right] = F\left(-\delta\right) + \left[1 - F\left(\delta\right)\right]$$

and so the statistical significance of any observed departure of Δ_{it} from zero may be calculated.

There are two main categories of primary target quantities L_{it}, those that depend on observations only up to time $t + 1$, and those that depend on observations beyond that time. Call these Categories I and II respectively.

Category I targets will be "resolved" by time $t + 1$ in the sense that their estimates from time *t* will have been replaced by observations. Category II targets will have been only partially resolved and, to the extent that they depend on observations beyond time $t + 1$, will still be estimated at that time.

In the case of Category I, the "new estimate" $\hat{L}_{it|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right)$ is simply ℓ_{it} the value of L_{it} observed in data $\mathcal{F}_{t+1}$, and (4.1) degenerates to

$$\Delta_{it} = \ell_{it} - \hat{L}_{it|t}\left(\hat{\beta}\right). \tag{4.2}$$

4.1.2 Examples

To illustrate these concepts, consider for example, a projected case estimates ("PCE") model. It consists of development factor and payment factor sub-models, as described in Section 4.4 of Taylor (2000). There are various possibilities as to the model's primary targets. A few of them are considered in the following paragraphs.

Development factors as primary target

Development factors consistently close to unity indicate accurate case estimation. Factors chronically different from unity indicate poor case estimation. So, if the accuracy of case estimation is the test objective, the development factors may be the primary targets.

The development factor sub-model will have been subjected to testing of its micro-structure along the lines of Section 3. It might be tested at the macro-level by the computation of a portfolio-wide development factor for the post-valuation period and its comparison with the model forecast of this factor. A similar comparison might be made with respect to development factors for portfolio segments.

The observable post-valuation development factors are those relating to the interval $(t, t + 1]$. These depend on observations only up to time $t + 1$, and so the target belongs to Category I. Hence the macro-test statistic is given by (4.2).

Suppose that a development factor for the i-th claim over the time interval $(t, t + 1]$ takes the form

$$D_{i,t+1} = \left(C_{i,t+1} + Y_{i,t+1}\right) / C_{it} \tag{4.3}$$

Where $C_{i,t+s}$ denotes the case estimate at time $t + s$ and $Y_{i,t+1}$ denotes claim payments during the interval $(t, t + 1]$.

Then, for any subset $\mathcal{A}$ of the portfolio of claims, the aggregate development factor is

$$D_{t+1}\left(\mathcal{A}\right) = \sum_{i \in \mathcal{A}} \left(C_{i,t+1} + Y_{i,t+1}\right) / \sum_{i \in \mathcal{A}} C_{it} \tag{4.4}$$

The statistics $D_{t+1}\left(\mathcal{A}\right)$ take the role of ℓ_{it} in (4.2).

The subset $\mathcal{A}$ may be chosen to be the entire portfolio of claims. It may also be chosen to be any proper subset of interest. If, as is typical, the model prediction of $E\left[D_{i,t+1}\right]$ depends heavily on development period (ie age of claims), then the subsets $\mathcal{A}$ might be chosen to partition the portfolio with respect to development period.

Claim payments as primary target

Suppose that the same PCE valuation model is used to forecast claim payment cash flows for the purpose of asset allocation. The observable post-valuation claim payments are those relating to the interval $(t, t + 1]$. These depend on observations only up to time $t + 1$, and so the target belongs to Category I. Hence the macro-test statistic is given by (4.2). Such forecasts will be provided by the payment factor sub-model. Suppose they take the form

$$\hat{Y}_{i,t+1} = C_{it} \hat{\pi}_{i,t+1} \tag{4.5}$$

where $\pi_{i,t+1}$ is the payment factor in respect of the i-th claim over the time interval $(t, t + 1]$ and $\hat{\pi}_{i,t+1}$ is its forecast from the payment factor sub-model.

In fact, the use of a PCE model for the forecast of claim payments would be relatively unusual. This fact will be set aside, however, for the sake of the present example.

Again, comparisons of actual and forecast are made, both portfolio-wide and for subsets. The aggregate forecast of claim payments for subset $\mathcal{A}$ of claims is

$$\hat{Y}_{t+1}\left(\mathcal{A}\right) = \sum_{i \in \mathcal{A}} C_{it} \hat{\pi}_{i,t+1} \tag{4.6}$$

The corresponding payments actually observed over (t, t + 1] take the role of ℓ_{it} in (4.2).

Liability valuation as primary target

Suppose that the primary use of the PCE model has been the estimation of the amount of incurred but unpaid liabilities at time t. Let L_{it} denote the liability in respect of the i-th claim and let $\hat{L}_{it|t}$ denote its estimate on the basis of data up to time t. Note that, due to IBNR claims, the existence of claims associated with some values of i may be unknown, and the number of values of i may be stochastic.

The liability L_{it} depends on the observed cost of claims over the time interval (t,∞) and so belongs to Category II of targets. Hence macro-testing needs to proceed on the basis of the general test statistic (4.1).

For this purpose

$$\hat{L}_{it|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right)=L_{i,t+1|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right)+Y_{i,t+1} \tag{4.7}$$

where $Y_{i,t+1}$ denotes claim payments during (t, t + 1] in respect of the i-th claim.

Substitution of (4.7) into (4.1) yields

$$\Delta_{it}=Y_{i,t+1}+\left[L_{i,t+1|t+1}\left(\hat{\beta}+\delta\tilde{\beta}\right)-L_{it|t}\left(\hat{\beta}\right)\right] \tag{4.8}$$

which may be interpreted as the amount of claims incurred during (t, t + 1] when loss reserves are established according to the valuation model, since the right side of (4.8) takes the form:

payments in (t, t + 1] + change in estimated liability between t and t + 1.

As in the previous examples, it may be useful to calculate $\Delta_{it}\left(\mathcal{A}\right)$ for subsets $\mathcal{A}$ of the portfolio. It would be common for the subsets to be defined in terms of accident periods but other definitions are possible, eg in terms of injury severity in the case of a bodily injury portfolio.

The statistics $\Delta_{it}\left(\mathcal{A}\right)$ would be tested for significance according to the hypotheses and significance tests set out in Section 4.1.1.

4.2 Extra-model monitoring

Section 3.1 expressed the monitoring of a model's fine structure in the form of hypothesis testing with null hypothesis $\theta=\hat{\theta}$, where $\hat{\theta}$ is a q-vector of parameter estimates.

It might be remarked that the selection of the q parameters excludes many other potential parameters. It is possible, therefore, to contemplate an augmented model with (q + r) – vector of parameters θ^+ and null hypothesis

$$H_0^+ : \hat{\theta}^+ = \theta^+ = \left(\theta^T, 0^T\right)^T$$

where 0 here denotes an r-vector of zeros.

There might be specific covariates of interest for inclusion in the additional r dimensions of the augmented model. They would include covariates such as superimposed inflation which, while currently excluded from the model, are perennially at risk of inclusion.

Testing of such covariates would require their formal inclusion in the model at time t + 1 and significance testing of the hypothesis H_0^+.

5. Conclusion

Claims experience monitoring is widely practiced by insurance companies, especially in relation to pricing and valuation models. However, there is no body of theory to guide it. This creates a risk that it will be misdirected or incomplete. The present paper is an attempt to formulate a theoretical basis.

The form of monitoring to be applied to a model depends partly on the purpose of that model. Primary targets need to be identified among the multiplicity of forecasts generated by the model.

Claims experience monitoring has been viewed here as comprising three components:

- **Macro-testing**: the testing of the extent to which post-model experience is consistent with the primary targets.
- **Micro-testing**: the testing of the fine structure of the model, with individual tests of the extent to which post-model experience is consistent with individual model parameters.
- **Testing for missing covariates**: testing for covariates that have been omitted from the model but whose omission is not supported by post-model experience.

All of these components of testing are formulated in hypothesis testing terms. If the model is a GLM, then the micro-testing may be formulated as a collection of likelihood ratio tests, one per model parameter. When all model parameters are covered in this way, the monitoring is said to have full coverage.

References

Berry J, Hemming G, Matov G & Morris O (2009). Report of the Model Validation and Monitoring in Personal Lines Pricing Working Party. Paper presented to the **GIRO Convention**, Edinburgh, October 2009. Institute of Actuaries and Faculty of Actuaries.

De Jong P & Zehnwirth B (1983). Claims reserving state space models and the Kalman filter. **Journal of the Institute of Actuaries**, 110, 157-181.

McCullagh P & Nelder JA (1989). **Generalised linear models** (2nd ed.). Chapman & Hall, London UK.

Mack T (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. **Astin Bulletin**, 23(2), 213-225.

Nelder JA & Wedderburn RWM (1972). Generalised linear models. **Journal of the Royal Statistical Society**, Series A, 135, 370-384.

Taylor G (2000). **Loss reserving: an actuarial perspective.** Kluwer Academic Publishers, Dordrecht, Netherlands.

Taylor G (2010). Maximum likelihood and forecast efficiency of the chain ladder. Submitted for publication.

Wright TS (1990). A stochastic method for claims reserving in general insurance. **Journal of the Institute of Actuaries**, 117, 677-731.