REGRESSION MODELS IN CLAIMS ANALYSIS I: THEORY

GREG C. TAYLOR

Abstract

This paper considers the application of regression techniques to the analysis of claims data. Examples are given to indicate why, in certain circumstances, this might be preferable to traditional actuarial methods.

The various errors of prediction which occur when loss reserves are estimated by regression are classified and discussed.

Formal procedures are discussed for determining which of the available predictors will be entered into a regression, and the drawbacks of these procedures.

Various approaches to the estimation of uncertainty associated with loss reserves estimated by regression are considered.

The effect on regression techniques of outlying data points, and hence the subject of robust/resistant regression, is considered briefly.

1. INTRODUCTION

Regression models have not been prevalent in claims analysis leading to loss reserving. This is evident from a survey of claims reserving methods (Taylor, [23]).

The scarcity arises from the suspicion with which many actuaries regard such models. Their use does not have the "hands on" nature characteristic of methods based on age-to-age factors, for example, with which actuaries tend to feel at ease. There is a feeling of abstractness and loss of control in the estimation of parameters from the data.

This skepticism is justified by the countless misapplications of regression methods which occur in practice. Despite this, it appears that regression techniques have a very definite place in the actuarial repertoire. But they will serve their users effectively only if it is realized that blind and mechanical application of simple least squares regression will, in certain circumstances, be statistically inefficient.

In these circumstances, regression becomes a delicate tool rather than the crude bludgeon as which it is often regarded, and in which role it is even more often used. A proposition which is all too often neglected in practice is that a user can expect effective performance of any body of methodology only if the user is aware of its general properties, its strengths and weaknesses, the circumstances in which it should and should not be applied, the response of its output to input anomalies, the whole array of quirks and pitfalls awaiting the unwary, how to "tune" the model building procedure for maximum results, and so on.

The intention of this paper is to canvass briefly the various aspects of regression modelling. Within this larger purpose, there are two intentions. First, some of the grosser abuses of such modelling will be suitably exposed. Second, from a more positive viewpoint, it is hoped that the exposure of the causes of anomalous regression output will set the procedures in a perspective from which their beneficial aspects can be more clearly seen.

The following sections deal very briefly with such questions as:

(i) Why use regression models as opposed to the "traditional" actuarial ones such as those using age-to-age factors?

(ii) Precisely what criteria are to be satisfied, and how should the extent to which they are satisfied be assessed?

(iii) How many of the available predictors should be included in a regression model, and how should the choice be made?

(iv) What procedures, other than ordinary least squares regression, are available for fitting the selected model to data?

(v) How might the impact on the fitting of isolated rogue data points be assessed, and how might the fitting procedures be modified to reduce this impact?

2. MOTIVATING EXAMPLES

Consider first a relatively complex example. A simpler one will be presented shortly.

In what follows, let

i = year of occurrence of claim;
j = development year, i.e., number of years after year of occurrence;
N_i = number of claims incurred in year of occurrence i;
N_{ij} = number of claims settled in (i,j);
C_{ij} = amount of claim payments (adjusted for claims escalation) in (i,j);
S_{ij} = C_{ij}/N_{ij} = average claim payment per settlement in (i,j);
F_{ij} = N_{ij}/N_i = rate of settlement in (i,j);
$t_i(j)$ = $\Sigma^j_{k=0} N_{ik}/N_i$ = proportion of claims from year of occurrence i settled by the end of development year j;
$\bar{t}_{ij}(k)$ = $\min(\frac{1}{2}[t_i(j) + t_i(j + 1)],u_k)$ for some partition $\{u_0, \ldots, u_{n+1}\}$ of $[0,1]$.

Suppose that the following model has been suggested:

$$S_{ij} = a + \sum_{k=0}^{n} b_k\bar{t}_{ij}(k) + c/F_{ij} + e_{ij}, \tag{2.1}$$

where a, $b_0, \ldots, b_n$, and c are unknown parameters and e_{ij} is a random error term. This is the invariant see-saw model (Taylor, [22]).

Formula (2.1) expresses S_{ij} as a linear function of the observations $\bar{t}_{ij}(0), \ldots, \bar{t}_{ij}(n)$, $1/F_{ij}$ and a random error. Evidently, the unknown parameters may be determined by some form of linear regression of the S_{ij} on these observations.

Indeed, how else might the parameter estimation be carried out? Note that the parameter values a, $b_0, \ldots, b_n$, c are common to all cells (i,j). In contrast with the example below involving age-to-age factors, there is no simple transformation of the dependent variable S_{ij} which will isolate any one of the parameters.

In this example, the very "shape" of the model, the intertwining of dependent and independent variables, virtually demands regression for parameter estimation.

In the next example, a much simpler model is considered but the situation is somewhat subtler. Using the same notation as before, let

$$C_{ij} = N_i u_i r_j + e_{ij}, \tag{2.2}$$

where

u_i = average claim size (adjusted for claims escalation) experienced in year of occurrence i;

r_j = the average proportion of claim payments (again adjusted for claims escalation) deriving from year of origin i which are payable in development year j.

The model (2.2) may be rewritten in the form:

$$\log C_{ij} = \log (N_i u_i) + \log r_j + f_{ij}, \tag{2.3}$$

where f_{ij} is a new random error term. The transformed model (2.3) is linear in the parameters $\log (N_i u_i)$ and $\log r_j$ which may therefore be estimated by regression methods. This indeed is the basis of Kremer's [13] ANOVA approach.

Note also, however, that (2.2) is the prototype for development of age-to-age factors (e.g., Skurnick, [20]; Berquist and Sherman, [3]). This is because it implies

$$C_{i,j+1}/C_{ij} = r_{j+1}/r_j + \text{error term}, \tag{2.4}$$

or more commonly,

$$A_{i,j+1}/A_{ij} = \sum_{k=0}^{j+1} r_k / \sum_{k=0}^{j} r_k + \text{error term}, \tag{2.5}$$

where

$$A_{ij} = \sum_{k=0}^{j} C_{ik} = \text{total claim payments (adjusted for}$$

claims escalation) made in respect of year of occurrence i up to the end of development year j,

and the r_{j+1}/r_j in (2.4), or the $\sum_{k=0}^{j+1} r_k / \sum_{k=0}^{j} r_k$ in (2.5), are the age-to-age factors.

This example is subtler than the previous one in the sense that one has a *choice* as to the method of estimation of its parameters. This choice should be made against reasonable criteria, and therefore one needs to specify these.

Consider, for example, the following three possibilities:

(i) proceed with regression estimation of the log (N_iu_i) and log r_j via (2.3) after appropriate specification of f_{ij};

(ii) ignore the error term in (2.5) and estimate $\sum_{k=0}^{j+1} r_k / \sum_{k=0}^{j} r_k$

by $A_{i,j+1}/A_{ij}$;

(iii) assume the vector (log r_0, log r_1, . . .) to lie within some finite-dimensional vector space spanned by $g_1, g_2, \ldots, g_s$ where $g_m = (g_{m0}, g_{m1}, \ldots)$, and so use regression methods to fit the following adaptation of (2.3):

$$\log C_{ij} = \log (N_iu_i) + \sum_{m=1}^{s} b_m g_{mj} + f_{ij}. \tag{2.6}$$

It is instructive to consider the number of parameters to be estimated in each case.

In case (i), there are $I + J - 1$ parameters if I values of i are considered and r_j is assumed zero for $j = J, J + 1$, etc. The -1 arises from the constraint

$$\sum_{k=0}^{\infty} r_j = 1, \tag{2.7}$$

by definition.

In case (ii), there are again $I + J - 1$ parameters, the only difference between the two cases being that the former approaches parameter estimation in a formal manner whereas the latter takes an *ad hoc* approach.

In case (iii), where the g_m are fully specified in advance, the parameters $b_1, \ldots, b_s$ and the log (N_iu_i) number just $I + s$.

Note that in the last case the number of parameters is independent of J. This contrasts with the first two cases in which increasing J without limit increases the number of model parameters also without limit. For example, consider the case $I = J = 10$, $s = 3$. The numbers of model parameters are:

Case (i):	19
Case (ii):	19
Case (iii):	13.

Case (iii) involves only two-thirds as many parameters as cases (i) and (ii).

Note that the number of parameters in case (iii) may be reduced further by treating the log $(N_i u_i)$ in the same way as the log r_j and representing them in some vector space of reduced dimension.

Every actuary is aware, intuitively at least, of the dangers of over fitting, i.e., fitting a model involving more parameters than are justified by the volume of available data. Generally, the fitting of models which are parsimonious in their use of parameters smoothes out the roughness inherent in the raw observations. Increasing the number of model parameters diminishes this smoothing effect until ultimately, when there are enough parameters that they are in one-to-one correspondence with the observations, the instability of the parameter estimates is equal to that of the observations themselves.

Many actuaries have a distaste for models like (2.6) on the ground that the parameters b_m under estimation are too abstract, that they do not correspond sufficiently with real world objects. This is what I meant in referring at the start of Section 1 to the "hands on" nature of the more traditional actuarial models.

The formal objection to models such as (2.6) is likely to take the form: "What if basis vectors $b_1, \ldots, b_s$ cannot be found (for s sufficiently small to be useful) which capture the more subtle features of the r_j?"

The answer is that any such losses of accuracy cannot be considered in isolation from possible gains in stability accruing from a reduction in the number of model parameters requiring estimation. In formal terms, the approximation of (2.3) by (2.6) may introduce some *bias* into the model, but this bias must be weighed against any reduction in *variability* of the model's predictions.

The distaste for abstraction that individuals may experience is perhaps understandable, but ultimately the relative merits of competing models must be assessed by the models' objective performance, rather than the users' preferences or prejudices.

The above remarks concerning questions of bias versus stability do no more than state the intuitively obvious. However, it is possible, and useful, to formalize the concepts involved so that model selection (such as the choice between cases (i), (ii) and (iii) dealt with above) can proceed on a more rigorous basis.

These matters are pursued in Section 4. A helpful preliminary to this is an examination and classification of the types of error that arise in the prediction of future observations on the basis of a model fitted to past data. This forms the subject of Section 3.

3. ERRORS OF PREDICTIONS

3.1. *Illustrative example*

Again it will be useful to consider an example which is highly simplified but nevertheless illustrative of the wrong turns that can be taken in a slipshod approach to errors of prediction. Though oversimplified, the essence of the model corresponds to some of the approaches which I have seen in practice.

Suppose it is assumed for the model (2.2) that:

$$V[C_{ij}|\{u_i, r_j\}] = N_i u_i^2 \sigma^2, \tag{3.1.1}$$

where σ^2 is independent of both i and j.

Suppose also, that estimates $\hat{u}_i$, $\hat{r}_j$ of the u_i, r_j have been obtained in the manner described in case (ii) of Section 2. Hence, estimates $\hat{C}_{ij}$ corresponding to the observations C_{ij} have been found. More particularly, though, predictions $\hat{P}_i = \Sigma_{j=T+1}^{\infty} \hat{C}_{ij}$ have been obtained of the future claim payments $P_i = \Sigma_{j=T+1}^{\infty} C_{ij}$ arising from year of occurrence i, where C_{iT} is the latest observation on that year of occurrence.

Suppose that one seeks:

$$V[\hat{P}_i] = \sum_{j=T+1}^{\infty} V[\hat{C}_{ij}] + \text{covariances}. \tag{3.1.2}$$

In practice, the estimation of the covariances may prove awkward. However, let us concentrate for the moment on some of the pitfalls involved in the estimation of the $V[\hat{C}_{ij}]$.

An argument that seems to appeal to some practitioners begins by considering the scaled residuals $(C_{ij} - \hat{C}_{ij})/N_i^{1/2}\hat{u}_i$. If $\hat{C}_{ij}$ is regarded as replaceable by $E[C_{ij}]$, assuming $\hat{C}_{ij}$ to be unbiased, the squares of these residuals become estimators of σ^2, i.e.,

$$\hat{\sigma}^2 = n^{-1} \sum_{i,j} [(C_{ij} - \hat{C}_{ij})^2 / N_i \hat{u}_i^2], \tag{3.1.3}$$

the summation running over the n pairs i,j for which observations exist, and perhaps with some reduction of n to reflect loss of degrees of freedom. The required estimate of $V[\hat{C}_{ij}]$ can then be obtained by means of (3.1.1) as:

$$n^{-1} N_i \hat{u}_i^2 \sum_{k,j} [(C_{kj} - \hat{C}_{kj})^2 / N_k \hat{u}_k^2]. \tag{3.1.4}$$

While this procedure may appear a reasonable practical solution to the problem, uncluttered by the quibbles of purists, it is suggested that it is in fact far from the truth. It is suggested further to contain a major error of reasoning likely to carry substantial numerical consequences. $V[\hat{P}_i]$ is not even the second moment of interest. Even if it were, material contributions to it have been omitted.

Essentially, the difficulties arise from the cavalier approach to the problem. A more careful and organized approach is required.

3.2. Component errors of prediction

To achieve the requirement of the previous subsection, let us drop the particular problem we have been considering and consider a generalized problem instead. Let Y denote an observable n-vector whose ith component is, apart from random noise, some function of observable quantities $X_{i1}, \ldots, X_{ip}$:

$$Y = f(X) + e, \tag{3.2.1}$$

where X is the $n \times p$ matrix with X_{ij} as (i,j)-element, f: $R^{np} \rightarrow R^n$ has the particular (possibly non-linear) form described above, and e is a random error term with zero mean.

Suppose that the functional form f is unknown in this context and consider linear approximations Xb to $f(X)$ where b is a p-vector of parameters. Then (3.2.1) becomes:

$$Y = Xb + [f(X) - Xb] + e. \tag{3.2.2}$$

Suppose further that the exact set of independent variables on which Y depends (the columns of X) is unknown, and that as a consequence Y is modelled as a linear function of a subset of Y, i.e., Y_i is modelled by:

$$\sum_{j \in A} X_{ij} b_j \tag{3.2.3}$$

for some $A \subset \{1, 2, \ldots, p\}$ instead of by $\sum_{j=1}^{p} X_{ij} b_j$.

Let (3.2.3) be denoted by $X_A b_A$, whereupon (3.2.2) decomposes as:

$$Y = X_A b_A + X_B b_B + [f(X) - Xb] + e, \tag{3.2.4}$$

where B denotes the set $\{1, 2, \ldots, p\} - A$.

Let $\hat{b}_j$ denote the regression estimate of b_j, where the term "regression estimate" is deliberately left vague for the moment. Let X^* denote an $m \times p$ matrix, each column of which represents m further values of the relevant predictor. The task is to predict the m-vector

$$Y^* = f^*(X^*) + e^*, \tag{3.2.5}$$

where now $f^*: R^{mp} \longrightarrow R^m$.

Corresponding to (3.2.4):

$$Y^* = X_A^* b_A + X_B^* b_B + [f^*(X^*) - X] + e^* \tag{3.2.6}$$

Let $\hat{Y}^*$ be the regression prediction of Y^*:

$$\hat{Y}^* = X_A^* \hat{b}_A, \tag{3.2.7}$$

so that the prediction error is:

$$\begin{aligned} Y^* - \hat{Y}^* &= X_A^*(b_A - \hat{b}_A) + X_B^* b_B + [f^*(X^*) - X] + e^* \\ &= X_A^*(E\hat{b}_A - \hat{b}_A) + [X_A^*(b_A - E\hat{b}_A) + X_B^* b_B] \\ &\quad + [f^*(X^*) - X] + e^*. \end{aligned} \tag{3.2.8}$$

In many applications X represents observation of the predictors in the past, and X^* represents values to be assumed by the same predictors in the future.

At this point it is convenient to stop and consider the components of prediction error appearing on the right side of (3.2.8). They are:

(i) the *specification error* $[f^*(X^*) - X]$ essentially due to unmodeled nonlinearity;
(ii) the *selection error* $[X_B^* b_B + X_A^*(b_A - E\hat{b}_A)]$ due to incorrect selection of predictors;
(iii) the *estimation error* $X_A^*(E\hat{b}_A - \hat{b}_A)$ arising from the fact that even the most efficient estimators of the regression coefficients are still only random variables; and,
(iv) the *statistical error* e^* reflecting the inherent random noise in the process.

The terminology in (i), (iii), and (iv) is taken from Bartholomew [2]. The terminology in (ii) is taken from Miller [15].

By the first version of (3.2.8), it might appear simpler to regard $X_A^*(b_A - \hat{b}_A)$ as estimation error and $X_B b_B^*$ as selection error. Note, however, that the selection of the set A of (linear) predictors instead of $A \cup B$ introduces a bias in $\hat{b}_A$ as estimator of b_A. For example, in the case of ordinary least squares regression with no specification error,

$$\hat{b}_A = (X_A^T X_A)^{-1} X_A^T Y,$$

whence (3.2.4) yields

$$E\hat{b}_A = b_A + (X_A^T X_A)^{-1} X_A^T X_B b_B. \tag{3.2.9}$$

In the case of claims analysis, it is possible to characterize the four contributions to prediction error as follows.

In the fitting of a model to past claims data, the wrong algebraic model structure may be chosen. This will lead to specification error.

Suppose that the true underlying model is in fact linear, and all of the relevant predictors are identified, so that there is no specification error. Still it will usually be necessary to use past data (incorporating its random noise) to decide which of the available predictors are included in the model. The noise in the process may lead to wrong decisions; relevant predictors may be omitted, and irrelevant ones included. This will result in selection bias.

Suppose that the true underlying model is linear and is correctly selected, so that there is neither specification nor selection error. Still it will be necessary to estimate the parameters of the linear model by reference to past data. As these data contain random noise, so will the parameter estimates. The deviation of these estimates from their true values constitutes estimation error.

Suppose that, by some unspecified means, it were possible to select the correct (linear) model form and estimate its parameters precisely, so that there were no specification, selection, or estimation error. Even then future claims experience could not be predicted with precision because the inherent randomness of the claims process would generate deviations of experience from expected values. These deviations constitute statistical error.

3.3. Prediction bias and mean square error of prediction

Let us now consider the *prediction bias* $E\hat{Y}^* - EY^*$ and the *mean square error of prediction* (MSEP)

$$E(Y^* - \hat{Y}^*)^2 = E(Y^* - \hat{Y}^*)^T (Y^* - \hat{Y}^*).$$

By (3.2.6) and (3.2.7) the prediction bias is:

$$E\hat{Y}^* - EY^* = X_A(E\hat{b}_A - b_A) - X_B - [f^*(X^*) - X]. \tag{3.3.1}$$

In finding an expression for the MSEP, it will be advantageous to decompose the prediction error as:

$$\begin{aligned} Y^* - \hat{Y}^* &= (Y^* - EY^*) - (\hat{Y}^* - E\hat{Y}^*) - (E\hat{Y}^* - EY^*) \\ &= (Y^* - EY^*) - (\hat{Y}^* - E\hat{Y}^*) - \text{prediction bias}. \qquad (3.3.2) \end{aligned}$$

In any form of linear regression of Y on X_A, Y^* and $\hat{Y}^*$ will be uncorrelated (easily checked from first principles since the former depends on future observations and the latter on past), so that (3.3.2) yields:

$$\begin{aligned} \text{MSEP} &= E(Y^* - EY^*)^2 + E(\hat{Y}^* - E\hat{Y}^*)^2 + (\text{prediction bias})^2 \\ &= E(e^*)^2 + E[X^*(\hat{b}_A - E\hat{b}_A)]^2 + (\text{prediction bias})^2, \qquad (3.3.3) \end{aligned}$$

by (3.2.5) and (3.2.7).

The MSEP is thus seen to comprise three identifiable contributions deriving from:

(i) statistical error;
(ii) estimation error; and,
(iii) prediction bias (incorporating specification error and selection error).

It is convenient at this point to revert to the example of Section 3.1, recalling particularly the critical remarks made at the end of that section.

With the benefit of the more formal analysis of Section 3.2 and the present subsection, it is possible to recognize that the expression (3.1.4) for $V[\hat{P}_i]$ is essentially only estimation error. Both statistical error and prediction bias are omitted.

3.4 Components of selection error

Section 3.2 defined selection error as the term $[X_B^* b_B + X_A^*(b_A - E\hat{b}_A)]$ in (3.2.8). As seen in (3.3.1), this is the part of prediction bias *not* arising from nonlinearity. It was shown in Section 3.2 that the first member represents the bias introduced directly by the omission of the set B of predictors; the second member is the bias in $\hat{b}_A$ arising from this omission.

It must now be recognized that E$\hat{b}_A$ has been implicitly regarded as an unconditional expectation in the above. This would be appropriate if the set A were chosen without reference to the data Y. In practice, however, and particularly in claims analysis, this will not be the case. Usually, A will be chosen because it produces a better fit of model to data than certain other sets.

In this case,

$$E\hat{b}_A = E[\hat{b}_A|P] + \{E\hat{b}_A - E[\hat{b}_A|P]\}, \tag{3.4.1}$$

where $E\hat{b}_A$ is now explicitly the unconditional expectation of $\hat{b}_A$ and P denotes the procedure for subset selection. Substitution of (3.4.1) in the expression for selection error given at the start of this subsection yields:

$$\text{selection bias} = X_B^* b_B + X_A^*\{b_A - E[\hat{b}_A|P]\} + X_A^* \{E[\hat{b}_A|P] - E\hat{b}_A\}. \tag{3.4.2}$$

There are now three contributions to selection bias:

(i) *omission bias,* consisting of the first two members on the right of (3.4.2), and representing the bias due to the omission of the set B of predictors;
(ii) *stopping rule bias,* consisting of that part of the final member of (3.4.2) which arises from the limitation imposed by P on the number of predictors included in A; and
(iii) *competition bias,* consisting of that part of the final member of (3.4.2) which, for a given size of set A, arises from the manner in which P selects A from subsets of $A \cup B$ of that size.

These components of selection error are discussed in some detail by Miller [15] (pp. 400–405), who gives various other references.

Miller also gives a simple example of competition bias in a case in which:

(i) $A \cup B$ consists of just 2 predictors;
(ii) A consists of just a single predictor;
(iii) P consists of selection of the single predictor according to ordinary least squares;
(iv) $E\hat{b}_1 = 1$ and $V[\hat{b}_1] = V[\hat{b}_2]$; and,
(v) the size of the sample of observations is large (presumably, results will be worse otherwise).

Values of $E[\hat{b}_1|\text{variable 1 selected}]$ are calculated for varying values of $E\hat{b}_2$, $V[\hat{b}_i]$, and $C[\hat{b}_1,\hat{b}_2]$, and range from 1.02 to 1.53, compared with $E\hat{b}_1 = 1$. The value of $E[\hat{b}_1|$ variable 1 selected] increases with increase in each of the variables $E\hat{b}_2$, $V[b_i]$, and $C[\hat{b}_1,\hat{b}_2]$.

Thus, it is apparent that selection bias may be substantial. Miller suggests a number of possible remedies for competition bias, though not for stopping rule bias. They are:

(i) using half of the data to select predictors and the other half to fit the model;
(ii) using jackknife or bootstrap methods (see Section 5 of this paper);
(iii) using shrunken estimators of the ridge or Stein type;
(iv) using simulation to estimate bias; or,
(v) using maximum likelihood estimation of regression coefficients taking the subset selection procedure P into account in the likelihood.

4. SUBSET SELECTION

4.1. General

Consider the method by which the subset A of predictors (in the terminology of Section 3) might be chosen. What criterion might be adopted?

Starting at the most naive end of reasoning, consider (3.2.8) and the identification of the different types of prediction error in the passage immediately following. If all available predictors are included in A, then B is empty and selection error falls to zero.

However, such a suggestion is likely to introduce the very practical problem (and, we shall see shortly, the theoretically objectionable fact) that the number of predictors runs literally into hundreds. Moreover, the evidence may be that the majority are statistically insignificant.

Alternatively, then, one might consider including in A only those predictors which can be demonstrated as statistically significant, and specifically as significant not only in isolation but also in conjunction with the other members of A. This, typically, is the type of procedure followed by stepwise regressions (Efroymson, [6]).

Certainly, this alternative procedure might reduce selection error to quite tolerable levels. It is necessary to recognize, however, that reduction of this type of error does not of itself result in efficient prediction. High efficiency in fact requires a low MSEP.

Recall from (3.3.3) and the text just following it that the MSEP consists of three components, only one of which is selection error. Of the remaining two, statistical error is independent of the model selected. Hence, an examination of the prediction efficiency of various models amounts to an examination of the respective effects of increasing the number of predictors on:

(i) selection error (as noted above, this decreases); and,
(ii) estimation error.

It turns out that, broadly, estimation error increases as the set of predictors increases. This is intuitive. The more predictors that need to be fitted to a fixed number of data points, the more difficult the fitting becomes. As the number of predictors becomes too large, the phenomenon of over fitting mentioned in Section 2 becomes more in evidence.

In the extreme case in which the numbers of data points and predictors are roughly equal, the whole fitting procedure is concentrated on achieving adherence of the model to past observation. The model is then being fitted to the random noise of past observation as well as the underlying signal, with consequent loss of predictive power. That is, estimation error is increased.

The opposite effects on selection error and estimation error of increasing the number of predictors are illustrated by Exhibit I.

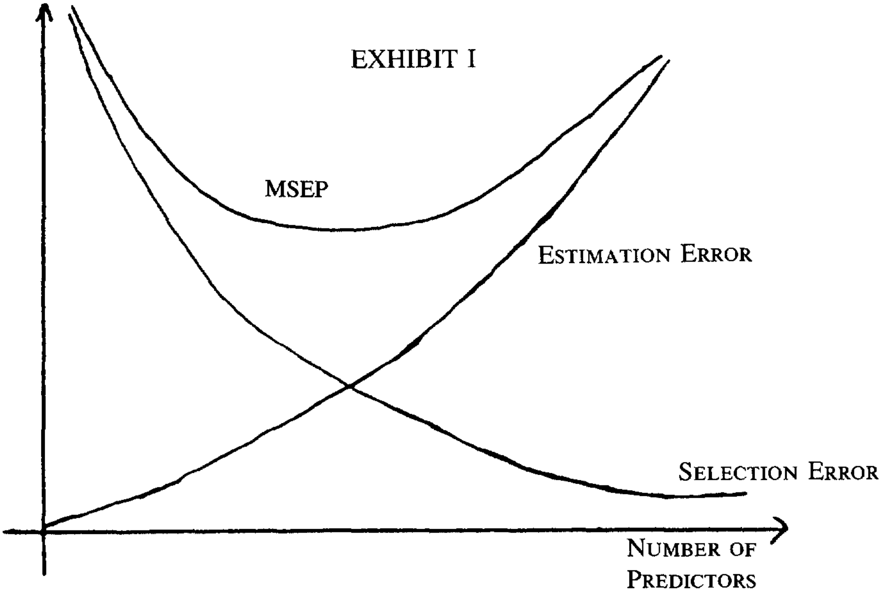


EXHIBIT I

This indicates the existence of an optimal subset of available predictors in the sense of minimizing MSEP. The next couple of subsections deal with simple statistics aimed at facilitating the selection of the subset which is optimal or, more realistically, which is not too far sub-optimal.

4.2 Mallows' C_p statistic

Consider once again the situation introduced in Section 3.2, but assume now the underlying algebraic structure $f(.)$ in (3.2.1) is linear. In this case (3.2.2) becomes:

$$Y = Xb + e, \tag{4.2.1}$$

where, as before, e has zero mean, and is further assumed to have stochastically independent components all with equal variance σ^2.

Recall the decomposition of MSEP:

$$E(Y^* - \hat{Y}^*)^2 = E(Y^* - EY^*)^2 + E(\hat{Y}^* - E\hat{Y}^*)^2 + (\text{prediction bias})^2. \tag{3.3.3}$$

A somewhat simplified version of this is:

$$\begin{aligned}\Delta = E(EY^* - \hat{Y}^*)^2 &= E(\hat{Y}^* - E\hat{Y}^*)^2 + (\text{prediction bias})^2 \\ &= \text{estimation error} + \text{prediction error}.\end{aligned} \tag{4.2.2}$$

The left side of (4.2.2) is a measure of deviation of the expected values of future observations from predictions, whereas MSEP is a measure of deviation of the actual values of future observations from predictions.

The difference between the two measures is the statistical error $\mathrm{E}(e^*)^2$. Since this is independent of the model chosen, subset selection according to minimum MSEP is the same as minimizing Δ. This is the basis of Mallows' C_p statistic introduced by Mallows [14] and discussed by Seber ([19], pp. 364–369).

In the following, let a subscript q indicate that the quantity under consideration relates to a model based on q of the available predictors (one of them representing a constant term, i.e., a constant column of X). Seber shows that:

$$\Delta_q = q\sigma^2 + (PB)_q^2, \tag{4.2.3}$$

with PB denoting prediction bias.

Now the usual definition of *residual sum of squares* (RSS) is:

$$\mathrm{RSS} = (Y - \hat{Y})^2,$$

and as is well-known,

$$E(\mathrm{RSS}_q) = (n - q)\sigma^2 + (PB)_q^2. \tag{4.2.4}$$

By (4.2.3) and (4.2.4),

$$E(\mathrm{RSS}_q) + (2q - n)\sigma^2 = \Delta_q.$$

Therefore, if

$$C_q = \mathrm{RSS}_q/\hat{\sigma}^2 + 2q - n, \tag{4.2.5}$$

with $\hat{\sigma}^2$ a suitable estimator of σ^2, C_q will be an approximately unbiased estimator of Δ_q/σ^2. Then minimization of MSEP, equivalently of Δ_q, will be approximately achieved by selection of the subset of predictors which minimizes C_q defined by (4.2.5).

In the case in which the number of predictors included in the model is denoted by p (recall that this symbol has been reserved for the total number of available predictors), (4.2.5) becomes C_p. This is the name by which it is usually known—Mallows' C_p statistic.

4.3. *Breiman and Freedman S_p statistic*

Breiman and Freedman [4] consider a situation similar to that of Section 4.2. In their case, however, the elements of the design matrix X in (4.2.1) are random variables.

It is assumed, in addition to the assumptions of Section 4.2, that e and the columns of X are jointly normal with zero mean and that e is stochastically independent of the columns of X. As before σ^2 denotes $V(e^*)$, and in addition we adopt the notation:

$$\sigma_q^2 = V[X_B b_B | X_A], \tag{4.3.1}$$

where X_A, X_B, have the same meaning as in Section 3, the set A now containing q predictors.

Just as in Section 4.2, the quality of the regression is assessed by reference to the MSEP, though in the presence of random variation of X this requires further definition. Breiman and Freedman define

$$\text{MSEP} = E[E[(Y^* - \hat{Y}^*)^2|X,Y]], \tag{4.3.2}$$

where the outer expectation operator is unconditional, i.e., averages over the data X, Y. The algebra is developed in terms of the case $m = 1$ (i.e., the vector Y^* has a single component) though this does not result in any loss of generality in the S_p statistic presented below.

The algebraic development is rather similar to that of Section 4.2. The extended MSEP (4.3.2) may be written in the form, parallel to (3.3.3):

$$\begin{aligned}\text{MSEP} = \text{statistical error} &+ E[\text{estimation error}|X,Y] \\ &+ E[(\text{prediction bias})^2|X,Y].\end{aligned} \tag{4.3.3}$$

Now, apart from the averaging over data, the final two terms of (4.3.3) are those appearing as Δ on the right side of (4.2.2). Hence, (4.3.3) becomes:

$$\begin{aligned}\text{MSEP} &= \sigma^2 + E[(PB)^2|X,Y] + E[\text{estimation error}|X,Y] \\ &= \sigma^2 + \sigma_q^2 + E[\hat{\mathrm{b}}_A - E\hat{\mathrm{b}}_A)^T X_A^T X_A(\hat{b}_A - E\hat{b}_A)|X,Y],\end{aligned} \tag{4.3.4}$$

where use has been made of (4.3.1).

With a little further development, Breiman and Freedman show that:

$$\text{MSEP} = (\sigma^2 + \sigma_q^2)\ [1 + q/(n - 1 - q)]. \tag{4.3.5}$$

The first bracketed term on the right is estimated by $(n - q)^{-1}(\text{RSS})$, whence MSEP is estimated by

$$S_q = (n - q)^{-1}(\text{RSS})[1 + q/(n - 1 - q)]. \tag{4.3.6}$$

The paper by Breiman and Freedman goes on to demonstrate certain optimality properties of S_q.

In the case in which the number of predictors included in the model is denoted by p (recall that in the present paper this symbol has been reserved for the total number of available predictors), (4.3.6) becomes S_p. This is the name by which it is usually known—Breiman and Freedman S_p statistic.

In application of S_p, the subset of regression predictors is selected from those available in such a way as to minimize S_p.

The choice between C_p and S_p in regressions arising from claims analysis is not always easy. In the first example of Section 2, the values entering the design matrix, $\bar{t}_{ij}(k)$ and $1/F_{ij}$ will indeed be random variables, as allowed by S_p (but not C_p). On the other hand, however, their mean values will be necessarily non-zero, contrary to the assumption underlying S_p (but not C_p).

This will be particularly true of any constant term in the regression equation, such as a in (2.1). It is perhaps desirable to examine the behavior of both C_p and S_p as the subset of predictors entered into the regression is varied.

Miller [15] (pp. 406–407) suggests in strong terms that the efficacy of stopping rules such as those based on C_p and S_p is very much limited by the existence of competition bias (Section 3.4):

> *"the vast literature on stopping rules . . . is an irrelevant academic exercise until the problems of estimation have been overcome."*

He points out that competition bias can easily be of the order of two standard errors when the same data set is used for subset selection and parameter estimation. He provides a simulated example in which the true MSEP is compared with that estimated, ignoring competition bias, by the formula:

$$\text{MSEP (false)} = [1 + (q + 1)/n]\ \text{RSS}/(n - 1 - q),$$

for a model containing q predictors and a constant term. The results were as shown in Exhibit II.

4.4. *Spjøtvoll's goodness-of-fit*

Spjøtvoll [21] provides a test of the goodness-of-fit of one subset of predictors relative to another. This is dealt with in reasonable detail by Miller [15] (pp. 397–399).

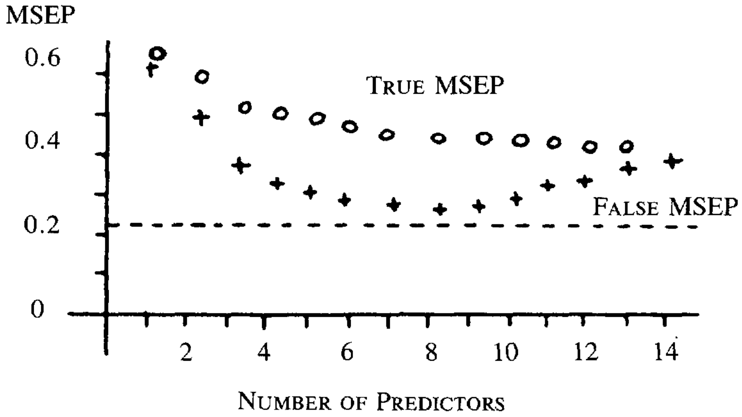
Spjøtvoll's measure of goodness-of-fit is:

$$(Xb - X_A E\hat{b}_A)^T\,(Xb - X_A E\hat{b}_A) = (Xb)^T(Xb) - (Xb)^T\,X_A(X_A^T X_A)^{-1} X_A^T(Xb). \tag{4.4.1}$$

Since the first member of this last expression is independent of the subset of predictors selected, Spjøtvoll chose to use just:

$$(Xb)^T\,X_A(X_A^T X_A)^{-1} X_A^T(Xb). \tag{4.4.2}$$

EXHIBIT II



Miller points out that, if goodness-of-fit is to be assessed for prediction purposes, (4.4.1) might reasonably be modified by the inclusion of a statistical error term. (See Section 3.2 for explanation.) Then (4.4.1) is replaced by:

$$\begin{aligned}(Xb - X_A\hat{b}_A)^T(Xb - X_A\hat{b}_A) &= (Xb)^T(Xb) - (Xb)^T X_A(X_A^TX_A)^{-1}X_A^T(Xb)\\ &\quad + \sigma^2 \text{ trace } [X_A(X_A^TX_A)^{-1}X_A^T],\end{aligned}$$

where $\sigma^2 I = Ve$. This extra term is equal to $q\sigma^2$ (just as in (4.2.3)) when there are q linear predictors including a constant term, so that (4.4.1) is replaced by:

$$\begin{aligned}(Xb - X_A\hat{b}_A)^T(Xb - X_A\hat{b}_A) &= (Xb)^T(Xb) - (Xb)^T X_A(X_A^TX_A)^{-1}X_A^T(Xb)\\ &\quad + q\sigma^2, \qquad (4.4.3)\end{aligned}$$

and (4.4.2) by:

$$(Xb)^TX_A(X_A^TX_A)^{-1}X_A^T(Xb) - q\sigma^2. \qquad (4.4.4)$$

Note that (4.4.3) is identical to Δ_q defined in (4.2.2) in the development of Mallows' C_p with the exception that in the latter case it is based on the future design matrix X^* whereas (4.4.3) is based on the past X.

By (4.4.4), different subsets of predictors, say M and N, are compared by means of the statistic:

$$G_{MN} = (Xb)^T[X_M(X_M^TX_M)^{-1}X_M^T - X_N(X_N^TX_N)^{-1}X_N^T](Xb) - (q_M - q_N)\sigma^2,$$
$$= b^TC_{MN}b - (q_M - q_N)\sigma^2, \tag{4.4.5}$$

where C_{MN} is the appropriate $p \times p$ matrix. We note that the final member of this expression was not used by Spjøtvoll.

Spjøtvoll goes on (summarized by Miller) to develop maximum and minimum values for G_{MN} conditional upon b lying within a $(1 - \alpha)$ confidence set of the form:

$$\Pr[(b - \hat{b})^TX^TX^T(b - \hat{b}) \leqslant k] = 1 - \alpha,$$

where $\hat{b}$ is the regression estimate of b in the full model.

These limits on G_{MN} may be used to test whether M provides a significantly better or worse fit than N to the data.

5. METHODS OF ESTIMATION OF SECOND MOMENTS OF LOSS RESERVES

5.1. *General*

This section will consider methods by which MSEP of loss reserves can be estimated.

First note that this will not consist merely of estimating (3.3.3). Typically, Y^* will be some vector of future claim payments, subdivided for example according to year of occurrence and development year. In such a case, the estimated loss reserve would be:

$$\hat{R} = 1^T\hat{Y}^*, \tag{5.1.1}$$

where 1 is an m-vector with every component equal to unity.

Then (3.3.3) is replaced by:

$$\text{MSEP}(R) = 1^T\text{e}(e^*)^2\ 1 + 1^TE[X^*(\hat{b}_A - E\hat{b}_A)]^2\ 1 + (\text{prediction bias})^2. \tag{5.1.2}$$

This last equation shows that the MSEP of loss reserve R consists of separate terms representing statistical error, estimation error and prediction bias respectively.

There is little that can be said as to the formal inclusion of the last of these components in any estimate of MSEP. To the extent that it is perceptible, it should be removed from the estimated loss reserve (i.e., first moment thereof) rather than allowed for in MSEP estimation. Some components of prediction bias, e.g., specification error (Section 3.2), are by their very nature, likely to defy any reliable formal evaluation.

The usual situation is therefore that the first two members on the right of (5.1.2) can be evaluated in systematic manner, but only informal allowance can be made for the third, bearing in mind Miller's remarks quoted in Section 4.3.

There are several approaches to this evaluation. They are discussed in detail by Ashe [1]. Brief details are given in the next few subsections.

5.2. Parametric estimation

The linear model (4.2.1) will be referred to here as the parametric model—parametric in the sense that the error term e is assumed to have certain (usually parametric) properties.

If e is well-defined, then its parameters (e.g., σ^2) may be estimated from the data, and hence the first two components of MSEP(R) in (5.1.2) estimated. Logically, this is straightforward even if the algebraic manipulation involved may be cumbersome occasionally. The algebraic details are provided by Taylor and Ashe [24].

The calculations involved in this procedure are quite manageable with just about any reputable regression package. Naturally, the results are reliable only to the extent that the parametric assumptions underlying the procedure may be relied upon. Care is therefore necessary in dealing appropriately with the covariance structure of e. See, for example, the weighting procedure used by Taylor and Ashe [24] in their regressions.

5.3. Jackknife

The jackknife algorithm was introduced by Quenouille [17] and is now found in many standard texts, e.g., Mosteller and Tukey [16]. The purpose of the algorithm was to reduce bias in parameter estimates based on limited data.

An outline of the method is as follows. Suppose that some parameter θ is estimated by a statistic S. This statistic may be a complicated function of the data. The precise properties of S are either unknown or difficult to compute. It is known, however, that the bias contained in S is of order n^{-1} for sample size n.

Let S be denoted by $S(n)$ for sample size n. Now, for each $i = 1, 2, \ldots,$ n, define $S_i(n)$ as the value of S based on the $(n - 1)$-sample obtained by deletion of the ith observation. Then define a pseudo-value:

$$P_i(n) = nS(n) - (n - 1)\, S_i(n),\ i = 1, 2, \ldots, n. \tag{5.3.1}$$

By assumption,

$${}_ES(n) = \theta + a/n + o(n^{-1}).$$

Hence

$${}_EP_i(n) = \theta + o(n^{-1}),$$

and so

$$\overline{P}(n) = \sum_{i=1}^{n} P_i(n)/n = \theta + o(n^{-1}) + \text{error term} \tag{5.3.2}$$

contains a bias of order less than n^{-1} as an estimator of θ.

The variance of $\overline{P}(n)$ is estimated by (Mosteller and Tukey, 1977, p. 135):

$$\{\overline{P}^2(n) - [\overline{P}(n)]^2\}/(n - 1), \tag{5.3.3}$$

where

$$\overline{P}^2(n) = \sum_{i=1}^{n} P^2(n)/n. \tag{5.3.4}$$

This algorithm may be applied to the present context by setting $S(n)$ equal to the estimated loss reserve obtained from a regression claims model based on n data points (a single data point being, for example, the observed claim payments in a given development year of a given year of occurrence). This can be generalized by taking $S(n)$ to be the vector of loss reserves for the different years of occurrence; or the vector of claim payments projected for each of the years of run-off; or, indeed, any one of the many cross-sections which might be taken from the regression forecast of future cash flows according to year of occurrence and development year.

In practical application, it might seem reasonable to adapt the jackknife estimates (5.3.1) to (5.3.4) to weighted regression. Possible replacement formulas are:

$$P_i(n) = [WS(n) - (W - w_i)S_i(n)]/w_i \tag{5.3.1a}$$

where w_i is the weight applied to observation i in the weighted regression and

$$W = \sum_{i=1}^{n} w_i;$$

$$\overline{P}(n) = \sum_{i=1}^{n} w_i P_i(n)/W; \qquad (5.3.2a)$$

$$\{\overline{P}^2(n) - [\overline{P}(n)]^2\} \times \sum_{i=1}^{n} w_i^2/W^2 \qquad (5.3.3a)$$

$$\overline{P}^2(n) = \sum_{i=1}^{n} w_i P_i^2(n)/W. \qquad (5.3.4a)$$

Despite the seeming reasonableness of (5.3.1a) to (5.3.4a), Ashe [1] (p. S108) points out that the response of the weighted jackknife to his particular numerical examples is wild. It is possible that the bias assumption underlying the jackknife is incorrect and that the adoption of unequal weights w_i magnifies this in $\overline{P}(n)$. Indeed, Miller [15] (p. 404) provides a semi-rigorous argument that competition bias is of order $n^{-1/2}$, not n^{-1} as required for the jackknife to be valid.

Ashe [1] (p. S110) points out the usefulness of the pseudo-values in their own right as providing an indication of the influence of individual data points. A deviant value of $P_i(n)$ indicates that the whole regression is strongly influenced by data point i. Further discussion of the influence function and the appropriate response to it will appear in Section 6.

There are two shortcomings of the jackknife.

First, the entire procedure is dependent on the assumption that bias in the statistic S is of order n^{-1}. In practical applications, this may not be known with any certainty.

Secondly, variance estimates (5.3.3) and (5.3.3a) are in fact estimates of estimation error only. Presumably, regression estimates $\hat{\sigma}^2(n)$ of statistical error could also be jackknifed. The results would however be dubious since the assumption of a bias of order n^{-1} would be even more uncertain in the case $\hat{\sigma}^2(n)$ than $S(n)$.

5.4. Bootstrap

The bootstrap (Efron, [5]) is a procedure which makes use of data resampling. Application of the technique to regression problems is discussed by Freedman and Peters [7].

Consider the model:

$$Y = Xb + e, \tag{5.4.1}$$

where X is a given design matrix and e is a random vector with mean zero and covariance matrix V.

As in previous sections, let $\hat{b}$ denote the regression estimate of b. Then let:

$$\hat{e} = Y - X\hat{b}, \tag{5.4.2}$$

and

$$\xi = V^{-1/2}\hat{e}, \tag{5.4.3}$$

where the meaning of $V^{-1/2}$ is the conventional one for a positive definite matrix V.

Note that the components of ξ are independent, identically distributed (i.i.d). Let $F(.)$ denote the empirical distribution function obtained by assigning equal masses to them. It is now possible to generate pseudo-data sets:

$$Y^{(i)} = X\hat{b} + e^{(i)}, \ i = 1, 2, \ldots \tag{5.4.4}$$

where

$$e^{(i)} = V^{1/2}\xi^{(i)},$$

and $\{\xi^{(i)}\}$ is a random sample drawn from $F(.)$. Each set of pseudo-data leads to a new estimate $\hat{b}^{(i)}$ of b.

Let X^*, Y^* have the same meaning as in earlier sections. Then each estimate $\hat{b}^{(i)}$ leads to an estimate $\hat{Y}^{*(i)}$ of Y^* where

$$\hat{Y}^{*(i)} = X^*\hat{b}^{(i)}, \ i = 1, 2, \ldots \tag{5.4.5}$$

The collection $\{\hat{Y}^{*(i)}\}$ provides an empirical distribution of the random variable

$$\hat{Y}^* = X^*\hat{b}. \tag{5.4.6}$$

This distribution may be used to study the mean, variance, non-normality, confidence limits, etc. of (5.4.6). Note that:

$$\hat{Y}^{*(i)} - X^*b = X^*[\hat{b}^{(i)} - b], \tag{5.4.7}$$

which contains only estimation error. More pertinent to forecasting is a collection of forecasts of $\hat{Y}^*$ which contains statistical error also.

This is obtained by replacing (5.4.5) and (5.4.6) by:

$$\hat{Y}^{*(i)} = X^{*}\hat{b}^{(i)} + e^{*(i)}, \tag{5.4.5a}$$

$$\hat{Y}^{*} = X^{*}\hat{b} + e^{*}, \tag{5.4.6a}$$

where

$$e^{*} = W^{1/2}\xi^{*}, \tag{5.4.8}$$

$$e^{*(i)} = W^{1/2}\xi^{*(i)}, \tag{5.4.9}$$

for a known matrix W, the components of ξ, ξ^{*} are i.i.d., and $\{\xi^{(i)}, \xi^{(i)*}\}$ is a random sample drawn from $F(.)$, i.e., a particular $\xi^{(i)}$ and $\xi^{(i)*}$ are stochastically independent.

In this case (5.4.7) is replaced by:

$$\hat{Y}^{*(i)} - X^{*}b = X^{*}[\hat{b}^{(i)} - b] + e^{*(i)}, \tag{5.4.7a}$$

which includes both estimation and statistical error.

Freedman and Peters [7] (p. 99) deal with the case in which V is unknown and provide an iterative scheme for its estimation simultaneously with the generation of pseudo-data.

It is to be emphasized that the whole procedure assumes the validity of the basic model (5.4.1). If the model is invalid, estimates of second moments will probably be enlarged but not necessarily in the correct way.

For example, if prediction bias is present in model (5.4.1), it will be absorbed into $\hat{e}$ of (5.4.2) and hence ξ of (5.4.3). The components of ξ will then have non-zero mean and will not in general be identically distributed as assumed in the generation of pseudo-data (5.4.4).

5.5. Comparison of the estimation procedures

The advantages and disadvantages of the three estimation procedures considered in Sections 5.2 to 5.4 are summarized by Ashe [1] (p. S112) as follows:

Parametric estimation
- small number of calculations
- estimation error and statistical error available
- accurate if the parametric assumptions are correct

Jackknife:
- influence of individual data points on the estimate is available
- only estimation error is available
- estimate of loss reserve possible has reduced bias

Bootstrap:

- non-parametric
- estimation error and statistical error available
- distribution of loss reserve given

6. ROBUSTNESS

6.1. *Influence function*

The concept of an influence curve was introduced by Hampel [9]. It is discussed by Mosteller and Tukey [16] (pp. 351–356). A generalization to an influence function, a multi-dimensional version of the influence curve, is discussed by Rey [18] (pp. 15, 16).

The influence function of data points $y_1, \ldots, y_n$ on statistic $S(y_1, \ldots, y_n)$ is defined as the vector,

$$I(y_1, \ldots, y_n) = \frac{\partial S}{\partial \boldsymbol{y}} (y_1, \ldots ,y_n), \tag{6.1.1}$$

with $\boldsymbol{y}$ denoting the vector $(y_1, \ldots, y_n)$. It indicates the influence on S of small variations in the data points.

A single component $\partial S/\partial y_i$ of (6.1.1), plotted as a function of y_i, with $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$ fixed at their observed values, provides the influence curve of y_i.

In the context of loss reserving by regression methods $S(y_1, \ldots, y_n)$ may be taken as the forecast (5.1.1):

$$\hat{R} = 1^T\hat{Y}^* = 1^T X^* \hat{b}_A, \tag{6.1.2}$$

where

$$\hat{b}_A = \hat{b}_A(Y_1, \ldots, Y_n)$$

is the regression estimate of b_A as in (3.2.7) and is a function of the data vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$.

Since $I_i(.)$ measures the effect of small variations of y_i on S, and the jackknife pseudo-estimate $P_i(n)$ measures the effect of removing y_i from the data, the two are related, as foreshadowed in Section 5.3. As suggested there, the pseudo-values perhaps serve as some kind of proxy for the influence function.

6.2. *Robust regression*

Regression need not be carried out by means of least squares, weighted or unweighted. Indeed, the importance of least squares regression derives, through the Gauss-Markov theorem (Graybill, [8]), from the oft-made assumption that random error terms in the data are normally distributed. When this assumption does not hold, least squares regression may not be appropriate.

There is no doubt that most classes of insurance involve long tailed claim size distributions. The basic data of any claims analysis, such as claim payments subdivided by year of occurrence and year of development, are therefore likely to incorporate error terms with long tailed distributions. Under weighted least squares regression, one or two rogue data points might well drag the entire regression away from the estimates which it would otherwise provide.

Robust regression encompasses procedures for fitting linear models whose properties are relatively insensitive to the distribution of these error terms. *Resistant* regression includes procedures leading to estimates which are not greatly distorted by extreme cases.

The latter of these two concepts is evidently related to the influence function. The smaller the influence function of a particular data point, the more resistant the regression to outlying values at that point.

Various methods have been used to reduce the influence function from that associated with least squares regression. For a summary, see Huber [11], [12]. An actuarial reference is Hogg [10]. Most of these methods can be viewed as fairly simple modifications of weighted least squares regression.

Consider the model,

$$Y = Xb + e, \tag{6.2.1}$$

where the notation is as in previous sections and, in particular, e is not necessarily normal although it is assumed to have zero mean. Under weighted least squares regression, b is estimated by that $\hat{b}$ which minimizes the weighted sum of squares (WSS):

$$\text{WSS} = (Y - X\hat{b})^T W (Y - X\hat{b}), \tag{6.2.2}$$

for some $n \times n$ matrix W which is independent of Y. Under resistant regression (6.2.2) is replaced by:

$$\text{WSS} = (Y - X\hat{b})^T W(\hat{Z})(Y - X\hat{b}), \tag{6.2.3}$$

where the weight matrix W depends on an estimate $\ddot{Z}$ of the vector of standardized residuals,

$$\hat{Z} = \text{diag}\,(\hat{\sigma}_1^{-1}, \ldots, \hat{\sigma}_n^{-1})(Y - X\hat{b}), \tag{6.2.4}$$

with $\hat{\sigma}_i^2$ an estimate of $V[Y_i]$.

Most commonly, the form of $W(\hat{Z})$ is:

$$\begin{aligned} w_{ij} &= h_i(\hat{Z}_i),\ j = i; \\ &= 0, \qquad j \neq i; \end{aligned} \tag{6.2.5}$$

for some function h_i which decreases as $\hat{Z}_i$ departs from zero. Thus, outlying observations, generating large values of $\hat{Z}_i$, are assigned little weight in WSS.

Typical choices of the attenuating function $h_i(.)$ are:

$$\begin{aligned} h_i(z) &= w_i, \qquad |z|\ ;\leqslant 2; \\ &= 4w_i/|z|^2,\ |z| \geqslant 2, \end{aligned} \tag{6.2.6}$$

where diag $(w_1, \ldots, w_n)$ is the weight matrix which would have been used for weighted least squares regression; or alternatively,

$$\begin{aligned} h_i(z) &= w_i z^{-1} \sin\,(2z/3),\ |z| \leqslant 3\pi/2; \\ &= 0, \qquad |z| \geqslant 3\pi/2; \end{aligned} \tag{6.2.7}$$

or again,

$$\begin{aligned} h_i(z) &= w_i\,[1 - (z/5)^2]^2,\ |z| \leqslant 5; \\ &= 0, \qquad |z| \geqslant 5. \end{aligned} \tag{6.2.8}$$

It is apparent that any system (6.2.3) in which the weight matrix $W(\hat{Z})$ depends on $\hat{b}$ renders WSS non-quadratic in $\hat{b}$. Then the solution b is nonlinear in the data Y. It will usually be necessary, therefore, for (6.2.3) to be minimized iteratively. At each iteration, the $\hat{\sigma}_i^2$ need to be recalculated on the basis of the residuals at the preceding iteration. Then $W(\hat{Z})$ can also be calculated on the basis of the same residuals, and (6.2.3) minimized with the new $W(\hat{Z})$ treated as independent of $\hat{b}$.

REFERENCES

[1] F. R. Ashe, "An Essay at Measuring the Variance of Estimates of Outstanding Claim Payments," *Astin Bulletin,* Vol. 16, 1986, p. S99.

[2] D. J. Bartholomew, "Errors of Prediction in Markov Chains," *Journal of the Royal Statistical Society,* Series B, Vol. 37, 1975, p. 444.

[3] J. R. Berquist and R. E. Sherman, "Loss Reserve Adequacy Testing: A Comprehensive Systematic Approach," *PCAS* LXIV, 1977, p. 123.

[4] L. Breiman and D. Freedman, "How Many Variables Should be Entered in a Regression Equation?" *Journal of the American Statistical Association,* Vol. 78, 1983, p. 131.

[5] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics,* Vol 7, 1979, p. 1.

[6] M. A. Efroymson, "Multiple Regression Analysis," *Mathematical Methods for Digital Computers,* Vol. 1 (A. Ralston and H. S. Wilf, eds), 1960, p. 191. John Wiley & Sons, New York.

[7] D. A. Freedman and S. C. Peters, "Bootstrapping a Regression Equation: Some Empirical Results," *Journal of the American Statistical Association,* Vol. 79, 1984, p. 97.

[8] F. A. Graybill, *An Introduction to Linear Statistical Models,* Vol. 1, 1961, McGraw-Hill.

[9] F. E. Hampel, "The Influence Curve and its Role in Robust Regression," *Journal of the American Statistical Association,* Vol. 69, 1974, p. 383.

[10] R. V. Hogg, "Statistical Robustness: One View of its Use in Applications Today," *Actuarial Research Clearing House,* 1979.3 Issue, p. 9.

[11] P. J. Huber, "Robust Statistics: A Review," *Annals of Mathematical Statistics,* Vol. 43, 1972, p. 1041.

[12] P. J. Huber, "Robust Statistical Procedures," *Society of Industrial and Applied Mathematics,* 1977.

[13] E. Kremer, "IBNR-Claims and the Two-Way Model of ANOVA," *Scandinavian Actuarial Journal,* 1982, p. 47.

[14] C. L. Mallows, "Some Comments on C_p," *Technometrics,* Vol. 15, 1973, p. 661.

[15] A. J. Miller, "Selection of Subsets of Regression Variables," *Journal of the Royal Statistical Society, Series A,* Vol. 147, 1984, p. 389.

[16] F. Mosteller and J. W. Tukey, *Data Analysis and Regression,* Addison-Wesley Publishing Company, Reading, Massachusetts, 1977.

[17] M. H. Quenouille, "Notes on Bias in Estimation," *Biometrika,* Vol. 43, 1956, p. 353.

[18] W. J. J. Rey, "Robust Statistical Methods," *Lecture Notes in Mathematics,* Vol. 690, 1979, Springer-Verlag, Berlin.

[19] G. A. F. Seber, *Linear Regression Analysis,* John Wiley and Sons, Inc., New York, 1977.

[20] D. Skurnick, "A Survey of Loss Reserving Methods," *PCAS* LX, 1973, p. 16.

[21] E. Spjøtvoll, "Multiple Comparison of Regression Functions," *Annals of Mathematical Statistics,* Vol. 43, 1972, p. 1067.

[22] G. C. Taylor, "An Invariance Principle for the Analysis of Non-Life Insurance Claims," *Journal of the Institute of Actuaries,* Vol. 110, 1983, p. 205.

[23] G. C. Taylor, *Claims Reserving in Non-Life Insurance,* North-Holland, Amsterdam, 1986.

[24] G. C. Taylor and F. R. Ashe, "Second Moments of Estimates of Outstanding Claims," *Journal of Econometrics,* Vol. 23, 1983, p. 37.