Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

SELECTING THE BEST LINEAR SIMULATION METAMODEL

Russell Cheng

School of Mathematics
University of Southampton
Southampton, SO17 1BJ, UK

ABSTRACT

We consider the output of a simulation model of a system about which little is initially known. This output is often dependent on a large number of factors. It is helpful, in examining the behaviour of the system, to find a statistical metamodel containing only those factors most important in influencing this output. The problem is therefore one of selecting a parsimonious metamodel that includes only a subset of the factors, but which nevertheless adequately describes the behaviour of the output. The total number of possible submodels from which we are choosing grows exponentially with the number of factors, so a full examination of all possible submodels rapidly becomes intractable. We show how resampling can provide a simple solution to the problem, by allowing potentially good submodels to be rapidly identified. This resampling approach also allows a systematic statistical comparison of good submodels to be made.

1 INTRODUCTION

Discrete-event simulation is often used in the exploratory study of a complex system. We suppose that the behaviour of the system is summarised by some performance measure which we shall take to be the output of interest.

This output is often dependent on a large number factors. We consider the situation where little is known about the system but it is expected that, though the number of factors is large, possibly only a few will be really important. We therefore wish to construct a parsimonious statistical metamodel that includes these important factors but we would like some assurance that such a submodel adequately represents system behaviour.

We shall only consider the simplest situation, where a linear model is to be fitted to a sample of output obtained from a set of simulation runs made at different factor settings. Ideally these runs will be based on a designed experiment, but our discussion does not require this to be so.

We suppose that there are P factors. There are thus ${}^{p}C_{k}$ distinct submodels in which k out of the p factors are present in the linear model. Summing over all possible k, from 1 to P there are thus a total of 2^{p}-1 possible submodels.

Statistically this is the classic problem of model selection. Though this problem is well known, the usually accepted methods of handling it are not always satisfactory. Wu and Hamada (2000) have discussed this problem at length. They considered the very well-known backward, forward and stepwise factor selection methods and also more sophisticated Bayesian strategies, employing Gibbs sampling. The main problems with these methods are as follows.

The backward, forward and stepwise selection methods are all sequential, in which factors are considered one at a time for possible inclusion, or elimination. It is therefore possible, with non-orthogonally designed experiments, simply because of the order in which factors are considered, to end up with a selected model that does not include all those factors that are important.

Use of a Bayesian approach avoids this difficulty, but a prior distribution for factor coefficient values has to be chosen and there is also the technical implementation issue of deciding when sufficient sampling has been carried out to ensure that adequate convergence to the posterior distribution has taken place.

In this paper we shall consider resampling methods. These work by generating, through bootstrap resampling, a large number of data sets each with the *same* statistical distributional properties as the original data set, at least asymptotically. We can therefore deploy whatever method we wish for selecting the model using the original data sample, but then gauge the adequacy of the selected model by studying how consistently it is selected in the bootstrap samples.

Whatever the method used for the selection process, there is also the additional issue of deciding on the selection criterion to be used in choosing between different models. Also one should have some way of checking whether the selected model is a sufficiently good fit.

There is the possibility that more than one model provides an adequate representation of the relationship be-

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

tween the output and factors. We need therefore to have some means for gauging the adequacy of fitted models. In this last regard, of the existing methods that we have already mentioned, the Bayesian approach seems most satisfactory in that a posterior distribution is obtained for the possible models, so that it will be clear whether there is one single best model choice or whether several competing models are equally or nearly as good. The Bayesian approach is not entirely satisfactory in that it does not provide immediate information on whether the models with the highest posterior probabilities are adequate or not. This of course depends on what purpose the metamodel will be used for, something which we may not be entirely sure of in exploratory studies. It would however be useful to have some criterion for assessing the goodness of fit of the model, at least in some general sense.

In the next section we describe the linear statistical model that we will use and discuss selection criteria for choosing between models. In Section 3 we set out two methods of bootstrap resampling for model selection and analysis. Numerical examples are given in Section 4 and a summary is provided in Section 5.

2 THE LINEAR MODEL

We consider the (full) linear model

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{12} & X_{13} & . & X_{1P} \\ 1 & X_{22} & X_{23} & . & X_{2P} \\ . & . & . & . & . \\ 1 & X_{n2} & X_{n3} & . & X_{nP} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ . \\ b_P \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ \varepsilon_n \end{bmatrix} \qquad (1)$$

where Y_i, $i = 1, 2, ..., n$ are the observed output values obtained from n simulation runs; X_{ij} are the factor values in each of the n runs; $b_j, j = 1, 2, ..., P$ are the unknown coefficients corresponding to each of the P factors; and $\varepsilon_i,\ i = 1, 2, ..., n$ are random errors. We have taken $X_{i1} = 1,\ i = 1, 2, ..., n$ so that b_1 corresponds to a general mean. We thus treat the mean as a coefficient, so that, as far as the model selection and fitting process is concerned, we do not treat it differently from the other coefficients. In what follows, when we refer to a 'factor' it is to be understood that this includes the general mean.

We shall assume that the $\varepsilon_i, i = 1, 2, ..., n$ are identically distributed with mean zero and variance

$$Var(\varepsilon) = \sigma^2 . \qquad (2)$$

Such random errors are often assumed to be normally distributed, but we do not assume that this is necessarily so in our formulation.

We shall, whenever convenient, write (1) in the alternative matrix form

$$\mathbf{Y} = \mathbf{Xb} + \boldsymbol{\varepsilon} . \qquad (3)$$

Equation (1) is the full model in which all factors are included. We shall define a *submodel* as

$$k = \{j_1, j_2, ..., j_p\} \qquad (4)$$

with

$$j_1 < j_2 < ... < j_p,\ p \le P ,$$

if (and only if)

$$b_{j_1} \ne 0, b_{j_2} \ne 0, ..., b_{j_p} \ne 0, \text{and all other } b_j = 0 .$$

We shall write the observations corresponding to this submodel as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1j_1} & X_{1j_2} & . & X_{1j_p} \\ X_{2j_1} & X_{2j_2} & . & X_{2j_p} \\ . & . & . & . \\ X_{nj_1} & X_{nj_2} & . & X_{nj_p} \end{bmatrix} \begin{bmatrix} b_{j_1} \\ b_{j_2} \\ . \\ b_{j_p} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ \varepsilon_n \end{bmatrix} \qquad (5)$$

or in the matrix form

$$\mathbf{Y} = \mathbf{X}(k)\mathbf{b}(k) + \boldsymbol{\varepsilon} . \qquad (6)$$

Where necessary we shall also write

$$p(k) = p \qquad (7)$$

for the number of unknown coefficients in the model k. Also we will denote the full model by K, so that $p(K) = P$.

When we fit the model k we shall use the least squares estimates (see Searle, 1971, for example)

$$\hat{\mathbf{b}}(k) = [\mathbf{X}^T(k)\mathbf{X}(k)]^{-1}\mathbf{X}(k)^T\mathbf{Y} \qquad (8)$$

for the unknown coefficient values, and

$$\begin{aligned} \hat{\sigma}^2(k) &= [n - p(k)]^{-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= [n - p(k)]^{-1}[\mathbf{Y} - \mathbf{X}(k)\hat{\mathbf{b}}(k)]^T[\mathbf{Y} - \mathbf{X}(k)\hat{\mathbf{b}}(k)] \end{aligned} \qquad (9)$$

for the unbiased estimate of the variance of the ε_i .

For model selection, we need a criterion for choosing between submodels. We consider two well known criteria.

The first is the C_p statistic proposed by Mallows (1973). This is defined as

$$C_p(k) = (n - p(k))\hat{\sigma}^2(k) / \hat{\sigma}^2(K) + 2p(k) - n \, . \quad (10)$$

An alternative statistic is the Akaike Information Criterion (Akaike, 1970), which for the linear model reduces to

$$AIC(k) = -2n \log[\hat{\sigma}^2(k)] + 2p(k) \, . \quad (11)$$

Asymptotically C_p and AIC have essentially the same distribution (see Nishii, 1984). However C_p is perhaps more satisfactory for our purpose. It will be seen that if the model k is satisfactory then the expected value of C_p is close to p. Thus once all important factors are included C_p will increase linearly with p. However if not all important factors are included the expected value of C_p will be larger than p. Thus, our selection method will be simply to select from amongst all possible submodels, that for which C_p is minimum. Moreover we would expect the choice to be satisfactory if the minimized value of C_p is p or smaller.

In summary the basic model selection method is therefore simply to:

(i) Consider each of the $2^P - 1$ possible submodels of (1) and for each submodel k calculate $C_p(k)$ from (10) or $AIC(k)$ from (11).

(ii) Select as the best model that k for which $C_p(k)$, or $AIC(k)$, is minimum.

3 BOOTSTRAP ANALYSIS

In this section we shall for simplicity assume that the selection criterion is C_p.

The methodology of the previous section is straightforward to apply in principle. However there are two aspects of concern.

Dimensionality Problem. Because the total number of submodels, $2^P - 1$, grows exponentially with P, inspection of all submodels is possible only when P is small. Thus even with just 20 factors there are already 1,048,575 submodels. We need therefore to be able to identify promising submodels in a way that is much more selective than the exhaustive search of all submodels.

We give below two methods utilising resampling that allows selective examination of submodels.

Quality of the Selected SubModel. Once a best submodel (as measured by smallest $C_p(k)$ value) has been determined, there is the question of how good this choice is. This is of especial concern if there are several models with values of $C_p(k)$ close to that of the best. This question would be answered if we had many samples and not just one, as we could determine the best submodel for each sample and see if the same submodel is best for all the samples. The resampling methods to be described do precisely this. The method and results are thus easily explained, even when they are not especially versed in mathematical statistics.

The Bayesian approach works similarly, gauging the relative merits of competing submodels in a precise way by assigning a posterior probability to each model. However the Bayesian methodology is arguably more technical and requires some understanding of prior and posterior probabilities by the non-expert.

We now describe our two proposed resampling methods highlighting how they handle the above two problems.

3.1 Bootstrap Samples

Both methods require the generation of bootstrap samples with precisely the same form as (1). The standard way of doing this is described, for example, by Davison and Hinkley (1997). We take the modified residuals

$$r_i = (Y_i - \hat{Y}_i)/(1 - h_{ii})^{1/2}, \; i = 1, 2, ..., n \quad (12)$$

obtained from the fitting the full model K to the original data, where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}$ and h_{ii} is the ith main diagonal entry in the 'hat' matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \, .$$

We then centre these so that their average is zero:

$$e_i = r_i - \bar{r}, \; i = 1, 2, ..., n \, . \quad (13)$$

A bootstrap sample is then obtained by forming

$$Y_i^* = \hat{Y}_i + e_i^*, \; i = 1, 2, ..., n \quad (14)$$

where the e_i^*, $i = 1, 2, ..., n$ are a random sample obtained by sampling with replacement from the e_i, $i = 1, 2, ..., n$.

A second way of resampling, *parametric bootstrapping*, is possible, if it can be assumed that the random errors ε_i, $i = 1, 2, ..., n$ in (1) are normally distributed and independent. The bootstrap sample still takes the same form as (14) only now the e_i^*, $i = 1, 2, ..., n$ are a random sample from the fitted normal distribution, i.e.:

$$e_i^* \sim N(0, \hat{\sigma}^2), \; i = 1, 2, ..., n \quad (15)$$

We are now in a position to describe the two bootstrap analysis methods.

3.2 First Bootstrap Analysis Method

This first method requires an upper limit p ($p \le P$), to be placed on the number of factors that a submodel can have (remembering that we interpret the mean as being a factor). The analysis then works as follows:

Step(1) Fit all submodels k with $p(k) \le p$ to the original data. Calculate $C_p(k)$ for each k. The model with the smallest $C_p(k)$ is the estimated best model (assuming the best model satisfies $p(k) \le p$). Also select a threshold factor α. We then retain all submodels k for which $C_p(k) \le \alpha p(k)$. In practice it turns out, assuming the linear model is appropriate in the first case, many submodels will give a fit with $C_p(k) \le p(k)$. Thus the value of α does not seem very critical. If keeping the number of submodels considered is a very serious issue then $\alpha = 1$ is probably acceptable. To be on the safe side we have used $\alpha = 3$ in the numerical examples of Section 4. This set of submodels, denoted by S, constitutes the set of *promising models*.

Step(2) Generate B bootstrap samples each of the form (14). For each bootstrap sample, j, fit each of the promising submodels in the set S, and select the submodel $k(j)$ for which $C_p(k)$ is smallest.

Step(3) Display the submodels of S, ranked in order of the proportion of times that they are selected as being the best model in the B bootstrap samples, displaying these proportions as well.

It will be seen that, provided the upper limit p in Step(1) can be set sufficiently small, this first methoddeals with the problem of dimensionality by limiting the total number of submodels that need to be fitted to the original data. The further imposition that only fitted models satisfying $C_p(k) \le \alpha p(k)$ be retained will additionally reduce the number of submodels that have to be considered in Step(2). Thus, even if the number of submodels that have to be fitted to the original data set in step(1) cannot be made small (because it may not be possible to set p small), the number of submodels in the bootstrapping of Step(2) can still be controlled by taking a small value for the scaling factor α.

Step(2) handles the problem of adequacy of the best fitted submodel found in Step(1) by seeing how often it is selected as the best submodel in the bootstrap samples.

A good additional check can be carried out by ranking the submodels of S according to their $C_p(k)$ value as calculated from the original data, and then comparing this ranking with the percentage of times they are found to be the best fit when fitted to the bootstrap samples.

The method requires an upper limit p ($p \le P$) to be placed on the number of factors that a submodel can have. Ideally this should be based on prior knowledge that the practitioner may have of the system under study. It may be that some of factors under consideration are not what are termed *main effects* but are *interactions*. In this case one can use the hereditary and hierarchical principles discussed by Wu and Hamada (2000), where an interaction between two main effects can be omitted if both main effects are thought unlikely to be important.

The number of models that have to be fitted to the original data set in this first bootstrap analysis method can still be intractably large, especially for carrying out Step(1), if p, the largest number of factors that there can be in a submodel, cannot be assumed manageably small. Our second method is designed to avoid this problem.

3.3 Second Bootstrap Analysis Method

The first method considers all submodels containing no more than p factors. The second method does not restrict the number of factors that a submodel can have, but keeps the number of submodels considered manageable by a selective process using bootstrapping that focuses on models that show some evidence of being possibly a good fit.

Step(1). Generate B bootstrap samples each of the form (14). For the original sample and for each bootstrap sample, carry out the following selection process:

Step(1.1) Fit the full model, K, to the sample and determine the *p-value* of each of the fitted coefficients, $\hat{b}_j, j = 1, 2, ..., P$, by calculating the so-called *t-value*

$$t_j = \hat{b}_j / \sqrt{\hat{\sigma}^2 d_j} \tag{16}$$

where d_j is the jth entry in the main diagonal of the dispersion matrix, i.e.

$$d_j = (\mathbf{X}^T\mathbf{X})^{-1}_{jj}. \tag{17}$$

If the true value of b_j is $b_j = 0$ then t_j has Student's t-distribution with $n - P$ degrees of freedom. If we therefore denote the complementary distribution function for the absolute value $|t_j|$ by $\overline{T}_{n-P}(\cdot)$, the *p-value* is then:

$$q_j = \overline{T}_{n-P}(|t_j|). \tag{18}$$

Step(1.2) Order the coefficients by their q_j values:

$$q_{j_1} \le q_{j_2} \le \ldots \le q_{j_P}\,, \tag{19}$$

so that $\hat{b}_{j_1}$ is the most significant, and select a significance level q, say.

Step(1.3) We then set a critical p-value, which we denote by q (the same q is used for all the samples), and select the submodels

$$\begin{aligned} k_1 &= \{j_1\} \\ k_2 &= \{j_1, j_2\} \\ &\ldots \\ k_m &= \{j_1, j_2, \ldots, j_m\} \end{aligned} \tag{20}$$

where

$$q_{j_m} \le q < q_{j_{m+1}}\,. \tag{21}$$

Thus the submodel k_i is the one where the i most significant factors have been retained, with a cutoff that only factors with significance level higher than q (i.e. with p-value less than q) are allowed in a submodel.

There is some flexibility in the choice of the value of q that might be used. If q is set small then this means that only factors that are really significant are likely ever to be considered. However it is perfectly reasonable to set q = 1. This simply means that a submodel with p factors will be considered for all values for all values of p, i.e. for p = 1, 2, …, P.

Step(1.4) Collect together all the distinct submodels obtained from the previous step for the original and for all the bootstrap samples. This set of distinct submodels comprises our set of *promising submodels*, which we denote by S. Note that the least stringent value we can take for q (in the sense of restricting the number of submodels considered) is q = 1. Even in this case at most P submodels are selected for consideration from each sample. There are thus at most $(B+1)P$ distinct submodels in S. In fact there are likely to be significantly fewer submodels because if the original and bootstrap samples are consistent then the same submodels tend to provide the best fit in all the samples, so that the same models will tend to be repeatedly identified for inclusion in S.

Step(2) For each bootstrap sample, fit each of the submodels in the set S and identify the best model k as the one with the smallest $C_p(k)$ value.

Step(3) Display the submodels of S, ranked in order of the proportion of times that they are selected in Step(2) as being the best model in the B bootstrap samples, displaying these proportions as well.

A final comparison can be carried out by ranking the submodels of S according to the C_p values obtained from fitting these submodels to the original data, and seeing how this ranking compares with that of Step(3).

This completes the second bootstrap analysis method.

4 NUMERICAL EXAMPLES

We give two examples. The first is a very small data set, but that allows the resampling methods to be demonstrated without unwieldy tabulations. The second is more the kind of problem for which the methods are actually intended.

Though neither of the data sets considered here stem directly from simulation experiments we have used them because they are readily accessible. The output of many large scale simulations have a similar structure. For example, Kleijnen et al. (2006) discuss an interesting supply chain simulation involving some 92 factors. Kleijnen et al. actually discuss the selection of important factors using sequential bifurcation. However, though the problem is large, it is actually possible to carry out a full experimental study of the supply chain. Though not reported here because of space limitations, the methods of this paper have been tested on data obtained from this supply chain simulation example resulting in a very similar set of factors being identified as being important to that found by Kleijnen etal.

4.1 Cement Hardening Example

The first example involves a well-known, but very awkward, data set originally reported by Hald (1952) and also discussed by Krzanowski (1998). The data is given in Table 1 and shows the chemical composition (X_1, X_2, X_3, X_4) of 13 cement samples and the heat evolved, $Y_,$ from each when hardening. The data was assumed to have the form (1), with a mean added, so that, with this included, there are five factors.

Krzanowksi gives the all-subsets analysis assuming however that the mean is always fitted. Using the C_p criterion the best fit models are

$$\begin{aligned} &(X_0, X_1, X_2) \quad\;\; \text{with } C_p = 2.68, \\ &(X_0, X_1, X_2, X_4) \text{ with } C_p = 3.02 \\ &(X_0, X_1, X_2, X_3) \text{ with } C_p = 3.04 \\ &(X_0, X_1, X_3, X_4) \text{ with } C_p = 3.50. \end{aligned}$$

where we have used the notation X_0 to indicate that the mean has been included.

As these results indicate, the choice of best model is not clear cut. Two further models: (X_0, X_1, X_2, X_3, X_4) with C_p = 5.00 and (X_0, X_1, X_2, X_4) with C_p = 5.50 are not unsatisfactory. Part of the problem is that the mean coefficient is nowhere near statistically significant in the full model, and it is not clear if always including it is a good thing to do.

Table 1: Cement Hardening Data

Sample	Heat Y	X1	X2	X3	X4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

We carried out an analysis using both bootstrap methods previously discussed, using the C_p statistic for selection criterion. In the first method, the α factor was set, (somewhat arbitrarily) at $\alpha = 3$. In the second method, the critical q was set at q = 1. This choice of q simply means that for each bootstrap sample exactly P submodels, i.e. m = P in (20), were selected. Table 2 shows the percentage of time each of the submodels in S, the set of promising submodels, was the best fit in a bootstrap sample.

Table 2: Bootstrap Analysis of Cement Data

First Method	%	Second Method	%
(X_0,X_1,X_2)	26	(X_0,X_1,X_2)	26
(X_1,X_2,X_3,X_4)	19	(X_1,X_2,X_3,X_4)	18
(X_0,X_1,X_2,X_3)	13	(X_0,X_1,X_2,X_3)	13
(X_0,X_1,X_4)	12	(X_0,X_1,X_4)	12
(X_0,X_1,X_2,X_4)	12	(X_0,X_1,X_2,X_4)	12
(X_0,X_1,X_3,X_4)	12	(X_0,X_1,X_3,X_4)	11
(X_0,X_2,X_3,X_4)	6	(X_0,X_2,X_3,X_4)	6
		(X_0,X_1,X_2,X_3,X_4)	2

It will be seen that the two methods give almost identical results. Moreover these results corroborate most of the findings reported by Krzanowski. There are two differences of note. Firstly there is some evidence that (X_0, X_1, X_4) is a reasonable submodel. Secondly, there is some evidence that the mean can be dropped, with the submodel (X_1, X_2, X_3, X_4) being picked as the best nearly 20% of the time amongst the bootstrap samples.

4.2 Bank Data Example

The second example is taken from Makridakis et al. (1998, Table 6-8). The data is monthly, The variable of interest, Y, is the first difference, D(EOM), between the successive end of month (EOM) balances of a mutual savings bank. There are three primary X variables: X_1 is a composite triple bond rate (AAA), X_2 is a composite (3-4) year US Government bond rate, X_3 is D(3-4), the monthly change in X_2. There were in addition 11 monthly seasonal factors (D1-D11), and three further variables, time t and its square and cube t^2, t^3, making 17 initial factor variables. We do not reproduce the data here as the three key variables, (EOM), (AAA) and (3-4), for 60 months, are downloadable from the website <http://www-personal.buseco.monash.edu.au/~hyndman/TSDL>.

In our analysis we followed Makridakis et al. (1998, Table 6-8) and express Y in thousands of dollars and analysed only the first 53 months of data. We have also added a general mean X_0 as an additional variable so that we work with 18 factors. There are thus $2^{18} - 1 = 262{,}143$ distinct submodels to select from; manageable, but still a somewhat large number of models to comfortably work through.

Using a best subset analysis with an adjusted coefficient of determination, $\bar{R}^2$, for selection criterion Makridakis et al. found the best model overall was

$$X_0\, X_1\, X_2\, X_3\, D_2\, D_3\, D_4\, D_5\, D_6\, D_7\, D_8\, D_9\, D_{10}\, D_{11}\, t^3 \tag{22}$$

and, using a stepwise regression, that the best model was

$$X_0\, X_1\, X_2\, X_3\, D_2 \quad D_4 \quad D_6\, D_7\, D_8\, D_9\, D_{10}\, D_{11}\, t^3. \tag{23}$$

However as the list in their Table 6-10 shows, there are many competing models with similar values for $\bar{R}^2$.

Such uncertainty in the best final model seems typical in such multivariate data, and it is difficult to come to a firm conclusion without further statistical analysis.

We have carried out such an analysis using the second bootstrap analysis method, with a p-value of q = 0.2. This led to the identification of a set of just 362 promising models at the end of Step(1).

Before we carried out the bootstrap analysis of Step(2) we fitted the 362 promising models to the original data sample. The best three models were (with the best first):

$$\begin{array}{c} X_1\, X_2\, X_3\, D_2 \quad D_4 \quad D_6\, D_7\, D_8\, D_9\, D_{10}\, D_{11}\, t\, t^2 \\ X_1\, X_2\, X_3\, D_2 \quad D_4 \quad D_6\, D_7\, D_8\, D_9\, D_{10}\, D_{11}\, t^2\, t^3 \\ X_0\, X_1\, X_2\, X_3\, D_2 \quad D_4 \quad D_6\, D_7\, D_8\, D_9\, D_{10}\, D_{11}\, t^3 \end{array} \tag{24}$$

so that the stepwise regression model (23) comes third using the C_p criterion.

In Step(2) of the bootstrap analysis all 362 promising submodels were fitted to each of the bootstrap samples. Only 50 of these submodels were ever selected as being the best fit to a bootstrap sample. The ten submodels selected most often as being the best fit to a bootstrap ssmple are listed in Table 3.

Table 3: Top Ten Selected Submodels using the Second Method for the Bank Data Sample. All Submodels include the factors X_1, X_2, X_3, D_2, D_4, D_6, D_7, D_8, and D_{10} which are therefore not listed

X_0	D_1	D_3	D_5	D_9	D_{11}	t	t^2	t^3	%
				x	x	x	x		11
				x	x		x	x	6
x					x	x			5
x	x	x	x	x				x	4
x		x	x	x	x	x			4
	x			x	x	x	x		4
			x	x	x	x	x		4
	x	x	x	x					4
				x	x				4
x		x	x	x	x				4

The results suggest that, whether the mean is fitted or not is not very important. In fact, when the full model is fitted to the original sample, the p-value for the mean is 0.66, showing that the general mean is not at all close to being statistically significant from zero for the original data.

For all the 50 submodels that were ever selected, the three main factors X_1 (AAA), X_2 (3-4), X_3 D(3-4) were always included, as were the seasonal variables D_2, D_4, D_6, D_7, D_8 and D_{10}. Of the others D_9 and D_{11} seemed marginally less important. The remaining three D_1, D_3, D_5 did not seem very important. It seemed worth including a time variable, but it is unclear if any one of them is to be preferred. That t^3 appears in both (22) and (23) seems fortuitous when one looks at the way that the rather random way that different time variables appear in the different models listed in Table 6-10 of Makridakis et al. (1998).

Though the details are a little different, in broad terms the bootstrap results are very similar to the results reported by Makridakis et al.

Finally it is interesting to see how the submodels (24), selected as being the best fit to the original data, came out in the bootstrap analysis. The top two were also selected most often by the bootstrap analysis as being the best fit to a bootstrap sample, but the third did not quite appear in the top ten in the bootstrap analysis and was only 12th best.

5 CONCLUSIONS

We have presented two methods using bootstrapping to analyse the selection and fitting of linear models in multiple regression. The second method in particular seems attractive in enabling promising models to be tractably selected out of the full set of all possible submodels when the number of factors is large.

The bootstrapping allows an assessment to be made of how stable the submodel estimated as being the best fit to the original actually is, in the sense of seeing how often that model is selected as being the best when a large number of promising models are fitted to a number of bootstrap samples with the same form as the original data. Such information is not available using a standard best subset analysis or a stepwise regression analysis.

An Excel workbook implementing both bootstrap methods is available from the author.

REFERENCES

Akaike, H. 1970. Statistical predictor identification. *Ann. Inst. Statist. Math.* 22 203-217.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

Hald, A. 1952. *Statistical Theory with Engineering Applications.* New York: Wiley.

Kleijnen, J.P.C., B. Bettonvil, and F. Persson. 2006. Screening for important factors in large discrete-event simulation models: sequential bifurcation and its applications. *In Screening Methods for Experimentation in Industry, Drug Discovery, and Genetics. Eds A. Dean and S. Lewis.* New York: Springer.

Krzanowski, W. J. 1998. *An Introduction to Statistical Modelling.* London: Arnold.

Makridakis, S., S. C. Wheelwright, and R. J. Hyndman. 1998. *Forecasting Methods and Applications, 3rd Ed.* New York: Wiley.

Mallows, C. L. 1973. Some comments on C_p. *Technometrics*. 15 661-675.

Nishii, R. 1984. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12, 758-765.

Searle, S. R. 1971. *Linear Models.* New York: Wiley.

Wu, C. F. J., and M. Hamada. 2000. *Experiments Planning, Analysis, and Parameter Design Optimization.* New York: Wiley.

AUTHOR BIOGRAPHY

RUSSELL C. H. CHENG is Emeritus Professor of Operational Research at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from Cambridge University, England. He obtained his Ph.D. from Bath University. He is a former Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society and the British Computer Society, Member of the Operational Research Society. His research interests include: variance reduction methods and parametric estimation methods. He was a Joint Editor of the *IMA Journal of Management Mathematics*. His email and web addresses are <R.C.H.Cheng@maths.soton.ac.uk> and <www.maths.soton.ac.uk/staff/Cheng>.