Computer Intensive Statistical Model Building

Russell Cheng

Abstract We consider resampling techniques in multiple linear regression where the objective is to identify a subset of the full set of explanatory variables that best captures the behaviour of the dependent variable, but using as few explanatory variables as possible. The total number of possible subsets or models grows exponentially with the number of explanatory variables, so a full examination of all possible models rapidly becomes intractable. The standard approach to this problem is to use a sequential selection procedure which avoids having to examine all subsets. When the number of explanatory variables is large there is a possible concern that good models might be missed. It is also important to examine whether the selected "best" model is the only good choice or whether other models might be equally satisfactory. We show how bootstrap resampling can handle both concerns in a simple way. In particular resampling enables a tractably small subset of good possible models to be selected as well as providing a method for comparing these models systematically. We describe the methodology and provide two numerical examples.

1 Introduction

This paper discusses the use of bootstrap (BS) resampling for tackling the well-known, but awkward, problem of model selection in multiple regression, when the number of possible explanatory variables is large. Our claim is that BS resampling is a simple and effective approach for this problem with distinct advantages over standard sequential methods that are often advocated and employed.

The ideas discussed in this paper were originally suggested for the exploratory study of a complex system using discrete event simulation. The basic methods were discussed by Cheng (2008) in that context. In this paper we discuss the methodology in more detail and more generally. In particular we consider the rationale of the methodology more fully and how to use it with the Mallows C_p criterion which is often suggested for handling this problem (see Krzanowski 1998 or Wu

R. Cheng (✉)
School of Mathematics, University of Southampton, Southampton, SO17 1BJ, UK
e-mail: r.c.h.cheng@soton.ac.uk

C. Alexopoulos et al. (eds.), *Advancing the Frontiers of Simulation: A Festschrift in Honor of George Samuel Fishman*, International Series in Operations Research & Management Science 133, DOI 10.1007/b110059_3, © Springer Science+Business Media, LLC 2009

and Hamada 2000, for example). The theoretical underpinning of bootstrapping in multiple regression is well-established; we will collect together the key results to underpin the methodology that we propose.

We suppose the dependent variable of interest is a (scalar) continuous random variable denoted by Y and that Y is linearly dependent on P explanatory variables X_j, $j = 1, 2, \ldots, P$. We are concerned with the *model selection problem* where we are interested in identifying simpler models in which some of the explanatory variables are omitted because they are actually unimportant. To avoid confusion we shall, from now on, use the term "model" to indicate that we are selecting a subset of explanatory variables, or *factors*, from the full set available, and use the term "full model" to indicate when all P explanatory variables are included. There are a total of 2^P distinct subsets of the explanatory variables, so that this is the number of models that we can choose from.

(Many authors exclude the null model $\mathbf{y} = \boldsymbol{\varepsilon}$ and so take the total number of distinct models to be $2^P - 1$ rather than 2^P. However, though the null model is very unlikely to be the best fit in applications, there seems no real reason for excluding it, and its exclusion can lead to misinterpretation of results if it happens to be the most appropriate model. We therefore do not exclude this possibility even if it is remote, and so take the total number of possible models as 2^P throughout this paper.)

Though the model selection 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 explanatory variable selection methods and also Bayesian strategies. The main problems with these methods are as follows.

The backward, forward and stepwise selection methods are all sequential, in which explanatory variables 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 explanatory variables are considered, to end up with a selected model that does not include all those explanatory variables that are important.

Use of a Bayesian approach avoids this difficulty, but a prior distribution for explanatory variable coefficient values has to be chosen and there are also technical implementation issues, such as deciding on the length of "burn-in" period and deciding when sufficient sampling has been carried out to ensure that adequate convergence to the posterior distribution has taken place.

In this paper we consider the use of BS resampling methods to generate a large number of data sets each with the *same* statistical distributional properties, at least asymptotically, as the original data set. We can therefore deploy whatever method we wish for selecting the model that best fits the original data sample (in some sense, to be defined), and then gauge the adequacy of the selected model by studying how consistently it is selected as the best fit in the BS samples, and how well it fits these samples.

We shall use the C_p statistic introduced by Mallows (1973, 1995) as the selection criterion for choosing between different models, as it is readily calculated in terms of ANOVA sums of squares and has a direct interpretation in terms of the prediction error, making it easy to understand and use. Several other criteria are asymptotically equivalent (see Nishii 1984).

We also consider in this paper the problem of checking whether the selected model is a sufficiently good fit. We are especially interested in the situation where there are a large number of factors. There is the strong possibility that there will be a number of models that are a satisfactory fit to the data. We need therefore to have some means for gauging the adequacy of competing fitted models. 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.

In this paper we propose an alternative approach, based on bootstrapping, to gauge the adequacy of selected models, as bootstrapping provides a natural way of demonstrating when there is little to choose amongst several, possibly many, models. It is of interest to note that such methods are now beginning to be recognized as very appropriate for model selection in simulation work. A good example is given by Fishman (2006, Section 2.8).

In Section 2 we describe the linear statistical model that we will use and discuss selection criteria for choosing between models. In Section 3 we discuss the Mallows C_p statistic and model selection. In Section 4 we discuss two ways of generating BS samples. We also give two methods using BS resampling for identifying a small but targeted number of promising models out of the full set of 2^P possible models for fitting to the original data. We also show how bootstrapping can also be used to assess the quality of models that seem to be a good fit to the original sample. Two numerical examples are given in Section 5, and a summary is provided in Section 6.

2 The Linear Model

We consider the (full) linear model

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

where Y_i, i =1, 2, ..., n are the observed output values obtained from n simulation runs; X_{ij} are the explanatory variable values in each of the n runs; b_j, j =1, 2, ..., P are the unknown coefficients corresponding to each of the P explanatory variables; and ε_i, $i = 1, 2, \dots, n$ are random errors. We have taken $X_{i1} = 1$, $i = 1, 2, \dots, n$ 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 ε_i, $i = 1, 2, \ldots, n$ are identically distributed with mean zero and variance

$$\mathrm{Var}(\varepsilon) = \sigma^2 \ . \tag{2}$$

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, where convenient, write (1) in the alternative matrix form

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon} \ . \tag{3}$$

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

$$m = \{j_1, j_2, \ldots, j_p\} \tag{4}$$

containing just the factor indices

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

if (and only if)

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

We shall write the observations corresponding to this model as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1j_1} & X_{1j_2} & \ldots & X_{1j_P} \\ X_{2j_1} & X_{2j_2} & \ldots & X_{2j_P} \\ \vdots & \vdots & \ddots & \vdots \\ X_{nj_1} & X_{nj_2} & \ldots & X_{nj_P} \end{bmatrix} \begin{bmatrix} b_{j_1} \\ b_{j_2} \\ \vdots \\ b_{j_P} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{5}$$

or in the matrix form

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

Where necessary we shall also write

$$p(m) = p \tag{7}$$

for the number of nonzero coefficients in the model m. Also we will denote the full model by M, so that $p(M) = P$.

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

$$\hat{\mathbf{b}}(m) = \left[\mathbf{X}^{\mathrm{T}}(m)\mathbf{X}(m)\right]^{-1}\mathbf{X}^{\mathrm{T}}(m)\mathbf{Y} \tag{8}$$

for the unknown coefficient values, and

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

for the unbiased estimate of the variance of the ε_i.

3 Methods for Selecting the Best Model

3.1 "Min C_p" and "Unbiased Min p" Selection Methods

In this section we consider various issues involved in selecting which model we think is the best. The first is the statistic to be used for measuring how well a given model fits the sample. One of the most popular is the C_p statistic proposed by Mallows (1973) which is an estimate of the expected prediction error taking into account the variance and bias of the fitted model. It is defined as

$$C_p(m) = [n - p(m)]\hat{\sigma}^2(m)/\hat{\sigma}^2(M) + 2p(m) - n\ . \tag{10}$$

An alternative statistic is the Akaike Information Criterion (Akaike 1970), which for the linear model reduces to $\mathrm{AIC}(m) = n\log[\hat{\sigma}^2(m)] + 2p(m)$, up to a constant depending on n but not on m. Asymptotically C_p and AIC have the same distribution (see Nishii 1984). However C_p is perhaps more satisfactory for our purpose because of its ease of interpretation. Mallows (1973) shows that if the model m (with p factors) is satisfactory in the sense that it has no bias, then the expected value of C_p is close to p, that is:

$$C_p \approx p\ . \tag{11}$$

However, if not all important factors are included, the expected value of C_p will be larger than p. A simple selection method is therefore the following.

"Min C_p" Model Selection Method

(i) Consider each of the 2^P possible models of (1) and for each model m calculate $C_p(m)$.
(ii) Select as the best model that m for which $C_p(m)$ is minimum, with the expectation that this model will be satisfactory if $C_p(m) \le p$.

This provides a simple selection method if we are able to examine all possible models.

As mentioned previously, an exhaustive search of all possible models can be avoided by using a sequential procedure, several of which are cited by Mallows (1995). Mallows points out that if the "min C_p" method is used in a sequential procedure, and if m^+ is a model containing one factor additional to those already in a model m, then the extra factor would be worth including if

$$C_{p+1}(m^+) - C_p(m) = 2 - (S_1/\hat{\sigma}^2(M)) < 0 \ ,$$

where S_1 is the 1-df sum of squares due to the additional factor. This criterion for inclusion is therefore equivalent to carrying out a t-test, with the factor included if

$$t^2 = S_1/\hat{\sigma}^2(M) > 2 \ . \tag{12}$$

In the non-orthogonal case the final selected model is dependent on the order in which factors are considered, but for an orthogonal design the sum of squares S_1 corresponding to each factor does not depend on the model fitted. Thus there is no need for a sequential procedure in this latter case. The minimum C_p is easily obtained by fitting the full model and then applying the test (12) in "blanket" fashion, i.e. simultaneously, to every factor sum of squares. The "min C_p" model then includes just those factors that satisfy (12).

The attraction of the orthogonal case is that the inclusion or exclusion of each factor is decided just from fitting the full model. We shall consider use of the same procedure in the non-orthogonal case, together with an adjustment to deal with the problem of including too many unimportant factors. We still fit the full model, and for each factor j calculate the so called t-value of its fitted coefficient $\hat{b}_j$:

$$t_j = \hat{b}_j/s_j \ , \tag{13}$$

where $s_j = \sqrt{d_j\hat{\sigma}^2(M)}$ is the estimated standard deviation of $\hat{b}_j$, with d_j the jth entry in the main diagonal of the dispersion matrix, i.e.

$$d_j = \left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right]_{jj} \ . \tag{14}$$

Our selected model is the one that includes only those factors j for which

$$|t_j| > a \ , \tag{15}$$

where a is a chosen critical level. 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 $\bar{T}_{n-P}(\cdot)$, then the probability of success of the test (15) under the assumption that $b_j = 0$, is

$$\pi_a = \Pr\{|t_j| > a\} = \bar{T}_{n-P}(a) \ .$$

A common alternative way of carrying out this test is to report the so-called p-value of the estimate $\hat{b}_j$, namely $\bar{T}_{n-P}(|t_j|)$, so that the factor j is retained if

$$\bar{T}_{n-P}(|t_j|) \ < \pi_a \ . \tag{16}$$

It will be seen that (12) is the special case of (15) or (16) where $a = \sqrt{2}$, with a corresponding critical p-value in (16), when $n - P$ is large, of $\pi_a = 0.1573$. This highlights a problem with using the "min C_p" method for selecting a model when the initial number of factors under consideration is large but where the (unknown) true values of many coefficients are at or near zero, as the selection test (16) would then include nearly 16% of such negligible coefficients in the model.

The effect of varying a can be seen more fully by considering the asymptotic probability that a factor with coefficient of size $b_j = bs_j$ is selected, when we allow b to vary also. For simplicity we assume that $n - P$ is large as then s_j can be treated essentially as being a known constant, so that $\hat{b}_j \sim N\left(bs_j, s_j^2\right)$. The probability we would include the factor is then

$$\begin{aligned}
\Pr\{\text{Factor } j \text{ is included in model}\} &= 1 - \Pr\{-as_j < \hat{b} < as_j\} \\
&= 1 - \Pr\{-a - b < (\hat{b} - bs_j)/s_j < a - b\} \\
&= 1 - \Phi(a - b) + \Phi(-a - b) \ , \qquad (17)
\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Figure 1 shows how this probability varies as a function of b for different selected a. It will be seen that somewhat larger values than $a = \sqrt{2}$ in (12), such as $a = \sqrt{6}$ or $a = 3$ might be more appropriate in exploratory studies where we are only interested in identifying

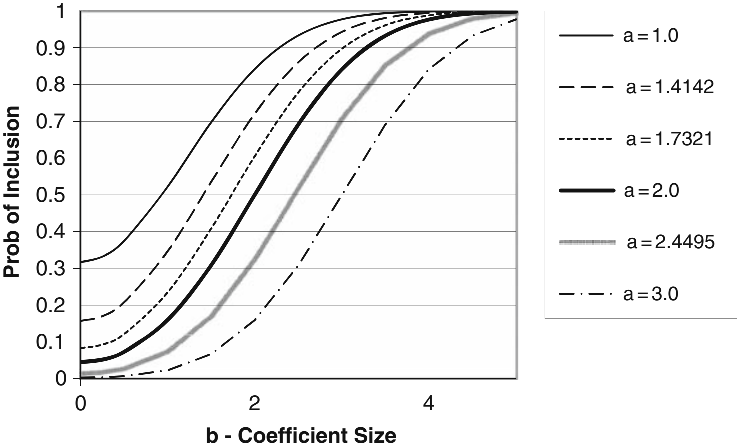


Fig. 1 Asymptotic probability that a factor with coefficient size b is included in a model using the t test with critical level a

significantly large b and would prefer the probability of retaining a zero coefficient to be much smaller than 16%.

The problem of unnecessarily including factors with zero coefficients is controlled by choice of a suitably large a. We can underline this choice by using a modified form of the "min C_p" model selection method. From (11) we know it is undesirable to select a model for which $C_p > p$. This suggests the following model selection procedure:

"Unbiased Min *p*" Model Selection Method

(i) Find the smallest p for which there are models m satisfying $C_p(m) \leq p$ and let

$$p_0 = \min\{p : C_p(m) \leq p\} \, . \tag{18}$$

(ii) Amongst all such models m, with $p(m) = p_0$, find the one for which $C_p(m)$ is minimum.

A simple way, at least in principle, of identifying this model is to plot C_p versus p for all possible models and look at the lower envelope of this scatterplot of points. For the orthogonal case where there are a large number of factors with coefficient values uniformly distributed in the neighbourhood of zero with density λ, Mallows (1995) has shown that the scatterplot has a lower boundary that is the (convex) cubic polynomial in p

$$C_p - P \approx \frac{(P-p)^3}{12\lambda^2} - 2(P-p) \, , \tag{19}$$

and that this boundary intersects the line $C_p = p$ at $P - p = 2\sqrt{3}\lambda$. Figure 2 depicts the scatterplot for the first example involving epoxide bonding that we will be discussing in Section 5, and this boundary and its intersection with the line $C_p = p$ are clearly distinguishable.

Our selection method (18) will clearly select a model corresponding to a point near this intersection. Specifically (18) requires finding the smallest p, p_0, for which there are points of the scatterplot below the line $C_p = p$ and then finding amongst those models with $p = p_0$, the one with minimum C_p.

In the orthogonal case, models at, or near, this intersection point will tend to include just those factors for which (15) is satisfied with $a = \sqrt{3}$, which is equivalent to using (16) to include just those factors whose estimated coefficients have p-value less than $\pi_a = 0.083$.

The condition (11) that $C_p \approx p$, obtains when the model contains no bias so that the model is completely appropriate whilst having the smallest p possible. For this reason we call (18) the "unbiased min p" method.

We delay discussion of how precisely to implement this method of model selection until we have discussed bootstrapping, as our proposed implementation will involve bootstrapping intimately.

3.2 Dimensionality Problem

As previously mentioned a critical issue that arises in model selection is the *dimensionality problem.* Because the total number of possible models, 2^P, grows exponentially with P, inspection of all models is tractable only when P is small. Thus even with just 20 explanatory variables there are already 1,048,576 models. Our approach is to identify a set of *promising models* using bootstrap resampling. The number of models in this set is easily controlled and so can be made much smaller than 2^P. But we shall show that it will almost certainly contain many good candidate models. It is thus satisfactory to select a "best" model from this subset.

We discuss bootstrapping in the next section.

4 Bootstrap Analysis

We shall use bootstrapping for two distinct purposes. Firstly, as already mentioned in the previous section, it can be used for identifying a set of promising models. However we shall also use bootstrapping to deal with the following second problem.

Once a model has been selected as being the best fit to a data set, we have the problem of determining what might be termed the *quality of the selected model*. For example, if we have used the "min C_p" method to select the model, there may be several models with values of $C_p(m)$ close to that of the best, so that we may not be sure which model really is the best. This question would be answered if we had many (independent but identically distributed) data samples and not just the one original sample, as we could determine the best model for each sample and see if the same model is best for all the samples. BS resampling enables such additional data samples to be generated.

We first outline how BS samples are generated in the next subsection, before going on to describe our two distinct uses of bootstrapping.

4.1 Bootstrap Samples

We describe first two ways of generating BS samples that asymptotically have the same form as (1). The standard way is described, for example, by Davison and Hinkley (1997). We take the modified residuals

$$r_i = (Y_i - \hat{Y}_i)/(1 - h_{ii})^{1/2}, \quad i = 1, 2, \ldots, n \tag{20}$$

obtained from the fitting the full model M 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

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

We then centre these so that their average is zero:

$$e_i = r_i - \bar{r}, \quad i = 1, 2, \ldots, n. \tag{21}$$

A BS sample is then obtained by forming

$$Y_i^* = \hat{Y}_i + e_i^*, i = 1, 2, \ldots, n \ , \tag{22}$$

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

A second way of resampling, *parametric bootstrapping*, is possible, if it can be assumed that the random errors ε_i, $i = 1, 2, \ldots, n$ in (1) are normally distributed and independent. The BS sample still takes the same form as (22) only now the e_i^*, $i = 1, 2, \ldots, n$ are a random sample from the fitted normal distribution, i.e.

$$e_i^* \sim N\big(0, \hat{\sigma}^2\big), \quad i = 1, 2, \ldots, n \ . \tag{23}$$

In either case we write $\hat{\mathbf{b}}^*$ and $\hat{\sigma}^{*2}$ for the estimates (8) and (9) obtained from fitting the model (1) to the BS observations (22).

The justification for bootstrapping is provided by Freedman (1981, Theorem 2.2). Assume that (1) and (2) hold and that $\mathbf{X}(n)$ is not random with

$$\frac{1}{n}\mathbf{X}^{\mathrm{T}}(n)\mathbf{X}(n) \to \mathbf{V} \text{ which is positive definite} \tag{24}$$

as $n \to \infty$. Then

$$\sqrt{n}\big\{\hat{\mathbf{b}}^*(n) - \hat{\mathbf{b}}(n)\big\} \text{ converges to } N\big(\mathbf{0}, \sigma^2\mathbf{V}^{-1}\big) \tag{25}$$

and

$$\hat{\sigma}^*(n) \text{ converges to a point mass at } \sigma \ . \tag{26}$$

The above result assumes that P is fixed as $n \to \infty$. We shall tacitly assume this in what follows. However a more refined treatment would allow P to become large as $n \to \infty$. Shibata (1981) has discussed the selection of factors for this regression problem. We shall not discuss this case explicitly here.

4.2 Bootstrap Generation of a Set of Promising Models

The "unbiased min p" method of selecting a best model does not require consideration of all 2^P models but only those near p_0, as defined in (18). Our first use of bootstrapping is therefore to generate a set of *promising models*. The number of models in this set does not need to be anywhere near 2^P, but it does need to be large enough to enable the lower boundary (19) to be clearly identified, at least near its intersection with $C_p = p$. Ideally it needs to contain all the models with scatterplot points near this intersection point. With these considerations in mind, our first proposed methods is:

"One Model per Sample" Generation of Promising Models by Bootstrapping

Step (1) Fit the full model to the original data and use this fitted full model to generate B BS samples each of the form (22).

Step (2) Set a critical t-value, a, (in view of our discussion in Section 3.1, we used $a = \sqrt{3}$) and construct one *promising* model for each BS sample as follows.

(i) Fit the full model, M, to the sample and calculate the t-value, t_j, as defined in (13), of each of the fitted coefficients, $\hat{b}_j$, $j = 1, 2, \ldots, P$.

(ii) Include in the promising model just those factors with t_j satisfying

$$|t_j| \geq a \ ;$$

see (15) above. Not all the promising models obtained in the above process will be distinct (in the sense of each model containing a subset of factors that is different from those of all other selected models). Denote the set of distinct models by S.

The above method produces at most B promising models, but can be far fewer, if the same model is repeatedly obtained from different BS Samples. If it were felt that the number of models needs to be increased, especially as we would want to include most if not all models satisfying $p = p_0$ and $C_p \approx p$, then the following variant of the "One model per sample" adds models in a straightforward way.

"Many Models per Sample" Generation of Promising Models by Bootstrapping

Step (1) Fit the full model to the original data and use this fitted full model to generate B BS samples, each of the form (22).

Step (2) For each BS sample:

(i) Fit the full model, M, to the sample and determine, as defined in (13), the t-value, t_j, of each of the fitted coefficients, $\hat{b}_j$, $j = 1, 2, \ldots, P$.

(ii) Order the coefficients by their $|t_j|$ values:

$$|t_{j_1}| \geq |t_{j_2}| \geq \cdots \geq |t_{j_P}|, \tag{27}$$

so that $\hat{b}_{j_1}$ is the most significant.

(iii) Set a critical t-value, a (we used $a = \sqrt{3}$ as before), and include all the following models in the promising set S:

$$\begin{aligned} m_1 &= \{j_1\} \\ m_2 &= \{j_1, j_2\} \\ &\vdots \\ m_k &= \{j_1, j_2, \ldots, j_k\}\ , \end{aligned} \tag{28}$$

where the last factor j_k satisfies

$$|t_{j_k}| \geq a > |t_{j_{k+1}}|. \tag{29}$$

Thus the model m_i is the one where the i most significant factors have been retained, with a cutoff that only factors with t-level greater than a are allowed in a model. So the last model, m_k in (28) is the one that includes just those coefficients with $|t|$-value a or greater, this being the sole model selected in Step (2) of the "One model per sample" method.

4.3 Bootstrap Quality Assessment of Selected Best Model

Once a set of promising models has been obtained, we can use the "unbiased min p" method to select the "best" model. That is we fit each promising model to the original data set, calculating C_p for each model; then we identify p_0 as in (18), and select as the best model the one with the smallest C_p subject to $p \leq p_0$ (checking that it satisfies the condition $C_p \leq p$).

We can now use bootstrapping to study the quality of the selected model. This is most easily done by adding the following steps to either of the bootstrap methods proposed in the previous section for generating a set of promising models.

Bootstrap Assessment of Selected Best Model:

Step (3) For each of the B BS samples, fit the set S of promising models, subject to the restriction that only models where $p \leq p_0$ are considered (we shall denote this restricted set of promising models by S_0) and calculate the C_p value for each model, selecting as the best model for this sample, that which minimizes C_p.

Step (4) Display the models of S_0, ranked in order of the frequency with which they are selected as being the best model in the B BS samples, displaying these frequencies as well.

Step (5) Display the empirical distribution functions of the C_p values of a selected number of those models in S_0 most frequently selected as being the best.

Let $\alpha(m)$ be the probability that model m will be selected as the best model in the sense of minimizing C_p amongst all models with $C_p \leq p_0$. Step (3) estimates these probabilities by fitting all the models in the restricted set S_0 of promising models to each of the BS samples and then selecting the best model (for the given BS sample) from this set. Note that, out of the full set of 2^p models, those that are not a good fit will have very little probability of being included in the set S, because of the way S is constructed. Hence they would not be considered for possible inclusion in S_0. Nevertheless every model has a *positive* probability of being included in S. Thus asymptotically, as $B \to \infty$, the restricted set S_0 of promising models considered in Step (3) above must tend to the full set of all models with $p \leq p_0$. This holds for either method of generating the set of promising models described in Section 4.2. Thus, as $B \to \infty$, Step (3) will converge to the exact situation where every model satisfying $p \leq p_0$ is considered for possible selection as the best. Hence, for each model m, $\alpha(m)$ can reasonably be estimated from the frequency with which m is selected as being the best model in Step (3).

The version of Step (3) given above concentrates on models with $p \leq p_0$, where p_0 is calculated from the original sample, in order to check how well this value p_0 performs. Different variants of Step (3) are possible. An alternative would be not to impose the condition $p \leq p_0$ at all, but instead simply to apply the "unbiased min p" selection procedure to each BS sample separately, using the full set S of promising models with each BS sample.

In Step (4) we simply display those models that have been most frequently selected as being the best fit.

The point of Step (5) is to assess the behaviour of the C_p values of those models that have been most frequently selected as being the best fit. For such a model to be satisfactory one would expect the distribution of its C_p value, over the BS samples, to be concentrated mainly in the region where $C_p \leq p$.

5 Numerical Examples

We give two examples. Both involve readily accessible real data samples. The first is a data set where the design matrix is orthogonal. As already remarked immediately after (12), the obvious strategy in this case of applying a test such as (12) simultaneously to each estimated coefficient gives an unambiguous selection strategy and cannot really be bettered. The analysis is thus straightforward in this case. However we include the example simply to demonstrate the way the resulting bootstrap analysis works. The second was discussed by Cheng (2008). Here we discuss the selection method more fully.

5.1 Epoxide Bond Example

The first example is data given by Williams (1968) and reproduced in Wu and Hamada (2000, Table 8.6). This measured the adhesion of an epoxide bonding system in an orthogonally designed experiment with, including a general mean, 25 factors, and 28 observations.

Our analysis is in two stages as set out in Sections 4.2 and 4.3.

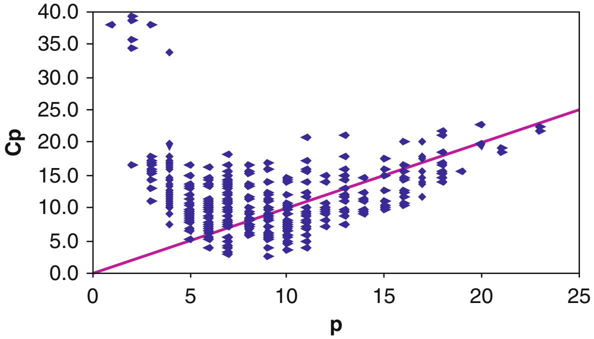


Fig. 2 C_p versus p plot of 320 promising models found for the Epoxide Bond data using the "One model per sample" BS method. The C_p values are those obtained when the promising models are fitted to the original sample

The first stage generates a set of promising models. We used the "One model per sample" BS method of Section 4.2 with B =500, $\pi_a = 0.083$, and with no limit placed on the maximum number of factors that can be included in a fitted model.

Step (1) of the analysis produced an initial subset of 320 promising models. The C_p values of these models when they were fitted to the *original* data are plotted against p in Fig. 2. Applying (18) gives $p_0 = 6$.

In the second stage we could have just used the previously generated promising models, but removing those with $p > p_0$. Instead we increased the number of promising models by using the "Many models per sample" BS method with $B = 500$, $\pi_a = 0.083$ but with the maximum number of factors permitted in a model limited to $p_0 = 6$. This yielded a set, S_0, of 488 promising models. Then, as described in Section 4.3, the models of S_0 were fitted to each BS sample, and the model with the minimum C_p was selected as being the best model for that BS sample. This yielded 235 different best models. The plot of the C_p values obtained by fitting each of these models to the *original* data set is given in Fig. 3.

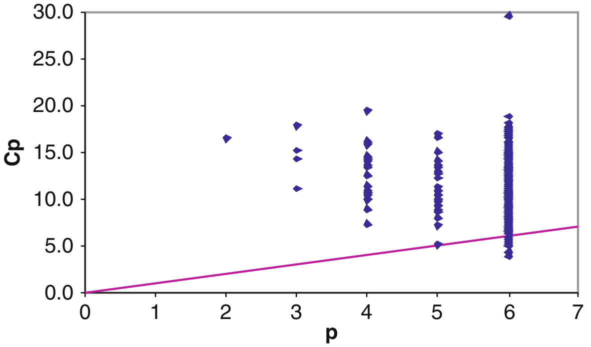


Fig. 3 C_p versus p for the 235 "best" models found by the Bootstrap Quality Assessment Method described in Section 4.3. The C_p values are those when the models are fitted to the original data

The frequency with which each of the models of S_0 was selected as being the best varies with model. The 25 models selected most frequently are displayed in Table 1. The model selected most frequently was

$$X_0,\ \ X_4,\ \ X_{14},\ \ X_{15},\ \ X_{16},\ \ X_{19}\ ,$$

where X_0 is the mean. This model was selected as being the best model (i.e., with the smallest C_p) in 14 of the BS samples.

The model with the lowest C_p value in Fig. 3 was

$$X_0,\ \ X_4,\ \ X_{15},\ \ X_{16},\ \ X_{19},\ \ X_{21},$$

which was the 4th best in terms of the number of times it was selected as best in the BS samples. There is little to choose between the top few models. Taken together it is fairly clear that factors X_0, X_{15}, X_{19} are the most important followed by X_4, X_{16}, and X_{21}.

5.2 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 explanatory variables (D1–D11), and three further variables, time t and its square and cube t^2, t^3, making 17 initial explanatory 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 Web site `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 factor so that we work with 18 explanatory factors. There are thus $2^{18} = 262{,}144$ distinct models to select from; a somewhat large number of models to comfortably work through.

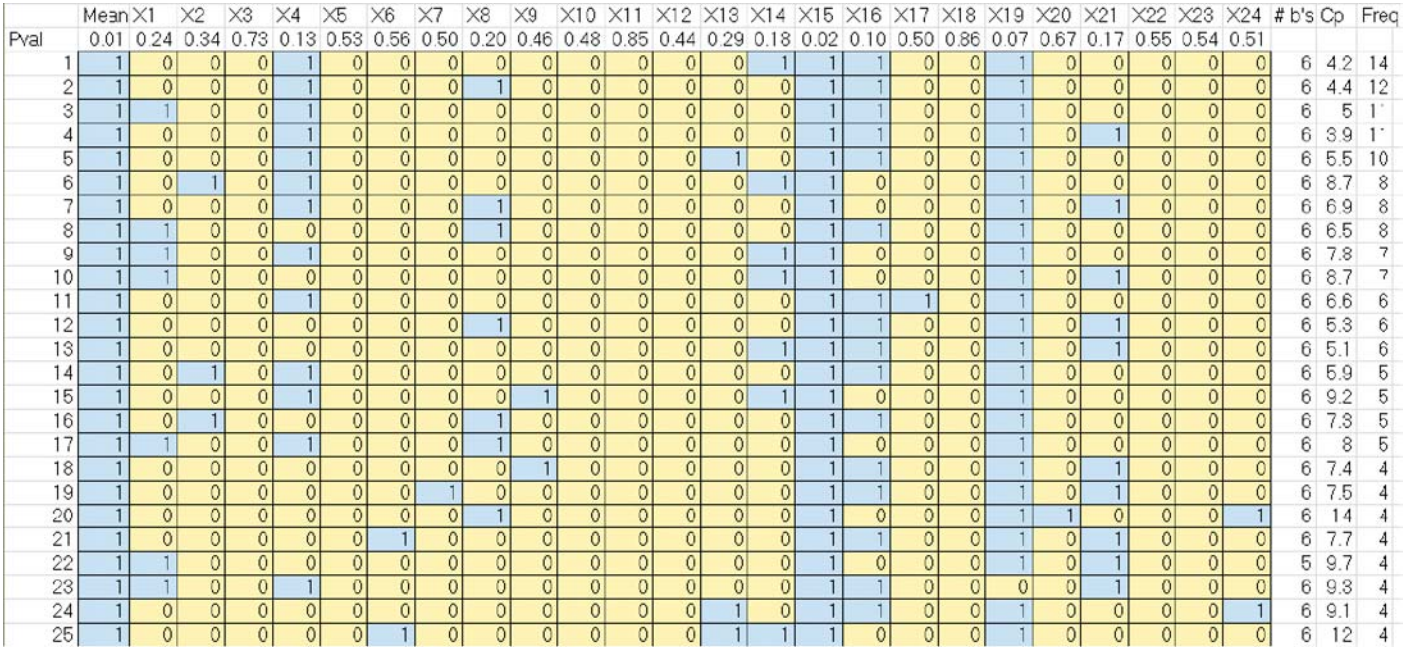
Using a best subset analysis with an adjusted coefficient of determination, $\bar{R}^2$, for selection criterion Makridakis et al. found the best model overall was

$$X_0\ X_1\ X_2\ X_3\ D_2\ D_3\ D_4\ D_5\ D_6\ D_7\ D_8\ D_9\ D_{10}\ D_{11}\ t^3 \tag{30}$$

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

$$X_0\ X_1\ X_2\ X_3\ D_2\ D_4\ D_6\ D_7\ D_8\ D_9\ D_{10}\ D_{11}\ t^3. \tag{31}$$

Table 1 Epoxide bond data, top 25 of final selection of 235 "best" models



	Mean	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	# b's	Cp	Freq
Pval	0.01	0.24	0.34	0.73	0.13	0.53	0.56	0.50	0.20	0.46	0.48	0.85	0.44	0.29	0.18	0.02	0.10	0.50	0.86	0.07	0.67	0.17	0.55	0.54	0.51			
1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	6	4.2	14
2	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	6	4.4	12
3	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	6	5	1
4	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	6	3.9	1
5	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	6	5.5	10
6	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	6	8.7	8
7	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	6	6.9	8
8	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	6	6.5	8
9	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	6	7.8	7
10	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0	6	8.7	7
11	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	6	6.6	6
12	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	6	5.3	6
13	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	1	0	0	0	6	5.1	6
14	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	6	5.9	5
15	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	6	9.2	5
16	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	6	7.3	5
17	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	6	8	5
18	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	6	7.4	4
19	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	6	7.5	4
20	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	1	6	14	4
21	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	6	7.7	4
22	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	5	9.7	4
23	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	6	9.3	4
24	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	1	6	9.1	4
25	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	6	12	4

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

We have carried out the same analysis as in the epoxide bond example. We used the “One model per sample” method to generate a set of promising models, with $B = 500$, a critical p-value of $\pi_a = 0.083$, and with no limit to the number of factors that could be included in the model fitted to each BS sample. This led to the generation of a set of just 189 promising models at the end of Step (1). The C_p versus p plot of these models fitted to the original data is shown Fig. 4. From this we took the “unbiased min p” as $p_0 = 13$.

As in the epoxide bond example, to increase the number of promising models with $p \le p_0$ that we examine, we used the “Many models per sample” method to generate a fresh set of promising models only with p restricted to being no greater than $p_0 = 13$, but still with $\pi_a = 0.083$ and $B = 500$. This gave a set 598 promising models. We then used the “Bootstrap Assessment of Selected Best Model” method of Section 4.3 to fit all the promising models to each BS sample, selecting the “min C_p” model for each sample, as the “best” model for that sample. This yielded just 74 “best” models. The C_p versus p plot of these models when fitted to the original data is shown in Fig. 5. The top 25 models are listed in Table 2 with their original C_p values as well as the number of times they were selected as the best model in the BS samples.

The most frequently selected model was

$$X_1\ X_2\ X_3\ D_2\ D_4\ D_6\ D_7\ D_8\ D_9\ D_{10}\ D_{11}\ t\ t^2\ .$$

This was selected as the best model in 75 of the 500 BS samples. This was also the model with the smallest C_p amongst all 74 “best” models when fitted to the original data.

The stepwise regression model

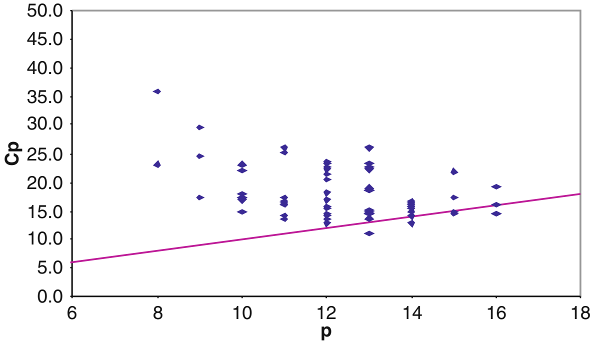
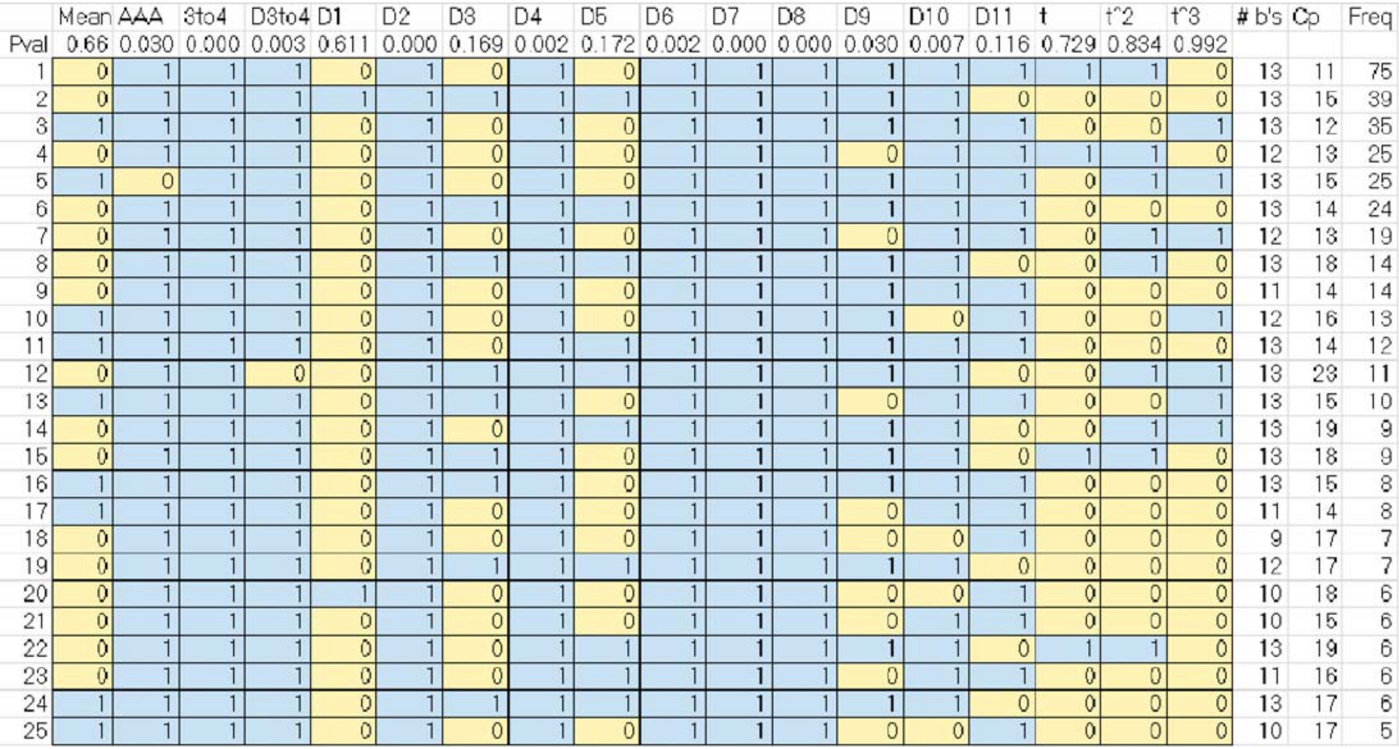


Fig. 4 C_p versus p plot of 320 promising models found for the Bank data using the “One model per sample” BS method. The C_p values are those obtained when the promising models are fitted to the original sample

Table 2 Bank data, top 25 of final selection of 235 “best” models



	Mean	AAA	3to4	D3to4	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	t	t^2	t^3	# b's	Cp	Freq
Pval	0.66	0.030	0.000	0.003	0.611	0.000	0.169	0.002	0.172	0.002	0.000	0.000	0.030	0.007	0.116	0.729	0.834	0.992			
1	0	1	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	0	13	11	75
2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	13	15	39
3	1	1	1	1	0	1	0	1	0	1	1	1	1	1	1	0	0	1	13	12	35
4	0	1	1	1	0	1	0	1	0	1	1	1	0	1	1	1	1	0	12	13	25
5	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	0	1	1	13	15	25
6	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0	13	14	24
7	0	1	1	1	0	1	0	1	0	1	1	1	0	1	1	0	1	1	12	13	19
8	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	0	13	18	14
9	0	1	1	1	0	1	0	1	0	1	1	1	1	1	1	0	0	0	11	14	14
10	1	1	1	1	0	1	0	1	0	1	1	1	1	0	1	0	0	1	12	16	13
11	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	0	0	0	13	14	12
12	0	1	1	0	0	1	1	1	1	1	1	1	1	1	0	0	1	1	13	23	11
13	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	0	0	1	13	15	10
14	0	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	1	1	13	19	9
15	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	0	13	18	9
16	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	0	0	13	15	8
17	1	1	1	1	0	1	0	1	0	1	1	1	0	1	1	0	0	0	11	14	8
18	0	1	1	1	0	1	0	1	0	1	1	1	0	0	1	0	0	0	9	17	7
19	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	0	12	17	7
20	0	1	1	1	1	1	0	1	0	1	1	1	0	0	1	0	0	0	10	18	6
21	0	1	1	1	0	1	0	1	0	1	1	1	0	1	1	0	0	0	10	15	6
22	0	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	0	13	19	6
23	0	1	1	1	0	1	0	1	1	1	1	1	0	1	1	0	0	0	11	16	6
24	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	0	13	17	6
25	1	1	1	1	0	1	0	1	0	1	1	1	0	0	1	0	0	0	10	17	5

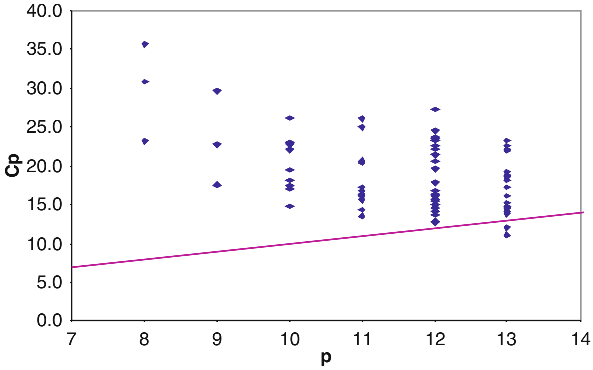


Fig. 5 C_p versus p for the 74 "best" models found by the Bootstrap Quality Assessment Method described in Section 4.3. in the Bank Data Example. The C_p values are those when the models are fitted to the original data

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

was the third most frequently selected, being selected 35 times.

The results suggest that it is not very important whether the mean is fitted or not. 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 significantly different from zero for the original data.

For all the 74 "best" models that were selected, the three main explanatory variables X_1 (AAA), X_2 (3-4), X_3 D(3-4) were clearly important, as were the seasonal variables D_2, D_4, D_6, D_7, D_8 and of the others D_9, D_{10} 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 given the rather random way that different time variables appear in the different models; this is similar to variations listed in Table 6-10 of Makridakis et al. (1998).

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

Finally it is interesting to see how the C_p values of the top-performing models in Table 2 varied across all 500 BS samples to which they were fitted. Figure 6 shows the empirical distribution function of the sample of 500 C_p values for the top 5 models. The result shows the inherent variability of the statistic for data of this type.

6 Conclusions

We have discussed how bootstrapping can be used to analyse the selection and fitting of linear models in multiple regression. We have shown how bootstrapping can be used for two purposes.

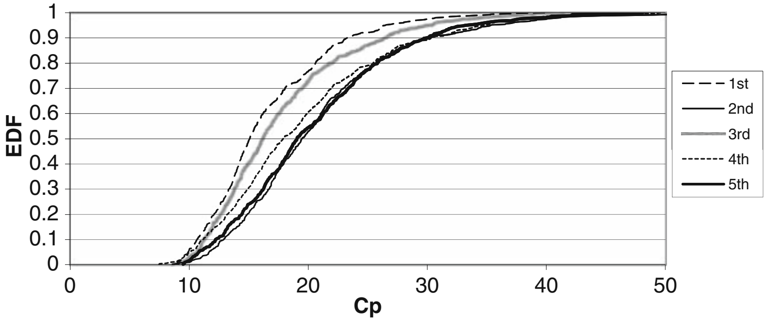


Fig. 6 Empirical Distribution Functions of the C_p value of the top 5 "best" models when fitted to 500 BS samples of the Bank Data

First it can be used to identify promising models out of the set of 2^P possible models. The "One model per sample" method yields just one model per BS sample, so that the largest number of possible models is B, the number of BS samples generated, though because of duplication, the number of distinct models (i.e., whose subsets of factors are different) is likely to be rather smaller. The "Many models per sample" method produces a maximum of BP models, though again duplication means the number of distinct models is usually significantly smaller.

The way that the set of promising models is constructed means that models with a small C_p value are likely to be identified, as is borne out in the two numerical examples. Thus bootstrapping seems attractive in enabling promising models to be tractably identified out of the full set of all possible models when the number of factors is large.

The bootstrapping also allows an assessment to be made of how stable the models estimated as being the best, or a good fit to the original data, actually are, 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 BS 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 at `http://www.personal.soton.ac.uk/rchc/BestLinModel.htm`.

References

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

Cheng, R. C. H. 2008. Selecting the best linear simulation metamodel. In *Proceedings of the 2008 Winter Simulation Conference*, eds. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, pp. 371–378. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online via `www.informs-sim.org/wsc08papers/043.pdf` [accessed December 29, 2008].

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

Fishman, G. S. 2006. *A First Course in Monte Carlo.* Australia: Thomson Brooks/Cole.

Freedman, D. A. 1981. Bootstrapping regression models. *Ann. Statist.* 9:1218–1228.

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

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 edition New York: Wiley.

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

Mallows, C. L. 1995. More comments on C_p. *Technometrics* 37:362–372.

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.

Shibata, R. 1981. An optimal selection of regression variables. *Biometrika* 68:45–54.

Williams, K. R. 1968. Designed experiments. *Rubber Age* 100:65–71.

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