British Journal of Mathematical and Statistical Psychology (2001), **54**, 201–225 *Printed in Great Britain*

Criticality of predictors in multiple regression

Razia Azen*

Department of Educational Psychology, University of Wisconsin-Milwaukee, USA

David V. Budescu

Department of Psychology, University of Illinois at Urbana-Champaign, USA

Benjamin Reiser

Department of Statistics, University of Haifa, Israel

A new method is proposed for comparing all predictors in a multiple regression model. This method generates a measure of predictor criticality, which is distinct from and has several advantages over traditional indices of predictor importance.

Using the bootstrapping (resampling with replacement) procedure, a large number of samples are obtained from a given data set which contains one response variable and p predictors. For each sample, all $2^p - 1$ subset regression models are fitted and the best subset model is selected. Thus, the (multinomial) distribution of the probability that each of the $2^p - 1$ subsets is 'the best' model for the data set is obtained.

A predictor's criticality is defined as a function of the probabilities associated with the models that include the predictor. That is, a predictor which is included in a large number of probable models is critical to the identification of the best-fitting regression model and, therefore, to the prediction of the response variable.

The procedure can be applied to fixed and random regression models and can use any measure of goodness of fit (e.g., adjusted R^2, C_p, AIC) for identifying the best model. Several criticality measures can be defined by using different combinations of the probabilities of the best-fitting models, and asymptotic confidence intervals for each variable's criticality can be derived. The procedure is illustrated with several examples.

1. Introduction

Multiple regression (MR) models are used to predict a single criterion variable from several predictors. Consider the MR model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}$ is an $n \times 1$ data vector (the criterion), $\mathbf{X}$ is an $n \times (p + 1)$ full-rank data matrix (the predictors); $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of unobservable 'error' terms and $\boldsymbol{\beta}$ is a $(p + 1) \times 1$ vector of parameters, estimated from the data by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Here p represents the

* Requests for reprints should be addressed to Dr Razia Azen, Department of Educational Psychology, University of Wisconsin-Milwaukee, PO Box 413, Milwaukee, WI 53201, USA.

number of predictors in the 'full' model (i.e., the maximum number of variables available for predicting the criterion variable), n represents the number of observations in a data set, and k represents the number of predictors in a particular (subset) regression model ($1 \leq k \leq p$).

Two main concerns, related to two distinct stages of MR analysis, often arise. First, at the model selection stage, one wishes to determine if all predictors are necessary, or if there is a *subset* of predictors which can adequately predict the criterion. Second, at the interpretation stage, once a subset of predictors is chosen one often wishes to determine how to rank-order them in terms of how important, useful, necessary or relevant they are for predicting the criterion.

There are two distinct bodies of statistical literature related to these two questions. The first focuses on the development, validation, justification and implementation of various model selection procedures. A portion of this literature is concerned with various sequential procedures (forward inclusion, backward exclusion, stepwise methods). More recent work is concerned with the choice of optimal measures for identification of the best subset models. In addition to the traditional measures (R^2, adjusted R^2, mean square error), most statistical textbooks discuss (and most statistical software packages implement) other meaningful measures such as C_p (Mallows, 1973), Akaike's (1973) information criterion (AIC), and so on. For partial reviews of this literature, the reader is referred to Hocking (1976) or Judge, Griffiths, Hill, Lütkepohl & Lee (1985). Regardless of the measure of optimality chosen by a researcher, and its mode of implementation, the final product of this stage of the analysis is the selection of a subset of k ($k \leq p$) predictors that constitute the best model for a given data set.

The second literature focuses on the comparison of the predictors included in the model in an effort to rank and scale them in terms of their importance. Many measures have been proposed to quantify the concept of 'predictor importance' in MR. A review of the different measures can be found in Budescu (1993). Conceptually, importance refers to a predictor's ability to predict the criterion such that more important predictors are those that contribute more to the overall prediction of the response. In practice, however, goodness of prediction can be quantified in a wide variety of ways and interpreted in almost as many. Most measures of predictor importance attempt to partition some overall goodness-of-fit measure (such as R^2) among all the predictors in the model. Many of the problems with these measures generally stem from the fact that 'importance' is not clearly defined or agreed upon and therefore these measures fail to capture the connotations and implications of 'importance' in a scientific theory. For example, most measures of importance are model-dependent: they are not invariant across all subsets of the predictors. Clearly, most scientists examining all subset models would be puzzled by a conclusion such as 'X_1 is more important than X_2 in the presence of X_3, but X_2 is more important than X_1 when X_3 is absent from the model'. Such conclusions are particularly hard to accept when X_1, X_2 and X_3 are meaningful predictors embedded in a well-understood theoretical context.

It is interesting to note that these two literatures are almost independent of each other. Nothing in the work on predictor importance depends on, or is sensitive to, the method and criterion used to select the variables in the model (i.e., those variables which are to be ranked in terms of their importance). On the other hand, the literature on model selection is not concerned at all with the nature, identity and interrelations among the variables included in the model. In fact, it is fair to say that this literature implicitly assumes that, in some sense, all variables in the selected model are equally important. This conclusion is sensible and easy to justify: after all, each predictor in the subset is *necessary* to make the chosen model the best. If a variable is omitted (regardless of how unimportant it may be according to regular measures of importance), the reduced subset may be outperformed by another set of predictors and no longer be the best. In addition, one may argue that all predictors in the best model are equally important because if the model is misspecified many of the attractive and optimal properties of the parameters do not hold. For example, the estimates may no longer be unbiased and/or have minimum variance.

The notion that all predictors in the best (selected) model are equally important implies that there is, in fact, a *single true best model* for a given phenomenon or behaviour, and that this best model can be identified unequivocally based on the sample evidence.

There are, of course, situations in which MR analysis is used as a theory-testing tool and its results are used to make inferences about the truth or falsehood of a given theory. In psychological research these cases often involve questions about the necessity of interactive terms; examples are given by Busemeyer and Jones (1983), Ganzach (1997) and Lubinski and Humphreys (1990). However, in many situations the assumption that there is a true model underlying the data is patently false. This is particularly true when MR is used as a predictive (not necessarily theory-driven) tool and/or in exploratory research. Consider the following illustrations:

- A financial forecasting firm wants to develop a new ‘market index’, that is a linear combination of a relatively small number of individual stocks, $X_1, \ldots, X_p$, that can predict accurately the performance of the stock market, Y, on a weekly basis.
- A political scientist tries to predict the likelihood that individuals would vote in the next election, Y, as a function of the voters’ answers to a long list of questions about social, political and economic issues, $X_1, \ldots, X_p$.
- An industrial psychologist wishes to predict the prevalence of workers’ absenteeism in a large organization, Y, based on a subset of items $X_1, \ldots, X_p$ in a large personality inventory.

In these (and in many similar) examples, standard model selection tools can be used to identify a subset of predictors that is best, in the sense that it optimizes the researcher’s criterion of choice (R^2, C_p, AIC, Bayesian information criterion, etc.) in the calibration sample. However, it is unlikely that any researcher would consider the selected subset as a realization of the true model. In particular, the researcher would not be surprised (a) if another subset of predictors performed almost, or exactly, as well as the one chosen, (b) if certain predictors are included in, or absent from, the selected subset, or (c) if a different random sample from the same population led to choosing a different subset of predictors, using the same methodology. In fact, in these cases, instead of discussing ‘the single true model’ it is more sensible to concede that there are ‘multiple reasonable models’ and try to estimate the probability that each of the candidates is the ‘best-fitting model’ (BFM). More specifically, assume that, given p predictors, one can specify $2^p - 1$ distinct models.[1] We are seeking to estimate $P^{(c)}(M_j|\mathbf{Y})$, the probability that model M_j ($j = 1, \ldots, 2^p - 1$) fits the data ($\mathbf{Y}$) best according to a criterion c.

[1] For the sake of simplicity we do not consider models involving polynomials or interactive terms. However, there is nothing in the nature of the proposed procedure that excludes this.

Draper and Guttman (1987) describe a Bayesian approach to the problem of assigning a probability to each of the subset models, by which prior probabilities are placed on the regression coefficients (β_i) and the error variance (σ^2) for all 2^p subset models (including the intercept-only model). The resulting posterior probabilities, $P(M_j|\mathbf{Y})$, ($j = 1, 2, \ldots, 2^p$), can then be determined, and the 'best' (i.e., most probable) subset model can be identified. The idea of generating a probability distribution across all subset models is appealing; however, specification of prior probabilities is always controversial and usually restricts the conclusions in other ways as well. For example, Draper and Guttman's results are not invariant under scale changes in either dependent or independent variables. The alternative approach to this problem, which we describe in this paper, is to estimate the empirical probability distribution of the subset models by a non-parametric method (the bootstrap), and use it to estimate both model probabilities and predictor criticalities.

2. Predictor criticality

The realization that there is no single true model suggests a new possible and important dimension along which predictors can be ranked, namely their criticality. Loosely speaking, a predictor, X_i, ($i = 1, \ldots, p$), is said to be critical as a function of its likelihood of being included in the BFM.

Consider, for example, random samples of n observations on one criterion variable (Y) and three predictors (X_1, X_2 and X_3). Within each sample the best subset model can be identified according to a criterion, c. If a large number of different random samples of size n were available, and in each of these samples the best subset model included X_1, then X_1 would be considered a highly critical predictor variable. This does not necessarily imply that Y could not be predicted accurately in the absence of X_1—this would be the claim made by a measure of predictor importance—but it is critical in the sense that in its absence the best subset for the prediction of Y would not have been identified.

More precisely, predictor criticality is defined as the probability that a predictor is included in the best subset model for a given population. In contrast to the assignment of importance ranks, which is based on the contribution each variable makes to predicting the criterion in a given model, the assignment of criticality ranks is based on predictability across multiple models. A highly critical variable is one which is necessary for the identification of the best model, where 'best' refers to a subset model that is chosen with high probability according to some well-defined goodness-of-fit criterion, c.

The four steps in the determination of predictor criticality are: (1) bootstrap (resample with replacement) the original data set in order to obtain a large number, B, of quasi-replications; (2) for each of the B data sets select the best-fitting subset model according to some criterion, c; (3) obtain the relative frequencies (out of B) with which each subset model was selected (the empirical probability distribution of the subset models); and (4) derive measures of criticality which represent the probability of inclusion of each predictor in the empirical probability distribution of the subset models. In this section we discuss these steps in detail.

2.1. *Data collection: bootstrapping*

The original data set is a random sample of size n from a certain population. It is of interest to examine as many random samples of size n, from the same population, as possible. Since collecting a large number of random samples is not feasible, the bootstrapping procedure can be used to resample the original observations, with replacement, thereby producing samples of size n that can be considered to have come from the same population as the original observations. Efron (1979) originally proposed the bootstrap as a technique for non-parametric estimation of the sampling distribution of a data-dependent random variable. Estimation proceeds by randomly sampling the observations of the data sample, with replacement, computing the value of the target random variable using each of the resamples, and thereby producing its empirical sampling distribution. Since the best-fitting model can be considered a random variable, the bootstrapping procedure is employed to estimate its sampling distribution.

There are two general resampling methods, which correspond to the two types of regression models. Case resampling corresponds to the random (or correlation) model, and residual resampling corresponds to the fixed (or classic regression) model (Efron, 1979; Freedman, 1981; Mooney & Duval, 1993; Shao, 1996; Stein, 1996).

Case resampling assumes that the predictors are random variables and can take on any value within their distributions. Each case or observation is a $(p + 1)$-dimensional vector, consisting of a criterion and its corresponding p predictors. The method proceeds by taking a random sample of n observations, with replacement, where each observation consists of the original criterion–predictors vector. Therefore, if $\mathbf{D}$ is the $n \times (p + 2)$ data matrix, $\mathbf{D} = [\mathbf{Y}|\mathbf{X}]$, each resampled data matrix $\mathbf{D}_i^*$, or each replication i $(i = 1, 2, \ldots, B)$, consists of n rows of $\mathbf{D}$ chosen randomly with replacement. In other words, on the ith replication n rows of $\mathbf{D}$ are randomly selected with replacement, and this constitutes the ith data resample, $\mathbf{D}_i^*$.

Residual resampling assumes that the predictor variables can take on only (experimentally) fixed values. This is the assumption typically made in the classic regression model. Under this assumption, the predictors themselves need not be resampled (because they are fixed). The residuals, on the other hand, are random variables. Residual resampling proceeds by fitting the full regression model to the original data set, and obtaining the n residuals. Then, n of the residuals are randomly chosen, with replacement, and they are added to the n expected criterion values (i.e., the predictor values do not change). Thus, the data resample consists of the original predictor values and the residual-modified criterion values. If $\mathbf{D}$ is again the $n \times (p + 2)$ data matrix, $\mathbf{D} = [\mathbf{Y}|\mathbf{X}]$, let

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta},$$

and let

$$\mathbf{res} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

Let $\mathbf{res}^*$ be a vector of n residuals randomly chosen, with replacement, from $\mathbf{res}$, and let $\mathbf{Y}^* = \hat{\mathbf{Y}} + \mathbf{res}^*$. Then the resampled data matrix on replication i $(i = 1, 2, \ldots, B)$ is the $n \times (p + 2)$ matrix $\mathbf{D}_i^* = [\mathbf{Y}^*|\mathbf{X}]$.

In most data sets collected in the social sciences the majority of the predictors are random, so residual resampling is often inappropriate. In addition, if a regression model is inappropriate for some other reason (e.g., a linear model is used where a polynomial model is appropriate), using the residuals from the misspecified regression analysis would be highly problematic. These types of problems are less likely to arise with case resampling, since the relationship between $\mathbf{Y}$ and its predictors does not need to be specified in order for case resampling to proceed.

2.2. *Model selection*

In each of the B bootstrapped samples, all $2^p - 1$ subset regression models are fitted to the data. For each of the resamples, $\mathbf{D}_i^*$ $(i = 1, 2, \ldots, B)$, one of these $2^p - 1$ regression models can be chosen as the BFM according to some goodness-of-fit criterion. In principle, the method can be applied with any arbitrary choice of criterion. For illustration purposes, three distinct criteria were chosen here: adjusted R^2, AIC and Mallow's C_p. For general reviews of various model selection criteria, see Hocking (1976) and Weisberg (1985).

2.3. *Probability distribution of BFMs*

The number of times each of the $2^p - 1$ possible models was selected as the BFM according to the selection criterion, c, defines the empirical sampling distribution of the BFM. This distribution is a non-parametric alternative to $P^{(c)}(M_j|\mathbf{Y})$ generated by the Draper and Guttman (1987) method, and one could use it as a model selection procedure (i.e., pick the most probable BFM). However, we are interested in an analysis of the predictors rather than the models. The criticality of each predictor is a function of the membership of X_i $(i = 1, \ldots, p)$ in each BFM as well as the frequency of each BFM in the empirical sampling distribution.

2.4. *Predictors' criticality*

Consider all $(2^p - 1)$ subset models of p predictors. Let F_j be the frequency with which the jth model was identified as a BFM in the B bootstrapping runs, and P_j be the empirical probability of each BFM $(\sum_j F_j = B)$. The subset models represent the $2^p - 1$ outcomes of a multinomial distribution with their associated probabilities, P_j, where $\sum_j P_j = 1$ $(j = 1, 2, \ldots, 2^p - 1)$.

To quantify the criticality of each predictor, X_i $(i = 1, 2, \ldots, p)$, we define the measure C_i as the linear combination

$$C_i = \sum_j a_{ij} P_j, \quad (i = 1, 2, \ldots, p, j = 1, 2, \ldots, 2^p - 1),$$

where

$$a_{ij} = \begin{cases} 1 & \text{if } X_i \text{ is in model } j, \\ 0 & \text{otherwise.} \end{cases}$$

Thus C_i represents the expected value (probability of inclusion in a BFM) of X_i. As C_i is the probability of membership in the BFM, a (minimum) value of $C_i = 0$ indicates that X_i is never included in a BFM and a (maximum) value of $C_i = 1$ indicates that X_i is included in all BFMs. The value of C_i represents the expected probability of model misspecification, or misidentification, when X_i is excluded from the analysis. Thus C_i rank-orders the predictors according to their likelihood of being included in the best model over B bootstrapped samples, and represents the criticality of the ith predictor. Furthermore, $\sum_i C_i$ represents the average number of predictors in the BFMs, or the expected size of the BFM. Table 1 describes three hypothetical examples in which $p = 3$. The top panel in the table describes the three hypothetical distributions over all possible models, and the second panel tabulates the criticality measures. The other panels will be discussed later in the paper.

Note that the average number of predictors in the BFM is about 2 for each example.

Table 1. Three hypothetical examples (with $p = 3$) of the BFM probability distributions and their associated criticality values

		P_j		
BFM	k	Example 1	Example 2	Example 3
X_1	1	0	0.3	0.3
X_2	1	0	0	0
X_3	1	0	0	0
X_1X_2	2	0	0.2	0
X_1X_3	2	0	0.2	0
X_2X_3	2	1.0	0	0.3
$X_1X_2X_3$	3	0	0.3	0.4

	C_i		
X_i	Example 1	Example 2	Example 3
X_1	0	1.0	0.7
X_2	1.0	0.5	0.7
X_3	1.0	0.5	0.7
Total	2.0	2.0	2.1

	wC_i		
X_i	Example 1	Example 2	Example 3
X_1	0	0.6	0.433
X_2	0.5	0.2	0.283
X_3	0.5	0.2	0.283
Total	1.0	1.0	1.0

Example 1:					Example 2:					Example 3:				
	X_1	X_2	X_3	dC_i		X_1	X_2	X_3	dC_i		X_1	X_2	X_3	dC_i
X_1	·	−1	−1	−1	X_1	·	1	1	0.5	X_1	·	0	0	0
X_2	1	·	0	0.5	X_2	−1	·	0	−0.5	X_2	0	·	0	0
X_3	1	0	·	0.5	X_3	−1	0	·	−0.5	X_3	0	0	·	0

Example 1 represents the extreme case in which the same model was chosen in all B runs. This model contains both X_2 and X_3, and these two predictors should therefore be given equal, and the highest possible, criticality values. This is indeed the case, as $C_1 = 0$, $C_2 = 1.0$, and $C_3 = 1.0$. In example 2, X_1 is included in all of the models that were chosen, while X_2 and X_3 are only included in a subset of these models. Therefore, we should expect that X_1 will receive the highest possible criticality value, which is indeed the case ($C_1 = 1.0$, $C_2 = 0.5$, and $C_3 = 0.5$). Example 3 is more interesting, in that it results in equally critical predictors ($C_1 = C_2 = C_3 = 0.7$).

3. Distribution theory for the criticality measures

Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{2^p-1})'$ be a vector of the multinomial class probabilities associated with each of the subset models in the population. Let the jth element of the vector $\mathbf{p} = \{p_j\}$,

$j = 1, 2, \ldots, 2^p - 1$, be the maximum likelihood estimate of π_j (the jth element of $\boldsymbol{\pi}$), based on the B bootstrapped samples.

For any linear combination of the probabilities, $\hat{\mathbf{L}} = \mathbf{A} \times \mathbf{p}$, where $\mathbf{A}$ is a matrix of constants, and rank $(\mathbf{A}) = t < 2^p - 1$, the limiting distribution of $\hat{\mathbf{L}}$ is the t-dimensional normal distribution, with mean $E(\hat{\mathbf{L}}) = \mathbf{L} = \mathbf{A} \times \mathbf{p}$ and covariance matrix $\boldsymbol{\Sigma}_{\hat{\mathbf{L}}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{p}}\mathbf{A}'$ (Chernoff, 1956). The elements $\{\sigma_{ij}\}$ of the matrix $\boldsymbol{\Sigma}_{\mathbf{p}}$ are given by (see, for example, Lindeman, Merenda & Gold, 1980):

$$\sigma_{ij} = \begin{cases} \pi_i(1 - \pi_i)/B, & i = j, \\ (-\pi_i \pi_j)/B, & i \neq j. \end{cases}$$

Thus $\boldsymbol{\Sigma}_{\mathbf{p}}$ can be estimated by $\mathbf{S}_{\mathbf{p}} = \{s_{ij}\}$, where

$$s_{ij} = \begin{cases} p_i(1 - p_i)/B, & i = j, \\ (-p_i p_j)/B, & i \neq j. \end{cases}$$

and $\boldsymbol{\Sigma}_{\hat{\mathbf{L}}}$ can be estimated by $\mathbf{S}_{\hat{\mathbf{L}}} = \mathbf{A}\mathbf{S}_{\mathbf{p}}\mathbf{A}'$. When B is large, the quadratic form $\mathbf{Q}(\hat{\mathbf{L}}) = (\hat{\mathbf{L}} - \mathbf{L})'\boldsymbol{\Sigma}_{\hat{\mathbf{L}}}^{-1}\ (\hat{\mathbf{L}} - \mathbf{L})$ has a χ^2 distribution with t degrees of freedom (Chernoff, 1956), and can be estimated by $\hat{\mathbf{Q}}(\hat{\mathbf{L}}) = (\hat{\mathbf{L}} - \mathbf{L})'\mathbf{S}_{\hat{\mathbf{L}}}^{-1}\ (\hat{\mathbf{L}} - \mathbf{L})$, which has the same limiting distribution (Wald, 1943).

The criticality measures are linear combinations of the multinomial distribution of the BFMs, and these results can be used to perform statistical tests involving variable criticality. For instance, L can be a criticality 'parameter', θ_i, which is defined as a linear combination of multinomial probabilities

$$L = \theta_i = \mathbf{a}_i \times \boldsymbol{\pi},$$

estimated by

$$\hat{L} = C_i = \mathbf{a}_i \times \mathbf{p},$$

where $\mathbf{a}_i$ is a $1 \times (2^p - 1)$ row vector of constants and $\mathbf{p}$ is a $(2^p - 1) \times 1$ vector of the probabilities from the BFM distribution. Therefore, asymptotic confidence intervals for the θ_i can be constructed using the C_i. Alternatively, L can be defined as a linear combination of the criticality parameters (θ_i), such that

$$L = \mathbf{b} \times \boldsymbol{\theta} = \mathbf{D} \times \boldsymbol{\pi},$$

where $\mathbf{b}$ is a $1 \times p$ row vector of p constants, $\boldsymbol{\theta}$ is a $p \times 1$ vector containing the θ_i, $\mathbf{D} = \mathbf{b} \times \mathbf{A}$ where $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p)'$, and the $\mathbf{a}_i$ are $1 \times (2^p - 1)$ row vectors of constants. Then L can be estimated by $\hat{L} = \mathbf{b} \times \mathbf{C} = \mathbf{D} \times \mathbf{p}$ (where $\mathbf{C}$ is a $p \times 1$ vector containing the C_i), and the statistical significance of the differences between C_i (defined by the vector $\mathbf{b}$) can be tested using the fact that $\hat{\mathbf{Q}}(\hat{\mathbf{L}}) = (\hat{\mathbf{L}} - \mathbf{L})'\mathbf{S}_{\mathbf{L}}^{-1}\ (\hat{\mathbf{L}} - \mathbf{L})$ is χ^2.

4. Examples

In this section we illustrate the procedure with artificial (simulated) and real data sets. In all cases we used data sets of moderate size ($10 \leq n \leq 100$) and with a small number of potential predictors ($p \leq 5$). Criticality analysis can be applied with any criterion for choosing BFMs, and, for illustration purposes, we employed AIC, R^2_{adj} and C_p. These three measures were chosen because they are commonly used in the social sciences and each of them includes a penalty for the number of predictors (k) included in the model (j). We first describe the three criteria and their use in this context.

The R^2_{adj} for subset model j is written as

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}_j/(n-k-1)}{\text{SSTO}/(n-1)} = 1 - \frac{n-1}{n-k-1}(1 - R^2_j)$$

(Neter, Wasserman, & Kutner, 1990), where SSE_j is the error sum of squares of model j, SSTO is the total sum of squares of the full model, and R^2_j is the squared multiple correlation of (i.e., proportion of variance of Y reproduced by) model j. The larger the value of R^2_{adj} the better the model fits the data. Thus, the BFM is the subset model with the highest R^2_{adj}.

The AIC value (Akaike, 1973) for subset model j is written as

$$\text{AIC} = n \ln\left(\frac{\text{SSE}_j}{n}\right) + 2(k+1).$$

Since n is fixed, the quantity $n \ln(\frac{\text{SSEj}}{n})$ is smallest for the model with the smallest error sum of squares (SSE_j). Thus, the BFM is the subset model with the lowest AIC.

The C_p value (Mallows, 1973) for subset model j is written as

$$C_p = \frac{\text{SSE}_j}{\text{MSE}} + 2(k+1) - n,$$

where MSE is the mean squared error for the full model. C_p can also be rewritten as

$$C_p = (p-k)(F_j - 1) + (k+1),$$

where F_j is the F-statistic for testing the hypothesis that the predictors left out of the jth subset model, but included in the full model, all have zero coefficients (Weisberg, 1985). In the full model ($k = p$), $C_p = k + 1 = p + 1$. Moreover, if the predictors left out of the jth subset model do indeed have zero coefficients, the expected value of F_j will be approximately 1. Thus, good models should have C_p values that are approximately equal to $k + 1$, and $C_p \leq p + 1$ (corresponding to $F_j \leq 2$) (Weisberg, 1985).[2] Therefore, the model with the smallest positive value of the difference $k + 1 - C_p$ is identified as the BFM according to the C_p criterion. If none of the models produces a positive value for $k + 1 - C_p$, then the model with $k + 1 - C_p = 0$ (i.e., the full model) is considered to be the BFM.

4.1. Simulations

To examine the effect of the size of the correlations among the predictors (i.e., the degree of multicolinearity) on the criticality measures, several simulations were conducted. These simulations consist of:

1. specifying a certain pattern of correlations between the predictors and the criterion;
2. generating a multivariate normal data set based on the specified correlation pattern;
3. determining predictor criticalities using B resamples of the data set; and
4. performing R repetitions of steps 2 and 3 to eliminate any bias that might exist within any one data set.

[2] If $C_p > k + 1$ then $F_j > 1$, and the difference between $k + 1$ and C_p is due to bias; if $C_p < k + 1$ then $F_j < 1$, and the difference between $k + 1$ and C_p is due to random error (Neter *et al.*, 1990).

For each simulation, the mean criticality and its standard error over R runs were recorded under each model selection method (maximum R^2_{adj}, minimum AIC and minimum positive $k + 1 - C_p$).

4.1.1. Simulation set A

This set consists of four simulations in which the correlations between Y and the predictors are kept constant, while the (equal) correlation among all six pairs of predictors is systematically varied. Specifically, for each of four predictors ($p = 4$), the correlation between Y and the ith predictor (X_i) was specified to be $\rho(Y, X_i) = 0.1 + 0.2(i - 1)$ in all four simulations, while the common correlation among the predictors was specified to be $\rho(X_i, X_j) = 0.75, 0.5, 0.25$ or 0.0 (for all $i \neq j = 1, \ldots, p$) for simulation A1, A2, A3 and A4, respectively. The four correlation matrices are shown in the left panels of Table 2.

Importance measures for each simulation, based on the population correlations, are presented in Table 2. They include squared correlations, standardized regression coefficients and squared partial and semi-partial correlations. The mean criticalities and their standard errors across $R = 40$ repetitions, based on $n = 50$ and $B = 500$, are presented in Table 3 under the three model selection methods. The distributions of the BFMs, for each simulation and each selection method, are presented in Table 4. (The models are listed according to the frequency with which they were identified as BFMs by the specific selection criterion). Note that the ranking of the predictors based on criticality is similar, though not identical, to the ranking based on (most of) the population importance measures.

This set of simulations highlights the difference between the interpretation of criticality and importance. For example, in simulation A1, all predictors are equally and maximally critical (although this is not the case using the population importance measures). The reason for this is that the correlations among the predictors are quite high and the full model is identified as best fitting in each and every case. Thus, removing any of the predictors would result in misidentification of the best-fitting model (even though not all $p = 4$ predictors are equally correlated with Y).

The effects of the inter-predictor correlations, $\rho(X_i, X_j)$, on the results are evident, though not always easy to characterize. Note, first, that the size of the selected BFM (as measured by ΣC_i) is smallest for low values of $\rho(X_i, X_j)$ and that the two predictors that correlate highly with Y (X_4 and X_3) are assigned almost equal (and invariably high) criticalities by all selection methods in all cases. However, there are wide differences, and order reversals, between the criticalities of X_1 and X_2 across the various conditions. It is instructive to consider first the two extreme cases. Simulation A1 is an extreme case of multicollinearity (very close to singularity) where all methods select the full model. Since in such cases the predictors are almost interchangeable, they are identified as equally critical. In the case of uncorrelated predictors (simulation A4) the four criticalities are monotonically related to the predictors' correlation with the criterion. As $\rho(X_i, X_j)$ increases, the criticality of X_2 (a predictor that has a 0.3 correlation with Y) decreases steadily, while the criticality of X_1 (a predictor that has a 0.1 correlation with Y) increases at an even faster rate. In fact, all selection methods rate X_1 as more critical than X_2 when $\rho(X_i, X_j) \geq 0.25$. This seems to indicate that when a predictor is only slightly correlated with Y, a high correlation with other predictors will increase its criticality. This was explored to some extent in simulation set B.

Table 2. Simulations A1–A4: Importance measures in the population

		Population Correlation matrix						Importance measures (population)			
Simulation		Y	X_1	X_2	X_3	X_4	R^2 of full model	$\rho^2(y, x_i)$	β_i	Partial ρ^2	Semi-partial ρ^2
A1	Y	1					0.9969				
	X1	0.1	1					0.01	−1.0769	0.9919	0.3769
	X2	0.3	0.75	1				0.09	−0.2769	0.8901	0.0249
	X3	0.5	0.75	0.75	1			0.25	0.5231	0.9666	0.0889
	X4	0.7	0.75	0.75	0.75	1		0.49	1.3231	0.9946	0.5689
A2	Y	1					0.6560				
	X1	0.1	1					0.01	−0.4400	0.2602	0.1210
	X2	0.3	0.5	1				0.09	−0.0400	0.0029	0.0010
	X3	0.5	0.5	0.5	1			0.25	0.3600	0.1906	0.0810
	X4	0.7	0.5	0.5	0.5	1		0.49	0.7600	0.5121	0.3610
A3	Y	1					0.6324				
	X1	0.1	1					0.01	−0.1714	0.0654	0.0257
	X2	0.3	0.25	1				0.09	0.0952	0.0211	0.0079
	X3	0.5	0.25	0.25	1			0.25	0.3619	0.2377	0.1146
	X4	0.7	0.25	0.25	0.25	1		0.49	0.6286	0.4846	0.3457
A4	Y	1					0.8400				
	X1	0.1	1					0.01	0.1000	0.0588	0.0100
	X2	0.3	0	1				0.09	0.3000	0.3600	0.0900
	X3	0.5	0	0	1			0.25	0.5000	0.6098	0.2500
	X4	0.7	0	0	0	1		0.49	0.7000	0.7538	0.4900

Table 3. Simulations A1–A4: Criticalities(C_i) and their standard errors over $R = 40$ replications

			C_i (SE of C_i) by criterion $n = 50, B = 500, R = 40$		
Simulation	$\rho(X_i, X_j)$	Predictor	AIC	Adjusted R^2	C_p
A1	0.75	X_1	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		X_2	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		X_3	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		X_4	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		Sum	4.00	4.00	4.00
A2	0.5	X_1	0.965 (0.004)	0.978 (0.003)	0.977 (0.003)
		X_2	0.359 (0.019)	0.450 (0.020)	0.450 (0.010)
		X_3	0.943 (0.007)	0.967 (0.004)	0.964 (0.005)
		X_4	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		Sum	3.27	3.40	3.39
A3	0.25	X_1	0.627 (0.016)	0.729 (0.015)	0.758 (0.015)
		X_2	0.525 (0.019)	0.644 (0.018)	0.668 (0.018)
		X_3	0.963 (0.005)	0.981 (0.003)	0.979 (0.004)
		X_4	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)
		Sum	3.11	3.35	3.40
A4	0.0	X_1	0.622 (0.019)	0.730 (0.017)	0.731 (0.017)
		X_2	0.989 (0.001)	0.994 (0.001)	0.994 (0.001)
		X_3	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		X_4	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		Sum	3.61	3.72	3.73

4.1.2. Simulation set B

This set consists of five simulations, each involving three predictors ($p = 3$). The correlation matrices were specified such that the correlation between Y and the predictors was kept constant across the five simulations, with $\rho(Y, X_1) = 0.6$ and $\rho(Y, X_2) = \rho(Y, X_3) = 0$. Two predictors, X_2 and X_3, were also uncorrelated in all cases; i.e., $\rho(X_2, X_3) = 0$. For the matrix to be positive semi-definite its elements are constrained by the inequality $-[1 - \rho(Y, X_1)^2], \leq [\rho(X_1, X_2)^2 + \rho(X_1, X_3)^2] \leq [1 - \rho(Y, X_1)^2]$. Thus in three of the simulations $\rho(X_1, X_2) = \rho(X_1, X_3) = 0$, $\sqrt{0.1}$, and $\sqrt{0.2}$ (these are simulations B1, B3, and B5, respectively). In the other two simulations $\rho(X_1, X_2) = 0$ while $\rho(X_1, X_3) = \sqrt{0.2}$ or $\sqrt{0.4}$ (these are simulations B2 and B4, respectively). The five correlation matrices, along with the population importance measures, are presented in Table 5.

The goal of these simulations was to examine how the systematic variation in $\rho(X_1, X_2)$ and $\rho(X_1, X_3)$ affects the criticality of X_2 and X_3. The mean criticalities and their standard errors for $n = 50$, $B = 500$ and $R = 40$ are presented in Table 6, and the BFM distributions obtained for each simulation are presented in Table 7 (for each selection criterion, the models are listed in descending order of observed frequency).

Not surprisingly, the results (in Table 6) show that X_1 is always highly critical (in this case, it is always essential). Our primary interest is in the criticalities of X_2 and X_3, which

Table 4. Simulations A1–A4: The BFM frequency distribution for each simulation, based on 20000 samples ($B = 500$ bootstraps$\times R = 40$ replications)

	AIC		Adjusted R^2		C_p	
Simulation	Predictors in the BFM	f(BFM)	Predictors in the BFM	f(BFM)	Predictors in the BFM	f(BFM)
A1	1234	20000	1234	20000	1234	20000
Total:		20000		20000		20000
A2	134	12082	134	9902	134	9829
	1234	6194	1234	9017	1234	9017
	14	528	123	385	124	524
	124	491	234	365	234	370
	234	437	14	244	14	157
	34	138	34	52	34	66
	4	65	24	25	24	22
	24	61	4	8	4	12
	123	3	123	2	123	2
	13	1			13	1
Total:		20000		20000		20000
A3	1234	6424	1234	9217	1234	9217
	134	5767	134	5136	134	5640
	34	3565	234	3341	234	3776
	234	3496	34	1930	34	946
	24	348	24	156	124	255
	124	213	124	155	24	101
	14	135	14	60	14	48
	4	40	4	2	4	14
	23	8	123	2	23	2
	123	3	23	1	123	1
	2	1				
Total:		20000		20000		20000
A4	1234	12315	1234	14522	1234	14522
	234	7472	234	5361	234	5357
	134	117	134	85	134	97
	34	94	34	32	34	24
	4	1				
	24	1				
Total:		20000		20000		20000

Note: Models are ordered according to their frequency.

are uncorrelated with Y. The results indicate that there is a clear monotonic relationship between the predictors' correlations with X_1 and their respective criticality measures. Consider, for example, C_3 (the criticality of X_3), under the AIC selection criterion. As $\rho(X_1, X_3)$ increases from 0 (simulation B1) to $\sqrt{0.1}$ (B3), to $\sqrt{0.2}$ (B2 and B5), and to $\sqrt{0.4}$ (simulation B4), C_3 increases gradually from 0.348 to 0.999. A similar trend can be observed for the other selection criteria and for C_2, as $\rho(X_1, X_2)$ increases from 0 (simulations B1, B2 and B4) to $\sqrt{0.1}$ (B3), and to $\sqrt{0.2}$ (B5). In fact, in B5, where the correlations of both X_2 and X_3 with X_1 are maximal (under the constraint of this example), the full model is

Table 5. Simulations B1–B5: Importance measures in the population

	Population correlation matrix						Importance measures (population)			
Simulation		Y	X_1	X_2	X_3	R^2 of full model	$\rho^2(y,x_i)$	β_i	Partial ρ^2	Semi-partial ρ^2
B1	Y	1				0.3600				
	X_1	0.6	1				0.36	0.6000	0.3600	0.3600
	X_2	0.0	0.0	1			0.00	0.0000	0.0000	0.0000
	X_3	0.0	0.0	0.0	1		0.00	0.0000	0.0000	0.0000
B2	Y	1				0.4500				
	X_1	0.6	1				0.36	0.7500	0.4500	0.4500
	X_2	0.0	0.0	1			0.00	0.0000	0.0000	0.0000
	X_3	0.0	$\sqrt{0.2}$	0.0	1		0.00	–0.3354	0.1406	0.0900
B3	Y	1				0.4500				
	X_1	0.6	1				0.36	0.7500	0.4500	0.4500
	X_2	0.0	$\sqrt{0.1}$	1			0.00	–0.2372	0.0833	0.0500
	X_3	0.0	$\sqrt{0.1}$	0.0	1		0.00	–0.2372	0.0833	0.0500
B4	Y	1				0.6000				
	X_1	0.6	1				0.36	1.0000	0.6000	0.6000
	X_2	0.0	0.0	1			0.00	0.0000	0.0000	0.0000
	X_3	0.0	$\sqrt{0.4}$	0.0	1		0.00	–0.6325	0.3750	0.2400
B5	Y	1				0.6000				
	X_1	0.6	1				0.36	1.0000	0.6000	0.6000
	X_2	0.0	$\sqrt{0.2}$	1			0.00	–0.4472	0.2727	0.1500
	X_3	0.0	$\sqrt{0.2}$	0.0	1		0.00	–0.4472	0.2727	0.1500

Table 6. Simulations B1–B5: Criticalities (C_i) and their standard errors over 40 replications

				C_i (SE of C_i) by criterion $n = 50$, $B = 500$, $R = 40$		
Simulation	$\rho(X_1, X_2)$	$\rho(X_1, X_3)$	Predictor	AIC	Adjusted R^2	C_p
B1	0	0	X_1	0.999 (0.000)	1.000 (0.000)	0.999 (0.000)
			X_2	0.366 (0.020)	0.512 (0.021)	0.582 (0.021)
			X_3	0.348 (0.020)	0.496 (0.021)	0.569 (0.021)
			Sum	1.71	2.01	2.15
B2	0	$\sqrt{0.2}$	X_1	0.999 (0.000)	1.000 (0.000)	0.999 (0.000)
			X_2	0.356 (0.019)	0.498 (0.020)	0.520 (0.021)
			X_3	0.857 (0.011)	0.909 (0.009)	0.911 (0.009)
			Sum	2.21	2.41	2.43
B3	$\sqrt{0.1}$	$\sqrt{0.1}$	X_1	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
			X_2	0.642 (0.017)	0.749 (0.015)	0.764 (0.015)
			X_3	0.728 (0.017)	0.824 (0.014)	0.830 (0.014)
			Sum	2.37	2.57	2.59
B4	0	$\sqrt{0.4}$	X_1	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
			X_2	0.311 (0.020)	0.461 (0.021)	0.462 (0.021)
			X_3	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)
			Sum	2.31	2.46	2.46
B5	$\sqrt{0.2}$	$\sqrt{0.2}$	X_1	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
			X_2	0.966 (0.004)	0.982 (0.003)	0.983 (0.003)
			X_3	0.970 (0.004)	0.985 (0.002)	0.985 (0.002)
			Sum	2.94	2.97	2.97

selected practically always, and the three variables are almost equally critical (just as in simulation A1).[3]

In all simulations there is a high level of agreement between the criticality measures based on the three distinct model selection criteria. However, the measures are not identical. In particular, note that AIC favours smaller models.

4.2. Real data

Suh, Diener, Oishi and Triandis (1998) collected the responses of over 7000 people in 41 countries on numerous variables pertaining to satisfaction with life.[4] For this analysis, we selected a subset of variables and averaged their values by country; therefore, we used 41 (averaged) observations on five predictors and one criterion. The criterion, Y, is the score on a ‘satisfaction with life’ scale consisting of five items. The five predictors are responses to

[3] The results of simulation B1 show that variables X_2 and X_3 are critical about 30% of the time, even though they are uncorrelated with the criterion or each other. This is due to sampling error, which exists in our simulations as well as in real data sets. Note that we do not necessarily recommend the use of the criticality analysis for model selection; rather, we propose its use for ranking predictors once a reasonable choice for a model has been made. In the case of simulation B1, for example, variable X_1 is clearly the most critical variable as its criticality value is twice that of X_2 and X_3. Thus one can safely say that X_1 is the most critical of the three available predictors.

[4] We thank these authors for sharing their data with us.

Table 7. Simulations B1–B5: The BFM frequency distribution for each simulation, based on 20 000 samples ($B = 500$ bootstraps $\times R = 40$ replications)

	AIC		Adjusted R^2		C_p	
Simulation	Predictors in the BFM	f(BFM)	Predictors in the BFM	f(BFM)	Predictors in the BFM	f(BFM)
B1	1	8378	123	5120	12	6517
	12	4653	12	5110	13	6240
	13	4283	1	4974	123	5121
	123	2665	13	4787	1	2105
	3	12	3	5	3	10
	2	6	2	2	23	6
	23	3	23	2	2	1
Total:		20000		20000		20000
B2	13	10771	123	9212	123	9214
	123	6366	13	8964	13	8997
	1	2113	1	1080	12	1170
	12	739	12	737	1	604
	2	7	2	5	2	6
	3	4	3	1	23	6
			23	1	3	3
Total:		20000		20000		20000
B3	123	9786	123	12515	123	12515
	13	4773	13	3967	13	4067
	12	3042	12	2457	12	2758
	1	2394	1	1058	1	646
	2	4	2	2	23	3
	3	1	3	1	2	2
Total:		20000		20000		20000
B4	13	13760	13	10769	13	10767
	123	6223	123	9226	123	9227
	1	11	1	4	1	3
	12	6	12	1	12	3
Total:		20000		20000		20000
B5	123	18851	123	19381	123	19381
	13	549	13	312	13	311
	12	463	12	267	12	281
	1	137	1	40	1	27
Total:		20000		20000		20000

Note: Models are ordered according to their frequency.

single items (different from the global scale) pertaining to domain-specific reported levels of satisfaction. The domains are X_1 = health, X_2 = financial situation, X_3 = family, X_4 = housing and X_5 = self.

Standard results from the multiple regression analysis of the original data set are presented in Table 8. They include, for each predictor, zero-order correlations with Y, standardized regression coefficients, partial and semi-partial correlations with Y as well as the averaged results of the dominance analysis (Budescu, 1993). These results all pertain to

Table 8. Multiple regression analysis results for the satisfaction with life data set

Correlations between the predictors

	X_1	X_2	X_3	X_4	X_5
X_1	1.0000				
X_2	0.418	1.000			
X_3	0.556	0.216	1.000		
X_4	0.602	0.362	0.381	1.000	
X_5	0.606	0.345	0.258	0.396	1.000

Predictor importance measures

	X_1	X_2	X_3	X_4	X_5
$r(y, x_i)$	0.4832	0.4436	0.2578	0.7279	0.5291
$r^2(y, x_i)$	0.2335	0.1968	0.0665	0.5298	0.2800
Standardized β_i	−0.1028	0.1522	−0.2430	0.6556	0.4337
$r^2(y, x_i{\cdot}x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$	0.0135	0.0512	0.0868	0.4406	0.2150
$r^2(y(x_i{\cdot}x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p))$	0.0046	0.0182	0.0321	0.2663	0.0926
Average dominance	0.0714	0.0759	0.0259	0.3473	0.1379

predictor importance, and appear to indicate that in the full model (including all five predictors), X_4 (housing) is the most important predictor of the score on the satisfaction with life scale.

The bootstrapping analysis was performed using the case resampling method since the predictors are random variables. Resampling was performed for $B = 250$, 500 and 1000 to investigate the degree to which the number of resamples affects the results. Within each of these analyses, the BFM was selected according to each of the three criteria described earlier (maximum R^2_{adj}, minimum AIC and minimum positive $k + 1 - C_p$).

To examine the level of agreement between the model selection methods in identifying the BFM, Cohen's κ (see, for example, Fleiss, 1973) was computed as a measure of the degree of agreement beyond chance. For perfect agreement $\kappa = 1$, and for chance agreement $\kappa = 0$. The large-sample approximation for the standard error of κ (Fleiss, Cohen & Everitt, 1969) was also computed.

The BFM frequency distributions, for each value of B and each model selection method, are presented in Table 9, along with pairwise measures of agreement between the various model selection methods. There is a high level of agreement between the models selected by the R^2_{adj} and AIC criteria, but a lower (though significantly larger than 0) agreement with the models identified by the C_p criterion. This pattern is consistent across all three values of B. In addition, for each selection method there is a high level of agreement between the distributions of the BFMs obtained for various numbers of bootstraps (B).

In Table 10 all $2^5 - 1 = 31$ subset models are listed in descending order according to their R^2_{adj} values in the original data set. This table also presents the probabilities (and their corresponding ranks) obtained from the empirical probability distributions (with $B = 250$, 500, 1000) under the maximum R^2_{adj} model selection criterion. At the bottom of the table we present the Pearson and Kendall rank-order correlations between the four columns. The results show a very high level of agreement between the three bootstrapped solutions, but

Table 9. Best-fitting model frequency distributions for the satisfaction with life data, with $B = 250$, 500 or 1000 and according to three model-selection methods

	$B = 250$			$B = 500$			$B = 1000$		
	Selection method			Selection method			Selection method		
Model	AIC	R^2_{adj}	C_p	AIC	R^2_{adj}	C_p	AIC	R^2_{adj}	C_p
X_1	0	0	0	0	0	0	0	0	0
X_2	0	0	0	0	0	0	0	0	0
X_3	0	0	0	0	0	0	0	0	0
X_4	1	0	0	2	0	0	7	0	2
X_5	0	0	0	0	0	0	0	0	0
X_1X_2	0	0	0	0	0	0	0	0	0
X_1X_3	0	0	0	0	0	0	0	0	0
X_1X_4	3	1	2	1	0	0	2	1	0
X_1X_5	0	0	0	0	0	0	1	0	0
X_2X_3	0	0	0	0	0	0	0	0	0
X_2X_4	5	1	2	8	3	4	39	14	3
X_2X_5	0	0	0	1	0	0	0	0	0
X_3X_4	0	0	0	2	2	1	2	1	2
X_3X_5	0	0	0	0	0	0	0	0	0
X_4X_5	11	2	1	17	1	6	40	6	16
$X_1X_2X_3$	0	0	0	0	0	0	0	0	0
$X_1X_2X_4$	10	5	3	16	12	3	37	26	12
$X_1X_2X_5$	0	0	0	0	0	0	1	0	1
$X_1X_3X_4$	1	0	0	0	0	2	0	2	3
$X_1X_3X_5$	0	0	0	0	0	0	1	0	0
$X_1X_4X_5$	8	3	5	22	15	12	49	16	33
$X_2X_3X_4$	1	0	1	2	2	0	4	4	4
$X_2X_3X_5$	0	0	0	0	0	0	1	1	0
$X_2X_4X_5$	12	10	8	43	23	33	60	40	38
$X_3X_4X_5$	68	46	16	137	102	32	248	165	56
$X_1X_2X_3X_4$	2	4	4	0	2	7	7	10	15
$X_1X_2X_3X_5$	0	0	0	0	0	0	1	2	1
$X_1X_2X_4X_5$	9	13	24	26	35	49	46	70	88
$X_1X_3X_4X_5$	26	33	52	54	67	98	107	134	163
$X_2X_3X_4X_5$	71	80	80	124	142	159	254	279	334
$X_1X_2X_3X_4X_5$	22	52	52	45	94	94	93	229	229
Total	250	250	250	500	500	500	1000	1000	1000

	$B = 250$		$B = 500$		$B = 1000$	
Agreement between:	κ	SE(κ)	κ	SE(κ)	κ	SE(κ)
R^2_{adj} and AIC	0.6251	0.027	0.6540	0.019	0.5961	0.013
R^2_{adj} and C_p	0.4950	0.028	0.4394	0.019	0.4892	0.014
C_p and AIC	0.4367	0.024	0.4372	0.017	0.4035	0.012

somewhat less agreement with the measures from the original data set, highlighting again the distinction between criticality and model fit.

The criticality measures (and their 95% confidence limits) under each model selection method and for each value of B are presented in Table 11. The $100(1 - \alpha)\%$ confidence

Table 10. Comparison of the rank of the subset models in the satisfaction with life data set and in the BFM distributions, according to the maximum R^2_{adj} criterion

		Probability (rank) of subset models in BFM distribution according to maximum R^2_{adj} criterion		
Subset model	R^2_{adj} values in original data set	$B = 250$	$B = 500$	$B = 1000$
$X_2X_3X_4X_5$	0.619193	0.320 (1)	0.283 (1)	0.279 (1)
$X_1X_2X_3X_4X_5$	0.613590	0.208 (2)	0.188 (3)	0.229 (2)
$X_3X_4X_5$	0.612835	0.184 (3)	0.204 (2)	0.165 (3)
$X_1X_3X_4X_5$	0.604065	0.132 (4)	0.134 (4)	0.134 (4)
$X_1X_2X_4X_5$	0.588638	0.052 (5)	0.070 (5)	0.070 (5)
$X_2X_4X_5$	0.585974	0.040 (6)	0.046 (6)	0.040 (6)
X_4X_5	0.577600	0.008 (10)	0.002 (13)	0.006 (11)
$X_1X_4X_5$	0.574334	0.012 (9)	0.030 (7)	0.016 (8)
X_2X_4	0.544437	0.004 (11.5)	0.006 (9)	0.014 (9)
$X_2X_3X_4$	0.533743		0.004 (11)	0.004 (12)
$X_1X_2X_4$	0.532148	0.020 (7)	0.024 (8)	0.026 (7)
$X_1X_2X_3X_4$	0.521430	0.016 (8)	0.004 (11)	0.010 (10)
X_4	0.517775			
X_1X_4	0.508452	0.004 (11.5)		0.001 (16)
X_3X_4	0.505550		0.004 (11)	0.001 (16)
$X_1X_3X_4$	0.498011			0.002 (13.5)
$X_1X_2X_3X_5$	0.328409			0.002 (13.5)
$X_1X_2X_5$	0.323998			
X_2X_5	0.323498			
$X_2X_3X_5$	0.316001			0.001 (16)
$X_1X_3X_5$	0.295931			
X_1X_5	0.286021			
X_1X_2	0.267648			
X_5	0.267648			
X_3X_5	0.252804			
$X_1X_2X_3$	0.247910			
X_1	0.213871			
X_1X_3	0.193356			
X_2X_3	0.183503			
X_2	0.176197			
X_3	0.042531			

Correlations between R^2_{adj} values in original data and probabilities from bootstrapping

		Probabilities		
	R^2_{adj} values	$B = 250$	$B = 500$	$B = 1000$
R^2_{adj} values	–	0.5207	0.5478	0.5451
$B = 250$	0.7350	–	0.9921	0.9923
$B = 500$	0.7580	0.8508	–	0.9881
$B = 1000$	0.8013	0.8474	0.8725	–

Note: values above the diagonal are Pearson’s correlations; values below the diagonal are *Kendall’s correlations.*

Table 11. Criticality measures (and their 95% confidence limits) for the satisfaction with life data with $B = 250$, 500 or 1000, according to three model-selection methods

		Model selection method					
		AIC		R^2_{adj}		C_p	
B	Predictor	C_i	95% limits	C_i	95% limits	C_i	95% limits
250	X_1	0.324	(0.266, 0.382)	0.444	(0.382, 0.506)	0.568	(0.507, 0.629)
	X_2	0.528	(0.466, 0.590)	0.660	(0.601, 0.719)	0.696	(0.639, 0.753)
	X_3	0.764	(0.711, 0.817)	0.860	(0.817, 0.903)	0.820	(0.772, 0.868)
	X_4	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)
	X_5	0.908	(0.872, 0.944)	0.956	(0.931, 0.981)	0.952	(0.926, 0.979)
	Total	3.52		3.92		4.04	
500	X_1	0.328	(0.287, 0.369)	0.450	(0.406, 0.494)	0.530	(0.486, 0.574)
	X_2	0.530	(0.486, 0.574)	0.626	(0.584, 0.668)	0.698	(0.658, 0.738)
	X_3	0.728	(0.689, 0.767)	0.822	(0.788, 0.856)	0.786	(0.750, 0.822)
	X_4	0.998	(0.994, 1.002)	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)
	X_5	0.938	(0.917, 0.959)	0.958	(0.940, 0.976)	0.966	(0.950, 0.982)
	Total	3.52		3.86		3.98	
1000	X_1	0.345	(0.316, 0.374)	0.490	(0.459, 0.521)	0.545	(0.514, 0.576)
	X_2	0.543	(0.512, 0.574)	0.675	(0.646, 0.704)	0.725	(0.697, 0.753)
	X_3	0.718	(0.690, 0.746)	0.827	(0.804, 0.850)	0.807	(0.783, 0.831)
	X_4	0.995	(0.991, 0.999)	0.997	(0.994, 1.000)	0.998	(0.995, 1.001)
	X_5	0.902	(0.884, 0.920)	0.942	(0.928, 0.956)	0.959	(0.947, 0.971)
	Total	3.50		3.93		4.03	

interval for each criticality parameter ($L = \theta_i = \mathbf{a}_i \times \mathbf{p}$) was constructed using its estimate, $\hat{L} = C_i = \mathbf{a}_i \times \mathbf{p}$ ($i = 1, 2, 3, 4$), and using the fact that $\mathbf{p}$ (the multinomial probabilities vector), and linear combinations ($\hat{L}$) of $\mathbf{p}$, are asymptotically multivariate normal. Therefore, the 95% confidence interval for the criticality parameter θ_i is

$$C_i \pm 1.96\sqrt{\mathbf{S}_{C_i}} = (\mathbf{a}_i \times \mathbf{p}) \pm 1.96\sqrt{\mathbf{a}_i\, \mathbf{S}_p \mathbf{a}_i'},$$

where $\mathbf{S}_{\hat{L}}$ is as discussed previously. Of course, simultaneous confidence intervals could be obtained by appropriate adjustments of the probability levels.

There are very few (and small) differences between the mean values obtained with different numbers of bootstraps (B), indicating that for practical purposes $B = 250$ resamples are sufficient. Similarly, there is a high level of agreement between the three model selection methods (note that C_p tends to choose the largest, and AIC the smallest, subsets). In all cases, X_4 (satisfaction with housing) is identified as the most critical variable, followed by X_5 (satisfaction with self), and X_1 (satisfaction with health) is always the least critical predictor.

5. Extensions—alternative measures of criticality

In this section we propose a few variations on the main theme discussed so far. In particular, we propose alternative measures of predictor criticality that can be computed from the

BFM distribution derived through bootstrapping and model selection. To motivate these alternative measures, consider again the three hypothetical examples in Table 1.

Note that in the third example X_1 *by itself* is chosen as a BFM with probability greater than zero, but this is not the case for X_2 or X_3, which always appear jointly. One may wish to develop an alternative criticality measure to highlight this distinction. This can be achieved by assigning different values to the coefficient a_{ij} in the computation of C_i; for example, the a_{ij} scores could be inversely proportional to the complexity of (or number of predictors in) the model. The *weighted criticality* measure, wC_i, is defined as

$$wC_i = \sum_j \frac{a_{ij}}{k_j} P_j, \ (i = 1, 2, \dots, p, j = 1, 2, \dots, 2^p - 1),$$

where

$$a_{ij} = \begin{cases} 1 & \text{if } X_i \text{ is in model } j, \\ 0 & \text{otherwise,} \end{cases}$$

and k_j is the number of predictors in the jth subset model. Evidently wC_i overweights (underweights) simple (complex) models. The weights used here are somewhat arbitrary, and were chosen because of their simplicity, but any other scheme that preserves a monotonic relationship between the weights and the number of predictors in the model would also have this property. Weighted criticality ranges between 0 and 1, where a 0 indicates (as in C_i) that X_i is not in any BFM, and a 1 indicates that predictor X_i appears in all BFMs alone (i.e., the model chosen is *always* X_i). The weighted criticality values for the predictors in the three examples are presented in the third panel of Table 1. Note that $\sum_i wC_j = 1$, and so the weighted criticality of a given predictor is in some sense a criticality 'percentage', or the proportion of the maximum criticality that is allocated to each predictor, where the allocation depends on model complexity. However, unlike C_i, wC_i can no longer be interpreted as the expected probability of model misidentification.

Occasionally, one may be interested in comparing the criticality of a pair of predictors without involving the other $p - 2$ predictors in this comparison. To compare the criticality of X_i to that of X_h, we can define $C_{i(h)}$ as the linear combination

$$C_{i(h)} = \sum_j a_{ij} P_j, \quad (h, i = 1, 2, \dots, p, j = 1, 2, \dots, 2^p - 1),$$

where

$$a_{ij} = \begin{cases} 1 & \text{if } X_i \text{ is in model } j \text{ and } X_h \text{ is not in model } j, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly,

$$C_{h(i)} = \sum_j a_{ij} P_j, \quad (h, i = 1, 2, \dots, p, j = 1, 2, \dots, 2^p - 1),$$

where

$$a_{ij} = \begin{cases} 1 & \text{if } X_h \text{ is in model } j \text{ and } X_i \text{ is not in model } j, \\ 0 & \text{otherwise.} \end{cases}$$

The difference $C_{i(h)} - C_{h(i)}$ is a measure of the relative criticality of X_i compared to X_h which, by excluding models that include both X_i and X_h, is independent of the relationship between these two predictors, and $C_{i(h)} - C_{h(i)} = C_i - C_h$. More sensitive measures can be obtained by standardizing this difference through division by factors such as the sum $C_{i(h)} + C_{h(i)}$ or the number of bootstrapping runs (B). Using these values, a pattern of pairwise dominance can be established which may then be used to rank the criticality of the predictors. The last panel of Table 1 displays the pairwise dominance matrices, where the entry in row i and column h represents the value of $(C_{i(h)} - C_{h(i)})/(C_{i(h)} + C_{h(i)})$, or 0 when $C_{i(h)} = C_{h(i)} = 0$. The row averages rank-order the predictors in terms of relative criticality, while the matrix entries can be used to determine the criticality dominance pattern between any pair of predictors. It can be shown that these pairwise dominance relations are transitive (see Appendix). Thus, it is meaningful to rank-order the predictors according to the row averages, which define the *dominance criticality* measure, dC_i:

$$dC_i = \left(\frac{1}{p-1}\right) \sum_{h \neq i = 1}^{p} \left[\frac{C_{i(h)} - C_{h(i)}}{C_{i(h)} + C_{h(i)}}\right].$$

This is a measure of the dominance of the ith predictor over all other predictors ($h \neq i = 1, \ldots, p$). It has a minimum value of -1 (when $C_{i(h)}$ is 0 and $C_{h(i)}$ is 1 for all $h \neq i$) and a maximum value of $+1$ (when $C_{i(h)}$ is 1 and $C_{h(i)}$ is 0 for all $h \neq i$). Of course, these bounds identify the cases in which X_i is strictly dominated, or strictly dominates, all other predictors in the model. A value of $dC_i = 0$ indicates that $C_{i(h)}$ dominates and is dominated to the same degree by the other predictors.

6. Summary

We have proposed a new approach to the old problem of comparing predictors in multiple regression models (Kruskal, 1987; Budescu, 1993) by defining the *criticality* of the predictors. Traditional measures of predictor importance are conditional on the choice of a model and its overall goodness of fit. They seek to rank-order and/or scale the predictors on a scale that reflects their contributions towards the prediction of the response by a model, or in terms of the average of such measures across all subset models. Unlike these measures of importance, the criticality analysis proposed here does not depend on the choice of a particular model. A predictor's criticality is defined as the probability that its omission from a model would result in the misspecification of the model. Thus, by its definition, the analysis considers all possible models. Criticalities are extracted from, and expressed in terms of, the conditional probability distribution of best-fitting models for a given data set.

The criticality analysis relies on resampling (using the bootstrap), and can be used for data sets containing fixed and/or random predictors. Residual resampling is recommended for the fixed model, and case resampling is suggested when the predictors are random variables. Data sets containing a mixture of fixed and random predictors can also be analysed, and in such a case one would be advised to 'default' to case resampling.

Criticality analysis can be performed with any reasonable criterion of model fit. Our results (simulations and real data) indicate that criticality analysis is relatively insensitive to the choice of this criterion in the ordinal sense. In all our simulations we found a high level of agreement among the ranking of the p predictors from the most to the least critical according to the three selection criteria employed. Indeed, it is important to point out that the criticalities of predictors can be compared directly only if they are based on the same selection criterion. Comparisons across selection criteria are difficult, unless one has a good deal of experience in converting measures from one scale to another. Therefore we recommend that they be avoided. Recall that all criticality measures are based on the frequency with which certain models are identified as best-fitting, and this varies from one criterion to another. To illustrate this point, consider the results from simulation B1 (see Tables 5, 6 and 7). One would expect that $C_1 = 1$ and $C_2 = C_3 = 0$ (X_2 and X_3 are correlated neither with Y nor with X_1). Yet, in all cases, both C_2 and C_3 are clearly greater than 0. Furthermore, note that their values vary considerably and systematically across the selection methods, with the AIC measures being the smallest and the C_p-based measures the largest. These differences reflect the inherent biases of the various selection methods. For example, in this case C_p tended to select larger models (average size of 2.15 predictors) than AIC (average size 1.71), and this tendency affects the various predictors' criticalities.[5]

The major innovation of the proposed method is the linkage between the distribution of best-fitting *models* to the criticality of single *predictors*. A variety of measures of a predictor's criticality can be defined through different linear combinations of the probability distributions across the models. These combinations determine the specific weight to be associated with each model in the assessment of any given predictor. We have discussed in this paper three such measures. The first measure (C_i) is simply the sum of the probabilities associated with all models containing the predictor, and can be interpreted as the expected probability of model misidentification resulting from excluding the predictor. Weighted criticality (wC_i) is defined as a weighted sum of these probabilities, where the weights are inversely proportional to the number of predictors in the model. This measure gives models with fewer (more) predictors greater (less) weight in the determination of criticality. It also has the convenient normalizing property that the sum of the criticalities (over all predictors) is one. The third criticality measure ($C_{i(h)}$) is defined in a pairwise fashion. It is based on the sum of the probabilities associated with models containing one predictor (X_i) and excluding another (X_h), and can be described as a measure of comparative criticality. By aggregating across all pairs, one can also define an overall index (dC_i) that captures the relative dominance of X_i over all other predictors. We consider the three measures to be intuitively compelling and easy to justify and interpret, but realize that numerous alternative functions can be used in this context. For example, since there is a nesting structure for the subset models (for instance, the model '$X_1X_2X_4$' contains six subset models: 'X_1', 'X_2', 'X_4', 'X_1X_2', 'X_1X_4' and 'X_2X_4'), it may be of interest to develop a linear combination that takes this nesting into account. Another intriguing possibility is to calculate all the criticality measures based on a subset of the bootstrapped samples, which consists of only those models which were identified as BFMs with a sufficiently high probability.

Most criteria of goodness of fit are invariant under linear transformations (of the response and predictors). Consequently, the criticality measures are also unaffected by such transformations. In addition, the simple asymptotic distribution theory for these measures allows one to test hypotheses about the difference in the magnitude of criticality values.

We conclude by pointing out that this method is not necessarily restricted to multiple

[5] To illustrate this point, we also ran a simulation where all $p = 4$ predictors were (a) mutually uncorrelated and (b) uncorrelated with Y. All predictors were equally critical, but their average criticality varied across selection methods from 0.4 for adjusted R^2 to 0.65 for C_p.

regression. With minor changes this approach to model selection and predictor ranking can easily be extended to other multi-factor procedures such as log-linear and logistic regression models with categorical data, discriminant analysis and multivariate multiple regression.

7. Software

A SAS macro that implements the procedures described in this paper can be obtained by writing to the first author, or from the web page at http://www.uwm.edu/~ azen/critmacro.html.

Acknowledgement

The authors wish to thank the editor and three anonymous reviewers for their helpful comments and suggestions.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki (Eds.), *Proceedings of the 2nd International Symposium on Information Theory* (pp. 267–281). Budapest Akadémiai Kiado.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*, 542–551.

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, *93*, 549–562.

Chernoff, H. (1956). Large sample theory: Parametric case. *Annals of Mathematical Statistics*, *27*, 1–22.

Draper, N. R., & Guttman, I. (1987). A common model selection criterion. In R. Viertl (Ed.), *Probability and Bayesian statistics*. New York: Plenum Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. New York: Wiley.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*, 323–237.

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics*, *9*, 1218–1228.

Ganzach, Y. (1997). Misleading interaction and curvilinear terms. *Psychological Methods*, *2*, 235–247.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, *32*, 1–49.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., & Lee, T. C. (1985). *The theory and practice of econometrics* (2nd ed.). New York: Wiley.

Kruskal, W. (1987). Relative importance by averaging over orderings. *American Statistician*, *41*, 6–10.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Glenview, IL: Scott, Foresman & Co.

Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious 'moderator effects': Illustrated substantively with the hypothesized ('synergistic') relation between spatial and mathematical ability. *Psychological Bulletin*, *107*, 385–393.

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

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). *Applied linear statistical models* (3rd ed.). Homewood, IL: Irwin.

Shao, J. (1996). Bootstrap model selection. *Journal of the American Statistical Association*, *91*, 655–665.

Stein, R. (1996, July). *Re-sampling techniques: Jackknife and bootstrap.* Workshop conducted at the Inter-university Consortium for Political and Social Research, Ann Arbor, MI.

Suh, E., Diener, E., Oishi, S., & Triandis, H. C. (1998). The shifting basis of life satisfaction judgments across cultures—emotions versus norms. *Journal of Personality and Social Psychology, 74,* 482–493.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.

Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York: Wiley.

Received 18 August 1999; revised version received 5 October 2000

Appendix: Transitivity of dominance matrices

Let $p = 3$, let the three predictors be X_h, X_i, and X_j, and let $P(X)$ be the probability that the model containing predictor(s) X is identified as a best-fitting model. Let $C_{h(i)} > C_{i(h)}$ and $C_{i(j)} > C_{j(i)}$. Then

$$P(X_h) + P(X_hX_j) > P(X_i) + P(X_iX_j)$$

$$P(X_i) + P(X_iX_h) > P(X_j) + P(X_jX_h).$$

Therefore,

$$P(X_h) + P(X_hX_j) + P(X_i) + P(X_iX_h) > P(X_i) + P(X_iX_j) + P(X_j) + P(X_jX_h),$$

$$P(X_h) + P(X_iX_h) > P(X_j) + P(X_iX_j)$$

and $C_{h(j)} > C_{j(h)}$. Therefore, the dominance pattern is transitive. This can be similarly extended to $p > 3$.