Best-Subset Selection Criteria for Multiple Linear Regression

Gordon P. Brooks **Pornchanok Ruengvirayudh**
Ohio University

This study presents comparisons of subset selection criteria used to help determine the "best" regression model in multiple linear regression. No such criteria can replace a researcher's knowledge of theory to help choose useful models, but such criteria may help in exploratory research. Relationships among variables can be more complicated than expected and may require adjustment to theory based on empirical models. Knowing how well each subset selection method performs can be useful in such cases. Monte Carlo simulations were performed to compare a number of well-known criteria (e.g., AIC, BIC, PRESS, adjusted R^2) to some less well-known criteria (e.g., AICc, AICu, GCV, cross-validity R^2). We found that although none of the criteria work well to identify a single correct model across a large number of coefficient and multicollinearity patterns, AIC and adjusted R^2 work reasonably well enough to recommend, in combination, to identify a model within-one of the correct model.

Whenever possible, researchers would prefer to allow theory to drive their decisions about what predictors constitute a "best" regression model. Sometimes, however, lack of strong theory requires researchers to allow the data to speak. In such cases in multiple linear regression, a researcher might use statistical methods to help determine which predictors are most useful and most worthy of continued future empirical consideration. A number of methods exist for this process in regression, including stepwise methods (e.g., forward, backward, stepwise), cross-validation strategies (e.g., leave-one-out, k-fold), and best-subset regression techniques.

In most cases, some criterion is required to help determine which predictors and/or which models are best among the options provided by the data. Most commonly, fit criteria of some kind are used in multiple linear regression. Such criteria include methods to determine the maximum strength of relationship fit and methods based on minimizing error variation, most of which include penalties for models that include more predictors of little value (and sometimes combined with penalties for smaller samples).

The primary purpose of this study was to investigate how well various subset selection criteria identify the correct regression model under varying conditions. In particular, the paper compared both the absolute and relative success of several selection methods for identifying the correct model from among several predictors. However, after some early results, emergent design based on those results led us to compare how effectively the selection criteria worked to identify a model "within-one" of the correct model.

Theoretical Perspectives

This study applies to standard multiple linear regression analysis, where all possible subsets of predictors are entered in order to create all possible models. In all subsets regression, researchers attempt to identify, based on some criterion, the "best" subset model of predictors. That is, researchers attempt to identify a subset model that works well, perhaps even as well or better than the full model of all available predictors. Some statistical programs have procedures available for such an analysis.

One approach to best-subset regression is to compare all possible regression models on the chosen criterion. There are 2^k-1 possible models with predictors that can be created from a set of k predictors. Another approach is to compare the best subset of predictors of each possible model size. For k predictors, there are k possible "best" models, one model of each size from 1 to k. Researchers choose a criterion statistic to use for the model comparisons. For example, by default SPSS provides several statistics in its "Model Summary" table that might be used for such a purpose: R^2, adjusted R^2, standard error of the estimate (*SEE*), and perhaps even R^2 change and F change statistics. With the "selection" syntax option included, SPSS also will provide the Akaike Information Criterion (AIC), the Schwarz Bayesian Criterion (BIC), the Amemiya Prediction Criterion (APC), and Mallows' Prediction Criterion (C_P). Newer editions of SPSS (e.g., version 22) include an implementation of all subsets regression, but it is an interactive procedure without the same detailed results as its traditional regression procedure.

Whatever methods and criteria are used, researchers must always remember that our scholarly confidence in regression models must be based on cross-validation. Therefore, the techniques described in this paper are intended to represent approaches to exploratory use of regression modeling and model

building (letting the data speak). Further, researchers are reminded that, due often to correlation among predictors, there may not be a single best model. Learning this is not something to be feared, but too many want statistical analyses to provide the single correct answer. Research is about theory, not statistics. Information gained through comparison of multiple models may help uncover complicated relationships among the variables (especially predictors) under investigation – especially when theory is sparse. Subset selection methods and criteria can help us understand potential theoretical models for further study.

Variable Selection Methods

Variable selection generally refers to systematic techniques for selecting a subset of predictor variables, from among a specified set of predictors, that adequately explains or predicts the given criterion variable (Weisberg, 1985). Generally, the most important tool in selecting a subset of variables for a multiple linear regression model is careful logical analysis based on the analyst's knowledge of theory and research in the area of study (Gordon, 1968). Not all researchers see the same relationships, same order of variable importance, and same models – perhaps based on different levels of theoretical expertise, experience, and creativity. Weisberg (1985) noted, however, that often a point is reached at which it becomes necessary to use data to determine a best subset of predictors. Afifi and Clark (1990) similarly suggested that the researcher may have prior justification for some but not all of the variables studied. Further, variable selection methods are often performed when a large number of candidate variables are under consideration, with theoretical rationale, but a priori knowledge does not provide clear understanding of their relevance (Flack & Chang, 1987; Huberty, 1989). For example, Weisberg (1985) noted that a smaller set of selected variables that provide nearly the same information as the original full set of variables can help focus future research in the area and simplify analysis.

Variable selection constitutes strategies by which a subset of "better" variables is chosen from among a "larger constellation of predictors" (Thompson, 1995, p. 525). Breiman (1995) suggested that these subsets are useful for two primary reasons: variance reduction and simplicity (i.e., parsimony). More regression coefficients increase the overall variance and the prediction errors. Huberty (1989) and Thompson, like Breiman, noted that while stepwise analyses may be used to assess relative importance of the predictor variables, the more accepted reason is to select a more parsimonious set of predictor variables for a final model. Weisberg (1985) also indicated that deletion of predictors from a prediction model can improve it and reduce apparent multicollinearity. Two more common approaches have evolved by which such subsets can be obtained in regression: stepwise methods and best-subsets.

In stepwise regression procedures, linear models are developed in a sequence of steps by adding and/or deleting predictor variables. That is, a path through possible models is chosen based on an appropriate statistical criterion, identifying first a subset of one size and then adding/deleting predictors until a final model is reached based on some stopping criterion. Stepwise regression techniques differ from all-subsets regression techniques because only a limited number of models of each size are examined. All-subsets regression, on the other hand, provides analysis of certain criterion statistics computed for every possible model of every size. Best-subsets regression, which we focus on in this study, then identifies the best model at each size, from among all possible models of that size, based on that criterion.

Best-Subsets Regression

Whereas in stepwise methods successive models are limited by variables already in the model from previous steps of the analysis, all-subsets regression provides analysis of certain statistical criteria (e.g., AIC, adjusted R^2, Mallow's C_P) computed for every possible model of every size (Weisberg, 1985). Then, in the best-subsets approach, for each model of a given size the best-subset model of predictors is chosen based on the chosen statistical criterion. The number of total models of all sizes is 2^k-1 (where k is the number of predictors), while the number of best-subset models is equal to the number of predictors. For example, for five predictors there will be 31 possible regression models for all subsets, but only five best-subset models based on some criterion such as the highest adjusted R^2 or lowest AIC: one best one-predictor model, one best two-predictor model, one best three-predictor model, one best four-predictor model, and the one full, five-predictor model. Because best-subsets approaches are computer intensive, because all possible regressions must be created, it is not always feasible to use the method, especially as k increases.

Concerns about Variable Selection Methods

Most scholars express concern over the use of data-based variable selection methods generally, and stepwise methods specifically (e.g., Harrell, 2015; Keith, 2005). We will not reiterate all the arguments against these methods in this paper, but we will highlight some more common concerns. Many concerns are related to using these methods in explanatory research.

All models, whether developed through stepwise or all subset procedures, are limited due to sampling error. Further, all models are susceptible to model specification error. Variable selection methods, however, remain common techniques in multiple linear regression. The most critical objection to variable selection methods is that, even when used for prediction, they may often fail to select the optimal subset of predictors, particularly when multicollinearity is present (Fox, 1991). Variable selection methods also cannot guarantee that the best variable set for any given size will be selected (Hocking, 1976; Thompson, 1995). Indeed, it should be noted that because of relationships among the variables, different variable selection approaches cannot always be expected to produce the same subset models at each step. Further, Hocking noted that excellent models may be missed when using stepwise methods because of the restriction of adding only a single variable at each step. Thompson described this concern as "a linear series of conditional decisions not unlike the choices one makes in working through a maze. An early mistake in the sequence will corrupt the remaining choices" (p. 532). Huberty (1989) noted that multicollinearity may result in a particular combination of variables chosen for the final model in one sample, but may result in a different combination in another very similar sample. Such mistakes in the sequence and relationships among predictors are often due to sample-specific variation, which is one of the reasons stepwise results often do not generalize. Derksen and Keselman (1992) determined that sample size impacts the number of authentic variables included in the final models. All-subsets regression and best-subsets regression have become more popular as computing power has made it more practical to examine all possible models. However, Berk (1978) reported that "in the sample, the all-subsets procedure always produces that best set for each subset size. However, this need not be the case in the population" (p. 3).

Potential Usefulness of Variable Selection Methods

Some scholars have recommended processes for using variable selection methods in a reasonable manner. For example, Wilkinson (1979) indicated that cross-validation should be performed in lieu of statistical significance testing for reduced models. That is, the results of variable selection analyses must be cross-validated in a new sample and only conclusions that can be drawn from both samples should be made. Huberty (1989) also recommended that results should only be considered valid when results can be shown to replicate in another sample.

Other scholars acknowledge that variable selection methods may be useful to help develop better prediction models or to manage multicollinearity (e.g., Herzberg, 1969). Fox (1991) also noted that selection techniques seem well-suited for prediction problems, so long as reasonable data generalizability conditions are met. That is, even badly biased coefficients may produce good estimates of the criterion variables. Similarly, Roecker (1991) indicated that predictive accuracy and model parsimony are reasonable motivations for subset selection. Copas (1983) reminded us that a good prediction equation may include predictors that are not individually statistically significant and exclude others that are significant. Consequently, Copas argued that several subsets should be examined prior to any determination of the best model.

It is hard to understand why the recommendation from Copas (1983) would not also make sense for any exploratory research. That is, the best model from a theoretical perspective may include predictors that are not statistically significant after controlling the other predictors, but do contribute to a statistically significant – and more importantly, a theoretically significant – model. In particular, such results due to unusual or unexpected correlation patterns among the predictors (e.g., suppressor relationships) may be theoretically valuable for either prediction or explanation. To this end, some scholars recommend all possible regressions, so all possible models for a given set of predictors can be compared. As predictors increase, however, this becomes more difficult and the attractiveness of other best-subsets approaches increases.

Flack and Chang (1987) noted that the best set of predictors should be theory-driven: "strictly speaking, variables should not be selected solely on the basis of statistical data analysis" (p. 84). Huberty (1989) argued that large predictor pools be reduced and that theory and prior experience should provide guidance for initially screening out many of the variables during the study design process. Wilkinson (1979) suggested that subset selection analyses can be almost as effective as biased estimation techniques (e.g., ridge regression) in minimizing both prediction errors and coefficient errors when the predictors are highly correlated.

Thompson (1995) and others have suggested guidelines for safer use of stepwise analysis, which may apply also to best-subsets regression. In particular, less sampling error tends to be present in data (a) based on larger samples, (b) with fewer predictors, and (c) larger effect sizes. Thompson suggested that stepwise analyses with more orthogonal predictors may distort the analysis less and may be "somewhat less sinful" (p. 533). Cohen and Cohen (1983) suggested that for stepwise regression to be useful, the analyses should (a) be used primarily for predictive purposes and only secondarily for explanation, (b) be based on very large samples, and (c) be cross-validated. Similarly, Derksen & Keselman (1992) explained that most problems affecting results of stepwise analyses are due to multicollinearity, smaller sample sizes, and larger numbers of predictor variables in the analysis. They argued that compensating for these factors may provide more acceptable stepwise results.

Model Selection Criteria

A number of criteria have been developed over the years to help researchers choose the best, or at least better, models. These criteria are calculated for each model and compared. Some allow statistical comparison (particularly when the models are nested) but most are used without statistical significance testing. Consequently, theoretical knowledge is useful to help determine which models are most useful when criteria are very similar.

Adjusted R^2 is often used to help identify the best model because, unlike R^2, it penalizes additional predictors that do not help. When adjusted R^2 is used as the criterion, the model with the largest adjusted R^2 is considered the best. Standard error of the estimate (*SEE*), or root mean squared error (*MSE*), is also commonly used. Because it is based on error, when *SEE* is used, the best model has the smallest *SEE*.

A class of criteria based on *MSE* are commonly used in subset selection decisions. Most commonly, perhaps, is AIC and a number of modified or related criteria (e.g., AICc, AICu, and RIC). Each modification provides a different penalty based on the number of predictors and/or sample size. Another approach, similar in calculation but derived differently – from a Bayesian perspective – is the BIC. Another criterion is the prediction error sum of squares (PRESS) statistic, which is based on prediction for each case when it is left out of the model. The Mallows C_P statistic is based on the error for each reduced model as compared to error for the full model of available predictors. The generalized cross validation (GCV) criterion is also based on model error.

Finally, this study included a number of statistics that have typically not been used for model selection purposes, but seem reasonable because a number of scholars have suggested cross-validation approaches to model selection. Several cross-validity R^2 statistics will be tested for model selection: one statistic due to Stein (1960) and Darlington (1968), one due to Lord (1950) and Nicholson (1960), and one due to Browne (1975). These statistics attempt to estimate the ability of a model to predict in another sample from the same population. R^2_{PRESS} is also often used for this purpose.

Purpose of the Study

We agree that variable selection methods used for explanatory purposes may be problematic for reasons addressed above. We argue, however, that using variable selection methods thoughtfully for exploratory model building and model comparison can be an efficient and useful methodology. All-subsets and best-subsets regressions allow researchers to compare multiple models.

Monte Carlo research has shown that neither stepwise selection nor all possible regressions may find the correct subset of predictors, depending on the level of multicollinearity in the data (e.g., Olejnik, Mills, & Keselman, 2000). But even if we cannot be sure which models are correct, the illumination provided by the various models may provide useful information about these complicated relationships among predictors and, indeed, all variables being studied.

Methods and Data Source

We used Monte Carlo simulation methods to investigate the research problem. A computer program was written in the statistical programming language, R, and used the *leaps* package for all-subsets regression analyses. The core of the program was tested and the output was verified in multiple ways. For example, single datasets were examined to verify that the accuracy of results as compared to hand calculations and to results from other programs and procedures. Also, we examined small numbers of replications (e.g., 10 and 100) to verify that the cumulative results across samples were being stored correctly by the program. For both single sample and small replication results, all variables were examined to ensure correct or reasonable results were obtained. Sensitivity testing was performed to ensure that the program and function logic worked correctly for all conditions that were varied. Stress testing was also performed to verify that strange values did not cause unexpected problems (e.g., division by zero). For example, correlation matrices used to generate data were verified to be legitimate (e.g., positive definite). Data generation techniques were verified to ensure they produced reasonable samples for the population values set.

Multivariate normal data were generated for 10,000 samples for each condition described below. In particular, five predictor variables (k) were used, with varying levels of explained variation (R^2), varying multicollinearity (i.e., none and moderate, based on variance inflation factor, or VIF), and three sample sizes (i.e., n = 50, n = 250, n = 500) loosely based on examination of cross-validation methods such as Park and Dudycha (1974) and Brooks and Barcikowski (2012). For each sample, the best subsets for five predictors were obtained using the exhaustive (all possible) regressions approach.

For data generation, we varied the correlation coefficients between the predictors and the dependent variable (the correlation pattern) from .0 to .8 with a .2 increment to cover a wide range of possible predictor correlation values with the outcome. For example, the first model created this way had 0, 0, 0, 0, and .2 as correlation between predictors and the outcome, while the last model had correlations of 0, .2, .2, .4, and .8. Some models had only one non-zero correlation (e.g., the first model), while some (e.g., .2, .4, .4, .4, and .6) had five non-zero correlations. This resulted in 47 total correlation matrices for each multicollinearity condition. Of course, certain patterns of these correlation coefficients were not possible (e.g., uncorrelated predictors with correlations with the outcome of 0, .2, .4, .6, and .8 would result in $R^2 > 1.0$ and would not be legitimate). We used a maximum R^2 of .90 with no multicollinearity as our ceiling (e.g., uncorrelated predictors with 0, .2, .4, .6, and .6 correlations with the outcome would result in $R^2 = .92$ and therefore that pattern was not included). We also included four patterns having just a .1 increment, from (0, 0, 0, .1, .2) to (.1, .2, .3, .4, .5), to explore a set of smaller predictor correlations with the dependent variable and very small overall R^2 values with multiple predictors.

We varied the patterns of correlations among predictors from no multicollinearity (i.e., all correlations among predictors equal 0, called M0 here) to moderate multicollinearity (based on VIF). The M1 multicollinearity pattern represented a situation where all predictors are correlated at r = .2 (where all VIF = 1.1 for each predictor) and M2 represented all predictors correlated at r = .4 (where all VIF = 1.4). Table 1 shows the multicollinearity patterns for M3 and M4, which were not consistent across all predictors like M1 and M2. It should be noted that as multicollinearity increased, it became increasingly difficult to identify correlation matrices that were positive definite for all 51 sets of predictor-outcome correlation patterns (e.g., we were not able to set all predictor correlations at r = .6). Several higher multicollinearity conditions were attempted until two could be identified that successfully produced positive definite matrices in combination with all 51 predictor-outcome correlation patterns. This resulted in only two particular patterns of higher multicollinearity used in the study. Further, it is important to note that the levels of multicollinearity represented by these matrices would be considered relatively mild (e.g., no VIF higher than 5.0). The combination of correlation patterns and multicollinearity produced population R^2 values of .04 to .96 across all conditions (some R^2 values increased above our original criterion of .90 as multicollinearity increased).

For each sample, selection criteria were used to identify the "best" model from among the five possible best-subset models (e.g., the model with the minimum AIC or the maximum adjusted R^2). We recorded whether each criterion identified the correct model. More specifically, we determined whether the criterion chose the model of the correct size (that is, the number of predictors in the model). Selection criteria statistics were calculated by the authors and verified where possible using built-in R functions.

Table 1. Multicollinearity Conditions and Associated Variance Inflation Factors (VIF) Values for the M3 and M4 Correlation Matrices Used in Simulated Samples

Multicollinearity		Correlations among Predictors					
Condition	Predictors	X1	X2	X3	X4	X5	VIF
M3	Y	0	.2	.2	.4	.6	-
	X1		.8	.6	.4	.2	3.0
	X2			.6	.4	.2	3.0
	X3				.4	.2	1.7
	X4					.2	1.3
	X5						1.1
M4	Y	0	0	.1	.2	.3	-
	X1		.8	.8	.4	.2	3.5
	X2			.8	.4	.2	3.5
	X3				.6	.2	4.6
	X4					.2	1.6
	X5						1.1

Note. Correlations with the outcome Y are examples of the 51 patterns we used.

We made a practical decision in order to determine which model was correct. There are at least three competing definitions for the "true" or "correct" population model. First, we could have chosen to consider the correct model to be the one that includes only the predictors which have non-zero correlations with the dependent variable. This rule, however, would neglect the fact that, due to multicollinearity, partial population regression coefficients may be non-zero even if they have zero correlation with the outcome (or vice versa). Second, we could have chosen to consider the correct model to be the one that includes only predictors with non-zero population regression coefficients. This rule, however, would almost exclusively have resulted in the full model being considered the correct model in our simulated conditions when there was multicollinearity. Therefore, we chose a practical solution and rounded population coefficients to the nearest tenth. Rates of success were calculated for each selection criterion method using this rule.

Phase 1 Results

We began by investigating which selection criterion most effectively identified the correct model. However, we saw that, as other researchers have, none performed all that well. They all performed most successfully with no multicollinearity. We learned that some criteria work better for some conditions and some work better for others.

More specifically, we calculated how frequently each selection criterion identified the correct model out of the 10,000 samples for each of the 765 conditions (i.e., 3 sample sizes, 5 multicollinearity conditions, and 51 regression coefficient patterns). We determined that BIC was most frequently correct across more of the conditions. That is, BIC identified the correct model in at least 80% of the samples (i.e., 8000 of the 10,000 samples) in each condition for 271 conditions (35.4% of 765 total conditions). AICu was correct at least 80% of the time for 235 conditions, AIC for 197, and adjusted R^2 for 186 conditions. BIC was correct in 90% of the samples for 203 conditions. Disappointed in the low accuracy rate, we decided to alter our investigation to study how frequently the selection criteria identified a model within one predictor (in size) of the correct model (described in Phase 2 results). That is, if the correct model had three predictors, we would consider a criterion as being correct "within one" if it identified the two-predictor model or the four-predictor model.

We also determined that a number of selection criteria we tested were highly correlated. That is, when examined across the 51 conditions for each condition, we discovered that methods were correlated in how frequently they identified the correct model. Table 2 shows the minimum correlations among the selection criteria across the 51 regression coefficient patterns for the 15 sample size and multicollinearity conditions. For example, out of 15 correlation matrices calculated in this way, AIC and AICc are always correlated at or above r = .96 (i.e., for all conditions, AIC and AICc are very highly correlated).

The following criteria are represented in the table:

BIC = Schwarz Bayesian Information Criterion (Schwarz, 1978)
CAIC = Conditional AIC (Bozdogan, 1987)
AICu = Unbiased AIC (McQuarrie, Shumway, & Tsai, 1997)
AICc = Corrected AIC (Hurvich & Tsai, 1991)
AIC = Akaike Information Criterion
MinCP = Minimum Mallow's C_P (Mallows, 2000)
CP = Mallow's C_P (Mallows, 2000)
Rc2SD = Stein-Darlington Cross-Validity R^2 (Darlington, 1968; Stein, 1960)
Rc2LN = Lord-Nicholson Cross-Validity R^2 (Lord, 1950; Nicholson, 1960)
Rc2B = Browne Cross-Validity R^2 (Browne, 1975)
PRESS = Prediction Error Sum of Squares
RIC = Risk Inflation Criterion Corrected (Leng, 2013)
GCV = Generalized Cross-Validation (Takezawa, 2014)
Amemiya = Amemiya Prediction Criterion (Amemiya, 1980)
Ra2 = Adjusted R^2 (Ezekiel, 1930; Wherry, 1931)
SEE = Standard Error of the Estimate
Ra2OP = Olkin-Pratt Adjusted R^2 (Olkin & Pratt, 1958)

Based on these correlations, we decided to reduce the number of selection criteria in the study to the following five: BIC, AICu, AICc, AIC, and adjusted R^2 (and indeed, because the results for AICc were not so different from AICu and AIC, we focused later on just the other four criteria).

The final result of note from Phase 1 was the determination that methods were often correct for different conditions. For example, of the 271 conditions where BIC was correct at least 80% of the time, adjusted R^2 was correct 80% of the time for only 108 of them and AIC was correct for only 142. For example, there were 78 conditions where adjusted R^2 was correct 80% of the time but BIC was not. In succeeding phases, we attempted to identify rules that researchers might be able to use to maximize the correct choice of criterion. For example, if a researcher can identify when to use BIC and when to use adjusted R^2, then we can increase the correct choice from 271 conditions to 349 conditions.

Table 2. Minimum Correlations among Correct Identifications of Selection Criteria across 15 Cells for N and Multicollinearity (Correlations above .95 are Highlighted)

	BIC	CAIC	AICu	AICc	AIC	MinCp	Cp	Rc2SD	Rc2LN	Rc2B	PRESS	RIC	GCV	Amemiya	Ra2	SEE	Ra2OP
BIC	1.00	0.98	0.55	0.30	0.29	0.29	0.29	0.29	0.29	0.30	0.29	0.29	0.29	0.29	0.10	0.10	0.10
CAIC	0.98	1.00	0.48	0.23	0.22	0.22	0.22	0.22	0.22	0.23	0.22	0.22	0.22	0.22	0.02	0.02	0.02
AICu	0.55	0.48	1.00	0.85	0.83	0.83	0.83	0.83	0.83	0.84	0.84	0.83	0.83	0.83	0.31	0.31	0.31
AICc	0.30	0.23	0.85	1.00	0.96	0.98	0.98	0.98	0.96	0.96	0.97	0.98	0.98	0.96	0.53	0.53	0.52
AIC	0.29	0.22	0.83	0.96	1.00	1.00	1.00	0.99	1.00	0.99	0.99	1.00	1.00	1.00	0.74	0.74	0.74
MinCp	0.29	0.22	0.83	0.98	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.67	0.67	0.67
Cp	0.29	0.22	0.83	0.98	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.67	0.67	0.67
Rc2SD	0.29	0.22	0.83	0.98	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.67	0.67	0.67
Rc2LN	0.29	0.22	0.83	0.96	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.74	0.74	0.74
Rc2B	0.30	0.23	0.84	0.96	0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.71	0.71	0.71
PRESS	0.29	0.22	0.84	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.70	0.70	0.70
RIC	0.29	0.22	0.83	0.98	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.67	0.67	0.67
GCV	0.29	0.22	0.83	0.98	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.68	0.68	0.68
Amemiya	0.29	0.22	0.83	0.96	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.74	0.74	0.74
Ra2	0.10	0.02	0.31	0.53	0.74	0.67	0.67	0.67	0.74	0.71	0.70	0.67	0.68	0.74	1.00	1.00	1.00
SEE	0.10	0.02	0.31	0.53	0.74	0.67	0.67	0.67	0.74	0.71	0.70	0.67	0.68	0.74	1.00	1.00	1.00
Ra2OP	0.10	0.02	0.31	0.52	0.74	0.67	0.67	0.67	0.74	0.71	0.70	0.67	0.68	0.74	1.00	1.00	1.00

Phase 2 Results

Because we were trying many different combination and multi-step rules in Phase 2, we changed to 1,000 replications per condition. We ran many Monte Carlo attempts to try to identify rules and needed to be able to process results more quickly than 10,000 iterations would allow. However, we increased the sample sizes to include an n = 100 condition, which increased our total number of conditions to 1,020 (4 sample sizes by 5 multicollinearity by 51 correlation-coefficient patterns). Finally, while we continued to verify the "actual correct" results from Phase 1, in Phase 2 we focused on the "within-one" correctness (for both correctness percentages and correlations among criteria).

We chose to investigate within-one for a theoretical reason as well. Knowing that, like us, previous researchers have generally found that subset selection criteria do not find the correct model at high rates of accuracy, we decided to use the within-one rule to get a sense of how frequently the methods were close to the correct model (in terms of number of predictors). That is, we believe, as did a number of scholars cited above, that researchers must bring theoretical knowledge to bear on subset selection methods in multiple regression. Knowing that a selection criterion can provide accuracy within-one can help a researcher identify the most promising candidate models from a statistical perspective (based on their data). This may help maximize their ability to determine a best model based on the other knowledge they bring to the problem.

We present Table 3 as an example of the kind of results we examined. Several examples of interesting outcomes from the n = 250 and M1 condition are highlighted. For example, the row identified as Condition 19 shows that all selection criterion identified models within-one of the correct model size in 100% of the samples. Condition 25 shows that none of the selection criterion were correct within-one at least 80% of the time (indeed, the maximum was 37.8% percent). In Conditions 21 and 23, adjusted R^2 is correct more frequently than any of the other three criteria shown here (i.e., BIC, AICu, and AIC). In the MAX column, the .727 reflects that adjusted R^2 is the maximum value for Condition 21 (which was not above the 80% correctness rule) and .938 for Condition 23 (in Condition 23, however, AIC is also above the 80% rule). Finally, Condition 29 shows a condition where BIC, AICu, and AIC are all above the 80% correctness rule, but adjusted R^2 only identified the within-one correct model 67% of the time.

Condition 17 (not highlighted) shows two regression coefficients (both 0.03) that would be considered 0.0 when rounded to the nearest tenth; therefore, the correct model in Condition 17 is considered a three-predictor model. Note that the original correlations are in the far right columns, but because of the multicollinearity, the population coefficients have changed (with no multicollinearity, the population coefficients would be the same as the correlations with the outcome). Table 3 also shows some of the combination rules we tried. For example, we examined how many times the three more common criteria (BIC, AIC, and adjusted R^2) agreed. That is, Table 3 shows columns for when adjusted R^2, BIC, and AIC, all correctly agreed in the samples (R=S=A), meaning that the maximum adjusted R^2, the minimum BIC, and the minimum AIC all pointed to the same model and this model was a correct within-one model (the three may also point to the same model, but it may not be a correct within-one model). We recorded how frequently this agreed-upon model was within-one of the correct model size. We also recorded when adjusted R^2 and AIC agreed (R=A), when BIC and AIC agreed (S=A), and when adjusted R^2 and BIC agreed (but R=S is not shown in the example).

Further, Table 3 shows an example of one of the many two-step rules we tried. Because the selection criteria did not frequently agree to identify the correct model within-one (when paired as described above), we tried two-step rules that would follow disagreement with another choice of selection criterion. For example, in Table 3, sample results are shown for the rule that first tested whether BIC and AIC agreed and then, if they did not agree, would use adjusted R^2 to identify the model. This rule is abbreviated as (S=A|R). We had hoped that such rules would take advantage of the agreement of the criteria in the conditions in which they performed well, but then revert to another criterion (here, adjusted R^2) for the conditions where it performed better. Sadly, the rules continued to agree even in conditions where they performed poorly and did not revert to the second option enough to improve the results. For example, in Condition 24 (not highlighted), BIC reached the 80% rule in 13.8% of the samples, and AIC reached it in 66.1% of the samples. Adjusted R^2, however, reached it in 84.8% of the samples, meeting our 80% correctness rule. The difficulty is trying to find a rule that will tell a researcher to use adjusted R^2 for that condition, in order to maximize correctness. We had hoped the two-step rule would allow a

Table 3. Sample Results from the Monte Carlo Simulations

Cond	RSQ	b1	b2	b3	b4	b5	MAX	BIC	AICu	AIC	Ra2	R=S=A	R=A	S=A	S=A|R	ry1	ry2	ry3	ry4	ry5
15	0.40	-0.17	-0.17	0.33	0.33	0.33	1.000	0.977	0.997	1.000	1.000	0.674	0.964	0.707	1.000	0.0	0.0	0.4	0.4	0.4
16	0.38	-0.19	0.06	0.31	0.31	0.31	0.996	0.884	0.970	0.987	0.996	0.517	0.823	0.664	0.987	0.0	0.2	0.4	0.4	0.4
17	0.34	0.03	0.03	0.28	0.28	0.28	1.000	1.000	0.990	0.960	0.875	0.433	0.679	0.659	0.907	0.2	0.2	0.4	0.4	0.4
18	0.44	-0.22	0.28	0.28	0.28	0.28	1.000	0.998	1.000	1.000	1.000	0.981	0.999	0.982	1.000	0.0	0.4	0.4	0.4	0.4
19	0.40	0.00	0.25	0.25	0.25	0.25	1.000	1.000	1.000	1.000	1.000	0.692	0.859	0.828	1.000	0.2	0.4	0.4	0.4	0.4
20	0.44	0.22	0.22	0.22	0.22	0.22	1.000	0.999	1.000	1.000	1.000	0.900	0.989	0.911	1.000	0.4	0.4	0.4	0.4	0.4
21	0.40	-0.08	-0.08	-0.08	-0.08	0.67	0.727	0.038	0.258	0.483	0.727	0.015	0.343	0.025	0.647	0.0	0.0	0.0	0.0	0.6
22	0.41	-0.11	-0.11	-0.11	0.14	0.64	0.944	0.315	0.657	0.835	0.944	0.099	0.649	0.142	0.897	0.0	0.0	0.0	0.2	0.6
23	0.41	-0.14	-0.14	0.11	0.11	0.61	0.938	0.294	0.639	0.833	0.938	0.115	0.659	0.153	0.899	0.0	0.0	0.2	0.2	0.6
24	0.40	-0.17	0.08	0.08	0.08	0.58	0.848	0.138	0.452	0.661	0.848	0.063	0.513	0.084	0.771	0.0	0.2	0.2	0.2	0.6
25	0.38	0.06	0.06	0.06	0.06	0.56	0.378	0.000	0.027	0.134	0.378	0.000	0.110	0.000	0.337	0.2	0.2	0.2	0.2	0.6
26	0.51	-0.14	-0.14	-0.14	0.36	0.61	1.000	0.902	0.980	0.992	1.000	0.426	0.899	0.496	0.992	0.0	0.0	0.0	0.4	0.6
27	0.50	-0.17	-0.17	0.08	0.33	0.58	0.996	0.766	0.939	0.978	0.996	0.400	0.828	0.506	0.980	0.0	0.0	0.2	0.4	0.6
28	0.48	-0.19	0.06	0.06	0.31	0.56	0.849	0.280	0.522	0.700	0.849	0.154	0.532	0.207	0.715	0.0	0.2	0.2	0.4	0.6
29	0.44	0.03	0.03	0.03	0.28	0.53	0.996	0.996	0.950	0.868	0.669	0.243	0.488	0.504	0.749	0.2	0.2	0.2	0.4	0.6
30	0.58	-0.19	-0.19	0.31	0.31	0.56	1.000	1.000	1.000	1.000	1.000	0.969	1.000	0.969	1.000	0.0	0.0	0.4	0.4	0.6
31	0.54	-0.22	0.03	0.28	0.28	0.53	1.000	0.999	1.000	1.000	1.000	0.645	0.840	0.805	1.000	0.0	0.2	0.4	0.4	0.6
32	0.50	0.00	0.00	0.25	0.25	0.50	1.000	1.000	0.992	0.970	0.879	0.467	0.659	0.726	0.918	0.2	0.2	0.4	0.4	0.6

researcher to see that BIC and AIC did not agree and therefore use adjusted R^2. However, the rule was only correct 77.1% of the time. We found that using the better individual criterion rules (e.g., AIC by itself) nearly always outperformed these multi-step combination rules. That is, even though the multi-step combination rules we tried did achieve a balance between individual criteria as we had hoped, one of the individual rules was nearly always more accurate than the combination rules, thereby making it difficult to recommend combination rules or multi-step rules that use multiple selection criteria.

In particular, AIC and adjusted R^2 provided the most frequent correct within-one results. In fact, we found that AICu was correct within-one most frequently using a 95% rule (i.e., within-one of the correct model in at least 95% of the samples in a condition), but that it was only correct for 54.1% of the 1,020 conditions. AIC was most accurate at both a 90% rule (for 59.9% of the conditions) and an 80% rule (for 70.6% of the conditions). Adjusted R^2 was most accurate within-one for all correctness rules below 80% (e.g., 70%, 60%, 50%) and, indeed, was correct more frequently for all conditions where none of the methods reached the correctness rule (e.g., recall Condition 25 in Table 3).

Table 4 shows the percentages across the 1,020 conditions of each individual method being correct within-one. For example, for no multicollinearity M0 and n = 50, AIC was correct at least 80% of the time in 42 (82.4%) of the 51 conditions. The column labeled maxPC represents the percentage correct if we could identify the best method to use in every condition. In that same M0 and n = 50 condition, if we knew which method to use in every one of the 51 conditions, we could be at least 80% accurate within-one for 48 (94.1%) of the 51 conditions. Consequently, AIC would help us identify the correct within-one model size in 42 (87.5%) of those 48 conditions. Adjusted R^2 would identify the other six conditions at 80% correctness within-one. We could use other criteria to help us identify the correct model within-one, but the accuracy rate would be lower than80% of the time for three of the conditions (the results show the lowest to be 52%). In total, if we were able to determine which criterion to use across all conditions, we could correctly identify models within-one in 819 (80.3%) of the 1,020 conditions, but using only AIC always, we could do so in 720 (70.6%) of the conditions.

We also found that we could not reach 50% within-one correctness in 99 (9.7%) of the conditions. Most commonly, the conditions that resulted in poor identification of the correct within-one models were those with smaller sample sizes, multicollinearity, and several very small regression coefficients (i.e., 0.1 or 0.2). As sample size increased, however, these selection criteria did indeed perform more effectively. For example, with n = 100 and M1 (which resulted in regression coefficients of 0.1, 0.1, 0.1, 0.1, and

0.6), AIC identified a correct model within-one in 4% of the samples; with $n = 500$, however, AIC identified a correct within-one model 37.1% of the time (adjusted R^2 increased from 17.9% when $n = 100$ to 65.5% when $n = 500$).

The yellow highlights in Table 4 identify the most effective criteria within each cell (20 combinations of sample size and multicollinearity). The green highlights indicate the most effective criteria at the margins for either sample size (4 sample sizes) or multicollinearity (5 levels of multicollinearity). The blue highlights show the most effective across all 1,020 cell combinations. In recognition of the relatively small number of replications (i.e., 1,000), highlights show the most effective criterion as well as any other criterion within 3% of that most effective result.

Table 4. Percentages of 80% Within-One Correctness in Conditions.

Mean

N	MC	maxPC	SBC80	ICU80	AIC80	RA280
50	0	.941	.686	.686	.824	.765
	1	.431	.196	.216	.333	.353
	2	.490	.314	.314	.373	.412
	3	.549	.314	.314	.373	.471
	4	.627	.314	.353	.451	.588
	Total	.608	.365	.376	.471	.518
100	0	1.000	.804	.863	.961	.784
	1	.588	.353	.431	.490	.510
	2	.608	.431	.451	.490	.510
	3	.784	.451	.490	.588	.706
	4	.765	.490	.549	.647	.725
	Total	.749	.506	.557	.635	.647
250	0	1.000	1.000	1.000	1.000	.765
	1	.804	.529	.588	.627	.686
	2	.843	.549	.647	.745	.745
	3	.941	.588	.843	.882	.843
	4	.882	.706	.804	.804	.804
	Total	.894	.675	.776	.812	.769
500	0	1.000	1.000	1.000	1.000	.745
	1	.922	.627	.765	.843	.784
	2	.941	.725	.843	.863	.824
	3	.961	.843	.941	.941	.843
	4	.980	.804	.863	.882	.902
	Total	.961	.800	.882	.906	.820
Total	0	.985	.873	.887	.946	.765
	1	.686	.426	.500	.574	.583
	2	.721	.505	.564	.618	.623
	3	.809	.549	.647	.696	.716
	4	.814	.578	.642	.696	.755
	Total	.803	.586	.648	.706	.688

Table 4 shows clearly that with any multicollinearity (recall that the multicollinearity we used in the study was relatively minimal), we cannot use these selection criteria to choose models even within-one very well when sample sizes are small. That is, when $n = 50$ (which would generally be considered small for a multiple regression) and $n = 100$ (which would often be considered acceptable for a multiple regression with five predictors), the criteria often more frequently than not did not reach 80% correctness within-one (i.e., most proportions with multicollinearity are below .5). The adjusted R^2 criterion was most useful when multicollinearity existed, but clearly was not the best choice with no multicollinearity, which was AIC. As multicollinearity increased slightly to M3 and M4 with smaller sample sizes, adjusted R^2 became the clear choice. As sample sizes increased to $n = 250$ and $n = 500$ (both of which would be very large for most multiple regressions with five predictors), AIC becomes a relatively clear choice based on an 80% correct rule. Also recall that AIC was generally better across the same conditions with an 80% as well as 90% correctness within-one rule.

Discussion

Using the results presented above, we believe new rules can be developed based on sample size and multicollinearity. First, as a general principle, we need to seriously consider much larger sample sizes in multiple regression. We almost always have some collinearity among predictors in multiple regressions performed in social sciences. Second, if a sample has almost no multicollinearity (recall that M1 had $r = .2$ for all predictors), we would recommend AIC at all sample sizes (see Table 4). Third, if a sample is small (e.g., $n \leq 100$) and has any multicollinearity, we recommend adjusted R^2. We added simulation for $n = 150$ and $n = 200$ to comparing several new rules across additional sample sizes based on this combination of criteria. We found that the rule "Use adjusted R^2 when $n \leq 200$ and the largest VIF > 2, otherwise use AIC" increased the number of 80% correct within-one results by approximately 20 conditions better than the AIC results. This was the best of the several rules we tested.

We also learned that BIC performed very well with multiple non-zero population regression coefficients, but like other criteria suffered with small regression coefficients. Consequently, we are not

certain at this time how to use this result in a sample-based rule, since even zero population coefficients are likely to be small and non-zero in samples.

Certainly, our study is limited by the multivariate normal data, number of predictors, and multicollinearity patterns we included. We used only five predictors in these simulations. However, we varied the correlations of the predictors with the dependent variables broadly. Additional study certainly must be done to examine these results with more predictors. There is a good chance that the within-one correctness will decrease as additional predictors are added, especially as multicollinearity increases and becomes more complicated. But the rules may continue to work equivalent to the relative effectiveness found here.

Our study is also limited by the definition of "correct" model we used. We were able to determine that there was little difference in within-one correctness effectiveness for the selection criteria when we used a correctness rule with rounding to the nearest hundredth instead of nearest tenth. However, using the correlation or true zero regression coefficients impacted within-one correctness (as well as exact correctness) of the criteria.

We often hope or think that statistical methods will just give us "the" answer. Rarely is it that easy, especially in exploratory research. The best alternative may sometimes be to obtain exploratory information from multiple statistical methods to make theoretically reasonable decisions about relationships based on empirical data obtained from samples. For example, identifying within-one models may help us discover and make sense of more complicated relationships among the predictors in relation to the outcomes that interest us. Finally, with variable selection methods just like all statistical methods, researchers using regression must remember the importance of testing assumptions, checking for influential cases, using sufficient sample size, and cross-validating results.

References

Afifi, A. A., & Clark, V. (1990). *Computer-aided multivariate analysis* (2nd ed.). New York: Van Nostrand Reinhold.

Amemiya, T. (1980). Selection of regressors. *International Economic Review, 21*, 331-354.

Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics, 20*, 1-6.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics, 37*, 373-384.

Brooks, G. P., & Barcikowski, R. S. (2012). The PEAR method for sample sizes in multiple linear regression. *Multiple Linear Regression Viewpoints, 38*(2), 1-16.

Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology, 28*, 79-87.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society Series B, 45*, 311-354.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69*, 161-182.

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology, 45*(2), 265-282.

Ezekiel, M. (1930). *Methods of correlational analysis*. New York: Wiley.

Flack, V. F., & Chang, P. C. (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *American Statistician, 41*, 84-86.

Fox, J. (1991). *Regression diagnostics* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-079). Thousand Oaks, CA: Sage.

Gordon, R. A. (1968). Issues in multiple regression. *American Journal of Sociology, 73*, 592-616.

Harrell, F. E. (2015). *Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). New York: Springer.

Herzberg, P. A. (1969). The parameters of cross-validation. *Psychometrika Monograph Supplement*, 34 (2, Pt. 2).

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

Huberty, C. J. (1989). Problems with stepwise methods—better alternatives. In B. Thompson (Ed.), *Advances in social science methodology: A research annual* (Vol. 1, pp. 43-70). Greenwich, CT: JAI.

Hurvich, C. M., & Tsai, C. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika, 78,* 499-509.

Keith, T. Z. (2005). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling* (2nd ed.). New York: Taylor & Francis.

Leng, C. (2013). *The residual information criterion, corrected.* Retrieved from http://arxiv.org/abs/0711.1918v1

Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample* (Research Bulletin No. 50-40). Princeton, NJ: Educational Testing Service.

McQuarrie, A., Shumway, R., & Tsai, C. (1997). The model selection criterion AICu. *Statistics and Probability Letters, 34,* 285-292.

Mallows, C. L. (2000). Some comments on CP. *Technometrics, 42,* 87-94.

Nicholson, G. E. (1960). Prediction in future samples. In I. Olkin et al. (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 322-330). Palo Alto, CA: Stanford University.

Olejnik, S., Mills, J., & Keselman, H. (2000). Using Wherry's adjusted R^2 and Mallow's C_P for model selection from all possible regressions. *Journal of Experimental Education, 68,* 365-380.

Olkin, I., & Pratt, J. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29,* 201-211.

Park, C. N., & Dudycha, A. L. (1974). A cross-validation approach to sample size determination for regression models. *Journal of the American Statistical Association, 69,* 214-218.

Roecker, E. B. (1991). Prediction error and its estimation for subset-selected models. *Technometrics, 33,* 459-468.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Stein, C. (1960). Multiple regression. In I. Olkin et al. (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp.425-443). Palo Alto, CA: Stanford University.

Takezawa, K. (2014). *Learning regression analysis by simulation*. New York: Springer.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55,* 525-534.

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

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2,* 440-451.

Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin, 86,* 168-174.

Send correspondence to:
Gordon P. Brooks
Ohio University
Email: brooksg@ohio.edu