Comparing Stepwise Regression Models to the Best-Subsets Models, or, the Art of Stepwise

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

Stepwise regression methods are widely recognized as undesirable for explanatory purposes. As exploratory methods, however, they may provide efficient means for researchers to examine multiple models for further investigation. This study used Monte Carlo methods to examine the use of forward and backward regression without stopping criteria (i.e., with customized stepping criteria) in order to create all possible stepwise models (i.e., from one predictor to the fully specified model). Using these stepwise methods without stopping criteria often produced the same models as best-subset regression when there was little multicollinearity. Agreement was still good, but it declined, as multicollinearity became stronger. These results may help us understand the value of stepwise methods for exploratory model building and comparison. These results also suggest the need to examine models using multiple variable selection methods, because when they do not agree, they each may expose different aspects of the complicated theoretical relationships among predictors.

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). One might call this the art of science, for 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 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 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 a strategy 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 therefore 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.

Stepwise Regression Methods

Three commonly used variable selection methods in regression are forward selection, backward elimination, and stepwise regression, which are often collectively called stepwise methods (Keith, 2006). Stepwise methods process each regression model step-by-step by either adding or deleting one variable at a time based on stepping criteria (often p to enter, or PIN, and/or p to remove, or POUT). Details about stepwise methods can be found in most regression textbooks. We will focus on forward and backward stepwise methods in this paper.

Briefly, the forward selection process starts with no predictors in the model. In a common approach to forward regression, the first predictor chosen for entry into the model is the one with the largest simple correlation with the outcome variable if is considered to be statistically significant (i.e., PIN, or p-in, $\alpha = .05$). Later, each predictor added into the model has the highest part correlation (Stevens, 1999) or partial correlation (Meyers, Gamst, & Guarino, 2013) that is considered statistically significant (i.e., the

predictor contributing the largest statistically significant increase to R^2). By default, the procedure terminates when no predictor adds a statistically significant contribution to explained variation of the outcome variable (or all predictors have been entered). It should be noted that while many statistics programs use this R2 improvement (or partial F significance) approach, some programs process stepwise models based on other criteria, such as Akaike's Information Criterion (AIC). That is, some programs will continue to add predictors as long as the AIC, which is a function of prediction error, continues to decrease. Also, a tolerance inhibiting rule is often used to prevent the entry of highly correlated predictors into the regression model.

In the backward elimination procedure, the process starts with simultaneous entry of the full model (i.e., all predictors specified for the regression analysis). Next, each predictor deleted from the model has the smallest, non-significant part or partial correlation (i.e., POUT $\alpha = .10$), and therefore contributes the smallest amount to R^2. This predictor's removal will reduce R^2 the least from the larger, previous model. The default procedure terminates when all remaining predictors would produce a statistically significant reduction in R^2 and therefore none is removed (or all predictors have already been removed).

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, 2006). 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.

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. Additionally, the F statistic used for testing the significance of the steps in stepwise analyses is biased (Wilkinson, 1979) and may result in a greater likelihood of spurious statistical significance and inflated Type I error (Thompson, 1995). This is related to the fact that multiple hypothesis testing may result in inflated Type I error (Cohen, 1994; Krueger, 2001; Trafimow & Marks, 2015). Further, Fox (1991) indicated that reduced models produced through stepwise analyses are respecified models that may not address the research questions originally posed in a given study. Finally,

Wilkinson reported that "researchers encouraged by a significant multiple correlation from a stepwise analysis are often surprised to find how much it shrinks under cross-validation" (p. 168).

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). Indeed, a number of the same concerns over stepwise methods apply also to best-subsets regression.

Potential Usefulness of Variable Selection Methods

Most of the concerns raised by scholars about stepwise methods are based on concerns over its use in explanatory research. For example, Derksen and Keselman (1992) wrote, "stepwise methods were not designed to find 'best' models or to indicate the relative importance of variables" (p. 268). Further, many of the concerns are related to allowing a stopping rule, rather than theory, to choose which variables are included in the final recommended model. That is, many concerns are related to stopping the stepwise process based on statistical significance or some other statistical stopping criterion (e.g., PIN $\alpha = .05$ or POUT $\alpha = .10$, AIC increasing).

Some Recommendations for Thoughtful Use of Stepwise Methods

Some scholars have recommended processes for using stepwise methods in a reasonable manner. For example, Wilkinson (1979) indicated that cross-validation should be performed in lieu of statistical significance testing for stepwise models. That is, the results of stepwise 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 of stepwise regression 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 stepwise regression 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 stepwise analyses. 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 stepwise 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) has suggested guidelines for safer use of stepwise analysis. 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 therefore 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. Berk (1978) reported that stepwise algorithms provide better estimates of population parameters than does all-subsets regression (see also Olejnik, Mills, & Keselman, 2000). Similarly, Derksen and 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.

Purpose of the Study

We agree that stepwise methods used for explanatory purposes may be problematic for reasons addressed above. We argue, however, that using stepwise variable selection methods thoughtfully for exploratory model building and model comparison can be an efficient and useful methodology. Like all-subsets and best-subsets regression, stepwise regression techniques can allow researchers to compare multiple models. Therefore, we recommend using stepwise methods with no stopping criteria (or more precisely, customized stepping criteria) such that the maximum possible subset models are created (i.e., from one predictor to the fully specified model). Such an approach may trace back to Jennrich (1977), who described stepwise regression as a technique that, in the process of computing an ordinary least squares regression on p predictors, "obtains, at essentially no additional expense, p intermediate regressions which may provide useful insight about functional relations between Y and selected subsets of the total set of predictors" (p. 58). Such an approach considers it inappropriate to stop, or short-circuit, the process based on some algorithmic decision rule.

We could find no extant research that investigated the use of customized PIN and POUT stepping criteria to force the creation of the maximum number of models in stepwise methods. The primary goal is to use these methods to help build more regression models for comparison. If using forward and backward procedures without stopping criteria can result in producing the same models as best subset models for any given number of predictors, this may make the stepwise methods more acceptable for use from an exploratory perspective in multiple linear regression analyses. The "create-all-models" stepwise approach taken here, using either AIC or customized PIN/POUT criteria, does not rely on statistical significance in the decision processes used within stepwise regression, which addresses one concern raised by many scholars about stepwise methods. It should be noted, however, that this approach does not address all the concerns raised by scholars about the use of stepwise methods.

The create-all-models stepwise strategy does, however, allow researchers to visualize the dynamic and "interactive" process of entering all predictors into the model. That is, researchers using this approach can see the changes in the sign and magnitude of regression coefficients as new variables are entered into (or removed from) the model, which may illustrate complex relationships among predictors. Simply seeing static results of the final models, even best-subsets models, makes it more difficult (if not impossible) to see such changes and relationships.

Although the maximum number of stepwise models will be created when no stopping criterion is used, we could find no research that compared how well these stepwise models of various sizes agreed with the corresponding best-subset models of those sizes. For these reasons, this study investigated how frequently models created by forward selection and backward elimination methods without stopping criteria matched the best-subsets models for a given model size (i.e., number of predictors) when using an all-possible (i.e., exhaustive) regressions strategy. It should be noted that we did not study how frequently either approach identified the correct model. 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 et al., 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 the AIC criterion rather than the partial correlation criterion described above. In typical, default forward regression with AIC used as a stopping criterion, predictors are added so that AIC will be reduced the most at each step. Then when no predictors will reduce AIC at a given step, the procedure stops. Similarly, with backward regression, the process stops when the minimum AIC has been reached while removing predictors (that is, the procedure stops before removing a predictor that increases AIC). However, we used the AIC criterion with no stopping criterion for predictor selection in both the forward

and backward stepwise regression methods. That is, in forward regression, predictors were added at each step that reduced AIC the most or, when no predictors would reduce AIC, that increased AIC the least. In this way, models of all possible sizes were created and the final model was always the full model, not a reduced model. For example, with five predictors, we would create a one-predictor model, a two-predictor model, and so forth, up to the maximum five-predictor model. In backward regression, predictors were removed that, while included in the model before removal, increased AIC the most or decreased it the least, ending with a single predictor as the final model. In all possible regressions, the best subset of each possible size was determined by the smallest AIC.

In programs such as SPSS, which use partial correlation (or R^2 change) as the statistical criterion, this process is functionally equivalent to using a PIN $\alpha = .999$ (with POUT $\alpha = 1.000$) in the forward selection procedure. This customized PIN stepping criterion results in most, if not all, predictors adding a statistically significant (i.e., $p < .999$) amount of explained variation to R^2. Therefore, all predictors are usually eventually entered into the forward selection process – such that the final model will often include all predictors specified for the analysis. In the backward elimination procedure in SPSS, the POUT $\alpha = .002$ (with PIN $\alpha = .001$) is used as the customized POUT stepping criterion for predictors to be removed from the model. Consequently, all predictors are usually removed from the model because they do not always explain a statistically significant (i.e., $p < .002$) change in explained variation in the outcome. However, because of the limitations of these statistical significance tests (e.g., there is no guarantee that all models from 1 to k predictors will be created), using AIC as the criterion enabled us to guarantee that models of all possible sizes were created through the forward and backward processes. Yamashita, Yamashita, and Kamimura (2007) have shown "that the stepwise AIC method and the stepwise methods using Partial *F*, Partial Correlation, and Semi-Partial Correlation lead to the same method as Partial *F*" (p. 2403).

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 stepwise 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 small (i.e., $n = 50$) and large (i.e., $n = 250$) sample sizes 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 forward selection, backward elimination, and exhaustive (all possible) regressions approaches.

The forward and backward models were compared to the best subsets from exhaustive regressions. That is, in each sample, the one predictor models from forward and backward were compared to the one predictor model identified as the best (lowest AIC) single predictor model from the exhaustive approach. Similarly, the two-predictor models from forward and backward regression were compared to the best two-predictor model from all possible regressions, and so forth, up to the five-predictor models created from all three regression approaches. Information was collected from each sample about the agreement of the three methods. That is, data were stored about whether the three approaches matched (i.e., produced five models with the same predictors). We also collected data about how frequently forward models matched backward, forward matched exhaustive, and backward matched exhaustive.

For example, Table 1 shows example results where the three methods do not agree within a single sample. It is evident from the table that the data from this particular sample produced different results for

Table 1. An Example of Which Predictors are Included in Best-Subset Results from Best-Subset Exhaustive Regressions, Forward Regression, and Backward Regression

Number of Predictors in Model	Predictor Variable in the Model				
	X1	X2	X3	X4	X5
1					E, B, F
2			B	E, F	E, B, F
3		E, B, F	E, B	F	E, B, F
4	E, B	E, B, F	E, B, F	F	E, B, F
5	E, B, F	E, B, F	E, B, F	E, B, F	E, B, F

Note. E indicates that a given predictor was in the model of that size for best-subset exhaustive regression, *F* represents forward section, and B represents backward elimination.

all three regression approaches. That is, all three methods of course produced the same five-predictor model, and also produced the same one-predictor model (X5 alone). However, the best subset from exhaustive regressions for two predictors was X4 and X5, which matched forward but not backward, which identified X3 and X5 as the best two-predictor model. Both backward and exhaustive regressions resulted in the same four-predictor model, but forward differed. Finally, the best-subset results show that one predictor (X4) was added or removed in ways that would not be permitted by forward or backward regression.

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. 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 as our ceiling. For example, with no multicollinearity the R^2 is simply the sum of the squared predictor correlations with the outcome (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 would not be 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 2 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 $r = .6$). Therefore, 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).

Results

The results show that, in general, stepwise methods (i.e., forward and backward) match the exhaustive (i.e., all possible) regressions quite well when there is no multicollinearity. That is, the forward, backward, and exhaustive methods generally produced the same one-predictor, two-predictor, three-predictor, four-predictor, and five-predictor best-subset models when there was little or no multicollinearity (M0 and M1). There are progressively more model mismatches as multicollinearity increased to M2, which had relatively low multicollinearity, and to M3 and M4, which had moderate but more complicated multicollinearity (e.g., largest VIF was 4.6).

Table 2. 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	X2	X3	X4	X5	VIF
M3	X1	.8	.6	.4	.2	3.0
	X2		.6	.4	.2	3.0
	X3			.4	.2	1.7
	X4				.2	1.3
M4	X1	.8	.8	.4	.2	3.5
	X2		.8	.4	.2	3.5
	X3			.6	.2	4.6
	X4				.2	1.6

Note. For both M3 and M4, the VIF for X5 is 1.1.

Table 3, which shows only those predictor-outcome correlation patterns for which the three variable selection methods did not agree in at least 95% of the samples, shows that when there is no multicollinearity, six patterns of predictor-outcome correlations resulted in model agreement in fewer than 90% of the samples for $n = 50$. That is, when there was no multicollinearity, 45 of the 51 (88.2%) correlation-pattern conditions resulted in agreement across all three regression approaches (F=B=E) in over 90% of the samples. For example, Table 3 shows that for the predictor-outcome correlation pattern (.1, .2, .3, .4, .5), all three methods (i.e., forward, backward, and exhaustive) produced exactly the same five subset models for the associated number of predictors in a total of 9,474 (94.7%) samples when $n = 50$ (the "All" column, also called F=B=E below). Forward and backward produced the same five subset models in 9,479 of the 10,000 samples (F=B), forward and exhaustive produced the same subset models in 9,475 samples (F=E), and backward and exhaustive produced the same five subset models in 9,613 samples (B=E). Further, the lowest level of agreement was 70% in just one condition, while all others were above 82% agreement. The increase in sample size from 50 to 250 improved the number of matches in all conditions except one: the (.4, .4, .4, .4, .4) condition, which decreased from F=B=E in 7,071 samples when $n = 50$ to 7,035 samples when $n = 250$. With $n = 250$, 44 of the 51 (86.3%) conditions resulted in over 95% agreement of the regression variable selection approaches.

We provide Table 3 as an example of the results we examined. However, for the other multicollinearity conditions, we will summarize the results a little differently because the disagreement increased as multicollinearity increased. Otherwise, the tables associated with M2, M3, and M4 would be too long and detailed to be useful summaries. Further, we will only report the agreement among all three methods (F=B=E) because we found that the associations among F=B, F=E, B=E, and F=B=E agreement levels over the 51 correlation-pattern conditions were extremely high for all multicollinearity conditions. For example, for the 51 correlation patterns in the $n = 50$ sample size conditions, the correlation between the F=B and F=B=E was $r > .999$. That is, the numbers of samples with method agreement for F=B and the number of samples with method agreement such that F=B=E showed strong correlation across the 51 predictor-outcome conditions. Even as multicollinearity increased, the lowest such correlation for any pair of method-agreement frequencies was $r = .952$. These strong correlations suggested that in conditions where any two of the regression methods (i.e., either F=B, F=E, or B=E) have high levels of agreement, all three methods tend to have high levels of subset-model agreement.

Further analysis confirmed that, beyond large correlations, actual agreement among the three methods was generally high, regardless of the size of the correlation. For example, in the (.1, .2, .3, .4, .5) correlation-pattern condition (see Table 3), the 9,613 samples where B=E were the same 9,474 samples where F=B=E, but with only 139 additional samples that matched between backward and exhaustive but not forward. But in only one sample did forward and backward match where they did not also match exhaustive. Under the M4 multicollinearity condition, one example predictor-outcome correlation pattern resulted in F=B=E agreement in 6,413 samples, but F=E agreement in 7,083 samples. Results showed that of the 3,587 samples that lacked complete agreement (i.e., not F=B=E, which was 10,000-6,413=3,587), over 75% of those samples actually showed complete disagreement (i.e., F≠B, F≠E, and B≠E in 2,693 of

Table 3. The Number of Matches Out of 10,000 Where Agreement Was Lower Than 95% Among Methods for No Multicollinearity (M0) Sorted By n = 50 When All Three Methods Match

Correlation Pattern					Sample Size Condition							
					n = 50				*n* = 250			
r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	F=B	F=E	B=E	All	F=B	F=E	B=E	All
0.4	0.4	0.4	0.4	0.4	7236	7368	7460	7071	7172	7335	7429	7035
0.2	0.4	0.4	0.4	0.4	8288	8255	8591	8223	8681	8649	8935	8649
0.2	0.4	0.4	0.4	0.6	8602	8564	8751	8564	9227	9226	9229	9226
0.0	0.4	0.4	0.4	0.4	8619	8576	8885	8575	8684	8661	8942	8661
0.0	0.4	0.4	0.4	0.6	8824	8793	8963	8793	9190	9188	9192	9188
0.2	0.2	0.4	0.4	0.4	8983	8975	9217	8958	9622	9622	9740	9622
0.2	0.2	0.2	0.2	0.8	9154	9133	9133	9133	9204	9186	9186	9186
0.2	0.2	0.2	0.4	0.4	9255	9253	9377	9241	9837	9838	9848	9837
0.2	0.2	0.2	0.2	0.2	9290	9345	9418	9277	9295	9361	9395	9279
0.2	0.2	0.2	0.2	0.4	9313	9326	9407	9299	9675	9672	9676	9671
0.0	0.2	0.4	0.4	0.4	9317	9307	9494	9302	9611	9611	9747	9611
0.2	0.2	0.4	0.4	0.6	9447	9436	9540	9436	9995	9995	9998	9995
0.0	0.2	0.2	0.2	0.2	9445	9462	9556	9437	9726	9724	9800	9723
0.2	0.2	0.2	0.2	0.6	9469	9460	9475	9459	9551	9540	9540	9540
0.1	0.2	0.3	0.4	0.5	9479	9475	9613	9474	9985	9985	9995	9985

Note. F=B indicates the number of samples in which the forward selection method and the backward elimination method agreed for all five models created (F=E for forward and exhaustive, B=E for backward and exhaustive, and All for agreement all three methods). The column label r_{12} indicates the correlation between the outcome (variable 1) and the first predictor (variable 2). Similarly, r_{16} represents the correlation between the outcome and the fifth predictor (variable 6).

the 3,587 samples, but exactly two of the three variable selection method comparisons agreed in 894 samples). This level of complete-disagreement-if-any-disagreement was maintained across most of the conditions examined, and was generally over 80%, or even 90%. Therefore, in addition to levels of agreement being high for all method comparisons (i.e., when any agreed all tended to agree), the level of disagreement of any two methods was generally associated with complete disagreement among all comparisons.

Table 4, below, shows all predictor-outcome correlation patterns and the number of samples in which all three variable selection methods agreed (i.e., F=B=E). From Table 4, we can see that the results for multicollinearity pattern M1, where all predictor correlations are $r = .2$, show 48 of 51 (94.1%) predictor-outcome correlation patterns, when $n = 50$, that result in the number of matches (i.e., level of agreement) over 90%, and 27 (52.9%) conditions with F=B=E agreement over 95%. The increase in sample size from 50 to 250 increased the number of method matches in all the M1 correlation pattern conditions. For $n = 250$, 46 of the 51 (90.2%) correlation conditions resulted in at least 95% agreement. Like the M0 and M1 conditions, M2 agreement levels were generally above 80%, with only one correlation pattern below 70% agreement and three below 80%. There were fewer predictor-outcome correlation conditions, however, where agreement exceeded 90% (38 or 74.5%) or 95% agreement (14 or 27.5%) in the 51 M2 $n = 50$ conditions. A similar pattern held for $n = 250$, where 44 (86.3%) conditions were over 90% and 34 (66.7%) conditions were over 95%.

When the multicollinearity is higher and more complicated, even fewer correlation patterns resulted in agreement over 90%. Table 4 shows that with $n = 50$ only 11 (21.6%) patterns resulted in the number of matches over 90% for M3 and only 4 (7.8%) for M4. Additionally, for M3 there are 41 (80.4%) conditions with levels of agreement over 70% in either the $n = 50$ or $n = 250$ condition, or both, and 36 (70.6%) conditions with levels of agreement over 80%. For M4, only 32 (62.7%) of the correlation patterns resulted in levels of agreement greater than 70% and only 18 (35.3%) above 80% agreement (indeed, nine of the correlation patterns resulted in agreement less than 60%).

Table 4. The Number of Matches for Each Predictor-Outcome Correlation Pattern for the F=B=E Results Where All Three Variable Selection Methods Produced the Same Five Subset Models

					Multicollinearity Condition									
Correlation Pattern					M0		M1		M2		M3		M4	
r12	r13	r14	r15	r16	n=50	n=250	n=50	n=250	n=50	n=250	n=50	n=250	n=50	n=250
0	0	0	0	0.2	***9590***	***9952***	*9418*	***9883***	*9199*	***9880***	8236	*9045*	7676	8573
0	0	0	0.2	0.2	***9591***	***9968***	*9373*	***9861***	*9137*	*9418*	8336	*9307*	7506	8274
0	0	0.2	0.2	0.2	***9591***	***9908***	*9318*	***9662***	*9030*	*9099*	7925	7013	5853	3434
0	0.2	0.2	0.2	0.2	*9437*	***9723***	*9272*	***9572***	*9121*	*9220*	6422	3555	5990	4124
0.2	0.2	0.2	0.2	0.2	*9277*	*9279*	*9381*	***9730***	*9282*	***9776***	8440	*9210*	7830	8566
0	0	0	0	0.4	***9753***	***9964***	***9625***	***9931***	***9604***	***9883***	8811	*9317*	8239	8972
0	0	0	0.2	0.4	***9734***	***9986***	***9533***	***9813***	*9322*	***9554***	8746	***9669***	7987	8199
0	0	0.2	0.2	0.4	***9689***	***9990***	*9447*	***9702***	*9235*	***9606***	8261	8093	6036	5962
0	0.2	0.2	0.2	0.4	***9537***	***9884***	*9444*	***9701***	*9289*	***9904***	6639	5039	6034	5919
0.2	0.2	0.2	0.2	0.4	*9299*	***9671***	***9574***	***9906***	*9469*	***9899***	8748	***9543***	8153	*9042*
0	0	0	0.4	0.4	***9870***	***9989***	***9574***	***9961***	8710	*9129*	8873	***9756***	8104	8967
0	0	0.2	0.4	0.4	***9738***	***9997***	*9463*	***9902***	8817	***9584***	8457	*9433*	6713	8577
0	0.2	0.2	0.4	0.4	***9540***	***9985***	*9422*	***9936***	*9174*	***9873***	6810	8015	6304	7827
0.2	0.2	0.2	0.4	0.4	*9241*	***9837***	***9566***	***9971***	*9487*	***9968***	8883	***9608***	8508	***9882***
0	0	0.4	0.4	0.4	***9595***	***9615***	*9069*	***9598***	7986	7946	6224	4297	3147	2596
0	0.2	0.4	0.4	0.4	*9302*	***9611***	*9178*	***9709***	8706	8796	6514	5095	3715	3089
0.2	0.2	0.4	0.4	0.4	8958	***9622***	*9391*	***9786***	*9279*	***9783***	8608	*9405*	6518	5653
0	0.4	0.4	0.4	0.4	8575	8661	8607	8817	8076	7797	3021	2136	3551	3482
0.2	0.4	0.4	0.4	0.4	8223	8649	*9066*	*9299*	*9150*	*9371*	6413	4402	6481	6058
0.4	0.4	0.4	0.4	0.4	7071	7035	8622	8664	*9183*	*9320*	8422	8362	7906	7459
0	0	0	0	0.6	***9798***	***9964***	***9719***	***9903***	***9592***	***9648***	*9085*	***9589***	8668	*9082*
0	0	0	0.2	0.6	***9839***	***9991***	***9663***	***9668***	***9544***	***9759***	*9069*	*9300*	8160	6317
0	0	0.2	0.2	0.6	***9827***	***9997***	***9623***	***9651***	***9682***	***9986***	8830	8629	7501	8080
0	0.2	0.2	0.2	0.6	***9752***	***9878***	***9688***	***9794***	***9767***	***9971***	7766	7420	7711	8320
0.2	0.2	0.2	0.2	0.6	*9459*	***9540***	***9731***	***9941***	***9644***	***9906***	8977	*9421*	8520	*9015*
0	0	0	0.4	0.6	***9955***	***9996***	***9614***	***9906***	8459	8362	*9281*	***9731***	8473	***9816***
0	0	0.2	0.4	0.6	***9924***	***10000***	***9668***	***9965***	*9107*	*9080*	*9022*	***9506***	7146	8422
0	0.2	0.2	0.4	0.6	***9832***	***10000***	***9746***	***9991***	***9626***	***10000***	7802	8764	7090	8480
0.2	0.2	0.2	0.4	0.6	***9571***	***9855***	***9825***	***9981***	***9583***	***9941***	*9209*	***9638***	*9292*	***9963***
0	0	0.4	0.4	0.6	***9790***	***9997***	*9105*	***9567***	8435	8318	6395	4746	2946	2037
0	0.2	0.4	0.4	0.6	***9624***	***9995***	*9421*	***9855***	*9214*	***9854***	7106	5642	3855	2661
0.2	0.2	0.4	0.4	0.6	*9436*	***9995***	***9706***	***9997***	*9460*	***9575***	8947	***9765***	6517	5184
0	0.4	0.4	0.4	0.6	8793	*9188*	8829	8970	8909	***9570***	2603	1642	3127	2147
0.2	0.4	0.4	0.4	0.6	8564	*9226*	*9385*	***9830***	*9349*	*9460*	6608	4406	6530	5931
0	0	0	0.6	0.6	***9946***	***9997***	***9754***	***9839***	7767	*9136*	*9293*	*9455*	8726	*9405*
0	0	0.2	0.6	0.6	***9968***	***10000***	***9905***	***9997***	8803	***9849***	*9499*	***9938***	***9660***	***10000***
0	0.2	0.2	0.6	0.6	***9958***	***10000***	***9948***	***10000***	***9528***	***9978***	***9836***	***10000***	***9521***	***9992***
0.2	0.2	0.2	0.6	0.6	***9641***	***9660***	***9922***	***9976***	***9816***	***9940***	***9699***	***9901***	***9905***	***10000***
0	0	0.4	0.6	0.6	***9749***	***9996***	*9307*	*9399*	8398	***9827***	7238	7978	800	9
0	0	0	0	0.8	***9834***	***9959***	***9649***	***9666***	8284	8287	*9247*	***9566***	8895	8564
0	0	0	0.2	0.8	***9886***	***9990***	***9555***	***9593***	*9376*	*9394*	*9075*	8988	7470	4423
0	0	0.2	0.2	0.8	***9917***	***10000***	***9654***	***9820***	***9982***	***10000***	8834	8916	8190	8498
0	0.2	0.2	0.2	0.8	***9679***	***9780***	***9736***	***9969***	***9910***	***9987***	8546	***9688***	8743	***9834***
0.2	0.2	0.2	0.2	0.8	*9133*	*9186*	***9736***	***9936***	***9718***	***9855***	8961	*9269*	8434	8839
0	0	0	0.4	0.8	***9932***	***9989***	*9160*	***9579***	6892	6483	8879	*9303*	8012	*9382*
0	0	0.2	0.4	0.8	***9996***	***10000***	***9588***	***9923***	*9490*	***9927***	8994	*9418*	7019	8091
0	0.2	0.2	0.4	0.8	***9957***	***10000***	***9845***	***10000***	***9901***	***10000***	8408	*9294*	7394	*9133*
0.1	0.2	0.3	0.4	0.5	*9474*	***9985***	***9563***	***9974***	*9370*	***9893***	8420	*9354*	6551	4983
0	0.1	0.2	0.3	0.4	***9615***	***9986***	*9484*	***9927***	*9201*	***9861***	8113	8863	6216	3505
0	0	0.1	0.2	0.3	***9646***	***9983***	*9442*	***9899***	*9131*	***9679***	8370	*9217*	7346	7368
0	0	0	0.1	0.2	***9602***	***9958***	*9349*	***9852***	*9226*	***9728***	8298	*9110*	7452	8234

Note. *With frequencies above 95% marked in bold italics and those above 90% in italics.*

Conclusions

To summarize, using the customized stepping criteria, in order to create all possible stepwise models, generally resulted in the same stepwise models as the best-subset approach to variable selection when multicollinearity is not present or is relatively low. It is important to note that our results do not support the use of stepwise methods with default stepping and stopping criteria because we did not study that. Our results simply suggest that stepwise methods without stopping criteria generally produce the same subset models as best-subsets regression except for conditions with unusual correlation and multicollinearity patterns. This may be useful to researchers who use SPSS, for example, which does not provide a native all-possible regressions or best-subsets regression procedure.

However, the methods begin to produce different sets of subset models as correlation among the predictors increases (honestly, it may not be appropriate to use the term multicollinearity in this study because none of the VIF values were over 5.0). Table 5 summarizes the number of the 51 predictor-outcome correlation patterns where all three regression approaches matched (i.e., F=B=E or All) across all multicollinearity patterns. Clearly, as multicollinearity increases, the levels of agreement across the variable selection methods decrease. This finding argues strongly for the need to examine results from multiple variable selection methods, especially when performing exploratory analyses. Each approach may bring focus to a different aspect of the complicated and idiosyncratic patterns of correlations in the data.

None of the predictor-outcome correlation patterns consistently produced the highest levels of agreement across all multicollinearity conditions, but 13 conditions resulted in most agreement regardless of sample sizes across most of the multicollinearity conditions. Results shown in Table 6 show the patterns with agreement over 90% across the most multicollinearity and sample size conditions. Interestingly, while some of these patterns resulted in the highest agreement only for the low multicollinearity conditions (e.g., the pattern 0, 0, 0, .2, .6), some produced high agreement even for high multicollinearity or across varied multicollinearity conditions (e.g., the pattern 0, .2, .2, .6, .6). Table 6 shows that the highest levels of agreement tend to occur when the predictor-outcome correlations combine .2 correlations with larger correlations.

Seven conditions, shown in Table 7, showed the least agreement across the variable selection methods for the conditions studied (i.e., lower than 80% agreement across the most predictor-outcome conditions). Table 7 shows that the predictor-outcome correlations that are largely in the .4 range resulted in the lowest agreement across methods.

Another interesting finding is that as multicollinearity increases and gets a bit more complicated, the forward selection and backward elimination methods tended to agree most frequently (e.g., for at least 30 predictor-outcome correlation patterns in M4) while with no or low multicollinearity, backward and exhaustive best-subsets agreed most frequently (e.g., 40 patterns for $n = 50$ and 26 for $n = 250$ under M0). Table 8 shows the gradual shift in these trends.

Generally an increase in sample size helps the three methods agree on which predictors are in each of the five models. Under M0 and M1 multicollinearity conditions, increased sample size improved agreement across variable selection methods. Even as predictor correlations increased further, however, increased sample size generally increased agreement levels. For example, 27.5% of the models agreed across all three methods for M2 when $n = 50$, but 66.7% of the models agreed for $n = 250$ (see Table 5). Based on the complete results from which Table 3 is a subset, Table 8 shows that as sample size increased to $n = 250$, all three variable selection methods tended to agree more frequently across the 51 predictor-outcome correlation patterns. For example, in M4, F=B decreased from 35 conditions with better agreement at $n = 50$ to 30 conditions at $n = 250$, but because F=B when F=B=E there was actually a net increase of conditions where F=B from 35 to 45 (so overall, F=B had higher agreement across more conditions with larger sample size).

We found, however, that there were specific conditions for which an increase in sample size did not always improve the number of matches, especially when multicollinearity is more worrisome. When all predictor correlations increased to $r = .4$ (the M2 condition), six predictor-outcome correlation patterns resulted in lower agreement for larger sample sizes (some of these can be seen in Table 7). There were 16

Table 5. The Percentage of Correlation Patterns (51 in Total) When All Three Regression Approaches Match Under Different Threshold Levels of Agreement in All Multicollinearity Patterns

Level of Agreement	Multicollinearity Patterns									
	M0		M1		M2		M3		M4	
	n=50	*n*=250	*n*=50	*n*=250	*n*=50	*n*=250	*n*=50	*n*=250	*n*=50	*n*=250
> 70%	100.0	100.0	100.0	100.0	98.0	98.0	80.4	80.4	62.7	62.7
> 80%	98.0	98.0	100.0	100.0	94.1	94.1	70.6	74.5	35.3	56.9
> 90%	88.2	94.1	94.1	94.1	74.5	86.3	21.6	58.8	7.8	25.5
> 95%	70.6	86.3	52.9	90.2	27.5	66.7	3.9	27.5	5.9	13.7
Average	95.2	97.5	94.8	97.6	91.5	94.7	81.0	81.7	70.2	70.7

Note. Average represents the average number of samples (out of 10,000) in which the three variable selection methods agreed across the 51 predictor-outcome correlation conditions.

Table 6. The Predictor-Outcome Conditions that Resulted in the Highest Levels of Agreement for the Variable Selection Methods for the Most Sample Size and Multicollinearity Conditions with Non-Zero Multicollinearity

Correlation Pattern					Multicollinearity and Sample Size Condition							
					M1		M2		M3		M4	
r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	*n*=50	*n*=250	*n*=50	*n*=250	*n*=50	*n*=250	*n*=50	*n*=250
0.2	0.2	0.2	0.2	0.4	9574	9906	9469	9899	8748	9543	8153	9042
0.2	0.2	0.2	0.4	0.4	9566	9971	9487	9968	8883	9608	8508	9882
0.0	0.0	0.0	0.0	0.6	9719	9903	9592	9648	9085	9589	8668	9082
0.0	0.0	0.0	0.2	0.6	9663	9668	9544	9759	9069	9300	8160	6317
0.2	0.2	0.2	0.2	0.6	9731	9941	9644	9906	8977	9421	8520	9015
0.0	0.0	0.2	0.4	0.6	9668	9965	9107	9080	9022	9506	7146	8422
0.2	0.2	0.2	0.4	0.6	9825	9981	9583	9941	9209	9638	9292	9963
0.0	0.0	0.0	0.6	0.6	9754	9839	7767	9136	9293	9455	8726	9405
0.0	0.0	0.2	0.6	0.6	9905	9997	8803	9849	9499	9938	9660	10000
0.0	0.2	0.2	0.6	0.6	9948	10000	9528	9978	9836	10000	9521	9992
0.2	0.2	0.2	0.6	0.6	9922	9976	9816	9940	9699	9901	9905	10000
0.0	0.2	0.2	0.2	0.8	9736	9969	9910	9987	8546	9688	8743	9834
0.0	0.2	0.2	0.4	0.8	9845	10000	9901	10000	8408	9294	7394	9133

Note. The column label r_{12} indicates the correlation between the outcome (variable 1) and the first predictor (variable 2). This follows similarly for the other column labels through r16, which represents the correlation between the outcome and the fifth predictor (variable 6).

of the M3 correlation conditions, however, that had lower agreement for the larger sample size, and of those, eight had agreement levels with relatively large differences for larger *n* (over 1000 samples with less agreement as compared to smaller *n*). Also in the M4 condition, 8 of the 22 predictor-outcome correlation conditions with lower agreement at $n = 250$ were relatively large differences. Because the patterns of predictor correlations differed across these multicollinearity conditions, there was no consistency for which predictor-outcome correlation patterns resulted in the decreases.

Another interesting result is that as multicollinearity increases, the change in variable selection method agreement was more varied in how it improves and worsens the agreement levels. That is, for M0 and M1, the additional number of samples that showed agreement due to sample size was rarely above 500 samples. For M2 the change in the number of samples with agreement across conditions varied from roughly -400 to +1400 as sample size changed from $n = 50$ to $n = 250$. For M3 and M4, however, the change in agreement varied from decreases of roughly 3,000 samples to increases over 1,500 when sample size changed from $n = 50$ to $n = 250$. It appears that more complicated patterns of correlations among predictors, in combination with the given predictor-outcome correlation patterns, had complicated impacts on the ability of the variable selection methods to produce the same subset models (due probably to smaller standard errors as sample size increased).

Table 7. The Predictor-Outcome Conditions that Resulted in the Lowest Levels of Agreement for the Variable Selection Methods for the Most Sample Size and Multicollinearity Conditions with Non-Zero Multicollinearity

Correlation Pattern					Multicollinearity and Sample Size Condition							
					M1		M2		M3		M4	
r12	r13	r14	r15	r16	*n*=50	*n*=250	*n*=50	*n*=250	*n*=50	*n*=250	*n*=50	*n*=250
0	0	0.4	0.4	0.4	9069	9598	7986	7946	6224	4297	3147	2596
0	0.2	0.4	0.4	0.4	9178	9709	8706	8796	6514	5095	3715	3089
0	0.4	0.4	0.4	0.4	8607	8817	8076	7797	3021	2136	3551	3482
0.2	0.4	0.4	0.4	0.4	9066	9299	9150	9371	6413	4402	6481	6058
0.4	0.4	0.4	0.4	0.4	8622	8664	9183	9320	8422	8362	7906	7459
0	0	0.4	0.4	0.6	9105	9567	8435	8318	6395	4746	2946	2037
0	0.4	0.4	0.4	0.6	8829	8970	8909	9570	2603	1642	3127	2147

Note. The column label r12 indicates the correlation between the outcome (variable 1) and the first predictor (variable 2). This follows similarly for the other column labels through r16, which represents the correlation between the outcome and the fifth predictor (variable 6).

Table 8. The Number of Predictor-Outcome Correlation Patterns (51 in Total) for Which Agreement Among Variable Selection Methods was Highest

Method	Multicollinearity Patterns									
	M0		M1		M2		M3		M4	
Agreement	*n*=50	*n*=250	*n*=50	*n*=250	*n*=50	*n*=250	*n*=50	*n*=250	*n*=50	*n*=250
F=B	7	3	15	16	24	20	24	28	35	30
F=E	0	0	2	2	8	4	19	9	9	4
B=E	40	26	33	16	14	7	8	5	7	2
F=B=E	4	22	1	17	5	20	0	9	0	15

Note. F=B represents when Forward equaled Backward the most, F=E for when Forward was the same as Exhaustive most frequently, and B=E represents when Backward matched Exhaustive most often. F=B=E indicates when all were equal with the same frequency for that predictor-outcome correlation pattern.

Discussion

Our results suggest that researchers might be relatively confident using stepwise methods with customized stepping criteria (i.e., no stopping criteria) when they do not have access to all-possible regressions procedures to obtain best-subsets models and they have relatively low multicollinearity. With lower multicollinearity, the three variable selection approaches tended to produce the same models at all possible numbers of predictors at a relatively high rate. That is, all three approaches (forward, backward, exhaustive) frequently produced the same one-predictor, two-predictor, three-predictor, four-predictor, and five-predictor models in the 10,000 samples that were tested for each condition. When multicollinearity is higher and the pattern of correlations is more complicated, however, there is less model agreement among forward, backward, and exhaustive best-subset methods. This result brings further concern to the use of any of the three approaches alone to identify best-subset models, even when using customized stepping criteria that result in all stepwise models being built. Perhaps a better way to think about it is that these results strongly suggest the need to run all approaches to build all possible competing models, because they disagree so frequently when multicollinearity becomes stronger. Each method may bring to light different dynamics of the relationships among the variables in models, especially the relationships among the predictors, which may be theoretically interesting.

Certainly, our study is limited by the multivariate normal data, number of predictors, and types of 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. There is little reason to believe that additional predictors will result in better agreement across the methods, particularly as multicollinearity increases and becomes more complicated. Indeed, some informal Monte Carlo analyses were run with 1,000 replications, nine predictors, $n = 50$, and predictor-outcome correlations created in a

similar way to this study. Only 5 of 83 predictor-outcome correlation patterns showed F=B=E method agreement for more than 90% of the samples in the no multicollinearity condition. On average, the variable selection methods agreed in 76.8% of the M0 samples for nine predictors, as compared to 95.2% with five predictors (see Table 5).

Further, the lack of agreement across methods even with the relatively low multicollinearity studied here (i.e., no VIF over 5) suggests that researchers must exercise extreme caution when building models with more severe multicollinearity. Further, it was also clear that larger sample size was not always a cure for the multicollinearity problem. Indeed, larger samples sometimes made the variable selection approaches agree at a lower rate, depending on the idiosyncrasies of the correlations. While we only reported results for $n = 50$ and $n = 250$, we also ran simulations with $n = 100$, $n = 150$, and $n = 200$. Given what we learned from the $n = 50$ and $n = 250$ conditions, the results from the other sample size conditions tended to be as expected in a relatively linear fashion across the multicollinearity conditions.

It is important to note again that we did not investigate whether these approaches are able to identify the correct population model. Agreement among the methods does not guarantee that the correct regression model has been identified, but disagreement among these notoriously imperfect methods certainly does not provide a researcher much confidence in any model chosen. Previous research has suggested that none of the three approaches with default criteria (i.e., forward, backward, exhaustive) is able to identify the correct model consistently. Perhaps the best course of action is to recognize the exploratory nature of these methods and simply conclude that a number of models may be worth exploring in future research based on the researcher's best understanding of the variables being studied. Running more possible models for comparison, using the create-all-models stepwise methods presented here, allows researchers to compare empirical evidence for more candidate models. Researchers can then artistically bring their content and theoretical knowledge to bear on decisions regarding the usefulness of the various candidate models.

Sometimes we want 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, rather than argue over whether stepwise methods or best-subsets produce better results or whether they produce unworthy results, we can recognize that they all analyze the correlations among variables (and perhaps more importantly, partial correlations) and may produce different useful and interesting results. Therefore, we can run multiple methods to examine and compare more complete exploratory results. That is, perhaps running all three types of variable selection techniques will help us discover and make sense of more complicated relationships among the predictors in relation to the outcomes that interest us.

Most of the predictor-outcome correlations patterns across all levels of multicollinearity showed agreement levels greater than 70%, but perhaps it may be just as interesting theoretically when the three variable selection methods do not agree on the same models (certainly less confirmatory, but perhaps more interesting from an exploratory perspective). For example, the dynamic nature of stepwise approaches may help researchers understand the complicated relationships among predictors more completely – rather than simply throwing away a nonsignificant predictor. That is, it could be the multicollinearity itself that it theoretically interesting. Notably, here, either forward or backward (or both) always produced the same best-subsets models as exhaustive under the conditions we studied (see Table 8), even under higher multicollinearity. If future research supports this finding with more predictors and more multicollinearity, it may be sufficient to run just forward and backward methods without also running best-subsets in order to obtain attractive models for comparison. That is, either forward or backward or both always provided the same set of models as best-subsets regression in this study. 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.

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

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.
Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.
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.
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.
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.
Jennrich, R. I. (1977). Stepwise regression. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Mathematical methods for digital computers. Volume III: Statistical methods for digital computers* (pp. 58-75). New York: Wiley-Interscience.
Keith, T. Z. (2006). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling* (2nd ed.). New York: Taylor & Francis.
Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56*(1), 16-26.
Meyers, L. S., Gamst, G., & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation* (2nd ed.). Thousand Oaks, CA: Sage.
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. *The Journal of Experimental Education, 68*(4), 365-380.
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.
Stevens, J. P. (1999). *Intermediate statistics: A modern approach* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55,* 525-534.
Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*, 1-2.
Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York: John Wiley & Sons.
Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin, 86*, 168-174.
Yamashita, Y., Yamashita, K., & Kamimura, R. (2007). A stepwise AIC method for variable selection in linear regression. *Communications in Statistics – Theory and Methods, 36,* 2395-2403.

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