United States
Department
of Agriculture

Forest Service

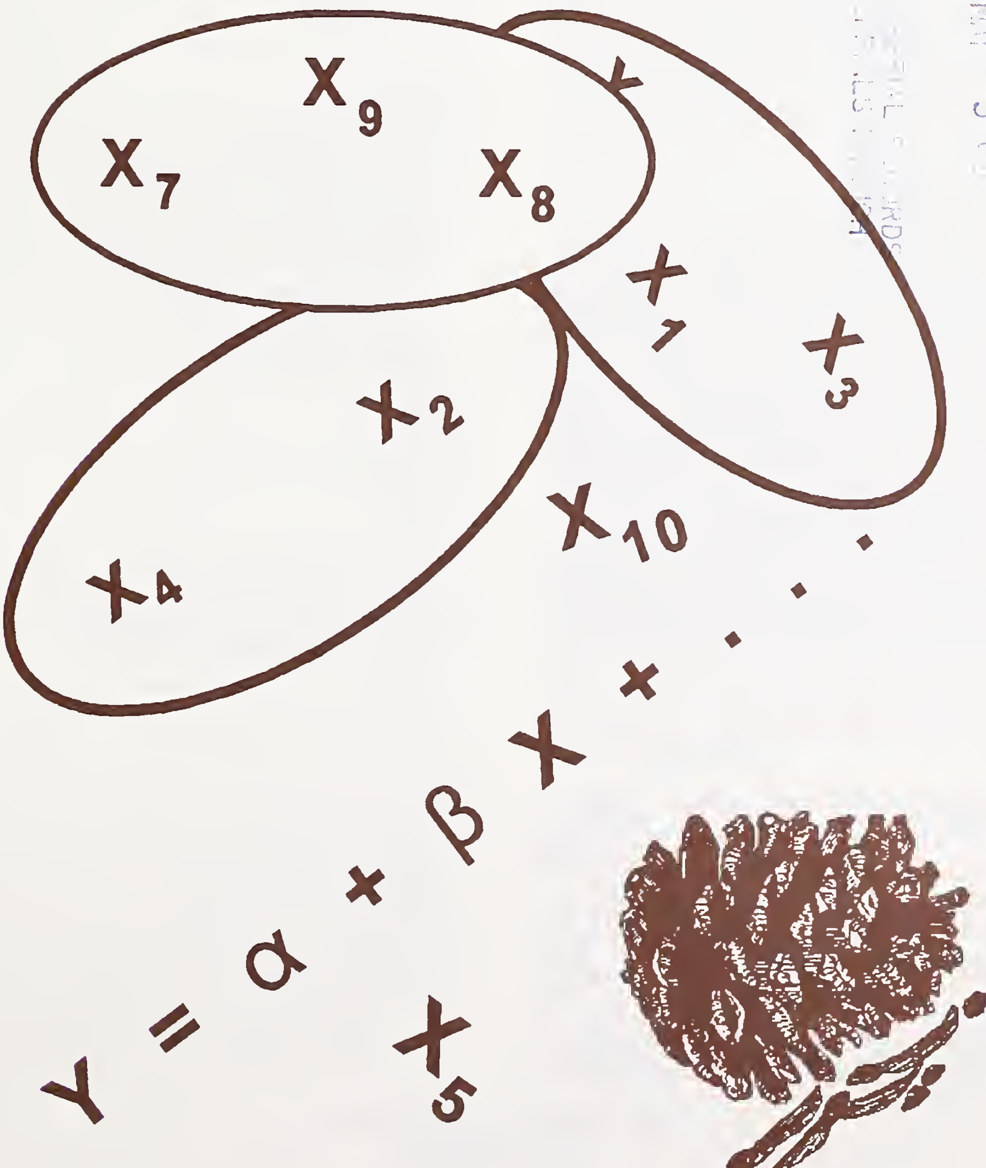
**Intermountain
Research Station**

Research Paper
INT-470

January 1994

Identifying Proxy Sets in Multiple Linear Regression: An Aid to Better Coefficient Interpretation

Gordon D. Booth
Michael J. Niccolucci
Ervin G. Schuster



THE AUTHORS

GORDON D. BOOTH is Project Leader of the Statistical Aspects of Monitoring Research Work Unit at the Intermountain Research Station in Ogden, UT. He came to the Station in 1981 from the Agricultural Research Service where he had worked since 1967. Before that he worked as a statistical consultant in private industry. He holds a B.S. degree in statistics from Brigham Young University and M.S. and Ph.D. degrees in statistics from Iowa State University.

MICHAEL J. NICCOLUCCI is an economist with the Economics Research Work Unit, Intermountain Research Station, Forestry Sciences Laboratory, Missoula, MT. He received academic training in economics at the University of Montana, where he earned his M.A. degree. His research includes timber sale design and appraisal.

ERVIN G. SCHUSTER is research forester and Project Leader of the Economics Research Work Unit. He studied forestry at the University of Minnesota and Iowa State University, where he received a Ph.D. degree in forest economics. His research includes modeling timber harvest and timber sale design, and economic impact analysis.

RESEARCH SUMMARY

Introduced here is the concept of a proxy set which we define to be a collection of potential explanatory variables linearly related to one another. Therefore, each member of the proxy set conveys some, and perhaps much, of the same information as other members of the same proxy set if they are included in a linear regression model together. Therefore, interpreting a coefficient in a multiple regression equation can be misleading if proxy-set membership is ignored. All potential explanatory variables should be examined *before* a linear regression model is constructed to see if some variables belong to proxy sets. Accounting for those proxies not included in the model as well as those that are included permits a more realistic interpretation of the coefficients in the final regression model. Seven diagnostic techniques are discussed: the correlation matrix method, the iterative variance inflation factor method (introduced here for the first time), the variance decomposition method, principal components without rotation, principal components with rotation, factor analysis, and cluster analysis. The effectiveness of these seven methods in identifying proxy sets is examined using data with known proxy set structure. The iterative variance inflation factor and the variance decomposition methods were the best overall performers; factor analysis was the worst.

ACKNOWLEDGMENTS

We are grateful to research forester Clinton E. Carlson, Forestry Sciences Laboratory, Missoula, MT, for permitting us to use his data on bract and scale widths of larch cones.

CONTENTS

Intermountain Research Station
324 25th Street
Ogden, UT 84401

Identifying Proxy Sets in Multiple Linear Regression: An Aid to Better Coefficient Interpretation

Gordon D. Booth
Michael J. Niccolucci
Ervin G. Schuster

INTRODUCTION

Multiple linear regression is used extensively when analyzing data from natural resource studies. Many natural resource workers do not limit use of the regression equations to prediction. Rather, they interpret the estimated regression coefficients, and these interpretations become the basis for far-reaching policy recommendations and management decisions.

Regression analyses are usually expressed in terms of a dependent variable, which we call a response variable. Likewise, we use the term "explanatory variable" to indicate what is often called an independent variable. Our reason for not using the terms "dependent" and "independent" is to avoid confusion when we discuss mathematical independence (orthogonality), a condition that plays a critical role in this article.

Regression coefficients are commonly interpreted as representing the change in the response variable caused by a one-unit increase in the corresponding explanatory variable, with all other explanatory variables held constant. This is tantamount to taking a partial derivative of an equation with respect to a specific explanatory variable and interpreting it. This procedure has several serious problems: (1) a cause-and-effect relationship is not inherent in regression analysis; (2) some explanatory variables (such as weather) cannot be held constant; (3) interpretation of regression coefficients must take into account other explanatory variables in the model (the traditional collinearity problem); and (4) interpretation of the regression coefficients must also take into account other explanatory variables that are *not* in the model. Common mistakes involving items (1) and (2) are discussed at length in the literature (Hocking 1976; Mosteller and Tukey 1977; Draper and Smith 1981). Belsley and others (1980) discuss the traditional collinearity problem, item (3). In this article, we concentrate on the importance of explanatory variables that are not in the model, item (4), although item (3) is inherent in our discussions.

THE PROBLEM

The explanatory variables finally included in a regression model are often selected from a larger set of potential explanatory variables. Some are truly independent of (orthogonal to) each other and pose no difficulty either in model building or interpretation. Other explanatory variables go together, occurring as packages or associations. We call these packages *proxy sets*. We define a proxy set as a collection of explanatory variables, any one of which conveys some of the same information as any of the other variables in the set. In a sense, each variable in a proxy set represents all other variables in the same set, at least partially; each such variable could conceivably serve as a proxy for the entire set. For example, average leaf length and average leaf width may belong to a proxy set conveying the effect of leaf size on a response variable such as total biomass. If an explanatory variable is not a member of any proxy set, we call it a nonproxy variable. Proxy-set membership has far-reaching implications for the interpretation of regression coefficients.

An example illustrates the problem of proxy-set membership. Consider the effects of four explanatory variables on the ratio of bract width to scale width (RATIO) in larch cones: wet cone length (WETLEN), wet cone width (WETWID), dry cone length (DRYLEN), and dry cone width (DRYWID). At this point, we wish to draw attention to only two of these four potential explanatory variables: WETLEN and DRYLEN. The model considered is:

$$\text{RATIO} = (\text{Bract width})/(\text{Scale width}) = \beta_0 + \beta_1 \text{WETLEN} + \beta_2 \text{DRYLEN} + \beta_3 \text{WETWID} + \beta_4 \text{DRYWID} + \varepsilon$$

The estimate of β_1, the coefficient of WETLEN, is 0.000705 ($t = 0.47$ with $P = 0.636$). The estimate of β_2, the coefficient of DRYLEN, is –0.001780 ($t = -1.13$ with $P = 0.261$). The small t-values and corresponding large probabilities for the coefficient estimates of both WETLEN and DRYLEN would seem to suggest that neither of these variables has a potential for explaining the bract/scale ratio. In fact, faced with comparable evidence, many published linear regression analyses have drawn similar conclusions.

In reality, we would be in error if we accepted the interpretation that neither WETLEN nor DRYLEN is important. This is seen easily by fitting the same model as before, but with DRYLEN removed from the model. The estimate of the coefficient of WETLEN becomes –0.000898 ($t = -2.06$ and $P = 0.0404$), a significant result, in direct contrast to the nonsignificant result obtained when both WETLEN and DRYLEN were in the model. Likewise, fitting the model with DRYLEN present, but with WETLEN removed, produces an estimated coefficient for DRYLEN of –0.001064 ($t = -2.30$ and $P = 0.0220$). Again, this result is in direct contrast with the result obtained when both WETLEN and DRYLEN were included in the model.

Either WETLEN or DRYLEN can describe the bract/scale ratio if the other is absent from the model. It would have been a mistake to have concluded that neither variable is useful in describing the bract/scale ratio. WETLEN and DRYLEN are members of the same proxy set. Either could be used to represent the effect of length on the bract/scale ratio. In such cases, a given regression coefficient might be thought of as representing the effect of the entire proxy set on the response variable. This example illustrates three characteristics of proxy sets:

1. If more than one member of a proxy set is included in a regression model, all members of that set included in the model appear less important and may even be statistically "nonsignificant."
2. To some extent, each member of a proxy set is capable of "standing in" for all members of the set. Thus, if a proxy set is important, at least one (and possibly only one) member of the set is needed in the regression model.
3. If a member of a proxy set is included in a model, its statistical "significance" implies that the entire proxy set is "significant." This includes all variables in the set—whether they are included in the model or not.

Traditional regression diagnostic tools focus exclusively on the explanatory variables included in the proposed model. Here, *we focus on the problems of interpreting regression coefficients corresponding to explanatory variables that are proxies of other explanatory variables, some of which may not have been included in the model.* This leads to two important questions: (1) "What techniques can be used to identify proxy sets?" and (2) "Do the techniques correctly identify proxy sets?"

PROXY-SET IDENTIFICATION METHODS

Variables that are collinear with one another are members of the same proxy set. Therefore, the problem of identifying proxy sets is essentially the problem of identifying collinear variables. Often a variable included in a model is a member of a proxy set containing several variables *that do not appear in the model.* The point we emphasize is that these proxy variables that are not in the final model must also be considered when interpreting the coefficients that remain in the model. The usual methods for identifying collinearity among model variables *after the model is fitted* fail to consider possible proxy variables that did not make it into the model.

Over the years, several methods have been proposed for diagnosing collinearity in a fitted model. Some of these methods can also be used to identify proxy sets containing proxies that did not make it into the final model. We will discuss several of these methods and will introduce a particularly powerful method that has not appeared previously in the literature. In all cases, proxy sets are to be identified *before fitting the model.* Thus, the methods we propose will consider variables that never become part of the final model.

The effectiveness of each of the proposed methods depends on the decision criteria the user employs. These criteria take the form of numerical limits or cutoffs which, when exceeded, indicate the presence of proxies. Unfortunately, there is no firm theoretical reason for choosing specific cutoffs. We agree with Baskerville and Toogood (1982) that "it is difficult and perhaps inappropriate to give general rules since a preliminary exploratory analysis should be flexible." Nevertheless, the identification of proxy sets demands that we choose some cutoffs. We have attempted to follow recommendations in the literature when possible, but in cases where no recommendations can be found, we suggest some numerical cutoffs we find useful. In fact, the cutoffs we suggest are the ones we used to obtain the results presented in this article. Our suggested cutoffs appear at the end of the description of each method. The cutoffs we did not find in the literature were determined during our research when we knew, by careful construction of the data sets, which variables belonged to proxy sets. The cutoffs we suggest are those that gave the specific identification method a fair opportunity of identifying the proxy sets we knew to be correct. In some cases, we chose cutoffs based on other considerations. Nevertheless, our suggested cutoffs are strictly empirical, and users of these methods may wish to select their own numerical cutoffs. More details on our choices of cutoffs are found in appendix A.

The Correlation Matrix Method

Examining the elements of the correlation matrix, **R**, is probably the oldest method for detecting linear relationships among variables. It has been used extensively. Its main limitation is its inability to detect relationships between more than two variables at a time. Inspection of correlation coefficients often fails to detect relationships involving several variables, especially when all relationships between pairs of variables are fairly weak. Another shortcoming is the difficulty of keeping track of many variables at once. On the other hand, the correlation matrix method (CORR) is easy to use and is generally available. In practice, users simply look for large positive or negative correlation coefficients.

If the absolute value of a correlation coefficient was greater than 0.5, we considered the two variables to be proxies of one another, belonging to the same proxy set. Likewise, if the absolute values of all correlation coefficients involving a specific variable were less than 0.32, we considered that variable to be a nonproxy. Variables with correlation coefficients between 0.32 and 0.5 were not specified as proxies or as nonproxies.

We cannot identify proxy sets of more than two variables by using CORR alone. However, we established a rule that if CORR identified *all* lower-order proxy sets, we considered it to have identified the higher-order proxy set as well. For example, a three-variable proxy set might involve variables x, y, and z. If the proxy sets $\{x,y\}$, $\{x,z\}$ and $\{y,z\}$ are all identified, we considered the three-variable set $\{x,y,z\}$ to be identified also.

Iterative Variance Inflation Factor Method

Variance inflation factors (VIF's) are the diagonal elements of $\mathbf{R}^{-1}$, the inverse of the correlation matrix (Belsley 1991, pp. 27-28). Here we introduce, for the first time, an algorithm based on a modification of the VIF. This algorithm is particularly useful for identifying proxy sets, but it can also be used during the fitting stage of regression model-building. When used in the latter capacity, it permits the user to determine which sets of explanatory variables are collinear with one another. Thus, it aids the user in building models with "relatively unrelated" explanatory variables.

The VIF, is usually applied in a static manner. Then it is capable only of identifying which explanatory variables are involved in collinearities. It does not specify which variables are collinear with which other variables. A VIF-based method capable of helping us identify the specific groups of variables that are collinear with one another would be even more useful. Such a method would have to use the VIF in a dynamic, iterative manner. We present the details of such a method, the iterative variance inflation factor (IVIF) method, in appendix B. In general, the method is based on entering variables one at a time, so the behavior of the entire set of VIF's is evaluated.

If the VIF of any variable (already in the model) jumped to a value greater than 1.5 when a new variable was introduced, we considered it to be a proxy of the newly entered variable. Those variables with VIF's near 1 can be considered nonproxy variables.

Variance Decomposition Method

Variance decomposition (VDC) is a regression diagnostics technique that is becoming more readily available in statistical analysis computer programs each year. It has proven itself to be of considerable value in detecting collinearity. As described in Belsley and others (1980), VDC specifically identifies those variables that are linearly related to one another.

VDC expresses each variance component as a proportion of the total variance for a given regression coefficient. Therefore, the total of all the proportions for a given regression coefficient will equal 1. This method can readily identify proxy sets containing more than two variables.

Following the suggestion of Belsley and others (1980) for identification of collinear variables, when two or more proportions corresponding to the same eigenvalue were greater than 0.5, we considered the corresponding variables to be proxies. If the variance proportions were all less than 0.5, we considered the corresponding variable to be a nonproxy. Thus, if there were four variables with proportions greater than 0.5 for the same eigenvalue, all four were considered members of the same proxy set.

Principal Components Methods

The use of principal components analysis (PC) is not new in natural resources (Isebrands and Crow 1975). However, the usual applications emphasize the principal components corresponding to the largest eigenvalues. These principal components leave the least unexplained variability. In contrast, we suggest focusing on the principal components associated with the *smallest* eigenvalues in order to identify proxy sets. Small eigenvalues are associated with linearly related variables; therefore, smaller eigenvalues result from collinearity.

Principal components can be rotated by applying a linear transformation to them. Sometimes, this type of rotation can help us interpret the components. Many rotations could be performed, but we consider only the Varimax rotation, which leaves the rotated principal components (RPC) orthogonal to one another. Among the many sources of additional information on PC and RPC are Morrison (1967) and Dillon and

Goldstein (1984). Latent root regression (Hawkins 1973; Sharma and James 1981; Webster and others 1974) is based on methods closely related to our use of principal components. Baskerville and Toogood (1982) propose a classification system for variables and suggest use of latent root regression as part of a whole model-building philosophy they call "guided regression."

Without rotation, one merely examines the coefficients within each principal component and looks for possible groupings. After rotation, one looks at the coefficients within a given principal component; any variables with a coefficient greater than 0.32 are considered members of the same proxy set. If either PC or RPC identified only a single variable, it was defined to be a nonproxy.

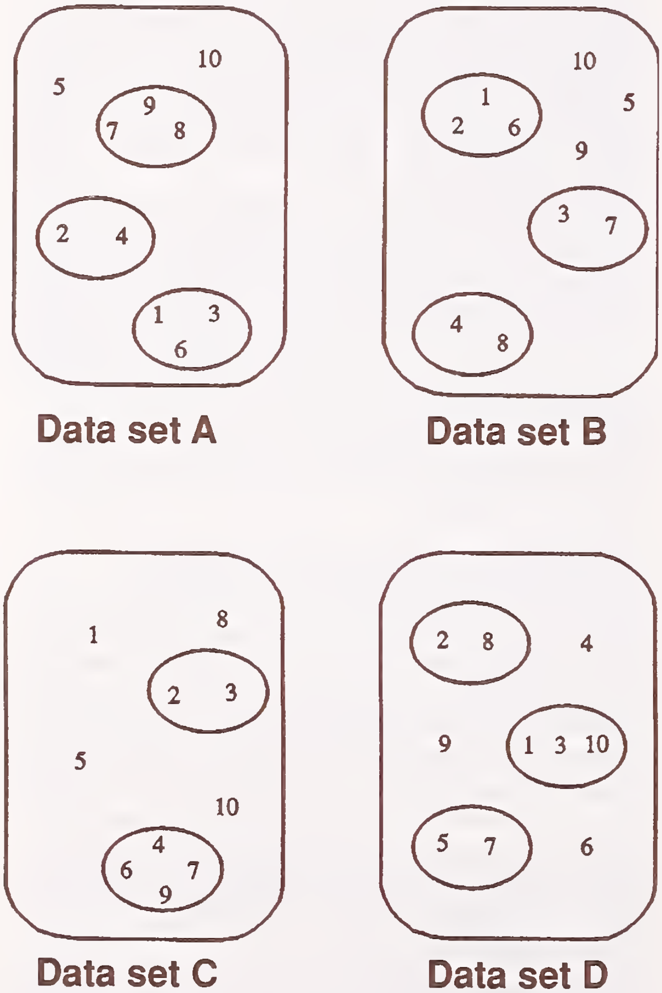


Figure 1—Graphic representation of the structure of the four data sets used in comparing the efficiencies of the seven proxy-set identification methods. There are 10 variables in each data set. Variable X_1 is represented by 1, X_2 is represented by 2, etc. Variables in the same proxy set are enclosed in the same ellipse.

Factor Analysis Method

Most factor analyses are performed on the correlation matrix, **R** (Dillon and Goldstein 1984, pp. 63-68). We recommend doing so when identifying proxy sets. A factor analysis is performed on all the variables that are candidates for inclusion in the regression model.

As with PC, factor analysis (FA) also allows rotation of the axes to permit better identification of some aspects of the data. When using factor analysis to identify members of a proxy set, the first factors, corresponding to the largest eigenvalues, are most likely to produce usable proxy sets. Within a given factor, any variables with a loading greater than 0.32 were considered members of the same proxy set. We only recommend the use of FA with the Varimax rotation.

Cluster Analysis Method

Cluster analysis (CLUSTER) is not a single method, but a collection of methods (Hartigan 1975; Romesburg 1984). These methods are commonly used in their traditional way in natural resources research (Turner 1974). Their traditional use is to attempt to group, or "cluster," individuals on the basis of similarities in a set of measurements made on each individual. Using CLUSTER to identify proxy sets requires exchanging the roles of the individuals and the variables. In other words, the variables are grouped, or "clustered," on the basis of their response on the various individuals. We always use the hierarchical method of clustering. Cluster analysis output identifies proxy sets directly. If a variable was not found in a proxy set (that is, it was not found in a cluster), it was considered a nonproxy.

STUDY DESIGN

This study was designed to determine the effectiveness of seven identification techniques used in proxy-set identification. We define effectiveness as the percentage of existing proxy sets identified correctly by an identification technique.

The effectiveness of an identification technique can only be measured if we know which proxy sets really exist. Otherwise, there is no basis for comparison. Therefore, all seven techniques were applied to each of four different data sets. Each data set was generated to contain several known proxy sets. Proxy sets had one, two, three, or four variables. Each of the four data sets consisted of 100 observations on 10 variables. The proxy-set structures are illustrated in figure 1, where variable X_1 is indicated by the number 1, X_2 by 2, and so forth. All variables with identification numbers enclosed in the same ellipse are members of the same proxy set. For example, there are three proxy sets in data set A: two sets with three members each, and one set with two members. In addition, there are two nonproxy variables.

We devised a scoring system to compare the different methods' efficiencies in identifying proxy sets correctly. To develop the score, all possible subsets of variables within each true proxy set were enumerated. For example, the proxy set in data set A containing variables X_1, X_3, and X_6, which we denote by $\{X_1, X_3, X_6\}$, has four subsets of interest: $\{X_1, X_3\}$, $\{X_1, X_6\}$, $\{X_3, X_6\}$, and $\{X_1, X_3, X_6\}$. These are the subsets in which a given variable appears correctly identified with at least one other member of its true proxy set. We refer to these subsets as "enumeration sets." Also, we refer to a list of all these subsets, from all the true proxy sets in a data set, as the "true enumeration list" for the data set. All nonproxy variables are also included in the enumeration list, because it is important to identify them as well. As an example, data set A includes two proxy sets that contain three variables each. Four enumeration sets are obtained from each of these two proxy sets (a total of eight enumeration sets). Also, data set A includes a proxy set containing only two variables. A single enumeration set is obtained from this two-variable proxy set. Finally, there are two nonproxy variables that contribute one variable each to the enumeration list. Thus, the enumeration list contains a total of 11 enumeration sets (8 + 1 + 1 + 1).

The variables for the data sets were generated by one of the authors. A different author who did not know which variables belonged to proxy sets applied the seven identification techniques.

For example, each of the identification methods was applied to data set A. A separate enumeration list was prepared for each of the methods. Each list was compiled from the proxy sets identified by the corresponding method. A method scored one point for each set in its enumeration list corresponding to a set in the true enumeration list of data set A. The rationale of the scoring system is that a point is given for every partially correct proxy-set identification. The scoring system simply gives the identification method a score equal to the percentage of enumeration sets in data set A that were correctly found when the technique was applied. Data sets B, C, and D were evaluated in the same way.

The proxy sets generated for this study contained variables having known linear relationships with one another that ranged from weak to strong. Therefore, some measure of the strength of relationships within a proxy set seemed useful. We chose the mean of the absolute values of the correlation coefficients for all possible pairs of variables within a proxy set. We call this measure the "average correlation." As an example, for a proxy set containing X_7, X_8, and X_9 we can compute three correlation coefficients: between X_7 and X_8, between X_7 and X_9, and between X_8 and X_9. Our measure of strength for the relationships among the variables in this proxy set is the mean of the absolute values of these three correlation coefficients.

Logistic regression (Hosmer and Lemeshow 1989; McCullagh and Nelder 1989) was used to describe two relationships: (1) between the proportion of enumeration sets identified correctly and average correlation and (2) between the proportion of enumeration sets identified correctly and the number of variables in the proxy set. For this analysis, nonproxies were not used.

RESULTS

We computed the effectiveness of the seven methods for identifying proxy sets, based on the four data sets generated with 100 observations each. The four data sets studied included 23 known associations: one four-variable proxy set, four three-variable proxy sets, six two-variable proxy sets, and 12 nonproxy variables. From these proxy sets, we constructed the enumeration sets defined earlier. From data set A we obtained 11 enumeration sets, from data set B we obtained 9 sets, from data set C we obtained 16 sets, and from data set D we obtained 9 sets. This gives us a total of 45 enumeration sets. We define overall effectiveness to be the percent of the enumeration sets that were correctly identified out of the total of 45 possible.

Table 1 shows substantial differences in effectiveness for the seven proxy-set identification techniques. The "Total" column provides an overall measure of each method's effectiveness in correctly identifying proxy sets. A method that identified all proxy sets correctly would have a score of 100 percent. In this study, effectiveness scores ranged from a high of about 91 percent for the IVIF method down to 47 percent for FA.

Our results are not extensive enough to differentiate clearly among methods that perform about equally well. Therefore, to provide practical guidelines, we chose to place the seven methods in four groups based on their effectiveness scores. The IVIF method stands alone with a score of 91.1 percent. The VDC method had the second highest effectiveness score (71.1 percent). CORR (57.8 percent), PC (60 percent), and RPC (60 percent) form the third group, correctly identifying an average of 59.3 percent correct. The final group consists of CLUSTER (51.1 percent) and FA (46.7 percent), for an average of only 48.9 percent correct.

Ideally a proxy-set identification method should be effective regardless of the composition of the data set. A method should find a high percentage of proxy sets, and it should identify different types of proxy sets. Table 1 presents the effectiveness scores for each of the seven methods when applied to each of the four data sets: A, B, C, and D. Effectiveness scores for individual combinations (of method and data set) range

Table 1—Effectiveness scores (percent correct) of seven proxy-set identification methods

Method[1]	Data set A	Data set B	Data set C	Data set D	Proxy-set size: Non-proxies	Proxy-set size: Two variables	Proxy-set size: Three variables	Proxy-set size: Four variables	Total
	Percent correct								
CORR	63.6	77.8	43.8	55.6	100.0	58.3	0	0	57.8
IVIF	100.0	100.0	100.0	55.6	100.0	87.5	87.5	100.0	91.1
VDC	63.6	44.4	100.0	55.6	100.0	58.3	62.5	100.0	71.1
PC	81.8	55.6	75.0	11.1	8.3	75.0	87.5	100.0	60.0
RPC	63.6	77.8	50.0	55.6	100.0	62.5	0	0	60.0
FA	45.5	55.6	25.0	77.8	16.7	75.0	12.5	0	46.7
CLUSTER	63.6	66.7	31.3	55.6	8.3	75.0	50.0	0	51.1

[1]CORR = the correlation matrix method; IVIF = the iterative variance inflation factor method; VDC = the variance decomposition method; PC = the principal components method; RPC = the rotated principal components method; FA = the factor analysis method; CLUSTER = the cluster analysis method.

from several perfect scores (100 percent) for the IVIF and VDC methods down to 11 percent for the PC method applied to data set D. However, most methods had difficulty finding the correct proxy sets in data set D, which contained some rather weak relationships within the proxy sets. Data set C, which included the only proxy set containing four variables, also presented problems for several of the methods.

The four data sets varied in their composition of one-, two-, three-, and four-variable proxy sets. A method that identifies complex proxy sets should perform better than other methods on data consisting of complex proxy sets. However, not all proxy sets are complex. The different sizes of proxy sets in data sets A, B, C, and D provide the variability needed to indicate a particular method's performance under different proxy-set structures.

Effectiveness scores varied greatly based on the size of the proxy set (table 1). Most methods identified two-variable proxy sets correctly, with scores ranging from 58.3 percent to 87.5 percent effective. If readers inspect only the two-variable column of table 1, they might conclude that the techniques are all relatively successful. However, it is only at the two-variable proxy-set size where this conclusion can be reached. For larger proxy sets, effectiveness ranged from 0 to 100 percent.

The IVIF and PC methods did a particularly good job of identifying the three- and four-variable proxy sets. Although VDC identified the four-variable set, it had some difficulty with the three-variable sets. The CLUSTER method did a fair job of identifying the three-variable proxy sets, but missed the single four-variable proxy set entirely. The CORR and RPC methods did not find any of the three- or four-variable proxy sets. Also, the FA method failed to find the four-variable set and did poorly on the three-variable proxy sets as well.

The CORR, IVIF, VDC, and RPC methods found all nonproxy variables, but the PC, FA, and CLUSTER methods missed all but a few of them. Although the PC method had high scores for the three- and four-variable proxy sets, it failed to find most of the nonproxy variables, which might be thought of as simple proxy sets. Nevertheless, nonproxy variables are very important when interpreting regression coefficients, because correct identification of the nonproxy variables permits relatively clear interpretation of the regression coefficients associated with them.

We modeled the proportion of enumeration sets identified correctly, using both "average correlation" and "number of variables in the proxy set" as the explanatory variables in a logistic regression. The results are presented in table 2. The fitted equations are plotted in figure 2.

The IVIF, VDC, and PC methods had fairly large probabilities associated with the fitted slope when average correlation was used as a predictor (table 2). This indicates lack of conclusive evidence of a relationship between proportion of enumeration sets identified correctly and average correlation (a measure of the strength of relationships among variables in the same proxy set). This is shown graphically in figure 2 by relatively flat curves for these three methods.

The other four proxy-set identification methods display strong relationships between the proportion of enumeration sets identified and average correlation.

Only the FA, RPC, and CLUSTER methods have probabilities small enough to indicate detectable relationships between the proportion of enumeration sets identified correctly and the number of variables in a proxy set (table 2). Figure 3 contains a graphical representation of the fitted relationships. Of course, the failure to find such a relationship may be the result of too little data. We had only a single four-variable proxy set.

Table 2—Results of logistic regression fitting of proportion of proxy sets identified correctly on average correlation and on number of variables in the proxy set

	Average correlation model			Number of variables in proxy set model		
Method[1]	Intercept	Slope	P-value[2]	Intercept	Slope	P-value[2]
CORR	−21.6	40.1	0.0211	1.4	0.2	0.8286
IVIF	1.7	.6	.7980	1.4	.2	.8286
VDC	.2	.6	.7321	−.6	.5	.5248
PC	2.7	−3.0	.1724	−1.4	1.2	.2909
RPC	−12.6	24.3	.0076	3.8	−1.3	.0776
FA	−5.3	11.7	.0037	6.6	−2.9	.0127
CLUSTER	−1.4	4.7	.0212	3.8	−1.3	.0776

[1]CORR = the correlation matrix method; IVIF = the iterative variance inflation factor method; VDC = the variance decomposition method; PC = the principal components method; RPC = the rotated principal components method; FA = the factor analysis method; CLUSTER = the cluster analysis method.
[2]P-value associated with the test on the slope of the logistic regression.

DISCUSSION

The methods used to identify proxy sets are valuable because they flag potential problems when interpreting multiple linear regression coefficients. An ideal method would have to be totally effective (identifying every real proxy correctly). Unfortunately, none of the methods we evaluated met that standard.

Any one of the seven methods could be useful if it were the only one available. Therefore, we do not recommend discarding any of the seven methods. However, some of the methods are of limited value in identifying proxy sets containing weakly related variables. The data analyst should use the best technique available.

When interpreting regression coefficients, it would be a serious error to treat a variable that is a member of a proxy set as if it were a nonproxy, independent of all other variables. Without any analysis to detect proxy sets, all variables would be regarded as nonproxies and would be interpreted accordingly. We feel analysis with any of the seven methods is better than no analysis at all, but some caution must be exercised when using any of the methods.

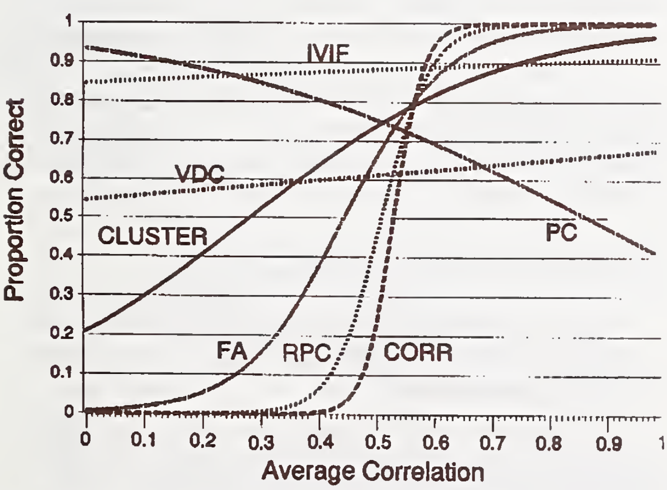


Figure 2—Plotted relationships from fitted logistic regressions of proportion of proxy sets correctly identified on average correlation, for seven identification methods: the correlation matrix method (CORR), the iterative variance inflation factor method (IVIF), the variance decomposition method (VDC), the principal components method (PC), the rotated principal components method (RPC), the factor analysis method (FA), and the cluster analysis method (CLUSTER).

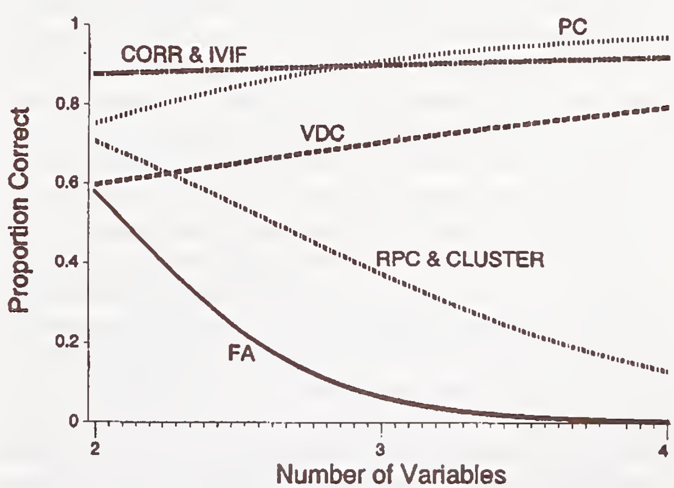


Figure 3—Plotted relationships from fitted logistic regressions of proportion of proxy sets correctly identified on number of variables in the proxy set, for seven identification methods: the correlation matrix method (CORR), the iterative variance inflation factor method (IVIF), the variance decomposition method (VDC), the principal components method (PC), the rotated principal components method (RPC), the factor analysis method (FA), and the cluster analysis method (CLUSTER).

Consider the explanatory variables, WETLEN and DRYLEN, addressed earlier. We began by discussing a situation with both variables in the model, where neither was statistically significant. If either one were included without the other, however, the remaining variable was highly significant. Analysis with IVIF indicated that these variables were members of the same proxy set, which we might call "length." Traditional diagnostics on a final regression model containing either of these variables may well have concluded "no collinearity problem," leaving the analyst free to interpret the coefficient without restraint. But traditional collinearity diagnostics fail to detect proxies that are not included in the model. Most of the seven proxy-set identification methods we used identified WETLEN and DRYLEN as members of the same proxy set. Such identification of an existing proxy set should cause us to exercise caution when interpreting the variables that are included in the model and to consider their relationships with their proxies that are not included in the model.

Analysis of proxy-set membership should take place before attempting to build the model, and should become an integral part of the overall model-building process. For example, knowledge of proxy-set membership might permit an analyst to include only a single member of each proxy set in the model. In fact, this may be the most common course to follow. However, not all variables in a proxy set perform equally well in a regression model because each variable probably conveys only part of the information conveyed by the other members of the same set; therefore, the choice of which member (or members) of the proxy set to include in the model is an important one. Such a choice may have to be based on non-statistical considerations, such as the availability of measurements, the cost of obtaining measurements, and so forth. If it is necessary to include more than one variable from the same proxy set in a regression model, the analyst may wish to use a biased regression procedure such as ridge regression, first proposed by Hoerl (1962). Further work by Hoerl and Kennard (1970) and work in forestry by Bare and Hann (1981) make this technique a useful tool once the proxy sets have been identified.

Of course, once the coefficients have been estimated, the analyst can only test for collinearity among those variables that were included in the final model. Though important, that type of postfitting diagnostic analysis cannot provide insight beyond those variables that were included in the final version. Traditional collinearity diagnostic procedures, applied to the final model, are totally ineffectual in identifying proxy-set variables that are not in the model. Yet, variables that are not in the model are often implicitly interpreted as having no importance in understanding the behavior of the response variable, an interpretation that may be totally incorrect.

This article focuses on proxy sets as they affect the interpretation of regression coefficients. Proxy-set membership should be analyzed whenever regression coefficients are interpreted or the explanatory variable of concern is a policy-related variable intended to be manipulated. For example, econometric analyses (such as supply-demand modeling) commonly interpret coefficients and then make policy recommendations based on the interpretation. Similarly, growth and yield models often use explanatory variables that could be members of proxy sets, but are sometimes treated as if they were not. The obvious danger of incorrectly acting as if a variable were a nonproxy is that resultant policy actions may well fail, or "scientific knowledge" allegedly gained may be misleading or erroneous.

Figure 2 illustrates the strengths and weaknesses of the seven proxy-set identification methods. The IVIF method is clearly the best choice because it correctly identifies a high proportion of the enumeration sets, regardless of the strength of relationships within the proxy set (as measured by average correlation). The VDC method also does quite well. The PC method does better when there are weak relationships in the proxy set than when there are strong ones. The CLUSTER method does better identifying strong relationships than weak relationships. The FA, RPC, and CORR methods do poorly identifying weak relationships, but well identifying strong ones. This information may help when selecting a method that might be suitable for a specific application.

The IVIF method performs well over the entire range of average correlation. It also has the advantage of being widely available because any regression program that computes VIF can be used interactively to produce the desired results. A step-by-step procedure for identifying proxy sets and using them in interpretation of coefficients in multiple linear regression is presented in the next section.

A SUGGESTED PROCEDURE

We recommend the following 7-step procedure for using proxy sets to help interpret the coefficients in multiple linear regression:

1. Obtain data on all variables that are possible candidates for inclusion as explanatory variables in the final multiple linear regression model.
2. Identify all proxy sets among the candidate variables in step 1.
3. Choose one variable from each proxy set identified in step 2. Each such variable is the initial representative of its proxy set. This choice can be made on practical grounds; we may wish to choose the variable that is the most easily available, the most inexpensive, or that is available in the most timely manner.
4. Use the customary model-building techniques on the representatives of each proxy set along with

the nonproxy variables. This produces the basic model.

5. If the coefficient of the representative variable from any proxy set was not statistically significant when tried in the model, attempt to enter a different member of the same proxy set instead. If no member variable of a certain proxy set produces a statistically significant coefficient, perhaps that proxy set does not help describe the response variable and does not need to be represented in the model. Continue this procedure until all proxy sets or nonproxy variables have had a chance to be represented in the model.

6. Now try to enter additional variables from the proxy sets to see if the fit improves. If a second (or even third) variable from the same proxy set produces a statistically significant coefficient, decide whether this new variable should become a part of the model. If you decide to include one or more variables of this type, final interpretation of the coefficients must take the duplicate nature of these variables into account. As a duplicate variable is introduced into the model, all its proxies that were already in the model appear less important.

7. Consider the proxy sets when interpreting the coefficients of the final regression model. Each variable in the model represents not only itself but also all other members of its proxy set. Therefore, the absence of a particular variable in the final model may not mean it is unimportant. Its absence may only mean that it is represented in the model by a proxy, another member of the same proxy set.

REFERENCES

Bare, B. B.; Hann, D. W. 1981. Applications of ridge regression in forestry. Forest Science. 27: 339-348.

Baskerville, J. C.; Toogood, J. H. 1982. Guided regression modeling for prediction and exploration of structure with many explanatory variables. Technometrics. 24: 9-17.

Belsley, D. A. 1991. Conditioning diagnostics. New York: John Wiley and Sons. 396 p.

Belsley, D. A.; Kuh, E.; Welsch, R. E. 1980. Regression diagnostics. New York: John Wiley and Sons. 292 p.

Dillon, William R.; Goldstein, Matthew. 1984. Multivariate analysis: methods and applications. New York: John Wiley and Sons. 587 P.

Draper, N.; Smith, H. 1981. Applied regression analysis. 2d ed. New York: John Wiley and Sons. 709 p.

Hartigan, J. A. 1975. Clustering algorithms. New York: John Wiley and Sons. 351 p.

Hawkins, D. M. 1973. On the investigation of alternative regressions by principal component analysis. Applied Statistics. 22: 275-286.

Hawkins, D. M.; Eplett, W. J. R. 1982. The Cholesky factorization of the inverse correlation or covariance matrix in multiple regression. Technometrics. 24: 191-198.

Hocking, R. R. 1976. The analysis and selection of variables in linear regression. Biometrics. 32: 1-51.

Hoerl, A. E. 1962. Applications of ridge analysis to regression problems. Chemical Engineering Progress. 58: 54-59.

Hoerl, A. E.; Kennard, R. W. 1970. Ridge regression: biased estimation for non-orthogonal problems. Technometrics. 12: 55-67.

Hosmer, D. W.; Lemeshow, S. 1989. Applied logistic regression. New York: John Wiley and Sons. 307 p.

Isebrands, J. G.; Crow, T. R. 1975. Introduction to uses and interpretation of principal components analysis in forest biology. Gen. Tech. Rep. NC-17. St. Paul, MN: U.S. Department of Agriculture, North Central Forest Experiment Station. 19 p.

Mansfield, E. R.; Helms, B. P. 1982. Detecting multicollinearity. The American Statistician. 36: 158-160.

McCullagh, P.; Nelder, J. A. 1989. Generalized linear models. 2d ed. London: Chapman and Hall. 511 p.

Morrison, D. F. 1967. Multivariate statistical methods. New York: McGraw-Hill. 338 p.

Mosteller, F.; Tukey, J. W. 1977. Data analysis and regression. Reading, MA: Addison-Wesley. 588 p.

Romesburg, H. C. 1984. Cluster analysis for researchers. Belmont, CA: Lifetime Learning Publications. 334 p.

Sharma, S.; James, W. L. 1981. Latent root regression: an alternate procedure for estimating parameters in the presence of multicollinearity. Journal of Marketing. 18: 154-161.

Turner, B. J. 1974. Applications of cluster analysis in natural resources research. Forest Science. 20: 343-349.

Webster, J. T.; Gunst, R. F.; Mason, R. L. 1974. Latent root regression analysis. Technometrics. 16: 513-522.

APPENDIX A: CHOOSING CUTOFF VALUES FOR IDENTIFYING PROXIES

With correlation coefficients and some other methods we chose 0.32 ($= \sqrt{0.1}$) as a cutoff because it represents an R^2 of only 10 percent. We felt that anything smaller would not represent a relationship of sufficient strength to be worth consideration. However, we did not choose a cutoff value for any identification method unless it produced about as good a classification into proxy sets as was achievable with that method. Of course, users of any of the methods discussed here may wish to choose other numerical cutoff values that are more meaningful to them.

APPENDIX B: THE ITERATIVE VARIANCE INFLATION FACTOR METHOD (IVIF)

The IVIF method requires that the variance inflation factors (VIF's) be examined after *each* new explanatory variable is entered in the model. If the addition of a new variable causes a dramatic increase (to a value greater than 1.5) in the VIF's of some other variable or variables already in the model, the new variable and the variables with the increased VIF's are collinear and thus belong to the same proxy set.

If the most recently entered variable causes the VIF of other variables already in the model to jump, it should be removed immediately. Then the next variable under consideration should be entered. This procedure of entering and possibly removing variables is followed until the analyst has attempted to enter each of the candidate variables in the model. After all variables have been entered into the model one time and some have been removed, all variables still included in the model should have VIF's between 1.0 to 1.5. From this point, each entry of one of the previously removed variables will cause a jump in the VIF's of some variables already in the model. Those variables that experience VIF jumps are collinear with the variable just entered. Proxy sets containing more than two explanatory variables are easily identified in this way.

A simple algorithm allows straightforward identification of proxy sets—even when they contain more than two explanatory variables. The algorithm follows:

Step 1A: Fit the model with only the first candidate explanatory variable. This will always produce a single VIF of 1.0.

Step 1B: Add the next candidate variable unless there are no others, in which case go to step 3A.

Step 2A: If there was no marked increase in any VIF, say to a value greater than 1.5, then return to step 1B.

Step 2B: If there was a jump in any of the previous VIF's, remove the last variable entered in step 1B. All variables in the model should then have VIF's near 1.0. The variable that was just removed belongs to the same proxy set as the previous variable or variables whose VIF jumped in size. Two members of a proxy set have now been identified. Return to step 1B.

Step 3A: At this point, all variables should have been entered into the model once. All variables still in the model should have small VIF's—say less than about 1.5. Each of the variables that was entered into the model and then removed will belong to some proxy set. You should keep a list of all such sets.

Step 3B: Reenter the first variable that was removed from the model. This should cause an immediate increase in the VIF of the variable that has already been identified as belonging to the same proxy set as the one just reentered. Note that other variables in the model may also display a jump in their VIF. They belong to the same proxy set as the ones identified before. It is this reentry procedure that allows identification of proxy sets with more than two members.

Step 3C: Repeat step 3B until all previously removed variables have been reentered into the model. There may now be several large VIF's when all variables are in the model, but you should already have identified the proxy sets.

An Example

An example will illustrate use of the IVIF method. The data consist of 25 observations on 11 candidate explanatory variables and a response variable, total biomass. The data are presented in table 3.

The only thing required to use the IVIF method is any computer program that prints the VIF's after the model is fitted. Table 4 indicates that a total of 15 passes of the regression program are required to obtain the desired result. Actually, fewer than 15 passes are required when one has become familiar with the process, but to avoid confusion for someone learning the method we recommend that all 15 passes be used. Because this process is iterative, a completely interactive computer program makes the work much simpler.

When we use this method, we add (or remove) an explanatory variable on each pass. We can introduce the explanatory variables in whatever order is convenient. We chose to add the variable, leaf density, first. Therefore, on pass 1 we fit a regression with only a single explanatory variable, leaf density. Whatever variable we choose to fit first, the VIF of that single

Table 3—The hypothetical data used to illustrate the iterative variance inflation factor method of identifying proxy sets and groups of collinear explanatory variables with total biomass as the response variable

No.	Leaf density	Stem length	Stem diameter	Growth rate	Crown diameter	Percent sucrose	Leaf length	Time to 2-cm height	Firm- ness	Infrared reflectance	Ratio Ca/N	Total biomass
1	6.9	11.64	7.3	3.0	3.2	7.8	5.35	2.70	6.41	4.5	1.5	27.6
2	.6	11.39	6.4	3.1	4.1	8.3	5.08	2.59	6.10	4.5	.8	23.3
3	6.8	9.56	5.1	4.7	3.0	8.8	5.28	1.49	6.04	4.1	1.8	24.7
4	5.1	11.93	7.4	3.0	2.9	7.4	5.33	3.06	6.45	3.6	1.2	26.8
5	2.5	9.28	3.5	4.6	2.5	9.1	4.76	1.17	5.67	4.1	.9	19.6
6	7.5	8.64	.5	4.2	3.5	8.6	3.40	1.85	6.18	4.3	.8	25.6
7	2.6	8.78	3.0	4.0	3.1	5.7	4.74	2.01	5.98	4.0	1.3	19.9
8	5.6	11.01	6.2	2.9	2.1	8.3	5.15	3.56	5.53	4.1	1.0	25.3
9	6.1	11.25	6.4	1.1	3.6	7.4	5.14	4.68	5.76	4.1	1.2	30.8
10	.4	11.06	5.5	1.7	2.8	8.7	4.70	4.10	6.13	3.7	1.7	21.9
11	7.1	9.26	3.7	3.0	3.2	8.2	4.73	2.75	5.67	4.0	1.5	25.9
12	5.9	9.53	4.2	2.3	2.4	7.8	4.93	3.70	5.92	4.6	1.1	25.9
13	3.1	8.21	3.8	3.0	2.7	7.5	5.24	2.91	6.27	4.0	.7	21.1
14	7.7	9.76	4.9	1.7	3.8	8.3	5.03	4.23	6.01	3.9	.5	28.6
15	4.1	7.29	1.7	2.6	3.1	6.6	4.71	3.21	5.95	3.8	1.1	21.9
16	4.8	8.59	4.2	1.9	2.8	10.0	5.13	3.68	6.05	3.6	.8	23.0
17	5.3	6.24	1.4	1.3	2.6	8.3	5.15	4.83	6.24	4.2	1.1	22.4
18	5.2	10.01	5.2	3.5	3.1	7.5	5.30	2.29	5.78	3.9	.8	25.1
19	5.2	11.69	5.2	3.8	2.4	8.3	4.21	2.22	6.71	4.0	.5	25.4
20	6.4	11.10	6.4	3.5	3.0	8.5	5.20	2.65	6.06	3.4	1.5	26.1
21	5.6	11.65	7.9	2.1	2.6	7.8	5.58	4.03	6.07	4.2	.4	27.2
22	4.6	11.57	3.5	2.6	3.2	7.8	3.52	3.83	5.54	3.6	.7	26.9
23	7.0	9.64	5.9	3.7	2.9	6.7	5.60	2.15	6.05	3.5	.7	24.2
24	4.1	9.01	7.1	4.4	3.7	8.8	6.49	1.55	5.84	3.9	.6	22.9
25	5.2	11.89	7.3	3.2	2.5	7.5	5.22	2.66	5.59	4.3	.8	26.9

variable will always be 1.0. We have just completed step 1A.

Suppose we choose to enter stem length in pass 2. From table 4, we see that the VIF for leaf density is unaffected by the introduction of stem length. Therefore the two variables do not belong to the same proxy set. Note, however, that this does not imply that either of these variables is free of proxy-set membership with other variables. We have now completed step 2A, and we must return to step 1B to enter the next variable.

When stem diameter is entered in pass 3, we note an immediate jump in the VIF for stem length. This indicates that stem diameter and stem length are members of the same proxy set. This is step 2B and we must remove the most recently added variable,

Table 4—Pass-by-pass results (VIF's) of the iterative variance inflation factor method used to determine proxy sets for variables in table 3

Pass	Leaf density	Stem length	Stem diameter	Growth rate	Crown diameter	Percent sucrose	Leaf length	Time to 2-cm height	Firm- ness	Infrared reflectance	Ratio Ca/N
1	1.0										
2	1.0	1.0									
3	1.0	2.5	2.5								
4	1.0	1.0									
5	1.0	1.0		1.0							
6	1.0	1.0		1.0	1.0						
7	1.0	1.0		1.0	1.0	1.0					
8	1.0	1.0		1.0	1.0	1.0	1.0				
9	1.1	1.0		27.4	1.1	1.0	1.1	27.8			
10	1.0	1.0		1.0	1.0	1.0	1.0				
11	1.1	1.0		1.0	1.0	1.0	1.0		1.0		
12	1.0	1.0		1.0	1.0	1.0	1.0		1.0	1.0	
13	1.0	1.0		1.0	1.0	1.0	1.0		1.0	1.0	1.0
14	1.0	158.1	264.1	1.0	1.0	1.2	106.7		1.1	1.1	1.1
15	1.1	161.9	269.5	29.6	1.1	1.2	108.2	29.7	1.2	1.1	1.1

stem diameter, from the model. Thus, pass 4 produces the same results as pass 2.

Passes 5 to 8 identify no proxy sets. Pass 5 introduces growth rate into the model. Because no VIF's jump, we are at step 2A again and must go to step 1B to enter another variable. Pass 6 enters crown diameter into the model. Again, no VIF's jump, so we can enter another variable. Pass 7 introduces percent sucrose, without causing any VIF's to jump. This means percent sucrose is not a proxy of any variables already in the model. Likewise, pass 8 enters leaf length without a jump in any VIF.

Pass 9 introduces time to 2-cm height (the time it takes seedlings to grow to a height of 2 cm) into the model, and we see an immediate jump in the VIF of growth rate. We have now identified another proxy set that contains at least the two variables, growth rate and time to 2-cm height. We are at step 2B again and must remove the last variable entered, namely time to 2-cm height. The result is pass 10.

Passes 11 to 13 identify no proxy sets. Pass 11 adds the variable, firmness, to the model with no jump in the VIF for any variable already in the model. Neither infrared reflectance in pass 12 nor the calcium-to-nitrogen ratio in pass 13 caused a jump in any VIF. Because we have now entered all candidate variables once, we are at step 1B, which directs us to go to step 3A.

We have removed two variables from the model: stem diameter (which we know is a proxy of stem length) and time to 2-cm height (which we know is a proxy of growth rate). We are now at step 3B and must reenter the first variable that was removed from the model, stem diameter. When we do so (pass 14), the VIF's of both stem length and leaf length jump. Therefore, we have identified a proxy set containing three explanatory variables: stem diameter, stem length, and leaf length.

We are now at step 3C, and we reenter the remaining variable that we removed earlier, time to 2-cm height (pass 15). Upon entry into the model, this variable caused a jump in only a single variable, growth rate. Therefore, this is a second proxy set containing only two variables: time to 2-cm height and growth rate.

Note the important difference between pass 14 and pass 15. In pass 14, a jump occurred in the VIF's of two variables, while in pass 15 only a single VIF increased.

This method, the iterative variance inflation factor method, is dynamic and allows identification of proxy sets of more than just two members. If followed systematically, it is an effective means of identifying proxies.

Here are a few time-saving suggestions. Because pass 1 will always yield a VIF of 1.0, it is unnecessary. We can start with pass 2, entering two variables initially and remembering that if the resulting VIF's are large (that is greater than 1.5), the two variables should be considered proxies and one of the variables should be removed from the model. Note also that pass 4 is identical to pass 2 and pass 10 is identical to pass 8. Therefore, we could have gone directly from pass 3 to pass 5 by removing stem diameter and adding growth rate in a single pass. Likewise, we could have gone directly from pass 9 to pass 11. These suggestions could have reduced the number of passes needed from 15 down to 12.

*U.S. GOVERNMENT PRINTING OFFICE 1993-0-573-029/81021

Booth, Gordon D.; Niccolucci, Michael J.; Schuster, Ervin G. 1994. Identifying proxy sets in multiple linear regression: an aid to better coefficient interpretation. Res. Pap. INT-470. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station. 12 p.

Introduced here is the concept of a proxy set, defined to be a collection of potential explanatory variables linearly related to one another. Therefore, each member of the proxy set conveys some, and perhaps much, of the same information as other members of the same proxy set if they are included in a multiple linear regression together. Interpreting a coefficient in a multiple linear regression equation can be misleading if proxy-set membership is ignored, even if the final regression model does not include some of the variables in the proxy set. This study compares the effectiveness of seven different methods of identifying proxy sets.

KEYWORDS: regression coefficients, statistical models, collinearity, statistical methods

INTERMOUNTAIN RESEARCH STATION

The Intermountain Research Station provides scientific knowledge and technology to improve management, protection, and use of the forests and rangelands of the Intermountain West. Research is designed to meet the needs of National Forest managers, Federal and State agencies, industry, academic institutions, public and private organizations, and individuals. Results of research are made available through publications, symposia, workshops, training sessions, and personal contacts.

The Intermountain Research Station territory includes Montana, Idaho, Utah, Nevada, and western Wyoming. Eighty-five percent of the lands in the Station area, about 231 million acres, are classified as forest or rangeland. They include grasslands, deserts, shrublands, alpine areas, and forests. They provide fiber for forest industries, minerals and fossil fuels for energy and industrial development, water for domestic and industrial consumption, forage for livestock and wildlife, and recreation opportunities for millions of visitors.

Several Station units conduct research in additional western States, or have missions that are national or international in scope.

Station laboratories are located in:

Boise, Idaho

Bozeman, Montana (in cooperation with Montana State University)

Logan, Utah (in cooperation with Utah State University)

Missoula, Montana (in cooperation with the University of Montana)

Moscow, Idaho (in cooperation with the University of Idaho)

Ogden, Utah

Provo, Utah (in cooperation with Brigham Young University)

Reno, Nevada (in cooperation with the University of Nevada)