ED 412 247	TM 027 498
AUTHOR	Brooks, Gordon P.; Barcikowski, Robert S.
TITLE	A New Sample Size Formula for Regression.
PUB DATE	1994-04-00
NOTE	55p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1994).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Effect Size; Monte Carlo Methods; *Prediction; *Regression (Statistics); *Sample Size; Selection; Simulation
IDENTIFIERS	Cross Validation; *Power (Statistics); *Precision (Mathematics)

ABSTRACT

The focus of this research was to determine the efficacy of a new method of selecting sample sizes for multiple linear regression. A Monte Carlo simulation was used to study both empirical predictive power rates and empirical statistical power rates of the new method and seven other methods: those of C. N. Park and A. L. Dudycha (1974); J. Cohen (1988); C. Gatsonis and A. R. Sampson (1989); S. B. Green (1991); E. J. Pedhazur and L. P. Schmelkin (1991); and J. Stevens (1992). [illegible] power rates of the new method were found to be superior, both relatively and absolutely, to other methods across most conditions examined. The results also demonstrate both the importance of using an effect size for determining regression sample sizes and the relative importance of predictive power over statistical power for regression. The new method of sample size selection developed in this paper provides a relatively simple means to account for both concerns. (Contains 9 tables and 87 references.) (Author/SLD)

A New Sample Size Formula for Regression

Gordon P. Brooks

Robert S. Barcikowski

Ohio University

Correspondence: Gordon Brooks
601 Courtland Lane
Pickerington, OH 43147

Paper presented at the annual meeting of the American Educational Research Association, April 1994, New Orleans, LA.

A New Sample Size Formula for Regression

ABSTRACT

The focus of this research was to determine the efficacy of a new method of selecting sample sizes for multiple linear regression. A Monte Carlo simulation was used to study both empirical predictive power rates and empirical statistical power rates of seven methods: the new method, Park and Dudycha (1974), Cohen (1988), Gatsonis and Sampson (1989), Green (1991), Pedhazur and Schmelkin (1991), and Stevens (1992). The power rates of the new method were found to be superior, both relatively and absolutely, to other methods across most conditions examined. The results also demonstrate both the importance of using an effect size for determining regression sample sizes and the relative importance of predictive power over statistical power for regression. The new method of sample size selection developed in this paper provides a relatively simple means to account for both concerns.

OBJECTIVES

Most researchers who use regression analysis to develop prediction equations are not only concerned with whether the multiple correlation coefficient or some particular predictor is significant, but they are also especially concerned with the generalizability of the regression model developed. However, the process of maximizing the correlation between the observed and predicted criterion scores requires mathematical capitalization on chance; that is, the correlation obtained is a maximum only for the particular sample from which it was calculated. If the estimate of the population multiple correlation decreases too much in a second sample, the regression model has little value. Because of this maximization on chance variation, researchers must ensure that their studies have adequate predictive power so that results will generalize. The best way to ensure predictive power in regression is to use a sufficiently large sample. This paper introduces and tests a new method for selecting appropriate sample sizes.

Why A New Method?

First of all, despite encouragement from many scholars, most notably Cohen (1977, 1988), many researchers continue to ignore power in their studies. This unfortunate fact has been documented on several occasions (Cohen ,1992; Sedlmeier & Gigerenzer, 1989; Stevens, 1992b). Indeed, this situation is compounded for multiple regression research. As Olejnik noted in 1984 and was confirmed during the current research, many regression textbooks avoid the issue of sample size selection completely (e.g. Dunn & Clark, 1974; Kleinbaum, Kupper, & Muller, 1987; Montgomery & Peck, 1992; Weisberg, 1985) and most simply provide a rule-of-thumb as a sidebar to a discussion of cross-validation (e.g. Cooley & Lohnes, 1971; Harris, 1985; Kerlinger & Pedhazur, 1973; Tabachnick & Fidell, 1989).

Next, several well-known methods exist for determining sample sizes for multiple linear regression. These methods can be grouped loosely into three categories: subject-to-variable ratio rules-of-thumb, statistical power analysis, and cross-validation analysis. Unfortunately, there are faults and contradictions among the various methods. For example, how does one reconcile differences between a 15-subjects-per-variable ratio (Stevens, 1986) and a 30-subjects-per-variable ratio (Pedhazur & Schmelkin, 1991)? Furthermore, Cohen's (1988) methods are derived from a fixed model, and statistical power, approach to regression; however, it is often suggested that a random model, cross-validation approach is most appropriate in social sciences. Additionally, Park and Dudycha's (1974) work is largely ignored, perhaps because the tables they provide are both incomplete and complicated.

Therefore, the purpose of this paper is to validate, through a Monte Carlo power study, a new and accessible method for calculating adequate sample sizes for multiple linear regression analyses. The sample size formula developed in this paper is not simply a rule-of-thumb; indeed, the new method differs from rules-of-thumb because it requires the use of an effect size. Furthermore, the new method will work from a random model perspective and does not require the researcher to use an incomplete or complicated set of tables for estimating necessary sample sizes.

PERSPECTIVES

Multiple Linear Regression

Multiple regression is a general and flexible data analytic technique that can be used either to predict or to explain phenomena (Browne, 1975; Cohen, 1968; Cohen & Cohen, 1983). The difference between the two purposes is in the interpretation of the results. Multiple regression can be used to help explain the variance of a dependent variable, or criterion, by using information from at least two independent variables, or predictors. The emphasis is on testing theoretical models and the relative importance of individual independent variables is especially meaningful (Kerlinger & Pedhazur, 1973). Furthermore, the degree of relationship between the predictors and the criterion is of interest (Cattin, 1980a). Practical application is the main emphasis of regression analysis used in prediction studies. A researcher desires to develop an efficient regression equation that optimally combines predictor scores in order to predict a subject's score on a particular criterion variable. The choice of predictors is determined primarily by their potential effectiveness in enhancing the prediction of the dependent variable.

The statistical techniques are the same for both situations. Weights are derived for a set of independent variables such that the resulting linear combination of predictors is maximally correlated with the dependent variable. In linear regression this is accomplished through the criterion of least squares; that is, the sum of the squared errors of prediction is minimized. The coefficient of multiple correlation is obtained by correlating the dependent variable scores with the optimally weighted set of predictors.

As noted above, the process of maximizing the correlation between the observed and predicted criterion scores requires mathematical capitalization on chance probabilities. When the regression equation is used with a second sample from the same population, the model will not predict as well as it did in the original sample. Consequently, the estimate of the population multiple correlation will decrease in the second sample. Researchers can ensure adequate predictive power of their regression models, and thus stable regression weights, by choosing appropriately large sample sizes. Therefore, the remainder of this paper examines the issue of sample size in multiple regression. Before sample size selection and power can be discussed in more detail, however, it is important to understand the basic perspective of regression analysis from which the authors proceed.

Two Regression Models

There are two models that can be used in regression analyses (Brogden, 1972; Sampson, 1974). The fixed model, also called the regression model or the conditional model, assumes that the researcher is able to select or control the values of the independent variables before measuring subjects on the random dependent variable. In the random model, also called the correlation model or the unconditional model, both the predictors and the criterion are sampled together from a joint multivariate normal distribution. The more stringent random model assumption of a joint distribution of the variables subsumes the fixed model assumptions of conditional normality and equal variances (Dunn & Clark, 1974). This random model is more useful to social scientists because they typically measure random subjects on predictors and the criterion simultaneously, thus they are not able to fix the values for the independent variables (Brogden, 1972; Cattin, 1980b; Claudy, 1972; Drasgow, Dorans, & Tucker, 1979; Herzberg, 1969; Park & Dudycha, 1974; Stevens, 1986, 1992a).

Mathematically, the two models are identical under a true null hypothesis of zero multiple correlation; but when the null hypothesis is not true the models differ (Herzberg, 1969). The distributional theory, and thus the mathematics, is much more complicated under the random model because the random model recognizes and accounts for the variation in both the criterion and the predictors (Claudy, 1972; Drasgow et al., 1979; Gatsonis & Sampson, 1989; Herzberg, 1969). Consequently, most research has used assumptions of the fixed model (Drasgow et al., 1979), "hoping that there will be little practical difference between the two models" (Herzberg, 1969, p. 2). Indeed, most textbook authors fail to distinguish between the two models (Barcikowski, 1980; Cummings, 1982), choosing instead to discuss, either explicitly or implicitly, a fixed model regression (e.g. Cohen & Cohen, 1983; Draper & Smith, 1966; Kerlinger & Pedhazur, 1973). Claudy (1972) noted, however, that misapplication of the fixed model to random model data causes biased estimates of the population parameters. Therefore, it is imperative that researchers understand the differences between the models. For more complete discussion of the two models, the reader is referred to Dunn and Clark (1974), Johnson and Leone (1977), and Sampson (1974).

The null hypotheses that are tested differ between the two models. Because researchers "fix" the values of the independent variables when using the fixed model approach, the hypotheses of interest are often written in terms of the predictors. More specifically, the hypotheses test the weights (called partial regression coefficients) that are given to the predictors. Because fixed predictor values are often mutually exclusive and uncorrelated, the fixed model can be used to perform analysis of variance. Indeed, in standard multiple regression (as opposed to hierarchical or stepwise regression), the same hypothesis tests the predictor weight (b_i), the partial correlation (pr_i), and the semi-partial correlation (sr_i) (Tabachnick & Fidell, 1989). The predictors need not be independent to use a fixed model approach, however; this is one reason Cohen (1968) called regression a more general and flexible statistical technique than analysis of variance.

Using a random model approach, a researcher is primarily interested in the relationship between the set of predictor variables and the criterion. Therefore, the null hypothesis tests the multiple correlation, ρ, to test the significance of the entire model. With only a single predictor, the two hypotheses, H_0: $\rho=0$ and H_0: $\beta=0$, are equivalent and thus the two models are equivalent (Cohen & Cohen, 1983; Dunn & Clark, 1974; Kraemer & Thiemann, 1987). Additionally, it should be noted that both models can be used for prediction, but only the fixed model can be used appropriately for explanation (in an experimental sense). Most of the discussion that follows is approached from a random model perspective.

Cross-Validation and Shrinkage

Because of the capitalization on chance sample covariation, the multiple correlation calculated in the sample is necessarily an overestimate of the population multiple correlation (Huberty & Mourad, 1980). That is, the expected value of the multiple correlation is larger than the true population value, or E(R) > ρ (Herzberg, 1969). Morrison (1976) reported that when ρ=0, E(R^2) = p/N-1, where p is the number of predictors and N is the sample size; however, an unbiased estimate would yield E(R^2)=0 when ρ=0. Consequently, researchers have employed a number of methods to "shrink" R^2 and thereby provide better estimates of true population multiple correlations. One method used by researchers requires empirical cross-validation, or data-splitting (Cattin, 1980a; Kerlinger & Pedhazur, 1973; Mosier, 1951; Picard & Cook, 1984). That is, the researcher builds the regression model using a derivation sample (usually half a random sample) and then applies the prediction equation to a validation sample to determine how well it predicts the criterion variable in the second sample (Stevens, 1992a). The correlation between the observed and predicted scores in the validation sample serves as an estimate of the population cross-validated multiple correlation (Cattin, 1980a, 1980b). Other variations of this method have been suggested, including double-cross-validation and jackknifing (Cummings, 1982; Huberty & Mourad, 1980; Kerlinger & Pedhazur, 1973).

Unfortunately, when using empirical cross-validation, the regression equation is not built using the entire sample. Therefore, formula methods of shrinkage are typically preferred to empirical cross-validation so that the entire sample may be used for model-building. Many such formulas have been proposed for the fixed model (Ezekiel, 1930; Lord, 1950; Nicholson, 1960; Rozeboom, 1978; Wherry, 1931) and also for the random model (Browne, 1975; Darlington, 1968; Herzberg, 1969; Stein, 1960). Indeed, several formula estimates have been shown superior to the empirical cross-validation techniques (Cattin, 1980a; 1980b; Kennedy, 1988; Murphy, 1982; Schmitt, Coyle, & Rauschenberger, 1977). Some confusion exists about the use of these "shrinkage" formulas, however.

Two types of formulas have been developed: shrinkage estimates and cross-validity estimates (see Table 1). Shrinkage formulas are used to estimate more accurately the squared population multiple correlation, ρ^2, also called the coefficient of determination; cross-validity formulas provide more accurate estimates of the squared population cross-validity coefficient, ρ_c^2. The values of R_c^2, the sample estimates of cross-validity, will vary from sample to sample; however, the expected value of R_c^2 (the average over many samples), approximates ρ_c^2. The cross-validity coefficient can be thought of as the squared correlation between the actual population criterion values and the criterion scores predicted by the sample regression equation when applied to the population or to another sample (Kennedy, 1988; Schmitt et al., 1977).

Insert Table 1 about here

Because R^2 is a positively biased estimator of both ρ^2 and ρ_c^2, such that $E(R^2) > \rho^2 > \rho_c^2$, researchers must report an appropriate shrunken R^2 (Herzberg, 1969). An estimate of ρ^2 is rarely useful for a researcher interested in developing a regression equation to be used for prediction; for prediction purposes, researchers should report $\hat{\rho}_c^2$ (Cattin, 1980b; Huberty & Mourad, 1980). It should be noted, however, that as the number of subjects increases relative to the number of predictors, both R^2 and ρ_c^2 converge toward ρ^2, and therefore the amount of shrinkage decreases (Cattin, 1980a). Similarly, the overestimation of ρ^2, and thus the shrinkage, is greatest when both N and R^2 are small (Cohen & Cohen, 1983; Schmitt et al., 1977).

The most common estimate of shrinkage reported in the literature is an adjusted R^2 that is attributed most frequently to Wherry (1931). This formula for "adjusted" R^2,

$\mathbf{R_a^2 = 1 - (1-R^2)(N-1)/(N-p-1)}$, is an unbiased estimate of ρ^2, however, not of ρ_c^2. Therefore, not only is Wherry's estimate not practical for prediction studies, but it also overestimates ρ_c^2 and thus is incorrect (Darlington, 1968). This is unfortunate because the most popular computer statistical packages, SAS, SPSS, and BMDP, use shrinkage estimates of ρ^2 rather than cross-validity estimates of ρ_c^2 (Kennedy, 1988; Stevens, 1992a). Indeed, Uhl and Eisenberg (1970) found that a cross-validity estimate (which they attribute to Lord, 1950) was consistently more accurate than Wherry's shrinkage formula. Therefore, researchers should report a cross-validity coefficient in prediction studies. Some of the more familiar cross-validity formulas are those by Stein (1960), Lord (1950), Nicholson (1960), and Browne (1975).

Methods for Selecting Sample Sizes

The distinction between the fixed and random models becomes particularly relevant when analyzing power and selecting sample sizes. Indeed, only Park and Dudycha (1974), Sawyer (1982), and Gatsonis and Sampson (1989) have discussed the random model directly for sample size calculations. Cohen and Cohen (1983), Kraemer and Thiemann (1987), Cohen (1988), and others have approached power from a fixed model approach. The differences between the two models in regard to power and sample size selection are explained in more detail in the following section. First, however, some comments are made about the many rules-of-thumb.

Rules-of-Thumb for Selecting Sample Size

The most extensive literature regarding sample sizes in regression analysis is in the area of cross-validation. Many scholars have suggested rules-of-thumb for choosing sample sizes that they claim will provide reliable estimates of the population regression coefficients. This ability of regression coefficients to generalize to other samples from the same population may be considered predictive power. The most common method for analyzing this predictive power is through shrinkage or cross-validation estimates. A review of Table 1 shows that all shrinkage estimates are functions of the number of subjects and the number of predictors.

Therefore, scholars have recommended subject-to-variable ratios that purportedly will provide an "acceptable" amount of shrinkage. That is, with a large enough ratio of subjects to predictors, the estimated regression coefficients will be reliable and will closely reflect the true population parameters since shrinkage will be slight (Miller & Kunce, 1973; Pedhazur & Schmelkin, 1991; Tabachnick & Fidell, 1989). Put another way, a larger subject-to-variable ratio is required for higher predictive power. These rules-of-thumb typically take the form of a subject-to-predictor (N/p) ratio. For example, Table 2 shows that statisticians have recommended using as small a ratio as 10 subjects to each predictor and as large a ratio as 40:1. Harris (1985) noted that ratio rules-of-thumb clearly break down for small numbers of predictors. That is, if only two predictors are used in a study, a 10:1 rule would require only 20 subjects, usually far too few for estimating reliable regression coefficients.

Insert Table 2 about here

Because of such obvious shortcomings of subject-to-variable ratios, some scholars have suggested that a minimum of 100, or even 200, subjects is necessary regardless of the number of predictors (e.g. Kerlinger & Pedhazur, 1973). Harris indicated that no systematic studies have been performed to analyze the use of a ratio rule versus a difference rule, say **N - p > 50**. More recently, however, Green (1991) did find that a combination formula such as **N > 50 + 8p** was much better than a subject-to-variable ratio. More complex rules-of-thumb have been developed by Green (1991) and Sawyer (1982) that do account for some measure of effect size. Unfortunately, Green simply has developed a formula for Cohen's (1988) tables. Sawyer's method, although mathematically more elegant, uses an unintuitive "inflation factor" for mean squared error rather than a more recognizable effect size. Finally, perhaps the most widely used rule-of-thumb was described by Olejnik (1984): "use as many subjects as you can get and you can afford" (p. 40).

Fixed Model Approach

"The power of a statistical test is the probability that it will yield statistically significant results" (Cohen, 1988, p. 1). That is, statistical power is the probability of rejecting the null hypothesis when the null hypothesis is indeed false. Statistical power analysis requires the consideration of four parameters: level of significance, power, effect size, and sample size. These four parameters are related such that when any three are fixed, the fourth is mathematically determined (Cohen, 1992). Therefore, it becomes obvious that it is necessary to consider power, alpha, and effect size when attempting to determine a proper sample size. The following section examines sample size requirements from the statistical power, fixed model framework from which Cohen (1988) and Cohen and Cohen (1983) proceed. Recall that this fixed model approach, however, is most useful when researchers use regression as a means to explain the variance of a phenomenon in lieu of analysis of variance.

Overall Test of the Regression Model. In any statistical analysis, there are three strategies for choosing an appropriate effect size: (a) use effect sizes found in previous studies, (b) decide on some minimum effect that will be practically significant, or (c) use conventional small, medium, and large effects (Cohen & Cohen, 1983). Cohen (1988) defined effect size in fixed model multiple regression as a function of the squared multiple correlation, specifically $\mathbf{f^2 = R^2/(1-R^2)}$. Since R^2 can be used in the formulas directly, Cohen also defined effect sizes in terms of R^2 such that small effect R^2=.02, medium effect R^2=.13, and large effect R^2=.26.

In order to calculate the required sample size to reach a desired level of power for testing the significance of R^2, Cohen's (1988) Case 0, a researcher needs the following information: (a) level of significance, α, (b) degrees of freedom for the numerator of the F ratio, u, which is the number of independent variables, (c) degrees of freedom for the denominator of the F ratio, v, and (d) desired power. Sample size is calculated as $\mathbf{N = \lambda(1-R^2)/R^2}$, where λ is the noncentrality parameter required for the noncentral F-distribution. Cohen's (1988) tables provide the λ needed for the sample size formula.

It should be noted that computing the degrees of freedom for the denominator of the F ratio (parameter v) can be problematic, since v is a function of the yet-to-be-determined N (i.e., v = N-u-1). However, Cohen (1988) suggested that a trial value of v=120 will usually yield sufficient accuracy. Furthermore, it should be noted that the newer Cohen (1988) tables differ from previous tables in this regard. Neither Cohen (1977) nor Cohen and Cohen (1983) required the determination of the parameter v. Examination of the tables reveals that the earlier tables are equivalent to the Cohen (1988) tables with v fixed at ∞ (see footnote 3 in Cohen, 1988, p. 551, for an explanation). Therefore, in the earlier editions, the sample size formula differs slightly from the newer formula: $\mathbf{N = [L(1-R^2)/R^2]+u+1}$, where L is the noncentrality parameter and u is the number of predictors. Cohen (1992) presented an abbreviated sample size table that provides sample sizes directly.

Green (1991) developed a rule-of-thumb based on Cohen's power analysis approach to sample size selection. However, Green's rule is valid only for Power=.80 and α=.05, and is most effective for moderate R^2 estimates. Sample size is defined using the Cohen formula: $\mathbf{N \geq L(1\text{-}R^2)/R^2}$ (Green, 1991, p. 504). The rule-of-thumb developed by Green is a method for approximating L, so that researchers can estimate sample size without having to consult a table for λ. The method for approximating L follows these steps: (a) for the first predictor, L=8, (b) for the second through tenth predictors, L increases with each additional predictor by 1.5, 1.4, 1.3, 1.2, 1.1, 1.0, 0.9, 0.8, and 0.7, respectively (algebraically, $L=6.4+1.65m-0.05m^2$ for m<11), (c) for each additional predictor after the tenth, L increases by 0.6. Through comparison of a variety of rules-of-thumb, Green concluded that the rule-of-thumb of $\mathbf{N \geq 50 + 8p}$ was more accurate than the simpler rules (e.g. N ≥ 10p) for the case of medium effect with Power=.80. However, Green found his approximation of Cohen's tables to be the most accurate rule-of-thumb.

Test of the Individual Predictors. Other scholars have addressed the issue of power in regard to the test of the relationship between the individual independent variables and the dependent variable. These scholars have taken an approach to selecting sample size based on the statistical power of the t-test used to test the partial regression coefficients (Kraemer & Thiemann, 1987; Milton, 1986; Neter, Wasserman, & Kutner, 1990). Cohen (1988) called this circumstance Case 1-1, where the null hypothesis tested is concerned with the unique contribution of a single independent variable to R^2. Cohen reminds us that the results of the F-test for proportion of variance of a particular predictor are identical to the results of the t-test performed on the partial regression coefficient; each provides a test of the independent variable's unique contribution to the criterion. Cohen (1988) also provides several other power analyses for partial regression models and coefficients.

Tests of the individual predictors may be useful in selecting predictors to include in a final model or in a regression analysis performed to analyze variance. However, these tests are not useful for those social scientists who wish to predict scores on some criterion or simply to describe an overall relationship. Therefore, the random model approach, which is more useful in prediction studies, is addressed in the next section.

Random Model Approach.

The random model of regression recognizes and accounts for extra variability because, in another replication, different values for the independent variables will be obtained (Gatsonis & Sampson, 1989). Because it is not known which specific values for the independent variables will be sampled on successive replications, Park and Dudycha (1974) took a cross-validation approach to calculating sample sizes. They noted that such a cross-validation approach is applicable to both the random and the fixed models of regression; however, because the fixed model poses no practical problems, they emphasized the random model. Park and Dudycha derived the following sample size formula: $\mathbf{N \geq [(1-\rho^2)\delta_1^2/\rho^2] + p + 2}$, where δ_1^2 is the noncentrality parameter for the t-distribution. The fixed model formula they derived was $\mathbf{N \geq (1-\rho^2)\delta_1^2/\rho^2}$ (note the similarity to Cohen's formula, with only a difference in the noncentrality parameter used in derivation). Furthermore, the random model only differs from the fixed model by (p+2); that is, the random model requires (p+2) more subjects than the fixed model according to the Park and Dudycha calculations.

The basic premise of Park and Dudycha's (1974) method of sample size selection is that researchers can estimate how close they want to estimate ρ from ρ_c. That is, researchers determine the probability with which they want to approximate ρ within some chosen error tolerance. The formula for this probability is: $P(\rho - \rho_c \leq \epsilon) = \gamma$. The researcher chooses (a) an expected ρ^2 as the effect size, (b) the error willing to be tolerated, ϵ, and (c) the probability of being within that error bound, γ. The tables provided by Park and Dudycha (most of which were reprinted in Stevens, 1986, 1992a) can then be consulted with these values. It should be noted that Park & Dudycha's tables were one factor that led Stevens to suggest a 15:1 subject-to-variable ratio as a rule of thumb.

Even though Gatsonis and Sampson (1989) calculated tables for the random model, they concluded that Cohen's (1977) approximation of the conditional model was also an adequate approximation to the unconditional model. After comparing their own tables to Cohen's (1977) tables, Gatsonis and Sampson determined that Cohen's approximations generally underestimated the exact required sample sizes only slightly. Gatsonis and Sampson recommended that researchers add five to Cohen's tabled values, especially for models with up to 10 predictors.

The Predictive Power Method for Selecting Sample Sizes

The most profound problem with many rules-of-thumb advanced by regression scholars is that they lack any measure of effect size. It is generally recognized that an estimated effect size must precede the determination of appropriate sample size. Effect size enables a researcher to determine in advance not only what will be necessary for statistical significance, but also what is required for practical significance (Hinkle & Oliver, 1983).

The problem with Cohen's (1988) method, and Green's (1991) formula based on Cohen's method, is that it is designed for use from a fixed model, statistical power approach. And although Gatsonis and Sampson (1989) use the random model approach, their method is also based on a statistical power approach to sample size determination. Unfortunately, statistical power to reject a null hypothesis of zero multiple correlation does not inform us how well a model will predict in other samples. That is, adequate sample sizes for statistical power tell us nothing about the number of subjects needed to obtain stable, meaningful regression weights (Cascio, Valenzi, & Silbey, 1978).

On the other hand, Park and Dudycha (1974) take a random model, cross validation approach. However, their tables are limited to only a few possible combinations of sample size, squared correlation, and epsilon; and unfortunately, their math is too complex for most researchers to derive the information they would need for the cases not tabulated. Additionally, there is no clear rationale for how to determine the best choice of either epsilon or the probability to use when consulting the tables (although Stevens, 1992a, implicitly offered .05 and .90, respectively, as acceptable values).

By combining elements from the cross-validation and shrinkage literature, the rules-of-thumb literature, and statistical power analysis literature, it was possible to devise a new sample size formula that does include effect size as part of its calculation. Recall that shrinkage and cross-validation formulas do include R^2, an adequate effect size value according to Cohen (1988). Therefore, the literature was explored for an acceptable cross-validation formula that could be manipulated algebraically to become a sample size formula. Additionally, a basic premise which allows researchers to decide how closely to estimate ρ_c^2 from expected R^2 was adapted from Park and Dudycha's (1974) method. Finally, a new method for selection of sample sizes in multiple linear regression was developed.

The formula developed by Rozeboom (1978) was determined to be the most adequate for present purposes. Rozeboom's cross-validation formula is a version of a widely-accepted formula developed independently by Lord (1950) and Nicholson (1960) that is "even tidier" (Rozeboom, 1978, p. 1350) because it is linear in all parameters. According to Rozeboom, the Lord-Nicholson formula works well in practice as applied from either the fixed model or the random model perspective (see Table 2 for the Lord-Nicholson formula). Additionally, at least one respected regression text has offered the Rozeboom formula as the recommended cross-validation formula (Cohen & Cohen, 1983). The Rozeboom (1978, p. 1350) formula is:

$$\rho_c^2 = 1 - \{ [(N+p)(1-R^2)] / (N-p) \}, \qquad (1)$$

where N is sample size, p is the number of predictors, and R^2 is the actual sample value. Manipulation of this formula to solve for N yields:

$$N = [p(2 - \rho_c^2 - R^2)] / (R^2 - \rho_c^2), \qquad (2)$$

where p is the number of predictors, R^2 is the expected sample value, and ρ_c^2 is the estimated population cross-validity value. The quantity $(R^2 - \rho_c^2)$ is the amount of shrinkage that will occur if the N calculated in equation (2) is used to calculate shrinkage with equation (1). Substituting

($\epsilon = R^2 - \hat{\rho}_c^2$) and therefore ($\hat{\rho}_c^2 = R^2 - \epsilon$), into formula (2) to represent acceptable shrinkage, we get

$$N \geq \{p\ [2 - (R^2 - \epsilon) - R^2]\} / \epsilon, \tag{3}$$

where p is the number of predictors, R^2 is the expected sample value, and ϵ is the acceptable absolute amount of shrinkage. Finally, simplifying formula (3) gives us

$$N \geq [p\ (2 - 2R^2 + \epsilon)] / \epsilon. \tag{4}$$

Because this formula is based on a cross-validity formula, it is expected to provide good predictive power when used to calculate sample sizes. Like Park and Dudycha's (1974) method, the new method allows flexibility in the choice of acceptable shrinkage; that is, one can substitute any appropriate value for ϵ into formula (4) such as an absolute value (like .03 or .05) or a proportional value (like $.8R^2$). For example, if a researcher wanted an estimate of $\hat{\rho}_c^2$ not less than 80% of the sample R^2 value, the formula can be reformulated such that $\epsilon = .2R^2$:

$$N \geq [p\ (2 - 1.8R^2)] / 0.2R^2. \tag{5}$$

If the researcher did not want the sample R^2 to decrease by more than .05 no matter what the expected value of R^2, formula (4) simplifies to

$$N \geq 20p\ (2.05 - 2R^2); \tag{6}$$

or if the researcher did not want the sample R^2 to decrease by more than .03, then

$$N > 33p\ (2.03 - 2R^2). \tag{7}$$

Although there is intuitive appeal to such a simple method for determining sample sizes in multiple linear regression, there is no way to compare this method to current methods mathematically. Therefore, a Monte Carlo power study was performed to determine the efficacy of the new method as compared to existing methods. Additionally, an attempt was made to make sense of the conflicting values provided by several preeminent methods. The next section describes this study in detail.

METHODS

Ideally, a mathematical proof would be provided that would compare directly the efficacies of existing sample size methods and the new method offered in this paper. However, the several sample size selection methods compared here are based on different probability distributions, making direct comparison problematic. For example, Park and Dudycha (1974) base their work on the probability density function of ρ_c^2, Cohen (1988) bases his material on the noncentral χ^2 distribution (Gatsonis & Sampson, 1989), Gatsonis and Sampson (1989) base their method on the distribution of R_{xy}, and rules-of-thumb are not based on probability distributions at all.

Fortunately, meaningful comparisons among the power rates of these methods can be accomplished through a Monte Carlo study. Monte Carlo methods use computer assisted simulations to provide evidence for problems that cannot be solved mathematically. In Monte Carlo power studies, random samples are generated and used in a series of simulated experiments in order to calculate empirical power rates. That is, many random samples are generated such that the null hypothesis is known to be false (e.g. the multiple correlation is non-null) and then the actual number of tests that are correctly rejected are counted. After all samples are completed, a proportion is calculated that represents the actual power rate.

While several scholars have used the term predictive power (e.g. Cascio et al., 1978; Kennedy, 1988), only Cattin (1980a) has provided a formal definition. Cattin (1980a) noted that the two common measures of predictive power are the mean squared error of prediction and the cross-validated multiple correlation. However, Cattin was discussing predictive power in regard to the comparison and selection of competing regression models. Although Cattin's definition could be applied to the current circumstances, it would not provide a measure at all similar to our general understanding of power. Therefore, for present purposes, predictive power was defined as $\boldsymbol{\rho_c^2/R^2}$ or **1 - pd**, where pd is the percentage decrease in the squared correlation after an appropriate cross-validity shrinkage estimate is made. For example, a predetermined acceptable level of shrinkage of 20% provides predictive power equal to .80. Indeed, the method which produces the highest predictive power using the current definition will also yield the largest average cross-validity coefficient, thereby satisfying Cattin's definition as well.

The Stein (1960) cross-validity formula (sometimes attributed to Darlington, 1968 and Herzberg, 1969) was used for ρ_c^2, because it is has been recommended by many scholars who have investigated cross-validation techniques from a random model perspective (e.g. Claudy, 1978; Huberty & Mourad, 1980; Kennedy, 1988; Schmitt et al., 1977; Stevens, 1986, 1992a). It should be noted that the authors are aware that the Stein formula is not uniformly regarded as the best cross-validation formula (e.g. Cattin, 1980a; Drasgow et al., 1979; Rozeboom, 1978). Statistical power was calculated as the proportion of total number of correct rejections to the total tests performed for each testing situation.

Because a variety of factors influence predictive power, several testing situations were considered. Four factors were manipulated and fully crossed for the present study. First, three effect sizes were used which represented the expected R^2, that is, the assumed population ρ^2: .10, .25, and .50. The .10 and .25 values were chosen because they are found in Park and Dudycha's (1974) tables and because they are very close to Cohen's (1988) medium and large effect sizes of

.13 and .26, respectively. The .50 value was chosen because Stevens (1992a) recommends it as "a reasonable guess for social science research" (p. 125). Second, data were generated for five sets of predictors: 2, 3, 4, 8, 15. Again, these numbers were chosen for ready comparison with tables provided by both Park and Dudycha (1974) and Gatsonis and Sampson (1989). Third, five separate ranges for the true population ρ^2 were used: .001-.04, .04-.16, .16-.36, .36-.64, and .64-.999. Correlation matrices were created with R^2 values in these ranges using a procedure described below. These particular ranges were chosen primarily because three of their midpoints are very close to the effect sizes chosen above.

Finally, seven sample size selection methods were compared: the new method offered in this paper, Park and Dudycha (1974), Cohen (1988), Gatsonis and Sampson (1989), the 30:1 subject-to-variable ratio from Pedhazur and Schmelkin (1991), the **N > 50 + 8p** formula from Green (1991), and the 15:1 ratio from Stevens (1992a). The relevant sample size tables from both Park and Dudycha and Gatsonis and Sampson were stored as data for access by the computer program, as were the appropriate tables for Cohen's lambda values. Because of the resultant simplicity of formulas (6) and (7), the amount of acceptable shrinkage for the new method, ϵ, was set absolutely. Both the new method and Park and Dudycha's tables were accessed with ϵ=.03 for expected $R^2 \leq .10$ and ϵ=.05 for expected $R^2 > .10$; additionally for Park and Dudycha, $P(\rho-\rho_c \leq \epsilon)$= .90. Both Cohen's and Gatsonis and Sampson's tables were entered using power=.90. Turbo Pascal code was written to calculate the sample sizes for the new method, the ratio methods, the combination method from Green, and Cohen's method (after looking up tabulated lambda values). It should be noted that for the case of expected R^2=.50, the tabulated sample size for ρ=.70 from Gatsonis and Sampson was used; for the case of expected

R^2=.10, the ρ=.30 value was chosen. Because in each of these cases the ρ value used was less than the square root of the expected R^2, the sample sizes chosen for the Gatsonis and Sampson method were slightly larger than exact values would have provided. The seven methods do provide a variety of suggested sample sizes, sometimes drastically different (see Table 3).

Insert Table 3 about here

A Turbo Pascal 6.0 program was written that generated and tested 10,000 samples for each of these 525 conditions. The program was run as a MS-DOS 6.2 application under Windows 3.1 on a computer equipped with an Intel DX2/40 processor, which has a built-in numeric coprocessor. Extended precision floating point variables, providing a range of values from $3.4\text{x}10^{-4932}$ to $1.1\text{x}10^{4932}$ with 19 to 20 significant digits, were used. For each sample, the program performed a standard regression analysis (all predictors entered simultaneously), calculated the F-statistic and its probability, tested the null hypothesis of zero correlation at a .05 significance level, and calculated Wherry (1931) shrinkage and Stein (1960) cross-validity coefficients. Because the null hypothesis (H_0: ρ=0) was known to be false in each sample, each rejection at a .05 significance level qualified as a correct rejection and was recorded as such. For each condition, then, empirical statistical rates were calculated simply as the proportion of the 10,000 tests that were correctly rejected. Also for each condition, average shrinkage and average cross-validity were calculated. Additionally, predictive power for each condition was calculated as the ratio of the average Stein cross-validity coefficient to the average sample R^2. Finally, these summary data were compared to determine how well the sample size methods performed both absolutely and relatively. Simulated samples were chosen randomly to test program function by comparison with results provided by SPSS/PC+ version 5.0.1.

DATA SOURCE

Because this research focused on power for the random model of regression, data were generated to follow a joint multivariate normal distribution. The first step was to create population correlation matrices that met the criteria required by this study, namely, appropriate numbers of predictors and appropriate ρ^2 values. These correlation matrices were then used to generate multivariate normal data following a procedure recommended by several scholars (Chambers, 1977; Collier, Baker, Mandeville, & Hayes, 1967; International Mathematical and Statistical Library, 1985; Karian & Dudewicz, 1991; Kennedy & Gentle, 1980; Keselman, Keselman, & Shaffer, 1991; Morgan, 1984; Ripley, 1987; Rubinstein, 1981).

For each range of ρ^2 and number of predictors (25 total conditions), a correlation matrix was created using the following procedure. Uniform random numbers between 0.0 and 1.0 were generated using an algorithm suggested by Knuth (1981) and coded in Pascal by Press, Flannery, Teukolsky, and Vetterling (1989). These values were entered as possible correlations into a matrix and the squared multiple correlation, R^2, was calculated. If the R^2 value fell in the required range, the matrix was then tested to determine whether it was positive definite. Press, Teukolsky, Vetterling, and Flannery (1992) suggested that the Cholesky decomposition is an efficient method for performing this test -- if the decomposition fails, the matrix is not positive definite. The algorithm for the Cholesky decomposition used in this study was adapted from Nash (1990). This procedure was repeated until the necessary 25 matrices were created. These correlation matrices were then used to generate the random samples as described below. It is worthwhile to note that with given values of R^2, sample size, and numbers of predictors, the distribution of the squared cross-validity coefficient does not depend on the particular form of the population covariance, or in this case correlation, matrix (Drasgow et al., 1979).

The Cholesky decomposition of a matrix produces a lower triangular matrix, L, such that $LL^T = \Sigma$, where Σ is a symmetric, positive definite matrix such as a covariance or correlation matrix. This lower triangular matrix, L, can be used to create multivariate pseudorandom normal variates through the equation

$$Z_{ij} = \mu_j + XL^T \tag{7}$$

where $\mathbf{Z}_{ij}$ is the multivariate normal data matrix, $\boldsymbol{\mu}_j$ is the mean vector, and $\mathbf{X}$ contains vectors of independent, standard normal variates. When $\boldsymbol{\mu}_j=\mathbf{0}$, the multivariate pseudorandom data is distributed with mean vector zero and covariance matrix Σ. Independent pseudorandom normal vectors, $\mathbf{X}_j$, with means, zero, and variances, unity, were generated using an implementation of the Box and Muller (1958) transformation adapted from Press, Flannery, Teukolsky, and Vetterling (1989). The Box and Muller algorithm converts randomly generated pairs of numbers from a uniform distribution into random normal deviates.

RESULTS AND CONCLUSIONS

Results

The seven methods of sample size selection were compared for three cases: (1) where expected R^2, $E(R^2)$, fell in the same range as the population ρ^2, (2) where $E(R^2) > \rho^2$, and (3) where $E(R^2) < \rho^2$. Data were collapsed over the number of predictors for practical reasons: (a) to provide a manageable number of comparisons, (b) because an acceptable method must be viable for any number of predictors, and (c) because within each $E(R^2)$ level the relative rankings for the methods that include effect size were fairly consistent across numbers of predictors[1]. After the results have been summarized, the methods will be discussed in terms of both relative effectiveness and absolute efficacy.

Expected $R^2 \approx \rho^2$

Beginning with Stevens' (1992a) recommended assumption of ρ^2=.50, all methods produced adequate predictive power over .80 except Cohen (COHEN) and Gatsonis and Sampson (GS). The five other methods were significantly different from COHEN and all others except the 15:1 rule (NP15) also were different from GS (see Table 4). Further, none of the remaining five methods were significantly different from another of the five. Although there was not as much discrepancy in the empirical statistical power rates as there was in the predictive power rates, COHEN differed significantly from the other methods (see Table 5). All other methods provided empirical statistical power rates over .95.

Insert Table 4 about here

Using nearly what Cohen (1988) called a large effect, $E(\rho^2)$=.25, the results were less favorable for all methods (see Table 4). The only method to provide predictive power over .80 was the new method (BB). However, a 95% confidence interval shows that BB is not significantly better than NP15, Park and Dudycha (PD), the 30:1 rule (NP30), or the **$N \geq 50 + 8p$** rule (COMBO). Once again, though, COHEN was significantly lower than four other methods (BB, NP30, PD, and COMBO) and GS was significantly lower than both BB and NP30. All methods showed statistical power greater than .85 and no two differed significantly (see Table 5).

Insert Table 5 about here

Finally, using what is often considered the smallest shared variance to be of practical significance, 10% or $E(R^2)$=.10, the relative ordering of the methods changed dramatically (see Table 4). NP15, which does not take effect size into account, showed significantly less predictive power than the four methods that do consider effect size (BB, PD, COHEN, GS). Only BB and PD were significantly different from COMBO, which also does not take effect size into account. The remaining methods were not significantly different in their predictive power. Even though NP15 and COHEN were the only methods with statistical power rates below .80, they were not significantly lower than the other methods (see Table 5).

Expected $R^2 > \rho^2$

When expected R^2 is greater than the true population ρ^2, all methods fail miserably at all levels of $E(R^2)$. An examination of Table 4 shows that the only method with predictive power for any $E(R^2)$ over .40 is NP30; Table 5 shows that the only methods with statistical power over .60 are BB and NP30. It is important to recognize that the results presented in Table 4 and Table 5 for $E(R^2)>\rho^2$ are aggregated for all values of ρ^2 below the relevant $E(R^2)$. For example, when $E(R^2)$=.25, two ranges of true ρ^2 fall below $E(R^2)$. Therefore, individual conditions are examined in the following paragraphs.

The results are most dramatic for the lowest range, $.001<\rho^2<.04$. Table 6 shows that for all methods and both $E(R^2)$=.50 and $E(R^2)$=.25, predictive power rates are less than .02! Indeed, examination of each predictor level at $E(R^2)$=.50 reveals that predictive power is zero for all cases except NP30 with 15 predictors; at $E(R^2)=.25$, only BB and NP30 had any predictive power rates greater than zero, again when there are 15 predictors. Even for $E(R^2)=.10$, where we might expect somewhat better results, BB shows predictive power of only .22 and PD shows .07; all others are under .03. Similarly, statistical power rates are below .40 for all methods under the conditions of $E(R^2)=.50$ and $E(R^2)=.25$. Although, BB provides statistical power of .61 in the $E(R^2)=.10$ case, all others fall below .50 (see Table 7).

Insert Table 6 about here

As the true ρ^2 range gets closer to the $E(R^2)$ value, the results improve, slightly. Predictive power rates for all methods remain under .50 for the case where $E(R^2)=.50$ and $.04<\rho^2<.16$ (see Table 6). More specifically, NP30 has predictive power greater than .46 and BB is over .33, but the rest remain below .30. Statistical power improves significantly for this condition, however, as NP30 has power over .80, BB and COMBO over .70, and PD over .62. Indeed, all methods show statistical power over .50 except COHEN and GS, which are significantly lower than the rest. For the condition of $E(R^2)=.25$ and the range $.04<\rho^2<.16$, BB and NP30 have predictive power over .48 and .46, respectively; PD is above .34, but all the rest are below .30. Once again, statistical power is better than predictive power as shown by power rates of .81 for both BB and NP30, .73 for PD, and .71 and .60 for COMBO and NP15, respectively (see Table 7).

Insert Table 7 about here

Finally, for the condition where $E(R^2)$=.50 and $.16<\rho^2<.36$, things improve even further. NP30 has predictive power over .79, BB is over .71, and COMBO is over .69. Both PD and NP15 have predictive power over .60, while COHEN and GS are again significantly lower than then rest, both being below .13 (see Table 6). All statistical power rates climb to over .90 except GS and COHEN, which show empirical power rates of .51 and .27, respectively (see Table 7).

Expected $R^2 < \rho^2$

The final case concerns conditions where the expected R^2 was less than the true population ρ^2. In this case, more subjects usually will be sampled than necessary. Indeed, all methods show averaged results with acceptable predictive power and adequate statistical power (only COHEN falls below .80, for predictive power when $E(R^2)$=.50). Indeed, examination of specific results shows that as ρ^2 increases relative to $E(R^2)$, power increases to a point where one is relatively certain to get both sufficient statistical power and reliability of regression coefficients.

Discussion

It was expected that the sample size methods that use effect size would be most appropriate for the situations where $E(R^2)$ fell in the same range as the true population ρ^2. Indeed, this was the case and is a critical finding of the current research. Subject-to-variable ratios, and other rules-of-thumb that do not account for any measure of effect size, show adequate power both relatively and absolutely for moderate-to-large $E(R^2)$. For example, Stevens' (1986, 1992a) suggestion of a 15:1 subject-to-predictor ratio does indeed provide empirical power rates similar to Park and Dudycha's (1974) method -- when $E(R^2) \approx .50$. However, as one expects a smaller R^2, the rules-of-thumb become inadequate. For example, with $E(R^2)=.50$, NP30 provided the highest empirical power ranking; however, as $E(R^2)$ was reduced to .25 and .10, NP30 was only the second best and then the fifth best, respectively (see Table 8). Contrast this with the COHEN method which moved from the bottom ranking when $E(R^2)=.50$ and $E(R^2)=.25$ to fourth, better than all methods without effect size, when $E(R^2)=.10$. Additionally, the miserable results for all methods when $E(R^2)>\rho^2$, and also the overkill when $E(R^2)<\rho^2$, attest to the importance of effect size. Therefore, the first conclusion to be made is that researchers cannot ignore effect size in determining sample size in multiple regression analysis any more than they can for any other statistical design.

Insert Table 8 about here

When one can make a reasonable guess at the population ρ^2, either from past research, personal experience, or based on practical significance, one has a good chance of determining the number of subjects necessary to have adequate power to detect that value as significant in a sample. More importantly perhaps, one has a much better chance of deriving regression coefficients that will be meaningful and stable when the regression model is used for prediction purposes. Indeed, these two types of power go hand-in-hand; it was determined empirically in the present study that predictive power and statistical power have a correlation of .9083 ($p<.001$). Still, the absolute values of the two types of power differ relative to what might be considered an acceptable minimum level; specifically, statistical power rates were often above .80 for cases where predictive power was still much less. The second main conclusion to be drawn in the present research, then, is that, given the high correlation between the two types of power, if researchers choose sample sizes to meet predictive power guidelines, they will be practically assured of sufficient statistical power for their study.

Finally, there is the question of which method best ensures predictive, and therefore statistical, power. First, the sample size selection method developed in this paper compared favorably to methods from Cohen (1988), Gatsonis and Sampson (1989), Park and Dudycha (1974), and three rules-of-thumb. In fact, the new method was the most consistent, best performer: it had either the highest or second highest mean rank for both predictive and statistical power for each level of $E(R^2)$. Further, the new method ranked first overall; specifically, for 58 of the 75 combinations of $E(R^2)$, predictors, and range of ρ^2, it ranked first or second for predictive power. However, just as important as how it performed relative to the other methods is how the new method performed absolutely.

Given the large number of replications (10,000) performed for each condition, one could argue that the empirical power rates observed for each method represent close approximations to true power rates. For several reasons, results were averaged across number of predictors; however, if the results are analyzed from an absolute perspective, one finds that the new method most often exceeds acceptable power levels. For example, the most important situation concerns the cases where $E(R^2) \approx \rho^2$. In these conditions, the new method reached .80 for predictive power in 10 out of the 15 cases; it exceeded .70 in all but one situation (see Table 9). For comparison, NP30 exceeded .80 in eight cases, COMBO six, and PD and NP15 five; both PD and NP30 exceeded .70 in 10 cases (however, NP30 never reached .70 for $E(R^2)$=.10 and PD only did once).

Insert Table 9 about here

The new method is preferable to subject-to-variable ratios and other rules-of-thumb for three reasons: (a) rules-of-thumb do not account for effect size and therefore have limited value as ρ^2 decreases, (b) despite Stevens' (1986, 1992a) suggestion of using ρ^2=.50, it may be more likely that ρ^2=.50 is an upper bound for social science research (Rozeboom, 1981), and (c) the new method is just as simple to use in its formula (6) form and provides consistently better results. The new method is preferable to both Cohen's method and Gatsonis and Sampson's method for all conditions tested here. Although both COHEN and GS performed relatively well for $E(R^2)$=.10, the new method still performed better relatively and was the only acceptable method absolutely.

Thus, the only remaining comparison is between the new method and Park and Dudycha's (1974) method. Because preliminary work showed that the new formula approximated Park and Dudycha's values at between 90% and 95% probability for most predictor tables, there was hope that the method would perform well in this study. Indeed, this favorable comparison was crucial for two reasons: (a) because both methods are based on a cross-validation approach and (b) because the Park and Dudycha method has mathematical derivations to support it. Indeed, the Monte Carlo results found here suggest that both methods perform very well both absolutely and also relative to other methods. The only other method that came close was NP30, at the cost of overly large samples in the $E(R^2)$=.50 case (in fact, most methods provided sample sizes too large for that situation). The advantages that the new method provides in consistently higher power rates are offset sometimes by consistently higher recommended sample sizes. As Tabachnick and Fidell (1989) wrote, "for both statistical and practical reasons, then, one wants to measure the smallest number of cases that has a decent chance of revealing a significant relationship if, indeed, one is there" (p. 129). Of course, sometimes these larger sample sizes are required. In particular, as $E(R^2)$ decreases, the larger samples required by the new method are necessary.

The researcher must balance the two concerns for the particular problem at hand. However, the preliminary work done for this study suggested that, for larger expected R^2 values, it may be possible to use a less conservative value for ϵ, say .075 or .10, and still approximate Park and Dudycha's (1974) tables at nearly the 90% probability level. Indeed, a supplementary analysis using formula (5) showed that when $E(R^2)=.50$, and therefore $\epsilon=.10$, the new method always provided empirical predictive power over .70 (over .80 in four of five cases) with only 52% of the sample size required. Indeed, formula (5) had power over .80 in 11 of the 15 cases. Unfortunately, the formula lost power rapidly as $E(R^2)$ increased beyond .50; therefore, it appears that .10 is a maximum threshold for ϵ.

The new method thus provides not only better recommendations but also quite a bit more flexibility than any other method presented in this paper. Acceptable levels of shrinkage can be approached two ways, as absolute levels (as in formulas 6 and 7) or as percentage decreases (as in formula 5). Note that although this level of shrinkage will always underestimate the actual shrinkage for the Stein and Lord-Nicholson cross-validity formulas, it does provide the most reasonable approximation. Therefore, the final recommendation, based on this research, is that either the new method developed here or Park and Dudycha's (1974) method will provide adequately powerful sample size for most situations. Indeed, they are the only two methods that work relatively well in all circumstances. However, when a reasonable effect size cannot be approximated or when $E(R^2)$ is small, the new method will provide more conservative, and therefore -- some would argue -- better, results.

Suggestions for future research

Although this study has limitations typical of Monte Carlo studies, it has provided important insight into sample size selection and power analysis in multiple regression. Further research can be performed in this area using similar methods, or using real data. For example, a fixed-model perspective can be tested. One would expect similar results for predictive power, but the fixed-model methods might prove more useful. Also, the data can be manipulated in several ways. For example, using non-normal data or multicollinear or suppressor variable relationships. In this study, the standard regression approach was used in which all predictors are entered simultaneously and each is tested as if it were the last to enter; in future studies hierarchical models can be studied. It was assumed that predictors were selected a priori; that is, it was known in advance of building the regression model which independent variables would be included in the analysis. However, it may be possible to study sample sizes required for the use of preselection processes (e.g. stepwise regression used to select the best subset of predictors).

EDUCATIONAL IMPORTANCE OF THE STUDY

The research presented in this paper is important for the reasons mentioned at the outset. In particular, sample sizes for multiple linear regression must be chosen so as to provide adequate power for both statistical significance and generalizability of results. It is well-documented and unfortunate that researchers do not heed this guideline. Type I errors are treated with extreme caution, but Type II errors are all but ignored until the research fails to find results; then lack of power often is used to rationalize the failure.

For whatever reason, empirical study into power for multiple regression has been lacking. Rules-of-thumb have existed for many decades, but little empirical or mathematical support has been offered for them. Indeed, this study has found very limited value for rules-of-thumb. Park and Dudycha (1974) provided one of the two best methods for sample size selection, but has been all but ignored in both textbooks and research methodology. Indeed, Stevens (1986, 1992a) is the only multivariate text or regression text author even to cite the work in his bibliography! Perhaps the results of this research and the new method will bring more attention to the issue in general and Park and Dudycha's work in particular. (Park and Dudycha's method still has the advantage of mathematical theory although it proved no better empirically.) Additionally, it is hoped that researchers will recognize that cross-validity shrinkage is more important than the adjusted R^2 printed out by the major statistical packages. Even though the average difference between adjusted R^2 and the given population ρ^2 was only .004 (standard deviation of .006) for the 525 cases, adjusted R^2 is not appropriate for prediction studies -- a cross-validity formula must be used.

Further, it is hoped that the evidence presented to recommend the new method developed in this paper, along with its simplicity, will encourage researchers to consider power a priori, as should be the case. Although power in regression may have a slightly different meaning than in other statistical designs, it is no less important. The authors believe that absent any meaningful guess for the true ρ^2, the researcher must consider carefully what will be practically significant. An appropriate sample size must be chosen not only so that the study will have adequate power to find that degree a relationship if it exists, but also that if a significant model is found, its regression coefficients will be meaningful and stable if applied to another sample from the population. The reader should recognize the potential danger in selecting an expected R^2 that is much above the true ρ^2. When in doubt, a conservative approach would suggest using a low-to-moderate value for $E(R^2)$, say .25, and therefore a larger sample. Finally, no statistical analysis can repair damage caused by an inadequate sample. The reader must remember that a sample must not only be large enough, but also must be random and appropriately representative of the population to which the research will generalize (Cooley and Lohnes, 1971; Miller & Kunce, 1973).

References

Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park, CA: Sage.

Barcikowski, R. S. (1980). Regression analysis. Unpublished manuscript.

Box, G. E. P., & Muller, M. E. (1958). A note on generation of normal deviates, AMS, 28, 610-611.

Brogden, H. E. (1972). Some observations on two methods in psychology. Psychological Bulletin, 77, 431-437.

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

Cascio, W. F., Valenzi, E. R., & Silbey, V. (1978). Validation and statistical power: Implications for applied research. Journal of Applied Psychology, 63, 589-595.

Cattin, P. (1980a). Estimation of the predictive power of a regression model. Journal of Applied Psychology, 65, 407-414.

Cattin, P. (1980b). Note on the estimation of the squared cross-validated multiple correlation of a regression model. Psychological Bulletin, 87, 63-65.

Chambers, J. M. (1977). Computational methods for data analysis. New York: John Wiley & Sons.

Claudy, J. G. (1972). A comparison of five variable weighting procedures. Educational and Psychological Measurement, 32, 311-322.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

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

Collier, R. O., Baker, F. B., Mandeville, G. K., & Hayes, T. F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. Psychometrika, 32, 339-353.

Cooley, W. W., & Lohnes, P. R. (1971). Multivariate Data Analysis. New York: John Wiley & Sons.

Cummings, C. C. (1982). Estimates of multiple correlation coefficient shrinkage. Paper presented at the meeting of the American Educational Research Association, New York, NY. (ERIC Document Reproduction Service No. ED 220 517)

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

Dixon, W. J. (1990). BMDP statistical software manual to accompany the 1990 software release (Vol. 1). Berkeley, CA: University of California.

Draper, N. R., & Smith, H. (1966). Applied regression analysis. New York: John Wiley & Sons.

Drasgow, F., Dorans, N. J., & Tucker, L. R. (1979). Estimators of the squared cross-validity coefficient: A Monte Carlo investigation. Applied Psychological Measurement, 3, 387-399.

Dunn, O. J., & Clark, V. A. (1974). Applied statistics: Analysis of variance and regression. New York: John Wiley & Sons.

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

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. Psychological Bulletin, 106, 516-524.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? Multivariate Behavioral Research, 26, 499-510.

Halinski, R. S., & Feldt, L. S. (1970). The selection of variables in multiple regression analysis. Journal of Educational Measurement, 7, 151-157.

Harris, R. J. (1985). A primer of multivariate statistics (2nd ed.). Orlando, FL: Academic Press.

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

Hinkle, D. E., & Oliver, J. D. (1983). How large should a sample be? A question with no simple answer? Or.... Educational and Psychological Measurement, 43, 1051-1060.

Howell, D. C. (1987). Statistical methods for psychology (2nd ed.). Boston: PWS-Kent.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. Educational and Psychological Measurement, 40, 101-112.

International Mathematical and Statistical Library. (1985). Stat/PC Library. Houston, TX: Author.

Johnson, N. L., & Leone, F. C. (1977). Statistics and experimental design in engineering and the physical sciences. New York: John Wiley & Sons.

Karian, Z. A., & Dudewicz, E. J. (1991). Modern statistical systems, and GPSS simulation: The first course. New York: Computer Science Press.

Kennedy, E. (1988). Estimation of the squared cross-validity coefficient in the context of best subset regression. Applied Psychological Measurement, 12, 231-237.

Kennedy, W. J., Jr., & Gentle, J. E. (1980). Statistical computing. New York: Marcel Dekker.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart, and Winston.

Keselman, H. J., Keselman, J. C., & Shaffer, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violations of multisample sphericity. Psychological Bulletin, 110, 162-170.

Kleijnen, J. P. C. (1974). Statistical techniques in simulation: Part I. New York: Marcel Dekker.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1987). Applied regression analysis and other multivariate methods (2nd ed.). Boston: PWS-Kent.

Knuth, D. E. (1981). The art of computer programming: Vol. 2. Seminumerical algorithms (2nd ed.). Reading, MA: Addison-Wesley.

Kraemer, H. C. (1985). A strategy to teach the concept and application of power of statistical tests. Journal of Educational Statistics, 10, 173-195.

Kraemer, H. C., & Thiemann, S. (1987). How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage.

Lerner, J. V., & Games, P. A. (1981). Maximum R^2 improvement and stepwise multiple regression as related to over-fitting. Psychological Reports, 48, 979-983.

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

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

McNemar, Q. (1962). Psychological statistics (3rd ed.). New York: John Wiley & Sons.

Miller, D. E., & Kunce, J. T. (1973). Prediction and statistical overkill revisited. Measurement and evaluation in guidance, 6, 157-163.

Milton, S. (1986). A sample size formula for multiple regression studies. Public Opinion Quarterly, 50, 112-118.

Montgomery, D. C., & Peck, E. A. (1992). Introduction to linear regression analysis (2nd ed.). New York: John Wiley & Sons.

Morgan, B. J. T. (1984). Elements of simulation. New York: Chapman and Hall.

Morrison, D. F. (1976). Multivariate statistical methods. New York: McGraw-Hill.

Mosier, C. I. (1951). Problems and designs of cross-validation. Educational and Psychological Measurement, 11, 5-11.

Murphy, K. R. (1982, August). Cost-benefit considerations in choosing among cross-validation methods. Paper presented at the meeting of the American Psychological Association, Washington, D.C. (ERIC Document Reproduction Service No. ED 223 701)

Nash, J. C. (1990). Compact numerical methods for computers: Linear algebra and function minimisation (2nd ed.). New York: Adam Hilger.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). Applied linear statistical models: Regression, analysis of variance, and experimental designs (3rd ed.). Homewood, IL: Irwin.

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

Norusis, M. J. (1988). SPSS-X advanced statistics guide (2nd ed.). Chicago: SPSS.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Olejnik, S. F. (1984). Planning educational research: Determining the necessary sample size. Journal of Experimental Education, 53, 40-48.

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.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum Associates.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. Journal of the American Statistical Association, 79, 575-583.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). Numerical recipes in Pascal: The art of scientific computing. New York: Cambridge University.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). Numerical recipes in FORTRAN: The art of scientific computing (2nd ed.). New York: Cambridge University.

Ray, A. A. (1982). SAS user's guide: Statistics, 1982 edition. Cary, NC: SAS Institute.

Ripley, B. D. (1987). Stochastic simulation. New York: John Wiley & Sons.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlations: A clarification. Psychological Bulletin, 85, 1348-1351.

Rozeboom, W. W. (1981). The cross-validational accuracy of sample regressions. Journal of Educational Statistics, 6, 179-198.

Rubinstein, R. Y. (1981). Simulation and the Monte Carlo method. New York: John Wiley & Sons.

Sampson, A. R. (1974). A tale of two regressions. Journal of the American Statistical Association, 69, 682-689.

Sawyer, R. (1982). Sample size and the accuracy of predictions made from multiple regression equations. Journal of Educational Statistics, 7, 91-104.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. (1977). A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. Psychological Bulletin, 84, 751-758.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

Stein, C. (1960). Multiple regression. In I. Olkin et al. (Eds.), Contributions to probability and statistics. Palo Alto, CA: Stanford University.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevens, J. (1992a). Applied multivariate statistics for the social sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevens, J. (1992b, October). What I have learned (up to this point) or ruminations on twenty years in the field. Paper presented at the meeting of the Midwestern Educational Research Association, Chicago, IL.

Subkoviak, M. J., & Levin, J. R. (1977). Fallibility of measurement and the power of a statistical test. Journal of Educational Measurement, 14, 47-52.

Tabachnick, B. G., & Fidell, L. S. (1989). Using multivariate statistics (2nd ed.). New York: HarperCollins.

Thorndike, R. M. (1978). Correlational procedures for research. New York: Gardner.

Uhl, N., & Eisenberg, T. (1970). Predicting shrinkage in the multiple correlation coefficient. Educational and Psychological Measurement, 30, 487-489.

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

West, L. J. (1990). Distinguishing between statistical and practical significance. Delta Pi Epsilon Journal, 32(1), 1-4.

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

Wherry, R. J. (1975). Underprediction from overfitting: 45 years of shrinkage. Personnel Psychology, 28, 1-18.

Footnotes

[1] It should be noted that for lower expected R^2 (especially $E(R^2)=.10$), the results were not as consistent across predictors as they were for higher expected R^2 values. That is, for all predictor values at $E(R^2)=.50$, the relative rankings for both statistical and predictive power were BB highest, PD, GS, and COHEN lowest. However, for $E(R^2)=.25$ and two predictors, GS very slightly better than PD, with BB as the most powerful and COHEN the least; but with $E(R^2)=.25$ and more than two predictors the rankings were the same as mentioned above for $E(R^2)=.50$. Finally, for $E(R^2)=.10$, the GS method showed the highest relative power for two predictors, followed by BB, COHEN, and then PD. For three predictors, the rankings were BB highest, then GS, PD, and COHEN. For four, eight, and fifteen predictors, BB was most powerful, followed by PD, GS, and COHEN. Indeed, a review of Table 3 shows that the GS and COHEN methods require larger sample sizes for both a small number of predictors and small values of expected R^2.

Table 1

Examples of Cross-Validation and Shrinkage Formulas

Formula	Attributed To:
$R_a^2 = 1 - \frac{(N-1)(1-R^2)}{(N-p)}$	Wherry (1931)

$R_a^2 = 1 - \frac{(N-1)(1-R^2)}{(N-p-1)}$	Wherry (1931); Ezekiel (1930);
	McNemar (1962);
	Lord & Novick (1968);
	Ray (1982, p. 69) [SAS]

$R_a^2 = R^2 - \frac{p(1-R^2)}{(N-p-1)}$	Norusis (1988, p. 18) [SPSS]
$R_a^2 = R^2 - \frac{p(1-R^2)}{(N-p')}$	Dixon (1990, p. 365) [BMDP][1]
$R_a^2 = R^2 - \frac{(p-2)(1-R^2)}{(N-p-1)} - \frac{2(N-3)(1-R^2)}{(N-p-1)(N-p+1)}$	Olkin & Pratt (1958)
$R_c^2 = 1 - \frac{(N-1)(N+p+1)(1-R^2)}{(N-p-1)N}$	Nicholson (1960) Lord (1950)
$R_c^2 = 1 - \frac{(N-1)(N-2)(N+1)(1-R^2)}{(N-p-1)(N-p-2)N}$	Stein (1960) Darlington (1968)
$R_c^2 = 1 - \frac{(N+p)(1-R^2)}{(N-p)}$	Rozeboom (1978)
$R_c^2 = 1 - \frac{(N+p+1)(1-R^2)}{(N-p-1)}$	Uhl & Eisenberg (1970) Lord (1950)

Note: R_a^2 represents an estimate of ρ^2; R_c^2 is an estimate of ρ_c^2.

[1] p'=p+1 with an intercept, p'=p if the intercept=0.

Table 2

Rules-of-Thumb for Sample Size Selection

Rule	Author(s)
$N \geq 10p$	Miller & Kunce, 1973, p. 162 Halinski & Feldt, 1970, p. 157 (for prediction if $R \geq .50$) Neter, Wasserman, & Kutner, 1990, p. 467
$N \geq 15p$	Stevens, 1992, p. 125
$N \geq 20p$	Tabachnick & Fidell, 1989, p. 128 ($N \geq 100$ preferred) Halinski & Feldt, 1970, p. 157 (for identifying predictors)
$N \geq 30p$	Pedhazur & Schmelkin, 1990, p. 447
$N \geq 40p$	Nunnally, 1978 (inferred from text examples) Tabachnick & Fidell, 1989, p. 129 (for stepwise regression)
$N \geq 50 + p$	Harris, 1985, p. 64
$N \geq 10p + 50$	Thorndike, 1978, p. 184
$N > 100$	Kerlinger & Pedhazur, 1973, p. 442 (preferably $N>200$)
$N \geq \frac{(2K^2-1) + K^2p}{(K^2-1)}$	Sawyer, 1982, p. 95 (K is an inflation factor due to estimating regression coefficients)

Note: In the formulas for sample size above, N represents the suggested sample size and p represents the number of predictors (independent variables) used in the regression analysis.

Table 3

Sample Sizes Suggested by Each Method for Each Level of Expected R^2

Number of			Sample Size for	
Predictors	Method	$E(R^2)=.50$	$E(R^2)=.25$	$E(R^2)=.10$
2	New Method (BB)	42	62	122
	Park & Dudycha (PD)	31	45	85
	Cohen (COHEN)	13	38	115
	Gatsonis & Sampson (GS)	20	45	135
	30:1 (NP30)	60	60	60
	50 + 8p (COMBO)	66	66	66
	15:1 (NP15)	30	30	30
3	New Method (BB)	63	93	183
	Park & Dudycha (PD)	50	71	133
	Cohen (COHEN)	14	44	130
	Gatsonis & Sampson (GS)	23	51	151
	30:1 (NP30)	90	90	90
	50 + 8p (COMBO)	74	74	74
	15:1 (NP15)	45	45	45
4	New Method (BB)	84	124	244
	Park & Dudycha (PD)	66	93	173
	Cohen (COHEN)	16	48	144
	Gatsonis & Sampson (GS)	25	55	165
	30:1 (NP30)	120	120	120
	50 + 8p (COMBO)	82	82	82
	15:1 (NP15)	60	60	60
8	New Method (BB)	168	248	488
	Park & Dudycha (PD)	124	171	311
	Cohen (COHEN)	20	61	183
	Gatsonis & Sampson (GS)	32	69	205
	30:1 (NP30)	240	240	240
	50 + 8p (COMBO)	114	114	114
	15:1 (NP15)	120	120	120
15	New Method (BB)	315	465	915
	Park & Dudycha (PD)	214	292	524
	Cohen (COHEN)	26	78	235
	Gatsonis & Sampson (GS)	42	88	256
	30:1 (NP30)	450	450	450
	50 + 8p (COMBO)	170	170	170
	15:1 (NP15)	225	225	225

Table 4

Empirical Predictive Power Rates (averaged across number of predictors)

$E(R^2)$	Method	$E(R^2) \approx \rho^2$	$E(R^2) > \rho^2$	$E(R^2) < \rho^2$
.50	New Method (BB)	.9137 (.03)	.3496 (.32)	.9652 (.03)
	Park-Dudycha (PD)	.8835 (.03)	.2807 (.28)	.9529 (.04)
	Cohen (COHEN)	.3650 (.27)	.0000 (.00)	.7724 (.11)
	Gatsonis-Sampson (GS)	.6206 (.19)	.0427 (.10)	.8814 (.06)
	30:1 N/p ratio (NP30)	.9402 (.02)	.4258 (.36)	.9761 (.02)
	$N \geq 50 + 8p$ (COMBO)	.9020 (.04)	.3238 (.30)	.9700 (.02)
	15:1 N/p ratio (NP15)	.8780 (.04)	.2663 (.28)	.9498 (.04)
.25	New Method (BB)	.8028 (.05)	.2477 (.29)	.9596 (.03)
	Park-Dudycha (PD)	.7274 (.07)	.1754 (.23)	.9432 (.04)
	Cohen (COHEN)	.4594 (.19)	.0122 (.03)	.8782 (.10)
	Gatsonis-Sampson (GS)	.5193 (.18)	.0321 (.06)	.8944 (.09)
	30:1 N/p ratio (NP30)	.7966 (.05)	.2385 (.29)	.9582 (.03)
	$N \geq 50 + 8p$ (COMBO)	.6916 (.09)	.1387 (.16)	.9362 (.05)
	15:1 N/p ratio (NP15)	.6082 (.10)	.0964 (.14)	.9140 (.05)
.10	New Method (BB)	.7052 (.14)	.2229 (.17)	.9524 (.04)
	Park-Dudycha (PD)	.5863 (.17)	.0682 (.07)	.9295 (.06)
	Cohen (COHEN)	.5097 (.10)	.0107 (.02)	.9076 (.09)
	Gatsonis-Sampson (GS)	.5574 (.11)	.0265 (.06)	.9179 (.08)
	30:1 N/p ratio (NP30)	.4665 (.23)	.0113 (.02)	.9041 (.09)
	$N \geq 50 + 8p$ (COMBO)	.2789 (.10)	.0000 (.00)	.8543 (.13)
	15:1 N/p ratio (NP15)	.1926 (.15)	.0000 (.00)	.8120 (.16)

Note: Standard deviations are in parentheses after the means.

Table 5

Empirical Statistical Power Rates (averaged across number of predictors)

$E(R^2)$	Method	$E(R^2)\approx\rho^2$	$E(R^2)>\rho^2$	$E(R^2)<\rho^2$
.50	New Method (BB)	1.0000 (.00)	.6483 (.35)	1.0000 (.00)
	Park-Dudycha (PD)	.9999 (.00)	.5824 (.36)	.9999 (.00)
	Cohen (COHEN)	.7294 (.15)	.1483 (.10)	.9591 (.05)
	Gatsonis-Sampson (GS)	.9550 (.05)	.2625 (.20)	.9971 (.00)
	30:1 N/p ratio (NP30)	1.0000 (.00)	.7257 (.33)	1.0000 (.00)
	$N \geq 50 + 8p$ (COMBO)	1.0000 (.00)	.6333 (.35)	1.0000 (.00)
	15:1 N/p ratio (NP15)	.9997 (.00)	.5678 (.36)	.9999 (.00)
.25	New Method (BB)	.9899 (.02)	.6019 (.33)	1.0000 (.00)
	Park-Dudycha (PD)	.9671 (.07)	.5054 (.33)	1.0000 (.00)
	Cohen (COHEN)	.8701 (.06)	.2651 (.17)	.9997 (.00)
	Gatsonis-Sampson (GS)	.9180 (.05)	.3030 (.20)	.9999 (.00)
	30:1 N/p ratio (NP30)	.9876 (.03)	.5951 (.33)	1.0000 (.00)
	$N \geq 50 + 8p$ (COMBO)	.9878 (.02)	.4571 (.30)	1.0000 (.00)
	15:1 N/p ratio (NP15)	.9046 (.15)	.3998 (.31)	.9999 (.00)
.10	New Method (BB)	.9294 (.15)	.6108 (.33)	.9999 (.00)
	Park-Dudycha (PD)	.8810 (.21)	.4783 (.28)	.9993 (.00)
	Cohen (COHEN)	.8923 (.14)	.3284 (.11)	.9999 (.00)
	Gatsonis-Sampson (GS)	.9248 (.12)	.3731 (.12)	1.0000 (.00)
	30:1 N/p ratio (NP30)	.8032 (.26)	.3816 (.26)	.9957 (.02)
	$N \geq 50 + 8p$ (COMBO)	.7065 (.20)	.2091 (.07)	.9959 (.01)
	15:1 N/p ratio (NP15)	.6049 (.32)	.1986 (.12)	.9677 (.09)

Note: Standard deviations are in parentheses after the means.

Table 6

Predictive Power for the Case $E(R^2)>\rho^2$ (averaged across number of predictors)

$E(R^2)$	Method	$.16<\rho^2<.25$	$.04<\rho^2<.16$	$.001<\rho^2<.04$
.50	New Method (BB)	.7153 (.07)	.3334 (.20)	.0000 (.00)
	Park-Dudycha (PD)	.6242 (.10)	.2178 (.14)	.0000 (.00)
	Cohen (COHEN)	.0000 (.00)	.0000 (.00)	.0000 (.00)
	Gatsonis-Sampson (GS)	.1280 (.16)	.0000 (.00)	.0000 (.00)
	30:1 N/p ratio (NP30)	.7969 (.05)	.4679 (.22)	.0126 (.03)
	$N \geq 50 + 8p$ (COMBO)	.6914 (.09)	.2800 (.11)	.0000 (.00)
	15:1 N/p ratio (NP15)	.6087 (.10)	.1901 (.16)	.0000 (.00)
.25	New Method (BB)		.4803 (.22)	.0150 (.03)
	Park-Dudycha (PD)		.3508 (.20)	.0000 (.00)
	Cohen (COHEN)		.0244 (.05)	.0000 (.00)
	Gatsonis-Sampson (GS)		.0642 (.08)	.0000 (.00)
	30:1 N/p ratio (NP30)		.4658 (.23)	.0112 (.03)
	$N \geq 50 + 8p$ (COMBO)		.2775 (.11)	.0000 (.00)
	15:1 N/p ratio (NP15)		.1927 (.15)	.0000 (.00)
.10	New Method (BB)			.2229 (.17)
	Park-Dudycha (PD)			.0682 (.06)
	Cohen (COHEN)			.0107 (.02)
	Gatsonis-Sampson (GS)			.0265 (.06)
	30:1 N/p ratio (NP30)			.0113 (.03)
	$N \geq 50 + 8p$ (COMBO)			.0000 (.00)
	15:1 N/p ratio (NP15)			.0000 (.00)

Note: Standard deviations are in parentheses after the means.

Table 7

Statistical Power for the Case $E(R^2)>\rho^2$ (averaged across number of predictors)

$E(R^2)$	Method	$.16<\rho^2<.25$	$.04<\rho^2<.16$	$.001<\rho^2<.04$
.50	New Method (BB)	.9605 (.08)	.7087 (.30)	.2758 (.19)
	Park-Dudycha (PD)	.9175 (.14)	.6259 (.31)	.2037 (.12)
	Cohen (COHEN)	.2684 (.07)	.1146 (.02)	.0619 (.01)
	Gatsonis-Sampson (GS)	.5126 (.09)	.1963 (.04)	.0787 (.01)
	30:1 N/p ratio (NP30)	.9876 (.03)	.8069 (.26)	.3827 (.27)
	$N \geq 50 + 8p$ (COMBO)	.9869 (.02)	.7069 (.20)	.2061 (.07)
	15:1 N/p ratio (NP15)	.9045 (.15)	.5976 (.33)	.2014 (.13)
.25	New Method (BB)		.8114 (.26)	.3924 (.27)
	Park-Dudycha (PD)		.7327 (.29)	.2780 (.17)
	Cohen (COHEN)		.4122 (.10)	.1180 (.02)
	Gatsonis-Sampson (GS)		.4722 (.11)	.1338 (.03)
	30:1 N/p ratio (NP30)		.8056 (.26)	.3846 (.27)
	$N \geq 50 + 8p$ (COMBO)		.7065 (.20)	.2078 (.07)
	15:1 N/p ratio (NP15)		.5998 (.32)	.1998 (.13)
.10	New Method (BB)			.6108 (.33)
	Park-Dudycha (PD)			.4783 (.28)
	Cohen (COHEN)			.3284 (.11)
	Gatsonis-Sampson (GS)			.3731 (.12)
	30:1 N/p ratio (NP30)			.3816 (.26)
	$N \geq 50 + 8p$ (COMBO)			.2091 (.07)
	15:1 N/p ratio (NP15)			.1986 (.12)

Note: Standard deviations are in parentheses after the means.

Table 8

Mean Power Ranks for Methods

Method	$E(R^2)=.50$	$E(R^2)=.25$	$E(R^2)=.10$	Overall
Predictive Power				
New Method (BB)	2.82 (0.95)	1.64 (1.11)	1.28 (0.68)	1.91 (1.13)
Park-Dudycha (PD)	3.90 (0.46)	3.70 (0.79)	2.70 (0.89)	3.43 (0.90)
Cohen (COHEN)	6.22 (1.16)	6.24 (1.14)	4.28 (0.75)	5.58 (1.38)
Gatsonis-Sampson (GS)	5.70 (0.88)	5.28 (0.85)	3.08 (1.33)	4.69 (1.55)
30:1 N/p ratio (NP30)	1.64 (1.11)	2.48 (0.77)	4.42 (1.19)	2.85 (1.56)
$N \geq 50 + 8p$ (COMBO)	3.38 (1.51)	3.72 (1.40)	5.96 (0.90)	4.35 (1.72)
15:1 N/p ratio (NP15)	4.34 (0.75)	4.94 (1.08)	6.28 (0.85)	5.19 (1.21)
Statistical Power				
New Method (BB)	2.60 (0.63)	2.36 (1.23)	2.58 (1.24)	2.51 (1.06)
Park-Dudycha (PD)	3.62 (0.63)	3.54 (0.79)	3.18 (0.85)	3.45 (0.78)
Cohen (COHEN)	6.88 (0.60)	5.84 (1.43)	3.86 (0.67)	5.53 (1.59)
Gatsonis-Sampson (GS)	5.82 (0.63)	4.86 (1.03)	3.14 (1.06)	4.61 (1.44)
30:1 N/p ratio (NP30)	1.92 (0.94)	2.76 (0.94)	4.16 (1.05)	2.95 (1.34)
$N \geq 50 + 8p$ (COMBO)	3.02 (1.36)	3.68 (1.22)	5.26 (1.32)	3.99 (1.59)
15:1 N/p ratio (NP15)	4.14 (0.86)	4.96 (1.38)	5.82 (1.38)	4.97 (1.39)

Note: Standard deviations are in parentheses after the means. Lower rank means better power.

Table 9

Frequency of Minimum Power Values for Each Method

	$E(R^2)$=.50		$E(R^2)$=.25		$E(R^2)$=.10	
Power Method	.7	.8	.7	.8	.7	.8
Predictive Power						
New Method (BB)	5	5	5	3	4	2
Park-Dudycha (PD)	5	5	4	0	1	0
Cohen (COHEN)	0	0	0	0	0	0
Gatsonis-Sampson (GS)	2	1	1	0	0	0
30:1 N/p ratio (NP30)	5	5	5	3	0	0
N ≥ 50 + 8p (COMBO)	5	5	2	1	0	0
15:1 N/p ratio (NP15)	5	5	0	0	0	0
Statistical Power						
New Method (BB)	5	5	5	5	4	4
Park-Dudycha (PD)	5	5	5	5	4	4
Cohen (COHEN)	3	3	5	4	4	4
Gatsonis-Sampson (GS)	5	5	5	5	5	4
30:1 N/p ratio (NP30)	5	5	5	5	4	3
N ≥ 50 + 8p (COMBO)	5	5	5	5	2	2
15:1 N/p ratio (NP15)	5	5	4	4	2	2

merican Educational Research Association
pril 4-8, 1994

TM027498

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

ERIC®

I. DOCUMENT IDENTIFICATION:

Title: A New Sample Size Formula for Regression	
Author(s): Gordon P. Brooks, Robert S. Barcikowski	
Corporate Source: Ohio University	Publication Date: April 1994

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ ◀ **Sample sticker to be affixed to document**

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document ▶ □

or here
Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Gordon P Brooks	Position: Ph.D. Student
Printed Name: Gordon P. Brooks	Organization: Ohio University
Address: 601 Courtland Lane Pickerington, OH 43147	Telephone Number: (614) 833-3791
	Date: 6/17/97

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

March 1994

Dear AERA Presenter,

Congratualations on being a presenter at AERA. The ERIC Clearinghouse on Assessment and Evaluation would like you to contribute to ERIC by providing us with a written copy of your presentation. Submitting your paper to ERIC ensures a wider audience by making it available to members of the education community who could not attend the session or this year's conference.

Abstracts of papers that are accepted by ERIC appear in RIE and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of RIE. Your contribution will be accessible through the printed and electronic versions of RIE, through the microfiche collections that are housed at libraries around the country and the world, and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse and you will be notified if your paper meets ERIC's criteria. Documents are reviewed for contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

To disseminate your work through ERIC, you need to sign the reproduction release form on the back of this letter and include it with **two** copies of your paper. You can drop of the copies of your paper and reproduction release form at the ERIC booth (#227) or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1994/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE