Mid-Western Educational Researcher

Volume 9
Issue 4 *Regression Analyses*

Article 5

1996

Precision Power and Its Application to the Selection of Regression Sample Sizes

Gordon P. Brooks
Ohio University

Robert S. Barcikowski
Ohio University

Precision Power and Its Application to the Selection of Regression Sample Sizes

Gordon P. Brooks and Robert S. Barcikowski, Ohio University

Abstract

Because of contradictions among the various methods, sample size selection in multiple regression has been problematic. For example, how does one reconcile the difference between a 15:1 subject-to-variable rule and a 30:1 rule? The purpose of this paper is to analyze the advantages and disadvantages of the various methods of selecting sample sizes in regression. A discussion of the importance of cross-validity to prediction studies will be followed by descriptions of the three categories of sample size methods: cross-validation approaches, rules-of-thumb, and statistical power methods. A rationale will then be developed for the application of precision power to multiple regression, leading to the presentation, through multiple examples, of the precision power method for sample size selection in prediction studies.

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 for prediction. Because of this possibility, researchers must ensure that their studies have adequate power so that results will generalize; the best way to ensure this power, and therefore stable regression weights, is to use a sufficiently large sample.

Despite encouragement from scholars, many researchers continue to ignore power in their studies (Cohen, 1992; Sedlmeier & Gigerenzer, 1989; Stevens, 1992b). This situation is compounded for multiple regression research even though several methods exist for choosing sample sizes for power. These methods can be grouped loosely into three categories: rules-of-thumb, statistical power methods, and cross-validation methods. Unfortunately, as Olejnik noted in 1984 and was confirmed recently (Brooks & Barcikowski, 1994), many regression textbooks do not discuss the issue of sample size selection (e.g., Dunn & Clark, 1974; Kleinbaum, Kupper, & Muller, 1987; Montgomery & Peck, 1992; Weisberg, 1985) or simply provide a rule-of-thumb (e.g., Cooley & Lohnes, 1971; Harris, 1985; Kerlinger & Pedhazur, 1973; Tabachnick & Fidell, 1989), possibly because there are problems and contradictions among the various methods.

For example, how does one reconcile differences between a statistical power method that suggests 16 subjects and a 15:1 subject-to-variable ratio rule that recommends 60? Furthermore, the many rules-of-thumb lack any measure of effect size, which is generally recognized as a critical element in the determination of sample sizes. Cohen's (1988) methods are derived from a fixed model and statistical power approach to regression; however, a random model and cross-validation approach, like Park and Dudycha's (1974), may be more appropriate in the social sciences, where a prediction function is often desired. This is because generalizability is the primary consideration for the development of a prediction model, whereas statistical power is the main concern when regression is used to test hypotheses about relationships between variables.

Therefore, the purpose of this paper is to analyze the advantages and disadvantages of the various methods of selecting sample sizes in regression. A discussion of the importance of cross-validity to prediction studies will be followed by descriptions of the three categories of sample size methods: cross-validation approaches, rules-of-thumb, and statistical power methods. A rationale will then be developed for the application of precision power to multiple regression, leading to the presentation, through multiple examples, of the precision power method for sample size selection in regression studies designed to develop prediction models.

Cross-Validation and Shrinkage

Because the expected value of the sample multiple correlation (i.e., an average correlation over many samples) is an overestimate of the population multiple correlation, researchers have employed a number of methods to "shrink" R^2 and thereby provide better estimates of true population multiple correlations. Formula methods of shrinkage are typically preferred to empirical cross-validation (data-splitting) so that the entire sample may be used for model-build-

ing. Indeed, several common formula estimates have been shown superior to empirical cross-validation techniques (Cattin, 1980a; 1980b; Kennedy, 1988; Murphy, 1982; Schmitt, Coyle, & Rauschenberger, 1977).

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. The multiple correlation, ρ, is the correlation between the criterion and the regression function if both are measured in the population (Herzberg, 1969; Stevens, 1992a). For example, a researcher who calculates a sample $R^2 = .3322$ with 121 subjects and 3 predictors might use an adjusted R^2 formula to conclude that, in the population, the multiple correlation between the criterion and the predictors is approximately $\rho = .5613$, since $R_a^2 = .3151$.

Cross-validity formulas, which are based on estimates of the mean squared error of prediction, 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 (that is, the average over many samples) approximates ρ_c^2. This cross-validity coefficient can be thought of as the squared correlation between the actual population criterion values and the scores predicted by the sample regression equation when applied to the population or to another sample (Kennedy, 1988; Schmitt et al., 1977). For example, a researcher who calculates a sample $R^2 = .3322$ with 121 subjects and 3 predictors might use a cross-validity formula to calculate the sample cross-validity coefficient as $R_c^2 = .2916$. This cross-validity coefficient implies that the researcher would explain 29%, not 33%, of the variance of the criterion when applying the sample regression function to future samples.

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
$R_c^2 = 1 - \frac{(N-1)(N+p+1)(1-R^2)}{(N-p-1)N}$	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) (who cite Lord, 1950)

Note: R_a^2 represents an estimate of r^2; R_c^2 is an estimate of r_c^2. $p' = p + 1$ with an intercept, $p' = p$ if the intercept = 0.

The most common estimate of shrinkage reported in the literature (and in statistical packages) is an adjusted R^2 that is attributed most frequently to Wherry (1931). However, when researchers are interested in developing a regression model to predict for future subjects, they should report both R_a^2 (for descriptive purposes) and R_c^2, which indicates how well their sample equation may predict in subsequent samples (Cattin, 1980b; Huberty & Mourad, 1980). 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 in this regard. Some of the more familiar cross-validity formulas are those by Stein (1960), Darlington (1968), Lord (1950), Nicholson (1960), and Browne (1975).

Multiple Regression Sample Size Methods

There are three primary types of sample size methods available for multiple linear regression: cross-validation approaches, rules-of-thumb, and statistical power approaches. The following sections describe each briefly, with emphasis on the aspects of each that pertain to the precision power method described later.

Cross-Validation Approach to Sample Sizes

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. In the random model, both the predictors and the criterion are sampled together from a joint multivariate distribution. The fixed model, on the other hand,

assumes that the researcher is able to select or control the values of the independent variables before measuring subjects on the random dependent variable. The random model is usually more appropriate to social scientists, because they typically measure subjects on predictors and the criterion simultaneously and therefore 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). It is important to recognize that the misapplication of fixed model data to the random model may cause biased estimates of the population parameters (Claudy, 1972). For a more complete discussion of the random and fixed models, the reader is referred to Afifi and Clark (1990), Brogden (1972), Dunn and Clark (1974), Johnson and Leone (1977), and Sampson (1974).

Park and Dudycha (1974) derived the following sample size formula:

$$N \geq \frac{(1 - \rho^2)\,\delta_1^{\,2}}{\rho^2} + p + 2\,,$$

where ρ is the anticipated population correlation, and $\delta_1^{\,2}$ is the noncentrality parameter for the t-distribution. Researchers determine the probability with which they want to approximate ρ within some chosen error tolerance. The formula for this probability is:

$$P(\rho - \rho_c \leq \varepsilon) = \gamma$$

The researcher chooses (a) an assumed ρ^2 as the effect size, (b) the absolute 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) then can be consulted with these values. Unfortunately, their tables are limited to only a few possible combinations of sample size, squared correlation, and epsilon. Also 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 ε or the probability to use when consulting the tables (although Stevens, 1992a, implied through examples that .05 and .90, respectively, are acceptable values).

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 Neter, Wasserman, & Kutner, 1990, p. 467
$N \geq 15p$	Stevens, 1992, p. 125
$N \geq 20p$	Tabachnick & Fidell, 1989, p. 128 Halinski & Feldt, 1970, p. 157 (for identifying predictors)
$N \geq 30p$	Pedhazur & Schmelkin, 1990, p. 447
$N \geq 40p$	Nunnally, 1978 (inferred from 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
$N > 100$ (or 200)	Kerlinger & Pedhazur, 1973, p. 442
$N \geq \frac{(2K^2-1) + K^2p}{(K^2-1)}$	Sawyer, 1982, p. 95 (K is an inflation factor due to estimating 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.

Rules-of-Thumb for Selecting Sample Sizes

The most extensive literature regarding sample sizes in regression analysis is in the area of experiential rules. Many scholars have suggested rules-of-thumb for choosing sample sizes that they claim will provide reliable estimates of the population regression coefficients. 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). This is true because as the number of subjects increases relative to the number of predictors, both R^2 and $\rho_c^{\,2}$ converge toward ρ^2, and therefore the amount of shrinkage decreases (Cattin, 1980a).

Rules-of-thumb typically take the form of a subject-to-predictor (N/p) ratio. 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. For example, Stevens (1986) recommended a 15:1 subject-to-variable ratio, which he based primarily on an analysis of Park and Dudycha's (1974) tables. Harris (1985) noted, however, that ratio rules-of-thumb clearly break down for small numbers of predictors. 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). Indeed, Green (1991) found that a combination formula such as $N > 50 + 8p$ was much better than subject-to-variable ratios alone. Additionally, Sawyer (1982) developed a formula based on limiting the inflation of mean squared error. Sawyer's formula, however, easily simplifies into a combination rule once the inflation factor, k, is chosen. 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).

The most profound problem with many rules-of-thumb advanced by regression scholars is that they lack any mea-

sure of effect size. Indeed, even Sawyer's inflation factor is not an 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 next section includes a more complete discussion of effect size and its importance in power analysis.

Statistical Power Approach to Sample Size

"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. Several scholars have proposed regression sample size methods based on statistical power (e.g., Cohen, 1988; Cohen & Cohen, 1983; Gatsonis & Sampson, 1989; Kraemer & Thiemann, 1987; Milton, 1986; Neter, Wasserman, & Kutner, 1990).

Statistical power analysis requires the consideration of at least 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. This is a fixed model approach to regression, however, and is most useful when researchers use regression as a means to explain the variance of a phenomenon in lieu of analysis of variance or to determine the importance of individual predictors. It is useful, though, to discuss effect size regardless of the approach to regression that is taken.

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

$$f^2 = \frac{R^2}{1 - R^2}.$$

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 $R^2 = .26$. Cohen's (1988) sample size is calculated as

$$N = \frac{\lambda (1 - R^2)}{R^2},$$

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.

For prediction studies, the fundamental 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 may 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). Therefore, selecting a sample size based on statistical power tests may be useful in selecting predictors to include in a final model, but it will not ensure adequate sample size to allow a regression equation to generalize to other samples from the given population.

Precision Power

While several scholars have used the term *predictive power* (e.g., Cascio et al., 1978; Kennedy, 1988; Nunnally, 1978; Stevens, 1986, 1992a), 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. Stevens (1992a), who discussed predictive power as an aspect of model validation, used the term to mean how well a derived regression equation will predict in other samples from the same population. Therefore, a "loss in predictive power" to Stevens is simply the size of the decrease in the sample R^2 when an appropriate shrinkage or cross-validity formula is applied.

Although both Cattin's and Steven's definitions of predictive power could be applied to the problem of sample size in some fashion, neither would provide any sense of the magnitude of error as compared to the original R^2 value. For example, a loss in predictive power (as Stevens defines it) of .20 suggests drastically different results if the sample R^2 is .50 than if the sample R^2 is .25. Because they desire a regression model that predicts well in subsequent samples, researchers hope to limit shrinkage as much as possible relative to the sample R^2 value they attained. Therefore, a concept is required that provides more information about the magnitude of shrinkage relative to sample values.

The term *precision power* is proposed to indicate how well a regression function is expected to perform if applied to future samples. The term is adapted from Darlington (1990), who used the phrase "precision of estimates" to oppose the "power of hypothesis tests" (i.e., statistical power) while introducing a chapter on choosing sample sizes (p. 379). Precision power is defined more precisely as R_c^2/R^2, which can be inferred and adapted from an example used by Stevens (1992a, p. 100). With a larger sample, this fraction

would be larger because less shrinkage occurs with larger samples, all else remaining constant. Using Stevens' example, a 61.8% shrinkage from $R^2 = .50$ to $R_c^2 = .191$ occurs with a sample size of 50; when the sample is increased to 150, there is only a 15.8% shrinkage from $R^2 = .50$ to $R_c^2 = .421$. The precision power in the first case would be .191/.50 = .382, and precision power in the second case is .421/.50 = .842.

The formulaic definition of precision power,

$$PP = \frac{R_c^2}{R^2}, \qquad (1)$$

can be manipulated algebraically into the formula

$$PP = 1 - \frac{(R^2 - R_c^2)}{R^2}. \qquad (2)$$

The fraction, $(R^2 - R_c^2)/R^2$, can be interpreted as the proportional decrease, or proportional shrinkage (PS), in the squared multiple correlation after an appropriate cross-validity estimate is made. Therefore, 1 - PS provides an estimate of the precision power, and therefore generalizability, of the regression equation. For example, if sample $R^2 = .50$ and $R_c^2 = .10$, the precision power for that regression model would be 1 - (.40/.5) = .20; this suggests very little generalizability for the regression model because the R^2 value shrank by 80%. A precision power value of .90, on the other hand, would indicate a highly generalizable model.

Precision power thus describes how well a regression equation will predict in other samples relative to its ability to predict in the derivation sample. Because the term power has special meaning in the research literature, a word of warning may be prudent at this time. Precision power as defined here, 1 - PS, is similar in form to the theoretical definition of statistical power, 1 - ß, where ß is the probability of a Type II error. However, PS is not the probability of error but the tolerance level for error, or more precisely, cross-validity shrinkage. Furthermore, the term statistical power is used in reference to a test of a hypothesis; the term precision power, on the other hand, applies not to a statistical test, but to an evaluation of the generalizability of a regression equation.

Table 3

Sample Sizes Suggested by Several Methods

K	Method	E(R^2) .75	.50	.25	.10
4	Precision Power ($\varepsilon = .2R^2$)	22	55	155	455
	Precision Power ($\varepsilon = .05$)	55	105	155	305
	Park & Dudycha (p = .90)	37	66	93	173
	Sawyer	22	30	55	130
	30:1	120	120	120	120
	50 + 8p	82	82	82	82
	15:1	60	60	60	60
	Cohen	8	16	48	144
	Gatsonis & Sampson	14	25	55	165
8	Precision Power ($\varepsilon = .2R^2$)	39	99	279	819
	Precision Power ($\varepsilon = .05$)	99	189	279	549
	Park & Dudycha (p = .90)	68	124	171	311
	Sawyer	38	53	98	233
	30:1	240	240	240	240
	50 + 8p	114	114	114	114
	15:1	120	120	120	120
	Cohen	12	20	61	183
	Gatsonis & Sampson	19	32	69	205

Note: K represents the number of predictors in the model.

The methods described earlier in the paper (a) provide contradictory sample size recommendations (see Table 3), (b) either oversimplify the issue or are too mathematically complex for many researchers to use, and (c) are not all based on the random model. Indeed, a Monte Carlo study that examined several of the methods from a precision power perspective found that none of the methods provided consistently accurate power rates (Brooks & Barcikowski, 1994). Therefore, the precision power method was developed and verified (Brooks & Barcikowski, 1995). The precision power method was determined to be both consistent and accurate across all levels of expected R^2, numbers of predictors, and actual ρ^2.

Precision Power Method

The theory underlying the precision power sample size method is that the researcher, knowing shrinkage is likely to occur, can set a limit as to the amount of shrinkage that will result. Algebraic manipulation and simplification of a cross-validity formula provides the tool needed to limit this expected shrinkage (Brooks & Barcikowski, 1995). Restructuring the cross-validity formula to solve for sample size yields:

$$N \geq \frac{(p + 1)(2 - 2R^2 + \varepsilon)}{\varepsilon} \qquad (3)$$

where p is the number of predictors, R^2 is the expected sample value (i.e., an effect size), and ε is an acceptable amount of shrinkage, $\varepsilon = R^2 - R_c^2$. This value of ε allows researchers to decide how closely to estimate ρ_c^2 from expected R^2: either as an absolute amount of acceptable shrinkage (e.g., $\varepsilon = .05$) or a proportional decrease (e.g., $\varepsilon = .2R^2$, which represents shrinkage of 20%). This is similar to the method employed by Park and Dudycha (1974).

Through changes in the shrinkage tolerance, ε, the precision power formula has the capacity for simplification. For example, if the researcher does not want the sample R^2 to decrease by more than .05 no matter what the expected value of R^2, formula (3) simplifies to

$$N \geq 20 (p + 1)(2.05 - 2R^2) ;$$

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

$$N \geq 33\ (p + 1)\ (2.03 - 2R^2)\ .$$

For example, if there are four predictors in the model and expected $R^2 = .50$, N should be chosen greater than $33*5*(2.03 - 2*.50) = 170$. If a researcher wants an estimate of ρ_c^2 not less than 80% of the sample R^2 value, formula (3) can be reformulated using $\varepsilon = R^2 - .8R^2 = .2R^2$, such that

$$N \geq \frac{(p + 1)\ (2 - 1.8\ R^2)}{.2\ R^2}\ ;$$

or if the researcher wants a ρ_c^2 estimate not less than 75% of the sample R^2 value, the formula can be reformulated such that $\varepsilon = .25R^2$:

$$N \geq \frac{(p + 1)\ (2 - 1.75\ R^2)}{.25\ R^2}\ .$$

As an example, with five predictors and an anticipated R^2 of .40, at least 78 subjects should be used to attain expected precision power of .75.

Other values for ε can be chosen by substituting ε for the quantity $(R^2 - R_c^2)$ in formula (2). Formula (2) can be rewritten as

$$PP = 1 - \frac{\varepsilon}{R^2} \qquad (4)$$

and therefore

$$\varepsilon = R^2 - (PP * R^2)\ . \qquad (5)$$

For example then, if researchers wanted the R_c^2 after shrinkage to be no less than 87% (i.e., a decrease in R^2 of no more than 13%) of the expected sample R^2 of .53 with four predictors, they would set PP = .87, and calculate $\varepsilon = .069$ to use in sample size formula (3). Plugging the values into formula (3) provides a sample size of

$$N \geq \frac{5\ (2 - 2\ (.53) + .069)}{.069} = 73$$

Thus, 73 subjects should provide a large enough sample so that expected $R_c^2 > .46$, which is 87% of the assumed $\rho^2 = .53$.

Conclusions

The seriousness of concern about sample sizes and precision power in regression is not obvious--after all, researchers have shrinkage and cross-validity formulas available to "correct" for inadequate sample sizes. However, a prediction model produced using a larger sample size will better estimate both ρ^2 (using R_a^2) and ρ_c^2 (using R_c^2); more importantly, it will provide more stable regression weights. Therefore, such a model will predict better in future samples because the efficiency of a prediction model depends not on the estimates of ρ^2 and ρ_c^2, but on the stability of the regression coefficients.

The primary goal of precision power analysis is to reduce the upward bias of R^2, thereby better estimating both ρ^2 and ρ_c^2, so that results are not sample specific. The precision power method provides researchers with a means to determine the optimum sample size for prediction studies. Assuming the researcher can make a reasonable estimate of the population ρ^2, the precision power method provides the most consistent precision power rates of all existing methods. It should be noted that Brooks and Barcikowski's (1995) results apply only to standard regression analysis, where all predictors are entered into the model simultaneously. Many researchers agree, however, that even larger samples are required when preselection or best subset regression analyses are used (Halinski & Feldt, 1970; Nunnally, 1978; Tabachnick & Fidell, 1989).

Unfortunately, no sample size method can eliminate all problems. When researchers choose an expected R^2 that overestimates ρ^2 (either explicitly by choice of an inflated effect size or implicitly by use of an inappropriate rule-of-thumb), power rates are unacceptably low. Similarly, when researchers choose an expected R^2 which is much lower than the population ρ^2, power rates are unnecessarily high (more subjects than necessary are recommended). Therefore, if the researcher cannot make a reasonable estimate of ρ^2, no sample size method will work well. In other words, effect size is just as critical when choosing sample sizes in multiple regression as it is with other statistical methods, because all methods are inadequate when expected R^2 deviates too far from ρ^2.

Researchers who hope to develop an efficient prediction model using multiple regression must be concerned with the size of their derivation samples, starting with an appropriate effect size, probably in the form of an expected R^2. It may be worth noting that although Stevens (1992a) suggested an effect size of $\rho^2 = .50$ as a reasonable guess for the social sciences when a better estimate is unavailable, Rozeboom (1981) believes that $\rho^2 = .50$ may be an upper bound and Cohen (1988) offers $\rho^2 = .26$ as a large effect size. Of course, the best choice of effect size is based on evidence from the research literature or from past research experience. Clearly, effect size impacts the selection of sample size in complex ways. Such discrepancies make it more obvious why some scholars have recommended sample sizes of 100, 200, and even 500, no matter how many predictors, and others have suggested subject-to-variable ratios as large as 40:1 (e.g., Kerlinger & Pedhazur, 1973; Nunnally, 1978; Pedhazur, 1982; Tabachnick & Fidell, 1989).

Another concern that researchers must consider is the question of a priori precision power rate. It is useful to remember that "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" (Tabachnick & Fidell, 1989, p. 129). Given the current state of the research, there are no clear guidelines as to what precision power rate to choose. Similar to choices regarding statistical power and Type I error rates, the importance of generalizability to a study must be considered by researchers. For example, if it is critical that the expected R^2 value not shrink much, the researcher may wish to choose a very high precision power rate.

Summary

Sample sizes for multiple linear regression, particularly when used to develop prediction models, must be chosen so as to provide adequate power both for statistical significance and also for generalizability of the model. It is well-documented and unfortunate that many researchers do not heed this guideline, probably often choosing instead to abide by the rule cited by Olejnik (1984): use as many subjects as you can get. Possibly more tragic are the cases where researchers have used a groundless rule-of-thumb to choose their sample sizes or have neglected to report an appropriate "shrunken" R^2; these studies probably provide inaccurate conclusions regarding the topics under investigation.

For whatever reasons, empirical study into power for multiple regression has been lacking. Rules-of-thumb have existed for decades with little empirical or mathematical support. Indeed, both studies by Brooks and Barcikowski (1994, 1995) have found very limited value for rules-of-thumb in regression. Additionally, sample size methods offered by Park and Dudycha (1974), Cohen (1988), Gatsonis and Sampson (1989), and Sawyer (1982) were each found lacking in some way. The only method which provided consistently accurate power for generalizability was the precision power method.

It is hoped that the information presented within this paper encourages researchers to consider more seriously the issues of power and sample size for multiple linear regression studies. Because power in prediction studies has more meaning than for other statistical designs, it is an even more important consideration. Researchers must recognize the potential danger of choosing an inappropriate effect size (either implicitly or explicitly) or ignoring effect size entirely. Further, no statistical analysis or correction (such as an adjusted R^2) can repair damage caused by an inadequate sample. Researchers must remember that a sample must not only be large enough, but that it must also be random and appropriately representative of the population to which the research will generalize (Cooley & Lohnes, 1971; Miller & Kunce, 1973).

References

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

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

Brooks, G. P., & Barcikowski, R. S. (April, 1994). *A new sample size formula for regression.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Brooks, G. P., & Barcikowski, R. S. (October, 1995). *Precision power method for selecting regression sample sizes.* Paper presented at the meeting of the Mid-Western Educational Research Association, Chicago, IL.

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.

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

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.

Cooley, W. W., & Lohnes, P. R. (1971). *Multivariate data analysis.* New York: John Wiley & Sons.

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

Darlington, R. B. (1990). *Regression and linear models.* New York: McGraw-Hill.

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

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.

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

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

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

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

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

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

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.

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)

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. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart, & Winston.

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

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

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.

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.

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.

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