The PEAR Method for Sample Sizes in Multiple Linear Regression

Gordon P. Brooks **Robert S. Barcikowski**

Ohio University

When multiple linear regression is used to develop prediction models, sample size must be large enough to ensure stable coefficients. If derivation sample sizes are inadequate, the models may not generalize well beyond the current sample. The precision efficacy analysis for regression (PEAR) method uses a cross-validity approach to select sample sizes such that models will predict as well as possible in future samples. The purposes of this study are (a) to verify further the PEAR method for regression sample sizes and (b) to extend the analysis to include an investigation of the effects of multicollinearity on coefficient estimates.

"I have so heavily emphasized the desirability of working with few variables and large sample sizes that some of my students have spread the rumor that my idea of the perfect study is one with 10,000 cases and no variables. They go too far." (Cohen, 1990, p. 1305). Although Darlington (1990), among others, has noted that the best rule for choosing sample sizes is simply that more is better, 10,000 may be just a few more than typically necessary. Indeed, for both statistical and practical reasons, researchers should choose for their sample size "the smallest number of cases that has a decent chance of revealing a relationship of a specified size" (Tabachnick & Fidell, 2001, p. 117). When generalizability of a regression model is the concern, as it is when regression is used to develop prediction models; however, this concept translates as the smallest sample that will provide the reliability of results required across multiple samples. Especially in multiple linear regression, which is used for many purposes, necessary sample size depends heavily on the goals and design of the analysis. For example, "at one extreme, the null hypothesis $\rho = 0$ can often be tested powerfully with only a few dozen cases. At the other extreme, hundreds or thousands of cases might be needed to accurately estimate the sizes of higher-order collinear interactions" (Darlington, 1990, p. 380). Consequently, the selection of adequate and appropriate sample sizes is not always an easy matter in regression.

The purpose of this study was to examine further the efficiency of the precision efficacy analysis for regression (PEAR) method for calculating appropriate sample sizes in regression studies where generalizability is a concern. Even though several methods currently exist to help researchers choose regression sample sizes, none use the straightforward approach taken here by the PEAR method that essentially uses an effect size within a single formula to determine the subject-to-variable ratio appropriate for the squared multiple correlation expected in a given study. The PEAR method, which is based on the algebraic manipulation of an accepted cross-validation formula, uses a cross-validity approach to sample size selection that limits the amount of expected shrinkage in R^2 so that regression models will predict as well as possible for future subjects.

Background

Unfortunately, many researchers apparently hold erroneous beliefs that smaller calculated probability values mean that "increasingly greater confidence can be vested in a conclusion that sample results are replicable" (Thompson, 1996, p. 27). Statistical significance indicates neither the magnitude nor the importance of a result (Shafer, 1993). Indeed, with a large enough sample size, a statistically significant result may be obtained even though there is very little relationship between the criterion and the predictor variables (Asher, 1993).

In particular, ordinary least squares multiple linear regression can result in a model being statistically significant, but with that model providing unrealistic estimates for the relationships under investigation. The process of maximizing the correlation between the observed and predicted criterion scores involves mathematical capitalization on chance sampling error variation. When the regression equation is used with a second sample, or future cases from the same population, it is most likely that the model will not perform as well as it did in the original sample; consequently, the estimate of the population multiple correlation will decrease in the second sample.

Sample sizes for multiple 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. From a statistical power perspective, a study with insufficient sample size stands a large chance of committing a Type II error. From a generalizability viewpoint, an insufficient sample leads to results that may apply only to the current sample and will not be useful for application to other samples. In either case, time, effort, and money would have been spent arriving at results that are inconclusive or useless. Probably most unfortunate are the cases where researchers use what they think is a good sample size rule when, in fact, it is nothing more than a groundless convention that ignores effect size completely or a statistical power approach that will not provide the sample sizes necessary for generalization.

Existing Sample Size Methods for Regression

Historically, there are three primary types of sample size methods available for multiple linear regression: conventional rules, statistical power approaches, and cross-validation approaches. The following sections describe each briefly, with emphasis on problems associated with each.

Conventional Rules. Because cross-validity formula estimates are primarily functions of sample size and the number of predictors, conventions have evolved that are based on the premise that with a large enough ratio of subjects to predictors, the sample regression coefficients will be reliable and will closely estimate the true population values (Miller & Kunce, 1973; Pedhazur & Schmelkin, 1991; Tabachnick & Fidell, 2001). For example, Stevens (2002) suggested a ratio of 15 subjects for each predictor (e.g., with 3 predictors 45 subjects are required) and Pedhazur and Schmelkin (1991) recommended $N \geq 30k$, where k is the number of predictors. Others have provided rules that combine some minimum value with a subject-to-predictor ratio, including $N \geq 30 + 10k$ (Knapp & Campbell-Heider, 1989), $N \geq 50 + k$ (Harris, 1985), and $N \geq 50 + 8k$ (Green, 1991). Sawyer (1982) developed a formula where setting an inflation factor to a constant of 5% results in a sample size recommendation of $N \geq 10.8k + 11.8$.

Unfortunately, because most of these rules lack any measure of effect size, they can only be effective at specific—usually unknown—effect sizes. For example, a $15{:}\,1$ subject-to-predictor ratio is acceptable only if the population squared multiple correlation is moderately large (i.e., over .40); otherwise, as the true squared multiple correlation decreases, expected cross-validity shrinks so much as to make the prediction model worthless (Brooks, 1998). For example, Stevens (2002) is explicit in describing how he derived his recommendation of a $15{:}\,1$ ratio based on Park and Dudycha's (1974) tables, but others are not so clear. Over time, the evolution of these rules causes their origins and rationales to become fuzzy. For example, someone who recommended a $10{:}\,1$ rule may have analyzed datasets that coincidentally all had an R^2 near .50.

Statistical Power Methods. 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; Green, 1991; Kraemer & Thiemann, 1987; Milton, 1986). From a statistical power perspective, multiple linear regression provides several alternative statistical significance tests that can be the basis for sample size selection. Two statistical tests are most common in practice: (a) the test of the full model (i.e., the overall or omnibus test), and (b) the test of the individual regression coefficients in the model.

Unfortunately, sample sizes that provide adequate statistical power to reject a regression null hypothesis may not provide the stable regression coefficients required for prediction and model-building. Therefore, choosing a sample size based on statistical power may not ensure that a regression function will generalize to other samples from the target population, which is the crucial factor in determining the validity of regression models used for prediction. That is, adequate sample sizes for statistical power tell us nothing about the number of subjects needed to obtain precise estimates of stable, meaningful regression weights (Cascio, Valenzi, & Silbey, 1978; Darlington, 1990). Although Gatsonis and Sampson (1989), Darlington (1990), and Maxwell (2000) have proposed methods for cross-validation using a random model approach, their methods are also based on a statistical power approach to sample size determination rather than a cross-validity approach.

Cross-Validity Methods. From a random model perspective, both the predictors and the criterion are sampled together from what is usually assumed to be a joint multivariate normal distribution. 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). The random model is usually more appropriate for social scientists because they typically measure random subjects on predictors and a criterion simultaneously and; therefore, are not able to fix the values for the independent variables (Darlington, 1990; Drasgow, Dorans, & Tucker, 1979; Herzberg, 1969; Park & Dudycha, 1974). Park and Dudycha (1974) noted that such a cross-validation approach is applicable to both the random and the fixed models of multiple linear regression; however, they emphasized the random model, cross-validation approach. Unfortunately, they published tables that were limited to only a few possible combinations of squared correlation and number of predictors. Additionally, there is no clear rationale for how to determine the best choice of either tolerance or the probability to use when consulting the tables (although Stevens, 2002, implied through example that .05 and .90, respectively, are acceptable values). More recently, Algina and Keselman (2000) used Monte Carlo methods to develop tables that list the sample sizes required to ensure that cross-validity estimates maintain a desired level relative to sample R^2 statistics. Their theoretical development of the study appears to be similar to the work done by Park and Dudycha and; indeed, their tables provide essentially the same sample sizes as those created by Park and Dudycha. The PEAR method follows a similar theoretical approach as both these methods.

Other Methods**.** Darlington (1990) has provided a method to calculate the sample size necessary for a validation sample; that is, Darlington's Fisher *z* method provides recommendations for the second sample used to verify the regression model derived from the original sample rather than for the initial, derivation sample. Although several scholars have proposed methods to determine sample sizes for better estimation of the population squared multiple correlation or change in squared multiple correlation (Algina & Keselman, 2000; Algina, Keselman, & Penfield, 2007; Algina & Moulder, 2001; Algina & Olejnik, 2000, 2003; Darlington, 1990; Knofczynski & Mundfrom, 2008) or squared semi-partial correlations (Algina, Keselman, & Penfield, 2008; Algina, Moulder, & Moser, 2002), this type of approach is not useful for determining sample sizes needed for better estimation of cross-validity coefficients. Kelley and Maxwell (2003) have developed a sample size method based on obtaining regression coefficients based on what they have called accuracy in parameter estimation (AIPE). Although it relies on precision of parameter estimates rather than statistical power, the AIPE approach does not address cross-validity directly.

Development of the PEAR Method

The primary goal of the PEAR method is to reduce the upward bias in R^2, thereby enhancing the cross-validity potential of the regression model so that results are less likely to be sample specific. That is, the PEAR method answers the question: "What is the minimum sample size that will yield a regression equation that will cross-validate with minimal shrinkage in another sample and thereby assure more stable regression coefficients?"

In a sense, the PEAR method can be viewed as cross-validation in reverse. That is, instead of determining by how much the sample R^2 will shrink due to the sample size, the PEAR method determines how large a sample is required to keep R^2 from shrinking too much. Like the work done by Park and Dudycha (1974), the theory underlying the PEAR method for sample size selection is that the researcher, knowing that cross-validation R^2 values are typically lower than the sample R^2 statistics (the difference is typically called shrinkage), can set a tolerance limit as to the amount of shrinkage expected to occur. The concepts of cross-validity, precision efficacy, effect size, and shrinkage tolerance serve as the foundation for using the PEAR method to, in Stevens' (2002) terms, "keep the shrinkage fairly small" (p. 146).

Cross-Validity Shrinkage

"Although we may determine from a sample R^2 that the population R^2 is not likely to be zero, it is nevertheless not true that the sample R^2 is a good estimate of the population R^2" (Cohen & Cohen, 1983, p. 105). Population R^2, or ρ^2, is the unknowable squared multiple correlation that would be obtained between the criterion variable and the specified linear combination of predictors if both were measured in the population. Because this parameter is useful in estimating the strength of the relationship between the criterion variable and a set of regressors in the population, it is of particular interest in descriptive and explanatory research (Kromrey & Hines, 1995). Sample R^2 is a positively biased estimator of ρ^2; however, such that the expected value of R^2 given by Herzberg (1969) is:

$$E(R^2) = \rho^2 + \frac{k}{N-1}(1-\rho^2)\ , \qquad (1)$$

where k is the number of predictors and N is the sample size. For example, if the null hypothesis is true and therefore $\rho^2 = 0$ with 5 predictors and $N = 11$, we would expect a sample R^2 to be near .50 rather than near 0; if $\rho^2 = .50$ with 5 predictors and $N = 11$, we would expect a sample R^2 close to .75. To account for this bias, most computer packages report an adjusted R^2 calculated using a formula most frequently attributed to Wherry:

$$R_A^2 = 1 - \left(\frac{N-1}{N-k-1}\right)(1 - R^2) \tag{2}$$

where k is the number of predictors and N is the sample size. For example, a researcher who computed sample $R^2 = .40$ with 60 subjects and 4 predictors might use the adjusted R^2 formula to conclude that, in the population, the proportion of variation accounted for in the criterion by the predictors is actually closer to $R_A^2 = .3564$.

Although R_A^2 is usually appropriate and frequently reported for questions concerning explanation and description, most problems of prediction and generalizability require a different type of correction formula, commonly called cross-validity formulas (Darlington, 1990). Herzberg (1969) noted that for application "one is more interested in how effective the *sample* regression function is in *other* samples" (p. 4). Indeed, the development of prediction models that are useful in future samples is one of the most common and most important uses of regression equations in the social sciences (Huberty, 1989; Weisberg, 1985). From this perspective of generalizability, an insufficient sample may lead to results that, even though statistically significant, apply only to the current sample and will not be useful for application to other samples. Mosteller and Tukey (1968) wrote that "users have often been disappointed by procedures, such as multiple regression equations, that 'forecast' quite well for the data on which they were built. When tried on fresh data, the predictive power of these procedures fell dismally" (p. 110). Researchers, therefore, should use and report strategies that evaluate the replicability of their results, which is necessary to provide confidence in the results; one way to gauge this generalizability is through an estimate of cross-validity.

Cross-validity shrinkage is the size of the decrease in the sample R^2 when an appropriate cross-validity formula is applied. Cross-validity shrinkage, ε, is defined as the difference between sample R^2 and the cross-validity estimate, R_C^2:

$$\varepsilon = R^2 - R_C^2 \tag{3}$$

Darlington (1990) has defined shrinkage as the difference between a regression model's apparent validity, as measured by R^2, and its actual predictive cross-validity. Essentially, shrinkage is what some authors (e.g., Cattin, 1980; Mosteller & Tukey, 1968; Stevens, 2002) call the loss in predictive power that occurs when small samples cause a reduction in R^2 such that $E(R^2) > \rho^2 > \rho_C^2$ (Herzberg, 1969). The squared cross-validity coefficient, ρ_C^2, is considered to be the squared multiple correlation between the actual population criterion values and the scores predicted by the sample regression equation when applied either to the population or to another sample (Huberty & Mourad, 1980; Schmitt, Coyle, & Rauschenberger, 1977).

Cross-validity formulas, which are symbolized by R_C^2 and based on estimates of the mean squared error of prediction (Darlington, 1968; Herzberg, 1969), provide more accurate estimates of ρ_C^2 than do the sample R^2 values. A number of cross-validity formulas have been proposed (e.g., Browne, 1975; Darlington, 1968; Herzberg, 1969; Lord, 1950; Nicholson, 1960; Rozeboom, 1978; Stein, 1960). Formula-based methods of cross-validity estimation are often preferred to empirical cross-validation (e.g., data splitting) so that the entire sample may be used for model-building. Indeed, several formula estimates have been shown to be at least as accurate as empirical cross-validation techniques (Cattin, 1980; Drasgow et al., 1978; Morris, 1981; Rozeboom, 1978; Schmitt et al., 1977). When cross-validity is estimated using a formula, any finite sample size will result in a cross-validity estimate, R_C^2, smaller than the sample R^2.

For example, using the random model cross-validity formula developed independently by Stein (1960) and Darlington (1968),

$$R_C^2 = 1 - \left(\frac{N-1}{N-k-1}\right)\left(\frac{N-2}{N-k-2}\right)\left(\frac{N+1}{N}\right)(1 - R^2)\ , \tag{4}$$

where k is the number of predictors and N is the sample size, a researcher who computed sample $R^2 = .40$ with 60 subjects and 4 predictors would calculate a shrunken $R_C^2 = .2972$ (note that $R_A^2 = .3564$ for these conditions). Although the researcher might explain 40% of the dependent variable variance in the current sample, or perhaps 36% in the population, the cross-validity estimate suggests that less than 30% of the variance will be explained when the same regression model is applied to future samples, which is shrinkage of 10%. The cross-validity estimates result in more shrinkage because they must correct for the sampling error present not only in the derivation sample but also in some future sample (Snyder & Lawson, 1993).

It should be noted that researchers are often interested in both prediction and description; that is, they would like to know not only how well the regression model will work in future samples, but also would like a relatively accurate estimate of the population ρ^2 value. In such cases, researchers should report both a cross-validity estimate and an adjusted R^2 estimate. These formulas estimate different parameters and are not interchangeable. In large normally distributed samples, the mean, median, and mode converge—but no one would argue that these are equivalent measures of central tendency. Similarly, R^2, R_A^2, and R_C^2 converge with large samples, but R_A^2 provides better estimates of the true population ρ^2 than does any cross-validity estimate (e.g., Carter, 1979). As Stevens (2002) indicated, however, “use of the Wherry formula would give a misleadingly positive impression of the cross validity predictive power of the [regression] equation” (p. 118).

Finally, because researchers have these cross-validity formulas available to correct for inadequate sample sizes, the importance of sample size for generalizability of regression results is not immediately obvious. By limiting the upward bias of sample R^2 values; however, a regression model produced using a larger, more adequate sample size will better estimate both ρ^2 and ρ_C^2. For example, the true population ρ_C^2 in the example cited above is probably larger than .2972; indeed, the true ρ^2 may be larger than .3564 because the small sample size limited the accuracy of these estimates.

Precision Efficacy

The term precision efficacy (PE) is proposed to describe how well a regression model is expected to perform when applied to future subjects relative to its effectiveness in the derivation sample. The formal definition of precision efficacy proposed here is:

$$PE = \frac{R_C^2}{R^2}, \qquad (5)$$

where R^2 is the sample proportion of variation accounted for and R_C^2 is the sample cross-validity estimate. Precision efficacy can be considered a measure of the cross-validitional power of a regression model: higher precision efficacy indicates more efficiency in term of cross-validity.

Because they desire regression models that generalize well to other samples, researchers who develop prediction models should hope to limit shrinkage as much as possible relative to the sample R^2 they obtained. Using an example from Stevens (2002, p. 118), 62% shrinkage from a sample $R^2 = .50$ to $R_C^2 = .191$ occurs based on Formula 4 with a sample size of 50; but if the sample size had been 150, there would only have been 16% shrinkage to $R_C^2 = .421$. Precision efficacy in the first case would be $PE = .382$ and in the second case, $PE = .842$. Consequently, even if the R^2 value were statistically significant in the former case, the results may not be expected to perform well enough for the model to be useful with future samples. Larger precision efficacy values imply that a regression model is expected to generalize better for future samples; that is, high PE values indicate that the regression model will work in other samples nearly as well as it did in the derivation sample.

Effect Size

In multiple regression research, perhaps the most common effect size is the squared multiple correlation, R^2. Effect size enables a researcher to decide *a priori* not only what size relationship will be necessary for statistical significance, but also what relationship should be considered for practical significance (Hinkle & Oliver, 1983). For example, because under 10% explained variance may not provide any new, useful knowledge in a field, a researcher may choose a minimum practical effect size of 20%. In multiple regression; however, the researcher must remember the effects of shrinkage. That is, if a researcher chooses 20% explained variance (i.e., $\rho^2 = .20$) as an effect size worthy of study, that researcher does not want a corrected sample estimate (e.g., R_A^2 or R_C^2) to be .05. No matter how it is chosen, an expected effect size must be determined *a priori*. In many cases, the researcher may have some basis for deciding the smallest correlation that would be interesting to find, based perhaps on experience, previous research, pilot studies, or practical significance. Researchers should remember, though, that "meta-analyses often reveal a sobering fact: effect sizes are not nearly as large as we all might hope" (Light, Singer, & Willett, 1990, p. 195).

Stevens (2002) emphasized that the magnitude of the population squared multiple correlation "strongly affects how many subjects will be needed for a reliable regression equation" (p. 146). Stevens demonstrated that 15 subjects per predictor are needed to keep shrinkage small if .40 is used as R^2 in the Stein cross-validity formula, but that fewer are needed if $R^2 = .70$. Similarly, Huberty (1994) noted that "the magnitude of R^2 should be considered in addition to N/p ratios when assessing the percent of shrinkage of R^2 that would result in the estimation process. That is, a general rule of thumb for a desirable N/p ratio (say, $10/1$) may not be applicable across many areas of study" (p. 356). Indeed, all sample size methods that account for effect size agree: as effect size decreases, sample size must increase proportionately (e.g., Cohen, 1988; Darlington, 1990; Milton, 1986; Park & Dudycha, 1974; Gatsonis & Sampson, 1989). Therefore, the first task in a sample size analysis for regression must be the identification of the magnitude of the multiple correlation expected in the population. Unfortunately, as Schafer (1993) noted, "if one knew the answer to that question one would not need to do the study, but a value is needed anyway" (p. 387).

When researchers have no empirical basis for deciding on an effect size, Light et al. (1990) offered as a starting point that effect size should be "the minimum effect size you consider worthy of your time" (p. 194). Stevens (2002) has suggested that $\rho^2 = .50$ is a reasonable guess for social science research; Rozeboom (1981), however, believed that $\rho^2 = .50$ is an upper limit for reasonable effect sizes. Cohen (1988) suggested that $\rho^2 = .26$ is a large regression effect size ($\rho^2 = .13$ and $\rho^2 = .02$ are Cohen's medium and small effect sizes, respectively). Indeed, because an effect of $\rho^2 = .25$ seemed unreasonably large to Schafer (1993), he recommended that it serve as an upper limit only as a last resort, when no other rationale is available.

Shrinkage Tolerance

Simply put, shrinkage is the size of the decrease in the sample R^2 when an appropriate cross-validity formula is applied. Shrinkage tolerance, an *a priori* definition of acceptable shrinkage, can also be calculated using Equation 3, but using *a priori* estimates of effect size rather than estimates calculated from the sample. Here, shrinkage tolerance (ε, but set *a priori*) can be considered either absolute or relative. In an absolute sense, ε can be set to a specific value regardless of the effect size expected in a given study. That is, no matter what ρ^2 is to be used as an effect size, the researcher may wish that the expected shrinkage stay within a given distance from the sample R^2 value. For example, if R^2 turns out to be .50 and the researcher has chosen $\varepsilon = .10$, R_C^2 is desired to be near .40; but if $R^2 = .25$, the researcher is willing to accept an R_C^2 of .15 when $\varepsilon = .10$. Algina and Keselman (2000) and Park and Dudycha (1974) used this idea of absolute shrinkage tolerance as an accuracy criterion. Although useful in some contexts, the absolute loss in predictive power does not provide any sense of the magnitude of loss as compared to the original sample R^2. For example, a loss in predictive power of .10 suggests drastically different implications for generalizability if $R^2 = .50$, where R_C^2 would be .40 (a proportional decrease of 20%), than if $R^2 = .25$, where $R_C^2 = .15$ (a proportional decrease of 40%). Therefore, in a relative sense, shrinkage tolerance can be set *a priori* to some proportional decrease in the sample R^2. For example, if sample $R^2 = .50$ with *a priori* $\varepsilon = .2\rho^2$, R_C^2 will be expected to shrink only by 20% to $R_C^2 = .40$ (i.e., R_C^2 will be 80% as large as R^2), or $R^2 = .25$ will shrink only to $R_C^2 = .20$.

Proportional Shrinkage. Proportional shrinkage (*PS*) is defined as the amount of shrinkage relative to R^2 that occurs after a cross-validity estimate is calculated from the data. Proportional shrinkage is calculated as

$$PS = \frac{R^2 - R_C^2}{R^2} \tag{6}$$

For example, if sample $R^2 = .50$ and $R_C^2 = .26$, the proportional shrinkage would be calculated to be .48, suggesting limited generalizability for the regression model because R^2 shrank by almost half. Lower proportional shrinkage values imply better generalizability.

Employing Equation 3, the formula for calculating precision efficacy, Equation 5, can be written as $1 - PS$, or

$$PE = 1 - \left(\frac{\varepsilon}{R^2}\right) \qquad (7)$$

For example, setting the predetermined acceptable shrinkage level at $\varepsilon = .2\rho^2$ provides $PE = .80$. Solving Equation 7 for ε and replacing the actual sample R^2 with ρ^2, which is the estimated *a priori* effect size chosen by the researcher, results in the formula

$$\varepsilon = \rho^2 - (PE)(\rho^2) \qquad (8)$$

Using this formula, when an effect size has been chosen along with a desired level of precision efficacy, the acceptable shrinkage tolerance can be determined. For example, if the researcher wishes to obtain a cross-validity estimate expected to be not less than 80% of the sample R^2, *a priori* precision efficacy would be .80. If the effect size is chosen to be $\rho^2 = .40$, then shrinkage tolerance for this example would be calculated to be $\varepsilon = .08$.

The PEAR Formula

The formula used in the PEAR sample size method was developed based on a cross-validity formula by Lord (as cited in Uhl & Eisenberg, 1970):

$$R_C^2 = 1 - \left(\frac{N + k + 1}{N - k - 1}\right)(1 - R^2) \qquad (9)$$

where N is sample size and k is the number of predictors. Uhl and Eisenberg (1970, p. 489) found this "relatively unknown formula" (their interpretation of Lord's 1950 paper differs from others) to give accurate estimates of "cross-sample" shrinkage, regardless of sample size and number of predictors. Algebraic manipulation of Equation 9 to solve for sample size yields the formula at the foundation of the PEAR method:

$$N \geq \left(\frac{2 - 2\rho^2 + \varepsilon}{\varepsilon}\right)(k + 1) \qquad (10)$$

where ε is the *a priori* shrinkage tolerance, k is the number of predictors, and ρ^2 is the *a priori* effect size. Shrinkage tolerance allows researchers to decide how closely to estimate ρ_C^2, whether absolutely (e.g., $\varepsilon = .05$) or relatively (e.g., $\varepsilon = .2\rho^2$). The level of precision efficacy itself is embedded within the shrinkage tolerance value in Equation 10 (through Equations 7 and 8). Note that when a proportional definition of shrinkage tolerance is used, Equation 10 simplifies slightly; for example, if $\varepsilon = .2\rho^2$ is used for desired $PE = .80$, then Equation 10 simplifies to

$$N \geq \left(\frac{2 - 1.8\rho^2}{0.2\rho^2}\right)(k + 1) \qquad (11)$$

Examples. If a researcher has three predictors and wants to ensure that R_C^2 is within .05 of the expected sample R^2 value of .50 (i.e., the effect size is $\rho^2 = .50$), ε would be set at .05 and Equation 10 will provide a recommended sample size of 84. The implied precision efficacy of this result is based on the shrinkage tolerance value; using Equation 7, $PE = .90$. If the researcher wishes to maintain shrinkage of about 20% with six predictors and an effect size of .40, Equation 11 provides a recommended sample size of 112. Finally, if a researcher wants an R_C^2 estimate to be at least 87% of the expected effect size of $\rho^2 = .53$ with four predictors, precision efficacy should be set to .87 and, based on Equation 8, $\varepsilon = .069$. Substituting these values into Equation 10 calculates a necessary sample size of 73.12. Therefore, at least 74 subjects should provide a large enough sample so that R_C^2 is expected to be at least .46, which is 87% of the expected R^2 of .53.

It is worth noting that the PEAR method formula essentially results in subject-to-variable ratios appropriate for given effect sizes (i.e., the PEAR method results in a subject-to-variable ratio rather than the more commonly used subject-to-predictor conventional rules). For example, using the criteria in Table 1, where $\varepsilon = \rho^2 - (PE - .1PS)\rho^2$ and *PS=1-PE,* at an effect size of expected $\rho^2 = .40$, the PEAR method suggests a subject-to-variable ratio of approximately $15:1$ for PE = .80. With the same criteria at an expected $\rho^2 = .20$, however, the number of cases required per variable increases to $37:1$ (see the $PE = .80$ column in Table 1).

Because previous work (Brooks, 1998; Brooks & Barcikowski, 1994, 1995, 1999) has found the PEAR method to be superior to a number of regression sample size methods (e.g., Cohen, 1988; Gatsonis & Sampson, 1989; Green, 1991; Park & Dudycha, 1974; Pedhazur & Schmelkin, 1991; Sawyer, 1982; Stevens, 1996) in limiting cross-validity shrinkage to given acceptable *a priori* levels of PE, the current study examines impact of the PEAR method sample sizes on the variance of the regression coefficients. First, does the PEAR method recommend sample sizes that enable the derivation of reliable regression coefficients (that is, coefficients with small standard errors)? In order to examine the stability of the coefficients, the standard errors of the coefficients are of primary interest. One would expect that a model based on a proper sample size will provide more reliable regression weights and; therefore, predict better for future subjects. Second, the impact of multicollinearity is investigated as it relates to sample sizes recommended by the PEAR method.

Method

A Monte Carlo analysis of precision efficacy rates was performed. The three PEAR method *a priori* precision efficacy levels of .60, .70, and .80 (which correspond to squared cross-validity estimates expected to be at least 60%, 70%, and 80% of the sample R^2 values, respectively) were considered to be individual methods for the analysis. That is, sample sizes were calculated using these PE levels with the PEAR method. Comparisons of the varying precision efficacy levels of the PEAR method helped to determine the effects of larger and smaller sample sizes on the regression coefficients.

Three factors were manipulated to comprise the testing situations for the study. First, three effect sizes that represent simultaneously the estimated population squared multiple correlation ρ^2 and the true population ρ^2 were set at: .10, .25, and .40. The numbers of predictors used to define the models in this study were 3 predictors (i.e., 4 variables including the criterion), 7, 11, and 15 predictors. Finally, four multicollinearity conditions were explored in the study: (1) *extensive* multicollinearity was defined as over one-half of the predictors with $VIF_j > 5.0$, (2) *moderate* multicollinearity was defined as one-quarter of the predictors involved in such a multicollinear relationship, (3) for all predictors in the *trivial* multicollinearity condition, $VIF_j < 3.0$, and (4) the correlation matrix for the *orthogonal* condition contained zero correlations among all predictors.

Table 1. Subjects per Variable[a] Sample Size Ratios from the PEAR Method and the 15:1 Ratio

	Precision Efficacy (*PE*)			
ρ^2	.60	.70	.80	*15:1* ratio
.05	87.4	116.2	173.7	15.0
.10	41.9	55.5	82.8	15.0
.15	26.8	35.3	52.5	15.0
.20	19.2	25.2	37.4	15.0
.25	14.6	19.2	28.3	15.0
.30	11.6	15.1	22.2	15.0
.35	9.4	12.3	17.9	15.0
.40	7.8	10.1	14.6	15.0
.45	6.6	8.4	12.1	15.0
.50	5.5	7.1	10.1	15.0
.55	4.7	6.0	8.4	15.0
.60	4.0	5.0	7.1	15.0
.65	3.4	4.3	5.9	15.0
.70	2.9	3.6	4.9	15.0
.75	2.5	3.0	4.0	15.0

Note. Here, $\varepsilon=\rho^2-(PE-.1PS)\rho^2$, where $PS=1-PE$ and ρ^2 is the estimated population value. To calculate N, multiply the number of variables by the tabled value and round to the next larger integer if necessary.

[a] number of variables is $(k+1)$, where k is the number of predictors.

A Delphi Pascal program was created for an original algorithm used to create 48 population correlation matrices (i.e., 3x4x4) to meet the above criteria required by this study. These correlation matrices were treated as population correlation matrices from which joint multivariate normal data from a random model perspective were generated for each sample in the study. Delphi Pascal procedures were developed to generate sample data through a process that converted uniformly distributed pseudorandom numbers created by the L'Ecuyer (1988) combined multiplicative congruential generator (translated from Press, Teukolsky, Vetterling, & Flannery, 1992) into multivariate-normally distributed data using the Box-Muller transformation (adapted from Press, Flannery, Teukolsky, and Vetterling, 1989) and the Cholesky decomposition (adapted from Nash, 1990; recommended by several scholars including Bratley, Fox, & Schrage, 1987; Chambers, 1977; Kennedy & Gentle, 1980; Morgan, 1984; Ripley, 1987; Rubinstein, 1981). Finally, these procedures were incorporated into a Delphi Pascal program that performed the Monte Carlo simulation with 10,000 iterations.

During program execution, several statistics were computed and recorded. For each sample, the program performed a standard multiple linear regression analysis. The program first calculated the necessary information from the full-model regression with all predictors entered simultaneously for each sample (e.g., PE, R^2, Wherry R_A^2, Stein R_C^2, B_j, SE_{Bj}). Both R_A^2 and R_C^2 were set equal to zero when they were negative, as recommended by Cohen and Cohen (1983) and Darlington (1990). These data were averaged over the number of iterations for each condition. Finally, counts were made for several statistics regarding their significance or accuracy. For example, statistical significance at $\alpha = .05$ was tested for both the full regression model and the regression coefficients, as was the accuracy of PE and R_C^2.

In addition to these raw statistics, the appropriate calculations were made and data were collected as required for bias, root mean squared error (*RMSE*), Relative Efficiency, and the standard deviations of several key estimates. Statistical bias is defined as

$$Bias = E\left(\hat{\theta}\right) - \theta \qquad (12)$$

where θ is the known population parameter and $E(\hat{\theta})$ is the expected value of the sample statistic, which is the average of the statistic over infinite samples (Drasgow et al., 1979; Kromrey & Hines, 1995; Mooney, 1997). *RMSE* provides an indication of the statistic's variability:

$$RMSE\left(\hat{\theta}\right) = \sqrt{\sum \frac{\left(\hat{\theta}_\imath - \theta\right)^2}{N}} \qquad (13)$$

where θ is the known population parameter (as set in the computer algorithm), θ_i is the estimate of that parameter obtained in sample *i* of the Monte Carlo simulation, and *N* is the total number of samples taken in the Monte Carlo study (Darlington, 1996; Drasgow et al., 1979; Mooney, 1997). Mooney defined Relative Efficiency as the ratio of two *RMSE* values, multiplied by 100 to convert it to a percentage:

$$Relative\ Efficiency = 100 * \left(\frac{RMSE\left(\widehat{\theta_A}\right)}{RMSE\left(\widehat{\theta_B}\right)}\right) \qquad (14)$$

where $\widehat{\theta_A}$ and $\widehat{\theta_B}$ are two different estimates the same parameter. Values under 100 would indicate the superiority of estimator $\widehat{\theta_A}$ (i.e., $\widehat{\theta_A}$ with smaller *RMSE*).

Results

The PEAR method recommended sample sizes that provided reliable regression coefficients. More specifically, higher *PE* levels provided more stable coefficients. For the conditions with three predictors, Table 2 provides the standard errors of the coefficients for the four sample size methods; similarly, Table 3 provides this information for seven predictor models. These tables show that the *PE* levels that recommended larger samples consistently resulted in smaller standard errors of the coefficients, regardless of the number of predictors or effect size. Although the problem of multicollinearity was not cured by the PEAR method, higher levels of *PE* do indeed help to alleviate the effects. The results not presented showed similar patterns for the 11 and 15 predictor cases as well. That is, across all studied conditions, multicollinearity did increase standard errors as expected, but larger sample sizes tended to mitigate the impact.

Table 4 provides the relative efficiency of the methods compared for all numbers of predictors, all multicollinearity levels, and all effect sizes. For this table, the standard errors for the individual predictors were used for comparison because, for unbiased estimates such as the regression coefficients, *RMSE*

Table 2. Average Standard Errors of the Standardized Coefficients (SE_{Bj}) for Three Predictors

Multicollinearity	ρ^2	Method	N	SE_{B1}	SE_{B2}	SE_{B3}
Orthogonal	.40	PE = .80	59	.102	.103	.094
		PE = .70	40	.126	.126	.118
		PE = .60	31	.147	.147	.136
	.25	PE = .80	113	.080	.080	.079
		PE = .70	77	.098	.099	.097
		PE = .60	59	.114	.113	.111
	.10	PE = .80	331	.052	.052	.050
		PE = .70	222	.064	.064	.062
		PE = .60	168	.074	.073	.071
Trivial	.40	PE = .80	59	.108	.108	.096
		PE = .70	40	.134	.135	.120
		PE = .60	31	.155	.155	.139
	.25	PE = .80	113	.139	.082	.136
		PE = .70	77	.170	.100	.166
		PE = .60	59	.195	.115	.193
	.10	PE = .80	331	.071	.066	.055
		PE = .70	222	.089	.083	.068
		PE = .60	168	.101	.095	.079
Moderate	.40	PE = .80	59	.202	.254 [a]	.140
		PE = .70	40	.254	.312 [a]	.173
		PE = .60	31	.295	.365 [a]	.201
	.25	PE = .80	113	.154	.213 [a]	.146
		PE = .70	77	.189	.260 [a]	.177
		PE = .60	59	.218	.302 [a]	.210
	.10	PE = .80	331	.114	.151 [a]	.090
		PE = .70	222	.140	.187 [a]	.113
		PE = .60	168	.160	.213 [a]	.128
Extensive	.40	PE = .80	59	.183	.264 [a]	.308 [a]
		PE = .70	40	.228	.327 [a]	.382 [a]
		PE = .60	31	.264	.387 [a]	.453 [a]
	.25	PE = .80	113	.129	.381 [a]	.407 [a]
		PE = .70	77	.158	.466 [a]	.499 [a]
		PE = .60	59	.179	.537 [a]	.573 [a]
	.10	PE = .80	331	.128 [a]	.124 [a]	.065
		PE = .70	222	.156 [a]	.152 [a]	.080
		PE = .60	168	.180 [a]	.176 [a]	.093

[a] indicates predictors with *VIF*> 5.0 (i.e., predictors involved in multicollinearity).

approximates the standard error. To create Table 4, the relative efficiency of each predictor was calculated and then those values were averaged for the predictor set. It would not have been appropriate to average the results for Table 4 across predictors if the relative efficiency results had not been so consistent (as can be calculated from Table 2).

There is a striking similarity between the relative efficiency statistics in Table 4 and those found by Brooks (1998) for the correlation statistics. Specifically, the relative efficiency statistics show that, regardless of multicollinearity level, the magnitude of the standard errors of the coefficients from the $PE = .80$ level were, on average, about 19% smaller than those from the $PE = .70$ level. Similarly, Relative Efficiency comparisons of the $PE = .70$ and $PE = .60$ levels showed $PE = .70$ to be approximately 13% more efficient in terms of standard errors.

Table 3. Average Standard Errors of the Standardized Coefficients (SE_{Bj}) for Seven Predictors

Multicollinearity	ρ^2	Method	N	SE_{B1}	SE_{B2}	SE_{B3}	SE_{B4}	SE_{B5}	SE_{B6}	SE_{B7}
Orthogonal	.40	PE = .80	117	.074	.073	.073	.068	.073	.074	.074
		PE = .70	81	.091	.089	.089	.083	.089	.091	.091
		PE = .60	63	.105	.103	.102	.098	.105	.104	.105
	.25	PE = .80	226	.058	.058	.056	.056	.058	.058	.058
		PE = .70	153	.071	.072	.069	.069	.071	.072	.071
		PE = .60	117	.084	.083	.080	.080	.082	.083	.082
	.10	PE = .80	663	.036	.037	.037	.036	.037	.037	.037
		PE = .70	444	.044	.044	.045	.045	.046	.046	.045
		PE = .60	335	.051	.052	.053	.052	.052	.053	.052
Trivial	.40	PE = .80	117	.100	.102	.100	.097	.109	.091	.081
		PE = .70	81	.123	.124	.123	.119	.135	.111	.099
		PE = .60	63	.142	.144	.141	.138	.156	.126	.116
	.25	PE = .80	226	.070	.085	.070	.064	.090	.071	.079
		PE = .70	153	.087	.105	.086	.078	.109	.086	.098
		PE = .60	117	.099	.121	.098	.089	.127	.099	.113
	.10	PE = .80	663	.043	.042	.050	.061	.052	.057	.054
		PE = .70	444	.053	.052	.060	.075	.065	.069	.066
		PE = .60	335	.060	.061	.071	.086	.075	.080	.076
Moderate	.40	PE = .80	117	.192[a]	.137	.141	.177[a]	.154	.094	.130
		PE = .70	81	.236[a]	.170	.174	.219[a]	.188	.116	.158
		PE = .60	63	.270[a]	.191	.200	.249[a]	.215	.132	.182
	.25	PE = .80	226	.129	.089	.130[a]	.079	.080	.074	.180[a]
		PE = .70	153	.159	.109	.160[a]	.097	.099	.092	.220[a]
		PE = .60	117	.184	.126	.187[a]	.113	.114	.107	.258[a]
	.10	PE = .80	663	.086[a]	.043	.098[a]	.083	.060	.047	.041
		PE = .70	444	.103[a]	.052	.120[a]	.101	.072	.058	.051
		PE = .60	335	.121[a]	.061	.139[a]	.118	.084	.066	.059
Extensive	.40	PE = .80	117	.118	.131	.166[a]	.168[a]	.256[a]	.228[a]	.132
		PE = .70	81	.143	.161	.199[a]	.204[a]	.306[a]	.273[a]	.158
		PE = .60	63	.167	.187	.233[a]	.236[a]	.359[a]	.318[a]	.184
	.25	PE = .80	452	.093	.168[a]	.147[a]	.150[a]	.097	.121	.147[a]
		PE = .70	307	.113	.207[a]	.181[a]	.184[a]	.118	.150	.179[a]
		PE = .60	234	.131	.237[a]	.207[a]	.213[a]	.139	.173	.205[a]
	.10	PE = .80	663	.153[a]	.136[a]	.083	.106[a]	.063	.047	.207[a]
		PE = .70	444	.185[a]	.164[a]	.101	.129[a]	.076	.058	.250[a]
		PE = .60	335	.213[a]	.187[a]	.114	.150[a]	.087	.066	.287[a]

[a] indicates predictors with *VIF*> 5.0 (i.e., predictors involved in multicollinearity).

Conclusions

The primary goal of precision efficacy analysis for regression is to provide a means by which the researcher can assess the predictive power potential (i.e., generalizability) of a regression model relative to its performance in the derivation sample. As Cohen (1990) stated, “the investigator is not interested in making predictions for that sample---he or she *knows* the criterion values for those cases. The idea is to combine the predictors for maximal prediction for *future* samples” (p. 1306). The PEAR method has been shown through a line of research (Brooks, 1998; Brooks & Barcikowski, 1994, 1995, 1996, 1999) to be a viable method for this generalizability analysis.

The PEAR method appears to fill an important gap in the regression literature in that it recommends sample sizes for prediction based not only on the number of predictors in a study, but also on the size of the effect expected. Indeed, most sample size methods in other areas of statistics, including fixed model regression, consider effect size to be an essential part of the calculation. The PEAR method provides a means by which researchers can use a straightforward formula to choose samples by setting *a priori* effect

Table 4. Average Relative Efficiency of the Standardized Coefficients Across Predictors

ρ^2	k	Method Comparison	Orthogonal	Trivial	Moderate	Extensive
.40	3	RMSE(.80) / RMSE(.70)	80.8	80.2	80.6	80.5
		RMSE(.80) / RMSE(.60)	69.5	69.5	69.2	68.5
		RMSE(.70) / RMSE(.60)	86.1	86.6	85.9	85.1
	7	RMSE(.80) / RMSE(.70)	81.7	81.6	81.3	82.9
		RMSE(.80) / RMSE(.60)	70.5	70.6	71.2	71.1
		RMSE(.70) / RMSE(.60)	86.3	86.6	87.6	85.8
	11	RMSE(.80) / RMSE(.70)	81.4	81.8	81.6	80.5
		RMSE(.80) / RMSE(.60)	70.7	70.4	70.8	70.5
		RMSE(.70) / RMSE(.60)	86.8	86.1	86.7	87.6
	15	RMSE(.80) / RMSE(.70)	81.7	81.5	80.4	81.9
		RMSE(.80) / RMSE(.60)	70.7	70.6	69.8	70.7
		RMSE(.70) / RMSE(.60)	86.5	86.7	86.9	86.3
.25	3	RMSE(.80) / RMSE(.70)	81.3	81.9	82.0	81.7
		RMSE(.80) / RMSE(.60)	70.7	71.0	70.2	71.3
		RMSE(.70) / RMSE(.60)	87.0	86.7	85.7	87.4
	7	RMSE(.80) / RMSE(.70)	81.2	81.5	81.2	81.6
		RMSE(.80) / RMSE(.60)	70.0	71.0	69.9	70.7
		RMSE(.70) / RMSE(.60)	86.2	87.1	86.1	86.6
	11	RMSE(.80) / RMSE(.70)	81.4	81.6	81.6	81.4
		RMSE(.80) / RMSE(.60)	70.5	70.8	70.6	71.1
		RMSE(.70) / RMSE(.60)	86.6	86.8	86.5	87.3
	15	RMSE(.80) / RMSE(.70)	81.8	81.2	81.0	81.4
		RMSE(.80) / RMSE(.60)	71.2	70.6	70.2	70.5
		RMSE(.70) / RMSE(.60)	87.0	86.9	86.8	86.5
.10	3	RMSE(.80) / RMSE(.70)	81.0	80.1	80.6	81.6
		RMSE(.80) / RMSE(.60)	70.6	69.8	70.8	70.5
		RMSE(.70) / RMSE(.60)	87.2	87.2	87.9	86.4
	7	RMSE(.80) / RMSE(.70)	81.9	81.6	82.1	82.4
		RMSE(.80) / RMSE(.60)	70.4	70.5	70.6	72.0
		RMSE(.70) / RMSE(.60)	86.0	86.4	86.0	87.4
	11	RMSE(.80) / RMSE(.70)	81.1	81.9	81.8	81.7
		RMSE(.80) / RMSE(.60)	70.4	70.9	71.2	70.6
		RMSE(.70) / RMSE(.60)	86.8	86.6	87.1	86.5
	15	RMSE(.80) / RMSE(.70)	81.0	80.9	81.7	81.1
		RMSE(.80) / RMSE(.60)	70.2	70.4	70.7	70.7
		RMSE(.70) / RMSE(.60)	86.6	87.1	86.5	87.2

sizes, shrinkage tolerance, and precision efficacy levels. Brooks (1998) and Brooks and Barcikowski (1995) have shown that prediction models produced using appropriately large sample sizes will better estimate ρ_C^2 and will also provide necessary statistical power. The most important argument for the PEAR method is that a model based on a proper sample size, as suggested by the PEAR method, will provide more reliable regression weights. Therefore, these models will predict better for future subjects because, ultimately, the efficiency of a prediction model depends not only on correlation statistics such as R^2 and R_C^2, but also on the stability of the regression coefficients used to calculate predicted scores.

From the relative efficiency statistics, it would seem that the $PE = .80$ level used with the PEAR method usually would be most desirable. However, rather than rely on such a generalization, researchers must consider the needs of each project. For example, at lower population ρ^2 effect sizes, the statistics based on the methods become rather close in absolute value. For example, at $\rho^2 = .10$ with three predictors, R_C^2 was .088 and SE_{Bj} averaged 0.05 for the $PE = .80$ level, but $R_C^2 = .077$ with average $SE_{Bj} = 0.07$ for $PE = .60$. The $PE = .80$ level required 331 subjects to obtain its larger R_C^2, whereas the $PE = .60$ level only required 168 subjects to obtain a value that many researchers might find acceptable

(Brooks, 1998). Other researchers may determine; however, that the additional subjects recommended by the $PE = .80$ level are well worth the added precision efficacy. These dramatic differences in sample sizes must be balanced against the expected gain in precision and R_C^2, particularly at lower effect sizes. The sample size differences are not quite so striking at higher effect sizes, but still must be considered. For example, at $\rho^2 = .40$ and three predictors, the extra 28 subjects recommended by the $PE = .80$ ($N \geq 59$) level as compared to the $PE = .60$ level ($N \geq 31$) resulted in the more noticeable difference in average R_C^2 of .350 versus .294, respectively, and SE_{Bj} of 0.10 and 0.14, also respectively. Indeed, higher PE (e.g., .90) might be desirable under certain circumstances. Fortunately, thoughtful adjustments to the *a priori* precision efficacy or shrinkage tolerance enable researchers to use the PEAR method to make such choices.

Some may argue that effect sizes required by the PEAR method are too difficult to determine---"if one knew the answer to that question one would not need to do the study. . ." (Schafer, 1993, p. 387)---but blind adherence to conventional subject-to-predictor ratios certainly cannot be better research practice. Further, research in the evolution of the PEAR method has determined that when expected R^2 overestimates the actual ρ^2 value by too much (e.g., based on an effect size too large or due to an inappropriate conventional rule), no regression sample size method will recommend appropriate sample sizes for generalizability. For example, Brooks and Barcikowski (1995) found that when expected $\rho^2 = .25$, but actual $\rho^2 = .10$, precision efficacy rates were in the .47 to .50 range even for desired $PE = .80$. This reinforces the need for carefully chosen effect sizes in regression---as Schafer (1993) continued, ". . . but a value is needed anyway" (p. 387). When effect sizes are difficult to determine, pilot studies, meta-analyses, and careful interpretation of previous research play a critical role in the research process. Fortunately, because the PEAR method has performed well at a variety of effect sizes, numbers of predictors, shrinkage tolerance levels, and levels of multicollinearity, it seems to be well-suited to a variety of research situations.

Developing a model with good precision efficacy should be considered only a first step in the model validation process. The use of mathematical cross-validity formulas does not supersede the need for the validation of regression models in other samples. The cross-validity formulas suggest how well a model should perform in other samples, assuming that the sample from which it was derived was reasonably representative of the population; however, any given sample can deviate from what would be expected or representative. Further, no matter what the precision efficacy, a model that does not predict well in a derivation sample also probably will not predict well in any other samples. Finally, empirical cross-validation does not depend upon the assumptions required for use of the cross-validity equations, thus providing a possible substitute when the assumptions are not met (Darlington, 1990; Wherry, 1975).

Therefore, the safest way to determine that a model will generalize to future subjects is to test it with new data. Indeed, replication is basic to all science and is essential to confidence in both the stability and the generalizability of results. Additionally, Darlington (1990) and Montgomery and Peck (1992) have reminded us of the importance not only of model validation, but also of model adequacy, which requires residual analyses for violations of assumptions, searching for high leverage or overly influential observations, and other analyses that test the fit of the regression model to the available data. Darlington noted, however, that robustness to certain violations of assumptions continues to increase as sample size increases.

It is hoped that both the evidence presented and the relative simplicity of the PEAR method will encourage researchers to consider more carefully the issues of sample size, effect size, and generalizability for regression research. Because generalizability may be an even more important issue than statistical power in much regression research, an assessment technique such as precision efficacy analysis for regression appears beneficial to a more complete understanding of regression results.

References

Algina, J., & Keselman, H. J. (2000). Cross-validation sample sizes. *Applied Psychological Measurement, 24*(2), 173–179.

Algina, J., Keselman, H. J., & Penfield, R. J. (2007). Confidence intervals for an effect size measure in multiple linear regression. *Educational and Psychological Measurement, 67*, 207-218.

Algina, J., Keselman, H. J., & Penfield, R. J. (2008). Note on a confidence interval for the squared semipartial correlation coefficient. *Educational and Psychological Measurement, 68*, 734-741.

Algina, J., & Moulder, B. C. (2001). Sample sizes for confidence intervals on the increase in the squared multiple correlation coefficient. *Educational and Psychological Measurement, 61*, 633-649.

Algina, J., Moulder, B. C., & Moser, B. K. (2002). Sample size requirements for accurate estimation of squared semi-partial correlation coefficients. *Multivariate Behavioral Research, 37*(1), 37-57.

Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research*, *35*(1), 119-137.

Algina, J., & Olejnik, S. (2003). Sample size tables for correlation analysis with applications in partial correlation and multiple regression analysis. *Multivariate Behavioral Research, 38*(3), 309-323.

Asher, W. (1993). The role of statistics in research. *Journal of Experimental Education*, *61*, 388-393.

Bratley, P., Fox, B. L., & Schrage, L. E. (1987). A guide to simulation (2nd ed.). New York: Springer-Verlag.

Brooks, G. P. (1998, October). *Precision Efficacy Analysis for Regression*. Paper presented at the meeting of the Mid-Western Educational Research Association, Chicago, IL.

Brooks, G. P., & Barcikowski, R. S. (1994, April). *A new sample size formula for regression*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 412247).

Brooks, G. P., & Barcikowski, R. S. (1995, October). *Precision power method for selecting sample sizes*. Paper presented at the meeting of the Mid-Western Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 412246).

Brooks, G. P., & Barcikowski, R. S. (1996). Precision power and its application to the selection of regression sample sizes. *Mid-Western Educational Researcher, 9*(4), 10-17.

Brooks, G. P., & Barcikowski, R. S. (1999, April). *The Precision Efficacy Analysis for Regression sample size method*. Paper presented at the meeting of the American Educational Research Association, Montreal, Quebec, Canada. (ERIC Document Reproduction Service No. ED449177).

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

Carter, D. S. (1979). Comparison of different shrinkage formulas in estimating population multiple correlation coefficients. *Educational and Psychological Measurement, 39*, 261-266.

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

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

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

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

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.

Darlington, R. B. (1996). Estimating the true accuracy of regression predictions. *Mid-Western Educational Researcher, 9*(4), 29-31.

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.

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.

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

Huberty, C. J. (1994). A note on interpreting an R^2 value. *Journal of Educational and Behavioral Statistics, 19*, 351-356.

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

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8*, 305-321.

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

Knapp, T. R., & Campbell-Heider, N. (1989). Numbers of observations and variables in multivariate analyses. *Western Journal of Nursing Research, 11*, 634-641.

Knofczynski, G. T., & Mundfrom, D. (2008). Sample sizes when using multiple linear regression for prediction. *Educational and Psychological Measurement, 68*, 431-442.

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

Kromrey, J. D., & Hines, C. V. (1995). Use of empirical estimates of shrinkage in multiple regression: A caution. *Educational and Psychological Measurement, 55*, 901-925.

L'Ecuyer, P. (1988). Efficient and portable combined random number generators. *Communications of the ACM, 31*, 742-749, 774.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University.

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.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5*, 434-458.

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.

Mooney, C. Z. (1997). *Monte Carlo simulation* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-116). Thousand Oaks, CA: Sage.

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

Morris, J. D. (1981). Updating the criterion for regression predictor variable selection. *Educational and Psychological Measurement, 41*, 777-780.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology. Volume Two: Research methods* (2nd ed., pp. 80-203). Reading, MA: Addison-Wesley.

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

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

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.

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.

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.

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

Schafer, W. D. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education*, *61*, 383-387.

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.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size parameters. *Journal of Experimental Education, 61*, 334-349.

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

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

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., Sr. (1975). Underprediction from overfitting: 45 years of shrinkage. *Personnel Psychology, 28,* 1-18.

Send correspondence to:

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