Educational and Psychological Measurement
Volume 68 Number 3
June 2008 431-442
© 2008 Sage Publications
10.1177/0013164407310131
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Sample Sizes When Using Multiple Linear Regression for Prediction

Gregory T. Knofczynski
Armstrong Atlantic State University
Daniel Mundfrom
University of Northern Colorado

When using multiple regression for prediction purposes, the issue of minimum required sample size often needs to be addressed. Using a Monte Carlo simulation, models with varying numbers of independent variables were examined and minimum sample sizes were determined for multiple scenarios at each number of independent variables. The scenarios arrive from varying the levels of correlations between the criterion variable and predictor variables as well as among predictor variables. Two minimum sample sizes were determined for each scenario, a good and an excellent prediction level. The relationship between the squared multiple correlation coefficients and minimum necessary sample sizes were examined. A definite relationship, similar to a negative exponential relationship, was found between the squared multiple correlation coefficient and the minimum sample size. As the squared multiple correlation coefficient decreased, the sample size increased at an increasing rate. This study provides guidelines for sample size needed for accurate predictions.

Keywords: *subject predictor ratio; Monte Carlo; simulation; sample size; squared multiple correlation coefficient; multiple linear regression*

As the popularity of multiple linear regression (MLR) has increased, the question of how large a sample is required to produce reliable results has become increasingly more important to address. “As with any statistical analysis that is computed using sample data, the size of the sample (n) in large part determines the ‘value’ of the statistical results of a multiple regression analysis” (Gross, 1973, p. 17). When the researcher has an accurate estimate of the overall model effect size, R^2, this research will provide some guidelines as to the minimum sample size needed for accurate predictions.

Authors’ Note: Please address correspondence to Gregory T. Knofczynski, Department of Mathematics, Armstrong Atlantic State University, University Hall 297, Savannah, GA 31419; e-mail: knofczgr@mail.armstrong.edu.

Although there are numerous uses for MLR, generally they can be sorted into one of two groups: prediction and explanation (Gross, 1973; Kerlinger & Pedhazur, 1973; Pedhazur, 1997, Pedhazur & Schmelkin, 1991). The minimum size of the necessary sample depends on which of these two applications the researcher will utilize.

> Sample size tables and procedures used to determine sample size for hypotheses tests should not be used for estimation because providing evidence that a parameter is not equal to some specific value is a fundamentally different task than accurately estimating the parameter. (Algina & Olejnik, 2000, p. 119)

Maxwell (2000) states that "sample size will almost certainly have to be much larger for obtaining a useful prediction equation than for testing the statistical significance of the multiple correlation coefficient" (p. 435). Also, authors and researchers tend to agree that the sample sizes needed to reject a null hypothesis will usually not be adequate for prediction purposes (Brooks & Barcikowski, 1995, 1996; Casciok, Valenzi, & Silbey, 1978; Darlington, 1990; Gross, 1973; Pedhazur, 1997; Tabachnik & Fidell, 2001).

The answer to the sample size question appears to depend in part on the objectives of the researcher, the research questions that are being addressed, and the type of model being utilized. Although there are several research articles and textbooks giving recommendations for minimum sample sizes for multiple regression, few agree on how large is large enough and not many address the prediction side of MLR. This study addressed random models used exclusively for prediction purposes.

Algina and Keselman (2000), Park and Dudycha (1974), Brooks and Barcikowski (1995, 1996), and Gross (1973) investigated the sample size issue for minimizing the difference between the squared cross-validity correlation coefficient and the squared multiple correlation coefficient. Algina and Keselman (2000) and Park and Dudycha (1974) approached the sample size issue for predictive studies by limiting the difference between the squared cross-validity correlation coefficient, ρ_c^2, and the squared multiple correlation coefficient, ρ^2. The recommended sample sizes from both pairs of authors ensured that the difference between the squared cross-validity coefficient and the multiple correlation coefficient was no more than the difference selected by the researcher, with a specified confidence level.

Although both pairs of authors determined minimum sample sizes for similar situations, the method used by Algina and Keselman (2000) was a Monte Carlo simulation whereas Park and Dudycha (1974) utilized a theoretical approach. These two methods resulted in similar sample size recommendations.

Brooks and Barcikowski (1995) created a method for calculating minimum sample sizes called the "precision efficacy method." This method estimated the sample sizes needed to provide consistent specified precision efficacy rates when the primary purpose of the regression equation was prediction.

If researchers are to use the minimum sample sizes recommended by Algina and Kesselman (2000), Park and Dudycha (1974), or Brooks and Barcikowski (1995), they will need to have an estimate of the squared multiple correlation coefficient, ρ^2, along with other specific information such as the number of predictor variables in the model and the maximum desired difference between ρ_c^2 and ρ^2. Through the results of their research, Brooks and Barcikowski (1995, 1996) determined that if a researcher does not estimate the squared multiple correlation coefficient closely, then the recommended sample size of any method will probably be inaccurate.

Although the aforementioned researchers utilized theory and simulations to devise sample size recommendations for minimizing shrinkage of R^2, other authors simply state rules of thumb, some of which may be inconsistent with others. To provide minimal shrinkage of R^2, Pedhazur and Schmelkin (1991) state that a substantial subject to predictor ratio is 30 to 1 whereas Miller and Kunce (1973) suggest that a ratio of 10 to 1 is sufficient. One reason for so many different sample size recommendations is the numerous applications of MLR. However, even when the statistical analysis for which the sample size was recommended is the same, some researchers and authors may disagree on the minimum required sample sizes because they tend to judge their models on different criteria or to give no reasoning for their recommendations.

Method

Unlike the previously mentioned studies, this study is not concerned with finding an accurate value of the squared multiple correlation coefficient or minimizing the shrinkage of the squared multiple correlation coefficient. Instead, this research attends to the task of finding sample regression models that predict similarly to population regression models. More precisely, what sample size is needed to ensure, with a desired amount of accuracy, that the sample regression equation will perform similarly to the population regression equation? These minimum sample sizes were determined by conducting a series of Monte Carlo simulations. This study determines minimum sample sizes for a wide range of population correlation structures.

The similarity between the sample predicted values and the population predicted values was determined by examining the correlation coefficients between the predicted values obtained when using the population regression coefficients and the corresponding predicted values obtained when using the sample regression coefficients from numerous replications. If a large enough percentage of the replications for a given sample size had Pearson correlation coefficients at or above a specified level, denoted τ, then the given sample size was considered sufficient.

The population correlation structures examined in this research cover a wide range of cases. Cases involving one dependent variable and two, three, four, five, seven, and nine independent variables were utilized. Because of the methodology of this study, no sample size recommendations were possible when only one independent variable was present.

For the cases involving two independent variables, all correlation coefficients were allowed to be both positive and negative. For cases involving more than two independent variables, all correlation coefficients were limited to nonnegative values in order to minimize the number of possible correlation matrices to be examined.

The population correlation coefficients between the dependent variable and independent variables were examined at all possible combinations of high, medium, and low correlation coefficients. The population correlation coefficients among the independent variables were examined at medium, medium-low, and low levels. By omitting any high correlation coefficients among the independent variables, severe multicollinearity was avoided. High correlation coefficients were defined as .7, .8, and .9. Medium correlation coefficients were defined as .4, .5, and .6. Low correlation coefficients were defined as 0, .1, .2, and .3.

The adequacy of a specified sample size was determined by examining if the two different tolerance levels used in this study were met. The first tolerance level required that at least 95% of the correlation coefficients meet or exceed $\tau = .92$. This criterion level was called the good prediction level. The second tolerance level required that at least 95% of the correlation coefficients meet or exceed $\tau = .98$. This criterion level was called the excellent prediction level.

The number of replications in this Monte Carlo simulation refers to the number of samples, each having the desired sample size, that were created from a given population correlation matrix for the purpose of the analysis. The number of replications used in this study for each scenario was 2,000.

Computer programs were written in the SAS PROC Interactive Matrix Language. The correlation matrices used in this study contained both the correlation coefficients between independent variables and the dependent variable and the correlation coefficients among the independent variables. Matrices of a specified size were created, and correlation coefficients were placed into the matrix until an appropriate correlation matrix was constructed.

A separate program was written for the case with only two independent variables, but both programs consist of five parts. The first part of both programs created and tested for appropriateness the population correlation matrices for a specified number of independent variables. It is in this first part that the two programs are different. For the case with two independent variables, population correlation matrices were created by allowing the correlation coefficients between any independent variable and the dependent variable to increment systematically through a range from −.9 to .9 at increments equal to 0.1. The correlation coefficients among the dependent variables were systematically increased at increments of 0.1 from −.6 to .6. For the cases with more than two independent variables, the population correlation matrices were randomly generated by computer with the specified low, medium, and high levels of correlation coefficients.

Regardless of the number of independent variables, it was necessary to test each correlation matrix to determine if it was a proper and useable correlation matrix. For a correlation matrix to be proper and useable, it must be positive definite and have a squared multiple correlation coefficient in the interval (0, 1]. If a generated matrix was not a proper correlation matrix, it was determined to be unusable and was omitted.

For the cases with two independent variables, if a matrix was deemed unusable, the process of incrementing by 0.1 continued and the next matrix was tested. For the cases with more than two independent variables, if a matrix was deemed unusable, another matrix with the same levels of correlation coefficients and number of predictor variables was randomly generated and tested. This process was repeated until 10 usable matrices, which were used as population matrices, were found for each combination of level of correlation coefficients between criterion variable and predictor variables, level of correlation coefficients among predictor variables, and number of predictor variables.

In the second part of the program, the population regression coefficients were calculated from the population correlation matrix and stored for later use. Because the created data have a multivariate normal distribution with means equal to zero and standard deviations equal to one, the parameter β_0, the *y*-intercept, was equal to zero for all correlation structures.

Third, the program used random seeds to generate matrices of standardized multivariate normal data. Each matrix was representative of 2,000 random samples of data from the given population correlation matrix, each consisting of the specified sample size. This was accomplished by using part of the algorithm presented by Johnson (1987). This process, which uses the lower Cholesky root for generating multivariate normal data, has been recommended by many researchers (Bratley, Fox, & Schrage, 1987; Karian & Dudewicz, 1991; Mooney, 1997; Ripley, 1987). This large matrix was then divided into 2,000 smaller $nx(k+1)$ matrices representing the 2,000 samples of data, where n is the sample size and k is the number of independent variables. The first row of each smaller matrix represents the data of the dependent variable, and the other k rows correspond to the data for the k independent variables.

The fourth part of the program calculated the sample regression coefficients, the sample predicted values, and the population predicted values. The program then calculated the correlation coefficients between these two sets of predicted values.

The fifth and final part of the program determined if a large enough proportion of the correlation coefficients were at or above the set criterion, $\tau = .92$ or $\tau = .98$. To claim that the sample size was adequately large enough, 95% of the correlation coefficients must meet or exceed these levels. Simply counting the number of correlation coefficients that were greater than or equal to $\tau = .92$ and counting the number of correlation coefficients that were greater than or equal to $\tau = .98$ for the sample size currently under investigation accomplished this.

For each scenario examined, a specified starting sample size was determined and then the sample size was continually increased until these tolerance levels were met. If at least 95% were at or above $\tau = .92$, then the sample size was recorded as the necessary minimal sample size for the good prediction level with the present population correlation structure. If at least 95% were at or above $\tau = .98$, then the sample size was recorded as the necessary minimal sample size for the excellent prediction level with the present correlation structure. Precautionary steps were taken to avoid reaching the desired 95% level for $\tau = .92$ or $\tau = .98$ inadvertently.

As the sample size increased incrementally, the size of the increments also increased. The increments and the sample sizes up to which the increments were used, with the latter in parentheses, are 1 (30), 5 (100), 10 (200), 20 (500), 50 (1,000), and 100 (3,000). That is, 1 (30), 5 (100) indicates that sample sizes were incremented up by 1 up to a sample size of 30, then the sample sizes were incremented by 5 up to 100. The maximum sample size allowed was 3,000. Any minimum sample size recommendations not obtained by this point were left unanswered.

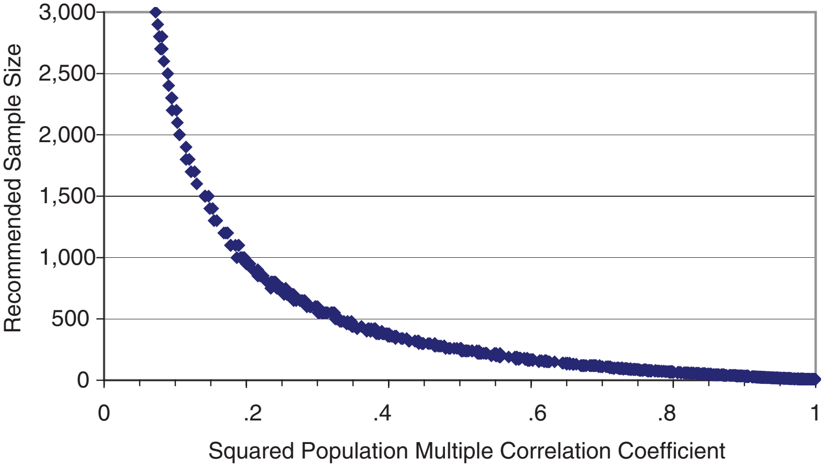
Results

The results of this analysis are based on more than 23,000,000 computer-generated samples, and many more samples were examined when the sample sizes were inadequate to reach the tolerance levels. The minimum necessary sample sizes, *n*, obtained from this study appear to have a relationship with the squared population multiple correlation coefficients, ρ^2. Because it is not practical to list all 12,668 sample size recommendations from this study, the relationship between the minimum sample sizes and the squared population multiple correlation coefficient is partially demonstrated graphically in Figure 1. Figure 1 displays approximately 1/12th of the total recommended sample sizes as it represents the relationship for five predictor variables at the excellent prediction level, $\tau = .98$.

A total of 12 such figures could be displayed resulting from the six levels of number of predictor variables—two, three, four, five, seven, and nine predictor variables—at each of the two criterion levels, $\tau = .92$ and $\tau = .98$. The 11 figures not shown had similar shapes to Figure 1. For models with larger numbers of predictor variables, the sample size recommendations increase more quickly and sooner than for models with fewer predictor variables as the squared multiple correlation coefficient decreased from 1 to 0. The sample size recommendations are also larger for the excellent prediction level than for the good prediction level, in some cases dramatically larger.

This relationship between the minimum recommended sample sizes and the number of predictor variables is illustrated in Table 1, which lists the minimum recommended sample sizes for the varying numbers of predictor variables at selected values of ρ^2. This information at the good prediction level and at the excellent prediction level is presented graphically in Figures 2 and 3, respectively. Figures 2 and 3 reveal how the minimum sample sizes are affected when increasing the number of

Figure 1
Minimum Sample Sizes for Five Predictor Variables at the Excellent Prediction Level



predictor variables while holding the value of ρ^2 constant. There is no recommended sample size for the excellent prediction level when $\rho^2 = .1$ with nine predictor variables because the maximum sample size of 3,000 was inadequate to reach the $\tau = .98$ tolerance level.

Table 2 enhances the usability of these sample size recommendations by listing the minimum recommended sample size to predictor ratio for varying numbers of predictor variables at selected values of ρ^2 at the good and excellent prediction levels.

Discussion

When using multiple regression for prediction purposes, the minimum recommended sample size and the sample size to predictor ratio definitely appear to have a relationship with the squared multiple correlation coefficient, ρ^2. Figure 1 demonstrates the relationship between minimum sample size and the squared multiple correlation coefficient. As the squared multiple correlation coefficient decreases, the sample size increases. The sample size increases slowly as the squared multiple correlation coefficient, ρ^2, departs from one, and then increases more quickly as ρ^2 approaches zero.

Table 1
Sample Size Recommendations at Selected Levels of Squared Population Multiple Correlation Coefficients for Varying Numbers of Predictor Variables

	Number of Predictor Variables					
ρ^2	2	3	4	5	7	9
			Good prediction level			
.10	240	380	440	550	700	900
.15	160	220	280	340	440	550
.20	110	170	200	260	320	400
.25	85	120	150	180	240	300
.30	65	95	130	150	190	240
.40	45	65	80	95	120	150
.50	35	45	55	65	85	100
.70	15	21	25	35	40	50
.90	7	9	10	11	14	16
			Excellent prediction level			
.10	950	1,500	1,800	2,200	2,800	—
.15	600	850	1,200	1,400	1,800	2,200
.20	420	650	800	950	1,300	1,500
.25	320	460	600	750	950	1,200
.30	260	360	480	600	800	1,000
.40	160	260	300	380	480	600
.50	110	130	220	230	320	400
.70	50	70	95	110	140	170
.90	15	21	29	35	40	50

Table 1 and Figures 2 and 3 show the relationships between the minimum sample size and the number of predictor variables and also the relationship between the minimum sample size and the squared multiple correlation coefficient. For a set squared multiple correlation coefficient, as the number of predictor variables increases, the sample size increases, but the required additional increase in sample size does not change drastically as the number of predictor variables increases. This relationship is evident by the nearly linear relationship between sample size and number of predictor variables for each specified level of squared correlation coefficient in Figures 2 and 3. Additionally, as the value of ρ^2 decreases and as the number of predictor variables increases, the recommended sample size increases at an increasing rate. For example, at the good prediction level, with two predictor variables and ρ^2 decreasing from .7 to .2, the difference in recommended sample sizes is 95 $(110 - 15)$. With seven predictor variables, this same decrease in ρ^2 requires an increase in sample size of 280 $(320 - 40)$. This relationship is also evident in

Figure 2
Recommended Sample Sizes for Different Numbers of Predictor Variables for Selected Values of Squared Population Multiple Correlation Coefficient at the Good Prediction Level

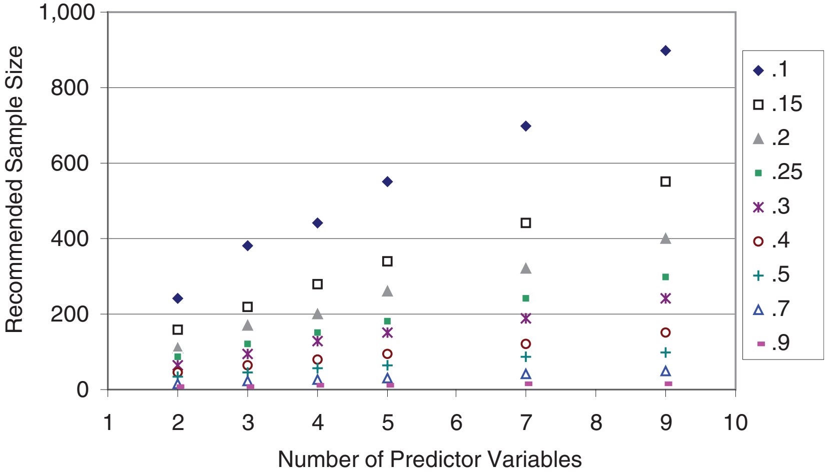


Figure 3
Recommended Sample Sizes for Different Numbers of Predictor Variables for Selected Values of Squared Population Multiple Correlation Coefficient at the Excellent Prediction Level

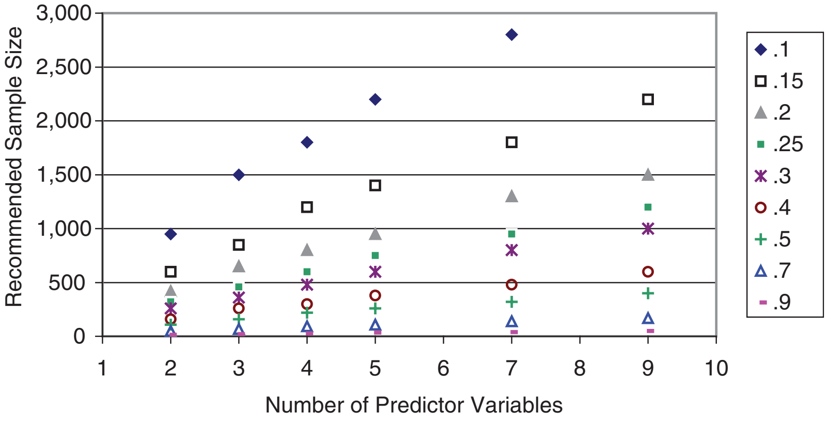


Table 2
Sample Size to Predictor Ratio Recommendations at Selected Levels of Squared Population Multiple Correlation Coefficients for Varying Numbers of Predictor Variables

	Number of Predictor Variables					
ρ^2	2	3	4	5	7	9
			Good prediction level			
.10	120	127	110	110	100	100
.15	80	73	70	68	63	61
.20	55	57	50	52	46	44
.25	43	40	38	36	34	33
.30	33	32	33	30	27	27
.40	23	22	20	19	17	17
.50	18	15	14	13	12	11
.70	8	7	6	7	6	6
.90	4	3	3	2	2	2
			Excellent prediction level			
.10	475	500	450	440	400	—
.15	300	283	300	280	257	244
.20	210	217	200	190	186	167
.25	160	153	150	150	136	133
.30	130	120	120	120	114	111
.40	80	87	75	76	69	67
.5	55	43	55	46	46	44
.70	25	23	24	22	20	19
.90	8	7	7	7	6	6

Figures 2 and 3, where the linear trend between sample size and number of predictor variables is steeper for smaller values of the squared multiple correlation coefficient.

Table 2 reveals the relationships among the sample size to predictor ratio, the number of predictor variables, and the squared multiple correlation coefficient. As the squared multiple correlation coefficient decreases, the recommended sample size to predictor ratio increases at an increasing rate for any set number of predictors. Also, as the number of predictor variables increases, the sample size to predictor ratio generally decreases.

If researchers have sufficiently reviewed previous research, are knowledgeable in their field, and have carefully planned their current study, they will know the number of predictor variables to be included in their regression model and have a good estimate of the squared multiple correlation coefficient. If both of these are known, the results of this research will give an estimate of the minimum sample size or the minimum sample size to predictor ratio at either the good prediction level or the excellent prediction level. Researchers may find it hard to get an accurate sample size recommendation from the figures because the *y*-axes include large ranges of sample sizes, namely, 3,000. Therefore, researchers are advised to use Tables 1 and 2 to assist in determining appropriate sample sizes or sample size to predictor ratios.

Researchers need to estimate their squared multiple correlation coefficients as accurately as possible. Underestimating ρ^2 will result in recommending a larger sample size than necessary. Although collection of unnecessary information is not desirable, the results of a study using a sample size that is slightly larger than necessary are not likely to differ from the results obtained when utilizing a smaller yet appropriate sample size. Overestimating ρ^2, on the other hand, will recommend sample sizes that are too small, possibly resulting in inaccurate or false results.

When a researcher suspects a small value for ρ^2, extra precaution should be used in estimating ρ^2 versus times when the researcher expects larger values of ρ^2. When dealing with large estimates of the squared multiple correlation coefficient, slight overestimation or slight underestimation of the squared multiple correlation coefficient will not have a large effect on the overall recommended sample size or the sample size to predictor ratio. However, when dealing with small estimates of the squared multiple correlation coefficient, it is more important that a researcher estimate the squared multiple correlation coefficient as accurately as possible. When ρ^2 is small, slight variations in the value of ρ^2 can cause major disparity regarding sample size recommendations. Slight overestimates result in sample size recommendations that are a great deal too small, and underestimation results in sample size recommendations that are unnecessarily much too large. For example, by means of Table 2, using the excellent prediction level with four predictor variables and a population squared multiple correlation of .2, the recommended sample size to predictor ratio is 200. If the estimated squared multiple correlation is underestimated at .15, the recommended sample size to predictor ratio is 50% larger at 300, whereas if the estimated squared multiple correlation coefficient is overestimated at .25, an inappropriately small sample size to predictor ratio of 150 will be recommended.

When utilizing MLR for prediction purposes, any author or researcher who does not take some aspect of the relationship between variables into consideration when making a sample size recommendation will seldom determine an appropriate sample size needed for the study. Also, the number of predictor variables is an important factor in determining the minimum required sample size. Authors and researchers who do not use the number of predictor variables as a determining factor when selecting appropriate sample sizes will probably end up with sample sizes that are too small or too large.

We recommend using the sample sizes presented in this article as a guideline when using multiple regression for predictive purposes. The results of this study are not recommended when using multiple regression for purposes other than prediction. Different applications of multiple regression usually require different minimum sample sizes (Brooks & Barcikowski, 1995, 1996; Casciok et al., 1978; Darlington, 1990; Gross, 1973; Pedhazur, 1997; Tabachnik & Fidell, 2001).

References

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

Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research*, *35*, 119-136.

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

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

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

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

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

Gross, A. L. (1973). How large should a sample size be in a regression analysis? *Proceedings of the Annual Convention of the American Psychological Association*, 17-18.

Johnson, M. E. (1987). *Multivariate statistical simulation*. New York: John Wiley.

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

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

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.

Mooney, C. Z. (1997). *Monte Carlo simulation*. London: Sage.

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. (1997). *Multiple regression in behavioral research* (3rd ed.). Fort Worth, TX: Harcourt Brace.

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

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

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