Sample Sizes for Predictive Regression Models and Their Relationship to Correlation Coefficients

Gregory T. Knofczynski †

Abstract

This article provides recommended minimum sample sizes for multiple linear regression models. Specifically, for regression models used for predictive purposes where the researcher has an accurate estimate of the largest correlation coefficient between the dependent variable and a predictor variable, regardless of the correlation structure of the rest of the variables included in the regression model.

Introduction

With advances in statistical software, it is becoming more common for researchers to conduct more complicated statistical analyses. One such analysis is multiple linear regression. According to Gross [7], when utilizing multiple linear regression, it is important to know that the accuracy of the results may be greatly affected by the sample size employed in the analysis. Unfortunately, researchers may not realize the importance of having an adequate sample size, may underestimate the required sample size, or may have a difficult time determining how large of a sample is necessary to achieve a specific level of precision in their study.

Multiple linear regression is a commonly used method of analysis and has several different applications, but most applications can be categorized into one of two categories, prediction and explanation [7] [14] [15]. When utilizing multiple linear regression, it is well established that the minimum sample size necessary for reliable results varies greatly for predictive purposes compared to explanation purposes [2] [4] [5] [6] [8] [14] [16]. Maxwell [11] 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). Other factors affecting necessary sample sizes are the number of predictor variables in the model and the level of desired accuracy [2] [10]. Additionally, both the strength of the linear association between the predictor variables and the dependent variable, and the strength of the association among the predictor variables themselves will be factors affecting the necessary minimum sample size [10].

To utilize current sample size recommendations for multiple linear regression models used for prediction purposes, a researcher is required to estimate at least one attribute of their regression model. Current research recommending minimum sample sizes for regression models used for prediction purposes relies on the researcher knowing or estimating attributes about their

model. If researchers are to use the minimum sample sizes recommended by Algina and Kesselman [1], Park and Dudycha [13], Brooks and Barcikowski [4], or Knofczynski and Mundfrom [10], they will need to estimate the squared multiple correlation coefficient, ρ^2, which can be difficult. The actual calculation of ρ^2 requires a researcher to know the population correlation structure of all variables included in the model. However, if the researcher can accurately estimate just the largest correlation coefficient between the dependent variable and a predictor variable, they can estimate the minimum sample size for prediction purposes using the results of this research.

Methods

This research attends to the task of determining how large of a sample is needed to ensure reliable predictions from a sample regression model, when only the largest correlation coefficient between the dependent variable and a predictor variable is known. A series of Monte Carlo simulations were used to determine the minimum sample sizes needed to find sample regression coefficients that predict similarly to the population regression coefficients. This study determined these minimum sample sizes for a wide range of population correlation structures when the largest correlation coefficient between the dependent variable and a predictor variable, max-ρ_{xy}, was known.

SAS PROC IML (Interactive Matrix Language) was used to conduct this study. Population correlation matrices were created and tested for appropriateness. In order to be considered a population correlation matrix, the matrix had to be positive definite and have a squared multiple correlation coefficient in the interval (0, 1].

The correspondence of the predicted values from the sample regression model and the predicted values from the population regression model was measured by calculating Pearson's correlation coefficients for the two sets of predicted values for numerous replications. If a high percent of the replications had correlation coefficients at or above a specified level, denoted τ, the sample size was considered adequate. Two different tolerance levels were considered in determining if a specified sample size was adequate or not. One tolerance level, called the good prediction level, was considered met if at least 95% of the correlation coefficients met or exceed $\tau = 0.92$. The second tolerance level, called the excellent prediction level, was considered met if at least 95% of the correlation coefficients exceeded $\tau = 0.98$. If either of the prediction levels were not met, the sample size was incrementally increased and new population correlation matrices were created.

All correlation coefficients used in this study were classified into one of three categories, high, medium or low. The high categories consisted of values of 0.7, 0.8 and 0.9. The medium category consisted of values of 0.4, 0.5, and 0.6. The low category consisted of values of 0, 0.1, 0.2, and 0.3. The values of

max-ρ_{xy} examined in this study ranged from 0.1 to 0.9 at increments of 0.1. The population correlations coefficients among the predictor variables were allowed to range in the medium and low categories. Higher correlations among the predictor variables were omitted in this study to prevent high levels of multicollinearity from existing. For models with only two predictor variables, both positive and negative correlation coefficients were utilized, but in order to make this study manageable, only positive correlation coefficients were utilized for all models with three or more predictor variables.

For each specified value of the max-ρ_{xy} several correlation structures were examined. A single population correlation structure for a given value of the max-ρ_{xy} would contain three structure qualities: (1) the number of predictor variables, (2) the level of correlation coefficients between the dependent variable and the remaining predictor variables with the restriction that these values were not allowed to exceed the value of the max-ρ_{xy}, and (3) the level of correlation coefficients among the predictor variables. The possible number of predictor variables were two, three, four, five, seven and nine. The correlation coefficient between the dependent variable and the remaining predictor variables were selected from one or any combination of the high, medium and low categories, with the limiting factor of the value of max-ρ_{xy}. The correlation coefficients among the predictor variables were selected from medium, low, or a combination of medium and low categories. The correlation coefficients among the predictor variables were not limited by the value of the max-ρ_{xy}. For each and every possible combination of the three correlation matrix structure qualities, a usable population correlation matrix was created. Then using the population correlation matrix 2,000 samples of the specified size were generated using part of the algorithm presented by Johnson [8], which uses the lower Cholesky root for generating multivariate normal data. This algorithm has been recommended by many researchers including Bratley, Fox and Schrage [3], Karian and Dudewicz [9], and Mooney [12]. The samples of data were used to create 2,000 sample regression equations. These sample regression equations were tested to determine if either the good prediction level or excellent prediction levels were met. If not, the sample size was incrementally increased and new usable population correlation matrices were created. Precautionary steps were taken to avoid reaching either the good or excellent prediction levels prematurely.

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 are given here with the latter in parentheses. 1 (30), 5 (100), 10 (200), 20 (500), 50 (1,000), 100 (3,000). (i.e. 1 (30) 5 (100) indicates that sample sizes were incremented up by ones up to a sample size of thirty, 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 minimum sample size recommendations recorded in this study were the most conservative (largest) sample sizes for each value of the max-ρ_{xy} examined. Therefore, regardless of the associations among the dependent variable and the other predictor variables, or the association among the predictor variables, the minimum sample sizes recommended in this study should provide a sample size large enough for either prediction level when the value of the max-ρ_{xy} and the number of predictor variables are known. The minimum sample sizes required to reach each of the two tolerance levels were not determined for all values of max-ρ_{xy} investigated in this study. For the good prediction level, the scenarios for which results were not obtained comprised of the cases when the max-ρ_{xy} was only 0.1 and four or more predictor variables were included in the model. For the excellent prediction level, regardless of the number of predictor variables, no results were found when the max-ρ_{xy} was 0.1. Additionally, at the excellent prediction level, when the max-ρ_{xy} was 0.2, no results were found for models with three or more predictor variables. For these cases the minimum necessary sample size exceeded the maximum sample size used in this study, viz. 3,000.

The results that were obtained indicate a relationship between the minimum sample sizes and the value of the max-ρ_{xy}. Table 1 displays the minimum sample sizes for the good prediction level for specified values of max-ρ_{xy} for regression models used for prediction purposes. Table 2 displays the minimum sample sizes for the excellent prediction level.

The relationship between the minimum sample sizes and the max-ρ_{xy} for the varying number of predictors in the model is shown graphically in Figure 1 and Figure 2. Figure 1 shows the relationship at the good prediction level. Figure 2 displays the relationship at the excellent prediction level.

Figures 1 and 2 both show that if the value of the max-ρ_{xy} is held constant, the necessary sample size increases as the number of predictor variables increases. This association was expected as it agrees with most other researchers' and authors' sample size recommendations. Though, as the value of the max-ρ_{xy} increases, the change in minimum sample size recommendations was not as great. For example, when the value of the max-ρ_{xy} is 0.3, at the good prediction level, the change in sample size recommendations from two predictor variables to nine predictor variables is 370 (650 – 280 = 370) compared to an increase of only 31 (50 – 19 = 31) when the value of the max-ρ_{xy} is set at 0.8.

Figures 1 and 2 also both show that for a specified number of predictor variables, as the value of the max-ρ_{xy} increases, the sample size decreases. This association also agrees with other researchers like Algina and Kesselman [1], Brooks and Barcikowski [4], Knofczynski and Mundfrom [10], and Park and Dudycha [13] who state that larger squared multiple correlation coefficients

correspond with smaller minimum sample size recommendations. This is not surprising as larger correlation coefficients between dependent variables and predictor variables correspond with larger squared multiple correlation coefficients.

When comparing Figure 1 to Figure 2 it appears that, for a specified value of the max-ρ_{xy}, the increases in minimum sample size recommendations is less dramatic at the good prediction level compared to the excellent prediction level when the number of predictor variables increases. For example, when the value of max-ρ_{xy} is set at 0.5, and the number of predictor variables increases from two to nine, the increase in the minimum recommended sample size at the good prediction level is 135 (220 – 85) while for the excellent prediction level the increase is 630 (950 – 320).

When the results of this study are compared to the research that gives sample size recommendations for regression models used for explanation purposes, it is clear that the sample sizes needed for prediction are much greater. This agrees with similar statements made by Brooks & Barcikowski [5], Darlington [6], Gross [7], Pedhazur [14] and Tabachnik & Fidell [16]. In addition to knowing the purpose of a regression model, Figures 1 and 2 make it clear, that knowing only the number of predictor variables in the model that is to be used for predictive purposes is inadequate information to determine a minimum sample size. More information is needed. If the additional information is limited to only knowing an accurate estimate of the largest correlation coefficient between the dependent variable and a predictor variable, then the results of this research can be used to find a recommended sample size for either the good or excellent prediction levels. Since Figures 1 and 2 are based on a wide sample size scale, from 0 to 3,000, it would be hard to obtain accurate sample size recommendations from these figures. Instead, it is recommended that Table 1 and Table 2 be used to find approximate recommended minimum sample sizes for either the good or excellent prediction levels respectively.

Future Studies

Additional studies on this subject could expand on the results of this study by including negative correlation coefficients both among the predictor variables and between the dependent variable and the predictor variables when there are at least three predictor variables. Also, a study could investigate the effect on the recommended minimum sample sizes when the second largest correlation coefficient between the dependent variable and a second predictor variable is known. Non-normal distributions could also be investigated.

† *G. Knofczynski, Ph.D., Armstrong State University, Savannah, GA, USA*

Table 1. Minimum recommended sample sizes for the good prediction level for values of the largest correlation between the dependent variable and a predictor variable

Good Prediction Level						
	Number of Predictor Variables					
max-ρ_{xy}	2	3	4	5	7	9
0.1	2600	2700				
0.2	650	950	1100	1200	1300	1300
0.3	280	380	500	550	600	650
0.4	140	220	260	300	340	340
0.5	85	120	160	180	220	220
0.6	55	75	95	110	120	120
0.7	35	45	55	70	80	85
0.8	19	26	35	40	40	50
0.9	10	13	16	17	19	21

Table 2. Minimum recommended sample sizes for the excellent prediction level for values of the largest correlation between the dependent variable a predictor variable

Excellent Prediction Level						
	Number of Predictor Variables					
max-ρ_{xy}	2	3	4	5	7	9
0.2	2400					
0.3	1100	1600	1900	2300	2400	2400
0.4	600	900	950	1200	1200	1700
0.5	320	460	600	750	800	950
0.6	190	300	360	420	460	480
0.7	120	170	220	240	300	340
0.8	60	95	120	140	160	170
0.9	29	45	55	60	65	80

Figure 1. Minimum sample size recommendations when only the largest correlation coefficient between the criterion variable and any predictor variable is known for the good prediction level.

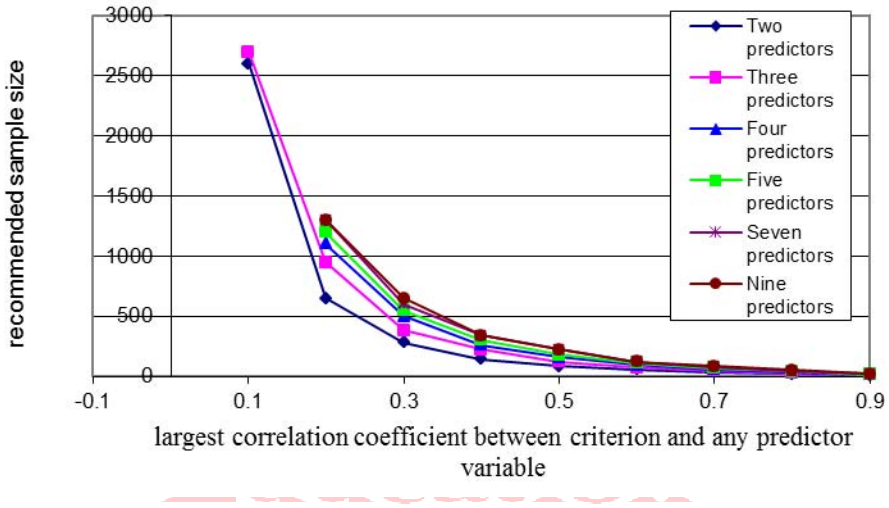
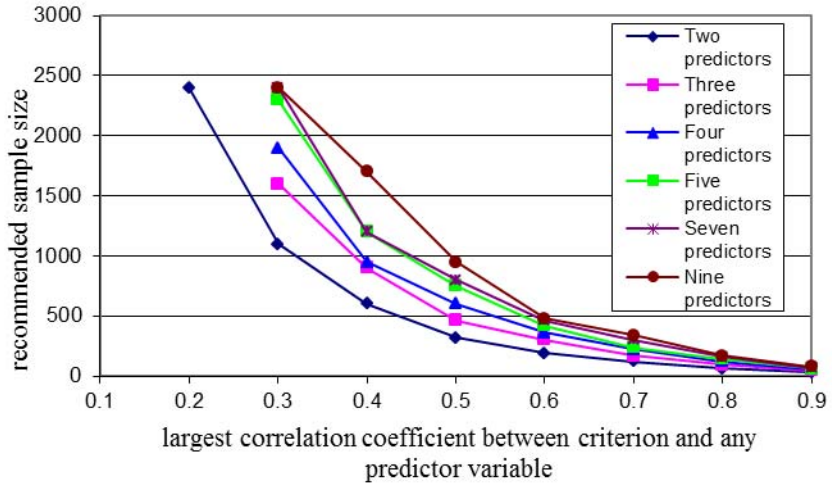


Figure 2. Minimum sample size recommendations when only the largest correlation coefficient between the criterion variable and any predictor variable is known for the excellent prediction level.



References

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

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

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

[4] 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, IL.

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

[6] Darlington, R. B. (1990). *Regression and Linear models*. New York: McGraw Hill.

[7] 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.

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

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

[10] Knofczynski, G. T., & Mundfrom, D. (2008). Sample Sizes When Using Multiple Linear Regression for Prediction. *Educational and Psychological Measurement*, 68, 431-442.

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

[12] Mooney, C. Z. (1997). *Monte Carlo simulation: Monte Carlo simulation*. London: Sage Publications Inc.

[13] 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.

[14] Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Fort Worth, TX: Harcourt, Brace & Company.

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

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