5-1999

Estimating R2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods

Ping Yin
Utah State University

Recommended Citation

Yin, Ping, "Estimating R2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods" (1999). *All Graduate Theses and Dissertations*. 6147.
https://digitalcommons.usu.edu/etd/6147

ESTIMATING R^2 SHRINKAGE IN MULTIPLE REGRESSION:

A COMPARISON OF DIFFERENT ANALYTICAL METHODS

by

Ping Yin

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Psychology

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

1999

ABSTRACT

Estimating R^2 Shrinkage in Multiple Regression:

A Comparison of Different Analytical Methods

by

Ping Yin, Master of Science

Utah State University, 1999

Major Professor: Xitao Fan, Ph.D.
Department: Psychology

This study investigated the effectiveness of various analytical methods used for estimating R^2 shrinkage in multiple regression analysis. Two categories of analytical formulae were identified: estimators of the population squared multiple correlation coefficient (ρ^2), and estimators of the population cross-validity coefficient (ρ_c^2). To avoid possible confounding factors that might be associated with a real data set such as data nonnormality, lack of precise population parameters, different degrees of multicollinearity among the predictor variables, and so forth, the Monte Carlo method was used to simulate multivariate normal sample data, with prespecified population parameters such as the squared multiple correlation coefficient (ρ^2), number of predictors, different sample sizes, known degree of multicollinearity, and controlled data normality conditions. Five hundred replicates were simulated within each cell of the sampling conditions. Various analytical formulae were applied to the simulated data in each sampling condition, and the "adjusted" coefficients were obtained and then compared to their corresponding population parameters (ρ^2 and ρ_c^2).

Analysis of the results indicates that the currently most widely used (in both SAS and SPSS) "Wherry" formula is probably not the most effective analytical formula in estimating ρ^2. Instead, the Pratt formula appeared to outperform other analytical formulae across most of these sampling conditions. Among the analytical formulae designed to estimate ρ_c^2, the Browne formula appeared to be the most effective and stable in minimizing statistical bias across different sampling conditions. The study also concludes that it is the *n/p* (sample size/number of predictor variables) ratio that affects the performances of these analytical formulae the most; different degrees of multicollinearity among predictor variables do not have dramatic influence on the performances of these analytical formulae. Further replications on both real and simulated data are still needed to investigate the effectiveness of these analytical formulae.

(136 pages)

ACKNOWLEDGMENTS

First I would like to thank all my committee members for their expertise, insights, and the help I needed to complete this thesis. I would especially like to thank Dr. Xitao Fan, my major professor. He has provided me with the knowledge and understanding of the quantitative research field in education and psychology, and helped me with his exceptional intelligence, experience, and patience. Without your encouragement, support, and prodding, I could not have completed this thesis. Thank you for guiding me through this whole process. You are the best instructor, teacher, and mentor I have ever met.

I would like to also thank Dr. Gary Mauk, my committee member and always my best friend. Thank you for your kindness and support. I would like to give special thanks also to Dr. Randy Jones, my committee member and special friend. Thank you for your understanding, kindness, support, and your great sense of humor.

Special thanks to Matt Shalala for his talents, warmth, and kindness. Also thanks for Randy and Yusnita for their love and support.

Finally and most importantly, I would like to thank my family for their support, especially my parents for their love and understanding.

Ping Yin

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND PROBLEM STATEMENT

To answer many research questions in the social and behavioral sciences, it is often useful to examine the relationship between a dependent (or criterion) variable and a set of independent (or predictor) variables at the same time. Statistically, with multiple regression, a dependent variable can be predicted from a set of independent variables. To do so, a linear combination of the independent variables is maximally correlated with the dependent variable. *Ordinary least squares* (OLS) is a method widely used to minimize the sum of squared errors of prediction, which is equivalent to maximizing the correlation between the observed and the predicted dependent variable. The maximized Pearson correlation coefficient between the dependent variable and the set of independent variables is called the multiple R (Stevens, 1996, p. 72).

In the process of optimizing the weighting of the independent variables for a sample, sampling chance or random error tends to be capitalized. This optimizing process from which the multiple regression equation is derived causes the sample multiple correlation coefficient *(R)* to be systemically higher than the corresponding population parameter ρ. When the equation is applied to an independent sample other than the one from which the equation is obtained (i.e., cross-validation), the predictive power drops off. This phenomenon is what the term "statistical bias" in multiple regression refers to (Glass & Hopkins, 1996; Stevens, 1996). The smaller the sample size and the more independent or predictor variables used, the greater the shrinkage in sample multiple R when applied to a new sample (Cohen & Cohen, 1983; Stevens, 1996).

To determine the generalizability or the predictive power of a sample regression equation, different approaches of model validation have been developed (Cohen & Cohen, 1983; Darlington, 1968; Herzberg, 1969). There are two major categories: empirical methods and analytical methods. The empirical methods usually involve the estimation of average predictive power of a sample regression equation on other samples (cross-validation). Typical empirical methods for this purpose are data splitting, multicross-validation, jackknife, and bootstrap methods (Ayabe, 1985; Cummings, 1982; Kromrey & Hines, 1995; Krus & Fuller, 1982). Analytical methods include several analytical correction formulae for adjusting the statistical bias and yield corrected R^2. Some major correction formulae designed for this purpose are the Smith formula (presented by Ezekiel, 1929), the Ezekiel formula (Ezekiel, 1929), the Darlington/Stein formula (Darlington, 1968; Stein, 1960), the Browne formula (1975), the Olkin/Pratt formula (1958), the Nicholson/Lord formula (Lord, 1950; Nicholson, 1960), and the Wherry formula (1931).

However, there is little consensus in the literature on which method is most appropriate under what circumstances for estimating "statistical bias" in multiple regression. Some studies suggest that the Browne formula may be superior to other estimates for estimating shrinkage in multiple regression (Kromrey & Hines, 1996), while other studies suggest that both the Nicholson/Lord formula and the Olkin/Pratt formula work equally well (Huberty & Mourad, 1980). Also, there are studies suggesting multicross validation "to be the method of choice" (Ayabe, 1985, p. 450). Few studies had specifically investigated these inconsistencies.

Several factors contribute to the inconsistent findings. In the literature, considerable confusion exists over various analytical formulae. For example, in several studies the Ezekiel formula was mistakenly cited as the Wherry formula (Ayabe, 1985; Kennedy, 1988; Krus & Fuller, 1982; Schmitt, 1982; Stevens, 1996). In other studies, authors failed to distinguish between ρ^2 (the population squared multiple correlation coefficient, or the population coefficient of determination) and ρ_c^2 (the population squared multiple correlation coefficient obtained with a specific sample equation, or the coefficient of cross-validation). Such distinction between the two parameters is important because an analytical method for shrinkage estimate of one of the two parameters might not be an accurate estimate for the other.

Beyond those discrepancies, there are some problematic methodological issues for estimating statistical bias in multiple regression. One problematic issue is that different studies have employed different types of shrinkage estimates: one study only used analytical formulae (Uhl & Eisenberg, 1970), while other studies used both analytical and empirical methods (Claudy, 1978; Huberty & Mourad, 1980; Kromrey & Hines, 1996). Different conclusions might have been drawn due to the limited shrinkage estimates that an individual study utilized. Another problematic issue concerns using real data to evaluate the performance of different estimating methods. One major limitation with real data set is that there might be a combination of confounding factors that the researcher could not control, such as different forms of data nonnormality, lack of precise population parameters, different degrees of multicollinearity among the predictor variables, and so forth. Therefore, a better assessment of the performance of different analytical methods would be to use simulated data with prespecified parameters, known degree of multicollinearity, and controlled data normality conditions.

Because of time constraints and project manageability, the present study focused on comparing the effectiveness of different *analytical* formulae in estimating shrinkage in multiple regression analysis. More specifically, the objectives in the present study were:

1. To compare the accuracy and usefulness of various analytical formulae for estimating ρ^2 (the population squared multiple correlation coefficient).

2. To compare the accuracy and usefulness of various analytical formulae for estimating ρ_c^2 (the population squared coefficient of cross-validation).

3. To assess the effects of sample size (n), number of predictor variables (p), the n/p ratio, and the degree of multicollinearity among the predictors on the accuracy and variability of the performances of the analytical formulae in estimating R^2 shrinkage.

CHAPTER II

LITERATURE REVIEW

Multiple Regression

In multiple regression, the linear relationship between one dependent variable and a set of independent variables is being modeled. The general multiple regression model with p independent variables could be explained as:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \epsilon_i \qquad [1]$$

where p stands for the number of predictors, β_0 is the regression constant, $\beta_1, \ldots, \beta_p$ are population regression weights to be estimated, and ϵ_i is the error of prediction.

In the model above, the criterion of least squares is used to establish the regression line, in which the sample regression parameter estimates (b_0 and $b_1, \ldots, b_p$) are selected so that the sum of squared residuals (e_i), that is, the sample counterpart of the population error term ϵ_i , is as small as possible. Such a procedure minimizes the sum of squared errors of prediction, which is equivalent to maximizing the correlation between the observed dependent variable (Y_i) and the predicted value $\hat{Y}_i$ (Stevens, 1996). The multiple R is a measure of association between the dependent variable and a set of two or more independent variables. The coefficient of determination (R^2) measures the proportion of total variance in the predicted variable that is associated with the set of predictor variables in the regression model (Stevens, 1996).

Statistical Bias

There are two major reasons for researchers to apply the multiple regression procedure (Claudy, 1978): (a) to estimate the population multiple correlation coefficient from a sample and (b) to predict the same criterion variable for new samples from the same population other than the one from which the regression weights are derived. It has long been recognized by quantitative researchers that when a multiple correlation coefficient is derived from a given sample, its value tends to be "deceptively" large, and it is a "positively biased" estimate of the population multiple correlation coefficient (Cohen & Cohen, 1983; Larson, 1931). Furthermore, when such a multiple regression equation is applied to an independent sample other than the original one, it usually would not fit a new sample as well as it did for the sample from which it was derived (Cohen & Cohen, 1983; Larson, 1931; Stevens, 1996). If the regression equation from a sample could neither estimate the population parameter accurately nor predict well when applied to other samples, the purposes of multiple regression are not fulfilled. Corresponding to the two research purposes of multiple regression, there are also two types of "shrunken *R*"s discussed in the literature. These two types will be described in the following section.

Estimates of Population Multiple Correlation Coefficient (ρ)

One type of shrinkage occurs when estimating the population ρ^2 from a sample R^2. For this purpose of multiple regression, a linear model is utilized to model the relationship between a dependent variable Y and the optimal linear composite of p independent variables $X_1, X_2, \ldots, X_p$, (which could also be represented by the vector variate **X**) in the population as a whole. Matrix algebra gives a compact explanation of multiple regression model (Stevens, 1996):

$$Y = X\beta + \epsilon \qquad [2]$$

where Y is the vector of the criterion variable, **X** is the $n \times (p + 1)$ matrix, with one intercept and p independent variables, β is the vector of regression weights, and ϵ is the vector of errors.

OLS is the statistical principle widely used to model the linear relationship between the dependent variable and the set of independent variables. One of the basic assumptions of the multiple regression model is that the values of the independent variables are known constants and fixed by the researcher prior to the experiment. Only the dependent variable is free to vary from sample to sample. Residuals in the regression model are assumed to be *i.i.d.*: (a) identically distributed with mean of zero and equal variance, (b) independent to each other, and (c) normally distributed (Hamilton, 1991). This widely used regression model is also called the *fixed linear regression model* (Cohen & Cohen, 1983; Park & Dudycha, 1974).

However, in applied situations in social and behavioral sciences, those assumptions are rarely met completely: the values of independent variables are rarely fixed by the researchers, and they are also subject to random errors. Therefore, Park and Dudycha (1974) suggested a second regression model for applications in the behavioral sciences, which is called the *random model (*or *correction model)*. In this model, the independent variables are allowed to vary freely, and the joint distribution of both dependent and independent variables is multivariate normal. However, this random model is so complex that more research is needed before it can be accepted as the commonly used fixed linear regression model. Therefore, the fixed model is usually applied even if the assumptions are not met completely (Claudy, 1978). Such applications of the fixed regression model with assumptions violated would cause "over-fitting" because of the random error introduced from the less-than-perfect data. Also, the sample multiple correlation coefficient obtained this way would tend to overestimate the real population multiple correlation (Claudy, 1978; Cohen & Cohen, 1983; Cummings, 1982).

Estimates of Coefficient of Cross-Validation (ρ_c)

The second type of shrinkage occurs when we want to predict the criterion variable for new samples from the same population, but other than the one from which the regression weights are derived. The cross-validity of the population ρ_c is defined as the population multiple correlation coefficient obtained with a specific sample equation. When the regression weights derived from one sample are applied to a new sample from the same population, a multiple correlation coefficient is obtained, and it is called R_c. R_c is the validity estimate of the original sample regression equation in another sample, and it is an estimator of the population cross-validity coefficient ρ_c. The expected value of R_c [E (R_c)] over many samples would approach or equal ρ_c [E (R_c) ≈ ρ_c] (Claudy, 1978; Cummings, 1982; Herzberg, 1969).

Because the population regression equation in the population will usually function better than the sample regression equation in the population, the value of ρ would tend to be greater than ρ_c ($\rho_c < \rho$). Also, the sample multiple correlation coefficient is a positively biased estimator of the population multiple correlation coefficient ($\rho < R$). Thus, the relationship between values of the two population parameters (ρ and ρ_c) and two sample estimates R and R_c) could be summarized as (Claudy, 1978; Cummings, 1982; Herzberg, 1969):

$$E\,(R_c) \approx \rho_c < \rho < R$$

As it is generally known, the sample multiple correlation coefficient R is used as the estimator for both ρ_c and ρ, but it is actually larger than either ρ_c or ρ. R is a positively biased estimator of ρ, and an even more positively biased estimator of ρ_c (Cummings, 1982). Therefore, the estimator R must be "shrunken" or "corrected" to adjust for the positive bias for estimating either parameter in multiple regression analysis.

Estimating R^2 Shrinkage in Multiple Regression

Estimating R^2 shrinkage and correcting for the statistical bias in sample multiple regression have been suggested in many studies (Browne, 1975; Cohen & Cohen, 1983; Huberty & Mourad, 1980; Krus & Fuller, 1982; Larson, 1931; Stevens, 1996; Wherry, 1931). These methods could be classified into two categories: empirical methods and analytical methods (Kromrey & Hines, 1995).

A review of literature located 11 such studies involving applications of the empirical and/or the analytical methods in estimating R^2 shrinkage in multiple regression. The following two major study characteristics were identified for these studies:

1. Estimating methods: studies may have used empirical (cross-validation, double

cross-validation, multicross-validation, jackknife, bootstrap) and/or analytical (formula) methods; and

2. Validation methods: studies may differ in terms of method used, data set selection, sample size, number of predictor variables used, and population parameters.

Estimating Methods

Empirical Methods

Empirical methods for correcting statistical bias for sample multiple R in multiple regression include the following approaches: cross-validation, double cross-validation, multicross-validation, jackknife, and bootstrap. All these approaches share the logic of cross-validation; that is, to estimate the shrinkage by applying the regression equation derived from one sample to new data in the same population. For these approaches, usually the squared population cross-validity coefficient (ρ_c^2) is what is being estimated.

Cross-validation. In cross-validation, the regression weights generated in one sample (derivation or screening sample) are used to predict values for the same dependent variable in another sample (validation or calibration sample). A cross-validation multiple R_c could thus be computed in the validation sample by correlating the observed dependent variable (Y) with the predicted dependent variable ($\hat{Y}$) obtained. It is important to note this cross-validated R_c is not an estimator of the population multiple correlation coefficient ρ, but rather the "cross-validated" ρ_c that tends to be smaller than ρ (Huberty & Mourad, 1980; Kromrey & Hines, 1995).

Cross-validation requires two equivalent samples (derivation and validation samples) that come from the same population. However, in applied situations usually only one sample is available for the researchers. In order to apply the cross-validation approach, it had been suggested that the sample be split into two subsamples, and one subsample of the data would be held in reserve while deriving the regression equation from the other subsample. Cross-validity could be estimated by applying the regression weights to the reserved subsample and calculate the cross-validation multiple R_c (Cummings, 1982). Typically, the reserved data is one third or one half of the total sample (Cummings, 1982).

It has been noted that the major problem associated with such cross-validation method is that the splitting of the data into two parts requires that some of the data be withheld from the derivation of the regression equation. The regression weights are, therefore, based upon *part* of the available data. It is well-known that the stability of the regression weights would tend to decrease as the ratio of the sample size to the number of variables decreases. Thus, not including all the available data in deriving a multiple regression equation would probably lead to a significant loss of information, and, therefore, introduce more instability into the regression equation (Huberty & Mourad, 1980; Newman, McNeil, Garver, & Seymour, 1979). The application of cross-validation is restricted especially when the sample size is small.

Of the previous studies reviewed, four studies used the cross-validation procedure in estimating population cross-validity coefficient (Cummings, 1982; Kromrey & Hines, 1995, 1996; Newman et al., 1979). Cummings (1982) indicated that the use of the cross-validation procedure tended to underestimate the population cross-validity coefficient ρ_c^2.

Newman et al. (1979) concluded that cross-validation method "forces one to split the sample in half which tends to produce less stability than one would get using the entire sample", and the results from cross-validation "shows no advantage over analytical methods" (p. 11). It was also not recommended as a reliable estimate by Kromrey and Hines (1995, 1996).

Double cross-validation. Double cross-validation was first developed by Mosier (1951) to address the instability in simple cross-validation while deriving multiple regression equations from only part of the available sample. In Mosier's double cross-validation, the available sample is first split in half and the regression equations are calculated for both halves of the sample. The regression equation derived from one half of the sample is then applied to the other half, and the cross-validation multiple R_c is calculated. The same procedure is repeated for the other half. Thus, two subsample cross-validation multiple R_cs are then obtained. The double cross-validation coefficient could then be calculated by averaging the two cross-validation multiple R_cs. The formula can be stated as:

$$\hat{\rho}_c = (R_{c1} + R_{c2})/2 \qquad [3]$$

where R_{c1} and R_{c2} stand for the cross-validation multiple $R_c s$ for both halves of the sample, and $\hat{\rho}_c$ stands for the estimation of ρ_c .

Claudy (1978) also developed a new double cross-validation procedure based on the Mosier's double cross-validation method. In Claudy's double cross-validation, first the regression equation is calculated within one half sample and then apply to the other half sample, and vice versa. The difference is that both the two cross-validity indices (R_{c1}

and R_{c2}) and the two sample multiple correlation coefficients (R_1 and R_2) are averaged to provide an estimate of the population ρ. It is important to note that Claudy's procedure intends to estimate the population multiple correlation coefficient ρ, not the cross-validity coefficient ρ_c. The formula can be written as:

$$\hat{\rho} = (R_1 + R_2 + R_{c1} + R_{c2})/4 \qquad [4]$$

where R_1 and R_2 stand for the two subsample multiple correlation coefficients, $R_{c1}\hat{\rho}$and R_{c2} stand for the sample cross-validity coefficients for both halves of the sample, and stands for the estimation of ρ.

Claudy also developed another variation of double cross-validation to estimate ρ_c, which was called "double shrinkage estimate" (Claudy, 1978):

$$\hat{\rho}_c = \frac{R_1 + R_2 + R_{c1} + R_{c2}}{2} - R \qquad [5]$$

where R_1 and R_2 stand for the two subsample multiple correlation coefficients, R_{c1} and R_{c2} stand for the sample cross-validity coefficients for both halves of the sample, R is the sample multiple regression coefficient, and$\hat{\rho}_c$ stands for the estimation of ρ_c.

Of the previous studies reviewed, three studies used a double cross-validation procedure (Claudy, 1978; Cummings, 1982; Kennedy, 1988). Claudy's study showed that the Claudy's double cross-validation procedure yielded more accurate estimates than the analytical formulae for estimating the population multiple correlation ρ (Claudy, 1978). In Cummings' study, Mosier's double cross-validation was found to underestimate ρ_c^2, and the estimation also appeared to have excessive amount of variation (Cummings, 1982).

Also, Claudy's double cross-validation procedure showed no advantage over analytical methods in estimating ρ. Finally in Kennedy's study, no advantage was found for Mosier's double cross-validation over analytical methods (Kennedy, 1988).

Multicross-validation. Krus and Fuller (1982) first introduced multicross validation as an extension of Mosier's double cross-validation. The technique is based on repeated double cross-validations to select subsamples of the data randomly. Regression weights are then calculated in each subsample and used for predicting the criterion variable of the other subsample. Cross-validated multiple R_cs are then computed between the actual and the predicted values of the criterion variable in each subsamples.

The cross-validated multiple R_cs are then normalized through Fisher-Z transformation:

$$Z = \tanh^{-1} R_c \qquad [6]$$

After each iteration, the mean of the Fisher Z-transformed cross-validated multiple R_c and its corresponding standard error are computed. The procedure is repeated until a prespecified number of iterations is reached, or after the mean of the cross-validated R_cs appears to converge; that is, the difference between consecutive normalized cross-validated multiple R_cs is less than an arbitrarily selected constant used as the criterion for convergence.

At the termination of the iteration process, the resulting normalized cross-validated multiple R_c is transformed back to its original correlation scale as:

$$R_c = \tanh Z \qquad [7]$$

The mean of cross-validated R_cs at convergence or after the last iteration is defined as the multicross-validated R_c (Krus & Fuller, 1982)

The multicross-validation approach gives more analytical power to the researcher with small data set, although this technique usually requires a large amount of computing. Of the previous studies reviewed, four studies included the multicross validation procedure (Ayabe, 1985; Kromrey & Hines, 1995, 1996; Krus & Fuller, 1982). Krus and Fuller (1982) suggested that for random data sets, the multicross validation procedure gave a more accurate estimate of the population multiple correlation coefficient ρ than analytical formulae. They further suggested that an empirical rather than an analytical approach should be used when data sets are small. Ayabe (1985) confirmed the findings by Krus and Fuller, and suggested that the multicross validation method produced "comparable or superior estimates of the analytical formula methods" (Ayabe, 1985, p. 449). And the multicross validation method also performed better than the jackknife method. In the study by Kromrey and Hines (1995), the only condition that the performance of multicross validation was superior to both jackknife and bootstrap method is when the population squared multiple correlation coefficient ρ^2 is very small (0.04). Otherwise, both jackknife and bootstrap methods performed better than the multicross validation procedure. In their 1996 study, there was no obvious advantage found for multicross validation over the Browne formula, and it was found to be more difficult to compute a multicross validity coefficient than to use the Browne formula.

Jackknife procedure. The jackknife procedure, first introduced by Quenouille (1949), is a technique to reduce bias in estimation and to assess the stability or accuracy of an empirically estimated parameter. The jackknife procedure first estimates the cross-validity coefficient R_c by splitting the sample into two subsamples, where one of the subsamples usually contains only one individual observation. The regression equation is derived in the large sample which has *n-1* subjects, and the regression weights are applied to the sample with one observation to yield a predicted value. The procedure is repeated *n* times with the exclusion of one different observation for each time to obtain the regression weights and to calculate the predicted value for that observation. Thus each observation has a predicted value on the criterion variable based on the regression equation derived from the remaining *n-1* subjects. A correlation coefficient between the original criterion variable and the predicted values for the criterion variables is then calculated. The cross-validity coefficient can then be calculated by either averaging the *n* obtained coefficients, or by using the same Fisher-Z transformation in the multicross validation method (Kromrey & Hines, 1995). Another name for the jackknife technique was descriptively termed as the "leave-one-out" method (Huberty & Mourad, 1980).

One variation of the jackknife method is called predicted residual/error sum of squares (PRESS), that was discussed by Stevens (1996), to assess the external predictive power in multiple regression. However, no empirical study utilizing the PRESS method for estimating cross-validity was found in the literature. Like jackknife, the PRESS first predicts each subject's criterion score based on the regression equation generated from the other *n-1* observations (Stevens, 1996). Then the PRESS residuals are calculated using the following formula:

$$\hat{e}_{(-i)} = y_i - \hat{y}_{(-i)} \qquad [8]$$

where $\hat{y}_{(-i)}$ is the predicted value for subject *i*, when that subject is not used in the derivation of the regression equation.

The PRESS statistic can be calculated using the following formula:

$$R^2_{PRESS} = 1 - \frac{\Sigma \hat{e}^2_{(-i)}}{\Sigma (y_i - \bar{y})^2} \qquad [9]$$

This PRESS value is a R^2-like statistic that estimates the squared population cross-validity coefficient ρ_c^2 (Stevens, 1996).

The jackknife procedure can be applied to a variety of situations including small sample size, and this procedure is also highly dependent on intensive computation. Of the previous studies reviewed, four studies included the jackknife procedure in estimating shrinkage in multiple regression (Ayabe, 1985; Huberty & Mourad, 1982; Kromrey & Hines, 1995, 1996). The results in Ayabe's study showed that the jackknife method did not perform as well as multicross validation. The reason for this was jackknife's "inadequacy in handling outliers" (Ayabe, 1985, p. 449). In Huberty and Mourad's study, the "leave-one-out" method was used (Huberty & Mourad, 1982). This method was found to be equally accurate to the Nicholson/Lord formula and the Darlington formula in estimating ρ_c^2, but tended to overestimate shrinkage slightly. Such a method was also found to be very difficult to calculate in practice, and "leave-one-out" was suggested to be "tentatively dropped as an estimator of ρ_c^2" (Huberty & Mourad, 1982, p. 108). In the study by Kromrey and Hines (1995), the normalized or transformed jackknife was shown to provide the best estimate when the sample size was relatively large (> 100). In their 1996 study, the jackknife performed less well than analytical formulae, and the normalized jackknife tended to overestimate ρ_c^2.

Bootstrap method. The bootstrap method was developed by Efron (1979). This method is designed to assess the statistical accuracy or stability from an empirically derived estimation of a population parameter. In the bootstrap method, many random samples of sample size *n* are repeatedly drawn with replacement from the original sample (Fan & Wang, 1996). Because of sampling with replacement, a typical bootstrap sample could leave out some cases from the original data and include other cases more than once. According to Kromrey and Hines (1995, 1996), to implement the bootstrap method in estimating the cross-validity coefficient in multiple regression, for each random bootstrap sample, the regression equation is computed and then applied to the original sample to yield the predicted values for the criterion variable. A standard Pearson correlation coefficient is then computed between the original and predicted values of the criterion variable. The process is repeated for each bootstrap sample to generate a distribution of the coefficients obtained from all the bootstrap samples, and the mean of the distribution of all the bootstrap estimates is defined as the bootstrap multiple R_c (Kromrey & Hines, 1995, 1996).

Of the previous studies reviewed, only two studies included bootstrap in estimating shrinkage in multiple regression (Kromrey & Hines, 1995, 1996). In their 1995 study, the bootstrap method only yielded acceptable estimate when sample size was relatively large (> 100). In their 1996 study, the bootstrap performed less well than analytical formulae, and it also tended to overestimate ρ_c^2.

Analytical Methods

An alternative to the empirical approach is the analytical approach represented by various "shrinkage" formulae. All of these mathematical formulae are based on the entire sample so that they would provide more stable results compared to those methods that are only based on part of the sample (e.g., cross-validation). Different shrinkage formulae have been proposed to estimate either ρ^2 (the population squared multiple correlation coefficient) or $\rho_c{}^2$ (the population squared coefficient of cross-validation).

In the literature, there has been some confusion about both the origins and the purposes of these different formulae (Cummings, 1982; Huberty & Mourad, 1980; Kromrey & Hines, 1996; Newman et al., 1979). For example, the popular "Wherry formula" actually was not proposed by Wherry himself (Wherry, 1931). Also in some studies, the Ezekiel formula was mistakenly cited as the Wherry formula (Ayabe, 1985; Kennedy, 1988; Krus & Fuller, 1982; Schmitt, 1982; Stevens, 1996).

The present review of literature has identified 14 such shrinkage formulae. These formulae have been categorized into two groups: estimator of ρ^2 and $\rho_c{}^2$.

Estimator of ρ^2: (a) the Smith formula (Wherry, 1931); (b) the Wherry formula-1 (1931); (c) the Wherry formula-2 (1931); (d) the Olkin and Pratt formula (1958); (e) the Pratt formula (cited in Claudy, 1978); and (f) the Claudy-3 formula (1978).

Estimator of $\rho_c{}^2$ or ρ_c: (a) the Lord formula-1 (1950); (b) the Lord formula -2 (1950); (c) the Burket formula (1964); (d) the Darlington formula (1968); (e) the Browne formula (1975); (e) the Claudy formula-1 (1978); (f) the Claudy formula-2 (1978); (g) the Rozeboom formula-1 (1981); and (h) the Rozeboom formula-2 (1978).

These formulae are presented and reviewed based on the parameters they are estimating. In the following presentation of these analytical formulae, N is the sample size; R is the sample multiple correlation coefficient; p is the number of predictor variables; ρ is the population multiple correlation coefficient; ρ_c is the population cross-validity coefficient; and $\hat{R}$ is the "corrected" R obtained from the analytical formula.

Estimator of ρ^2. The Smith formula takes the form:

$$\hat{R}^2 = 1 - \frac{N}{N - P}(1 - R^2) \qquad [10]$$

The formula was originally developed by Smith, and presented by Ezekiel in 1928 (Wherry, 1931). Larson (1931) empirically tested the formula on real data. The regression equation derived from one group of subjects was used to predict the criterion scores of a second group. However, the results indicated that the Smith formula tended to result in greater shrinkage. Because the formula was originally proposed as an estimator of ρ^2, the Larson study was actually cross-validation that was estimating ρ_c^2 instead. This probably could explain why the Smith formula showed greater shrinkages in Larson's study.

Of the previous studies reviewed, only one study included this formula in estimating shrinkage in multiple regression. In Cummings' study (1982), no advantage was found for the Smith formula over other analytical methods in estimating shrinkage in multiple regression.

The Wherry formula-1 (1931) - estimator of ρ^2, can be stated as:

$$\hat{R}^2 = 1 - \frac{N - 1}{N - P - 1}(1 - R^2) \qquad [11]$$

The formula was actually proposed by Ezekiel as an estimator of ρ^2 (Ayabe, 1985; Cohen & Cohen, 1983; Cummings, 1982; Huberty & Mourad, 1980; Kromrey & Hines, 1996; Newman et al., 1979). However, in the literature, it has been cited widely with different names, mostly as the Wherry formula (Ayabe, 1985; Kennedy, 1988; Krus & Fuller, 1982; Schmitt, 1982; Stevens, 1996), secondly as the Ezekiel formula (Huberty & Mourad, 1980; Kromrey & Hines, 1996), the Wherry/McNemer formula (Newman et al., 1979), and, finally, the Cohen/Cohen formula (Kennedy, 1988). It was also cited in one study as estimator for cross-validation (Kennedy, 1988). One study mistakenly cited this formula as "the analytical formula used in the most popular statistical programs (SPSS, SAS, BMDP) to correct sample bias" (Kromrey & Hines, 1996, p. 242). However, this is not the analytical formula used in both SAS and SPSS.

This formula is the most frequently used analytical method in the studies reviewed. However, none of the studies recommended it as the most effective method in estimating ρ^2 in multiple regression. Kennedy (1988) found that the formula gave the most biased estimate in most situations. Cummings (1982) found it tended to overestimate ρ^2 but was less variable. Only Huberty and Mourad (1980) and Kromrey and Hines (1996) suggested that it gave a reasonable estimate of ρ^2.

The Wherry formula-2 (1931) - estimator of ρ^2, can be stated as:

$$\hat{R}^2 = 1 - \frac{N - 1}{N - P}(1 - R^2) \qquad [12]$$

This formula is actually currently being implemented by both SAS and SPSS for computing the adjusted R^2 in multiple regression procedures (SAS/STAT User's Guide, 1990, p. 1354). This formula was presented by Wherry (Wherry, 1931), but it was cited in one study as the McNemer formula (Newman et al., 1979). In the literature, it is usually confused with the Wherry formula-1 (formula [11]) above. Few studies have correctly cited it as the Wherry formula (Cummings, 1982; Huberty & Mourad, 1980; Kromrey & Hines, 1996; Uhl & Eisenberg, 1970). The formula was also developed as an estimator of ρ^2.

Of the previous studies reviewed, three studies included this formula in estimating ρ^2 in multiple regression. It was found to be less accurate than other analytical methods in two of the studies (Cummings, 1982; Uhl & Eisenberg, 1970). Newman et al. (1979), however, found it to be a relatively stable estimate for ρ^2.

The Olkin and Pratt Formula (1958) - estimator of ρ^2 is:

$$\hat{R}^2 = R^2 - \frac{P - 2}{N - p - 1}(1 - R^2) - \frac{2(N - 3)}{(N - P - 1)(N - p + 1)}(1 - R^2)^2 \quad [13\text{-}1]$$

or

$$\hat{R}^2 = 1 - \frac{(N - 3)(1 - R^2)}{(N - p - 1)} - \frac{2(N - 3)(1 - R^2)^2}{(N - p - 1)(N - p + 1)} \quad [13\text{-}2]$$

or

$$\hat{R}^2 = 1 - \frac{(N - 3)(1 - R^2)}{(N - P - 1)}\left\{1 + \frac{2(1 - R^2)}{N - p + 1}\right\} \quad [13\text{-}3]$$

Equation [13-1], [13-2], and [13-3] are basically the same equation in different forms, and they are all approximations of the Olkin and Pratt's (1958) unbiased estimate of the squared multiple correlation ρ^2. The original formula for the unbiased estimate developed by Olkin and Pratt (1958) is:

$$\hat{R}^2 = 1 - \frac{(N - 3)(1 - R^2)}{(N - p - 1)} F(1,1;\frac{N - p + 1}{2};1 - R^2) \qquad [13\text{-}4]$$

where F is the hypergeometric function:

$$F(\alpha,\beta;\gamma;x) = \sum_{k=0}^{\infty} \frac{\Gamma(\alpha + k)\Gamma(\beta + k)\Gamma(\gamma)x^k}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma + k)k!}$$

Formulae [13-1] to [13-3] have been cited as the Olkin and Pratt formula in several studies (Ayabe, 1985; Claudy, 1978; Huberty & Mourad, 1980; Krus & Fuller, 1982) and erroneously cited as Herzberg formula in one study (Cummings, 1982).

Of the previous studies reviewed, five studies used formula [13] in estimating ρ^2 in multiple regression (Ayabe, 1985; Claudy, 1978; Cummings, 1982; Huberty& Mourad, 1980; Krus & Fuller, 1982). In two of these studies, results from this formula were found to be less accurate than multicross validation (Ayabe, 1985; Krus & Fuller, 1982). In Huberty and Mourad's study (1980), the formula was found to be accurate in estimating ρ^2.

The Pratt formula (1964) - estimator of ρ^2, another approximation of the unbiased estimate has been used in two studies (Claudy, 1978; Cummings, 1982):

$$\hat{R}^2 = 1 - \frac{(N - 3)(1 - R^2)}{(N - P - 1)}\left\{1 + \frac{2(1 - R^2)}{N - p - 2.3}\right\} \qquad [14]$$

Of the previous studies reviewed, two studies included this formula in estimating ρ^2 in multiple regression (Claudy, 1978; Cummings, 1982). Both of these studies showed that this formula gave the most accurate estimate for ρ^2 in multiple regression.

The Claudy formula-3 was introduced in Claudy's study (Claudy, 1978).

$$\hat{R}^2 = 1 - \frac{(N - 4)(1 - R^2)}{(N - P - 1)}\left\{1 + \frac{2(1 - R^2)}{N - p + 1}\right\} \qquad [15]$$

This formula was very similar to the Pratt approximation of the Olkin and Pratt formula (formula [13-3]), except for some differences in the second term.

Of the previous studies reviewed, only one study used this formula in estimating ρ^2 in multiple regression. Claudy (1978) suggested that this formula gave a better estimation of the population multiple correlation coefficient than both the Pratt and the Herzberg approximations of the Olkin and Pratt formula for estimating ρ^2.

Estimator of ρ_c^2 or ρ_c. The Lord formula-1 (1950) can be represented as:

$$\hat{R}^2 = 1 - \frac{N + p + 1}{N - P - 1}(1 - R^2) \qquad [16]$$

This formula was developed to estimate the population cross-validity coefficient ρ_c^2 (Newman et al., 1979; Uhl & Eisenberg, 1970). It had been cited mostly as the Lord formula (Newman et al., 1979; Uhl & Eisenberg, 1970); however, in one study it was referred to as the Uhl and Eisenberg formula (Cummings, 1982).

From the previous studies reviewed, three studies included this formula in estimating ρ_c^2 in multiple regression (Cummings, 1982; Newman et al., 1979; Uhl & Eisenberg, 1970). All three studies found that it usually gave an accurate estimate of ρ_c^2.

The Lord formula -2 (1950) - estimator of ρ_c^2

$$\hat{R}^2 = 1 - \frac{(N + p + 1)(N - 1)}{(N - P - 1)N}(1 - R^2) \qquad [17]$$

was developed by both Lord and Nicholson independently, and it had been cited as either the Lord formula (Kennedy, 1988; Newman et al., 1979) or the Nicholson formula (Schmitt, 1982). It was also erroneously cited as the Herzberg formula in one study (Cummings, 1982). This formula was developed also as an estimator for the population cross-validity coefficient ρ_c^2 .

Of the previous studies reviewed, six studies employed this method in estimating ρ_c^2 in multiple regression (Claudy, 1978; Cummings, 1982; Huberty & Mourad, 1980; Kennedy, 1988; Newman et al., 1979; Schmitt, 1982). Schmitt (1982) found that it did not provide an accurate estimate when the squared population multiple correlation coefficient (ρ^2) is less than .6. Huberty and Mourad (1980) found that it was one of the most accurate estimates for ρ_c^2, but it tended to overestimate shrinkage. The other four studies showed that its performance was neither excellent nor poor (Claudy, 1978; Cummings, 1982; Kennedy, 1988; Newman et al.,1979).

The Burket formula (1964)- estimator of ρ_c follows:

$$\hat{R} = \frac{NR^2 - p}{R(N - p)} \qquad [18]$$

This formula was first presented by Burket (1964) as a direct estimate of the population validity coefficient rather than the squared population cross validity coefficient ρ_c^2. The formula was also called “weight validity.”

Of the previous studies reviewed, two studies employed this formula in estimating ρ_c in multiple regression. (Claudy, 1978; Cummings, 1982). No significant advantage was found for this formula than other analytical methods in estimating ρ_c in multiple regression.

The Darlington (1968) or Stein formula (1960) - estimator of ρ_c^2 is:

$$\hat{R}^2 = 1 - \left(\frac{N - 1}{N - P - 1}\right)\left(\frac{N - 2}{N - p - 2}\right)\left(\frac{N + 1}{N}\right)(1 - R^2) \qquad [19]$$

The formula was developed as an estimator of cross-validation coefficient ρ_c^2 and it has been referred to as either the Darlington formula or the Stein formula (Cummings, 1982; Huberty & Mourad, 1980; Kennedy, 1988; Kromrey & Hines, 1996; Newman et al., 1979; Schmitt, 1982; Stevens, 1996).

Six studies employed this formula in estimating ρ_c^2 in multiple regression (Cummings, 1982; Huberty & Mourad, 1980; Kromrey & Hines, 1995,1996; Newman et al., 1979; Schmitt, 1982). Newman et al. (1979) found it to be a "fairly decent estimate of the population ρ^2, but tends to underestimate the population parameter" (p. 10). Kennedy (1988) found that it yielded the best estimate of ρ_c^2. Huberty and Mourad (1980) also noticed that it tended to slightly overestimate shrinkage. Schmitt (1982) found that it failed to give accurate shrinkage estimates for low levels of multiple correlation ($R^2 < .6$). Kromrey and Hines (1996) did not find any advantage of this formula over other analytical methods.

The Browne formula (1975) can be stated as:

$$\hat{R}^2 = \frac{(N - p - 3)\rho^4 + \rho^2}{(N - 2p - 2)\rho^2 + p} \qquad [20]$$

where ρ^2 is the squared population multiple correlation coefficient. It was suggested that

ρ^2 to be estimated by either the Wherry formula-1(formula [11]), or the Olkin and Pratt formula (formula [13]; Schmitt, 1982).

Compared to the original Browne formula, only the first part of the original formula was used here (Browne, 1975). The original Browne formula is lengthy and complicated:

$$\mathscr{E}(w^2) = \frac{(N-p-3)\rho^4+\rho^2}{(N-2p-2)\rho^2+p} - \frac{2(N-p-2)(N-2p-6)(p-1)\rho^4(1-\rho^2)^2}{(N-p-4)\{(N-2p-2)\rho^2+\rho^3\}^3} + o(\{N-p\}^{-1}) \qquad [21]$$

It was noted by Cattin (1980) that the second part of the formula only yields negligible values compared to the first part, and Darlington (1968) also stated that the first part is more valuable when the sample is small, which is applicable in social and behavioral sciences.

The Browne formula was developed as an estimator for cross-validity coefficient ρ_c^2. It has been cited as the Browne formula with only the first part in two studies (Kennedy, 1988; Kromrey & Hines, 1996), as the Cattin formula in one study (Schmitt, 1982), and as the Browne formula as the original form (formula [21]) in the same study (Schmitt, 1982).

Of the previous studies reviewed, three studies employed this formula in estimating ρ_c^2 in multiple regression. Both Schmitt (1982) and Kromrey and Hines (1996) concluded that this formula was the most appropriate estimator of ρ_c^2 with the Wherry formula-1 as the estimator for ρ^2. Kromrey and Hines (1996) also noted that the performance of the Browne formula was excellent when sample size was relatively large (> 100). On the contrary, Kennedy (1988) did not find that the Browne formula yielded estimates as accurate as that of the Darlington formula. No advantage was found for the original Browne formula (formula [21) over the commonly used Browne formula (formula [20]) in estimating ρ_c^2 (Schmitt, 1982).

The Claudy formula -1 (1978) - estimators of ρ_c^2 is shown below. Claudy (1978) proposed three different formulae for estimating either the population ρ^2 or ρ_c^2. The Claudy formula-1 takes the form:

$$\hat{R}^2 = (2\rho - R)^2 \qquad [22]$$

The formula was first introduced by Claudy as an estimator of ρ_c^2 (Claudy, 1978). It was also suggested that ρ be estimated by the Wherry formula-1 (formula [11]) (Cummings, 1982).

Of the previous studies reviewed, only one study employed this formula in estimating ρ_c^2 in multiple regression. Cummings (1982) found that it was the most accurate and least variable estimate of ρ_c^2 with the Wherry formula-1 as the estimator of ρ. However, it had a slight tendency to overestimate ρ_c^2.

The Claudy formula -2 (1978) - estimators of ρ_c^2 is shown below. In the same study, Claudy proposed another formula for estimating either the population ρ_c^2.

$$\hat{R}^2 = 1 - \left(\frac{N-1}{N-P-1}\right)\left(\frac{N-2}{N-p-2}\right)\left(\frac{N-1}{N}\right)(1-R^2) \qquad [23]$$

In the original study, this formula was presented as "the Darlington formula" (Claudy, 1978). Compared to the original formula in Darlington's study and several other similar studies (equation [19]), the only difference between equation [23] and [19] is the minus |(-)| or plus |(+)| sign in the second part. It is very likely such difference is due to either misprint or miscitation.

Two studies used this Claudy formula-2 in estimating ρ_c^2 in multiple regression. Claudy (1978) concluded that this formula yielded the most accurate estimate of ρ_c. Kennedy (1988), however, did not find that it yielded an estimate as accurate as that of the Darlington formula.

In the literature, there are two forms of the Rozeboom formula which were developed as estimators of cross-validity coefficient ρ_c^2. The Rozeboom formula-1 (1981) takes the form:

$$\hat{R}^2 = 1 - \frac{N + p}{N - P}(1 - R^2) \qquad [24]$$

Of the previous studies reviewed, 2 studies used this formula in estimating ρ_c^2 in multiple regression. Kennedy (1988) found that it did not yield estimate as accurate as that of the Darlington formula. Huberty and Mourad (1980) concluded it gave an estimate as precise as the Darlington formula.

The Rozeboom formula-2 (1978) takes the form:

$$\hat{R}^2 = \rho^2 \left[1 + (\frac{p}{N-P-2})\left(\frac{1-\rho^2}{\rho^2}\right)\right]^{-1} \qquad [25]$$

Of the previous studies reviewed, only one study used the Rozeboom formula-2 in estimating ρ_c^2 in multiple regression (Schmitt, 1982). However, it was found to be less satisfactory than the Browne formula.

After reviewing those various analytical formulae for correcting the statistical bias, there are two possible reasons for the confusion in the literature about different analytical formulae. The first reason is the large number of correction formulae and the names associated with them. There are 14 formulae reviewed in the present study. For some of those formulae, more than one name was found to be associated with the same formula in the literature and more than one formula was associated with the same name. The second reason is that some of the formulae are developed as the estimate of ρ^2, and some of them are developed as the estimate of ρ_c^2. But the distinction, however, is sometimes not clearly made.

Validation Methods

Statistical Methods and Data Set Selection

Five of the studies reviewed utilized the Monte Carlo technique in the validation procedure (Claudy, 1978; Kennedy, 1979; Kromrey & Hines, 1995, 1996; Newman et al., 1979), and the remainder of the studies did not use the Monte Carlo method in the validation procedure. However, suggestions for future Monte Carlo simulation studies had been explicitly made in two such studies (Ayabe, 1985; Huberty & Mourad, 1980).

Three studies used simulated data for the estimating purpose (Claudy, 1978; Kennedy, 1988; Newman et al., 1979). Four studies utilized real data (Cummings, 1982; Huberty & Mourad, 1989; Kromrey & Hines, 1996; Uhl & Eisenberg, 1970). Two studies used both prestructured data (adapted from other studies) and random data (simulated) (Ayabe, 1985; Krus & Fuller, 1982). One study did not specify the origin of the data set (Schmitt, 1982).

Sample Size

Sample sizes range from 14 to 325 in the studies reviewed. In most of the studies, the number of the sample size was within 200. Sample sizes of 20, 40 or 60 or 80, 100, and 200 were the commonly selected sample sizes in most of the studies reviewed, and such a sample size was selected to be reasonably representative of sample sizes in current applied multiple regression research (Kromrey & Hines, 1995, 1996; Schmitt, 1982).

Of the studies reviewed, Kromrey and Hines concluded that the estimation of ρ_c^2 was very poor for any of the analytical methods utilized in their study when the sample size was smaller than 100 (Kromrey & Hines, 1996). Kennedy (1988) also concluded that sample size was a primary factor, rather than the number of predictor variables, that influenced R^2 shrinkage in multiple regression the most. On the contrary, Newman et al. (1979) did not find the association between large sample size and better estimate.

Population Squared Multiple Correlation Coefficient

The population squared multiple correlation coefficients vary from .02 to .9, which covers almost the entire possible range of the coefficient. In most of the studies, the population squared multiple correlation coefficients were quite small, mostly lower than .5 (Kromrey & Hines, 1995, 1996; Newman et al., 1979). Results from Kromrey and Hines' study (1996) showed that as the population ρ^2 increases, more estimating methods gave better estimates for the population parameters.

Number of Predictors

The number of predictors in multiple regression ranged from 2 to 25, and most of the studies included fewer than 10 predictors. In the field of psychological and educational research, 2 , 3, 4, and 5 were shown to be representative of the number of predictors for real data (Claudy, 1978). No specific conclusion was found in the previous studies on the implications of the number of the predictor variables on the performance of these estimating methods in multiple regression analysis.

Multicollinearity

Collinearity refers to the linear correlation between two independent variables. Multicollinearity, a more general term, refers to linear relationships between two or more independent variables. In the presence of strong multicollinearity, the regression weights from multiple regression are less useful in prediction because a strong relationship implies redundancy. Stevens (1996) summarized three major problems with multicollinearity for the researchers: (a) it limits the range of multiple correlation coefficient; (b) it confounds the importance of a given independent variable; and (c) it increases the variances of the regression coefficients.

Moderate to high multicollinearity among independent variables is not uncommon in social and behavioral sciences. However, only two of the studies reviewed investigated the performance of those analytical methods under the influence of multicollinearity. One study did indicate that the intercorrelation among the independent variables in the psychological and educational literature ranged from .01 to .65, but the effects of different degrees of multicollinearity on the performance of these analytical formulae were not clearly discussed (Claudy, 1978). In the other study, multicollinearity *r* was selected to range from .13 to .82, with approximate .15 as the interval (Newman et al., 1979). The author later concluded that multicollinearity had almost no detectable effect on the accuracy of the shrinkage estimates (Newman et al., 1979).

Summary of Literature Review

The study characteristics and conclusions for all the studies reviewed previously are summarized in Appendix A. This literature review has revealed little consensus regarding which method is the most appropriate under what specific conditions for estimating statistical bias in multiple regression. The inconsistencies in the studies' results are possibly due to: (a) inconsistent terminology of analytical formulae, (b) lack of distinction of the two population parameter ρ^2 and ρ_c^2 and their corresponding sample estimates, and (c) different characteristics of the real data sets utilized in individual study. Because of time constraints and project manageability, the present study only focused on the *analytical methods* for estimating the population squared correlation coefficient ρ^2 and the population cross-validity coefficient ρ_c^2.

CHAPTER III

METHODOLOGY AND PROCEDURE

Analytical Formulae

To compare the effectiveness of different analytical formulae in estimating both the population squared correlation coefficient ρ^2 and the population cross-validity coefficient ρ_c^2 in multiple regression, the analytical formulae reviewed in previous chapters are categorized into two groups: estimators of ρ^2 and ρ_c^2. To avoid confusion associated with different names, the respective formula numbers in the present study are also provided in Table 1.

Validation Method

The Monte Carlo simulation is a method widely used to evaluate substantive hypotheses and statistical estimators by: (a) developing a computer algorithm to simulate a statistical population with specified parameters, (b) drawing random samples from the population, and (c) evaluating the behaviors of the sample estimates for the population parameters (Johnson, 1987).

One of the major features in the Monte Carlo procedure is the control of relevant population factors that includes the choice of population distributions and their parameters, sample sizes, and other related variables. This feature usually could not be easily obtained in real data sets because of the potential confounding effects from

Table 1

Analytical Formulae Analyzed in the Present Study

Estimator	Analytical formulae	Formula number
ρ^2	1. the Smith formula	[10]
	2. the Wherry formula-1	[11]
	3. the Wherry formula-2	[12]
	4. the Olkin and Pratt formula	[13-1], [13-2], [13-3]
	5. the Pratt formula	[14]
	6. the Claudy-3 formula	[15]
ρ_c^2 or ρ_c	1. the Lord formula-1	[16]
	2. the Lord formula -2	[17]
	3. the Burket formula	[18]
	4. the Darlington formula	[19]
	5. the Browne formula-1[a]	[20]
	5. the Browne formula-2 [b]	[20]
	6. the Claudy formula-1	[22]
	7. the Claudy formula-2	[23]
	8. the Rozeboom formula-1	[24]
	9. the Rozeboom formula-2	[25]

[a] The Browne formula with ρ^2 being estimated by the Wherry formula-1 (formula [11]).

[b] The Browne formula with ρ^2 being estimated by the Olkin and Pratt formula (formula [15]).

multiple extraneous factors (Johnson, 1987). One major limitation with real data is that there might be a combination of confounding factors the researcher cannot control, such as different forms of data nonnormality, lack of precise population parameters, different degrees of multicollinearity among the predictor variables, and so forth. Such confounding of multiple extraneous factors may make it very difficult, or nearly impossible, for the researcher to assess the performance of different analytical methods under different data conditions.

For this reason, it is often easier to assess the effectiveness of different analytical methods if simulated data are used that have prespecified population parameters, known degrees of multicollinearity, and controlled data normality conditions. Therefore, the Monte Carlo method is employed in this study to simulate statistical populations with prespecified parameters. Potential factors considered in the study include different population ρ^2, sample sizes, number of predictor variables, and different conditions of multicollinearity among the predictor variables.

Simulation Design of Population Parameters

Squared Population Correlation Coefficient ρ^2

From the literature reviewed for this study, the possible range of ρ^2 has been from .1 to .9 in previous studies. The squared population correlation coefficient ρ^2, or what is also called the coefficient of determination, can also be interpreted as a measure of strength or the magnitude of the relationships between the dependent and predictor variables. That is, it can also be considered as a measure of effect size. According to Cohen's specification of small, medium, and large effect sizes in the form of squared correlation coefficient based on typical findings in social and behavioral research studies, .1 is usually considered to be small, .25 (.2 to .3) is considered to be medium, and .5 is considered to be relatively large (Cohen, 1988). In the present study, the squared population multiple correlation coefficients were selected to be .2, .5, and .8 to represent what is considered to be the magnitude of between small and medium, relatively large, and very large in the areas of social and behavioral research.

Number of Independent Variables (p)

From the literature reviewed for this study, most studies included fewer than 10 predictors in the regression analyses. Also with respect to representativeness of the real data and the project manageability, in the present study, the numbers of independent variables were selected to be 2, 4, and 8.

Sample Size (n)

From the previous studies reviewed, the size of most samples selected was within 200. It was also noted that, in social and behavioral sciences, many applied studies that utilized multiple regression analysis used relatively small samples (Claudy, 1978). Based on the previous studies, and to represent the research characteristics as reported in the psychological and educational literature, samples with sample sizes of 20, 40, 60, 100, and 200 were randomly selected from the simulated populations. Sample size (n), number of predictor variables (p), and the n/p ratio to be simulated are summarized in Table 2.

Multicollinearity

From the literature reviewed, the typical intercorrelation among the independent variables in the psychological and educational literature ranged from .01 to .65 (Claudy, 1978). As can be suspected, most of the independent variables in the regression analysis in education and psychological research are related in a variety of ways to different degrees. However, because of time constraints and project manageability, in the present study, three conditions of intercorrelation among the independent variables (.1, .3, .5)

Table 2

Summary of Sample Size (*n*), Number of Predictor Variables (*p*), and *n/p* Ratio

Number	Sample size (*n*)				
of predictors (*p*)	20	40	60	100	200
2	10	20	30	50	100
4	5	10	15	25	50
8	2.5	5	7.5	12.5	25

were simulated to represent typical multicollinearity conditions in the real data. Also considering program manageability, the degree of multicollinearity among all the independent variables is specified to be equal; that is, the correlation coefficients among all the independent variables are the same.

Replications

From the previous studies that used simulated data, the number of replications were chosen to be 100 (Kennedy, 1988; Newman et al., 1979) and 1000 (Kromrey & Hines, 1995, 1996). In order to obtain stable estimates of sample statistics, a certain number of replications are needed in the simulation process. In the present study, 500 samples were drawn under each of the cell conditions, which will be discussed later in detail.

Simulation Design

The fully crossed experimental design of three conditions of population squared multiple correlation coefficients (ρ^2= .2, .5, .8), three conditions of predictor numbers (p= 2, 4, 8), five conditions of sample sizes (n= 20, 40, 60, 100, 200), and three conditions of multicollinearity (.1, .3, .5) entails 135 cell conditions (3×3×5×3). Within each cell condition, 500 samples were randomly drawn. This makes the total number of replications in the study 67,500 [(3×3×5×3)× 500].

The simulation design for *one of the three* multicollinearity conditions is graphically illustrated in Figure 1.

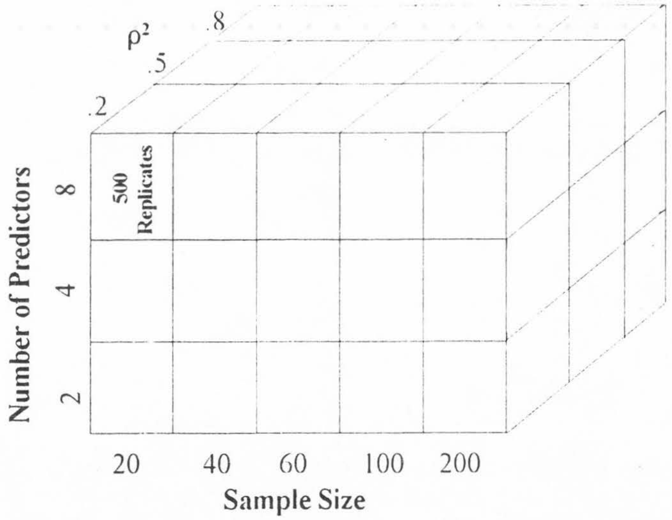


Figure 1. Simulation design for one of the three multicollinearity conditions.

Data Generation

Generating Correlated Multivariate Normal Data

Matrix decomposition procedure is used to generate correlated multivariate normal data (Kaiser & Dickman, 1962). Using matrix decomposition, a specified correlation matrix can be imposed on a set of random normal variables to yield correlated multivariate normal data. In the present study, to generate multivariate normal data within each cell condition, the following steps were implemented:

First, for each of the three multicollinearity situations (.1, .3, .5), the population correlation coefficients among the independent variables were set to be either .1, .3, or .5. Next, the correlation coefficients among the dependent variable and independent variables were chosen to yield the desired population squared multiple correlation coefficient ρ^2 (.2, .5, .8). In total, 27 population intercorrelation matrices (3×3×3), three multicollinearity conditions, three conditions of predictor number, and three population squared multiple correlation coefficients were obtained. The SAS program files included the population intercorrelation matrices and the output of the squared multiple correlation coefficients ρ^2 are listed in Appendix B.

Second, within each cell condition, uncorrelated random normal variables for the required number of independent variables and required sample sizes were generated. The SAS pseudorandom number generator (rannor) and the SAS MACRO language were used for this purpose. This procedure was conducted through the IML (Interactive Matrix Language) software of SAS (SAS/IML, 1990) .

Third, premultiply the uncorrelated data matrix generated in step 2 with the principal component loadings matrix, which was obtained by applying the principal component factorization to the population intercorrelation matrix obtained in step 1. The resultant data matrix became a matrix of correlated multivariate normal data, which was equivalent to data randomly sampled from a population with specified intercorrelation patterns (Kaiser & Dickman, 1962). This procedure was also conducted through the IML software of SAS (SAS/IML, 1990). The SAS program files for step 2 and step 3 are selectively listed in Appendix C.

Estimating the Population Cross-Validity ρ_c^2

Although the desired population squared multiple correlation coefficient ρ^2 can be prespecified, the population cross-validity coefficient ρ_c^2 is always unknown. As a result, it can only be empirically estimated through repeated sampling from a prespecified statistical population. In this study, the population cross-validity coefficient ρ_c^2 was estimated through the procedure similar to double cross-validation (Mosier, 1951). Cross-validation needs two equivalent samples that came from the same population, and the regression equation derived from one sample was applied to the other sample to predict the dependent variable, and obtained a sample cross-validity coefficient. To implement the estimation procedure through repeated sampling, the following steps were followed.

First, the steps for generating correlated multivariate normal data with the matrix decomposition procedure described above were followed to generate random samples of correlated multivariate data.

Second, within each cell condition (shown in Figure 1), 500 random samples were drawn from the corresponding simulated population with specified population parameters.

Third, the 500 random samples were randomly assigned into 250 pairs of random samples. For each pair of random samples, regression analysis was conducted in each sample and the sample regression weights were obtained.

Fourth, for each pair of random samples, the sample regression weights obtained from one sample were then applied to the other sample to predict the corresponding dependent variable, and vice versa.

Fifth, for each pair of random samples, Pearson correlation coefficients between the predicted values obtained in step 4 above and the actual values of the dependent variable were calculated for the two samples as the sample cross-validity coefficient. Two Pearson correlation coefficients were obtained for each pair of these random samples. The 250 pairs of random samples yielded 500 such sample cross-validity coefficients.

Sixth, the obtained sample cross-validity coefficients were squared. The average of these squared coefficients was the estimate of the population squared cross-validity coefficient ρ_c^2.

The procedure was also conducted through the SAS/IML software (SAS/IML, 1990). The SAS program files for estimating the population squared cross-validity coefficient (ρ_c^2) are selectively listed in Appendix D.

Obtaining Sample Adjusted $\hat{\rho}^2$ and $\hat{\rho}_c^2$

Estimator of the Population ρ^2

Six analytical formulae were designed to estimate the squared population correlation coefficient ρ^2. These analytical formulae were applied to the multivariate normal data generated with known population parameters (ρ^2, number of predictors, sample sizes, multicollinearity among predictors). The corrected or adjusted R^2s based on each of these six formulae were then obtained. The procedure was also conducted through the SAS system, and the SAS program files for calculating the corrected R^2 estimate ρ^2 are listed in Appendix E.

Estimator of the Population ρ_c^2

Nine analytical formulae were designed to estimate the squared population cross validity coefficient ρ_c^2. These analytical formulae are also applied to the simulated data with known population parameters (ρ^2, number of predictors, sample sizes, multicollinearity among predictors) and the corrected or adjusted R_c^2s based on these formulae were obtained. The procedure was also conducted through the SAS system, and the SAS program files for calculating corrected R_c^2 are listed in Appendix E.

CHAPTER IV

RESULTS

Descriptive Statistics

Means and standard deviations of the 500 replicates for the sample adjusted R^2 based on these six analytical formulae for estimating ρ^2, and the sample adjusted R_c^2 based on the 10 analytical formulae for estimating ρ_c^2 in all the specified sampling conditions (i.e., population squared correlation coefficient, number of predictors, different degree of multicollinearity, and sample sizes) are obtained. To save space, means obtained from the 16 analytical formulae (for the 15 *n/p* ratio conditions across the three conditions of multicollinearity), together with the population ρ^2 and estimated population ρ_c^2, are summarized in Tables 3, 4, and 5.

Estimating Statistical Bias

To guide the evaluation of the estimates of statistical bias, an unbiased estimate was operationally defined as having means based on the 500 replicates to be within ±.01 of the corresponding population parameters (Kromrey & Hines, 1996).

Population ρ^2 and Unadjusted Sample R^2

The bias in the unadjusted sample R^2 across the 135 sampling conditions was obvious, especially when *n/p* ratio was small. The sample R^2 was almost always consistently larger than the corresponding population ρ^2. Only 2 out of 135 conditions

Table 3

Summary of Means of the Adjusted R^2 and R_c^2 (Multicollinearity r = .1)

N/p	*p*	n	ρ^2	R^2	1	2	3	4	5	6	ρ_c^2	7	8	9	10	11	12	13	14	15	16
2.5	8	20	.2	.530	.216	.187	.255	.215	.195	.261	.119	-.241	-.179	.190	-.536	.127	.145	.165	-.390	-.098	.119
			.5	.693	.494	.476	.519	.505	.496	.543	.315	.200	.240	.378	.008	.318	.347	.340	.103	.292	.296
			.8	.874	.790	.783	.801	.801	.800	.812	.681	.668	.685	.717	.589	.666	.691	.698	.628	.706	.649
5	4	20	.2	.354	.193	.182	.233	.209	.195	.256	.159	-.076	-.022	.180	-.104	.152	.172	.162	.001	.032	.142
			.5	.594	.493	.486	.518	.515	.509	.544	.408	.324	.358	.421	.307	.415	.445	.407	.373	.392	.391
			.8	.829	.786	.784	.797	.802	.800	.813	.767	.715	.729	.747	.708	.742	.763	.740	.736	.744	.728
5	8	40	.2	.358	.197	.192	.217	.202	.199	.223	.120	-.016	.010	.134	-.049	.126	.135	.125	.002	.036	.117
			.5	.595	.494	.491	.507	.505	.503	.518	.414	.361	.377	.414	.339	.402	.417	.399	.372	.393	.390
			.8	.833	.791	.790	.796	.798	.798	.804	.749	.736	.743	.752	.727	.744	.755	.748	.740	.749	.738
7.5	8	60	.2	.298	.190	.188	.203	.194	.193	.208	.141	.050	.066	.130	.042	.130	.136	.120	.074	.819	.122
			.5	.564	.496	.495	.505	.504	.503	.513	.449	.409	.419	.438	.405	.436	.445	.432	.424	.429	.427
			.8	.823	.796	.795	.799	.801	.800	.804	.772	.761	.765	.770	.759	.768	.774	.768	.767	.769	.764
10	2	20	.2	.266	.184	.179	.225	.207	.194	.253	.205	.007	.056	.171	.031	.166	.188	.166	.123	.103	.152
			.5	.542	.491	.488	.517	.518	.512	.546	.481	.381	.412	.448	.396	.462	.492	.443	.453	.440	.436
			.8	.805	.784	.782	.794	.801	.800	.812	.780	.737	.750	.763	.743	.770	.789	.760	.767	.762	.756
10	4	40	.2	.278	.198	.196	.218	.206	.203	.228	.153	.072	.095	.151	.079	.156	.166	.146	.123	.118	.143
			.5	.540	.489	.488	.502	.501	.500	.515	.449	.409	.424	.444	.413	.449	.463	.439	.442	.438	.436
			.8	.811	.790	.790	.795	.798	.798	.804	.781	.757	.763	.770	.759	.772	.781	.768	.771	.769	.766
12.5	8	100	.2	.263	.199	.198	.207	.202	.201	.210	.160	.117	.126	.153	.118	.155	.158	.146	.136	.135	.149
			.5	.536	.495	.495	.500	.500	.500	.505	.466	.444	.449	.458	.445	.459	.465	.456	.456	.455	.454
			.8	.813	.797	.797	.799	.800	.800	.802	.785	.776	.778	.781	.776	.781	.785	.780	781	.781	.779
15	4	60	.2	.253	.200	.198	.213	.205	.204	.219	.168	.117	.132	.161	.125	.167	.174	.156	.153	.146	.157
			.5	.523	.488	.488	.497	.497	.496	.505	.471	.436	.445	.457	.441	.462	.471	.454	.459	.452	.453
			.8	.811	.798	.797	.801	.803	.803	.806	.784	.777	.780	.784	.779	.786	.792	.784	.786	.784	.783
20	2	40	.2	.233	.192	.191	.212	.202	.199	.223	.183	.108	.131	.165	.125	.176	.187	.166	.168	.152	.161
			.5	.514	.489	.488	.502	.502	.500	.515	.487	.436	.450	.465	.446	.475	.489	.463	.473	.463	.462
			.8	.810	.800	.799	.805	.808	.807	.813	.787	.779	.784	.790	.783	.794	.802	.789	.793	.790	.788
25	4	100	.2	.228	.196	.196	.204	.199	.199	.207	.183	.147	.155	.169	.153	.174	.178	.166	.170	.164	.167
			.5	.509	.489	.489	.494	.494	.494	.499	.488	.458	.463	.469	.461	.463	.478	.468	.472	.468	.468
			.8	.805	.797	.797	.799	.800	.800	.802	.790	.784	.786	.789	.786	.790	.794	.788	.790	.789	.788
25	8	200	.2	.227	.195	.195	.199	.197	.196	.201	.176	.154	.158	.168	.157	.170	.172	.165	.165	.163	.166
			.5	.517	.497	.497	.499	.499	.499	.502	.482	.472	.474	.478	.473	.479	.482	.477	.478	.477	.477
			.8	.806	.798	.798	.799	.800	.800	.800	.790	.788	.789	.790	.788	.791	.792	.790	.790	.790	.790
30	2	60	.2	.226	.199	.198	.212	.205	.204	.219	.194	.144	.158	.177	.156	.187	.193	.174	.184	.172	.175
			.5	.510	.493	.493	.502	.502	.501	.510	.493	.458	.467	.477	.466	.484	.493	.476	.483	.476	.475
			.8	.798	.791	.791	.795	.797	.797	.800	.795	.777	.781	.784	.780	.788	.793	.784	.787	.784	.784

(table continues)

N/p	*p*	n	ρ^2	R^2	1	2	3	4	5	6	ρ_c^2	7	8	9	10	11	12	13	14	15	16
50	2	100	.2	.201	.192	.192	.200	.196	.195	.204	.195	.159	.168	.178	.167	.184	.188	.177	.183	.176	.177
			.5	.506	.496	.496	.501	.501	.501	.506	.497	.475	.480	.486	.480	.490	.496	.485	.490	.486	.485
			.8	.799	.795	.795	.797	.799	.799	.801	.796	.787	.789	.791	.789	.793	.797	.791	.793	.791	.791
50	4	200	.2	.213	.197	.197	.201	.199	.199	.203	.191	.173	.177	.182	.176	.186	.187	.182	.185	.181	.182
			.5	.501	.497	.496	.499	.499	.499	.502	.492	.481	.484	.487	.484	.489	.491	.486	.489	.486	.486
			.8	.800	.796	.796	.797	.798	.798	.799	.798	.790	.791	.792	.791	.793	.795	.792	.793	.792	.792
100	2	200	.2	.208	.200	.200	.204	.201	.201	.205	.196	.184	.188	.192	.187	.196	.197	.192	.196	.192	.192
			.5	.502	.497	.497	.500	.500	.500	.503	.498	.487	.490	.492	.490	.495	.497	.492	.495	.492	.492
			.8	.800	.798	.798	.799	.799	.799	.800	.797	.794	.795	.796	.795	.797	.798	.796	.797	.796	.796

Note. *N/p: N/p* Ratio. *p:* Number of predictor variables. n: Sample size. ρ^2: Squared population multiple correlation coefficient. R^2: Sample R^2 without adjustment. 1: the Smith formula. 2: the Wherry formula-1. 3: the Wherry formula-2. 4: the Olkin and Pratt formula. 5: the Pratt estimation of the Olkin and Pratt formula. 6: the Claudy-3 formula. ρ_c^2: (Estimated) population squared cross-validity coefficient. 7: the Lord formula-1 8: the Lord formula-2. 9: the Burket formula. 10: the Darlington formula. 11: the Browne formula-1 with ρ^2 estimated by the Wherry-1 formula. 12: the Browne formula-2 with ρ^2 estimated by the Olkin and Pratt formula. 13: the Claudy formula-1. 14: the Claudy formula-2. 15: the Rozeboom formula-1 . 16: the Rozeboom formula-2.

Table 4

Summary of Means of the Adjusted R^2 and R_c^2 (Multicollinearity r = .3)

N/p	*p*	n	ρ^2	R^2	1	2	3	4	5	6	ρ_c^2	7	8	9	10	11	12	13	14	15	16
2.5	8	20	.2	.532	.221	.193	.260	.220	.201	.266	.121	-.232	-.170	.183	-.526	.128	.145	.155	-.380	-.090	.118
			.5	.700	.500	.481	.524	.511	.502	.539	.322	.208	.248	.383	.019	.323	.353	.344	.113	.299	.301
			.8	.877	.795	.787	.805	.805	.803	.816	.680	.675	.691	.723	.598	.672	.697	.704	.636	.712	.656
5	4	20	.2	.367	.209	.198	.248	.226	.212	.271	.156	-.055	-.002	.182	-.082	.163	.184	.179	.021	.051	.152
			.5	.578	.472	.465	.499	.495	.489	.525	.417	.296	.332	.396	.278	.390	.421	.381	.347	.367	.365
			.8	.823	.803	.802	.813	.819	.818	.830	.752	.760	.772	.784	.766	.790	.808	.782	.789	.784	.778
5	8	40	.2	.354	.193	.188	.312	.198	.195	.220	.125	-.021	.005	.130	-.055	.122	.131	.118	-.003	.032	.113
			.5	.598	.498	.494	.510	.508	.507	.521	.409	.365	.381	.417	.344	.406	.421	.403	.376	.397	.393
			.8	.936	.795	.794	.800	.802	.802	.807	.753	.741	.747	.756	.732	.749	.759	.753	.745	.754	.742
7.5	8	60	.2	.299	.191	.189	.204	.195	.194	.209	.131	.051	.067	.133	.043	.132	.138	.122	.075	.083	.124
			.5	.564	.497	.495	.505	.504	.504	.513	.440	.410	.420	.439	.405	.436	.445	.432	.424	.430	.427
			.8	.822	.794	.794	.798	.799	.799	.803	.769	.759	.763	.768	.757	.766	.772	.766	.765	.767	.762
10	2	20	.2	.271	.190	.185	.231	.213	.200	.259	.205	.014	.063	.012	.378	.174	.195	.171	.129	.109	.162
			.5	.525	.473	.470	.499	.499	.494	.528	.493	.358	.390	.429	.373	.443	.473	.421	.433	.420	.416
			.8	.801	.779	.778	.790	.797	.796	.809	.780	.731	.745	.758	.738	.765	.784	.755	.763	.757	.752
10	4	40	.2	.270	.189	.187	.209	.197	.194	.219	.157	.062	.085	.148	.068	.151	.161	.144	.114	.108	.140
			.5	.539	.488	.487	.501	.500	.499	.513	.452	.408	.423	.443	.412	.448	.462	.438	.441	.437	.435
			.8	.816	.806	.806	.811	.814	.814	.819	.775	.786	.791	.796	.790	.800	.809	.796	.800	.796	.795
12.5	8	100	.2	.262	.198	.197	.206	.201	.200	.209	.153	.116	.125	.152	.117	.154	.157	.144	.135	.134	.148
			.5	.536	.495	.495	.500	.500	.500	.505	.466	.444	.449	.458	.444	.459	.464	.456	.455	.455	.454
			.8	.813	.797	.796	.799	.800	.800	.802	.787	.776	.778	.781	.776	.781	.784	.780	.780	.780	.779
15	4	60	.2	.253	.200	.199	.213	.206	.204	.219	.175	.118	.132	.162	.125	.168	.174	.156	.154	.147	.157
			.5	.526	.492	.492	.501	.500	.500	.509	.479	.440	.449	.461	.445	.466	.475	.459	.463	.458	.457
			.8	.806	.799	.799	.802	.805	.804	.808	.784	.785	.789	.793	.788	.796	.801	.792	.795	.793	.792
20	2	40	.2	.238	.198	.197	.218	.208	.205	.229	.183	.115	.137	.171	.131	.182	.193	.172	.174	.158	.167
			.5	.519	.494	.494	.507	.507	.506	.520	.478	.442	.456	.471	.452	.480	.494	.468	.488	.469	.467
			.8	.804	.793	.793	.798	.801	.801	.807	.797	.772	.777	.783	.776	.787	.796	.782	.787	.783	.782
25	4	100	.2	.228	.196	.195	.204	.199	.199	.207	.178	.147	.155	.169	.153	.174	.178	.166	.170	.164	.167
			.5	.513	.493	.493	.498	.498	.498	.503	.480	.462	.468	.473	.466	.476	.483	.473	.477	.473	.472
			.8	.803	.799	.799	.801	.802	.802	.804	.790	.791	.793	.795	.793	.797	.800	.795	.797	.795	.795
25	8	200	.2	.229	.197	.197	.201	.198	.198	.202	.177	.156	.161	.170	.159	.172	.174	.167	.167	.165	.168
			.5	.517	.497	.497	.500	.499	.500	.502	.479	.472	.474	.478	.473	.479	.482	.477	.478	.477	.477
			.8	.806	.798	.798	.799	.800	.800	.801	.793	.788	.789	.790	.789	.791	.792	.790	.791	.790	.790
30	2	60	.2	.217	.190	.190	.204	.196	.195	.210	.198	.135	.149	.168	.147	.178	.185	.168	.175	.163	.167
			.5	.508	.491	.491	.500	.500	.499	.508	.497	.456	.465	.474	.464	.482	.491	.473	.481	.474	.473
			.8	.806	.799	.799	.802	.804	.804	.808	.791	.785	.789	.792	.788	.795	.801	.792	.795	.792	.792

(table continues)

N/p	*p*	n	ρ^2	R^2	1	2	3	4	5	6	ρ_c^2	7	8	9	10	11	12	13	14	15	16
50	2	100	.2	.216	.200	.200	.208	.204	.203	.212	.201	.168	.176	.186	.175	.192	.196	.185	.191	.184	.185
			.5	.501	.491	.490	.496	.496	.495	.501	.491	.470	.475	.481	.475	.485	.491	.480	.485	.480	.480
			.8	.800	.796	.796	.798	.799	.799	.801	.795	.787	.790	.792	.789	.794	.797	.792	.793	.792	.791
50	4	200	.2	.214	.198	.197	.202	.199	.199	.203	.198	.173	.177	.183	.177	.186	.188	.182	.185	.181	.182
			.5	.511	.501	.501	.503	.503	.503	.506	.496	.486	.488	.491	.488	.493	.496	.491	.493	.491	.491
			.8	.804	.802	.802	.803	.803	.803	.804	.797	.798	.799	.800	.799	.801	.802	.800	.801	.800	.800
100	2	200	.2	.206	.198	.198	.202	.200	.200	.204	.198	.182	.186	.190	.186	.194	.196	.190	.194	.190	.190
			.5	.506	.501	.501	.503	.503	.503	.506	.498	.490	.493	.496	.493	.498	.501	.496	.498	.496	.495
			.8	.802	.800	.800	.801	.802	.802	.803	.798	.796	.797	.798	.797	.799	.801	.798	.799	.798	.798

Note. *N/p: N/p* Ratio. *p:* Number of predictor variables. n: Sample size. ρ^2: Squared population multiple correlation coefficient. R^2: Sample R^2 without adjustment. 1: the Smith formula. 2: the Wherry formula-1. 3: the Wherry formula-2. 4: the Olkin and Pratt formula. 5: the Pratt estimation of the Olkin and Pratt formula. 6: the Claudy-3 formula. ρ_c^2: (Estimated) population squared cross-validity coefficient. 7: the Lord formula-1 8: the Lord formula-2. 9: the Burket formula. 10: the Darlington formula. 11: the Browne formula-1 with ρ^2 estimated by the Wherry-1 formula. 12: the Browne formula-2 with ρ^2 estimated by the Olkin and Pratt formula. 13: the Claudy formula-1. 14: the Claudy formula-2. 15: the Rozeboom formula-1 . 16: the Rozeboom formula-2.

Table 5

Summary of Means of the Adjusted R^2 and R_c^2 (Multicollinearity r = .5)

N/p	*p*	n	ρ^2	R^2	1	2	3	4	5	6	ρ_c^2	7	8	9	10	11	12	13	14	15	16
2.5	8	20	.2	.522	.203	.174	.243	.201	.181	.248	.109	-.261	-.198	.184	-.561	.123	.140	.162	-.412	-.116	.116
			.5	.698	.496	.478	.521	.507	.498	.536	.326	.203	.243	.383	.128	.324	.353	.357	.107	.294	.302
			.8	.874	.790	.782	.800	.800	.788	.812	.673	.668	.684	.717	588	.666	.691	.698	.628	.706	.649
5	4	20	.2	.353	.192	.181	.232	.208	.194	.255	.146	-.078	-.024	.166	-.106	.148	.168	.157	-.001	.030	.137
			.5	.596	.495	.488	.520	.517	.511	.546	.426	.326	.360	.422	.309	.415	.446	.404	.375	.394	.390
			.8	.829	.786	.783	.797	.801	.800	.813.	.763	.715	.729	.746	.707	.741	.762	.739	.735	.743	.727
5	8	40	.2	.353	.191	.186	.212	.197	.193	.219	.133	-.022	.003	.131	-.056	.122	.131	.119	-.005	.030	.113
			.5	.592	.490	.487	.503	.500	.499	.514	.407	.355	.371	.409	.334	.398	.413	.395	.366	.388	.385
			.8	.834	.792	.791	.797	.799	.799	.805	.753	.727	.744	.753	.728	.746	.756	.749	.741	.750	.739
7.5	8	60	.2	.305	.198	.195	.211	.202	.201	.216	.130	.059	.075	.137	.051	.136	.142	.124	.082	.091	.128
			.5	.567	.500	.499	.508	.508	.507	.516	.445	.414	.424	.442	.409	.440	.449	.436	.428	.433	.431
			.8	.823	.796	.795	.799	.801	.801	.804	.769	.760	.764	.769	.758	.768	.774	.768	.766	.768	.764
10	4	40	.2	.267	.185	.183	.206	.194	.191	.215	.167	.057	.081	.142	.064	.146	.156	.135	.110	.104	.134
			.5	.540	.489	.488	502	.501	.500	.515	.463	.409	.424	.444	.413	.449	.463	.439	.442	.438	.436
			.8	.812	.791	.791	.796	.799	.799	.805	.772	.758	.764	.771	.760	.773	.782	.769	.772	.770	.767
10	2	20	.2	.282	.202	.198	.242	.225	.213	.271	.180	.029	.077	.223	.052	.182	.205	.174	.142	.122	.166
			.5	.532	.480	.477	.506	.507	.501	.536	.450	.367	.399	.437	.383	.451	.481	.432	.441	.428	.425
			.8	.801	.779	.777	.790	.796	.795	.808	.793	.730	.744	.757	.747	.764	.783	.754	.762	.756	.751
12.5	8	100	.2	.264	.200	.199	.208	.203	.202	.211	.160	.118	.127	.154	.119	.156	.160	.147	.137	.136	.150
			.5	.540	.500	.500	.505	.505	.505	.510	.461	.449	.455	.464	.450	.464	.470	.461	.461	.460	.459
			.8	.814	.798	.798	.800	.801	.801	.803	.787	.777	.779	.782	.777	.782	.786	.781	.782	.782	.780
15	4	60	.2	.255	.202	.201	.215	.208	.206	.222	.173	.120	.134	.164	.128	.170	.177	.159	.156	.149	.160
			.5	.527	.494	.493	.502	.502	.501	.511	.479	.442	.451	.462	.446	.467	.476	.460	.465	.460	.459
			.8	.804	.790	.790	.793	.795	.795	.799	.787	.768	.772	.776	770	.778	.784	.776	.779	.776	.774
20	2	40	.2	.227	.186	.185	.207	.196	.193	.218	.184	.102	.124	.158	.119	.170	.180	.157	.162	.146	.155
			.5	.506	.480	.479	.493	.493	.491	.506	.486	.426	.440	.455	.436	.466	.479	.453	.464	.454	.452
			.8	.804	.794	.794	.800	.802	.802	.807	.792	.773	.778	.784	.777	.788	.797	.783	.788	.784	.783
25	4	100	.2	.231	.198	.198	.206	.202	.201	.201	.180	.150	.158	.172	.156	.177	.181	.170	.172	.166	.170
			.5	.518	.498	.498	.503	.503	.503	.508	.483	.468	.473	.479	.472	.483	.488	.478	.482	.478	.478
			.8	.805	.797	.797	.799	.800	.800	.802	.791	.785	.787	.789	.786	.790	.794	.789	.790	.789	.788
25	8	200	.2	.231	.199	.199	.203	.200	.200	.205	.176	.158	.163	.172	.161	.174	.176	.169	.169	.167	.170
			.5	.519	.499	.498	.501	.501	.501	.503	.482	.473	.476	.480	.475	.481	.483	.479	.480	.478	.478
			.8	.806	.798	.798	.799	.799	.799	.800	.794	.788	.789	.790	.788	.790	.792	.790	.790	.790	.789
30	2	60	.2	.224	.197	.196	.210	.203	.202	.217	.191	.142	.156	.175	.154	.184	.191	.173	.182	.170	.173
			.5	.515	.498	.498	.506	.506	.506	.515	.488	.463	.472	.482	.471	.489	.498	.481	.488	.481	.480
			.8	.799	.792	.792	.795	.797	.797	.801	.792	.778	.781	.785	.781	.788	.794	.785	.787	.785	.784

(table continues)

N/p	*p*	n	ρ^2	R^2	1	2	3	4	5	6	ρ_c^2	7	8	9	10	11	12	13	14	15	16
50	2	100	.2	.226	.211	.210	.218	.214	.214	.222	.195	.178	.187	.196	.186	.203	.207	.195	.202	.195	.195
			.5	.507	.497	.497	.502	.502	.502	.507	.493	.477	.482	.487	.482	.492	.497	.487	.492	.487	.487
			.8	.802	.798	.798	.800	.801	.801	.803	.797	.790	.792	.794	.792	.796	.799	.794	.796	.794	.795
50	4	200	.2	.214	.198	.198	.202	.199	.199	.203	.191	.173	.178	.183	.177	.186	.188	.182	.185	.182	.182
			.5	.509	.499	.499	.502	.502	.502	.504	.493	.484	.487	.489	.486	.492	.494	.489	.491	.489	.489
			.8	.802	.798	.798	.799	.799	.799	.800	.797	.792	.793	.794	.793	.795	.796	.794	.795	.794	.794
100	2	200	.2	.205	.197	.197	.201	.199	.199	203	.196	.181	.185	.190	.185	.193	.195	.189	.193	.189	.189
			.5	.505	.500	.500	.502	.502	.502	.505	.498	.490	.492	.494	.492	.497	.500	.495	.497	.495	.495
			.8	.801	.799	.799	.800	.800	.800	.801	.799	.795	.796	.797	.796	.798	.799	.797	.798	.797	.797

Note. *N/p: N/p* Ratio. *p:* Number of predictor variables. n: Sample size. ρ^2: Squared population multiple correlation coefficient. R^2: Sample R^2 without adjustment. 1: the Smith formula. 2: the Wherry formula-1. 3: the Wherry formula-2. 4: the Olkin and Pratt formula. 5: the Pratt estimation of the Olkin and Pratt formula. 6: the Claudy-3 formula. ρ_c^2: (Estimated) population squared cross-validity coefficient. 7: the Lord formula-1 8: the Lord formula-2. 9: the Burket formula. 10: the Darlington formula. 11: the Browne formula-1 with ρ^2 estimated by the Wherry-1 formula. 12: the Browne formula-2 with ρ^2 estimated by the Olkin and Pratt formula. 13: the Claudy formula-1. 14: the Claudy formula-2. 15: the Rozeboom formula-1 . 16: the Rozeboom formula-2.

where the sample R^2s were minimally smaller than the corresponding population ρ^2 : (a) multicollinearity $r = .1$, $n/p = 30$ ($p = 2$, $n = 60$) , population $\rho^2 = .8$, sample $R^2 = .798$; and (b) multicollinearity $r = .1$, $n/p = 50$ ($p = 2$, $n = 100$), population $\rho^2 = .8$, sample $R^2 = .799$. From these results, it was obvious that the statistical bias in multiple regression was almost always positive, although not in every single case. Such results confirmed the common concept of positive bias from previous studies (Cummings, 1982; Huberty & Mourad, 1980; Kromrey & Hines, 1995), but differed in the sense that the "bias" was not *always* positive.

Population ρ_c^2 and Unadjusted Sample R^2

From these tables, all the unadjusted sample R^2s were greater than their corresponding estimated population cross-validity coefficient ρ_c^2 across the 135 sampling conditions. Such results also confirmed the findings from previous numerous studies (e.g., Claudy, 1978; Cummings, 1982; Herzberg, 1969).

Population ρ^2 and Population ρ_c^2

From these tables, it was also observed that the estimated population cross-validity coefficient ρ_c^2 was almost consistently smaller than the corresponding population ρ^2. Only three instances where the estimated population cross-validity coefficient ρ_c^2s were minimally greater than the corresponding population ρ^2s: (a) for population $\rho^2 = .2$, $\rho_c^2 = .205$, while multicollinearity $r = .1$, and $n/p = 10$ ($p = 2$, $n = 20$); (b) for population $\rho^2 = .2$, $\rho_c^2 = .205$, while multicollinearity $r = .3$, and $n/p = 10$ ($p = 2$, $n = 20$); and (c) for population $\rho^2 = .2$, $\rho_c^2 = .201$, while multicollinearity $r = .1$, and $n/p = 50$ ($p = 2$, $n = 100$).

Such results confirmed the results from the previous studies (Claudy, 1978; Cummings, 1982; Herzberg, 1969), although ρ_c^2 maybe larger than ρ^2 in a few rare cases.

Overall Summary

To help evaluate the performance of individual formula under different sampling conditions, summary of frequencies of each analytical formula as an "unbiased estimate" across different degrees of multicollinearity, population ρ^2, and *n/p* ratio are listed in Table 6. Because there were too many *n/p* ratio conditions, for the sake of clarity, only 5 *n/p* ratio conditions (5, 10, 25, 50, and 100) are presented in this table.

Best Estimator(s) of the Population ρ^2

Based on Table 6 and the relative rankings of percentages of unbiased estimates, for the six analytical formulae estimating the population ρ^2, several observations are made:

1. Across the three different conditions of multicollinearity, approximately 91% to 98% of the time the Pratt formula gave unbiased estimates of the population ρ^2 that gave the best performance among the six analytical formulae.

2. Across the three different conditions of population ρ^2, approximately 93% to 96% of the time the Pratt formula gave unbiased estimates of population ρ^2. Still, its performance was the best among the six analytical formulae.

3. Across the five different conditions of *n/p* ratio, approximately 83% to 100% of the time both the Pratt formula gave unbiased estimates of population ρ^2. Again the performance of the Pratt formula was the best among the six analytical formulae.

Table 6

Percentages of Cell Conditions in Which Unbiased Estimates Are Observed Across Multicollinearity Conditions, Population ρ^2, and n/p Ratio--Estimators of ρ^2

	Multicollinearity				Population ρ^2				n/p Ratio					
Formula	Rank1[b]	.1	.3	.5	Rank 2[c]	.2	.5	.8	Rank3[d]	5	10	25	50	100
Smith	3	82.22	86.67	82.22	3	84.44	77.78	88.89	3	100	38.89	94.44	94.44	100
Wherry-1	5	62.22	77.78	68.89	4	66.67	62.22	80.00	5	66.67	27.78	94.44	100	100
Wherry-2	4	71.11	80.00	80.00	5	53.33	86.67	93.33	4	66.67	66.67	100	94.44	100
Olkin/Pratt	2[a]	93.33	86.67	93.33	2[a]	91.11	91.11	95.56	2[a]	100	77.78	100	94.44	100
Pratt	1[a]	97.78	91.11	91.11	1[a]	95.56	93.33	93.33	1[a]	100	83.33	100	94.44	100
Claudy-3	6	57.78	60.00	57.78	6	42.22	53.33	80.00	6	33.33	22.22	100	89.89	100

[a] Indicates the best two rankings.

[b] Rank1: performance ranking of the analytical formulae across different conditions of multicollinearity; lower ranking indicates better performances.

[c] Rank2: performance ranking of the analytical formulae across different conditions of population ρ^2; lower ranking indicates better performances.

[d] Rank3: performance ranking of the analytical formulae across different n/p ratio; lower ranking indicates better performances.

Table 7

Percentages of Cell Conditions in Which Unbiased Estimates Are Observed Across Multicollinearity Conditions, Population ρ^2, and *n/p* Ratio -- Estimators of ρ_c^2

	Multicollinearity				Population ρ^2				*n/p* Ratio					
Formula	Rank1[b]	.1	.3	.5	Rank 2[c]	.2	.5	.8	Rank3[d]	5	10	25	50	100
Lord-1	9	22.22	82.89	24.44	10	0	20.00	51.11	10	0	0	50.00	44.44	66.67
Lord-2	8	28.89	26.67	31.11	8	2.22	22.22	60.00	7	33.33	5.56	55.56	55.56	77.78
Burket	5	44.44	57.78	21.11	3	44.44	57.78	64.44	3	77.78	22.22	88.89	77.78	100
Darlington	10	24.44	22.22	22.22	9	4.44	20.00	44.44	9	0	0	50.00	55.56	77.78
Browne-1	1[a]	73.33	77.78	77.78	1[a]	75.56	71.11	80.00	1[a]	77.78	50.00	94.44	88.89	100
Browne-2	2[a]	71.11	75.56	75.56	2[a]	66.67	71.11	84.44	2[a]	55.56	50.00	100	94.44	100
Claudy-1	4	44.44	51.11	57.78	5	40.00	46.67	64.44	4	66.67	22.22	77.78	72.22	100
Claudy-2	3	48.89	46.67	57.78	6	31.11	46.67	73.33	5	22.22	22.22	83.33	89.89	100
Rozeboom-1	7	31.11	35.56	48.89	7	15.56	33.33	66.67	6	33.33	5.56	66.67	72.22	100
Rozeboom-2	6	37.78	44.44	57.78	4	42.22	33.33	60.00	8	11.11	11.11	83.33	61.11	100

[a] Indicates the best two rankings.

[b] Rank1: performance ranking of the analytical formulae across different conditions of multicollinearity; lower ranking indicates better performances.

[c] Rank2: performance ranking of the analytical formulae across different conditions of population ρ^2; lower ranking indicates better performances.

[d] Rank3: performance ranking of the analytical formulae across different *n/p* ratio; lower ranking indicates better performances.

The overall performance of the most commonly used (in both SPSS and SAS) Wherry-2 formula was approximately the fourth or the fifth best among the six analytical formulae. The Pratt formula, Olkin and Pratt formula, and the Smith formula all outperformed the Wherry-2 formula. Based on the results obtained for estimating the population ρ^2, the Wherry-2 formula did not demonstrate any advantage over those formulae mentioned above.

Best Estimator(s) of the Population Cross-Validity Coefficient ρ_c^2

From Table 7 and the relative rankings of percentages of unbiased estimates, for the nine analytical formulae estimating population cross-validity coefficient ρ_c^2, the following observations were made:

1. Across the three different degrees of multicollinearity, approximately 73% to 78% of the time the Browne formula (with ρ^2 estimated by the Wherry formula-1) gave an unbiased estimate of the population cross-validity coefficient ρ_c^2 that gave the best performance among the nine analytical formulae.

2. Across the three different conditions of population ρ^2, approximately 76% to 80% of the time the Browne formula (with ρ^2 estimated by the Wherry formula-1) gave an unbiased estimate of the population cross-validity coefficient ρ_c^2. Still its performance was the best.

3. Across the five different conditions of *n/p* ratio, approximately 50% to 100% of the time the Browne formula (with ρ^2 estimated by either the Wherry formula-1 or the Olkin\Pratt formula) gave an unbiased estimate of the population cross-validity coefficient

ρ_c^2. Again, the performance of the Browne formula is the best among the nine analytical formulae.

In all, the overall performance of the Pratt formula was the best among the six analytical formulae estimating the population ρ^2. The Browne formula (with ρ^2 estimated by either the Wherry formula-1 or the Olkin/Pratt formula) was the most effective estimator of the population cross-validity coefficient ρ_c^2.

Descriptive Statistics for the Bias

Means and standard deviations for the biases from the 500 replicates across the specified sampling conditions (population ρ^2, *n/p* ratio, and multicollinearity) were obtained. Because the amount of information obtained was large, these descriptive statistics are presented in Appendix F, rather than in a table in the body of the text.

From the tables in Appendix F, the biases for these analytical formulae were obvious, especially when the *n/p* ratio was relatively small. And most of the time, means and standard deviations for the biases from the analytical formulae that estimated population ρ^2 were much smaller than for those from the analytical formulae that estimated population cross-validity coefficient ρ_c^2. This indicated that the formulae estimating population ρ^2 tended to give a better estimate than those that estimated population cross-validity coefficient ρ_c^2.

For an ideal analytical formula, the means of these biases approached zero (accuracy) and the standard deviations was the smallest (stable), if it was effective in adjusting for the R^2 shrinkage in multiple regression. Across each of the sampling conditions, frequencies for each analytical formula with mean bias closest to zero and the smallest standard deviation were recorded. The total frequencies for each analytical formula were then summarized. Based on the frequency rankings obtained, the best analytical formulae with means of bias closest to zero and the smallest bias standard deviations were selected as the "recommended formulae" and summarized across different sample sizes and number of predictor variable in Tables 8 and 9. The results indicated that the Pratt formula was the best estimate among the analytical formulae estimating population ρ^2, especially when the n/p ratio was relatively small. Still the Browne formula gave the best estimate for the population cross-validity coefficient ρ_c^2 across almost all these different n/p ratio conditions.

Also based on the means and standard deviations from the sample statistical biases obtained, analytical formulae with the largest mean biases and the largest bias standard deviations were selected as the worst formulae and summarized across different sample sizes and number of predictor variables in Tables 10 and 11 . The results indicated that the Claudy-3 formula was the least effective analytical formula estimating the population ρ^2, while the Darlington formula and Lord-1 formula performed the worst in estimating the population cross-validity coefficient ρ_c^2. Cautions should be warranted in using these analytical formulae estimating statistical bias, and preferably, using the most effect analytical formulae instead.

Table 8

Recommended Formulae for Estimating Population ρ^2 across Different Sample Size (n) and Number of Predictor Variables (p)

Number	Sample size (n)				
of predictors (p)	20	40	60	100	200
2	Pratt formula	Pratt formula	Pratt formula & Claudy-3 formula	Wherry-2 formula	Smith formula & Wherry-1 formula
4	Pratt formula	Olkin/Pratt formula	Wherry-1 formula	Pratt formula	Claudy-3 formula
8	Pratt formula	Olkin/Pratt formula & Pratt formula	Pratt formula	Pratt formula	Wherry-2 formula

Table 9

Recommended Formulae for Estimating Population Cross-Validity Coefficient ρ_c^2 across Different Sample Size (n) and Number of Predictor Variables (p)

Number	Sample size (n)				
of predictors (p)	20	40	60	100	200
2	Browne-2 formula	Browne-2 formula	Browne-2 formula	Browne-2 formula	Claudy-2 formula
4	Browne-2 formula	Browne-2 formula	Browne-2 formula	Browne-2 formula	Browne-2 formula
8	Browne-1 formula & Rozeboom-2 formula	Browne-2 formula	Burket formula	Browne-2 formula	Browne-2 formula

Table 10

Worst Formulae for Estimating ρ^2 Across Different Sample Size (n) and Number of Predictor Variables (p)

Number	Sample Size (n)				
of predictors (p)	20	40	60	100	200
2	Claudy-3 formula	Claudy-3 formula	Claudy-3 formula	Claudy-3 formula	Claudy-3 formula
4	Claudy-3 formula	Claudy-3 formula	Claudy-3 formula	Claudy-3 formula	Claudy-3 formula
8	Claudy-3 formula	Wherry-1 formula	Claudy-3 formula	Claudy-3 formula	Smith formula & Wherry-1 formula

Table 11

Worst Formulae for Estimating ρ_c^2 Across Different Sample Size (n) and Number of Predictor Variables (p)

Number	Sample Size (n)				
of predictors (p)	20	40	60	100	200
2	Lord-1 formula	Lord-1 formula	Lord-1 formula	Lord-1 formula	Lord-1 formula
4	Darlington formula	Lord-1 formula	Lord-1 formula	Lord-1 formula	Lord-1 formula
8	Darlington formula	Darlington formula	Darlington formula	Lord-1 formula	Lord-1 formula

Visual Representation--Boxplots of the Estimates

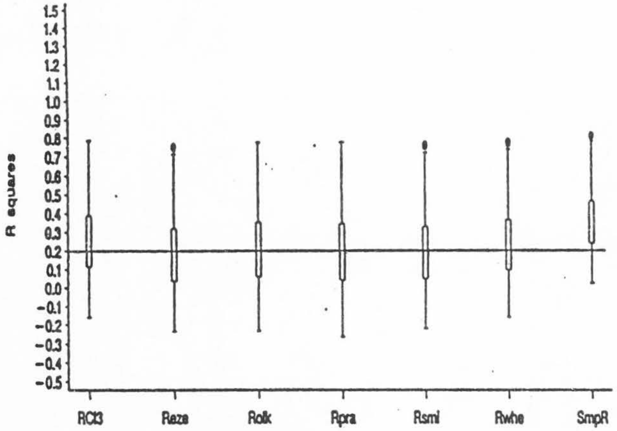
Visual representation (side-by-side modified boxplots) comparing all the analytical formulae across the 500 replicates were produced using the GPLOT option in the SAS graphic procedure. The boxplot was chosen because it provided distributional information of sample estimates. For the boxplot presented in the study, the box length equaled IQR (Interquartile Range), with the lower end equal to the first quartile (25th percentile) and the upper ends equal to the third quartile (75th percentile). The two lines (whiskers) outside the box extend to $1.5\times\text{IQR}$ beyond the quartiles, and any observations beyond the range of these whiskers were considered outliers, and were plotted as individual dots (Moore, 1993). In the boxplots shown in Figure 2, population parameters were indicated by the horizontal lines.

Estimators of the Population ρ^2

One hundred thirty-five box plots were produced for the six analytical formulae estimating population ρ^2 across different conditions of multicollinearity, population ρ^2, and *n/p* ratio. Again, for the sake of clarity, only three *n/p* ratio conditions (5, 25, and 50) are selectively presented in Figure 2. The numbers of sample size to the numbers of predictor variables are 20/4, 100/4, and 200/4, respectively.

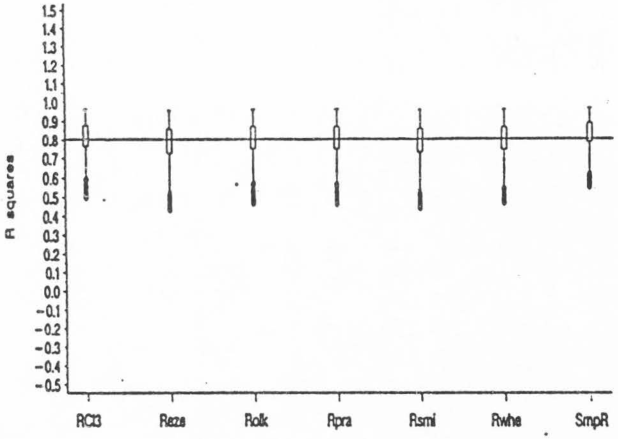
From these boxplots, it is obvious that the sample multiple R^2 was almost consistently greater than the corresponding population ρ^2. This was shown from these graphs that the third quartiles were almost always higher than the horizontal lines. Several observations were made from these boxplots:

Boxplot of Population Rsq Estimates (n=20)



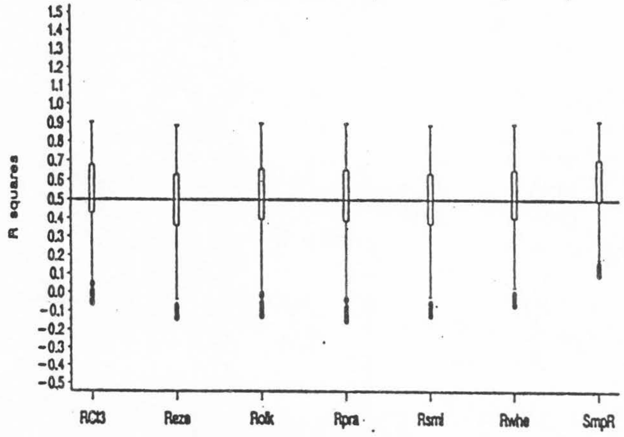
Population R square=.2; Multicollinearity r=.1; 4 predictors

Boxplot of Population Rsq Estimates (n=20)



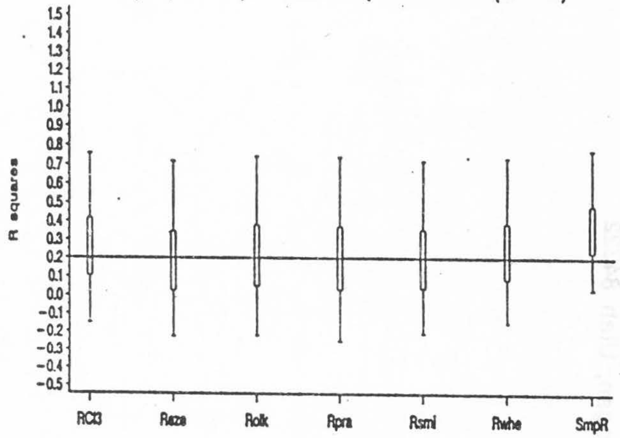
Population R square=.8; Multicollinearity r=.1; 4 predictors

Boxplot of Population Rsq Estimates (n=20)



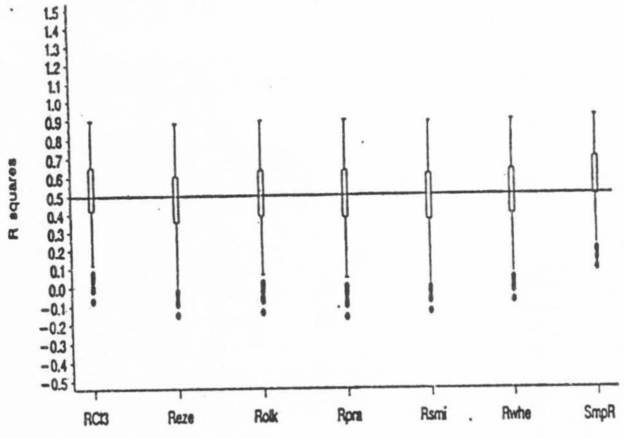
Population R square=.5; Multicollinearity r=.1; 4 predictors

Boxplot of Population Rsq Estimates (n=20)



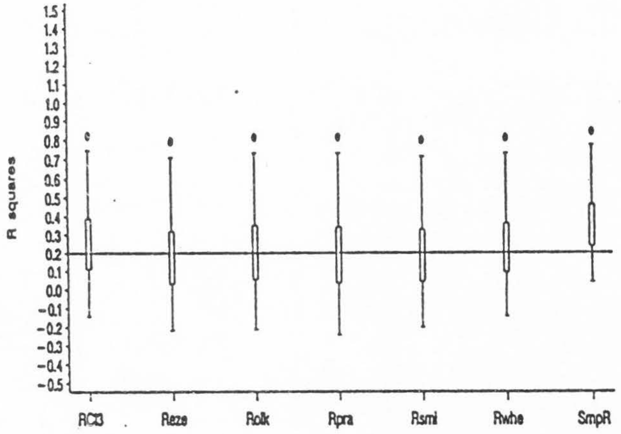
Population R square=.2; Multicollinearity r=.3; 4 predictors

Boxplot of Population Rsq Estimates (n=20)



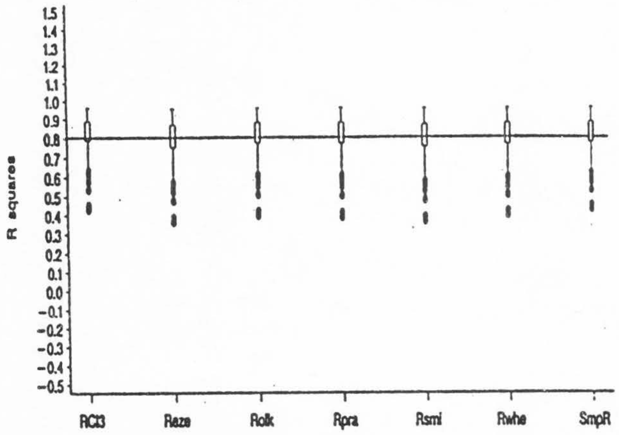
Population R square=.5; Multicollinearity r=.3; 4 predictors

Boxplot of Population Rsq Estimates (n=20)



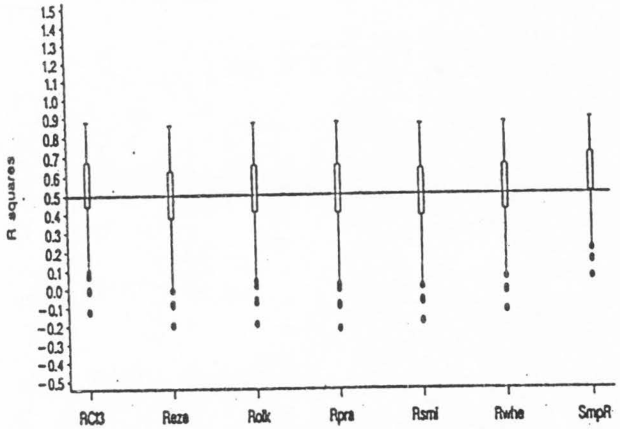
Population R square=.2; Multicollinearity r=.5; 4 predictors

Boxplot of Population Rsq Estimates (n=20)

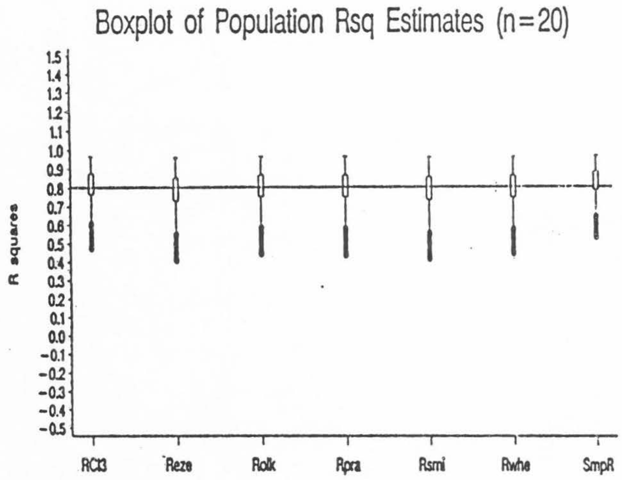


Population R square=.8; Multicollinearity r=.3; 4 predictors

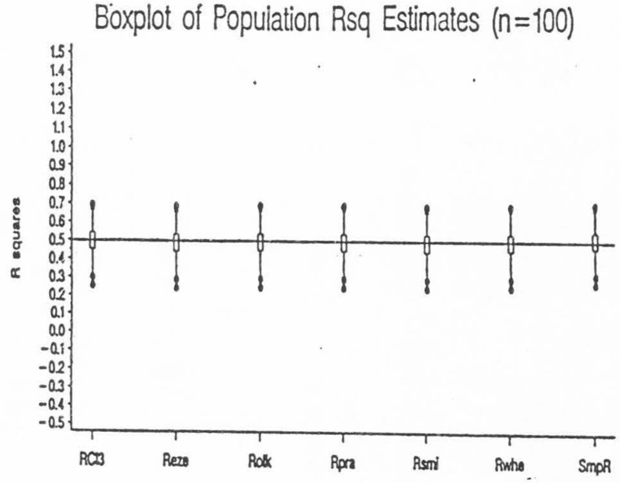
Boxplot of Population Rsq Estimates (n=20)



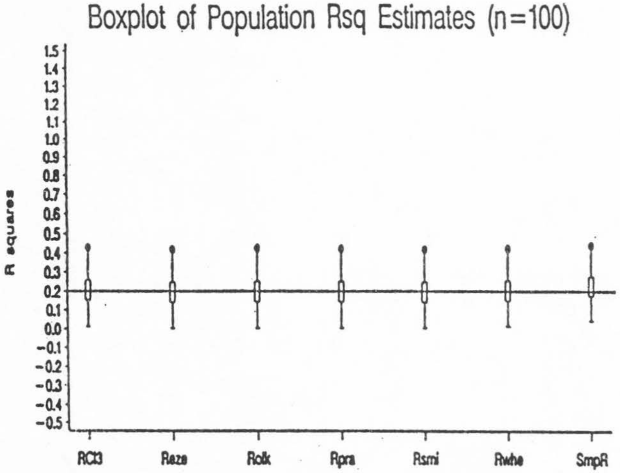
Population R square=.5; Multicollinearity r=.5; 4 predictors


Boxplot of Population Rsq Estimates (n=20)

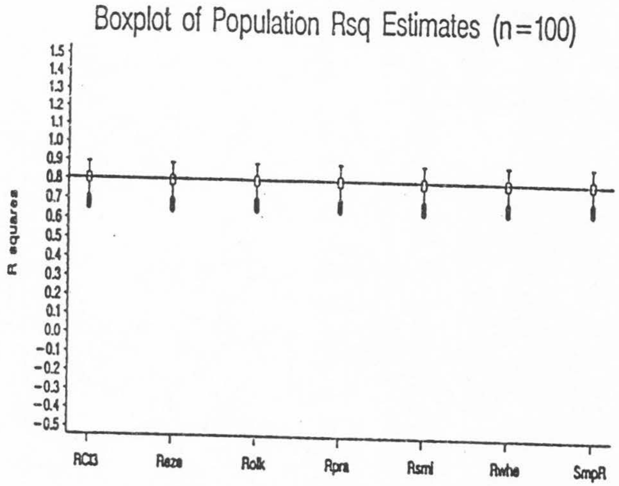
Population R square=.8; Multicollinearity r=.5; 4 predictors


Boxplot of Population Rsq Estimates (n=100)

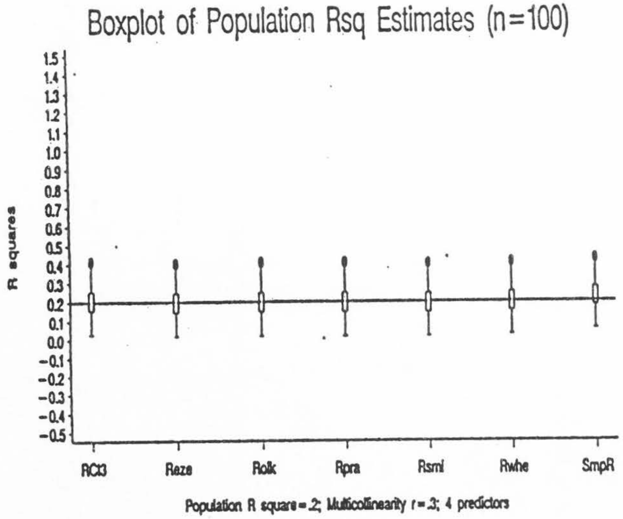
Population R square=.5; Multicollinearity r=.1; 4 predictors


Boxplot of Population Rsq Estimates (n=100)

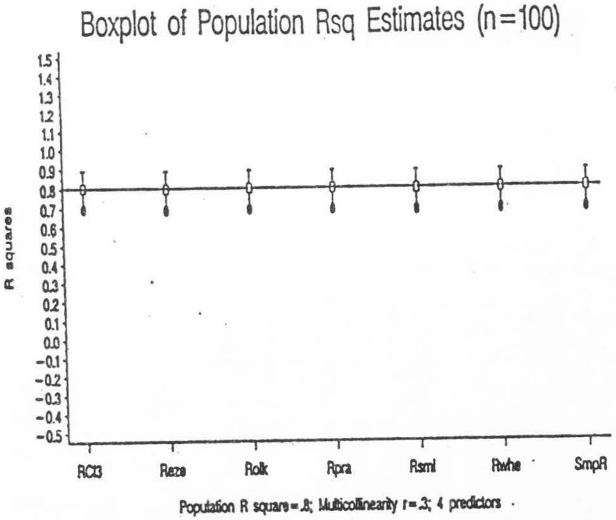
Population R square=.2; Multicollinearity r=.1; 4 predictors


Boxplot of Population Rsq Estimates (n=100)

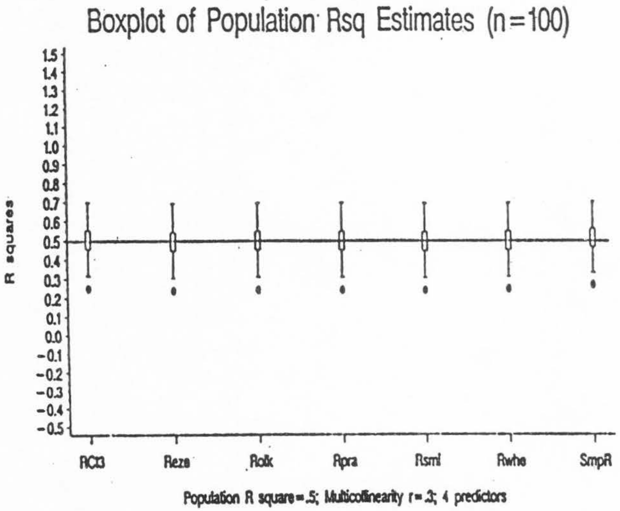
Population R square=.8; Multicollinearity r=.1; 4 predictors


Boxplot of Population Rsq Estimates (n=100)

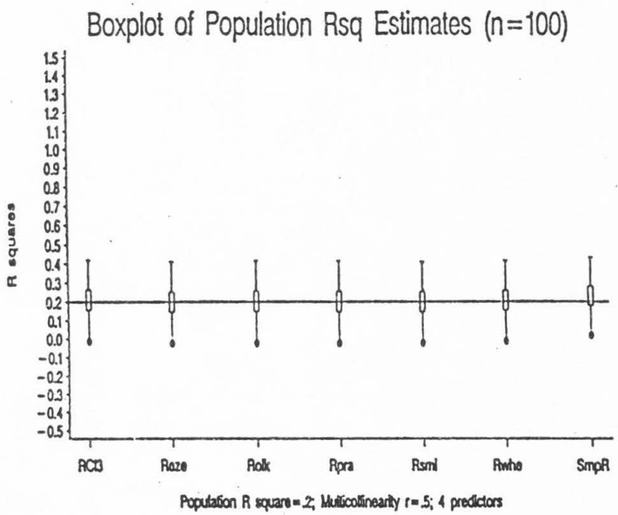
Population R square=.2; Multicollinearity r=.3; 4 predictors


Boxplot of Population Rsq Estimates (n=100)

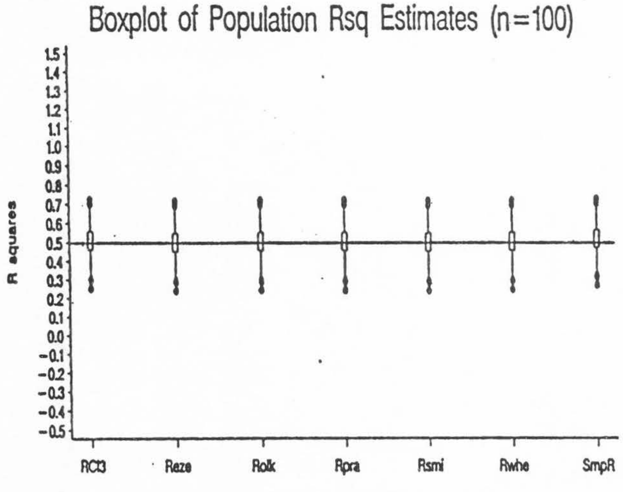
Population R square=.8; Multicollinearity r=.3; 4 predictors


Boxplot of Population Rsq Estimates (n=100)

Population R square=.5; Multicollinearity r=.3; 4 predictors

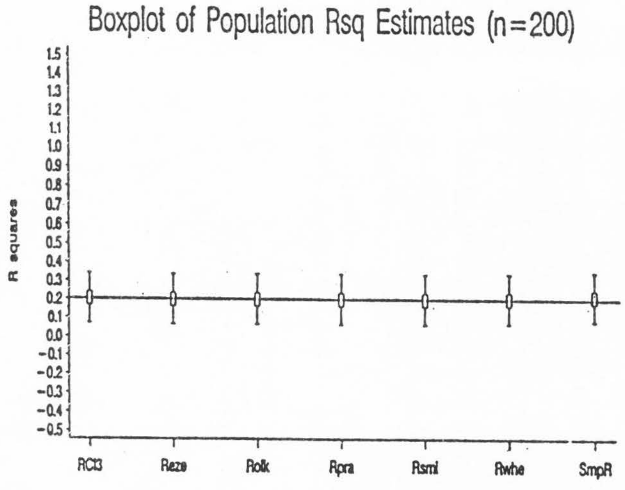

Boxplot of Population Rsq Estimates (n=100)

Population R square=.2; Multicollinearity r=.5; 4 predictors



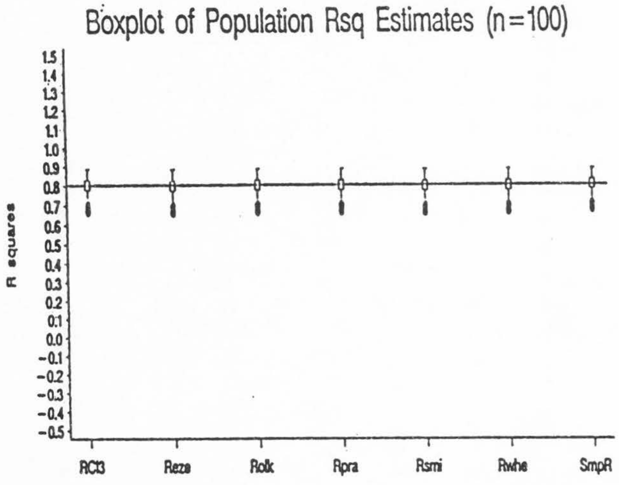
Boxplot of Population Rsq Estimates (n=100)

Population R square=.5; Multicollinearity r=.5; 4 predictors



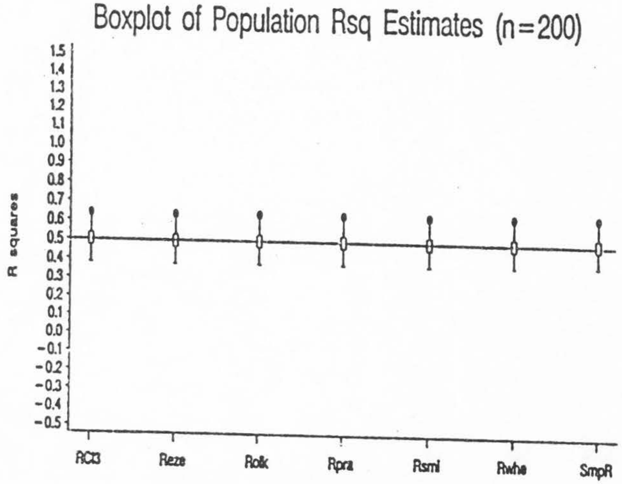
Boxplot of Population Rsq Estimates (n=200)

Population R square=.2; Multicollinearity r=.1; 4 predictors



Boxplot of Population Rsq Estimates (n=100)

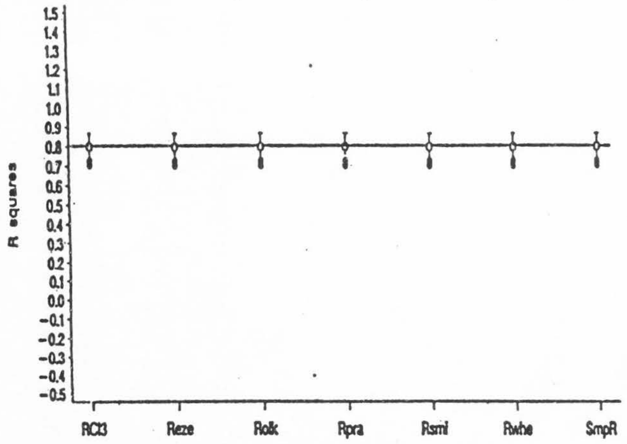
Population R square=.8; Multicollinearity r=.5; 4 predictors



Boxplot of Population Rsq Estimates (n=200)

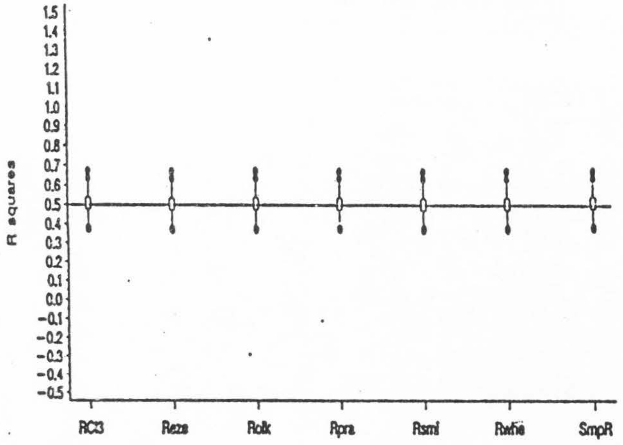
Population R square=.5; Multicollinearity r=.1; 4 predictors

Boxplot of Population Rsq Estimates (n=200)



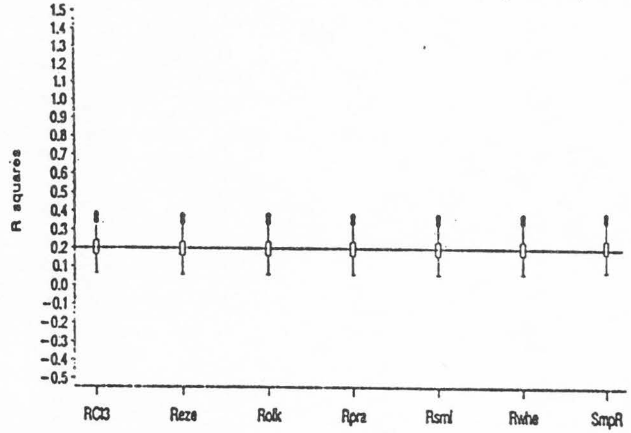
Population R square=.8; Multicollinearity r=.1; 4 predictors

Boxplot of Population Rsq Estimates (n=200)



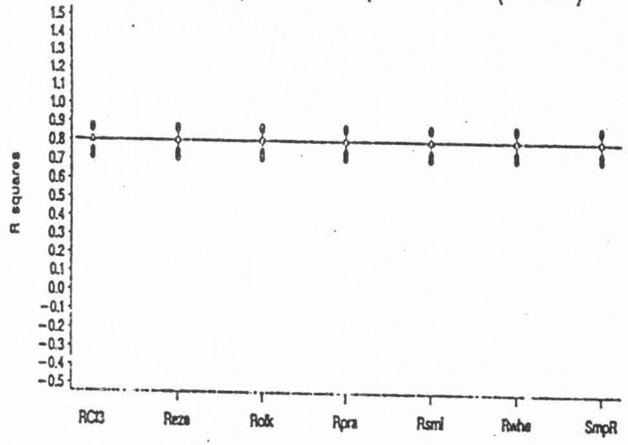
Population R square=.5; Multicollinearity r=.3; 4 predictors

Boxplot of Population Rsq Estimates (n=200)

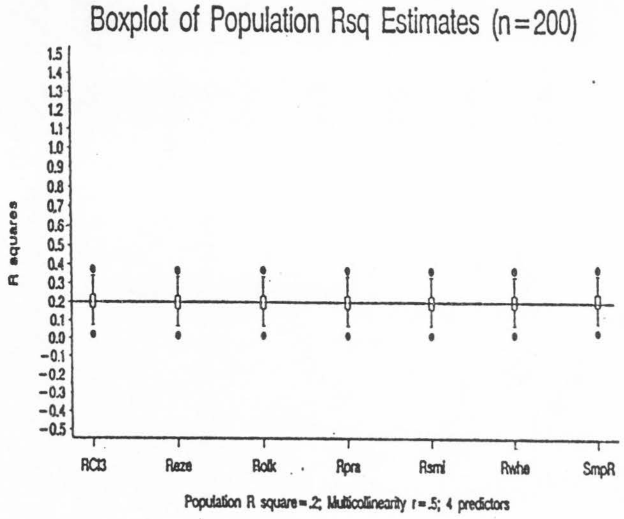


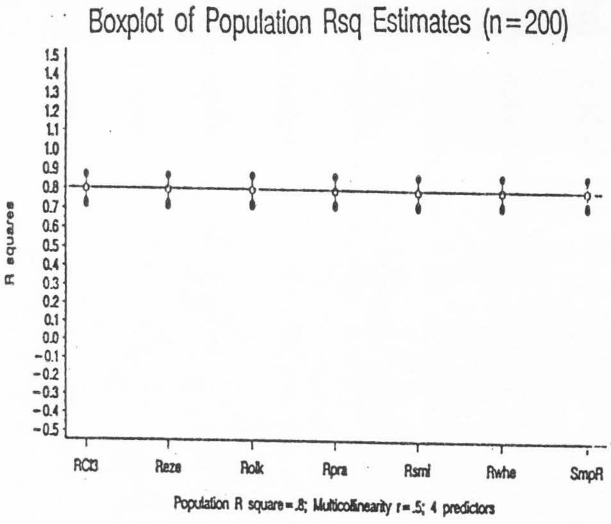
Population R square=.2; Multicollinearity r=.3; 4 predictors

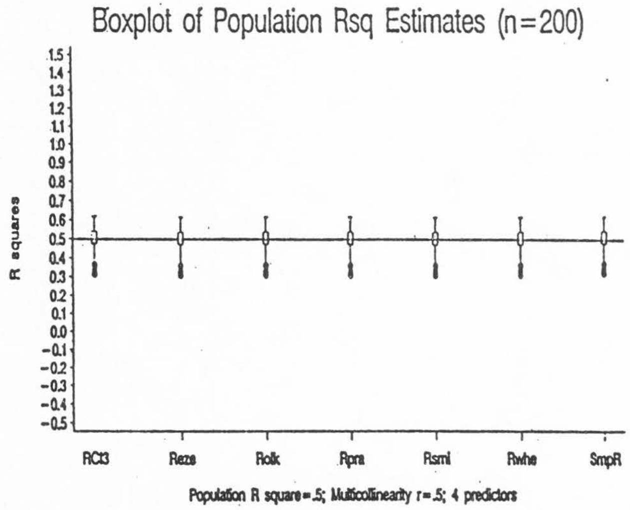
Boxplot of Population Rsq Estimates (n=200)



Population R square=.8; Multicollinearity r=.3; 4 predictors


Boxplot of Population Rsq Estimates (n=200)
R squares
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
RC13
Reze
Rolk
Rpra
Rsmi
Rwhe
SmpR
Population R square=.2; Multicollinearity r=.5; 4 predictors


Boxplot of Population Rsq Estimates (n=200)
R squares
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
RC13
Reze
Rolk
Rpra
Rsmi
Rwhe
SmpR
Population R square=.8; Multicollinearity r=.5; 4 predictors


Boxplot of Population Rsq Estimates (n=200)
R squares
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
RC13
Reze
Rolk
Rpra
Rsmi
Rwhe
SmpR
Population R square=.5; Multicollinearity r=.5; 4 predictors

1. Across the three different conditions of n/p ratio, as n/p ratio increased (from 5 to 50), the IQRs for the six estimates decreased, which indicated that their performances became more stable.

2. Across the three different conditions of population ρ^2, the performances of these formulae were comparable when ρ^2 was either small (.2) or moderate (.5). When ρ^2 was relatively large (.8), the IQRs for the six estimates were the smallest, which indicated that the performances were the most stable.

3. Across the three different conditions of multicollinearity, all the boxplots were similar in shapes, which indicated that multicollinearity did not seem to have any significant effects on the distributions of these estimates.

Estimator of the Population Cross-Validity Coefficient ρ_c^2

One hundred thirty-five modified box plots were also produced for the 10 analytical formulae estimating the population cross-validity coefficient ρ_c^2 across different conditions of multicollinearity, population ρ^2, and n/p ratio. Again, for the sake of clarity, only three n/p ratio conditions (5, 25, and 50) are selectively presented in Figure 3. The numbers of sample sizes to the numbers of predictor variables are 20/4, 100/4, and 200/4 respectively.

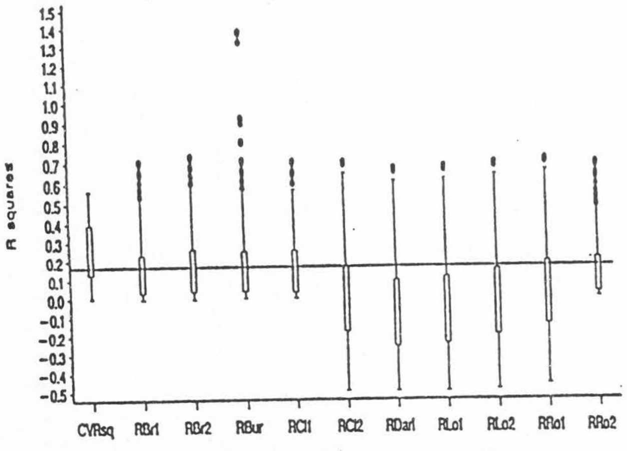
1. Across the three different conditions of n/p ratio, as the n/p ratio increased (from 5 to 50), the IQRs for the 10 estimates decreased, which indicated that their performances became more stable.

2. Across the three different conditions of population ρ^2, the performances of these formulae were comparable when ρ^2 was either small (.2) or moderate (.5). And when ρ^2 was relatively large (.8), the IQRs for the six estimates were the smallest, which indicated that the performances were the most stable.

Note that when ρ^2 was .2 and *n/p* ratio was 5, all the outliers from these distributions were located at the upper end. This indicated when the population ρ^2 and the *n/p* ratio were both relatively small, there was some tendency for these analytical formulae to overestimate the population cross-validity coefficient ρ_c^2. Among the 10 formulae, the Burket formula produced more extreme large outliers.

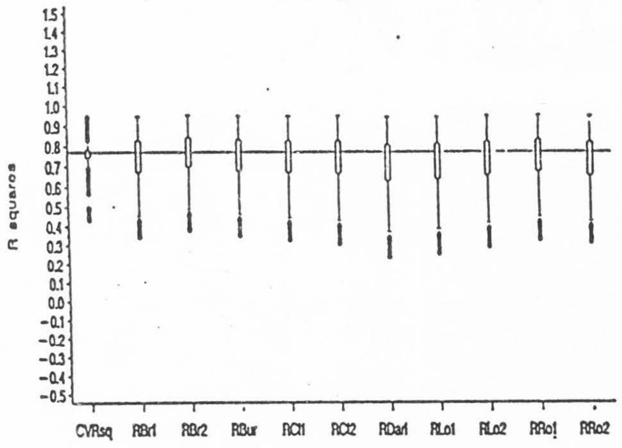
When the population ρ^2 is either .2 or .5, the 10 analytical formulae could be categorized into two groups: overestimator and underestimator of the population cross-validity coefficient ρ_c^2. For the overestimators, these formulae tended to have more large positive outliers; the upper whiskers were longer than the lower whiskers, and usually the 75th percentiles were also larger than the population ρ_c^2. The Browne formula, the Burket formula, and the Claudy formula-1 all belong to this category. For the underestimators, these formulae tended to have more large negative outliers; the lower whiskers were sometimes longer than the upper whiskers, and sometimes the 75th percentiles are smaller than the population ρ_c^2. The Claudy formula-2, the Darlington formula, the Lord formula-1 and -2, and the Rozeboom formula-1 and -2 all belong to this category. Such a distinction cannot be clearly made when the population ρ^2 is relatively large (.8).

Boxplot of CrossValidity Estimates (n=20)



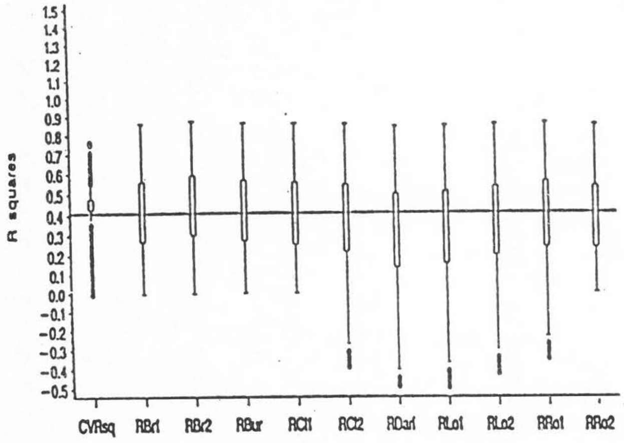
Population R square=.2; Multicollinearity r=.1; 4 predictors

Boxplot of CrossValidity Estimates (n=20)



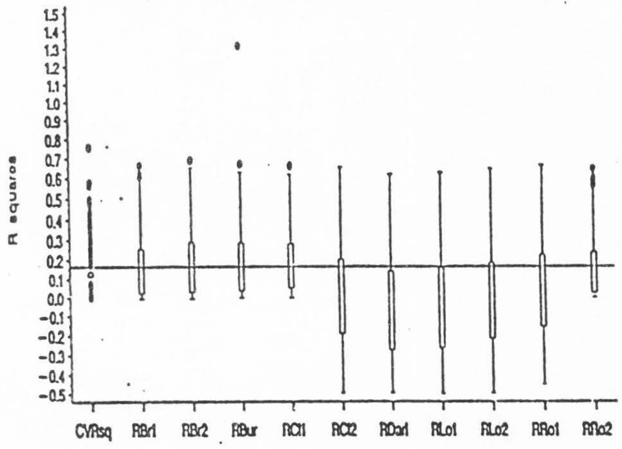
Population R square=.8; Multicollinearity r=.1; 4 predictors

Boxplot of CrossValidity Estimates (n=20)

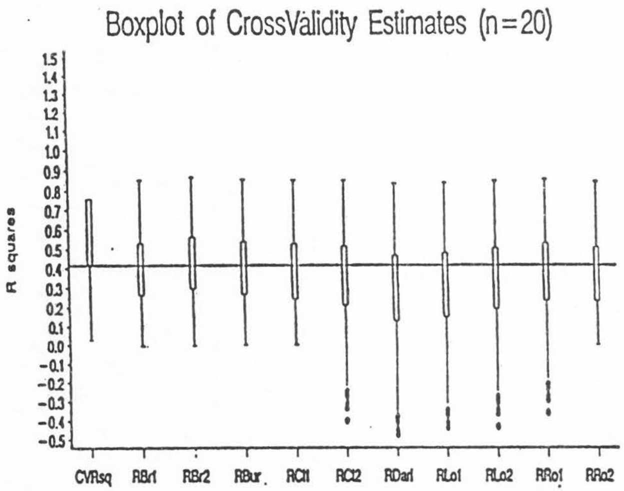


Population R square=.5; Multicollinearity r=.1; 4 predictors

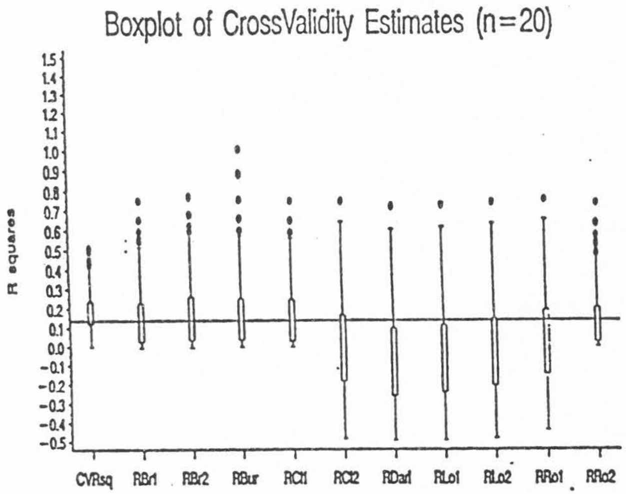
Boxplot of CrossValidity Estimates (n=20)



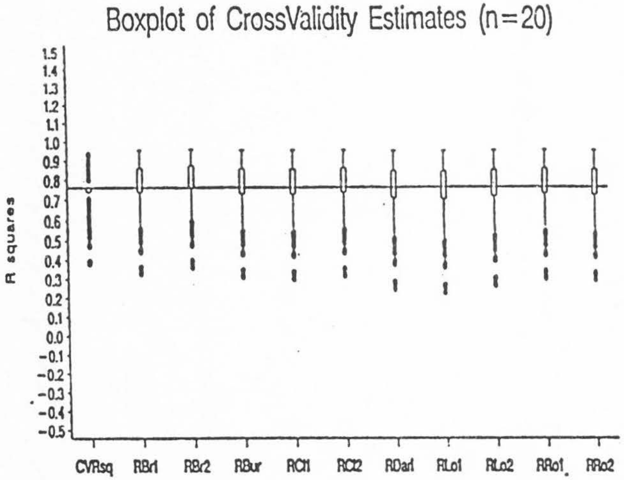
Population R square=.2; Multicollinearity r=.3; 4 predictors



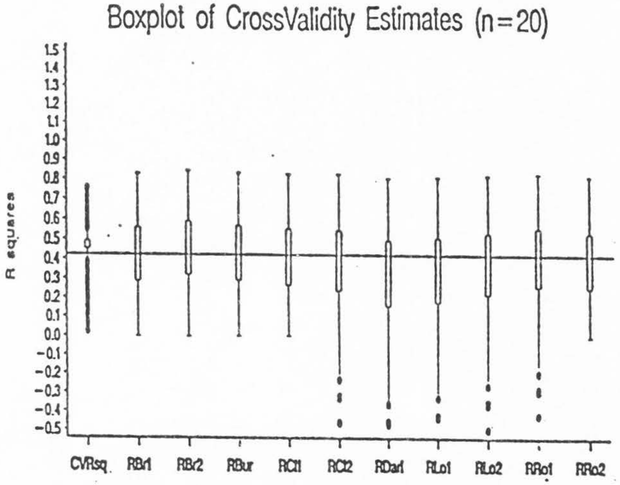
Population R square=.5; Multicollinearity r=.3; 4 predictors



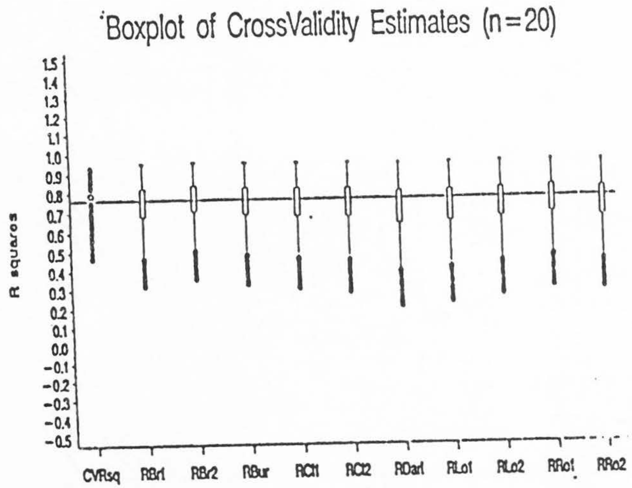
Population R square=.2; Multicollinearity r=.5; 4 predictors



Population R square=.8; Multicollinearity r=.3; 4 predictors

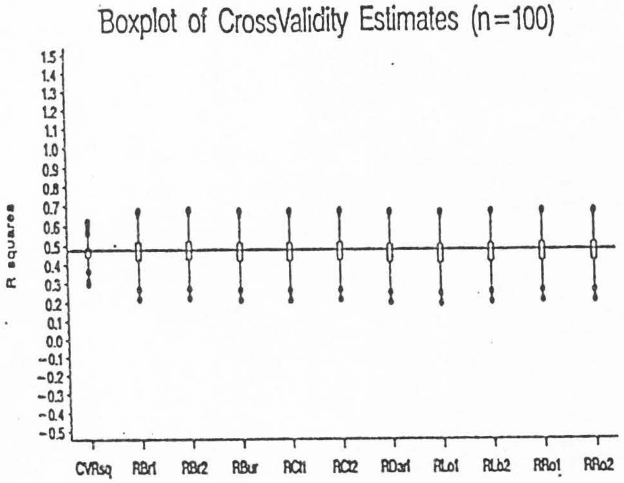


Population R square=.5; Multicollinearity r=.5; 4 predictors



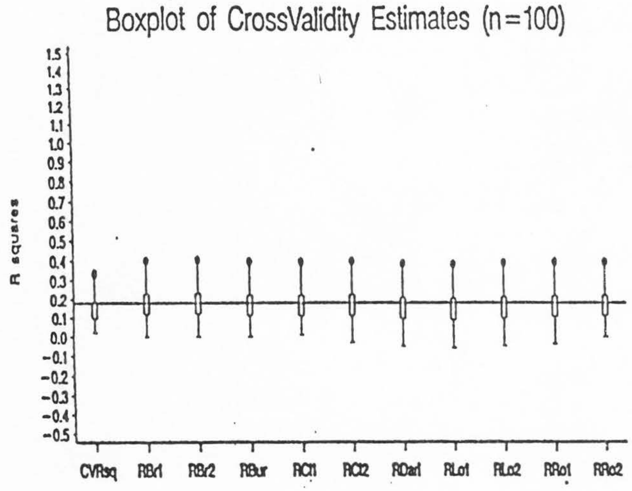
Boxplot of CrossValidity Estimates (n=20)

Population R square=.8; Multicollinearity r=.5; 4 predictors



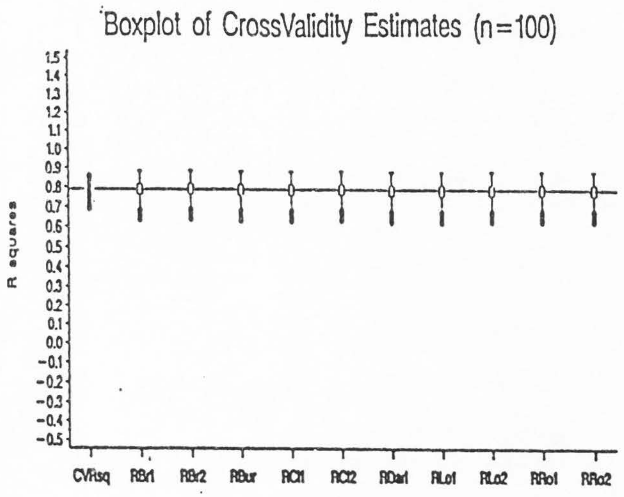
Boxplot of CrossValidity Estimates (n=100)

Population R square=.5; Multicollinearity r=.1; 4 predictors



Boxplot of CrossValidity Estimates (n=100)

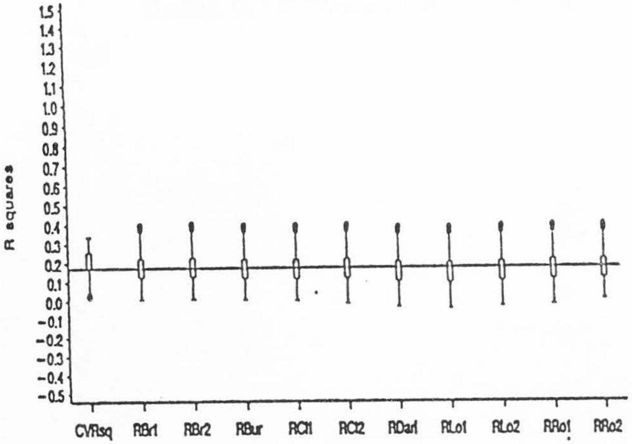
Population R square=.2; Multicollinearity r=.1; 4 predictors



Boxplot of CrossValidity Estimates (n=100)

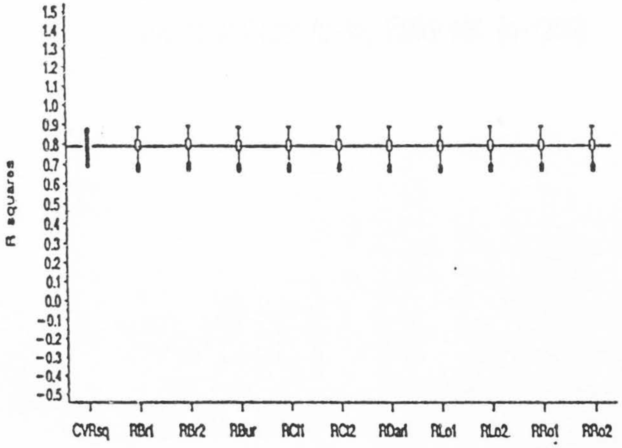
Population R square=.8; Multicollinearity r=.1; 4 predictors

Boxplot of CrossValidity Estimates (n=100)



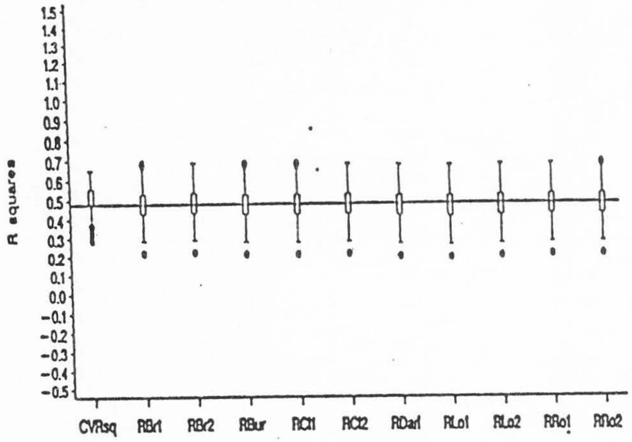
Population R square=.2; Multicollinearity r=.3; 4 predictors

Boxplot of CrossValidity Estimates (n=100)



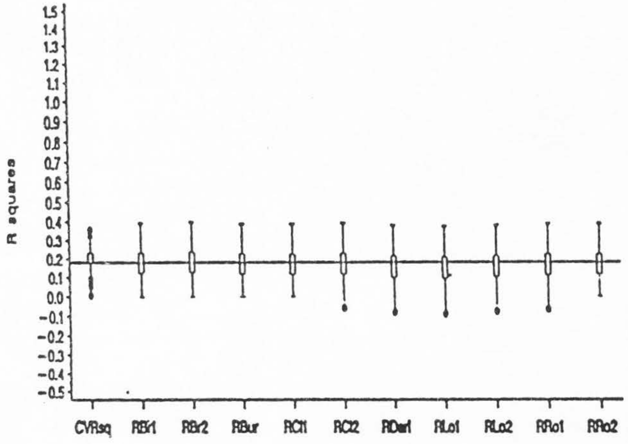
Population R square=.8; Multicollinearity r=.3; 4 predictors

Boxplot of CrossValidity Estimates (n=100)



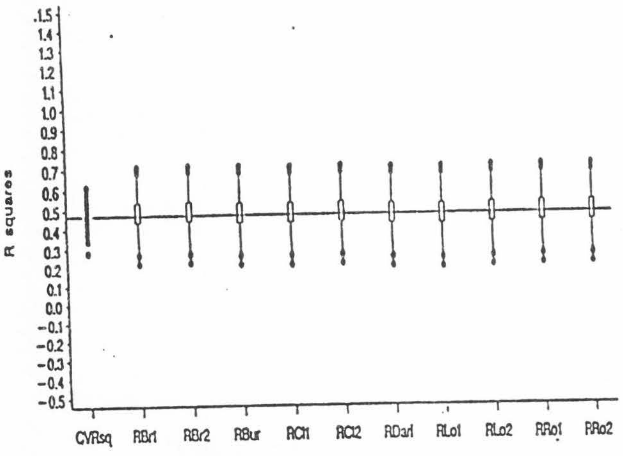
Population R square=.5; Multicollinearity r=.3; 4 predictors

Boxplot of CrossValidity Estimates (n=100)



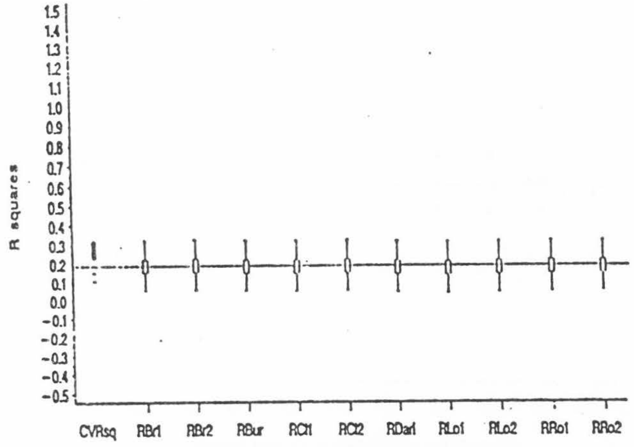
Population R square=.2; Multicollinearity r=.5; 4 predictors

Boxplot of CrossValidity Estimates (n=100)



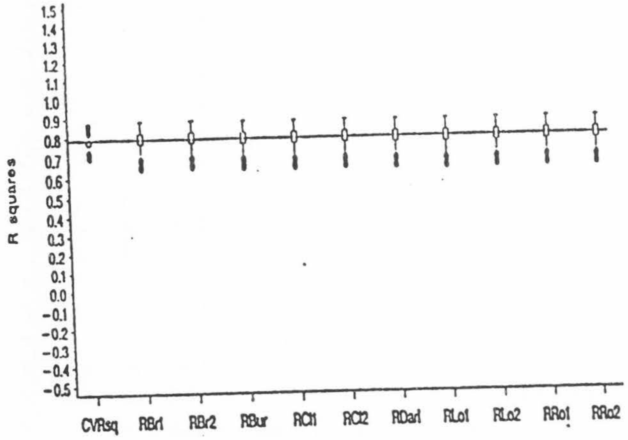
Population R square=.5; Multicollinearity r=.5; 4 predictors

Boxplot of CrossValidity Estimates (n=200)



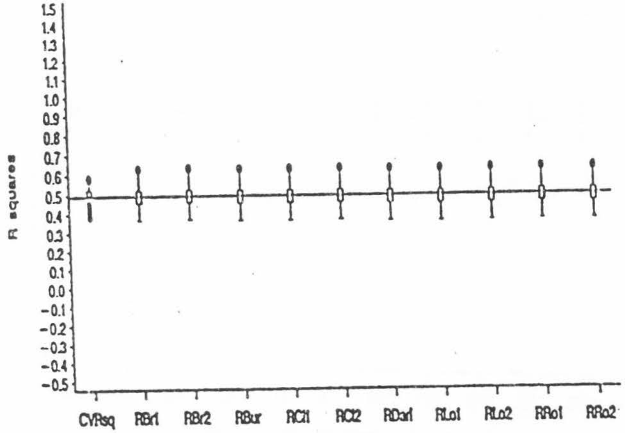
Population R square=.2; Multicollinearity r=.1; 4 predictors

Boxplot of CrossValidity Estimates (n=100)

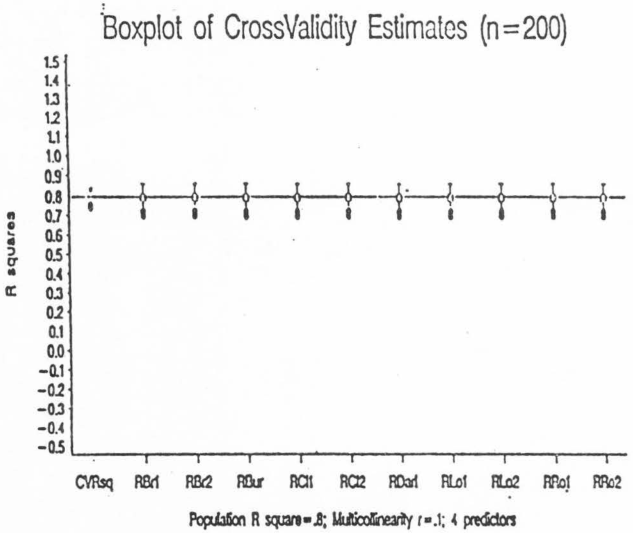


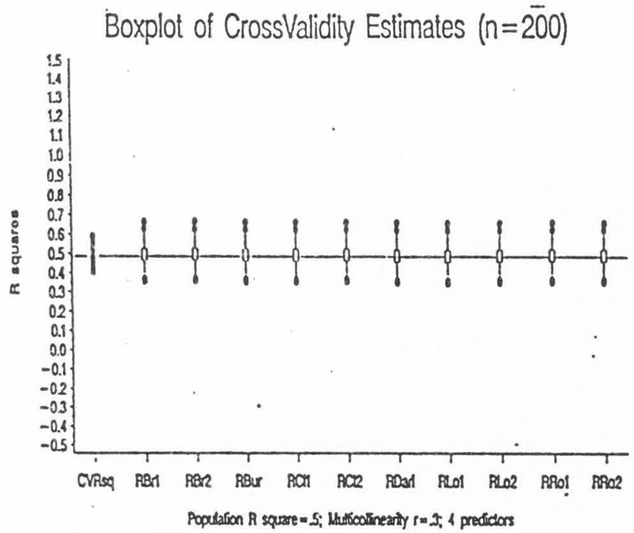
Population R square=.8; Multicollinearity r=.5; 4 predictors

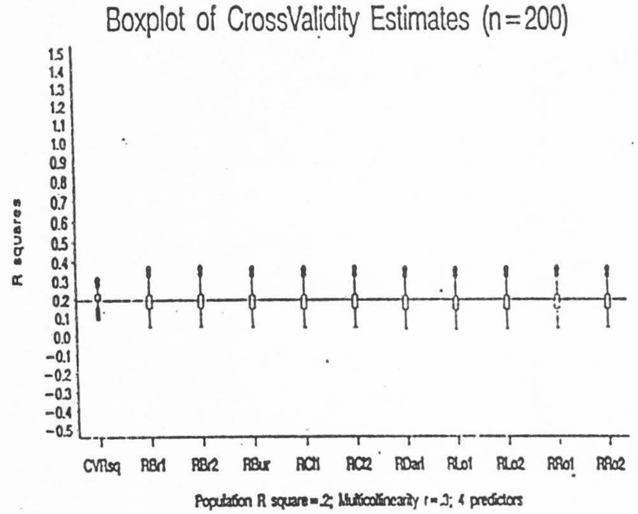
Boxplot of CrossValidity Estimates (n=200)

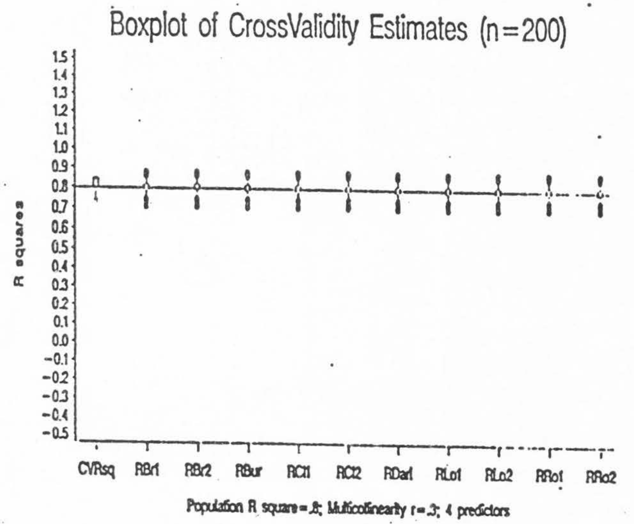


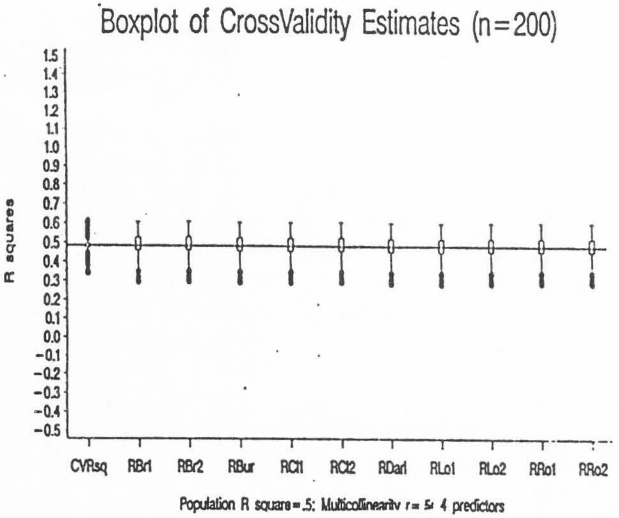
Population R square=.5; Multicollinearity r=.1; 4 predictors

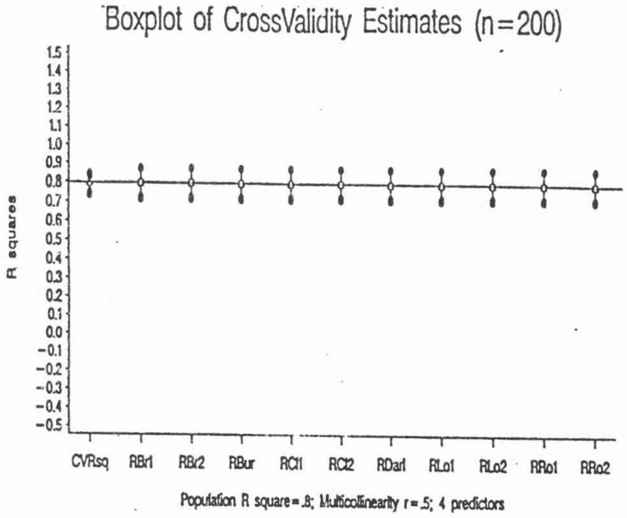

Boxplot of CrossValidity Estimates (n=200)
R squares
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
CVRsq
RBr1
RBr2
RBur
RCl1
RCl2
RDar1
RLo1
RLo2
RRo1
RRo2
Population R square=.8; Multicollinearity r=.1; 4 predictors

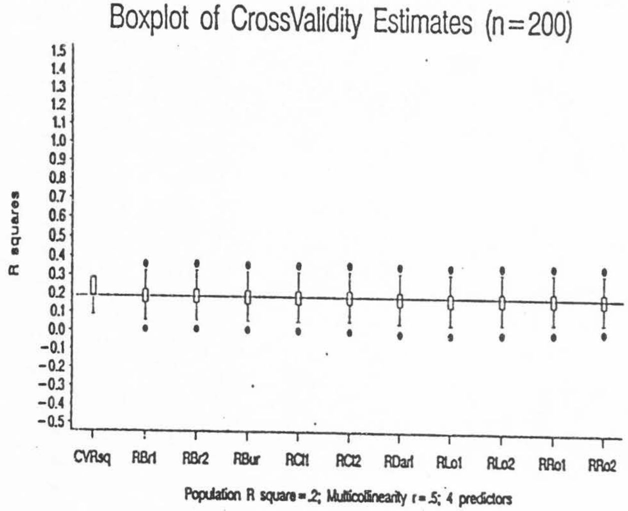

Boxplot of CrossValidity Estimates (n=200)
R squares
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
CVRsq
RBr1
RBr2
RBur
RCl1
RCl2
RDar1
RLo1
RLo2
RRo1
RRo2
Population R square=.5; Multicollinearity r=.3; 4 predictors


Boxplot of CrossValidity Estimates (n=200)
R squares
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
CVRsq
RBr1
RBr2
RBur
RCl1
RCl2
RDar1
RLo1
RLo2
RRo1
RRo2
Population R square=.2; Multicollinearity r=.3; 4 predictors


Boxplot of CrossValidity Estimates (n=200)
R squares
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
CVRsq
RBr1
RBr2
RBur
RCl1
RCl2
RDar1
RLo1
RLo2
RRo1
RRo2
Population R square=.8; Multicollinearity r=.3; 4 predictors


Boxplot of CrossValidity Estimates (n=200)
R squares
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
CVRsq
RBr1
RBr2
RBur
RCl1
RCl2
RDarl
RLo1
RLo2
RRo1
RRo2
Population R square=.5; Multicollinearity r=.5; 4 predictors


Boxplot of CrossValidity Estimates (n=200)
R squares
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
CVRsq
RBr1
RBr2
RBur
RCl1
RCl2
RDarl
RLo1
RLo2
RRo1
RRo2
Population R square=.8; Multicollinearity r=.5; 4 predictors


Boxplot of CrossValidity Estimates (n=200)
R squares
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
CVRsq
RBr1
RBr2
RBur
RCl1
RCl2
RDarl
RLo1
RLo2
RRo1
RRo2
Population R square=.2; Multicollinearity r=.5; 4 predictors

3. Across the three different conditions of multicollinearity, the distributions for the 10 analytical formulae across different multicollinearity conditions were similar in shapes, which indicated that multicollinearity did not seem to have any dramatic effects on the performances of these estimates.

Explaining the Variations of Sample Estimate Biases

Bias is defined as the difference between the corrected R^2 obtained by applying each analytical formula to the sample and the population parameters; that is, the population ρ^2 or the population cross-validity coefficient ρ_c^2. For the six analytical formulae designed for estimating the population ρ^2, biases were calculated by subtracting the prespecified population ρ^2 (.2, .5, .8) from the corrected R^2 obtained from each formula. And for the 10 analytical formulae designed for estimating the population cross-validity coefficient ρ_c^2, biases were calculated by subtracting the estimated population cross-validity coefficient ρ_c^2 from the corrected R_c^2.

Factors that might have influenced the biases of these analytical formulae were investigated using the analysis of variance (ANOVA) model to partition the variances of sample estimated biases to different sources. These factors included: sample size, population ρ^2, degree of multicollinearity among the predictor variables, and number of predictor variables. The two-way, three-way, and four-way interactions among these factors were also considered potential sources in the analysis. Tables 12 and 13 present the results of partitioning the variance of different sources of sample estimate biases for the analytical formulae for either the population ρ^2 or for the population cross-validity

Table 12

Eta-Squares for Different Sources of Variance for the Analytical Formulae Estimating Population ρ^2

Source	Smith	Wherry-1	Wherry-2	Olkin\Pratt	Pratt	Claudy-3
Sample size	.061	.284	.335	.0602	.002	1.322
Population ρ^2	.029	.010	.435	.0185	.004	.475
Sample size ×population ρ^2	.066	.019	.364	.0482	.001	.410
Multicollinearity	.000	.000	.000	.0003	.000	.000
Sample size ×multicollinearity	.020	.019	.020	.0191	.019	.020
Population ρ^2 ×multicollinearity	.004	.004	.003	.0034	.003	.003
Sample size ×population ρ^2 ×multicollinearity	.031	.031	.030	0302	.030	.029
Number of predictors (p)	.022	.001	.021	.0008	.003	.001
Sample size × p	.062	.003	.059	0027	.005	002
Population ρ^2 × p	.003	.006	.004	.0056	.009	.006
Sample size ×population ρ^2 × p	.019	.014	.019	0138	.018	014
Multicollinearity × p	.002	.002	.001	.0016	.002	.001
Sample size ×multicollinearity × p	.018	.018	.016	.0181	.018	.018
Population ρ^2 ×multicollinearity× p	.011	.011	.011	.0109	.011	.011
Sample size ×population ρ^2 ×multicollinearity × p	.046	.046	.046	.0457	.046	.045

Table 13

Eta-Squares for Different Sources of Variance for Analytical Formulae Estimating Population Cross-Validity Coefficient ρ_c^2

Source	Lord-1	Lord-2	Burket	Darlington	Browne[a]	Browne[b]	Claudy1	Claudy2	Rozeboom1	Rozeboom2
Sample size	6.593	3.991	.018	1.183	.049	.163	.069	6.400	1.634	.373
Population ρ^2	3.747	2.737	.005	3.660	.005	.006	.042	2.073	1.761	.049
Sample size × population ρ^2	3.300	2.544	.024	4.522	.046	.043	.165	3.159	1.732	.128
Multicollinearity	.001	.001	.003	.001	.004	.004	.004	.001	.001	.003
Sample size × multicollinearity	.024	.026	.012	.016	.048	.047	.060	.020	.029	.048
Population ρ^2 × multicollinearity	.005	.005	.007	.003	.009	.009	.009	.004	.005	.009
Sample size × population ρ^2 × Multicollinearity	.048	.051	.029	.034	.074	.073	.077	.040	.055	.074
Number of predictors (p)	.182	.154	.002	2.140	.026	.027	.271	2.083	.070	.039
Sample size × p	.191	.150	.007	4.186	.081	.079	.619	3.961	.041	.090
Population ρ^2 × p	.561	.585	.007	1.297	.016	.008	.014	1.407	.553	.033
Sample size × population ρ^2 × p	.777	.802	.030	2.322	.031	.023	.039	2.457	.711	.041
Multicollinearity × p	.009	.009	.006	.007	.011	.012	.010	.008	.010	.010
Sample size × multicollinearity × p	.027	.029	.025	.022	.030	.031	.026	.025	.031	.028
Population ρ^2 × multicollinearity × p	.012	.013	.014	.009	.017	.017	.016	.011	.014	.016
Sample size × population ρ^2 × Multicollinearity × p	.055	.059	.057	.040	.082	.081	.083	.047	.064	.082

[a] The Browne formula with ρ^2 estimated by the Wherry formula-1.

[b] The Browne formula with ρ^2 estimated by the Olkin\Pratt formula.

coefficient ρ_c^2. In the tables, eta-square was used as the percentage of variation accounted for by a source, and the eta-square was obtained through:

$$\eta^2 = [(\text{sum of squares due to a source})/(\text{toal sum of squares})] \times 100$$

From these tables, the amount that each source explained ranged from nearly zero (.0002%) to 6.59% of the total variance in the sample estimate biases obtained from these analytical formulae. The overall small amount of variance accounted for by different sources in the model indicated that the variation of sample estimate biases was mainly due to random variation. Factor(s) or interactions that accounted for less than .1% of the total variance were omitted from discussion because of the insignificant amount of variance explained by these factors in the model.

Sample Size

Sample size contributed the most to the variation of three out of six analytical formulae estimating population ρ^2, and 5 out of the 10 analytical formulae estimating the population cross-validity coefficient ρ_c^2. The proportion of variance accounted for by sample size ranged from .06% to 1.32% for the formulae estimating population ρ^2, and from .16% to 6.59% for the formulae estimating the population cross-validity ρ_c^2. It appeared that sample size might be of some importance to the total variation for the bias obtained from the Wherry formula-2, the Claudy formula-3, the Lord formula-1 and -2, the Browne formula (with ρ^2 estimated by the Olkin\Pratt formula), the Claudy formula-2, and the Rozeboom formula-2.

Population ρ^2

Variances accounted for by the population ρ^2 ranged from .004% to .47% for the formulae estimating population ρ^2, and from .005% to 3.75% for the formulae estimating the population ρ_c^2. Among all the sources of variation, this factor contributed the most to the total variation for the Rozeboom formula-1 (1.76%) that estimated population cross-validity ρ_c^2.

The Interaction Between Sample Size and the Population ρ^2

Variances accounted for by the interaction term between sample size and the population ρ^2 ranged from .01% to .41% for the formulae estimating population ρ^2, and from .04% to 4.52% for the formulae estimating the population ρ_c^2. Among all the sources of variation, this interaction term contributed the most to the total variation for the Smith formula (.07%) and to the Darlington formula (4.52%). For the six formulae estimating population ρ^2, it accounted for less than .4% of the total variation. For the Rozeboom-1 and Rozeboom-2 formula that estimated population cross-validity coefficient ρ_c^2, it contributed the second most to the total variation (1.73% and .13%, respectively).

Number of Predictors (p)

The variances explained by the number of predictors accounted for less than .03% of the total variation for the six analytical formulae estimating population ρ^2. For the 10 analytical formulae estimating population cross-validity coefficient ρ_c^2, the variance accounted for by this factor ranged from .002% to 2.14% of the total variance. However, the amount of variation this factor explained was relatively small among all the other sources.

The Interaction Between Sample Size and Number of Predictors (p)

The variances explained by the interaction term between sample size and number of predictors accounted for less than .06% of the total variation for the six analytical formulae estimating population ρ^2, and it ranged from .007% to 4.2% for the 10 analytical formulae estimating population cross-validity coefficient ρ_c^2. The interaction term accounted the most for the total variation for the Claudy formula-1 (.62%), and the second most for the Darlington formula (4.19%) and the Claudy formula-2 (3.96%).

The Interaction Between Sample Size, the Population ρ^2, and Number of Predictors (p)

The three-way interaction term explained less than .02% of the total variance for the analytical formulae estimating population ρ^2, and it ranged from .02% to 2.46% for the analytical formulae estimating the population cross-validity coefficient ρ_c^2. The effect of this interaction term might be more related to the analytical formulae estimating the population cross-validity coefficient ρ_c^2 than for those estimating the population ρ^2. However, the overall percentage for this interaction term was relatively small, and no definite conclusion can be drawn from the results.

CHAPTER V

CONCLUSIONS

When estimating R^2 shrinkage in multiple regression, there is considerable confusion and little consensus in the literature about which analytical formula should be utilized under what circumstances. The present study utilized a Monte Carlo simulation to generate correlated multivariate random data, and investigated the effectiveness of various analytical formulae designed to estimate R^2 shrinkage in multiple regression under the influence of commonly encountered confounding factors such as different degrees of multicollinearity among the predictor variables, population squared multiple correlation conditions, number of predictors, and sample sizes. Five hundred replicates were simulated within each cell of the sampling conditions. Then analytical formulae were applied to the simulated data in each sampling condition, and the adjusted R^2s and R_c^2s were obtained and then compared to their corresponding population parameters (ρ^2 and ρ_c^2).

Discussion for Objective 1

The first objective of the study was to compare the accuracy and usefulness of various analytical formulae for estimating the population ρ^2 in the population from which the sample was drawn. Among the six analytical formulae designed to estimate the population ρ^2, the performances of the Pratt formula were found to be the most stable and satisfactory, especially when n/p ratio is relatively small. When n/p ratio was relatively large (e.g., 100), almost all of the six analytical formulae gave unbiased estimates across all these sampling conditions. The commonly known Wherry formula (the Wherry-2 formula in the present study), which is also the currently used "shrinkage formula" in both SAS and SPSS, only performed as well as other analytical formulae when the *n/p* ratio was relatively large (e.g., 100). Small *n/p* ratio is not uncommon in social and behavioral researches. The results indicated that it might need more consideration in choosing the most effective shrinkage formula for estimating R^2 shrinkage in multiple regression analysis, especially when there were relatively large numbers of predictor variables, and at the same time the sample size was relatively small. Practically, all these analytical formulae were relatively easy to calculate and straightforward to apply.

Discussion for Objective 2

The second objective of this study was to compare the accuracy and usefulness of various analytical formulae for estimating R^2 shrinkage for cross-validation purpose in multiple regression. Among the 10 analytical formulae designed to estimate the population cross-validity coefficient ρ_c^2, the Browne formula (with ρ^2 estimated either by the Olkin\Pratt formula or the Wherry formula-1) gave the best and most stable estimate across different conditions of population ρ^2, multicollinearity, and *n/p* ratio. Biases obtained from the Browne formula with ρ^2 estimated by Olkin/Pratt formula were slightly less than the Browne formula with ρ^2 estimated by the Wherry formula-1. When *n/p* ratio was relatively small or moderate, the Browne formula with ρ^2 estimated by the Wherry formula-1 gave a slightly better estimate than the Browne formula with ρ^2 estimated by the

Olkin/Pratt formula. Such results supported the conclusions from studies by Schmitt (1982) and Kromrey and Hines (1996), that the Browne formula was the most appropriate estimator of ρ_c^2. When *n/p* ratio was relatively large (e.g., 100), more analytical formulae gave unbiased estimates across all these sampling conditions.

To calculate some of these analytical formulae (the Browne formula, the Claudy formula-1, and the Rozeboom formula-2), two steps were needed, because there are requirements for obtaining the population ρ or ρ^2 first. However, the overall application of these analytical formulae was also relatively simple and straightforward.

Discussion for Objective 3

The third objective of the study was to assess the effects of sample size, number of predictor variables, and degree of multicollinearity among the predictors on the accuracy and variability of the performances of the analytical formulae in estimating R^2 shrinkage in multiple regression. The results suggested that *n/p* ratio, instead of either the number of predictors or the sample size alone, was the most influential factor that affected the performance of these analytical formulae. Both the accuracy and stability of these adjusted *R*s increased as *n/p* ratio increased, especially when *n/p* ratio was relatively large (e.g., 100). Most of these analytical formulae give unbiased estimates across all these sampling conditions.

Variance partitioning was performed for the sample biases obtained from these analytical formulae based on the factors considered in the study (e.g., sample size, number of predictors, degree of multicollinearity, and population ρ^2). Although sample size seemed to be the most important factor in explaining the variation in the sample biases for most of these analytical formulae, the amount of variance accounted for by all these factors was relatively small, and thus no definite conclusion can be drawn from the results. However, for those analytical formulae that performed relatively well across different sampling conditions (e.g., the Olkin/Pratt formula and the Browne formula), the amount of variation each factor accounted for was much smaller than for those analytical formulae where performance was not very satisfactory (e.g., the Lord formula-1, the Lord formula-2, and the Darlington formula). It could be inferred that the performances of those analytical formulae such as the Lord-1 and -2 formula might indeed be related, to some degree, to the confounding factors investigated in the study. Nevertheless, the greatest amount of variation accounted for by any factor was only 6.59%. Random error appeared to account for the majority of the fluctuation in the performances of these analytical formulae. Results from both boxplots and variance partitioning analysis indicated that multicollinearity did not seem to play an important role in affecting the performances of these analytical formulae in the present study.

Study Limitations

One limitation of the present study is that only multivariate normal data were generated and analyzed, which might have simplified the usually nonnormal and more complex distributions that researchers usually expect from real data. In the future, generating multivariate nonnormal distributions may provide data that could be more representative of real research data. Another limitation about the data generation design is that only three of the simplest conditions of multicollinearity were simulated. Also, all possible correlations among independent variables were assumed to be equal. With real data, different degrees of correlations among different independent variables are more likely to be expected. In future studies, a more complex multicollinearity pattern may provide researchers with a better understanding of the influence of multicollinearity on the performance of these analytical methods. Besides, only three types of population ρ^2 were generated in this study, which might only represent part of what might be expected from the real data. Also the fixed linear regression model was used in the present study. As it is known, the assumptions of the fixed linear regression model usually cannot be met completely. In the future, more complex regression models will be useful in handling distributions for which these assumptions are not met, and providing researchers with more insights when working with real data. Another approach to deal with this issue is to replicate the study under different situations in which these assumptions are violated, and to investigate the robustness of the fix linear regression model under these conditions.

Another limitation of this present study is that only analytical methods are investigated in estimating R^2 shrinkage in multiple regression analysis due to time limit and project manageability. A comparative study of both the empirical and analytical methods will provide more comprehensive and complete information on all the available methods for estimating R^2 shrinkage in multiple regression. Further replications on both real and simulated data are still needed to investigate the effectiveness of these analytical formulae.

Recommendations for Applications in Social and Behavioral Sciences

Studies of relationships among variables are common in social and behavior sciences. Psychologists, educational researchers, and sociologists have been using multiple regression extensively to answer different research questions about relationships, with the ease and availability brought by the popular statistical software (e.g., SPSS and SAS; Cohen & Cohen, 1983; Huberty & Mourad, 1980). As mentioned earlier, there are two major reasons to apply the multiple regression procedure: to estimate the population multiple correlation coefficient from a sample, or to predict the same dependent variable for new samples from the same population but other than the one from which the regression weights are derived. From the results in this study, the following recommendations for applications in social and behavioral sciences can be made.

1. The purpose of the application should be clearly defined before using the multiple regression procedure. Such a distinction is needed because each analytical formula is designed for only one of the two purposes. An effective analytical formula for one purpose might not be accurate for the other.

2. The commonly used statistical software only provides an adjusted R^2 without distinction between the two parameters based upon the two research purposes. Also the currently used Wherry-2 formula for calculating the adjusted R^2 was not found to be the most effective analytical formula. Therefore, it is recommended that to obtain a more accurate adjusted R^2, instead of simply relying on the statistical software, researchers use the Pratt formula for the first purpose, the Browne formula for the second purpose, or

refer to the more detailed "recommended formulae" across sample sizes and number of predictors in Tables 8 and 9 of the present study.

3. The ratio of sample size to the number of predictors appears to be a major factor that affects the performance of these analytical formulae. Therefore, it is recommended that sufficient sample size and relatively few predictors be used in the multiple regression procedure in order to obtain a relatively accurate and stable estimate of the population parameter.

REFERENCES

Ayabe, C. R. (1985). Multicrossvalidation and the Jackknife in the estimation of shrinkage of the multiple coefficient of correlation. *Educational and Psychological Measurement, 45,* 445 - 451.

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

Burket, G. R. (1964). A study of reduced rank models for multiple prediction. *Psychometric Monograph* (12, serial No. 65).

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

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement, 2,* 595-607.

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

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

Cummings, C. C. (1982, March). *Estimates of multiple correlation coefficient shrinkage.* Paper presented at the Annual Meeting of American Educational Research Association, New York.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69*(3), 161-182.

Efron, B. (1979). Bootstrap method: Another look at the jackknife. *The Annals of Statistics, 7,* 1-26.

Ezekiel, M. (1929). The application of the theory of error to multiple and curvilinear correlation. *Proceeding Supplement, American Statistical Association Journal, 24,* 99-104.

Fan, X., & Wang, L. (1996). Comparability of Jackknife and Bootstrap results: An investigation for a case of canonical correlation analysis. *The Journal of Experimental Education, 64*(2), 173-189.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology.* Needham Height, MA: Allyn & Bacon.

Hamilton, L. C. (1991). *Regression with graphics: A second course in applied statistics.* Belmont, CA: Dubury Press.

Herzberg, P. A. (1969). The parameters of cross-validation. *Psychometrika Monograph Supplements, 16,* 1-10.

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

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

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27,* 179-182.

Kennedy, E. (1988). Estimation of the squared cross-validity coefficient in the context of best subset regression. *Applied Psychological Measurement, 12*(3), 231-237.

Kromrey, J. D., & Hines, C. V. (1995). Use of empirical estimates of shrinkage in multiple regression: A caution. *Educational and Psychological Measurement, 55*(6), 901-925.

Kromrey, J. D., & Hines, C. V. (1996). Estimating the coefficient of cross-validity in multiple regression: A comparison of analytical and empirical methods. *The Journal of Experimental Education, 64(3)*, 240-266.

Krus, D. J., & Fuller, E. A. (1982). Computer-assisted multicross-validation in regression analysis. *Educational and Psychological Measurement, 40,* 101-112.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology, 22,* 45-55.

Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample.* Princeton, NJ: Educational Testing Service.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11,*1-11.

Newman, I., McNeil, K., Garver, T., & Seymour, G. (1979, April). *A Monte Carlo evaluation of estimated parameter of five shrinkage estimate formuli.* Paper presented at the Annual Meeting of Americal Educational Research Association, San Francisco, CA.

Nicholson, G. E. (1960). Prediction in future samples. In I. Olkin (Ed.), *Contribution to probability and statistics* (pp. 350). Stanford, CA: Stanford University Press.

Olkin, E., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29,* 201-211.

Park, C. N., & Dudycha, A. L. (1974). A cross-validation approach to sample size: determination for regression model. *Journal of the American Statistical Association, 69*, 214-218.

Quenouille, M. H. (1949). Approximate test of correlation in the time-series. *Journal of the Royal Statistical Society, 11* (Series B), 68-84.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: a clarification. *Psychological Bulletin, 85, 29*, 201-211.

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

SAS/IML [Computer software] (1990). *Usage and reference, Version 6,* 1st ed. Cary, NC: SAS institute Inc.

SAS/STAT. (1990). User's guide, Vol. 2 Cary, NC: SAS.

Schmitt, N. (1982, August). *Formula estimation of cross-validated multiple correlation.* Paper presented at the Annual Meeting of the Americal Psychological Association, Washington, DC.

Stein, C. (1960). Multiple Regression. In I. Olkin (Ed.). *Contributions to probability and statistics* (pp. 264 - 305). Stanford, CA: Stanford University Press.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Erlbaum.

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

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2*, 440-457.

APPENDICES

Appendix A

Summary of Studies on Estimating R^2 Shrinkage in Multiple Regression Analysis

Summary of Studies on Estimating R^2 Shrinkage in Multiple Regression Analysis

Author/Year	Estimating Method		Study Design	
	Analytical methods (formula)	Empirical methods	Statistical methods	Data set
1. Uhl & Eisenberg (1970)	1. Wherry 2. Modified Wherry 3. Lord-1	None	Regression and Prediction	Test scores from Army Classification Battery and Navy General Classification Test
2. Claudy (1978)	1. Larson/ Smith/ Wherry 2. Olkin /Pratt 3. Pratt 4. Herzberg approximation 4. Lord-2 /Nicholson 5. Darlington 6. Burket 7. Claudy-3	1. Mosier's Double cross-validation 2. Claudy's Double shrinkage estimate	Monte Carlo Study	Computer generated data with parameter chosen to be representative in social and behavioral sciences
3. Newman (1979)	1. McNemar/Wherry 2. Wherry/McNemer/Ezekiel 3. Lord-1 4. Darlington 5. Lord-2	Cross-validation	Monte Carlo Study	Artificially generated data set with known parameters
4. Huberty & Mourad (1980)	1. Smith[1] 2. Ezekiel 3. Wherry 4. Olkin/ Pratt 5. Nicholson/Lord-2 6. Darlington/Stein 7. Rozeboom-1	"Leave-one-out"	Regressions and prediction	Real data set from freshmen (A) and college students (B) at the University of Georgia in 1968-69

(to be continued)

Summary of Studies on Estimating R^2 Shrinkage in Multiple Regression Analysis

	Study Design			
Author/Year	Sample Size (N)	Population Parameters	Number of Predictors (*p*)	Results and Conclusions
1. Uhl & Eisenberg (1970)	50, 100, 150, 250, 325	Calculate Composite R from sample	2 through 13	The Lord-1 formula gave more accurate estimates of shrinkage during cross-validation, regardless of sample size and number of predictors.
2. Claudy (1978)	20, 40, 80, 160	16 independent multivariate normal population of 500 sets of observations with parameters similar to psychological and educational literature; with 400 samples drawn for each sizes	2, 3, 4, 5	1. To estimate ρ^2, the double cross validity estimate was the most accurate in the empirical methods. The Herzberg approximation of Olkin/Pratt formula performed almost equally well. 2. To estimate ρ_c^2, the Darlington formula yielded the most accurate estimate.
3. Newman (1979)	14, 30, 50, 100	ρ_1^2=.06, .07, .06, .08 ρ_2^2=.31, .32, .33, .34 ρ_3^2=.45, .47, .46, .55 with 100 replications for each conditions	4	1. The McNemar/Wherry formula and the Wherry/McNemer/Ezekiel formula are more stable for different sample sizes. 2. Cross-validation shows no advantage over analytical methods. 3. The results might due to artificially generated data in the present Monte Carlo Study.
4. Huberty & Mourad (1980)	50	Determined for the population (A), (B)	(A) 9, 3 (B) 4	1. The Ezekiel formula and the Olkin/ Pratt formula are almost equally accurate in estimating ρ^2. 2. The Nicholson/Lord-2 formula, the Darlington/ Stein formula, and "Leave-one-out" method are nearly accurate in estimating ρ_c^2. 3. "Leave-one-out" method is less practically useful.

Summary of Studies on Estimating R^2 Shrinkage in Multiple Regression Analysis

Author/Year	Estimating Method		Study Design	
	Analytical methods (formula)	Empirical methods	Statistical methods	Data set
5. Schmitt (1982)	1. Wherry/Ezekiel 2. Nicholson/Lord-2 3. Darlington 4. Rozeboom-2 5. Cattin/Browne 6. Browne	None	Regression and prediction	Not specified
6. Cummings (1982)	1. Larson/Smith 2. Wherry 3. Ezekiel/Wherry 4. Olkin/Pratt/Herzberg 5. Pratt 6. Barten 7. Lord-2/Nicholson 8. Darlington 9. Uhl/Eisenberg/Lord-1 10. Burket 11. Claudy-1 12. 1 and 10 13. 2 and 10 14. 3 and 10 15. 4 and 10 16. 5 and 10 17. 6 and 10 18. 7 and 10	1. Half-sample cross validation 2. One- third cross validation 3. Mosier's Double cross validation 4. Claudy's Double cross validation	Regression and prediction	Real data set from freshman at a large university
7. Krus & Fuller (1982)	1. Wherry/Ezekiel 2. Olkin/Pratt	Multicross validation	Regression and prediction	1. Prestructured data set (Thurstone's box) 2. Random data

(to be continued)

Summary of Studies on Estimating R^2 Shrinkage in Multiple Regression Analysis

Author/Year	Study Design			Results and Conclusions
	Sample Size (N)	Population Parameters	Number of Predictors (*p*)	
5. Schmitt (1982)	40 to 240 (40, 80, 240)	.1 to .9 (.1, .2, .4, .6, .8, .9)	5 to 25 (5, 10, 25)	1. When n/*p* ratio increases, the estimations from those analytical formulae become less stable. 2. The Browne formula is more appropriate for cross-validation purpose.
6. Cummings (1982)	30, 60, 120	Calculated with BMDP and SPSS	4, 8	1. Of the double cross-validation methods, Mosier's method is more accurate than Claudy's estimate. 2. To estimate ρ_c^2, for multiple regression, the combination of the Ezekiel formula and the Claudy-1 formula is the most accurate; for stepwise regression, the combination of the Barten formula and the Claudy-1 formula is the most accurate. 3. To estimate ρ^2, for multiple regression, the Darlington formula is the most accurate; for stepwise regression, the Smith formula, the Ezekiel formula, and the Barten formula are almost equally accurate, but all tend to over-estimate ρ^2.
7. Krus & Fuller (1982)	Random data: 100x20 matrix Thurstone's data: 20x4 matrix	1. Random data: ρ=.462 2. Thurstone's box: ρ=.917	1. Random data: not specified 2. Thurstone's box: 3	1. For Thurstone's data set, both the analytical formulae and multicross-validation work almost equally well. 2. For random data, multicross validation estimate is more accurate than the analytical methods.

Summary of Studies on Estimating R^2 Shrinkage in Multiple Regression Analysis

Author/Year	Estimating Method		Study Design	
	Analytical methods (formula)	Empirical methods	Statistical methods	Data set
8. Ayabe (1985)	1. Wherry/Ezekiel 2. Olkin/Pratt	1. Jackknife 2. Multicross validation	Regression and prediction	1. Prestructured data set (Thurstone's box) 2. Random data
9. Kennedy (1988)	1. Wherry/Ezekiel 2. Browne 3. Claudy-2 4. Lord-2/Nicholson 5. Darlington/Stein 6. Rozeboom-1 7. Cohen/Cohen/Ezekiel	Double-cross validation	Monte Carlo Study	Hypothetically generated data from a nationally representative sample of high school students
10. Kromrey & Hines (1995)	None	1. Cross-validation 2. Multicross validation 3. Jackknife 4. Bootstrap	Monte Carlo Study	Survey data from the National Educational Longitudinal Study
11. Kromrey & Hines (1996)	1. Browne 2. Darlington 3. Ezekiel	1. Cross-validation 2. Multicross validation 3. Jackknife 4. Bootstrap	Monte Carlo Study	Survey data from the National Educational Longitudinal Study

(to be continued)

Summary of Studies on Estimating R^2 Shrinkage in Multiple Regression Analysis

	Study Design			
Author/Year	Sample Size (N)	Population Parameters	Number of Predictors (p)	Results and Conclusions
8. Ayabe (1985)	Random data: 100x20 matrix Thurstone's data: 20x4 matrix	1. Random data: ρ=.409 2. Thurstone's box: ρ=.878	1. Random data: not specified 2. Thurstone's box: 3	1. Multicrossvalidation method produces comparable or superior estimates to the analytical formulae methods. 2. Both the empirical and analytical methods show greater shrinkage for the random data than the prestructured data; multicross validation is better than the others for random data.
9. Kennedy (1988)	30, 70, 150	2,000 simulated subjects for each conditions ρ_1^2=.12 ρ_2^2=.20 100 random samples for each conditions	7, 6, 5	1. The Ezekiel formula gives the most biased estimate in most situations. 2. The Darlington/Stein formula performs better than the Browne formula. 3. Sample size is a primary factor in shrinkage than the number of predictors.
10. Kromrey & Hines (1995)	20, 40, 60, 100, 200	ρ^2=.04, .125, .25, .50 1000 random samples for each conditions	2, 4, 6, 8, 10	None of the empirical estimates consistently provide unbiased estimates and analytical methods thus recommended
11. Kromrey & Hines (1996)	20, 40, 60, 100, 200	ρ^2=.04, .125, .25, .50 1000 random samples for each conditions	2, 4, 6, 8, 10	1. The Browne formula appears to provide the best estimate of ρ_c^2 compared to other methods. 2. The Ezekiel formula is an effective estimate of ρ^2, but not ρ_c^2. 3. The estimation of ρ_c^2 is very poor when sample size is less than 100 for both analytical and empirical methods.

[1] The formula is actually $\hat{R}^2=1-\frac{N}{N-p-1}(1-R^2)$ which is mistakenly used as the Smith formula.

Appendix B

Population Correlation Matrices for Data Simulation

Multicollinearity r=.1; Population ρ^2=.2; 2 predictor variables

```
SAS Program                Output
options linesize=80;
libname lib                                       Analysis of Variance
'defdsk:[sas.monte]';
data a (type=corr);                                  Sum of          Mean
 _type_='corr';            Source      DF           Squares        Square    F Value      Prob>F
 input x1 x2 y;            Model        2        2000.25250    1000.12625   1249.978      0.0001
cards;                     Error     9997        7998.74750       0.80011
1.00   .     .             C Total   9999        9999.00000
 .10  1.00   .
 .3317 .3317 1.00;              Root MSE          0.89449     R-square        0.2000
proc reg;                       Dep Mean          0.00000     Adj R-sq        0.1999
 model y=x1 x2;                 C.V.                    .
 run;
```

Multicollinearity r=.1; Population ρ^2=.5; 2 predictor variables

```
SAS Program [a]            Output [b]
1.00     .      .                R-square        0.5000
 .10    1.00    .
 .5244   .5244 1.00;
```

Note. a. To avoid repetition, the rest of the programs are omitted from the table (the rest of Appendices B).
b. To avoid repetition, the rest of the outputs are omitted from the table (the rest of Appendices B).

Multicollinearity r=.1; Population ρ^2=.8; 2 predictor variables

```
SAS Program                Output
1.00      .       .
 .10     1.00     .              R-square        0.8000
 .66334   .66334 1.00;
```

Multicollinearity r=.3; Population ρ^2=.2; 5 predictor variables

```
SAS Program                Output
1.00     .      .
 .30    1.00    .                  R-square        0.2000
 .3606   .3606 1.00;
```

Multicollinearity r=.3; Population ρ^2=.5; 2 predictor variables

```
SAS Program                Output
1.00     .     .
 .30  1.00     .                 R-square        0.5000
 .5701 .5701 1.00;
```

Multicollinearity r=.3; Population ρ^2=.8; 2 predictor variables

```
SAS Program                Output
1.00     .     .
 .30  1.00     .                   R-square        0.8000
 .7211 .7211 1.00;
```

Multicollinearity r=.5; Population ρ^2=.2; 2 predictor variables

```
SAS Program                Output
1.00     .      .
 .50    1.00    .          R-square        0.2000
 .3873   .3873 1.00;
```

Multicollinearity r=.5; Population ρ^2=.5; 2 predictor variables

```
SAS Program                    Output
1.00    .     .
 .50  1.00    .                    R-square        0.5000
 .6124 .6124 1.00;
```

Multicollinearity r=.5; Population ρ^2=.8; 2 predictor variables

```
SAS Program                    Output
1.00      .     .
 .50    1.00    .                  R-square        0.8000
 .7746   .7746 1.00;
```

Multicollinearity r=.1; Population ρ^2=.2; 4 predictor variables

```
SAS Program                                            Output
1.00     .       .       .      .
 .1     1.00     .       .      .                      R-square        0.2000
 .1      .1     1.00     .      .
 .1      .1      .1     1.00    .
 .25495 .25495 .25495 .25495 1.00;
```

Multicollinearity r=.1; Population ρ^2=.5; 4 predictor variables

```
SAS Program                                            Output
1.00    .      .      .     .
 .1    1.00    .      .     .                          R-square        0.5000
 .1     .1    1.00    .     .
 .1     .1     .1    1.00   .
 .4031 .4031 .4031 .4031 1.00;
```

Multicollinearity r=.1; Population ρ^2=.8; 4 predictor variables

```
SAS Program                                            Output
1.00    .       .       .      .
 .1    1.00     .       .      .                       R-square        0.8000
 .1     .1     1.00     .      .
 .1     .1      .1     1.00    .
 .5099 .5099   .5099   .5099 1.00;
```

Multicollinearity r=.3; Population ρ^2=.2; 4 predictor variables

```
SAS Program                                            Output
1.00    .       .       .      .
 .3    1.00     .       .      .                       R-square        0.2000
 .3     .3     1.00     .      .
 .3     .3      .3     1.00    .
 .3082 .3082   .3082   .3082 1.00;
```

Multicollinearity r=.3; Population ρ^2=.5; 4 predictor variables

```
SAS Program                                            Output
1.00     .       .       .      .
 .3     1.00     .       .      .                      R-square        0.5000
 .3      .3     1.00     .      .
 .3      .3      .3     1.00    .
 .48735 .48735 .48735 .48735 1.00;
```

Multicollinearity r=.3; Population ρ^2=.8; 4 predictor variables

```
SAS Program                                       Output
1.00    .      .      .      .
 .3    1.00    .      .      .                    R-square        0.8000
 .3     .3    1.00    .      .
 .3     .3     .3    1.00    .
 .61645 .61645 .61645 .61645 1.00;
```

Multicollinearity r=.5; Population ρ^2=.2; 4 predictor variables

```
SAS Program                                       Output
1.00    .      .      .      .
 .5    1.00    .      .      .                    R-square        0.2000
 .5     .5    1.00    .      .
 .5     .5     .5     1.00   .
 .35355 .35355 .35355  .35355 1.00 ;
```

Multicollinearity r=.5; Population ρ^2=.5; 4 predictor variables

```
SAS Program                                       Output
1.00  .    .    .    .
 .5  1.00  .    .    .                            R-square        0.5000
 .5   .5  1.00  .    .
 .5   .5   .5  1.00  .
 .559 .559 .559 .559 1.00;
```

Multicollinearity r=.5; Population ρ^2=.8; 4 predictor variables

```
SAS Program                                       Output
1.00    .     .     .     .
 .5   1.00    .     .     .                       R-square        0.8000
 .5    .5   1.00    .     .
 .5    .5    .5   1.00    .
 .7071 .7071 .7071 .7071 1.00 ;
```

Multicollinearity r=.1; Population ρ^2=.2; 8 predictor variable

```
SAS Program                                                             Output
1.00    .      .      .      .      .      .      .      .
 .1    1.00    .      .      .      .      .      .      .             R-square
 .1     .1    1.00    .      .      .      .      .      .               0.2000
 .1     .1     .1    1.00    .      .      .      .      .
 .1     .1     .1     .1    1.00    .      .      .      .
 .1     .1     .1     .1     .1    1.00    .      .      .
 .1     .1     .1     .1     .1     .1    1.00    .      .
 .1     .1     .1     .1     .1     .1     .1    1.00    .
 .20615 .20615 .20615 .20615 .20615 .20615 .20615 .20615 1.00;
```

Multicollinearity r=.1; Population ρ^2=.5; 8 predictor variables

```
SAS Program                                                             Output
1.00    .      .      .      .      .      .      .      .
 .1    1.00    .      .      .      .      .      .      .             R-square
 .1     .1    1.00    .      .      .      .      .      .               0.5000
 .1     .1     .1    1.00    .      .      .      .      .
 .1     .1     .1     .1    1.00    .      .      .      .
 .1     .1     .1     .1     .1    1.00    .      .      .
 .1     .1     .1     .1     .1     .1    1.00    .      .
 .1     .1     .1     .1     .1     .1     .1    1.00    .
 .32595 .32595 .32595 .32595 .32595 .32595 .32595 .32595 1.00;
```

Multicollinearity r=.1; Population ρ^2=.8; 8 predictor variables

```
SAS Program                                                                  Output
1.00    .       .       .       .       .       .       .       .
 .1    1.00     .       .       .       .       .       .       .            R-square
 .1     .1     1.00     .       .       .       .       .       .              0.8000
 .1     .1      .1     1.00     .       .       .       .       .
 .1     .1      .1      .1     1.00     .       .       .       .
 .1     .1      .1      .1      .1     1.00     .       .       .
 .1     .1      .1      .1      .1      .1     1.00     .       .
 .1     .1      .1      .1      .1      .1      .1     1.00     .
 .4123  .4123   .4123   .4123   .4123   .4123   .4123   .4123  1.00;
```

Multicollinearity r=.3; Population ρ^2=.2; 8 predictor variables

```
SAS Program                                                                  Output
1.00    .      .      .      .      .      .      .      .
 .3    1.00    .      .      .      .      .      .      .                   R-square
 .3     .3    1.00    .      .      .      .      .      .                     0.2000
 .3     .3     .3    1.00    .      .      .      .      .
 .3     .3     .3     .3    1.00    .      .      .      .
 .3     .3     .3     .3     .3    1.00    .      .      .
 .3     .3     .3     .3     .3     .3    1.00    .      .
 .3     .3     .3     .3     .3     .3     .3    1.00    .
 .2784 .2784 .2784 .2784 .2784 .2784 .2784 .2784 1.00;
```

Multicollinearity r=.3; Population ρ^2=.5; 8 predictor variables

```
SAS Program                                                                  Output
1.00     .       .       .       .       .       .       .       .
 .3     1.00     .       .       .       .       .       .       .           R-square
 .3      .3     1.00     .       .       .       .       .       .             0.5000
 .3      .3      .3     1.00     .       .       .       .       .
 .3      .3      .3      .3     1.00     .       .       .       .
 .3      .3      .3      .3      .3     1.00     .       .       .
 .3      .3      .3      .3      .3      .3     1.00     .       .
 .3      .3      .3      .3      .3      .3      .3     1.00     .
 .44019 .44019 .44019 .44019 .44019 .44019 .44019 .44019 1.00;
```

Multicollinearity r=.3; Population ρ^2=.8; 8 predictor variables

```
SAS Program                                                                  Output
1.00     .       .       .       .       .       .       .       .
 .3     1.00     .       .       .       .       .       .       .           R-square
 .3      .3     1.00     .       .       .       .       .       .             0.8000
 .3      .3      .3     1.00     .       .       .       .       .
 .3      .3      .3      .3     1.00     .       .       .       .
 .3      .3      .3      .3      .3     1.00     .       .       .
 .3      .3      .3      .3      .3      .3     1.00     .       .
 .3      .3      .3      .3      .3      .3      .3     1.00     .
 .55678 .55678 .55678 .55678 .55678 .55678 .55678 .55678 1.00;
```

Multicollinearity r=.5; Population ρ^2=.2; 8 predictor variables

```
SAS Program                                                                  Output
1.00    .      .      .      .      .      .      .      .
 .5    1.00    .      .      .      .      .      .      .                   R-square
 .5     .5    1.00    .      .      .      .      .      .                     0.2000
 .5     .5     .5    1.00    .      .      .      .      .
 .5     .5     .5     .5    1.00    .      .      .      .
 .5     .5     .5     .5     .5    1.00    .      .      .
 .5     .5     .5     .5     .5     .5    1.00    .      .
 .5     .5     .5     .5     .5     .5     .5    1.00    .
 .3354 .3354 .3354 .3354 .3354 .3354 .3354  .3354 1.00;
```

Multicollinearity r=.5; Population ρ^2=.5; 8 predictor variables

```
Program                                                                   Output
1.00    .      .      .      .      .      .      .      .
 .5    1.00    .      .      .      .      .      .      .              R-square
 .5     .5    1.00    .      .      .      .      .      .                0.5000
 .5     .5     .5    1.00    .      .      .      .      .
 .5     .5     .5     .5    1.00    .      .      .      .
 .5     .5     .5     .5     .5    1.00    .      .      .
 .5     .5     .5     .5     .5     .5    1.00    .      .
 .5     .5     .5     .5     .5     .5     .5    1.00    .
 .53035 .53035 .53035 .53035 .53035 .53035 .53035 .53035 1.00;
```

Multicollinearity r=.5; Population ρ^2=.8; 8 predictor variables

```
SAS Program                                                Output
1.00    .     .     .     .     .     .     .     .
 .5   1.00    .     .     .     .     .     .     .        R-square          0.8000
 .5    .5   1.00    .     .     .     .     .     .
 .5    .5    .5   1.00    .     .     .     .     .
 .5    .5    .5    .5   1.00    .     .     .     .
 .5    .5    .5    .5    .5   1.00    .     .     .
 .5    .5    .5    .5    .5    .5   1.00    .     .
 .5    .5    .5    .5    .5    .5    .5   1.00    .
 .6708 .6708 .6708 .6708 .6708 .6708 .6708 .6708 1.00;
```

Appendix C

Basic SAS Program for Simulating Sample Data

Multicollinearity r=.1; Population ρ^2=.2, .5, .8; 2 predictor variables

SAS Program

```
options linesize=80 nonumber nodate;
libname lib 'c:\ping\sas\formula';
proc printto log='c:\ping\sas\formula\logfile.tmp';

/* Monte Carlo simulation for 2 predictors conditions*/
/* A22, A52, A82; population r quare=.2, .5, .8; p=2; coll r=.1 */
/* n=20, 40, 60, 100, 200; replicate=500 */

data t (type=corr);
 _type_='corr';
 input x1 x2 y;

cards;
```

Insert the intercorrelation matrices
from Appendix A

```
/*Generate factor pattern*/
proc factor n=3 outstat=FACOUT;
data pattern;
set FACOUT ;
 if _TYPE_='PATTERN';
 drop _TYPE_ _NAME_;
run;

/*start regress module*/
proc iml;
start regress;

%macro a22;
%let N=500;

%do b=1 %to 5;
 %if &b=1 %then %do; %let smpln=20;  %end;
 %if &b=2 %then %do; %let smpln=40;  %end;
 %if &b=3 %then %do; %let smpln=60;  %end;
 %if &b=4 %then %do; %let smpln=100; %end;
 %if &b=5 %then %do; %let smpln=200; %end;
%do I=1 %to &N;

/*Define necessary variables for analysis*/
 nov=3;                                                   /*Number of variables*/
 mcol=.1;                                                 /*Multicollinearity r */
                                                          /*Population multiple R
square*/
```

/*Population ρ^2=.2*/ pmr=.2;	/*Population ρ^2=.5*/ pmr=.5;	/*Population ρ^2=.8*/ pmr=.8;

```
 smpsize=&smpln;
 NAMES={rsq smpsize nov mcol pmr};
 con=j(&smpln,1,1);

/*Generate data*/
 use pattern;

 read ALL VAR _NUM_ INTO F;
 F=F`;
 data=rannor(j(&smpln, 3, 0));
 data=data`;
 Z=F*data;
 Z=Z`;
 Z=con||Z;

 x=Z[,{1 2 3}];
 y=Z[,4];
```

```
/*Calculate sample R square*/
 b=inv(x`*x)*x`*y;
 yhat=x*b;
 r=y-yhat;
 sse=ssq(r);
 dfe=nrow(x)-ncol(x);
 mse=sse/dfe;
 cssy=ssq(y-sum(y)/&smpln);
 rsq=(cssy-sse)/cssy;

/*Generate output matrix*/
 tempdata=rsq||smpsize||nov||mcol||pmr;

 if &b=1 & &i=1 then outp=tempdata;
 else outp=outp//tempdata;

%end;
%end;
%mend a22;
%a22;

/*Create SAS data file*/
/*lib.a22, lib.a52, lib.a82*/
create lib.a22 from outp [colname=NAMES];
append from outp;

/*finish regress module*/
finish;
run regress;
quit; [a]
```

Note. a. To avoid repetition, the rest of the programs are omitted from the table.

Appendix D

Basic SAS Program for Estimating Population Cross-Validity Coefficient (ρ_c^2)

Multicollinearity r=.1; Population ρ^2=.2, .5, .8; 2 predictor variables

```
options linesize=80 nonumber nodate;
libname lib 'c:\ping\sas\cross';
proc printto log='c:\ping\sas\cross\cv2abc.tmp';

/*Cross-validation with 2 indepent samples*/
/* CVA22, CVA52, CVA82; population r quare=.2, .5, .8; p=2; coll r=.1 */
/* n=20, 40, 60, 100, 200; replicate=500 */

data t (type=corr);
 _type_='corr';
 input x1 x2 y;

cards;

 Insert the intercorrelation matrices
 from Appendix A

/*Generate factor pattern*/
proc factor n=3 outstat=FACOUT;
data pattern;
set FACOUT ;
 if _TYPE_='PATTERN';
 drop _TYPE_ _NAME_;
run;

/*start regress module*/
proc iml;
start regress;

%macro cva22;
%let N=250;

%do b=1 %to 5;
 %if &b=1 %then %do; %let smpln=20;  %end;
 %if &b=2 %then %do; %let smpln=40;  %end;
 %if &b=3 %then %do; %let smpln=60;  %end;
 %if &b=4 %then %do; %let smpln=100; %end;
 %if &b=5 %then %do; %let smpln=200; %end;
%do i=1 %to &N;

/*Define necessary variables for analysis*/
 nov=3;                                                /*Number of variables*/
 mcol=.1;                                              /*Multicollinearity r */
                                                       /*Population multiple R
square*/
    /*Population ρ²=.2*/      /*Population ρ²=.5*/      /*Population ρ²=.8*/
    pmr=.2;                   pmr=.5;                   pmr=.8;

 smpsize=&smpln;

/*create intercept matrix*/
 con=j(&smpln,1,1);

 /*create sample size matrix*/
 smpsize=smpsize#con;

 /*generate 2 groups of random data*/
 use pattern;
 read ALL VAR _NUM_ INTO F;
 F=F`;
 data1=rannor(j(&smpln, 3, 0));
 data2=rannor(j(&smpln, 3, 0));

 data1=data1`;
 data2=data2`;

 Z1=F*data1;
 Z2=F*data2;
 Z1=Z1`;
 Z2=Z2`;

 /*add intercept*/
 Z1=con||Z1;
 Z2=con||Z2;

 /*define dependent and independent variables*/
 x1=Z1[,{1 2 3}];
 y1=Z1[,4];
 x2=Z2[,{1 2 3}];
 y2=Z2[,4];

 /*calculate regression weights for each groups*/
 b1=inv(x1`*x1)*x1`*y1;
 b2=inv(x2`*x2)*x2`*y2;

 /*apply regression weights from one sample to another*/
 /*calculate predicted y*/
 /*yhat12=predicted y1 from regression weights derived from sample 2*/
 /*yhat21=predicted y2 from regression weights derived from sample 1*/
 yhat12=x1*b2;
 yhat21=x2*b1;

 /*generate output matrices with predicted and original dependent variable*/
 outp=yhat12||y1||yhat21||y2||smpsize;
 n=nrow(outp);

 /*calculate sum of cross-product of y and predicted y */
  yhat12y1=yhat12#y1;
  sumyy1=yhat12y1[+,];
  yhat21y2=yhat21#y2;
  sumyy2=yhat21y2[+,];

 /*calculate sum of each column*/
  s=outp[+,];
  sum1=s[,1];
  sum2=s[,2];
  sum3=s[,3];
  sum4=s[,4];

  /*calculate sum of squares, standard deviations*/
  ss=outp[##,];
  sq=(s##2)/n;
  ssq=ss-sq;
  v=ssq/(n-1);
  sd=sqrt(v);
  syhat12=sd[,1];
  sy1=sd[,2];
  syhat21=sd[,3];
  sy2=sd[,4];

  /*calculate correlation coefficient*/
  ssyy1=sumyy1-sum1*sum2/n;
  ssyy2=sumyy2-sum3*sum4/n;
  r1=ssyy1/((n-1)*syhat12*sy1);
  r2=ssyy2/((n-1)*syhat21*sy2);

  /*square correlation coefficient*/
  rsq1=r1##2;
  rsq2=r2##2;

  /*calculate the average r square*/
  rsqbar=(rsq1+rsq2)/2;

  smpsize=&smpln;
  nov=3;                                                        /*Number of variables*/
  mcol=.1;                                                      /*Multicollinearity r */
                                                                /*Population multiple R
square*/
```

```
/*Population ρ²=.2*/
pmr=.2;
```

```
/*Population ρ²=.5*/
pmr=.5;
```

```
/*Population ρ²=.8*/
pmr=.8;
```

```
   /*create output matrix with estimated cross validity r square*/
   tempr=rsq1||rsq2||rsqbar||smpsize||nov||mcol||pmr;

  if &b=1 & &i=1 then out=tempr;
   else out=out//tempr;

%end;
%end;
%mend cva22;
%cva22;

/*Create SAS data file*/
/*lib.cva22, lib.cva52, lib.cva82*/
create lib.cva22 from out[colname={rsq1 rsq2 rsqbar smpsize nov mcol pmr}];
append from out;

/*finish regress module*/
finish;
run regress;
quit; [a]
```

Note. a. To avoid repetition, the rest of the programs are omitted from the table.

Appendix E

Calculating Adjusted R^2 and R_c^2 with Analytical Formulae with SAS

Multicollinearity r=.1; Population ρ^2=.2, .5, .8; 2 predictor variables

```
/*apply correction formulas to different sample conditions */
options linesize=80;
libname lib 'c:\ping\sas\data';
data lib.outa22; set lib.a22;
/*define necessary variables*/
R=1-RSQ;
N=SMPSIZE;
P=NOV-1;
mcol=.1;
pmr=.2;

/*apply analytical formulae*/
RSMITH=1-N*R/(N-P);                                           /*the Smith formula  */
REZEK=1-(N-1)*R/(N-P-1);                                      /*the Ezekiel formula*/
RWHERRY=1-(N-1)*R/(N-P);                                      /*the Wherry formula */
RLORD1=1-(N+P+1)*R/(N-P-1);                                   /*the Lord-1 formula */
RLORD2=1-(N+P+1)*(N-1)*R/((N-P-1)*N);                         /*the Lord-2 formula */
ROLKIN=1-(N-3)*R*(1+2*R/(N-P+1))/(N-P-1);                     /*the Olkin formula  */
RPRATT=1-(N-3)*R*(1+2*R/(N-P-2.3))/(N-P-1);                   /*the Pratt formula  */
RBURKET= ((N*RSQ-P)/(sqrt(rsq)*(N-P)))**2;                    /*the Burket formula */
RDARLIN=1-(N-1)*(N-2)*(N+1)*R/((N-P-1)*(N-P-2)*N);            /*the Darlington form*/
RBROWNE1=((N-P-3)*REZEK**2+REZEK)/((N-2*P-2)*REZEK+P);        /*the Browne+Ezekiel */
RBROWNE2=((N-P-3)*ROLKIN**2+ROLKIN)/((N-2*P-2)*ROLKIN+P);     /*the Browne+Olkin   */
RCLAUDY1=(2*sqrt(REZEK)-sqrt(RSQ))**2;                        /*the claudy1 formula*/
RCLAUDY2=1-(N-1)*(N-2)*(N-1)*R/((N-P-1)*(N-P-2)*N);           /*the Claudy2 formula*/
RCLAUDY3=1-(N-4)*R*(1+2*R/(N-P+1))/(N-P-1);                   /*the Claudy3 formula*/
RROZE1=1-(N+P)*R/(N-P);                                       /*the Rozeboom1 formu*/
RROZE2=REZEK*(1+P*(1-REZEK)/((N-P-2)*REZEK))**-1;             /*the Rozeboom2 formu*/

data lib.outa221;
                                                                 /*output n=20
                                                              */

set lib.outa22;
 if smpsize=20;
 proc means data=lib.outa221; run;

data lib.outa222;
                                                                 /*output n=40
                                                              */

 set lib.outa22;
 if smpsize=40;
 proc means data=lib.outa222; run;

data lib.outa223;
                                                                 /*output n=60
                                                              */

 set lib.outa22;
 if smpsize=60;
 proc means data=lib.outa223; run;

data lib.outa224;
                                                                 /*output n=100
                                                              */

 set lib.outa22;
 if smpsize=100;
 proc means data=lib.outa224; run;

data lib.outa225;
                                                                 /*output n=200
                                                              */

 set lib.outa22;
 if smpsize=200;
 proc means data=lib.outa225; run;
```

Appendix F

Means and Standard Deviations of Bias Obtained from Analytical Formulae

Means and Standard Deviations of the Bias Obtained from Analytical Formulae (Multicollinearity r = .1)

N/p	*p*	n	ρ^2		Bsm	Bzee	Bwh	Bolk	Bpra	Bcl3	ρ_c^2	Blo1	Blo2	Bbur	Bdar	Bbr1	Bbr2	Bcl1	Bcl2	Bro1	Bro2
2.5	8	20	.2	x̄	.0158	-.0128	.0550	.0146	-.0051	.0608	.1192	-.3598	-.2977	.0701	-.6553	.0079	.0253	.0460	-.5090	-.2172	-.0007
				sd	.2540	.2633	.2413	.2703	.2822	.2544	.1089	.4162	.3968	.2740	.5092	.1816	.1901	.1989	.4630	.3718	.1760
			.5	x̄	-.0060	-.0244	.0193	.0052	-.0035	.0343	.3165	-.1151	-.0751	.0638	-.3058	.0034	.0326	.0271	-.0211	-.0231	-.0191
				sd	.2073	.2149	.1970	.2129	.2199	.2003	.1561	.3789	.3640	.2801	.4514	.2619	.2654	.2742	.4153	.3449	.2607
			.8	x̄	-.0098	-.0175	.0007	.0007	-.0009	.0124	.6806	-.0125	.0041	.0368	-.0916	-.0146	.0108	.0179	-.0524	.0257	-.0312
				sd	.1062	.1100	.1009	.1036	.1053	.0975	.1096	.1992	.1923	.1729	.2337	.1834	.1783	.1816	.2164	.1834	.1879
5	4	20	.2	x̄	-.0071	-.0178	.0333	.0092	-.0051	.0557	.1592	-.2353	-.1815	.0209	-.2632	-.0069	.0126	.0003	-.1581	-.1277	-.1069
				sd	.2029	.2056	.1928	.2112	.2177	.1987	.1251	.3006	.2884	.2450	.3070	.1999	.2077	.2069	.2831	.2762	.1948
			.5	x̄	-.0069	-.0137	.0184	.0153	.0092	.0438	.4083	-.0842	-.0504	.0125	-.1017	.0064	.0367	-.0015	-.0357	-.0166	-.0175
				sd	.1936	.1961	.1839	.1939	.1984	.1825	.1590	.2987	.2879	.2508	.3044	.2504	.2508	.2546	.2832	.2772	.2516
			.8	x̄	-.0136	-.0164	-.0029	.0016	.0005	.0133	.7667	-.0516	-.0373	-.0201	-.0580	-.0251	-.0039	-.0270	-.0311	-.0231	-.0390
				sd	.0988	.1001	.0938	.0940	.0950	.0884	.0839	.1526	.1471	.1392	.1554	.1403	.1350	.1423	.1448	.1418	.1448
5	8	40	.2	x̄	-.0031	-.0083	.0170	.0023	-.0011	.2390	.1199	-.1354	-.1100	.0143	-.1693	.0064	.0150	.0052	-.1181	.0836	-.0030
				sd	.1451	.1460	.1414	.1492	.1504	.1451	.0870	.2098	.2057	.1526	.2155	.1465	.1501	.1474	.2071	.2014	.1435
			.5	x̄	-.0056	-.0089	.0070	.0047	.0033	.0180	.4137	-.0531	-.0371	-.0004	-.0744	-.0114	.0034	-.0144	-.0422	-.0205	-.0240
				sd	.1235	.1243	.1204	.1239	.1245	.1205	.1029	.1815	.1783	.1641	.1858	.1638	.1641	.1688	.1793	.1750	.1649
			.8	x̄	-.0088	-.0102	-.0036	-.0016	-.0019	.0038	.7493	-.0134	-.0068	.0025	-.0222	-.0049	.0054	-.0013	-.0089	.0001	-.0113
				sd	.0653	.0658	.0637	.0638	.0639	.0620	.0606	.1031	.1014	.0984	.1053	.0998	.0980	.0996	.1019	.0997	.1012
7.5	8	60	.2	x̄	-.0101	-.0122	.0034	-.0058	-.0072	.0083	.1407	-.0910	-.0751	-.0107	-.0990	-.0112	-.0055	-.0211	-.0676	-.0592	-.0195
				sd	.1047	.1049	.1030	.1067	.1071	.1048	.0763	.1422	.1405	.1168	.1431	.1159	.1177	.1168	.1397	.1388	.1145
			.5	x̄	-.0036	-.0049	.0048	.0038	.0033	.0125	.4495	-.0400	-.0302	-.0111	-.0450	-.0139	-.0044	-.0177	-.0254	-.0203	-.0223
				sd	.0983	.0985	.0966	.0985	.0987	.0968	.0885	.1463	.1448	.1389	.1471	.1388	.1389	.1410	.1441	.1433	.1396
			.8	x̄	-.0041	-.0046	-.0007	.0009	.0008	.0044	.7722	-.0115	-.0075	-.0025	-.0135	-.0042	.0020	-.0041	-.0056	-.0035	-.0082
				sd	.0512	.0513	.0503	.0503	.0504	.0494	.0429	.0751	.0743	.0730	.0755	.0732	.0724	.0734	.0074	.0734	.0740
10	2	20	.2	x̄	-.0157	-.0205	.0250	.0067	-.0057	.0534	.2051	-.1984	-.1487	-.0340	-.1743	-.0392	-.0172	-.0337	-.0820	-.1024	-.0528
				sd	.1660	.1670	.1577	.1716	.1763	.1615	.1278	.2478	.2392	.2023	.2437	.2005	.2089	.1950	.2279	.2313	.1901
			.5	x̄	-.0087	-.0117	.0167	.0176	.0125	.0460	.4812	-.1007	-.0697	-.0328	-.0857	-.0190	.0113	-.0389	-.0281	-.0408	-.0450
				sd	.1705	.1715	.1620	.1693	.1726	.1593	.1252	.2386	.2299	.2142	.2343	.2120	.2109	.2157	.2183	.2183	.2155
			.8	x̄	-.0163	-.0176	-.0055	.0005	-.0005	.0123	.7805	-.0439	-.0308	-.0177	-.0376	-.0111	.0081	-.0206	-.0131	-.0185	-.0241
				sd	.0935	.0941	.0889	.0883	.0892	.0831	.0764	.1381	.1334	.1284	.1358	.1262	.1215	.1296	.1272	.1291	.1305

(to be continued)

N/p	*p*	n	ρ^2		Bsm	Bzee	Bwh	Bolk	Bpra	Bcl3	ρ_c^2	Blo1	Blo2	Bbur	Bdar	Bbr1	Bbr2	Bcl1	Bcl2	Bro1	Bro2
10	4	40	.2	$\bar{x}$	-.0021	-.0044	.0180	.0064	.0034	.0278	.1534	-.0816	-.0584	-.0025	-.0749	.0030	.0129	-.0058	-.0300	-.0357	-.0101
				sd	.1244	.1248	.1213	.1273	.1282	.1238	.0940	.1676	.1646	.1436	.1667	.1443	.1470	.1423	.1610	.1617	.1421
			.5	$\bar{x}$	-.0108	-.0122	.0020	.0013	.0001	.0148	.4486	-.0396	-.0248	-.0044	-.0354	.0002	.0144	-.0098	-.0067	-.0104	-.0129
				sd	.1212	.1216	.1182	.1213	.1219	.1181	.1013	.1767	.1738	.1674	.1758	.1665	.1665	.1693	.1703	.1710	.1681
			.8	$\bar{x}$	-.0098	-.0104	-.0045	-.0018	-.002	.0036	.7815	-.0243	-.0182	-.0116	-.0225	-.0098	-.0005	-.0131	-.0108	-.0123	-.0158
				sd	.0621	.0623	.0605	.0604	.0605	.0588	.0515	.0864	.0850	.0833	.0860	.0828	.0813	.0837	.0833	.0836	.0841
12.5	8	100	.2	$\bar{x}$	-.0011	-.0018	.0069	.0019	.0014	.0101	.1601	-.0429	-.0341	-.0070	-.0419	-.0052	-.0017	-.0148	-.0244	-.0253	-.0114
				sd	.0753	.0753	.0745	.0761	.0762	.0753	.0564	.0984	.0977	.0900	.0983	.0898	.0906	.0917	.0970	.0971	.0895
			.5	$\bar{x}$	-.0046	-.0051	.0004	.0001	-.0001	.0053	.4656	-.0127	-.0161	-.0073	-.0211	-.0064	-.0010	.0097	-.0101	-.0106	-.0115
				sd	.0726	.0727	.0189	.0727	.0727	.0719	.0553	.0974	.0967	.0949	.0973	.0978	.0948	.0955	.0960	.0961	.0953
			.8	$\bar{x}$	-.0032	-.0034	-.0012	-.0001	-.0001	.0019	.7847	-.0086	-.0064	-.0038	-.0084	-.0035	.0000	-.0043	-.0039	-.0042	-.0057
				sd	.0374	.0374	.0370	.0370	.0370	.0366	.0287	.0501	.0498	.0494	.0501	.0493	.0489	.0495	.0494	.0495	.0496
15	4	60	.2	$\bar{x}$	-.0006	-.0016	.0127	.0050	.0037	.0189	.1682	-.0513	-.0366	-.0072	-.0435	-.0009	.0055	-.0123	-.0148	-.0222	-.0113
				sd	.0992	.0993	.0975	.1008	.1011	.0991	.0773	.1370	.1355	.1261	.1362	.1263	.1279	.1268	.1333	.1340	.1256
			.5	$\bar{x}$	-.0116	-.0122	-.0031	-.0034	-.0040	.0054	.4709	-.0352	-.0258	-.0140	-.0302	-.0092	-.0003	-.0164	-.0012	-.0166	-.0179
				sd	.0969	.0970	.0953	.0970	.0972	.0953	.0689	.1314	.1299	.1272	.1306	.1265	.1265	.1279	.1277	.1284	.1276
			.8	$\bar{x}$	-.0024	-.0027	.0010	.0028	.0028	.0063	.7837	-.0070	-.0033	.0006	-.0050	.0026	.0084	.0000	.0022	.0004	-.0011
				sd	.0472	.0473	.0464	.0463	.0463	.0455	.0366	.0652	.0644	.0636	.0648	.0633	.0625	.0638	.0634	.0638	.0640
20	2	40	.2	$\bar{x}$	-.0077	-.0088	.0125	.0019	-.0010	.0234	.1830	.0747	-.0524	-.0177	-.0580	-.0067	.0036	-.0167	-.0154	-.0311	-.0216
				sd	.1142	.1144	.1114	.1168	.1176	.1137	.0844	.1470	.1444	.1325	.1451	.1345	.1370	.1320	.1402	.1420	.1322
			.5	$\bar{x}$	-.0112	-.0119	.0016	.0016	.0005	.0151	.4873	-.0507	-.0375	-.0224	-.0411	-.0214	.0014	-.0245	-.0141	-.0240	-.0256
				sd	.1184	.1186	.1155	.1183	.1189	.1151	.0849	.1557	.1529	.1493	.1536	.1477	.1475	.1500	.1484	.1503	.1498
			.8	$\bar{x}$	-.0004	-.0007	.0046	.0076	.0074	.0128	.7869	-.0082	-.0026	.0028	-.0040	.0069	.0154	.0022	.0066	.0027	.0014
				sd	.0558	.0559	.0544	.0541	.0542	.0527	.0458	.0763	.0751	.0739	.0754	.0730	.0716	.0740	.0731	.0739	.0742
25	4	100	.2	$\bar{x}$	-.0042	-.0046	.0038	-.0010	-.0014	.0073	.1829	-.0362	-.0277	-.0138	-.0301	-.0086	-.0049	-.0168	-.0133	-.0193	-.0156
				sd	.0737	.0737	.0729	.0745	.0746	.0737	.0610	.0972	.0966	.0939	.0967	.0938	.0944	.0947	.0955	.0960	.0938
			.5	$\bar{x}$	-.0112	-.0115	-.0061	-.0063	-.0065	-.0011	.4880	-.0304	-.0250	-.0188	-.0265	-.0150	-.0097	-.0196	-.0159	-.0197	-.0102
				sd	.0709	.0709	.0702	.0709	.0710	.0702	.0567	.0930	.0924	.0915	.0930	.0912	.0912	.0917	.0914	.0918	.0917
			.8	$\bar{x}$	-.0034	-.0034	-.0013	-.0002	-.0002	.0019	.7901	-.0058	-.0037	-.0015	-.0043	.0001	.0035	-.0017	.0000	-.0016	-.0021
				sd	.0381	.0381	.0377	.0377	.0377	.0373	.0293	.0492	.0489	.0485	.0490	.0483	.0480	.0486	.0484	.0486	.0486

(to be continued)

N/p	*p*	n	ρ^2		Bsm	Bzee	Bwh	Bolk	Bpra	Bcl3	ρ_c^2	Blo1	Blo2	Bbur	Bdar	Bbr1	Bbr2	Bcl1	Bcl2	Bro1	Bro2
25	8	200	.2	$\bar{x}$	-.0050	-.0051	-.0009	-.0035	-.0036	.0006	.1758	-.0214	-.0172	-.0080	-.0190	-.0058	-.0041	-.0106	-.0106	-.0130	-.0094
				sd	.0518	.0518	.0515	.0521	.0521	.0518	.0410	.0683	.0681	.0665	.0682	.0665	.0667	.0671	.0677	.0679	.0665
			.5	$\bar{x}$	-.0031	-.0032	-.0005	-.0006	-.0007	.0019	.4818	-.0102	-.0076	-.0041	-.0087	-.0027	.0000	-.0047	-.0035	-.0049	-.0052
				sd	.0500	.0500	.0498	.0500	.0501	.0498	.0405	.0666	.0664	.0660	.0065	.0659	.0659	.0661	.0661	.0662	.0661
			.8	$\bar{x}$	-.0021	-.0021	-.0011	-.0005	-.0005	.0005	.7904	-.0026	-.0016	-.0004	-.0020	.0002	.0018	-.0006	.0001	-.0005	-.0009
				sd	.0263	.0263	.0262	.0262	.0262	.0261	.0195	.0328	.0326	.0325	.0327	.0325	.0323	.0325	.0325	.0325	.0326
30	2	60	.2	$\bar{x}$	-.0012	-.0017	.0121	.0048	.0036	.0188	.1944	-.0505	-.0362	-.0176	-.0386	-.0078	-.0013	-.0197	-.0109	-.0224	-.0195
				sd	.0964	.0965	.0948	.0980	.0983	.0963	.0733	.1266	.1252	.1211	.1254	.1211	.1224	.1215	.1227	.1238	.1210
			.5	$\bar{x}$	-.0069	-.0072	.0015	.0015	.0010	.0103	.4928	-.0344	-.0253	-.0159	-.0269	-.0087	.0002	-.0168	-.0094	-.0166	-.0174
				sd	.0957	.0957	.0941	.0956	.0957	.0939	.0717	.1240	.1226	.1210	.1228	.1200	.1198	.1212	.1202	.1212	.1212
			.8	$\bar{x}$	-.0086	-.0087	-.0051	-.0031	-.0032	.0004	.7950	-.0179	-.0105	-.0105	-.0148	-.0074	-.0017	-.0108	-.0076	-.0106	-.0111
				sd	.0495	.0495	.0486	.0485	.0486	.0477	.0325	.0637	.0622	.0622	.0631	.0616	.0607	.0622	.0616	.0622	.0623
50	2	100	.2	$\bar{x}$	-.0079	-.0081	.0002	-.0045	-.0049	.0038	.1953	-.0360	-.0276	-.0178	-.0285	-.0109	.0074	-.0184	-.0120	-.0194	-.0185
				sd	.0703	.0703	.0696	.0711	.0703	.0703	.0523	.0896	.0890	.0876	.0891	.0875	.0881	.0875	.0879	.0885	.0876
			.5	$\bar{x}$	-.0043	-.0044	.0007	.0007	.0005	.0059	.4968	-.0216	-.0164	-.0110	-.0169	-.0064	-.0011	-.0113	-.0066	-.0112	-.0115
				sd	.0722	.0722	.0715	.0722	.0722	.0714	.0536	.0930	.0924	.0917	.0925	.0912	.0912	.0918	.0913	.0918	.0918
			.8	$\bar{x}$	-.0046	-.0046	-.0026	-.0013	-.0014	.0007	.7955	-.0085	-.0063	-.0042	-.0065	-.0023	.0010	-.0043	-.0024	-.0042	-.0044
				sd	.0372	.0372	.0369	.0368	.0368	.0364	.0283	.0489	.0486	.0483	.0487	.0480	.0477	.0483	.0480	.0483	.0483
50	4	200	.2	$\bar{x}$	-.0028	-.0029	.0012	-.0012	-.0013	.0029	.1910	-.0180	-.0139	-.0085	-.0145	-.0052	-.0035	-.0093	-.0063	-.0098	-.0090
				sd	.0500	.0500	.0497	.0503	.0503	.0500	.0376	.0644	.0642	.0636	.0642	.0635	.0637	.0638	.0638	.0640	.0636
			.5	$\bar{x}$	-.0035	-.0035	-.0009	-.0010	-.0010	.0016	.4920	-.0106	-.0080	-.0053	-.0084	-.0031	-.0050	-.0055	-.0033	-.0055	-.0056
				sd	.0487	.0488	.0485	.0487	.0488	.0485	.0371	.0642	.064	.0637	.0640	.0635	.0635	.0637	.0636	.0638	.0637
			.8	$\bar{x}$	-.0038	-.0038	-.0028	-.0022	-.0022	-.0012	.7975	-.0075	-.0065	-.0054	-.0066	-.0045	-.0029	-.0055	-.0046	-.0054	-.0056
				sd	.0265	.0265	.0264	.0263	.0263	.0262	.0181	.0318	.0317	.0315	.0317	.0315	.0313	.0316	.0315	.0316	.0316
100	2	200	.2	$\bar{x}$	-.0003	-.0003	.0037	.0014	.0013	.0054	.1964	-.0129	-.0088	-.0044	-.0090	-.0007	.0010	-.0046	-.0009	-.0047	-.0046
				sd	.0513	.0513	.0510	.0516	.0516	.0513	.0367	.0646	.0644	.0641	.0644	.0639	.0642	.0642	.064	.0642	.0641
			.5	$\bar{x}$	-.0025	-.0026	.0000	.0000	-.0001	.0025	.4981	-.0107	-.0082	-.0056	-.0083	-.0032	-.0006	-.0057	-.0032	-.0056	-.0057
				sd	.0498	.0498	.0498	.0496	.0498	.0496	.0346	.0616	.0613	.0611	.0614	.0609	.0609	.0611	.0609	.0611	.0611
			.8	$\bar{x}$	-.0022	-.0022	-.0012	-.0006	-.0006	-.0004	.7970	-.0033	-.0022	-.0012	-.0023	-.0002	.0014	-.0012	-.0002	-.0012	-.0013
				sd	.0267	.0267	.2656	.0265	.0265	.0264	.0183	.0321	.0312	.0319	.0320	.0317	.0316	.0319	.0317	.0319	.0319

Means and Standard Deviations of the Bias Obtained from Analytical Formulae (Multicollinearity r = .3)

N/p	*p*	n	ρ^2		Bsm	Bzee	Bwh	Bolk	Bpra	Bcl3	ρ_c^2	Blo1	Blo2	Bbur	Bdar	Bbr1	Bbr2	Bcl1	Bcl2	Bro1	Bro2
2.5	8	20	.2	$\bar{x}$	.0211	-.0072	.0601	.0205	.0010	.0664	.1205	-.3525	-.2909	.0630	-.6461	.0072	.0249	.0347	-.5008	-.2109	-.0021
				sd	.2499	.2590	.2374	.2659	.2776	.2503	.1009	.4028	.3937	.2065	.4945	.1675	.1760	.1833	.4490	.3590	.1612
			.5	$\bar{x}$	-.0007	-.0189	.0244	.0106	.0021	.0394	.3221	-.1141	-.0745	.0610	-.3028	.0013	.0306	.0227	-.2094	-.0231	-.0212
				sd	.2071	.2147	.1968	.2124	.2194	.2000	.1508	.3565	.3417	.2505	.4285	.2435	.2467	.2591	.3926	.3228	.2427
			.8	$\bar{x}$	-.0054	-.0129	.0049	.0050	.0034	.0165	.6798	-.0048	.0115	.0431	-.0822	-.0077	.0174	.0246	-.0439	.0326	-.0241
				sd	.1039	.1076	.0987	.1009	.1025	.0950	.1243	.2025	.1961	.1794	.2347	.1901	.1849	.1875	.2185	.1880	.1945
5	4	20	.2	$\bar{x}$	.0089	-.0017	.0484	.0257	.0119	.0712	.1557	-.2106	-.1578	.0262	-.2380	.0077	.0281	.0220	-.1349	-.1051	-.0039
				sd	.2063	.2091	.1960	.2146	.2212	.2019	.1291	.3011	.2887	.2116	.3076	.2002	.2086	.1993	.2834	.2765	.1929
			.5	$\bar{x}$	-.0277	-.0347	-.0013	-.0049	-.0113	.0249	.4175	-.1210	-.0859	-.0212	-.1393	-.0271	.0039	-.0372	-.0706	-.0507	-.0521
				sd	.1794	.1818	.1704	.1800	.1841	.1694	.1541	.2820	.2721	.2387	.2872	.2378	.2386	.2429	.2678	.2623	.2389
			.8	$\bar{x}$	-.0213	-.0243	-.0102	-.0058	-.0070	.0063	.7524	-.0475	-.0328	-.0147	-.0552	-.0197	.0019	-.0219	-.0264	-.0180	-.0340
				sd	.1062	.1076	.1009	.1015	.1027	.0956	.0884	.1669	.1610	.1516	.1701	.1527	.1474	.1552	.1584	.1551	.1573
5	8	40	.2	$\bar{x}$	.0070	-.0122	.0131	-.0017	-.0051	.0200	.1247	-.1453	-.1198	.0051	-.1793	-.0023	.0063	-.0063	-.1279	-.0932	-.0116
				sd	.1423	.1432	.1387	.1463	.1475	.1424	.0876	.1984	.1944	.1405	.2038	.1371	.1405	.1388	.1957	.1902	.1343
			.5	$\bar{x}$	-.0024	-.0056	.0102	.0080	.0066	.2123	.4009	-.0443	-.0284	.0082	-.0655	-.0032	.0116	-.0061	-.0334	-.0119	-.0158
				sd	.1225	.1233	.1194	.1228	.1235	.1195	.1058	.1881	.1849	.1712	.1924	.1710	.1712	.1758	.1859	.1816	.1722
			.8	$\bar{x}$	-.0051	-.0064	.0000	.0020	.0018	.0074	.7527	-.0120	-.0056	.0035	-.0207	-.0037	.0064	-.0002	-.0076	.0012	-.0101
				sd	.0632	.0636	.0617	.0617	.0619	.0601	.0616	.1012	.0996	.0966	.1033	.0980	.0963	.0978	.1001	.0977	.0993
7.5	8	60	.2	$\bar{x}$	-.0091	-.0112	.0044	-.0048	-.0062	.0093	.1312	-.0799	-.0641	.0016	-.0879	.0006	.0063	-.0082	-.0565	-.0482	-.0076
				sd	.1051	.1054	.1034	.1072	.1076	.1053	.0750	.1431	.1413	.1144	.1439	.1138	.1156	.1127	.1405	.1396	.1122
			.5	$\bar{x}$	-.0033	-.0047	.0050	.0041	.0036	.0128	.4398	-.0300	-.0202	-.0011	-.0350	-.0038	.0056	-.0077	-.0155	-.0103	-.0124
				sd	.1001	.1003	.0984	.1003	.1005	.0985	.0788	.1399	.1383	.1323	.1407	.1322	.1323	.1344	.1375	.1367	.1330
			.8	$\bar{x}$	-.0059	-.0065	-.0025	-.0009	-.0010	.0026	.7687	-.0101	-.0011	-.0011	-.0122	-.0028	.0035	-.0026	-.0042	-.0021	-.0067
				sd	.0509	.0510	.0500	.0500	.0500	.0491	.0428	.0749	.0728	.0728	.0753	.0731	.0722	.0732	.0737	.0732	.0738
10	2	20	.2	$\bar{x}$	-.0098	-.0146	.0307	.0127	.0004	.0590	.2050	-.1910	-.1417	.8611	-.1672	-.0311	-.0100	-.0356	-.0755	-.0957	-.0427
				sd	.1729	.1739	.1642	.1783	.1831	.1678	.1281	.2436	.2347	17.96	.2393	.1972	.2042	.1946	.2230	.2266	.1947
			.5	$\bar{x}$	-.0274	-.0305	-.0010	-.0009	-.0064	.0286	.4971	-.1393	-.1072	-.0685	-.1237	-.0542	-.0238	-.0766	-.0641	-.0772	-.0809
				sd	.1730	.1940	.1643	.1720	.1754	.1619	.1231	.2419	.2329	.2168	.2375	.2145	.2136	.2196	.2210	.2246	.2180
			.8	$\bar{x}$	-.0206	-.0219	-.0095	-.0034	-.0045	.0085	.7800	-.0486	-.0352	-.0218	-.0421	-.0151	.0044	-.0248	-.0171	-.0226	-.0283
				sd	.0946	.0952	.0900	.0900	.0906	.0843	.0728	.1303	.1256	.1204	.1280	.1182	.1138	.1216	.1193	.1212	.1224

(to be continued)

N/p	*p*	n	ρ^2		Bsm	Bzec	Bwh	Bolk	Bpra	Bcl3	ρ_c^2	Blo1	Blo2	Bbur	Bdar	Bbr1	Bbr2	Bcl1	Bcl2	Bro1	Bro2
10	4	40	.2	$\bar{x}$	-.0110	-.0134	.0092	-.0028	-.0059	.0189	.1567	-.0952	-.0717	-.0087	-.0885	-.0053	.0043	-.0139	-.0430	-.0488	-.0170
				sd	.1263	.1267	.1232	.1296	.1305	.1261	.0844	.1675	.1644	.1387	.1666	.1387	.1419	.1384	.1605	.1613	.1359
			.5	$\bar{x}$	-.0117	-.0132	.0011	.0004	-.0009	.0139	.4519	-.0440	-.0292	-.0088	-.0398	-.0042	.0101	-.0142	-.0111	-.0148	-.0173
				sd	.1201	.1205	.1171	.1202	.1208	.1170	.0980	.1711	.1682	.1621	.1702	.1612	.1612	.1639	.1648	.1655	.1629
			.8	$\bar{x}$	-.0047	-.0053	.0004	.0031	.0029	.0084	.7745	-.0115	-.0055	.0008	-.0098	.0026	.0117	-.0006	.0017	.0002	-.0032
				sd	.0589	.0591	.0574	.0573	.0574	.0557	.0481	.0831	.0817	.0801	.0827	.0797	.0781	.0805	.0801	.0804	.0810
12.5	8	100	.2	$\bar{x}$	-.0021	-.0028	.0059	.0008	.0003	.0090	.1535	-.0374	-.0286	-.0014	-.0364	.0005	.0040	-.0094	-.0189	-.0198	-.0057
				sd	.0795	.0795	.0787	.0804	.0805	.0795	.0594	.1048	.1041	.0965	.1047	.0963	.0970	.0982	.1033	.1033	.0960
			.5	$\bar{x}$	-.0047	-.0052	.0003	.0000	-.0002	.0051	.4657	-.0219	-.0164	-.0076	-.0213	-.0067	-.0012	-.0100	-.0103	-.0108	-.0118
				sd	.0713	.0713	.0706	.0713	.0714	.0706	.0586	.0968	.0962	.0945	.0968	.0944	.0944	.0951	.0955	.0956	.0948
			.8	$\bar{x}$	-.0035	-.0037	-.0015	-.0004	-.0004	.0167	.7873	-.0115	-.0094	-.0068	-.0113	-.0065	-.0030	-.0073	-.0069	-.0071	-.0087
				sd	.0405	.0405	.0401	.0401	.0401	.0396	.0318	.0524	.0520	.0516	.0523	.0515	.0511	.0517	.0516	.0517	.0518
15	4	60	.2	$\bar{x}$	.0008	-.0009	.1342	.0057	.0044	.0196	.1749	-.0572	-.0425	-.0133	-.0495	-.0071	-.0006	-.0188	-.0208	-.0282	-.0175
				sd	.0948	.0949	.0932	.0964	.0967	.0947	.0761	.1337	.1323	.1234	.1329	.1236	.1250	.1242	.1302	.1209	.1229
			.5	$\bar{x}$	-.0077	-.0084	.0007	.0004	-.0001	.0092	.4788	-.0389	-.0296	-.0179	-.0339	-.0131	-.0040	-.0203	-.0157	-.0204	-.0218
				sd	.0938	.0939	.0922	.0938	.0940	.0921	.0749	.1298	.1284	.1260	.1291	.1253	.1253	.1266	.1264	.1271	.1264
			.8	$\bar{x}$	-.0080	-.0082	-.0045	-.0026	-.0027	.0009	.7838	-.0132	-.0093	-.0053	-.0111	-.0033	.0026	-.0060	-.0037	-.0056	-.0071
				sd	.0519	.0519	.0510	.0509	.0510	.0500	.0383	.0696	.0688	.0679	.0692	.0675	.0666	.0680	.0676	.0680	.0682
20	2	40	.2	$\bar{x}$	-.0017	-.0028	.0183	.0079	.0051	.0293	.1828	-.0679	-.0458	-.0117	-.0514	-.0005	.0100	-.0106	-.0090	-.0246	-.0153
				sd	.1198	.1200	.1168	.1225	.1233	.1192	.0830	.1587	.1559	.1438	.1566	.1456	.1482	.1249	.1512	.1532	.1434
			.5	$\bar{x}$	-.0058	-.0065	.0068	.0069	.0058	.0203	.4779	-.0364	-.0224	-.0074	-.0259	.0025	.0162	-.0095	.0008	-.0090	-.0106
				sd	.1244	.1246	.1213	.1243	.1249	.1210	.0892	.1715	.1686	.1647	.1693	.1630	.1628	.1655	.1638	.1658	.1652
			.8	$\bar{x}$	-.0067	-.0070	-.0015	.0015	.0013	.0068	.7973	-.0255	-.0198	.0142	-.0212	-.0099	-.0013	.0148	-.0103	-.0143	-.0156
				sd	.0620	.0620	.0604	.0602	.1603	.0586	.0442	.0844	.0829	.0815	.0833	.0804	.0789	.0817	.0805	.0816	.0818
25	4	100	.2	$\bar{x}$	-.0043	-.0046	.0037	-.0010	-.0015	.0072	.1780	-.0314	-.0229	-.0091	-.0253	-.0038	-.0002	-.0120	-.0085	-.0145	-.0108
				sd	.0717	.0718	.0710	.0725	.0726	.0718	.0601	.0968	.0962	.0935	.0963	.0933	.0940	.0943	.0952	.0956	.0934
			.5	$\bar{x}$	-.0068	-.0070	-.0018	-.0019	-.0021	.0033	.4796	-.0173	-.0120	-.0058	-.0135	-.0021	.0032	-.0067	-.0029	-.0067	-.0072
				sd	.0701	.0702	.0694	.0701	.0702	.0694	.0572	.0939	.0933	.0925	.0935	.0921	.0921	.0927	.0923	.0928	.0926
			.8	$\bar{x}$	-.0051	-.0052	-.0031	-.0019	-.0019	.0002	.7903	-.0079	-.0057	-.0035	-.0062	-.0019	.0015	-.0037	-.0020	-.0036	-.0041
				sd	.0365	.0365	.0361	.0361	.0361	.0357	.0280	.0486	.0483	.0479	.0484	.0477	.0474	.0480	.0477	.0480	.0480

(to be continued)

N/p	*p*	n	ρ^2		Bsm	Bzee	Bwh	Bolk	Bpra	Bcl3	ρ_c^2	Blo1	Blo2	Bbur	Bdar	Bbr1	Bbr2	Bcl1	Bcl2	Bro1	Bro2
25	8	200	.2	$\bar{x}$	-.0031	-.0033	.0009	-.0016	-.0017	.0024	.1769	-.0206	-.0163	-.0071	-.0182	-.0049	-.0032	-.0097	-.0098	-.0121	-.0085
				sd	.0579	.0579	.0576	.0583	.0583	.0580	.0426	.0728	.0725	.0706	.0726	.0705	.0708	.0712	.072	.0723	.0705
			.5	$\bar{x}$	-.0029	-.0031	-.0004	-.0005	-.0006	.0020	.4795	-.0078	-.0052	-.0017	-.0064	-.0003	.0023	-.0024	-.0011	-.0026	-.0028
				sd	.0495	.0495	.0493	.0495	.0496	.0493	.0418	.0659	.0657	.0653	.0658	.0652	.0652	.0654	.0654	.0655	.0654
			.8	$\bar{x}$	-.0019	-.0019	-.0009	-.0003	-.0003	.0007	.7930	-.0051	-.0040	-.0029	-.0045	-.0023	-.0006	-.0030	-.0024	-.0030	-.0033
				sd	.0252	.0252	.0250	.0250	.0250	.0249	.0178	.0324	.0323	.0321	.0323	.0321	.0319	.0322	.0321	.0322	.0322
30	2	60	.2	$\bar{x}$	-.0098	-.0103	.0037	-.0039	-.0051	.0102	.1978	-.0630	-.0486	-.0293	-.0510	-.0196	-.0132	-.0295	-.0230	-.0346	-.0311
				sd	.0928	.0928	.0912	.0944	.0946	.0927	.0669	.1204	.1190	.1141	.1193	.1146	.1160	.1138	.1166	.1177	.1140
			.5	$\bar{x}$	-.0089	-.0092	-.0004	-.0004	-.0009	.0094	.4966	-.0403	-.0313	-.0218	-.0328	-.0145	-.0056	-.0227	-.0152	-.0225	-.0233
				sd	.0889	.0889	.0874	.0888	.0890	.0873	.0656	.1176	.1163	.1147	.1165	.1138	.1137	.1149	.1140	.1150	.1149
			.8	$\bar{x}$	-.0001	-.0011	.0024	.0044	.0043	.0078	.7911	-.0058	-.0022	.0013	-.0028	.0043	.0099	.0011	.0041	.0013	.0007
				sd	.0503	.0504	.0495	.0494	.0494	.0485	.0364	.0677	.0069	.0661	.0671	.0655	.0647	.0662	.0656	.0662	.0063
50	2	100	.2	$\bar{x}$	.0020	.0000	.0082	.0037	.0033	.0119	.2013	-.0335	-.0252	-.0156	-.0261	-.0088	-.0051	-.0166	-.0097	-.0171	-.0163
				sd	.0700	.0700	.0693	.0708	.0708	.0700	.0550	.0903	.0897	.0886	.0898	.0883	.0889	.0888	.0887	.0892	.0886
			.5	$\bar{x}$	-.0094	-.0095	-.0043	-.0043	-.0045	.0009	.4912	-.0213	-.0160	-.0106	-.0166	-.0059	-.0007	-.0109	-.0062	-.0108	-.0111
				sd	.0695	.0695	.0688	.0695	.0695	.0688	.0531	.0895	.0889	.0883	.0890	.0878	.0878	.0883	.0878	.0883	.0883
			.8	$\bar{x}$	-.0043	-.0043	-.0023	-.0011	-.0011	.0010	.7954	-.0080	-.0059	-.0038	-.0061	-.0019	.0015	-.0039	-.0019	-.0039	-.0040
				sd	.0364	.0364	.0360	.0360	.0360	.0356	.0265	.0465	.0462	.0459	.0462	.0456	.0453	.0459	.0456	.0459	.0459
50	4	200	.2	$\bar{x}$	-.0025	-.0025	.0015	-.0009	-.0010	.0032	.1978	-.0246	-.0204	-.0150	-.0210	-.0117	-.0100	-.0158	-.0128	-.0163	-.0155
				sd	.0542	.0542	.0539	.0545	.0545	.0542	.0416	.0682	.0680	.0673	.0680	.0672	.0675	.0675	.0676	.0677	.0673
			.5	$\bar{x}$	.0008	.0007	.0033	.0033	.0032	.0058	.4961	-.0104	-.0078	-.0051	-.0082	-.0029	-.0003	-.0053	-.0031	-.0053	-.0054
				sd	.0496	.0496	.0494	.0496	.0496	.0494	.0353	.0634	.0632	.0630	.0633	.0628	.0628	.0630	.0628	.0630	.0630
			.8	$\bar{x}$	-.0003	-.0003	.0007	.0013	.0013	.0023	.7967	-.0030	-.0020	-.0010	-.0021	-.0001	.0016	-.0010	-.0001	-.0010	-.0011
				sd	.0241	.0241	.0240	.0240	.0240	.0239	.0182	.0302	.0301	.0300	.0301	.0299	.0298	.0300	.0299	.0300	.0299
100	2	200	.2	$\bar{x}$	-.0019	-.0020	.0021	-.0003	-.0004	.0038	.1983	-.0164	-.0123	-.0079	-.0125	-.0042	-.0025	-.0082	-.0044	-.0083	-.0081
				sd	.0505	.0505	.0503	.0508	.0508	.0505	.0365	.0646	.0644	.0641	.0644	.0639	.0641	.0641	.0640	.0642	.0641
			.5	$\bar{x}$	.0005	.0005	.0030	.0031	.0030	.0056	.4978	-.0074	-.0048	-.0022	-.0049	.0001	.0027	-.0023	.0001	-.0023	-.0024
				sd	.0518	.0518	.0516	.0518	.0518	.0516	.0378	.0665	.0663	.0661	.0663	.0659	.0659	.0661	.0659	.0661	.0661
			.8	$\bar{x}$	.0004	.0004	.0014	.0020	.0020	.0030	.7976	-.0012	-.0002	.0008	-.0003	.0018	.0034	.0008	.0017	.0008	.0007
				sd	.0266	.0266	.0265	.0265	.0265	.0263	.0176	.0326	.0325	.0324	.0325	.0323	.0322	.0324	.0323	.0324	.0324

Means and Standard Deviations of the Bias Obtained from Analytical Formulae (Multicollinearity r = .5)

N/p	*p*	n	ρ^2		Bsm	Bzee	Bwh	Bolk	Bpra	Bcl3	ρ_c^2	Blo1	Blo2	Bbur	Bdar	Bbr1	Bbr2	Bcl1	Bcl2	Bro1	Bro2
2.5	8	20	.2	$\bar{x}$	.0031	-.0259	.0429	.0010	-.0194	.0480	.1085	-.3691	-.3061	.0758	-.6694	.0148	.0315	.0555	-.5208	-.2242	.0075
				sd	.2580	.2673	.2451	.2750	.2872	.2588	.1022	.4245	.4047	.2549	.5194	.1746	.1843	.1891	.4723	.3792	.1672
			.5	$\bar{x}$	-.0040	-.0223	.0212	.0069	-.0019	.0359	.3255	-.1228	-.0829	.0579	-.3128	-.0013	.0276	.0316	-.2187	-.0312	-.0232
				sd	.2177	.2257	.2069	.2241	.2317	.2109	.1620	.3828	.3673	.2652	.4585	.2552	.2596	.2625	.4207	.3473	.2531
			.8	$\bar{x}$	-.0101	-.0178	.0004	.0004	-.0013	.0121	.6732	-.0055	.0111	.0439	-.0847	-.0072	.0180	.0249	-.0455	.0327	-.0239
				sd	.1087	.1127	.1033	.1058	.1076	.0996	.1236	.2095	.2026	.1838	.2438	.1940	.1890	.1927	.2266	.1939	.1985
5	4	20	.2	$\bar{x}$	-.0084	-.0192	.0320	.0080	-.0062	.0546	.1463	-.2242	-.1703	.0217	-.2522	.0014	.0212	.0101	-.1469	-.1164	-.0096
				sd	.1950	.1976	.1853	.2033	.2097	.1913	.1167	.2829	.2711	.1924	.2891	.1822	.1906	.1850	.2660	.2594	.1751
			.5	$\bar{x}$	-.0052	-.0119	.0200	.0173	.0114	.0457	.4260	-.0997	-.0660	-.0045	-.1172	-.0112	.0196	-.0215	-.0514	-.0323	-.0356
				sd	.1828	.1852	.1736	.1830	.1870	.1721	.1531	.2821	.2719	.2388	.2874	.2374	.2378	.2415	.2675	.2619	.2389
			.8	$\bar{x}$	-.0141	-.0169	-.0034	.0012	.0000	.0129	.7627	-.0481	-.0338	-.0166	-.0555	-.0216	-.0003	.0235	-.0276	-.0196	-.0356
				sd	.0970	.0983	.0922	.0923	.0934	.0869	.0813	.1528	.1474	.1394	.1557	.1405	.1353	.1425	.1450	.1420	.1450
5	8	40	.2	$\bar{x}$	-.0085	-.0137	.0117	-.0032	-.0066	.0185	.1331	-.1555	-.1299	-.0020	-.1896	-.0110	-.0023	-.0164	-.1381	-.1033	-.0201
				sd	.1405	.1414	.1370	.1447	.1459	.1408	.0926	.1982	.1943	.1454	.2034	.1479	.1411	.1403	.1956	.1903	.1356
			.5	$\bar{x}$	-.0098	-.0131	.0029	.0005	-.0009	.0140	.4070	-.0517	-.0356	.0021	-.0732	-.0093	.0055	-.0124	-.0407	-.0188	-.0220
				sd	.1208	.1215	.1177	.1213	.1219	.1180	.1108	.1815	.1785	.1650	.1856	.1648	.1651	.1693	.1794	.1754	.1658
			.8	$\bar{x}$	-.0079	-.0093	-.0027	-.0007	-.0010	.0047	.7526	-.0156	-.0090	.0002	-.0243	-.0071	.0031	-.0035	-.0111	-.0022	-.0135
				sd	.0646	.0651	.0630	.063	.0633	.0614	.0629	.1046	.1030	.0999	.1068	.1013	.0996	.1012	.1035	.1013	.1027
7.5	8	60	.2	$\bar{x}$	-.0024	-.0045	.0110	.0020	.0006	.1601	.1300	-.0708	-.0552	.0066	-.0788	.0061	.0119	-.0064	-.0476	-.0394	-.0024
				sd	.1056	.1059	.1039	.1076	.1079	.1057	.0744	.1398	.1381	.1162	.1407	.1153	.1171	.1172	.1372	.1363	.1141
			.5	$\bar{x}$	.0001	-.0012	.0084	.0075	.0070	.0162	.4450	-.0312	-.0214	-.0027	-.0362	-.0054	.0041	-.0092	-.0168	-.0116	-.0139
				sd	.0990	.0992	.0973	.0991	.0993	.0974	.0774	.1405	.1389	.1330	.1413	.1328	.1329	.1351	.1382	.1373	.1337
			.8	$\bar{x}$	-.0043	-.0048	-.0009	.0007	.0006	.0042	.7694	-.0089	-.0049	.0001	-.0109	-.0017	.0046	-.0015	-.0030	-.0009	-.0056
				sd	.0489	.0481	.0481	.0481	.0481	.0472	.0434	.0683	.0676	.0664	.0687	.0667	.0659	.0668	.0672	.0668	.0674
10	2	20	.2	$\bar{x}$	.0022	-.0025	.0421	.0254	.1346	.7093	.1796	-.1511	-.1025	.0436	-.1276	.0021	.0251	-.0054	-.0372	-.0572	-.0136
				sd	.1668	.1678	.1585	.1720	.1766	.1619	.1214	.2339	.2253	.5457	.2297	.1899	.1968	.1866	.2139	.2174	.1863
			.5	$\bar{x}$	-.0196	-.0226	.0064	.0067	.0013	.0357	.4500	-.0827	-.0511	-.0127	-.0674	.0012	.0313	-.0185	-.0086	-.0216	-.0250
				sd	.1781	.1792	.1692	.1768	.1803	.1664	.1356	.2621	.2528	.2366	.2576	.2342	.2329	.2378	.2407	.2444	.2381
			.8	$\bar{x}$	-.0214	-.0227	-.0104	-.0043	-.0053	.0077	.7919	-.0615	-.0481	-.0347	-.0550	-.0279	-.0084	-.0376	-.0300	-.0355	-.0412
				sd	.0977	.0983	.0928	.0925	.0934	.0870	.0661	.1368	.1316	.1261	.1343	.1236	.1185	.1274	.1248	.1268	.1282

(to be continued)

N/p	*p*	n	ρ^2		Bsm	Bzcc	Bwh	Bolk	Bpra	Bcl3	ρ_c^2	Blo1	Blo2	Bbur	Bdar	Bbr1	Bbr2	Bcl1	Bcl2	Bro1	Bro2
10	4	40	.2	$\bar{x}$	-.0145	-.0169	.0058	-.0064	-.0096	.0154	.1665	-.1091	-.0855	-.0247	-.1023	-.0203	-.0108	-.0328	-.0567	-.0625	-.0326
				sd	.1217	.1220	.1186	.1247	.1256	.1213	.0985	.1686	.1657	.1452	.1678	.1451	.1477	.1459	.1623	.1630	.1431
			.5	$\bar{x}$	-.0107	-.0122	.0020	.0013	.0001	.0148	.4630	-.0540	-.0392	-.0188	-.0497	-.0142	.0001	-.0242	-.2111	-.0248	-.0272
				sd	.1199	.1202	.1169	.1200	.1206	.1167	.1065	.1722	.1695	.1633	.1715	.1625	.1625	.1651	.1662	.1669	.1640
			.8	$\bar{x}$	-.0089	-.0095	-.0036	-.0009	-.0011	.0045	.7724	-.0141	-.0081	-.0016	-.0124	.0003	.0095	-.0030	-.0007	-.0022	-.0057
				sd	.0587	.0589	.0573	.0571	.0572	.0556	.0493	.0872	.0858	.0841	.0868	.0837	.0822	.0846	.0841	.0845	.0850
12.5	8	100	.2	$\bar{x}$	.0000	-.0001	.0080	.0029	.0025	.0111	.1597	-.0413	.0325	-.0055	-.0403	-.0036	-.0001	-.0131	-.0229	-.0237	-.0098
				sd	.0770	.0770	.0762	.0079	.0780	.0771	.0569	.0976	.0969	.0889	.0975	.0887	.0894	.0901	.0962	.0962	.0883
			.5	$\bar{x}$	.0003	-.0001	.0053	.0050	.0048	.0101	.4614	-.0121	-.0066	.0021	-.0114	.0030	.0084	-.0002	-.0006	-.0011	-.0021
				sd	.0750	.0751	.0743	.0750	.0751	.0743	.0589	.1025	.1018	.1000	.1024	.0999	.0999	.1007	.1011	.1012	.1004
			.8	$\bar{x}$	-.0022	-.0023	-.0001	.0009	.0009	.0030	.7868	-.0096	-.0074	-.0048	-.0094	-.0045	-.0010	-.0053	-.0050	-.0052	-.0067
				sd	.0371	.0371	.0367	.0367	.0367	.0363	.0258	.0482	.0479	.0474	.0482	.0474	.0470	.0475	.0475	.0475	.0477
15	4	60	.2	$\bar{x}$	.0020	.0010	.0153	.0076	.0064	.2151	.1734	-.0536	-.0390	-.0092	-.0459	-.0031	.0034	-.0148	-.0173	-.0246	-.0133
				sd	.0986	.0970	.0970	.1003	.1006	.0986	.0831	.1403	.1389	.1293	.1396	.1296	.1311	.1303	.1367	.1375	.1288
			.5	$\bar{x}$	-.0063	-.0069	.0021	.0018	.0013	.0105	.4787	-.0372	-.0279	-.0163	-.0323	-.0115	-.0024	-.0187	-.0141	-.0188	-.0201
				sd	.0994	.0995	.0977	.0994	.0996	.0976	.0748	.1285	.1271	.1244	.1277	.1238	.1237	.1251	.1249	.1256	.1249
			.8	$\bar{x}$	-.0100	-.0103	-.0065	-.0046	-.0047	-.0010	.7873	-.0189	-.0151	-.0110	-.0169	-.0090	-.0030	-.0117	-.0094	-.0113	.0128
				sd	.0482	.0483	.0474	.0473	.0474	.0465	.0387	.0662	.0655	.0647	.0658	.0643	.0635	.0648	.0644	.0648	.0650
20	2	40	.2	$\bar{x}$	-.0136	-.0147	.0067	-.0042	-.0071	.1756	.1838	-.0821	.0597	-.0257	-.0653	-.0139	-.0037	-.0267	-.0223	-.0381	-.0292
				sd	.1148	.1149	.1119	.1174	.1181	.1142	.0839	.1496	.1470	.1367	.1476	.1380	.1404	.1369	.1427	.1445	.1363
			.5	$\bar{x}$	-.0203	-.0210	-.0073	-.0074	-.0086	.0063	.4864	-.0608	-.0464	-.0310	-.0500	-.0208	-.0069	-.0331	-.0225	-.0326	-.0342
				sd	.1155	.1157	.1126	.1154	.1160	.1123	.0811	.1562	.1534	.1499	.1541	.1482	.1480	.1505	.1490	.1508	.1503
			.8	$\bar{x}$	-.0058	-.0061	-.0007	.0023	.0021	.0077	.7920	-.0193	-.0136	-.0080	-.0150	-.0038	.0049	-.0086	-.0041	-.0081	-.0094
				sd	.0601	.0601	.0586	.0584	.0585	.0568	.0438	.0796	.0782	.0768	.0786	.0758	.0744	.0770	.0760	.0769	.0772
25	4	100	.2	$\bar{x}$	-.0015	-.0019	.0065	.0018	.0013	.0100	.1797	-.0301	-.0216	-.0075	-.0240	-.0023	.0013	-.0097	-.0073	-.0132	-.0093
				sd	.0755	.0755	.0747	.0763	.0764	.0956	.0580	.0950	.0944	.0909	.0946	.0910	.0917	.0914	.0933	.0938	.0909
			.5	$\bar{x}$	-.0016	-.0019	.0034	.0033	.0031	.0084	.4827	-.0150	-.0097	-.0035	-.0111	.0001	.0054	-.0044	-.0007	-.0044	-.0050
				sd	.0742	.0742	.0734	.0742	.0742	.0734	.0556	.0966	.0959	.0950	.0961	.0946	.0946	.0952	.0949	.0953	.0952
			.8	$\bar{x}$	-.0031	-.0032	-.0011	.0001	.0001	.0216	.7909	-.0644	-.0043	-.0021	-.0049	-.0005	.0028	-.0023	-.0007	-.0022	-.0027
				sd	.0392	.0392	.0388	.0387	.0388	.0384	.0271	.0479	.0476	.0472	.0477	.0470	.0466	.0473	.0470	.0472	.0473

(to be continued)

N/p	*p*	n	ρ^2		Bsm	Bzee	Bwh	Bolk	Bpra	Bcl3	ρ_c^2	Blo1	Blo2	Bbur	Bdar	Bbr1	Bbr2	Bcl1	Bcl2	Bro1	Bro2
25	8	200	.2	x̄	-.0010	-.0012	-.0012	.0030	.0005	.0046	.1761	-.0175	-.0133	-.0043	-.0152	-.0021	-.0004	-.0068	-.0068	-.0092	-.0057
				sd	.0530	.0530	.0530	.0527	.0533	.0533	.4087	.0690	.0687	.0671	.0688	.0670	.0673	.0678	.0684	.0685	.0671
			.5	x̄	-.0015	-.0016	.0010	.0009	.0008	.0034	.4824	-.0093	-.0066	-.0032	-.0078	-.0018	.0008	-.0038	-.0026	-.0040	-.0043
				sd	.0546	.0546	.0543	.0546	.0546	.0543	.0374	.0692	.0689	.0685	.0691	.0684	.0684	.0686	.0686	.0687	.0686
			.8	x̄	-.0022	-.0022	-.0012	-.0006	-.0006	.0004	.7941	-.0064	-.0054	-.0042	-.0058	-.0036	-.0020	-.0044	-.0037	-.0043	-.0047
				sd	.0256	.0256	.0255	.0255	.0255	.0253	.0195	.0344	.0343	.0341	.0343	.0341	.0340	.0342	.0341	.0342	.0342
30	2	60	.2	x̄	-.0032	-.0036	.0102	.0029	.0017	.1688	.1911	-.0492	-.0349	-.0164	-.0373	-.0065	.0001	-.0181	-.0096	-.0210	-.0182
				sd	.0939	.0939	.0923	.0955	.0958	.0938	.0693	.1200	.1186	.1149	.1189	.1148	.1161	.1156	.1163	.1173	.1148
			.5	x̄	-.0022	-.0025	.0062	.0063	.0058	.0150	.4884	-.0249	-.0160	-.0067	-.0175	.0005	.0095	-.0075	-.0001	-.0073	-.0081
				sd	.0925	.0926	.0910	.0925	.0927	.0909	.0721	.1235	.1222	.1206	.1224	.1196	.1195	.1208	.1198	.1209	.1208
			.8	x̄	-.0081	-.0082	-.0046	-.0026	-.0027	.0009	.7919	-.0143	-.0106	-.0069	-.0112	-.0038	.0019	-.0072	-.0040	-.0070	-.0075
				sd	.0514	.0514	.0505	.0504	.0504	.0495	.3575	.0671	.0663	.0656	.0665	.0650	.0641	.0657	.0650	.0656	.0657
50	2	100	.2	x̄	.0105	.0104	.0184	.0141	.0137	.0222	.1948	-.0164	-.0082	.0013	-.0090	.0080	.0118	.0003	.0071	-.0001	.0006
				sd	.0757	.0757	.0749	.0765	.0766	.0757	.0536	.0951	.0944	.0931	.0945	.0928	.0935	.0934	.0932	.0938	.0931
			.5	x̄	-.0027	-.0028	.0023	.0023	.0021	.0074	.4926	-.0158	-.0105	-.0052	-.0110	-.0001	.0047	-.0055	-.0078	-.0054	-.0057
				sd	.0701	.0701	.0694	.0701	.0701	.0693	.0520	.0887	.0881	.0874	.0882	.0870	.0869	.0875	.0870	.0875	.0875
			.8	x̄	-.0021	-.0021	-.0001	.0011	.0011	.0032	.7960	-.0072	-.0051	-.0030	-.0053	-.0012	.0021	-.0031	-.0012	-.0031	-.0033
				sd	.0367	.0367	.0363	.0363	.0363	.0359	.0264	.0461	.0457	.0454	.0458	.0451	.0448	.0454	.0452	.0454	.0455
50	4	200	.2	x̄	-.0023	-.0024	.0017	-.0007	-.0008	.0034	.1909	-.0174	-.0133	-.0079	-.0139	-.0046	-.0029	-.0087	-.0057	-.0092	-.0084
				sd	.0522	.0522	.0519	.0525	.0525	.0522	.0395	.0686	.0684	.0677	.0684	.0676	.0679	.0679	.0680	.0682	.0678
			.5	x̄	-.0008	-.0009	.0017	.0016	.0016	.0042	.4929	-.0089	-.0063	-.0036	-.0067	-.0014	.0012	-.0038	-.0016	-.0038	-.0039
				sd	.0489	.0489	.0486	.0489	.0489	.0486	.0388	.0640	.0638	.0635	.0638	.0634	.0634	.0635	.0634	.0636	.0635
			.8	x̄	-.0022	-.0022	-.0012	-.0006	-.0006	.0004	.7968	-.0051	-.0041	-.0030	-.0042	-.0021	-.0005	-.0031	-.0022	-.0030	-.0032
				sd	.0261	.0261	.0259	.0259	.0259	.0258	.0186	.0331	.0330	.0329	.0331	.0328	.0327	.0329	.0328	.0329	.0329
100	2	200	.2	x̄	-.0026	-.0027	.0014	-.0010	-.0011	.0031	.1960	-.0148	-.0107	-.0063	-.0109	-.0025	-.0008	-.0065	.0028	-.0066	-.0065
				sd	.0530	.0530	.0527	.0533	.0533	.0530	.0396	.0663	.0661	.0658	.0661	.0656	.0659	.0659	.0657	.0659	.0658
			.5	x̄	-.0003	-.0003	.0022	.0022	.0022	.0048	.4983	-.0087	-.0061	-.0035	-.0062	-.0011	.0014	-.0036	-.0012	-.0036	-.0036
				sd	.0517	.0517	.0514	.0517	.0517	.0514	.0370	.0636	.0634	.0632	.0634	.0630	.0630	.0632	.0630	.0632	.0632
			.8	x̄	-.0012	-.0012	-.0002	.0004	.0004	.0014	.7994	-.0047	-.0037	-.0027	-.0037	-.0017	-.0001	-.0027	-.0017	-.0027	-.0027
				sd	.0259	.0259	.0258	.0257	.0257	.0256	.0180	.0323	.0322	.0321	.0322	.0320	.0319	.0321	.0320	.0321	.0321

Note. *N/p: N/p* Ratio. *p:* Number of predictor variables. n: Sample Size. ρ^2: Squared population multiple correlation coefficient. Smr: Squared sample multiple correlation coefficient. Bsm: Bias for the Smith formula. Beze: Bias for the Ezekiel formula. Bwh: Bias for the Wherry formula. Bolk: Bias for the Olkin and Pratt formula. Bpra: Bias for the Pratt estimation of the Olkin/Pratt formula. Bcl3: Bias for the Claudy-3 formula. ρ_c^2: (Estimated) population squared cross-validity coefficient. Blo1: Bias for the Lord formula-1. Blo2: Bias for the Lord formula - 2. Bbur: Bias for the Burket formula. Bdar: Bias for the Darlington/Stein formula. Bbr1: Bias for the Browne formula with ρ^2 estimated by the Ezekiel formula. Bbr1: Bias for the Browne formula with ρ^2 estimated by the Olkin/Pratt formula. Bcl1: Bias for the Claudy formula-1. Bcl2: Bias for the Claudy formula-2. Bro1: Bias for the Rozeboom formula-1. Bro2: Bias for the Rozeboom formula -2.