5-2015

Estimation of the Squared Population Cross-Validity Under Conditions of Predictor Selection

Andrew J. Kircher
Western Kentucky University, andrew.kircher538@topper.wku.edu

ESTIMATION OF THE SQUARED POPULATION CROSS-VALIDITY UNDER CONDITIONS OF PREDICTOR SELECTION

A Thesis
Presented to
The Faculty of the Department of Psychological Sciences
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

By
Andrew Kircher

May 2015

ESTIMATION OF THE SQUARED POPULATION CROSS-VALIDITY UNDER CONDITIONS OF PREDICTOR SELECTION

Date Recommended April 10, 2015

Reagan D. Brown, Director of Thesis

Elizabeth L. Shoenfelt

Amber N. Schroeder

Dean, Graduate Studies and Research 4-21-15 Date

This thesis is dedicated to my parents, Scott and Teresa Moore, for their continuous support and to Camry Krupp for motivating me to continue to persist when I struggled the most.

ACKNOWLEDGMENTS

I wish to thank my committee members who have provided me with both their time and expertise. I would also like to give special thanks to Dr. Reagan Brown, my committee chairman for his constructive feedback, his time, and his assistance throughout the entire process. Thank you Elizabeth Shoenfelt and Amber Schroeder for agreeing to serve on my committee.

CONTENTS

LIST OF TABLES

ESTIMATION OF THE SQUARED POPULATION CROSS-VALIDITY UNDER CONDITIONS OF PREDICTOR SELECTION

Andrew Kircher May 2015 26 Pages

Directed by: Reagan Brown, Elizabeth Shoenfelt, and Amber Schroeder

Department of Psychological Sciences Western Kentucky University

The current study employed a Monte Carlo design to examine whether sample-based and formula-based estimates of cross-validated R^2 differ in accuracy when predictor selection is and is not performed. Analyses were conducted on three datasets with 5, 10, or 15 predictors and different predictor-criterion relationships. Results demonstrated that, in most cases, a formula-based estimate of the cross-validated R^2 was as accurate as a sample-based estimate. The one exception was the five predictor case wherein the formula-based estimate exhibited substantially greater bias than the estimate from a sample-based cross validation study. Thus, formula-based estimates, which have an enormous practical advantage over a two sample cross validation study, can be used in most cases without fear of greater error.

Introduction

Individuals in fields related to business, education, health, and psychology often engage in research in which variables are used to forecast the outcome of a given criterion (Punch, 2009; Saks & Allsop, 2007; Sekaran & Bougie, 2013; Spatz & Kardas, 2008). The most common analytic technique for creating a predictive model is Ordinary Least Squares (OLS) linear regression analysis. An OLS linear regression uses a sample from the target population with the intent to create a model that accurately predicts the criterion variable. One of the results obtained from this analysis is an estimate of the predictive power (R^2) of the model. One of the unfortunate consequences of using a model developed on a sample of data is that the model is overly customized to the sample of data on which it was derived (Pedhazur, 1997; Raju, Bilgic, Edwards, & Fleer, 1999; Schmitt & Ployhart, 1999). In other words, the model will not predict as well when applied to other samples derived from the same population. In order to correct for overfitting, researchers calculate an estimate of R^2 that reflect how well the predictors, as weighted in the regression equation, predict the criterion variable when applied to future samples of data. The estimate of the reduced, or shrunken, R^2 can be computed through either empirical cross-validation or formula-based methods. Because of the relative ease of formula-based methods, these methods are often preferred over empirical cross-validation.

When conducting predictive research it is important for the model to be practical as well as accurate (Pedhazur, 1997). For practical application, a model may be simplified by removing predictors that only marginally improve the accuracy of the model as a whole. However, empirical processes for selecting which variables to include

in a model result in an increase of overfitting (Babyak, 2004). This study will examine how formula-based methods compare to empirical cross-validation in their ability to estimate the shrunken R^2 accurately when predictors have been selected. In order to develop a comprehensive understanding, this paper will provide a conceptual background to review key concepts.

Conceptual Background

The goal of predictive research is to optimize the prediction of a given criterion (Pedhazur, 1997). In predictive research, variables are chosen *a priori* or are selected after an examination of the data based on their overall contribution to criterion prediction (Pedhazur, 1997). Often, these predictions are made using a linear regression analysis. A linear regression analysis uses a linear model to estimate the relationship between a criterion variable and a predictor variable. When a linear regression model has only one predictor variable, the model is a simple linear regression. The population simple linear regression model is as follows:

$$Y_i = \alpha + \beta_1 x_i + \varepsilon_i \qquad (1)$$

Where:

Y_i is the criterion variable.

x_i is the predictor variable.

β is the beta weight; the amount of change in Y_i for every one-unit increase in x_i.

α is the constant; the value of Y_i when the value of x_i is zero.

ε_i is the error; the variability in Y_i not related to the predictor in the model.

In most cases, one predictor alone cannot accurately forecast the outcome of a criterion variable; better prediction is possible with multiple predictors. A model with

multiple predictors is referred to as a multiple linear regression model. The population multiple linear regression model is as follows:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad (2)$$

Where:

k is the number of predictors with in the model.

Both the simple and multiple regression equations represent the population. However, it is often difficult, if not impossible, to obtain data from the entire target population. Therefore, researchers often rely on a sample of the population; the multiple regression model for a sample is represented by the following model:

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdots b_k x_{ik} + e_i \qquad (3)$$

Where:

y_i is the sample criterion value.

a is the estimated constant.

b is the estimated beta weight.

e is the estimated error.

Ordinary Least Squares

One method used by researchers to determine the value of the parameters a and b is Ordinary Least Squares (OLS). In the OLS model, parameters are differentially weighted for each predictor variable to minimize the Sum of Squares Error (SSE). The parameters chosen to minimize the SSE are the best fitting parameters for that set of data. The sample regression equation for the prediction of scores on *Y* given scores on various *X* variables (i.e., the prediction equation) is:

$$y'_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdots b_k x_{ik} \qquad (4)$$

Where:

y'_i is the predicted criterion value.

The prediction equation allows for the computation of a predicted *Y* score for each person given that person's scores on the various *X* variables. It does not yield actual scores on *Y*. The difference between an actual *Y* score and a predicted *Y* score is the error of prediction, *e* (literally: $e_i = y_i - y'_i$).

In the OLS regression, *b*s are weighted based on a given predictor variable's relationship with both y_i and the other predictor variables. The model then weighs b in a way that minimizes the difference between y_i and y'_i. These weights, called optimal weights, may lead to problems when a model derived on one sample is applied to other samples from the same population.

Squared Multiple Correlation Coefficient

To understand how well the model predicts the criterion, researchers calculate the squared multiple correlation coefficient (R^2). R^2 is determined by dividing the sum of squares regression (SSR) by the sum of squares total (SST):

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \qquad (5)$$

R^2 can also be computed by computed as one minus the ratio of SSE to SST:

$$R^2 = 1 - \frac{SSE}{SST} \qquad (6)$$

The values of R^2 range from zero to one, with a value of zero indicating that there is not a linear relationship, and a value of one indicating that there is a perfect linear relationship. Although R^2 is both useful and important to regression, several problems can cause a misrepresentation of R^2.

The first problem is that the multiple regression model is derived from the sample that was used to generate the model; this process causes the model to be tailored to that particular sample, a phenomenon referred to as overfitting. All sample-based regression models have some degree of overfitting, causing the predictive power of the model (indexed by R^2) to decrease when the model is applied to another sample from the same population; that is, the overall model will not predict the criterion as well in future samples as it did in the first sample.

A second problem is that R^2 typically increases when the number of predictor variables used in a model increases (even when the added variables are not significant). This increase in R^2 occurs because in a given sample, the correlations differ from their true population values due to sampling error. Sampling errors that result in the inflation of sample correlations can result in overestimated regression coefficients and R^2 values. Pedhazur (1997) noted that when in the population R^2 is zero, the sample R^2 is equal to $k/(N-1)$ (where k is the number of predictors and N is the sample size). In other words, a sample R^2 will have a value of one (i.e., a perfect correlation) when the number of predictors is equal to the sample size minus one, while the actual population R^2 has a value of zero. Schmitt and Ployhart (1999) suggested that, in order to reduce the magnitude of the inflation in R^2, the N:k ratio should be at least 10:1.

Predictor Selection

In order to avoid having too many predictor variables, researchers should only select the predictors that produce a significant increase in R^2 in conjunction with the other predictor variables. There are many approaches available for selecting predictors. Three of the most common are forward selection, backward elimination, and stepwise selection. Forward selection begins by entering the predictor variable that has the highest zero-order correlation with the criterion in an empty model. The next predictor variable entered is the one that produces the greatest increase to R^2 relative to the rest of the predictors in the model. Predictors will continue to be entered in the model until no more of the available predictors can add a significant increase in R^2. In contrast to forward selection is backward elimination, in which predictor variables are removed one at a time from a model containing all of the predictors. The predictor variable first removed is the one that will lead to the smallest (and non-significant) reduction in R^2 relative to the rest of the predictors in the model. Predictor variables will continue to be removed until removing a predictor causes a significant reduction in R^2. Finally, stepwise regression is a combination of both forward selection and backward elimination. After each variable is entered into the model using forward selection, backward elimination is used to determine if the variable should stay in the model. Predictor selection techniques such as forward, backward, and stepwise selection may seem ideal; however, these selection techniques are likely to increase overfitting problems. Predictor selection leads to overfitting because the process is influenced by the unique characteristics of the sample, which allows both the model and R^2 to have a greater chance of being tailored to the sample. To be specific, the likelihood of retention for a given predictor variable is greater

for sampling error inflated correlations and lessor for sampling error deflated correlations. Thus, the resultant regression equation is likely to contain a set of predictors that are not the best at predicting the criterion. Therefore, it is suggested that researchers exercise caution when reviewing research that does not validate its model after utilizing predictor selection techniques (Society for Industrial and Organizational Psychology, Inc., 2003).

Shrunken R^2

To correct for the effects of overfitting, researchers must adjust the sample R^2 (Society for Industrial and Organizational Psychology, Inc., 2003); this adjusted R^2 is referred to as the shrunken R^2. There are two methods used to estimate the shrunken R^2, empirical (or sample-based) cross-validation and formula-based methods. In order to empirically cross-validate the data, the results from a regression analysis performed on one sample must be applied to a second sample so that predicted *Y* scores can be computed for each case in that sample. These predicted *Y* scores within that second sample are then correlated with the actual *Y* scores. The resultant correlation, once squared, is the cross-validated squared multiple correlation. The cross-validated squared multiple correlation serves as the estimate of the squared population cross-validity (p_c^2). A major limitation of empirical cross-validation is the requirement of a second sample; attaining a second sample can be extremely difficult, time consuming, and costly.

The alternate method to correct for overfitting is the formula-based method. With this approach various formulas are used to estimate the squared population cross-validity. These methods do not require a second sample and are therefore more time and cost effective than empirical cross-validation. Although the benefits offered by these methods

are tempting, the very nature of a statistical estimate (as opposed to an actual application to a second sample) may inspire skepticism regarding the relative accuracy of these formulas.

Cross-Validation Research

Raju et al. (1999) found that when using an OLS model, formula-based methods work just as well as empirical cross-validation in estimating p_c^2. The Raju et al. study also compared different formula-based methods and found that the Burket (1964) equation performed at least as well as other more complicated equations (e.g. Cattin, 1980). The Burket equation is as follows:

$$R_{Burket} = \frac{(NR^2 - k)}{R(N - k)} \qquad (7)$$

Following the research of Raju et al. (1999), Schmitt and Ployhart (1999) conducted a study to determine which formula-based method produced the best estimate of p_c^2 after predictor selection. Each of the formulas was calculated with either all the predictors (k_{full}) or only the remaining predictors after selection (k_{step}). In addition, each of the formulas was calculated with either an R^2 that used all the predictors (R^2_{full}) or only the remaining predictors after selection (R^2_{step}). These formulas were applied to three different data sets that varied in sample size, population validity, and the number of predictors. Based on the data gathered, it appeared the Burket_{full} equation (computed using both k_{full} and R^2_{full}) produced the least biased estimates of p_c^2. Although the Schmitt and Ployhart study did address the effects of predictor selection on formula-based estimates of p_c^2, it did not address empirical cross-validation. That is, Schmitt and Ployhart identified the best formula from a group of possible formulas, but they did not

compare the effectiveness of formula-based estimates of p_c^2 to the effectiveness of empirical cross-validation. It is possible that even the best of these formulas is inferior to an empirical cross-validation in predictor selection situations.

The Current Study

Without a comparison of the effectiveness of formula-based methods to sample-based cross-validation under circumstances involving predictor selection, it is unclear if these formula-based estimates are substantially less accurate than an empirical cross-validation in the estimation of ρ_c^2. If the formula-based methods were found to estimate ρ_c^2 at least as well as the empirical cross-validation under conditions involving predictor selection, it would be far more efficient to use the formula-based methods. The current study is designed to test if there is a difference between the two methods.

Hypothesis: When predictors are selected via forward selection, the accuracy of estimates of the cross-validated R^2 will differ between empirical and formula-based estimates.

The current study will employ a Monte Carlo design. A Monte Carlo procedure is optimal for this study because it allows for both the generation and manipulation of large datasets with known parameters. Having access to a population will allow for the actual population cross-validity (ρ_c^2) to be calculated and compared to the estimates derived from the two techniques.

This study will examine three datasets with a multiple regression equation developed with predictor selection and without predictor selection. In order to prevent confusion, statistical terms that pertain to conditions without predictor selection will be

denoted by the subscript "(*ns*)," while conditions with predictor selection will be denoted by the subscript "(*s*)" unless stated otherwise.

In conditions without predictor selection:

$\rho^2_{c(ns)}$ is the squared population cross-validity for the full regression equation (i.e., without predictor selection).

$k_{(ns)}$ is the total number of predictors (i.e., number of predictors before predictor selection).

$R^2_{(ns)}$ is the squared sample multiple correlation coefficient for the full regression equation (i.e., without predictor selection).

$R^2_{c(ns)}$ is the squared sample cross-validity for the full regression equation (i.e., without predictor selection).

Burket$_{(ns)}$ is the Burket adjustment to the sample squared multiple correlation computed with both $k_{(ns)}$ and $R^2_{(ns)}$. This equation is equivalent to Burket$_{full}$ equation used in the Schmitt and Ployhart (1999) study.

In conditions with predictor selection:

$\rho^2_{c(s)}$ is the squared population cross-validity for the selected regression equation (i.e., with predictor selection).

$R^2_{(s)}$ is the squared multiple correlation coefficient for the selected regression equation (i.e., with predictor selection).

$k_{(s)}$ is the number of selected predictors.

$R^2_{c(s)}$ is the squared sample cross-validity from the predictor selected equation.

Burket$_{(s)}$ is the Burket adjustment to the sample squared multiple correlation computed with both $k_{(s)}$ and $R^2_{(s)}$. This equation is equivalent to Burket$_{step}$ equation used in the Schmitt and Ployhart (1999) study.

Burket$_{(hyb)}$ is the Burket equation computed with both $k_{(ns)}$ and $R^2_{(ns)}$. Although Burket$_{(hyb)}$ is redundant with Burket$_{(ns)}$, the statistics are given different names to indicate a crucial difference in how they are assessed for accuracy. Accuracy of estimates of the cross-validated R^2 are always determined by a comparison to a population cross-validity (obtained by an application of the sample regression equation to the population). The difference between the two statistics lies in which sample regression is applied. For Burket$_{(ns)}$, the regression equation developed on all of the predictors (i.e., no selection) is applied to the population. For Burket$_{(hyb)}$, it is the selected regression equation that is applied to the population. The Burket$_{(hyb)}$ is a true hybrid model: the Burket equation uses terms from the no selection condition to estimate the cross-validated R^2, but it is the selected equation that is of interest; it is the selected equation that is cross-validated on the population. As a final note, Burket$_{(hyb)}$ is equivalent to Burket$_{full}$ equation used in the Schmitt and Ployhart (1999).

Schmitt and Ployhart (1999) found that when predictor selection is performed, the Burket$_{(hyb)}$ equation produced the least biased estimator of ρ^2_c. Therefore, the present study will use the Burket$_{(hyb)}$ equation when calculating the formula-based method for estimating $\rho^2_{c(s)}$. To provide a more comprehensive understanding of effects of predictor selection on the Burket equation, this study will include the Burket$_{(s)}$ equation as well as the Burket$_{(ns)}$ equation for the full regression equation.

Method

Population Generation

Three datasets, each representing a population consisting of 1,000,000 cases, were generated. Predictors in all three datasets were generated to have the same population multiple correlation of .50 with the criterion variable. Additionally, all predictor variables in each dataset were created to have intercorrelations of .30 (Appendix A, Appendix B, and Appendix C) . Consistent population multiple correlations and predictor intercorrelations allow for a more direct comparison of results between datasets.

In a manner similar to Schmitt and Ployhart (1999), each dataset differed in the number of predictors and in the predictor-criterion relationship. The first dataset (D1) consisted of five predictors with individual predictor-criterion relationships ranging from .10 to .40. The second population dataset (D2) consisted of 10 predictors with individual predictor-criterion relationships ranging from .00 to .40. The third dataset (D3) consisted of 15 predictors with individual predictor-criterion relationships ranging from -.10 to .40. Appendices A-C list the correlation matrices for each dataset. Means and standard deviations for each variable were set to zero and one, respectively. For each dataset, samples were randomly selected from the population with a sample size of 150 cases, a sample size typical of personnel selection research (Schmitt & Ployhart, 1999).

Procedure

The following procedure was used to generate sample R^2 values, formula-based estimates of cross-validities, sample-based cross-validities, and squared population cross-validities for regression equations developed without predictor selection (i.e., all

predictors included) and with predictor selection (i.e., only significant predictors included).

1. A sample of 150 cases was randomly selected from the population.
2. A multiple regression equation, using all of the predictors, was generated from the sample data, yielding $R^2_{(ns)}$.
3. Forward selection (probability of entry = .05) was applied to the same sample data, yielding a second regression equation and an $R^2_{(s)}$.
4. The $R^2_{(ns)}$ obtained from Step 2 was adjusted using the Burket$_{(ns)}$ equation, yielding a formula estimate of $\rho^2_{c(ns)}$.
5. The $R^2_{(s)}$ obtained from Step 3 was adjusted using the Burket$_{(s)}$ and Burket$_{(hyb)}$ equations, yielding formula estimates of $\rho^2_{c(s)}$.
6. A second sample of 150 cases, serving as the sample for a sample-based empirical cross-validation, was randomly drawn from the population.
7. The OLS models from Steps 2 and 3 were applied to the sample from Step 6 to obtain predicted criterion scores in this second sample. The squared correlations between the predicted criterion scores and the criterion scores in the second sample were computed to obtain empirical estimates of $\rho^2_{c(ns)}$ and $\rho^2_{c(s)}$. That is, these squared correlations are the empirical cross-validated R^2 without predictor selection ($R^2_{c(ns)}$) and with predictor selection ($R^2_{c(s)}$).
8. The OLS models from Steps 2 and 3 were applied to the entire population to obtain the actual ρ^2_c without predictor selection ($\rho^2_{c(ns)}$) and with predictor selection ($\rho^2_{c(s)}$).

9. The uncorrected sample R^2 values as well as the various estimates of the cross validated R^2 (the sample cross validated R^2 values and the Burket estimates of the population cross validated R^2) were compared to the actual cross validated R^2 values ($\rho^2_{c(ns)}$ for the no selection condition and $\rho^2_{c(s)}$ for the selected condition) to assess the accuracy of the corrected and uncorrected coefficients. Bias, the signed difference between the actual ρ^2_c and its respective estimate, and squared bias, an index of the variability of the bias estimate, were computed.
10. The process described in Steps 1-9 was repeated until it yielded 1000 complete samples for each dataset (i.e., D1, D2, D3). Samples are considered valid if they retained at least one predictor variable after selection. In the event that all predictor variables were removed after selection, both the sample in the selection condition and the corresponding sample without selection were replaced with the next computed sample.
11. The results were then averaged across the 1000 samples, yielding a Mean Bias (MB) and a Mean Squared Bias (MSB) for each estimator.
12. Cohen's *d* was computed to assess the effect size for comparisons of various corrections for the estimates of ρ^2_c (e.g., Burket$_{(hyb)}$ versus $R^2_{C(s)}$). Cohen's (1988) standards for effect sizes of *d* are .2 for small, .5 for medium, and .8 for large.

Results

All samples for datasets D1 and D2 yielded valid (i.e., at least one predictor selected) results for the predictor selection portion of the analysis. For dataset D3 one of the 1000 samples resulted in zero predictors selected via forward selection; the results from this sample were deleted. A new sample was drawn; the results from the analysis of this new sample were retained in place of the original sample.

Estimating ρ_c^2: MB

Table 1 shows mean and SD of Bias for estimates of $\rho_{c(ns)}^2$ and $\rho_{c(s)}^2$ for each of the three datasets. There were several trends that were found to be consistent in both ρ_c^2 conditions (i.e., $\rho_{c(ns)}^2$ and $\rho_{c(s)}^2$). First, the uncorrected squared multiple correlation coefficients (i.e., $R_{(ns)}^2$ and $R_{(s)}^2$) were found to produce the greatest amount of bias across all three datasets (R^2 overestimated by .04 at a minimum). These results were no surprise and are the reason why cross-validation exists. Second, in most conditions, bias was greater for datasets with more predictors. Other factors held constant, more predictors in a model increases the likelihood and impact of sampling error. Third, both the sample cross-validation and the Burket equation are effective at reducing bias. Fourth, when predictor selection is performed, Burket$_{(hyb)}$ exhibits less bias than Burket$_{(s)}$. Last of all, in both conditions, a sample-based cross-validation exhibits less bias than any of the Burket corrected values; however, the magnitude of that difference was trivial for datasets D1 and D2.

Table 1

Mean and SD of Bias

			D1		D2		D3	
	Variable	N	M	SD	M	SD	M	SD
	$R^2_{(ns)}$	1000	-0.043	0.060	-0.087	0.057	-0.133	0.060
$\rho^2_{c(ns)}$	$R^2_{c(ns)}$	1000	0.001	0.060	-0.005	0.058	-0.004	0.057
	Burket$_{(ns)}$	1000	0.005	0.063	0.004	0.062	-0.002	0.066
	$R^2_{(s)}$	1000	-0.049	0.059	-0.063	0.058	-0.078	0.068
	$R^2_{C(s)}$	1000	0.001	0.058	-0.005	0.058	-0.003	0.058
$\rho^2_{c(s)}$	Burket$_{(s)}$	1000	-0.022	0.058	-0.032	0.056	-0.047	0.064
	Burket$_{(hyb)}$	1000	-0.020	0.059	-0.009	0.059	-0.004	0.068

Note: All bias statistics in the $\rho^2_{c(ns)}$ condition represent the difference between the population cross-validated R^2 of the regression equation based on the all predictors and the named variable. All bias statistics in the $\rho^2_{c(s)}$ condition represent the difference between the population cross-validated R^2 of the regression equation based on the selected predictors and the named variable.

Estimating ρ^2_c: MSB

Table 2 shows mean and SD of Squared Bias of $\rho^2_{c(ns)}$ and $\rho^2_{c(s)}$ for each of the three datasets. In both the ρ^2_c conditions, the uncorrected squared multiple correlation coefficients (i.e., $R^2_{(ns)}$ and $R^2_{(s)}$) were found to produce the greatest amount of variably in bias across all three datasets. For D1, the differences MSB values across all conditions were small and consistent. For D2 and D3, uncorrected R^2 was worse than any method for estimating ρ^2_c. All methods for estimating ρ^2_c performed about the same.

Table 2

Mean and SD of Squared Bias

			D1		D2		D3	
	Variable	N	M	SD	M	SD	M	SD
$\rho^2_{c(ns)}$	$R^2_{(ns)}$	1000	0.005	0.008	0.011	0.011	0.021	0.017
	$R^2_{c(ns)}$	1000	0.003	0.004	0.003	0.005	0.003	0.004
	Burket$_{(ns)}$	1000	0.004	0.006	0.004	0.005	0.005	0.006
$\rho^2_{c(s)}$	$R^2_{(s)}$	1000	0.006	0.008	0.007	0.009	0.011	0.012
	$R^2_{C(s)}$	1000	0.003	0.004	0.003	0.005	0.003	0.004
	Burket$_{(s)}$	1000	0.004	0.006	0.004	0.006	0.006	0.008
	Burket$_{(hyb)}$	1000	0.004	0.006	0.004	0.005	0.005	0.006

Effect Size Analysis

Rather than compute significance tests for the above comparisons, tests that have no meaning in a Monte Carlo analysis, the differences between various cross-validation techniques were assessed using effect sizes. Table 3 shows the effect size for the differences in bias between various cross-validation techniques. Within the no selection condition, bias values for a sample-based cross-validation and a Burket estimate of the cross-validated R^2 were similar; the largest difference in bias was only .06 standard deviations (Cohen's *d*). Thus, consistent with Raju et al. (1999), a formula-based estimate of the cross-validated R^2 is as accurate as a sample-based cross-validation study.

For the predictor selection condition, sample cross-validation was more accurate than Burket$_{(s)}$, with Cohen's *d* values ranging from .35 to .71. Sample cross-validation

was also more accurate than Burket$_{(hyb)}$, but only for D1 (d = .32). As the number of predictors increased from 5 to 10 (and beyond), the difference between the two techniques was trivial (ds < .10). Finally, consistent with Schmitt and Ployhart (1999), Burket$_{(hyb)}$ exhibited less bias than Burket$_{(s)}$ for datasets D1 and D2 (ds ranged from .39 to .64).

Table 3

Effect Size Estimates for Differences in Bias

Comparison	Cohen's d		
	D1	D2	D3
Burket$_{(ns)}$ vs. $R^2_{C(ns)}$	0.055	-0.013	-0.029
Burket$_{(s)}$ vs. $R^2_{C(s)}$	0.356	0.474	0.711
Burket$_{(hyb)}$ vs. $R^2_{C(s)}$	0.318	0.075	0.016
Burket$_{(s)}$ vs. Burket$_{(hyb)}$	0.035	0.393	0.644

Discussion

The purpose of the current study was to determine whether the accuracy of estimates of the cross-validated R^2 differed between empirical and formula-based methods when predictors are selected via forward selection. The results of the study found that when predictor selection is performed, a sample-based cross-validation is superior to a Burket$_{(s)}$ (i.e., the Burket adjustment to the sample squared multiple correlation computed with both $k_{(ns)}$ and $R^2_{(ns)}$) estimate of the cross validated R^2 across all conditions. However, when predictor selection is performed, a sample-based cross-validation is superior to a Burket$_{(hyb)}$ (i.e., the Burket equation computed with both $k_{(ns)}$ and $R^2_{(ns)}$) estimate of the cross validated R^2 only when there are five predictors, most of which are useful (on average, 80% of the five predictors were selected). For situations in which there are many predictors, most of which are not useful (on average 30% or fewer of the predictors in the 10 and 15 predictor datasets were selected), Burket$_{(hyb)}$ is as accurate as a sample-based cross-validation and is more accurate than Burket$_{(s)}$. Thus, Burket$_{(hyb)}$ should be preferred to a sample-based cross-validation unless there are very few predictors, most of which are retained.

When predictor selection is not performed, Burket's equation provides an accurate estimate of the cross-validated R^2. Estimates from Burket's equation are as accurate as a sample-based cross-validation study. These findings are consistent with the results found by Raju et al. (1999). Given the vast difficulty of obtaining a second sample for a sample-based cross-validation as well as the inherent problems with sample splitting techniques (Murphy, 1983), the Burket equation should be the preferred method.

Limitation and Future Research

Considering this study only examined one sample size, 150, it is unclear whether the empirical and formula-based estimates of the cross-validated R^2 would produce similar results at other samples sizes. This is a possible limitation because smaller samples sizes lead to an increase in sampling error. Therefore, it is recommended that future studies replicate this study using various samples sizes.

In order to allow for better comparisons between the datasets, all of the predictor variables were set to have intercorrelations of .3. Future studies may want to replicate this study with different intercorrelations because stronger intercorrelations may lead to erroneous predictors being selected. This study also used a population multiple correlation between the criterion and predictor variables of .5 across all datasets. In doing so, it led to the five predictor model retaining more predictors (as a percent of the predictors) than the 10 and 15 predictor models. (The five predictor model had four predictors with correlations greater than .20, whereas the 10 and 15 predictor models had two and one, respectively.). Future studies should be conducted to test different multiple correlations across several datasets containing five predictor models.

This study examined three variations in the subjects to predictor ratio (i.e., *N*:*k* ratio), 10:1, 15:1, and 30:1. It is worth noting that the only case in which Burket$_{(hyb)}$ exhibited substantially greater bias than a sample-based cross-validation was when the subjects to predictor ratio was 30:1. Future studies should test to see if these results would hold true as the *N*:*k* ratio varies. Furthermore, whereas the *N*:*k* rule of 10:1 is better than no guideline at all, it still leads to inefficiencies in determining the desirable *N* size (Green, 1991). Green (1991) suggested that it would be more appropriate to conduct a

power analysis to determine the appropriate *N*. After conducting a power analysis, researchers will be able to determine the sample size required to detect a given effect size within a given degree of confidence. Perhaps future studies could incorporate power analyses to determine how estimates of ρ_c^2 are affected.

Finally, Raju et al. (1999) suggested that when compared to equal weights models, OLS models are more prone to overfitting due to their use of optimal weighting. It is unknown if an equal weights procedure, combined with predictor selection, cross-validates as well as an optimal weighting procedure. Researchers should consider addressing this issue in a future study.

Conclusion

In summary, when predictor selection is performed, a sample-based cross-validation is superior to a Burket estimate of the cross validated R^2 when there are only five predictors, most (on average 80%) of which are useful. For situations in which there are many predictors, very few (on average, a maximum of 30%) of which are useful, Burket$_{(hyb)}$ is as accurate as a sample-based cross-validation and is more accurate than Burket$_{(s)}$. Thus, Burket$_{(hyb)}$ should be preferred to a sample-based cross-validation, unless there are very few predictors, most of which are retained. In addition, when all predictors are retained, the Burket equation estimates the cross-validated R^2 as well as a sample-based cross-validation study. Given the costs associated with a sample-based cross-validation study and the efficiency of the Burket estimators in most situations, there are strong reasons to prefer them over the sample-based effort.

References

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66,* 411-421. doi:10.1097/00006842-200405000-00021

Burket, G. R. (1964). A study of reduced rank models for multiple prediction. *Psychometric Monograph, 12,* 1-66.

Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology, 65,* 407-414. doi:10.1037//0021-9010.65.4.407

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

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499-510. doi:10.1207/s15327906mbr2603_7

Murphy, K. R. (1983). Fooling yourself with cross-validation: Single sample designs. *Personnel Psychology, 36,* 111-118. doi:10.1111/j.1744-6570.1983.tb00507.x

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth: Wadsworth/Thomson Learning.

Punch, K. (2009). *Introduction to research methods in education*. Los Angeles: Sage.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement, 23,* 99-115. doi:10.1177/01466219922031220

Saks, M. & Allsop, J. (2007). *Researching health: Qualitative, quantitative, and mixed methods*. Los Angeles: SAGE Publications.

Schmitt, N. & Ployhart, R. E. (1999). Estimates of cross-validity for stepwise regression and with predictor selection. *Journal of Applied Psychology, 84,* 50-57. doi:10.1037//0021-9010.84.1.50

Sekaran, U. & Bougie, R. (2013). *Research methods for business: A skill-building approach*. Chichester: Wiley.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). New York: APA.

Spatz, C. & Kardas, E. P. (2008). *Research methods in psychology: Ideas, techniques, and reports*. Boston: McGraw-Hill.

APPENDIX A:
Dataset D1: Intercorrelations for Predictors and Criterion

	x1	x2	x3	x4	x5
x1	1.0				
x2	.30	1.0			
x3	.30	.30	1.0		
x4	.30	.30	.30	1.0	
x5	.30	.30	.30	.30	1.0
y	.10	.26	.31	.33	.40

APPENDIX B:
Dataset D2: Intercorrelations for Predictors and Criterion

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	1.0									
x2	.30	1.0								
x3	.30	.30	1.0							
x4	.30	.30	.30	1.0						
x5	.30	.30	.30	.30	1.0					
x6	.30	.30	.30	.30	.30	1.0				
x7	.30	.30	.30	.30	.30	.30	1.0			
x8	.30	.30	.30	.30	.30	.30	.30	1.0		
x9	.30	.30	.30	.30	.30	.30	.30	.30	1.0	
x10	.30	.30	.30	.30	.30	.30	.30	.30	.30	1.0
y	.00	.05	.05	.10	.10	.10	.15	.19	.30	.40

APPENDIX C:
Dataset D3: Intercorrelations for Predictors and Criterion

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15
x1	1.0														
x2	.30	1.0													
x3	.30	.30	1.0												
x4	.30	.30	.30	1.0											
x5	.30	.30	.30	.30	1.0										
x6	.30	.30	.30	.30	.30	1.0									
x7	.30	.30	.30	.30	.30	.30	1.0								
x8	.30	.30	.30	.30	.30	.30	.30	1.0							
x9	.30	.30	.30	.30	.30	.30	.30	.30	1.0						
x10	.30	.30	.30	.30	.30	.30	.30	.30	.30	1.0					
x11	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	1.0				
x12	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	1.0			
x13	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	1.0		
x14	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	1.0	
x15	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	1.0
y	-.10	.00	.00	.05	.05	.05	.05	.05	.10	.10	.10	.10	.14	.15	.40