Multicollinearity's Effect on Regression Prediction Accuracy with Real Data Structures

John D. Morris **Mary G. Lieberman**

Florida Atlantic University

Recommendations from popular statistics texts regarding avoidance of predictor variable multicollinearity in the use of multiple regression are considered from the perspective of the alternate purposes of explanation and prediction. As opposed to prior studies that consider the effect of multicollinearity on prediction accuracy by varying a constant proportion eigenvalue decrement, a method for manipulating multicollinearity while maintaining a real data set's eigenvalue structure is used. For 21 data sets examined, it is shown that multicollinearity has no effect in respect to either relative or absolute prediction accuracy.

As mentioned in a prior publication in this journal (Morris & Lieberman, 2015), texts used in multiple regression courses (Belesley, Kuh, & Welch, 1980; Brook & Arnold, 1985; Chatterjee & Price, 1977; Cliff, 1987; Cohen, Cohen, West & Aiken, 2003; Gnanadesikan, 1977; Kerlinger & Pedhazur, 1973; Kleinbaum, Kupper, Nizan, & Rosenberg, 2013; Lomax & Hahs-Vaughn, 2012; Meyers, Gamst, & Guarino, 2006; Pedhazur, 1982; Stevens, 2009; Tabachnick & Fidell, 2012) provide a consistent warning to avoid the dangers of multicollinearity among predictor variables.

However, two different goals in multiple regression modeling may be of interest: prediction or explanation. These have been distinguished elsewhere (Kerlinger, 1973, p. 9-10; Kerlinger & Pedhazur, 1973, p. 48-49; Morris & Lieberman, 2015). Briefly, the distinction is that, in the case of explanation, one's interest is in parameter estimation, whereas in prediction, interest is in model accuracy. Consideration of the potentially different effect of multicollinearity on prediction and explanation analyses does not appear to be considered in these mainstream texts.

The aforementioned recommendations regarding the **prediction** performance of regression models over a wide span of multicollinearity conditions has been examined (Morris & Lieberman, 2015). Multicollinearity depends upon the structure of the predictor variable intercorrelation matrix (**R**) eigenvalues (λ). The λs of **R** will, other than with artificially orthogonalized experimental data, necessarily decrease from first to last, but it is the severity of that decrease that determines multicollinearity. It is this dependence of multicollinearity on severity of eigenvalue decrement that Darlington (1978) used, and that was also used in the subsequent examination mentioned, to manipulate multicollinearity. As originally manipulated by Darlington, eigenvalues were set to decrease with a constant proportion (e.g., $\lambda_2/\lambda_1 = \lambda_3/\lambda_2 = \lambda_4/\lambda_3$ … etc. = .50). Such "eigenvalue-ratios" of .30, .40, .50, .65, .80, and .95 have been examined in a variety of studies (Morris, 1982; Morris & Huberty, 1987; Morris & Lieberman, 2015). Results have been that OLS regression is insensitive to multicollinearity if the criterion is prediction accuracy. Surprise, and some degree of disbelief, regarding these results have been evident, which is consistent with the advice in the aforementioned texts.

Although the predictor variable λs are necessarily a monotonically decreasing function, they do not necessarily decrease with a constant proportion, nor according to any other prescribed mathematical pattern; they decrease, in any particular data set, as nature "decides." The purpose of this study was to extend these results to the λ decrement structures of real data.

Theoretical Framework

Several indices of multicollinearity are currently in use, including the Variance Inflation Factor (*VIF*, and its redundant inverse, the "Tolerance") the *MI* indices (Thisted & Morris, 1980), and Condition Indices (Belesley, 1991). The Condition Index (*CI*) will be of principal use herein. For the entire prediction problem, it is defined as:

$$CI = \sqrt{\lambda_{max}}/\sqrt{\lambda_{min}}, \quad (1)$$

where λ_{max} and λ_{min} are the largest and smallest λs of **R**. As the first λ is necessarily the largest, and the last λ is necessarily the smallest from a full-rank **R**, for a *p* variable predictor **R,** we can equivalently use the definition of:

$$CI = \sqrt{\lambda_1}/\sqrt{\lambda_p} \quad (2)$$

To further investigate the effect of multicollinearity on regression performance with arguably more realistic characteristics, the present study used real data to determine the eigenvalue structure, rather than prescribing a constant proportional decline. Of course, to examine the effect of multicollinearity on prediction accuracy, it must be manipulated, but a specific data set has only one degree of multicollinearity, regardless of how one measures it. So, in order to manipulate multicollinearity, but still, as much as possible, maintain the λ "scree" pattern of decrement of the real data, the following mechanism was used.

To make multicollinearity larger, in respect to that manifested in the original data, a portion, X_{shift}, of the sum of the second through p_{th} eigenvalues was added to the first eigenvalue, and to lessen multicollinearity, a portion of the first eigenvalue was distributed evenly across all other (second through p_{th}) eigenvalues. This way, rank was maintained, and multicollinearity was manipulated from that of the original data, but, as the "shift" of component variance from first to remaining, or remaining to first, λ was constant, the pattern of λ decrease was parallel to that of the original data from component 2 through *p*, thus closely mimicking the real data structure. That is, in using this multicollinearity manipulation strategy, the only λ for which the "scree" is not parallel to that of the original data is λ_1, wherein multicollinearity was manipulated. Component-criterion correlations were maintained as in the original data, thus, in turn, the data set R^2 was maintained.

Method

To cover a very wide multicollinearity range, the *CI* was increased or decreased through several levels from that of the native data. One can show that the variance proportion, X_{shift}, needed for the movement of eigenvalue variance from, or to, λ_1, to or from, the remaining *p*-1 λs to achieve any desired *CI* is given by:

$$X_{shift} = (p - 1)(\lambda_1 - \lambda_p CI^2)/(\lambda_1(CI^2 + p - 1)), \quad (3)$$

with all symbols previously defined. Then, to achieve the desired *CI*, λs are modified as:

$$\lambda_1(\text{modified}) = \lambda_1 - X_{shift}\lambda_1, \text{ and} \quad (4)$$

$$\lambda_i(\text{modified}), (i>1) = \lambda_i + X_{shift}\lambda_1/(p\text{-}1) \quad (5)$$

If the desired *CI* is greater than that of the original data, X_{shift} becomes negative (thus moving component variance to λ_1 from the remaining *p*-1 λs) and positive (doing the exact opposite) when the desired *CI* is less than that of the original data. Recognizing that the λs must decrease in magnitude gives rise to an upper bound for X_{shift}, and a corresponding lower bound for an accomplishable *CI* as:

$$CI_{lb} = \sqrt{(\lambda_1 + \lambda_2(p - 1))}/\sqrt{(p\lambda_p + \lambda_1 - \lambda_2)} \quad (6)$$

To illustrate, a plot (see Figure 1) is included for a well-known simple four predictor variable data set from Kerlinger and Pedhazur (1973, p. 292). The original data manifests a CI_o of 2.5 – trivial collinearity. The decline in λs is: from λ_1 to λ_2, 32%; from λ_2 to λ_3, 81%; and from λ_3 to λ_4, 60%. Note that after the decrement from λ_1 to λ_2, the proportion of decline lessens dramatically, as opposed to the aforementioned simulation studies wherein there is a constant proportion decline. Another way to look at this is that in those previous studies the most and least severe degrees of collinearity considered were represented by constant proportions of .30 and .95, respectively, both of which are approximated in this real data set. The purpose then, herein, is to allow the natural decline in eigenvalues to obtain to more closely approximate real data conditions.

With the formula mentioned above, one can add or subtract a portion, X_{shift}, of λ_1 to, or from, the second through fourth λs to obtain whatever *CI* one wishes for this data set. As the latter (2 through *p*) λ decline is mapped as it is in the original data, any concern over the artificial nature of the simulated constant proportion decrease of λs in previous studies should be alleviated. Figure 1 shows the original λ decline pattern (CI_o=2.5), and the decline calculated via the X_{shift} that gives rise to a *CI* of 1.5 (thus, less collinearity than that of the original data) and 30 (far more collinearity than that of the original data). The pattern of decline is, as one can see, necessarily perfectly mapped regardless of *CI*; the plots will always be perfectly parallel to the pattern of the real data for λs 2 through *p*, regardless of native pattern or *p*.

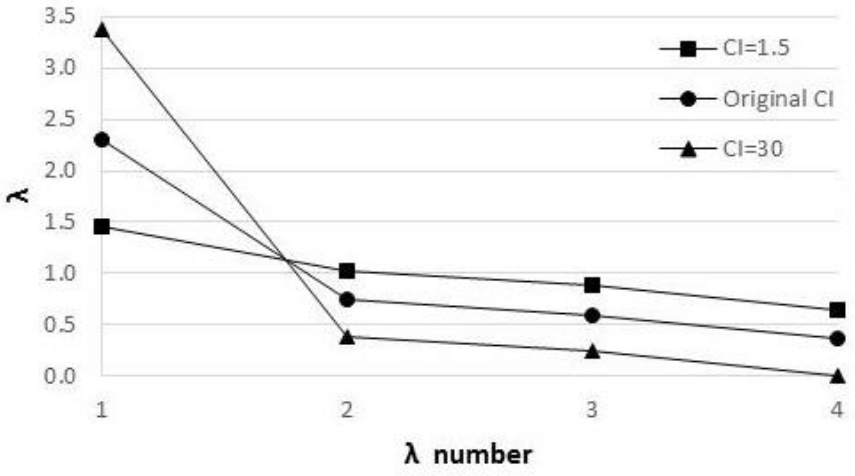

Figure 1. λs for original and modified Kerlinger and Pedhazur data.

Data Source

For each data set, multicollinearity was manipulated over a **very** wide range, up and down from the original CI_o (also presented). The smallest target CI_{low} was 1.5. As a *CI* of 1.0 represents predictor variable orthogonality, a *CI* of 1.5 represents **very** low collinearity. For some data sets, a *CI* of 1.5 was not possible due to the aforementioned lower limit, and in those cases the approximate lowest *CI* available was calculated in 10% increments from 1.5, as 1.65, 1.8, 1.95, etc, until an obtainable CI_{low} was achieved. In addition, *CIs* of 15, 30, 75, and 150, representing a 10, 20, 50 and 100-fold increase, respectively, were used. A *CI* of 30 is often posited as a basis for concern. In addition, a *CI* of 150 was chosen as the infamous Longley (1967) data -- the standard-bearer for multicollinearity -- manifests a *CI* of about 110. Thus, even greater multicollinearity is included for all data sets in this study than that of the Longley data.

For each real data set, a population of 10,000 subjects was created (Morris, 1975; Morris, 1982) that manifested each of the desired *CIs*, with the eigenvalue decline pattern created from the original data by the appropriate X_{shift} for each *CI*.

Samples of the original size represented in the data set were selected. The sample regression weights were then cross-validated by using them to predict the criterion for all 10,000 population subjects; this was replicated 10,000 times with the mean performance presented. Both relative accuracy, as measured by the cross-validated correlation between predicted and actual criterion score, $R_{cv,}$ and absolute accuracy, as measured by the MSE_{cv} $[\Sigma(Z_y-\hat{y})^2/n]$ (predicting the standardized Z_y so that results were scaled similarly across data sets), were included.

A Fortran 90 computer program compiled by Intel Parallel Studio XE 2018 was used to accomplish all simulations. The random normal deviates required were created by the "Rectangle-Wedge-Tail" method (Marsaglia, MacLauren, & Bray, 1964), with the required uniform random numbers generated by the "shuffling" Algorithm M recommended by Knuth (1969, p. 30). Dolker and Halperin (1982) found this combination to perform most satisfactorily in creating random normal deviates.

Results

The source for data were the original 21 data sets used in a comparison of OLS and ridge regression (Morris, 1986). Those data sets seem appropriate herein as they were selected to vary widely in respect to multicollinearity, n/p and R^2. The program, however, is completely general, thus can be used with any real data set. Citations for the data set source for those that are published are provided; the three data sets that are not published are available from the authors.

Figure 2 illustrates regression performance for the aforementioned Kerlinger and Pedhazur data (1973). [The CI_o of the original data is designated with an *.] As well, *VIF* ranges for the *CI* levels are also provided. Multicollinearity was varied extremely via the CIs as specified, and if you prefer *VIFs*, you can see that they ranged from representing near orthogonality (< 1.2) in the *CI*=1.5 condition, to VERY extreme multicollinearity (2822) in the *CI*=150 condition. Regarding the effect of multicollinearity on prediction accuracy, as can be seen, relative (R_{cv}) and absolute (MSE_{cv}) accuracy were totally unaffected

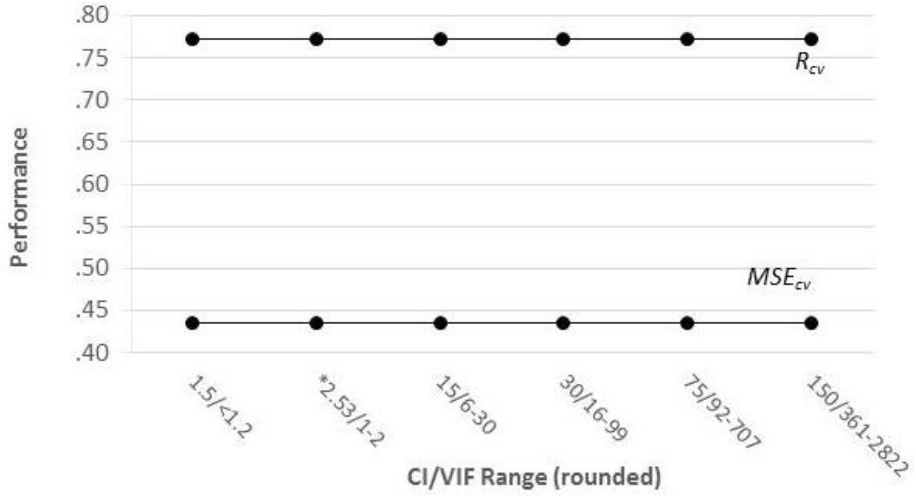

Figure 2. R_{cv} and MSE_{cv} performance by *CI*/*VIF* for Kerlinger and Pedhazur data.

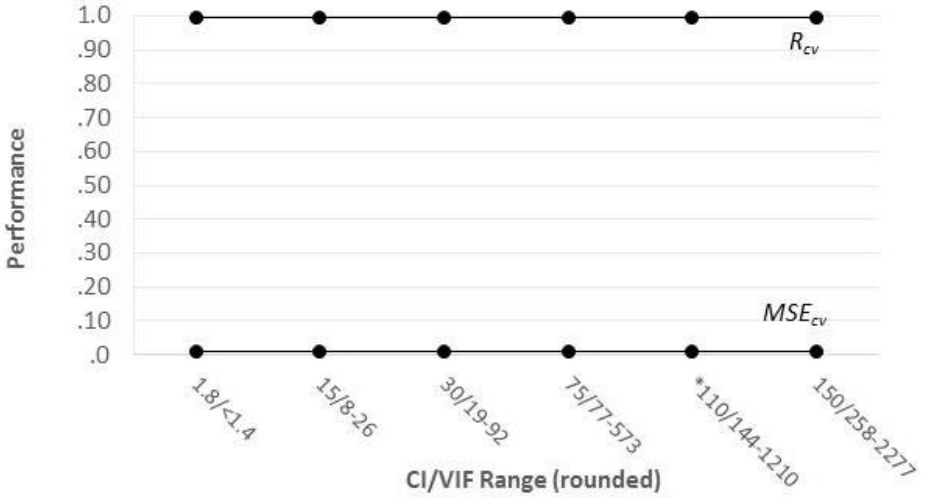

Figure 3. R_{cv} and MSE_{cv} performance by *CI*/*VIF* for Longley data.

by multicollinearity – flat with no trend. Not only is this visually apparent from the Figure, but the mean R_{cv} varied across multicollinearity conditions by no more than .0001, and to four significant digits, there was zero variance for MSE_{cv}.

Figure 3 illustrates the same information for the aforementioned Longley (1967) data. These data were posited by Longley as constituting a regression problem so ill-conditioned as to be a challenge for digital computers to solve. With a native CI_o of around 110, it can be seen from the Figure that the *VIFs* of the original data vary from about 144 to 1210. Through the mechanism mentioned, *CIs* were manipulated, as in the former data set, from near orthogonality, with *CI*=1.8, and *VIFs*<1.4, to even more collinearity than in the original Longley data, with *CI*=150, and *VIFs* that ranging to 2277. Figure 3 shows the same visual effect – accuracy was flat with no trend in respect to multicollinearity. In addition, neither R_{cv} nor MSE_{cv} varied by as much as .0001 as multicollinearity increased. [It may appear from the plot that the R_{cv} is "1" and MSE_{cv} is "zero," but the respective values across all six multicollinearity conditions were respectively .9965 and .0084.]

The results for the remaining 19 data sets are presented in tabular form (see Table 1) as the results are essentially identical. There, the data characteristics are included (p, *n*, R^2, CI_o) as well as the R_{cv} and MSE_{cv}. As can be seen, the CI_{low} obtainable was often 1.5, and when not, close enough to afford little collinearity, except in data set #18. The eigen-structure of this data set is unusual in the proximity of λ_1 and λ_2 (3.29 and 3.16, respectively) producing a CI_{low} of around 13. For all other data sets, a CI_{low} was attainable that none would argue represents collinearity. As well, for data sets # 13, 14 and 15, the CI_o was smaller than 1.5, thus the CI_o was used as the CI_{low}. In all cases, the remaining *CIs*, representing an increase to extreme collinearity (15, 30, 75, and 150), were used.

For all data sets, R_{cv} did not vary in more that the ten-thousandths decimal place across multicollinearity conditions, most often not even there. MSE_{cv} was even more stable with only one data set showing any variance, even in the ten-thousandths place. Regarding such fluctuations, they were random given the simulation; for the conditions that did manifest these very small differences, the largest collinearity condition was as likely to achieve greater accuracy as less. Therefore, the R_{cv} and MSE_{cv} figures (displayed to thousandths precision) in Table 1, represent the accuracy across all multicollinearity conditions. Thus, further strong evidence is afforded regarding the warnings regarding multicollinearity leveled by previously mentioned texts; they do not apply to the objective of **prediction**, whether considering a constant eigenvalue decline as has been previously documented, or that of real data structures as is included in this paper.

Table 1. Data Sets, Characteristics and Cross-validation Performance

Data Set Number and Description	*p*	*n*	R^2	CI_o	CI_{low}	R_{cv}	MSE_{cv}
1 Marquardt & Snee Acetylene (1975)	3	16	.92	7.0	1.80	.952	.108
2 Chew & Morris Lollipop (1984)	5	293	.94	24.6	1.95	.969	.062
3 Hoerl Kansas Corn Yield (1982)	6	51	.80	4.7	2.10	.882	.233
4 Draper & Smith Chemical (1966, p. 204)	3	21	.91	3.2	1.50	.951	.107
5 Drehmer & Morris Example (1981)	9	14	.82	15.3	3.45	.751	.793
6 Golf Score from Task Perf	4	120	.84	1.9	1.50	.915	.165
7 Draper & Smith Chemical (1966, p. 366)	4	13	.98	37.3	3.30	.988	.030
8 Hocking & Dunn Example (1982)	3	20	.62	11.5	1.50	.757	.478
9 Hoerl & Kennard Example (1981)	5	15	.99	18.2	1.50	.990	.024
10 Kerlinger & Pedhazur GPA (1973, p. 292)	4	30	.64	2.5	1.50	.773	.444
11 Longley Collinearity Example (1967)	6	16	.99	110.	1.80	.997	.008
12 Morris, et al. Selection Example (1980)	4	83	.47	8.1	1.50	.674	.560
13 Rulon, et al. Mechanics (1967, p. 399)	3	93	.26	1.3	1.30	.495	.772
14 Rulon, et al. OC Agents (1967, p. 402)	3	66	.32	1.1	1.10	.549	.721
15 Rulon, et al. Pass Agents (1967, p. 396)	3	86	.19	1.4	1.40	.415	.849
16 Retention from Demos & WISC	10	29	.31	4.6	2.55	.395	1.103
17 Piers-Harris from IQ & Ach	7	55	.18	4.2	1.50	.340	.954
18 Draper & Smith Steam (1966, p. 352)	9	25	.92	28.0	13.0	.941	.131
19 Draper & Smith Chemical (1966, p. 233)	4	16	.82	3.9	1.50	.876	.275
20 Cooley & Lohnes Talent (F) (1971, p. 353)	12	271	.33	6.1	1.80	.552	.703
21 Cooley & Lohnes Talent (M) (1971, p. 349)	12	234	.41	7.8	1.95	.619	.624

Note. From Morris, J. D. (1986), *Educational and Psychological Measurement, 46,* 853-867.

References

Belesley, D. A. (1991). *Conditioning diagnostics: Collinearity and weak data in regression*. New York: Wiley.

Belesley, D. A., Kuh, E., & Welch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.

Brook, R. J., & Arnold, G. C. (1985). *Applied regression analysis and experimental design*. New York: Dekker.

Chatterjee, S., & Price, B. (1977). *Regression analysis by example*. New York: Wiley.

Chew, A. L., & Morris, J. D. (1984). Validation of the Lollipop Test: A diagnostic screening test of school readiness. *Educational and Psychological Measurement, 44*, 987-991.

Cliff, N. (1987). *Analyzing multivariate data*. Orlando: Harcourt.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cooley, W. W, & Lohnes, P. R. (1971). *Multivariate data analysis.* New York: Wiley.

Darlington, R. B. (1978). Reduced variance regression. *Psychological Bulletin, 85*, 1238-1255.

Dolker, M., & Halperin, S. (1982). Comparing inverse, polar, and rectangle-wedge-tail Fortran routines for pseudo-random normal number generation. *Educational and Psychological Measurement, 42*, 223-236.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis*. New York: Wiley.
Drehmer, D. E., & Morris, G. W. (1981). Cross-validation with small samples: An algorithm for computing Gollob's estimator. *Educational and Psychological Measurement, 41*, 195-200.
Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: Wiley.
Hocking, R. R., & Dunn, M. R. (1982, October). *Collinearity, influential data and ridge regression.* Paper presented at the Ridge Regression Symposium. University of Delaware.
Hoerl, A. E. (1982, October). *Computational examples: Kansas corn yield*. Paper presented at the Ridge Regression Symposium. University of Delaware.
Hoerl, A. E., & Kennard, R. W. (1981). Ridge regression – 1980: Advances, algorithms, and applications. *American Journal of Mathematical and Management Sciences*, *1*, 5-83.
Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). New York: Holt.
Kerlinger, F. N., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt.
Kleinbaum, D. G., Kupper, L. L., Nizan, A., & Rosenberg, E. S. (2013). *Applied regression analysis and other multivariable methods* (5th ed.). Boston: Cengage.
Knuth, D. E. (1969). *The art of computer programming* (Vol. 2: Seminumerical algorithms). Reading, MA: Addison-Wesley.
Lomax, R. G., & Hahs-Vaughn, D. L. (2012). *Statistical concepts: A second course* (4th ed.). New York: Routledge.
Longley, J. W. (1967). An appraisal of least-squares programs from the point of view of the user. *Journal of the American Statistical Association, 62*, 819–841.
Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician, 29*, 3-20.
Marsaglia, G., MacLaren, D., & Bray, T. A. (1964). A fast procedure for generating random normal variables. *Communications of the ACM, 7*, 4-10.
Meyers, L. S., Gamst, G., & Guarino A. J. (2006). *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: Sage Publications.
Morris, J. D. (1975). A computer program to create a population with any desired centroid and covariance matrix. *Educational and Psychological Measurement, 35*, 707-710.
Morris, J. D. (1982). Ridge regression and some alternative weighting techniques: A comment on Darlington. *Psychological Bulletin, 91*, 203-210.
Morris, J. D. (1986). Selecting a predictor weighting method by PRESS. *Educational and Psychological Measurement, 46*, 853-869.
Morris, J. D., & Huberty, C. J (1987). Selecting a two-group classification weighting algorithm. *Multivariate Behavioral Research, 22*, 211-232.
Morris, J. D., & Lieberman, M. G. (2015). Prediction, explanation, multicollinearity, and validity concentration in multiple regression. *General Linear Model Journal, 41*, 29-35.
Morris, J. D., Morgan, F. B., & Maynor, W. (1980). On selecting the best set of regression predictors. *Journal of Experimental Education, 48*, 100-103.
Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt.
Rulon, P. J., Tiedeman, D. V., Tatsuoka, M. M., & Langmuir, C. R. (1967). *Multivariate statistics for personnel classification*. New York: Wiley.
Stevens, J. P. (2009). *Applied multivariate statistics for the behavioral sciences* (5th ed.). New York: Routlege.
Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). New York: Pearson.
Thisted, R. A., & Morris, C. N. (1980). *Theoretical results for adaptive ordinary ridge regression estimators* (Tech. Rep. No. 94). Chicago: University of Chicago.

Send correspondence to:
John D. Morris
Florida Atlantic University
Email: jdmorris@fau.edu