AIR Professional File

Association for Institutional Research
Enhancing knowledge. Expanding networks.
Professional Development, Informational Resources & Networking

Number 92, Summer, 2004

Ridge Regression as an Alternative to Ordinary Least Squares: Improving Prediction Accuracy and the Interpretation of Beta Weights

David A. Walker
Assistant Professor of Educational
Research and Assessment
Northern Illinois University

Abstract

This article looked at non-experimental data via an ordinary least squares (OLS) model and compared its results to ridge regression models in terms of cross-validation predictor weighting precision when using fixed and random predictor cases and small and large p/n ratio models. A majority of the time with two random predictor cases, ridge regression accuracy was superior to OLS in estimating beta weights. Thus, ridge regression was very useful under this condition. However, when the fixed predictor case was reviewed, OLS was much more precise at estimating predictor weights than the ridge techniques regardless of the p/n ratio. In determining the cross validation accuracy of the ridge estimated weights in respect to the OLS estimated weights, ridge models were superior for improving the accuracy of model prediction.

Introduction

Ridge regression is not a new idea within the education field. It has been applied as a non-ordinary least squares (OLS) alternative predictor weighting technique. However, ridge regression analyses within educational research appear to be sporadic. The current study is not intended to argue in support of or against ridge regression. This goal was accomplished in the literature (Darlington, 1978; Dempster, Schatzoff, & Wermuth, 1977; Hoerl & Kennard, 1970; Kennedy, 1988; Laughlin, 1978; Morris, 1982; Pagel & Lunneborg, 1985; Rozeboom, 1979). This article looks at non-experimental data via an OLS model and compares its results to ridge models in terms of cross-validation predictor weighting precision when using fixed and random predictor cases and small and large p/n ratio models (i.e., p = the number of predictors and n = the number of observations). A supplementary, pervading function of this article is to initiate, or elucidate, a conversation with faculty, practitioners, and graduate students concerning some of the fundamentals of ridge regression.

Research Questions

There appears to be a void in the literature pertaining to performance comparisons of OLS and one-parameter ridge regression models using both fixed and random predictor cases. This article is intended to fill the chasm in the educational literature concerning the prediction accuracy and beta weight estimation performance of these two models by answering the following questions:

1. If the ridge technique is an improvement from the OLS model in terms of accuracy of model prediction, what is the magnitude of the absolute gain of the improvement when the ridge estimate has a large p/n ratio and a small p/n ratio?

2. If cross validation accuracy of the ridge estimated weights compared to the OLS estimated weights is established, then, when examining fixed regressor case(s) and random regressor case(s), can the ridge technique out perform OLS concerning the estimation of the importance of population beta weights?

Review of the Literature

OLS

The OLS regression models are conducted to identify independent variables that yield the most parsimonious variable, x are the independent variables, and b are the regression coefficients (Gall, Borg, & Gall, 1996).

However, OLS estimates of beta weights have been found untrustworthy in the presence of multicollinearity (Cohen & Cohen, 1983). Multicollinearity is caused by highly correlated independent variables or by variables that are nearly linearly dependent, which do not provide exclusive information to explain the model (Cohen & Cohen, 1983). In addition, multicollinearity can produce high standard errors and imprecise parameter estimates, which can abate the stability of a model and its prediction power (Kidwell & Brown, 1982). Darlington (1968) found that the presence of multicollinearity causes variance to increase in standardized coefficients, which diminishes the power of a statistical test. Tate (1988) reiterated the matter of substandard statistical power by adding that multicollinearity is a concern particularly with non-experimental designs, causing regression beta coefficients to have inflated standard errors.

Furthermore, the coefficient of determination (R^2), which is the percentage of variance in the dependent variable explained by the linear combination of diverse weightings of predictor variables, has been found in OLS models to overestimate model effect sizes when R^2 is $\leq .80$. This overestimation of the internal accuracy of the sample squared multiple correlation causes a miscalculation, in a bias upward, of the population value, thus overrating the regression equation effectiveness in the population and future samples (Agresti & Finlay, 1997; Morris & Meshbane, 1995; Pedhazur, 1997). The diminished predictive accuracy of a regression equation has been termed "validity shrinkage" or the propensity for correlations, specifically the squared multiple correlation, to decrease when a regression equation is replicated in another research study (Gall et al., 1996; Synder & Lawson, 1993).

Ridge Regression

Ridge regression is a method that attempts to render more precise estimates of regression coefficients and minimize shrinkage, than is found with OLS, when cross-validating results (Darlington, 1978; Hoerl & Kennard, 1970; Marquardt & Snee, 1975). As Faden and Bobko (1982) stated, "The technique of ridge regression is considered as a device which may limit validity shrinkage, while maintaining absolute levels of predictability which are higher than that of OLS regression" (p. 73). As with OLS, ridge regression produces an "R^2" statistic, which is not the usual R^2 found in OLS, but rather the percentage of criterion variance accounted for by the full and reduced models of interest using the biased ridge weights (Morris, 1983).

To calculate ridge weights, Hoerl and Kennard (1970) recommended that a biased ridge estimator

$$\beta^* = (R_{xx} + kI)^{-1}R_{xy}$$

β^* = the vector of standardized ridge regression weights
R_{xx} = predictor intercorrelation matrix
R_{xy} = predictor criterion correlation vector
I = p-dimensional identity matrix
k = a biasing parameter (typically $0 < k < 1$)

be employed to diminish the error influence, for example, introduced through multicollinearity or minute validity coefficients, between sample estimates and population weights thus producing estimates with smaller MSE than found with a typical OLS estimator, where $b = (R_{xx})^{-1}R_{xy}$ (Faden & Bobko, 1982; Kennedy, 1988). In addition, ridge regression can be considered a penalization technique where an optimum, biasing parameter (k) or "penalty factor" is added to the variance/covariance matrix preceding the calculation of the regression equation to yield the lowest MSE for the equation, less multicollinearity with predictors, and a better fitting model in terms of prediction power (Darlington, 1978). It should be noted that k is solved for iteratively until the MSE is minimized using a Newton-Raphson minimization algorithm.

Ridge regression is not a panacea for estimating the importance of beta weights or selecting the exact degree of shrinkage, and is a biased estimate that, periodically, may not be correlated favorably with the population parameters (Morris, 1982; Pagel & Lunneborg, 1985; Rozeboom, 1979). Yet, many times it displayed an ability to reduce multicollinearity in the inverted matrix and provides better predictive power than OLS (Barker & Brown, 2001; Pasternak, Schmilovitch, Fallik, & Edan, 2001). It should be mentioned, though, that ridge regression is not the only shrinkage method used as an option to OLS. Principal-components analysis and partial least squares regression (PLS) are noted techniques that have incurred mixed results (Butler & Denham, 2000; Foucart, 2000; Jonathan, Krzanowski, & McCarthy, 2000).

Method

Methodologically this study was not intended to compare OLS to ridge regression under all possible conditions, but to factor into the research known elements that affect results such as sample size and distributional asymmetry. Further, this study will elaborate more on which predictor weighting procedure affords the greatest absolute increase in prediction accuracy and is more appropriate, a majority of the time, rather than the traditional discussion limited to determining the most "efficient" predictors of a particular criterion.

Thus, when considering the unique contribution of each variable to a model, in the sense of partial slope, ridge regression is viewed from a different perspective than the traditional OLS model, which determines if a variable adds to the predictive accuracy given the remaining variables in the model. The manner in which this issue should be considered is by looking at the difference in R^2s between all of the models (Morris, 1983). Therefore, it is noted that the subsequent software used in this study to perform

ridge regression does not produce statistics such as t-tests or the standard error of beta because researchers usually are trying to determine the cross validation accuracy of the ridge estimated weights in respect to the OLS estimated weights and, thus, the criterion of performance is for the total model. If cross validation accuracy of the ridge and OLS estimated weights is not the intent of a study, but statistics, such as the standard error of beta, t-tests, and tests of significance, are of interest to assist in answering a specific question when conducting research via a ridge regression, FORTRAN programs are available to calculate these (cf. Morris, 1983; Morris, 1986).

Instrument

Data for this article come from the 1999-2000 National Association of Student Personnel Administrators (NASPA) Survey implemented during the fall of 1999 and conducted biennially. This study focused on four-year public and private institutions and extracted data related only to senior-level administrators at higher education institutions from the larger NASPA data set.

Sample

Participants included student affairs administrators at NASPA member institutions. Surveys were mailed to 1,198 United States higher education institutions. Respondents returned 419 surveys, a 35% response rate. Although the current response rate is about 10 to 15% lower than in previous years, the overall sample is very representative demographically of past NASPA populations (i.e., a similar sample composition) (NASPA Research Division, 1996; 1998).

Variables

For this exploratory study, the dependent variable was respondent salary (SAL). The independent variables were: age of respondent (AGE), length of time the respondent has been employed in his or her current position (POS), and the length of time the respondent has been employed at the institution (INS). The SAL, POS, and INS variables were reported as continuous variables (i.e., random regressor cases). AGE was coded as an ordinal variable where 1 = 21 to 25, 2 = 26 to 35, 3 = 36 to 45, 4 = 46 to 55, and 5 = 56 to 70 (i.e., fixed regressor case).

Population Parameters

During the course of the last 20 to 25 years that this survey has sampled its population of interest biennially, characteristic population parameters for the three continuous variables, SAL, POS, and INS, have been established (NASPA Research Division, 1996; 1998; 2000). For personnel in charge of counseling services, the population parameters for SAL extend from a minimum of $10,528 to a maximum of $143,472. For the variable POS, parameters are from 0 to 42 years. Finally, for the variable INS, the parameters are from a minimum of 0 to a maximum of 44 years.

Variance Inflation Factor

The premise of the variance inflation factor (VIF) is based on the fact that multicollinearity causes the variance of regression coefficients to increase, which in non-experimental research such as the current study's design, produces regression beta coefficients to have inflated standard errors (Darlington, 1968; Tate, 1988). Thus, the identification of multicollinearity within models can be detected through the VIF.

It is at the discretion of the researcher concerning how much VIF to tolerate before considering the presence of multicollinearity. In the present study, VIF values > 2.000 were deemed to be multicollinear. Therefore, multicollinearity is known to be present with POS (private only) and INS (public and private).

Distribution

A series of boxplots and histograms indicated that the dependent variable SAL was distributed normally and there were no influential observations for the 162 respondents from public institutions and the 122 from private institutions who were operationalized as having major responsibility for the area of student affairs termed "counseling services."

Limitations

The VIF cut points for determining multicollinearity within regression models are at the discretion of the researcher. It is understood that there may be honest disagreement with this study's choice of a VIF cut point established at > 2.000. Dually, it is noted that using a continuous variable, such as AGE as an ordinal, fixed regressor, does sacrifice some of the variance within this variable (Gall et al., 1996).

Code

The SPSS (Statistical Package for the Social Sciences) code used for the current research was a macro program for ridge regression. Note that the variables are particular to this research and will change with your data set. A variant of this macro can be downloaded from http://pages.infinit.net/rlevesqu.

```
INCLUDE 'C:\Program Files\SPSS\Ridge regression.sps'.
RIDGEREG DEP=counsal /ENTER = counsf to counsh
/DEBUG='Y'
/START=0 /STOP=1 /INC=0.05.
```

In addition, see Appendix A for the complete ridge regression syntax version provided in SPSS software for personal computer use (SPSS, 2002). When accessing SPSS, go to File, Open, Other, and then find Ridge Regression.

Analyses

The predictability of SAL for counseling services directors at both public and private institutions was studied through the independent variables AGE, POS, and INS. An OLS model was conducted separately for public (n = 162) and private institutions (n = 122). It is important when fitting models for prediction accuracy to "resample" via a cross-validation technique to confirm the results indicated initially and also to ascertain estimates of generalization error. Thus, to determine the amount of model improvement concerning prediction, or lack thereof, the regression equations from the OLS models were compared when cross-validated on two different sets of observations (n = 50 and n = 12). These subsamples were drawn randomly without replacement from the larger sets of data.

The absolute shrinkage value in the R^2 (i.e., $R^2_{sample} - R^2_{population}$) was calculated for both public and private institutions (Faden & Bobko, 1982). It has been noted that in ridge regression, when n is large in comparison to the number of predictors (p), the total gain in prediction accuracy is often very minor, approximately .000 ≤ .010 percentage points, between OLS and ridge estimates. In contrast, when the ratio between n and p is very small, the absolute increase in prediction accuracy can be considerable between the two methods (Dempster et al., 1977; Faden & Bobko, 1982). For the present study, the p (3) to n (50) ratio for one of the cross-validation samples was considered large at 1/17 and the second sample was deemed small at 1/4 (p = 3 and n = 12).

Results and Discussion

As noted in previous studies (Dempster et al., 1977; Kennedy, 1988; Pasternak et al., 2001), the overall utility, in terms of improving the accuracy of model prediction, of the ridge regression technique compared to OLS appears to be warranted. As Table 1 indicates, in every instance, ridge regression, regardless of p/n ratio, surpassed OLS in reducing shrinkage.

Table 1
Validity Shrinkage

	Public				Private			
Model	$R2_{sample}$	90% CI	SE	Absolute Gain	$R2_{sample}$	90% CI	SE	Absolute Gain
OLS	.158	(.069, .240)	(1.483)	-.001	.170	(.065, .265)	(1.481)	.005
Ridge p = 3, n = 50	.161	(.000, .300)	(1.482)	-.002	.199	(.023, .344)	(1.476)	.034
Ridge p = 3, n = 12	.195	(.000, .417)	(1.476)	.036	.221	(.000, .453)	(1.471)	.056
Note: The $R2_{population}$ for Public = .159 (SE = 1.483) and for Private = .165 (SE = 1.482).								

Further, the absolute gain between the ridge models and OLS was consistent with previous studies (Dempster et al., 1977; Faden & Bobko, 1982). At public institutions, the $R^2_{population}$ = .159 and at private institutions the $R^2_{population}$ = .165. The only absolute gain for the OLS model was = .005 at private institutions, while at publics there was no gain, but a loss = -.001.

A discernable trend in the data indicates that for validity shrinkage, the ridge model with a small p/n ratio was superior to the OLS estimators in terms of absolute gains (i.e., .036 and .056 for public and private institutions, respectively). Further, when the p/n ratio was large, the cross-validated ridge estimators proceeded to out perform the OLS R^2 with gains = .002 and .034 at public and private institutions, respectively.

With cross validation accuracy of the ridge and OLS estimated weights established, which concluded that the R^2s of the ridge models surpassed OLS in reducing shrinkage, estimation accuracy will be reviewed. For estimation accuracy, when the VIF is > 2.000, the accuracy of the ridge estimates is a noticeable improvement to OLS in all cases except where the two estimators were equal. For instance, at public institutions, ridge estimates for INS were 1.5 and 1.6 times better than the OLS estimate with n = 50 and n = 12, respectively. However, when the VIF is ≤ 2.000, the OLS estimators always out performed the ridge estimators. At private institutions, the OLS estimator for AGE was 1.4 and 5.9 times better than the ridge estimators with n = 50 and n = 12, respectively. As Kennedy (1988) detected, when the ridge models endured further substandard conditions, for example smaller sample size and higher VIF, the performance of the estimators to the OLS estimators was much more marked.

Pagel and Lunneborg (1985) noted that when the regressor is fixed so that all true values of the predictor can be identified, the performance of ridge regression for approximating specific beta weights should not be the foremost intention of the research. Table 2 shows this condition with the fixed predictor case AGE at public institutions, where the OLS model out performed both ridge regressions in terms of estimating the importance of population beta weights. The OLS model was 2.1 times better than the ridge estimate with the large p/n ratio and 1.5 times better than the ridge model with a small p/n ratio. This tendency also followed for private institutions.

As was found by Kennedy (1988), when the regressors are random, the performance of ridge regression for estimating specific beta weights may be considered as a primary function of the research a majority of the time. The current research illustrates this inclination, but also adds the caveat of having a VIF > 2.000. For example, the predictors INS and POS with a VIF > 2.000 were equal to or appreciably more improved than those in the OLS

Table 2
OLS and Ridge Beta Weight Estimator Accuracy

	Public				Private			
Estimator	OLS Beta Coefficients	Ridge Beta p = 3, n = 50	Ridge Beta p = 3, n = 12	VIF	OLS Beta Coefficients	Ridge Beta p = 3, n = 50	Ridge Beta p = 3, n = 12	VIF
INS	.105	.162 (+1.5)	.164(+1.6)	2.428	.242	.250 (1.0)	.268 (+1.1)	2.587
POS	.056	.036 (-1.6)	.010 (-5.6)	1.839	.005	.044 (+8.8)	.059 (+11.8)	2.055
AGE	.279	.135 (-2.1)	.184 (-1.5)	2.000	.224	.156 (-1.4)	.038 (-5.9)	1.543
Note: In parenthesis, the + or - ratio indicates how many times better, or not, the ridge estimator was in comparison to OLS.								

model. At private institutions, the predictor POS with the small p/n ratio was 11.8 times better than the OLS model, while the large p/n ratio ridge model was 8.8 times better than OLS in estimating specific beta weights.

Future Research

It would be of interest to conduct a simulation-based study very comparable to the current research. This type of study, which was conducted in the research from Morris (1982), would allow for numerous simulated populations to produce a myriad of replications of double-crossed validations for random samples selected from the population of interest. This perspective of cross-validated prediction accuracy would provide more evidence if the ridge technique performs OLS or the contrary.

In terms of another non-experimental study, an extremely large sample of personnel in charge of counseling services at public and private institutions could be drawn from one of the National Center for Education Statistics (NCES) data sets. Several smaller samples from the NCES data set could be drawn to conduct OLS and ridge regression models similar to those in the present study.

Implications

An intention of this study was to emphasize via a research example how ridge regression could be used as an alternative to OLS to address important issues such as multicollinearity, validity shrinkage, and prediction accuracy. The current study used a non-experimental research situation as a mode to examine the overall function of ridge regression and explain some of its subtleties.

The findings confirmed that a major advantage of ridge regression to OLS appears when the research interest lies in interpreting coefficients from random predictor cases, which rendered ridge as the superior of the two techniques (cf. Kennedy, 1988). In addition, the OLS technique provided less than favorable solutions pertaining to prediction accuracy. However, using the same data, ridge procedures yielded more improved accuracy of model prediction (cf. Dempster et al., 1977; Kennedy, 1988). Which method is correct? The interpretation is context driven and within the purview of the researcher. Yet, when confronted with multicollinearity, and validity shrinkage and estimation accuracy are consequential, OLS should not be the definitive technique. As this study observed, an alternative, such as ridge regression, should be implemented as a comparison. Concerning this issue, Price (1977) remarked:

> Application of ridge regression does not necessarily produce the correct answer. However, as an exploratory technique it clearly identifies the presence of multicollinearity problems....[and] suggest[s] directions for further investigation that may not be apparent from the regular least squares solution (p. 765).

References

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Barker, L., & Brown, C. (2001). Logistic regression when binary predictor variables are highly correlated. *Statistics in Medicine, 20*, 1431-1442.

Butler, N. A., & Denham, M. C. (2000). The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society, 62*, 585-593.

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

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

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

Dempster, A. P., Schatzoff, M., & Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association, 72*, 77-91.

Faden, V., & Bobko, P. (1982). Validity shrinkage in ridge regression: A simulation study. *Educational and Psychological Measurement, 42*, 73-85.

Foucart, T. (2000). A decision rule for discarding principal components in regression. *Journal of Statistical Planning and Inference, 89*, 187-195.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics, 12*, 55-67.

Jonathan, P., Krzanowski, W. J., & McCarthy, W. V. (2000). On the use of cross-validation to assess performance in multivariate prediction. *Statistics and Computing, 10*, 209-229.

Kennedy, E. (1988). Biased estimators in explanatory research: An empirical investigation of mean error properties of ridge regression. *Journal of Experimental Education, 56*, 135-141.

Kidwell, J., & Brown, L. (1982). Ridge regression as a technique for analyzing models with multicollinearity. *Journal of Marriage and the Family, 44*, 287-299.

Laughlin, J. E. (1978). Comment on estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 85*, 247-253.

Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician, 29*, 3-20.

Morris, J. D. (1982). Ridge regression and some alternative weighting techniques: A comment on Darlington. *Psychological Bulletin, 91*, 203-210.

Morris, J. D. (1983). Stepwise ridge regression: A computational clarification. *Psychological Bulletin, 94*, 363-366.

Morris, J. D. (1986). Calculating a stepwise ridge regression. *Educational and Psychological Measurement, 46*, 151-156.

Morris, J. D., & Meshbane, A. (1995). Selecting predictor variables in two-group classification problems. *Educational and Psychological Measurement, 55*, 438-441.

NASPA Research Division. (1996). *NASPA salary survey 1995-1996: Comprehensive report.* Washington, DC: Author.

NASPA Research Division. (1998). *NASPA salary survey 1997-1998: Comprehensive report.* Washington, DC: Author.

NASPA Research Division. (2000). *NASPA salary survey 1999-2000: Comprehensive report.* Washington, DC: Author.

Pagel, M. D., & Lunneborg, C. E. (1985). Empirical evaluation of ridge regression. *Psychological Bulletin, 97*, 342-355.

Pasternak, H., Schmilovitch, Z., Fallik, E., & Edan, Y. (2001). Overcoming multicollinearity in near infrared analysis for lycopene content estimation in tomatoes by using ridge regression. *Journal of Testing and Evaluation, 29*, 60-66.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Price, B. (1977). Ridge regression: Application to nonexperimental data. *Psychological Bulletin, 84*, 759-766.

Rozeboom, W. W. (1979). Ridge regression: Bonanza or beguilement? *Psychological Bulletin, 86*, 242-249.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61*, 334-349.

SPSS. (2002). SPSS (Version 11.01). [Computer software]. Chicago: Author.

Tate, R. L. (1988). Ridge regression for interactive models. *Florida Journal of Educational Research, 30*, 15-33.

Appendix A

```
preserve.
set printback=off.
define ridgereg (enter=!charend('/')
        /dep = !charend('/')
        /start=!default(0) !charend('/')
        /stop=!default(1) !charend('/')
        /inc=!default(.05) !charend('/')
        /k=!default(999) !charend('/')
        /debug=!DEFAULT ('N')!charend('/') ).

preserve.
!IF ( !DEBUG !EQ 'N') !THEN
set printback=off mprint off.
!ELSE
set printback on mprint on.
!IFEND .
SET mxloops=200.

*-------------------------------------------------------.
* Save original active file to give back after macro is done.
*-------------------------------------------------------.
!IF (!DEBUG !EQ 'N') !THEN
SET RESULTS ON.
DO IF $CASENUM=1.
PRINT / "NOTE: ALL OUTPUT INCLUDING ERROR MESSAGES HAVE BEEN TEMPORARILY"
/ "SUPPRESSED. IF YOU EXPERIENCE UNUSUAL BEHAVIOR, RERUN THIS"
/ "MACRO WITH AN ADDITIONAL ARGUMENT / DEBUG='Y'."
/ "BEFORE DOING THIS YOU SHOULD RESTORE Y
/ "THIS WILL FACILITATE FURTHER DIAGNOSIS OF ANY PROBLEMS.".
END IF.
!IFEND .

save outfile='rr__tmp1.sav'.

*-------------------------------------------------------
* Use CORRELATIONS to create the correlation matrix.
*-------------------------------------------------------

* DEFAULT:  SET RESULTS AND ERRORS OFF TO SUPPRESS CORRELATION PIVOT TABLE *.
!IF (!DEBUG='N') !THEN
set results off errors off.
!IFEND

correlations variables=!dep !enter /missing=listwise/ matrix out(*).
set errors on results listing .

*-------------------------------------------------------.
* Enter MATRIX.
*-------------------------------------------------------.
matrix.

*-------------------------------------------------------.
* Initialize k, increment, and  number of iterations. If k was not
* specified, it is 999 and looping will occur. Otherwise, just the one
* value of k will be used for estimation.
*-------------------------------------------------------.

do if (!k=999).
. compute k=!start.
. compute inc=!inc.
. compute iter=trunc((!stop - !start ) / !inc ) + 1.
. do if (iter <= 0).
.   compute iter = 1.
. end if.
else.
. compute k=!k.
. compute inc=0.
. compute iter=1.
end if.
*-------------------------------------------------------.
* Get data from working matrix file.
*-------------------------------------------------------.

get x/file=*/names=varname/variable=!dep !enter.

*-------------------------------------------------------.
* Third row of matrix input is the vector of Ns. Use this to compute number
* of variables.
*-------------------------------------------------------.

compute n=x(3,1).
compute nv=ncol(x)-1.

*-------------------------------------------------------.
* Get variable names.
*-------------------------------------------------------.

compute varname=varname(2:(nv+1)).

*-------------------------------------------------------.
* Get X'X matrix (or R, matrix of predictor correlations) from input data
* Also get X'Y, or correlations of predictors with dependent variable.
*-------------------------------------------------------.

compute xpx=x(5:(nv+4),2:(nv+1)).
compute xy=t(x(4,2:(nv+1))).
```

```
*————————————————————————————————————.
* Initialize the keep matrix for saving results, and the names vector.
*————————————————————————————————————.

    compute keep=make(iter,nv+2,-999).
    compute varnam2={'K','RSQ',varname}.

*————————————————————————————————————.
* Compute means and standard deviations. Means are in the first row of x and
* standard deviations are in the second row. Now that all of x has been
* appropriately stored, release x to maximize available memory.
*————————————————————————————————————.

    compute xmean=x(1,2:(nv+1)).
    compute ybar=x(1,1).
    compute std=t(x(2,2:(nv+1))).
    compute sy=x(2,1).
    release x.

*————————————————————————————————————.
* Start loop over values of k, computing standardized regression
* coefficients and squared multiple correlations. Store results
*————————————————————————————————————.

    loop l=1 to iter.
    . compute b = inv(xpx+(k &* ident(nv,nv)))*xy.
    . compute rsq= 2* t(b)*xy - t(b)*xpx*b.
    . compute keep(l,1)=k.
    . compute keep(l,2)=rsq.
    . compute keep(l,3:(nv+2))=t(b).
    . compute k=k+inc.
    end loop.

*————————————————————————————————————.
* If we are to print out estimation results, compute needed pieces and
* print out header and ANOVA table.
*————————————————————————————————————.

    do if (!k <> 999).
    .!let !rrtitle=!concat('****** Ridge Regression with k = ',!k).
    .!let !rrtitle=!quote(!concat(!rrtitle,' ****** ')).
    . compute sst=(n-1) * sy **2.
    . compute sse=sst * ( 1 - 2* t(b)*xy + t(b)*xpx*b).
    . compute ssr = sst - sse.
    . compute s=sqrt( sse / (n-nv-1) ).
    . print /title=!rrtitle /space=newpage.
    . print {sqrt(rsq);rsq;rsq-nv*(1-rsq)/(n-nv-1);s}
      /rlabel='Mult R' 'RSquare' 'Adj RSquare' 'SE'
      /title=' '.
    . compute anova={nv,ssr,ssr/(nv);n-nv-1,sse,sse/(n-nv-1)}.
    . compute f=ssr/sse * (n-nv-1)/(nv).
    . print anova
      /clabels='df' 'SS','MS'
      /rlabel='Regress' 'Residual'
      /title='          ANOVA table'
      /format=f9.3.
    . compute test=ssr/sse * (n-nv-1)/nv.
    . compute sigf=1 - fcdf(test,nv,n-nv-1).
    . print {test,sigf} /clabels='F value' 'Sig F'/title=' '.

*————————————————————————————————————.
* Calculate raw coefficients from standardized ones, compute standard errors
* of coefficients, and an intercept term with standard error. Then print
* out similar to REGRESSION output.
*————————————————————————————————————.

    . compute beta={b;0}.
    . compute b= ( b &/ std ) * sy.
    . compute intercpt=ybar-t(b)*t(xmean).
    . compute b={b;intercpt}.
    . compute xpx=(sse/(sst*(n-nv-1)))*inv(xpx+(k &* ident(nv,nv)))*xpx*
                  inv(xpx+(k &* ident(nv,nv))).
    . compute xpx=(sy*sy)*(mdiag(1 &/ std)*xpx*mdiag(1 &/ std)).
    . compute seb=sqrt(diag(xpx)).
    . compute seb0=sqrt( (sse)/(n*(n-nv-1))+xmean*xpx*t(xmean)).
    . compute seb={seb;seb0}.
    . compute rnms={varname,'Constant'}.
    . compute ratio=b &/ seb.
    . compute bvec={b,seb,beta,ratio}.
    . print bvec/title='————————————Variables in the Equation————————————'
      /rnames=rnms /clabels='B' 'SE(B)' 'Beta' 'B/SE(B)'.
    . print /space=newpage.
    end if.

*————————————————————————————————————.
* Save kept results into file. The number of cases in the file will be
* equal to the number of values of k for which results were produced. This
* will be simply 1 if k was specified.
*————————————————————————————————————.

    save keep /outfile='rr__tmp2.sav' /names=varnam2.

*————————————————————————————————————.
```

```
* Finished with MATRIX part of job.
*-----------------------------------------------------------.

    end matrix.

*-----------------------------------------------------------.
* If doing ridge trace, get saved file and produce table and
plots.
*-----------------------------------------------------------.

    !if (!k = 999) !then

    get file='rr__tmp2.sav'.
    print formats k rsq (f6.5) !enter (f8.6).
    report format=list automatic
     /vars=k rsq !enter
     /title=center 'R-SQUARE AND BETA COEFFICIENTS
FOR ESTIMATED VALUES OF K'.

    plot
      /format=overlay /title='RIDGE TRACE'
       /horizontal 'K'
      /vertical 'RR Coefficients'
       /plot !enter with k
       /title='R-SQUARE VS. K'
       /horizontal 'K'
      /vertical 'R-Square'
       /plot rsq with k.

    !ifend.

*-----------------------------------------------------------.
* Get back original data set and restore original settings.
*-----------------------------------------------------------.

    get file=rr__tmp1.sav.
    restore.
    !enddefine.
    restore.
```

THE AIR PROFESSIONAL FILE—1978-2004

A list of titles for the issues printed to date follows. Most issues are "out of print," but microfiche or photocopies are available through ERIC. Photocopies are also available from the AIR Executive Office, 222 Stone Building, Florida State University, Tallahassee, FL 32306-4462, $3.00 each, prepaid, which covers the costs of postage and handling. Please do not contact the editor for reprints of previously published Professional File issues.

Organizing for Institutional Research (J.W. Ridge; 6 pp; No. 1)
Dealing with Information Systems: The Institutional Researcher's Problems and Prospects (L.E. Saunders; 4 pp; No. 2)
Formula Budgeting and the Financing of Public Higher Education: Panacea or Nemesis for the 1980s? (F.M. Gross; 6 pp; No. 3)
Methodology and Limitations of Ohio Enrollment Projections (G.A. Kraetsch; 8 pp; No. 4)
Conducting Data Exchange Programs (A.M. Bloom & J.A. Montgomery; 4 pp; No. 5)
Choosing a Computer Language for Institutional Research (D. Strenglein; 4 pp; No. 6)
Cost Studies in Higher Education (S.R. Hample; 4 pp; No. 7)
Institutional Research and External Agency Reporting Responsibility (G. Davis; 4 pp; No. 8)
Coping with Curricular Change in Academe (G.S. Melchiori; 4 pp; No. 9)
Computing and Office Automation—Changing Variables (E.M. Staman; 6 pp; No. 10)
Resource Allocation in U.K. Universities (B.J.R. Taylor; 8 pp; No. 11)
Career Development in Institutional Research (M.D. Johnson; 5 pp; No 12)
The Institutional Research Director: Professional Development and Career Path (W.P. Fenstemacher; 6pp; No. 13)
A Methodological Approach to Selective Cutbacks (C.A. Belanger & L. Tremblay; 7 pp; No. 14)
Effective Use of Models in the Decision Process: Theory Grounded in Three Case Studies (M. Mayo & R.E. Kallio; 8 pp; No. 15)
Triage and the Art of Institutional Research (D.M. Norris; 6 pp; No. 16)
The Use of Computational Diagrams and Nomograms in Higher Education (R.K. Brandenburg & W.A. Simpson; 8 pp; No. 17)
Decision Support Systems for Academic Administration (L.J. Moore & A.G. Greenwood; 9 pp; No. 18)
The Cost Basis for Resource Allocation for Sandwich Courses (B.J.R. Taylor; 7 pp; No. 19)
Assessing Faculty Salary Equity (C.A. Allard; 7 pp; No. 20)
Effective Writing: Go Tell It on the Mountain (C.W. Ruggiero, C.F. Elton, C.J. Mullins & J.G. Smoot; 7 pp; No. 21)
Preparing for Self-Study (F.C. Johnson & M.E. Christal; 7 pp; No. 22)
Concepts of Cost and Cost Analysis for Higher Education (P.T. Brinkman & R.H. Allen; 8 pp; No. 23)
The Calculation and Presentation of Management Information from Comparative Budget Analysis (B.J.R. Taylor; 10 pp; No. 24)
The Anatomy of an Academic Program Review (R.L. Harpel; 6 pp; No. 25)
The Role of Program Review in Strategic Planning (R.J. Barak; 7 pp; No. 26)
The Adult Learner: Four Aspects (Ed. J.A. Lucas; 7 pp; No. 27)
Building a Student Flow Model (W.A. Simpson; 7 pp; No. 28)
Evaluating Remedial Education Programs (T.H. Bers; 8 pp; No. 29)
Developing a Faculty Information System at Carnegie Mellon University (D.L. Gibson & C. Golden; 7 pp; No. 30)
Designing an Information Center: An Analysis of Markets and Delivery Systems (R. Matross; 7 pp; No. 31)
Linking Learning Style Theory with Retention Research: The TRAILS Project (D.H. Kalsbeek; 7 pp; No. 32)
Data Integrity: Why Aren't the Data Accurate? (F.J. Gose; 7 pp; No. 33)
Electronic Mail and Networks: New Tools for Institutional Research and University Planning (D.A. Updegrove, J.A. Muffo & J.A. Dunn, Jr.; 7pp; No. 34)
Case Studies as a Supplement to Quantitative Research: Evaluation of an Intervention Program for High Risk Students (M. Peglow-Hoch & R.D. Walleri; 8 pp; No. 35)
Interpreting and Presenting Data to Management (C.A. Clagett; 5 pp; No. 36)
The Role of Institutional Research in Implementing Institutional Effectiveness or Outcomes Assessment (J.O. Nichols; 6 pp; No. 37)
Phenomenological Interviewing in the Conduct of Institutional Research: An Argument and an Illustration (L.C. Attinasi, Jr.; 8pp; No. 38)
Beginning to Understand Why Older Students Drop Out of College (C. Farabaugh-Dorkins; 12 pp; No. 39)
A Responsive High School Feedback System (P.B. Duby; 8 pp; No. 40)
Listening to Your Alumni: One Way to Assess Academic Outcomes (J. Pettit; 12 pp; No. 41)
Accountability in Continuing Education Measuring Noncredit Student Outcomes (C.A. Clagett & D.D. McConochie; 6pp; No. 42)
Focus Group Interviews: Applications for Institutional Research (D.L. Brodigan; 6 pp; No. 43)
An Interactive Model for Studying Student Retention (R.H. Glover & J. Wilcox; 12 pp; No. 44)
Increasing Admitted Student Yield Using a Political Targeting Model and Discriminant Analysis: An Institutional Research Admissions Partnership (R.F. Urban; 6 pp; No. 45)
Using Total Quality to Better Manage an Institutional Research Office (M.A. Heverly; 6 pp; No. 46)
Critique of a Method For Surveying Employers (T. Banta, R.H. Phillippi & W. Lyons; 8 pp; No. 47)
Plan-Do-Check-Act and the Management of Institutional Research (G.W. McLaughlin & J.K. Snyder; 10 pp; No. 48)
Strategic Planning and Organizational Change: Implications for Institutional Researchers (K.A. Corak & D.P. Wharton; 10 pp; No. 49)
Academic and Librarian Faculty: Birds of a Different Feather in Compensation Policy? (M.E. Zeglen & E.J. Schmidt; 10 pp; No. 50)
Setting Up a Key Success Index Report: A How-To Manual (M.M. Sapp; 8 pp; No. 51)
Involving Faculty in the Assessment of General Education: A Case Study (D.G. Underwood & R.H. Nowaczyk; 6 pp; No. 52)

THE AIR PROFESSIONAL FILE—1978-2003

Using a Total Quality Management Team to Improve Student Information Publications (J.L. Frost & G.L. Beach; 8 pp; No. 53)
Evaluating the College Mission through Assessing Institutional Outcomes (C.J. Myers & P.J. Silvers; 9 pp; No. 54)
Community College Students' Persistence and Goal Attainment: A Five-year Longitudinal Study (K.A. Conklin; 9 pp; No. 55)
What Does an Academic Department Chairperson Need to Know Anyway? (M.K. Kinnick; 11 pp; No. 56)
Cost of Living and Taxation Adjustments in Salary Comparisons (M.E. Zeglen & G. Tesfagiorgis; 14 pp; No. 57)
The Virtual Office: An Organizational Paradigm for Institutional Research in the 90's (R. Matross; 8 pp; No. 58)
Student Satisfaction Surveys: Measurement and Utilization Issues (L. Sanders & S. Chan; 9 pp; No. 59)
The Error Of Our Ways; Using TQM Tactics to Combat Institutional Issues Research Bloopers (M.E. Zeglin; 18 pp; No. 60)
How Enrollment Ends; Analyzing the Correlates of Student Graduation, Transfer, and Dropout with a Competing Risks Model (S.L. Ronco; 14 pp; No. 61)
Setting a Census Date to Optimize Enrollment, Retention, and Tuition Revenue Projects (V. Borden, K. Burton, S. Keucher, F. Vossburg-Conaway; 12 pp; No. 62)
Alternative Methods For Validating Admissions and Course Placement Criteria (J. Noble & R. Sawyer; 12 pp; No. 63)
Admissions Standards for Undergraduate Transfer Students: A Policy Analysis (J. Saupe & S. Long; 12 pp; No. 64)
IR for IR–Indispensable Resources for Institutional Researchers: An Analysis of AIR Publications Topics Since 1974 (J. Volkwein & V. Volkwein; 12 pp; No. 65)
Progress Made on a Plan to Integrate Planning, Budgeting, Assessment and Quality Principles to Achieve Institutional Improvement (S. Griffith, S. Day, J. Scott, R. Smallwood; 12 pp; No. 66)
The Local Economic Impact of Higher Education: An Overview of Methods and Practice (K. Stokes & P. Coomes; 16 pp; No. 67)
Developmental Education Outcomes at Minnesota Community Colleges (C. Schoenecker, J. Evens & L. Bollman: 16 pp; No. 68)
Studying Faculty Flows Using an Interactive Spreadsheet Model (W. Kelly; 16 pp; No. 69)
Using the National Datasets for Faculty Studies (J. Milam; 20 pp; No. 70)
Tracking Institutional leavers: An Application (S. DesJardins, H. Pontiff; 14 pp; No. 71)
Predicting Freshman Success Based on High School Record and Other Measures (D. Eno, G. W. McLaughlin, P. Sheldon & P. Brozovsky; 12 pp; No. 72)
A New Focus for Institutional Researchers: Developing and Using a Student Decision Support System (J. Frost, M. Wang & M. Dalrymple; 12 pp; No. 73)
The Role of Academic Process in Student Achievement: An Application of Structural Equations Modeling and Cluster Analysis to Community College Longitudinal Data[1] (K. Boughan, 21 pp; No. 74)
A Collaborative Role for Industry Assessing Student Learning (F. McMartin; 12 pp; No. 75)
Efficiency and Effectiveness in Graduate Education: A Case Analysis (M. Kehrhahn, N.L. Travers & B.G. Sheckley; No.76)
ABCs of Higher Education-Getting Back to the Basics: An Activity-Based Costing Approach to Planning and Financial Decision Making (K. S. Cox, L. G. Smith & R.G. Downey; 12 pp; No. 77)
Using Predictive Modeling to Target Student Recruitment: Theory and Practice (E. Thomas, G. Reznik & W. Dawes; 12 pp; No. 78)
Assessing the Impact of Curricular and Instructional Reform - A Model for Examining Gateway Courses[1] (S.J. Andrade; 16 pp; No. 79)
Surviving and Benefitting from an Institutional Research Program Review (W.E. Knight; 7 pp; No. 80)
A Comment on Interpreting Odds-Ratios when Logistic Regression Coefficients are Negative (S.L. DesJardins; 7 pp; No. 81)
Including Transfer-Out Behavior in Retention Models: Using NSC EnrollmentSearch Data (S.R. Porter; 16 pp; No. 82)
Assessing the Performance of Public Research Universities Using NSF/NCES Data and Data Envelopment Analysis Technique (H. Zheng & A. Stewart; 24 pp; No. 83)
Finding the 'Start Line' with an Institutional Effectiveness Inventory[1] (S. Ronco & S. Brown; 12 pp; No. 84)
Toward a Comprehensive Model of Influences Upon Time to Bachelor's Degree Attainment (W. Knight; 18 pp; No. 85)
Using Logistic Regression to Guide Enrollment Management at a Public Regional University (D. Berge & D. Hendel; 14 pp; No. 86)
A Micro Economic Model to Assess the Economic Impact of Universities: A Case Example (R. Parsons & A. Griffiths; 24 pp; No. 87)
Methodology for Developing an Institutional Data Warehouse (D. Wierschem, R. McBroom & J. McMillen; 12 pp; No. 88)
The Role of Institutional Research in Space Planning (C.E. Watt, B.A. Johnston. R.E. Chrestman & T.B. Higerd; 10 pp; No. 89)
What Works Best? Collecting Alumni Data with Multiple Technologies (S. R. Porter & P.D. Umback; 10 pp; No. 90)
Caveat Emptor: Is There a Relationship between Part-Time Faculty Utilization and Student Learning Outcomes and Retention? (T. Schibik & C. Harrington; 10 pp; No. 91)

The AIR Professional File is intended as a presentation of papers which synthesize and interpret issues, operations, and research of interest in the field of institutional research. Authors are responsible for material presented. The *File* is published by the Association for Institutional Research.

Editor:
Gerald W. McLaughlin
Director of Planning and Institutional Research
DePaul University
1 East Jackson, Suite 1501
Chicago, IL 60604-2216
Phone: 312/362-8403
Fax: 312/362-5918
gmclaugh@depaul.edu

Associate Editor:
Dr. Jessica S. Korn
Director of Institutional Research
Eckerd College
4200 54th Avenue North
Saint Petersburg, FL 33711
Phone: 727/864-7677
Fax: 727/964-1877
kornjs@eckerd.edu

Managing Editor:
Dr. Terrence R. Russell
Executive Director
Association for Institutional Research
222 Stone Building
Florida State University
Tallahassee, FL 32306-4462
Phone: 850/644-4470
Fax: 850/644-8824
air@mailer.fsu.edu

AIR *Professional File* Editorial Board

Authors interested in having their manuscripts considered for the *Professional File* are encouraged to send four copies of each manuscript to the editor, Dr. Gerald McLaughlin. Manuscripts are accepted any time of the year as long as they are not under consideration at another journal or similar publication. The suggested maximum length of a manuscript is 5,000 words (approximately 20 double-spaced pages), including tables, charts and references. Please follow the style guidelines of the *Publications Manual of the American Psychological Association, 4th Edition.*