HANDLING MULTICOLLINEARITY WITH SAS IML® SOFTWARE: RIDGE REGRESSION ON THE PC

GARY WILLIAM CARR, NYNEX
GERALD TLAPA, BELLCORE

1.0 Introduction.

Analyses of economic data must consider the likely presence of multicollinearity. If one is interested in studying Gross Domestic Product (GDP) versus other macroeconomic infrastructure components, multicollinearity will be anticipated since components of the infrastructure generally march in harmony with GDP. For example, many correlation studies have shown a strong direct relationship between GDP and Energy. (1) In particular, this relationship has been analyzed by Janosi and Grayson with a resulting $R^2 \geq 0.9$ in thirty-two of the thirty-four cases studied. These results were obtained by using a log-log model assuming constant rates of continuous growth. (2).

As the number of variables increases within an econometric model, the concern with multicollinearity increases since the probability that some variables measure similar phenomena increases. Presence of multicollinearity violates the assumption that the explanatory variables in a regression model are not strongly interrelated. The assumption is one of the three fundamental assumptions of regression analysis. Consequence of violating this assumption leads to low precision of individual regression coefficient estimates, which in turn can lead to erroneous inferences.

One specific focus of this paper addresses multicollinearity using the ridge regression technique. Ridge regression provides a statistically robust method for overcoming data problems frequently encountered in econometric modeling using Ordinary Least Squares methodologies.

2.0 Analysis.

Assume that the relationships of any given country's infrastructure components relative to its GDP are sought such that they can be expressed in terms of a general linear regression model:

$$Y = X\beta \quad \varepsilon \qquad (1)$$

where Y is a vector of observations (Ln GDP in subsequent analyses of this study), X is a matrix of infrastructure variable observations (infrastructure components), β is a vector of parameters, and ε is a vector of errors normally distributed with expected value of $E(\varepsilon) = 0$, Var $E\{\varepsilon\} = \sigma^2$. In this case the elements of variance are uncorrelated.

Since $E(\varepsilon) = 0 \quad E(y) = E(X\beta)$ (2)

and

$$\varepsilon'\varepsilon = (Y - Xb)'(Y - Xb) \qquad (3)$$

where b is the LSE of β and $(\ )'$ indicates the matrix transpose.

The least squares estimate (LSE) of β is the value b which minimizes the error sum of squares, $\varepsilon'\varepsilon$. This provides the normalized equation:

$$(X'X)b = X'Y \qquad (4)$$

where (X'X) is the correlation matrix and is non-singular.

$$b = (X'X)^{-1} X'Y \qquad (5)$$

where $(X'X)^{-1}$ is the inverse of the correlation matrix. The solution b has the following properties:

1) It estimates β with minimized error sum of squares irrespective of the distribution function of the errors.
2) The elements of b are linear functions of the observations $Y_1, Y_2, \ldots Y_n$ and provide unbiased estimates of the elements of β which have the minimum variances irrespective of distribution functions of the errors.

Normal use of ordinary least squares regression assumes that the input variables are uncorrelated. In addition, the researcher wants as many observations as possible per variable to insure that a purely random component will be less likely to affect inferences about the deterministic portion of the equation. Economic data, however, are often sparse, especially with regard to developing countries and may, to a significant degree, measure the same basic phenomena. Economic data therefore, almost always displays multicollinearity. This is a common problem that must be addressed when modeling economic effects. Minimizing multicollinearity maximizes the explanatory power of any model chosen to describe causual economic relationships.

Multicollinearity can be defined as a property of the correlation matrices where the off diagonals (independent regressions) approach 1. (3). When significant multicollinearity exists, it is impossible to determine the importance of each independent (regressor) variable in explaining a dependent variable based on R^2. (4). Consider the variables of energy, telecommunications, domestic investment, airline travel and savings that produce a correlation matrix, (X'X), below.

CORRELATION MATRIX (X'X) FOR CHINA (1961-1985)

	Ln(ENG)	Ln(TEL)	Ln(AIR)	Ln(INV)	Ln(SAV)	Ln(GDP)
Ln(ENG)	1.0000	0.9183	0.8786	0.9131	0.9110	0.8911
Ln(TEL)	0.9183	1.0000	0.9257	0.9313	0.9349	0.9259
Ln(AIR)	0.8786	0.9257	1.0000	0.9618	0.9649	0.9902
Ln(INV)	0.9131	0.9313	0.9618	1.0000	0.9992	0.9813
Ln(POP)	0.9110	0.9349	0.9649	0.9992	1.0000	0.9844
Ln(GDP)	0.8911	0.9259	0.9902	0.9813	0.9844	1.0000

ENG = Level of Energy consumption expressed in million tons of coal equivalent. (UN).
TEL = Number of telephones installed. (AT&T).
AIR = Number of airline passengers carried. (UN)
INV = Domestic investment. (World Bank).
SAV = Domestic savings. (World Bank).
GDP = Gross domestic investment. (IMF).

Each "independent" infrastructure variable shows a high correlation to each other as well as to Ln(GDP). Alternatively, multicollinearity can be observed in the inverse of this correlation matrix, $(X'X)^{-1}$ measured by the high diagonal values above 1. (See inverse correlation matrix below). While independent variables should ideally not be correlated with each other, economic data on infrastructure components will show some degree of correlation to each other. Collinearity among the independent variables must therefore be kept at acceptable levels in the regression model. The diagonal elements of the inverse matrix are called Variance Inflation Factors (VIF).

$$VIFi = 1/(1-R^2i) \qquad (6)$$

R^2i is the coefficient of determination of the i-th independent variable regressed on all other independent variables. (5). As an example, using the China data as the independent variable Ln(SAV) regressed on all other independent variables, one obtains a R^2 value of .9987. When this R^2 value is introduced into equation 6 above, the VIF_i = 778.084. When the VIF_i value exceeds the value of 10 (as identified by Freund and Littel) (obtained by substituting the R-squared value of the total regression results into the VIF formula) the presence of unacceptable multicollinearity is identified. (6). The analyst must now select which VIF element value is most closely related and not independent from the other independent variables. Correction of the data is now required to eliminate the presence of collinearity.

Consider once again the,

INVERSE CORRELATION MATRIX $(X'X)^{-1}$ FOR CHINA (1961 - 1985)

	Ln(ENG)	Ln(TEL)	Ln(AIR)	Ln(INV)	Ln(SAV)
Ln(ENG)	8.225	-5.036	1.035	-21.951	18.150
Ln(TEL)	-5.036	11.768	-3.227	24.936	-28.216
Ln(AIR)	1.035	-3.227	116.113	15.484	-28.945
Ln(INV)	-21.951	24.936	15.484	722.240	-739.911
Ln(POP)	18.150	-28.216	-28.945	-739.911	778.084

Several remedies are frequently suggested to correct poorly conditioned data. The general approach is usually to collect more data. This answer rarely helps the econometrician, who typically has short data series and cannot wait for additional data to be obtained. In addition, the cost of obtaining additional data may be prohibitive and cannot guarantee a reduced collinearity sample. (7). Another common approach is to reduce the number of independent variables in the model if similar phenomena are being measured by many independent variables. This may not be feasible if one is trying to determine the importance of several variables influencing one dependent variable. A procedure more robust than ordinary least squares (OLS) regression is appropriate in this instance.

To overcome problems with data quality and improve the reliability of data analyses and forecasts, econometricians may (intentionally) introduce bias into their models. This serves the purpose of reducing standard errors and multicollinearity. Indirect introduction of bias into a model occurs when one of the variables which is collinear with another variable is dropped. By reducing the information input into a model, collinearity is reduced. Another procedure that can be used introduces bias directly. One such robust procedure is ridge regression and is becoming increasingly popular in econometric analyses. Hoerl and Kennard (1970) are cited frequently for their use of this technique. The guiding principles which they have followed are listed below. (8).:

1) As bias, k, is added to the diagonal elements of the correlation matrix, the resulting coefficients will stabilize and have characteristics similar to an orthogonal system.
2) The actual coefficients will have reasonable absolute values respective to the factors they represent.
3) The proper sign will be assigned to all coefficients.
4) The residual sum of squares will not be inflated to an unreasonable value. The amount of variance will not be large relative to the process generating the data.

The ridge regression procedure is intended to overcome multicollinearity problems where the correlation matrix is nearly 1.0, giving rise to unstable parameter estimates. Additional work on this method by G.M. Mullett has explained why an incorrectly signed coefficient becomes corrected. (9). G. Jelisavcic justifies this technique by demonstrating the lower mean squared error which it produces and its ability to choose the proper bias, k, required for the variables being analyzed. (10). The work of Chatterjee and Price support this technique by showing how it minimizes the mean squared error when a regression equation is used to predict future values (11). Draper and Smith confirm the use of ridge regression when prior knowledge of the parameters is known (lower coefficient values or the sign of a coefficient is incorrect), and also, when ridge regression is subject to restrictions on the parameters (a least square problem with the addition of restricted or constraints on external information). (12). T. H. Wonnacott demonstrates that in using ridge regression to avoid multicollinearity, the confidence intervals of relevant regressor variables are more precise. (13).

Ridge regression is not a panacea for all economic problems but in many instances it has led to improved understanding of available data.

The ridge regression methodology demonstrated here will add bias, k, to the trace elements of the correlation matrix. If any trace element of the inverse correlation matrix is less than one, no bias is added. This procedure is repeated until all trace values are equal to one. This insures that the collinearity of each variable is treated separately and only those variables demonstrating collinearity will receive bias. In addition, the relationships of the variables are maintained to assure correct "balance" thereby avoiding distortion of the original hypothesis. Another value of importance to be calculated is P^*, (P star), which determines the correct amount of bias for establishing the best regression equation. (14).

$$P^* = trI\{[(X'X+kI)^{-1}(X'X)-I]'[(X'X+kI)^{-1}(X'X)-I]\} \quad (7)$$

Since, I is a p * p matrix where p is the number of independent variables in the model, the trI = p. Clearly $P^* \leq p$, which suggest that the violation of assumption about interrelationship of the independent variables in the model has not been ignored, but rather accounted for. Quantity p^* could therefore be thoutht of as "effective" number of independent variables in the model.

Another important quantity to be calculated is trace of inverted design matrix, TR^*:

$$TR^* = trI(X'X + kI)^{-1}(X'X)(X'X + kI)^{-1}] \quad (8)$$

Both P^* and TR^* are calculated for a whole range of parameter k; starting with k = 0. Optional value for k is the value at which the following equality is reached, or approached as close as feasible:

$$TR^* = P^* \quad (9)$$

SAS IML® software program was used to compute results following the above procedures. This program appears in Appendix A to this paper. (15).

3.0 Results

Let us now apply ridge regression techniques to the China example of Section 2. First consider the ordinary least squares results operating upon the dependent variable of Ln(GDP):

ORDINARY LEAST SQUARE RESULTS
CHINA (1961 to 1985)

B0	Ln(ENG)	Ln(TEL)	Ln(AIR)	Ln(INV)	Ln(SAV)
5.715	0.0388	-0.0475	0.3047	-0.5911	0.9865

Using the above coefficients for each infrastructure component generates erroneous results. For a growing economy the signs of the coefficients should be positive. A negative coefficient value is expected only if a segment of an economy is flat or declines (negative slope). Also, the large coefficients are misleading as indicators of the contribution being made to the dependent variable. China's economy shows increasing growth based on the gross domestic product per capita:

CHINA (GDP PER CAPITA)

1960	1970	1980	1985
87.22	109.66	275.56	308.73

Using ridge regression, a different set of coefficients is obtained. Adding sectors of an economy while not exceeding the dependent variable, usually results in positive coefficients. Contributions made by one sector upon another should be removed to insure the ceteris paribus requirement thereby representing the independent contributions to the dependent variable. Imposing these conditions results in the following:

RIDGE REGRESSION RESULTS
CHINA (1961 to 1985)

B0	Ln(ENG)	Ln(TEL)	Ln(AIR)	Ln(INV)	Ln(SAV)
5.5153	0.1964	0.0635	0.1445	0.1764	0.1711
T for H_0					
	(3.265)	(4.873)	(11.72)	(12.88)	(13.90)
Probability>|T|					
	(.00407)	(.00001)	(.0001)	(.00001)	(.00001)

The above results show a level of significance below .004 in all independent variables. This measures the probability that a |T| statistic would obtain a value greater than the observed given that the true parameter is zero. The probability of the T-statistic for energy in China is 0.00407. This means that if we reject the null hypothesis (B = 0), there is a 0.4% probability that the null hypothesis is actually true.

The amount of ridge bias necessary to achieve the proper regression result was within "rule of thumb" parameters. (16). The maximum amount of bias added to any element was as follows:

BIAS REQUIRED FOR RIDGE REGRESSION
CHINA 1961-1985

k = .275

4.0 Conclusions:

Ridge Regression bias intentionally introduced can be kept at low levels and only added to those elements demonstrating collinearity. The mean square error term is kept at low levels, insuring improved results. Ridge regression corrects improper OLS signs, inflated parameter estimates and unstable coefficients.

This study dealt with five variables, which show varying degrees of multicollinearity. Ridge regression methodology was employed to deal with multicollinearity in determining model coefficients. One must, however, use caution when making general statements concerning these regression results beyond the variables discussed. Additional work should be undertaken with an expanded economic model of infrastructure components to develop a better understanding of each country.

Using SAS IML® on a personal computer to perform ridge regression of economic data streams provides increased flexibility to the analyst.

Care should be given to some of the PC's limitations using SAS IML®. SAS version 6.03 requires more memory than earlier versions and 640K bytes should be considered the minimum requirement. The amount of variables one uses should also be kept at a minimum and variables only added that are important to the economic model. Please note that in the attached program listing IML worksize will need to be adjusted upward as variables are added to the model.

In Appendix B, the SAS output window display of the CHINA data is included. It is important to observe the signs of the variables and the statistical measure of "press," "ESS" and "MSE." Appendix C includes the log window results. With each computational loop the effects of bias, k, on the diagonal elements can be observed by the analyst. No equation should be considered the best until all statistical parameters are reviewed. Intermediate knowledge of ridge regression procedural results and underlying data is always useful.

NOTES and REFERENCES

SAS/IML® software is the registered trademark of SAS Instutute Inc., Cary, NC, USA.

1. deJanosi, Peter E., and Grayson, Leslie E., "Patterns of Energy Consumption and Economic Growth and Structure", Journal of Development Studies. Vol. 8, 1972, pp. 241-249.
   See also:
   Toldaro, Michael P., Economic Development in the Third World. Longman, Ney York, 1985. pp.540-542.
2. Dowling, Edward T., Mathematics for Economists. Schaum's Outline Series, McGraw Hill Book Company, 1980.
3. Draper, N.R., and Smith, H., Applied Regression Analysis. John Wiley & Sons, New York, 1981. pp. 294-379.
4. Cassidy, Henry J., Using Econometrics: A Beginners Guide. Reston Publishing Co. Inc., New York, 1981. pp. 160-168.
5. Belsley, David A., Kuh, Edwin, and Welsch, Roy E., Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, New York, 1980. pp. 192-229.
6. Freund, Rudolph J. Ph.D., and Littel, Ramon C., Ph.D., SAS System for Regression. SAS Institute Inc, Cary, NC, 1986.
7. Belsley, op. cit.
8. Hoerl, A. E., and Kennard, R. W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics. No. 12, 1970. pp. 69-82.
9. Mullett, G. W., "Why Regression Coefficients Have the Wrong Sign", Journal of Quality Technology. No.8, 1976. pp. 112-126.
10. Jelisavcic, Gordana, "Mean Square Error as a Reliability Measure for Biased Estimators", Paper delivered at ASA Meeting, August 16-19, 1982, Cincinnati, Ohio.
11. Chatterjee, Samprit, and Price, Bertram, Regression Analysis by Example. John Wiley & Sons, New York, 1977. pp. 143-214.
12. Draper, op. cit.
13. Wonnacott, Thomas H., and Wonnacott, Ronald J., Regression: A Second Course in Statistics. John Wiley & Sons, New York, 1981. pp. 64-448.
14. Personal papers from Gordana Jelisavcic, Ph.D. who has done extensive work on ridge regression and published several papers and delivered speeches to America Statistical Association.
15. SAS IML® is a computer software product by SAS institute, Box 8000, Cary N.C. The original ridge program appeared in course notes of Principals of Regression Analysis and has been modified to include Gordana Jelisavcic, Ph.D. notes.
16. Ridge bias should not exceed the range of 0 to .30.

APPENDIX A

SAS IML® VERSION 6.03
RIDGE REGRESSION COMPUTER PROGRAM
FOR THE PERSONAL COMPUTER

```
OPTIONS NODATE;
DATA COUNTRY;
INFILE 'D:CHN.PRN';
LENGTH YAR AIR ENG GDI GDP GNS TEL 8;
INPUT YAR AIR ENG GDI GDP GNS TEL;
LAIR=LOG(AIR);
LENG=LOG(ENG);
LGDI=LOG(GDI);
LGDP=LOG(GDP);
LGNS=LOG(GNS);
LTEL=LOG(TEL);

TITLE1 'CHINA  DATA (5 VARABLES)';
/* THE MACRO VARIABLE VARLIST CONTAINS */
/* THE REGRESSOR VARIABLES */
%LET VARLIST=
  LAIR LENG LGDI LGNS LTEL ;
  /* THE MACRO VARIABLE DEPVAR CONTAINS THE */
  /* DEPENDENT VARIABLE */
```

```
%LET DEPVAR=LGDP;
/* THE MACRO VARIABLE DATA SET CONTAINS THE NAME */
/* OF THE SAS DATA SET */
%LET DATASET=COUNTRY;
/* THE MACRO VARIABLE OUTDSN CONSTAINS THE */
/* OUTPUT DATA SET */
%LET OUTDSN=RIDGE;
/* THE MACRO VARIABLE COEF CONTAINS THE */
/* LABELS FOR COEFFICIENTS */
%LET COEF=
 'BO' 'LAIR' 'LENG' 'LGDI' 'LGNS'
 'LTEL';
/* REQUIRED FOR CHARACTER MATRIX OF MODEL */
%LET LABL=CC;
/* THE MACRO VARIABLE BBNAMES CONTAINS THE */
/* NAMES OF THE COEFFICIENTS AND USEFUL STATISTICS */
%LET BBNAMES='PRESSSS' 'ESS' 'MSE' 'CK' 'K' 'TRHK'
 'BO' 'LAIR' 'LENG' 'LGDI' 'LGNS' 'LTEL' ;
 /* USED TO NAME CHARACTER MATRIX OF MODEL */
%LET CCNAMES='LABEL';
 PROC PRINT DATA=&DATASET;
/* PROC REG DATA=&DATASET;  */
 /* MODEL &DEPVAR=&VARLIST/VIF; */
PROC IML WORKSIZE=70;

START RIDGE;
 N=NROW(X);   /* NUMBER OF INPUTS PER VARIABLE   */
 P=NCOL(X);   /* NUMBER OF DEPENDENT VARIABLES   */
 JX=J(N,1,1)//X;
/* COMPUTE ANOVA EXTIMATE OF SIGMA2 */
SIGMA2=(Y-JX*INV(JX/*JX)*JX/*Y)/
   *(Y-JX*INV(JX/*JX)*JX/*Y)/(N-P-1);
XM=X[:,];
YM=Y[:,];
XC=(X-REPEAT(XM,N,1)); /* X'S CENTERED */
YC=Y-YM; /* Y'S CENTERED */
YCPYC=YC/*YC;
SSXC=XC[##,];
STDXC=SQRT(SSXC/);
XCS=XC*DIAG(1/STDXC); /* X'S ARE CENTERED AND SCALED */
XCSPXCS=XCS/*XCS; /* CORRELATION MATRIX */
XCSPYC=XCS/*YC;
LABL=J(1,1);      /* USED TO MAKE NUMERIC MATRIX */
LABL=CHAR(LABL);  /* USED TO CHANGE NUMERIC TO CHARACTER
                     MATRIX */
ZA=J(P,1,0);      /* USED TO MAKE NUMERIC MATRIX */
STARTT=0;   /* START OF BIAS TO BE SET */
ENDD=.50;   /* END OF BIAS LOOP TO BE SET */
INCREMNT=.025; /* INCREMENT OF BIAS TO BE SET */
DO K= STARTT TO ENDD BY INCREMNT;
/* RMAT=K#I(P); */
 IF K=0 THEN RMAT=K#I(P);
 ELSE RMAT=ZC;
ZB=VECDIAG(RMAT);
XCSPXCSK=XCSPXCS+RMAT;
MATINV=INV(XCSPXCSK); /* CORR MATRIX + BIAS INVERSED */
MATCOR=MATINV*XCSPXCS;/* VALUE OF M(Z) */
IDLSMAT=MATCOR-I(P);  /* VALUES OF MM(Z) */
TRMAT=IDLSMAT*IDLSMAT; /* VALUE OF MM(Z)*MM(Z) */
TROFMAT=DIAG(TRMAT);   /* DIAG OF MATRIX TR */
GARY=VECDIAG(TROFMAT); /* COLUMN VECTOR OF DIAG TR VALUES */
IDLSTRMA=I(P)-TROFMAT; /* VALUE OF P MATRIX FORM */
STRIG=VECDIAG(IDLSTRMA); /* COLUMN VECTOR OF DIAG P VALUES */
VIFMZMZK=MATINV*MATCOR;
MZMZKVIF=VECDIAG(VIFMZMZK); /* VIF (TR) VALUES */
RVALUE=MZMZKVIF/STRIG; /* RVALUE MUST BE KEPT AT UPPER
                          BOUND >1 */
TR=TRACE(VIFMZMZK);
PSTAR=TRACE(IDLSTRMA);
R=TR/P;
KBIAS=VECDIAG(RMAT);

IF ANY(RVALUE>1) THEN LABEL="OVER";
ELSE LABEL="UNDER";
LABL=LABEL;

PRINT RVALUE KBIAS PSTAR TR R LABEL ;

/* THIS DUE LOOP IS REQUIRED TO INCREASE THE K BIAS ON ONLY
   THOSE ELEMENTS WHICH THE VIARANCE INFLATION FACTOR
   (RVALUE) IS ABOVE 1 THEN ADDS EQUAL BIAS TO ALL ELEMENTS
   AFTER (RVALUE) LESS THAN 1 */

DO I=1 TO P BY 1;
IF RVALUE[I,1]>1 THEN ZA[I,1]=K+INCREMNT;
ELSE ZA[I,1]=ZB[I,1];
  IF ALL(RVALUE<1) THEN DO I=1 TO P BY 1; /* DO LOOP FOR
                                             RVALUE */
  ZA[I,1]=K+INCREMNT;                  /* ALL UNDER 1  */
  END;
ZC=DIAG(ZA);
RMAT=ZC;
END;
```

```
SBETA=MATINV*XCSPYC;
/* COMPUTE UNSTANDARDIZED REGRESSION COEFFICIENTS */
SXC=1/STDXC;
BETA=SXC#SBETA;
BO=BETA/*XM/;
INTERCPT=YM-BO;
BETA=INTERCPT//BETA;
/* PREPARE FOR COMPUTING A PRESS LIKE STATISTIC */
/* THIS STATISTIC PR(RIDGE) IS EASIER */
/* TO COMPUTE THAN PRESS */
RESID=Y-JX*BETA;
ESS=RESID[##,];
MSE=ESS/(N-P-1);
/* HAT MATRIX USING K */
HK=XCS*INV(XCSPXCSK)*XCS/;
TRHK=TRACE(HK);
/* COMPUTE CP TYPE STATISTIC */
CK=ESS/SIGMA2-N+2+2*TRHK;
PRESSRES=RESID/(1-VECDIAG(HK)-1/N);
PRESSSS=PRESSRES[##,];
BB=BB//(PRESSSS//ESS//MSE//CK//K//TRHK//BETA/);
CC=CC//(LABL);   /* CREATES A VERTICAL CHARACTER MATRIX OF
LABEL */
END;
/* BUILD ANOVA TABLE BASED ON SMALLEST PRESS */
A=NROW(CC);   /* READS THE AMOUNT OF COMPUTATIONS IN LABL */
FF=0;         /* USED AS STARTING POINT CALULATIONS */
CC[1,1]="O.L.S.";     /* CHANGES LABEL TO READ OLS */
DO I=1 TO A BY 1;     /* THIS DO LOOP IS REQUIRED */
IF CC[I,1]="UNDER" THEN FF=1+FF;  /* SHOW THE POINT WHERE */
IF FF=1 THEN CC[I-1,1]="BEST";/* MULTI-COLINEARITY IS OUT */
END;                          /* BASED ON R(VALUE) */
DO I=1 TO A BY 1;             /* DO LOOP REQUIRED TO READ */
IF CC[I,1]="BEST" THEN EE=I;  /* THE BEST DATA SET       */
END;
FINISH;
   START ANOVA;
   USE &OUTDSN VAR/K PRESSSS MSE ESS CK TRHK BO &VARLIST/;
   READ POINT EE;
   PRINT ,K PRESSSS MSE ESS CK TRHK BO, &VARLIST;
   XCSPXCSK=XCSPXCS+K#I(P);
   VARCOV=INV(XCSPXCSK)*(XCS/*XCS)*INV(XCSPXCSK);
   VIF=VECDIAG(VARCOV);
   SE=SQRT(MSE#VIF);
   SBETA=INV(XCSPXCSK)*XCSPYC;
   T=SBETA/SE;
   PVALUE=(1-PROBT(ABS(T),N-P-1))*2;
   COEF=/&COEF//;COEF=COEF[2:P+1,];
   PRINT ,COEF T PVALUE VIF;
FINISH;
/* MAIN PROGRAM */
USE &DATASET;
READ ALL VAR/&VARLIST/ INTO X;
READ ALL VAR/&DEPVAR/ INTO Y;
RUN RIDGE; /* RUN RIDGE PROGRAM */
/* SORT BB BY PRESSSS */
/* TBB=BB;
BB[RANK(BB[,1]),]=TBB;
FREE TBB; */
CREATE &OUTDSN FROM BB[COLNAME=/&BBNAMES/];
 APPEND FROM BB; /* CREATE OUTPUT DATA SET */
CREATE C FROM CC[COLNAME=/&CCNAMES/];
 APPEND FROM CC; /* CREATE CHARACTER OUTPUT DATA SET */
RUN ANOVA;   /* CREATE NEW T STATISTICS AND VIFS */
QUIT;
 DATA D; MERGE &OUTDSN C;
 PROC PRINT DATA=D;
 /*   PROC PLOT DATA=D;
      PLOT (&VARLIST)*K;    */
    RUN;
```

APPENDIX B

SAS OUTPUT
ORIGINAL DATA SET and RIDGE REGRESSION RESULTS

CHINA INPUT DATA (5 VARABLES)

OBS	YAR	AIR	ENG	GDI	GDP	GNS	TEL	LAIR	LENG	LGDI	LGDP	LGNS	LTEL
1	1961	116.2	265.1	9074.7	47136.2	9281.8	244.03	4.7553	5.5801	9.1132	10.7608	9.1358	5.4973
2	1962	136.4	264.2	4886.7	43716.0	5447.2	244.03	4.9156	5.5767	8.4943	10.6855	8.6029	5.4973
3	1963	146.5	285.3	8286.6	47306.8	8891.9	244.03	4.9870	5.6535	9.0224	10.7644	9.0929	5.4973
4	1964	153.2	306.0	11780.0	55146.6	12336.5	244.03	5.0317	5.7236	9.3742	10.9178	9.4203	5.4973
5	1965	176.4	316.8	16150.8	65586.2	16487.9	244.03	5.1728	5.7583	9.6897	11.0911	9.7104	5.4973
6	1966	188.9	349.0	20590.6	75054.8	20793.7	244.03	5.2412	5.8551	9.9326	11.2260	9.9424	5.4973
7	1967	201.0	245.5	14010.1	70290.0	14257.9	244.03	5.3033	5.5033	9.5475	11.1604	9.5651	5.4973
8	1968	242.0	325.0	13392.6	66873.8	13685.1	244.03	5.4889	5.7838	9.5025	11.1106	9.5241	5.4973
9	1969	320.0	350.2	15622.7	76415.6	16179.2	742.02	5.7683	5.8585	9.6565	11.2439	9.6915	6.6094
10	1970	398.0	347.7	26561.9	91006.6	26590.3	1240.02	5.9865	5.8513	10.1872	11.4187	10.1883	7.1229
11	1971	476.0	390.3	29125.0	98119.3	29856.2	1738.01	6.1654	5.9669	10.2794	11.4939	10.3041	7.4605
12	1972	554.0	408.5	28879.7	103943.5	29720.9	2236.01	6.3172	6.0125	10.2709	11.5516	10.2996	7.7125
13	1973	632.0	445.2	35890.0	120729.4	36485.8	2734.01	6.4489	6.0985	10.4882	11.7013	10.5047	7.9135
14	1974	710.0	477.0	39196.5	132154.6	38412.1	3232.00	6.5653	6.1675	10.5763	11.7917	10.5561	8.0809
15	1975	1000.0	501.2	46669.8	150201.4	46396.1	3412.00	6.9078	6.2170	10.7509	11.9197	10.7450	8.1351
16	1976	1050.0	526.4	42805.1	146834.7	43096.6	3629.00	6.9566	6.2661	10.6644	11.8971	10.6712	8.1967
17	1977	1140.0	602.4	48369.0	162860.9	48724.2	3833.00	7.0388	6.4009	10.7866	12.0007	10.7939	8.2514
18	1978	1540.0	669.9	65550.4	191617.4	64401.4	4059.00	7.3395	6.5071	11.0906	12.1633	11.0729	8.3087
19	1979	2519.0	688.7	74523.2	229102.9	71933.1	4220.00	7.8316	6.5348	11.2189	12.3419	11.1835	8.3476
20	1980	2568.0	562.8	81570.6	272007.6	77960.7	4432.00	7.8509	6.3329	11.3092	12.5136	11.2640	8.3966
21	1981	3236.0	556.6	79131.5	289224.2	79383.8	4634.00	8.0821	6.3219	11.2789	12.5750	11.2820	8.4412
22	1982	3942.0	464.8	81190.9	290673.9	86901.9	4907.00	8.2794	6.1416	11.3046	12.5800	11.3725	8.4984
23	1983	3836.0	498.5	84842.2	294417.4	88718.2	5161.00	8.2522	6.2116	11.3485	12.5928	11.3932	8.5489
24	1984	5000.0	550.2	93468.2	316327.8	95213.5	5536.00	8.5172	6.3103	11.4454	12.6645	11.4639	8.6190
25	1985	7300.0	587.7	124025.2	322714.5	109304.0	6134.00	8.8956	6.3762	11.7282	12.6845	11.6019	8.7216

CHINA RIDGE REGRESSION RESULTS (5 VARABLES)

OBS	PRESSSS	ESS	MSE	CK	K	TRHK	80	LAIR	LENG	LGDI	LGNS	LTEL	LABEL
1	0.20428	0.06375	0.003355	6.0000	0.000	5.0000	5.7150	0.3047	0.0388	-0.5911	0.9865	-0.0475	O.L.S
2	0.14182	0.09108	0.004794	10.6603	0.025	3.2575	6.1186	0.2597	-0.0019	0.1614	0.2130	0.0026	OVER
3	0.15998	0.11220	0.005905	16.0584	0.050	2.8105	5.9555	0.2281	0.0307	0.1741	0.2046	0.0207	OVER
4	0.18149	0.13335	0.007018	21.7777	0.075	2.5185	5.8365	0.2077	0.0627	0.1766	0.1996	0.0324	OVER
5	0.20299	0.15338	0.008073	27.3281	0.100	2.3087	5.7487	0.1931	0.0908	0.1763	0.1954	0.0407	OVER
6	0.22356	0.17210	0.009058	32.5878	0.125	2.1490	5.6833	0.1820	0.1149	0.1750	0.1916	0.0469	OVER
7	0.24302	0.18959	0.009978	37.5478	0.150	2.0226	5.6345	0.1732	0.1355	0.1733	0.1882	0.0517	OVER
8	0.26143	0.20601	0.010843	42.2363	0.175	1.9196	5.5985	0.1660	0.1531	0.1715	0.1852	0.0555	OVER
9	0.27894	0.22154	0.011660	46.6922	0.200	1.8337	5.5724	0.1600	0.1682	0.1697	0.1824	0.0586	OVER
10	0.28903	0.23044	0.012128	49.2265	0.225	1.7753	5.4986	0.1498	0.1675	0.1767	0.1899	0.0578	OVER
11	0.30256	0.24263	0.012770	52.7329	0.250	1.7113	5.4750	0.1452	0.1788	0.1557	0.2083	0.0596	OVER
12	0.32419	0.26141	0.013758	58.2034	0.275	1.6489	5.5153	0.1445	0.1964	0.1764	0.1711	0.0635	BEST
13	0.31820	0.25711	0.013532	56.8664	0.300	1.6209	5.5255	0.1456	0.1904	0.1784	0.1729	0.0607	UNDER
14	0.35788	0.29064	0.015297	66.7198	0.325	1.5512	5.5338	0.1399	0.2186	0.1615	0.1714	0.0676	UNDER
15	0.37271	0.30347	0.015972	70.4650	0.350	1.5123	5.5376	0.1371	0.2252	0.1601	0.1696	0.0687	UNDER
16	0.38739	0.31613	0.016638	74.1689	0.375	1.4771	5.5440	0.1345	0.2312	0.1587	0.1679	0.0696	UNDER
17	0.40200	0.32868	0.017299	77.8448	0.400	1.4451	5.5524	0.1321	0.2364	0.1574	0.1663	0.0703	UNDER
18	0.41656	0.34116	0.017956	81.5040	0.425	1.4158	5.5627	0.1300	0.2411	0.1561	0.1648	0.0710	UNDER
19	0.43111	0.35359	0.018610	85.1558	0.450	1.3887	5.5745	0.1280	0.2452	0.1549	0.1633	0.0715	UNDER
20	0.44568	0.36601	0.019264	88.8081	0.475	1.3637	5.5875	0.1261	0.2489	0.1537	0.1619	0.0720	UNDER

APPENDIX C

SAS LOG OUTPUT
1. THE LOG OUTPUT SHOWS EACH TIME BIAS (k) IS ADDED.
2. ANOVA RESULTS OF BEST RIDGE REGRESSION EQUATION AT THE END

RVALUE	KBIAS	PSTAR	TR	R	LABEL
16.109266	0	5	1527.2107	305.44202	OVER
8.2208294	0				
717.77584	0				
773.34088	0				
11.763261	0				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
7.2631094	0.025	3.8328391	23.013723	4.6027446	OVER
4.9220496	0.025				
5.7428669	0.025				
5.5193258	0.025				
6.287537	0.025				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
4.6674848	0.05	3.5128333	14.046961	2.8093923	OVER
3.6615162	0.05				
3.4635398	0.05				
3.2691576	0.05				
4.3891981	0.05				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
3.4032339	0.075	3.2492532	9.7431704	1.9486341	OVER
2.9171168	0.075				
2.4819302	0.075				
2.3331589	0.075				
3.3408918	0.075				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
2.6580896	0.1	3.0330128	7.2432591	1.4486518	OVER
2.4204417	0.1				
1.9182552	0.1				
1.8018389	0.1				
2.6771872	0.1				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
2.1682736	0.125	2.8537721	5.6380188	1.1276038	OVER
2.0640819	0.125				
1.5523229	0.125				
1.4587214	0.125				
2.2205525	0.125				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
1.8227115	0.15	2.7031865	4.5370252	0.907405	OVER
1.7956109	0.15				
1.2967520	0.15				
1.2197496	0.15				
1.8881955	0.15				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
1.5665515	0.175	2.5750145	3.7451394	0.7490279	OVER
1.5860831	0.175				
1.1091566	0.175				
1.0446158	0.175				
1.6362293	0.175				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
1.3695912	0.2	2.4646275	3.15459	0.630918	OVER
1.4181031	0.2				
0.9663416	0.2				
0.9114124	0.2				
1.4392048	0.2				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
1.1624102	0.225	2.3870757	2.7873265	0.5574653	OVER
1.2723201	0.225				
1.0126116	0.2				
0.9670189	0.2				
1.2724013	0.225				

RVALUE	KBIAS	PSTAR	TR	R	LABEL
1.0256471	0.25	2.3011845	2.4224651	0.484493	OVER
1.1561114	0.25				
0.489377	0.25				
1.2618844	0.2				
1.1364077	0.25				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
0.9643583	0.275	2.2171943	2.0899463	0.4179893	OVER
1.0643210	0.275				
0.8907362	0.25				
0.5604755	0.275				
1.0411158	0.275				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
0.9700943	0.275	2.1785343	1.9541674	0.3908335	UNDER
0.9742094	0.3				
0.9030201	0.25				
0.5717253	0.275				
0.9344435	0.3				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
0.8239372	0.325	2.0829715	1.636379	0.3272758	UNDER
0.9146512	0.325				
0.5806489	0.325				
0.5519799	0.325				
0.8781866	0.325				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
0.7610498	0.35	2.0289314	1.4760887	0.2952177	UNDER
0.8522179	0.35				
0.5375327	0.35				
0.5117989	0.35				
0.8120679	0.35				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
0.7066914	0.375	1.9798693	1.3406179	0.2681236	UNDER
0.7973455	0.375				
0.5005576	0.375				
0.4773343	0.375				
0.7546672	0.375				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
0.6593105	0.4	1.9351114	1.2250181	0.2450036	UNDER
0.7487820	0.4				
0.4685662	0.4				
0.4475075	0.4				
0.7044443	0.4				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
0.6177041	0.425	1.8941001	1.1255234	0.2251047	UNDER
0.7055347	0.425				
0.4406651	0.425				
0.4214874	0.425				
12.660195	0.425				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
0.5809254	0.45	1.8563696	1.0392253	0.2078451	UNDER
0.6668068	0.45				
0.4161595	0.45				
0.3986243	0.45				
0.6209648	0.45				
RVALUE	KBIAS	PSTAR	TR	R	LABEL
0.5482199	0.475	1.821528	0.9638473	0.1927695	UNDER
0.6319509	0.475				
0.3944935	0.475				
0.3784036	0.475				
0.5859872	0.475				

K	PRESSSS	MSE	ESS	CK	TRHK
0.275	0.3241925	0.0137583	0.2614074	58.203402	1.6488649

B0	LAIR	LENG	LGDI	LGNS	LTEL
5.5153418	0.1444754	0.1963999	0.1764421	0.1710571	0.0635369

COEF	T	PVALUE	VIF
LAIR	11.720654	3.853E-10	0.456007
LENG	3.2651576	0.0040725	0.643119
LGDI	12.881383	7.756E-11	0.2229594
LGNS	13.903939	2.075E-11	0.2042108
LTEL	4.8729964	0.0001056	0.5339643

Gary William Carr
NYNEX Corporation
335 Madison Ave Room 2104
New York, NY 10017
(212) 370-7459