SAS Macros for Bootstrapping and Cross-Validating Regression Equations

Richard T. Carson
University of California, Berkeley

Abstract

Measures of predictive success from regression equations such as $R^2$ and measures of parameter precision such as standard errors for coefficients tend to be overly optimistic, particularly, when the ratio of estimated parameters to observations is not large. Statisticians have recently proposed two non-parametric techniques, bootstrapping and cross-validation, which are resampling plans useful for obtaining more precise estimates of these statistics. The two techniques are applicable in a wide range of situations.

Bootstrapping is based on creating new data sets by drawing random samples (with replacement) from the original data set (or function of the original data set), while cross-validation is based on dividing the original data set into subsets and creating cross-validation data sets by leaving out one of the subsets for each pair of learning and test sets. Thus both bootstrapping and cross-validation set up (in different ways) artificial samples and populations from which to calculate the statistics of interest.

Both of these procedures are easily implemented in SAS using the MACRO facility and are useful additions to the statistical analyst's tool kit.

1. Introduction

Two popular techniques for estimating the "true" predictive power of regression equations and/or the standard errors of the regression coefficients in the presence of "overfitting" or non-normal error terms are the bootstrap and cross-validation. Both of these techniques are resampling procedures in which the original observations or the residuals from the original regression (or some function of both) are sampled and the statistics of interests are recalculated. This resampling and reestimation is usually repeated a number of times.

The bootstrap is generally the non-parametric maximum likelihood estimator of a statistic while cross-validation is a quadratic approximation to the bootstrap. The theorems proving the optimal properties of these two techniques use a high level of mathematics and are not presented here. A very intuitive explanation of why these procedures work is in the Scientific American article by Diaconis and Efron (1983) and a fairly non-technical survey of the uses and properties of these two procedures (as well as a related technique, the jackknife) is Efron and Gong (1982). For specific problems a particular variant of one of the techniques presented below may be optimal. The interested reader should consult the references given below before undertaking a bootstrap or cross-validation exercise.

2. Bootstrapping

The bootstrap can best be thought of as drawing samples from an initial distribution with replacement. This initial distribution can either be the original observations or more frequently the residuals from an initial regression equation. Just as the original observations (frequently referred to as the empirical distribution) are the result of a random sample from the parent population, the bootstrap observations represent a sample from the empirical distribution. Under fairly general assumptions, it can be shown that the bootstrap samples from the empirical distribution approximate random samples from the unobserved parent population.

As an illustration, let $\{X_1, \ldots, X_n\}$ be a random sample of observations. The mean value of these observations is simply $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. To calculate the bootstrap variance of $X$, draw with replacement from the original sample a bootstrap sample of the same number of observations as the original sample. Calculate the mean $\bar{X}_j$ from each of the $j$ bootstrap samples. These are now the observations. The bootstrap variance of $X$ is simply

$$\frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X})^2.$$ 

In the regression case, let $Y = X\beta$ be the model estimated by the researcher, with $Y$ and $X$ being observed. There are two basic ways to bootstrap. The most common way uses the residuals, $e$, from this initial regression. Drawing random observations (with replacements), one of the residuals is added to each $X$ to form a $Y$. A new set of coefficients $\beta$ is estimated by regressing $Y$ on $X$. This form of bootstrapping is most useful for calculating standard errors which is done in a fashion directly analogous to our earlier example using $B$ as the population parameter(s) and $B_{(m)}$ as the observations. The other way to bootstrap is to draw a random sample of the original observations $(X, Y)$, and to calculate $Y_m = X\beta$ where $\beta$ is obtained as above. The statistics of predictive success are based on the relationship between $Y_m$ and $Y$ can be easily calculated.

3. Cross-Validation

On fairly good way to think about cross-validation is that of of a technique for simulating out of sample predictions. This notion recalls an older variant of cross-validation where the researcher divided the data into two subsets (usually $1/2$ and $1/2$ or $2/3$ and $1/3$) estimating the parameters of the model of interest on one subset and then predicting the other subset to get an honest estimate of how well the model would predict "new" observations. This however is wasteful of the data. The problem lies in over optimistic estimates of predictive success. The solution is to use all of the data to estimate the coefficients and then use...
cross-validation to estimated how good the fit is. It is common now to divide the data into several subsets and to predict one of these subsets out of sample using all of the other subsets. The trick is not to make each subset so large that dropping one of them out noticeable affects the performance of the estimating equation with respect to its apparent "in sample" fit. The two extremes in cross-validation are to drop out half of the data and to drop out the observations one at a time. The two extremes in cross-validation are to drop out half of the data and to drop out the observations one at a time. A limited amount of work in this area (Breiman et al., 1984) has suggested that between 5 and 10 subsets is optimal. Unless the one observation at a time scheme is adopted the cross-validation exercise is usually repeated many times.

4. MACRO BOOTREG

The key to an efficient bootstrap macro is to use the SAS point operator in combination with a random number generator to select the observations before reading in the data. The macro below uses an input data set which contains residuals from an initial regression as well as the original dependent and independent variables. The output of the BOOTREG Macro below are data sets containing the coefficients estimated using the bootstrap samples and the predicted values from these regressions. The data sets can be gathered together and the statistics of interest estimated. If a very large number of bootstrap samples are planned or if the number of independent variables is large, the researcher may want to substitute PROC MATRIX for PROC REG where it would be necessary to calculate \( \text{INV}(X'X)X' \) only once.

%MACRO BOOTREG(DATIN=,PVAR=,IVARS=,NIVARS=,RVAR=,DVAR=NEWY,DCOEFF=BTCO,DPRED=BTPR,NPVAR=NPRED,REP=);
"MACRO ARGUMENTS:

DATIN: INPUT DATA SET
PVAR: PREDICTED VALUES FROM THE ORIGINAL ESTIMATION
IVARS: INDEPENDENT VARIABLES
NIVARS: NUMBER OF INDEPENDENT VARIABLES
RVAR: RESIDUALS FROM ORIGINAL ESTIMATION
DVAR: NEW DEPENDENT VARIABLE
DCOEFF: COEFFICIENTS FROM ESTIMATION USING BOOTSTRAP SAMPLE
NPVAR: PREDICTED VALUES USING BOOTSTRAP SAMPLE
DPRED: DATA SET CONTAINING NPVAR
OBS: NUMBER OF OBSERVATIONS IN ORIGINAL SAMPLE
REP: REPETITIONS (NUMBER OF BOOTSTRAP SAMPLES) DESIRED
"

%DO I=1 %TO &REP;
DATA BTBES&I;
DO J=1 TO &OBS;
IOBSN=INT(UNIFORM(0) , &OBS) + 1);
SET &DATIN  POINT=IOBSN;
OUTPUT END STOP;
KEEP &RVAR;
"DATA BTSMP&I;
MERGE &DATIN (KEEP=&PVAR &ivar)
BTRES&I;
ARVAR&ARVAR * (ACOBS/(ACOBS - ANIVARS - 2));
ADVAR=APVAR + &ARVAR;
PROC REG DATA=STSMPI OUTEST=DCOEFF;&I;
MODEL ADVARS=ivar/NOPRINT;
OUTPUT OUT=DPRED P=APVAR;
PROC DELETE DATA=BTSMPI BTRES&I;
%END;
"$END BOOTREG;

2. MACRO CVALIDR

The key to an efficient cross-validation macro is in subsetting the data. The method used below is quick but is not guaranteed to produce subsets of equal size which may be desirable in some cases. PROC RANK can be used in combination with a random variate if it is desirable to predetermine the exact size of each subset. The macro here takes the initial observations as input and the output data sets contain the coefficients from each regression and the predicted values for the subsets dropped out. These data sets can then be aggregated and the statistics of interest calculated.

%MACRO CVALIDR(DATAIN=,DVAR=,IVARS=,PVAR=PY,DVAR2=,GRP=,REP=);
"MACRO ARGUMENTS:

DATAIN: INPUT DATA SET
DVAR: DEPENDENT VARIABLE
IVARS: INDEPENDENT VARIABLES
PVAR: PREDICTED VALUES
DVAR2: COPY OF DEPENDENT VARIABLE
GRP: NUMBER OF GROUPS OR SUBSETS
REP: NUMBER OF REPETITIONS
"

%DO I=1 %TO &AREP;
DATA ORIG&I;
SET &DATAIN;
DVAR2=DVAR;
RANVAR=UNIFORM(0);
%DO J=1 %TO &GRP;
DATA CVSMP&I&J;
SET ORIG&I;
IF RANVAR GE (&J - 1)/&GRP)
RANVAR LT (&J/&GRP) THEN &DVAR=.;
DROP RANVAR;
PROC REG DATA=CVSMP&I&J;
MODEL &DVAR=&IVARS/NOPRINT;
OUTPUT OUT=PRED&I&J P=&PVAR;
DATA CVPR&I&J;
SET PRED&I&J;
IF &DVAR NE • THEN DELETE;
KEEP &DVAR2 &PVAR;
PROC DELETE DATA=ORIG&I;
%ENDj
%END CVALIDR;

A. A More Complicated Example

Obviously more complicated examples are possible. The researcher may have an autoregressive error structure (Freedman and Peters, 1984a,b), or want to look at a simultaneous system (Freedman, 1984), or want to examine different variable selection procedures (Breiman and Freedman, 1983). To give a flavor of these, we take an example from Carson (1984) which
shows the use of cross-validation to determine the mean square error and percent change root mean square error of the predictions of a well known small macro-economic model (Klein's Model I). This macro (CvKL) uses PROC SYSREG for the regression estimation of the ordinary least squares, two-stage least squares, and three stage least squares forms of the model. PROC SIMLIN uses the reduced forms (estimated using the coefficients from PROC SYSREG). PROC RANK is used to divided the data into seven equal sized subsets. The results of this macro for 100 complete repetitions along with the cross-validation R^2's are shown in table 1. The SAS code for this macro and table I are given in the appendix below.

FOOTNOTES

*SAS is the registered trademark of SAS Institute Inc., Cary, NC, USA.


2/The residuals should be multiplied by \(\sqrt{n} \) where \( n \) is the number of observations and \( \hat{\sigma} \) is the number of estimated parameters in the homoscedastic OLS case (see Freedman (1981) for a discussion).

3/The seminal references on cross-validation are Geisser (1975) and Stone (1974). The interested reader should also see Efron (1982, 1983).

4/The trade off is between bias and variance. Division into a large number of subsets tends to reduce the bias in the statistic calculated using the cross-validation results but increases the variance in those statistics. Efron (1983) discusses this issue in more detail.

5/Because cross-validation is fairly expensive, it may be desirable to use some type of stopping rule in terms of percent change in the statistic of interest after some minimum number of repetitions.

REFERENCES


APPENDIX

$MACRO CVKLEIN
$DO J=1 STO 50;
DATA RANKJ;
SET KLEIN;
RANVAR=UNIFORM(0);
PROC RANK GROUPS=7 OUT=RANKJ;
VAR RANVAR;
RANKS RANKJ;
$DO J=1 STO 7;
DATA KRK&J;
SET RANKJ;
RANKJ=UNIFORM(0);
PROC RANK GROUPS=7 OUT=RANKJ;
VAH RANVAR;
RANKS RANKJ;
%00 1=1 JTO 7;
DATA KRK&I&J;
SET RANKJ RANKJ;
RANKJ=RANKJ+RANKJ;
IF RANKJ=&I THEN C=.;
APPEND;
PROC SYSREG DATA=KRK&I&J OUTEST=B&I&J;
FIRST: BLOCK C P W I X WSUM K Y = KLAG PLAG XLAG WP W T YEAR;
CONSUME: MODEL C = P PLAG WSUM;
INVEST: MODEL I = P PLAG KLAG;
LABOR: MODEL W = X KLAG YEAR;
PRODUCT: IDENTITY X = C + I + G - T;
INCOME: IDENTITY Y = C + I + G;
PROFIT: IDENTITY P = X - W - T;
STOCK: IDENTITY K = KLAG + I;
WAGE: IDENTITY WSUM = W + WP;
PROC SIMLIN EST=BAI&J DATA=KRK&I&J TYPE='OLS' START=0 TOTAL;
ENDOGENOUS C P W I X WSUM K Y;
EXOGENOUS WP G T YEAR;
LAGGED KLAG K 1 PLAG P 1 XLAG X 1;
ID YEAR;
OUTPUT OUT=POLS&I&J P=P1C P1P P1W P1I P1X P1WSUM P1K P1Y;
PROC SIMLIN EST=BAI&J DATA=KRK&I&J TYPE='2SLS' START=0 TOTAL;
ENDOGENOUS C P W I X WSUM K Y;
EXOGENOUS WP G T YEAR;
LAGGED KLAG K 1 PLAG P 1 XLAG X 1;
ID YEAR;
OUTPUT OUT=P2SLS&I&J P=P2C P2P P2W P2I P2X P2WSUM P2K P2Y;
PROC SIMLIN EST=BAI&J DATA=KRK&I&J TYPE='3SLS' START=0 TOTAL;
ENDOGENOUS C P W I X WSUM K Y;
EXOGENOUS WP G T YEAR;
LAGGED KLAG K 1 PLAG P 1 XLAG X 1;
ID YEAR;
OUTPUT OUT=P3SLS&I&J P=P3C P3P P3W P3I P3X P3WSUM P3K P3Y;
$END;

DATA KOSLAJ;
SET POLS&J POLS&J POLS&J POLS&J POLS&J POLS&J;
IF C=.
KEEP YP P1C P1P P1W P1I P1X P1WSUM P1K P1Y;
PROC SORT; BY YR;
DATA K2SLS&J;
IF C=.
KEEP YP P2C P2P P2W P2I P2X P2WSUM P2K P2Y;
PROC SORT; BY YR;
DATA K3SLS&J;
IF C=.
KEEP YP P3C P3P P3W P3I P3X P3WSUM P3K P3Y;
PROC SORT; BY YR;
DATA KPREDJ;
MERGE KOSLAJ K2SLS&J K3SLS&J KLEIN (KEEP=YP C P W I X WSUM K Y);
BY YEAR;
ARRAY TRUE C P W I X WSUM K Y;
ARRAY XLAG LC LP LW LX LWSUM LL LY;
ARRAY P1I P1C P1P P1W P1I P1X P1WSUM P1K P1Y;
ARRAY P2I P2C P2P P2W P2I P2X P2WSUM P2K P2Y;
ARRAY P3I P3C P3P P3W P3I P3X P3WSUM P3K P3Y;

1067
ARRAY MSEOLS M1C M1P M1W M1I M1WSUM M1K M1Y;
ARRAY MSEOLS M2C M2P M2W M2I M2WSUM M2K M2Y;
ARRAY MSEOLS M3C M3P M3W M3I M3WSUM M3K M3Y;
ARRAY PCMOLS PCM1C PCM1P PCM1W PCM1I PCM1WSUM PCM1K PCM1Y;
ARRAY PCMOLS PCM2C PCM2P PCM2W PCM2I PCM2WSUM PCM2K PCM2Y;
ARRAY PCM3SLS PCM3C PCM3P PCM3W PCM3I PCM3WSUM PCM3K PCM3Y;
DO OVER TRUE;
  TLAG=LAG(TRUE);
  MSEOLS=(PD1 - TRUE)**2;
  MSEOLS=(PD2 - TRUE)**2;
  MSEOLS=(PD3 - TRUE)**2;
  PCMOLS=((PD1/TLAG) - (TRUE/TLAG))**2;
  PCMOLS=((PD2/TLAG) - (TRUE/TLAG))**2;
  PCMOLS=((PD3/TLAG) - (TRUE/TLAG))**2;
END;
PROC MEANS;
DATA COEFFCC;
IF .JDE="FIRST" AND C=-1 THEN _MODEL="COLS";
IF _MODEL="FIRST" AND P=-1 THEN _MODEL="POLS";
IF _MODEL="FIRST" AND W=-1 THEN _MODEL="WOLS";
IF _MODEL="FIRST" AND I=-1 THEN _MODEL="IOLS";
IF _MODEL="FIRST" AND K=-1 THEN _MODEL="KOLS";
IF _MODEL="FIRST" AND T=-1 THEN _MODEL="TOLS";
IF _MODEL="CONSUME" AND TYPE="2SLS" THEN _MODEL="2SLS";
IF _MODEL="LABOR" AND TYPE="2SLS" THEN _MODEL="2SLS";
IF _MODEL="LABOR" AND TYPE="3SLS" THEN _MODEL="3SLS";
IF _MODEL="LABOR" AND TYPE="IDENTITY" THEN DELETE;
DROP _TYPE _TYPEVAR;
END;
$END CVKLEIN;
DATA KLEIN;
*SEE SAS/ETS MANUAL UNDER SYSREG FOR A DESCRIPTION AND THE OBSERVATIONS IN THIS DATA SET;
$CVKLEIN
DATA COEFF;
SET COEFF1 COEFF2 COEFF3 COEFF4 COEFF5
   COEFF46 COEFF47 COEFF48 COEFF49 COEFF50;
PROC SORT; BY _MODEL;
PROC MEANS; BY _MODEL;
DATA KPREDS;
SET KPREDS1 KPREDS2 KPREDS3 KPREDS4 KPREDS5;
PROC MEANS;
<table>
<thead>
<tr>
<th></th>
<th>RMS Klein</th>
<th>RMS CV</th>
<th>SD-RMS Klein</th>
<th>SD-RMS CV</th>
<th>MIN RMS-CV</th>
<th>MAX RMS-CV</th>
<th>PRMS Klein</th>
<th>PRMS CV</th>
<th>$R^2$ Klein</th>
<th>$R^2$ CV</th>
<th>SD-$R^2$ CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>C(OLS)</td>
<td>1.663</td>
<td>3.332</td>
<td>.329</td>
<td>2.500</td>
<td>4.072</td>
<td>3.14</td>
<td>6.18</td>
<td>.938</td>
<td>.758</td>
<td>.048</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>1.300</td>
<td>2.800</td>
<td>.276</td>
<td>2.580</td>
<td>3.237</td>
<td>3.06</td>
<td>8.14</td>
<td>.950</td>
<td>.768</td>
<td>.042</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>1.956</td>
<td>2.966</td>
<td>.290</td>
<td>2.258</td>
<td>3.694</td>
<td>90.27</td>
<td>198.69</td>
<td>.807</td>
<td>.296</td>
<td>.191</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>2.969</td>
<td>6.200</td>
<td>.600</td>
<td>4.685</td>
<td>7.051</td>
<td>10.10</td>
<td>10.33</td>
<td>.918</td>
<td>.094</td>
<td>.068</td>
<td></td>
</tr>
<tr>
<td>WS</td>
<td>1.300</td>
<td>2.800</td>
<td>.276</td>
<td>2.360</td>
<td>3.637</td>
<td>3.39</td>
<td>7.01</td>
<td>.966</td>
<td>.839</td>
<td>.079</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>1.956</td>
<td>2.396</td>
<td>.298</td>
<td>2.250</td>
<td>3.091</td>
<td>.90</td>
<td>1.87</td>
<td>.977</td>
<td>.909</td>
<td>.018</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>2.960</td>
<td>6.200</td>
<td>.600</td>
<td>4.685</td>
<td>7.051</td>
<td>5.01</td>
<td>12.07</td>
<td>.905</td>
<td>.528</td>
<td>.090</td>
<td></td>
</tr>
<tr>
<td>C(2SLS)</td>
<td>3.995</td>
<td>4.665</td>
<td>.269</td>
<td>4.157</td>
<td>5.930</td>
<td>7.66</td>
<td>8.71</td>
<td>.694</td>
<td>.516</td>
<td>.052</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>3.753</td>
<td>4.120</td>
<td>.197</td>
<td>3.763</td>
<td>5.012</td>
<td>10.81</td>
<td>11.50</td>
<td>.628</td>
<td>.551</td>
<td>.042</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>2.707</td>
<td>2.079</td>
<td>.131</td>
<td>2.720</td>
<td>3.993</td>
<td>196.69</td>
<td>155.39</td>
<td>.390</td>
<td>.264</td>
<td>.054</td>
<td></td>
</tr>
<tr>
<td>WS</td>
<td>3.753</td>
<td>4.120</td>
<td>.197</td>
<td>3.762</td>
<td>5.012</td>
<td>9.50</td>
<td>10.20</td>
<td>.741</td>
<td>.667</td>
<td>.029</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>4.335</td>
<td>9.810</td>
<td>.439</td>
<td>3.792</td>
<td>6.093</td>
<td>2.06</td>
<td>2.38</td>
<td>.768</td>
<td>.751</td>
<td>.005</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>2.998</td>
<td>3.170</td>
<td>.198</td>
<td>2.812</td>
<td>3.587</td>
<td>252.35</td>
<td>190.50</td>
<td>.258</td>
<td>.166</td>
<td>.079</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>5.004</td>
<td>5.277</td>
<td>.110</td>
<td>4.180</td>
<td>6.352</td>
<td>2.42</td>
<td>2.64</td>
<td>.696</td>
<td>.701</td>
<td>.036</td>
<td></td>
</tr>
</tbody>
</table>