Improving the Rainbow Test: A Macro to Measure the Lack of Fit in Multiple Regression with the use of the Bootstrap.

Paul Johnson, University of California San Francisco, CA

ABSTRACT

A SAS® macro for measuring the lack of fit in multiple regression with the use of the bootstrap is presented. The macro compares fitting a multiple regression model over low leverage points with a fit over the entire data. This is the Rainbow test for measuring the lack of fit in regression.
The Bootstrap is used to estimate the bias and an improved estimate for lack of fit is calculated. The procedure is repeated for several different subsets of low leverage points. The final improved estimate incorporates Bayesian prior weight information for the central region of low leverage points. This macro requires base SAS and SAS/STAT software to run.

INTRODUCTION

Improving the Rainbow Test:

The model fit is:

$$y_i = \beta_0 + \beta_1 x_i + \sum_{h=1}^{q} \theta_h w_{hi} + \varepsilon_i$$

where the $\underline{\theta} = (\theta)_h$ is a q x 1 vector of unknown parameters and the w_{hi} are fixed observable real numbers.

The rainbow test for lack of fit (see Utts, 1982) is carried out. The test is: $H_0: \underline{\theta} = \mathbf{0}$ against $H_1: \underline{\theta} \neq \mathbf{0}$.

Let SSE_{FULL} be the error sum of squares from fitting the above model to the entire data set. A subset of the data is constructed by choosing those points that have low leverage. This results in constructing a data set primarily from the central region of the entire data set. Points with leverage less than 2/n were chosen. For the sample data set, n = 50, which meant the cutoff point for inclusion was 0.04.
The data was constructed by fitting the model:

$y_i = 1.50 + 0.5 * x_i + 0.00194 * x_i^2 + \varepsilon_i$, where $\varepsilon_i \overset{iid}{\sim} N(0,1)$.

The vector $\mathbf{x}$ (50 x 1) of observations consists of points:

[1.1 1.3 2.5 2.5 2.5 3.8 3.8 4.1 4.6 4.8 5.3 6.4 6.4 6.4 6.4 8.7 9.6 10.8 10.8 12.4 12.7 14.7 16.5 17.8 22.4 22.4 26.0 28.8 28.8 32.1 32.4 32.8 33.4 37.1 37.1 37.1 39.6 72.1 76.2 78.7 79.1 83.1 83.1 88.0 90.0 90.0 90.0 92.1 92.1 95.4] $_{50 \text{ x } 1}$

The vector $\mathbf{y}$ is constructed. The above model is fit and the central region of data points determined. Let m be the number of points in the central region. Let $SSE_{CENTRAL}$ be the residual sum of squares from fitting the above model to these m observations.
The rainbow test statistic is defined as:

$$F = \frac{\left(SSE_{FULL} - SSE_{CENTRAL}\right) / (n - m)}{SSE_{CENTRAL} / (m - 2)}.$$

The sum of squares of error due to the lack of fit is defined as:

$$SSE_{LOF} = SSE_{FULL} - SSE_{CENTRAL}.$$

Hence

$$F = \frac{SSE_{LOF} / (n - m)}{SSE_{CENTRAL} / (m - 2)} \overset{H_0}{\sim} F(n - m,\ m - 2).$$

A Bootstrap sample of size B is obtained and used to obtain bootstrap estimates of bias (see Efron and Tibshirani, 1993). These bias values for F are subtracted from the simple estimator to find an improved estimate of F. The p-value associated with this bias corrected estimate is obtained. Let the p-value associated with this bootstrap bias corrected estimate be equal to p-boot1. The B bootstrap sample of F's are then examined and two central regions of data are constructed. One using the cutoff points of the 10th and 90th percentiles. The other with the lower and upper quartiles.
The p-values associated with the average of these estimates are p-boot2 and p-boot3 respectively. Prior weights determine the final estimate. For the example weights of 0.3, 0.3 and 0.4 were chosen. In other words a little more weight was given to the estimates of lower leverage, when determined by examination of the p-values obtained from the B bootstrap samples.

The final estimate of the p-value associated with the bias corrected bootstrapping of the rainbow test is:

$$\text{final-p} = W_1 * \text{p - boot1} + W_2 * \text{p - boot2} + W_3 * \text{p - boot3}.$$

CONCLUSION

The results obtained for a bootstrap sample of 100 were:

Model: $y_i = 1.50 + 0.5 * x_i + 0.00194 * x_i^2 + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} N(0,1)$.

W_1	p-boot1	W_2	p-boot2	W_3	p-boot3	final-p
.3	.0000136	.3	.0000139	.4	.0000154	.0000144

A second model was fit, by fitting the model:

$y_i = 1.50 + 0.5 * x_i + 0.0001 * x_i^2 + \varepsilon_i$, where $\varepsilon_i \overset{iid}{\sim} N(0,1)$.

The results obtained for this model were:

W_1	p-boot1	W_2	p-boot2	W_3	p-boot3	final-p
.3	.22275	.3	.22976	.4	.23800	.23096

The Bootstrap simulations are printed. The Bootstrap estimates are printed. The improved estimates, incorporating the bias correction factor from the bootstrap, are printed. The alternative estimate for the bootstrap p-value, using the central region defined by the upper and lower quartiles, is printed. The alternative estimate for the bootstrap p-value, using the central region defined by the 10th and 90th percentiles, is printed. The alternative improved estimate of the p-value is printed (using the Q3 − Q1 Central Region).
The alternative improved estimate of the p-value is printed (using the P90 − P10 Central Region). A final estimate is obtained by using prior weights. The weights are factors for the Bootstrap (entire data), the Bootstrap based on the Q3 − Q1 central region, and the Bootstrap for the P90 − P10 Region.

In the example above the additional term in the model is $\beta_2 x^2$.

The test $H_0: \underline{\theta} = \mathbf{0}$ against $H_1: \underline{\theta} \neq \mathbf{0}$ is equivalent to testing

$H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$, for this example.

Suppose, without loss of generality, that q = 2, and $\theta_1 = \theta_2 = 1$ then

$$\sum_{h=1}^{q} \theta_h w_{h\,i} = w_{1i} + w_{2i} = \beta_2 x^2 .$$

In the above two examples for model 1, β_2 = 0.00194, and for model 2, β_2 = 0.0001. Since w_{1i} and w_{2i} are fixed observable real numbers, and the *x*'s are defined by the vector **x**, w_{1i} and w_{2i} can be solved for each *i* = 1, 2, ...n.

e.g., for x_1 = 1.1 and β_2 = 0.00194,

$\beta_2 x^2$ = 0.0023474 choose $w_{11} = 0.0013$ and $w_{21} = 0.0010474$.

There are repeat x_i values for several combinations. Hence the test for lack of fit using the pure error term as the denominator is carried out. PROC RSREG (see SAS/STAT, 1990) was used to test for the lack of fit for a linear term. The results for the rainbow test without the use of the bootstrap are also given for both models.

$y_i = 1.50 + 0.5 * x_i + 0.00194 * x_i^2 + \varepsilon_i$, where $\varepsilon_i \overset{iid}{\sim} N(0,1)$.

Lack of Fit Based on this model:

Pure Error		Rainbow F-test	
F-ratio	p-value	F-ratio	p-value
2.487	0.0316	6.0431	.000008451

$y_i = 1.50 + 0.5 * x_i + 0.0001 * x_i^2 + \varepsilon_i$, where $\varepsilon_i \overset{iid}{\sim} N(0,1)$.

Lack of Fit Based on this model:

Pure Error		Rainbow F-test	
F-ratio	p-value	F-ratio	p-value
0.775	0.7372	1.4378	0.1874

Neill and Johnson (1984) provide for a review in testing for the lack of fit in regression. The Rainbow test for a test of lack of fit in regression does not require replicates or a prior estimate of variance. The improved estimates, through the use of the bootstrap, correct for the bias. The final estimate (final-p) for the p-value provides an improved estimate incorporating prior weight information for the central region of low leverage points.
The SAS macro follows.

CODE

```
/*-------------------------------------------------------------------------*
|                                                                         |
|                                                                         |
|  This Program called LEVERAGE.SAS reads in the data set.                |
|  Improving the Rainbow Test: A macro to measure the lack                |
|  of fit in multiple regression with the use of the                      |
|  Bootstrap. The macro compares a fit over low leverage                  |
|  points with a fit over the entire data.                                |
|  The Bootstrap is used to estimate bias and an improved                 |
|  estimate for lack of fit is calculated. The procedure                  |
|  is repeated for several different subsets of low                       |
|  leverage points. Bayesian prior weights are used to                    |
|  obtain a final estimate.                                               |
|                                                                         |
 -------------------------------------------------------------------------*/

OPTIONS PAGESIZE = 60 LINESIZE =132;

data ctrl; input x @@;
cards;

1.1 1.3  2.5  2.5  2.5  3.8  3.8  4.1  4.6  4.8  5.3  6.4  6.4  6.4  6.4
8.7  9.6 10.8  10.8 12.4 12.7 14.7 16.5 17.8  22.4  22.4 26.0 28.8
28.8 32.1 32.4 32.8 33.4 37.1 37.1 37.1 39.6 72.1 76.2 78.7 79.1
83.1 83.1 88.0 90.0 90.0 90.0  92.1  92.1  95.4
;

data ctrl;set ctrl; y2 = 1.50 + 0.5*x + 0.00194*x*x;

proc means data = ctrl noprint; var x; output out = a n = n1;

data ctrl;set ctrl;rtest ='a';
data a;set a;rtest = 'a'; drop _type_ _freq_;

/*-------------------------------------------------------------------------*
 |                                                                         |
 |  y = 1.50 + 0.5 * x + 0.00194 * x² + ε ,                                |
 |   where ε ~ iid N(0,1).                                                 |
 |                                                                         |
 -------------------------------------------------------------------------*/

data b;merge ctrl a;by rtest; n_obs = _N_;
    seed=floor(1000000000*(sqrt(time())-floor(sqrt(time()))));
       y = y2 + rannor(int(seed*(n_obs+1)/n_obs));

proc sort data = b;by x;
proc rsreg data = b; model y = x/lackfit covar=1;
Title1 'Lack of Fit Test Based on Pure error (Test for Linear Fit)';

proc sort data = b;by x; proc rsreg data = b; model y = x /lackfit;
Title1 'Lack of Fit Test Based on Pure error
        (Test for Quadratic Fit)';

data b;set b;keep x y n1;

proc print data = b;var x y;
Title1 'The set of (x,y) data points'; Title2 'y has been found by:';
Title3 'y = 1.50 + 0.5*x + 0.00194*x*x + e, ';
Title4 'with e = sigma*N(0,std)';

/*-------------------------------------------------------------------------*
 |                                                                         |
 |          UNIVARIATE Statistics for the y variable                       |
 |                                                                         |
 -------------------------------------------------------------------------*/
```

```
proc univariate data = b normal plot;var y;
Title1 'Checking the Normality Assumption';

proc reg data = b outest = g; model y=x/selection = rsquare sse;
output out = c h = leverage;
Title1 'Fitting Liner Model for Full Data Model [SSE(Full)]';

proc reg data = b; model y=x;
Title1 'Full Model Fit of Linear Model Y = X';

 /*-------------------------------------------------------------------*
 |                                                                   |
 |     Keeping those data points with low leverage                   |
 |                                                                   |
 *-------------------------------------------------------------------*/

data c;set c; if leverage < 2/n1; keep x y leverage;

proc print data = c; var x y leverage;
Title1 'Central Region for Observations with Leverage < 2/n';

proc reg data = c outest = g2;
model y=x/selection = rsquare sse;
Title1 'FITTING Model for the Central Region';
Title2 '[SSE(Reduced) = SSE(Central)]';

data g;set g; sse_f = _sse_; edf_f = _edf_;
beta0_f=intercep; beta1_f = x; rtest = 'b';
keep sse_f edf_f beta0_f beta1_f rtest;

data g2;set g2; sse_c =_sse_; edf_c = _edf_ ;
beta0_c=intercep; beta1_c = x; rtest = 'b';
keep sse_c edf_c beta0_c beta1_c rtest;

proc sort data = g;by rtest; proc sort data = g2;by rtest;
data gg;merge g g2;by rtest;

 /*-------------------------------------------------------------------*
 |                                                                   |
 |     The Rainbow test statistic:                                   |
 |                                                                   |
 |      F = (SSE(Full) ― SSE(Central))/(n ―m)                        |
 |           ------------------------------------------              |
 |                SSE(Central)/(m ―2)                                |
 |                                                                   |
 *-------------------------------------------------------------------*/

data gg;set gg;drop rtest;
f_num = (sse_f-sse_c)/(edf_f-edf_c);
f_den = sse_c/edf_c; f = f_num/f_den;
df1 = (edf_f-edf_c); df2 = edf_c;  p_value = 1-probf(f,df1,df2);

proc print data = gg;
var edf_f sse_f edf_c sse_c beta0_f beta1_f f p_value;
Title1 'Rainbow F-Test for lack of Fit';

/*-------------------------------------------------------------------*
 |                                                                   |
 |     The Macro to perform the Bootstrap                            |
 |                                                                   |
 ---- -------------------------------------------------------------*/

%macro boot;

%do i=1 %to 100;

  data cboot1;scan: set b end=last; n+1;
    if not last then goto scan;
    do i=1 to n;
      seed=floor(1000000000*(sqrt(time())-floor(sqrt(time()))));
      k=ceil(ranuni(seed)*n); set b point=k;
      if _error_ then abort; output; end; stop; keep n x y;

  proc reg data = cboot1 outest = ggcc noprint;
   model y=x/selection = rsquare sse;
    output out = cc&i h = leverage;

  data cc&i;set cc&i;if leverage < 2/n;
   proc reg data = cc&i outest = gg&i noprint;
    model y=x/selection = rsquare sse;

  data ggcc;set ggcc;sse_f = _sse_;
     edf_f = _edf_;beta0_f=intercep;
       beta1_f = x; rtest = 'c';
          keep sse_f edf_f beta0_f beta1_f rtest;
  data gg&i;set gg&i; sse_c =_sse_;
       edf_c = _edf_ ;beta0_c=intercep;
           beta1_c = x; rtest = 'c';
             keep sse_c edf_c beta0_c beta1_c rtest;

  proc sort data = ggcc;by rtest;
  proc sort data = gg&i;by rtest;
  data cgg&i;merge ggcc gg&i;by rtest;

  data cgg&i;set cgg&i;drop rtest;
  f_num = (sse_f-sse_c)/(edf_f-edf_c);
  f_den = sse_c/edf_c; f = f_num/f_den;
  df1 = (edf_f-edf_c); df2 = edf_c; p_value = 1-probf(f,df1,df2);
%end;

%do j = 2 %to 100;
  proc append base = cgg1 data = cgg&j;
%end;

%mend boot;
%boot

proc print data = cgg1; Title1 'Bootstrap Simulations';

proc means data = cgg1 noprint; var f beta0_f beta1_f;
output out = boot mean = f_boot boot_b0 boot_b1;

data boot;set boot;drop _type_ _freq_;

proc print data = boot; Title1 'Bootstrap Estimates';

data boot;set boot;rtest = 'd'; data gg;set gg;rtest = 'd';
proc sort data = gg;by rtest; proc sort data = boot;by rtest;
data fboot;merge gg boot;by rtest;

 /*-------------------------------------------------------------------------*
 |                                                                         |
 |  Bootstrap Estimates of Bias used to provide Improved                   |
 |  Estimates                                                              |
 |                                                                         |
 *-------------------------------------------------------------------------*/

data fboot;set fboot;drop rtest; bias_f = f_boot - f;
p_boot = 1-probf(f_boot,df1,df2); bias_p = p_boot - p_value;
improvef = f - bias_f; improvep = p_value - bias_p;
bias_b0 = boot_b0 - beta0_f; bias_b1 = boot_b1 - beta1_f;
im_beta0 = beta0_f - bias_b0; im_beta1 = beta1_f - bias_b1;

proc print data = fboot; var f f_boot bias_f improvef p_value
p_boot bias_p improvep beta0_f boot_b0 bias_b0 im_beta0
beta1_f boot_b1 bias_b1 im_beta1;
Title1 'Improved Estimates';
```

```
proc univariate data=cgg1;var p_value;
output out = univ1 p10 = p10 p90 = p90 q1 = q1 q3 = q3;

data cgg1; set cgg1;rtest = 'e';
data univ1;set univ1;rtest = 'e';

proc sort data = cgg1;by rtest;
proc sort data = univ1;by rtest;

data univ2;merge cgg1 univ1;by rtest;

/*----------------------------------------------------------------------------*
 |                                                                            |
 |  Using the Upper and Lower Quartiles to Define the Central                 |
 | Region                                                                     |
 |                                                                            |
 ---- ---------------------------------------------------------------------*/

data univ3;set univ2;drop rtest; if p_value < q1 then delete;
if p_value > q3 then delete;

proc means data = univ3; var f; output out = v3 mean = f_boot2;
Title1 'Alternative Estimate for the Bootstrap p_value using';
Title2 'the Central Region defined by the Upper and Lower
        Quartiles';

/*----------------------------------------------------------------------------*
 |                                                                            |
 |  Using the 10th and 90th Percentiles to Define the Central                 |
 | Region                                                                     |
 |                                                                            |
 ---- ---------------------------------------------------------------------*/

data univ4;set univ2;drop rtest; if p_value < p10 then delete;
if p_value > p90 then delete;

proc means data = univ4; var f; output out = v4 mean = f_boot3;
Title1 'Alternative Estimate for the Bootstrap p_value using';
Title2 'the Central Region defined by the 10th and 90th
Percentiles';

data gg;set gg;rtest ='f'; data v3;set v3;rtest ='f';

proc sort data = gg;by rtest; proc sort data = v3;by rtest;

data fboot2;merge gg v3;by rtest;

data fboot2;set fboot2;drop rtest;
 p_boot2 = 1-probf(f_boot2,df1,df2);
 bias_p2 = p_boot2 - p_value; improvp2 = p_value - bias_p2;

proc print data = fboot2;var p_value p_boot2 bias_p2 improvp2;
Title1 'Alternative Improved Estimates (using Q3 - Q1 Central
Region)';

data v4;set v4;rtest ='f';

proc sort data = gg;by rtest; proc sort data = v4;by rtest;
data fboot3;merge gg v4;by rtest;

data fboot3;set fboot3;drop rtest;
p_boot3 = 1-probf(f_boot3,df1,df2);
bias_p3 = p_boot3 - p_value; improvp3 = p_value - bias_p3;

proc print data = fboot3;var p_value p_boot3 bias_p3 improvp3;
Title1 'Alternative Improved Estimates (using P90 - P10 Central
Region)';

data fboot;set fboot;rtest = 'g';
data fboot2;set fboot2;rtest = 'g';
data fboot3;set fboot3;rtest = 'g';

proc sort data = fboot;by rtest;
proc sort data = fboot2;by rtest;
proc sort data = fboot3;by rtest;
data fboot4;merge fboot fboot2 fboot3;by rtest;

/*----------------------------------------------------------------------------*
 |                                                                            |
 |  Using Bayesian Weights to obtain a final estimate based                   |
 |  upon the Bootstrap for the entire data, the Bootstrap based               |
 |  on the Q3-Q1 central region, and the Bootstrap for the                    |
 |  P90-P10 Region.                                                           |
 |                                                                            |
 ----------------------------------------------------------------------------*/

/*----------------------------------------------------------------------*
 |                                                                      |
 |  In this example weights of 0.3, 0.3 and 0.4 are used.               |
 |                                                                      |
 ---- ---------------------------------------------------------------*/

data fboot4;set fboot4;drop rtest; bayes_w1 = 0.30;
bayes_w2 = 0.30; bayes_w3 = 0.40;
final_p = bayes_w1*improvp2 + bayes_w2*improvp3 +
          bayes_w3*improvep;

proc print data = fboot4; var bayes_w1 improvp2 bayes_w2
improvp3 bayes_w3 improvep final_p;

Title1 'A Final Estimate is obtained by using Bayesian Prior
        Weights for';
Title2 'the Bootstrap (entire data), the Bootstrap based on the';
Title3 'Q3-Q1 central region, and the Bootstrap for P90-P10
        Region';

run;
```

REFERENCES

Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Neill, J.W., and Johnson, D.E. (1984), “Testing for Lack-of-fit in Regression: a Review,” *Communications in Statistics - Theory and Methods*, 13(4), 485-511.

Utts, J.M. (1982), "The Rainbow Test for Lack of Fit in Regression," *Communications in Statistics - Theory and Methods*, A11, 2801-2815.

SAS Institute Inc. (1990), *SAS/STAT User’s Guide, Version 6, Fourth Edition*, Cary, NC: SASInstitute INC.

ACKNOWLEDGMENTS

SAS, and SAS/STAT are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Paul Johnson.
P.O. Box 4146
Davis
CA 95617-4146

E-Mail: *JohnsonP12@aol.com*