DIAGNOSTICS FOR MODEL SELECTION IN NONLINEAR REGRESSION BASED ON THE CONCEPT OF CLOSENESS TO LINEARITY USING SAS/IML® SOFTWARE

M. S. Blanchard, University of Cincinnati
M. S. Levy, University of Cincinnati

ABSTRACT

The purpose of this paper is to present a SAS/IML® software program that can be used in conjunction with the NLIN procedure in the SAS/STAT® software to calculate Box's bias approximation and Hougaard's measure of skewness. As an example, Box's bias approximation and Hougaard's measure of skewness were calculated on a leaf production data set (Ratkowsky, 1983, 1990) using the model \( Y = B_1 - B_2B_3 \). This program is compatible with Kugler and Lee's (1990) SAS/IML software program for curvature measures.

INTRODUCTION

The least squares estimates of the parameters for a linear regression model \( (Y = \beta X + e; \beta \) is the vector of unknown parameters, and the \( e \)'s are \( N(0, \sigma^2) \) ) are linear combinations of the random variable. Consequently, the estimates are, unbiased, normally distributed minimum variance estimates. In the case of nonlinear models \( (Y = f(0, x) + e; \theta \) is the vector of unknown parameters and the \( e \)'s are \( N(0, \sigma^2) \) ) the parameter estimates are not linear combinations of the random variable and do not possess these qualities. The amount of bias, non-normality, and deviation from minimum variance depends on both the model and the data set in combination. One possible goal in selecting the "best" nonlinear model for a particular data set is to find the model that in combination with that data set behaves closest to linear (Ratkowsky, 1983, 1990).

A number of diagnostic measures of nonlinearity or closeness to linearity have been devised, including the measures of maximum relative curvature defined by Bates and Watts (1980). Bates and Watts (1980) divide the concept of nonlinearity into two parts, intrinsic nonlinearity and parameter effects nonlinearity. Intrinsic nonlinearity is defined as the curvature of the solution locus at the point \( (\hat{\theta}) \) and is an indicator of how well the planar assumption associated with the linear model is met (Bates and Watts, 1980; Ratkowsky, 1983). The parameter effects measure of maximum relative curvature is an indicator of how well the uniform coordinate assumption associated with the linear model is met (Bates and Watts, 1980) that is the amount of "unequal spacing and lack of parallelism of the parameter lines projected on the tangent plane to the solution locus (Ratkowsky, 1983, p.18)". For most models of interest, intrinsic nonlinearity is small, and not of practical significance (Bates and Watts, 1980; Clarke, 1997; Ratkowsky, 1983). However, parameter effects nonlinearity is important in model selection (Bates and Watts, 1980; Clarke, 1987; Ratkowsky, 1983). Cook and Witmer (1985) examined the parameter effects measure of nonlinearity and stated that "Bates-Watts criterion has merit but it falls short of being a reliable indicator of the accuracy of the tangent plane approximation (Cook and Witmer, 1985, p.877)". Since it is a measure of maximum curvature it is not surprising that it could overestimate the amount of curvature in the parameter space. Parameter effects measure of curvature can also underestimate the curvature since it is a comparison of the quadratic term to the linear term in the Taylor approximation, and does not consider the contribution of any of the higher order terms.

Both Cook and Witmer (1985) and Clarke (1987) suggested that one of the failings of the parameter effects curvature as a measure of nonlinearity is that it condenses a multidimensional phenomena into a single value. Another way of investigating closeness to linearity of a model is to examine the sampling distribution of the parameter estimates \( (\hat{\theta}) \). This type of study allows for separate analysis of each of the parameter estimates. Two diagnostic measures investigating the statistical properties of the estimates are Box's bias approximation (Cook, Tsai & Wei, 1986) and Hougaard's measure of skewness (Ratkowsky, 1990). Box's bias approximation looks at the difference between the linear and quadratic components of the Taylor approximation. It provides similar information to the parameter effects curvature measure of Bates and Watts (1980) for each of the parameters, separately. However, it is location and scale dependent and therefore can only be used to identify parameters that have the greatest nonlinearity within a model (Ratkowsky, 1990). Hougaard's measure of skewness capitalizes on the relationship between closeness to linearity and non-normality providing further information about the distribution of the estimates.

The purpose of this paper is to present a SAS/IML® software program (SAS/IML User's Guide, 1985) that can be used in conjunction with PROC NLIN from SAS/STAT® software (SAS User's Guide: Statistics, 1985) to calculate Box's bias approximation (Cook, Tsai, Wei, 1986) and Hougaard's measure of skewness (Ratkowsky, 1990). This program is compatible with Kugler and Lee's (1990) SAS/IML software program for curvature measures. The atanu and iter_gam modules from Kugler and Lee's (1990) SAS/IML software program for curvature measures can be appended to the program presented here to calculate Bates and Watts (1980) measures of maximum relative curvature.

PROGRAM DESCRIPTION

The program is divided into four sections. In the first section, the user provides the program with the data, the number of parameters \( (p) \), the vector of estimates \( (\hat{\theta}) \), and the sums of squares for error.

The second section is the module labeled deriv. In this module the user provides the first and second derivatives. This module defines the npm first derivative matrix \( (D2) \) and the npm second derivative matrix \( (F3) \). The second derivative matrix is stored in the form of an npm matrix, with each of the \( n \) npm faces laid one below the other. The expected difference between the quadratic and linear components of the Taylor approximation \( (H2) \) is also calculated.

The third section is the module labeled bias. In this module Box's bias approximation is calculated in multivariate form as given by Cook, Tsai & Wei (1986): Bias \( = (D2_2F2') - (D2_2F2) \), where \( H2=1/2\sigma^2[(D2_2F2') - (D2_2F2)] \). Ratkowsky (1983) suggested using an absolute value of greater than 1% as an indicator of nonlinear behavior.

The fourth section is the module labeled skew. The module skew calculates Hougaard's measure of skewness as defined by Ratkowsky (1990).
\[ E(\hat{\theta} - \hat{\theta})^2 = \text{VAR}(\hat{\theta}) \]

where \( \text{VAR}(\hat{\theta}) \) is calculated as

\[ E(\hat{\theta} - \hat{\theta})^2 = \text{VAR}(\hat{\theta}) \]

Ratkowsky gives the following rule of thumb for assessing nonlinearity of the parameter \( \theta \).

If \( g_\theta < 0.1 \) the estimator \( \hat{\theta} \) of the parameter \( \theta \) is very close-to-linear in behavior and, if \( 0.1 < g_\theta < 0.25 \), the estimator is reasonably close-to-linear. For \( g_\theta > 0.25 \), the skewness is very apparent, and \( g_\theta \) indicates considerable nonlinear behavior (Ratkowsky, 1990, p.28).

(Note: \( g_\theta \) can be negative in which case one would perform a lower tailed test using the absolute value of \( g_\theta \).)

**EXAMPLE**

The diagnostic measures will be calculated on the leaf production data set (Ratkowsky, 1983; 1990) using the model

\[ Y = B_1 + B_2 X + \epsilon \]

The parameter estimates and sums of squares for error were calculated using PROC NLIN from SAS/STAT software (SAS User's Guide: Statistics, 1985).

**Program:**

```sas
PROC IML;
*RESET PRINT;
OA=112.094.
23.119.
40.199.
92.260.
156.309.
215.331;
*(RATKOWSKY, 1983, P. 102; RATKOWSKY, 1990, P.230);
X=DAT(I.11); Y=DAT(I.21};
P=3; N_=NROW(X); NP=N_#P;
B=10.33502041. 0.29551211. 0.983829681;
*-LEAST SQUARES ESTIMATES;
S2=.000605635/(N_-P); *- MSE;
~~.o=SQRT(S2*P);
*-SCALING FACTOR (RATKOWSKY, 1983);
START DERIV;
*- CALCULATES FIRST DERIVATIVE MATRIX;
DO 1=1 TO NROW(X);
D1=1/RHO;
D2=-B(I3.I}##X(lI.I)/RHO;
D3=-B(I2.I}#X(II.I}#B(I3.1}##(X(II.I}-1 )/RHO;
IF 1=1 THEN DO;
D=D1I1D2I1D3;END;
ELSE DO; D=D//(D1I1D2I1D3);END;
END;
*-THE SIZE OF THE D MATRIX WILL VARY BASED ON B;
D2=RHO#D; *- UNSCALED 1ST DERIVATIVE MATRIX; *
PXP MATRIX;
PRINTB;
PRINT S2;
G=D2"D2;
GINV-INV(G); *- (PXP MATRIX);
*- CALCULATES 2ND DERIVATIVE MATRIX;
DO I=1 TO NROW(X);
D11=0;
D22=0;
D33~B([2,1]#X(I1.I)#X(II.I}-1 )#B([3,1]#X(II.I}-2)/RHO;
D12=0;
D13=0;
D23=X(I.I}#B([3,1]#X(II.I}-1 )/RHO;
F=(D11[D12][D13][C12][D21][D23])
//(C13[D23][D33]);
*THE SIZE OF THE F MATRIX WILL VARY BASED ON B;
F2=RHO*F; *- UNSCALED 2ND DERIVATIVES (PXP MATRIX);
H=-(1/2)*S2*TRACE(GINV*F2);
IF I=1 THEN DO;
H2=H;
HT=VECDIAG(GINV*F2*GINV); FU=F;
F3=F3/F2; *- NXXP MATRIX;
END;
END;
FINISH; *- DERIV;
RUN DERIV;
****;
*- CALCULATE BOX'S BIAS APPROXIMATION;
*- (COOK, TSAI, & WEI, 1986);
START BIAS;
BIAS=GINV*D2"H2;
PCBIAS=(BIAS*100)/B;
PRINT BIAS;
PRINT PCBIAS;
FINISH; *- BIAS;
RUN BIAS;
****;
*- CALCULATE HOUGAARD'S MEASURE OF SKEWNESS;
*- (RATKOWSKY, 1990);
START SKEW;
DO 1=1 TO P;
IF 1=1 THEN DO J=1 TO NROW(X);
IF J=1 THEN DO;
W=(D2(I.I}#F3(I.;P.I});
W1=W1+W;
ELSE DO;
W=(D2(I.I}#F3(I(J-1 )#P+1;J#P.I});
W1=W1+W;
W=W1/W;
END;
ELSE DO;
W=(D2(I.I}#F3(I.;P.I});
W1=W1+W;
END;
IF I=1 THEN DO;
M=0;
IF I=1 THEN DO J=1 TO P;
DO K=1 TO P;
DO L=1 TO P;
M=M+(GINV(I.I}#GINV(I.K)I#GINV(I.L)I)#
\[
(WN(I(J-l)#P+K,LI)+WN((K-1)#P+J,LI)+WN((L-1)#P+K,JI));
END;
END;
END;
E3=(S2)#2#M;
G1=E3/(S2#GINV(II,11)#(3/2);
END;
DO I=2 TO P; M=0;
DO J=1 TO P;
DO K=1 TO P;
DO L=1 TO P;
M=M+((GINV(II,JI)#GINV(II,KI)#GINV(II,LI)#
(WN(I(J-l)#P+K,LI)+WN((K-1)#P+J,LI)+WN((L-1)#P+K,JI)));
END;
END;
END;
E31=(S2)#2#M;
E3=E3/E31;
G1=G1/G11;
END;
PRINT E3;
PRINT G1;
FINISH; * SKEW;
RUN SKEW;

Output:

<table>
<thead>
<tr>
<th>COL1</th>
<th>COL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
</tr>
<tr>
<td>ROW1</td>
<td>0.3350</td>
</tr>
<tr>
<td>ROW2</td>
<td>0.2955</td>
</tr>
<tr>
<td>ROW3</td>
<td>0.9838</td>
</tr>
<tr>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>ROW1</td>
<td>2.0E-04</td>
</tr>
<tr>
<td>BIAS</td>
<td>COL1</td>
</tr>
<tr>
<td>ROW1</td>
<td>0.0029714</td>
</tr>
<tr>
<td>ROW2</td>
<td>0.0047364</td>
</tr>
<tr>
<td>ROW3</td>
<td>-2.2E-04</td>
</tr>
<tr>
<td>PCTBIAS</td>
<td>COL1</td>
</tr>
<tr>
<td>ROW1</td>
<td>0.8869</td>
</tr>
<tr>
<td>ROW2</td>
<td>1.6028</td>
</tr>
<tr>
<td>ROW3</td>
<td>-0.0222</td>
</tr>
<tr>
<td>E3</td>
<td>COL1</td>
</tr>
<tr>
<td>ROW1</td>
<td>3.7E-06</td>
</tr>
<tr>
<td>ROW2</td>
<td>3.8E-07</td>
</tr>
<tr>
<td>ROW3</td>
<td>-1.8E-08</td>
</tr>
<tr>
<td>G1</td>
<td>COL1</td>
</tr>
<tr>
<td>ROW1</td>
<td>0.7175</td>
</tr>
<tr>
<td>ROW2</td>
<td>0.0512</td>
</tr>
<tr>
<td>ROW3</td>
<td>-0.9533</td>
</tr>
</tbody>
</table>

REFERENCES


SAS, SAS/IML, and SAS/STAT are registered trademarks of SAS Institute Inc., Cary, NC, in the USA and other countries.

@Indicates USA registration.

For further information the authors may be contacted at the following addresses.

M.S. Blanchard
Statistical & Data Analysis Center
Harvard School of Public Health
665 Huntington Ave., Room 1206
Boston, MA 02115

M.S. Levy
Dept. of Quantitative Analysis
Carl H. Lindner Hall
Cincinnati, OH 45221