Computing Case Diagnostics for Structural Equation Models

Noel G. Cadigan, Northwest Atlantic Fisheries Centre
St. John's, Newfoundland, Canada

Abstract

A SAS® program is presented to compute case diagnostics for the maximum likelihood analysis of a LISREL type of structural equation model. The diagnostics computed are Cook's Distance, COVRATIO and Likelihood Distance. The SAS program essentially consists of a data step and a macro and is very easy to implement. An example is considered and the application of the diagnostic measures is briefly discussed.

Introduction

Structural equation models are increasingly becoming common analytical tools for studying linear relationships among variables, especially by researchers in the social sciences. The LISREL (Jöreskog 1973) formulation of a structural equation model is considered here. The essential information required to analyze data with structural equation models and a short history of the development of structural equation models is given in Bollen (1989).

It is often of interest when performing a statistical analysis (and a good practise in general) to compute case diagnostics. A case refers to the collection of observations for an individual, etc. The purpose of case diagnostics is to provide measures of the influence of cases in the analysis. A good treatment of influence is presented in Cook and Weisberg (1982). Tanaka et al (1991) present some diagnostics for confirmatory factor models that are easily extended to structural equation models. The purpose of this paper is to present a SAS program that computes these diagnostics.

Structural equation model

Structural equation models typically deal only with the analysis of covariance structures. Let $x$ and $y$ be vectors of observable random variables. The measurement models in a LISREL model are:

$$
\begin{align*}
    y &= A_{y} \eta + \varepsilon, \\
    x &= A_{x} \xi + \delta.
\end{align*}
$$

The vectors $\eta$ and $\xi$ are unobservable latent random vector variables. Let $COV(\xi) = \Phi$, $\varepsilon$ and $\delta$ are independent random error vectors in the measurement of $x$ and $y$ and have covariances $\Theta_{\varepsilon}$ and $\Theta_{\delta}$ respectively.

The structural part of the general structural equation model is

$$
\eta = B\eta + G\xi + \zeta.
$$

This portion of the model links the two equations in (1). $\zeta$ is a random vector of process errors with covariance $\Psi$. It is usually assumed that $\zeta$ is independent of $\xi, \varepsilon$ or $\delta$.

The population covariance matrix for $y$ and $x$ ($\Sigma$) may be partitioned as

$$
\Sigma = \begin{bmatrix}
    \Sigma_{yy} & \Sigma_{yx} \\
    \Sigma_{xy} & \Sigma_{xx}
\end{bmatrix}.
$$

From (1)

$$
\Sigma_{yy} = A_{y}E(\eta\eta')A_{y} + \Theta_{\varepsilon}.
$$
\[ \Sigma_{zz} = A_z \Phi A'_z + \Theta_\delta. \]

From (2) \( \eta = (I - B)^{-1} \Gamma \xi + \zeta \) so that
\[
\begin{align*}
\Sigma_{yy} &= A_y(I - B)^{-1}(\Gamma \Phi \Gamma') \\
&\quad + \Psi(I - B)^{-1}A'_y + \Theta_\varepsilon \quad \text{and} \\
\Sigma_{yx} &= A_y(I - B)^{-1}\Gamma \Phi A'_z.
\end{align*}
\]

The estimation of \( \Sigma \) involves a scalar fit function that measures the discrepancy between the observed variance-covariance matrix (S) and \( \Sigma \). The maximum likelihood fit function is the most popular, although other methods such as unweighted and generalized least squares (Bollen 1989) are commonly used. The fit function that is minimized in order to find the maximum likelihood estimates when the data are from a multivariate normal distribution is:

\[
\log | \Sigma | + \text{tr}(S \Sigma^{-1}) - \log(| S |) - p,
\]

where tr is the matrix trace and p is the dimension of S. This function is used here to produce parameter estimates.

### Influence

The influence statistics considered in Tanaka et al. (1991) for the ith case are:

1. **Cook's Distance (CD):**
   \[
   CD_i = (\hat{\theta}_{(i)} - \hat{\theta})' \hat{T}(\hat{\theta}_{(i)} - \hat{\theta}),
   \]
   where \( \hat{\theta}_{(i)} \) is the maximum likelihood estimate (\( \Sigma = \Sigma(\theta) \)) with the ith case deleted, \( \hat{T} = COV(\hat{\theta})^{-1} = -E(\tfrac{\partial^2 L}{\partial \theta \partial \theta'}) \) and \( L \) is the loglikelihood. The CALIS procedure in SAS/STAT\(^{\circledR} \) approximates \( \hat{T} \) and this is used in the program.

2. **Covariance Ratio (CVR):**
   \[
   CVR_i = \frac{| \hat{T} |}{| \hat{T}_{(i)} |},
   \]

### Implementation

The SAS code to compute the above influence statistics is presented in Appendix A. This program requires SAS/IML\(^{\circledR} \), SAS/STAT and SAS/MACRO\(^{\circledR} \) software and was developed on a unix workstation running SAS V6.07.02.

### Example. Fish data

This example was considered in Myers and Cadigan (1993) and involves estimating the significance of density-dependent mortality in a fish stock. The data used are estimates of abundance at age in the Georges Banks cod stock found off the east coast of the United States (Anon. 1991) and are presented in Table B.1 in Appendix B. The analysis in Myers and Cadigan (1993) involved fitting a one factor confirmatory factor model to the natural logarithm of the data. 1 > \( \lambda_2 > \lambda_3 > \lambda_4 \) is taken as evidence of density-dependent mortality (see Myers and Cadigan (1993) for details).

The estimates are presented in Table 1. They do not support density-dependent mortality although most of the stocks examined in Myers...
and Cadigan (1993) did.

Table 1. Estimates of parameters in the density-dependent mortality model (standard errors in parentheses).

<table>
<thead>
<tr>
<th>Rank</th>
<th>(\hat{\lambda})</th>
<th>(\hat{\phi})</th>
<th>(\hat{\theta})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>-</td>
<td>0.81(0.25)</td>
<td></td>
</tr>
<tr>
<td>1.01</td>
<td>0.75(0.39)</td>
<td>0.24(0.09)</td>
<td></td>
</tr>
<tr>
<td>1.04</td>
<td>-</td>
<td>0.06(0.06)</td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td>-</td>
<td>0.23(0.23)</td>
<td></td>
</tr>
</tbody>
</table>

\(\chi^2 = 8.52\) with 2 d.f.

The five most influential cases are presented in Table 2. CVR does not indicate that any case is extremely influential on the estimated standard errors of parameter estimates; however, both CD and LD indicate that case 6 (cohort 1967) is influential both in terms of parameter estimates and goodness-of-fit.

The estimates obtained with case 6 deleted are presented in Table 3. In this analysis a negative estimate of \(\theta_{3,3}\) resulted (not uncommon with small sample sizes) so all error variances were constrained to be greater than the minimum obtained in the full sample analysis. The estimates of \(\lambda\)'s now provide evidence for density-dependent mortality in Georges Banks cod which is consistent with many other fish stocks. Although the standard errors generally decrease with case 6 removed the \(\chi^2\) goodness-of-fit statistic increases substantially; this casts some doubt as to the usefulness of these measures in structural equation models with small samples.

Table 2. Influence diagnostics for the density-dependent mortality model (case number in parentheses).

<table>
<thead>
<tr>
<th>Rank</th>
<th>CD</th>
<th>CVR</th>
<th>LD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27.67(6)</td>
<td>4.50(11)</td>
<td>6.91(6)</td>
</tr>
<tr>
<td>2</td>
<td>6.08(12)</td>
<td>3.01(16)</td>
<td>-3.09(22)</td>
</tr>
<tr>
<td>3</td>
<td>2.91(22)</td>
<td>2.09(4)</td>
<td>-2.79(19)</td>
</tr>
<tr>
<td>4</td>
<td>2.35(19)</td>
<td>2.02(20)</td>
<td>1.73(28)</td>
</tr>
<tr>
<td>5</td>
<td>2.27(11)</td>
<td>0.08(6)</td>
<td>1.68(12)</td>
</tr>
</tbody>
</table>

Table 3. Estimates of parameters in the density-dependent mortality model without case 6 (standard errors in parentheses).

<table>
<thead>
<tr>
<th>(\lambda)</th>
<th>(\phi)</th>
<th>(\theta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00( - )</td>
<td>-</td>
<td>0.57(0.19)</td>
</tr>
<tr>
<td>0.92(0.16)</td>
<td>0.99(0.43)</td>
<td>0.06(0.06)</td>
</tr>
<tr>
<td>0.88(0.16)</td>
<td>-</td>
<td>0.15(0.07)</td>
</tr>
<tr>
<td>0.67(0.17)</td>
<td>-</td>
<td>0.34(0.11)</td>
</tr>
</tbody>
</table>

\(\chi^2 = 16.12\) with 2 d.f.

Bibliography


Appendix A. SAS Program

* data to be analyzed must be in a sas;
* dataset called dataset the input sas;
* data set is used to compute the SEM;
data input;set dataset;
ind = _N_;call symput('n',_N_);run;

* the macro 'influen' computes the;
  * diagnostics and outs them in a data;
  * set called infdiag. This file contains;
    * 1. N - the observation number;
    * 2. CookD and rCookD - Cook's Distance;
      * and the rank of Cook's Distance;
    * 3. CVR and rCVR - COVRATIO and the;
      * rank of COVRATIO;
    * 4. LD rLD - Likelihood Distance and;
      * the rank of Likelihood Distance;
    * the user must modify the PROC CALIS;
    * code for their model;

%macro influen;
*loop through observations;
%do i = 0 %to &n;
*create covariance matrix for PROC CALIS;
* to compute the SEM;
  proc corr cov data=input
   (where=(ind="&i» noprint
    outp=covm(type=cov);
  * Produce maximum likelihood estimates;
  * of the SEM;
  *sas data test outest and outram contain;
  * model estimates;
  proc calis cov data=covm method=ml
    outest=outest outram=outram
    nopr ridge;
    lineqs
      x1 = f1 + e1,
      x2 = 12 f1 + e2,
      x3 = 13 f1 + e3,
      x4 = 14 f1 + e4;
    std
e1-e4 = ve1-ve4,
f1 = phi;run;
  * save the parameter estimates and;
  * information matrix;
data estmi;set outest;
  if _TYPE_ = 'INFORMAT' or _TYPE_ = 'PARMS';
  ind = &i;run;
* save the loglikelihood (called;
* chisquare);
data chisqi;set outram;
  if _NAME_= 'CHISQUAR';
  rename _ESTIM_ = chisq;
  ind = &i;
  keep ind _ESTIM_; run;
* append the results in [dataset]i;
* to [dataset];
  %if &i = 0 %then %do;
  data estmi;set estmi;
  data chisqi;set chisqi;run;
  %end;
  %else %do;
    proc append base=estm
      data=estmi force;
    proc append base=chisq
      data=chisqi force;run;
  %end;
%end;
* compute the influence diagnostic;
* measures;
proc iml;
do i = 0 to &n;
  use estmi;
  read all into Inf
  where(_TYPE_ = 'INFORMAT')
    &(ind=i);
  read all into theta
  where(_TYPE_ = 'PARMS')
    &(ind=i));
close estmi;
use chisqi;
read all var[chisq] where(ind=i);
close chisqi;
p = ncol(theta)-1;
theta = shape(theta[1:p],p);
Inf = Inf[1:i:p];
if i = 0 then do;
  chisqf = chisq;thetaf = theta;
  Inff = Inf;
end;
k = (&n-1)/2;
if i = 1 then do;
CookD = k*t(theta-thetaf) *Inff*(theta-thetaf);
CVR = det(Inff)/det(Inf);
LD = chisq - chisqf;
end;
if i > 1 then do;
CookD = CookD/(k*t(theta-thetaf) *Inff*(theta-thetaf));
CVR = CVR/det(Inff)/det(Inf);
LD = LD/chisq - chisqf;
end;
end;
N = t(1:&n);
rCookD = rank(CookD);
rCVR = rank(abs(CVR-1));
rLD = rank(abs(LD));
create infdiag var{N CookD rCookD eVR rCVR LD rLD};
append;close infdiag;
quit;run;
%mend influen;

Appendix B. Example data

Table B.1. Data for the Georges Banks cod stock from the U.S.A. Northeast Fisheries Center spring and fall surveys. Units are in numbers per tow.

<table>
<thead>
<tr>
<th>cohort</th>
<th>SPRING</th>
<th>FALL</th>
<th>SPRING</th>
<th>FALL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 yr.</td>
<td>1 yr.</td>
<td>2 yr.</td>
<td>2 yr.</td>
</tr>
<tr>
<td></td>
<td>olds</td>
<td>olds</td>
<td>olds</td>
<td>olds</td>
</tr>
<tr>
<td>1962</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.778</td>
</tr>
<tr>
<td>1963</td>
<td>.</td>
<td>0.719</td>
<td>.</td>
<td>0.699</td>
</tr>
<tr>
<td>1964</td>
<td>.</td>
<td>0.640</td>
<td>.</td>
<td>0.998</td>
</tr>
<tr>
<td>1965</td>
<td>.</td>
<td>1.299</td>
<td>.</td>
<td>1.000</td>
</tr>
<tr>
<td>1966</td>
<td>.</td>
<td>1.693</td>
<td>.</td>
<td>1.334</td>
</tr>
<tr>
<td>1967</td>
<td>.</td>
<td>7.596</td>
<td>1.615</td>
<td>1.611</td>
</tr>
<tr>
<td>1968</td>
<td>0.136</td>
<td>0.314</td>
<td>0.546</td>
<td>0.622</td>
</tr>
<tr>
<td>1969</td>
<td>0.123</td>
<td>0.343</td>
<td>0.814</td>
<td>1.353</td>
</tr>
<tr>
<td>1970</td>
<td>0.381</td>
<td>1.688</td>
<td>0.819</td>
<td>0.632</td>
</tr>
<tr>
<td>1971</td>
<td>0.207</td>
<td>0.602</td>
<td>1.833</td>
<td>1.295</td>
</tr>
<tr>
<td>1972</td>
<td>2.902</td>
<td>7.443</td>
<td>11.64</td>
<td>6.070</td>
</tr>
</tbody>
</table>

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