**A program to calculate the empirical bias in autocorrelation estimators**

Jaume Amau and Roser Bono  
University of Barcelona

The statistical analysis of short time series designs is influenced by the presence of serial dependence. Therefore, it is important to correctly estimate the first-order autocorrelation in behavioral data. The empirical bias is an indicator generally used to evaluate the adequacy degree of an estimator. This paper presents the Bias program, a Monte Carlo simulation program that generates first-order autoregressive processes and calculates the bias in three autocorrelation estimators (r1, r+ and r1') for different values of the lag-one autocorrelation parameter and sample sizes. The program has been designed with MATLAB programming language and it runs on IBM-PC compatible computers with a 486 or later processor.

**Programa para el cálculo del sesgo empírico de estimadores de autocorrelación.** El análisis estadístico de los diseños de series temporales cortas está influido por la presencia de dependencia serial. De ahí la importancia de estimar correctamente la autocorrelación de primer orden en datos conductuales. El sesgo empírico es un indicador generalmente usado para evaluar el grado de precisión de un estimador. Este artículo presenta el programa Bias, un programa de simulación Monte Carlo que genera procesos autoregresivos de primer orden y calcula el sesgo de tres estimadores de autocorrelación (r1, r+ y r1') para diferentes valores del parámetro de autocorrelación de retardo uno y tamaños muestrales. El programa ha sido diseñado mediante el lenguaje de programación MATLAB y funciona en ordenadores IBM-PC con un procesador 486 o superior.

The study of autocorrelation in time series was considered from the end of the seventies (Bono & Amau, 1996, 2000; Busk & Marascuilo, 1988; Escudero & Vallejo, 2000; Glass, Willson, & Gottman, 1975; Gorsuch, 1983; Hartman et al., 1980; Huitema, 1985, 1988; Jones, Vaught, & Weinrott, 1977; Rosel, Jara, & Oliver, 1999; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987; Vallejo, 1994, among others). In the nineties, the works were oriented towards the correction of the bias generated in short time-series by the conventional autocorrelation estimator r1. Matyas and Greenwood (1991) verified that the difference between the mean of the conventional estimator r1 and the value of the autocorrelation parameter p1 increases with short time series. These authors concluded that the relationship between p1 and the mean r1 was nonlinear. The greatest coincidence between calculated mean r1 and p1 occurred when p1 was around -0.2 or 0.3. According to these values, when p1 increased positively, the calculated mean r1 underestimated r1 with a negative bias, while when the p1 increased negatively, the calculated mean r1 underestimated p1 but with a positive bias. Based on this simulation study, it was deduced that when small sample sizes were used, the r1 values were biased. Huitema and McKean (1991) proposed the modified estimator r1+. This estimator corrects the empirical bias for positive values of the autocorrelation parameter by adding 1/nt to r1. Arna (1999) and Amau and Bono (2001) proposed the estimator r1' which consists of the correction of r1 by the absolute value of a fitting model. This model is obtained from the polynomial function of the bias for each sample size. The estimator r1 reduces the bias for small sample sizes, both for positive as well as negative values of p1.

This article describes a computer program using MATLAB, Version 5.2 (1998). The program, called Bias, makes it possible to calculate the empirical bias in r1, r+ and r1' by Monte Carlo simulation. This program requires the specification of the parameter p1 and the sample size (n). The structure of the program is as follows (see the Appendix): simulation of lag-one autoregressive processes, calculation of the estimators r1 and r1+, polynomial models for different sample sizes, calculation of the estimator r1' and estimation of the empirical bias in r1, r+ and r1'. Bias runs under the Windows 3.1 or later operating system on IBM-PC compatible with a 486 or later processor, and 8 MB of extended memory.

In the first place, the program simulates first-order autoregressive processes. The *randn* function (Forsythe, Malcolm, & Moler, 1977) is used to generate observations from a lag-one autoregressive process:

\[ Y_t = p_1 Y_{t-1} + e_t \]  

(1)

where \( Y_t \) is the score on the response measure at time \( t \), \( p_1 \) is the autocorrelation parameter and \( e_t \) is a random normal variable with a mean of zero and a variance of one. Each time series starts with a normal variate \( Y_0 \) having zero mean and variance \( 1/(1-p_1^2) \). For each combination of \( n \) and \( p_1 \), 10,000 samples are generated. The first 30 observations of each series are drop-
ped to eliminate any dependency between them (DeCarlo & Tryon, 1993; Huitema & McKean, 1991, 1994a, 1994b). After this, the program calculates the estimators \( r_1 \) (equation 2) and \( r_{1+} \) (equation 3) for each simulated time series. The conventional estimator \( r_1 \) is defined by the following equation:

\[
    r_1 = \frac{\sum_{t=1}^{n-1} (Y_t - \bar{Y})(Y_{t+1} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2}
\]

where \( Y_t \) and \( Y_{t+1} \) are the data obtained in the intervals \( t \) and \( n+1 \) and \( \bar{Y} = \frac{1}{n} \sum_{t=1}^{n} Y_t \). The estimator \( r_1 \) is biased with small sample sizes because the numerator only includes \( n-1 \) terms instead of \( n \). Huitema and McKean (1991) proposed the modified estimator \( r_{1+} \) that corrects, to some extent, for the negative bias generated by positive autocorrelations

\[
    r_{1+} = r_1 + (1/n)
\]

The empirical bias function of the conventional estimator \( r_1 \) mildly drifts from the linearity (Arnaud, 1999; Arnaud & Bono, 2001; DeCarlo & Tryon, 1993; Huitema & McKean, 1991, 1994a, 1994b). Models with significant coefficients of the empirical bias function for \( n = 6, 10, 20 \) and 30 were derived with polynomial fitting. Within the MATLAB environment, another program that calculated the empirical bias of the estimator \( r_1 \) in function of \( \rho_1 \) and \( n \) by Monte Carlo simulation was designed (Arnaud, 1999; Arnaud & Bono, 2001). The polytool function of the Statistics Toolbox of the MATLAB was used for polynomial fitting. With the significant polynomial coefficients at 5% (Table 1) the following fitting models based on \( n \) were designed and were introduced into the Bias program.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficients of the polynomial functions of the bias</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( n )</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>CI (0.05)</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>CI (0.05)</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>CI (0.05)</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>CI (0.05)</td>
</tr>
</tbody>
</table>

CI = Confidence Interval

\[
    \text{Model}_{\text{lin}} = -0.1648 - 0.5643 r_1 - 0.0916 r_1^2
\]

(4)

\[
    \text{Model}_{\text{quad}} = -0.0972 - 0.3760 r_1 - 0.0676 r_1^2
\]

(5)

\[
    \text{Model}_{\text{poly}} = -0.0482 - 0.2028 r_1 - 0.0333 r_1^2
\]

(6)

\[
    \text{Model}_{\text{poly}} = -0.0373 - 0.1360 r_1
\]

(7)

For sample sizes other than those studied in this paper, it is previously necessary to calculate, by simulation, the empirical biases of the conventional estimator. Once the corresponding empirical biases have been found, the polynomial fitting is carried out with polytool function specifying the \( \rho_1 \) values, empirical biases previously obtained and the degree of the polynomial fit. In this way, the significant coefficients are obtained. These must be included in the polynomial models for different sample sizes statement indicating the value of \( n \).

Using the equations 4-7, the Bias program calculates the estimator \( r_{1+} \) by adding the corresponding polynomial model in absolute values to \( r_1 \).

\[
    r_{1+} = r_1 + \text{Fitting Model}
\]

(8)

Finally, for each simulated sample, the program calculates the empirical bias in \( r_1, r_{1+} \) and \( r_{1+} \), which is the difference between the mean value of the estimator and the value of the parameter \( \rho_1 \). The Bias program only calculates \( r_{1+} \) for four sample sizes (\( n = 6, 10, 20 \) and 30). For other \( n \) values, the polynomial fitting of the bias must be performed first in order to determine the corresponding correction model and to introduce it into the program.

### Running the program

The Bias program is a function M-file and the input variables in the MATLAB workspace, autoregressive parameter (\( \rho_0 \)) and sample size (\( n \)), must be introduced into it. As can be seen in the Appendix, the first line of the function M-file specifies the file name (bias), the input variables names (\( \rho_0 \) and \( n \)) and output variables names (\( r_1, r_{1+}, \) and \( r_{1+} \)). The following expression must be specified in the MATLAB workspace in order to run the program: \[ r_1, r_{1+}, r_{1+} \]= bias(\( \rho_0, n \)). The MATLAB executes the commands in Bias program and the results are shown in the MATLAB workspace.

### Availability

Interested users who prefer not to type the program can request a file of the program by e-mailing the second author (rbono@psi.ub.es). It has been designed with the Version 5.2 of MATLAB but it can also work with previous versions.

### Acknowledgements

The research was supported by Grant PS95-0228 from the General Council of Scientific and Technical Investigation, resolution of the Secretary of State of Universities and Investigation of the Ministry of Education and Culture (Spain).
APPENDIX
Program Listing

function [r1_bias, r1_plus_bias, r1_prime_bias]=bias(rho,n)

% **************************************************
% LAG-ONE AUTOREGRESSIVE PROCESSES SIMULATION
% **************************************************
% Number of simulations
for s=1:10000
% Number of observations
obs=30+n;
% Variance
variances=1/(1-rho^2);
% Standard deviation
sd=sqrt(variances);
% Series with mean 0 and fixed standard deviation
y0=normrnd(0,sd,obs,1);
% Random series with mean 0 and variance 1
y=randn(obs,1);
% First value of the series
y(1,1)=y0(1)+e(1);
% Next values of the series
for t=2:obs
y(t,1)=rho*y(t-1,1)+e(t);
t=t+1;
end
% Eliminations of the first 30 observations
y=y(31:obs);

% **************************************************
% CALCULATION OF r1
% **************************************************
% y_mean=mean(y);
% denominator=(y-y_mean).^2;
% numerator_sum=sum(denominator);
y_lag=y(2:n);
y=y(1:n-1);
% y_lag_diff=y_lag-y_mean;
% y_diff=y-y_mean;
% numerator=y_lag_diff.*y_diff;
% numerator_sum=sum(numerator);
% r1(s,1)=numerator_sum/denominator_sum;
end

% **************************************************
% CALCULATION OF r1+
% **************************************************
% r1_plus=r1+(1/n);

% **************************************************
% POLYNOMIAL MODELS FOR DIFFERENT SAMPLE SIZES
% **************************************************
if n==6
model = [(-0.1648)+r1.*-0.5643+(r1.^2).*-0.0916];
elseif n==10
model = [(-0.0972)+r1.*-0.3760+(r1.^2).*-0.0676];
elseif n==20
model = [(-0.0482)+r1.*-0.2028+(r1.^2).*-0.0333];
elseif n==30
model = [(-0.0373)+r1.*-0.1360];
end

% **************************************************
% EMPIRICAL BIAS IN r1, r1+ AND r1’
% **************************************************
r1_bias=mean(r1)-rho;
r1_plus_bias=mean(r1_plus)-rho;
r1_prime_bias=mean(r1_prime)-rho;

References


