

Nordic Informatics Network in Agricultural Sciences.  
 Computer Intensive Statistical Methods – with Biological Applications.  
 TUNE Landboskole

## Project B3 Estimation of bias in time series (Problem by John Öhrvik)

by

**Kirsten Bjørkestøl**

Norges Landbrukshøgskole and Høgskolen i Agder (email: Kirsten.Bjorkestol@hia.no)

A usual formula to calculate the (first) autocorrelation at lag 1 is

$$r(1) = \frac{\sum_{t=1}^{n-1} (x_t - \bar{x})(x_{t+1} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}$$

Here we have three AR(1)- processes, each of the form:

$$x_t - m = r(x_{t-1} - m) + e_t, \quad t = 1, \dots, n$$

where

$$\begin{aligned} E(e_t) &= 0 \quad \text{and} \quad E(e_t e_{t+k}) = \begin{cases} \sigma_e^2 & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases} \end{aligned}$$

$\rho$  is here the autocorrelation we want to estimate.

If we had used  $(n-1)/(n-2)r(1)$ , we would have (unbiased estimate of covariance) divided by (unbiased estimate of variance). Even in that situation we could not conclude with unbiased estimate of correlation. Generally dividing will not give unbiased results.

It is usual to use  $r(1)$  since we then have positive semidefinite correlation matrices.

Also in Minitab and S-plus they use  $r(1)$  as an estimate of first autocorrelation-coefficient,  $\rho$ . I have used the formula and the Minitab command `acf` to check that the results are the same.

In this project we shall try to estimate the bias of  $r(1)$ .

General: Let  $\theta$  be a parameter and  $\hat{q}$  an estimate.

Let  $\theta^*$  be bootstrap estimate calculated in the same way as  $\hat{q}$ .

Then **bootstrap assessment of bias** is

$$\text{bias} = \text{Mean of } (\theta^*) - \hat{q}$$

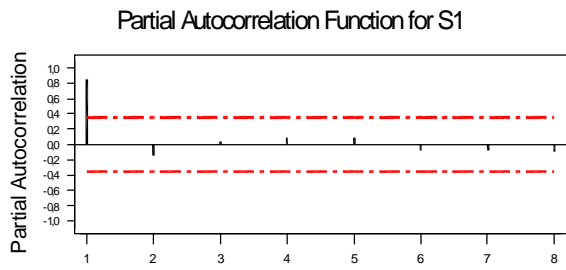
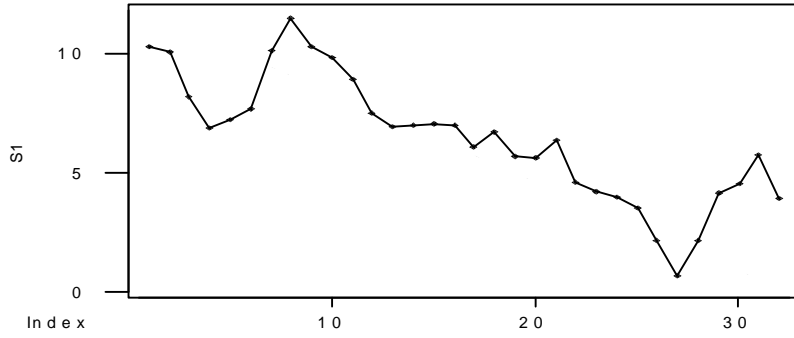
**Bias corrected estimate of  $\theta$**  is then

$$\bar{q} = \hat{q} - \text{bias} = 2 \cdot \hat{q} - \text{Mean of } (\theta^*)$$

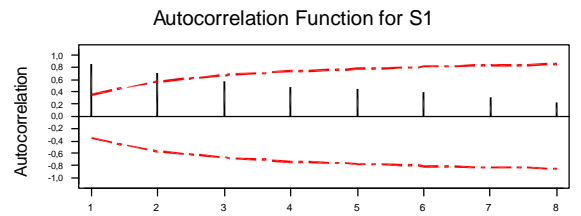
I will use 1000 replications for the bootstrap estimate.

First I will give some graphs to give pictures of the three series and their autocorrelations. For an AR(1) time serie the partial autocorrelation is 0 for lag greater than 1. The autocorrelation decrease exponential for increasing lag. These things seem "OK" here. In the graph of autocorrelation, we are here only interested in lag 1.

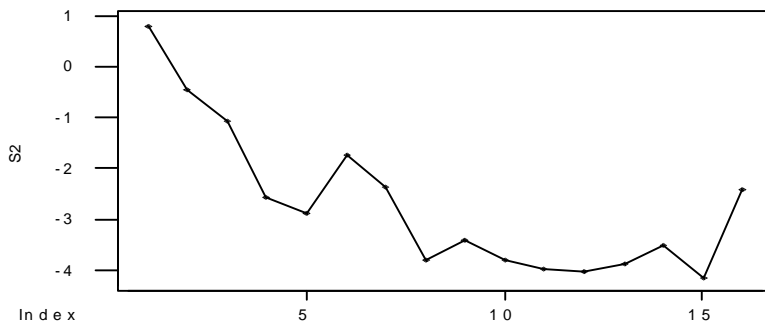
### Time series 1, 2 and 3



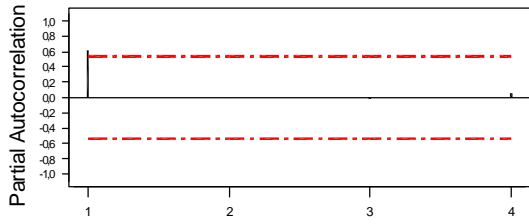
Lag	PAC	T	Lag	PAC	T
1	0,86	4,87	8	-0,10	-0,54
2	-0,15	-0,83			
3	0,03	0,17			
4	0,09	0,52			
5	0,08	0,48			
6	-0,08	-0,43			
7	-0,08	-0,43			



Lag	Corr	T	LBQ	Lag	Corr	T	LBQ
1	0,86	4,87	25,97	8	0,23	0,55	87,69
2	0,70	2,52	43,84				
3	0,57	1,74	56,19				
4	0,49	1,38	65,67				
5	0,45	1,19	73,96				
6	0,40	1,01	80,68				
7	0,33	0,80	85,29				

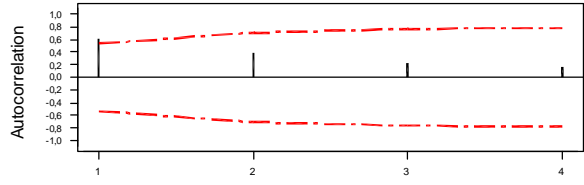


Partial Autocorrelation Function for S2

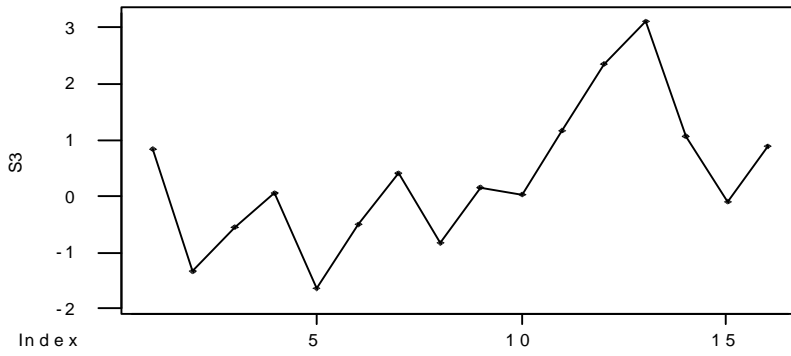


Lag	PAC	T
1	0,62	2,46
2	0,01	0,03
3	-0,02	-0,09
4	0,06	0,23

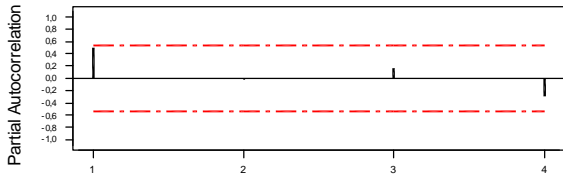
Autocorrelation Function for S2



Lag	Corr	T	LBQ
1	0,62	2,46	7,27
2	0,38	1,16	10,29
3	0,22	0,63	11,40
4	0,17	0,45	12,07

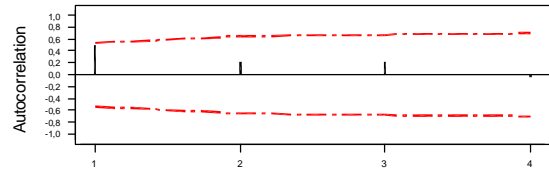


Partial Autocorrelation Function for S3



Lag	PAC	T
1	0,50	1,99
2	-0,02	-0,09
3	0,17	0,66
4	-0,30	-1,20

Autocorrelation Function for S3



Lag	Corr	T	LBQ
1	0,50	1,99	4,76
2	0,23	0,75	5,65
3	0,23	0,73	7,03
4	-0,04	-0,13	7,07

#Macro that find autocorrelation. The same formula as acf in Minitab

```
gmacro
  acorr1
  let k4=count(c9) -1
  let k5=mean(c9)
  do k6=1:k4
    let c10(k6)=(c9(k6)-k5)*(c9(k6+1)-k5)
  enddo
  let k7=sum(c10)
  let k8=k4*stdev(c9)**2
  let k8=k7/k8
  print k8
endmacro
```

The result is:

```
K1      0,860237      r(1) for S1
K2      0,615265      r(2) for S2
K3      0,497808      r(3) for S3
```

```
MTB > acf 1 c2 c6      give the result:      C6      0,860237
```

### Autocorrelation Function

ACF of S1

```
      -1,0 -0,8 -0,6 -0,4 -0,2  0,0  0,2  0,4  0,6  0,8  1,0
      +-----+-----+-----+-----+-----+-----+-----+-----+-----+
1      0,860                                     XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

We can also let Minitab to fit an AR(1) process with ARIMA command:

### Time serie 1:

```
MTB > arima 1 0 0 c5 c7 c8;
```

```
SUBC> const.
```

### ARIMA Model

ARIMA model for C5

Estimates at each iteration

Iteration	SSE	Parameters	
0	181,204	0,100	5,895
1	135,505	0,250	4,899
2	98,672	0,400	3,904
3	70,693	0,550	2,913
4	51,547	0,700	1,926
5	41,193	0,850	0,951
6	39,437	0,913	0,553
7	39,182	0,931	0,448
8	39,112	0,940	0,396
9	39,090	0,945	0,365
10	39,083	0,949	0,346
11	39,081	0,951	0,334
12	39,081	0,952	0,326
13	39,081	0,952	0,325

Relative change in each estimate less than 0,0010

Final Estimates of Parameters

Type	Coef	StDev	T	P
<b>AR 1</b>	<b>0,9521</b>	0,0756	12,59	0,000
Constant	0,3249	0,1994	1,63	0,114
Mean	6,780	4,160		

Number of observations: 32  
Residuals: SS = 38,0519 (backforecasts excluded)  
MS = 1,2684 DF = 30

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	12,6	36,3	*	*
DF	10	22	*	*
P-Value	0,248	0,028	*	*

So we can here see that the arima estimate of  $\rho$  is not the  $r(1)$ . I think that Minitab has done a correction since  $r(1)$  is not unbiased.

### Time serie 2:

```
MTB > ARIMA 1 0 0 'S2';
SUBC> Constant;
SUBC> Brief 2.
```

#### ARIMA Model

Final Estimates of Parameters

Type	Coef	StDev	T	P
<b>AR 1</b>	<b>0,9411</b>	0,1382	6,81	0,000
Constant	-0,0854	0,2692	-0,32	0,756
Mean	-1,450	4,575		

### Time serie 3

```
MTB > ARIMA 1 0 0 'S3';
SUBC> Constant;
SUBC> Brief 2.
```

#### ARIMA Model

Final Estimates of Parameters

Type	Coef	StDev	T	P
<b>AR 1</b>	<b>0,5121</b>	0,2316	2,21	0,044
Constant	0,1870	0,2817	0,66	0,518
Mean	0,3833	0,5774		

If the AR-coefficient is close to  $\pm 1$ , the series can be very unstable. If it is more than 1, we can “drive” away very much.

**Macro to estimate bias for autocorr. coeff. and sigma (with 95% conf.int)**

The thinking here is to fit an AR(1) process. To the fitted model I than add a sample of the residuals. I than get a new serie, called Serie\*. From this Serie\* I calculate  $r(1)^*$ . Mean of 1000 such  $r(1)^*$  is than an estimat of the expectation of  $r(1)$ .

```

gmacro
bscorr
# start time series in c5

let k1=count(c5)
let k2=1000
acf 1 c5 c6
let k3=c6(1) # r(1)
# arima 1 0 0 c5 calculate AR(1) from c5
# resid in c7 and fit in c8
arima 1 0 0 c5 c7 c8;
const;
brief 0.
do k4=1:k2
sample k1 c7 c9; # Resample residuals
replace.
let c10=c8+c9 # Bootsample time serie
acf 1 c10 c11 # Calculation of r(1)
let c12(k4)=c11(1)
let c13(k4)=stdev(c9) # Estimation of sd(residuals)
enddo
let k5=mean(c12) # b.s. estimate of E ( r(1))
let k6=k5 - k3 # estimated bias for r(1)
let k16=k3-k6 # bias corrected p estimate
print k3 k5 k6 k16
descr c12
let c13(k2+1)=stdev(c7) # estimated  $\sigma$ 
sort c13 c15
let k7=round((k2+1)*0.025 +0.49)
let k8=round((k2+1)*0.975 +0.49)
let k9=c15(k7) # 0.025 percentile for  $\sigma$ 
let k10=c15(k8) # 0.975 percentile for  $\sigma$ 
let k11=mean(c13) # bootstrap estim. for  $\sigma$ 
let k12=stdev(c13)
let k13=k11 - 1.96*k12
let k14=k11 + 1.96*k12
let k15=stdev(c7)
print k15 k11 k13 k14 k9 k10
endmacro

```

**Results from Time serie 1**

K3 0,860237 # **r(1) = est.  $\rho$**  from formula  
 K5 0,749999 # **bootstrap estimate of  $E(r(1))$**   
 K6 -0,110238 # **estimated bias for r(1)**  
 K16 0,970475 # **bias corrected  $\rho$**   
 31/30 \* r(1) = 0,8889 # (n-1)/(n-2) \* r(1); "unbiased / unbiased"

**Descriptive Statistics for bootstrap acf**

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C12	1000	0,75000	0,75507	0,75152	0,05076	0,00161

Variable	Minimum	Maximum	Q1	Q3
C12	0,54781	0,87827	0,71765	0,78645

K15 1,08926 # s from residual data  
 K11 1,05729 # **s from bootstrap s**  
 K13 0,809560  
 K14 1,30501 # s(bs)  $\pm$  1.96 sd s(bs) = [0.810 , 1.305]  
 K9 0,797481  
 K10 1,30274 # 95% percentile for s = [0.797 , 1.303]

**Results from Time serie 2**

K3 0,615265 # **r(1) = est.  $\rho$**  from formula  
 K5 0,575586 # **bootstrap estimate of  $E(r(1))$**   
 K6 -0,0396793 # **estimated bias for r(1)**  
 K16 0,6549443 # **bias corrected estimate of  $\rho$**   
 15 / 14 \* r(1) = 0,6592 # "unbiased / unbiased"

**Descriptive Statistics for bootstrap acf**

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C12	1000	0,57559	0,58406	0,57993	0,10404	0,00329

Variable	Minimum	Maximum	Q1	Q3
C12	0,00301	0,81507	0,51599	0,64575

K15 0,830607 # s from residual data  
 K11 0,784226 # s from bootstrap s  
 K13 0,486172  
 K14 1,08228 # s(bs)  $\pm$  1.96 sd s(bs) = [0.486 , 1.082]  
 K9 0,492297  
 K10 1,05742 # 95% percentile for s = [0.492 , 1.057]

### Results from Time serie 3

K3 0,497808 # **r(1) = est.  $\rho$**  from formula  
 K5 0,0925940 # **bootstrap estimate of  $E(r(1))$**   
 K6 -0,405214 # **estimated bias for r(1)**  
 K16 0,903022 # **Bias corrected estimate of  $\rho$**   
 15 / 14 \* r(1) = 0,5334

### Descriptive Statistics for bootstrap acf

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C12	1000	0,09259	0,10576	0,09466	0,22937	0,00725

Variable	Minimum	Maximum	Q1	Q3
C12	-0,57057	0,67310	-0,05918	0,25300

K15 1,08598 # **s from residual data**  
 K11 1,04013 # **s from bootstrap s**  
 K13 0,712371  
 K14 1,36789 # **s(bs)  $\pm$  1.96 sd s(bs) = [0.712 , 1.368]**  
 K9 0,692149  
 K10 1,35583 # **95% percentile for s = [0.692 , 1.356]**

The results for the r(1) and the bias seems very strange here. It is very large difference between r(1) and the bias corrected estimator. I did the simulations ones more, and I got similar results:

K3 0,497808 # r(1)  
 K5 0,0820506 # bootstrap estimate of  $E(r(1))$   
 K6 -0,415758 # bias  
 K16 0,913566 # bias corrected estimate of  $\rho$   
 K15 1,08598  
 K11 1,03764  
 K13 0,715246  
 K14 1,36002  
 K9 0,690438  
 K10 1,32512

This time serie seems special in the graph. In the first part the autocorrelation seems to be negative, while in the last part it seems positive. Maybe this is the reason for the large difference.

Efron and Tibshirani say in "An Introduction to the Bootstrap" (p. 138): "...*Bias correction can be dangerous in practice. Even if  $\bar{Q}$  ("the bias-corrected estimator") is less bias than  $\hat{Q}$ , it may have substantially greater standard error. . . . . the exact use of a bias estimate is often problematic. Biases is harder to estimate than standard error, . . . . . Correcting the bias may cause a larger increase in standard error, which in turn results in a larger mean squared error. . . . If bias is small compared to the estimated standard error se, than it is safer to use  $\hat{Q}$  than bias corrected estimate. If bias is large compared to se, then it may be an indication that the statistic  $\hat{Q} = s(\mathbf{x})$  is not an appropriate estimate of the parameter  $q$* ". In Time serie 3 se(mean bs corr) = 0,0016 while bias is -0,405. I Time serie 2 the values are 0,0033 and -0,040.

### Generalisations

We can use the same method to correct autocorrelations also for lag larger than 1. I think these will be "more unbiased" than lag 1. To get "unbiased / unbiased" for lag k we must use  $(n-1) / (n - k - 1) * r(k)$ , if we calculate r(k) in a similar way.