ESTIMATING THE VARIANCES OF SAMPLE CORRELATIONS

Alice Orcutt Nakamura and Masao Nakamura, University of Alberta

I. INTRODUCTION

Let

$$p(r|r_1,r_2,\rho_1,\rho_2) = \frac{p(r,r_1,r_2|\rho_1,\rho_2)}{p(r_1,r_2|\rho_1,\rho_2)} \quad (1)$$

denote the distribution of the sample product moment correlation coefficient, r, conditional on the first-lag sample autocorrelations, r_1 and r_2, and the first-lag population autocorrelations ρ_1 and ρ_2, where the population cross-correlation is $\rho = 0$. The conventional tests of significance for r rely on

$$p(r|\rho_1,\rho_2) = \iint p(r,r_1,r_2|\rho_1,\rho_2)dr_1dr_2, \quad (2)$$

Moreover Bartlett [1] and McGregor [5,6] have established that

$$p(r|\rho_1,\rho_2) \cong p(r|\rho_1\ \rho_2). \quad (3)$$

and that

$$V(r|\rho_1,\rho_2) \cong V(r|\rho_1\ \rho_2) \quad (4)$$

$$\cong \frac{1+\rho_1\rho_2}{n(1-\rho_1\rho_2)} - \frac{2\rho_1\rho_2(1-\rho_1{}^n\rho_2{}^n)}{n^2(1-\rho_1\rho_2)^2} \quad (5)$$

$$\cong \frac{1+\rho_1\rho_2}{n(1-\rho_1\rho_2)} - \frac{2\ \rho_1\rho_2}{n^2(1-\rho_1\rho_2)^2} \quad (6)$$

where n denotes the number of items in each of the series correlated.[1]

In a recent paper, Nakamura, Nakamura and Orcutt [7] note that (1) is more informative than (2), and argue that tests of significance for r should be based on (1) rather than (2). Moreover Monte Carlo evidence is presented that

$$V(r|r_1,r_2,\rho_1,\rho_2) = V(r|r_1,r_2). \quad (7)$$

Lacking Monte Carlo tabulations of (7) for small samples, a researcher can attempt to estimate (7) by substituting r_1 and r_2 for ρ_1 and ρ_2 in equation (6) as suggested by Orcutt and James [8], he can use (6) directly, or he can use

$$V(r|\rho_1\rho_2 = 0) \cong \frac{1}{n-3}{}^2. \quad (8)$$

As $n \to \infty$ clearly $V(r|r_1,r_2,\rho_1,\rho_2) = V(r|r_1,r_2)$

$$\to V(r|\rho_1,\rho_2).$$

In this paper, sampling methods are used to compare the percentage errors made in estimating (7) for series of length 30 using these three approaches.

II. METHODOLOGY

Our generating relationships were of the form

$$X_t = \rho_1X_{t-1} + u_t, \quad (9)$$

and

$$Y_t = \rho_2Y_{t-1} + v_t, \quad (10)$$

where u and v were generated by two Chen random normal number generators.[3] 1,000 pairs of series of length 30 were generated and saved for each of the following pairs of values of ρ_1 and ρ_2: (-.9,.9),(-.7,.7),(-.3,.3),(-.3,-.3),(.3,.3), (-.7,-.7),(.7,.7),(-.9,-.9), and (.9,.9).[4]

For each series the autoregressive parameter was estimated using least squares regression.[5] Also we calculated the Pearson product-moment correlation coefficient for each pair of series. Each set of 1,000 sample correlation coefficients was then classified according to the values of the products of the sample autocorrelation coefficients, r_1r_2, of the pairs of series correlated. Intervals of 0.1 were used. Finally the observed variance of the sample correlations was calculated for each cell for each of our 9 sets of 1,000 correlations.

For each cell in our product classification for each of our 9 sets of correlations we next approximated the variance of the sample correlations in that cell using the modified version of formula (6):

$$\text{var } r \cong \frac{1+\overline{r_1r_2}}{30(1-\overline{r_1r_2})} - \frac{2\ (\overline{r_1r_2})}{900(1-\overline{r_1r_2})}{}^2 \quad (11)$$

where $\overline{r_1r_2}$ stands for the cell mean of the products of the sample autoregressive coefficients. Secondly we estimated the cell variances using formula (6) with n = 30. As a third alternative, we estimated the cell variances using formula (8). We will call the estimates obtained for each cell using formulas (11), (6) and (8), estimates 1,2 and 3 respectively.[7]

We now calculated the percentage errors made in approximating the observed cell variances of our sample correlations using each of these three estimation methods. The formula used to obtain these percentage errors was

$$\%\text{ error } i = \frac{(\text{estimate } i)-(\text{observed cell variance})}{(\text{observed cell variance})} \quad (12)$$

$$i = 1,\ 2,\ 3.$$

The percentage errors are shown in Table 1, where the top number in each cell corresponds to the percentage error made using estimate 1, the next number to the percentage error made using estimate 2, and the third number to the percentage error made using estimate 3 for that cell. The cell frequencies - that is, the number of correlations in each cell - are shown in Table 2.

III. FINDINGS

Estimation method 1 results in smaller percentage errors in estimating our observed cell variances than either estimation methods 2 or 3 for 69% of our cells, and smaller percentage errors than method 2 for 80% of our cells. Looking only at those cells where the frequency, or number of correlations, is at least 30, and hence where the observed sample variances of the correlations in each cell can be regarded as a reasonable estimate of the population conditional variance for that cell, we see that method 1 results in smaller percentage errors than either methods 2 or 3 for 84% of these 49 cells.

Thus method 1 is seen to be a more efficient method of estimating $V(r|r_1,r_2)$ than either methods 2 or 3,[8] and is more operational than method 2 which requires knowledge of the population autoregressive parameters. Further experiments using (-.9,0),(-.7,0),(-.3,0),(0,0),(.3,0),(.7,0),(.9,0) for the values of ρ_1 and ρ_2 indicate that this result holds even when $\rho_1\rho_2 = 0$.

FOOTNOTES

1. This formalization of our problem was suggested to us by Professor Arthur S. Goldberger.
2. See Fisher [4], p. 191.
3. See Chen [2,3]. The initial values used for the starting integers were 748511649 and 147303541 for the u series and 180810529 and 536841077 for the v series. Satisfactory statistical properties are reported for random numbers generated using these initial numbers in Chen [3]. For both series the mean was 0 and the standard deviation was 25. We set $X_0 = Y_0 = 0$. The computer used was the IBM System/360 model 67 at the University of Alberta Computing Center.
4. To minimize the effect of the initial values used in generating u and v the first pair of series of length 30 generated for each pair of values of ρ_1 and ρ_2 was discarded. Also every other one of the subsequent pairs of series of length 30 generated was discarded.
5. Since in practice one would have no way of knowing the true value of the constant term, we estimated a constant term along with the autoregressive parameter.
6. Since
$$E(r|r_1r_2) = E(r) = 0$$
where r denotes the sample correlation coefficient [7],
$$\overline{r^2} = \frac{\Sigma r^2}{n} = \frac{\Sigma[r-E(r)]^2}{n}$$
is an unbiased estimate of the variance of r. This is the formula which we used in computing the cell variances.
7. In our abstract these three estimation methods are referred to in reverse order.
8. Stuart [9] presents a theoretical argument showing that given an estimator u of a parameter θ in a multiparameter distribution, one does not necessarily improve its efficiency by substituting true parameter values into u to replace estimators of them. For a discussion of estimating efficiency and the power of tests see Sundrum [10].

REFERENCES

[1] Bartlett, M.S., "Some Aspects of the Time Correlation Problem in Regard to Tests of Significance," Journal of the Royal Statistical Society, 98 (1935), 536-43.

[2] Chen, E.H., "A Random Normal Number Generator for 32-Bit-Word Computers," Journal of the American Statistical Association, 66 (1971), 400-3.

[3] ____________, "Supplement to 'A Random Normal Number Generator for 32-Bit-Word Computers,'" Health Sciences Computing Facility, University of California, Los Angeles, California, 1971.

[4] Fisher, R.A., Statistical Methods for Research Workers, twelfth edition, Edinburgh: Oliver and Boyd, 1954.

[5] McGregor, J.R., "The Approximate Distribution of the Correlation between Two Stationary Linear Markov Series," Biometrika, 49, Nos. 3 and 4 (1962), 379-88.

[6] ____________, and U.M. Bielenstein, "The Approximate Distribution of the Correlation between Two Stationary Linear Markov Series II," Biometrika, 52 (1965), 301-2.

[7] Nakamura, A.O., M. Nakamura, and G.H. Orcutt, "Testing for Relationships between Time Series," an unpublished paper.

[8] Orcutt, G.H., and S.T. James, "Testing the Significance of Correlation between Time Series," Biometrika, 35, Nos. 3 and 4 (1948), 397-413.

[9] Stuart, A., "A Paradox in Statistical Estimation," Biometrika, 42 (1955), 527-9.

[10] Sundrum, R.M., "On the Relation between Estimating Efficiency and the Power of Tests," Biometrika, 41 (1954) pp. 542-4.

Class intervals for $r_1 r_2$	Values of ρ_1 and ρ_2								
	(-.9,.9)	(-.7,.7)	(-.3,.3)	(-.3,-.3)	(.3,.3)	(-.7,-.7)	(.7,.7)	(-.9,-.9)	(.9,.9)
-1.1 to -1.0	-114 7 875								
-1.0 to -.9	-36 44 1221								
-.9 to -.8	15 35 1134	444 1638 5309							
-.8 to -.7	-6 -29 552	83 278 1076							
-.7 to -.6	6 -43 419	19 81 462							
-.6 to -.5	-3 -61 259	7 22 281							
-.5 to -.4	1 -68 190	-4 -14 169	22 140 217						
-.4 to -.3	-12 -78 100	2 -26 131	2 64 117						
-.3 to -.2	4 -78 97	-2 -42 80	2 37 82						
-.2 to -.1	-9 -85 39	13 -45 71	-6 3 36	25 91 78	-14 33 24				
-.1 to 0	298 -45 401	-18 -66 5	-4 -12 16	-3 21 13	-4 20 12		18 252 40		
0 to .1		292 35 320	-3 -22 3	0 7 0	-14 -5 -12	-40 45 42	-10 119 -13		-29 405 -30
.1 to .2			664 390 548	2 -8 -14	12 2 -5	-12 81 -28	-13 80 -28	9 570 -7	-32 292 -46
.2 to .3				-17 -39 -43	13 -14 -20	31 120 -13	-4 62 -36	-15 293 -46	-22 282 -47

1. PERCENTAGE ERRORS (cont.)

Class intervals for $r_1 r_2$	Values of ρ_1 and ρ_2								
	(-.9,.9)	(-.7,.7)	(-.3,.3)	(-.3,-.3)	(.3,.3)	(-.7,-.7)	(.7,.7)	(-.9,-.9)	(.9,.9)
.3 to .4				-5 -41 -45	220 81 69	10 51 -40	-9 27 -50	-8 262 -50	32 422 -28
.4 to .5				10 -47 -50		7 19 -53	-14 -6 -63	38 327 -41	-9 184 -61
.5 to .6						3 -11 -64	-10 -22 -69	5 155 -65	-12 115 -70
.6 to .7						4 -31 -73	60 5 -58	24 123 -69	-15 57 -78
.7 to .8						64 -19 -68	51 -25 -70	10 40 -80	-9 20 -83
.8 to .9						361 68 -33		16 -2 -86	-4 -15 -88
.9 to 1.0								43 -22 -89	55 -16 -88
1.0 to 1.1									

2. CELL FREQUENCIES

Class intervals for $r_1 r_2$	Values of $\rho_1 \rho_2$								
	(-.9,.9)	(-.7,.7)	(-.3,.3)	(-.3,-.3)	(.3,.3)	(-.7,-.7)	(.7,.7)	(-.9,-.9)	(.9,.9)
-1.1 to -1.0	1								
-1.0 to -.9	14								
-.9 to -.8	118	1							
-.8 to -.7	251	9							
-.7 to -.6	280	64							
-.6 to -.5	177	139							
-.5 to -.4	94	279	1						
-.4 to -.3	45	270	9						
-.3 to -.2	12	154	55						
-.2 to -.1	6	66	246	2	4				
-.1 to 0	2	16	544	99	186		4		
0 to .1		2	142	485	575	5	35		3
.1 to .2			3	307	199	36	117	1	13
.2 to .3				93	34	108	212	7	40
.3 to .4				13	2	205	288	21	84
.4 to .5				1		289	214	33	153
.5 to .6						226	107	91	249
.6 to .7						115	20	214	234
.7 to .8						15	3	308	173
.8 to .9						1		257	47
.9 to 1.0								68	4
1.0 to 1.1									