MISCELLANEA

CONTENTS

A LARGE-SAMPLE TEST FOR THE GOODNESS OF FIT OF AUTOREGRESSIVE SCHEMES

By M. H. QUENOUILLE

(Rothamsted Experimental Station)

BARTLETT (1946) has recently given formula for the variances and covariances of observed auto-correlation coefficients in terms of their theoretical values. These formula, which are independent of the underlying distribution error, have been used by Bartlett in their limiting forms to obtain a rough goodness-of-fit test for autoregressive schemes. The purpose of this paper is to demonstrate a more precise test for autoregressive schemes in large samples.

Bartlett's main formula for the covariance of two observed auto-correlations r_s and r_{s+t} in terms of the theoretical auto-correlations may be written—

$$\sigma^4 \operatorname{cov}(r_s, r_{s+t}) \sim \frac{1}{n} \sum_{\nu=-\infty}^{\infty} (\rho_\nu \rho_{\nu-t} + \rho_{\nu-s-t} \rho_{\nu+s} + 2\rho_s \rho_{s+t} \rho^2_\nu - 2\rho_s \rho_\nu \rho_{\nu-s-t} - 2\rho_{s+t} \rho_\nu \rho_{\nu-s}) \quad . \quad . \quad . \quad . \quad . \quad (1)$$

If we put $X_t = \sum_{\nu=-\infty}^{\infty} \rho_\nu \rho_{\nu-t}$ and

$$(1 + a_1 x + \quad . \quad . \quad . \quad + a_k x^k)^2 = \sum_{i=-\infty}^{\infty} A_i x^i$$

then by methods previously given (Quenouille, 1947*a*) it may be shown that

$$\left.\begin{aligned} &\sum_{i=-\infty}^{\infty} A_i X_{t-i} = 0 \qquad t > 0 \\ \text{and} \quad &\sum_{i=-\infty}^{\infty} A_i \rho_{t-i} = 0 \qquad t > k \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Using the formula (2) if we define $R_s = \sum_{i=-\infty}^{\infty} A_i r_{s-i}, \quad s = k+1, k+2, . \ . \ . \ .$

then it is not difficult to see that

$$\begin{aligned} cov\,(r_s, R_t) &\sim 0 \qquad\qquad t > s \\ &\sim \frac{1}{n\sigma^4} \sum_{i=-\infty}^{\infty} A_i X_i, \quad t = s \end{aligned}$$

from which we get the following important result:

For n large, the forms R_s are independently and normally distributed about zero with variances $\sum_{=-\infty}^{\infty} A_i X_i / n\sigma^4$. Furthermore the forms $r_s - \rho_s$ $s = 1, 2, \ . \ . \ . \ k$ are jointly distributed independent of the forms R_s. Thus, if we use the equations $\rho_s = r_s$ to fit an autoregressive scheme, we can use the forms R_s to test the adequacy of the fit. For example, for the Markoff scheme $u_{n+1} = \rho u_n + \varepsilon_{n+1}$, we have $\rho_s = \rho_s X_t = \rho^t \frac{1+\rho^2}{1-\rho^2} + t\rho^t$ so that the forms $r_1 - \rho$, $R_s = r_s - 2\rho r_{s-1} + \rho^2 r_{s-2}$, $s = 2, 3, \ . \ . \ .$ are independently and normally distributed with variances $(1-\rho^2)/n$ and $(1-\rho^2)^2/n$. To test whether any scheme deviates significantly from a Markoff scheme, we might set $\rho = r_1$, and test the remaining degrees of freedom given by the forms R_s to magnitude, homogeneity, etc. Alternatively we can use the forms $r_1 - \rho_1$, and R_s ($s = 2 \ldots q+1$) to obtain an optimum fit, and test the remaining q degrees of freedom together with the degrees of freedom given by R_s ($s = q + 2 \ldots$).

It is not difficult to see the forms R_s are related to the correlations r_i^1 of the ε_j, since

$$\begin{aligned} R_{k+j} &= \sum_i A_i\, r_{k+j-i} \\ &= \frac{\sum_i A_i \sum_l u_l\, u_{k+j+l-i}}{\Sigma u^2} \\ &\sim \frac{1}{\Sigma u^2} \Sigma\{u_l\, (\sum_{i=0}^{k} a_i\, \varepsilon_{k+l+j-i})\} \\ &\sim \frac{\Sigma \varepsilon^2}{\Sigma u^2} \sum_{l=0}^{\infty} B_l\, r^1_{j-l} \end{aligned} \qquad (3)$$

where

$$\frac{a_k + a_{k-1}t + \ldots + a_0 t^k}{a_0 + a_1 t + \ldots + a_k t^k} = \sum_{i=0}^{\infty} B_i t^i$$

From which, we deduce directly,

$$\sum_{i=0}^{\infty} B_i B_{i+l} \begin{array}{ll} = 0 & l \neq 0 \\ = 1 & l = 0 \end{array} \qquad (4)$$

Thus it appears that the R_s are moving averages of infinite extent of the r_i^1, so that each will provide a wider, as well as a simpler, test than the r_i^1.

From the formulae (3) and (4), it can be seen that

$$\text{var } R_s \sim \frac{1}{n}\left(\frac{\text{var } \varepsilon}{\text{var } u}\right)^2, \qquad s > 0 \qquad (5)$$

A similar application confirms that R_s and R_t are uncorrelated. Thus, if we wish to test the legitimacy of assuming an autoregressive scheme $u_{n+2} + au_{n+1} + bu_n = \varepsilon_{n+2}$, we can use the estimates

$$a = -\frac{r_1(1-r_2)}{1-r_1^2},\ b = \frac{1-r_2}{1-r_1^2} - 1 \text{ and test that}$$

$$r_{s+2} + 2a\, r_{s+1} + (a^2 + 2b) r_s + 2ab r_{s-1} + b^2 r_{s-2} = R_s^1,\ s = 1, 2, \ldots$$

are distributed with mean zero, and variance $\frac{1}{n}\left[\frac{(1-b)\{(1+b)^2 - a^2\}}{1+b}\right]^2$. This approximation can be improved still further by the use of $n-s$ instead of n.

To illustrate the method we shall consider firstly artificial series, for which a and b are known, and secondly series experienced in practice, for which a and b have been calculated. The adequacy of the usual methods of fitting an autoregressive scheme and its suitability for the representation of practical results will not be discussed here, but a discussion of these problems will be given elsewhere (Quenouille and Orcutt, 1947*b*).

Kendall (1946) has calculated auto-correlations of

(i) 480 terms of $u_{n+2} - 1\cdot 1\, u_{n+1} + 0\cdot 5\, u_n = \varepsilon_{n+2}$
(ii) 240 terms of $u_{n+2} - 1\cdot 2\, u_{n+1} + 0\cdot 4\, u_n = \varepsilon_{n+2}$
(iii) 240 terms of $u_{n+2} - 1\cdot 1\, u_{n+1} + 0\cdot 8\, u_n = \varepsilon_{n+2}$
(iv) 240 terms of $u_{n+2} + 1\cdot 0\, u_{n+1} + 0\cdot 5\, u_n = \varepsilon_{n+2}$

where the ε_n is a randomly-chosen number between -49 and $+49$. The values of R_s^1 together with the corresponding values of $\chi^2_{(1)}$ and $\Sigma\chi^2_{(1)}$ for each of the four series are given in Table 1.

It is seen that the values of $\Sigma\chi^2_{(1)}$ behave according to theory, the total for 90 degrees of freedom being 85·868. The values of $\chi^2_{(1)}$ can be compared with its distribution. This is done in Table 2. The expected number in each class is nine, so that the observed results again conform to theory.

TABLE 1.—*Values of* R_s^1, $\chi^2_{(1)}$, *and* $\Sigma\chi^2_{(1)}$ *for Kendall's Artificial Series.*

Series i

s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$	s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$
1	·02412	2·319	2·319	16	—·00341	0·045	13·363
2	·00417	0·068	2·387	17	—·00305	0·036	13·399
3	·01979	1·558	3·945	18	·00687	0·181	13·580
4	—·00812	0·261	4·206	19	·01558	0·931	14·511
5	—·00013	0·000	4·206	20	—·00955	0·349	14·860
6	·02026	1·619	5·825	21	—·01610	0·990	15·850
7	—·02247	1·987	7·812	22	·00604	0·139	15·989
8	—·02903	3·310	11·122	23	—·01464	0·815	16·804
9	—·01047	0·430	11·552	24	—·02046	1·589	18·393
10	—·01226	0·588	12·140	25	—·00996	0·376	18·769
11	·00747	0·218	12·358	26	—·02064	1·610	20·379
12	·01345	0·704	13·062	27	—·00083	0·003	20·382
13	—·00573	0·128	13·190	28	·00079	0·002	20·384
14	—·00517	0·104	13·294	29	·00224	0·019	20·403
15	·00249	0·024	13·318	30	·00895	0·300	20·703

Series ii

s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$	s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$
1	—·02440	2·865	2·865	11	—·00500	0·115	14·012
2	—·01156	0·640	3·505	12	—·03040	4·243	18·255
3	—·01056	0·532	4·037	13	·00912	0·380	18·635
4	·02704	3·474	7·511	14	—·02896	3·816	22·451
5	—·02398	2·721	10·232	15	·00792	0·284	22·735
6	—·00040	0·001	10·233	16	·00624	0·176	22·911
7	·01992	1·862	12·095	17	·00068	0·002	22·913
8	—·01924	1·730	13·825	18	·01600	1·145	24·058
9	·00320	0·048	13·873	19	·01132	0·570	24·628
10	·00228	0·024	13·897	20	·01204	0·642	25·270

Series iii

s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$	s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$
1	·01133	0·605	0·605	11	·00247	0·027	13·919
2	·00773	0·280	0·885	12	·03281	4·824	18·743
3	·00243	0·028	0·913	13	—·01769	1·396	20·139
4	—·01192	0·659	1·572	14	·02389	2·535	22·674
5	—·01941	1·740	3·312	15	·01053	0·490	23·164
6	·03032	4·228	7·540	16	·00415	0·076	23·240
7	·00178	0·015	7·555	17	·00347	0·053	23·293
8	—·02385	2·594	10·149	18	·02687	3·151	26·444
9	·00885	0·356	10·505	19	—·00180	0·014	26·458
10	·02737	3·387	13·892	20	·01367	0·808	27·266

Series iv

s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$	s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$
1	·02900	1·158	1·158	11	·01900	0·476	9·362
2	·02000	0·548	1·706	12	—·03725	1·822	11·184
3	·00000	0·000	1·706	13	—·01225	0·196	11·380
4	—·04150	2·341	4·047	14	·00200	0·005	11·385
5	·00000	0·000	4·047	15	—·02350	0·716	12·101
6	—·05050	3·437	7·484	16	·00550	0·039	12·140
7	·03125	1·311	8·795	17	—·00575	0·043	12·183
8	·00025	0·000	8·795	18	—·01025	0·134	12·317
9	·00725	0·070	8·865	19	·01375	0·241	12·558
10	·00400	0·021	8·886	20	·00750	0·071	12·629

TABLE 2.—*Observed Distribution of* $\chi^2_{(1)}$

P	*Number observed.*
0·0–0·1	11
0·1–0·2	12
0·2–0·3	9
0·3–0·4	6
0·4–0·5	8
0·5–0·6	11
0·6–0·7	5
0·7–0·8	8
0·8–0·9	9
0·9–1·0	11
	90

Before testing series which have been observed in practice, we shall be faced with the problem of superposed variation. If, in fact, a superposed variation, η_i, exists, then we shall observe a quantity $u_i + \eta_i$, and

$$\text{var } R_s \sim \frac{1}{n}\left(\frac{\text{var } \varepsilon}{\text{var } u + \text{var } \eta}\right)^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

Thus significant differences between proposed and existing autoregressive schemes will tend to be obscured by superposed variation. This difficulty can be overcome if we observe from equation (3) that such differences will, in general, yield significant values of R_s only for s small, while for s large, R_s will have a variance given by (6). It will therefore be possible to list the values of R_s for s large to indicate the presence of superposed variation, while the adequacy of the fitted autoregressive scheme can be tested by a comparison of the values of R_s for s small and for s large. In carrying out this latter test, we must remember that our choice of what constitutes a large or small value of s will necessitate using a higher level of significance than usual. Thus the choice of the first eight values of series i and the first seven values of series iv to be tested against the other values given in Table 1, indicates a significant superposed variation and a significant deviation in the autoregressive equation at the 5% level. These spurious effects arise from a choice of a low level of significance, in fact, the next ten values of $\chi^2_{(1)}$ for series i total 7·2038, while the corresponding total for the next five values of series iv is 8·5638.

If we decide that a superposed variation is present, then we will require fresh estimates of a and b together with an estimate of $\frac{\text{var } \eta}{\text{var } u}$. These may be obtained in several ways : we may minimize the χ^2 obtained from R_s, $s = 1 \ . \ . \ . \ q$, or we may use the equations $\rho_s = r_s$. This latter method will be equivalent to choosing $q = k + 1$. Thus, for example, for the linear autoregressive scheme $u_{n+2} + au_{n+1} + bu_n = \varepsilon_{n+2}$, if we put $\frac{\text{var } \eta}{\text{var } u} = c$, we have

$$-\frac{a}{1+b} = r_1(1+c)$$

$$\frac{a^2}{1+b} - b = r_2(1+c)$$

$$-\frac{a^3}{1+b} + ab + \frac{ab}{1+b} = r_3(1+c).$$

The solution of these equations gives

$$a^2\frac{r_1^2 - r_2^2}{r_1 - r_3} + 2ar_2 + r_3 = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

$$b = -\frac{ar_2 + r_3}{r_1} \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

$$c = -\frac{a}{r_1 - r_3 - ar_2} - 1 \quad . \quad . \quad (9)$$

where a is chosen from equation (7) so that c will be positive. This method of estimating a, b and c will not be the most efficient, and other methods are possible. Thus, for example, we might estimate a and b from

$$r_3 + ar_2 + br_1 = 0$$
$$r_4 + ar_3 + br_2 = 0$$

and these estimates will be independent of superposed variation.

We shall then have

$$\text{var } R_s \sim \frac{1}{n\,(1+c)^2}\left[\frac{(1-b)\,\{(1+b)^2-a^2\}}{1+b}\right]^2 \quad . \quad . \quad . \quad (10)$$

provided that we replace $r_o = 1$ by $(1+c)^{-1}$ in R_s.

If we had treated this scheme as if no superposed error were present, we should have estimated

$$\text{var } R_s \sim \frac{1}{n(1+c)^2}\left[\frac{(1-b)\,\{(1+b)^2-a^2\}}{1+b}\right.$$
$$\left.+\,c\left((1-b)+\frac{(1+b)\,(1+c+2b+bc)\,\{(1+b+c)\,(1+b)-a^2\}}{(1+b)^2\,(1+c)^2-a^2}\right)\right]^2 \quad . \quad . \quad (11)$$
$$\sim \frac{(1+\Delta)^2}{n\,(1+c)^2}\left[\frac{(1-b)\,\{(1+b)^2-a^2\}}{1+b}\right] \qquad \text{for } c \text{ small}$$

where $\Delta = \dfrac{2\,(1+b)\,(1-b^3)}{(1-b)^2\,\{(1+b)^2-a^2\}}$

Therefore the detection of any superposed variation will depend upon the value of $(1+\Delta)^2$. Thus, for example, in Kendall's series (i), we have

$(1+\Delta)^2 = (1+10\cdot 1c)^2 = \left(1+29\cdot 1\,\dfrac{\text{var }\eta}{\text{var }\varepsilon}\right)^2$, so that we should be able to detect easily a superposed variation of 5%.

If we multiply each of the coefficients of Kendall's series 1 by 0·9 this will roughly correspond to a superposed variation of 4%. We can fit an autoregressive scheme using the first two coefficients, giving $a = -0\cdot 8554$, $b = 0\cdot 2662$. The values of $R_s{}^1$, $\chi^2{}_{(1)}$ and $\Sigma\chi^2{}_{(1)}$ are given in Table 3.

TABLE 3.—*Values of $R_s{}^1$, $\chi^2{}_{(1)}$ and $\Sigma\chi^2{}_{(1)}$ for Kendall's Series i with Theoretical Damping*

s	$R_s{}^1$	$\chi^2{}_{(1)}$	$\Sigma\chi^2{}_{(1)}$	s	$R_s{}^1$	$\chi^2{}_{(1)}$	$\Sigma\chi^2{}_{(1)}$
1	−·04923	4·550	4·550	11	·00678	0·084	10·150
2	·00556	0·058	4·608	12	·01372	0·345	10·495
3	·01694	0·536	5·144	13	·01340	0·329	10·824
4	·00045	0·000	5·144	14	·00150	0·004	10·828
5	−·00249	0·011	5·155	15	−·00160	0·005	10·833
6	·01390	0·359	5·514	16	−·00799	0·116	10·949
7	·01463	0·397	5·911	17	−·00805	0·117	11·066
8	−·03471	2·229	8·140	18	·00245	0·011	11·077
9	−·02625	1·272	9·412	19	·01706	0·526	11·603
10	−·01884	0·654	10·066	20	·00127	0·003	11·606

If the first ten degrees of freedom from this table are tested against the second ten, the result is significant at the 1 per cent. level, while the latter ten degrees of freedom are 99 per cent. significant. Thus it appears that the estimates of a and b are biased towards zero, probably as a result of superposed variation. This conclusion is strengthened by the large value of $R_1{}^1$, due to the fact that $r_o = 1$ cannot be altered by superposed variation. For practical series, we use Beveridge's (1921) series of wheat-price index, and Kendall's (1943) series of wheat, barley and oats prices. Autoregressive schemes have been fitted to these series using the first two autocorrelations, and the fitted schemes are tested in Table 4.

TABLE 4.—*Values of R_s^1, $\chi^2_{(1)}$ and $\Sigma\chi^2_{(1)}$ for Practical Series*

Beveridge's wheat-price index

s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$	s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$
1	·02395	0·554	0·554	11	·02011	0·380	20·279*
2	—·02230	0·479	1·033	12	·04181	1·639	21·918*
3	—·02662	0·681	1·714	13	·02435	0·554	22·472*
4	—·07800	5·831	7·545	14	·06797	4·307	26·779*
5	—·07122	4·848	12·393*	15	—·02899	0·781	27·560*
6	—·07984	6·076	18·469†	16	—·01629	0·246	27·806*
7	·00118	0·001	18·470*	17	—·01360	0·171	27·977*
8	—·00760	0·055	18·525*	18	·01619	0·242	28·219
9	—·02433	0·560	19·085*	19	·00738	0·050	28·269
10	—·02939	0·814	19·899*	20	·06872	4·328	33·597*

Kendall's wheat prices

s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$	s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$
1	—·02173	0·110	0·110	11	—·09806	1·884	15·333
2	—·14697	4·936	5·046	12	·00078	0·000	15·333
3	—·04400	0·435	5·481	13	—·15599	4·590	19·923
4	—·12796	3·623	9·104	14	—·07101	0·933	20·856
5	·01950	0·083	9·187	15	·10241	1·902	22·758
6	·12078	3·122	12·309	16	—·00230	0·001	22·759
7	·03703	0·288	12·597	17	·09470	1·562	24·321
8	·01337	0·037	12·634	18	·04095	0·286	24·607
9	·04921	0·492	13·126	19	—·04107	0·282	24·889
10	—·03990	0·323	13·449	20	—·01493	0·036	24·925

Kendall's barley prices

s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$	s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$
1	—·20886	8·019	8·019†	11	·00340	0·002	21·142*
2	—·01528	0·042	8·061*	12	·01367	0·028	21·170*
3	—·01587	0·045	8·106*	13	·00017	0·000	21·170
4	—·02523	0·112	8·218	14	·01247	0·023	21·193
5	—·24002	9·928	18·146†	15	·09616	1·328	22·521
6	·05606	0·533	18·679†	16	·06852	0·661	23·182
7	·06002	0·600	19·279†	17	—·01177	0·019	23·201
8	—·05828	0·556	19·835*	18	—·01926	0·050	23·251
9	—·07066	0·803	20·638*	19	—·01130	0·017	23·268
10	—·05634	0·502	21·140*	20	—·05253	0·357	23·625

Kendall's oats prices

s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$	s	R_s^1	$\chi^2_{(1)}$	$\Sigma\chi^2_{(1)}$
1	—·15114	4·023	4·023*	11	—·09810	1·415	20·882*
2	—·20087	6·995	11·018†	12	·02960	0·128	21·010
3	·06577	0·738	11·756†	13	—·03101	0·138	21·148
4	—·15404	3·983	15·739†	14	·03699	0·192	21·340
5	·11821	2·307	18·046†	15	·12130	2·024	23·364
6	·07870	1·006	19·052†	16	·00706	0·007	23·371
7	—·03137	0·154	19·206†	17	—·00608	0·005	23·376
8	·01818	0·052	19·258*	18	·02389	0·074	23·450
9	·00783	0·009	19·267*	19	—·07951	0·800	24·250
10	—·03633	0·200	19·467*	20	—·05293	0·347	24·597

* Denotes 5 per cent. significance. † Denotes 1 per cent. significance.

It appears that, of the four series, only Kendall's wheat prices are fitted adequately by the chosen autoregressive scheme. That the other series differ from the fitted autoregressive schemes is shown by the high values of $\Sigma\chi^2_{(1)}$ and the correlation between successive values of R_s^1. For Kendall's wheat and oats prices, the estimates of a and b appear to be biased towards zero, and there is a strong suggestion of superposed variation, since the initial values are high, and $\chi^2_{(1)}$ rapidly decreases to a mean value of less than one.

It must be remembered that the above examples test particular autoregressive schemes, and not the most general autoregressive scheme. Thus it may be possible to a better fit by another second-order autoregressive scheme, or by a scheme of a higher order. This presents primarily a problem of fitting, which is being made the subject of another paper (Quenouille and Orcutt, 1947*b*).

Summary

A test has been proposed for the adequacy of representation of fitted autoregressive schemes, when the number of observations is large. This test has been found to work satisfactorily on artificial series. It has been shown how superposed variation can be detected by this method, and series experienced in practice have been used to demonstrate the test.

References

Bartlett, M. S. (1946), "On the theoretical specification and sampling properties of autocorrelated time-series," *J. Roy. Stat. Soc.*, **8**, 27.
Beveridge, Sir W. H. (1921)," Weather and harvest cycles," *Econ. J.*, **31**, 429.
Kendall, M. G. (1943)," Oscillatory movements in English agriculture," *J. Roy. Stat. Soc.*, **106**, 91.
—— (1946). *Contributions to the study of oscillatory time series.* Cambridge University Press.
Quenouille, M. H. (1947*a*)," Notes on the calculation of autocorrelations of linear autoregressive schemes," *Biometrika* (in press).
—— and Orcutt, G. H. (1947*b*)," The large-sample fitting of linear autoregressive schemes ".