Biometrika (1978), **65**, 2, *pp.* 297–303
With 1 *text-figure*
Printed in Great Britain

On a measure of lack of fit in time series models

BY G. M. LJUNG
College of Business Administration, University of Denver, Colorado
AND G. E. P. BOX
Department of Statistics, University of Wisconsin, Madison

SUMMARY

The overall test for lack of fit in autoregressive-moving average models proposed by Box & Pierce (1970) is considered. It is shown that a substantially improved approximation results from a simple modification of this test. Some consideration is given to the power of such tests and their robustness when the innovations are nonnormal. Similar modifications in the overall tests used for transfer function-noise models are proposed.

Some key words: Autoregressive-moving average model; Residual autocorrelation; Test for lack of fit; Transfer function-noise model.

1. INTRODUCTION

Consider a discrete time series $\{w_t\}$ generated by a stationary autoregressive-moving average model

$$\phi(B)\, w_t = \theta(B)\, a_t,$$

where $\phi(B) = 1-\phi_1 B-\ldots-\phi_p B^p$, $\theta(B) = 1-\theta_1 B-\ldots-\theta_q B^q$, $B^k w_t = w_{t-k}$, and $\{a_t\}$ is a sequence of independent and identically distributed $N(0, \sigma^2)$ random deviates. The w_t's can in general represent the d-th difference or some other suitable transformation of a non-stationary series $\{z_t\}$.

After a model of this form has been fitted to a series $w_1, \ldots, w_n$, it is useful to study the adequacy of the fit by examining the residuals $\hat{a}_1, \ldots, \hat{a}_n$ and, in particular, their autocorrelations

$$\hat{r}_k = \sum_{t=k+1}^{n} \hat{a}_t \hat{a}_{t-k} \Big/ \sum_{t=1}^{n} \hat{a}_t^2 \quad (k = 1, 2, \ldots).$$

An informal graphical analysis of these quantities combined with overfitting (Box & Jenkins, 1970, § 8.1) usually proves most effective in detecting possible deficiencies in the model. In addition, however, it is often worthwhile to look at an overall criterion of adequacy of fit. Box & Pierce (1970) noted that if the model were appropriate and the parameters were known, the quantity

$$\tilde{Q}(r) = n(n+2) \sum_{k=1}^{m} (n-k)^{-1} r_k^2, \tag{1·1}$$

where

$$r_k = \sum_{t=k+1}^{n} a_t a_{t-k} \Big/ \sum_{t=1}^{n} a_t^2,$$

would for large n be distributed as χ^2_m since the limiting distribution of $r = (r_1, \ldots, r_m)'$ is multivariate normal with mean vector zero (Anderson, 1942; Anderson & Walker, 1964),

$\text{var}(r_k) = (n-k)/\{n(n+2)\}$ and $\text{cov}(r_k, r_l) = 0$ $(k \neq l)$. Using the further approximation $\text{var}(r_k) = 1/n$, Box & Pierce (1970) suggested that the distribution of

$$Q(r) = n \sum_{k=1}^{m} r_k^2 \tag{1·2}$$

could be approximated by that of χ^2_m. Furthermore, they showed that when the $p+q$ parameters of an appropriate model are estimated and the $\hat{r}_k$'s replace the r_k's, then

$$Q(\hat{r}) = n \sum_{k=1}^{m} \hat{r}_k^2$$

would for large n be distributed as χ^2_{m-p-q} yielding an approximate test for lack of fit.

In applications of this test, suspiciously low values of $Q(\hat{r})$ have sometimes been observed, and studies by the present authors, reported in a University of Wisconsin technical report, and by Davies, Triggs & Newbold (1977) have verified that the distribution of $Q(\hat{r})$ can deviate from χ^2_{m-p-q}. This observation was also made by Prothero & Wallis (1976) in the discussion of their paper. The observed discrepancies could be accounted for by several factors, for instance departures from normality of the autocorrelations. It appears, however, that the main difficulty is caused by the approximation of (1·1) by (1·2). A modified test based on the criterion

$$\tilde{Q}(\hat{r}) = n(n+2) \sum_{k=1}^{m} (n-k)^{-1} \hat{r}_k^2$$

was recommended by the present authors but its usefulness was questioned by Davies *et al.* (1977) on the ground that the variance of $\tilde{Q}(\hat{r})$ exceeds that of the χ^2_{m-p-q} distribution. Our studies show however that the modified test provides a substantially improved approximation that should be adequate for most practical purposes.

2. Means and variances of $Q(r)$ and $\tilde{Q}(r)$

To examine the overall test, it is useful to consider initially the quantities $Q(r)$ and $\tilde{Q}(r)$ which involve the white noise autocorrelations r. Since the limiting distribution of r is $N(0, n^{-1} I_m)$, $Q(r)$ and $\tilde{Q}(r)$ are asymptotically distributed as χ^2_m and have expectation m and variance $2m$. For finite values of n, $\tilde{Q}(r)$ has expectation m, whereas

$$E\{Q(r)\} = n \sum_{k=1}^{m} E(r_k^2) = \frac{mn}{n+2}\left(1 - \frac{m+1}{2n}\right). \tag{2·1}$$

Clearly, unless n is large relative to m, $E\{Q(r)\}$ can be much smaller than m.

The variances are

$$\text{var}\{Q(r)\} = n^2 \sum_{k=1}^{m} \text{var}(r_k^2) + 2n^2 \sum_{k=1}^{m-1} \sum_{l=k+1}^{m} \text{cov}(r_k^2, r_l^2), \tag{2·2}$$

$$\text{var}\{\tilde{Q}(r)\} = n^2(n+2)^2 \sum_{k=1}^{m} (n-k)^{-2} \text{var}(r_k^2) + 2n^2(n+2)^2 \sum_{k=1}^{m-1} \sum_{l=k+1}^{m} (n-k)^{-1}(n-l)^{-1} \text{cov}(r_k^2, r_l^2),$$

where, for fixed n, $\text{cov}(r_k^2, r_l^2)$ is nonzero. The univariate and bivariate moments of the r_k's needed to evaluate (2·2) can be obtained using the identity

$$E(r_k^i r_l^j) = \frac{E\{(\Sigma a_t a_{t-k})^i (\Sigma a_t a_{t-l})^j\}}{E\{(\Sigma a_t^2)^{i+j}\}}, \tag{2·3}$$

which follows from independence of the r_k's and Σa_t^2 (Anderson, 1971, p. 304). Taking

$\text{var}(a_t) = 1$ without loss of generality, we have that Σa_t^2 is distributed as χ_n^2 and $E(\Sigma a_t^2)^{i+j} = n(n+2)\ldots(n+2i+2j-2)$. The term in the numerator of (2·3) can be evaluated by multiplying term by term and taking the expected value. It can thus be verified that for $k < \frac{1}{2}n$

$$\text{var}(r_k^2) = \frac{6(3n-5k)+3(n-k)^2}{n(n+2)(n+4)(n+6)} - \frac{(n-k)^2}{n^2(n+2)^2}, \tag{2·4}$$

$$\text{cov}(r_k^2, r_l^2) = \frac{(n-k)(n-l)+4(n-l)+8(n-k-l)}{n(n+2)(n+4)(n+6)} - \frac{(n-k)(n-l)}{n^2(n+2)^2}.$$

The exact variances of $Q(r)$ and $\tilde{Q}(r)$ are readily evaluated using (2·2) and (2·4). By ignoring terms of order higher than $1/n$ it may be shown that approximately, for n large relative to m,

$$\text{var}\{Q(r)\} = 2m\left(1+\frac{m-10}{n}\right), \quad \text{var}\{\tilde{Q}(r)\} = 2m\left(1+\frac{2m-5}{n}\right).$$

The variance of $\tilde{Q}(r)$ exceeds $2m$ but the absence of a location bias makes its distribution much closer to χ_m^2 than that of $Q(r)$. This is illustrated in Fig. 1 which compares Monte Carlo distributions of $Q(r)$ and $\tilde{Q}(r)$ based on 1000 replications to the χ_m^2 distribution for $m = 30$ and $n = 100$. The observed distribution of $Q(r)$ has mean 24·97 and variance 60·47; $\tilde{Q}(r)$ has mean 30·17 and variance 88·25. These values agree quite closely with the theoretical values 24·85, 63·15, 30·00 and 91·48, respectively. Also shown by dashed lines in Fig. 1 is a distribution of the form $a\chi_b^2$ for which both the mean and variance are adjusted to correspond with those of $\tilde{Q}(r)$. There is perhaps somewhat better agreement in the upper tail but the main improvement results from adjusting the mean.

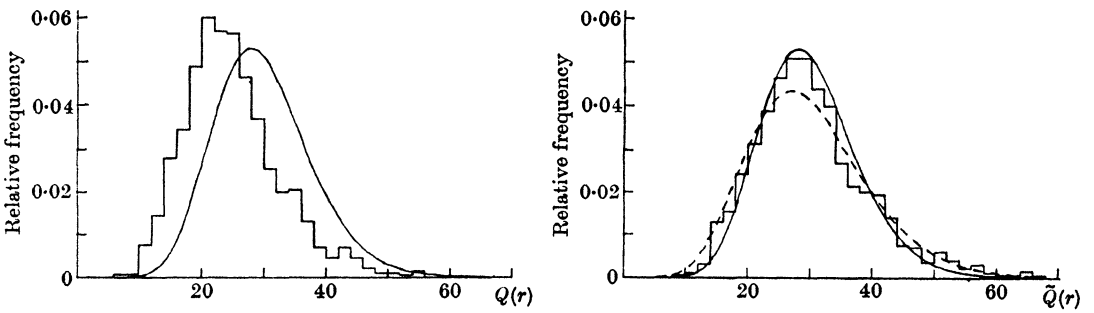

Fig. 1. Monte Carlo distributions of $Q(r)$ and $\tilde{Q}(r)$ and approximations; 1000 replications, $n = 100$ and $m = 30$; solid line, χ_{30}^2; dashed line, $a\chi_b^2$ ($a = 1{\cdot}52$, $b = 19{\cdot}68$).

3. The test statistics $Q(\hat{r})$ and $\tilde{Q}(\hat{r})$

Box & Pierce (1970) showed that the residual autocorrelations $\hat{r} = (\hat{r}_1, \ldots, \hat{r}_m)'$ from a correctly identified and fitted model can to a close approximation be represented as

$$\hat{r} \simeq (I-D)r,$$

where $I-D$ is an idempotent matrix of rank $m-p-q$. With this relationship the expectation of $Q(\hat{r})$ is

$$E\{Q(\hat{r})\} \simeq E\{nr'(I-D)r\} = \text{tr}\{n(I-D)C\},$$

where C is the exact covariance matrix of r. The matrix D has its largest elements in the upper left corner with the remaining elements d_{ij} decreasing to zero as i and/or j increases. The matrix DC is therefore nearly equal to $n^{-1}D$. Using this approximation and noting that

$E\{Q(r)\} = \text{tr}(nC)$, we have

$$E\{Q(\hat{r})\} \doteq E\{Q(r)\} - p - q. \tag{3·1}$$

On combining (2·1) and (3·1), the expected value of $Q(\hat{r})$ is approximately

$$E\{Q(\hat{r})\} \doteq \frac{mn}{n+2}\left(1 - \frac{m+1}{2n}\right) - p - q, \tag{3·2}$$

which indicates that the distribution of $Q(\hat{r})$ can deviate markedly from χ^2_{m-p-q} unless n is large relative to m. However, using the same approximations it can be shown that

$$E\{\tilde{Q}(\hat{r})\} \doteq E\{\tilde{Q}(r)\} - p - q = m - p - q.$$

It may be expected therefore that the distribution of $\tilde{Q}(\hat{r})$ might be approximated by the χ^2_{m-p-q} distribution.

The adequacy of this approximation was questioned by Davies *et al.* (1977) on the ground that the variance of $\tilde{Q}(\hat{r})$ exceeds $2(m-p-q)$. However, results from a simulation study reported in the next section suggest that the reduction in the location bias results as before in a markedly improved approximation that should be adequate for most practical purposes It also appears that the expression for the variance given by Davies *et al.*, which is not exact, overestimates the variance of $\tilde{Q}(\hat{r})$. For example, for fitting a first-order autoregressive model to white noise, Davies *et al.* obtain for $m = 20$ and $n = 50$, 100 and 200, $\text{var}\{\tilde{Q}(\hat{r})\} = 58·80$, 50·08 and 44·20, respectively, while our study gives $\text{var}\{\tilde{Q}(\hat{r})\} = 46·84$, 43·20 and 41·97, respectively.

4. Some numerical results

4·1. *Comparison of the overall tests*

A Monte Carlo study was conducted by generating 4000 sets of observations $\{w_1, \ldots, w_n\}$ from the first-order autoregressive model $w_t - \phi w_{t-1} = a_t$, estimating ϕ by the approximate maximum likelihood estimator

$$(n-2)(n-1)^{-1} \sum_{t=2}^{n} w_t w_{t-1} \Big/ \sum_{t=2}^{n-1} w_t^2$$

(Box & Jenkins, 1970, p. 279), and calculating autocorrelations of the residuals $\hat{a}_1 = (1-\hat{\phi}^2)w_1$, $\hat{a}_t = w_t - \hat{\phi} w_{t-1}$ $(t = 2, \ldots, n)$. The statistics $Q(\hat{r})$ and $\tilde{Q}(\hat{r})$ were then calculated.

Table 1 shows the proportion of $Q(\hat{r})$ and $\tilde{Q}(\hat{r})$ values exceeding the upper 5, 10 and 25 percentage points of the χ^2_{m-1} distribution for a few combinations of n and m and for $\phi = 0·5$. The table also gives the means and variances of the observed distributions. It seems clear that although the variance of $\tilde{Q}(\hat{r})$ exceeds $2(m-1)$ a test based on this statistic would for smaller sample sizes provide a considerable improvement over the previously used $Q(\hat{r})$ test.

Table 1. *Empirical means, variances and significance levels of the statistics* $Q(\hat{r})$ *and* $\tilde{Q}(\hat{r})$; *data generated from the model* $w_t - \frac{1}{2}w_{t-1} = a_t$

		$Q(\hat{r})$					$\tilde{Q}(\hat{r})$				
				% level					% level		
n	m	Mean	Var.	5	10	25	Mean	Var.	5	10	25
50	10	7·48	13·79	2·3	4·7	13·4	8·82	19·11	5·3	9·5	23·0
	20	13·96	27·50	1·3	2·3	6·4	18·58	47·76	6·1	10·4	23·2
100	10	8·14	16·04	3·4	7·0	18·2	8·83	18·88	5·0	9·9	23·1
	20	16·26	35·45	2·5	5·0	13·1	18·63	46·46	5·8	10·2	22·8
	30	23·53	55·74	1·7	3·6	9·1	28·58	81·71	7·2	11·6	23·4
200	10	8·57	16·76	4·2	8·3	21·5	8·92	18·16	5·0	9·8	23·9
	20	17·46	36·36	3·5	6·9	17·6	18·66	41·51	5·4	10·0	22·7
	30	26·11	56·01	2·9	5·6	14·2	28·66	67·37	5·9	10·5	23·8

4·2. *An alternative test based on* $Q(\hat{r})$

The above results suggest that a closer approximation to the distribution of $Q(\hat{r})$ should be obtainable by appropriate adjustment of the mean of the approximating distribution. Furthermore, Table 1 shows values of $\text{var}\{Q(\hat{r})\}$ which are nearly twice the mean, suggesting the approximation $Q(\hat{r}) \sim \chi^2_{E\{Q(\hat{r})\}}$ with $E\{Q(\hat{r})\}$ given by (3·2). Empirical significance levels obtained using this approximation and the criterion $\tilde{Q}(\hat{r})$ are compared in Table 2. The agreement is quite close. It may however be more convenient generally to use $\tilde{Q}(\hat{r})$, since the test based on $Q(\hat{r})$ will have noninteger degrees of freedom.

Table 2. *Empirical significance levels based on the approximations* $Q(\hat{r}) \sim \chi^2_{E\{Q(\hat{r})\}}$ *and* $\tilde{Q}(\hat{r}) \sim \chi^2_{m-1}$; *data generated from the model* $w_t - \phi w_{t-1} = a_t$

		$Q(\hat{r}) \sim \chi^2_{E\{Q(\hat{r})\}}$						$\tilde{Q}(\hat{r}) \sim \chi^2_{m-1}$					
		$m = 10$			$m = 20$			$m = 10$			$m = 20$		
		% level						% level					
n	ϕ	5	10	25	5	10	25	5	10	25	5	10	25
50	0·1	4·1	8·3	21·2	4·6	8·1	20·9	4·7	9·3	21·4	5·9	10·1	22·5
	0·4	4·3	8·5	22·1	4·6	8·6	21·6	5·1	9·3	22·8	6·0	10·3	23·0
	0·7	4·7	9·5	23·3	5·1	9·6	22·6	5·4	10·1	23·6	6·7	11·3	24·0
100	0·1	4·3	8·8	23·4	5·1	9·3	22·2	4·7	9·3	23·5	5·9	10·0	22·7
	0·4	4·4	8·5	23·5	5·3	9·1	22·7	4·8	9·3	23·5	6·0	10·0	23·0
	0·7	4·7	9·0	24·1	5·6	9·6	22·7	4·9	9·4	24·0	6·2	10·3	23·2
200	0·1	5·0	9·6	24·1	5·2	9·8	22·7	5·2	9·9	24·2	5·5	10·2	23·2
	0·4	4·8	9·5	23·8	5·1	9·6	22·5	5·1	9·8	24·0	5·4	10·1	22·8
	0·7	4·8	9·9	24·1	4·9	10·0	22·5	5·0	10·1	24·2	5·3	10·5	22·8

4·3. *A power calculation*

The two criteria $Q(\hat{r})$ and $\tilde{Q}(\hat{r})$ differ in the weighting which is applied to the autocorrelations $\hat{r}_k$ with $\tilde{Q}(\hat{r})$ giving more emphasis to later autocorrelations than $Q(\hat{r})$. This would perhaps be an advantage if serial correlation occurs at high lags k. However, for large n this difference should be rather small. If the type of discrepancies to be expected is known, tests specifically aimed at detecting these discrepancies should be used. Such specific tests will of course be much more powerful. This point is illustrated in Table 3 which empirically compares the power of the overall tests and the method of "overfitting" (Box & Jenkins, 1970). The results are based on data generated from a second-order autoregressive model, with a first-order model being fitted to obtain $Q(\hat{r})$ and $\tilde{Q}(\hat{r})$. As might be expected, the overall tests

Table 3. *Empirical power of the overall tests and the method of overfitting for* $n = 100$. *Assumed model*: $w_t - \phi w_{t-1} = a_t$; *true model*: $(1 - 0{\cdot}7B)(1 - G_2 B) w_t = a_t$. *Nominal significance level:* 5%

Test	m	$G_2 = 0$	$G_2 = 0{\cdot}1$	$G_2 = 0{\cdot}3$	$G_2 = 0{\cdot}5$	$G_2 = 0{\cdot}7$	$G_2 = 0{\cdot}9$
Overfitting		5·3	12·0	59·7	93·8	99·7	99·1
$Q(\hat{r}) \sim \chi^2_{E\{Q(\hat{r})\}}$	10	4·7	6·7	28·6	72·0	96·6	99·9
	20	5·6	7·3	24·4	62·8	93·7	99·7
	30	6·0	7·7	22·9	58·1	91·7	99·5
$\tilde{Q}(\hat{r}) \sim \chi^2_{m-1}$	10	4·9	7·0	28·9	71·6	96·2	99·9
	20	6·2	8·0	24·7	61·7	93·2	99·6
	30	7·0	9·0	23·7	57·0	90·5	99·3

are much less powerful than overfitting which tests the hypothesis that the second-order autoregressive coefficient is zero. A smaller value of m improves the power of the overall tests for this particular alternative.

4·4. *Effect of nonnormality of the a_t's*

In developing the overall test, it is assumed that the innovations a_t in the model are normally distributed. Circumstances occur where this assumption is not true. For example, it is known that stock price innovations often have highly leptokurtic distributions. Results by Anderson & Walker (1964) show that the asymptotic normality of the r_k's does not require normality of the a_t's, only that var (a_t) is finite. The overall test might therefore be expected to be insensitive to departures from normality of the a_t's. This is supported by Table 4, which shows the behaviour of $\tilde{Q}(\hat{r})$ when the a_t's have a double exponential and a uniform distribution. The results agree closely with those obtained under the normality assumption in Table 1.

Table 4. *Empirical means, variances and significance levels of $\tilde{Q}(\hat{r})$ when the innovations a_t have* (i) *a double exponential and* (ii) *a uniform distribution; data generated from the model*

$$w_t - \tfrac{1}{2}w_{t-1} = a_t$$

		(i) $a_t \sim$ double exponential					(ii) $a_t \sim$ uniform				
				% level					% level		
n	m	Mean	Var.	5	10	25	Mean	Var.	5	10	25
50	10	8·50	18·59	4·7	8·6	20·7	9·01	19·35	5·6	10·0	24·4
	20	17·77	47·00	5·4	8·8	19·6	18·95	52·39	7·3	12·1	24·3
100	10	8·80	18·70	5·0	9·1	22·4	9·11	19·41	5·5	10·8	25·3
	20	18·37	43·62	4·8	9·2	22·0	19·00	47·52	6·4	11·5	25·7
	30	27·94	76·60	6·3	10·1	21·9	28·98	81·72	7·5	12·4	25·3

5. EXTENSION TO TRANSFER FUNCTION NOISE MODELS

To check the adequacy of the transfer function in the model

$$w_t = \frac{\omega(B)}{\delta(B)}\alpha_t + \frac{\theta(B)}{\phi(B)}a_t,$$

where

$$\omega(B)/\delta(B) = (\omega_0 - \omega_1 B - \ldots - \omega_u B^u)/(1 - \delta_1 B - \ldots - \delta_v B^v)$$

and where the input series $\{\alpha_t\}$ is assumed to be white noise and independent of $\{a_t\}$, it is useful to examine the cross-correlations between $\{\alpha_t\}$ and the residuals $\{\hat{a}_t\}$

$$\hat{r}_k^* = \sum_{t=k+1}^{n} \alpha_{t-k}\hat{a}_t \Big/ \left(\sum_{t=1}^{n}\alpha_t^2 \sum_{t=1}^{n}\hat{a}_t^2\right)^{1/2} \quad (k = 0, 1, \ldots).$$

D. A. Pierce in a University of Wisconsin technical report, Box & Jenkins (1970, § 11.3) and Pierce (1972) propose an overall test for lack of fit based on approximating the distribution of

$$S(\hat{r}^*) = n\sum_{k=0}^{m}(\hat{r}_k^*)^2$$

by the χ^2_{m-v-u} distribution. However, on arguing as above, it appears that a criterion of the form

$$\tilde{S}(\hat{r}^*) = n^2\sum_{k=0}^{m}(n-k)^{-1}(\hat{r}_k^*)^2$$

would be more appropriate. The criterion $S(\hat{r}_k^*)$ is obtained by approximating the variance of the k-th sample cross-correlation between $\{\alpha_t\}$ and $\{a_t\}$ by $1/n$, while the actual variance is $(n-k)/n^2$.

The modification considered in the previous sections applies to the overall test for lack of fit in the noise model $\theta(B)/\phi(B)$ discussed by Box & Jenkins (1970, § 11.3).

This work was sponsored by the United States Army Research Office and the Air Force Office of Scientific Research.

References

Anderson, R. L. (1942). Distribution of the serial correlation coefficients. *Ann. Math. Statist.* **13**, 1–13.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. New York: Wiley.

Anderson, T. W. & Walker, A. M. (1964). On the asymptotic distribution of the autocorrelations of a sample from a linear stochastic process. *Ann. Math. Statist.* **35**, 1296–303.

Box, G. E. P. & Jenkins, G. M. (1970). *Time Series Analysis Forecasting and Control*. San Francisco: Holden-Day.

Box, G. E. P. & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J. Am. Statist. Assoc.* **65**, 1509–26.

Davies, N., Triggs, C. M. & Newbold, P. (1977). Significance levels of the Box–Pierce portmanteau statistic in finite samples. *Biometrika* **64**, 517–22.

Pierce, D. A. (1972). Residual correlations and diagnostic checking in dynamic-disturbance time series models. *J. Am. Statist. Assoc.* **67**, 636–40.

Prothero, D. L. & Wallis, K. F. (1976). Modelling macroeconomic time series (with discussion). *J. R. Statist. Soc.* A **139**, 468–500.

[*Received September* 1977. *Revised January* 1978]