CHAPTER 7

DIAGNOSTIC CHECKING

7.1 INTRODUCTION

In Chapter 5, a variety of useful graphical tools are presented for identifying one or more promising ARMA or ARIMA models to fit to a given time series. Subsequent to model identification, the method of maximum likelihood described in Chapter 6 can be employed for obtaining MLE's and SE's for the model parameters. When parameter estimates are calculated for more than one model, the AIC of Section 6.3, or another appropriate ASC mentioned in Section 6.3.6, can be used to select the overall best model. The objective of Chapter 7 is to ensure that this model adequately describes the time series under consideration by subjecting the calibrated model to a range of statistical tests which are referred to as *diagnostic checks*. The overall approach to model construction is displayed in Figure III.I while Figure 6.3.1 shows the ways in which the AIC can be used in conjunction with the model building stages.

One class of diagnostic checks is devised to test model adequacy by *overfitting*. This approach assumes that the possible types of model inadequacies are known in advance. The procedure of overfitting consists of including one or more extra parameters in the model to ascertain if an improved model can be designed (Box and Jenkins, 1976, Ch. 8; Granger and Newbold, 1977, Ch. 3). Section 7.2 explains how overfitting can be carried out in practice.

The most useful and informative diagnostic checks deal with determining whether or not the assumptions underlying the innovation series are satisfied by the residuals of the calibrated ARMA or ARIMA model. As pointed out in Section 3.4.5 and many other locations in the book, when fitting a model to a time series the estimated innovations or *residuals* are assumed to be independent, homoscedastic (i.e. have a constant variance) and normally distributed. Estimates for the a_t's are automatically calculated at the estimation stage along with MLE's and SE's for the model parameters (see Appendices A6.1 and A6.2).

Of the three innovation assumptions, *independence* and, hence, whiteness, is by far the most important. A data transformation cannot correct dependence of the residuals because the lack of independence indicates the present model is inadequate. Rather, the identification and estimation stages must be repeated in order to determine a suitable model. If the less important assumptions of *homoscedasticity* and *normality* are violated, they can often be corrected by a *Box-Cox transformation* of the data defined in [3.4.30].

Table 7.1.1 lists the main problems that can occur with the statistical properties of the residuals of a fitted model and how they can be corrected. Diagnostic checks for whiteness, normality and homoscedasticity of the residuals are presented in Sections 7.3 to 7.5, respectively, along with explanations regarding corrective actions that can be taken. Practical applications of applying these tests to a yearly riverflow series and sunspot numbers are presented in Section 7.6.

One should keep in mind that diagnostic checks only have meaning if the parameters of the model are efficiently estimated using the *maximum likelihood approach* of Chapter 6 at the estimation stage. If, for example, the method of moments were used to estimate the parameters of an ARMA model containing MA parameters, these moment estimates would be inefficient and

Table 7.1.1. Rectifying violations of the assumptions underlying the model residuals.

Violations of Residual Assumptions	Corrective Actions	Sections
Dependence and non-whiteness	Consider other models	7.3
Variance change or heteroscedasticity	Box-Cox data transformation	7.4
Non-normality	Box-Cox data transformation	7.5

probably quite different from the corresponding MLE's. Problems arising in the residuals of the ARMA model calibrated using moment estimates may be due to the inefficiency of the estimator rather than the specific parameters included in the model. Accordingly, for all of the diagnostic checks presented in Chapter 7 it is assumed that a maximum likelihood estimator is used to estimate the model parameters. For ARMA models, the only exception to this is the case of a pure AR model in [3.2.5]. Recall that for an AR model, both the method of moments using the *Yule-Walker equations* in [3.2.12] and the technique of maximum likelihood furnish efficient parameter estimates.

7.2 OVERFITTING

Overfitting involves fitting a more elaborate model than the one estimated to see if including one or more additional parameters greatly improves the fit. Extra parameters should be estimated for the more complex model only where it is feared that the simpler model may require more parameters. For example, the sample PACF and the IACF for an annual time series may possess decreasing but significant values at lags 1, 2, and 9. If an AR(2) model were originally estimated, then a model to check by overfitting the model would be

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_9 B^9)(w_t - \mu) = a_t$$

In Section 6.4.3, this is the type of model which is fitted to the square roots of the yearly sunspot numbers. Because, as shown in Table 6.4.3, the MLE of ϕ_9 is more than three times the value of its SE, this indicates that the more elaborate AR model containing ϕ_9 should be selected. Moreover, the AIC (Table 6.4.2) and diagnostic checks applied to the residuals of the constrained AR(9) model fitted to the square roots of the annual sunspot numbers (Section 7.6.3) confirm that the more complex model should be employed. Box and Newbold (1971, Section 3.6) as well as Box and Jenkins (1976, Section 8.1.2) show other interesting applications of overfitting.

The practitioner must take care to avoid *model redundancy* which could occur if the AR and MA components were simultaneously enlarged. For example, suppose that one initially fits an AR(1) model to a series but then expands the model by adding one more AR plus an additional MA parameter to form an ARMA(2,1) model. Suppose that the difference equation for an ARMA(2,1) model fitted to a given series given as w_t is

$$(1 - 0.80B + 0.12B^2)(w_t - 26) = (1 - 0.20B)a_t$$

Upon examining the SE's for some of the MLE's for the parameters one sees that they are very large. For example, the SE for ϕ_2 may be 0.22 which is much larger than $\hat{\phi}_2 = 0.12$, even though there are 200 entries in the series. The reason for a large SE is the instability introduced into the estimation algorithm due to parameter redundancy. Notice that the difference equation can be written as

$$(1 - 0.60B)(1 - 0.20B)(w_t - 26) = (1 - 0.20B)a_t$$

which simplifies to

$$(1 - 0.60B)(w_t - 26) = a_t$$

Therefore, the AR(1) model is more appropriate than the ARMA(2,1) model for fitting to the series.

Whenever one notices abnormally large SE's one should check for redundant or nearly redundant factors in a model due to overspecifying the model and then take corrective action by removing the redundant factors and fitting a simpler model. The over specification of the model parameters may cause rather large flat regions near the maximum point of the likelihood function and this in turn means that the SE's must be large (see Appendix A6.2). The large SE's suggest that a wide range of models could suitably model the data. However, in keeping with the principle of model parsimony, the simpler model should be chosen and, hence, redundancy should be avoided.

The problem of model redundancy provides an explanation as to why one cannot start out by fitting an overspecified model having many parameters and then reducing the number of parameters until an adequate model is found. Rather, one must begin with a fairly simple model and then carefully expand to a more complicated model, if necessary.

Another method of testing model adequacy by overfitting, which was originally suggested by Whittle (1952), is to fit a high-order AR model of order r where $20 < r < 30$. Suppose the original model has k estimated parameters plus the estimated residual variance, $\hat{\sigma}_a^2(k)$. Then it is shown (McLeod, 1974; Hipel et al., 1977) that the *likelihood ratio statistic* is

$$n \ln \left[\hat{\sigma}_a^2(k)/\hat{\sigma}_a^2(r)\right] \approx \chi^2(r-k) \qquad [7.2.1]$$

where $\hat{\sigma}_a^2(r)$ is the residual variance estimate for an AR process of order r. If the calculated $\chi^2(r-k)$ from [7.2.1] is greater than $\chi^2(r-k)$ from the tables at a chosen significance level, then a model with more parameters is needed.

The likelihood ratio test in [7.2.1] can also be used to determine if a model containing fewer parameters gives as good a fit as the full model. An application of this test is presented in Section 6.4.2 where three types of AR models are fitted to the average annual flows of the St. Lawrence River. The likelihood ratio test, as well as the AIC, select a constrained AR(3) without ϕ_2 as the best AR model to fit to the St. Lawrence flows.

When using the likelihood ratio test, the models being compared must be *nested*. Hence, the less complex model must be contained within the more complicated one. For instance, an AR(1) model is nested within an AR(k) model for $k \geq 2$. As pointed out in Section 6.3, when

using the AIC for model discrimination, the models do not have to be nested and one can compare any number of different kinds of models at the same time.

7.3 WHITENESS TESTS

7.3.1 Introduction

The a_t sequence for AR (see Section 3.2), MA (Section 3.3), ARMA (Section 3.4) and ARIMA (Section 4.3) models are assumed to be independently distributed in the theoretical definition of these models. This implies that the estimated innovations or residuals are uncorrelated or white. In the next subsections, a number of statistical tests are described for determining whether or not the residuals, represented as $\hat{a}_t$, $t = 1,2, \ldots, n$, are white.

7.3.2 Graph of the Residual Autocorrelation Function

The most informative approach to check for whiteness is to examine a graph of the *residual autocorrelation function (RACF)*. The RACF at lag k is calculated as

$$r_k(\hat{a}) = \sum_{t=k+1}^{n} \left(\hat{a}_t \hat{a}_{t-k} / \sum_{i=1}^{n} \hat{a}_i^2 \right) \qquad [7.3.1]$$

Because of the term in the denominator in [7.3.1], the values of the RACF can range between -1 and +1. Additionally, since the RACF is symmetric about lag zero, one can plot the RACF against lags for positive lags from lag one to about lag $n/4$.

When examining a plot of the RACF, one would like to know if a given value is significantly different from zero. Asymptotically, the RACF is normally distributed as $N(0,\frac{1}{n})$ for any lag. Therefore, to draw the 95% confidence interval, for example, one can plot $\frac{+1.96}{\sqrt{n}}$ and $\frac{-1.96}{\sqrt{n}}$ above and below, respectively, the lag axis. If a given value of the RACF is significantly different from zero, it will fall outside the confidence interval.

A more accurate derivation for the large sample distribution of the RACF is provided by McLeod (1978). Define the vector of the first L values of the RACF as

$$\mathbf{r}(\hat{a}) = [r_1(\hat{a}), r_2(\hat{a}), \ldots, r_L(\hat{a})]' \qquad [7.3.2]$$

Denote by $\psi_k(\phi)$ the coefficient of B^k in the Maclaurin series expansion of $[\phi(B)]^{-1}$, where $\phi(B)$ is the AR operator defined in [3.4.4] as $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$. Likewise, let $\psi_k(\theta)$ be the coefficient of B^k in the Maclaurin series expansion of $[\theta(B)]^{-1}$, where $\theta(B)$ is the MA operator given in [3.4.4] as $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$. Then it can be proven for large samples that the residuals in $\mathbf{r}(\hat{a})$ in [7.3.2] follow the multivariate normal distribution given as:

$$\mathbf{r}(\hat{a}) \sim N[\mathbf{0}, \frac{\mathbf{U}}{n}] \qquad [7.3.3]$$

where $\mathbf{U} = 1_L - X'I^{-1}X$, 1_L is the identity matrix, $I \simeq X'X$ is the large-sample information

matrix, and $X = [\psi_{i-j}(\phi), \psi_{i-j}(\theta)]$ are the i,j entries in the two partitions of the X matrix. The dimensions of the matrices X, $\psi_{i-j}(\phi)$, and $\psi_{i-j}(\theta)$ are, respectively, $L \times (p+q)$, $L \times p$, and $L \times q$.

Notice in [7.3.3] that U is a function of the AR and MA parameters in the ARMA model fitted to the original series. This is the reason why the findings are better than earlier work. Previously, Box and Pierce (1970) obtained [7.3.3] for an AR model but the result in [7.3.3] is valid for a more general ARMA model. Finally, equation [7.3.3] can be extended for use with seasonal ARIMA models (Section 12.3.4) as well as the other ARMA based models presented in Parts VI to IX.

To obtain the 95% confidence interval for the RACF at lag k, one calculates

$$\underset{\substack{\textit{confidence}\\ \textit{interval}}}{95\%} = \pm 1.96\sqrt{\frac{1}{n}U_{kk}}$$

where U_{kk} is the diagonal entry at location k,k in the matrix U in [7.3.3]. For each lag $k = 1,2,\ldots,L \simeq \frac{n}{4}$, one can determine the 95% confidence interval which can be plotted on a graph of the values of the RACF against lag k. Usually, the most important values of the RACF to examine are those located at the first few lags for nonseasonal data. If one or more of the values of the RACF fall outside the 95% confidence interval, this means that the current model is inadequate. The use of these confidence limits for checking model adequacy is discussed by Hipel et al. (1977), McLeod et al. (1977) and McLeod (1977).

When the present model is insufficient due to correlated residuals, one can use the results contained in a graph of the RACF to update the model. Suppose, for example, that an examination of the graph of the RACF reveals that the residuals of an AR(1) model fitted to the given w_t series are correlated at lag one. Hence, the inadequate model can be written as

$$(1 - \phi_1 B)(w_t - \hat{\mu}) = b_t$$

where ϕ_1 is the AR parameter, μ is the mean of the w_t series and b_t is the residual series that is correlated at lag one. Because the RACF has a significantly large value at lag one, the following MA(1) model can be fitted to the b_t series representing the correlated residuals:

$$b_t = (1 - \theta_1 B)a_t$$

where θ_1 is the MA parameter. By substituting b_t into the previous equation, one obtains the ARMA(1,1) model written as

$$(1 - \phi_1 B)(w_t - \mu) = (1 - \theta_1 B)a_t$$

Consequently, one can fit an ARMA(1,1) model to the original w_t series in order to obtain MLE's for the parameters when the parameters are all estimated together within the same ARMA(1,1) model framework. The residuals of the ARMA(1,1) can then be subjected to rigorous diagnostic checks in order to ascertain if further model modifications are required.

In the foregoing example for redesigning a model having correlated residuals, the form of the RACF clearly indicates how to expand the model. When this is not the case, other procedures can be employed for developing a more suitable model. One approach is to repeat the

identification and estimation stages of model construction shown in Figure III.1 in order to discover a more suitable model. Another alternative is to use the AIC in conjunction with the earlier stages of model construction by following an appropriate path in Figure 6.3.1.

7.3.3 Portmanteau Tests

Rather than examine the magnitude of the value of RACF at each lag as is done in the previous subsection, one could look at an overall test statistic which is a function of the RACF values from lags one to L in order to perform a significance test for whiteness. However, this type of test is less sensitive because the lag locations of significantly large correlations and their magnitudes are buried in the test statistic. When a test statistic indicates a correlation problem in the RACF, one must then examine the graph of the RACF in order to understand what is happening and, subsequently, take corrective action.

Box and Pierce (1970) developed a Portmanteau statistic given as

$$Q'_L = n\sum_{k=1}^{L} r_k^2(\hat{a}) \qquad [7.3.4]$$

which is χ^2 distributed on $(L-p-q)$ degrees of freedom. Later, Davies et al. (1977) and Ljung and Box (1978) derived an improved version of the Portmanteau statistic which is written as

$$Q''_L = n(n+2)\sum_{k=1}^{L} r_k^2(\hat{a})/(n-k) \qquad [7.3.5]$$

and is also χ^2 distributed on $(L-p-q)$ degrees of freedom. More recently, Li and McLeod (1981) devised another enhanced Portmanteau statistic to test for whiteness. Specifically, if L is large enough so that the weights $\psi_k(\phi)$ and $\psi_k(\theta)$ in [7.3.3] have damped out, then

$$Q_L = n\sum_{k=1}^{L} r_k^2(\hat{a}) + \frac{L(L+1)}{2n} \qquad [7.3.6]$$

where Q_L is χ^2 distributed on $(L-p-q)$ degrees of freedom, and L can be given a value from about 15 to 25 for nonseasonal time series where L is not greater than about $n/4$. A test of this hypothesis can be done for model adequacy by choosing a level of significance and then comparing the value of the calculated χ^2 to the actual χ^2 value for $(L-p-q)$ degrees of freedom from the tables. If the calculated value is greater, on the basis of the available data the present model is inadequate, and appropriate changes must be made by examining in detail a plot of the RACF and, perhaps, also identification graphs of the original w_t series.

The modified Portmanteau statistics in [7.3.5] and [7.3.6] are recommended for employment over the first version in [7.3.4]. Moreover, the statistic in [7.3.6] has advantages over the one defined in [7.3.5]. In particular, using simulation experiments, Kheoh and McLeod (1992) demonstrate that the Portmanteau test statistic in [7.3.6] has a more accurate significance level than the one in [7.3.5] and possesses about the same power as that statistic. Also, the test statistic in [7.3.6] can be naturally extended for use in the multivariate case as in [21.3.2].

7.3.4 Other Whiteness Tests

A range of other whiteness tests can be employed for checking whether or not the residuals of a fitted ARMA model are white. For example, one can use the *cumulative periodogram graph* of Section 2.6 to test for whiteness. However, when examining model residuals, it is known that this test is inefficient. Often the cumulative periodogram test fails to indicate model inadequacy due to dependence of the residuals unless the model is a very poor fit to the given data.

A quite different approach to whiteness tests is to examine the *autocorrelation function (ACF) of the squared model residuals*, $\hat{a}_t^2$, $t = 1,2,\ldots,n$, which is estimated at lag k as

$$r_k(\hat{a}^2) = \sum_{t=k+1}^{n} \left[(\hat{a}_t^2 - \hat{\sigma}_a^2)(\hat{a}_{t-k}^2 - \hat{\sigma}_a^2)\right] / \left[\sum_{t=1}^{n}(\hat{a}_t^2 - \hat{\sigma}_a^2)^2\right] \qquad [7.3.7]$$

where the variance of the residuals is calculated using

$$\hat{\sigma}_a^2 = \sum_{t=1}^{n} \hat{a}_t^2/n$$

Consider the vector of squared residuals given by

$$\mathbf{r}(\hat{a}^2) = [r_1(\hat{a}^2), r_2(\hat{a}^2), \ldots, r_L(\hat{a}^2)]^T \qquad [7.3.8]$$

For fixed L, McLeod and Li (1983) show that $\sqrt{n}\,\mathbf{r}(\hat{a}^2)$ is asymptotically multivariate normal with mean zero and unit covariance matrix. Hence, one could check for correlation of the squared residuals by examining a graph of $r_k(\hat{a}^2)$ against lag $k = 1,2,\ldots,L$, along with the 95% confidence limits. Furthermore, a significance test is provided by the Portmanteau statistic (Ljung and Box, 1978)

$$Q_L(\hat{a}^2) = n(n+2)\sum_{k=1}^{L} r_k^2(\hat{a}^2)/(n-k) \qquad [7.3.9]$$

which is asymptotically χ^2 distributed on $(L-p-q)$ degrees of freedom if the a_t are independent.

In some applications, the autocorrelation function of the squared residuals is more sensitive than the RACF for detecting residual dependence. In particular, the autocorrelation function of squared residuals have been found especially useful for detecting nonlinear types of statistical dependence in the residuals of fitted ARMA models (Granger and Andersen, 1978; Miller, 1979; McLeod and Li, 1983).

7.4 NORMALITY TESTS

7.4.1 Introduction

The theoretical definitions for AR, MA, ARMA and ARIMA models are presented in Sections 3.2.2, 3.3.2, 3.4.2, and 4.3.1, respectively. Recall that for each of these models it is assumed that the innovations, represented by the a_t's, are identically and independently distributed. This means that the disturbances must follow the same distribution, such as a Gamma or Gaussian distribution, and be independent of one another. As pointed out in Section 6.2, in order

to obtain estimates for the model parameters one must assume that the innovations follow a specific distribution. In particular, comprehensive maximum likelihood estimators for ARMA models have been developed for the situation where the a_t's are Gaussian or normally distributed. A maximum likelihood estimator which is both statistically and computationally efficient is described in Appendix A6.1.

A wide range of flexible tests are available for ascertaining whether or not the residuals of a fitted ARMA model follow a normal distribution. Some of these normality tests are described in the subsequent subsections. If, for example, tests reveal that the residuals are not normal, one can transform the given data using the Box-Cox transformation in [3.4.30]. After fitting an ARMA model to the transformed series, one can employ appropriate normality tests to check whether or not the residuals from this model are Gaussian.

In addition to the statistical tests presented in the next three subsections and elsewhere, one can employ graphical methods for visually detecting departures from normality. A range of graphical techniques for use in exploratory data analysis are presented in Section 22.3 and referred to in Section 5.3.2. Some of these graphs can be used as visual normality checks. For example, if the box and whisker graph in Section 22.3.3 for the given time series is fairly symmetric, one can argue that the data follow a symmetric distribution such as a normal distribution. In a plot of the series against time, one should not see a lot of extreme values if the w_t series is Gaussian.

7.4.2 Skewness and Kurtosis Coefficients

Let the residual series for the fitted ARMA or ARIMA model be given as $\hat{a}_t$, $t = 1,2,\ldots,n$. If the $\hat{a}_t$'s are normally distributed, they should possess no significant skewness. The *skewness coefficient* g_1 for the $\hat{a}_t$ series is calculated using

$$g_1 = \left(\frac{1}{n}\sum_{t=1}^{n}\hat{a}_t^3\right) \Big/ \left(\frac{1}{n}\sum_{t=1}^{n}\hat{a}_t^2\right)^{3/2} \qquad [7.4.1]$$

To test the null hypothesis that the data are normal and therefore possess no significant skewness, one must know the distribution of g_1. D'Agostino (1970) presents a method for transforming g_1 so that the transformed value is distributed as $N(0,1)$. This allows one to calculate the significant level for g_1.

The steps required in transforming g_1 to a random variable which is $N(0,1)$ are as follows (D'Agostino, 1970):

1. $Y = g_1\left[\dfrac{(n+1)(n+3)}{6(n-2)}\right]^{1/2}$ where g_1 is calculated from the $\hat{a}_t$ series using [7.4.1].
2. $B_2 = \dfrac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$
3. $W^2 = -1 + [2(B_2-1)]^{1/2}$
4. $\delta = (\ln W)^{-1/2}$

5. $\alpha = [2/(W^2 - 1)]^{1/2}$

6. $Z = \delta \ln[Y/\alpha + \left\{(Y/\alpha)^2 + 1\right\}^{1/2}]$

The random variable Z, which is a transformation of the skewness coefficient g_1, is distributed as $N(0,1)$.

After calculating Z and choosing a level of significance, one can refer to standard normal tables to determine whether or not Z is significantly large. If, for example, Z has a significance level which is less than 0.05 according to the tables, one can assume that based upon the current information the $\hat{a}_t$ series possesses significant skewness and is, therefore, not normally distributed.

The *kurtosis coefficient* for the $\hat{a}_t$ series is determined as

$$g_2 = \left(\frac{1}{n}\sum_{t=1}^{n}\hat{a}_t^4\right) / \left(\frac{1}{n}\sum_{t=1}^{n}\hat{a}_t^2\right)^2 - 3 \qquad [7.4.2]$$

If the given data are normal, the statistic g_2 is approximately distributed as $N(0,24/n)$. Hence, for an estimated g_2, one can calculate the significance level for testing the null hypothesis that the data are normally distributed.

7.4.3 Normal Probability Plot

As before, suppose that a residual series is given as $\hat{a}_t$, $t = 1,2,\ldots,n$. When the entries in the $\hat{a}_t$ series are ordered from smallest to largest, the sample order statistic is

$$\hat{a}_{(1)} \le \hat{a}_{(2)} \le \cdots \le \hat{a}_{(n)} \qquad [7.4.3]$$

Let the hypothesized cumulative distribution function of the transformed data be $F(\hat{a}/\hat{\sigma}_a)$. Also, let p_i, which is called the plotting position, be an estimate of $F(\hat{a}_{(i)}/\hat{\sigma}_a)$. Hence, $F^{-1}(p_i)$ is the theoretical standard quantile. To construct a probability plot, the $\hat{a}_{(i)}$ and $F^{-1}(p_i)$ are plotted as the abscissae and ordinates, respectively.

Following the recommendation of Looney and Gulledge (1985), for the case of a normal probability plot, the plotting position of Blom (1958) is recommended for use in practical applications. This plotting position is defined as

$$p_i = \frac{i - 0.375}{n + 0.25} \qquad [7.4.4]$$

When the $\hat{a}_t$'s are $N(0,\hat{\sigma}_a^2)$, a normal probability plot, consisting of the theoretical standard normal quantile $F^{-1}(p_i)$ being plotted against the empirical quantile $\hat{a}_{(i)}$, should form a straight line. The 95% Kilmogorov-Smirnov confidence interval (CI) can also be included with the normal probability plot. For a given plotting position, p_i, the two sides of the confidence interval are calculated using (Lilliefors, 1967)

$$95\%CI = \hat{a}_t + \hat{\sigma}_a \cdot F^{-1}\left[p_i \pm \frac{0.886}{\sqrt{n}}\right] \qquad [7.4.5]$$

The reader should keep in mind that this procedure is known to not be very sensitive to departures from normality, particularly in the tails. Additional research on probability plots includes contributions by Stirling (1982), Michael (1983) and Royston (1993).

7.4.4 Other Normality Tests

Besides those tests described in the previous two subsections, many other tests are available for determining whether or not a time series such as the sequence of model residuals is normally distributed. Normality tests are described in most standard statistical textbooks, statistical encyclopediae and handbooks, plus research papers. Shapiro et al. (1968), for instance, review and compare nine methods for testing for normality in a single sample. Two normality tests are briefly referred to below.

Shapiro-Wilk Test

The *Shapiro-Wilk test* for normality is based on the test statistic

$$W = b^2 / \sum_{t=1}^{n} \hat{a}_t^2 \qquad [7.4.6]$$

where b^2 is proportional to the best linear unbiased estimate of the slope of the linear regression of $\hat{a}_{(i)}$ in [7.4.3] on the expected value of the ith normal order statistic (Shapiro and Wilk, 1965). A general algorithm for calculating W and its significance level is given by Royston (1982). Simulation experiments suggest that the Shapiro-Wilk test is a good general omnibus test for normality in many situations. Finally, Filliben (1975) defines the normal probability plot correlation coefficient, which is closely related to the Shapiro-Wilk statistic, and compares the power of this test statistic for normality with six others.

Blom's Correlation Coefficient

Looney and Gulledge (1985) recommend the use of a correlation coefficient test for normality. The test, which is based upon Blom's plotting position, summarizes and objectively evaluates the information contained in a normal probability plot.

The test statistic for the composite test of normality is constructed using the Pearson product-moment correlation coefficient between $F^{-1}(p_i)$ and $\hat{a}_{(i)}$. As with the Shapiro-Wilk test, "large" values for the test statistic tend to support the assumption of normality. The significance range for the correlation coefficient test is obtained from the tabulated empirical percentage points printed in Looney and Gulledge's (1985) paper. Monte Carlo results indicate that this correlation coefficient test compares quite favourably to the Shapiro-Wilk test.

7.5 CONSTANT VARIANCE TESTS

7.5.1 Introduction

For the ARMA and ARIMA models of Chapters 3 and 4, respectively, as well as most of the other models in the book, the innovation series is assumed to have a constant variance, σ_a^2. The statistical word for constant variance is *homoscedasticity*. One would like the residuals of a fitted ARMA or ARIMA model to be homoscedastic.

If the variance of the innovations change, they are said to be *heteroscedastic*. Changing variance or heteroscedasticity can occur in a number of different ways. Firstly, the variance of the residuals may increase or decrease over time. Secondly, the variance may be a function of the magnitude of the series. For instance, the variance may be greater for higher values of the innovations and lower for smaller values. In the next section, tests are presented for checking for variance changes that occur over time and changes that are dependent upon level.

The plot of the Beveridge wheat price indices are shown in Figure 4.3.15. As can be seen, the variance or "spread" of the data is increasing over time. If an ARIMA model were fitted directly to the given time series, the variance of the residuals of the model would also become greater with increasing time. Consequently, as explained in Section 4.3.3, to alleviate problems with heteroscedasticity the wheat price indices are first transformed using the natural logarithmic transformation contained in [3.4.30] before fitting an ARIMA model to the series. In general, an appropriate Box-Cox transformation can often alleviate the problem of heteroscedasticity in the model residuals.

7.5.2 Tests for Homoscedasticity

The following tests were developed by McLeod (1974), and their application described by Hipel et al. (1977) and McLeod et al. (1977), are useful for determining whether a transformation of the data is needed by checking for changes in variance (heteroscedasticity) of the residuals. As is mentioned earlier, the variance of the normally independently distributed residuals is assumed to be constant (homoscedastic). Suppose that a_t is NID$[0,\sigma_a^2(t)]$ and that the variance changes with time as $\sigma_a^2(t)$. Let the stochastic random variable ζ_t be NID $(0,\sigma^2)$ and hence have constant variance. Suppose then that

$$a_t = \exp\left\{(\chi/2)[K(t) - \bar{K}]\right\}\zeta_t \qquad [7.5.1]$$

where χ is some constant to be estimated, $K(t)$ is a function of time to be specified, and $\bar{K}$ is the mean of $K(t)$ and equals $n^{-1}\sum_{t=1}^{n} K(t)$. The variance of the a_t residuals is then

$$\sigma_a^2(t) = E\left\{\exp[\chi(K(t) - \bar{K})]\zeta_t^2\right\}$$

$$= \exp\left\{\chi[K(t) - \bar{K}]\right\}\sigma^2 \qquad [7.5.2]$$

It can be shown that the natural logarithm of the likelihood Lh for σ^2 and χ is

$$Lh = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\left\{\exp[-\chi(K(t)-\bar{K})]a_t^2\right\} \qquad [7.5.3]$$

and

$$\frac{\partial Lh}{\partial \chi} = \frac{1}{\sigma^2}\sum_{t=1}^{n}\left\{K(t)\exp[-\chi(K(t)-\bar{K})]a_t^2\right\} \qquad [7.5.4]$$

One solves $\partial Lh/\partial\sigma^2 = 0$ exactly for σ^2, and substitutes for σ^2 into [7.5.4]. Next, equation [7.5.4] is set equal to zero, and the residual estimates $\hat{a}_t$ obtained from the estimation stage in Section 6.2 are used for a_t. This equation is then solved for a MLE of χ by using the Newton-Raphson method with an initial value of $\chi = 0$.

In order to carry out a test of the hypothesis, the first step is to postulate the null hypothesis that $\chi = 0$ and, therefore, to assume that the residuals have constant variance. The alternative hypothesis is that the residuals are heteroscedastic and that $\chi \neq 0$. By putting $K(t) = t$ in the previous equations, it is possible to test for trends in variance of the residuals over time. If $K(t) = w_t - \hat{a}_t$, then one can check for changes of variance depending on the current level of the w_t series in [4.3.3]. A likelihood ratio test of the null hypothesis is obtained by computing the MLE of χ and comparing it with its standard error. The variance for the MLE $\hat{\chi}$ for χ is calculated by using the equation

$$Var\hat{\chi} = -1/(\partial^2 Lh/\partial\chi^2) \qquad [7.5.5]$$

Because the MLE for χ is asymptotically normally distributed, after a level of significance is chosen it is a straightforward procedure to determine whether to accept or to reject the null hypothesis. This test is also valid for transfer function-noise, intervention, multivariate ARMA and regression models. In regression models, the test for heteroscedasticity can indicate whether an important covariate is missing (Anscombe, 1961; Pierce, 1971).

If model inadequacy is revealed by either of the tests, a simultaneous estimation procedure can be used to estimate the AR and MA parameters, σ^2, and χ. This would involve an enormous amount of computer time. However, in practice, the Box-Cox transformation in [3.4.30] will often stabilize the variance.

7.6 APPLICATIONS

7.6.1 Introduction

Tables 5.4.1 and 5.4.2 list ARMA and ARIMA models identified for fitting to five nonseasonal stationary and three yearly nonstationary time series, respectively. Detailed identification and estimation results are presented in Sections 5.4 and 6.4, respectively, for the average annual St. Lawrence riverflows and the yearly sunspot numbers. Likewise, in this section representative output from the diagnostic check stage of model construction is given for these same two annual geophysical time series. However, the reader should keep in mind that all of the models identified in Tables 5.4.1 and 5.4.2 passed the tests for whiteness, normality and homoscedasticity given in Sections 7.3 to 7.5, respectively.

7.6.2 Yearly St. Lawrence Riverflows

Figures 2.3.1 and 5.4.1 display the average annual flows of the St. Lawrence River (Yevjevich, 1963) in m^3/s at Ogdensburg, New York, from 1860 to 1957. Identification graphs in Figures 5.4.2 to 5.4.5 indicate that a constrained AR(3) model without ϕ_2 is the most appropriate AR model to fit to this series. Parameter estimates for this model along with their SE's are given in Table 6.4.1 while [6.4.2] is the difference equation for the calibrated model. Furthermore, both the likelihood ratio test (see [6.4.1] and [7.2.1]) and the AIC (see Section 6.3) select the constrained AR(3) model for describing to the St. Lawrence flows over the AR(1) and unconstrained AR(3) models, which are also listed in Table 6.4.1.

The St. Lawrence riverflow model in [6.4.2] is now subjected to rigorous diagnostic tests to ensure that the independence, normality and constant variance assumptions are satisfied. Figure 7.6.1 shows a plot of the RACF of Section 7.3.2 for the AR(3) model without ϕ_2. The 95% confidence limits in Figure 7.6.1 have jagged edges at low lags because the more accurate technique of Section 7.3.2, that is a function of both the fitted model parameters and the lag, is used to calculate these limits. Although the value of the RACF at lag 18 is rather large, it actually lies within the 1% significance interval. This larger value could be due to inherent random variation or to the length of the time series used to estimate it. However, the important values of the RACF for the lower lags all lie well within the 95% confidence interval. Therefore, the RACF indicates that the chosen model for the St. Lawrence River satisfies the whiteness assumption. This fact is also confirmed by the χ^2 distributed Portmanteau statistic Q_L in [7.3.6] whose calculated magnitude for Q_L is 13.46 for 18 degrees of freedom and is, therefore, not significant.

The less important assumptions of normality and homoscedasticity of the residuals are also satisfied. The skewness statistic g_1 in [7.4.1] has a value of -0.1482 and a SE of 0.3046. Because g_1 is much less than 1.96SE, there is no significant skewness and this indicates that the residuals are normally distributed. Likewise, the kurtosis coefficient in [7.4.2] confirms that the residuals are Gaussian. In particular, the kurtosis coefficient, g_2, has a value of -0.3240 which is less than its SE of 0.4974.

The χ statistic from Section 7.5.2 for changes in variance depending on the current level of the series has a magnitude of 0.000081 and a SE of 0.000341, while the χ statistic for trends in the variance over time possesses a value of 0.002917 with a corresponding SE of 0.00504. Because, in both instances, the SE's are greater than the χ statistics, based upon the information used, it can be assumed that the residuals are homoscedastic.

The flows used for the St. Lawrence River are in cubic meters per second. However, if the flows had been in cubic feet per second and a model had been fit to these data, all the AR parameters and SE's would have been identical with the metric model in [6.4.2]. Only the mean level of the series and $\hat{\sigma}_a^2$ would be different. In general, no matter what units of measurement are used the AR and the MA parameter estimates and the SE's will remain the same, while the mean level and $\hat{\sigma}_a^2$ will be different.

The type of model fit to the St. Lawrence River data reflects the actual physical situation. The Great Lakes all flow into the St. Lawrence River, and due to their immense size they are capable of over-year storage. If there is an unusually wet or an unusually dry year, the Great Lakes dampen the effect of extreme precipitation on the flows of the St. Lawrence River.

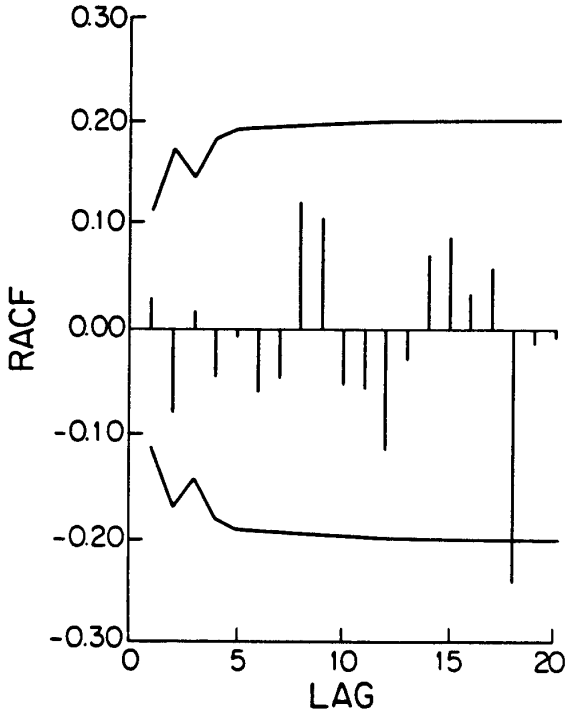

Figure 7.6.1. RACF and 95% confidence limits for the constrained AR(3) model without ϕ_2 fitted to the average annual flows of the St. Lawrence River from 1860 to 1957.

Because of this, the average annual flows are correlated, and the correct model is an AR process rather than white noise. For a general discussion of the employment of ARMA models in hydrology, the reader can refer to Section 3.6.

7.6.3 Annual Sunspot Numbers

Yearly Wolfer sunspot numbers are available from 1700 to 1960 (Waldmeier, 1961) and a plot of the series from 1770 to 1869 is shown in Figure 5.4.6. The identification graphs for this time series are presented in Figures 5.4.7 to 5.4.10. As explained in Section 5.4.3, these identification graphs in conjunction with diagnostic check output point out that an appropriate model to fit to the square roots of the sunspot series is a constrained AR(9) model without ϕ_3 to ϕ_8. In Section 6.4.3, the MAICE procedure also selects this model as the best overall ARMA model to describe the sunspot series. The finite difference equation for the best model is presented in [6.4.3] for the series of 100 sunspot values from 1770 to 1869 which is listed as Series E in Box and Jenkins (1976). In addition, the calibrated model for the entire sunspot series from 1700 to 1960 is written in [6.4.4].

The constrained AR(9) model in [6.4.4] without ϕ_3 to ϕ_8 satisfies all the modelling assumptions for the residuals. A plot of the RACF in Figure 7.6.2 shows that the residuals are uncorrelated. All of the estimated values of the RACF fall within the 5% significance interval. The χ^2 distributed portmanteau statistic Q_L in [7.3.9] has a value of 18.85 for 22 degrees of freedom. Therefore, the Q_L statistic in [7.3.6] also confirms that the residuals are not correlated. The

diagnostic checks for homoscedasticity and normality of the residuals reveal that these assumptions are also fulfilled. The model in [6.4.4], therefore, adequately models the yearly Wolfer sunspot numbers. Other types of constrained models were examined, but the AR(9) process with ϕ_3 to ϕ_8 constrained to zero is the only model that is found to be satisfactory.

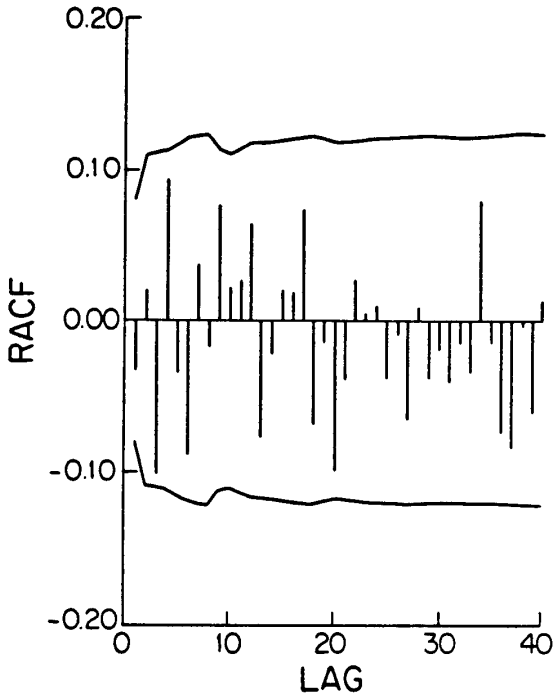

Figure 7.6.2. RACF and 95% confidence limits for the constrained AR(9) model without ϕ_3 to ϕ_8 fitted to the square roots of the yearly sunspot series from 1700 to 1960.

7.7 CONCLUSIONS

When fitting a time series model, such as an ARMA or ARIMA model, to a time series, one can follow the three stage procedure of model identification, estimation and diagnostic checking depicted in Figure III.I. The ways in which the AIC can enhance model construction are outlined in Figure 6.3.1. As explained in this and the previous two chapters, a variety of useful techniques are now available for allowing a practitioner to develop systematically and conveniently an appropriate model for describing a data set. The informative identification graphs of Section 5.3 permit a user to decide upon fairly quickly one or more tentative models to fit to the time series. These models can then be calibrated by using the method of maximum likelihood estimator presented in Appendix A6.1. When parameters for more than one model have been estimated, the AIC of Section 6.3 can be utilized to choose the overall best model. The model residuals can then be subjected to rigorous diagnostic checks to ascertain whether or not the residuals are white (Section 7.3), normally distributed (Section 7.4) and homoscedastic (Section 7.5). When the residuals are not white, then one must redesign the model by adding other parameters and, perhaps, eliminating unnecessary ones. The RACF of Section 7.3.2 is the best tool available for detecting nonwhiteness and assisting in developing a better model when the

residuals are correlated. If residual problems are caused by non-normality and/or heteroscedasticity, these can often by corrected by invoking a Box-Cox transformation from [3.4.30] and then refitting the model.

The average annual riverflows of the St. Lawrence River at Ogdensburg, New York (Yevjevich, 1963), and the yearly sunspot numbers (Waldmeier, 1961) are used throughout Part III to explain clearly how model building is executed in practice. Some model building results are also referred to in Parts II and III for the other annual time series listed in Tables 5.4.1 and 5.4.2. For the case of the St. Lawrence riverflows, model identification plots in Figures 5.4.2 to 5.4.5 efficiently identify a constrained AR(3) model without ϕ_2 as being the best model to fit to the flows. In Section 6.4.2, the MAICE procedure and the likelihood ratio test confirm this as the most appropriate model to describe the series. Finally, the choice of a constrained AR(3) is reinforced by the diagnostic checks carried out in Section 7.6.2.

When examining the yearly sunspot numbers, the identification graphs of Figures 5.4.7 to 5.4.10 do not clearly pinpoint the most suitable ARMA type model to fit to the series. Rather, the need for a square root data transformation as well as the parameters required in the model are iteratively decided upon in Section 5.4.3 by examining a range of models. The final selection is a constrained AR(9) model without ϕ_3 to ϕ_8 that is fitted to the square roots of the sunspot numbers. In Section 6.4.3, the MAICE procedure also chooses this model from many possible candidates. When the constrained AR(9) model undergoes diagnostic testing for whiteness, normality and homoscedasticity in Section 7.6.3, the results confirm that the model is adequate.

After iteratively developing a model according to the steps in Figures III.1 and 6.3.1, one can use the calibrated model for practical applications. Two important applications of time series models are forecasting and simulation, which are now described in Part IV of the book.

PROBLEMS

7.1 Select an average annual time series that is of interest to you. Following the three stages of model construction and using an available time series program such as the MH Package mentioned in Section 1.7, fit the most appropriate ARMA(p,q) model to the data set. Overspecify the fitted model by adding an additional MA or AR parameter. Estimate the parameters of the overspecified model and comment upon the size of the SE's. Employ the likelihood ratio test of [7.2.1] to ascertain if overfitting is needed to start with and also to determine if the overfitted model is better than the simpler model.

7.2 An ARMA(2,1) model is written as

$$z_t - 0.13z_{t-1} + 0.36z_{t-2} = a_t - 0.4a_{t-1}$$

where it is assumed that the mean of z_t is zero. Can this model be written in a more parsimonious fashion?

7.3 Deliberately fit an overspecified ARMA or ARIMA model to an annual time series by assuming the model is ARMA(3,4). Comment upon the size of the SE's for the parameter estimates. Try to roughly factor this model to discover parameter redundancy. Determine

the most appropriate model to fit to the time series.

7.4 Assume that one has an ARMA(1,1) model and $L = 5$ in [7.3.2] and [7.3.3]. Determine the entries of the matrix U in [7.3.3] for the distribution of the RACF.

7.5 Deliberately fit an underspecified ARMA or ARIMA model to a given annual time series. Based upon the RACF for this model, explain how the model can be expanded to provide a better fit to the series. If necessary, use other tools in your search for an improved model.

7.6 In Section 7.3.3, three versions of a Portmanteau statistic are presented for use in whiteness tests. By referring to appropriate references compare the relative advantages and drawbacks of the three statistics.

7.7 Explain why the autocorrelation function of the squared residuals is capable of detecting nonlinear statistical dependence in the residuals of fitted ARMA models.

7.8 The normality tests of Section 7.4 are described for use with the residual series from a fitted ARMA model. However, the tests can be employed with any given series such as the w_t series given in [4.3.3]. If the w_t series has a mean, then the mean should be subtracted from each w_t observation when calculating a given normality test statistic. Using a given annual series of your choice, determine if the series is Gaussian using the following tests:

(i) skewness coefficient,

(ii) kurtosis coefficient,

(iii) normality plot.

7.9 For a residual series obtained by fitting an ARMA model to a yearly time series, check for normality using the tests described in Sections 7.4.2 and 7.4.3.

7.10 Describe three additional normality tests beyond those given in Section 7.4.

7.11 A general test for homoscedasticity is described in Section 7.5.2. Assuming that one is checking for variance change over time and hence $K(t) = t$, describe in detail using equations how the test is carried out.

7.12 Select a yearly hydrological time series to model. Using a time series package, follow the three stages of model construction to ascertain the best ARMA or ARIMA model to fit to the data. Clearly explain all of your steps and show both identification and diagnostic check graphs.

7.13 In Figure 6.3.1, two main approaches are shown for using the AIC in model construction. Follow both of these approaches to find the most appropriate ARMA or ARIMA models for fitting to an annual riverflow series and also a yearly water demand series. Include both numerical and graphical results with your explanations of how you modelled the series.

REFERENCES

DATA SETS

Waldmeier, M. (1961). *The Sunspot Activity in the Years 1610-1960.* Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology paper no. 1, Colorado State University, Fort Collins, Colorado.

HOMOSCEDASTICITY TESTS

Anscombe, F. J. (1961). Examination of residuals. *In 4th Berkeley Symposium,* Berkeley, California.

Pierce, D. A. (1971). Distribution of residual autocorrelations in the regression model with autoregressive-moving average errors. *Journal of the Royal Statistical Society,* Series B, 33:140-146.

NORMALITY TESTS

Blom, G. (1958). *Statistical Estimates and Transformed Beta-Variables.* John Wiley, New York.

D'Agostino, R. B. (1970). Transformation to normality of the null distribution of g_1. *Biometrika,* 57:679-681.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics,* 17(4):111-520.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association,* 62:399-402.

Looney, S. W. and Gulledge Jr., T. R. (1985). Use of the correlation coefficient with normal probability plots. *The American Statistician,* 39(1):75-79.

Michael, J. R. (1983). The stabilized probability plot. *Biometrika,* 70(1):11-17.

Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics,* 31(2):115-124.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika,* 52:591-611.

Shapiro, S. S., Wilk, M. B. and Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association,* 63:1343-1372.

Royston, P. (1993). Graphical detection of non-normality by using Michael's Statistic. *Applied Statistics,* 42(1):153-158.

Stirling, W. D. (1982). Enhancements to aid interpretation of probability plots. *The Statistician,* 31(3):211-220.

OVERFITTING

Box, G. E. P. and Newbold, P. (1971). Some comments on a paper of Coen, Gomme and Kendall. *Journal of the Royal Statistical Society, Series A*, 2:229-240.

Granger, C. W. J. and Newbold, P. (1977). *Forecasting Economic Time Series*. Academic Press, New York.

Whittle, P. (1952). Tests of fit in time series. *Biometrika*, 39:309-318.

TIME SERIES ANALYSIS IN GENERAL

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Hipel, K. W., McLeod, A. I. and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 1, Model construction. *Water Resources Research*, 13(3):567-575.

McLeod, A. I. (1974). Contributions to applied time series. Master's thesis, Department of Statistics, University of Waterloo, Waterloo, Ontario.

McLeod, A. I. (1977). *Topics in Time Series and Econometrics*. Ph.D. Thesis, Department of Statistics, University of Waterloo, Waterloo, Ontario, Canada, 283 pp.

McLeod, A. I., Hipel, K. W. and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 2, Applications. *Water Resources Research*, 13(3):577-586.

WHITENESS TESTS

Box, G. E. P. and Pierce, D. A. (1970). Distribution of the residual autocorrelations in autoregressive integrated moving average models. *Journal of the American Statistical Association*, 65:1509-1526.

Davies, N., Triggs, C. M. and Newbold, P. (1977). Significance levels of the Box-Pierce Portmanteau statistics in finite samples. *Biometrika*, 64:517-522.

Granger, C. W. J. and Andersen, A. P. (1978). *An Introduction to Bilinear Time Series Models*. Vandenhoeck and Ruprecht, Gottingen.

Kheoh, T. S. and McLeod, A. I. (1992). Comparison of modified Portmanteau tests. *Journal of Computational Statistics and Data Analysis*, 14:99-106.

Li, W. K. and McLeod, A. I. (1981). Distribution of the residual autocorrelations in multivariate ARMA time series models. *Journal of the Royal Statistical Society*, Series B, 43(2):231-239.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65:297-303.

McLeod, A. I. (1978). On the distribution of residual autocorrelations in Box-Jenkins models. *Journal of the Royal Statistical Society*, Series B, 40(3):296-302.

McLeod, A. I. and Li, W. K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series*, 4(4):269-273.

Miller, R. B. (1979). Book review on "An introduction to bilinear time series models" by C. W. Granger and A. P. Anderson. *Journal of the American Statistical Association*, 74:927.