CHAPTER 3

MODEL SELECTION

In their book, Box and Jenkins (1970) propose a three step procedure which is and has been applied in many practical occasions, although alternative methods for model selection have also been developed. In section 3.1, I will briefly review a few of these proposals, and I will evaluate their performance with some simulation exercises. Based on their outcomes and on some philosophical arguments, I propose a new strategy in section 3.2, of which the three major steps roughly are: the determination of a set of models, the checking of these on eventual deficiencies, and finally, the selection of a model with criteria which are related to the model purpose. These three steps will be highlighted in the sections 3.3 through 3.5. The time series of the unemployment rate in the United States, 1920-1988, is chosen to illustrate several of the issues in this chapter. The observations are given in the appendix A.1. In each section however I consider some particular issue in more detail, and then other empirical series will be used for illustration.

3.1 BOX–JENKINS METHODOLOGY

The Box–Jenkins model building strategy consists of three steps, which are the identification of a tentative model, the estimation of this model, and diagnostic checking its adequacy and proposing modifications when necessary. As already stated in chapter 2, I will only consider the first and third step, and I assume for convenience that the estimation step can be carried out with the procedures in standard statistical packages like MicroTSP. In fact, this program, in its 6.53 (1989) version, is used for most of the computations in this chapter.
Identification

The identification stage typically uses the sample counterparts of the (partial) autocorrelations discussed in chapter 2, see (2.20) and below. If a model is to be identified as an MA(q) model then the ACF should show zero valued autocorrelations for \( k > q \). For the identification of an AR(p) model it applies that the PACF should show zero values for \( k > p \). If neither the ACF nor the PACF cut off after a certain \( k \), then a mixed ARMA model may be appropriate. Under several assumptions and regularity conditions, which are given in early work by Bartlett, Quenouille, R.L. Anderson or more recently in Anderson (1971) and Granger and Newbold (1977, 1986), it appears to be convenient to check whether the estimated autocorrelations are within the \( \pm 2n^{-1/2} \) region to establish their significance.

Once a tentative model has been estimated it is advised to investigate model adequacy with two tools. The first is to add extra, say one or two, terms to the model and to test for their significance. The second is to inspect the ACF of the estimated residuals given by

\[
r_s(\hat{\varepsilon}) = \frac{\sum_{t=k+1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2},
\]

(3.1)

to see whether there is any correlation left in the residuals. Given the definition of univariate time series models in section 2.1 it is clear that this is an important feature to test. Again under some assumptions it can be shown that the population equivalents \( r_s(\varepsilon) \) are uncorrelated and have variances \((n-r)/(n^2+2n)\) which is approximately \( n^{-1} \). Hence, a rough check may be to test whether the estimated autocorrelations lie within the \( \pm 2n^{-1/2} \) interval. In Box and Pierce (1970) it is shown that the \((m \times 1)\) vector \( \hat{\varepsilon} \) containing the first \( m \) \( r_s(\hat{\varepsilon}) \)'s can be written as a function of the \((m \times 1)\) vector \( \varepsilon \) containing the \( r_s(\varepsilon) \) via a matrix \( Q \) which is idempotent and has rank \( m-p-q \) in case of an ARMA\((p,q)\) model. This establishes the Box–Pierce test statistic

\[
BP(m) = n \sum_{r=1}^{m} r_s^2(\hat{\varepsilon}),
\]

(3.2)

which asymptotically follows a \( \chi^2(m-p-q) \) distribution under the null hypothesis of no residual autocorrelation provided that \( m/n \) is small and \( m \) is moderately large. This type of test statistics where the value of \( m \) is often set \( a \ priori \) and which should detect any kind of misspecification up
to \( m \) is usually called a portmanteau test, a name which should reflect its multifunctional use.

Despite its intuitive appeal, it is however found to be rather difficult in practice to make a good initial guess of a tentative model from the autocorrelation functions only. For example, consider an AR(1) with \( \phi_1=0.8 \), an AR(2) with \( \phi_1=0.6, \phi_2=0.2 \) and an ARMA(1,1) model with \( \phi_1=0.9, \beta_1=-0.4 \). The theoretical ACFs up to \( k = 3 \) are now given by (see (2.15), (2.17) and (2.25)), \{0.80, 0.64, 0.51\}, \{0.75, 0.65, 0.54\} and \{0.73, 0.66, 0.59\}, respectively. It is clear that theoretically it might be difficult to distinguish between these models, let it be the empirical case in which the ACF has to be estimated.

Several suggestions have been made to facilitate the initial guess. For example, in Beguin, Gourioux and Monfort (1980) it is proposed to identify an ARMA(\( p,q \)) model with the so-called corner method. This method uses the matrix of estimated autocorrelations, of which it checks whether certain determinants equal zero. The numbers of the row and column for which this property ceases to be valid yields an indication of the model order. It is however readily imaginable that again identification problems may arise for models which generate similar ACFs. Another suggestion which is often applied in practice is to select the model which minimizes, e.g., a criterion like

\[
C_{p,q} = \log \hat{\sigma}_e^2 + g(n,p,q),
\]

(3.3)

where \( \hat{\sigma}_e^2 \) is the estimated variance of the estimated residuals and where \( g(n,p,q) \) denotes a scalar function of \( n \) and the numbers of parameters. In case this function is \( 2(p+q)/n \) the \( C_{p,q} \) reflects the Akaike criterion, AIC, (Akaike 1974), and when it is \( (p+q)\log n/n \) the expression in (3.3) reduces to the Schwarz criterion, SC (Schwarz 1978). In section 3.5, I will treat these and many other criteria in more detail.

**Hannan–Rissanen approach**

An approach in time series model identification propagated in Granger and Newbold (1986) which uses criteria as in (3.3) is developed in Hannan and Rissanen (1982), to be denoted as HR. Consider the mixed ARMA(\( p,q \)) model as in (2.9). The first step of the HR approach consists of selecting an appropriate autoregressive model of order \( K \) for \( y_t \), where \( K \) does not exceed
some *a priori* assigned value $K^*$. The order is found by taking the value of $K$ which minimizes the AIC criterion. The second step uses the residuals, say $u_i$, from this AR($K$) model to estimate by OLS

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_{l-1} y_{t-l} + \epsilon_t + \theta_1 u_{t-1} + \ldots + \theta_m u_{t-m},$$

(3.4)

where for convenience $l \leq K$, and where $m$ also does not exceed some prescribed value, say $m^*$. The order $(l,m)$ of this model is now established by minimizing the Schwarz criterion, and these $l$ and $m$ are the estimates for the orders $p$ and $q$.

From the simulation experiments in HR it appears that the method is rather successful in case the $l$ is set equal to $m$ and when the number of observations is large, e.g., 500. It also appears that the orders can be overestimated. Therefore, in Hannan and Kavalieris (1984) (HK) a third step has been added. Denote $(l,m)$ as $(l_0,m_0)$, where the subscript reflects the iteration phase. The additional step consists of estimating model (3.4) again, but now taking the $u_t$ to be the new $u_t$ resulting in $(l_2,m_2)$. This step can be repeated until $(l_n,m_n)$ converges to some fixed value, although often $i=3$ may suffice. Simulation experiments in HK support the conclusion formulated in HR.

The major argument in Hannan and Rissanen (1982) for equalizing the $l$ and $m$ is that in engineering applications the $p$ may be often equal to $q$ in ARMA($p,q$) models. For economic time series it is however uncertain whether this applies too. It seems therefore reasonable to investigate the performance of the HR strategy when $l$ and $m$ can take different values. An additional reason for such an investigation is that assuming $p$ equal to $q$ can only reduce the number of candidate models. For example, when one finds that $l=m=1$, one might specify an AR(1), an MA(1), or an ARMA(1,1) model. For higher order AR models, this argument becomes even more important.

The Monte Carlo simulation results in tables 3.1 and 3.2 are based on 100 replications. The first 50 observations are always discarded to reduce possible starting-up effects. The simulations are done for $n$ is 50, 100 and 200, because such numbers of observations may appear in economics more than those of length 500. Furthermore, the orders $l$ and $m$ may be too high in the HR steps since the AIC criterion can permit too large an AR model. The cells in the tables report the number of times the DGP is selected. However in the case where it is assumed that $p=q$, or where $l$ is set equal to $m$, the values in the cells are the number of times the appropriate order is chosen.
Table 3.1  
Evaluation of the Hannan and Rissanen procedure based on 100 replications  
DGP is ARMA(1,1): $y_t = \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$ \(^{(1)}\)

<table>
<thead>
<tr>
<th>Method(^{(2)}) Criteria(^{(3)})</th>
<th>n</th>
<th>{.5,0}</th>
<th>{.9,0}</th>
<th>{0.5}</th>
<th>{0.9}</th>
<th>{.6,-.4}</th>
<th>{.6,.4}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HR, (p=q)</strong> AIC–SC</td>
<td>50</td>
<td>99</td>
<td>97</td>
<td>92</td>
<td>80</td>
<td>100</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>99</td>
<td>98</td>
<td>94</td>
<td>81</td>
<td>100</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>100</td>
<td>100</td>
<td>97</td>
<td>81</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td><strong>SC–SC</strong></td>
<td>50</td>
<td>99</td>
<td>97</td>
<td>95</td>
<td>86</td>
<td>99</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>99</td>
<td>99</td>
<td>95</td>
<td>79</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>100</td>
<td>100</td>
<td>97</td>
<td>82</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td><strong>HR, (\forall p, q)</strong> AIC–SC</td>
<td>50</td>
<td>69</td>
<td>89</td>
<td>65</td>
<td>57</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>83</td>
<td>95</td>
<td>70</td>
<td>69</td>
<td>3</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>95</td>
<td>98</td>
<td>84</td>
<td>74</td>
<td>6</td>
<td>70</td>
</tr>
<tr>
<td><strong>SC–SC</strong></td>
<td>50</td>
<td>67</td>
<td>91</td>
<td>66</td>
<td>66</td>
<td>3</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>82</td>
<td>96</td>
<td>72</td>
<td>72</td>
<td>2</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>94</td>
<td>98</td>
<td>85</td>
<td>85</td>
<td>5</td>
<td>74</td>
</tr>
<tr>
<td><strong>HK, (\forall p, q)</strong> AIC–SC</td>
<td>50</td>
<td>63</td>
<td>91</td>
<td>59</td>
<td>61</td>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>83</td>
<td>94</td>
<td>77</td>
<td>71</td>
<td>3</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>95</td>
<td>99</td>
<td>85</td>
<td>79</td>
<td>5</td>
<td>78</td>
</tr>
<tr>
<td><strong>SC–SC</strong></td>
<td>50</td>
<td>63</td>
<td>91</td>
<td>64</td>
<td>61</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>83</td>
<td>94</td>
<td>76</td>
<td>67</td>
<td>3</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>94</td>
<td>99</td>
<td>86</td>
<td>80</td>
<td>6</td>
<td>77</td>
</tr>
</tbody>
</table>

\(^{(1)}\) The values in the cells are the number of times the DGP is selected. When \(p=q\), these values correspond to the selection of the appropriate order. The DGP is represented by \(\{\phi_1, \theta_1\}\), \(\varepsilon_t \sim N(0,1)\).

\(^{(2)}\) The selection procedures are the Hannan and Rissanen (1982) method when \(p\) is set equal to \(q\), the same method when \(p\) and \(q\) can take several values, and the Hannan and Kavalieris (1984) modification, for all \(p\) and \(q\).

\(^{(3)}\) The selection criteria in the first and second step. The AIC and SC refer to the Akaike and Schwarz criteria.

So, the HR strategy will therefore also be considered when the SC criterion is used in the first step. The HR strategy has already been evaluated via simulations in Newbold and Bos (1982). This study however does not incorporate the HK step and the SC–SC variant. Since it sometimes also considers other parameter sets in the DGP, the present tables can be viewed as being complementary to those in Newbold and Bos (1982).
Table 3.2  
Evaluation of the Hannan and Rissanen procedure, 100 replications  
DGP is ARMA(2,2): $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$ \(^{(1)}\)

<table>
<thead>
<tr>
<th>Method (^{(2)})</th>
<th>Criteria (^{(3)})</th>
<th>n</th>
<th>(.9,-.4,0,0)</th>
<th>(1.2,-.6,0,0)</th>
<th>(0.0,-.8,3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR, (p=q)</td>
<td>AIC-SC</td>
<td>50</td>
<td>35</td>
<td>78</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>48</td>
<td>88</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>55</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>SC-SC</td>
<td></td>
<td>50</td>
<td>30</td>
<td>76</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>47</td>
<td>89</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>53</td>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td>HR, (\forall p,q)</td>
<td>AIC-SC</td>
<td>50</td>
<td>58</td>
<td>83</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>79</td>
<td>94</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>86</td>
<td>96</td>
<td>26</td>
</tr>
<tr>
<td>SC-SC</td>
<td></td>
<td>50</td>
<td>54</td>
<td>84</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>77</td>
<td>93</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>84</td>
<td>98</td>
<td>25</td>
</tr>
<tr>
<td>HK, (\forall p,q)</td>
<td>AIC-SC</td>
<td>50</td>
<td>49</td>
<td>72</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>75</td>
<td>93</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>86</td>
<td>96</td>
<td>25</td>
</tr>
<tr>
<td>SC-SC</td>
<td></td>
<td>50</td>
<td>45</td>
<td>72</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>74</td>
<td>93</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>85</td>
<td>98</td>
<td>25</td>
</tr>
</tbody>
</table>

\(^{(1)(2)(3)}\) See below table 3.1.

The general conclusion from the results of table 3.1 is that, in case of ARMA(1,1) models, the restriction \(l=m\) seems to be indispensable for success. In case this restriction is dropped it can be seen that one will still be able to detect a true AR(1) model, especially as sample size increases. This result carries over to a true MA(1) model, albeit to a lesser extent. For the ARMA(1,1) model however a deterioration can be witnessed, such that there may be combinations of parameters for which the data generating process will generally not be found. From table 3.2, where the DGP is an AR(2) or an MA(2) model, some other conclusions can be drawn. The AR(2) or its order can be found, but it seems that the \(l=m\) restriction may now even prevent correct model selection. On the other hand, the MA(2) model or its order faces the risk of never being detected. Combining the two tables it can be seen that the HK modification and the SC-SC variant do not yield large, if any, improvements. Furthermore, it emerges that model selection may heavily depend on the data generating process. This may be
explained by the fact that models can in some sense be 'close' to each other, such that it may be difficult to select between rival models, see, e.g., Sneek (1984). The above evaluation of the Hannan and Rissanen (1982) procedure seems to indicate that in case of simple AR models it may work reasonably well, but that in other instances its performance may not yield an appropriate initial model selection.

Recently a modification of the HR method has been proposed in Pukkila, Koreisha and Kallinen (1990). This so-called residual white noise AR order determination method modifies the estimation step in (3.4) by including an unobserved component, which causes that Generalized Least Squares may now be useful as an estimation method. The simulation results reported show that this method performs better that the HR method indeed, and hence may be more appropriate in guiding an initial model selection.

Diagnostic checking

Suppose for the moment that one is able to specify and estimate a tentative initial model. The third step in the Box-Jenkins approach is now to check its adequacy and eventually modify this model. One check is given by the BP(m) test statistic as in (3.2). The drawback of such a portmanteau test is however that it may be able to detect whether the model is inadequate, but that it may not be useful in indicating how the model should be modified when necessary.

An alternative is to consider the nested hypotheses tests developed in e.g., Godfrey (1979). In these tests procedures one nests the current model in a more general model and tests for the validity of the restrictions. It turns out that with the Lagrange Multiplier (LM) principle these tests are relatively easy to calculate. One convenience of these LM type tests is e.g., that in case of an AR(p) model, the test against an AR(p+r) or an ARMA(p,r) model is the same. This may however also be inconvenient since this can suggest two different routes for modification. Furthermore, in some occasions it can be shown that the LM and BP tests are equivalent, see Newbold (1980). A related suggestion in case of an inadequate model is to fit such a general model straight away and test for eventual parameter restrictions. This may prove to be tedious because the addition of extra terms to a correct model may lead to indeterminacy. It is then possible that multicollinearity causes that the point estimates and the estimated standard deviations are not useful (cf. Granger and Newbold 1977, pp. 91,
110). This implies that a simple to general strategy, as well as a general to simple strategy, may become hard to apply since modification can be carried out into the wrong direction and suboptimal models can be found.

3.2 AN ALTERNATIVE STRATEGY

There are several additional arguments for proposing an alternative model selection strategy. One of these is that it has regularly been found that models giving a good fit within the sample do not necessarily forecast well out of sample, see, e.g., Fildes (1979). It seems therefore natural to incorporate the purpose for which the model is initially built in the model selection process. More precisely, one may want to compare models built for forecasting purposes and those built for other purposes with distinct criteria. An example of such another purpose is the construction of a time series model as the first step in relating two or more variables in a dynamic econometric model, see, e.g., the transfer function analysis in Box and Jenkins (1970) and the prewhitening approach in Haugh and Box (1977), and the empirical modeling strategy in Davidson et al. (1978).

Another argument may be given by the apparent influence of the true, but unknown, data generating process on model selection, see also tables 3.1 and 3.2. Formal discussions of this phenomenon can be found in Sneek (1984), where it is argued that the 'distance' between models may determine time series model selection, and Franses (1989a), where similar results are established in the context of a linear regression model. To highlight this interesting issue, several simulations have been carried out. Only the salient results are displayed in the tables 3.3 and 3.4. The first table considers the choice between an AR(1) and an AR(2) model while an AR(3) model is the data generating process. The choice is based on both the AIC and the SC criteria. Furthermore, the number of times both incorrect models are rejected with a modified BF test, i.e. the LB test, is also reported. This LB test proposed in Ljung and Box (1978) follows a \( \chi^2(m-p-q) \) distribution under the null hypothesis of no residual autocorrelation, and it is given by

\[
LB(m) = n(n+2) \sum_{r=1}^{m} (n-r)^{-1} \hat{r}_r^2(\hat{e}).
\]

(3.5)

It corrects for the approximation of \( (n-r)/(n^2-2n) \) by \( n^{-1} \) as is done in the BF test. In Davies, Triggs and Newbold (1977) it has been shown that
especially in small samples this approximation can be rather crude. The simulations in tables 3.3 and 3.4 are based on 100 replications. The $\varepsilon_t$ have been generated from a standard normal distribution.

Table 3.3
Selection between AR(1) and AR(2) while AR(3):
$y_t=\phi_1 y_{t-1}+\phi_2 y_{t-2}+\phi_3 y_{t-3}+\varepsilon_t$ is DGP, $n=100$, 100 replications

<table>
<thead>
<tr>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$\phi_3$</th>
<th>AIC</th>
<th>SC</th>
<th>LR($^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>-0.3</td>
<td>81</td>
<td>95</td>
<td>100</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>-0.2</td>
<td>69</td>
<td>85</td>
<td>99</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>-0.1</td>
<td>59</td>
<td>85</td>
<td>94</td>
</tr>
<tr>
<td>0.7</td>
<td>0.3</td>
<td>-0.3</td>
<td>65</td>
<td>85</td>
<td>99</td>
</tr>
<tr>
<td>0.7</td>
<td>0.4</td>
<td>-0.3</td>
<td>28</td>
<td>50</td>
<td>96</td>
</tr>
<tr>
<td>0.7</td>
<td>0.5</td>
<td>-0.3</td>
<td>9</td>
<td>21</td>
<td>98</td>
</tr>
</tbody>
</table>

($^1$) The model is selected with Akaike’s information criterion, AIC, or with the Schwarz criterion, SC.

($^2$) The number of times that both models are rejected with the Ljung-Box test for residual autocorrelation.

From table 3.3 it can be seen that the selection of an AR(1) model may heavily depend on the parameter of the omitted AR(3) term, as well as on the value of the parameter of the included AR(2) term. On the other hand, it appears that the rejection of both incorrect models does not seem to be much affected by these parameter values.

In table 3.4 additional results are reported for the choice between an AR(1), AR(2) and an AR(3) model in case an AR(4) model is the data generating process. In constrast to the previous table, I also display the results for varying series lengths.
Table 3.4
Selection between AR(1), AR(2) and AR(3) while AR(4):
y_t=0.8y_{t-1}+\phi_2y_{t-2}-0.2y_{t-3}+0.4y_{t-4}+\epsilon_t$ is DGP, 100 replications

<table>
<thead>
<tr>
<th>n</th>
<th>$\phi_2$</th>
<th>AIC</th>
<th>SC</th>
<th>AIC</th>
<th>SC</th>
<th>LB(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>-0.30</td>
<td>5</td>
<td>19</td>
<td>58</td>
<td>65</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>-0.25</td>
<td>18</td>
<td>38</td>
<td>54</td>
<td>54</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>-0.20</td>
<td>30</td>
<td>44</td>
<td>52</td>
<td>47</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>-0.16</td>
<td>42</td>
<td>60</td>
<td>43</td>
<td>35</td>
<td>99</td>
</tr>
<tr>
<td>100</td>
<td>-0.30</td>
<td>5</td>
<td>13</td>
<td>57</td>
<td>68</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>-0.25</td>
<td>3</td>
<td>16</td>
<td>47</td>
<td>61</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>-0.20</td>
<td>20</td>
<td>40</td>
<td>44</td>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>-0.16</td>
<td>28</td>
<td>56</td>
<td>38</td>
<td>32</td>
<td>100</td>
</tr>
<tr>
<td>200</td>
<td>-0.30</td>
<td>0</td>
<td>1</td>
<td>30</td>
<td>58</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>-0.25</td>
<td>0</td>
<td>4</td>
<td>43</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>-0.20</td>
<td>6</td>
<td>25</td>
<td>37</td>
<td>46</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>-0.16</td>
<td>20</td>
<td>51</td>
<td>29</td>
<td>27</td>
<td>100</td>
</tr>
</tbody>
</table>

(1) See table 3.3.

From table 3.4 it appears that a larger number of observations causes that model selection moves away from the entirely incorrect AR(1) model to a larger model. Combining the results in both tables leads to the general impression that the unknown data generating process can influence model selection but that the rejection of incorrect models seems to be only weakly affected by changes in these true models. Although the results in tables 3.3 and 3.4 may seem to be somewhat exaggerated or far-fetched, I feel that an implication of the outcomes is that a more natural sequence of steps in model building seems to be to test models first and then to select those which pass these checks.

There are many reasons to start with a set of models. First, there are the problems discussed above with the identification of one model either from the autocorrelation functions or from more sophisticated selection strategies. Second, in practical circumstances there are reasons to believe that there may not be a true model, that if such a model would exist one might not be able to detect it, or that a certain model probably does not hold for all future times. Hence, it seems sensible to entertain a set of
tentative models which is considered again any time new observations become available. See Lempers (1971) and Marriott and Tremayne (1988) for similar ideas of considering a portfolio of models.

The models in the set can be limited to a manageable number by an inspection of the autocorrelation functions and of the data themselves. Given that the interpretation of these functions is facilitated under the assumption of a stationary time series, it is suggested to transform the raw data series to get rid of nonstationary behavior. This interpretation of the (partial) ACF is one reason for not considering models for transformed and untransformed series in the same set. An additional reason is that it is unclear how the diagnostic measures and the corresponding asymptotic distributions behave when the time series is nonstationary and not transformed. Furthermore, suppose one wants to select between

\[ \Delta y_t = \rho y_{t-1} + \psi_1 \Delta y_{t-1} + \epsilon_t \]  
(3.6)

and

\[ \Delta y_t = \psi_1 \Delta y_{t-1} + \epsilon_t \]  
(3.7)

with the AIC or the SC criteria, instead of the Dickey-Fuller procedure in (2.12). Since these criteria can be written as \( F \) statistics with other critical values, see \textit{inter alia} Amemiya (1980) and Franses (1989a), it is clear that they are related to Dickey-Fuller's \( \hat{\rho} \) statistic. Unreported simulation results show that with, e.g., the SC criterion in over 50 of the 100 replications model (3.6) is selected in case (3.7) is the DGP. Translated to an application of the \( \hat{\rho} \) test, this means that the empirical size of the test is about 0.5 instead of 0.05. All together, this suggests that it may be more appropriate to start with the application of a Dickey-Fuller test, and to use transformed series if necessary. Given that it is unclear which model is most appropriate, it may then be sensible to check for nonstationarity in a range of models.

In a second step, all models are submitted to a battery of diagnostic tests, which are designed to check for basic requirements. These are given by the assumptions underlying the considered models, such as white noise error processes, constancy of parameters, linearity, and absence of outliers. Some of the models may fail to fulfil these requirements, and hence will not be considered in the next stage of model selection. When new observations become available these models should not be automatically
excluded from an initial set, although repeated failure may indicate their persistent inadequacy. It may of course also occur that the set of adequate models is empty. A natural step is then to extend the model class to, e.g., nonlinear time series models or models with interventions.

In the model selection stage the remaining models are compared in a symmetrical way. This means that no model is considered to be the null hypothesis which should be defended, but that the models are evaluated by their description or forecasting performance as compared to the amount of parameters they incorporate. The purpose of the model determines the use of selection criteria. Some criteria are designed for forecasting models, while others originate from the idea that a model should fit in some sense close to the available data. It seems therefore natural to use different criteria for separate modeling purposes.

In conclusion, this alternative strategy differs in several respects from the conventional Box-Jenkins strategy. First, a set of models is constructed instead of only one model. Second, the models are checked on more features than only on eventual residual autocorrelation. Third, although the models are screened by these checks, the final symmetrical model selection is not based on testing outcomes, but it is left to criteria related to model use. In the subsequent sections of this chapter, I will treat each of these steps in more detail.

3.3 A SET OF MODELS

Before any formal modeling is started, it may be insightful to look at the graph of a series and also to gain some knowledge of the underlying economic process. A clarifying example may be given by the series depicted in figure 3.1 which are the stock, $S_t$, and sales, $Q_t$, of cars in the Netherlands. It is evident that a relationship between these series can be established by the fact that $S_t$ equals $S_{t-1}$ plus $Q_t$ and minus all cars scrapped between $t$ and $t-1$. When one is interested in fitting a dynamic model to $S_t$, it may be sensible to recall this relationship before any model is fitted. From figure 3.1 it can be seen that a univariate time series model for $S_t$ is likely to suffer from a structural break in 1981. This break is induced by the temporary decrease in new car sales because of the second oil crisis. One remedy may be to model $S_t$ with an explicit break, but a more obvious option is to consider $S_t$ only in relation to $Q_t$.
Figure 3.1
Stock, $S$, and sales, $Q$, of cars in the Netherlands, 1960–1989 (x1000)

In Franses (1991b) a detailed discussion of the construction of empirical bivariate models for $S_t$ and $Q_t$ is presented. From these results it emerges that some of these bivariate models do have constant parameters indeed.

Stationarity

A nonstationary time series can be recognized by having a mean which is not constant and an increasing conditional empirical variance anytime a new observation becomes available (cf. section 2.1). For example, the series $S_t$ clearly shows such behavior in the considered period. Two simple models for nonstationary series are the so-called trend stationary model.
\[ y_t = \mu + \tau T_t + \varepsilon_t, \]  
(3.8)

where \( T_t \) represents a deterministic trend, and the \( I(1) \) process

\[ y_t = \tau + y_{t-1} + \varepsilon_t. \]  
(3.9)

where \( \varepsilon_t \) is white noise. With respect to figure 3.1 it can be argued that, at least at first sight, the \( S_t \) and \( Q_t \) series may be representable by models like (3.8) and (3.9), respectively. The \( Q_t \) series can then be first order differenced to achieve stationarity.

![Graph showing two lines labeled RURS and RURQ](image)

**Figure 3.2 Recursive testing for unit roots**

To investigate whether there are unit roots in the models for the series \( S_t \) and \( Q_t \) with the Dickey-Fuller procedure as described in (2.12), it seems
suitable to consider an alternative hypothesis which contains a
deterministic trend. It may however be, and this may occur especially in
economic time series, that there has been a change in the slope of the
trend, see e.g., $S_t$ is figure 3.1. In that case one may want to model this
change explicitly and to incorporate it in the unit root tests, see
Perron (1989). An obvious issue is now the detection of the turning-point.
In these circumstances it may be useful to consider recursive testing for
unit roots, see also Hendry and Mizon (1990). This means that one
calculates the relevant statistic for the first $n_t$ observations, then one
takes the first $n_t+1$ observations, and so on.

The results of this recursive procedure for the statistic $r_t$ and for
the series $S_t$ and $Q_t$ are displayed in figure 3.2. For these series the $p$ in
(2.12) is set equal to 1 and 0, respectively. The effect of the structural
break in 1981 becomes even more evident from this figure. The value of the
test statistic for $S_t$ is steadily decreasing until 1980, hence indicating
that a model like (3.8) might have been appropriate indeed when the break
in 1981 would not have occurred. The $r_t$ values for $Q_t$ seem to indicate that
the new car sales series may contain a unit root as in (3.9), although the
effect of a break can also be noticed.

Unit roots in MA processes

The first order differencing filter $\Delta_t$ is applied to many economic time
series for it is argued that they often follow nonstationary patterns. The
basis for these arguments is usually provided by evidence from unit root
tests, or from the observation that the ACF does not die out rapidly, which
according to Box and Jenkins (1970) is a sign of nonstationary behavior.
These arguments are however not entirely beyond discussion, because one may
easily construct such ACF patterns by suitable choices of ARMA parameters.
Furthermore, the power of the unit root tests appears to be rather low, see
e.g., Dickey, Bell and Miller (1986).

One indication that the differencing filter is inappropriately used
can be given by inspection of the ACF of $\Delta_t y_t$. In the extreme case that a
white noise variable has been differenced, one would theoretically find a
value of $-0.5$ for the estimated first order autocorrelation. Consider

$$y_t = \epsilon_t$$
and apply the filter $\Delta$ to both sides, then

$$\Delta y_t = e_t - e_{t-1},$$

emerges. Applying the expression in (2.21) one can derive that for an MA(1) process

$$x_t = e_t + \theta_1 e_{t-1},$$

(3.10)

it holds that

$$\rho_1 = \theta_1/(1+\theta_1^2).$$

(3.11)

When $\theta_1$ equals $-1$, this $\rho_1$ becomes equal to $-0.5$.

One approach to test whether the parameter $\theta_1$ in (3.10) is equal to $-1$ is to investigate whether the estimated moving average parameter $\hat{\theta}_1$ equals $-1$. This estimate is however downward biased under the null hypothesis, and hence its usefulness may be disputable, see Plosser and Schwert (1977). The test procedure developed in Saikkonen and Luukkonen (1990) requires the estimation of an ARMA model. A further drawback of that procedure is, that when (3.10) contains an unknown mean $\mu$, things get more complicated. On the other hand, the empirical size and power properties of their test statistics appear to be reasonably satisfactory.

An alternative procedure, proposed in Franses and Kloek (1990), for testing for noninvertibility of MA processes is to base inference on functions of the sample autocorrelations of an MA process when these are estimated as in (2.20). This approach is particularly useful since it can be shown that for all noninvertible MA processes it holds that the sum of their theoretical autocorrelations at positive lags is equal to $-0.5$. Consider for example the MA(2) model

$$x_t = (1-B)(1+\theta_1 B) e_t,$$

(3.12)

for which it holds that

$$\rho_1 = -(1+2\theta_2 - \theta_1^2)/(2-2\theta_1^2 + 2\theta_2^2),$$

$$\rho_2 = -\theta_1/(2-2\theta_1^2 + 2\theta_2^2),$$
and hence that $\rho_1 + \rho_2 = -0.5$.

The distributional results for sample autocorrelations for moving average processes, given and proved in Anderson and Walker (1964), may now be useful. Consider $n$ observations on the zero mean linear process

$$w_t = \sum_{i=-\infty}^{+\infty} \eta_i e_{t-i}, \quad t = 0, \pm 1, \pm 2, \ldots \quad (3.13)$$

where $\sum_{i=-\infty}^{+\infty} |\eta_i| < \infty$ and $\sum_{i=-\infty}^{+\infty} |i\eta_i|^2 < \infty$. It can be shown that $n^{1/2}(r_k - \rho_k)$, with $1 \leq k \leq s$, asymptotically follows an $s$-variate normal distribution with mean zero and with covariances given by

$$\text{ncov}(r_k, r_l) = \sum_{j=-\infty}^{+\infty} (\rho_j \rho_{j+k-l} + \rho_j \rho_{j+k+l} + 2\rho_k \rho_{j+l} - 2\rho_k \rho_{j-k} - 2\rho_k \rho_{j+k}). \quad (3.14)$$

Note that the restrictions for $\eta_i$ apply in the MA cases considered here.

For ease of exposition, consider an MA(2) process for which the only nonzero autocorrelations are those at lags 0, ±1, and ±2, where $\rho_{\pm 1} = \rho_i$ for $i = 1, 2$. Application of (3.14) results, after some straightforward algebra, in

$$n^{1/2} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \sim N \left( \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right) \quad (3.15)$$

where

$$A_{11} = 1 + 2\rho_2 + 4\rho_1^2 - 3\rho_2^2 + 2\rho_1^2 + 4\rho_1^2 \rho_2^2 - 8\rho_2^2 \rho_2$$
$$A_{12} = A_{21} = 2\rho_1 - 2\rho_1^2 \rho_2 + 4\rho_1^3 \rho_2 - 4\rho_1^2 \rho_2^2 + 2\rho_1^3 - 4\rho_1 \rho_2^2$$
$$A_{22} = 1 + 2\rho_1^2 - 3\rho_2^2 + 4\rho_1^4 + 4\rho_1^3 \rho_2^2 - 4\rho_1^3 \rho_2.$$ 

This expression can be used to construct a test for the hypothesis $H_{un}$ that $y_t$ is a white noise process, or that

$$\Delta y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}, \quad (3.16)$$

is noninvertible since $\theta_1 = -1$. In that case it holds true that $\rho_1 = -0.5$ and $\rho_2 = 0$, so $A_{11}$ reduces to 0.5. The test statistic $T_{un}$ is then simply
given by

\[ T_{nm} = (2n)^\frac{1}{2}(r_1+0.5), \]  

which asymptotically follows a standard normal distribution under the \( H_{nm} \) null hypothesis. To see whether this statistic avoids the boundary problem as it is discussed in Plosser and Schwert (1977), consider figure 3.3 in which the empirical distribution of the variable \( r_1+0.5 \) is depicted when 200 times 100 observations are generated for the process \( x_t = \varepsilon_t - \varepsilon_{t-1} \) and the \( r_1 \) is calculated using (2.20). It is clear that it reflects a distribution which seems to come close to a normal distribution with mean zero.

![Histogram](image)

**Figure 3.3** Empirical distribution of \( r_1+0.5 \) when DGP is \( x_t = \varepsilon_t - \varepsilon_{t-1} \)

This figure also suggests that the \( T_{nm} \) test statistic may better be used in
a double-sided test.

The hypothesis $H_{n=1}$ is that $y_t$ follows an invertible MA(1) model, or that the model

$$\Delta \gamma_t = (1+\theta_1 B)(1+\theta_2 B)\epsilon_t,$$

(3.18)

is noninvertible since $\theta_1 = -1$, can be tested in an analogous manner. Under this null hypothesis it can be shown that $n^{1/2}(r_1 + r_2 - 0.5)$ asymptotically follows a univariate normal distribution with mean zero and variance $A_{11} + 2A_{12} + A_{22}$, which, when substituting $\rho_1 = -\rho_2 - 0.5$, can be written as $1 + 2\rho_1 + 2\rho_1^2$. Note that this variance is smallest, i.e. 0.5, in case $\rho_1 = -0.5$.

| Table 3.5 |
| Estimation results for some monthly U.S. macroeconomic series |

<table>
<thead>
<tr>
<th></th>
<th>MBASE</th>
<th>WAGE</th>
<th>LAB</th>
<th>EMP</th>
<th>SP500</th>
<th>P/E</th>
<th>D/P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>466</td>
<td>466</td>
<td>454</td>
<td>454</td>
<td>466</td>
<td>382</td>
<td>466</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lag</th>
<th>Autocorrelations of the deviations of first differences of logarithms from a time trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.09 0.07 -0.30 -0.16 0.02 -0.08 0.03</td>
</tr>
<tr>
<td>2</td>
<td>0.01 0.11 0.02 0.13 -0.04 0.05 -0.05</td>
</tr>
<tr>
<td>3</td>
<td>0.10 0.14 -0.05 0.05 0.03 0.08 0.07</td>
</tr>
<tr>
<td>4</td>
<td>0.02 0.07 0.00 0.10 0.08 0.04 0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lag</th>
<th>Autocorrelations of the second differences of series</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.54 -0.52 -0.62 -0.63 -0.47 -0.56 -0.46</td>
</tr>
<tr>
<td>2</td>
<td>0.00 0.00 0.16 0.16 -0.06 0.05 -0.10</td>
</tr>
<tr>
<td>3</td>
<td>0.08 0.06 -0.06 -0.06 0.01 0.03 -0.10</td>
</tr>
<tr>
<td>4</td>
<td>-0.05 -0.07 0.04 0.04 0.01 -0.05 -0.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimate of the MA parameter in MA(1,1) model for growth rates (standard error in parentheses)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA(1) (0.16)</td>
</tr>
</tbody>
</table>

Estimating $\rho_1$ by $r_1$ gives the test statistic for $H_{m1}$, or

$$T_{m1} = n^{1/2}(r_1 r_2 + 0.5)/(1 + 2r_1 r_2^{1/2}), \quad (3.19)$$

which asymptotically follows a standard normal distribution under its null hypothesis. Summarizing, small values of $T_{um}$ and $T_{m1}$ imply that the null hypotheses of overdifferencing in two special cases cannot be rejected.

To illustrate this test procedure, consider a subset of the variables from Schwert (1987). A conclusion in that article is that the growth rates of many US macroeconomic variables can be described by an IMA process of order $(1,1)$, or that one should doubly difference variables $y_t$ to reach stationarity. In the first few blocks of table 3.5 the time series properties of some of the series are given. The sample autocorrelations of the deviations of the first differences of logarithms from a time trend of $MBASE$, $WAGE$, $SP500$, $P/E$ and $D/P$ suggest that these growth rates may follow white noise processes, while those of $LAB$ and $EMP$ indicate that an MA(1) process may be appropriate. The sample autocorrelations of the second differences of the series suggest that all 7 series are overdifferenced in case the $\Delta_1$-filter is applied to the growth rates. The estimates of the first order MA parameter in table 3.5 may provide some additional evidence.

Table 3.6
Testing for noninvertibility of IMA(1,1) models for growth rates

<table>
<thead>
<tr>
<th></th>
<th>$MBASE$</th>
<th>$WAGE$</th>
<th>$LAB$</th>
<th>$EMP$</th>
<th>$SP500$</th>
<th>$P/E$</th>
<th>$D/P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test$^{(1)}$</td>
<td>$T_{um}$</td>
<td>$T_{um}$</td>
<td>$T_{m1}$</td>
<td>$T_{um}$</td>
<td>$T_{um}$</td>
<td>$T_{um}$</td>
<td>$T_{um}$</td>
</tr>
<tr>
<td>Value</td>
<td>-1.22</td>
<td>-0.61</td>
<td>1.17</td>
<td>0.88</td>
<td>-0.87</td>
<td>-1.66*</td>
<td>1.22</td>
</tr>
</tbody>
</table>

* Significant at a 10% level.

$^{(1)}$ The statistic $T_{um}$ is given in (3.17) and $T_{m1}$ in (3.19).

The hypotheses that the growth rates themselves may already be modeled using a white noise or an MA(1) model are now tested with the statistics in (3.17) and (3.19). The results are displayed in table 3.6. The values of the test statistics $T_{um}$ or $T_{m1}$ are insignificant at 5%. Hence, the null
hypotheses that some series can be modeled with a white noise process or an MA(1) process is not rejected, which clearly confirms the suggestions from table 3.5 that the IMA(1,1) model for some growth rates are misspecified.

Unit roots in univariate series, not in multivariate series

One drawback of the Dickey-Fuller procedure in (2.12) is that it may suffer from low power, which implies that the null hypothesis of a unit root is accepted too frequently. This conclusion, drawn from simulated evidence, holds for the case in which one is able to determine the exact order of the process. In practice this order is often unknown and has to be estimated. It is readily imaginable that too large or too small an order can establish serious size and power distortions for the test procedure. So, one may argue that the small rejection rate of the null hypothesis, see for example the results in Nelson and Plosser (1982) where almost all time series seem to have a unit root, can be due to an incorrect model. For example, a more suitable model could have been an ARMA model (cf. Choi 1990). Another explanation may be that one considers the time series each in a univariate context only, while it may be more appropriate to consider some of the series in a multivariate system. It is conceivable, and in practice one sometimes may encounter this phenomenon, that a multivariate time series model appears to be stationary, while some of its component series are not. It seems natural to start from a stationary multivariate model and to derive the conditions for a seemingly nonstationary univariate series. This strategy appears to yield simply interpretable results and I discuss it below for a bivariate time series.

Consider the bivariate first order autoregressive system

\[
\begin{bmatrix}
  x_t \\
  y_t
\end{bmatrix} =
\begin{bmatrix}
  a & b \\
  c & d
\end{bmatrix}
\begin{bmatrix}
  x_{t-1} \\
  y_{t-1}
\end{bmatrix} +
\begin{bmatrix}
  \nu_t \\
  \mu_t
\end{bmatrix},
\] (3.20)

where \((\nu_t, \mu_t)\)' is a zero mean innovation process with covariance matrix

\[
\begin{bmatrix}
  \sigma^2_\nu & \sigma_{\nu\mu} \\
  \sigma_{\nu\mu} & \sigma^2_\mu
\end{bmatrix}.
\]

This system can be written as

\[
\Pi
\begin{bmatrix}
  x_t \\
  y_t
\end{bmatrix} =
\begin{bmatrix}
  \nu_t \\
  \mu_t
\end{bmatrix},
\] (3.21)
with

\[
\Pi = \begin{bmatrix}
1-aB & -bB \\
-cB & 1-dB
\end{bmatrix}.
\]

A necessary condition for stationarity of this system is that the roots \( \xi_{1,2} \) of

\[
\xi^2 - (a+d)\xi + (ad-bc) = 0,
\]

are inside the unit circle, or

\[
|\xi_{1,2}| = |(a+d) \pm (a^2+d^2-2ad+4bc)^{1/2}|/2 < 1.
\]

The time series models for the univariate series, given model (3.20), can now be found from

\[
\Pi \begin{bmatrix} x_t \\ \nu_t \end{bmatrix} = \begin{bmatrix} 1-dB & bB \\ cB & 1-aB \end{bmatrix} \begin{bmatrix} \nu_t \\ \mu_t \end{bmatrix},
\]

see Zellner and Palm (1974). For \( x_t \) and \( y_t \) this results in

\[
(1 - (a+d)B + (ad-bc)B^2)x_t = (1-hB)\varepsilon_t,
\]

with

\[
(1-hB)\varepsilon_t = (1-dB)\nu_t + bB\mu_t,
\]

and

\[
(1 - (a+d)B + (ad-bc)B^2)y_t = (1-jB)\kappa_t,
\]

with

\[
(1-jB)\kappa_t = cB\nu_t + (1-aB)\mu_t,
\]

where \( \varepsilon_t \) and \( \kappa_t \) are zero mean innovation processes with variances \( \sigma^2_\varepsilon \) and \( \sigma^2_\kappa \). The values of \( h \) and \( j \) depend on the \( a, b, c, d \), on the variances \( \sigma^2_\nu, \sigma^2_\mu \) and on the covariance \( \sigma_{\nu\mu} \). This can be seen by equating the variances and first order autocovariances of both sides of, e.g., (3.26), see Granger and Newbold (1986, p.30), or

44
\[(1+h^2)\sigma_e^2 = (1+a^2)\sigma_e^2 + b^2\sigma_\mu^2 - \beta d\sigma_\nu^2,\]

and
\[-ha_e^2 = -a\sigma_e^2.\]

These two expressions can be combined as
\[(1+h^2)/h = (1+a^2)/d + (b^2/d)(\sigma_\mu^2/\sigma_e^2) - \beta(\sigma_\nu^2/\sigma_e^2). \quad (3.29)\]

For the variable \(y_t\) it can analogously be derived that
\[(1+j^2)/j = (1+a^2)/a + (c^2/a)(\sigma_\mu^2/\sigma_e^2) - c(\sigma_\nu^2/\sigma_e^2). \quad (3.30)\]

These expressions are used below to indicate that some chosen \(h\) and \(j\) are quite reasonable.

Most common test procedures for unit roots assume an autoregressive model. This is a crucial point because in case stationarity is investigated with model (3.25) there would theoretically be no roots differing from those of (3.20). It is the autoregressive approximation which causes that the roots can differ. Moreover, in some cases the roots of such univariate series may approach unity, while those of the multivariate series do not. To construct examples of such cases it appears not necessary to assume large differences between \(\sigma_e^2\) and \(\sigma_\mu^2\).

The ARMA(2,1) model for the \(x_t\) in (3.25) can be approximated by an AR(3) model given by
\[
(1 + (-a-d+h)B + (ad-bc-hd+h^2)B^2 + h(ad-bc-hd+h^2)B^3 + \ldots)x_t = \epsilon_t. \quad (3.31)
\]

Assume now that the value of \(h\) is small such that the \((-a-d+h)\) is not about equal to \((-a-d)\), but that \(h(ad-bc-hd+h^2)\) is so small that it may be approximated by 0, then (3.31) can be approximated by the AR(2) process
\[
(1 + (-a-d+h)B + (ad-bc-hd+h^2)B^2)x_t = \epsilon_t. \quad (3.32)
\]

Of course, this approximation may not be appropriate in some cases, but it is also fairly easy to construct examples in which this approximation may
well be sustained.

A necessary condition for the variable $x_t$ in (3.32) to be stationary is that the roots of

$$
\lambda^2 + (-a-d+h)\lambda + (ad-ha-hd+h^2) = 0,
$$

(3.33)

satisfy the inequality

$$
|\lambda_{1,2}| = |(a+2d-h) \pm (a^2+d^2-3h^2-2ad+4bc+2ah+2hd)^{1/2}|/2 < 1.
$$

(3.34)

For $y_t$, an analogous condition for the roots $\zeta_{1,2}$, can be derived, to be denoted (3.34a) for convenience. Given the expressions (3.23), (3.34) and (3.34a), it is now conceivable that there may exist combinations of values of $\sigma^2_\mu$, $\sigma^2_\mu$, $\alpha$, $\beta$, $c$, $d$, and hence $h$ and $j$, which establish that the values of $\lambda_{1,2}$ and/or $\zeta_{1,2}$ come close to unity, while the absolute values of $\xi_{1,2}$ are clearly distinct from 1. So, there may be two cases, i.e. the nonstationarity of both the $x_t$ and $y_t$ variables or of either $x_t$ or $y_t$ can not be rejected.

A numerical example may clarify these cases. Consider again the model in (3.20) and assume for convenience that $a=d$ and that $b=c$. The roots are now given by

$$
\xi_{1,2} = a \pm b
$$

(3.35)

$$
\lambda_{1,2} = -h/2 + a \pm (-3h^2+4b^2+4ah)^{1/2}/2
$$

(3.36)

$$
\zeta_{1,2} = -j/2 + a \pm (-3j^2+4b^2+4aj)^{1/2}/2.
$$

(3.37)

Furthermore, set $a$, $b$ and $h$ equal to 0.7, 0.2 and 0.2, respectively. Note that $a$ reflects the dependence of the $x_t$ and $y_t$ on their own past, that $b$ reflects their interdependence, and that these chosen values are quite reasonable. It is now easy to see that the largest $\xi$ root equals 0.9, while that of $\lambda$ is equal to 0.987. Note that with this combination of parameters the expression in (3.31) becomes

$$
(1 -1.2B + 0.21B^2 + 0.042B^3 + ..)x_t = \epsilon_t,
$$

(3.38)

which may well be approximated by an AR(2) process. Additionally, when the

46
\( \sigma^2 \) equals \( \sigma^2 \), it can be seen that the expressions in (3.29) and (3.30) are equivalent, yielding that \( j \) equals \( h \), or that both series may be recognized to be nonstationary. The substitution of the assumed parameter values in (3.29) yields that

\[
\left( \frac{\sigma^2}{\sigma^2} \right) = 53.755 + 3.500\left( \frac{\sigma_{\nu\mu}}{\sigma^2} \right).
\]  

(3.39)

In an extreme case that for example \( \sigma^2 = 100\sigma_{\nu\mu}^2 \), it can be calculated from expression (3.39) that \( \sigma_{\nu\mu}/\sigma^2 \) is 13.213, which substituted in (3.30) gives that \( j \) is now about equal to \( a \), implying that the roots of the \( y_t \) process do not approach unity. This example illustrates that, although the roots of a bivariate time series system are distinct from unity, the roots of one or both of the univariate models can approach unity. How many of the series do have a unit root may depend on the residual variances and covariance.

It is now apparent that these theoretical results may have an impact on testing for unit roots in practical applications. It may be that in some of these occasions the univariate autoregressive models as in (2.12) are incorrectly specified given that a multivariate model would suggest a much higher order model. One solution may be to entertain a set of these models, and to see for which orders significant results emerge. An obvious drawback is now that it becomes difficult to recognize whether rejection is due to the correctness of the model or to improper empirical size of the test. A more sensible alternative is of course to consider the full system like in (3.20) and to test for unit roots in this context. One approach to this is given in Johansen (1988) and Johansen and Juselius (1990), which was discussed in section 2.2. One of the possible outcomes of this cointegration approach is that although each of the series may seem to be \( I(1) \), the multivariate time series model is stationary. The test statistic is given by the rank of the matrix \( \Pi \), and full rank indicates that the system is stationary.

Another solution may be to consider testing for unit roots in an ARMA model like in (3.25) with the techniques developed in e.g., Said and Dickey (1985) or Hall (1989). This framework should yield more adequate results, at least theoretically. Again the question of the specification of the orders of this model may however be hard to answer. One of the reasons is that the AR and the MA polynomials may contain nearly common factors, and this may seriously blur inference.
Testing for unit roots in Lydia Pinkham's annual data

An example of the empirical occurrence of no unit roots in the multivariate series and unit roots in the univariate series is given by the well-known Lydia Pinkham data. These data consist of annual observations for a sales variable $S_t$ and an advertising variable $A_t$ for the period 1907–1960. The observations are given in Erickson (1981) and displayed in figure 3.4.

![Figure 3.4 Lydia Pinkham's sales, $S$, and advertising, $A$](image)

From this figure it appears that one can argue that both series may show a nonstationary pattern, although these patterns seem to correspond. There may be some outlying observations, and to reduce their possible impact the series are transformed by a natural logarithm to $s_t$ and $a_t$. Furthermore, the series do not seem to be governed by a deterministic trend.
The series have been frequently used, especially in the area of marketing research, to illustrate newly developed modeling techniques and to investigate causality issues, see, e.g., Fase (1987) for a survey. A striking feature is that sometimes the data are used in levels, while in other applications inference is drawn using first order differenced series. In particular, univariate time series models are often built for $S_t$ and $A_t$ in differences, see, e.g., the prewhitening step in the transfer function analysis as it has been applied in Helmer and Johansson (1977). On the other hand, multivariate time series models, see Fase (1987), Granger and Newbold (1986), and distributed lag models (cf. Erickson 1981), are constructed for the variables in levels. This phenomenon may not be as paradoxical as it seems in case both the presumed univariate unit roots disappear in a bivariate system.

To test for univariate unit roots I consider the test procedure as given in (2.12). The results for the test statistic $\hat{f}_{\mu}$ for a large number of model specifications are displayed in table 3.7. The critical values of the test statistics can be found in table 2.1. The sample is taken to be from 1907 to 1960.

Table 3.7

<table>
<thead>
<tr>
<th>$p$</th>
<th>$s_t$</th>
<th>$a_t$</th>
<th>$p$</th>
<th>$s_t$</th>
<th>$a_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1.628</td>
<td>-2.255</td>
<td>6</td>
<td>-2.522</td>
<td>-2.222</td>
</tr>
<tr>
<td>1</td>
<td>-2.402</td>
<td>-2.575</td>
<td>7</td>
<td>-3.235**</td>
<td>-2.288</td>
</tr>
<tr>
<td>2</td>
<td>-2.169</td>
<td>-1.888</td>
<td>8</td>
<td>-3.243**</td>
<td>-2.284</td>
</tr>
<tr>
<td>3</td>
<td>-2.656</td>
<td>-1.817</td>
<td>9</td>
<td>-3.350**</td>
<td>-2.702*</td>
</tr>
<tr>
<td>4</td>
<td>-2.968**</td>
<td>-2.226</td>
<td>10</td>
<td>-3.697**</td>
<td>-2.067</td>
</tr>
<tr>
<td>5</td>
<td>-2.921*</td>
<td>-2.406</td>
<td>11</td>
<td>-2.539</td>
<td>-3.359**</td>
</tr>
</tbody>
</table>

** Significant at a 5% level.
* Significant at a 10% level.

(1) The test equation is (2.12). The critical values are displayed in table 2.1. The value of $p$ corresponds with an assumed AR($p+1$) model.

(2) The underlined values indicate the smallest order of the model for which the residual autocorrelations do not indicate misspecification.
The usual procedure is to set the order $p$ equal to that value for which the model does not suffer from residual autocorrelation, and hence dynamic mis specification. From table 3.7 it is clear that this order would be 1 and 2 for $s_t$ and $a_t$, respectively, and that then the null hypothesis of a unit root can not be rejected in either case. Increasing the order of the model leads to mixed results. The larger the $p$, the more significant the $\hat{p}_p$ becomes for $s_t$, but only when $p$ is 11 this statistic is significant at a 5% level for the $a_t$ series. Hence, it appears that only at high lags, where most of the times the corresponding parameters are not significantly different from zero, the null hypothesis can be rejected. It is however unclear whether this is due to a size distortion or to other causes. Most practitioners would certainly not specify that large the models for these short series, see Helmer and Johansson (1977).

Table 3.8
An application of the Johansen procedure to the annual Lydia Pinkham series

<table>
<thead>
<tr>
<th>Statistics$^{(1)}$</th>
<th>VAR(3)</th>
<th>VAR(4)</th>
<th>VAR(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>0.234</td>
<td>0.269</td>
<td>0.289</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>0.133</td>
<td>0.165</td>
<td>0.146</td>
</tr>
<tr>
<td>$Tr(1)$</td>
<td>7.261</td>
<td>9.028$^*$</td>
<td>7.751$^*$</td>
</tr>
<tr>
<td>$Tr(0)$</td>
<td>20.876$^{**}$</td>
<td>24.709$^{**}$</td>
<td>24.471$^{**}$</td>
</tr>
</tbody>
</table>

** Significant at a 5% level.
* Significant at a 10% level.
(1) Critical values for $Tr(r)$ can be found in table 2.2, last two columns.

Given that the series obviously seem to constitute a bivariate system, and that there are no indications (Fase 1987) that there may be relevant omitted variables, it seems appropriate to consider a multivariate model for testing for stationarity. In Granger and Newbold (1986, p.253-256) and Fase (1987) it is shown that a VAR(3) model and a VARMA(2,1) model seem to be the most adequate for the levels of the series. The estimation and testing results can be found in these references. The issue of interest here is the application of the cointegration test procedure given in Johansen and Juselius (1990) and reviewed in section 2.2. From Boswijk and
Franses (1991) it emerges that it is a sensible strategy to check for cointegration in a range of models, and hence I report the results for a VAR(3), VAR(4) and a VAR(5) for the series $s_t$ and $q_t$ in table 3.8. From these outcomes it appears that the bivariate system seems to be stationary indeed, given that one is willing to adopt a 10% significance level and assuming that a VAR(4) model is the most appropriate. The results after the removal of an outlying observation in the advertising series do not appear too strongly at variance with this. The residuals of the two cointegration relations in the VAR(4) are depicted in figure 3.5, and they look rather stationary. A suitable normalization of these relations gives the equilibrium relations $s_t = 0.546q_t + 3.785$ and $q_t = 0.748s_t + 1.202$, both of which may well be rationalized with theoretical arguments.

![Figure 3.5 Two cointegration relationships for the Lydia Pinkham data](image-url)

Figure 3.5 Two cointegration relationships for the Lydia Pinkham data
The univariate time series models which can be derived from a bivariate VAR(3) model are at most ARMA(6,3) models. Any approximating AR models may have been of order 9 or even higher. So, if one would have started the univariate analysis with such large models, one would also have concluded that stationarity of the univariate series is not unacceptable. Since these orders are unknown in case the bivariate model is not specified a priori, it seems unlikely that they would be established empirically.

The outcomes of this empirical exercise, together with its preceding theoretical results, seem to lead to the conclusion that testing for stationarity should be preferably carried out in a multivariate framework. Related experience with the Johansen (1988) method shows that its rejection rate is not unsatisfactory provided that one selects an adequate model, see Boswijk and Franses (1991). When such a model is not readily available it may be sensible to consider a range of univariate models for sensitivity analysis purposes.

A set of models for U.S. unemployment rate

Now that several of the issues related to the determination whether a time series is stationary or not have been discussed above, I turn to the time series which is supposed to provide an illustration of the alternative model selection strategy proposed in the previous section. The graph of the annual series of the log of U.S. unemployment rate, $U_t$, for the period 1919–1988 is depicted in figure 3.6. In the sequel the last 8 observations will be discarded for they will be used in a forecasting performance exercise. The figure for $U_t$ seems to show a rather stationary behavior and no dependence on a deterministic trend.

<table>
<thead>
<tr>
<th>Table 3.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autocorrelations and partial autocorrelations of U.S. unemployment rate</td>
</tr>
<tr>
<td>Lag</td>
</tr>
<tr>
<td>ACF</td>
</tr>
<tr>
<td>PACF</td>
</tr>
</tbody>
</table>

** Significant at a 5% level.
* Significant at a 10% level.
Furthermore, there is some visual evidence for the presence of cycles. This implies that the order of the autoregressive part of eventual ARMA models for this series should at least be 2. The values of the ACF and PACF up to lag 8 are displayed in table 3.9.

![Graph showing the annual unemployment rate in the U.S., 1919-1988](Image)

**Figure 3.6** The annual unemployment rate in the U.S., 1919-1988

The ACF seems to be indicate that $U_t$ may be described by an autoregressive process, and the values of the PACF support this conjecture.

To formally investigate whether the series is stationary or not, the Dickey-Fuller procedure given in (2.12) is applied. The results for models up to $p = 7$ in (2.12) are displayed in table 3.10. Higher orders do not give very distinct results. Again it appears that it matters which model order is assumed, although the general outcome seems to be that the null hypothesis of nonstationarity can be rejected. Recursive results as in
figure 3.2 are not reported for they do not yield any insights with respect to structural breaks.

Table 3.10
Testing for unit roots in the U.S. unemployment series with Dickey–Fuller's $\hat{f}_\mu$ statistic\(^{(1)}\)

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\hat{f}_\mu$</th>
<th>$p$</th>
<th>$\hat{f}_\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-3.038**</td>
<td>4</td>
<td>-3.291**</td>
</tr>
<tr>
<td>1</td>
<td>-3.053**</td>
<td>5</td>
<td>-2.134</td>
</tr>
<tr>
<td>2</td>
<td>-2.218</td>
<td>6</td>
<td>-2.534</td>
</tr>
<tr>
<td>3</td>
<td>-2.840*</td>
<td>7</td>
<td>-2.687*</td>
</tr>
</tbody>
</table>

** Significant at a 5% level.
* Significant at a 10% level.

\(^{(1)}\) This test statistic is given in (2.12). Critical values are displayed in table 2.1. The $p$ corresponds to the assumed AR($p+1$) model in (2.12).

The series $U_t$ will therefore not be differenced to obtain stationarity. Given the above arguments I decide to consider AR models with nonzero mean only. The lowest and highest order will be set equal to 1 and 6, respectively. The year of the first observation is 1925 to have identical sample size for all models. Furthermore, I will limit the analysis to complete AR models, which implies that an AR($p$) model incorporates $p$ terms. An extension to include subset AR models (McClave 1975) is possible, but this may not contribute to the clarity of further exposition.

Parameter constancy will be tested to check for the presence of structural breaks. Finally, from the literature on unemployment it emerges that this variable may show nonlinear behavior, see Neftci (1984) and Leukkonen and Teräsvirta (1990), which may be established by the periods of increasing unemployment which may be shorter than those of decreases. Their empirical results however have been established for seasonally adjusted quarterly data, and it may well be that nonlinearity is not prevalent in annual data. Anyhow, a test for nonlinearity will certainly be included in the battery of diagnostic checks of the basic requirements.
3.4 DIAGNOSTICS

Diagnostic checking of a time series model implies the investigation of the sample equivalents of the assumptions underlying the model. The most important assumption for an ARMA model is that the \( \{ \varepsilon_t \} \) is a white noise process. Since the mean of the estimated residuals is automatically equal to 0, it is important to check whether these residuals are uncorrelated and whether their variances do not depend on time. Next, from (2.1) one can see that the considered class of models is linear, and this provides another feature for which one may want to test. The parameters are also assumed to be time independent, and hence testing for parameter constancy may also be useful. Although the \( \varepsilon_t \) are not necessarily assumed to be Gaussian, it may be insightful to check for the normality of the estimated residuals. Rejection of the null hypothesis of normality may indicate the presence of outliers, or of structural breaks. In that case it may also be useful to formally check whether there are outlying observations. In the sequel of this section test procedures for each of these assumptions will be briefly reviewed and applied.

The general viewpoint taken here is that each assumption is tested individually. An obvious drawback of this approach is that the properties of most test statistics are derived given that all other requirements are fulfilled. For example, when one faces large outliers, convenient asymptotic results for autocorrelation test statistics cease to hold. One strategy may now be to construct test procedures which may detect any, or more than one, type of misspecification, see, e.g., Bierens (1987), Bera and Jarque (1982). Although the coefficients in auxiliary regressions can give some insights, an obvious disadvantage of these tests is that in case of misspecification, the modeling exercise may be terminated without having strong indications for subsequent steps. Furthermore, it is my experience that a certain misspecification will usually be reflected in the values of several test statistics at the same time. For example, a sequence of outliers can have an effect on parameter constancy tests, on autocorrelation checks, on homoscedasticity checks, and also on tests for linearity.

Uncorrelated residuals

The well-known portmanteau test statistics for residual autocorrelation of order \( m \), i.e. \( BP(m) \) and \( LB(m) \), have already been given in (3.2) and (3.5). These tests are called pure significance tests in the sense that there is
no explicit formulation of an alternative hypothesis, see for a survey and
Both test statistics are extensively used in practice, and also several
empirical size and power investigations have been carried out, see Davies
and Newbold (1977) and Clarke and Godolphin (1982). From these studies it
emerges that in small samples the statistics may lack power, although the
LB test is not always performing that disappointingly. In moderately large
samples the behavior of the tests is rather similar.

Nested and nonnested hypothesis tests, as developed in Godfrey (1979)
and McAleer et al. (1988), respectively, design a specific alternative
model. As expected, the empirical powers of these tests can be higher than
the portmanteau tests, although the sizes of all tests may vary across the
different generating processes (Hall and McAleer 1989). In some cases the
Lagrange Multiplier (LM) test advocated in Godfrey (1979) can be written as
$\alpha$ times the $R^2$ of an auxiliary regression. For example, to test an AR($p$)
model against an AR($p+r$) or an ARMA($p, r$) model (see also Poskitt
and Tremayne 1980), the LM test is found by estimating

$$
\hat{\epsilon}_t = \alpha_0 \hat{Y}_{t-1} + \ldots + \alpha_p \hat{Y}_{t-p} + \alpha_{p+1} \hat{Y}_{t-p-1} + \ldots + \alpha_{p+r} \hat{Y}_{t-p-r} + \hat{u}_t,
$$

(3.40)

where $\hat{\epsilon}_t$ are the estimated residuals of the AR($p$) model. The statistic
LM($r$), when calculated as $nR^2$, is asymptotically $\chi^2(r)$ distributed under
the null hypothesis that the AR($p$) model is adequate. A procedure which is
related to the equation in (3.40) is to use the cross correlation function
between the observed series and future values of the estimated residuals
(Hokstad 1983). The general idea is that these should not be correlated in
an adequate model.

A modification to the portmanteau test statistic has been proposed by
Godolphin (1980). This test is however not easy to calculate, although its
power seems to be rather satisfactory (Clarke and Godolphin 1982). Yet
another approach is to fit ARMA models to the estimated residuals and to
see whether the order of the selected model is unequal to zero (Pukkila
and Krishnaiah 1988). In table 3.11 only the results for the BP and LM tests
will be displayed since the results of other test procedures did not yield
different outcomes. The values in the arguments of these test statistics
are those used in Hall and McAleer (1989). Below I report results that
indicate that there are no outliers. Hence it is not expected that the
results of the autocorrelation checks are affected by such data points.
Table 3.11
Testing AR models for U.S. unemployment: Autocorrelation\(^{(1)}\)

<table>
<thead>
<tr>
<th>Model</th>
<th>BP(5)</th>
<th>BP(10)</th>
<th>LM(1)</th>
<th>LM(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td>7.925*</td>
<td>12.766</td>
<td>5.577**</td>
<td>5.585*</td>
</tr>
<tr>
<td>AR(2)</td>
<td>3.106</td>
<td>6.803</td>
<td>0.009</td>
<td>7.016**</td>
</tr>
<tr>
<td>AR(3)</td>
<td>3.187</td>
<td>6.939</td>
<td>7.008**</td>
<td>7.517**</td>
</tr>
<tr>
<td>AR(4)</td>
<td>2.555</td>
<td>8.871</td>
<td>0.581</td>
<td>2.312</td>
</tr>
<tr>
<td>AR(5)</td>
<td>na(^{(2)})</td>
<td>8.387</td>
<td>1.749</td>
<td>2.141</td>
</tr>
<tr>
<td>AR(6)</td>
<td>na(^{(2)})</td>
<td>7.259</td>
<td>0.405</td>
<td>1.160</td>
</tr>
</tbody>
</table>

** Significant at a 5% level  
* Significant at a 10% level.

\(^{(1)}\) The expressions for the statistics can be found from (3.2) and (3.40).
\(^{(2)}\) The distribution of the test statistic is not available in this case.

The results in this table indicate that the AR(1), AR(2) and AR(3) models do not seem to pass the autocorrelation checks. Since the assumption of uncorrelated residuals is crucial in this type of models, they will not be used in further testing.

**Normality and outliers**

Under the assumption of uncorrelated residuals, the test statistics for skewness, \(LO_{SK}\), and kurtosis, \(LO_{K}\), given in Lomnicki (1961) reduce to

\[
LO_{SK} = (n/6)^{\frac{1}{2}} \frac{m_3}{m_2^{\frac{3}{2}}} 
\]

(3.41)

and

\[
LO_{K} = (n/24)^{\frac{1}{2}} \left( m_4/m_2^2 - 3 \right),
\]

(3.42)

where

\[m_j = n^{-1} \sum_{i=1}^{n} \bar{e}_j^i \]

These statistics follow standard normal distributions under the hypothesis
of normality. The results for these normality test statistics can be found in columns 2 and 3 of table 3.12. It is clear that the null hypothesis cannot be rejected.

Although there seems to be no indication for the presence of outliers, it may always be rather instructive to consider statistics for the detection of influential observations. In the time series literature several suggestions for the detection, and eventual removal, of such observations have been given, see e.g., Fox (1972), Abraham and Chuang (1989) and Bruce and Martin (1989). A simple detection device for AR models is given in Abraham and Yatawara (1988). Consider the AR model as in (2.6), or \( \phi_p(B)\eta_t = \varepsilon_t \), for which two types of models can be distinguished. When there is an aberrant observation (AO) at time \( \tau \), this can be modeled as

\[
z_t = y_t + \Xi_1 \delta_{t,\tau} = \phi_p(B)^{-1} \varepsilon_t + \Xi_1 \delta_{t,\tau}, \tag{3.43}
\]

or

\[
\varepsilon_t = \phi_p(B)(z_t - \Xi_1 \delta_{t,\tau})
\]

where \( \delta_{t,\tau} \) equals 1 if \( t=\tau \), and 0 elsewhere, and where \( \Xi_1 \) is the magnitude of the shift at \( t=\tau \), and \( z_t \) is the observed time series. The aberrant innovation (AI) model is given by

\[
z_t = y_t + \phi_p(B)^{-1} \Xi_2 \delta_{t,\tau} = \phi_p(B)^{-1} \varepsilon_t + \Xi_2 \delta_{t,\tau}, \tag{3.44}
\]

or

\[
\varepsilon_t = \phi_p(B)z_t - \Xi_2 \delta_{t,\tau}
\]

where \( \Xi_2 \) is the magnitude of the shift in the innovation at time \( \tau \). It is clear from (3.43) and (3.44) that, depending on \( \phi_p(B) \), in case of an AO the residuals in future times are affected, and in case of an AI the residual at time \( \tau \) and future observations are influenced.

To test the joint null hypothesis that \( \Xi_1=\Xi_2=0 \) at time \( t=\tau \), a Lagrange Multiplier test statistic has been derived in Abraham and Yatawara (1988), which is given by

\[
LM_\tau = (\hat{\varepsilon}_t^2/\hat{\sigma}^2) + \left( \sum_{i=1}^p \phi_i \hat{\varepsilon}_{t-i}^2 \right)/ (\hat{\sigma}^2 \sum_{i=1}^p \hat{\phi}_i^2), \tag{3.45}
\]
where \( r \) runs from \( t=p+1,\ldots,n \), and where \( \hat{\sigma}^2 \) is the maximum likelihood estimate of the residual variance, or \( \text{RSS}/n \), with \( \text{RSS} \) denoting the residual sum of squares. In case the values of the parameters and the value of \( r \) are known, it can be shown that the LM\(_r\) statistic follows a \( \chi^2(2) \) distribution. When the parameters have to be estimated as in (3.45) and when the position of \( r \) is also unknown and, additionally, when there may be more than one influential data point, things get more complicated. In practice however it may still be useful to compare the values with the \( \chi^2 \) table, as well as relative to each other, to get an impression of the possible presence of influential observations. Abraham and Yatawara (1988) also discuss two test statistics for the separate null hypotheses that there is an AO or an AI. These statistics follow \( \chi^2(1) \) distributions and can be used to distinguish between the two outlier types. The values of the test statistic in (3.45) for the AR(4) model for the \( U_t \) series are displayed in Figure 3.7.

![Figure 3.7 Testing for outliers in an AR(4) model](image)

Figure 3.7 Testing for outliers in an AR(4) model
From this figure it emerges that the null hypothesis of no influential observations can not be rejected. Similar patterns show up for the other AR models. Note that this hypothesis considers such observations in single models only. It may however be that a comparison between the models is influenced by one or more observations, see Franses (1990a). In section 3.5 this will be investigated in more detail.

Homoscedasticity

One of the properties of a white noise error process is that the variance of the residuals does not depend on time. An example of an error process for which the assumption of homoscedasticity is not valid is the autoregressive conditional heteroscedasticity (ARCH) process. One version of such a process is given and studied in Weiss (1984,1986),

$$E(e_t^2|I_{t-1}) = h_t$$

$$= \alpha_0 + \sum_{i=1}^{k} \alpha_i e_{t-i}^2 + \delta_0 (\hat{\gamma}_t - \mu)^2 + \sum_{i=1}^{I} \delta_i (y_{t-i} - \mu)^2,$$  \hspace{1cm} (3.46)

where $\alpha_i, \delta_i \geq 0$, where $I_{t+1}$ denotes the information set including all past observations on $y_t$, and where $\hat{\gamma}_t$ denotes the fitted values for $y_t$. Further stability conditions for model (3.46) are discussed in Weiss (1984). The test procedure for the null hypothesis that all parameters other than $\alpha_0$ are equal to zero can be based on the Lagrange Multiplier principle, and is then given by $n^2 \hat{\delta}^2$ of the auxiliary regression with the sample equivalents from elements in (3.46). This test statistic follows a $\chi^2(k+I+1)$ distribution under the null hypothesis of no ARCH. Note that this test procedure includes the procedures proposed in Engle (1982) for simple ARCH models, and the heteroscedasticity test developed in Breusch and Pagan (1979). The results when setting $k=I=1$ for the AR models for the $U_t$ series are displayed in the fourth column of table 3.12. The outcomes for higher values of $k$ and $I$ are similar and hence not reported.

Linearity

The test procedures for linearity can also be divided into two types, i.e. the pure significance tests and the tests which explicitly consider an alternative hypothesis. The empirical sizes and powers of many of these
test statistics have been studied in Chan and Tong (1986), Lee, White and Granger (1989), and Luukkonen, Saikkonen and Teräsvirta (1988). From their results it emerges that the test developed in Keenan (1985) performs better than the other pure significance tests as, e.g., the McLeod and Li (1983) test, and that the Lagrange Multiplier test for bilinearity given in Weiss (1986) also shows satisfactory behavior. Both these test statistics will be considered in the present application.

Table 3.12
Testing some AR models for U.S. unemployment: Other features

<table>
<thead>
<tr>
<th>Model</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>ARCH</th>
<th>Linearity</th>
<th>Bilinearity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(4)</td>
<td>1.225</td>
<td>-0.096</td>
<td>0.764</td>
<td>3.684*</td>
<td>1.800</td>
</tr>
<tr>
<td>AR(5)</td>
<td>1.044</td>
<td>-0.235</td>
<td>0.903</td>
<td>2.845*</td>
<td>1.484</td>
</tr>
<tr>
<td>AR(6)</td>
<td>1.496</td>
<td>-0.090</td>
<td>0.975</td>
<td>2.533</td>
<td>1.927</td>
</tr>
</tbody>
</table>

* Significant at a 10% level.

(1) The $N(0,1)$ test statistics for skewness and excess kurtosis are given in (3.41) and (3.42), respectively. The $\chi^2(3)$ test statistic for ARCH can be found using (3.46). The $F(1, 56-2p-2)$ test for linearity, where $p$ is the order of the AR model, is displayed in (3.47). The $\chi^2(1)$ test statistic for bilinearity is constructed as $nR^2$, where $R^2$ results from estimating (3.48).

The Keenan (1985) test amounts to three auxiliary regressions. First, an autoregressive model of order $M$ is fitted to the original time series $y_t$, giving fitted values $\hat{y}_t$ and residuals $\hat{\varepsilon}_t$. Then, $\hat{y}_t^2$ is regressed on $1, y_{t-1}, \ldots, y_{t-M}$, which gives the residuals $\xi_t$. Finally, the $\xi_t$ is regressed on $\hat{\varepsilon}_t$, yielding a regression coefficient $\hat{\eta}_p$. Denoting $RSS_\varepsilon$ and $RSS_\xi$ as the sums of corresponding squared residuals, the test statistic is calculated by

\[ Keenan = (n-2M-2)\hat{\eta}_p^2RSS_\xi/(RSS_\varepsilon - \hat{\eta}_p^2RSS_\xi), \]

(3.47)

which asymptotically follows an $F(1, n-2M-2)$ distribution under the null
hypothesis of linearity. From (3.47) one can observe that this test can be viewed as a special case of the well-known RESET test. The choice of a appears to be of importance. For example, incorrect sizes may occur in M is too small, since the test will then also respond to remain in autocorrelation in the residuals because of similarly correlated residual processes \( \epsilon_t \) and \( \eta_t \). The results for this test statistic, where \( M \) is corresponding to the hypothesized orders of the AR models, are given in the fifth column of table 3.12. It appears that there may be some evidence the presence of nonlinearity. Unreported sensitivity analysis reveals that most nonlinearity seems to be present in the pre-war years, and that series may seem to be linear after 1960.

The LM test procedure developed in Weiss (1986) considers a bilinear model as an explicit alternative. The bilinear process discussed in Granger and Anderson (1978) which nests an AR(p) model is given by

\[
\phi_p(B) y_t = \sum_{i=1}^{P} \sum_{j=1}^{Q} \beta_{ij} y_{t-i} \epsilon_{t-j},
\]

(3.48)

where \( \epsilon_t \) is again defined as in chapter 2. A \( \chi^2(PQ) \) test statistic can be derived similar to the test for ARCH, see (3.46), and the results for \( P=Q=1 \) are displayed in column 6 of table 3.12. Extending \( P \) and \( Q \) to 2 showed rather significant values, but again this significance appeared to be largely dependent on a subset of the sample. In summary, it may be reasonable to accept linearity for the AR models considered, although subsamples show some nonlinear behavior.

Parameter constancy

A version of a Chow (1960) test, which was originally developed for linear regression models, is given by

\[
\text{CHOW} = \left( \frac{\text{RSS}_1/\text{RSS}_n}{\text{RSS}_n - 1} \right) \left( \frac{(n_1-k)}{n_2} \right),
\]

(3.49)

where \( n_1 \) refers to the subset containing the first \( n_1 \) observations, and where \( n = n_1 + n_2 \), and \( k \) is equal to the number of parameters. This statistic, which is particularly useful because it does not necessitate model estimation over \( n_2 \) observations, asymptotically follows an \( F(n_2, n_1-k) \) distribution under the null hypothesis of parameter constancy. This CHOW test has been calculated for \( n_1 = 6, \ldots, 55 \), which yields 51 test outcomes.
The highest value of these is 1.332, and the lowest is 0.160. Hence, it can be concluded that parameter constancy can not be rejected.

Summarizing these results and those in table 3.12, the three models all pass a battery of diagnostic checks. In the next section, these models will be considered in two model selection exercises, one in which the final model is used for forecasting, and one in which it is used for econometric model building.

What if no model is adequate?

It is of course imaginable that one may end up with all models being inadequate with respect to one or more of the criteria reviewed above. An example is when there is an observation which is an outlier for all models in the set. The best way to proceed then is to take care of this observation, e.g., by using a single dummy variable, and to start the whole modeling exercise again. In general, it may be argued that putting a dummy variable in the middle of a model building exercise may bias model selection outcomes. In case all models suffer from nonlinearities or ARCH in the residuals, the most sensible thing to do seems to be to start model building with an extended class of models in which the linear ARMA models are special cases. Examples of such classes are given in (3.46) and (3.48). Models in these classes are somewhat more complicated, which is mostly reflected by the parameter estimation procedures, but model selection may proceed along similar lines as proposed in section 3.2.

Although there are many test procedures available to check for several phenomena, there is always a judgmental aspect involved. When exactly is an observation an outlier? How many of such observations are there? For example, it may be rather difficult to distinguish whether a certain observation and not its nearest neighbor is an outlier. The same argument applies to the detection of structural breaks. In the stock of cars example in figures 3.1 and 3.2 this appeared to be fairly easy. Also, the remedy, which is to consider a bivariate series, was rapidly available. In other cases, this might be not that straightforward, and an option can be to incorporate dummy variables for entire periods.

The difficulty now is of course the determination of the turning-points. In fact, it can occur that some slight changes in such dummies can yield an entirely different model selection. An example of such an occurrence is given in Broersma and Franses (1990). Starting point in that
study is the moving planning horizon model proposed in Winder and Palm (1989), which implies that an ARIMA(1,1,1) model for quarterly consumption may be appropriate. When fitted to real data, this model contains three dummy variables for subperiods and four additional dummies for the single observations right before and after these breaks. The exact position of these variables is determined by the pattern of an income variable and the assumption that consumption follows income with one quarter. However, two slight changes in the dummies in the empirical model for consumption can be defended. The first is that the income series seems flexible enough to allow another specification for the turning-points. The second is that it appears from the figure in which both series are plotted together, that consumption may just as well follow income with three quarters. The two new univariate ARIMA(1,1,1) models for consumption also pass several diagnostic checks. Surprisingly, however, it turns out that both these models can be simplified to an ARIMA(0,1,0) model, which was the model to be rejected by Winder and Palm (1989) in the first place.

3.5 MODEL SELECTION

Once the set of models has been identified, and its elements have been screened for possible deficiencies, the selection of the model for further use will be based on the subset of models which did pass the checks for the basic requirements. I will consider the occasions in which the model is built for forecasting or for description purposes. Of course, there may be other models suitable or explicitly designed for forecasting, such as exponential smoothing models or formal judgmental procedures. But, I will confine the current analysis to ARMA models. The descriptive purpose must be viewed in the light of dynamic econometric model building. For some strategies initial models for the individual time series are required, see, e.g., Haugh and Box (1977) and Davidson et al. (1978). For convenience, I will label such models to be descriptive models. The division of the purposes of modeling into two categories is largely influenced by the often recognized phenomenon that models with good fit or description do not necessarily predict well, and vice versa. Since there are also model selection criteria designed for each of these two purposes, it seems natural to select a forecasting model with forecasting criteria and a descriptive model with description criteria.
In the time series literature one can find many surveys of model selection criteria, see, e.g., De Gooijer et al. (1985) and Shibata (1985), and the discussion below will therefore not be that extensive and mostly highlight some interesting issues and recent advances. One of the aspects of the model selection strategy adopted in this chapter is that no model is considered to be true, or a priori preferable, and hence the models will be symmetrically treated. In general, this implies that the model is selected which minimizes the value of a certain criterion function, and not when a test statistic fails to reject a null hypothesis. Of course, it may be possible to express the criterion functions in terms of test statistics, see Teräsvirta and Mellin (1986) and Franses (1989a), but then one should anyhow have to fix a certain significance level.

**Forecasting criteria**

The class of forecasting model selection criteria can be divided in two subclasses, depending on whether one makes an explicit use of a validation sample or not. Such a sample contains observations which are not used for estimation, but which are reserved for model evaluation. However, I will only consider the criteria for which no such sample is needed and which mostly make use of asymptotic arguments or of recursive calculations within the sample. The years 1981–1988 for the unemployment series in the example will be used to evaluate the empirical forecast errors related to the model selection.

The Final Prediction Error (FPE) criterion developed in Akaike (1969) is often applied. Consider an AR(p) model that is used to predict one step ahead at time $t$, then the forecast is

$$f_{t+1} = \phi_1 y_t + \ldots + \phi_p y_{t-p+1}.$$  \hspace{1cm} (3.50)

The estimated forecast $\hat{f}_{t+1}$ can be calculated by substituting the estimates of the parameters in (3.50). The several AR models in the previous section are estimated with the same $n$ observations, and this is equivalent to assuming, as in Akaike (1969), that the pre-sample data are set equal to zero. In Davisson (1965) it is shown that the asymptotic mean-square error of this one-step forecast is given by

$$E(y_{t+1} - \hat{f}_{t+1})^2 = \sigma_e^2 (1 + p/n).$$  \hspace{1cm} (3.51)
An unbiased estimator of \( \sigma_x^2 \) in case of AR models is the least squares estimator

\[
\hat{\sigma}_{LS}^2 = \text{RSS}/(n-p),
\]

(3.52)

where RSS denotes the residual sum of squares. The model selection based on these expressions is to choose the model which minimizes the FPE criterion

\[
\text{FPE}_{LS}(k) = \hat{\sigma}_{LS}^2 (1 + k/n),
\]

(3.53)

where \( k \) is the order of the AR model, with \( k \in [1, \ldots, p] \). When the variance of the residuals is estimated with the maximum likelihood estimator

\[
\hat{\sigma}_{ML}^2 = \text{RSS}/n,
\]

(3.54)

the FPE criterion can be written as

\[
\text{FPE}_{ML}(k) = \hat{\sigma}_{ML}^2 ((n+k)/(n-k)),
\]

(3.55)

to yield a similar unbiased criterion. It is the variant in (3.55) which is often applied in practice, although one may also find the approximation

\[
\text{FPE}_{ML}(k) \approx \hat{\sigma}_{ML}^2 (1 + 2k/n),
\]

(3.56)

in the time series literature. Note that the penalty for including too many regressors is higher for (3.55) than for (3.56), and that in small samples such differences may become important. Moreover, note that (3.56) is related to the information criterion proposed in Akaike (1974).

The PC criterion for model selection in the linear regression context which is proposed in Amemiya (1980) is identical to the FPE criterion, see also Amemiya (1980, footnote 4 on page 335). This result corresponds to the findings in Söderström (1977) that the FPE criterion can be stated independently of the model structure. An assumption for these criteria is that in turn each model is assumed to be true. When it is assumed that the largest model is the true model, criteria as those developed in Allen (1974) and Mallows (1973) can emerge, see, e.g., Aznar (1989) for a survey.

Several modifications to the FPE criteria have been proposed. The FPE criterion as in (3.55) can show inconsistent behavior implying that it
often selects too large a model. Therefore the version

\[ \text{FPE}_{ML, \gamma}(k) = \hat{\theta}^2_{ML} \left( \frac{(n+k-1-\gamma)}{(n-k)} \right), \]  

(3.57)

where \( 0 < \gamma < 1 \) is proposed in Akaike (1970), for which it is easy to see that its value exceeds that of (3.55). The determination of the value of \( \gamma \) is however not straightforward. In Akaike (1970) it is argued that when \( \gamma = 0.25 \) this criterion has the tendency to underestimate the model order. Another modification concerns a generalization of (3.56),

\[ \text{FPE}_{ML}(k) = \hat{\theta}^2_{ML} \left( 1 + \alpha k/n \right), \]  

(3.58)

where \( \alpha \) is a positive fixed constant. In Bhansali and Downham (1977) it is concluded that it is not possible to give a simple rule for choosing the value of \( \alpha \), although the restriction of \( \alpha \) to one particular value is not advised. Confidence in a model increases in case the choice of this model remains constant when \( \alpha \) is varied. When the criteria are in conflict it is suggested to adopt the choice made for \( \alpha = 4 \). The FPE criterion for the selection between MA models is given in Bhansali (1983)

From (3.51) it is clear that the FPE criterion has been derived for predicting one-step ahead only. In many occasions however the model will be used for the prediction of more than one step. A useful model selection criterion can now be constructed with the results in Bhansali (1974). There it is derived that for an AR(\( p \)) process the asymptotic mean square error of predicting \( h \) steps ahead, where \( h > 1 \), is

\[ \text{E}(y_{t+h} - \hat{y}_{t+h})^2 = V(h)(1 + p/n), \]  

(3.59)

where \( V(h) \) is defined as in (2.30). Rewriting a stationary AR model in its MA representation gives coefficients, say, \( \eta_i \). The expression in (3.59) can be used to construct a new model selection criterion for \( h \)-step forecasting which was already suggested in a remark in Bhansali (1974), but which to my knowledge has never been applied in practical occasions, is given by

\[ \text{HFPE}_{h, ML}(k) = \hat{\theta}^2_{ML} \left( (n+k)/(n-k) \right) \sum_{h} \hat{\eta}_j^2, \]  

(3.60)

where the \( \hat{\eta}_j \) are calculated from the estimated AR coefficients.

The forecasting criteria discussed so far only consider the predictions
made at time $t$. It might however also be sensible to see how the models performed in previous periods with respect to forecasting. This boils down to fitting the models for subsamples, to out-of-sample forecasting 1 or $h$ periods ahead, and to evaluating their forecasting performance with the available data at hand. The idea is that models with a successful past forecasting record may also perform well in future instances. It is clear that criteria based on recursive forecast errors are related to the cross-validatory criteria in regression as given in, e.g., Allen (1974) and Stone (1974) and the sample re-use methods as the bootstrap and the jackknife, see for example Dijkstra (1988). The Predictive Least Squares criterion $PLS_h$ developed in Rissanen (1986) is such a criterion. In case of AR models and when the recursive estimation principle is applied to the whole sample it can be shown that the $PLS_1$ criterion is consistent, see Wax (1988) and Hemerly and Davis (1989). These results are obtained by assuming that there is a fixed upper bound $P$. Recently, Hemerly and Davis (1991) prove that consistency is maintained when $P$ increases with the sample size. The $PLS_h$ criterion selects the model for which the sum of the squared $h$-step ahead prediction errors along the sequence are minimized. Equivalently, one can consider the average of the root mean squared errors (RMSE). In practice it may be convenient to consider only a recent subperiod instead of the entire sequence of observations.

### Table 3.13

<table>
<thead>
<tr>
<th>Criteria(^{(1)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>AR(4)</td>
</tr>
<tr>
<td>AR(5)</td>
</tr>
<tr>
<td>AR(6)</td>
</tr>
</tbody>
</table>

\(^{(1)}\) FPE criterion calculated via (3.55), the \(HFPE\) criterion as in (3.60), the \(PLS_1\) criterion is calculated as the mean of the RMSE of forecasting one period ahead for 1971–1980, the \(PLS_h\) for 8 periods ahead, and RMSE is the root mean squared error of forecasts for the period 1981–1988.
The results of the selection of a forecasting model for the $U_t$ series are displayed in Table 3.13. The various FPE criteria all yield a similar outcome, and hence only the one in (3.55) is given. When the goal is to forecast 8 steps ahead, the values of the criterion as in (3.60) suggest that an AR(4) model is most appropriate. For both the PLS criteria the same outcome emerges. The final column of Table 3.13 displays the true RMSE for the period 1981 through 1988. Unfortunately, it appears that now the AR(6) seems to have the best forecasting performance. Close inspection of the observations and the forecasts reveals that in 1981 a slight structural break upwards can be witnessed and that the estimated mean of the series by the AR(6) is higher than those of its competitors. Note that judgmental forecasting procedures could have gained from this information.

The forecasting criteria reviewed and applied thus far are measures of the expected variance of forecast errors, i.e. FPE and HFE, and of the true squared forecast errors observed for a validation sample, i.e. PLS and RMSE. In the terminology of Aznar (1989), the first two criteria measure the precision of the forecasts and the other two criteria are designed to measure the accuracy of the predictions. Aznar (1989) proposes a model selection strategy which combines measures of precision and accuracy. Its empirical relevance is yet to be investigated.

Description criteria

Descriptive models are designed to be descriptive about the underlying data generating process. One can imagine however that there are several ways to define an appropriate description measure, and often there are as many model selection criteria. Extensive derivations of the criteria can be found in the original studies, and since there are already some excellent surveys on the subject, see for example De Gooijer et al. (1985), I will highlight only the major issues and some recent advances.

The criterion for the selection of an AR model given in Parzen (1974) formalizes the idea that any AR($p$) model is an approximation to an AR($\infty$) model. Hence, the adequacy of such approximations can be evaluated with a certain distance measure. After several modifications in subsequent studies by Parzen and others the resulting Criterion Autoregressive Transfer Function (CAT) is now given by

$$\text{CAT}(k) = (n^{-1} \sum_{j=0}^{k} \hat{\sigma}_M^2(j)) - \hat{\sigma}_M^2(k).$$  \hspace{1cm} (3.61)
The major drawback of this criterion is that it is not readily extendable to MA or ARMA processes.

A criterion with a much wider range of applicability is proposed in Akaike (1974). The information measure chosen is a measure of entropy, which, loosely speaking, is a goodness of fit measure of the model relative to the true data generating process. Maximizing this entropy is equivalent to minimizing the Kullback and Leibler (1951) mean information measure. With some additional assumptions, approximations and rewriting, the often applied Akaike Information Criterion (AIC) in case of autoregressive models can be given by

\[ \text{AIC}(k) = n \log \delta_{ML}^2 + 2k. \]  

(3.62)

This expression clearly reflects a balance between the number of parameters and the goodness of fit. This may be important since it is fairly easy to show that too small as well as too large a model can yield improper predictions, see, e.g., Steerneman (1987). Since the entropy measure can be formulated for several types of models, the expression in (3.62) can be modified for application in, e.g., ARMA models, multivariate time series, and regression models. For comparability purposes, i.e. with the expressions (3.51) through (3.60), it may be convenient to rewrite (3.62) as

\[ \text{AIC}^*(k) = \delta_{ML}^2 \exp(2k/n). \]  

(3.63)

Note that limiting the expansion of \( \exp(2k/n) \) to its first two terms results in the criterion given in (3.56). So, it can be expected that the AIC and the FPE criterion will often yield the same model selection. The problem with the AIC criterion is that, if one is prepared to assume that the true model is amongst the considered models, it is not consistent (cf. Shibata 1976). From Monte Carlo studies it often appears that the model selected by the AIC is too large. For a further critical evaluation of the AIC criterion, see Chow (1981).

The measure of description, to be called discrepancy, proposed in Linhart and Zucchini (1982, 1986) is related to those measures for the CAT and AIC criteria, although there are some conceptual differences. Again it is assumed that there is an unknown true data generating process, and that from a set of models the model 'closest' to this true model is to be preferred. The total discrepancy between these two models consists of the
discrepancy due to approximation of the 'best' model to the true model, and of the discrepancy due to estimation of the former. An operating model is defined to be a stochastic model with a structure that is similar to the structure of the data. One is assumed to be able to find a reasonable operating model, which in principle can be estimated from the data, albeit most of the times poorly. Because of the complexity of the operating model \( M \) a simpler model \( M(\theta) \) has to be fitted. The quality of this fit is judged by the discrepancy, which is defined by

\[
\Delta\{M(\theta), M\} \geq 0. \tag{3.64}
\]

The choice of the function \( \Delta \) depends on the purpose of the statistical analysis.

It is assumed that there exists a unique \( \theta^* \) where the discrepancy in (3.64) obtains a minimum value. The discrepancy between \( M(\theta^*) \), i.e. the 'best' model of all possible \( M(\theta) \), and \( M \) is called the discrepancy due to approximation, or

\[
\Delta\{M(\theta^*), M\}. \tag{3.65}
\]

With an estimator \( \hat{\theta} \) for \( \theta^* \) one can select a model from the family \( M(\theta) \) with the available data. The discrepancy due to estimation between the ultimately fitted model and the best model in the set is denoted by

\[
\Delta\{M(\hat{\theta}), M(\theta^*)\}. \tag{3.66}
\]

The overall discrepancy \( \Delta\{M(\hat{\theta}), M\} \) is the sum of (3.65) and (3.66). Inherent in this measure is again the balance between the number of parameters and the goodness of fit, because as the number of parameters increases, the expression in (3.65) decreases, but generally the discrepancy in (3.66) will increase. The measure \( \Delta\{M(\hat{\theta}), M\} \) includes the unknown \( M \), so it is proposed to work with the expected overall discrepancy. The latter may again depend on \( M \), but in many applications it can be estimated. Final model selection decisions will then be based on an estimator of this expected \( \Delta\{M(\hat{\theta}), M\} \), which is called a criterion.

The choice of the function \( \Delta \) largely determines the final selection criterion. When it comes to AR model selection, Linhart and Zucchini (1986) consider a measure which is based on spectral densities, and they end up
with a criterion similar to (3.56). In Hurvich and Tsai (1989) it is shown that the choice for the Kullback and Leibler (1951) measure may yield other results. In that case the criterion for AR(p) model selection is

\[ \text{AIC}_C(k) = n \log \hat{\sigma}^2_{ML} + n(n+k)/(n-k-2). \]  

(3.67)

From the results of several small sample simulation experiments it emerges that this AIC$_C$ criterion closely follows the true discrepancy while the AIC and many other criteria fail to follow especially when the true model order is large. Surprisingly it appears that this modification is similar to the one proposed in Sugiura (1978), albeit based on distinct derivations. From (3.67) it is easy to see that the penalty of too many parameters of AIC$_C$ (or AIC$_C^*$) exceeds that of AIC (or AIC$^*$).

The derivations of the above description criteria all assume a true, but unknown, data generating process, and hence all necessitate some approximation to this model. The criteria based on Bayesian methods, which consider model selection between pairs of models, circumvent this problem. See Zellner (1978) for a comparison of AIC and a Bayesian criterion. For Bayesian inference it is necessary to a priori have some knowledge of the parameters in the model. Together with the information in the data one may now be able to calculate the posterior probabilities of the models. The well-known criterion developed in Schwarz (1978) however ignores the part related to the prior density in the likelihood function, see Chow (1981) for a simpler derivation of this criterion. The SC criterion is given by

\[ \text{SC}(k) = n \log \hat{\sigma}^2_{ML} + k \log n, \]  

(3.68)

or

\[ \text{SC}^*(k) = \hat{\sigma}^2_{ML} n^{k/n}, \]  

(3.69)

the latter of which is the form in which it is comparable with (3.63). It appears that the criterion proposed in Rissanen (1978) is the same although it is derived from computer technical arguments.

Comparing the expressions in (3.62) and (3.68) it is clear that when \( n \geq 8 \) the SC criterion penalizes the inclusion of regressors more than the AIC does. This means that the model selected with the SC criterion is usually smaller than or equal to the model selected with the AIC. There are other criteria, some of which are also based on Bayesian arguments, which
can be viewed as intermediate cases between these two criteria. Generally, alternative criteria have penalty terms for the number of parameters other than 2k or klogn. For example, in Hannan and Quinn (1979) it is proposed to consider the penalty term k(1+clogn), although there is no indication which value of c is most convenient. From the results in Franses (1989a) it appears that in most instances, i.e. most combinations of values of k and n, the AIC and SC criteria cover the range of penalties of model selection criteria. In a linear regression context it is argued in Franses (1989a) that when a model is preferred by both criteria, it can be shown that possibly omitted variables, or in fact the unknown true model, has the least effect on model selection.

As to the example of selecting a descriptive AR model for the $U_t$ series, things have become rather easy. With the FPE criterion one would already have selected an AR(4) model. Given the similarity between the AIC and the FPE criteria, and since this AR(4) model is the smallest amongst the candidate models, it is clear that also with all description criteria the AR(4) model will be preferred, as can be seen from the first three columns of table 3.14.

<table>
<thead>
<tr>
<th>Model</th>
<th>All Observations</th>
<th>Without 1932,1947,1954</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AIC*</td>
<td>AICc*</td>
</tr>
<tr>
<td>AR(4)</td>
<td>0.1052</td>
<td>0.3027</td>
</tr>
<tr>
<td>AR(5)</td>
<td>0.1079</td>
<td>0.3133</td>
</tr>
<tr>
<td>AR(6)</td>
<td>0.1083</td>
<td>0.3181</td>
</tr>
</tbody>
</table>

The expressions for AIC* and SC* are (3.63) and (3.69), while that for AICc* can be derived from (3.67).

It turns out that in this case one and the same model is preferred for forecasting as well as for description. Note that if model selection would have been carried out using a holdout sample, say 1981-1988, the AR(6) model would have been chosen. On the other hand, it may also be that if the
holdout sample was extended to, say, 1990 again different conclusions were drawn. So, it does matter at which point in time one starts selecting models. This may provide an additional argument for initiating the model selection exercise anytime new observations become available.

Influential observations and model selection

An issue which may also have an effect on model selection is the presence of influential observations. In Franses (1990a) it is shown that these observations are not necessarily outliers or influential observations in the sense that they can be found using the techniques developed, e.g., in Belsley, Kuh and Welsch (1980). The detection of choice influencing observations in the linear regression context is studied in Weisberg (1981) and Chatterjee and Hadi (1988). Weisberg (1981) proposes to partition a model selection criterion into individual components to assess the influence of observations on subset model selection. In Chatterjee and Hadi (1988) the impact of simultaneously omitting one variable and one observation is considered. In Franses (1990a) a simple method is proposed which tries to extend these approaches since model selection can involve more than one variable and more than one selection criterion. This method has been designed for the regression case, but when applied with caution, and while other statistics have indicated that there are no outliers in the individual models (see figure 3.6), it may be useful in AR model selection.

Suppose there are two rival models $M_1$ and $M_2$, where the first involves $k_1 + k_2$ variables and the second $k_2$, and that $M_2$ is nested in $M_1$. The choice between $M_1$ and $M_2$ can be made using model selection criteria belonging to the class of criteria which are functions of the estimated residual sums of squares (RSS). A typical form of these criteria is to choose $M_1$ if

$$RSS_2 > q.RSS_1,$$  \hspace{1cm} (3.70)

where $q$ is a function of $n$, $k_1$ and $k_2$. A summary of $q$ values for several criteria is given in Franses (1990a, table 1). Given the expressions in (3.63) and (3.69), it is easy to derive that the $q$ values for the Akaike and Schwarz criteria can be given by $\exp(2k_2/n)$ and $n^{k_2/n}$, respectively. From (3.70) it is easy to see that a larger $q$ reduces the probability of choosing $M_1$.

With respect to the influence of a single observation two distinct
situations can be recognized. The first is that $M_1$ is chosen with $n$
observations and $M_2$ with $n-1$ observations (case 1). The second is that $M_2$
is chosen on the whole data set and that $M_1$ is preferred after deletion of
one observation (case 2). Denote $z_{i(1)}$ as the scalar which is computed
without using the $i$-th row of the data set, and $z_i$ as the $i$-th element of a
vector. Furthermore, denote $h_{ii}$ as the $(i,i)$-th element of the hat matrix
$X'(X'X)^{-1}X'$, where $X$ contains the explanatory variables in $M_1$, and $h_{2ii}$ as
the corresponding element for $M_2$. Case 1 occurs for observation $i$ when

$$RSS_2 > q.RSS_1,$$
and

$$RSS_{2(1)} < q_{(i)}RSS_{1(1)}.$$  

(3.71)

Since $q_{(i)}$ and $q$ exceed 1, and

$$RSS_{j(i)} = RSS_j - \hat{e}^2_{ji}/(1-h_{ji}), \quad j = 1,2,$$  

(3.72)

and where $\hat{e}^2_{ji}$ denotes the squared estimated $i$-th residual for model $j$,
(3.71) can be transformed into

$$RSS_2 > q.RSS_1 > (q/q_{(i)})RSS_2 + q_{(i)}f_{1i} - (q/q_{(i)})f_{2i},$$  

(3.73)

where $f_{ji} = \hat{e}^2_{ji}/(1-h_{ji})$. This inequality can not hold when $RSS_2$ does not
exceed the expression entirely on the right hand side of (3.73), or

$$f_{2i} < (1 - (q_{(i)}/q))RSS_2 + q_{(i)}f_{1i}.$$  

(3.74)

A similar result can be obtained for case 2, where now the $f_{2i}$ should
exceed the expression on the right hand side of (3.74).

The line $L = \{(f_1,f_2)\in(R_0,R)_2 | f_2 = (1 - (q_{(i)}/q))RSS_2 + q_{(i)}f_{1i}\}$ can now
be drawn. For case 1, all points that lie under this line $L$, i.e. those
that satisfy (3.74), are not influential for model choice. All $(f_{1i},f_{2i})$
lying above $L$ deserve attention for they may be influential. For these it
will be true that observations with the largest distance to $L$ will have the
largest probability of being influential. For case 2 the reverse argument
holds. Of course, one could just calculate these distances and skip the
graphical part, but I feel that a graphical device is often insightful.

The examples in Franses (1990a) to illustrate this method are all
taken from well-known data sets in the outlier detection literature. For an example in Belsley, Kuh and Welsch (1980) it can be argued that a previously unrecognized observation appears to be influential for model selection. Furthermore, with examples taken from Weisberg (1980) and Chatterjee and Hadi (1988) it emerges that it matters with which criteria a model is selected, and that the method is able to detect more than one influential observation for model choice. One should however be aware that inference on this last issue may be blurred by, e.g., highly correlated observations. The first issue ensures that some inference can be made with respect to the robustness and preferable use of model selection criteria, see also Franses (1989b) and Jungelges (1989). For case 1 it applies that the criterion with minimum q value is most robust in the sense that model choice is the least effected by influential observations. Analogously this applies to case 2 and the criterion with maximum q value. Intuitively, one would therefore be more confident in model selection results when a certain model is chosen with all criteria, and when there are no observations influential enough to change this. So, in case m possibly influential observations can be depicted, it seems preferable to choose the same model on the basis of n and n-m observations with all criteria.

The selection of an AR model for the $U_t$ series appears to be an example of case 2, i.e. the smaller model is preferred. I have chosen to display the plot for the selection between the AR(4) and AR(6) models in figure 3.8 below, i.e. the combinations ($f_{AR6}$/$f_{AR4}$). The line corresponds to the AIC criterion, and the line for SC is very similar. From this figure it emerges that the observations in 1932, 1947 and 1954 may be influential. Unreported calculations show that the individual observations do not matter, but that the three observations simultaneously can have an effect on model selection, as can be seen from the second part of table 3.14. From these results, which have been found by including three dummy variables into the regressions, it is also clear that now it matters which criterion is used when selecting a model. The SC criterion yields the same choice, while the other criteria show different outcomes. Since not all criteria prefer another model, there is reasonable confidence in the selection of the AR(4) model as the most descriptive model.

This example illustrates the possible fragility of model selection. An implication, which may be useful for empirical dynamic multivariate model building where univariate time series models are required, is to consider more than one initial model for each variable, as is done for example in
tions disappear in the presence of other variables. Model selection in such econometric models can then be carried out along similar lines as in this chapter.

![Figure 3.8](image)

The detection of observations possibly influential for model selection

The example of the selection of an AR model for the U.S. unemployment series turns out to be a simple example. Model selection can become more complicated in case also MA terms are involved. Some choice criteria can easily be extended just by including the number of MA parameters in $k$, but others are not available in that case. Another problem is that models are not necessarily nested. Furthermore, the method to detect influential observations is not extendable to ARMA models in another way than via an AR approximation.
Testing an AR(1) model versus an MA(1) model

Things can become even more complicated when the number of parameters is equal in the models and hence the balance between fit and parameters can not be made. An example is the choice between an AR(1) and an MA(1) model, which has received some attention recently. This selection can be important in several occasions, e.g., using one model while the other is appropriate can have an effect on forecasting (cf. Magnus and Pesaran 1989), and on modeling disturbances in a linear regression model (cf. King and McAleer 1987). The latter study also provides a test procedure to distinguish between AR(1) and MA(1), in case one is prepared to assume that the errors are positively correlated. In Burke, Godfrey and Tremayne (1990) an alternative test procedure is proposed which is based on the sample autocorrelations. From various simulations however it emerges that its empirical size and power can be rather unsatisfactory. In Smith and Tremayne (1990) some finite sample results for the test statistics in Burke et al. (1990) is provided. Monte Carlo repetitions indicate that now the power can be somewhat improved by small sample modifications, although the gain is not impressive. An alternative model choice device may therefore be useful. Below, I will propose such a method which is based on the idea to fit both models to the data and to compare the estimated residual sums of squares. Note that in case the models are used for forecasting one can simply use, e.g., the PLS or HFPE criteria.

Consider the AR(1) and the MA(1) models for a zero mean time series \( y_t \) which can be given by

\[
y_t - \phi y_{t-1} = \varepsilon_t \quad (3.75)
\]

\[
y_t = \varepsilon_t + \theta \varepsilon_{t-1} \quad (3.76)
\]

and assume that \( 0 < |\phi|, |\theta| < 1 \), which excludes the white noise model. The theoretical autocorrelations of the AR(1) model are \( \rho_j = \phi^j \), for \( j = 0, 1, 2, \ldots \), and the only nonzero autocorrelations for the MA(1) model are \( \rho_0 = 1 \) and \( \rho_j = \theta / (1 + \theta^2) \). Define the multiple correlation coefficient \( R^2 \) as the proportion of the variability of \( y_t \) explained by a stationary invertible time series model as

\[
R^2 = 1 - \frac{\sigma^2}{\sigma^2_y} \quad (3.77)
\]
where $\sigma^2_\gamma$ denotes the variance of $\gamma_t$. To investigate whether the use of the $R^2$ yields proper selection, one may consider the asymptotic distribution of the estimated $R^2$ under both hypotheses. The $R^2$ in (3.77) can be estimated by

$$\hat{R}^2 = 1 - \frac{\sigma^2_e}{\sigma^2_\gamma}$$  \hspace{1cm} (3.78)

Hosking (1979) derives that for a zero mean series this $\hat{R}^2$ asymptotically follows a normal distribution given by

$$\hat{R}^2 \sim N(R^2, 4(\sigma^2_e/\sigma^2_\gamma)\sum_{r=1}^\infty \rho^2_r/n),$$  \hspace{1cm} (3.79)

where $n$ denotes the number of observations.

The asymptotic distribution of the $\hat{R}^2$ for the AR(1) model is easily found. For this model it applies that

$$\sigma^2_\gamma = \sigma^2_e/(1-\phi^2),$$  \hspace{1cm} (3.80)

$$R^2_{AR} = \phi^2,$$  \hspace{1cm} (3.81)

and that

$$\sum_{r=1}^\infty \rho^2_r = \rho^2_1 + \rho^2_2 + \rho^2_3 + \ldots \hspace{1cm} (3.82)$$

$$= \rho^2_1 + \phi^2 \rho^2_1 + \phi^4 \rho^2_1 + \ldots$$

$$= \phi^2/(1-\phi^2),$$

and so

$$\text{var} \hat{R}^2 = 4(\sigma^2_e/\sigma^2_\gamma)\sum_{r=1}^\infty \rho^2_r/n = 4\phi^2/n,$$

or

$$\hat{R}^2 \sim N(\phi^2, 4\phi^2/n).$$  \hspace{1cm} (3.83)

Given that under the AR(1) null hypothesis, $\phi^2$ equals $\rho_2$, and hence that no model has to be fitted to establish this distribution, it can be observed from (3.83) that it is convenient to design a test procedure for an AR(1) versus an MA(1) model. In fact, the distribution of the estimated $R^2$ for an
MA(1) model would include $\rho_1$ as well as $\theta$ (see (3.88)). From (3.83) it is clear that under the null hypothesis of an AR(1) model, the statistic

$$ Q = n\nu_1 (\hat{R}^2 - \rho_2) / 2\sigma_1, $$

(3.84)

asymptotically follows a $N(0, 1)$ distribution.

An empirical test statistic for an AR(1) model versus an MA(1) model may be given by the sample analogue of $Q$. Given that the $\hat{R}^2$ under the null hypothesis of an AR(1) model can be estimated by $r_1^2$, the appropriate test statistic is now

$$ Q_{AR} = n\nu_2 (r_1^2 - r_2) / 2r_1, $$

(3.85)

which is a two-sided test statistic. Note that the assumption of no white noise is crucial. Large absolute values of $Q_{AR}$ imply the rejection of the AR(1) null hypothesis.

The test statistic for the same hypothesis which is presented in Burke, Godfrey and Tremayne (1990) is based on the second order partial autocorrelation coefficient $\varphi_{22}$ which can be written as

$$ \varphi_{22} = (\rho_2 - \rho_1^2) / (1 - \rho_2), $$

(3.86)

(cf. Box and Jenkins 1970, p.65). For an AR(1) model is can be shown that approximately

$$ \hat{\varphi}_{22} \sim N(0, n^{-1}). $$

The resulting test statistic is now given by

$$ t_{AR} = n\nu_2 (r_2 - r_1^2) / (1 - r_2), $$

(3.87)

and it should follow an asymptotic $N(0, 1)$ distribution. Note that this $t_{AR}$ implies one-sided testing since under the MA(1) hypothesis the expression $(\rho_2 - \rho_1^2)$ is smaller than zero and $(1 - \rho_2)$ is always positive.

Before comparing these two test statistics, it is convenient to look again at the issue of interest. The $R^2$ of an AR(1) model has already been given in (3.81). The $R^2$ of an MA(1) model is

$$ R^2_{MA, \theta} = \frac{\theta^2}{1 + \theta^2} = \theta \rho_1. $$

(3.88)
It may now be interesting to carry out some misspecification analysis, i.e. to see what are the effects of considering one model while the other is the DGP. In case the MA(1) model is true, and an AR(1) model is used, then it is clear that its $R^2$, say $R^2_{\text{ARIMA}}$, can be derived from the $\rho_1$ in (3.88), or $R^2_{\text{MA1AR}} = \rho_1^2$. Since for the MA(1) model it applies that always $|\theta| > |\rho_1|$, it is easy to recognize that $R^2_{\text{MA1AR}} > R^2_{\text{ARIMA}}$; see also Nelson (1976). When the AR(1) is the DGP the $R^2_{\text{AR1AR}}$ is equal to $\rho_1^2$. In case an MA(1) is erroneously used then

$$R^2_{\text{MA1AR}} = \theta \rho_1 = \frac{1}{2} \pm \frac{1}{2} (1-4\rho_1^2)^{1/2},$$

(3.39)

and then if $|\rho_1| \leq 0.5$ it applies that $R^2_{\text{MA1AR}} \geq R^2_{\text{AR1AR}}$. In summary, if $|\rho_1|$ exceeds 0.5, an MA(1) model can not be appropriate. And, for smaller values of $|\rho_1|$ it appears that the MA(1) model has a better fit anyhow, i.e. whether it is the DGP or not.

Table 3.15
Empirical size and power of statistics for testing an AR(1) versus an MA(1)

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Parameter values</th>
<th>Nominal Size</th>
<th>Rejection rate$^{(1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_t=\phi y_{t-1}+\varepsilon_t$ (size)</td>
<td>$\phi = 0.9$</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>$y_t=\varepsilon_t+\theta \varepsilon_{t-1}$ (power)</td>
<td>$\theta = -0.9$</td>
<td>0.05</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>-0.5</td>
<td>0.05</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>-0.2</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>0.26</td>
</tr>
</tbody>
</table>

$^{(1)}$ The test statistics are given in (3.87) and (3.85), respectively.
These results should have an implication on the design and performance of the test statistics. An ideal situation would be that when the parameter $|\phi|$ of the AR(1) approaches unity, or in fact exceeds 0.5, the size of a test should go to zero. From (3.84) it can be seen that it may be the case for the test statistic $Q_{AR}$, but from (3.80) it appears that this is obviously not the case for the $t_{AR}$ statistic. When in an AR(1) model the parameter is close to 1, the $\rho_2$ will also be close to 1, thus providing that the denominator $(1-\rho_2)$ becomes small and the value of $t_{AR}$ large. The latter may induce that the null hypothesis can be rejected too often in case the null is correct. This is also recognized in Burke et al. (1990), and therefore they propose an additional step, which is to check whether $|\rho_1| > 0.5$. Empirical evidence for this size problem for the $t_{AR}$ test can be found in Burke et al. (1990, p.141) and in Smith and Tremayne (1990, tables 1 through 3, columns 4). On the other hand, it can be deduced from the expression in (3.89) that for an AR(1) model with a small value for $|\phi|$ the size of the $Q_{AR}$ test is likely to increase.

With respect to the power of the statistics it can be argued that for all $\theta$ the denominator in $t_{AR}$ would theoretically be equal to 1, while that of $Q_{AR}$ will decrease in case $|\theta|$ gets smaller. Since for smaller absolute values of $\theta$ it is likely that estimation biases will increase, it can be expected that the power of $t_{AR}$ will decrease when $|\theta|$ goes to zero, while that of $Q_{AR}$ may not deteriorate that much.

To investigate the performance of the $Q_{AR}$ test relative to the $t_{AR}$ test, several simulations have been carried out. For a sample size of 100, 100 replications of AR(1) and MA(1) processes have been generated. From the simulation results in Smith and Tremayne (1990) it appears that empirical size and power results are symmetrical across positive or negative values for the parameters, so only a selection of possible DGP's is considered. The results are displayed in table 3.15. They show patterns which are close to what could have been expected. Of course, one could do more simulations for a larger number of DGP's, but the general conclusion to be drawn from the results in table 3.15 is that the $Q_{AR}$ test may be preferred above the $t_{AR}$ test.