Contents

Chapter 6

Model Building and Prediction

6.1 Introduction

- In Chapter 3 we investigated the identification of time series models using the autocorrelation and partial correlation functions.

- In Chapter 5, we discussed testing methodology that can be applied in the context of ARMA models

- These together with the portmanteau tests and prediction error evaluation introduced later in this chapter constitute what is commonly called **Box-Jenkins analysis** (Identification, Estimation and Diagnostic Checking).

- The basic approach is to use the autocorrelation and partial correlation functions to get a list of candidate ARIMA models.

- These models are then held as a tentative or candidate models which are subjected to some diagnostic or other model selection criteria.

- In econometrics there has been a sizeable development of tools to evaluate an estimated specification.

- Presently, these techniques are being brought to bear on time series models.

- The idea is simple: is the estimated specification an **adequate representation** of the time series process? What are the properties of the errors?; the forecasts?; the stability of the parameters?; etc.

- Certainly, all of the traditional questions we apply to regression models can be raised for pure time series specifications.

- An important consideration in univariate time series modelling is the stationarity of the series. The prescribed method of dealing with nonstationary time series is first differencing until stationarity is achieved.

- Identifying whether a time series is **first difference** stationary was left more as an art form.
- Basically, the technique was to look at both the size of the autocorrelation coefficient and the persistence in the correlations.
- Unit root time series have an infinite memory and hence do not die out at longer lags.
- The techniques were at best informal and the idea seemed to evolve that as the researcher got more familiar with time series he/she would get the 'feel' for when a particular time series needs to be first differenced.

6.2 Model Selection Criteria and Diagnostic Testing

- For the discussion below, we will presume that the series are stationary and leave the discussion of formal testing for nonstationary time series later.

6.2.1 Parsimony

- One very important criterion in model selection is parsimony.
- The idea is simple: the preferred model is one in which the key properties of the data (the time dependencies) are captured 'adequately' in as small (parsimonious) a model as possible.
- The intention in Box-Jenkins has always been to build small (few parameters) models.
- Hence, everything else equal, we tend to favour smaller models.
- Not only are they simpler to handle, but when their performance breaks down, it is typically an easier task to figure out why.

6.2.2 Randomness of the Disturbances

- Since the goal in model building has always been to uncover the systematic part of the dependent variable, a natural place to consider success is in the disturbance term.
- In time series modeling, we are especially interested in capturing the time dependencies in the data and hence one obvious criterion is randomness on the errors.

- There are generally 2 types of tests for randomness applied in the literature.
- Let us assume that we have a tentative ARMA model estimated of some chosen order in p and q.
- Denote the residuals at time t as e_t, which obviously depend on $\hat{\theta}$ and $\hat{\phi}$, the estimated coefficients.
- Our goal is to test whether the residuals are **white noise** or random.
- If they are not, we interpret this as evidence that our model is **misspecified** and that further modeling or re-thinking is required.
- As usual we have to make allowance for the fact that we are dealing with residuals e_t rather than the actual (unobservable) errors ϵ_t. We will have more to say on this.

6.2.3 A Test for Randomness using ϵ_t

- For the moment, suppose we had observations directly on the ϵ_t.
- The basic idea is to examine the correlations in these and determine whether these are significantly different from zero.
- $$\begin{aligned} H_0 &: \text{ errors are random} \\ H_1 &: \text{ errors are not} \end{aligned}$$

  Notice that the alternative is extremely general and that there are a number of alternatives that could give rise to a rejection of the null hypothesis (for instance, structural breaks or time-dependent variances).
- Hence these kinds of general tests are called **portmanteau-tests.**
- If H_0 is true, then the sample correlation coefficient $r(h)$ of the errors (this is just the regression of the ϵ_t on ϵ_{t-h} for ($h = 1, 2, \ldots$) has a limiting normal distribution with mean zero and variance:

  $$\sigma^2 p \lim(T^{-1} \sum \varepsilon_{t-h}^2)^{-1} = 1$$
- Hence we may test **individually** whether $\rho(h)$ is zero by

  $$T^{\frac{1}{2}} r(h) \sim N(0,1) \text{ under } H_0 \text{ (asymptotically)}$$
- That is at the 5% level $\mid T^{\frac{1}{2}} r(h) \mid >$ 1.96 then reject H_0.
- Of course, if we were to do 20 of these ($h = 1, \ldots 20$), we should then expect to reject one out of the 20 even if the null were true and if the tests were independent.

6.2.4 A Problem with using Residuals

- A difficulty with this kind of test is we do not have the actual disturbances but only the residuals (e_t) and unfortunately these are not the same.

- For example, if the model is $AR(1)$, the first-order autocorrelation of the residuals has a variance of $\frac{\phi^2}{T}$, compared against $\frac{1}{T}$ just described for white noise.

- The situation improves somewhat for more complex ARMA(p,q) models, since the bias falls at higher lags for $r(h)$.

- The rule of thumb is that the test results individually tend to have overstated variances and hence reject the null too infrequently.

6.2.5 Box-Pearce Q Statistic

- A common test (called the Box-Pearce Q statistic) is:

$$Q \;=\; T\sum_{h=1}^{M} r^2(h) \sim \chi^2_M \text{ (under } H_0) \qquad (6.1)$$

  where M is the number of autocorrelations (based on the residuals) calculated.

- This is selected by the researcher (although common choices are $M \;=\; 12, 24, 36$).

- The issue in the choice of M is power.

- If too high a choice is made and the error is really, say and $MA(1)$, then we will introduce a number of correlations that are zero and lose power.

- On the other hand, too small an M might overlook some correlations at higher h, particularly if there are seasonal effects.

- Note that for this test, we have taken advantage of the fact that the $r(h)$ under the null is standardized normal and independent.

- As indicated in (6.1), the test is asymptotic and unfortunately the small sample properties are poor.

- A **modified** Box-Pearce statistic (due to Ljung and Box) is:

$$Q^* \;=\; T(T+2)\sum_{h=1}^{M}(T-h)^{-1}r^2(h) \sim \chi^2_M \text{ (under } H_0) \qquad (6.2)$$

  where the idea is $(T-h)/T(T+2)$ gives a better approximation to the variance than does $1/T$.

- In fact, these tests be motivated as an Lagrange Multiplier (LM) test of an $AR(M)$ or $MA(M)$ against white noise.

- The alternative hypothesis is quite general and does not distinguish among AR, MA, or $ARMA$ alternatives (see Harvey).

- If the ARMA(p, q) model is first estimated and the residuals are tested for randomness. In this case, the rule is to choose the degrees of freedom to be $M - p - q$ for the χ^2.

- Both statistics (6.1) and (6.2) tend to favour larger models.

6.2.6 Cumulative Periodogram

- This test can be found in Harvey pp. 150 and Fuller pp. 285). In Chapter 8, we shall see that the white noise process has a flat spectrum and that the contributions from all frequencies are the same.

- Hence take the periodogram

$$I_T(\omega_k) = I_k \quad (\omega_k = 2\pi k, \quad k \;=\; 1, \ldots .n)$$

and construct the **normalized cumulative periodogram**:

$$C_k \;=\; \frac{\sum_{i=1}^{k} I_i}{\sum_{i=1}^{n} I_i} \qquad k \;=\; 1, \ldots n. \tag{6.3}$$

- Under the null hypothesis that the errors are white noise, the periodogram ordinates are multiples of independent χ^2 with 2 degrees of freedom. Thus, there are a number of goodness of fit type tests.

- The normalized cumulative periodogram for $k \;=\; 1, \ldots n$ has the same distribution as the ordered sample of size $n-1$ selected from the uniform distribution (0,1) whose distribution is given in Durbin, 1969.

- Notice that if we graphed C_k against k under H_0, the line should be a 45^o one.

- Departures indicating greater (or less) contributions of some frequencies.

- Construct two parallel lines to the 45^o determined by: $C \;=\; \pm c_0 + k/n$ where c_0 is the critical value from Durbin (1969).

- If the line from C_k goes outside the two lines, then we reject H_0 and conclude that the errors are not white.

6.2.7 Diagnostic Checking

- There is nothing stopping us from applying standard hypothesis tests to pure time series models (see Chapter 5).

- For instance, we could proceed from the most general model to the specific in a manner similar to standard regression models.

- There are several problems with this approach.

- One is the question of parsimony. Standard hypothesis tests have been found in finite sample to favour the larger models.

- If a goal of time series analysis is to capture the process with as small a model as possible, then religious adherence to classical hypothesis tests is likely to lead to large models.

- Two other problems with classical hypothesis tests for determining the order of the process are model redundancy (it is easy to define models that are redundant) and nonnested hypotheses (*i.e.* an $AR(1)$ and $MA(1)$ are not nested).

- We can test an ARMA(p, q) against and ARMA$(p + M, q)$ process:

$$H_0 : \phi_{p+1} \;=\; \phi_{p+2} \;=\; \ldots \;=\; \phi_{p+M} \;=\; 0 \tag{6.4}$$

- For this test, the LR procedure is extremely straightforward.

- The LM procedure requires that we regress the residual e_t obtained from estimating the ARMA(p, q) model against the full set of $p + M + q$ derivatives and under H_0, the resulting $T \times R^2$ (uncentered) is χ^2 with m degrees of freedom.

- Harvey demonstrates the ARMA(1,1) case against ARMA$(p + M, 1)$. Notice if q is zero (a pure AR model) then this is just a regression of e_t onto $y_{t-1} \ldots . y_{t-p-M}$.

- It may be easier to do either a Likelihood ratio test or a Wald test when q is non zero.

- There is, of course, also the issue of pretesting (for instance how did we first reach this stage of testing and model) and hence controlling for *Type I error.*

- A feature of selecting models on this basis is that the tests favour large models which leads to our model selection methods.

6.3 Model Selection or Information Criteria

- The basic idea of these kinds of criteria is to 'select' as opposed to test for the appropriate the order of the ARMA process (see the advanced Judge et al) .
- In a selection approach one model is chosen even though it could be very misleading and not supported by the data (it is simply the best of a poor lot)
- There are a number of criteria which typically have a Bayesian motivation.
- Let $k \; = \; p+q$ where k is the number of parameters estimated.

6.3.1 Final Prediction Error

- Based on Akaike(1969) the final prediction error (FPE) is:

$$FPE(k) = \frac{T+k}{T-k}\,\tilde{\sigma}_k^2 \tag{6.5}$$

- Under this criterion the order is chosen to: min $FPE(k)$.

6.3.2 Akaike Criterion (AIC)

- $$AIC(k) \;=\; \tilde{\sigma}_k^2 \exp\left[log \mid V \mid + \frac{2k}{T}\right] \tag{6.6}$$

  where V is the variance covariance matrix of $\boldsymbol{y}$.
- Most often $|V|$ is replaced by the identity martix I.
- Again we choose k to minimize AIC.
- It can be shown that :

  $$\ln FPE(k) \;=\; AIC + O(T^{-2}) \tag{6.7}$$

  and hence the two criteria are asymptotically equivalent. Notice the term $2k/T$ assigns a penalty to models that are not parsimonious.
- We typically take logs of criterion (6.6) to give :

  $$AIC(k) \;=\; \ln \tilde{\sigma}_k^2 + \frac{2k}{T} \tag{6.8}$$

- These criteria are also applied in multivariate models.

6.3.3 Schwarz Criterion (SC or BIC)

- $$SC(k) \;=\; -2ln(\text{ maximum likelihood }) + (lnT)k \tag{6.9}$$

which is to be minimized.

If the process is Gaussian then k is chosen to minimize:

$$SC(k) \;=\; ln\,\tilde{\sigma}_k^2 + \frac{k\;lnT}{T} \tag{6.10}$$

- Tsay (1984, Annals of Statistics, Vol 12 1425-1433) showed that for the general Information Criterion (IC) equation

  $$IC(k) = ln\,\tilde{\sigma}_k^2 + \frac{c_T}{T}k \tag{6.11}$$

  that if $c_T \rightarrow \infty$ and $\frac{c_T}{T} \rightarrow 0$ as $T \rightarrow \infty$, then the k minimizing (6.11) is consistent.

- This is true for both y_t: $I(0)$ and $I(1)$.

- In the case of AIC, we see that $c_T = 2$ and so is not consistent (leading to a nonzero probability of overfitting) whereas with SC $c_T = \ln(T)$ and is consistent.

- Of course, in finite samples there are not such hard results.

- A STATA ado file called select can be used to selct among ARMA models with a predetermined p and q with or without trend. This is a top down application

- Monte carlo results from a variety of authors have shown that there is not one dominating critieria.

- In univariate applications, both AIC and FPE have a tendency to overparameterize and consequently the SC is favoured.

- Interestingly, in the context of systems of equations (Vector Autoregressions-VAR) Gonzalo and Pitarakis (1999) have shown that the probability of overfitting is a decreasing function of the dimension of the system.

- Large systems will have less chance of overfitting.

- Their Monte carlo work establishes that AIC as the best preforming criterion for systems.

6.4 Automated Model ARIMA Selection Method

- In a paper Koreisha and Pukklia (1995, **Journal of Business and Economic Statistics**, pp. 127-131) reviewed some automated model selection criteria used for choosing the order of an ARMA(p, q) Box-Jenkin model. They also present their own method based on residual testing.

6.4.1 Hannan and Rissanen

First we consider a simple 3-stage procedure suggested by Hannan and Rissanen (1982, **Biometrika**, 81-94):

1. Estimate a 'long AR process' to obtain estimated residuals $\hat{\epsilon}_t$.
2. Establish a "preliminary maximum" by regressing y_t on p lags of y_t and q lags of $\hat{\epsilon}_t$, with the restriction $p = q = \tilde{p}$. Use SC to determine the "optimal" $\tilde{p}^*$.
3. Use a MLE procedure and SC to select the structure whose order is either in the set $\{p \ = \ \tilde{p}^*, q \leq \tilde{p}^*\}$ or $\{p \ \leq \ \tilde{p}^*, q = \tilde{p}^*\}$.

The Monte Carlo evidence from this procedure does suggest that it leads to too high a choice of $\tilde{p}^*$.

6.4.2 Residual White-Noise Autoregressive (RWNAR) Order-Determination Criterion

- Koreisha and Pukkila suggest a procedure that they claim is preferred to others and is appropriate for ARIMA(p, d, q) models.
- Their procedure is to systematically fit increasing orders of ARIMA structures using a three-stage generalized least squares method (we can use our standard MLE estimator).
- The idea is to choose the first model that is fit that delivers white noise errors.
- White noise errors are determined by fitting a high order $AR(k)$ process and determining what k minimizes:

  $$\delta(k) \ = \ n \ln \hat{\sigma}^2_{k,p,d,q} + k \ ln(n) \tag{6.12}$$

  where

  $$n \ = \ T - d - p - q,$$

  $\hat{\sigma}^2_{k,p,d,q}$ is an estimate of the residual variance of the residuals in the $AR(k)$ regression, $k \ = \ 0, 1, 2, \ldots k^*$, and k^* is an a priori upper bound that can be set as $k^* \ = \ T^{\frac{1}{3}}$.
- The sequence in which parameters are included in the model is $j = p + d + q$, where $j = 0, 1, \ldots$.
- We choose the model that gives $k = 0$.
- This procedure is called residual white-noise autoregressive (RWNAR) order-determination criterion.
- Notice that this approach also identifies the degree of differencing necessary to induce stationarity in the data.

- Whenever one or more models with the same j yield white noise, the selection criterion is the one minimizing:

$$\delta(p,d,q) = n \ \ ln \ \hat{\sigma}_{p,d,q} + (p+q)\ln T, \tag{6.13}$$

  where $\hat{\sigma}_{p,d,q}$ is the estimated variance from the residuals of the ARIMA(p,d,q) fitted.

- A PhD student Fredrick Tremblay wrote an ado program tpo do this as his 2nd stage (impressive this was) called rwnar.ado

6.5 Prediction

- There is a long tradition in time series modeling that the proof of a model rests on its ability to predict.

- Keep in mind if the model is stationary the conditional forecast at distant horizons will approach the unconditional mean (WHY?)

- Certainly, a good or useful model is one in which there are accurate forecasts.

- Before considering the forecasting problem, we have to step back and develop the prediction decompositions.

- With this set-up, the analysis of forecasting and testing the forecast or prediction errors is direct

- See Chapter for for Kalman filtering, prediction error decomposition and exact likellihood

- Consider a set of dependent observations on $y : y_1, \ldots, y_T$ with $y \sim N(\mu, \sigma^2 V)$.

6.5.1 Deterministic Series

- Time series for which it is possible with observations $y_t, y_{t-1}, \ldots$ to predict the future observations with zero mean squared error are called **deterministic** or **singular** (in terms of the variance covariance matrix).

- Time series which have a non zero mean squared error are called **nondeterministic** or **regular.**

- We will presently investigate prediction within the context of ARMA models. First we should highlight the assumptions for the analysis.

6.6 Three Assumptions for Calculating MSE

1. The disturbances are NID. This implies that conditional expected value (a linear estimator) is the **optimal predictor**. See Judge (pp. 236) for a counter example.

2. The ARMA parameters are **known**. The same general rules which we develop for the case in which the parameter values are known applies in the estimated parameter situation. Of course, this assumption means we are ignoring another source of uncertainty.

3. We assume that all past and present disturbances are **known** $\epsilon_T, \epsilon_{T-1}, \ldots$ which implies an infinite realization of observations. If T is large, or the roots of the MA process are not near the unit circle, this is not restrictive.

- Let the ARMA(p, q) at time $T + k$ be given by

$$y_{T+k} \;=\; \phi_1 y_{T+k-1} + \ldots + \phi_p y_{T+k-p} + \epsilon_{T+k} + \ldots + \theta_q \epsilon_{T+k-q} \qquad (6.14)$$

- We assume that future values of ϵ_t are unknown and are not predictable (other than assigning them the value of zero which is their expected value).

6.6.1 Predictions for the $ARMA(p, q)$

- Predictions of y_t are through the recursion formula:

$$\tilde{y}_{T+k/T} \;=\; \phi_1 \tilde{y}_{T+k-1/T} + \ldots + \phi_p \tilde{y}_{T+k-p/T} + \tilde{\epsilon}_{T+k/T} + \ldots + \theta_q \tilde{\epsilon}_{T+k-q/T} \qquad k \;=\; 1, 2 \ldots \qquad (6.15)$$

where $\tilde{y}_{T+j/T} = y_{T+j}$ for all $j \leq 0$ and

$$\tilde{\epsilon}_{T+j/T} \;=\; \begin{cases} 0 & \text{for} \quad j > 0 \\ \epsilon_{T+j} & \text{for} \quad j \leq 0 \end{cases} \qquad (6.16)$$

6.6.2 Special Case: $AR(1)$

- Substituting in (6.15) and (6.16) for the $AR(1)$ process leads to :

$$\tilde{y}_{T+k/T} \;=\; \phi \tilde{y}_{T+k-1/T} \qquad k \;=\; 1, 2 \ldots \qquad (6.17)$$

- Since $\tilde{y}_{T/T} \;=\; y_T$, then we may substitute in (6.17) to give:

$$\tilde{y}_{T+k/T} \;=\; \phi^k y_T \qquad (6.18)$$

- Note that the predicted value declines geometrically in ϕ to zero (to the unconditional mean of zero).
- Also, the forecast function is similar to the autocovariance function.
- In general the $AR(p)$ process yields a **one step ahead linear predictor**

$$\tilde{y}_{T+1/T} = \sum_{i=1}^{p} \phi_i y_{T+1-i} \tag{6.19}$$

- The **two step ahead**
- $$\tilde{y}_{T+2/T} = \phi_1 \tilde{y}_{T+1/T} + \sum_{i=2}^{p} \phi_i y_{T+1-i} \tag{6.20}$$
- We may continue to feed in the forecasts until we arrive at the $k \geq p+1$ ahead forecast:

$$\tilde{y}_{T+k/T} = \sum_{i=1}^{p} \phi_i \tilde{y}_{T+k-i} \tag{6.21}$$

6.6.3 Special Case: $MA(1)$

- At time $T = 1$, the $MA(1)$ process is : $y_{T+1} = \epsilon_{T+1} + \theta\epsilon_T$.
- Since the ϵ_{T+1} is **unknown** it is set equal to zero and the forecast is:

$$\tilde{y}_{T+1/T} = \theta\epsilon_T \text{ and } \tilde{y}_{T+k/T} = 0 \text{ for all } k > 1. \tag{6.22}$$

- From this it should be obvious how to extend to a $MA(q)$ process and also how to handle mixed processes (see Fuller pp. 81-82 for the general ARMA process).

6.6.4 Special Case: $ARMA(2, 2)$

- For example the ARMA(2,2) leads to a one period ahead forecast:

$$\tilde{y}_{T+1/T} = \phi_1 y_T + \phi_2 y_{T-1} + \theta_1\epsilon_T + \theta_2\epsilon_{T-1} \tag{6.23}$$

- A two period ahead forecast:

$$\tilde{y}_{T+2/T} = \phi_1 \tilde{y}_{T+1/T} + \phi_2 y_T + \theta_2\epsilon_T \tag{6.24}$$

- And finally a $k > 2$ ahead forecast:

$$\tilde{y}_{T+k/T} = \phi_1 \tilde{y}_{T+k-1/T} + \phi_2 \tilde{y}_{T+k-2/T} \tag{6.25}$$

6.6.5 Determining MSE of Forecasts

- The trick to determining MSE, is to obtain the MA representation of the forecast and then simply calculate the variance.

- Given our 3 assumptions earlier, we have

$$MSE((\tilde{y}_{T+1/T}) = Var_T[(\tilde{y}_{T+1/T})]. \tag{6.26}$$

- That is the mean squared error of the prediction is the **same as the conditional variance $Var_T[(\tilde{y}_{T+1/T})]$.**

- For instance, the $MA(1)$ process we have the appropriate representation immediately and its easy to see using $\tilde{y}_{T+1/T} \ = \ \theta\epsilon_T$:

$$MSE(\tilde{y}_{T+1/T}) \ = \ E\left[(y_{T+1} - \tilde{y}_{T+1/T})^2\right] = E\left[\epsilon^2_{T+1}\right] \ = \ \sigma^2 \tag{6.27}$$

- For $k > 1$, since the forecast $\tilde{y}_{T+k/T} \ = \ 0$, then

$$MSE(\tilde{y}_{T+k/T}) \ = \ Var(y_{T+k}) \ = \ (1+\theta^2)\sigma^2$$

  which is the *unconditional variance.*

- This kind of calculation can be made for any $MA(q)$ process.

MSE for General $ARMA(p,q)$

- For mixed stationary ARMA processes we may obtain the infinite moving-average representation:

$$y_t \ = \ \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \tag{6.28}$$

  where we have previously discussed the determination of the $\psi_j's$ (see equations 3.21-3.25 for the $AR(2)$ and 3.46 for the general case).

- It would be very handy if STATA could approximate and provide directly the infinite moving average in (6.28)

- Given the form of (6.28), it follows that the optimal linear predictor is a linear function of the disturbances:

$$\tilde{y}_{T+k/T} \ = \ \psi^*_k \epsilon_T + \psi^*_{k+1}\epsilon_{T-1} + \ldots.. \tag{6.29}$$

  (again we assume that the future values of ϵ_T are unknown).

- The coefficients $\psi^*_k, \cdots$ are weights and if we wish to choose them to minimize the MSE it follows from:

  $$y_{T+k}-\tilde{y}_{T+k/T} \;=\; \epsilon_{T+k}+\psi_1\epsilon_{T+k-1}+\ldots+\psi_{k-1}\epsilon_{T+1}+(\psi_k-\psi^*_k)\epsilon_T+(\psi_{k+1}-\psi^*_{k+1})\epsilon_{T-1}+\ldots$$

  and

  $$MSE(\tilde{y}_{T+k/T}) \;=\; \sigma^2(1+\psi_1^2+\ldots+\psi_{k-1}^2)+\sigma^2\sum_{j=0}^{\infty}(\psi_{k+j}-\psi^*_{k+j})^2$$

  that the optimal choice is $\psi_{k+j} \;=\; \psi^*_{k+j}$ for all j.

- The **minimum mean square linear estimator** ($MMSLE$) is:

-
  $$\begin{aligned}\tilde{y}_{T+k/T} &= \sum_{j=0}^{\infty}\psi_{k+j}\epsilon_{T-j} &(6.30)\\ MSE(\tilde{y}_{T+k/T}) &= \sigma^2(1+\psi_1^2+\ldots+\psi_{k-1}^2)\end{aligned}$$

- Note that the predictions $\tilde{y}_{T+k/T}$ are the same as (6.15).

MSE of AR(1) from the MA Representation

- A simple example of this is the $AR(1)$ process (since its easy to obtain the MA representation)

  $$y_t \;=\; \phi y_{t-1}+\epsilon_t \;=\; \sum_{j=0}^{\infty}\phi^j\epsilon_{t-j} \qquad (6.31)$$

  so that in terms of the predictions above

  $$\psi_{k+j} \;=\; \phi^{k+j} \qquad (6.32)$$

  and

  $$\tilde{y}_{T+k/T} \;=\sum_{j=0}^{\infty}\phi^{k+j}\epsilon_{T-j}=\phi^k\sum_{j=0}^{\infty}\phi^j\epsilon_{T-j} \;=\; \phi^k y_T \qquad (6.33)$$

  Compare to (6.18).

- Also

-
  $$MSE(\tilde{y}_{T+k/T}) \;=\; \sigma^2(1+\phi^2+\ldots+\phi^{2(k-1)}) \;=\; \sigma^2\frac{1-\phi^{2k}}{1-\phi^2} \qquad (6.34)$$

- As we increase the forecast period ($k \to \infty$), this MSE converges to the unconditional variance of y_t: $MSE(\tilde{y}_{T+k/T}) \;=\; Var[y_{t+k}] \;=\; \sigma^2/(1-\phi^2)$ as $k \to \infty$

- Also since the ϵ_t is NID, the **confidence intervals for prediction** can be obtained using the asymptotic approximation:

$$y_{T+k} \;=\; \tilde{y}_{T+k/T} \pm Z_{\frac{\alpha}{2}} \;\sigma(1+\sum_{j=1}^{k-1}\psi_j^2)^{\frac{1}{2}} \tag{6.35}$$

- We can get Stata to do this calculation but we need to add observations outside sample (use $tsappend, add(\#)$) command and then calculate mse ($predict\ mse, mse$)

- Notice that the first observation forecasted the mse is equal to σ^2, the next will be the 2nd step MSE, and so on.

- This should *converge to the sample variance* of y_t

6.6.6 A word on Finite Sample Prediction

- In obtaining the infinite MA representations and doing the calculations, we have assumed essentially an infinite sample.

- Of course, this is not true in practice and we use $\epsilon_t(Y;\theta)$ (see 4.32 for the $MA(1)$ process and later section for an explanation on higher orders).

- For the $MA(1)$ process $y_t \;=\; \epsilon_t + \theta\epsilon_{t-1}$ we wrote $\epsilon_t = y_t - \theta\epsilon_{t-1}$ and through recursive substitution:

$$\epsilon_t(Y;\theta) \;=\; y_t + (-\theta)y_{t-1} + \theta^2 y_{t-2} + \ldots + (-\theta)^{t-1}y_1 + (-\theta)^t\epsilon_0 \tag{6.36}$$

  where

$$\epsilon_0 \;=\; \sum_{j=0}^{\infty}(-\theta)^j y_{-j} \tag{6.37}$$

  which we set to zero (its unconditional mean).

- Hence we replace ϵ_t with the recursive expression $\epsilon_t(Y;\theta)$ and $\epsilon_0 = 0$ (so that $\hat{y}_{T+1/T} = \theta\epsilon_T(Y;\theta)$), and substituting this in the MSE formula:

$$MSE(\hat{y}_{T+1/T}) \;=\; E\left[(\epsilon_{T+1} + \theta(\epsilon_T - \epsilon_T(\mathbf{Y};\theta)))^2\right] = \sigma^2\left(1+\theta^{2(T+1)}\right) \tag{6.38}$$

1. For large T and $\theta < 1$, the recursion formula with ϵ_0 set to zero does not lose much (*i.e.* MSE is converging to σ^2) and the difference between ϵ_t and $\epsilon_t(\mathbf{Y},\theta)$ is negligible.

2. For small T, the biases from using infinite realizations to develop MSE formulae and using the MSE for small data sets can be large.

3. Also the biases will depend on the "closeness" of θ to the unit root (*i.e.* how fast is the infinite MA dying out and hence how much are we losing by setting ϵ_0 to zero).

4. If $\theta = 1$ (the unit root on the MA case) then for the $MA(1)$

$$\hat{y}_{T+1} = \theta\epsilon_T(Y;\theta)$$

and

$$\begin{aligned} y_{T+1} - \hat{y}_{T+1} &= y_{T+1} - \theta\epsilon_T(Y;\theta) \\ &= \epsilon_{T+1} - (-\theta)^{T+1}\epsilon_0 \\ &= \epsilon_{T+1} \pm \epsilon_0 \quad \text{(since } \theta = 1) \end{aligned} \tag{6.39}$$

and remains dependent on ϵ_0 regardless of sample size. Hence for this example the MSE$(\hat{y}_{T+1/T}) = 2\sigma^2$ regardless of sample size.

5. The same problem arises for AR or mixed ARMA type processes. Kalman filtering methods or multi-step predictions methods discussed in (**??** - **??**) can be employed. That is, for given values of ϕ and θ see equation (4.37), we have the $\epsilon_t(Y;\phi,\theta)$ to substitute in the formula. We will return to this.

6.6.7 Prediction and Estimation

- There are really no critical differences when we move to the more realistic situation in which the parameter values of the model are unknown and have to be estimated.

- The forecasts are identical, except we substitute the estimated values for the actual.

- The substitution creates an additional source of uncertainty (the estimates are not equal to the actual). This is of course exactly the same problem in regression analysis. Fuller pp. 384 gives a general proof for consistency in the estimated parameter case. Fuller also suggests using the prediction variance formulae (based on known parameters) discussed earlier as approximations. STATA follows this suggestion.

- The finite sample properties of such approximations appear to be good for large T. Of course, the best procedure is to take into account the fact that the parameters are estimated (as we do in the standard case).

- Parametric bootstrap techniques are very useful to not only take into account the uncertainty from stochastic regressors but also to account for the uncertainy frpm parameter estimation. There are many instances documented for dramatic improvement in finite samples

- Keep in mind that we will be u*nderstating the standard errors* of prediction by ignoring the uncertainty associated with the estimated parameters.

- For example, the $AR(1)$ case the forecast (we will use ˆ to denote a forecast based upon estimated parameters denoted as ˜) is
-

$$\hat{y}_{T+k/T} \;=\; \tilde{\phi}^k y_T. \tag{6.40}$$

- In the general ARMA(p, q) structure (6.15), we just evaluate the expression at the MLE estimates and follow the forecasting rules developed earlier.
- We can decompose the **prediction error** for any ARMA model into two orthogonal components:

$$y_{T+k} - \hat{y}_{T+k/T} \;=\; (y_{T+k} - \tilde{y}_{T+k/T}) + (\tilde{y}_{T+k/T} - \hat{y}_{T+k/T}) \tag{6.41}$$

1. The first term is the prediction error for **known** parameters
2. The second term is the error due to estimation and is ignored in Stata.

- We can specialize (6.41) for specific processes: for instance the $AR(1)$ process is:

$$y_{T+k} - \hat{y}_{T+k/T} \;=\; (y_{T+k} - \tilde{y}_{T+k/T}) + (\phi^k - \tilde{\phi}^k)y_T \tag{6.42}$$

- We obtain the MSE for the one step-ahead predictor:

$$MSE(\hat{y}_{T+1/T}) \;= MSE(\tilde{y}_{T+1/T}) + y_T^2 E\left[(\phi - \tilde{\phi})^2\right] \tag{6.43}$$

  where we have **conditioned** on information available at time T, namely y_T.

- The last part of the second term may be replaced by its asymptotic variance and we have the approximation to the MSE:

$$MSE(\hat{y}_{T+1/T}) \cong \sigma^2 + \frac{y_T^2(1-\phi^2)}{T}. \tag{6.44}$$

- Some computer programs are based upon the OLS formula variance in which case the estimator looks like

$$mse(\hat{y}_{T+1/T}) \;=\; s^2\left[1 + \frac{y_T^2}{\sum_{t=2}^T y_{t-1}^2}\right]$$

- The relationship between the two is easy to establish as $T \to \infty$ then

$$\sum_{t=2}^T y_{t-1}^2 \to T\sigma^2/(1-\phi^2) \tag{6.45}$$

- If we want to see what the **average variance of the prediction error**, is we may replace y_T^2 by its expected value $\sigma^2/(1-\phi^2)$ and obtain the **asymptotic mean square error** ($AMSE$) which for the $AR(1)$ case is:

$$AMSE(\hat{y}_{T+1/T}) \;=\; \sigma^2(1+T^{-1}). \tag{6.46}$$

1. We note that the contribution arising from estimating the parameters of the model is $O(T^{-1})$ which is dominated by the terms for the optimal predictor for known parameters. This is just another restatement of the consistency result and is why we may use (asymptotically) the formulae with known parameter values.

2. Most programs just base the forecast variance (including multistep) on the OLS formulae for AR models

3. While asymptotically we can ignore the uncertainty in the prediction from using estimated parameters, finite sample results are improved when this is taken into account.

4. Bootstrapping procedures are being employed to jointly incorporate both the uncertainty from the errors and the parameter estimates.

6.7 Model Comparisons based upon Prediction

- Often we will evaluate a model on its ability to deliver "close" and "reliable" forecasts.

- One obvious feature of **stationary** models is that as the forecast horizon increases, the forecasts approach the unconditional mean. In some cases, this is just zero.

- Therefore, forecasting using time series models for the long term is **not a very sensible exercise.**

- These models have quite poor properties once the forecast period gets very long.

- In addition, if one is using selection criteria that favour large models there is a tendency for the forecast to be **dominated** by the long tails (the lags) of the model. It is this kind of problem that led some time series forecasters to adopt more Bayesian strategies in forecasting. Placing more weight on recent observations and down weighting (sometimes simply by a priori reasoning) more distant terms. We will return to this.

- Selecting models by their comparative ability to forecast is not straightforward. For instance how or what do we use for a loss function to assess forecast performance? What are the statistical bases for such functions and what is the relevant period over which forecast should be checked and compared? How should we incorporate the variability of the forecasts against the quality of the forecast predictions?

- One fact that reappears in time series analysis is that the standard errors are suspect. Forecasts often appear to have fairly tight confidence intervals and yet if we do an *ex-post* analysis, we find many more of the observations lying outside the region.

- Certainly if one is considering a variety of time series models it is always advisable to conduct a forecasting exercise. Often we will see that models that appear to be quite different in their time series structure will, lead to forecasts that are much more similar.

6.8 Testing Forecast Equality

- This test is based on a paper by Francis X. Diebold and Roberto S. Mariano ("Comparing Predictive Accuracy", Journal of Business and Economic Statistics, July 1995).

- Here we briefly discuss their approach.

- Examples of it are programmed using STATA as $forecast\_error.do$.

- Although the argument appears quite general McCracken (*Journal of Econometrics*, 2000) has shown that the a**symptotic distribution is only appropriate for nonnested models**.

- A refinement discussed below by Clark and West(2006, 2007) argue that most of the bias can be removed with an adjustment so that a gaussian distrobution comparison is still meanigful

- In the nested case, there is a singularity in the covariance matrix of the test statistic. For nested models using the normal asymptotic theory can in fact lead to substantial size distortion (toward the null hypothesis).

6.8.1 The Basic Idea

- The key idea here is to provide a means for comparison of forecasts using formal testing.

- Further, rather than simply seeking to minimize forecast error, any function of that forecast error can be used.

- This may be used to reflect the economic loss associated with the error, which may take any functional form.

- Given two series of forecasts $\{\hat{y}_{it}\}_{t=1}^{T}$, $\{\hat{y}_{jt}\}_{t=1}^{T}$ with corresponding errors $\{e_{it}\}_{t=1}^{T}$, $\{e_{jt}\}_{t=1}^{T}$ and corresponding loss functions $g(e_{it})$, $g(e_{jt})$, they test that the forecasts are equally accurate:

$$H_o : E[g(e_{it})] = E[g(e_{jt})] \tag{6.47}$$

or equivalently, defining:

$$d_t \equiv [g(e_{it}) - g(e_{it})] \tag{6.48}$$

$$H_o : E(d_t) = 0 \tag{6.49}$$

Asymptotic Distribution and Test Statistic

- Asymptotically under H_0,

$$\sqrt{T}(\bar{d}) \sim N(0, V_{\bar{d}}) \tag{6.50}$$

where:

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T}[g(e_{it}) - g(e_{jt})] \tag{6.51}$$

$$f_d(0) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma_d(\tau) \tag{6.52}$$

$$\gamma_d(\tau) = E[(d_t - \mu)(d_{t-\tau} - \mu)] \tag{6.53}$$

$f_d(0)$ is the spectral density of loss differential at frequency zero, and $\gamma_d(\tau)$ is the autocovariance of loss differential at displacement τ.

$$V_{\bar{d}} = 2\pi f_d(0) \tag{6.54}$$

- The large sample test statistic is:

$$S_1 = \frac{\bar{d}}{\sqrt{\frac{2\pi\hat{f}_d(0)}{T}}} \tag{6.55}$$

where $2\pi\hat{f}_d(0)$ is a consistent estimator of $2\pi f_d(0)$.

6.8.2 Estimation for Test Statistic

- A consistent estimate of $2\pi f_d(0)$ is:

$$2\pi\hat{f}_d(0) = \sum_{\tau=-(T-1)}^{T-1} w(\frac{\tau}{S(T)})\hat{\gamma}_d(\tau) \tag{6.56}$$

where:

$$\hat{\gamma}_d(\tau) = \frac{1}{T}\sum_{t=|\tau|+1}^{T}(d_t - \bar{d})(d_{t-|\tau|} - \bar{d}) \tag{6.57}$$

where $w(\frac{\tau}{S(T)})$ is the lag window, and $S(T)$ is the truncation lag.

- In our computer program, we use the Newey West tent window

6.8.3 Results of Tests for Forecast Equality

- From Monte Carlo analysis, Diebold and Mariano find that for Gaussian forecast errors, the test is robust to contemporaneous and serial correlation in large samples and oversized in small samples.
- With non-Gaussian forecast errors, other available tests are drastically missized even in large samples, while the one proposed above is approximately correct, even for moderate sample size.

6.8.4 Nested Model Forecast Equality

- As mentioned the tests of Diebold and Mariano are intended for nonested models since the denomiator of the test converges to zero under the null hypothesis.
- A statistic similar to the one above which can be used is to replace

  $$\bar{d} = \frac{1}{T}\sum_{t=1}^{T} e_{it}(e_{it} - e_{jt})$$

  and use the long-run variance estimats with this in (6.56)