Book Summary: *Introductory Econometrics for Finance, 2nd Edition*

Yan Zeng

Version 1.0, last revised on 2017-03-08

Abstract

Summary of Brooks [2].

Contents

1 Introduction

See Zeng [20] for the summary.

2 A brief overview of the classical linear regression model

See Zeng [20] for the summary.

3 Further development and analysis of the classical linear regression model

See Zeng [20] for the summary.

4 Classical linear regression model assumptions and diagnostic tests

See Zeng [20] for the summary.

5 Univariate time series modelling and forecasting

5.1 Introduction

Time series models are usually a-theoretical, implying that their construction and use is not based upon any underlying theoretical model of the behaviour of a variable. Time series models may be useful when a structural model is inappropriate. The book endeavours to answer the following two questions:

- "For a specified time series model with given parameter values, what will be its defining characteristics?"
- "Given a set of data, with characteristics that have been determined, what is a plausible model to describe that data?"

A series is **strictly stationary** if the distribution of its values remains the same as time progresses. A series is **weakly** or **covariance stationary** if it has constant mean, constant variance, and constant autocovariances for each given lag.

5.2 White noise and joint test of its autocorrelation

A **white noise process** $(u_t)_t$ is a weakly stationary series whose autocovariances are identically zero when the lag is greater than 0. If it is further assumed that u_t is distributed normally, then the sample autocorrelation coefficients are also approximately normally distributed

$$\hat{\tau}_s \sim N(0, 1/T)$$

where T is the sample size, and $\hat{\tau}_s$ denotes the autocorrelation coefficient at lag s estimated from a sample. This result can be used to conduct significance tests for the autocorrelation coefficients by constructing a non-rejection region for an estimated autocorrelation coefficient to determine whether it is significantly different from zero. It is also possible to test the joint hypothesis that all m of the τ_k correlation coefficients are simultaneously equal to zero using the **Box-Pierce statistic** [1]

$$Q = T \sum_{k=1}^{m} \hat{\tau}_k^2$$

where T is sample size, m is maximum lag length.[1]

However, the Box-Pierce test has poor small sample properties. A variant of the Box-Pierce test, having better small sample properties, is the **Ljung-Box statistic** [16]

$$Q^* = T(T+2)\sum_{k=1}^{m}\frac{\hat{\tau}_k^2}{T-k} \sim \chi_m^2.$$

This statistic is very useful as a portmanteau (general) test of linear dependence in time series.

Individual test vs. joint test. Sometimes an individual test caused a rejection of the null hypothesis (of a coefficient being zero) while the joint test did not. This unexpected result may have arisen as a result of the low power of the joint test when other individual autocorrelation coefficients are insignificant. Thus the effect of the significant autocorrelation coefficient is diluted in the joint test by the insignificant coefficients.

5.3 MA(q) and AR(p)

5.3.1 Characteristics

A q**th order moving average model**, denoted by MA(q), is defined as

$$y_t = \mu + \sum_{i=1}^{q}\theta_i u_{t-i} + u_t$$

where $(u_t)_t$ is a white noise process with zero mean and variance σ^2. Using the lag operator L with $Ly_t = y_{t-1}$, the defining equation can be written as

$$y_t = \mu + \theta(L)u_t$$

where $\theta(L) = 1+\theta_1 L+\theta_2 L^2+\cdots+\theta_q L^q$. Removing μ considerably eases the complexity of algebra involved, and is inconsequential for it can be achieved without loss of generality.

Then the distinguishing properties of the moving average process of order q given above are

(1) $E[y_t] = \mu$.

(2) $\mathrm{Var}(y_t) = \gamma_0 = (1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2)\sigma^2$.

(3) Covariances

$$\gamma_s = \begin{cases}(\theta_s+\theta_{s+1}\theta_1+\theta_{s+2}\theta_2+\cdots+\theta_q\theta_{q-s})\sigma^2 & \text{for } s=1,2,\cdots,q\\ 0 & \text{for } s>q\end{cases}$$

An **autoregressive model of order** p, denoted by AR(p), is defined as

$$y_t = \mu + \sum_{i=1}^{p}\phi y_{t-i} + u_t$$

or

$$\phi(L)y_t = \mu + u_t$$

where $\phi(L) = (1-\phi_1 L-\phi_2 L^2-\cdots-\phi_p L^p)$. It has the following distinguishing properties

(1) $E[y_t] = \frac{\mu}{1-\phi_1-\phi_2-\cdots-\phi_p}$.

(2) $\mathrm{Var}(y_t) = \gamma_0 = \left(\sum_{i=0}^{\infty}\psi_i^2\right)\sigma^2$, where ψ_i's are determined by $\psi(L) = \sum_{i=0}^{\infty}\psi_i L^i = \phi(L)^{-1}$.

(3) Autocorrelation functions obtained by solving the **Yule-Walker equations**: $\tau_s = \sum_{k=1}^{p}\phi_k\tau_{k-s}$, i.e.

$$\begin{cases}\tau_1 = \phi_1+\tau_1\phi_2+\cdots+\tau_{p-1}\phi_p\\ \tau_2 = \tau_1\phi_1+\phi_2+\cdots+\tau_{p-2}\phi_p\\ \qquad\vdots\ \vdots\ \vdots\\ \tau_p = \tau_{p-1}\phi+\tau_{p-2}\phi_2+\cdots+\phi_p\end{cases}$$

[1]The intuition goes as follows. Since the sum of squares of independent standard normal variates is itself a χ^2 variate with degree of freedom equal to the number of squares in the sum, it can be stated that the Q-statistic is asymptotically distributed as a χ_m^2 under the null hypothesis that all m autocorrelation coefficients are zero.

or in matrix form

$$\begin{bmatrix}\tau_1\\ \tau_2\\ \vdots\\ \tau_p\end{bmatrix} = \begin{bmatrix} 1 & \tau_1 & \tau_2 & \cdots & \tau_{p-2} & \tau_{p-1} \\ \tau_1 & 1 & \tau_1 & \cdots & \tau_{p-3} & \tau_{p-2} \\ \tau_2 & \tau_1 & 1 & \cdots & \tau_{p-4} & \tau_{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \tau_{p-1} & \tau_{p-2} & \tau_{p-3} & \cdots & \tau_1 & 1 \end{bmatrix} \begin{bmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_p\end{bmatrix}$$

For any AR model that is stationary, the autocorrelation function will decay geometrically to zero (this is called the *short memory property* of time series). For example, for an AR(1) model,

$$E[y_t] = \frac{\mu}{1-\phi_1},\ \mathrm{Var}(y_t) = \frac{\sigma^2}{1-\phi_1^2},\ \tau_s = \phi_1^s.$$

5.3.2 Stationarity of AR(p) and Wold's decomposition theorem

Without loss of generality, assume $\mu = 0$. Then it would be stated that **the $AR(p)$ process is stationary** if it is possible to write

$$y_t = \phi(L)^{-1} u_t$$

with $\phi(L)^{-1}$ converging to zero. This means the autocorrelations will decline eventually as the lag length is increased. When the expansion $\phi(L)^{-1}$ is calculated, it will contain an infinite number of terms, and can be written as an MA(∞). The condition for testing for the stationarity of a general AR(p) model is that the roots of the characteristic equation

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

all lie outside the unit circle.

Wold's decomposition theorem states that any stationary series can be decomposed into the sum of two unrelated processes, a purely deterministic part and a purely stochastic part, which will be an MA(∞) process. In the context of AR modelling, any stationary AR(p) process with no constant can be expressed as an MA(∞) process.

5.4 The partial autocorrelation function

The **partial autocorrelation function**, or PACF (denoted τ_{kk}), measures the correlation between an observation k periods ago and the current observation, after controlling for observations at intermediate lags.

The intuitive explanation of the characteristic shape of the PACF for an AR(p) process is that, there will be direct connections between y_t and y_{t-s} for $s \leq p$, but no direct connections for $s > p$. Hence the PACF will usually have non-zero partial autocorrelation coefficients for lags up to the order of the model, but will have zero partial autocorrelation coefficients thereafter.

To see the shape of the PACF for an MA(q) process, one would need to think about the MA model as being transformed into an AR model in order to consider whether y_t and y_{t-s}, $s = 1, 2, \cdots$, are directly connected. In fact, so long as the MA(q) process is invertible, it can be expressed as an AR(∞) process. An MA(q) model is typically required to have roots of the characteristic equation $\theta(z) = 0$ greater than one in absolute value. The invertibility condition prevents the model from exploding under an AR(∞) representation, so that $\theta^{-1}(L)$ converges to zero.

In sumary, the ACF of an AR model has the same basic shape as the PACF for an MA model, and the ACF for an MA model has the same shape as the PACF for an AR model.

5.5 ARMA processes

5.5.1 Characteristics in terms of ACF and PACF

An **ARMA(p,q) process** is defined as

$$\phi(L) y_t = \mu + \theta(L) u_t$$

where

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p,\ \theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q$$

with $E[u_t] = 0$, $E[u_t^2] = \sigma^2$, and $E[u_t u_s] = 0$ for $t \neq s$. The characteristics of an ARMA process will be a combination of those from the AR and MA parts, and the PACF is particularly useful in this context. The ACF alone can distinguish between a pure AR and a pure MA process. However, an ARMA process will have a geometrically declining ACF, as will a pure AR process. So the PACF is useful for distinguishing between an AR(p) process and an ARMA(p, q) process – the former will have a geometrically declining autocorrelation function, but a partial autocorrelation function which cuts off to zero after p lags, while the latter will have both autocorrelation and partial autocorrelation functions which decline geometrically.

In summary, an AR process has a geometrically decaying ACF and a number of non-zero points of PACF = AR order. An MA process has a number of non-zero points of ACF = MA order and a geometrically decaying PACF. An ARMA process has a geometrically decaying ACF and a geometrically decaying PACF.

5.5.2 The Box-Jenkins approach to ARMA estimation

The **Box-Jenkins approach** estimates an ARMA model in a systematic manner. This approach involves three steps:

(1) *Identification*. This step involves determining the order of the model. Graphical procedures can be used by plotting the ACF and the PACF to determine the most appropriate specification. But a more objective technique is to use information criteria.

Information criteria embody two factors: a term which is a function of the residual sum of squares (RSS), and some penalty for the loss of degrees of freedom from adding extra parameters. The object is to choose the number of parameters which minimises the value of the information criteria. The three most popular information criteria are Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (SBIC), and the Hannan-Quinn criterion (HQIC):

$$\begin{cases} AIC = \ln(\hat{\sigma}^2) + \frac{2k}{T} \\ SBIC = \ln(\hat{\sigma}^2) + \frac{k}{T}\ln T \\ HQIC = \ln(\hat{\sigma}^2) + \frac{2k}{T}\ln(\ln(T)) \end{cases}$$

where $\hat{\sigma}^2$ is the residual variance, $k = p + q + 1$ is the total number of parameters estimated, and T is the sample size. The information criteria are actually minimised subject to $p \leq \bar{p}$, $q \leq \bar{q}$, i.e. an upper limit is specified on the number of moving average ($\bar{q}$) and/or autoregressive ($\bar{p}$) terms that will be considered.

It is worth noting that SBIC embodies a much stiffer penalty term than AIC, while HQIC is somewhere in between. The adjusted R^2 measure can also be viewed as an information criterion, although it is a very soft one, which would typically select the largest models of all.

(2) *Estimation*. This step involves estimation of the parameters of the model specified in Step 1. This can be done using least squares or maximum likelihood.

(3) *Diagnostic checking*. This step involves model checking via two methods: overfitting and residual diagnostic. *Overfitting* involves deliberately fitting a larger model and check if any extra terms added to the ARMA model would be insignificant. *Residual diagnostics* imply checking the residuals for evidence of autocorrelation via the ACF, PACF, or Ljung-Box tests (rather than the whole barrage of tests outlined in Chapter 4), and this approach could only reveal a model that is underparameterised ("too small") and would not reveal a model that is overparameterised ("too big"). Examining whether the residuals are free from autocorrelation is much more commonly used than overfitting.

Which criterion should be preferred if they suggest different model orders? SBIC will asymptotically deliver the correct model order, while AIC will deliver on average too large a model, even with an infinite amount of data. On the other hand, the average variation in selected model orders from different samples within a given population will be greater in the context of SBIC than AIC. Overall, then, no criterion is definitely superior to others.

An **integrated autoregressive process** is one whose characteristic equation has a root on the unit circle. Typically researchers difference the variable as necessary and then build an ARMA model on those

differenced variables. *For the remainder of this chapter, it is assumed that the data used are stationary, or have been suitably transformed to make them stationary.*

5.6 Exponential smoothing

An **exponential smoothing model** imposes a geometrically declining weighting scheme on the lagged values of a series. The equation for the model is

$$S_t = \alpha y_t + (1 - \alpha) S_{t-1}$$

where α is the smoothing constant, with $0 < \alpha < 1$, y_t is the current realised value, S_t is the current smoothed value.

The forecasts from an exponential smoothing model are simply set to the current smoothed value, for any number of steps ahead, s

$$f_{t,s} = S_t, \ s = 1, 2, 3, \cdots$$

The exponential smoothing model can be seen as a special case of a Box-Jenkins model, an ARIMA(0, 1,1), with MA coefficient $(1 - \alpha)$ – see Granger and Newbold [9, p. 174].

Exponential smoothing has several advantages over the slightly more complex ARMA class of models. First, exponential smoothing is obviously very simple to use. Among the disadvantages of exponential smoothing is the fact that it is overly simplistic and inflexible. Also, the forecasts from an exponential smoothing model do not converge on the long-term mean of the variable as the horizon increases. The upshot is that long-term forecasts are overly affected by recent events in the history of the series under investigation and will therefore be sub-optimal.

5.7 Forecasting in econometrics

It is useful to distinguish between two approaches of forecasting:

• *Econometric (structural) forecasting* – relates a dependent variable to one or more independent variables. Such models often work well in the long run, since a long-run relationship between variables often arises from no-arbitrage or market efficiency conditions.

• *Time series forecasting* – involves trying to forecast the future values of a series given its previous values and/or previous values of an error term.

It is also worth distinguishing between *point forecasts* and *interval forecasts*, *in-sample forecasts* and *out-of-sample forecasts*, and *one-step-ahead forecast* and *multi-step-ahead forecasts*.

Time series models are generally better suited to the production of time series forecasts than structural models.

5.7.1 Forecasting with ARMA models

Let $f_{t,s}$ denote a forecast made using an ARMA(p, q) model at time t for s steps into the future for some series y. The forecasts are generated by what is known as a **forecast function**, typically of the form

$$f_{t,s} = \sum_{i=1}^{p} a_i f_{t,s-i} + \sum_{j=1}^{q} b_j u_{t+s-j},$$

where $f_{t,s} = y_{t+s}$, $s \leq 0$;

$$u_{t+s} = \begin{cases} 0 & s > 0 \\ u_{t+s} & s \leq 0 \end{cases}$$

and a_i and b_i are the autoregressive and moving average coefficients, respectively.

5.7.2 Determining whether a forecast is accurate or not

The mean square error (*MSE*) or mean square error (*MAE*) from one model would be compared with those of other models for the same data and forecast period, and the model(s) with the lowest value of the error measure would be argued to be the most accurate.

Let T be the total sample size (in-sample + out-of-sample), and T_1 the first out-of-sample forecast observation. Thus in-sample model estimation initially runs from observation 1 to $(T_1 - 1)$, and observations T_1 to T are available for out-of-sample estimation, i.e. a total holdout sample of $T - (T_1 - 1)$.

The **mean square error** (*MSE*) is defined as

$$MSE = \frac{1}{T - (T_1 - 1)} \sum_{t=T_1}^{T} (y_{t+s} - f_{t,s})^2.$$

The **mean absolute error** (*MAE*) measures the average absolute forecast error, and is given by

$$MAE = \frac{1}{T - (T_1 - 1)} \sum_{t=T_1}^{T} |y_{t+s} - f_{t,s}|.$$

The **mean absolute percentage error** (*MAPE*) has the attractive property compared to *MSE* that it can be interpreted as a percentage error, and furthermore, its value is bounded from below by 0. It is defined as

$$MAPE = \frac{100}{T - (T_1 - 1)} \sum_{t=T_1}^{T} \left| \frac{y_{t+s} - f_{t,s}}{y_{t+s}} \right|.$$

The **adjusted MAPE** (*AMAPE*) or symmetric *MAPE* corrects for the problem of asymmetry between the actual and forecast values

$$AMAPE = \frac{100}{T - (T_1 - 1)} \sum_{t=T_1}^{T} \left| \frac{y_{t+s} - f_{t,s}}{y_{t+s} + f_{t,s}} \right|.$$

Another criterion which is popular is **Theil's U-statistic**, which is defined as

$$U = \frac{\sqrt{\sum_{t=T_1}^{T} \left(\frac{y_{t+s} - f_{t,s}}{y_{t+s}} \right)^2}}{\sqrt{\sum_{t=T_1}^{T} \left(\frac{y_{t+s} - fb_{t,s}}{y_{t+s}} \right)^2}}$$

where $fb_{t,s}$ is the forecast obtained from a benchmark model (typically a simple model such as a naive or random walk). A U-statistic of one implies that the model under consideration and the benchmark model are equally (in)accurate, while a value of less than one implies that the model is superior to the benchmark, and vice versa for $U > 1$.

Statistical versus financial or economic loss functions. Many econometric forecasting studies evaluate the model's success using statistical loss functions such as those described above. However, it is not necessarily the case that models classed as accurate because they have small mean squared forecast errors are useful in practical situations. For example, it has recently been shown (Gerlow, Irwin and Liu [7]) that the accuracy of forecasts according to traditional statistical criteria may give little guide to the potential profitability of employing those forecasts in a market trading strategy.

On the other hand, models that can accurately forecast the sign of future returns, or can predict turning points in a series have been found to be more profitable (Leitch and Tanner [14]). Two possible indicators of the ability of a model to predict direction changes are

$$\% \text{ correct sign predictions} = \frac{1}{T - (T_1 - 1)} \sum_{t=T_1}^{T} z_{t+s}$$

where

$$z_{t+s} = \begin{cases} 1 & (y_{t+s} f_{t,s}) > 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\% \text{ correct direction change predictions} = \frac{1}{T - (T_1 - 1)} \sum_{t=T_1}^{T} z_{t+s}$$

where

$$z_{t+s} = \begin{cases} 1 & (y_{t+s} - y_t)(f_{t,s} - y_t) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

MSE penalises large errors disproportionately more heavily than small errors, *MAE* penalises large errors proportionately equally as heavily as small errors, while the sign prediction criterion does not penalise large errors any more than small errors.

Finance theory and time series analysis. An example of ARIMA model identification, estimation and forecasting in the context of commodity prices is given by Chu [3]. He finds ARIMA models useful compared with structural models for short-term forecasting, but also finds that they are less accurate over longer horizons. It also observed that ARIMA models have limited capacity to forecast unusual movements in prices.

Chu [3] argues that, although ARIMA models may appear to be completely lacking in theoretical motivation, and interpretation, ARIMA specifications quite often arise naturally as reduced form equations (see chapter 6) corresponding to some underlying structural relationships. In such a case, not only would ARIMA models be convenient and easy to estimate, they could also be well grounded in financial or economic theory after all.

6 Multivariate models

6.1 Simultaneous equations

One of the CLRM assumptions was that $\boldsymbol{X}$ and $\boldsymbol{u}$ are independent, and given also the assumption that $E[\boldsymbol{u}] = \boldsymbol{0}$. However, structural equations which are part of a simultaneous system often violate this assumption and application of OLS to these equations will lead to biased coefficient estimates. This is known as *simultaneity bias* or *simultaneous equations bias*. And the OLS estimator is not even consistent.

6.1.1 Identification

Identification is the issue of whether there is enough information in the reduced form equations[2] to enable the structural form coefficients to be calculated. There are two conditions that could be examined to determine whether a given equation from a system is identified – the *order condition*, a necessary but not sufficient condition, and the *rank condition*, a necessary and sufficient condition.

• An equation is **unidentified**, if structural coefficients cannot be obtained from the reduced form estimates by any means.

• An equation is **exactly identified (just identified)**, if unique structural form coefficient estimates can be obtained by substitution from the reduced form equations.

• An equation is **overidentified**, if more than one set of structural coefficients can be obtained from the reduced form.

Statement of the order condition. Let G denote the number of structural equations. An equation is just identified if the number of variables excluded from an equation is $G - 1$, where "excluded" means the number of all endogenous and exogenous variables that are not present in this particular equation. If more than $G - 1$ are absent, it is over-identified. If less than $G - 1$ are absent, it is not identified.

[2] A reduced form equation has only exogenous variables on the RHS. So the usual requirements for consistency and unbiasedness of the OLS estimator will hold. However, the values of the transformed coefficients are not of much interest; what is wanted are the original parameters in the structural equations.

6.1.2 The Hausman test

A variable is **strictly exogenous** if it is independent of all contemporaneous, future and past errors in that equation. Financial theory might suggest that there should be a two-way relationship between two or more variables. A **Hausman test** can be used to test whether a simultaneous equations model is necessary. To conduct a Hausman test for exogeneity, for each equation:

1) *Obtain the reduced form equations corresponding to the structural form equation.*

2) *Estimate the reduced form equations using OLS, and obtain the fitted values.*

3) *Run the regression of the structural form equation, at this stage ignoring any possible simultaneity.*

4) *Run the regression of the structural form equation again, but now also including the fitted values from the reduced form equations as additional regressors.*

5) *Use an F-test to test the joint restriction that the coefficients of the fitted values are zero. If the null hypothesis is rejected, then a simultaneous equations model is necessary; otherwise, the additional regressors can be treated as exogenous.*

A **recursive** or **triangular system** is a set of equations that looks like a simultaneous equations system, but isn't. In fact, there is not a simultaneity problem here, since the dependence is not bi-directional, for each equation it all goes one way.

6.1.3 Estimation procedures for simultaneous equations systems

Estimation procedures for simultaneous equations systems include indirect least squares (ILS), two-stage least squares (2SLS or TSLS), and instrumental variables. Other estimation techniques available for systems of equations include three-stage least squares (3SLS), full information maximum likelihood (FIML) and limited information maximum likelihood (LIML). For further technical details on each of these procedures, see Green [10, Chapter 15].

Indirect least squares (ILS). If the system is just identified, ILS involves estimating the reduced form equations using OLS, and then using them to substitute back to obtain the structural parameters. This method is not widely applied because:

1) Solving back to get the structural parameters can be tedious.

2) Most simultaneous equations systems are overidentified.

ILS estimators are consistent and asymptotically efficient, but in general they are biased. The bias arises from the fact that the structural form coefficients under ILS estimation are transformations of the reduced form coefficients.

2SLS. This method is applicable to estimating just identified and overidentified systems. Two-stage least squares (2SLS or TSLS) is done in two stages:

Stage 1: Obtain and estimate the reduced form equations using OLS. Save the fitted values for the dependent variables.

Stage 2: Estimate the structural equations using OLS, but replace any RHS endogenous variables with their stage 1 fitted values.

The 2SLS estimator is consistent, but not unbiased. In a simultaneous equations framework, it is still of concern whether the usual assumptions of the CLRM are valid or not. If the disturbances in the structural equations are autocorrelated, the 2SLS estimator is not even consistent. Thestandard error estimates also need to be modified compared with their OLS counterparts, but once this has been done, the usual t-tests can be used to test hypotheses about the structural form coefficients.

Instrumental variables. Recall that the reason that OLS cannot be used directly on the structural equations is that the endogenous variables are correlated with the errors. One solution to this would be to use some other variables instead. These other variables should be (highly) correlated with the endogenous variables, but not correlated with the errors – such variables would be known as *instruments*. If the instruments are the variables in the reduced form equations, then IV is equivalent to 2SLS, so that the latter can be viewed as a special case of the former.

What happens if IV or 2SLS are used unnecessarily? In other words, suppose that one attempted to estimate a simultaneous system when the variables specified as endogenous were in fact independent of one another. The consequences are similar to those of including irrelevant variables in a single equation OLS

model. That is, the coefficient estimates will still be consistent, but will be inefficient compared to those that just used OLS directly.

6.2 Vector autoregressive models

A **vector autoregressive model** (VAR) is a systems regression model (i.e. there is more than one dependent variable) that can be considered a kind of hybrid between the univariate time series models and the simultaneous equations models.VARs have often been advocated as an alternative to large-scale simultaneous equations structural models. An important feature of the VAR model is its flexibility and the ease of generalisation. Another useful facet of VAR models is the compactness with which the notation can be expressed:

$$\boldsymbol{y}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{y}_{t-1} + \boldsymbol{u}_t, \; \boldsymbol{y}_t = \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{p,t} \end{bmatrix}.$$

VAR models have several advantages compared with univariate time series models or simultaneous equations structural models:

• The researcher does not need to specify which variables are endogenous or exogenous – *all are endogenous*. Since Hausman-type tests are often not employed in practice when they should be, the specification of certain variables as exogenous, required to form identifying restrictions, is likely in many cases to be invalid. VAR estimation, on the other hand, requires no such restrictions to be imposed.

• VARs are more flexible than univariate AR models since they allow the value of a variable to depend on more than just its own lags or combination of white noise terms.

• Provided that there are no contemporaneous terms on the RHS of the equations, it is possible to simply use OLS separately on each equation. This arises from the fact that all variables on the RHS are pre-determined – that is, at time t, they are known. There is no possibility for feedback from any of the LHS variables to any of the RHS variables.

• The forecasts generated by VARs are often better than "traditional structural" models. It has been argued in a number of articles that large-scale structural models performed badly in terms of their out-of-sample forecast accuracy. This could perhaps arise as a result of the ad hoc nature of the restrictions placed on the structural models to ensure identification.

VAR models of course also have drawbacks and limitations relative to other model classes:

• VARs are *a-theoretical* and are less amenable to theoretical analysis and therefore to policy prescriptions. A hapless researcher could obtain an essentially spurious relationship by mining the data. It is also often not clear how the VAR coefficient estimates should be interpreted.

• How should the appropriate lag lengths for the VAR be determined?

• *So many parameters*! For relatively small sample sizes, degrees of freedom will rapidly be used up, implying large standard errors and therefore wide confidence intervals for model coefficients.

• Should *all of the components of the VAR be stationary*? Obviously, if one wishes to use hypothesis tests to examine the statistical significance of the coefficients, then it is essential that all of the components in the VAR are stationary. The purpose of VAR estimation is purely to examine the relationships between the variables, and that differencing will throw information on any long-run relationships between the series away.

6.2.1 Choosing the optimal lag length for a VAR

There are broadly two methods that could be used to arrive at the optimal lag length: cross-equation restrictions and information criteria.

Cross-equation restrictions for VAR lag length selection. One approach is to specify the same number of lags in each equation and to determine the model order as follows. Suppose that a VAR estimated using quarterly data has 8 lags of the two variables in each equation, and it is desired to examine a restriction that the coefficients on lags 5-8 are jointly zero. This can be done using a likelihood ratio test. Denote the

variance-covariance matrix of the residuals (given by $\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}'$), as $\hat{\boldsymbol{\Sigma}}$. The likelihood ratio test for this joint hypothesis is given by

$$LR = T\left[\log|\hat{\boldsymbol{\Sigma}}_r| - \log|\hat{\boldsymbol{\Sigma}}_u|\right]$$

where $|\hat{\boldsymbol{\Sigma}}_r|$ is the determinant of the variance-covariance matrix of the residuals for the restricted model (with 4 lags), $|\hat{\boldsymbol{\Sigma}}_u|$ is the determinant of the variance-covariance matrix of residuals for the unrestricted VAR (with 8 lags) and T is the sample size. The test statistic is asymptotically distributed as a χ^2 variate with degrees of freedom equal to the total number of restrictions. Intuitively, the test is a multivariate equivalent to examining the extent to which the RSS rises when a restriction is imposed. If $\hat{\boldsymbol{\Sigma}}_r$ and $\hat{\boldsymbol{\Sigma}}_u$ are "close together", the restriction is supported by the data.

Information criteria for VAR lag length selection. The likelihood ratio (LR) test explained above is intuitive and fairly easy to estimate, but has its limitations. Principally, one of the two VARs must be a special case of the other and, more seriously, only pairwise comparisons can be made. A further disadvantage of the LR test approach is that the χ^2 test will strictly be valid asymptotically only under the assumption that the errors from each equation are normally distributed. This assumption is unlikely to be upheld for financial data.

Information criteria require no such normality assumptions concerning the distributions of the errors. The multivariate versions of the information criteria can be defined as

$$\begin{cases} MAIC = \log|\hat{\boldsymbol{\Sigma}}| + 2k'/T \\ MSBIC = \log|\hat{\boldsymbol{\Sigma}}| + \frac{k'}{T}\log(T) \\ MHQIC = \log|\hat{\boldsymbol{\Sigma}}| + \frac{2k'}{T}\log(\log(T)) \end{cases}$$

where $\hat{\boldsymbol{\Sigma}}$ is the variance-covariance matrix of residuals, T is the number of observations and k' is the total number of regressors in all equations, which will be equal to $p^2k + p$ for p equations in the VAR system, each with k lags of the p variables, plus a constant term in each equation. The values of the information criteria are constructed for $0, 1, \cdots, \bar{k}$ lags (up to some pre-specified maximum $\bar{k}$), and the chosen number of lags i that number minimising the value of the given information criterion.

6.2.2 Three sets of statistics constructed for an estimated VAR model

One fundamental weakness of the VAR approach to modelling is that its a-theoretical nature and the large number of parameters involved make the estimated models difficult to interpret. In order to partially alleviate this problem, three sets of statistics are usually constructed for an estimated VAR model: block significance tests, impulse responses, and variance decompositions.

Block significance and causality tests. It is likely that, when a VAR includes many lags of variables, it will be difficult to see which sets of variables have significant effects on each dependent variable and which do not. In order to address this issue, tests are usually conducted that restrict all of the lags of a particular variable to zero.

Assuming that all of the variables in the VAR are stationary, the joint hypotheses can easily be tested within the F-test framework, since each individual set of restrictions involves parameters drawn from only one equation. The equations would be estimated separately using OLS to obtain the unrestricted RSS, then the restrictions imposed and the models reestimated to obtain the restricted RSS. The F-statistic would then take the usual form described in chapter 3. Thus, evaluation of the significance of variables in the context of a VAR almost invariably occurs on the basis of joint tests on all of the lags of a particular variable in an equation, rather than by examination of individual coefficient estimates.

The tests described here could also be referred to as *causality tests* (see Granger [8] and Sims [19]). Causality tests seek to answer simple questions of the type, "Do changes in y_1 cause changes in y_2?" The argument follows that if y_1 causes y_2, lags of y_1 should be significant in the equation for y_2. If this is the case and not vice versa, it would be said that y_1 "Granger-causes" y_2 or that there exists unidirectional causality from y_1 to y_2. On the other hand, if y_2 causes y_1, lags of y_2 should be significant in the equation for y_1. If both sets of lags were significant, it would be said that there was "bi-directional causality" or "bi-directional feedback". If y_1 is found to Granger-cause y_2, but not vice versa, it would be said that variable y_1 is strongly

exogenous (in the equation for y_2). If neither set of lags are statistically significant in the equation for the other variable, it would be said that y_1 and y_2 are independent. Finally, the word "causality" is somewhat of a misnomer, for Granger-causality really means only a correlation between the current value of one variable and the past values of others; it does not mean that movements of one variable cause movements of another.

VARs with exogenous variables (VARX). Consider the following specification for a VAR(1) where $\boldsymbol{X}_t$ is a vector of exogenous variables and $\boldsymbol{B}$ is a matrix of coefficients

$$y_t = \boldsymbol{A}_0 + \boldsymbol{A}_1 y_{t-1} + \boldsymbol{B}\boldsymbol{X}_t + \boldsymbol{e}_t.$$

The components of the vector $\boldsymbol{X}$ are known as exogenous variables since their values are determined outside of the VAR system. Such a model could be viewed as simply a restricted VAR where there are equations for each of the exogenous variables, but with the coefficients on the RHS in those equations restricted to zero. Such a restriction may be considered desirable if theoretical considerations suggest it, although it is clearly not in the true spirit of VAR modelling, which is not to impose any restrictions on the model but rather to "let the data decide".

Impulse responses and variance decompositions. Block F-tests and an examination of causality in a VAR will suggest which of the variables in the model have statistically significant impacts on the future values of each of the variables in the system. But F-test results will not reveal whether changes in the value of a given variable have a positive or negative effect on other variables in the system, or how long it would take for the effect of that variable to work through the system. This is where impulse responses and variance decompositions come into play.

Impulse responses trace out the responsiveness of the dependent variables in the VAR to shocks to each of the variables. *Variance decompositions* give the proportion of the movements in the dependent variables that are due to their "own" shocks, versus shocks to the other variables. In practice, it is usually observed that own series shocks explain most of the (forecast) error variance of the series in a VAR.

For calculating impulse responses and variance decompositions, the ordering of the variables is important, because in practice, the errors will have a common component that cannot be associated with a single variable alone.[3] Assuming a particular ordering is necessary to compute the impulse responses and variance decompositions, although the restriction underlying the ordering used may not be supported by the data. Again, ideally, financial theory should suggest an ordering.

7 Modelling long-run relationships in finance

7.1 Stationarity

Recall a **stationary series** can be defined as one with a constant mean, constant variance and constant autocovariances for each given lag. Stationarity is essential for the following reasons:

• The stationarity or otherwise of a series can strongly influences its behaviour and properties. As an illustration, for a stationary series, "shocks" to the system will gradually die away, while for non-stationary data, the persistence of shocks will always be infinite.

• The use of non-stationary data can lead to **spurious regressions**. If two variables are trending over time, a regression of one on the other could have a high R^2 even if the two are totally unrelated. So, if standard regression techniques are applied to non-stationary data, the end result could be a regression that "looks" good under standard measures (significant coefficient estimates and a high R^2), but which is really valueless. Such a model would be termed a "spurious regression".

• If the variables employed in a regression model are not stationary, then the standard assumptions for asymptotic analysis will not be valid – the usual "t-ratios" will not follow a t-distribution, and the F-statistic will not follow an F-distribution, and so on. By plotting the histogram of simulated t-ratios, we can often see intuitively if the "t-ratios" truly follow a t-distribution (fat tails often indicate a "No").

There are two types of non-stationarity: the **random walk model with drift**

$$y_t = \mu + y_{t-1} + u_t$$

[3]The usual approach to this difficulty is to generate *orthogonalised impulse responses*.

and the **trend-stationary process**

$$y_t = \alpha + \beta t + u_t.$$

The random walk model with drift could be generalised to the case where y_t is an explosive process

$$y_t = \mu + \phi y_{t-1} + u_t$$

where $\phi > 1$. Typically, this case is ignored and $\phi = 1$ is used to characterise the non-stationarity because $\phi > 1$ does not describe many data series in econometrics and finance, but $\phi = 1$ has been found to describe accurately many financial and economic time series.

The random walk with drift process is known as **stochastic non-stationarity** or a **unit root process**, and stationarity can be induced by "difference once": $\Delta y_t = \mu + u_t$. The trend-stationary process is known as **deterministic non-stationarity** and de-trending is required. A regression of the form $y_t = \mu + \phi y_{t-1} + u_t$ would be run, and any subsequent estimation would be done on the residuals from this equation.

Although trend-stationary and difference-stationary series are both "trending" over time, the correct approach needs to be used in each case. One possibility is to nest both cases in a more general model and to test that. For example, consider the model

$$\Delta y_t = \alpha_0 + \alpha_1 t + (\gamma - 1) y_{t-1} + u_t.$$

Although again, of course the t-ratios in the above regression will not follow a t-distribution.

This book will concentrate on the stochastic stationarity model since it is the model that has been found to best describe most non-stationary financial and economic time series. If a non-stationary series y_t must be differenced d times before it becomes stationary, then it is said to be **integrated of order** d, and is written as $y_t \sim I(d)$. An $I(d)$ series contains d unit roots. The majority of financial and economic time series contain a single unit root, although some are stationary and some have been argued to possibly contain two unit roots (series such as nominal consumer prices and nominal wages).

7.2 Testing for unit roots: ADF, PP, and KPSS

The ACF for a unit root process (a random walk) will often be seen to decay away very slowly to zero. Thus, such a process may be mistaken for a highly persistent but stationary process. Hence it is not possible to use the ACF or PACF to determine whether a series is characterised by a unit root or not. Furthermore, even if the true data generating process for y_t contains a unit root, the results of the tests for a given sample could lead one to believe that the process is stationary. Therefore, what is required is some kind of formal hypothesis testing procedure that answers the question, "given the sample of data to hand, is it plausible that the true data generating process for y contains one or more unit roots?"

The **Dickey-Fuller test** (Fuller [6], Dickey and Fuller [4]) examines the null hypothesis that $\phi = 1$ in

$$y_t = \phi y_{t-1} + u_t$$

against the one-sided alternative $\phi < 1$. Thus the hypotheses of interest are H_0: series contains a unit root versus H_1: series is stationary. In practice, the following regression is employed for ease of computation and interpretation

$$\Delta y_t = \psi y_{t-1} + u_t$$

so that a test of $\phi = 1$ is equivalent to a test of $\psi = 0$. The test statistics for the DF tests are defined as

$$\text{test statistic} = \frac{\hat{\psi}}{\hat{SE}(\hat{\psi})}$$

The test statistics do not follow the usual t-distribution under the null hypothesis, since the null is one of non-stationarity, but rather they follow a non-standard distribution.

A full set of Dickey-Fuller (DF) critical values is given in the appendix of statistical tables at the end of this book. Comparing these with the standard normal critical values, it can be seen that the DF critical

values are much bigger in absolute terms (i.e. more negative). Thus more evidence against the null hypothesis is required in the context of unit root tests than under standard t-tests.

The Dickey-Fuller (DF) tests are also known as τ-tests, and can be conducted allowing for an intercept, or an intercept and deterministic trend, or neither, in the test regression. The general form of the test regression is

$$y_t = \phi y_{t-1} + \mu + \lambda t + u_t$$

or

$$\Delta y_t = \psi y_{t-1} + \mu + \lambda t + u_t.$$

In another paper, Dickey and Fuller [5] provide a set of additional test statistics and their critical values for joint tests of the significance of the lagged y, and the constant and trend terms. They are not examined further here.

Back to the original DF tests, the tests above are valid only if u_t is white noise. If u_t is autocorrelated, the test would be "oversized", meaning that the true size of the test (Type I error) would be higher than the nominal size used (e.g. 5%). The solution is to "augment" the test using p lags of the dependent variable:

$$\Delta y_t = \psi y_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + u_t.$$

The lags of Δy_t now "soak up" any dynamic structure present in the dependent variable, to ensure that u_t is not autocorrelated. The test is known as an **augmented Dickey-Fuller** (ADF) test and is still conducted on ψ, and the same critical values from the DF tables are used as before.

Including too few lags will not remove all of the autocorrelation, thus biasing the results, while using too many will increase the coefficient standard errors. To determine the optimal number of lags of the dependent variable for the ADF test, there are two simple rules of thumb: the frequency of the data (e.g. 12 lags for monthly data and 4 lags for quarterly data) and an information criterion (the number of lags should minimise the value of the information criterion).

Testing for higher orders of integration. If the time series has higher orders of integration, an increasing ordering of the tests (i.e. testing for $I(1)$, then $I(2)$, and so on) is invalid. The theoretically correct approach would be to start by assuming some highest plausible order of integration (e.g. $I(2)$), and to test $I(2)$ against $I(1)$. If $I(2)$ is rejected, then test $I(1)$ against $I(0)$. In practice, however, to the author's knowledge, no financial time series contain more than a single unit root, so that this matter is of less concern in finance.

The **Philips-Perron (PP) tests** are similar to ADF tests, but they incorporate an automatic correction to the DF procedure to allow for autocorrelated residuals. The tests often give the same conclusions as, and suffer from most of the same important limitations as, the ADF tests.

Criticisms of Dickey–Fuller- and Phillips–Perron-type tests. The most important criticism that has been levelled at unit root tests is that their power is low if the process is stationary but with a root close to the non-stationary boundary. The source of this problem is that, under the classical hypothesis-testing framework, if the null hypothesis is never accepted, it is simply stated that it is either rejected or not rejected. This means that a failure to reject the null hypothesis could occur either because the null was correct, or because there is insufficient information in the sample to enable rejection. One way to get around this problem is to use a stationarity test as well as a unit root test, such as the KPSS test (see below).

The **KPSS test**. Stationary tests have stationarity under the null hypothesis, thus reversing the null and alternatives under the Dickey-Fuller approach. One such stationarity test is the KPSS test. The results of these tests can be compared with the ADF/PP procedure to see if the same conclusion is obtained. The null and alternative hypotheses under each testing approach are as follows:

ADF/PP	$KPSS$
$H_0 : y_t \sim I(1)$	$H_0 : y_t \sim I(0)$
$H_1 : y_t \sim I(0)$	$H_1 : y_t \sim I(1)$

There are four possible outcomes:

	ADF/PP		$KPSS$
(1)	Reject H_0	and	Do not reject H_0
(2)	Do not reject H_0	and	Reject H_0
(3)	Reject H_0	and	Reject H_0
(4)	Do not reject H_0	and	Do not reject H_0

For the conclusions to be robust, the results should fall under outcomes 1 or 2, while outcomes 3 or 4 imply conflicting results. The joint use of stationarity and unit root tests is known as *confirmatory data analysis*.

7.3 Cointegration

Let w_t be a $k \times 1$ vector of variables, then the components of w_t are **integrated of order (**d**,** b**)** if:
(1) All components of w_t are $I(d)$.
(2) There is at least one vector of coefficients α such that

$$\alpha' w_t \sim I(d-b).$$

In practice, many financial variables contain one unit root, and are thus $I(1)$, so that the remainder of this chapter will restrict analysis to the case where $d = b = 1$. In this context, a set of variables is defined as **cointegrated** if a linear combination of them is stationary.

Examples where cointegration might be expected to hold in theory include: spot and futures prices for a given commodity or asset, ratio of relative prices and an exchange rate, and equity prices and dividends. If there were no cointegration, there would be no long-run relationship binding the series together, so that the series could wander apart without bound.

An interesting question to ask is whether a potentially cointegrating regression should be estimated using the levels of the variables or the logarithms of the levels of the variables. Financial theory may provide an answer as to the more appropriate functional form, but fortunately even if not, Hendry and Juselius [12] note that if a set of series is cointegrated in levels, they will also be cointegrated in log levels.

7.4 Equilibrium correction or error correction models

When the concept of non-stationarity was first considered in the 1970s, a usual response was to independently take the first differences of each of the $I(1)$ variables and then to use these first differences in any subsequent modelling process. In the context of univariate modelling (e.g. the construction of ARMA models), this is entirely the correct approach. However, when the relationship between variables is important, such a procedure is inadvisable. While this approach is statistically valid, it does have the problem that pure first difference models have no long-run solution and it therefore has nothing to say about whether x and y have an equilibrium relationship.

Fortunately, there is a class of models that can overcome this problem by using combinations of first differenced and lagged levels of cointegrated variables. For example, consider the following equation

$$\Delta y_t = \beta_1 \Delta x_t + \beta_2 (y_{t-1} - \gamma x_{t-1}) + u_t.$$

This model is known as an **error correction model** or an **equilibrium correction model**, and $y_{t-1} - \gamma x_{t-1}$ is known as the **error correction term**. Provided that y_t and x_t are cointegrated with cointegrating coefficient γ , then $(y_{t-1} - \gamma x_{t-1})$ will be $I(0)$ even though the constituents are $I(1)$. It is thus valid to use OLS and standard procedures for statistical inference on the above equation.

Error correction models are interpreted as follows. y is purported to change between $t-1$ and t as a result of changes in the values of the explanatory variable(s), x, between $t-1$ and t, and also in part to correct for any disequilibrium that existed during the previous period.

Of course, an error correction model can be estimated for more than two variables. The **Granger representation theorem** states that if there exists a dynamic linear model with stationary disturbances and the data are $I(1)$, then the variables must be cointegrated of order $(1, 1)$. For details of this theorem, see Hamilton [11, p. 582].

7.5 Testing for cointegration in regression: a residual-based approach (EG, CRDW)

The model for the equilibrium correction term can be generalised further to include k variables (y and the $(k-1)$ x's)

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \cdots + \beta_k x_{kt} + u_t \tag{1}$$

u_t should be $I(0)$ if the variables y_t, x_{2t}, $\cdots$, x_{kt} are cointegrated, but u_t will still be non-stationary if they are not. Thus it is necessary to test the residuals of (1) to see whether they are non-stationary or stationary, hence the residual-based approach to testing cointegration.

The **Engle-Granger (EG) test** uses the DF or ADF test on $\hat{u}_t$, using a regression of the form

$$\Delta\hat{u}_t = \psi\hat{u}_{t-1} + v_t$$

with v_t an i.i.d. error term. Since the test is now operating on the residuals of an estimated model rather than on raw data, the critical values are changed compared to a DF or an ADF test on a series of raw data. The new set of critical values are larger in absolute value (i.e. more negative) than the DF critical values. The critical values also become more negative as the number of variables in the potentially cointegrating regression increases.

The **Cointegrating Regression Durbin Watson (CRDW) test** uses the Durbin-Watson (DW) test statistic or the Philips-Perron (PP) approach to test for non-stationarity of $\hat{u}_t$. Under the null hypothesis of a unit root in the errors, $CRDW \approx 0$, so the null of a unit root is rejected if the $CRDW$ statistic is larger than the relevant critical value (which is approximately 0.5).

Suitable model specification for the residual-based approach to cointegration testing. The null and alternative hypotheses for any unit root test applied to the residuals of a potentially cointegrating regression are

$$\begin{aligned} H_0: &\quad \hat{u}_t \sim I(1) \\ H_1: &\quad \hat{u}_t \sim I(0). \end{aligned}$$

If the null hypothesis is not rejected, there is no cointegration. The appropriate strategy for econometric modelling in this case would be to employ specifications in first differences only. Such models would have no long-run equilibrium solution, but this would not matter since no cointegration implies that there is no long-run relationship anyway.

If the null hypothesis is rejected, the variables would be classed as conintegrated. The appropriate strategy for econometric modelling in this case would be to form and estimate an error correction model, using a method described in Section 7.6.

Multiple cointegrating relationship. Suppose that there are k variables in a system (ignoring any constant term), denoted y_t, x_{2t}, $\cdots$, x_{kt}. In this case, there may be up to r linearly independent cointegrating relationship (where $r \leq k-1$). This potentially presents a problem for the OLS regression approach described above, which is capable of finding at most one cointegrating relationship no matter how many variables there are in the system. And if there are multiple cointegrating relationships, how can one know if there are others, or whether the "best" or strongest cointegrating relationship has been found? An OLS regression will find the minimum variance stationary linear combination of the variables, but there may be other linear combinations of the variables that have more intuitive appeal. The answer to this problem is to use a systems approach to cointegration, which will allow determination of all r cointegrating relationships. One such approach is Johansen' s method (see Section 7.6.3).

7.6 Methods of parameter estimation in cointegrated systems

What should be the modelling strategy if the data at hand are thought to be non-stationary and possibly cointegrated? There are (at least) three methods that could be used: Engle-Granger, Engle-Yoo and Johansen.

7.6.1 The Engle-Granger 2-step method

This is a single equation technique, which is conducted as follows:

Step 1. Make sure that all the individual variables are $I(1)$. Then estimate the cointegrating regression using OLS. Save the residuals of the cointegrating regression, $\hat{u}_t$. Test these residuals to ensure that they are $I(0)$, proceed to Step 2; if they are $I(1)$, estimate a model containing only first difference.

Step 2. Use the residuals obtained in Step 1 as one variable in the error correction model, e.g.

$$\Delta y_t = \beta_1 \Delta x_t + \beta_2 (\hat{u}_{t-1}) + v_t$$

where $\hat{u}_{t-1} = y_{t-1} - \hat{\tau} x_{t-1}$. It is now valid to perform inference in the second-stage regression, i.e. concerning the parameters β_1 and β_2, since all variables in this regression are stationary.

The Engle-Granger 2-step method suffers from a number of problems:
(1) The usual finite sample problem of a lack of power in unit root and cointegration tests (when there is a root close to the non-stationary boundary) discussed above.
(2) There could be a simultaneous equations bias if the causality runs in both directions, but this single equation approach requires the researcher to normalise on one variable (i.e. to specify one variable as the dependent variable and the others as independent variables).
(3) It is not possible to perform any hypothesis tests about the actual cointegrating relationship estimated at stage 1.

Problems 1 and 2 are small sample problems that should disappear asymptotically. Problem 3 is addressed by another method due to Engle and Yoo.

7.6.2 The Engle-Yoo 3-step method

The Engle and Yoo 3-step procedure takes its first two steps from Engle-Granger (EG). Engle and Yoo then add a third step giving updated estimates of the cointegrating vector and its standard errors. The Engle and Yoo (EY) third step is algebraically technical and additionally, EY suffers from all of the remaining problems of the EG approach. There is arguably a far superior procedure available to remedy the lack of testability of hypotheses concerning the cointegrating relationship – namely, the Johansen procedure. For these reasons, the Engle-Yoo procedure is rarely employed in empirical applications and is not considered further here.

7.6.3 The Johansen technique based on VARs

Suppose that a set of g variables ($g \geq 2$) are under consideration that are $I(1)$ and which are thought may be cointegrated. A VAR with k lags containg these variables could be set up:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + u_t.$$

In order to use the Johansen test, the VAR above needs to be turned into a vector error correction model (VECM) of the form

$$\Delta y_t = \Gamma y_{t-k} + \Gamma_1 \Delta y_{t-1} + \Gamma_2 \Delta y_{t-2} + \cdots + \Gamma_{k-1} \Delta y_{t-(k-1)} + u_t$$

where $\Gamma = \left(\sum_{i=1}^{k} \beta_i \right) - I_g$ and $\Gamma_i = \left(\sum_{j=1}^{i} \beta_j \right) - I_g$. This VAR contains g variables in first differenced from on the LHS, and $k-1$ lags of the dependent variables (differences) on the RHS, each with a Γ coefficient matrix attached to it. Note the comparability between this set of equations and the testing equation for an ADF test.

The Johansen test centres around an examination of the Γ matrix. The test for cointegration between y's is calculated by looking at the rank of the Γ matrix via its eigenvalues. The eigenvalues, denoted by λ_i, are put in ascending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_g$. If the λ's are roots, in this context they must be less than 1 in absolute value and positive. If the variables are not cointegrated, the rank of Γ will not be significantly different from zero, so $\lambda_i \approx 0 \; \forall i$.

There are two test statistics for cointegration under the Johansen approach, which are formulated as

$$\lambda_{trace}(r) = -T \sum_{i=r+1}^{g} \ln(1 - \hat{\lambda}_i)$$

and

$$\lambda_{max}(r, r+1) = -T \ln(1 - \hat{\lambda}_{r+1})$$

where r is the number of cointegrating vectors under the null hypothesis and $\hat{\lambda}_i$ is the estimated value for the ith ordered eigenvalue from the Γ matrix.

Intuitively, the larger is $\hat{\lambda}_i$, the more large and negative will be $\ln(1-\hat{\lambda}_i)$ and hence the larger will be the test statistic. Each eigenvalue will have associated with it a different cointegrating vector, which will be eigenvectors. A significantly non-zero eigenvalue indicates a significant cointegrating vector.

λ_{trace} is a joint test where the null is that the number of cointegrating vectors is less than or equal to r against an unspecified or general alternative that there are more than r. It starts with p eigenvalues, and then successively the largest is removed. $\lambda_{trace} = 0$ when all the $\lambda_i = 0$, for $i = 1, \cdots, g$.

λ_{max} conducts separate tests on each eigenvalue, and has as its null hypothesis that the number of cointegrating vectors is r against an alternative of $r + 1$.

Johansen and Juselius [13] provide critical values for the two statistics. The distribution of the test statistics is non-standard, and the critical values depend on the value of $g-r$, the number of non-stationary components and whether constants are included in each of the equations.

If the test statistic is greater than the critical value from Johansen's tables, reject the null hypothesis that there are r cointegrating vectors in favour of the alternative that there are $r + 1$ (for λ_{trace}) or more than r (for λ_{max}). The testing is conducted in a sequence and under the null, $r = 0, 1, \cdots, g-1$ so that the hypotheses for λ_{max} are

$$\begin{array}{lll} H_0 : r = 0 & \text{versus} & H_1 : 0 < r \leq g \\ H_0 : r = 1 & \text{versus} & H_1 : 1 < r \leq g \\ H_0 : r = 2 & \text{versus} & H_1 : 2 < r \leq g \\ \vdots & \vdots & \vdots \\ H_0 : r = g - 1 & \text{versus} & H_1 : r = g. \end{array}$$

The first test involves a null hypothesis of no cointegrating vectors (corresponding to Γ having zero rank). If this null is not rejected, it would be concluded that there are no cointegrating vectors and the testing would be completed. However, if $H_0 : r = 0$ is rejected, the null that there is one cointegrating vector (i.e. $H_0 : r = 1$) would be tested and so on. Thus the value of r is continually increased until the null is no longer rejected.

But how does this correspond to a test of the rank of the Γ matrix? r is the rank of Γ. Γ cannot be of full rank (g) since this would correspond to the original y_t being stationary. If Γ has zero rank, then by analogy to the univariate case, Δy_t depends only on Δy_{t-j} and not on y_{t-1}, so that there is no long-run relationship between the elements of y_{t-1}. Hence there is no cointegration. For $1 < \text{rank}(\Gamma) < g$, there are r cointegrating vectors. Γ is then defined as the product of two matrices, α and β', of dimension $(g \times r)$ and $(r \times g)$, respectively, i.e.

$$\Gamma = \alpha\beta'$$

The matrix β gives the cointegrating vectors, while α gives the amount of each cointegrating vector entering each equation of the VECM, also known as the "adjustment parameters".

8 Modelling volatility and correlation

8.1 Non-linear models

Linear structural (and time series) models are unable to explain a number of important features common to much financial data, including

• *Leptokurtosis* – the tendency for financial asset returns to have distributions that exhibit fat tails and excess peakedness at the mean.

• *Volatility clustering* or *volatility pooling* – the tendency for volatility in financial markets to appear in bunches).

• *Leverage effects* – the tendency for volatility to rise more following a large price fall than following a price rise of the same magnitude.

The modeling choice of linear versus non-linear should come at least in part from financial theory and may also be made partly on statistic grounds.

Regarding the tools to detect nonlinearity, "traditional" tools of time series analysis (such as estimates of the ACF or PACF, or "spectral analysis") are likely to be of little use. Useful tests for detecting non-linear patterns in time series can be broadly split into two types: general tests and specific tests.

General tests, also sometimes called "portmanteau" tests, are usually designed to detect many departures from randomness in data. Most applications of these tests conclude that there is non-linear dependence in financial asset returns series, but that the dependence is best characterised by a GARCH-type process.

• Perhaps the simplest general test for non-linearity is Ramsey's RESET test.

• One of the most widely used tests is known as the BDS test, which has as its null hypothesis that the data are pure noise and which has been argued to have power to detect a variety of departures from randomness.

• The bispectrum test.

• The bicorrelation test.

Specific tests, are usually designed to have power to find specific types of non-linear structure. Specific tests are unlikely to detect other forms of non-linearities in the data, but their results will by definition offer a class of models that should be relevant for the data at hand.

8.2 Models for volatility: EWMA, AR, ARCH

The **exponentially weighted moving average (EWMA)** model of volatility has two advantages over the simple historical model:

• Volatility is in practice likely to be affected more by recent events.

• The effect on volatility of a single given observation declines at an exponential rate, not abruptly once it falls out of the measurement sample.

The EWMA model of volatility has two important limitations:

• When the infinite sum in the theoretical formula of EWMA is replaced with a finite sum of observable data, the weights will now sum to less than 1. In the case of small samples, this could make a large difference and a correction may be necessary.

• The "mean-reverting" property of volatility time series is accounted for in GARCH volatility forecasting models, but not by EWMAs.

The idea of the **autoregressive (AR) volatility models** is that a time series of observations on some volatility proxy are obtained. The standard Box-Jenkins-type procedures for estimating autoregressive (or ARMA) models can then be applied to this series. The forecasts are also produced in the usual fashion discussed in Chapter 5 in the context of ARMA models.

The **autoregressive conditionally heteroscedastic (ARCH) models** capture the heteroscedasticity of the errors, as well as "volatility clustering" or "volatility pooling". Under the ARCH model, the "autocorrelation in volatility" is modelled by allowing the conditional variance of the error term, $\sigma_t^2 = E[(u_t - E[u_t])^2 | u_{t-1}, u_{t-2}, \cdots]$, to depend on the previous values of the squared error

$$\begin{cases} y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{it} + u_t \\ \sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_q u_{t-q}^2. \end{cases}$$

The test for "ARCH effects" is one of a joint null hypothesis that all q lags of the squared residuals have coefficient values that are not significantly different from zero. If the value of the test statistic is greater than the critical value from the χ^2 distribution, then reject the null hypothesis. The steps of the test are

(1) Run any postulated linear regression of y regressed against x's, saving the residuals, $\hat{u}_t$:

$$y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{it} + u_t$$

(2) Square the residuals, and regress them on q own lags to test for ARCH(q) effect:

$$\hat{u}_t^2 = \gamma_0 + \sum_{j=1}^{q} \gamma_j \hat{u}_{t-j}^2 + v_t$$

where v_t is an error term. Obtain R^2 from this regression.

(3) The test statistic is defined as TR^2 (the number of observations multiplied by the coefficient of multiple correlation) from the last regression, and is distributed as a $\chi^2(q)$.

(4) The null and alternative hypotheses are

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_q = 0$$
$$H_1 : \gamma_1 \neq 0 \text{ or } \gamma_2 \neq 0 \text{ or } \cdots \text{ or } \gamma_q \neq 0$$

Limitations of ARCH(q) models. ARCH models themselves have rarely been used in the last decade or more, since they bring with them a number of difficulties:

• How should the value of q be decided? One approach is the use of a likelihood ratio test, although there is no clearly best approach.

• The value of q might be very large.

• Non-negativity constraints might be violated. Everything else equal, the more parameters there are in the conditional variance equation, the more likely it is that one or more of them will have negative estimated values.

8.3 Models for volatility: Generalised ARCH (GARCH), GJR, EGARCH

The **GARCH model** allows the conditional variance to be dependent upon previous own lags. A GARCH(p, q) model is formulated as

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i u_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2.$$

Compared with ARCH models, GARCH models are more parsimonious and avoids overfitting. Consequently, the model is less likely to breach non-negativity constraints. In general a GARCH(1,1) model will be sufficient to capture the volatility clustering in the data, and rarely is any higher order model estimated or even entertained in the academic finance literature.

Estimation of ARCH/GARCH models. Since the GARCH model is no longer of the usual linear form, OLS cannot be used for GARCH model estimation, because OLS minimises the residual sum of squares. The *RSS* depends only on the parameters in the conditional mean equation, and not the conditional variance, and hence *RSS* minimisation is no longer an appropriate objective. Instead, the appropriate estimation method is *maximum likelihood*, which follows the following steps:

(1) Specify the appropriate equations for the mean and the variance – e.g. an AR(1)-GARCH(1,1) model

$$\begin{cases} y_t = \mu + \phi y_{t-1} + u_t, \ u_t \sim N(0, \sigma_t^2) \\ \sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \beta \sigma_{t-1}^2 \end{cases}$$

(2) Specify the log-likelihood function (*LLF*) to maximise under a normality assumption for the disturbances

$$L = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(\sigma_t^2) - \frac{1}{2}\sum_{t=1}^{T}(y_t - \mu - \phi y_{t-1})^2/\sigma_t^2$$

(3) The computer will maximise the function and generate parameter values that maximise the *LLF* and will construct their standard errors.

N.B.:

1) Note the unconditional variance of u_t is constant and given by

$$\text{Var}(u_t) = \frac{\alpha_0}{1 - (\alpha_1 + \beta)}$$

so long as $\alpha_1 + \beta < 1$. For $\alpha_1 + \beta \geq 1$, the unconditional variance of u_t is not defined, and this would be termed "non-stationarity in variance". $\alpha_1 + \beta = 1$ would be known as a "unit root in variance", also termed "**Integrated GARCH**" or IGARCH.

2) The *conditional normality* assumption for u_t is essential in specifying the likelihood function. In the context of non-normality, the usual standard error estimates will be inappropriate, and a different variance-covariance matrix estimator that is robust to non-normality, due to Bollerslev and Wooldridge, should be used. This procedure is known as **quasi-maximum likelihood**, or QML.

The **GJR model** is a simple extension of GARCH with an additional term added to account for possible asymmetries. It assumes the conditional variance is given by

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma u_{t-1}^2 I_{t-1}$$

where

$$I_{t-1} = \begin{cases} 1 & \text{if } u_{t-1} < 0 \\ 0 & \text{otherwise.} \end{cases}$$

The **exponential GARCH model** assumes

$$\ln(\sigma_t^2) = \omega + \beta \ln(\sigma_{t-1}^2) + \gamma \frac{u_{t-1}}{\sqrt{\sigma_{t-1}^2}} + \alpha \left[\frac{|u_{t-1}|}{\sqrt{\sigma_{t-1}^2}} - \sqrt{\frac{2}{\pi}} \right].$$

9 Switching models

Skipped for this version.

10 Panel data

The simplest way to deal with panel data would be to estimate a pooled regression, which would involve estimating a single equation on all the data together. This approach has some severe limitations. Most importantly, pooling the data in this way implicitly assumes that the average values of the variables and the relationships between them are constant over time and across all of the cross-sectional units in the sample.

One approach to making more full use of the structure of panel data would be to use the seemingly unrelated regression (SUR) framework. The idea behind SUR is essentially to transform the model so that the error terms become uncorrelated. If the correlations between the error terms in the individual equations had been zero in the first place, then SUR on the system of equations would have been equivalent to running separate OLS regressions on each equation.

However, the applicability of the SUR technique is limited because it can be employed only when the number of time-series observations per cross-sectional unit is at least as large as the total number of such units. A second problem with SUR is that the number of parameters to be estimated in total is very large. For these reasons, the more flexible full panel data approach is much more commonly used.

There are broadly two classes of panel estimator approaches: *fixed effects models* and *random effects models*.

10.1 The fixed effects model

Skipped for this version.

10.2 Time-fixed effects models

Skipped for this version.

10.3 The random effects model

Skipped for this version.

Fixed or random effects? The random effects model is more appropriate when the entities in the sample can be thought of as having been randomly selected from the population, but a fixed effect model is more plausible when the entities in the sample effectively constitute the entire population. The random effects approach has a major drawback which arises from the fact that it is valid only when the composite error term ω_{it} is uncorrelated with all of the explanatory variables. A test for whether this assumption is valid for the random effects estimator is based on a slightly more complex version of the Hausman test described in section 6.6.

11 Limited dependent variable models

11.1 Common limited dependent variable models

The situation where the explained variable is qualitative is referred to as a *limited dependent variable.*

The linear probability model (LPM).

$$P_i = p(y_i = 1) = \beta_0 + \sum_{l=1}^{k} \beta_l x_{li} + u_i,\ i = 1, \cdots, N.$$

The LPM also suffers from a couple of more standard econometric problems that we have examined in previous chapters. First, since the dependent variable takes only one or two values, for given (fixed in repeated samples) values of the explanatory variables, the disturbance term will also take on only one of two values. Hence the error term cannot plausibly be assumed to be normally distributed. Since u_i changes systematically with the explanatory variables, the disturbances will also be heteroscedastic. It is therefore essential that heteroscedasticity-robust standard errors are always used in the context of limited dependent variable models.

The logit model. Let $F(z) = \frac{1}{1+e^{-z}}$, and the model specification is

$$P_i = p(y_i = 1) = F\left(\beta_0 + \sum_{l=1}^{k} \beta_l x_{li} + u_i\right),\ i = 1, \cdots, N.$$

This model is not linear and thus is not estimable using OLS. Instead, maximum likelihood is usually used.

The probit model. Let $F(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma}}$, and the model specification is

$$P_i = p(y_i = 1) = F\left(\beta_0 + \sum_{l=1}^{k} \beta_l x_{li} + u_i\right),\ i = 1, \cdots, N.$$

For the majority of the applications, the logit and probit models will give very similar characterisations of the data because the densities are very similar. That is, the fitted regression plots will be virtually indistinguishable and the implied relationships between the explanatory variables and the probability that $y_i = 1$ will also be very similar. Both approaches are much preferred to the linear probability model. The only instance where the models may give non-negligibility different results occurs when the split of the y_i between 0 and 1 is very unbalanced.

11.2 Estimation of limited dependent variable models

Given that both logit and probit are non-linear models, they cannot be estimated by OLS. Maximum likelihood (ML) is invariably used in practice – we form the appropriate log-likelihood function and then the software package will find the values of the parameters that jointly maximise it using an iterative search procedure. Once the model parameters have been estimated, standard errors can be calculated and hypothesis tests conducted. While t-test statistics are constructed in the usual way, the standard error formulae used following the ML estimation are valid asymptotically only. Consequently, it is common to use the critical values from a normal distribution rather than a t-distribution with the implicit assumption that the sample size is sufficiently large.

11.3 Goodness of fit measures for linear dependent variable models

While it would be possible to calculate the values of the standard goodness of fit measures such as RSS, R^2 or $\bar{R}^2$ for linear dependent variable models, these cease to have any real meaning. The objective of ML is to maximise the value of the log-likelihood function (LLF), not to minimise the RSS. Moreover, R^2 and adjusted R^2, if calculated in the usual fashion, will be misleading because the fitted values from the model can take on any value but the actual values will be only either 0 and 1.

Two goodness of fit measures that are commonly reported for limited dependent variable models are *percent correct predictions* and *pseudo-*R^2.

Percent correct predictions. The percentage of y_i values correctly predicted, defined as $100\times$ the number of observations predicted correctly divided by the total number of observations:

$$\text{Percent correct predictions} = \frac{100}{N}\sum_{i=1}^{N}\left[y_i I(\hat{P}_i) + (1-y_i)(1-I(\hat{P}_i))\right]$$

where $I(\hat{y}_i) = 1$ if $\hat{y}_i > \bar{y}$ and 0 otherwise. This goodness of fit measure is not ideal, since it is possible that a "naive predictor" could do better than any model if the sample is unbalanced between 0 and 1. An alternative measure is proposed as the percentage of $y_i = 1$ correctly predicted plus the percentage of $y_i = 0$ correctly predicted:

$$\text{Percent correct predictions} = 100 \times \left[\frac{\sum y_i I(\hat{P}_i)}{\sum y_i} + \frac{\sum(1-y_i)(1-I(\hat{P}_i))}{N - \sum y_i}\right]$$

Pseudo-R^2. This measure is defined as

$$\text{pseudo-}R^2 = 1 - \frac{LLF}{LLF_0}$$

where LLF is the maximised value of the log-likelihood function for the logit and probit model and LLF_0 is the value of the log-likelihood function for a restricted model where all of the slope parameters are set to zero. Pseudo-R^2 does not have any intuitive interpretation.

11.4 Multinomial linear dependent variables

Multinomial logit/probit model.

Independence of irrelevant alternatives.

11.5 Ordered response linear dependent variables models

Since only the ordering can be interpreted with data and not the actual numerical values, OLS cannot be employed and a technique based on ML is used instead. The models used are generalisations of logit and probit, known as *ordered logit* and *ordered probit*.

11.6 Censored and truncated dependent variables

For both censored and truncated data, OLS will not be appropriate, and an approach based on maximum likelihood must be used, although the model in each case would be slightly different. Censored data occur when the dependent variable has been "censored" at a certain point so that values above (or below) this cannot be observed. Even though the dependent variable is censored, the corresponding values of the independent variables are still observable. A truncated dependent variable, meanwhile, occurs when the observations for both the dependent and the independent variables are missing when the dependent variable is above (or below) a certain threshold.

Censored dependent variable models. Two important limitations of tobit modelling should be noted. First, such models are much more seriously affected by non-normality and heteroscedasticity than are standard regression models, and biased and inconsistent estimation will result. Second, the tobit model requires it to be plausible that the dependent variable can have values close to the limit.

Truncated dependent variable models. For truncated data, a more general model is employed that contains two equations – one for whether a particular data point will fall into the observed or constrained categories and another for modelling the resulting variable. The second equation is equivalent to the tobit approach.

For more details on these two types of models, see Pedace [18].

12 Simulation methods

12.1 Variance reduction techniques

Antithetic variates. The use of low-discrepancy sequences leads the Monte Carlo standard errors to be reduced in direct proportion to the number of replications rather than in proportion to the square root of the number of replications. Thus,

Control variates.

Random number re-usage across experiments. Although of course it would not be sensible to re-use sets of random number draws within a Monte Carlo experiment, using the same sets of draws across experiments can greatly reduce the variability of the difference in the estimates across those experiments. Another possibility involves taking long series of draws and then slicing them up into several smaller sets to be used in different experiments.

Random number re-usage is unlikely to save computational time, for making the random draws usually takes a very small proportion of the overall time taken to conduct the whole experiment.

12.2 Bootstrapping

Suppose a sample of data, $y = y_1, y_2, \cdots, y_T$ are available and it is desired to estimate some parameter θ. An approximation to the statistical properties of $\hat{\theta}_T$ can be obtained by studying a sample of bootstrap estimators. This is done by taking N samples of size T with replacement from y and re-calculating $\hat{\theta}$ with each new sample. A series of $\hat{\theta}$ estimates is then obtained, and their distribution can be considered.

In econometrics, the bootstrap has been used in the context of unit root testing. Another important recent proposed use of the bootstrap is as a method for detecting data snooping (data mining) in the context of tests of the profitability of technical trading rules. The technique works by placing the rule under study in the context of a "universe" of broadly similar trading rules. This gives some empirical content to the notion that a variety of rules may have been examined before the final rule is selected. The bootstrap is applied to each trading rule, by sampling with replacement from the time series of observed returns for that rule. The null hypothesis is that there does not exist a superior technical trading rule. Sullivan, Timmermann and White show how a p-value of the "reality check" bootstrap-based test can be constructed, which evaluates the significance of the returns (or excess returns) to the rule after allowing for the fact that the whole universe of rules may have been examined.

Data snooping biases are apparent in other aspects of estimation and testing in finance. Lo and MacKinlay [15] find that tests of financial asset pricing models (CAPM) may yield misleading inferences when properties of the data are used to construct the test statistics. These properties relate to the construction of portfolios based on some empirically motivated characteristic of the stock, such as market capitalisation, rather than a theoretically motivated characteristic, such as dividend yield.

There are at least two situations where the bootstrap will not work well: *outliers in the data* and *non-independent data*.

13 Conducting empirical research or doing a project or dissertation in finance

In terms of econometrics, conducting empirical research is one of the best ways to get to grips with the theoretical material, and to find out what practical difficulties econometricians encounter when conducting research. Many web sites contain lists of journals in finance or links to finance journals. For some useful ones, see [2, page 588] for details

14 Recent and future developments in the modelling of financial time series

14.1 What was not covered in the book

Bayesian statistics. Under the classical approach, the researcher postulates a theory and estimates a model to test that theory. Tests of the theory are conducted using the estimated model within the "classical" hypothesis testing framework. Based on the empirical results, the theory is either refuted or upheld by the data.

Under a Bayesian approach, the researcher would start with an assessment of the existing state of knowledge or beliefs, formulated into a set of probabilities. These prior inputs or priors would then be combined with the observed data via a likelihood function. The beliefs and the probabilities would then be updated as a result of the model estimation, resulting in a set of posterior probabilities. Probabilities are thus updated sequentially, as more data become available.

The Bayesian approach to estimation and inference has found a number of important recent applications in financial econometrics, in particular in the context of *GARCH modelling*, *asset allocation*, and *portfolio performance evaluation*.

Note if the researcher set very strong priors, an awful lot of evidence against them would be required for the notion to be refuted. Contrast this with the classical case, where the data are usually permitted to freely determine whether a theory is upheld or refuted, irrespective of the researcher' s judgement.

Chaos in financial markets. Almost without exception, applications of chaos theory to financial markets have been unsuccessful. Academic and practitioner interest in chaotic models for financial markets has arguably almost disappeared. Financial markets are extremely complex, involving a very large number of different participants, each with different objectives and different sets of information – and, above all, each of whom are human with human emotions and irrationalities. The consequence of this is that financial and economic data are usually far noisier and "more random" than data from other disciplines, making the specification of a deterministic model very much harder and possibly even futile.

Neural network models. Artificial neural networks (ANNs) have been widely employed in finance for tackling time series and classification problems, including forecasting financial asset returns, volatility, bankruptcy„ and takeover prediction.

Neural networks have virtually no theoretical motivation in finance, but owe their popularity to their ability to fit any functional relationship in the data to an arbitrary degree of accuracy. They are likely to work best in situations where financial theory has virtually nothing to say about the likely functional form. However, their popularity has arguably waned over the past five years or so as a consequence of several perceived problems with their employment. First, the coefficient estimates from neural networks do not have

any real theoretical interpretation. Second, virtually no diagnostic or specification tests are available for estimated models to determine whether the model under consideration is adequate. Third, ANN models can provide excellent fits in-sample to a given set of "training" data, but typically provide poor out-of-sample forecast accuracy. Finally, the non-linear estimation of neural network models can be cumbersome and computationally time-intensive.

Long-memory models.

14.2 Financial econometrics: the future?

An excellent overview of the state of the art in a vast array of areas in econometrics is provided by Mills and Patterson [17].

Tail models. A mixture of normal distributions with difference variances will lead to an overall series that is leptokurtic. Second, a Student's t distribution can be used. Finally, the "stable" distributions that fall under the general umbrella of extreme value theory can be used.

Copulas and quantile regressions.

Market microstructure. "Market microstructure" may broadly be defined as the process whereby investors' preferences and desires are translated into financial market transactions. There has been considerable advancement in the sophistication of econometric models applied to microstructure problems. An important innovation was the Autoregressive Conditional Duration (ACD) model due to Engle and Russell.

Computational techniques for options pricing and other uses.

Higher moment models.

References

[1] G. E. P. Box and D. A. Pierce. "Distributions of Residual Autocorrelations in Autoregressive Integrated Moving Average Models", *Journal of the American Statistical Association* 65, 1509–26, 1970. 3

[2] Chris Brooks. *Introductory Econometrics for Finance*, 2nd Edition, Cambridge University Press, 2008. 1, 26

[3] S.-H. Chu. "Short-Run Forecasting of Commodity Prices: An Application of Autoregressive Moving Average Models", *IMF Staff Papers* 25, 90-111, 1978. 9

[4] D. A. Dickey and W. A. Fuller. "Distribution of Estimators fro Time Series Regressions with a Unit Root", *Journal of the American Statistical Association* 74, 427-31, 1979. 14

[5] D. A. Dickey and W. A. Fuller. "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root", *Econometrica* 49(4), 1057-72, 1981. 15

[6] W. A. Fuller. *Introduciton to Statistical Time Series*, Wiley, New York, 1976. 14

[7] M. E. Gerlow, S. H. Irwin, and T.-R. Liu. "Economic Evaluation of Commodity Price Forecasting Models", *International Journal of Forecasting* 9, 387-97, 1993. 8

[8] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods", *Econometrica* 37, 424–38, 1969. 12

[9] C. W. J. Granger and P. Newbold. *Forecasting Economic Time Series*, 2nd edn., Academic Press, San Diego, CA, 1986. 7

[10] William H. Greene. *Econometric Analysis*, 5th Edition, Prentice Hall, 2002. 10

[11] J. D. Hamilton. *Time Series Analysis*, Princeton University Press, 1994. 16

[12] D. F. Hendry and K. Juselius. "Explaining Cointegration Analysis: Part I", *Energy Journal* 21, 1–42, 2000. 16

[13] S. Johansen and K. Juselius. "Maximum Likelihood Estimatin and Inference on Cointegration with Applications to the Demand for Money", *Oxford Bulletin of Economics and Statistics* 52, 169-210, 1990. 19

[14] G. Leitch and J. E. Tanner. "Economic Forecast Evaluation: Profit Versus the Conventional Error Measures", *American Economic Review* 81(3), 580-90, 1991. 8

[15] Lo, A. W. and MacKinlay, C. A. (1990) "Data-Snooping Biases in Tests of Financial Asset Pricing Models", *Review of Financial Studies* 3, 431–67. 26

[16] G. M. Ljung and G. E. P. Box. "On a Measure of Lack of Fit in Time Series Models", *Biometrika* 65(2), 297–303, 1978. 4

[17] Mills, T. C. and Patterson, K. D. (eds.) (2006). *Palgrave Handbook of Econometrics Volume 1: Econometric Theory*, Palgrave Macmillan, Basingstoke, UK. 27

[18] Roberto Pedace. *Econometrics for dummies*. Hoboken, John Wiley & Sons Inc., 2013. 25

[19] C. A. Sims. "Money, Income, and Causality", *American Economic Review* 62(4), 540–52, 1972. 12

[20] Y. Zeng. "Classicial Linear Regression Model: Assumptions and Diagnostic Tests", study notes, 2016 3