9 Switching models

Learning Outcomes

In this chapter, you will learn how to

- Use intercept and slope dummy variables to allow for seasonal behaviour in time series
- Motivate the use of regime switching models in financial econometrics
- Specify and explain the logic behind Markov switching models
- Compare and contrast Markov switching and threshold autoregressive models
- Describe the intuition behind the estimation of regime switching models

9.1 Motivations

Many financial and economic time series seem to undergo episodes in which the behaviour of the series changes quite dramatically compared to that exhibited previously. The behaviour of a series could change over time in terms of its mean value, its volatility, or to what extent its current value is related to its previous value. The behaviour may change once and for all, usually known as a 'structural break' in a series. Or it may change for a period of time before reverting back to its original behaviour or switching to yet another style of behaviour, and the latter is typically termed a 'regime shift' or 'regime switch'.

9.1.1 What might cause one-off fundamental changes in the properties of a series?

Usually, very substantial changes in the properties of a series are attributed to large-scale events, such as wars, financial panics – e.g. a 'run

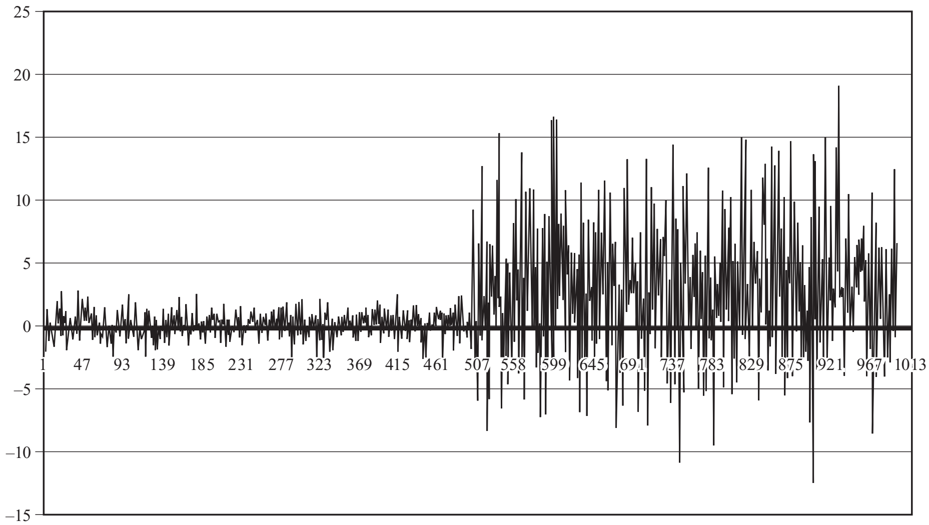


Figure 9.1
Sample time series plot illustrating a regime shift

on a bank', significant changes in government policy, such as the introduction of an inflation target, or the removal of exchange controls, or changes in market microstructure – e.g. the 'Big Bang', when trading on the London Stock Exchange (LSE) became electronic, or a change in the market trading mechanism, such as the partial move of the LSE from a quote-driven to an order-driven system in 1997.

However, it is also true that regime shifts can occur on a regular basis and at much higher frequency. Such changes may occur as a result of more subtle factors, but still leading to statistically important modifications in behaviour. An example would be the intraday patterns observed in equity market bid–ask spreads (see chapter 6). These appear to start with high values at the open, gradually narrowing throughout the day, before widening again at the close.

To give an illustration of the kind of shifts that may be seen to occur, figure 9.1 gives an extreme example.

As can be seen from figure 9.1, the behaviour of the series changes markedly at around observation 500. Not only does the series become much more volatile than previously, its mean value is also substantially increased. Although this is a severe case that was generated using simulated data, clearly, in the face of such 'regime changes' a linear model estimated over the whole sample covering the change would not be appropriate. One possible approach to this problem would be simply to split the data around the time of the change and to estimate separate models on each portion. It would be possible to allow a series, y_t to be drawn from two or more different generating processes at different times. For example, if it was thought an AR(1) process was appropriate to capture the relevant features of a particular series whose behaviour changed at observation 500, say, two models could be estimated:

$$y_t = \mu_1 + \phi_1 y_{t-1} + u_{1t} \quad \text{before observation 500} \tag{9.1}$$

$$y_t = \mu_2 + \phi_2 y_{t-1} + u_{2t} \quad \text{after observation 500} \tag{9.2}$$

In the context of figure 9.1, this would involve focusing on the mean shift only. These equations represent a very simple example of what is known as a piecewise linear model – that is, although the model is globally (i.e. when it is taken as a whole) non-linear, each of the component parts is a linear model.

This method may be valid, but it is also likely to be wasteful of information. For example, even if there were enough observations in each sub-sample to estimate separate (linear) models, there would be an efficiency loss in having fewer observations in each of two samples than if all the observations were collected together. Also, it may be the case that only one property of the series has changed – for example, the (unconditional) mean value of the series may have changed, leaving its other properties unaffected. In this case, it would be sensible to try to keep all of the observations together, but to allow for the particular form of the structural change in the model-building process. Thus, what is required is a set of models that allow all of the observations on a series to be used for estimating a model, but also that the model is sufficiently flexible to allow different types of behaviour at different points in time. Two classes of regime switching models that potentially allow this to occur are *Markov switching models* and *threshold autoregressive models*.

A first and central question to ask is: How can it be determined where the switch(es) occurs? The method employed for making this choice will depend upon the model used. A simple type of switching model is one where the switches are made deterministically using dummy variables. One important use of this in finance is to allow for 'seasonality' in financial data. In economics and finance generally, many series are believed to exhibit seasonal behaviour, which results in a certain element of partly predictable cycling of the series over time. For example, if monthly or quarterly data on consumer spending are examined, it is likely that the value of the series will rise rapidly in late November owing to Christmas-related expenditure, followed by a fall in mid-January, when consumers realise that they have spent too much before Christmas and in the January sales! Consumer spending in the UK also typically drops during the August vacation period when all of the sensible people have left the country. Such phenomena will be apparent in many series and will be present to some degree at the same time every year, whatever else is happening in terms of the long-term trend and short-term variability of the series.

9.2 Seasonalities in financial markets: introduction and literature review

In the context of financial markets, and especially in the case of equities, a number of other 'seasonal effects' have been noted. Such effects are usually known as 'calendar anomalies' or 'calendar effects'. Examples include open- and close-of-market effects, 'the January effect', weekend effects and bank holiday effects. Investigation into the existence or otherwise of 'calendar effects' in financial markets has been the subject of a considerable amount of recent academic research. Calendar effects may be loosely defined as the tendency of financial asset returns to display systematic patterns at certain times of the day, week, month, or year. One example of the most important such anomalies is the *day-of-the-week effect*, which results in average returns being significantly higher on some days of the week than others. Studies by French (1980), Gibbons and Hess (1981) and Keim and Stambaugh (1984), for example, have found that the average market close-to-close return in the US is significantly negative on Monday and significantly positive on Friday. By contrast, Jaffe and Westerfield (1985) found that the lowest mean returns for the Japanese and Australian stock markets occur on Tuesdays.

At first glance, these results seem to contradict the efficient markets hypothesis, since the existence of calendar anomalies might be taken to imply that investors could develop trading strategies which make abnormal profits on the basis of such patterns. For example, holding all other factors constant, equity purchasers may wish to sell at the close on Friday and to buy at the close on Thursday in order to take advantage of these effects. However, evidence for the predictability of stock returns does not necessarily imply market inefficiency, for at least two reasons. First, it is likely that the small average excess returns documented by the above papers would not generate net gains when employed in a trading strategy once the costs of transacting in the markets has been taken into account. Therefore, under many 'modern' definitions of market efficiency (e.g. Jensen, 1978), these markets would not be classified as inefficient. Second, the apparent differences in returns on different days of the week may be attributable to time-varying stock market risk premiums.

If any of these calendar phenomena are present in the data but ignored by the model-building process, the result is likely to be a misspecified model. For example, ignored seasonality in y_t is likely to lead to residual autocorrelation of the order of the seasonality – e.g. fifth order residual autocorrelation if y_t is a series of daily returns.

9.3 Modelling seasonality in financial data

As discussed above, seasonalities at various different frequencies in financial time series data are so well documented that their existence cannot be doubted, even if there is argument about how they can be rationalised. One very simple method for coping with this and examining the degree to which seasonality is present is the inclusion of dummy variables in regression equations. The number of dummy variables that could sensibly be constructed to model the seasonality would depend on the frequency of the data. For example, four dummy variables would be created for quarterly data, 12 for monthly data, five for daily data and so on. In the case of quarterly data, the four dummy variables would be defined as follows:

$D1_t = 1$ in quarter 1 and zero otherwise
$D2_t = 1$ in quarter 2 and zero otherwise
$D3_t = 1$ in quarter 3 and zero otherwise
$D4_t = 1$ in quarter 4 and zero otherwise

How many dummy variables can be placed in a regression model? If an intercept term is used in the regression, the number of dummies that could also be included would be one less than the 'seasonality' of the data. To see why this is the case, consider what happens if all four dummies are used for the quarterly series. The following gives the values that the dummy variables would take for a period during the mid-1980s, together with the sum of the dummies at each point in time, presented in the last column:

		$D1$	$D2$	$D3$	$D4$	Sum
1986	Q1	1	0	0	0	1
	Q2	0	1	0	0	1
	Q3	0	0	1	0	1
	Q4	0	0	0	1	1
1987	Q1	1	0	0	0	1
	Q2	0	1	0	0	1
	Q3	0	0	1	0	1

etc.

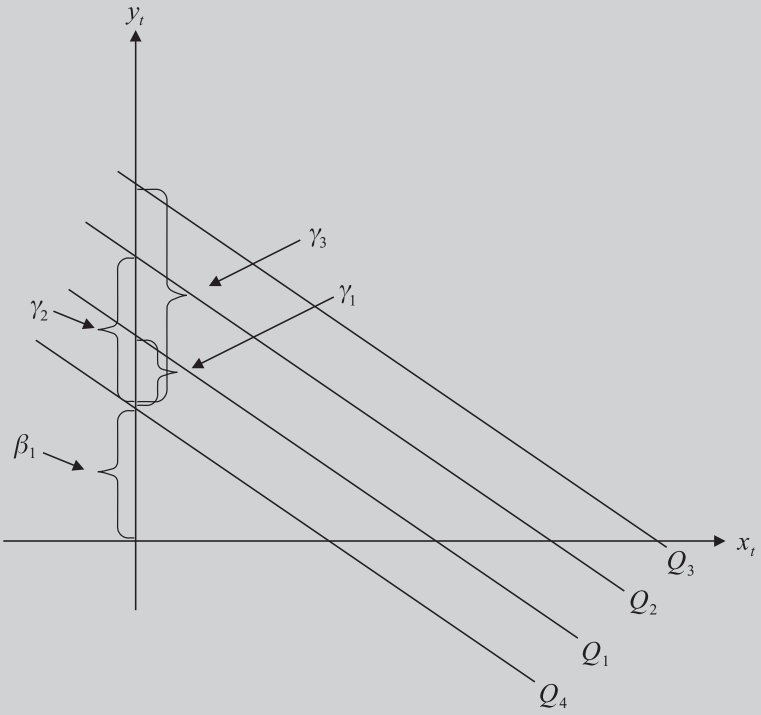
The sum of the four dummies would be 1 in every time period. Unfortunately, this sum is of course identical to the variable that is implicitly attached to the intercept coefficient. Thus, if the four dummy variables and the intercept were both included in the same regression, the problem would be one of perfect multicollinearity so that $(X'X)^{-1}$ would not exist and none of the coefficients could be estimated. This problem is known as the *dummy variable trap*. The solution would be either to just use three dummy variables plus the intercept, or to use the four dummy variables with no intercept.

The seasonal features in the data would be captured using either of these, and the residuals in each case would be identical, although the interpretation of the coefficients would be changed. If four dummy variables were used (and assuming that there were no explanatory variables in the regression), the estimated coefficients could be interpreted as the average value of the dependent variable during each quarter. In the case where a constant and three dummy variables were used, the interpretation of the estimated coefficients on the dummy variables would be that they represented the average deviations of the dependent variables for the included quarters from their average values for the excluded quarter, as discussed in the example below.

Box 9.1 How do dummy variables work?

The dummy variables as described above operate by *changing the intercept*, so that the average value of the dependent variable, given all of the explanatory variables, is permitted to change across the seasons. This is shown in figure 9.2.

Figure 9.2
Use of intercept dummy variables for quarterly data



Consider the following regression

$$y_t = \beta_1 + \gamma_1 D1_t + \gamma_2 D2_t + \gamma_3 D3_t + \beta_2 x_{2t} + \cdots + u_t \tag{9.3}$$

During each period, the intercept will be changed. The intercept will be:

- $\hat{\beta}_1 + \hat{\gamma}_1$ in the first quarter, since $D1 = 1$ and $D2 = D3 = 0$ for all quarter 1 observations
- $\hat{\beta}_1 + \hat{\gamma}_2$ in the second quarter, since $D2 = 1$ and $D1 = D3 = 0$ for all quarter 2 observations.
- $\hat{\beta}_1 + \hat{\gamma}_3$ in the third quarter, since $D3 = 1$ and $D1 = D2 = 0$ for all quarter 3 observations
- $\hat{\beta}_1$ in the fourth quarter, since $D1 = D2 = D3 = 0$ for all quarter 4 observations.

Example 9.1

Brooks and Persand (2001a) examine the evidence for a day-of-the-week effect in five Southeast Asian stock markets: South Korea, Malaysia, the Philippines, Taiwan and Thailand. The data, obtained from Primark Datastream, are collected on a daily close-to-close basis for all weekdays (Mondays to Fridays) falling in the period 31 December 1989 to 19 January 1996 (a total of 1,581 observations). The first regressions estimated, which constitute the simplest tests for day-of-the-week effects, are of the form

$$r_t = \gamma_1 D1_t + \gamma_2 D2_t + \gamma_3 D3_t + \gamma_4 D4_t + \gamma_5 D5_t + u_t \tag{9.4}$$

where r_t is the return at time t for each country examined separately, $D1_t$ is a dummy variable for Monday, taking the value 1 for all Monday observations and zero otherwise, and so on. The coefficient estimates can be interpreted as the average sample return on each day of the week. The results from these regressions are shown in table 9.1.

Briefly, the main features are as follows. Neither South Korea nor the Philippines have significant calendar effects; both Thailand and Malaysia have significant positive Monday average returns and significant negative Tuesday returns; Taiwan has a significant Wednesday effect.

Dummy variables could also be used to test for other calendar anomalies, such as the January effect, etc. as discussed above, and a given regression can include dummies of different frequencies at the same time. For example, a new dummy variable $D6_t$ could be added to (9.4) for 'April effects', associated with the start of the new tax year in the UK. Such a variable, even for a regression using daily data, would take the value 1 for all observations falling in April and zero otherwise.

If we choose to omit one of the dummy variables and to retain the intercept, then the omitted dummy variable becomes the reference category

Table 9.1 Values and significances of days of the week coefficients

	Thailand	Malaysia	Taiwan	South Korea	Philippines
Monday	0.49E-3	0.00322	0.00185	0.56E-3	0.00119
	(0.6740)	(3.9804)**	(2.9304)**	(0.4321)	(1.4369)
Tuesday	−0.45E-3	−0.00179	−0.00175	0.00104	−0.97E-4
	(−0.3692)	(−1.6834)	(−2.1258)**	(0.5955)	(−0.0916)
Wednesday	−0.37E-3	−0.00160	0.31E-3	−0.00264	−0.49E-3
	(−0.5005)	(−1.5912)	(0.4786)	(−2.107)**	(−0.5637)
Thursday	0.40E-3	0.00100	0.00159	−0.00159	0.92E-3
	(0.5468)	(1.0379)	(2.2886)**	(−1.2724)	(0.8908)
Friday	−0.31E-3	0.52E-3	0.40E-4	0.43E-3	0.00151
	(−0.3998)	(0.5036)	(0.0536)	(0.3123)	(1.7123)

Notes: Coefficients are given in each cell followed by *t*-ratios in parentheses; * and ** denote significance at the 5% and 1% levels, respectively.
Source: Brooks and Persand (2001a).

against which all the others are compared. For example consider a model such as the one above, but where the Monday dummy variable has been omitted

$$r_t = \alpha + \gamma_2 D2_t + \gamma_3 D3_t + \gamma_4 D4_t + \gamma_5 D5_t + u_t \tag{9.5}$$

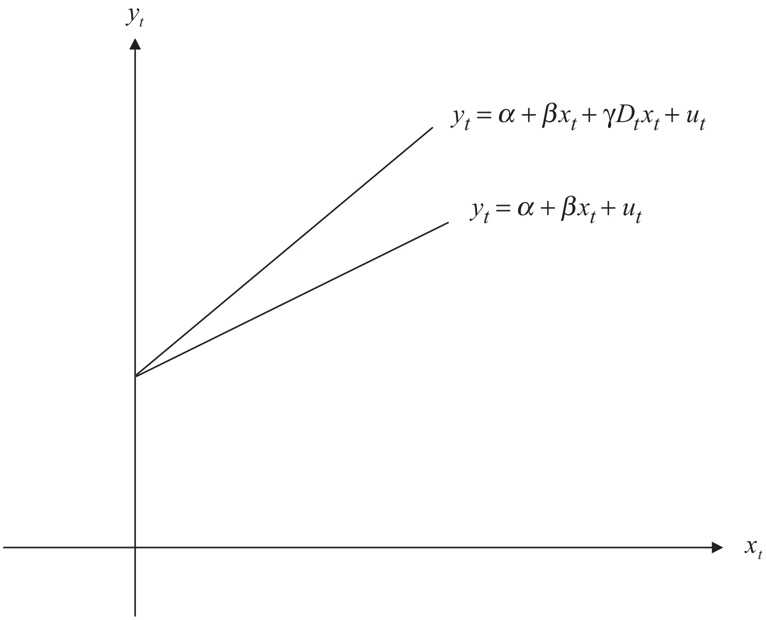
The estimate of the intercept will be $\hat{\alpha}$ on Monday, $\hat{\alpha} + \hat{\gamma}_{21}$ on Tuesday and so on. $\hat{\gamma}_2$ will now be interpreted as the difference in average returns between Monday and Tuesday. Similarly, $\hat{\gamma}_3, \ldots, \hat{\gamma}_5$ can also be interpreted as the differences in average returns between Wednesday, ..., Friday, and Monday.

This analysis should hopefully have made it clear that by thinking carefully about which dummy variable (or the intercept) to omit from the regression, we can control the interpretation to test naturally the hypothesis that is of most interest. The same logic can also be applied to slope dummy variables, which are described in the following section.

9.3.1 Slope dummy variables

As well as, or instead of, intercept dummies, slope dummy variables can also be used. These operate by changing the slope of the regression line, leaving the intercept unchanged. Figure 9.3 gives an illustration in the context of just one slope dummy (i.e. two different 'states'). Such a setup

Figure 9.3
Use of slope dummy variables


y_t
$y_t = \alpha + \beta x_t + \gamma D_t x_t + u_t$
$y_t = \alpha + \beta x_t + u_t$
x_t

would apply if, for example, the data were bi-annual (twice yearly) or bi-weekly or observations made at the open and close of markets. Then D_t would be defined as $D_t = 1$ for the first half of the year and zero for the second half.

A slope dummy changes the slope of the regression line, leaving the intercept unchanged. In the above case, the intercept is fixed at α, while the slope varies over time. For periods where the value of the dummy is zero, the slope will be β, while for periods where the dummy is one, the slope will be $\beta + \gamma$.

Of course, it is also possible to use more than one dummy variable for the slopes. For example, if the data were quarterly, the following setup could be used, with $D1_t \ldots D3_t$ representing quarters 1–3.

$$y_t = \alpha + \beta x_t + \gamma_1 D1_t x_t + \gamma_2 D2_t x_t + \gamma_3 D3_t x_t + u_t \tag{9.6}$$

In this case, since there is also a term in x_t with no dummy attached, the interpretation of the coefficients on the dummies (γ_1, etc.) is that they represent the deviation of the slope for that quarter from the average slope over all quarters. On the other hand, if the 4 slope dummy variables were included (and not βx_t), the coefficients on the dummies would be interpreted as the average slope coefficients during each quarter. Again, it is important not to include 4 quarterly slope dummies and the

βx_t in the regression together, otherwise perfect multicollinearity would result.

Example 9.2

Returning to the example of day-of-the-week effects in Southeast Asian stock markets, although significant coefficients in (9.4) will support the hypothesis of seasonality in returns, it is important to note that risk factors have not been taken into account. Before drawing conclusions on the potential presence of arbitrage opportunities or inefficient markets, it is important to allow for the possibility that the market can be more or less risky on certain days than others. Hence, low (high) significant returns in (9.4) might be explained by low (high) risk. Brooks and Persand thus test for seasonality using the empirical market model, whereby market risk is proxied by the return on the FTA World Price Index. Hence, in order to look at how risk varies across the days of the week, interactive (i.e. slope) dummy variables are used to determine whether risk increases (decreases) on the day of high (low) returns. The equation, estimated separately using time-series data for each country can be written

$$r_t = \left(\sum_{i=1}^{5} \alpha_i D_{it} + \beta_i D_{it} RWM_t \right) + u_t \tag{9.7}$$

where α_i and β_i are coefficients to be estimated, D_{it} is the ith dummy variable taking the value 1 for day $t = i$ and zero otherwise, and RWM_t is the return on the world market index. In this way, when considering the effect of market risk on seasonality, both risk and return are permitted to vary across the days of the week. The results from estimation of (9.6) are given in table 9.2. Note that South Korea and the Philippines are excluded from this part of the analysis, since no significant calendar anomalies were found to explain in table 9.1.

As can be seen, significant Monday effects in the Bangkok and Kuala Lumpur stock exchanges, and a significant Thursday effect in the latter, remain even after the inclusion of the slope dummy variables which allow risk to vary across the week. The *t*-ratios do fall slightly in absolute value, however, indicating that the day-of-the-week effects become slightly less pronounced. The significant negative average return for the Taiwanese stock exchange, however, completely disappears. It is also clear that average risk levels vary across the days of the week. For example, the betas for the Bangkok stock exchange vary from a low of 0.36 on Monday to a high of over unity on Tuesday. This illustrates that not only is there a significant positive Monday effect in this market, but also that the responsiveness of

Table 9.2 Day-of-the-week effects with the inclusion of interactive dummy variables with the risk proxy

	Thailand	Malaysia	Taiwan
Monday	0.00322 (3.3571)**	0.00185 (2.8025)**	0.544E-3 (0.3945)
Tuesday	−0.00114 (−1.1545)	−0.00122 (−1.8172)	0.00140 (1.0163)
Wednesday	−0.00164 (−1.6926)	0.25E-3 (0.3711)	−0.00263 (−1.9188)
Thursday	0.00104 (1.0913)	0.00157 (2.3515)*	−0.00166 (−1.2116)
Friday	0.31E-4 (0.03214)	−0.3752 (−0.5680)	−0.13E-3 (−0.0976)
Beta-Monday	0.3573 (2.1987)*	0.5494 (4.9284)**	0.6330 (2.7464)**
Beta-Tuesday	1.0254 (8.0035)**	0.9822 (11.2708)**	0.6572 (3.7078)**
Beta-Wednesday	0.6040 (3.7147)**	0.5753 (5.1870)**	0.3444 (1.4856)
Beta-Thursday	0.6662 (3.9313)**	0.8163 (6.9846)**	0.6055 (2.5146)*
Beta-Friday	0.9124 (5.8301)**	0.8059 (7.4493)**	1.0906 (4.9294)**

Notes: Coefficients are given in each cell followed by *t*-ratios in parentheses; * and ** denote significance at the 5% and 1%, levels respectively.
Source: Brooks and Persand (2001a).

Bangkok market movements to changes in the value of the general world stock market is considerably lower on this day than on other days of the week.

9.3.2 *Dummy variables for seasonality in EViews*

The most commonly observed calendar effect in monthly data is a *January effect*. In order to examine whether there is indeed a January effect in a monthly time series regression, a dummy variable is created that takes the value 1 only in the months of January. This is easiest achieved by creating a new dummy variable called JANDUM containing zeros everywhere, and then editing the variable entries manually, changing all of the zeros for January months to ones. Returning to the Microsoft stock price example of chapters 3 and 4, **Create this variable** using the methodology described above, and run the regression again including this new dummy variable as well. The results of this regression are:

```
Dependent Variable: ERMSOFT
Method: Least Squares
Date: 09/06/07 Time: 20:45
Sample (adjusted): 1986M05 2007M04
Included observations: 252 after adjustments

                     Coefficient    Std. Error    t-Statistic        Prob.

         C            −0.574717      1.334120     −0.430783         0.6670
     ERSANDP           1.522142      0.183517      8.294282         0.0000
       DPROD           0.522582      0.450995      1.158730         0.2477
     DCREDIT          −6.27E-05      0.000144     −0.435664         0.6635
   DINFLATION          2.162911      3.048665      0.709462         0.4787
      DMONEY          −1.412355      0.641359     −2.202129         0.0286
     DSPREAD           8.944002      12.16534      0.735203         0.4629
       RTERM           6.944576      2.978703      2.331409         0.0206
    FEB89DUM          −68.52799      12.62302     −5.428811         0.0000
    FEB03DUM          −66.93116      12.60829     −5.308503         0.0000
      JANDUM           6.140623      3.277966      1.873303         0.0622

R-squared              0.368162    Mean dependent var        −0.420803
Adjusted R-squared     0.341945    S.D. dependent var         15.41135
S.E. of regression     12.50178    Akaike info criterion      7.932288
Sum squared resid      37666.97    Schwarz criterion          8.086351
Log likelihood        −988.4683    Hannan-Quinn criter.       7.994280
F-statistic            14.04271    Durbin-Watson stat         2.135471
Prob(F-statistic)      0.000000
```

As can be seen, the dummy is just outside being statistically significant at the 5% level, and it has the expected positive sign. The coefficient value of 6.14, suggests that on average and holding everything else equal, Microsoft stock returns are around 6% higher in January than the average for other months of the year.

9.4 Estimating simple piecewise linear functions

The piecewise linear model is one example of a general set of models known as *spline techniques*. Spline techniques involve the application of polynomial functions in a piecewise fashion to different portions of the data. These models are widely used to fit yield curves to available data on the yields of bonds of different maturities (see, for example, Shea, 1984).

A simple piecewise linear model could operate as follows. If the relationship between two series, y and x, differs depending on whether x is smaller or larger than some threshold value x^*, this phenomenon can be captured using dummy variables. A dummy variable, D_t, could be defined, taking values

$$D_t = \begin{cases} 0 & \text{if } x_t < x^* \\ 1 & \text{if } x_t \geq x^* \end{cases} \tag{9.8}$$

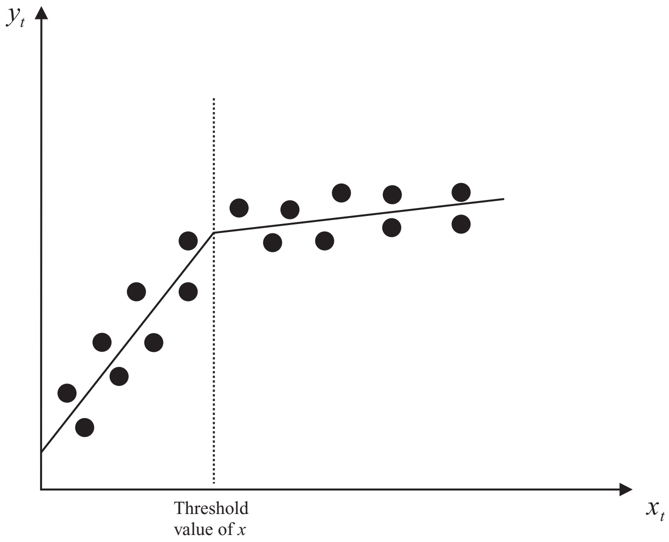
To offer an illustration of where this may be useful, it is sometimes the case that the tick size limits vary according to the price of the asset. For example, according to George and Longstaff (1993, see also chapter 6 of this book), the Chicago Board of Options Exchange (CBOE) limits the tick size to be \$(1/8) for options worth \$3 or more, and \$(1/16) for options worth less than \$3. This means that the minimum permissible price movements are \$(1/8) and (\$1/16) for options worth \$3 or more and less than \$3, respectively. Thus, if y is the bid–ask spread for the option, and x is the option price, used as a variable to partly explain the size of the spread, the spread will vary with the option price partly in a piecewise manner owing to the tick size limit. The model could thus be specified as

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 D_t + \beta_4 D_t x_t + u_t \tag{9.9}$$

with D_t defined as above. Viewed in the light of the above discussion on seasonal dummy variables, the dummy in (9.8) is used as both an intercept and a slope dummy. An example showing the data and regression line is given by figure 9.4.

Note that the value of the threshold or 'knot' is assumed known at this stage. Throughout, it is also possible that this situation could be

Figure 9.4
Piecewise linear model with threshold x^*



generalised to the case where y_t is drawn from more than two regimes or is generated by a more complex model.

9.5 Markov switching models

Although a large number of more complex, non-linear threshold models have been proposed in the econometrics literature, only two kinds of model have had any noticeable impact in finance (aside from threshold GARCH models of the type alluded to in chapter 8). These are the Markov regime switching model associated with Hamilton (1989, 1990), and the threshold autoregressive model associated with Tong (1983, 1990). Each of these formulations will be discussed below.

9.5.1 *Fundamentals of Markov switching models*

Under the Markov switching approach, the universe of possible occurrences is split into m states of the world, denoted s_i, $i = 1, \ldots, m$, corresponding to m regimes. In other words, it is assumed that y_t switches regime according to some unobserved variable, s_t, that takes on integer values. In the remainder of this chapter, it will be assumed that $m = 1$ or 2. So if $s_t = 1$, the process is in regime 1 at time t, and if $s_t = 2$, the process is in regime 2 at time t. Movements of the state variable between regimes are governed by a Markov process. This Markov property can be expressed as

$$\text{P}[a < y_t \leq b \mid y_1, y_2, \ldots, y_{t-1}] = \text{P}[a < y_t \leq b \mid y_{t-1}] \tag{9.10}$$

In plain English, this equation states that the probability distribution of the state at any time t depends only on the state at time $t-1$ and not on the states that were passed through at times $t-2, t-3, \ldots$ Hence Markov processes are not path-dependent. The model's strength lies in its flexibility, being capable of capturing changes in the variance between state processes, as well as changes in the mean.

The most basic form of Hamilton's model, also known as 'Hamilton's filter' (see Hamilton, 1989), comprises an unobserved state variable, denoted z_t, that is postulated to evaluate according to a first order Markov process

$$\text{prob}[z_t = 1|z_{t-1} = 1] = p_{11} \tag{9.11}$$

$$\text{prob}[z_t = 2|z_{t-1} = 1] = 1 - p_{11} \tag{9.12}$$

$$\text{prob}[z_t = 2|z_{t-1} = 2] = p_{22} \tag{9.13}$$

$$\text{prob}[z_t = 1|z_{t-1} = 2] = 1 - p_{22} \tag{9.14}$$

where p_{11} and p_{22} denote the probability of being in regime one, given that the system was in regime one during the previous period, and the probability of being in regime two, given that the system was in regime two during the previous period, respectively. Thus $1 - p_{11}$ defines the probability that y_t will change from state 1 in period $t-1$ to state 2 in period t, and $1 - p_{22}$ defines the probability of a shift from state 2 to state 1 between times $t-1$ and t. It can be shown that under this specification, z_t evolves as an AR(1) process

$$z_t = (1 - p_{11}) + \rho z_{t-1} + \eta_t \tag{9.15}$$

where $\rho = p_{11} + p_{22} - 1$. Loosely speaking, z_t can be viewed as a generalisation of the dummy variables for one-off shifts in a series discussed above. Under the Markov switching approach, there can be multiple shifts from one set of behaviour to another.

In this framework, the observed returns series evolves as given by (9.15)

$$y_t = \mu_1 + \mu_2 z_t + (\sigma_1^2 + \phi z_t)^{1/2} u_t \tag{9.16}$$

where $u_t \sim \mathrm{N}(0, 1)$. The expected values and variances of the series are μ_1 and σ_1^2, respectively in state 1, and $(\mu_1 + \mu_2)$ and $\sigma_1^2 + \phi$ in respectively, state 2. The variance in state 2 is also defined, $\sigma_2^2 = \sigma_1^2 + \phi$. The unknown parameters of the model $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p_{11}, p_{22})$ are estimated using maximum likelihood. Details are beyond the scope of this book, but are most comprehensively given in Engel and Hamilton (1990).

If a variable follows a Markov process, all that is required to forecast the probability that it will be in a given regime during the next period is the current period's probability and a set of transition probabilities, given for the case of two regimes by (9.11)–(9.14). In the general case where there are m states, the transition probabilities are best expressed in a matrix as

$$P = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1m} \\ P_{21} & P_{22} & \dots & P_{2m} \\ \dots & \dots & \dots & \dots \\ P_{m1} & P_{m2} & \dots & P_{mm} \end{bmatrix} \tag{9.17}$$

where P_{ij} is the probability of moving from regime i to regime j. Since, at any given time, the variable must be in one of the m states, it must be true that

$$\sum_{j=1}^{m} P_{ij} = 1 \forall i \tag{9.18}$$

A vector of current state probabilities is then defined as

$$\pi_t = [\pi_1 \quad \pi_2 \quad \dots \quad \pi_m] \tag{9.19}$$

where π_i is the probability that the variable y is currently in state i. Given π_t and P, the probability that the variable y will be in a given regime next period can be forecast using

$$\pi_{t+1} = \pi_t P \tag{9.20}$$

The probabilities for S steps into the future will be given by

$$\pi_{t+s} = \pi_t P^s \tag{9.21}$$

9.6 A Markov switching model for the real exchange rate

There have been a number of applications of the Markov switching model in finance. Clearly, such an approach is useful when a series is thought to undergo shifts from one type of behaviour to another and back again, but where the 'forcing variable' that causes the regime shifts is unobservable.

One such application is to modelling the real exchange rate. As discussed in chapter 7, purchasing power parity (PPP) theory suggests that the law of one price should always apply in the long run such that the cost of a representative basket of goods and services is the same wherever it is purchased, after converting it into a common currency. Under some assumptions, one implication of PPP is that the real exchange rate (that is, the exchange rate divided by a general price index such as the consumer price index (CPI)) should be stationary. However, a number of studies have failed to reject the unit root null hypothesis in real exchange rates, indicating evidence against the PPP theory.

It is widely known that the power of unit root tests is low in the presence of structural breaks as the ADF test finds it difficult to distinguish between a stationary process subject to structural breaks and a unit root process. In order to investigate this possibility, Bergman and Hansson (2005) estimate a Markov switching model with an AR(1) structure for the real exchange rate, which allows for multiple switches between two regimes. The specification they use is

$$y_t = \mu_{s_t} + \phi y_{t-1} + \epsilon_t \tag{9.22}$$

where y_t is the real exchange rate, $s_t, (t = 1, 2)$ are the two states, and $\epsilon_t \sim N(0, \sigma^2)$.[1] The state variable s_t is assumed to follow a standard 2-regime Markov process as described above.

[1] The authors also estimate models that allow ϕ and σ^2 to vary across the states, but the restriction that the parameters are the same across the two states cannot be rejected and hence the values presented in the study assume that they are constant.

Quarterly observations from 1973Q2 to 1997Q4 (99 data points) are used on the real exchange rate (in units of foreign currency per US dollar) for the UK, France, Germany, Switzerland, Canada and Japan. The model is estimated using the first 72 observations (1973Q2–1990Q4) with the remainder retained for out-of-sample forecast evaluation. The authors use 100 times the log of the real exchange rate, and this is normalised to take a value of one for 1973Q2 for all countries. The Markov switching model estimates obtained using maximum likelihood estimation are presented in table 9.3.

As the table shows, the model is able to separate the real exchange rates into two distinct regimes for each series, with the intercept in regime one (μ_1) being positive for all countries except Japan (resulting from the phenomenal strength of the yen over the sample period), corresponding to a rise in the log of the number of units of the foreign currency per US dollar, i.e. a depreciation of the domestic currency against the dollar. μ_2, the intercept in regime 2, is negative for all countries, corresponding to a domestic currency appreciation against the dollar. The probabilities of remaining within the same regime during the following period (p_{11} and p_{22}) are fairly low for the UK, France, Germany and Switzerland, indicating fairly frequent switches from one regime to another for those countries' currencies.

Interestingly, after allowing for the switching intercepts across the regimes, the AR(1) coefficient, ϕ, in table 9.3 is a considerable distance below unity, indicating that these real exchange rates are stationary. Bergman and Hansson simulate data from the stationary Markov switching AR(1) model with the estimated parameters but they assume that the researcher conducts a standard ADF test on the artificial data. They find that for none of the cases can the unit root null hypothesis be rejected, even though clearly this null is wrong as the simulated data are stationary. It is concluded that a failure to account for time-varying intercepts (i.e. structural breaks) in previous empirical studies on real exchange rates could have been the reason for the finding that the series are unit root processes when the financial theory had suggested that they should be stationary.

Finally, the authors employ their Markov switching AR(1) model for forecasting the remainder of the exchange rates in the sample in comparison with the predictions produced by a random walk and by a Markov switching model with a random walk. They find that for all six series, and for forecast horizons up to 4 steps (quarters) ahead, their Markov switching AR model produces predictions with the lowest mean squared errors; these improvements over the pure random walk are statistically significant.

Table 9.3 Estimates of the Markov switching model for real exchange rates

Parameter	UK	France	Germany	Switzerland	Canada	Japan
μ_1	3.554 (0.550)	6.131 (0.604)	6.569 (0.733)	2.390 (0.726)	1.693 (0.230)	−0.370 (0.681)
μ_2	−5.096 (0.549)	−2.845 (0.409)	−2.676 (0.487)	−6.556 (0.775)	−0.306 (0.249)	−8.932 (1.157)
ϕ	0.928 (0.027)	0.904 (0.020)	0.888 (0.023)	0.958 (0.027)	0.922 (0.021)	0.871 (0.027)
σ^2	10.118 (1.698)	7.706 (1.293)	10.719 (1.799)	13.513 (2.268)	1.644 (0.276)	15.879 (2.665)
p_{11}	0.672	0.679	0.682	0.792	0.952	0.911
p_{22}	0.690	0.833	0.830	0.716	0.944	0.817

Notes: Standard errors in parentheses.
Source: Bergman and Hansson (2005).
Reprinted with the permission of Elsevier Science.

9.7 A Markov switching model for the gilt–equity yield ratio

As discussed below, a Markov switching approach is also useful for modelling the time series behaviour of the gilt–equity yield ratio (GEYR), defined as the ratio of the income yield on long-term government bonds to the dividend yield on equities. It has been suggested that the current value of the GEYR might be a useful tool for investment managers or market analysts in determining whether to invest in equities or whether to invest in gilts. Thus the GEYR is purported to contain information useful for determining the likely direction of future equity market trends. The GEYR is assumed to have a long-run equilibrium level, deviations from which are taken to signal that equity prices are at an unsustainable level. If the GEYR becomes high relative to its long-run level, equities are viewed as being expensive relative to bonds. The expectation, then, is that for given levels of bond yields, equity yields must rise, which will occur via a fall in equity prices. Similarly, if the GEYR is well below its long-run level, bonds are considered expensive relative to stocks, and by the same analysis, the price of the latter is expected to increase. Thus, in its crudest form, an equity trading rule based on the GEYR would say, ‘if the GEYR is low, buy equities; if the GEYR is high, sell equities’. The paper by Brooks and Persand (2001b) discusses the usefulness of the Markov switching approach in this context, and considers whether profitable trading rules can be developed on the basis of forecasts derived from the model.

Brooks and Persand (2001b) employ monthly stock index dividend yields and income yields on government bonds covering the period January 1975 until August 1997 (272 observations) for three countries – the UK, the US and Germany. The series used are the dividend yield and index values of the FTSE100 (UK), the S&P500 (US) and the DAX (Germany). The bond indices and redemption yields are based on the clean prices of UK government consols, and US and German 10-year government bonds.

As an example, figure 9.5 presents a plot of the distribution of the GEYR for the US (in bold), together with a normal distribution having the same mean and variance. Clearly, the distribution of the GEYR series is not normal, and the shape suggests two separate modes: one upper part of the distribution embodying most of the observations, and a lower part covering the smallest values of the GEYR.

Such an observation, together with the notion that a trading rule should be developed on the basis of whether the GEYR is ‘high’ or ‘low’, and in the absence of a formal econometric model for the GEYR, suggests that a Markov switching approach may be useful. Under the Markov switching approach, the values of the GEYR are drawn from a mixture of normal

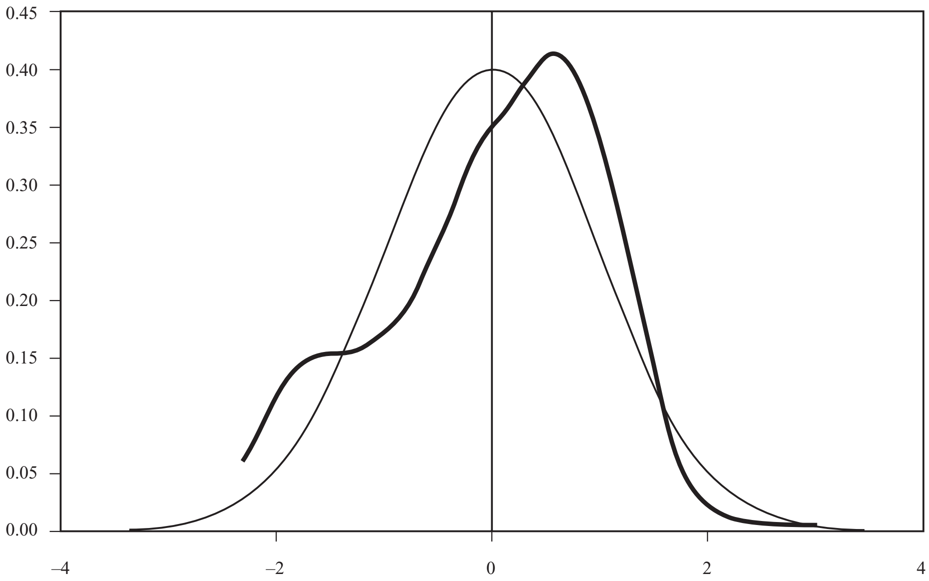


Figure 9.5 Source: Brooks and Persand (2001b). Unconditional distribution of US GEYR together with a normal distribution with the same mean and variance

Table 9.4 Estimated parameters for the Markov switching models

Statistic	μ_1 (1)	μ_2 (2)	σ_1^2 (3)	σ_2^2 (4)	p_{11} (5)	p_{22} (6)	N_1 (7)	N_2 (8)
UK	2.4293 (0.0301)	2.0749 (0.0367)	0.0624 (0.0092)	0.0142 (0.0018)	0.9547 (0.0726)	0.9719 (0.0134)	102	170
US	2.4554 (0.0181)	2.1218 (0.0623)	0.0294 (0.0604)	0.0395 (0.0044)	0.9717 (0.0171)	0.9823 (0.0106)	100	172
Germany	3.0250 (0.0544)	2.1563 (0.0154)	0.5510 (0.0569)	0.0125 (0.0020)	0.9816 (0.0107)	0.9328 (0.0323)	200	72

Notes: Standard errors in parentheses; N_1 and N_2 denote the number of observations deemed to be in regimes 1 and 2, respectively.
Source: Brooks and Persand (2001b).

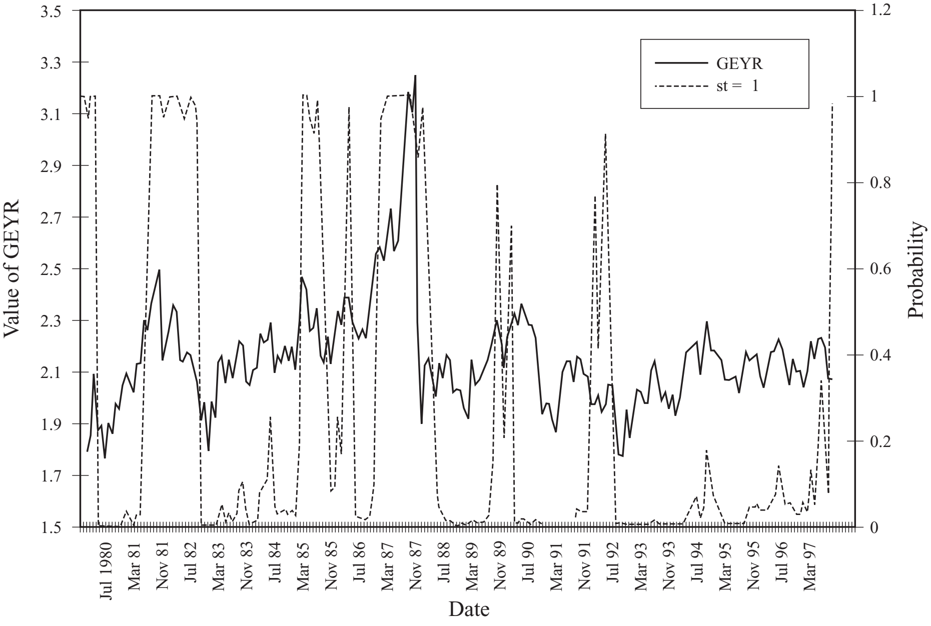
distributions, where the weights attached to each distribution sum to one and where movements between series are governed by a Markov process. The Markov switching model is estimated using a maximum likelihood procedure (as discussed in chapter 8), based on GAUSS code supplied by James Hamilton. Coefficient estimates for the model are presented in table 9.4.

The means and variances for the values of the GEYR for each of the two regimes are given in columns headed (1)–(4) of table 9.4 with standard errors associated with each parameter in parentheses. It is clear that the regime switching model has split the data into two distinct samples – one with a high mean (of 2.43, 2.46 and 3.03 for the UK, US and Germany, respectively) and one with a lower mean (of 2.07, 2.12, and 2.16), as was anticipated from the unconditional distribution of returns. Also apparent is the fact that the UK and German GEYR are more variable at times when it is in the high mean regime, evidenced by their higher variance (in fact, it is around four and 20 times higher than for the low GEYR state, respectively). The number of observations for which the probability that the GEYR is in the high mean state exceeds 0.5 (and thus when the GEYR is actually deemed to be in this state) is 102 for the UK (37.5% of the total), while the figures for the US are 100 (36.8%) and for Germany 200 (73.5%). Thus, overall, the GEYR is more likely to be in the low mean regime for the UK and US, while it is likely to be high in Germany.

The columns marked (5) and (6) of table 9.4 give the values of p_{11} and p_{22}, respectively, that is the probability of staying in state 1 given that the GEYR was in state 1 in the immediately preceding month, and the probability of staying in state 2 given that the GEYR was in state 2 previously, respectively. The high values of these parameters indicates that the regimes are highly stable with less than a 10% chance of moving from a low GEYR to a high GEYR regime and vice versa for all three series. Figure 9.6 presents a 'q-plot', which shows the value of GEYR and probability that it is in the high GEYR regime for the UK at each point in time.

Figure 9.6

Source: Brooks and Persand (2001b). Value of GEYR and probability that it is in the High GEYR regime for the UK



As can be seen, the probability that the UK GEYR is in the 'high' regime (the dotted line) varies frequently, but spends most of its time either close to zero or close to one. The model also seems to do a reasonably good job of specifying which regime the UK GEYR should be in, given that the probability seems to match the broad trends in the actual GEYR (the full line).

Engel and Hamilton (1990) show that it is possible to give a forecast of the probability that a series y_t, which follows a Markov switching process, will be in a particular regime. Brooks and Persand (2001b) use the first 60 observations (January 1975–December 1979) for in-sample estimation of the model parameters (μ_1, μ_2, σ_1^2, σ_2^2, p_{11}, p_{22}). Then a one step-ahead forecast is produced of the probability that the GEYR will be in the high mean regime during the next period. If the probability that the GEYR will be in the low regime during the next period is forecast to be more that 0.5, it is forecast that the GEYR will be low and hence equities are bought or held. If the probability that the GEYR is in the low regime is forecast to be less than 0.5, it is anticipated that the GEYR will be high and hence gilts are invested in or held. The model is then rolled forward one observation, with a new set of model parameters and probability forecasts being constructed. This process continues until 212 such probabilities are estimated with corresponding trading rules.

The returns for each out-of-sample month for the switching portfolio are calculated, and their characteristics compared with those of buy-and-hold equities and buy-and-hold gilts strategies. Returns are calculated as continuously compounded percentage returns on a stock (the FTSE in the UK, the S&P500 in the US, the DAX in Germany) or on a long-term government bond. The profitability of the trading rules generated by the forecasts of the Markov switching model are found to be superior in gross terms compared with a simple buy-and-hold equities strategy. In the UK context, the former yields higher average returns and lower standard deviations. The switching portfolio generates an average return of 0.69% per month, compared with 0.43% for the pure bond and 0.62% for the pure equity portfolios. The improvements are not so clear-cut for the US and Germany. The Sharpe ratio for the UK Markov switching portfolio is almost twice that of the buy-and-hold equities portfolio, suggesting that, after allowing for risk, the switching model provides a superior trading rule. The improvement in the Sharpe ratio for the other two countries is, on the contrary, only very modest.

To summarise:

- The Markov switching approach can be used to model the gilt-equity yield ratio

- The resulting model can be used to produce forecasts of the probability that the GEYR will be in a particular regime
- Before transactions costs, a trading rule derived from the model produces a better performance than a buy-and-hold equities strategy, in spite of inferior predictive accuracy as measured statistically
- Net of transactions costs, rules based on the Markov switching model are not able to beat a passive investment in the index for any of the three countries studied.

9.8 Threshold autoregressive models

Threshold autoregressive (TAR) models are one class of non-linear autoregressive models. Such models are a relatively simple relaxation of standard linear autoregressive models that allow for a locally linear approximation over a number of states. According to Tong (1990, p. 99), the threshold principle 'allows the analysis of a complex stochastic system by decomposing it into a set of smaller sub-systems'. The key difference between TAR and Markov switching models is that, under the former, the state variable is assumed known and observable, while it is latent under the latter. A very simple example of a threshold autoregressive model is given by (9.23). The model contains a first order autoregressive process in each of two regimes, and there is only one threshold. Of course, the number of thresholds will always be the number of regimes minus one. Thus, the dependent variable y_t is purported to follow an autoregressive process with intercept coefficient μ_1 and autoregressive coefficient ϕ_1 if the value of the state-determining variable lagged k periods, denoted s_{t-k} is lower than some threshold value r. If the value of the state-determining variable lagged k periods, is equal to or greater than that threshold value r, y_t is specified to follow a different autoregressive process, with intercept coefficient μ_2 and autoregressive coefficient ϕ_2. The model would be written

$$y_t = \begin{cases} \mu_1 + \phi_1 y_{t-1} + u_{1t} & \textit{if } s_{t-k} < r \\ \mu_2 + \phi_2 y_{t-1} + u_{2t} & \textit{if } s_{t-k} \geq r \end{cases} \tag{9.23}$$

But what is s_{t-k}, the state-determining variable? It can be any variable that is thought to make y_t shift from one set of behaviour to another. Obviously, financial or economic theory should have an important role to play in making this decision. If $k = 0$, it is the current value of the state-determining variable that influences the regime that y is in at time t, but in many applications k is set to 1, so that the immediately preceding value of s is the one that determines the current value of y.

The simplest case for the state determining variable is where it is the variable under study, i.e. $s_{t-k} = y_{t-k}$. This situation is known as a self-exciting TAR, or a SETAR, since it is the lag of the variable y itself that determines the regime that y is currently in. The model would now be written

$$y_t = \begin{cases} \mu_1 + \phi_1 y_{t-1} + u_{1t} \ \mathit{if} \ y_{t-k} < r \\ \mu_2 + \phi_2 y_{t-1} + u_{2t} \ \mathit{if} \ y_{t-k} \geq r \end{cases} \tag{9.24}$$

The models of (9.23) or (9.24) can of course be extended in several directions. The number of lags of the dependent variable used in each regime may be higher than one, and the number of lags need not be the same for both regimes. The number of states can also be increased to more than two. A general threshold autoregressive model, that notationally permits the existence of more than two regimes and more than one lag, may be written

$$x_t = \sum_{j=1}^{J} I_t^{(j)} \left(\phi_0^{(j)} + \sum_{i=1}^{p_j} \phi_i^{(j)} x_{t-i} + u_t^{(j)} \right), \ r_{j-1} \leq z_{t-d} \leq r_j \tag{9.25}$$

where $I_t^{(j)}$ is an indicator function for the jth regime taking the value one if the underlying variable is in state j and zero otherwise. z_{t-d} is an observed variable determining the switching point and $u_t^{(j)}$ is a zero-mean independently and identically distributed error process. Again, if the regime changes are driven by own lags of the underlying variable, x_t (i.e. $z_{t-d} = x_{t-d}$), then the model is a self-exciting TAR (SETAR).

It is also worth re-stating that under the TAR approach, the variable y is either in one regime or another, given the relevant value of s, and there are discrete transitions between one regime and another. This is in contrast with the Markov switching approach, where the variable y is in both states with some probability at each point in time. Another class of threshold autoregressive models, known as smooth transition autoregressions (STAR), allows for a more gradual transition between the regimes by using a continuous function for the regime indicator rather than an on–off switch (see Franses and van Dijk, 2000, chapter 3).

9.9 Estimation of threshold autoregressive models

Estimation of the model parameters (ϕ_i, r_j, d, p_j) is considerably more difficult than for a standard linear autoregressive process, since in general they cannot be determined simultaneously in a simple way, and the values chosen for one parameter are likely to influence estimates of the others. Tong (1983, 1990) suggests a complex non-parametric lag regression procedure to estimate the values of the thresholds (r_j) and the delay parameter (d).

Ideally, it may be preferable to endogenously estimate the values of the threshold(s) as part of the non-linear least squares (NLS) optimisation procedure, but this is not feasible. The underlying functional relationship between the variables is discontinuous in the thresholds, such that the thresholds cannot be estimated at the same time as the other components of the model. One solution to this problem that is sometimes used in empirical work is to use a grid search procedure that seeks the minimal residual sum of squares over a range of values of the threshold(s) for an assumed model. Some sample code to achieve this is presented later in this chapter.

9.9.1 Threshold model order (lag length) determination

A simple, although far from ideal, method for determining the appropriate lag lengths for the autoregressive components for each of the regimes would be to assume that the same number of lags are required in all regimes. The lag length is then chosen in the standard fashion by determining the appropriate lag length for a linear autoregressive model, and assuming that the lag length for all states of the TAR is the same. While it is easy to implement, this approach is clearly not a good one, for it is unlikely that the lag lengths for each state when the data are drawn from different regimes would be the same as that appropriate when a linear functional form is imposed. Moreover, it is undesirable to require the lag lengths to be the same in each regime. This conflicts with the notion that the data behave differently in different states, which was precisely the motivation for considering threshold models in the first place.

An alternative and better approach, conditional upon specified threshold values, would be to employ an information criterion to select across the lag lengths in each regime simultaneously. A drawback of this approach, that Franses and van Dijk (2000) highlight, is that in practice it is often the case that the system will be resident in one regime for a considerably longer time overall than the others. In such situations, information criteria will not perform well in model selection for the regime(s) containing few observations. Since the number of observations is small in these cases, the overall reduction in the residual sum of squares as more parameters are added to these regimes will be very small. This leads the criteria to always select very small model orders for states containing few observations. A solution, therefore, is to define an information criterion that does not penalise the whole model for additional parameters in one state. Tong (1990) proposes a modified version of Akaike's information criterion (*AIC*) that weights $\hat{\sigma}^2$ for each regime by the number of observations in that regime. For the two-regime case, the modified *AIC* would be written

$$AIC\,(p_1, p_2) = T_1 \ln \hat{\sigma}_1^2 + T_2 \ln \hat{\sigma}_2^2 + 2(p_1 + 1) + 2(p_2 + 1) \tag{9.26}$$

where T_1 and T_2 are the number of observations in regimes 1 and 2, respectively, p_1 and p_2 are the lag lengths and $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the residual variances. Similar modifications can of course be developed for other information criteria.

9.9.2 Determining the delay parameter, d

The delay parameter, d, can be decided in a variety of ways. It can be determined along with the lag orders for each of the regimes by an information criterion, although of course this added dimension greatly increases the number of candidate models to be estimated. In many applications, however, it is typically set to one on theoretical grounds. It has been argued (see, for example, Kräger and Kugler, 1993) that in the context of financial markets, it is most likely that the most recent past value of the state-determining variable would be the one to determine the current state, rather than that value two, three, . . . periods ago.

Estimation of the autoregressive coefficients can then be achieved using NLS. Further details of the procedure are discussed in Franses and van Dijk (2000, chapter 3).

9.10 Specification tests in the context of Markov switching and threshold autoregressive models: a cautionary note

In the context of both Markov switching and TAR models, it is of interest to determine whether the threshold models represent a superior fit to the data relative to a comparable linear model. A tempting, but incorrect, way to examine this issue would be to do something like the following: estimate the desired threshold model and the linear counterpart, and compare the residual sums of squares using an *F*-test. However, such an approach is not valid in this instance owing to unidentified nuisance parameters under the null hypothesis. In other words, the null hypothesis for the test would be that the additional parameters in the regime switching model were zero so that the model collapsed to the linear specification, but under the linear model, there is no threshold. The upshot is that the conditions required to show that the test statistics follow a standard asymptotic distribution do not apply. Hence analytically derived critical values are not available, and critical values must be obtained via simulation for each individual case. Hamilton (1994) provides substitute hypotheses for Markov switching model evaluation that can validly be tested using the standard hypothesis testing framework, while Hansen (1996) offers solutions in the context of TAR models.

This chapter will now examine two applications of TAR modelling in finance: one to the modelling of exchange rates within a managed floating environment, and one to arbitrage opportunities implied by the difference between spot and futures prices for a given asset. For a (rather technical) general survey of several TAR applications in finance, see Yadav, Pope and Paudyal (1994).

9.11 A SETAR model for the French franc–German mark exchange rate

During the 1990s, European countries which were part of the Exchange Rate Mechanism (ERM) of the European Monetary System (EMS), were required to constrain their currencies to remain within prescribed bands relative to other ERM currencies. This seemed to present no problem by early in the new millenium since European Monetary Union (EMU) was already imminent and conversion rates of domestic currencies into Euros were already known. However, in the early 1990s, the requirement that currencies remain within a certain band around their central parity forced central banks to intervene in the markets to effect either an appreciation or a depreciation in their currency. A study by Chappell *et al*. (1996) considered the effect that such interventions might have on the dynamics and time series properties of the French franc–German mark (hereafter FRF–DEM) exchange rate. ‘Core currency pairs’, such as the FRF–DEM were allowed to move up to ±2.25% either side of their central parity within the ERM. The study used daily data from 1 May 1990 until 30 March 1992. The first 450 observations are used for model estimation, with the remaining 50 being retained for out-of-sample forecasting.

A self-exciting threshold autoregressive (SETAR) model was employed to allow for different types of behaviour according to whether the exchange rate is close to the ERM boundary. The argument is that, close to the boundary, the respective central banks will be required to intervene in opposite directions in order to drive the exchange rate back towards its central parity. Such intervention may be expected to affect the usual market dynamics that ensure fast reaction to news and the absence of arbitrage opportunities.

Table 9.5 SETAR model for FRF–DEM

Model	For regime	Number of observations
$\hat{E}_t = 0.0222 + 0.9962 E_{t-1}$ $(0.0458) \quad (0.0079)$	$E_{t-1} < 5.8306$	344
$\hat{E}_t = 0.3486 + 0.4394 E_{t-1} + 0.3057 E_{t-2} + 0.1951 E_{t-3}$ $(0.2391) \quad (0.0889) \quad (0.1098) \quad (0.0866)$	$E_{t-1} \geq 5.8306$	103

Source: Chappell *et al.* (1996). Reprinted with permission of John Wiley and Sons.

Let E_t denote the log of the FRF–DEM exchange rate at time t. Chappell *et al.* (1996) estimate two models: one with two thresholds and one with one threshold. The former was anticipated to be most appropriate for the data at hand since exchange rate behaviour is likely to be affected by intervention if the exchange rate comes close to either the ceiling or the floor of the band. However, over the sample period employed, the mark was never a weak currency, and therefore the FRF–DEM exchange rate was either at the top of the band or in the centre, never close to the bottom. Therefore, a model with one threshold is more appropriate since any second estimated threshold was deemed likely to be spurious.

The authors show, using DF and ADF tests, that the exchange rate series is not stationary. Therefore, a threshold model in the levels is not strictly valid for analysis. However, they argue that an econometrically valid model in first difference would lose its intuitive interpretation, since it is the *value* of the exchange rate that is targeted by the monetary authorities, not its change. In addition, if the currency bands are working effectively, the exchange rate is constrained to lie within them, and hence in some senses of the word, it must be stationary, since it cannot wander without bound in either direction. The model orders for each regime are determined using *AIC*, and the estimated model is given in table 9.5.

As can be seen, the two regimes comprise a random walk with drift under normal market conditions, where the exchange rate lies below a certain threshold, and an AR(3) model corresponding to much slower market adjustment when the exchange rate lies on or above the threshold. The (natural log of) the exchange rate's central parity over the period was 5.8153, while the (log of the) ceiling of the band was 5.8376. The estimated threshold of 5.8306 is approximately 1.55% above the central parity, while the ceiling is 2.25% above the central parity. Thus, the estimated threshold is some way below the ceiling, which is in accordance with the authors'

Table 9.6 FRF–DEM forecast accuracies

	Steps ahead				
	1	*2*	*3*	*5*	*10*
Panel A: mean squared forecast error					
Random walk	1.84E-07	3.49E-07	4.33E-07	8.03E-07	1.83E-06
AR(2)	3.96E-07	1.19E-06	2.33E-06	6.15E-06	2.19E-05
One-threshold SETAR	1.80E-07	2.96E-07	3.63E-07	5.41E-07	5.34E-07
Two-threshold SETAR	1.80E-07	2.96E-07	3.63E-07	5.74E-07	5.61E-07
Panel B: Median squared forecast error					
Random walk	7.80E-08	1.04E-07	2.21E-07	2.49E-07	1.00E-06
AR(2)	2.29E-07	9.00E-07	1.77E-06	5.34E-06	1.37E-05
One-threshold SETAR	9.33E-08	1.22E-07	1.57E-07	2.42E-07	2.34E-07
Two-threshold SETAR	1.02E-07	1.22E-07	1.87E-07	2.57E-07	2.45E-07

Source: Chappell *et al.* (1996). Reprinted with permission of John Wiley and Sons.

expectations since the central banks are likely to intervene before the exchange rate actually hits the ceiling.

Forecasts are then produced for the last 50 observations using the threshold model estimated above, the SETAR model with two thresholds, a random walk and an AR(2) (where the model order was chosen by in-sample minimisation of *AIC*). The results are presented here in table 9.6.

For the FRF–DEM exchange rate, the one-threshold SETAR model is found to give lower mean squared errors than the other three models for one-, two-, three-, five- and ten-step-ahead forecasting horizons. Under the median squared forecast error measure, the random walk is marginally superior to the one threshold SETAR one and two steps ahead, while it has regained its prominence by three steps ahead.

However, in a footnote, the authors also argue that the SETAR model was estimated and tested for 9 other ERM exchange rate series, but in every one of these other cases, the SETAR models produced less accurate forecasts than a random walk model. A possible explanation for this phenomenon is given in section 9.13.

Brooks (2001) extends the work of Chappell *et al*. to allow the conditional variance of the exchange rate series to be drawn from a GARCH process which itself contains a threshold, above which the behaviour of volatility is different to that below. He finds that the dynamics of the conditional variance are quite different from one regime to the next, and that models allowing for different regimes can provide superior volatility forecasts compared to those which do not.

9.12 Threshold models and the dynamics of the FTSE 100 index and index futures markets

One of the examples given in chapter 7 discussed the implications for the effective functioning of spot and futures markets of a lead–lag relationship between the two series. If the two markets are functioning effectively, it was also shown that a cointegrating relationship between them would be expected.

If stock and stock index futures markets are functioning properly, price movements in these markets should be best described by a first order vector error correction model (VECM) with the error correction term being the price differential between the two markets (the basis). The VECM could be expressed as

$$\begin{bmatrix} \Delta f_t \\ \Delta s_t \end{bmatrix} = \begin{bmatrix} \pi_{11} \\ \pi_{21} \end{bmatrix} [\, f_{t-1} - s_{t-1}\,] + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{9.27}$$

where Δf_t and Δs_t are changes in the log of the futures and spot prices, respectively, π_{11} and π_{21} are coefficients describing how changes in the spot and futures prices occur as a result of the basis. Writing these two equations out in full, the following would result

$$f_t - f_{t-1} = \pi_{11}[\, f_{t-1} - s_{t-1}] + u_{1t} \tag{9.28}$$

$$s_t - s_{t-1} = \pi_{21}[\, f_{t-1} - s_{t-1}] + u_{2t} \tag{9.29}$$

Subtracting (9.29) from (9.28) would give the following expression

$$(\, f_t - f_{t-1}) - (s_t - s_{t-1}) = (\pi_{11} - \pi_{21})[\, f_{t-1} - s_{t-1}] + (u_{1t} - u_{2t}) \tag{9.30}$$

which can also be written as

$$(\, f_t - s_t) - (\, f_{t-1} - s_{t-1}) = (\pi_{11} - \pi_{21})[\, f_{t-1} - s_{t-1}] + (u_{1t} - u_{2t}) \tag{9.31}$$

or, using the result that $b_t = f_t - s_t$

$$b_t - b_{t-1} = (\pi_{11} - \pi_{21})b_{t-1} + \varepsilon_t \tag{9.32}$$

where $\varepsilon_t = u_{1t} - u_{2t}$. Taking b_{t-1} from both sides

$$b_t = (\pi_{11} - \pi_{21} - 1)b_{t-1} + \varepsilon_t \tag{9.33}$$

If the first order VECM is appropriate, then it is not possible to identify structural equations for returns in stock and stock index futures markets with the obvious implications for predictability and the two markets are indeed efficient. Hence, for efficient markets and no arbitrage, there should be only a first order autoregressive process describing the basis and no further patterns. Recent evidence suggests, however, that there are more dynamics present than should be in effectively functioning markets. In particular, it has been suggested that the basis up to three trading days prior carries predictive power for movements in the FTSE 100 cash index, suggesting the possible existence of unexploited arbitrage opportunities. The paper by Brooks and Garrett (2002) analyses whether such dynamics can be explained as the result of different regimes within which arbitrage is not triggered and outside of which arbitrage will occur. The rationale for the existence of different regimes in this context is that the basis (adjusted for carrying costs if necessary), which is very important in the arbitrage process, can fluctuate within bounds determined by transaction costs without actually triggering arbitrage. Hence an autoregressive relationship between the current and previous values of the basis could arise and persist over time within the threshold boundaries since it is not profitable for traders to exploit this apparent arbitrage opportunity. Hence there will be thresholds within which there will be no arbitrage activity but once these thresholds are crossed, arbitrage should drive the basis back within the transaction cost bounds. If markets are functioning effectively then irrespective of the dynamics of the basis within the thresholds, once the thresholds have been crossed the additional dynamics should disappear.

The data used by Brooks and Garrett (2002) are the daily closing prices for the FTSE 100 stock index and stock index futures contract for the period January 1985–October 1992. The October 1987 stock market crash occurs right in the middle of this period, and therefore Brooks and Garrett conduct their analysis on a ‘pre-crash’ and a ‘post-crash’ sample as well as the whole sample. This is necessary since it has been observed that the normal spot/futures price relationship broke down around the time of the crash (see Antoniou and Garrett, 1993). Table 9.7 shows the coefficient estimates for a linear AR(3) model for the basis.

The results for the whole sample suggest that all of the first three lags of the basis are significant in modelling the current basis. This result is confirmed (although less strongly) for the pre-crash and post-crash subsamples. Hence, a linear specification would seem to suggest that the basis is to some degree predictable, indicating possible arbitrage opportunities.

In the absence of transactions costs, deviations of the basis away from zero in either direction will trigger arbitrage. The existence of transactions costs, however, means that the basis can deviate from zero without actually triggering arbitrage. Thus, assuming that there are no differential transactions costs, there will be upper and lower bounds within which the basis can fluctuate without triggering arbitrage. Brooks and Garrett

Table 9.7 Linear AR(3) model for the basis

$b_t = \phi_0 + \phi_1 b_{t-1} + \phi_2 b_{t-2} + \phi_3 b_{t-3} + \varepsilon_t$			
Parameter	Whole sample	Pre-crash sample	Post-crash sample
ϕ_1	0.7051**	0.7174**	0.6791**
	(0.0225)	(0.0377)	(0.0315)
ϕ_2	0.1268**	0.0946*	0.1650**
	(0.0274)	(0.0463)	(0.0378)
ϕ_3	0.0872**	0.1106**	0.0421
	(0.0225)	(0.0377)	(0.0315)

Notes: Figures in parentheses are heteroscedasticity-robust standard errors; * and ** denote significance at the 5% and 1% levels, respectively.
Source: Brooks and Garrett (2002).

(2002) estimate a SETAR model for the basis, with two thresholds (three regimes) since these should correspond to the upper and lower boundaries within which the basis can fluctuate without causing arbitrage. Under efficient markets, profitable arbitrage opportunities will not be present when $r_0 \leq b_{t-1} < r_1$ where r_0 and r_1 are the thresholds determining which regime the basis is in. If these thresholds are interpreted as transactions costs bounds, when the basis falls below the lower threshold (r_0), the appropriate arbitrage transaction is to buy futures and short stock. This applies in reverse when the basis rises above r_1. When the basis lies within the thresholds, there should be no arbitrage transactions. Three lags of the basis enter into each equation and the thresholds are estimated using a grid search procedure. The one-period lag of the basis is chosen as the state-determining variable. The estimated model for each sample period is given in table 9.8.

The results show that, to some extent, the dependence in the basis is reduced when it is permitted to be drawn from one of three regimes rather than a single linear model. For the post-crash sample, and to some extent for the whole sample and the pre-crash sample, it can be seen that there is considerably slower adjustment, evidenced by the significant second and third order autoregressive terms, between the thresholds than outside them. There still seems to be some evidence of slow adjustment below the lower threshold, where the appropriate trading strategy would be to go long the futures and short the stock. Brooks and Garrett (2002) attribute this in part to restrictions on and costs of short-selling the stock that prevent adjustment from taking place more quickly. Short-selling of futures contracts is easier and less costly, and hence there is no action in the basis beyond an AR(1) when it is above the upper threshold.

Table 9.8 A two-threshold SETAR model for the basis

$$b_t = \begin{cases} {\phi_0}^1 + \sum_{i=1}^{3} \phi_i^1 b_{t-i} + \varepsilon_t^1 \ \mathit{if} \ b_{t-1} < r_0 \\ {\phi_0}^2 + \sum_{i=1}^{3} \phi_i^2 b_{t-i} + \varepsilon_t^2 \ \mathit{if} \ r_0 \le b_{t-1} < r_1 \\ {\phi_0}^3 + \sum_{i=1}^{3} \phi_i^3 b_{t-i} + \varepsilon_t^3 \ \mathit{if} \ b_{t-1} \ge r_1 \end{cases}$$

	$b_{t-1} < r_0$	$r_0 \le b_{t-1} < r_1$	$b_{t-1} \ge r_1$
Panel A: whole sample			
ϕ_1	0.5743**	−0.6395	0.8380**
	(0.0415)	(0.7549)	(0.0512)
ϕ_2	0.2088**	−0.0594	0.0439
	(0.0401)	(0.0846)	(0.0462)
ϕ_3	0.1330**	0.2267**	0.0415
	(0.0355)	(0.0811)	(0.0344)
$\hat{r}_0$		0.0138	
$\hat{r}_1$		0.0158	
Panel B: pre-crash sample			
ϕ_1	0.4745**	0.4482*	0.8536**
	(0.0808)	(0.1821)	(0.0720)
ϕ_2	0.2164**	0.2608**	−0.0388
	(0.0781)	(0.0950)	(0.0710)
ϕ_3	0.1142	0.2309**	0.0770
	(0.0706)	(0.0834)	(0.0531)
$\hat{r}_0$		0.0052	
$\hat{r}_1$		0.0117	
Panel C: post-crash sample			
ϕ_1	0.5019**	0.7474**	0.8397**
	(0.1230)	(0.1201)	(0.0533)
ϕ_2	0.2011*	0.2984**	0.0689
	(0.0874)	(0.0691)	(0.0514)
ϕ_3	0.0434	0.1412	0.0461
	(0.0748)	(0.0763)	(0.0400)
$\hat{r}_0$		0.0080	
$\hat{r}_1$		0.0140	

Notes: Figures in parentheses are heteroscedasticity-robust standard errors, * and ** denote significance at the 5% and at 1% levels, respectively.
Source: Brooks and Garrett (2002).

Such a finding is entirely in accordance with expectations, and suggests that, once allowance is made for reasonable transactions costs, the basis may fluctuate with some degree of flexibility where arbitrage is not profitable. Once the basis moves outside the transactions costs-determined range, adjustment occurs within one period as the theory predicted.

9.13 A note on regime switching models and forecasting accuracy

Several studies have noted the inability of threshold or regime switching models to generate superior out-of-sample forecasting accuracy than linear models or a random walk in spite of their apparent ability to fit the data better in sample. A possible reconciliation is offered by Dacco and Satchell (1999), who suggest that regime switching models may forecast poorly owing to the difficulty of forecasting the regime that the series will be in. Thus, any gain from a good fit of the model within the regime will be lost if the model forecasts the regime wrongly. Such an argument could apply to both the Markov switching and TAR classes of models.

Key concepts

The key terms to be able to define and explain from this chapter are

- seasonality
- slope dummy variable
- regime switching
- self-exciting TAR
- Markov process
- intercept dummy variable
- dummy variable trap
- threshold autoregression (TAR)
- delay parameter
- transition probability

Review questions

1. A researcher is attempting to form an econometric model to explain daily movements of stock returns. A colleague suggests that she might want to see whether her data are influenced by daily seasonality.
   (a) How might she go about doing this?
   (b) The researcher estimates a model with the dependent variable as the daily returns on a given share traded on the London stock exchange, and various macroeconomic variables and accounting ratios as independent variables. She attempts to estimate this model, together with five daily dummy variables (one for each day of the week), and a constant term, using EViews. EViews then tells her that it cannot estimate the parameters of the model. Explain what has probably happened, and how she can fix it.