2 Maximum likelihood estimation

In many ways the maximum likelihood (ML) approach forms the cornerstone of classical statistical methodology. The conceptual approach underlying maximum likelihood procedures is an appealing one which is much less '*ad hoc*' than other estimation procedures. To emphasise this point Hendry (1976) shows that many of the conventional estimation techniques, such as three-stage least squares, two-stage least squares, etc. can be interpreted as approximations to the maximum likelihood estimator. Generally speaking, an appropriate maximum likelihood estimation technique is both consistent and asymptotically efficient, so the ML approach forms a useful point of comparison for judging other estimators. We present the general issues behind maximum likelihood estimation and the associated test procedures (LR, Wald and LM) in section 2.1. Next we discuss numerical optimisation procedures and in section 2.3 we outline two special forms of the likelihood function frequently encountered in the empirical literature: the discrete switching model and various forms of the ARCH model. In section 2.4 we present two empirical examples, a model of the mortgage market and a model of time-varying risk premia in the foreign exchange markets.

2.1 The conceptual approach

ML is a very general procedure with the following common features. First we assume a particular probability distribution and calculate the probability of observing a particular outcome. This generally depends on some unknown parameters. Given our data set we then choose

those parameter estimates which maximise the probability of the observed outcome. These parameter estimates are then the maximum likelihood estimate of the unknown true parameter values.

An example may help to clarify this. Suppose we wish to test a consignment of goods for quality, we might take a *sample* of ten items and test these and find that five fail the quality check. What then is our estimate of the proportion of total goods which are faulty? The intuitive answer is, of course 0.5, but the ML procedure would approach the question rather differently. Consider first the probability distribution for the problem at hand. Suppose we draw a random sample of size n and the (unknown) probability of each item being defective is Π in the population. If we actually find B bad items, then the probability P of finding B bad items in our sample of n is given by the binomial formula (i.e. our probability distribution)

$$P = \frac{n!}{B!(n-B)!}\Pi^B(1-\Pi)^{n-B} \tag{2.1}$$

In the example above $n = 10$, $B = 5$. Given fixed n and B, from our *sample*, if we arbitrarily set $\Pi = 0.1$ then (2.1) yields $P = 0.0015$, if $\Pi = 0.2$ then $P = 0.0254$, etc. So in principle we could search over the whole range of Π and we would discover that P is maximised when we chose $\Pi = 0.5$ (which gives $P_{max} = 0.264$). The value of Π which maximises the probability of getting the *observed* sample outcome (i.e. $B = 5$ for $n = 10$) is therefore $\Pi = 0.5$. This is the ML estimate of the *true* population value of Π. We could of course maximise (2.1) analytically by setting its first derivative equal to zero.

$$\frac{\partial P}{\partial \bar{\Pi}} = B\left[\frac{n!}{B!(n-B)!}\bar{\Pi}^{B-1}(1-\bar{\Pi})^{n-B}\right]$$

$$-(n-B)\frac{n!}{B!(n-B)!}\bar{\Pi}^B(1-\bar{\Pi})^{n-B-1} = 0$$

$$B\bar{\Pi}^{B-1}(1-\bar{\Pi})^{n-B} = (n-B)\bar{\Pi}^B(1-\bar{\Pi})^{n-B-1}$$

$$B\bar{\Pi}^{-1} = (n-B)(1-\bar{\Pi})^{-1}$$

$$\bar{\Pi} = \frac{B}{n} \tag{2.2}$$

$\bar{\Pi}$ is the ML estimator and in our case $\bar{\Pi} = B/n = 5/10 = 1/2$.

There are many cases where we can define the probability density function but where the problem is too complex for analytical maximisation. In these cases some numerical technique for locating the maximum must be used but even in the most complex cases the conceptual approach remains that discussed above.

A general statement

Suppose we have a sample of $(X_1, X_2 \ldots X_n)$ which is drawn from a probability distribution $P(X|A)$ where A is a set of parameters which, together with the assumed structural form, define the density function of X. We further assume that the X_i are independent, each with probability distribution $P(X_i|A)$ and so the joint probability distribution of the whole set $X_1 \ldots X_n$ is given by:

$$P(X_1, X_2 \ldots X_n|A) = P(X_1|A).P(X_2|A) \ldots P(X_n|A) = \prod_{i=1}^{n} P(X_i|A) \tag{2.3}$$

We assume that the X_i are sample values, and therefore fixed. If we now ask what value of A maximises the probability of observing the sample values X_i we may restate (2.3) as the likelihood function

$$L(A) = \prod_{i=1}^{n} P(X_i|A) \tag{2.4}$$

It is often convenient to work in terms of the log of the likelihood function, which is simply

$$\log[L(A)] = \sum_{i=1}^{n} \log[P(X_i|A)] \tag{2.5}$$

The advantage of the ML approach is that it is a very general specification which can be applied to a wide range of models. It generally gives consistent parameter estimates which are asymptotically efficient. The main disadvantages are essentially practical. ML often produces highly complex non-linear optimisation problems and it also assumes an exact knowledge of the form of the probability distribution involved (up to a set of unknown parameters). This means that ML may be particularly sensitive to any structural misspecification in the model.

The likelihood function for a general non-linear model

If we write a non-linear model with N endogenous variables Y and M exogenous variables X, as

$$e = Y - f(X, \beta) \tag{2.6}$$

where β is a set of parameters and $e \sim N(0, \Theta)$ is a set of error terms which are normally distributed with zero mean and covariance matrix Θ, then the likelihood function evaluated for one period may be written as

$$L(\beta, \phi) = \frac{1}{(2\pi)^{1/2}|\Theta|^{1/2}} \exp\{(-1/2)[Y - f(X, \beta)]' \times \Theta^{-1}[Y - f(X, \beta)]\} \tag{2.7}$$

or the log form may be written (after dropping the constant and multiplying through by 2):

$$\log[L(\beta, \phi)] = -\log|\Theta| - [Y - f(X, \beta)]'\Theta^{-1}[Y - f(X, \beta)] \tag{2.8}$$

In the special case where the diagonal elements of Θ are constant and the off-diagonal elements are zero, that is all covariances are zero, then $\Theta = \sigma^2 I$ and (2.8) is maximised by setting B at the value which minimises the squared errors of the model. We see therefore that under the assumption that the model errors are independent and normally distributed then the least squares estimator for β, is equivalent to the maximum likelihood estimator. (The variance–covariance matrix for β is however different in the two cases, in small samples.)

If we now return to the general form of the log likelihood function (2.5), there are two particularly important matrices which can be derived from it. The first of these is the *efficient score* for A, defined as

both provide measure of variance of $\hat{\beta}$.

$$\frac{\partial \log(L(A))}{\partial A} = S(A) \tag{2.9}$$

So at the maximum likelihood estimate of A the efficient score is zero. The second matrix is the *information matrix*. It is defined:

$$E\left[-\left(\frac{\partial^2 \log(L(A))}{\partial A \partial A'}\right)\right] = I(A) \tag{2.10}$$

Where E is the expectations operator, $I(A)$ is a measure of the (average) curvature of the likelihood function. Under a suitable set of regularity conditions it may be shown that the variance of the ML estimator of A is given by the inverse of the information matrix.

$$\mathrm{Var}(\hat{A}_{ML}) = [I(\hat{A})]^{-1} \tag{2.11}$$

Equation (2.11) is a statement of the Cramer–Rao lower bound. As the ML estimator normally attains the Cramer–Rao lower bound in large samples, it is said to be asymptotically efficient.

Concentrating the likelihood function

In (2.5) the parameter vector A contains both the parameters associated with the equation and the unknown moments of the error distribution (in the case of (2.8) these are the elements of the covariance terms). It is often possible to deal with subsets of the total parameter vector however and this is termed 'concentrating the likelihood function'. This can be both analytically and numerically convenient.

Let A consist of two subvectors A_1 and A_2, then (2.5) can be written as $L(A_1, A_2)$. Now suppose we knew a value for A_1, then it is possible that we could derive an analytical formula for the maximum of $L(A_1, A_2)$ with respect to A_2 for that given value of A_1. If this formula can be represented by a function of the form $A_2 = g(A_1)$ then we could write the likelihood function as $L(A_1, g(A_1))$ which could be restated as $L^*(A_1)$, the concentrated likelihood function. As an example consider the single-equation version of (2.8). The likelihood function for the non-linear model evaluated over T periods now is given by

$$L(A) = -T\log(\sigma^2) - e'e/\sigma^2$$

where e is the $T \times 1$ vector of errors $e_t = Y_t - f(X_t, \beta)$ and A consists of both β and σ^2. We may concentrate this likelihood function with respect to σ^2 in the following way. The FOC for a maximum with respect to σ^2 is given by

$$\partial L/\partial\sigma^2 = -T/\sigma^2 + e'e/(2\sigma^2)^2 = 0$$

and so we can derive an expression for σ^2 which is dependent on β, that is $\sigma^2 = e'e/T$, (where e_t depends on the unknown β). The concentrated log likelihood function which depends only on β becomes

$$L^*(\beta) = -T - T\log(e'e/T)$$

The prediction error decomposition

In the likelihood functions specified above, [e.g. in (2.5) and (2.8)] we make the assumption that the observations are independent of each other over time. This assumption will not generally be true when we are dealing with dynamic time series models which include lagged dependent variables. The maximum likelihood approach may still be used, even in this case, by adopting the following factorisation called the prediction error decomposition (Harvey 1981). From the basic definition of conditional probability we know that

$$\Pr(\alpha, \beta) = \Pr(\alpha|\beta)\Pr(\beta) \tag{2.12}$$

That is the unconditional probability of event α occurring is given by the probability of α conditional on β, multiplied by the unconditional probability of β.

This condition may be applied directly to a general form of likelihood function, which is after all simply a particular form of probability function. Suppose we have a general log likelihood function $\log[L(Y)] = \log[L(Y_1, Y_2 \ldots Y_T)]$ where the observations at each time period are not independent due to the dynamic structure of the model. Then by using the log of (2.12) we may write

$$\log[L(Y)] = \log[L(Y_T|Y_1, Y_2 \ldots Y_{T-1})] + \log[L(Y_1, Y_2 \ldots Y_{T-1})] \tag{2.13}$$

The first term is simply the conditional probability of the final period Y_T given the past realisations of Y. The second term is the unconditional probability of $Y_1 \ldots Y_{T-1}$ occurring. This second term can of course be factorised again to give the conditional likelihood function for Y_{T-1} and the unconditional function for $Y_1 \ldots Y_{T-2}$. This process may be repeated for all periods to give

$$\log[L(Y)] = \sum_{i=0}^{T-2} \log[L(Y_{T-i}|Y_1, \ldots Y_{T-1-i})] + \log[L(Y_1)] \tag{2.14}$$

This decomposes the likelihood function into a set of one step ahead prediction errors, v_t.

$$v_t = Y_t - E(Y_t|Y_1 \ldots Y_{t-1}) \tag{2.15}$$

That is the prediction error is defined as actual Y_t minus the models forecast of Y_t conditional on all information up to period $t-1$. For the general non-dynamic model (2.8), $v_t = Y - f(X, \beta)$. Then the likelihood function may be restated for the dynamic case as

$$\log[L(\beta, \Theta)] = -\log|\Theta| - (v'\Theta^{-1}v) \tag{2.16}$$

where Θ is the covariance matrix of the residuals, which is here assumed to be time invariant.

Constructing asymptotic hypothesis tests

In general, the purpose of hypothesis testing is to construct a test statistic which has a well-defined distribution under both the null

(H_0) and the alternative (H_1) hypotheses but which does not depend on the set of *unknown* parameters A.

There are three major classes of test statistic available which allow the construction of such tests: the Wald test, the Lagrange multiplier test and the likelihood ratio test. All three tests rely on the ML procedure and may be regarded as utilising different transformations of the score function. The three procedures are asymptotically equivalent but there small sample properties differ (except when the likelihood function is quadratic in the unknown parameters). One difference between the three tests lies in the point estimate which is used to calculate the test statistics. The Wald test is evaluated using only an unrestricted estimate of the model, the Lagrange multiplier test is evaluated using only the restricted estimate of the model (i.e. under the null) and the likelihood ratio test uses information from both the restricted and unrestricted estimates. In practice therefore the choice between the procedures is often made on the grounds of which set of estimates is actually easiest to compute. All three test procedures are frequently used when the estimated system is non linear because they may give different inferences in small samples.

Suppose the *unrestricted* ML estimate of the true vector A is $\hat{A}$ then we may wish to test the general restriction $H_0{:}g(A)=0$ against the alternative $H_1{:}g(A)\neq 0$. The function $g(A)$ must be a function for which all the restricted parameters can be estimated; $g(A)$ must also be continuously differentiable and $(\partial g/\partial A)$ must have full rank in the neighbourhood of A. Gallant and Holly (1984) give a full set of conditions on g. The simplest forms for $g(A)$ for a single parameter are $a_1=0$ or $a_1=1$, etc. A linear restriction involving two parameters might be $g(A)=a_1+a_2-1=0$ or a joint hypothesis might be $a_1=0$ *and* $a_2-1=0$. A non-linear restriction would be $g(A)=a_1^2-4(a_2^2/a_1)=0$, for example.

The likelihood ratio test

The likelihood ratio test (LR) is the oldest of the three procedures, originating from the work of Neyman and Pearson (1928). It relies on the comparison between the value of the likelihood function at the unrestricted estimate $\hat{A}$ and the restricted estimate $[A^r|g(A)=0]$. It is clear that

$$\text{LR} = \frac{L(A^r)}{L(\hat{A})} < 1 \qquad (2.17)$$

since by definition $L(\hat{A}) > L(A^r)$. We need now to express this term in a form which will have a well-defined asymptotic distribution. This is done by taking a Taylor series expansion of $\log[L(A)]$ around the unrestricted estimate $\hat{A}$. (A suitable set of regularity assumptions is needed to justify this procedure.)

$$\log[L(A)] = \log[L(\hat{A})] + (\hat{A}-A)\left[\frac{\partial \log[L(A)]}{\partial A}\right] + \tfrac{1}{2}(\hat{A}-A)'\left[\frac{\partial^2 \log[L(A)]}{\partial A \partial A'}\right](\hat{A}-A) + 0(1) \qquad (2.18)$$

Where $O(1)$ refers to a set of terms which is asymptotically negligible. At $\hat{A}$:

$$\frac{\partial \log[L(A)]}{\partial A} = S(A) = 0 \qquad (2.19)$$

and

$$\frac{\partial^2 \log[L(A)]}{\partial A \partial A'} \xrightarrow{P} I(A) \qquad (2.20)$$

Dropping the term $O(1)$ we may state (2.18) (following Serfling 1980) as

$$\log[L(A)] = \log[L(\hat{A})] + \tfrac{1}{2}(\hat{A}-A)'I(\hat{A})(\hat{A}-A) \qquad (2.21)$$

From (2.17) we have that

$$-2\log(\text{LR}) = 2\{\log[L(\hat{A})] - \log[L(A^r)]\} \qquad (2.22)$$

and so from (2.21), replacing the 'unknown' A by A^r,

$$-2\log(\text{LR}) = (\hat{A}-A^r)'I(\hat{A})(\hat{A}-A^r) \qquad (2.23)$$

Also, under a reasonable set of regularity conditions it is known that asymptotically an ML estimate gives

$$\sqrt{n}(\hat{A}-A) \sim N(0, I(A)^{-1}) \qquad (2.24)$$

and that $(\hat{A}-A)'I(A)(\hat{A}-A)$ is $\chi^2(m)$, where m is the number of constraints. Hence using (2.23) we may write the likelihood ratio test statistic as LRT.

$$\text{LRT} = 2\{\log[L(\hat{A})] - \log[L(A^r)]\} \sim \chi^2(m) \qquad (2.25)$$

This is the usual form of the likelihood ratio test and simply states that the difference in the log-likelihoods (multiplied by 2) is $\chi^2(m)$. If the test statistic, LRT, exceeds the chosen critical value then we reject the restriction.

Three test procedures

The three general forms of test procedure used are the (LR) test (as described above), the Wald test (W) and Lagrange multiplier (LM) test. To illustrate the relationship (Buse 1982) between these three test procedures, suppose we wish to test the simple hypothesis on the scalar parameter A, namely $H_0{:}A = A_0$ against $H_1{:}A \neq A_0$. The LR test computes the value of the likelihood under *both* H_0 and H_1 and directly computes the distance (1/2)LR (Figure 2.1). The distance (1/2)LR depends on the distance $(\hat{A} - A_0)$, and the curvature of the log likelihood function which we define as $R(\hat{A}) = |(d^2 \log L)/dA^2|$ evaluated at $A = \hat{A}$. For a given distance $(A - A_0)$ the greater the curvature or 'steepness' of the likelihood function the larger is the distance (1/2)LR. Thus the 'precision' of the ML estimate $\hat{A}$ is greater for likelihood function L^1 (Figure 2.1) than for likelihood L^*. With the likelihood function L^1, we would tend to reject $A = A_0$ more often than with likelihood L^*. If the curvature $R(\hat{A})$ is large then the variability of A around its ML estimate is small: somewhat loosely the variance is inversely related to the curvature.

The Wald test uses only the unrestricted ML estimates. Intuitively in the Wald test we estimate the distance (1/2)LR by standing at point X, measuring the distance $(\hat{A} - A_0)$ and estimating the position of P_1 (or P_2) using the curvature $R(\hat{A})$ evaluated at the maximum point X. Thus we might define the Wald statistic for $H_0{:}A = A_0$ by

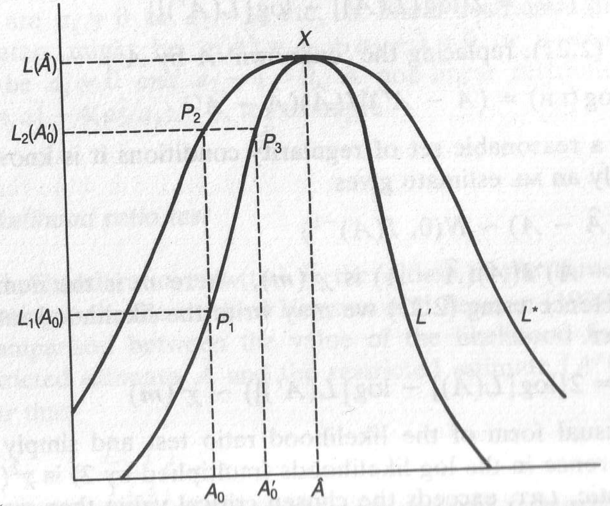

Figure 2.1 The three test procedures.

$W = (\hat{A} - A_0)^2 R(\hat{A})$. However, the Wald statistic uses the *average* curvature, as measured by the information matrix:

$$W = (\hat{A} - A_0)^2 I(\hat{A})$$

where $I(\hat{A})$ is defined in (2.10).

We can now generalise the above for a set of g non-linear restrictions $g(A) = 0$ on the k parameters $g < k$, and the Wald statistic (Wald 1943) is

$$W = [g(\hat{A})]'\{G(I(\hat{A}))^{-1}G'\}^{-1}g(\hat{A}) \qquad (2.26)$$

Where G is the $g \times k$ matrix of partial derivatives $\partial g(A)/\partial A$ evaluated at $\hat{A}$. Large values of W are generated by large deviations of $g(\hat{A})$ from zero and the deviations are 'weighted' by the average curvature of the log-likelihood. Hence for large values of W we reject H_0. The Wald statistic is distributed $\chi^2(m)$ where m is the number of restrictions in the vector g. For example, the hypothesis $H_0{:}A = A_0$ (where $A_0 = 1$ say) is a special case and here $g(A) = (A - A_0)$. Hence G is the identity matrix. It is easily seen that the standard t-test for a restriction on a single parameter in a linear regression is a particularly simple form of Wald test. Suppose we wish to test the restriction $\hat{\beta} = 0$ in a linear regression. Then $g(\beta) = \beta - 0$, the Wald test is:

$$W = \hat{\beta}(I(\hat{A}))^{-1}\hat{\beta} = \hat{\beta}^2/(\text{Var}\,\hat{\beta}) \sim \chi^2(1) \qquad (2.27)$$

where we have noted that the inverse of the information matrix is the ML estimate of the variance. The Wald test in this case is therefore simply the square of the standard t-test (and gives the same inference asymptotically as applying the χ^2 distribution as in (2.27)).

The Lagrange multiplier test, suggested by Aitchison and Silvey (1938) and a closely related test, the Rao statistic (Rao 1948) are both based solely on the restricted estimate of the model. The Lagrange multiplier test is sometimes referred to as the efficient score test as it is based on the asymptotic distribution of the score function

$$\frac{1}{\sqrt{n}} S(A) \sim N(0, I(A)) \qquad (2.28)$$

Intuitively the LM test *estimates* the distance (1/2LR$_2$) (Figure 2.1) but uses P_2 as its starting point. First the likelihood function is evaluated with the restriction $A = A_0$ imposed, that is, at point P_2. We then estimate the point X based on the curvature of L^* *at* P_2. The unrestricted ML estimate $\hat{A}$ satisfies the equation $S(\hat{A}) = \partial \log L/\partial A = 0$ where S is the score function. At $A = A_0$ the score function is *not* zero and $[S(A_0)]^2$ therefore gives a measure of the departure of A_0 from $\hat{A}$. However, two likelihoods can generate the same value for $[S(A_0)]^2$ but one has A_0 closer to the maximum. We therefore weight the 'squared slope' by the curvature of L^*. In fact the greater the curvature (i.e. L' as opposite to L^* in Figure 2.1) for any value of $\log[L(A_0)]$ the closer is the restricted estimate to $\hat{A}$ (compare points P_2 and P_3 in Figure 2.1, where A_0' is clearly closer to $\hat{A}$, and the curvature of L' is larger than that for L^*). We therefore weight by the inverse of the (expected value) of the curvature $[I(A_0)]^{-1}$ evaluated at the *restricted estimate* A_0. Our simple LM test statistic is therefore

$$\text{LM} = [S(A_0)]^2[I(A_0)]^{-1}$$

The generalised version is

$$\text{LM} = [S(A_0)]'[I(A_0)]^{-1}[S(A_0)] \sim \chi^2(m) \quad \textbf{(2.29)}$$

where m is again the number of restrictions.

The intuition behind the test is that if the restrictions hold exactly (i.e. $A^r = \hat{A}$) then $S(A^r) = 0$ as this is the first order condition for an unconstrained maximum. The departure of $S(A^r)$ from zero therefore measures the strength of the effect of the restriction, on the maximum likelihood value.

The relationship between the three procedures depends on how good an approximation the second derivative is able to give of the value of the likelihood function at the restricted or unrestricted estimates. If we are testing a simple linear restriction, as in the example above, and the likelihood function is quadratic then the second derivative would provide a perfect estimate of the global shape of the likelihood function. In this case all three test procedures produce exactly the same numerical value, W = LM = LR. When, however, the second derivative is not known with certainty but must be estimated, this equality disappears and instead we have that W ⩾ LR ⩾ LM as demonstrated by Berndt and Savin (1977).

A transformation of the LM test

The formula presented in (2.27) for the Lagrange multiplier test is not a particularly useful one in practice as it requires estimates of both the information matrix and the score matrix, evaluated with the restrictions imposed. Following Breusch and Pagan (1980) a transformation of (2.27) which is particularly easy to calculate and which is applicable to a wide class of problems may be performed. Suppose the model takes the form of the non-linear regression

$$Y_t = f(x_t; A) + e_t \equiv f_t + e_t \qquad t = 1 \ldots T \quad \textbf{(2.30)}$$

where the errors, e_t, are identically and independently normally distributed as $N(0, \sigma^2)$ and f_t is independent of all e_t. The parameter set A is split into two subsets: A_1, which will be restricted (fixed) and A_2 which are unrestricted. The log-likelihood function will have the general form of (2.8) and the information matrix will be block diagonal between the terms in A and σ^2 and so we can concentrate solely on the term due to A. The non-linear restrictions are $g(A) = 0$. Now to evaluate (2.29) we need

$$S = \frac{\partial \log L}{\partial A} = \sigma^{-2}G'e \quad \textbf{(2.31)}$$

and

$$V = E\left[\frac{\partial^2 \log(L)}{\partial A^2}\right]^{-1} = (\bar{\sigma}^2 E(G'G))^{-1} \quad \textbf{(2.32)}$$

where G is a matrix of partial derivatives of g with respect to the parameter A. We may then write (2.29) as

$$\tilde{\sigma}^{-2}\tilde{e}'\tilde{G}[\tilde{\sigma}^{-2}\tilde{E}(G'G)]^{-1}\tilde{\sigma}^{-2}\tilde{G}'\tilde{e} \quad \textbf{(2.33)}$$

where ~ denotes an estimate formed at the restricted parameter set $A_1 = \tilde{A}$. If $\tilde{E}(G'G)$ is estimated as $\tilde{G}'\tilde{G}$ then (2.33) may be simplified to

$$\tilde{\sigma}^2\tilde{e}'\tilde{e} \quad \textbf{(2.34)}$$

which may be interpreted as TR^2 where R^2 is the coefficient of determination in the regression of $\tilde{e}$ on $\tilde{G}$.

This procedure has found a range of applications, the most popular of which is the Lagrange multiplier test for serial correlation.

Suppose the unrestricted model is

$$Y = X\beta + u \quad \textbf{(2.35)}$$

$$u_t = \rho_i u_{t-i} + \varepsilon_t \quad \textbf{(2.36)}$$

This model may be transformed to give

$$Y_t = \rho_i Y_{t-i} + (X_t - X_{t-i}\rho_i)\beta + e_t \equiv f(\rho, \beta) + e_t \quad \textbf{(2.37)}$$

which puts it into the notation given above. Now if we wish to construct the LM test for $\rho_i = 0$ we proceed as above by identifying $\rho_i = 0$ with $A_1 = A$ and β as A_2. For $\rho_i = 0$ the restricted estimates of β are given by an OLS regression of Y on X. The residuals from this OLS regression $\hat{u} = Y - X\hat{\beta}$ may then be associated with $\tilde{e}$. To derive the elements of $\tilde{G}$ we note that

$$\frac{\partial f}{\partial \rho_i} = (Y_{t-1} - X_{t-i}\beta)' = \hat{u}'_{-i} \quad \textbf{(2.38)}$$

$$\frac{\partial f}{\partial B} = (X_t - X_{t-i}\rho)' \quad \textbf{(2.39)}$$

so that $\tilde{G} = (\hat{u}_{-i}, X_t)$.

Now under the null hypothesis that $\rho_i = 0$ and that X is strictly exogenous it may be shown that the LM statistic becomes Tr_i^2 where

$$r_i = (\hat{u}'\hat{u})^{-1}\hat{u}'_{-1}\hat{u}$$

An alternative form of this test may also be constructed by performing the regression

$$\hat{u}_t = X_t\delta + \sum_{i=1}^{m} \gamma_i \hat{u}_{t-i} \quad \textbf{(2.40)}$$

The R^2 from this regression may then be used to form an LM statistic as $TR^2 \sim \chi^2(m)$. This test may be constructed either for individual lagged errors or for a number of lagged errors considered jointly. If TR^2 exceeds $\chi^2(m)$ then we reject the null hypothesis that the restrictions are valid (i.e. that there is no serial correlation of order $1, 2, \ldots m$).

Both of these forms of the LM test have the same intuitive interpretation. On the assumption that $\rho_i = 0$ the correlation between u_t and u_{t-i} should be zero. The first test looks at this correlation directly and its interpretation is obvious. In the second test, because $\hat{u}_t$ are the OLS residuals, the R^2 of $\hat{u}_t$ regressed on X_t is zero, (recall that the normal equations for OLS imply $\hat{u}'X = 0$). So the R^2 of the auxiliary regression (2.40) measures the extra explanatory power given by the terms $\hat{u}_{t-i} (i = 1, 2, \ldots m)$. If R^2 from (2.40) is low then there is a low correlation between $\hat{u}_t$ and $\hat{u}_{t-i}$ and hence autocorrelation is unlikely to be present.

2.2 Non-linear optimisation procedures

In general the log-likelihood function is a non-linear function of the parameters of the model and often $\log(A)/\partial A = 0$ is not amenable to an analytical solution. There are, of course, exceptions to this statement the most important of which is the case of the general linear model where the ML estimate of β can be derived analytically and is numerically equal to the OLS estimator. If an analytic solution is not possible we must use some numerical method for finding the maximum likelihood parameter values. The techniques which may be used are applicable to maximising any objective function which is a non-linear function of a set of control variables and we will discuss them within this general framework.

We therefore view the parameters (A) of the model as a set of control variables, C, and similarly the likelihood function itself is viewed simply as any general non-linear function of those control variables. Our objective may then be described without loss of generality as

$$\text{Min } H(C) = -\log[L(A)] \quad \textbf{(2.41)}$$

where C is a vector of all the parameters of the system – such as Π in (2.1) and, in some cases, any unknown variance or covariance terms.

Practical computation

The numerous methods of solving a minimisation problem, such as (2.41), proceed along a broadly similar set of steps and may all be classified under the general heading of hill-climbing algorithms. From an initial, and arbitrary, guess of the optimal solution C^*, say C_1, they attempt to construct a sequence of vectors $C_1, C_2 \ldots, C_N$ such that at every point on the sequence $H(C_J) < H(C_{J-1}) < H(C_{j-2}) < \ldots$ etc. and as $N \to \infty$, $C_N \to C^*$.

The broad steps of achieving this sequence may be outlined as follows:

1. Set an arbitrary initial value for C_i.
2. Determine a *direction of movement* for C_i which will decrease the value of $H(C_i)$.
3. Determine a 'step length' for the change C_i and evaluate the objective function of C_{i+1}.
4. Examine a terminal criterion; if it is fulfilled, stop. If it is not fulfilled, set $i = i + 1$ and repeat the procedure from step 2.

A usual criterion for termination would be that $H(C_{i-1}) - H(C_i) < \varepsilon$ where ε is some small tolerance. Because of the possibility of the algorithm 'jamming' at some non-optimal point we might also examine

$$\frac{\partial H}{\partial C'_K} \qquad K = 1, 2, \ldots J$$

or

$$\left(\left|\frac{\partial H}{\partial C'}\right|\right)'\left(\left|\frac{\partial H}{\partial C'}\right|\right)$$

to see that both of these are close to zero.

Among the hill-climbing algorithms by far the most important group are those which base the optimisation procedure on the calculation of derivatives of the objective function. These algorithms are known collectively as gradient methods for example, the Newton method, Davidson-Fletcher-Powell, steepest descent, and quadratic hill climbing. The non-gradient, or derivative free methods, are generally of most use when the function to be minimised is extremely irregular. This class includes, for example, the Powell algorithm, the non-linear Simplex method and grid search methods.

Gradient methods

Given a current value C_i the gradient methods all proceed by constructing a sequence where

$$H(C_{i+1}) < H(C_i) \tag{2.42}$$

where C_{i+1} is defined as follows

$$C_{i+1} = C_i + s\,\mathrm{d}(C) = C_i + s[V(C).\partial(C)] \tag{2.42a}$$

s = the *step length* (a positive scalar), $\partial(C)$ is the ***gradient***, (we use '∂' as shorthand below) a vector of first-order partial derivatives of H with respect to the control variables (i.e. $\partial = \partial H/\partial C$). $V = V(C)$ is a function which varies depending on the gradient method used. $\mathrm{d}(C)$ is the direction vector and depends on the gradient *and* the function $V(C)$. The evaluation of both the first and indeed the second derivatives (see below) may be done either analytically or numerically. For analytical calculation the actual formulae for the derivatives must be coded into the computer program (for example, if $H(C) = 2C^2 - 4C$, $\partial H/\partial C = 4C - 4$, $\partial^2 H/\partial C^2 = 4$). In the case of complex functions it is often impossible to calculate derivatives analytically. In practice it is often satisfactory to use a numerical *approximation* to the derivatives, so that for the first derivative we use:

$$\frac{\partial H}{\partial C_K} = \frac{H(C_1C_2, C_K + \Delta, C_{K+1}, C_J) - H(C_1, C_2, C_J)}{\Delta} \tag{2.43}$$

where Δ is a suitably small number. To illustrate the calculation of the numerical derivative (2.43) consider the simple quadratic

$$H(C) = 2C^2 - 4C$$

In this simple case we can solve analytically for the first derivative, ∂, and second derivatives, $\partial^{(2)}$

$$\partial = \partial H/\partial C = 4C - 4$$

The numerical approximation ∂_a for arbitrary values $C_1 = 0.5$, $C_2 = 0.52$ and hence $\Delta = 0.02$ is

$$\partial_a = [H(C_2) - H(C_1)]/0.02 = [(-1.5392) - (-1.5)]/0.02$$
$$= -1.96$$

We can check this approximation by evaluating the 'true' slope using $\partial = 4C - 4$. For $C_1 = 0.5$, $\partial_1 = -2$, while for $C_2 = 0.52$, $\partial_2 = -1.92$, so the approximation ∂_a lies between the two analytic values as one would expect. Similarly the second derivative $\partial_a^{(2)}$ can be approximated by

$$\partial_a^{(2)} = (\partial_{a2} - \partial_{a1})/(C_2 - C_1)$$

Equation (2.43) is a 'one-sided' derivative calculation; improved accuracy can be achieved, at extra cost, by using a two-sided approximation. The choice of Δ embodies two considerations: an accurate derivative requires a small Δ, but if there is any inaccuracy in the objective function evaluation $H(C)$ itself then Δ must not become so small that the inaccuracy significantly affects the calculation of ∂.

The Newton method

The Newton (sometimes called Newton–Raphson) method is perhaps the most fundamental of the gradient methods. Many of the other methods are developments of it, or approximations to it, and are often called quasi-Newton methods. The Newton method makes use of the matrix of second derivatives of the objective function with respect to the control variables (the Hessian matrix) for V:

$$V = \left(\frac{\partial^2 H}{\partial C \partial C'}\right)^{-1} \tag{2.44}$$

and $s = 1$. If the function H were quadratic the Newton step procedure would reach the optimum point in one iteration. In essence the algorithm works by making a series of local quadratic approximations of H, solving this problem and then recomputing the approximation.

In order to give some intuitive understanding of the procedure consider the one control variable case where the minimum is given by C^*. A Taylor series expansion of $H(C)$ around the minimum C^* gives

$$H(C) = H(C^*) + (C - C^*)\partial(C^*) + (1/2)(C - C^*)^2\partial^{(2)}(C^*)$$

Differentiating with respect to C and noting that $H(C^*)$, $\partial^{(2)}(C^*)$ are constants (for a given $C = C^*$) then

$$\partial(C) = \partial(C^*) + (C - C^*)\partial^{(2)}(C^*)$$

Rearranging and noting that at the minimum $\partial(C^*) = 0$, we have

$$C^* = C - [\partial^{(2)}(C)]^{-1}\partial(C) \tag{2.45}$$

If $H(C)$ is quadratic then C could be set at *any* initial value and C^* would be given *exactly* by the RHS of (2.45) (see below). For more general functions the latter does not hold but (2.45) suggests an iterative scheme

$$C_2 = C_1 - [\partial^{(2)}(C_1)]^{-1}\partial(C_1) \tag{2.46}$$

In this single parameter case $V(C) = [\partial^2(C_1)]^{-1}$, the general case is shown in (2.44).

To illustrate some of the above points consider our quadratic example $H(C) = (2C^2 - 4C)$. Analytically the minimum is given by $\partial(C) = 4C - 4 = 0$, and hence $C^* = 1$. How would our iterative scheme handle this problem starting with an arbitrary starting value $C_1 = 2$ and the *analytic* derivatives $\partial = 4C - 4$ and $\partial^{(2)} = 4$ so that $\partial(C_1) = 4$, $\partial^{(2)}(C_1) = 4$, hence

$$C_2 = 2 - (1/4)4 = 1$$

The minimum is achieved in one iteration when $H(C)$ is quadratic and we utilise analytic derivatives. (This is because the curvature $\partial^{(2)}$ is constant for any quadratic.) Consider next a cubic for $H(C)$:

$$H(C) = C^3 - 3C^2 + 7$$

hence

$$\partial(C) = 3C^2 - 6C$$
$$\partial^{(2)}(C) = 6C - 6$$

If we begin with $C_1 = 1.5$, then $\partial = -2.25$, and $\partial^{(2)} = 3$, hence

$$C_2 = (1.5) - (-2.25)/3 = 2.25$$

The next iteration is

$$C_3 = (2.25) - (0.168)/7.5 = 2.02$$

Analytically we know the solution namely $\partial(C) = 0$, which implies $C^* = 2$ so our second iteration is close to the optimum. One problem with the Newton–Raphson method is that it may move *away from* the minimum if $\partial^{(2)}$ is not positive definite. For example if we had chosen $C_1 = 0.5$ then $\partial^{(2)} = -3$ (i.e. negative) and $C_2 = 0.5 - (-2.25)/(1.3) = -0.25$ and we move in the wrong direction. Some techniques modify the basic Newton–Raphson procedure to ensure that the gradient $\partial(C)$ is always multiplied by a positive definite matrix (see below).

Method of steepest descent

At the current point C_i, the direction which will improve the objective function most rapidly is given by the vector of first derivatives, ∂. The method of steepest descent therefore simply sets V equal to the identity matrix (or minus the identity matrix if the problem is being maximised). The important choice therefore becomes the determination of the step size. In this case some variant of the Armijo (1966) step procedure is generally used. This works as follows: a succession of steps is generated using

$$s_i = \lambda B^i \qquad i = 0, \ldots$$

where λ is some given maximum step size and B is a constant $0 < B < 1$. Some form of grid search may then be used over these step sizes to check for the best step size at each iteration.

The method of steepest descent avoids the costly computation of the Hessian matrix but its disadvantage is that convergence can often be slow and there are well-known examples where the algorithm will not reach a maximum.

Method of quadratic hill climbing

The method of quadratic hill climbing (Goldfeld *et al.* 1966) is a slight extension of the standard Newton algorithm to include a variable step size and to ensure a positive definite matrix. The iterative scheme is $C_2 = C_1 - sQ\partial$ where $Q = (V + uI)^{-1}$, with u a positive scalar. This may improve the performance of the algorithm when the function is non-concave or is not close to quadratic.

Quasi-Newton methods

In order to calculate the Hessian matrix required by the Newton method, either an expensive numerical procedure must be repeated at each iteration or the analytical second derivatives must be calculated and supplied by the user. The quasi-Newton methods are a family of algorithms which avoid this necessity by calculating an *approximation* to the Hessian matrix and continually update it and hence improve on the approximation (to the true matrix of second derivatives).

From (2.42) we can see that for iteration $i+1$, the inverse of the second derivative matrix at iteration i was

$$\left(\frac{\partial^2 H}{\partial C \partial C'}\right)_i^{-1} = (C_{i+1} - C_i)(\partial_{i+1} - \partial_i) \qquad \textbf{(2.47)}$$

so by comparing the parameter estimates (C_{i+1}, C_i) and the derivatives $(\partial_{i+1}, \partial_i)$ at two *succeeding iterations* we can estimate the Hessian at the last iteration. This may be compared with the estimate at iteration i, namely, E_i and then some correction based on the error can be made so that

$$E_{i+1} = E_i + f\left(E_i, \frac{\partial^2 H}{\partial C \partial C'}\right) \qquad \textbf{(2.48)}$$

where 'f' is a function of the Hessian evaluated at iteration i, the precise form of the correction determines the form of the quasi-Newton algorithm under consideration. One of the most common algorithms in this class is the Davidson–Fletcher–Powell method. Himmelblau (1972) presents a number of correction formulae.

Scoring

It is sometimes easier to obtain a numerical approximation to the *expectation* of the matrix of second derivatives, that is the *information* matrix, $I(C)$

$$I(C) = -E[\partial^2 H / \partial C \partial C']$$

The iterative scheme is then

$$C_{i+1} = C_i + [I(C)]^{-1} \partial(C)$$

and the procedure is known as the *method of scoring*. It is likely to have a slower rate of convergence than Newton–Raphson since $I(C)$ is only an approximation to the Hessian. However, the information matrix is easier to compute and will be estimated more quickly. A

variable step length is also often incorporated. If the model is identified then $I(C)$ is always positive definite.

Derivative-free techniques

Generally speaking, optimisation techniques which employ derivatives are fast and reliable when the function being maximised is well behaved. However the derivative-free techniques are recommended for highly non-linear functions or functions which are subject to discontinuities. In principle the reason for this is simple to understand: the gradient-based techniques work by examining the first and second derivatives at a single point and drawing an inference about the whole surface based on some simple regularity conditions. When a function is either discontinuous or highly non-linear the information given at a single point can be very misleading. (Consider trying to find the direction of Everest, based on the slope of a minor peak in the foot-hills of the Himalayas.) The derivative-free techniques generally derive their 'working information' by examining a larger area around a current point on a surface and so they are less likely to draw very fragile inferences about the shape of the surface being climbed. (Derivative-free techniques maybe likened to having a powerful pair of binoculars at the top of a local peak from which one can see Everest in the distance, although, of course, there is no guarantee of this.)

The two widely used algorithms in this class are the conjugate gradient method of Powell (1964) and the non-linear Simplex method suggested by Spendley, Hext and Himsworth (1962). The Powell technique works essentially by carrying out a set of linear searches in orthogonal pairs and deriving a direction of movement from this information. The Simplex technique constructs a simplex around some initial point and evaluates the objective function at the vertices of the simplex. A simplex is the simplest shape which has positive area in any given dimension; in the two-dimensional plane it is simply a triangle. The algorithm works by starting from an arbitrary simplex in the hill-climbing space and examining the value of the objective function at each corner; it then drops the least desirable corner and calculates a point which is a weighted average of the other corners. A line search is then conducted from the least desirable corner in the direction of the weighted point. The best point along this line then forms one of the corners of a new simplex which is completed by using the corners of the old simplex with the exception of the worst one which has already been dropped. The algorithm then repeats itself, thus moving the simplex around the n-dimensional space until the maximum is bracketed within the simplex and then collapsing the simplex around the maximum until all the corners lie arbitrarily close to the optimal point.

Inequality-constrained optimisation

In many cases we may wish to maximise an objective function $H(C)$ while obeying a set of inequality constraints (for example, that the probability of default in a loan is always greater than zero). This complicates the maximisation algorithm substantially. There are basically two approaches to dealing with this problem: the first involves adapting the objective function so as to penalise any violations of the constraint; the second adapts the optimisation algorithm.

When we adapt the objective function the technique is generally known as a barrier method. The idea is to define a barrier function which heavily penalises violation of the constraint but has a near-zero effect when the constraint is satisfied. If we have the following set of inequality constraints

$$G(C) \geqslant 0 \tag{2.49}$$

we create a set of barrier functions such as

$$B[G(C)] \tag{2.50}$$

Where $B[G(C)]$ is near zero for $G(C) \geqslant 0$ and is large for $B[G(C)] < 0$, a typical function for iteration i, might be

$$B_i[G(C)] = -\gamma \ln [G_i(C)]$$

where γ is a suitably chosen weighting factor. Since as $G(C) \to 0$ the log approaches minus infinity, this severely penalises moving $G_i(C)$ towards zero.

Disadvantages of this technique are: (a) a good barrier function should be highly non-linear and therefore makes the optimisation more difficult; (b) if the unconstrained optimum were near or on the constraint the barrier function will tend to distort the final solution. If a barrier function is to be used it is often advisable to experiment by sequentially dropping some or all of the constraints, to check which individual constraints do not hold in the unconstrained optimisation.

The second main approach to inequality constraints is to adapt the direction finding procedure so that the algorithm does not move in directions which violate the inequality constraints. This amounts to deriving a value '$V(C)\partial(C)$' in (2.42a) in such a way that it will not cause steps out of the feasible region. Algorithms which implement such procedures are collectively termed methods of feasible direction and a detailed survey of these techniques may be found in Polak (1972). A typical procedure would be to derive the gradient vector and then calculate the derivatives of any close inequality constraints. A linear-programming problem may then be formed which maximises the change in the objective function, given from the gradient vector, subject to non-violations of the constraints.

2.3 Special forms of likelihood functions

Qualitative response models

The basic idea which lies behind the qualitative response (QR) model (sometimes referred to as limited dependent variable model) is that there are times when we either have only partial information on a variable or the information is not continuous. For example in Tobin (1958), a model of the demand for cars is constructed using disaggregated data. The basic idea is that expenditure on cars is related to an individual's income. The problem is that some individuals choose not to buy cars at all, and so individuals are divided into two groups, G_1, those who had bought a car and G_2, those who did not. Hence we have only partial information. If we simply remove the group G_2 from the sample we would get a biased estimate of the income elasticity. Consider the second case where we have a non-continuous variable, in the simplest case 1 or 0 say. For example, we might know when an incomes policy is on '1' or off '0' but we have no continuous measure of the strength of the policy. Finally, a classic example from the field of biology is the testing of the effect of poison on insects. The hypothesis is that more poison increases the probability of death, but the observations on the individual insect come in the form of survivors '0' or deaths '1'.

The general approach to this class of problem is to assume a linear regression model:

$$Y_t = \beta X_t + u_t$$

We observe Y_t only if $Y_t > 0$, so that the model becomes

$$Y_t = \beta X_t + u_t \qquad \text{if } \beta X_t + u_t > 0$$
$$Y_t = 0 \qquad \text{otherwise}$$

If we attempt to estimate this equation by OLS using only the G_1 observations when $Y_t > 0$ then the resulting estimates would be biased and inconsistent since we cannot assert that the $E(u_t) = 0$ for all t. The approach in the QR model is to define the likelihood function for the model on the assumption that the error term follows a particular distribution. The assumption made by Tobin (1958) was that u_t has a normal distribution with zero mean and variance σ^2, this gives rise to the Tobit (or Tobin's probit) model. We may then write the likelihood function L for the model as

$$L = \prod_{Y \varepsilon G_1} \left\{ \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2\sigma^2} (Y_t - \beta X_t)^2 \right] \right\} \times \prod_{Y \varepsilon G_2} \left\{ \int_{-\infty}^{0} \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2\sigma^2} (Y_t - \beta X_t)^2 \right] dy \right\} \tag{2.51}$$

This may be maximised to give estimates of σ and β using the numerical techniques discussed earlier in this chapter.

Variations on the QR model generally involve alternative assumptions about the distribution of the error term. We can therefore present a compact form of the likelihood function by defining $f(.)$ to be a given density function and $F(.)$ to be the cumulative density function. Using this notation we may restate the likelihood function as

$$L = \prod_{Y \varepsilon G_1} \left[\frac{1}{\sigma} f[(Y_t - \beta X_t)/\sigma] \right] \prod_{Y \varepsilon G_2} [F(-\beta X_t/\sigma)] \tag{2.52}$$

An important alternative QR model arises when $F(.)$ is defined as the logistic function

$$F(w) = \frac{e^w}{1 + e^w}$$

This model is then known as a *logit model*. Its main advantage over the Tobit model is the ease of numerical calculation, as the logistic function is much easier to calculate than the cumulative normal function. (Amemiya, 1981 discusses the relative merits of the two functions at some length.)

In the case where we have only discrete observations on the dependent variable, then the likelihood function is a simplification of the general QR model. For example, suppose $Y = 1$ when a government is operating an incomes policy and $Y = 0$ if the policy is inoperative. If G_1 is the group when $Y = 0$ then the likelihood function involves only the cumulative density function of the following form:

$$L = \prod_{Y \varepsilon G_1} F(-\beta X_t) \prod_{Y \varepsilon G_2} [1 - F(-\beta X_t)] \tag{2.53}$$

The likelihood function would then represent a Probit model, when $F(.)$ is the cumulative normal (density) function and a logit model when it is in the logistic function.

Discrete switching disequilibrium models

A model closely related to the qualitative responses QR model is the discrete switching disequilibrium model. The link between them lies in the form of the likelihood function. Both the QR and disequilibrium model contain terms in the cumulative density function of the error terms of the model. The disequilibrium model contains equations for the 'notional' or 'desired' demand Y^d and supply Y^s of the good Y. These typically have the form

$$Y_t^d = \alpha_1 P_t + \beta_1' X_t + u_{1t} \tag{2.54}$$

$$Y_t^s = \alpha_2 P_t + \beta_2' X_t + u_{2t} \tag{2.55}$$

where P_t is the real price of the good, X_t is a vector of exogenous variables, β_1, β_2 are vectors of parameters and u_{1t} and u_{2t} are normally distributed error processes. Equations (2.54) and (2.55) are standard to any single market model. The distinguishing feature of the disequilibrium model is given by the method of determining Q_t the *actual* quantity of the good Y_t which is to be traded in each time period. The standard equilibrium model makes the assumption that $Q_t = Y_t^D = Y_t^S$ that is, the actual quantity Q_t is given by the intersection of demand and supply. Other assumptions which are sometimes made are, $Q_t = Y_t^d$, that is demand side dominance only, or $Q_t = Y_t^S$, supply side dominance only. The assumption made in disequilibrium models is $Q = \min(Y^d, Y^s)$, that is the *traded quantity* is determined by the smaller of the notional demand or supply.

The justification for this approach is based on the idea of voluntary exchange. A notional demand or supply curve may be thought of as defining the maximum amount of a good which will be exchanged voluntarily at any given price. If someone is offered a smaller quantity than he demands at a given price, he will generally accept this trade as profitable, but an individual will not generally purchase a larger quantity than indicated by his demand curve.

In order to close the disequilibrium model it is necessary to make some assumption about the determination of prices. The typical assumption is:

$$P_t = P_{t-1} + \gamma(Y_t^d - Y_t^s) + u_{3t}, \qquad \gamma > 0 \tag{2.56}$$

If demand is greater than supply, the real price will rise and vice-versa. Equations (2.54)–(2.56) then constitute a full statement of the single market disequilibrium (SMDM) model. Over time, the real price will tend to adjust to the market clearing price and the speed at which it does this is governed by γ. If γ becomes very large the disequilibrium model will closely approximate the equilibrium model. Alternatively, if γ is small the disequilibrium will persist for a considerable time. One of the advantages of using an empirical model based on (2.54)–(2.56) is that the estimate of γ will give us an indication of how closely the model approximates a market clearing model.

An early attempt to estimate a model of this type is due to Fair and Jaffee (1972). However, their work is not based on the maximum likelihood approach but makes the simplifying assumption that $u_{3t} = 0$ and utilises instrumental variable estimation of the model. The likelihood function for SMDM was developed by Maddala and Nelson (1974).

The derivation of the likelihood function begins by defining:

$$g(Y_t^D, Y_t^S) \tag{2.57}$$

as the joint probability density function of the unobserved random variables (Y_t^d, Y_t^s) and $h(Q_t)$ as the probability density function of Q_t, the *traded* quantity. We can then relate

$$h(Q_t) \text{ to } g(Y_t^D, Y_t^S) \tag{2.58}$$

in the following way:

$$h(Q_t) = f(Q|Y_t^D < Y_t^S)\Pr(Y_t^D < Y_t^S) + f(Q_t|Y_t^S \leq Y_t^D)\Pr(Y_t^S \leq Y_t^D) \tag{2.59}$$

That is to say, the PDF of Q_t is given by (a) the conditional PDF of Q_t when Q is demand constrained, multiplied by the probability of being demand constrained *plus* (b) the PDF of Q_t when Q_t is supply constrained, multiplied by the probability of being supply constrained. Now:

$$f(Q_t|Y_t^D < Y_t^S) = \int_{Q_t}^{\infty} g(Q_t, Y_t^S|Y_t^D < Y_t^S)\,dY_t^S = [1/\Pr(Y_t^D < Y_t^S)]\int_{Q_t}^{\infty} g(Q_t, Y_t^S)\,dY_t^S \tag{2.60}$$

and similarly $f(Q_t|Y_t^S \leq Y_t^D)$ may be expressed as

$$[1/\Pr(Y_t^S \leq Y_t^D)]\int_{Q_t}^{\infty} g(Y_t^D, Q_t)\,dY_t^D$$

The PDF of Q_t may therefore be written as

$$h(Q_t) = \int_{Q_t}^{\infty} g(Q_t, Y_t^s)\,dY_t^s + \int_{Q_t}^{\infty} g(Y_t^d, Q_t)\,dY_t^d \tag{2.61}$$

the likelihood function may then be specified as

$$L = \prod_t h(Q_t) \tag{2.62}$$

'L' is a function of all the parameters of the system and the covariance matrix of the errors, u_{1t}, u_{2t} and u_{3t}.

ARCH and GARCH likelihood functions

Our general statement of the likelihood function of the non-linear model (2.7) assumed that the error terms had a constant covariance structure Θ. In fact it is not obvious that the covariance matrix will always be constant over time and it is easy to further generalise (2.7) to take account of this. If the covariance matrix is known over time, that is Θ_t is a known series, then Θ_t may simply be entered into (2.7). If Θ_t is assumed to vary over time but its value is unknown then the problem is more complex and we cannot simply estimate all the elements of Θ_t as there can never be sufficient degrees of freedom to allow this.

One approach to estimating Θ_t lies in a suggestion made by Engle (1982) to model the expected (or conditional) covariance matrix as a function of observed past squared errors, this model is termed the autoregressive conditional heteroskedasticity (ARCH) model. The basic assumption is that H_t is a conditional expectation of Θ_t based on past information, thus

$$H_t = E(\Theta_t:\Omega_{t-1}) \tag{2.63}$$

where Ω_{t-1} is the relevant known information set. The specific assumption of the ARCH model is that

$$\mathrm{Vech}(H_t) = \gamma_0 + \sum_{i=1}^{N} \gamma_i\,\mathrm{Vech}(e_{t-i}e'_{t-i}) \tag{2.64}$$

where Vech(H) denotes column-stacking the lower triangular elements of a symmetric matrix H and γ_0 and γ_1 are suitably dimensioned vectors of parameters. A scalar version of (2.64) is simply $h_t = \alpha_0 + \sum \alpha_i e_{t-i}^2$; the conditional variance h_t depends on past squared forecast errors. H_t may then be substituted into (2.7) in place of Θ to produce the ARCH likelihood function.

A further extension to the ARCH model is the generalised autoregressive conditional heteroscedasticity model (GARCH) (due to Bollerslev 1986) which basically allows other terms to enter the ARCH equation beyond simply the lagged errors. One particularly useful form of GARCH model is when the lagged conditional expectation of the covariance term enters the equation. In this case

$$\text{Vech}(H_t) = \gamma_0 + \sum_{i=1}^{P} \gamma_{1i}\,\text{Vech}(H_{t-i}) + \sum_{j=1}^{N} \gamma_{2j}\,\text{Vech}(e_{t-j}e'_{t-j}) \tag{2.65}$$

In its general form this would be termed a GARCH(N, P) model, denoting the number of lags in H and ee' which feature in the model. Once again H_t can simply be substituted into (2.7) to produce the GARCH likelihood function. A simple scalar version of the above is e_{t-1}^2, which would be a GARCH(1, 1) model.

A final further elaboration of this type of model is due to Engle, Lilien and Robins (1987) who point out that many theoretical models, especially in finance theory, include terms in 'risk' which can be modelled by including conditional elements of the covariance matrix into the specification of the model. Thus (2.6), the structural equations of the model may include any elements of H_t as 'risk' terms: $y_t = f(x, \beta) + \delta H_t + e_t$. When this is done the models are then generally termed ARCH-in-Mean (ARCH-M) or GARCH-in-Mean (GARCH-M) models.

2.4 Empirical applications using maximum likelihood

A discrete switching disequilibrium model of the market for building society loans

In this section we present an example of maximum likelihood estimation of a disequilibrium model for mortgage lending from building societies (taken from Hall and Urwin 1989). The model involves formulating equations for the demand and supply for mortgage lending and the determination of mortgage interest rates. The model is estimated on the assumption that the short side of the market dominates and uses the discrete switching model discussed above.

The demand for mortgages

The demand for mortgages may be derived from a fairly simple utility maximisation problem. Suppose a representative household has a utility function $U(H, G)$ where H is housing services and G is an aggregate of other goods (in real terms). The household maximises this function subject to a total limit on disposable income of the form:

$$g(r^m, P^H)H + GP = DY \tag{2.66}$$

Where $g(r^m, P^H)$ is a cost function of servicing a mortgage which will provide housing services H. The cost function depends on r^m the rate of interest on mortgages and P^H the price of houses. DY is (nominal) disposable income and P is the general price level (of goods). Maximising $U(H, G)$ subject to (2.66) yields a demand function of the form:

$$H = f(r^m, P^H, DY, P) \tag{2.67}$$

Hall and Urwin then relate the demand for mortgages (M^D) to this basic function by introducing the number of owner-occupied houses (NOH). They then invoke adjustment costs to introduce lagged actual mortgage borrowing and a term for the effects of banks moving into the mortgage market (ZBL). This then gives the general demand function:

$$\begin{aligned}\log(M^D/P) = {} & A_0 + A_1\log(r^m) + A_2\log(P^H/P)\\ & + A_3\log(NOH) + A_4\log(DY/P)\\ & + A_5\log(P) + A_6\Delta\log(P)\\ & + A_7\log(ZBL) + A_8\log(M/P)_{t-1}\end{aligned} \tag{2.68}$$

where $A_1, A_7 < 0$; $A_2, A_4, A_8 > 0$

The supply of mortgage lending The supply of mortgages depends on two main factors. First, the supply of building society shares and deposits (primarily) from the personal sector and second the action of the building society when it carries out its role as an intermediary between depositors and lenders.

The supply of deposits is given by a fairly simple application of portfolio theory. The supply of deposits to building societies is given by the demand function of the personal sector for building society deposits. Deposits will therefore vary with income and relative returns between building society deposits and other assets (r^D/r^1).

They then introduce terms in the loan to value ratio of first time buyers (LV) and the loan to income rate of first time buyers (LY) as proxies for the willingness of societies to lend. To capture changes in the supply of mortgages they introduce a term for building societies borrowing in the wholesale money markets (ZWB). Lags are introduced to model adjustment costs giving the supply equation:

$$\log(M^S/P) = B_0 + B_1 \log(r_D/r^1) + B_2 \log(DY/P) + B_3 \Delta \log(P) + B_4 \log(LV) + B_5 \log(LY) + B_6 \log(ZWB) + B_7 \log(M/P)_{t-1} \quad \textbf{(2.69)}$$

where $B_1, B_2, B_5, B_6, B_7 > 0$

The interest rate adjustment equation Finally, in order close the model we need an interest-rate adjustment equation. We assume the change in $\log(r_D/r^1)$ is a function of excess demand or supply to which is added a set of other relevant interest rates. This part of the model is really of only minor interest. A simple '*ad hoc*' equation involves the change in the long-term consol rate (20*YC*), the change in the treasury bill yield (r^1) and a lagged dependent variable:

$$\Delta \log(r^D/r^1) = C_0 + C_1 \Delta \log(20YC) + C_2 \Delta \log(r^1) + C_3 \Delta \log(r^D/r^1)_{t-1} + C_4 \log(M^D/M^S) \quad \textbf{(2.70)}$$

Estimation of the model

The likelihood function for the discrete switching disequilibrium model is an extremely complex one. It is not available as part of any of the standard econometric computer programs and it is sufficiently ill-conditioned to present serious problems for any of the standard numerical maximisation procedures. Numerical optimisation is therefore achieved by the combined use of a non-linear simplex algorithm and a conventional quasi-Newton algorithm using analytical first derivatives. The non-linear simplex algorithm is used first as it is relatively robust to the presence of local maxima and discontinuities; its final convergence on the maximum point is, however, slow. When we are close to the maximum the quasi-Newton algorithm takes over the optimisation problem from the simplex procedure and it then efficiently pinpoints the true maximum. Verifying that a true maximum has actually been located is of course difficult. One check is to use a graphical search around the final solution, resulting in a set of line searches across the likelihood space, and these may be used to detect a failure to find a true maximum.

We outlined above the general form of the model to be estimated, but there is of course scope within this general framework for a wide range of dynamic specifications. In a 'general-to-specific' modelling exercise we start from a general model and 'test down' on the dynamics until a parsimonious form of the model is achieved. This is not a practical procedure for this type of system estimation as the general form would involve far too many parameters for successful optimisation. Even in the final form to be reported here the model involved maximising the likelihood function with respect to 26 parameters. The estimation procedure is therefore less systematic than one might like. It is also worth pointing out that the standard battery of diagnostic test procedures on the error process are not applicable to this model. The reason for this is that the observed error, $Q - \hat{Q}$ (where Q is the traded quantity of mortgages), cannot be uniquely associated with any of the structural error terms in the model. The observed error will be a combination of the errors on the notional supply and demand curves and as such it provides no formal evidence about the properties of the structural errors. We do not make the assumption that $Q - \hat{Q}$ is white noise and uncorrelated and so there is no point in testing this assumption. Residual tests may however be constructed in a number of complex ways, see Hall, Henry and Pemberton (1991).

Maximising the log likelihood then produces the results detailed in Table 2.1 which gives the parameter estimates of the preferred model

Table 2.1 Parameter estimates for the model (Asymptotic *t* statistics in parenthesis)

	Model 1		Model 2		Model 3	
A_0	−6.86	(5.1)	−6.92	(5.8)	−1.13	(5.5)
A_1	−0.045	(8.1)	−0.046	(8.8)	−0.08	(8.3)
A_2	0.03	(2.4)	0.03	(2.9)	0.005	(0.34)
A_3	0.75	(5.2)	0.78	(6.1)	–	
A_4	0.06	(1.8)	0.04	(1.7)	0.11	(3.4)
A_5	−0.12	(5.2)	−0.12	(6.0)	0.005	(0.6)
A_6	−0.79	(8.1)	−0.82	(10.6)	−0.73	(5.6)
A_7	−0.087	(3.7)	−0.09	(3.9)	−0.07	(1.6)
A_8	0.94	(23.1)	0.95	(27.7)	1.01	(85.9)
B_0	−1.03	(4.0)	−1.08	(5.2)	−0.99	(5.1)
B_1	0.003	(0.7)	0.002	(0.4)	0.003	(1.2)
B_2	1.10	(3.9)	0.13	(4.4)	0.11	(2.0)
B_3	−1.1	(16.7)	−1.1	(18.6)	−1.1	(2.0)
B_4	0.11	(3.6)	0.10	(4.2)	0.09	(1.9)
B_5	0.06	(3.2)	0.06	(3.5)	0.06	(2.9)
B_6	0.38	(2.3)	0.54	(2.2)	0.82	(1.9)
B_7	0.91	(44.1)	0.88	(35.0)	0.91	(16.9)
C_0	0.007	(0.9)	0.007	(0.8)	0.007	(0.9)
C_1	0.35	(2.2)	0.42	(3.0)	0.36	(2.2)
C_2	−0.95	(14.3)	−0.97	(15.4)	−0.955	(15.1)
C_3	−0.13	(2.1)	−0.06	(1.11)	−0.13	(2.2)
C_4	0.0004	(0.00002)	0.00004	(0.0)	0.00003	(0.0)
$\sigma(Q - \hat{Q})$	0.0031		0.0026		0.0034	
Likelihood	510.006		462.03		505.04	
Data period	6902–8601		6902–8401		6902–8601	

(model 1). This model estimated excluding the last eight data points which are then used the test structural stability (model 2). In model 3 we combine the terms in the stock of housing and the price of housing and use the *value* of owner-occupied housing. The term $\sigma(Q - \hat{Q})$ is the standard error of the observed forecast of the model which may be compared with the standard error of the Anderson and Hendry (1984) model of 0.0029 and the Wilcox (1985) model of 0.0029.

The preferred model 1 conforms with our prior views about the signs of the parameters. It produces a model which tracks the data reasonably well even in comparison with conventional OLS models. This is indicated by the standard deviation of the observed error which is of a size similar to that found in other studies of mortgage lending (although the data periods are quite different). The tendency of the models to move towards equilibrium is measured by the size of C_4, ($C_4 = 0$ implies equilibrium is never reached, $C_4 = \infty$ implies continuous market clearing) this parameter estimate suggests that there is only very slow adjustment and that for practical purposes disequilibrium may persist indefinitely. This conforms well with the conventional view of building society prior to 1986. Nevertheless the market for mortgages is not characterised by a very large degree of disequilibrium. Figure 2.2 shows the model's forecast for the stock of

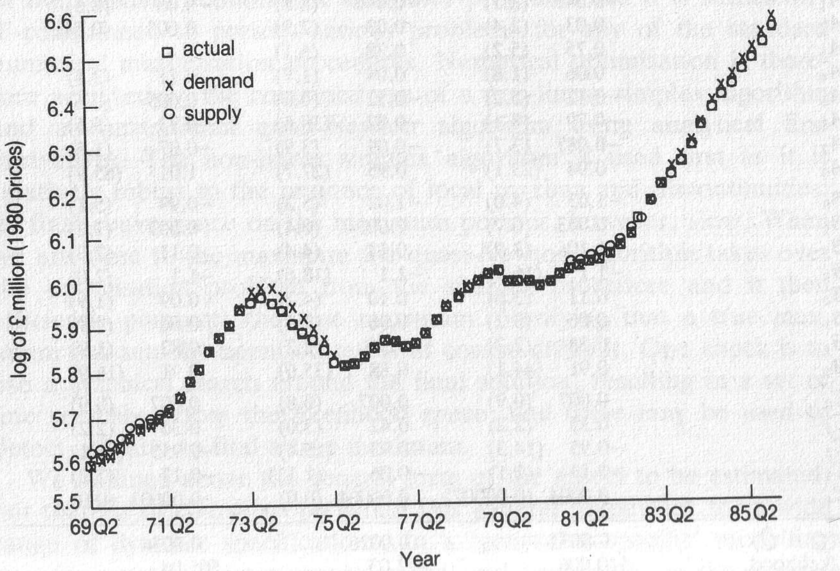

Figure 2.2 Demand and supply of mortgages.

mortgage demand and supply in contrast with the actual level of lending. It is quite clear from this figure that, by and large, the building societies are able to equate the demand and supply of mortgages fairly effectively.

However, this is not to suggest that disequilibrium is insignificant in this market. Figure 2.3 shows the deterministic model estimates of excess demand over the period 1969 Q2–1986 Q1. The degree of disequilibrium peaks in 1974 at around 4% of the mortgage stock. This represents a sizeable constraint on households borrowing. For example, in 1985 this would have implied a constraint in excess of £1,000 million. The overall pattern of excess demand corresponds remarkably closely with that estimated by Wilcox, although this model does not detect such strong excess demand in the period 1979–80. Unfortunately there is no time series available for the size and duration of mortgage queuing to compare with Figure 2.3.

Figure 2.3 also suggests that the incursions into the mortgage market of non-building society lenders, particularly the banks, have had a very significant impact on either the degree of excess supply or demand. The three periods (the start of the 1970s, 1981–3 and 1986–7) in which the banks' market share rose very rapidly are estimated to have been those in which the extent of rationing fell substantially, or even that conditions of excess supply prevailed. The

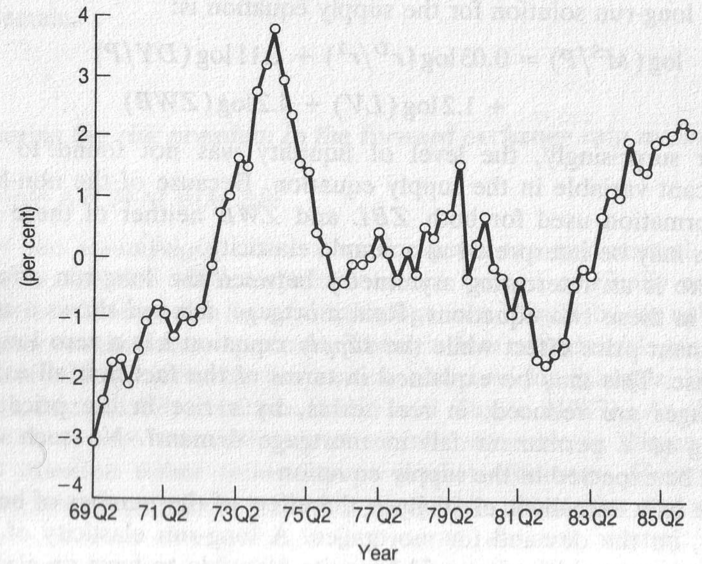

Figure 2.3 Excess demand for mortgages.

fact that lenders did not attempt to reduce their lending standards in order to eliminate excess supply suggests that competitive forces have had a relatively weak impact on such standards and that lenders have attempted to retain the appropriate prudential criteria. It is perhaps surprising that the degree of rationing is estimated to have been greater in 1984–5 than in the second half of the 1970s, as in the later period building societies were thought to have adopted a more flexible interest-rate policy. While this finding may not be consistent with general perceptions of the way of the mortgage market operated at that time, the results do indicate that over the period as a whole, societies' propensity to use interest rates to equilibriate the demand for and supply of mortgage was greater when competitive pressures were more intense.

The long-run properties of the demand and supply equations are fairly sensible. The long-run solution to the demand equation is:

$$\log(M^D/P) = -0.75\log(r^m) + 0.5\log(P^H/P) + 12.5\log(NOH) + 1.0\log(DY/P) - 2.0\log(P) - 1.4\log(ZBL)$$

These parameter estimates are all quite reasonable with the possible exception of the elasticity on the number in owner-occupied housing, NOH, which will be discussed further below.

The long-run solution for the supply equation is:

$$\log(M^S/P) = 0.03\log(r^D/r^1) + 1.11\log(DY/P) + 1.2\log(LV) + 4.2\log(ZWB)$$

Rather surprisingly, the level of liquidity was not found to be a significant variable in the supply equation. Because of the non-linear transformation used for both ZBL and ZWB neither of these coefficients may be interpreted as a simple elasticity.

There is an interesting asymmetry between the long-run effect of prices in these two equations. Real mortgage *demand* shows a strong permanent price effect while the *supply* equation has a zero long-run response. This may be explained in terms of the fact that all existing mortgages are reduced, in real terms, by a rise in the price level leading to a permanent fall in mortgage demand. No such effect would be expected in the supply equation.

The only unrealistic elasticity is the effect of the number of houses, NOH, on the demand for mortgages. A long-run elasticity of 12 is clearly unreasonable. It would be quite plausible to have an elasticity greater than one and we would certainly expect the elasticity on the

number of houses to be larger than that on house prices, as almost all houses which are additions to the owner-occupied stock have associated mortgages. None the less, a long-run figure of 12 is clearly implausible. There would seem to be two possible explanations. First, we may have failed to pick up the full dynamic effect and so we may have a plausible short-run effect from housing but a very poorly defined long-run. Second, there may be a trend factor in mortgage demand which we have failed to model but which is highly collinear with the housing stock. In this case, part of our long-run effect on NOH, may be due in part to this unidentified component.

In an attempt to investigate these possibilities we performed a number of experiments. First, lags in the housing stock were introduced to allow for the possibility of more complex dynamics. This did not change the long-run elasticity on NOH, to any great extent. However, it is possible that more complex dynamics are required but our data, which span less that 20 years, are simply not long enough to analyse a market where the average term to maturity of loans is about 7 years. Second, model 3 in Table 2.1 considers the effect of restricting the housing terms to be the *value* of the owner-occupied housing stock. This restriction when applied to the model has a number of undesirable features. In particular, the demand equation is dynamically unstable and so the long-run solution is no longer defined. We therefore conclude that this real-world application of the SMDM has yielded useful insights but clearly specification problems still remain.

Measuring the risk premium in the forward exchange rate market

A simple ARCH-M example

Under the assumptions that economic agents are risk neutral, there are no transaction costs, expectations are formed rationally and the market is efficient, the forward exchange rate should be an unbiased predictor for the future spot rate. There is now considerable empirical evidence which rejects this proposition however; for example, Hansen and Hodrick (1980), Hakkio (1981) and Taylor (1987). As the assumption of zero transaction costs seems reasonable in this case, we might question either rationality or market efficiency. However in neither case do we have a readily acceptable alternative and clearly another extreme assumption is that agents are risk neutral. Much recent work has concentrated on a search for the 'risk premium' such as Frankel (1982), Domowitz and Hakkio (1985), Fama (1984), Hodrick and Srivastava (1984), Nelson (1985), and Taylor (1987). An important element of this research has been the recognition that, in principle, the risk premium will vary over time depending on the degree of uncertainty in the system.

Our example will assume that the existence of a risk premium causes a differential between the forward rate and the expected future spot rate in accordance with a simple ARCH-M model.

If the log of the forward exchange rate i periods ahead is denoted as f_{t+i}, and the log of the spot exchange rate is denoted s_t then the risk premium ρ_t may be defined as

$$\rho_t = f_{t+i} - s^e_{t+i} \tag{2.71}$$

where s^e_{t+i} is the market expectation of the spot exchange rate at period $t+i$, based on information at time $t(\Omega_t)$. Under the rational expectations hypothesis

$$s^e_{t+i} = E(s_{t+i}|\Omega_t) = s_{t+i} + \varepsilon_{t+i} \tag{2.72}$$

where ε_{t+i} is the RE forecast error, hence

$$\rho_t = f_{t+i} - (s_{t+i} + \varepsilon_{t+i}) \tag{2.73}$$

A formal derivation of the risk premium is complex and will not be given here (see Grauer, Litzenberger and Stehlf 1976, or Stockman 1978). The important feature of the derivation which holds irrespective of the specific form of the model, is that the risk premium is determined by the degree of risk aversion of the market agents and the variances and covariances of the assets in the system. It is perhaps reasonable to assume that the degree of agents risk aversion is constant but the idea that *uncertainty* about asset returns and in particular exchange rate movements is constant is rather hard to accept.

As a simple first step towards recognising the importance of the time-varying nature of the risk premium, suppose that ε_{t+i} has zero mean but a time varying conditional variance, so that $\varepsilon_t \sim N(0, h_t)$. Agents form an expectation of the variance, h_t based on available information. A simple assumption is that the risk premium ρ_t is positively related to the conditional variance of the RE forecast errors $\rho_t = A_0 + A_1 h_t$. Using (2.73) we have,

$$(f_{t+i} - s_{t+i}) = A_0 + A_1 h_t + \varepsilon_t \tag{2.74}$$

This is a simple ARCH-M model. To complete the model we need to make h_t an explicit function of the information set, again a simple approach is to assume that the expected variance is a linear function of recent lagged squared errors:

$$h_t = B_0 + B_1\left(\sum C_i \varepsilon^2_{t-i}\right) \tag{2.75}$$

Then, conditional on the initial values of the data, the log likelihood function may be expressed as

$$\log(L) = \sum_{t=1}^{T} (-\log h_t - \varepsilon^2_t / h_t) \tag{2.76}$$

In order to simplify the estimation we assume that the weights C_i, decline linearly over eight months to zero and this leaves four parameters to be estimated: B_0, B_1, A_0, A_1. Note that if $B_0 \neq 0$ and $B_1 = 0$ then the risk premium is not time varying, so this model has the constant risk premium as a special case.

Estimation may be carried out using a numerical maximisation technique as described above, t statistics may be derived for the parameters of the system from the inverse of the Hessian of the likelihood function and standard likelihood ratio tests may be used to test special versions of the model.

Estimation results

The model outlined above was estimated using monthly data from 1973 M2 to 1987 M6 for the sterling-dollar spot rate and three-month forward rate. The parameter estimates are given below

$A_0 = 0.034$ (1.63)

$A_1 = -7.655$ (1.61)

$B_0 = 0.002$ (5.01)

$B_1 = 0.431$ (3.30)

Log likelihood = −806.22

$SEE(\varepsilon_t) = 0.059$

Normality test (see Chapter 4)$\chi^2(2) = 0.85$

The coefficient B_1 has a reasonable size and sign and is significantly different from zero, which suggests that there is an important ARCH component to the error process. The coefficients A_0 and A_1 both have sensible magnitudes but are not strictly significant; this suggests that either term may be dropped from the model and hence the risk premium may not be time varying. On balance there is *weak* evidence in favour of a time-varying risk premium (i.e. $A_1 \neq 0$) and clearly there is an ARCH process in the error term ($B_1 \neq 0$). This suggests that a more complex relationship determining the conditional variance is required, perhaps of the form

$$h_t = B_0 + B_1 h_{t-1} + B_2 \varepsilon_{t-1}^2 + B_3 Z_t \quad (2.77)$$

This GARCH(1, 1) process allows 'shocks' ε_{t-1} to have an impact on the conditional variance h_t in all future periods (but with declining weights). Z_t consists of other information which might influence the conditional variance (such as domestic and foreign interest rates or current account factors).

Although ARCH and GARCH type models have proved useful in modelling time-varying risk premia in financial markets (e.g. Chou 1988, or Hall *et al*. 1989), nevertheless the exact formulation of the ARCH equation (2.77) is often not well based in a formal framework where agents optimise some explicit objective function.

2.5 Summary

We have explained the basis of maximum likelihood estimation and discussed testing using the likelihood ratio test, the Wald test and the LM test. We have outlined several numerical optimisation techniques and demonstrated how certain 'non-standard' models may be examined in the maximum likelihood framework.