CHAPTER 2

2. Maximum Likelihood Estimation

Suppose we are interesting in estimating the parameters of a model with GARCH(p,q) disturbances. Let the conditional mean be

$$y_t = x_t' b + u_t$$

Here x_t denotes a vector of explanatory variables, which could include lagged values of y. The disturbances u_t are assumed to be:[1]

$$u_t \mid I_{t-1} \sim N(0, h_t)$$
$$h_t = v_t' \omega = a_0 + \sum_{i=1}^{q} a_i u_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}$$

Let $v_t' = (1, u_{t-1}^2, \ldots, u_{t-q}^2, h_{t-1}, \ldots, h_{t-p})$, $\omega' = (a_0, a_1, \ldots, a_q, \beta_1, \ldots, \beta_p)$, $\theta = (b', \omega')$ and $\theta \in \Theta \subseteq R^m$. Under normality the conditional distribution of y_t is Gaussian with mean $x_t' b$ and variance h_t:

$$f(y_t \mid I_{t-1}) = \frac{1}{\sqrt{2\pi h_t}} \exp\left(\frac{-(y_t - x_t' b)^2}{2h_t} \right).$$

The sample log likelihood function for a sample of T observations is:

$$L_T(\theta) = \sum_{t=1}^{T} \log(f(y_t \mid I_{t-1}; \theta)) =$$
$$= \frac{-T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \log(h_t) - \frac{1}{2} \sum_{t=1}^{T} \frac{(y_t - x_t' b)^2}{h_t}.$$

The first and second derivatives of the log of the conditional likelihood of the t^{th} observation with respect to the variance parameters are:

[1] I_{t-1} denotes any information available at time $t-1$

$$\frac{\partial L_t}{\partial \omega} = \frac{\partial \log(f(y_t \mid I_{t-1};\theta))}{\partial \omega} = \frac{1}{2h_t}\frac{\partial h_t}{\partial \omega}\left(\frac{u_t^2 - h_t}{h_t}\right),$$

$$\frac{\partial^2 L_t}{\partial \omega \partial \omega'} = \frac{\partial^2 \log(f(y_t \mid I_{t-1};\theta))}{\partial \omega \partial \omega'} = \left(\frac{u_t^2 - h_t}{h_t}\right)\frac{\partial}{\partial \omega'}\left[\frac{1}{2h_t}\frac{\partial h_t}{\partial \omega}\right] - \frac{1}{2h_t^2}\frac{\partial h_t}{\partial \omega}\frac{\partial h_t}{\partial \omega'}\frac{u_t^2}{h_t},$$

where

$$\frac{\partial h_t}{\partial \omega} = v_t + \sum_{i=1}^{p} \beta_i \frac{\partial h_{t-1}}{\partial \omega}.$$

The Information matrix corresponding to ω is given as:

$$I_{\omega\omega} = \frac{-1}{T}\sum_{t=1}^{T}\left(E\left(\frac{\partial^2 L_t}{\partial \omega \partial \omega'}\right)\right) = \frac{1}{2T}\sum_{t=1}^{T}\left(h_t^{-2}\frac{\partial h_t}{\partial \omega}\frac{\partial h_t}{\partial \omega'}\right).$$

The first and second derivatives of the log of the conditional likelihood of the tth observation with respect to the mean parameters are:

$$\frac{\partial L_t}{\partial b} = \frac{\partial \log(f(y_t \mid I_{t-1};\theta))}{\partial b} = u_t x_t h_t^{-1} + \frac{1}{2}h_t \frac{\partial h_t}{\partial b}\left(\frac{u_t^2 - h_t}{h_t}\right),$$

$$\frac{\partial^2 L_t}{\partial b \partial b'} = \frac{\partial^2 \log(f(y_t \mid I_{t-1};\theta))}{\partial b \partial b} =$$

$$= -h_t^{-1}x_t x_t' - \frac{1}{2}h_t^{-2}\frac{\partial h_t}{\partial b}\frac{\partial h_t}{\partial b'}\left(\frac{u_t^2}{h_t}\right) - 2h_t^{-2}u_t x_t \frac{\partial h_t}{\partial b} + \left(\frac{u_t^2 - h_t}{h_t}\right)\frac{\partial}{\partial b'}\left[\frac{1}{2}h_t^{-1}\frac{\partial h_t}{\partial b}\right],$$

where

$$\frac{\partial h_t}{\partial b} = -2\sum_{i=1}^{q} a_i x_{t-i} u_{t-i} + \sum_{j=1}^{p} \beta_j \frac{\partial h_{t-j}}{\partial b}.$$

The Information matrix corresponding to b is given as:

$$I_{bb} = \frac{-1}{T}\sum_{t=1}^{T}\left(E\left(\frac{\partial^2 L_t}{\partial b \partial b'}\right)\right) = \frac{1}{T}\sum_{t=1}^{T}\left(h_t^{-1}x_t x_t' + 2h_t^{-2}\sum_{i=1}^{q} a_i^2 u_{t-i}' x_{t-i} x_{t-i}' u_{t-i} + \frac{1}{2}\sum_{j=1}^{p}\beta_j^2\left(\frac{\partial h_{t-j}}{\partial b}\right)^2\right).$$

The elements in the off-diagonal block in the information matrix are zero:

$$I_{\omega b} = \frac{-1}{T}\sum_{t=1}^{T}\left(E\left(\frac{\partial^2 L_t}{\partial \omega \partial b'}\right)\right) = 0$$

The likelihood function can be maximized numerically using the BHHH algorithm (Berndt et al. (1974)[1]). Let $\theta^{(i)}$ denote the parameter estimates after the i[th] iteration. $\theta^{(i+1)}$ is the calculated from:

$$\theta^{(i+1)} = \theta^{(i)} + \lambda_i \left(\sum_{t=1}^{T} \frac{\partial L_t}{\partial \theta} \frac{\partial L_t}{\partial \theta'} \right)^{-1} \sum_{t=1}^{T} \frac{\partial L_t}{\partial \theta},$$

where $\partial L_t / \partial \theta$ is evaluated at $\theta^{(i)}$ and λ_i is a variable step length chosen to maximize the likelihood function in the given direction. The maximum likelihood estimate $\hat{\theta}_T$ is strongly consistent for θ_0 and asymptotically normal with mean θ_0 and covariance matrix $T^{-1}F^{-1} = -T^{-1}E\left(\frac{\partial^2 L_t}{\partial\theta\partial\theta'}\right)^{-1} = T^{-1}E\left(\frac{\partial L_t}{\partial\theta}\frac{\partial L_t}{\partial\theta'}\right)^{-1} = \left(\sum_{t=1}^{T}\frac{\partial L_t}{\partial\theta}\frac{\partial L_t}{\partial\theta'}\right)^{-1}$ from the last (T[th]) BHHH iteration:

$$\sqrt{T}\left(\hat{\theta}_T - \theta_0\right) \xrightarrow{D} N\left(0, F^{-1}\right).$$

Note that when the outer product is near singular we may use a ridge correction in order to handle numerical problems and improve the convergence rate. The Marquardt algorithm modifies the BHHH algorithm by adding a correction matrix to the sum of the outer product of the gradient vectors. The Marquardt updating algorithm is given by:

$$\theta^{(i+1)} = \theta^{(i)} + \lambda_i \left(\sum_{t=1}^{T} \frac{\partial L_t}{\partial \theta} \frac{\partial L_t}{\partial \theta'} - aI \right)^{-1} \sum_{t=1}^{T} \frac{\partial L_t}{\partial \theta},$$

where I is the identity matrix and a is a positive number (chosen by the algorithm). The effect of this modification is to push the parameter estimates in the direction of the gradient vector. The idea is that when we are far from the maximum, the local quadratic approximation to the function may be a poor guide to its overall shape, so we may be better off simply following the gradient. The correction may provide better performance at locations far from the optimum, and allows for computation of the direction vector in cases where the Hessian is near singular.

[1] The BHHH algorithm is similar to Newton-Raphson algorithm, but replaces the negative of the Hessian (second derivative of the log likelihood function with respect to the vector of unknown parameters) by an approximation formed from the sum of the outer product of the gradient vectors for each observation's contribution to the objective function. This approximation is asymptotically equivalent to the actual Hessian when evaluated at the parameter values, which maximize the function.

2.1 Maximum Likelihood Estimation under non-normality

The standard ARCH models assume that the disturbances of the model when divided by their true conditional standard deviation are standard normal variables:

$$z_t \equiv u_t \big/ \sqrt{h_t}$$

However the unconditional distribution of many financial time series seems to have fatter tails than allowed by the Gaussian distribution. Some of this can be explained by the presence of ARCH model. As we have already stated, even if z_t has a normal distribution, the unconditional distribution of u_t is non-normal with heavier tails than a normal distribution. Even so, there is a fair amount of evidence that the conditional distribution of u_t is often non-normal as well. Thus, Bollerslev (1987) proposed that z_t is drawn from a t-distribution with n degrees of freedom, where n is regarded as parameter to be estimated by maximum likelihood. The same approach is used with other distributions for z_t. Other distributions that have been employed are the Generalized Error distribution of Nelson (1991), the Normal-Poisson mixture distribution of Jorion (1988), the Generalized t-distribution of Bollerslev, Engle and Nelson 1994), the Power Exponential distribution of Baillie and Bollerslev (1989), the normal-log normal mixture of Hsieh (1989) and others.

2.2 Quasi Maximum Likelihood Estimation

Bollerslev and Wooldridge (1992) showed that the maximization of the Normal log likelihood function can provide consistent estimates of the parameter vector ω even when the distribution of u_t in non normal, provided that

$$\begin{aligned} u_t &= \sqrt{h_t} z_t \\ E(z_t \mid I_{t-1}) &= 0\,. \\ E(z_t^2 \mid I_{t-1}) &= 1 \end{aligned}$$

However, the standard errors have to be adjusted. Let $\hat{\theta}_T$ be the estimate that maximizes the normal log likelihood and let θ_0 be the true value that characterizes the linear representations:

$$
\begin{aligned}
y_t &= x_t' b + u_t \\
u_t \mid I_{t-1} &\sim N(0, h_t) \\
h_t &= v_t'\omega = a_0 + \sum_{i=1}^{q} a_i u_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}
\end{aligned}
$$

Then even when the z_t is non normal, under certain regularity conditions:

$$\sqrt{T}\left(\hat{\theta}_T - \theta_0\right) \xrightarrow{D} N\left(0, D^{-1} S D^{-1}\right),$$

where

$$S = \underset{T\to\infty}{p\lim}\, T^{-1} \sum_{t=1}^{T} \left(\frac{\partial L_t}{\partial \theta}\right)\left(\frac{\partial L_t}{\partial \theta}\right)',$$

$$D = \underset{T\to\infty}{p\lim}\, T^{-1} \sum_{t=1}^{T} -E\left[\frac{\partial^2 L_t}{\partial\theta\partial\theta'} \mid I_{t-1}\right].$$

The matrix S can be consistently estimated by:

$$\hat{S}_T = T^{-1} \sum_{t=1}^{T} \left(\left.\frac{\partial L_t}{\partial \theta}\right|_{\theta=\hat{\theta}_T}\right)\left(\left.\frac{\partial L_t}{\partial \theta}\right|_{\theta=\hat{\theta}_T}\right)',$$

and the matrix D can be consistently estimated by:

$$\hat{D}_T = T^{-1} \sum_{t=1}^{T} \left\{ \frac{1}{2} \hat{h}_t^{-2} \begin{bmatrix} -2\sum_{i=1}^{q} \hat{a}_i \hat{u}_{t-i} x_{t-i} + \sum_{j=1}^{p} \hat{\beta}_j \left.\frac{\partial h_{t-j}}{\partial b}\right|_{b=\hat{b}} \\ \hat{v}_t + \sum_{j=1}^{p} \hat{\beta}_j \left.\frac{\partial h_{t-j}}{\partial b}\right|_{b=\hat{b}} \end{bmatrix} \times \begin{bmatrix} -2\sum_{i=1}^{q} \hat{a}_i \hat{u}_{t-i} x_{t-i}' + \sum_{j=1}^{p} \hat{\beta}_j \left.\frac{\partial h_{t-j}}{\partial b}\right|_{b=\hat{b}} & \hat{v}_t' + \sum_{j=1}^{p} \hat{\beta}_j \left(\left.\frac{\partial h_{t-j}}{\partial b}\right|_{b=\hat{b}}\right)' \end{bmatrix} + \hat{h}_t^{-1} \begin{bmatrix} x_t x_t' & 0 \\ 0 & 0 \end{bmatrix} \right\}.$$

Standard errors for $\hat{\theta}_T$ that are robust to misspecification of the family of densities can thus be obtained from the square root of diagonal elements of

$$T^{-1} \hat{D}_T^{-1} \hat{S}_T \hat{D}_T^{-1}.$$

Recall that if the model is correctly specified so that the data were really generated by a Gaussian model, then $S = D$ and this specifies to the usual asymptotic variance matrix for maximum likelihood estimation:

$$T^{-1}\hat{S}_T^{-1}.$$