Chapter 8 Risk Measurement with High Frequency Data

April 21, 2014

Contents

1. Introduction

Once weigths of different assets on a given portfolio are chosen (possibly determined by the pattern of long-run predictability of returns) a relevant issue is the measurement of risk associ-

ated to the specific asset allocation. If strategic asset allocation depends on the low-frequency properties of the data, risk measurement depends on their high frequency fluctuations. Can econometrics be fruitfully used at high frequency to assess the risk of a given portfolio? Risk is commonly measured through VaR. The VaR of a position is the percentage loss obtained with a probability at most of α percent:

$$\Pr\left(R^p < -VaR_\alpha\right) = \alpha.$$

Econometrics is useful to compute VaR in that it is useful to predict the distribution of returns at high frequency. We already know that the best forecast for high frequency mean returns is zero, in this chapter we are going to argue that volatilities are predictable at high frequency and that econometric models of volatilities can be fruifully put at work to compute Value at Risk. The presence of conditional heteroskedastic patterns in financial returns is also intimately related to the fact that there is overwhelming evidence that the (unconditional) distribution of realized returns on most assets (not only stocks and bonds, but also currencies, real estate, commodities, etc.) tends to display considerable departures from the classical normality assumption. We shall document that conditional heteroskedasticity implies that the unconditional, long-run distribution of asset returns is non-normal.[1] This is well-known to be potentially responsible for strong departures of observed derivative prices from simple but still widely employed pricing frameworks that are built on the classical results by Black and Scholes (1973) that rely on normality of financial returns. However, one should not be mislead by the naive notion that because second moments change over time, this implies that the time series process characterized by such moments becomes "wild", in the sense of being non-stationary. On the contrary, under appropriate technical conditions, one can prove that even though the conditional variance may change in heteroskedastic fashion, the underlying time series process may still be stationary.[2] In practice, this means that even though the variance of a series may

[1] We shall define the technical terms later on, but for the time being, the unconditional distribution of a time series process is the overall, long-run distribution of the data generated by the process. Drawing on one familiar example, if $X_{t+1} = \phi X_t + \epsilon_{t+1}$ with $\epsilon_{t+1} \sim \mathcal{N}(0,1)$, it is clear that the conditional distribution of X_{t+1} at time t (i.e., given information observed at time t) is $\mathcal{N}(\phi X_t, 1)$; however, in the long-run, when one averages over infinite draws from the process, because (under stationarity, i.e., $|\phi| < 1$) $E[X_{t+1}] = 0$ and $Var[X_{t+1}] = 1/(1-\phi^2)$, you know already that $X_{t+1} \sim \mathcal{N}(0, 1/(1-\phi^2))$ so that conditional and unconditional distributions will differ unless $\phi = 0$.

[2] Heuristically, stationarity of a stochastic process $\{X_t\}$ means that for every $k \geq 0$, $\{X_t\}_{t=k}^{\infty}$ has the same distribution as $\{X_t\}_{t=1}^{\infty}$. In words, this means that whatever is the point from which one starts sampling a time series process, the resulting overall (unconditional) distribution is unaffected by the choice: under stationarity, the implied distribution of returns over the last 20 years is the same as the distribution over 20 years of data to be sampled 10 years from now, say. Intuitively, this is related to the concept that a stationary time series will display "stable" long-run statistical properties—as summarized by its *unconditional distribution*—over time. Here the opposition between the unconditional natural of a distribution and time-varying *conditional variance* is important.

go through high and low periods, the unconditional (long-run, average) variance may still exist and be actually constant.[3] We shall start by describing the main features of financial data at high frequency, we shall then introduce econometric modelling of volatilities, next we illustrate how models for volatility can be mapped in VaR, finally we discuss backtesting of VaR models.

2. **The Evidence from high frequency data**

2.1. *Correlation and Heteroskedasticity*

Consider the daily data from a portfolio equally invested in the German the UK and the US stock market index (this is a possible outcome of the solution of an asset allocation problem over a long horizon) Figure 1 plots the time series of the daily returns over the period Even though this is not among the practice time series to be used in this class, the series is similar to the typical ones that appear in most textbooks.[4] Figure 1 plots the time series of the portfolio

[3]However, if the unconditional variance of a time series is not constant, then the series is non-stationary.

[4]Do not worry: we shall take care of examining your typical class data during your MATLAB sessions as well as at the end of this chapter.

returns and its components.

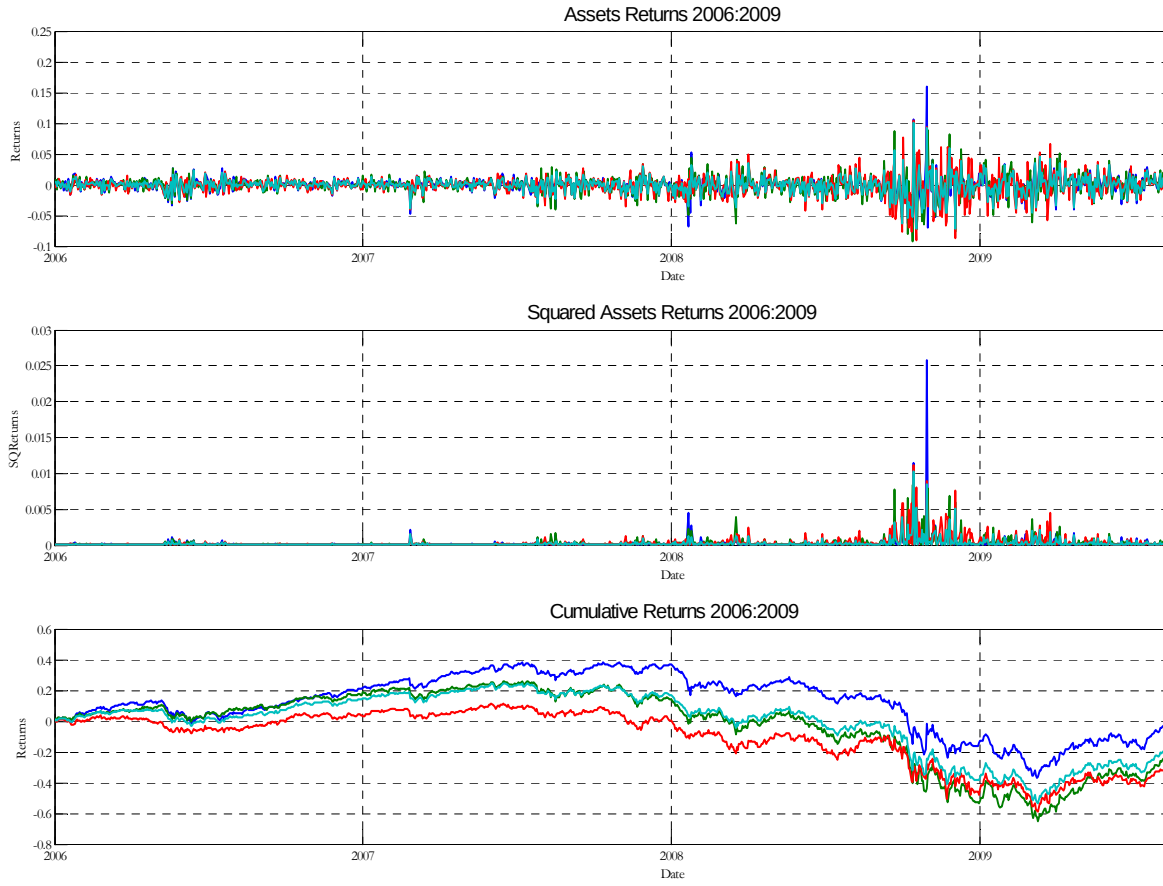


Figure 1: Daily Stock Returns

Visibly, volatility "clusters" in time: high (low) volatility tends to be followed by high (low) volatility. Casual inspection does have its perils, and formal testing is necessary to substantiate any first impressions. In fact, our objective in this chapter is to develop models that can fit this typical sequence of calm and turbulent periods. And especially forecast them.

Let's now examine the autocorrelogram of the series. Figure 2 shows the autocorrelogram functions of returns and squared returns (ACF) together with their asymptotical critical values. As you would expect of a series sampled at a relatively high frequency (such as monthly), there is weak serial correlation in U.S. stock returns. This lack of correlation means that, given past returns, the forecast of today's expected return is unaffected by knowledge of the past. However, more generally, the autocorrelation estimates from a standard ACF can be used to test the hypothesis that the process generating observed returns is a series of independent and identically distributed (IID) variables. The asymptotic (also called Bartlett's) standard error of

the autocorrelation estimator is approximately $1/\sqrt{T}$, where T is the sample size. In table 1, such a constant $\pm 2/\sqrt{T}$ 95% confidence interval boundary is represented as the blue horizontal lines that surround the red vertical lines that represent the sample autocorrelation estimates.[5]

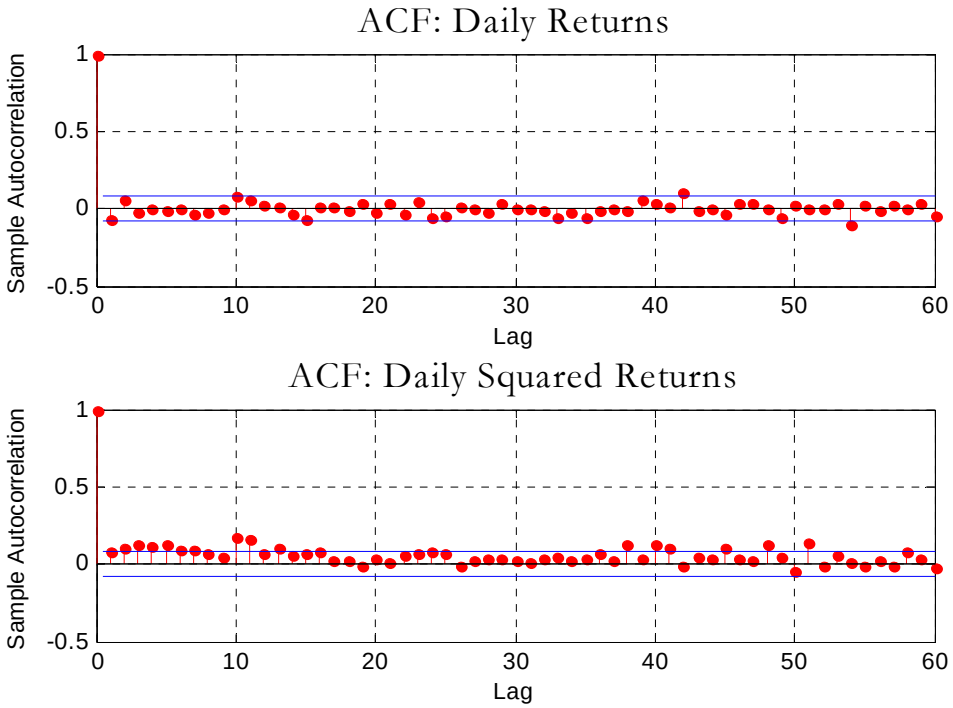


Figure 2: Serial correlation properties of daily stock returns and sqaured returns

Visibly, the autocrrelation function for returns lies consistently within the 95% confidence interval; the null hypothesis of absence of correlation cannot be rejected. This evidence does not apply to squared returns that shows much stronger sign of persistence with several violations of the 95 per cent confidence bounds. However, the absence of serial correlation is insufficient to establish independence.[6]

The independence hypothesis can more formally be tested using the Portmanteau Q-statistic

[5]To be precise, the 2 in the confidence interval statement $\pm 2/\sqrt{T}$ should be replaced by 1.96:

$$\Pr\{-1.96/\sqrt{T} \leq \rho_\tau \leq 1.96/\sqrt{T}\} = 0.95.$$

Notice that this confidence interval only obtains as an approximation, as $T \to \infty$.

[6]Note that the fact that $\{X_t\}$ is independently distributed (over time) implies that the all autocorrelation coefficients $\rho_\tau = 0$, $\forall \tau \geq 1$, does not imply the opposite: even though $\rho_\tau = 0$, $\forall \tau \geq 1$, independence does not follow.

of Box and Pierce (1970), $\hat{Q}_k$, calculated from the first k autocorrelations of returns as:[7]

$$\hat{Q}_k \equiv T \sum_{\tau=1}^{k} \hat{\rho}_\tau^2 \overset{a}{\sim} \chi_k^2 \quad \text{where} \quad \hat{\rho}_\tau \equiv \frac{\sum_{t=1}^{T-\tau}(R_t - \bar{R})(R_{t+\tau} - \bar{R})}{\sum_{t=1}^{T-\tau}(R_t - \bar{R})^2}$$

(where $\tau > 0$). Here the notation $\overset{a}{\sim}$ means that asymptotically, as $T \to \infty$, the distribution of the $\hat{Q}_k$ statistic, under the null of an IID process (i.e., assuming that the null hypothesis holds), is chi-square, with k degrees of freedom.[8]

The evidence from figure 1 and 2 leads us to conclude that the IID model is not appropriate for returns Note that

$$R_t \text{ is IID} \implies \hat{Q}_k \simeq 0 \text{ for all } k \geq 1$$

but the opposite does not hold:

$$\hat{Q}_k \simeq 0 \text{ for all } k \geq 1 \not\implies R_t \text{ is IID.}$$

The reason is that the definition of independence of a time series process has the following characterization:[9]

$$R_t \text{ is IID} \iff \hat{Q}_k^g \simeq 0 \text{ for all } k \geq 1$$

$$\hat{Q}_k^g \equiv T \sum_{\tau=1}^{k} (\hat{\rho}_\tau^g)^2 \overset{a}{\sim} \chi_k^2 \quad \text{where} \quad \hat{\rho}_\tau^g \equiv \frac{\sum_{t=1}^{T-\tau}(g(R_t) - \overline{g(R_t)})(g(R_{t+\tau}) - \overline{g(R_t)})}{\sum_{t=1}^{T-\tau}(g(R_t) - \overline{g(R_t)})}$$

and $g(\cdot)$ is any (measurable) function that satisfies appropriate "smoothness" conditions. For instance, one may consider $g(x) = x^d$, where d is any positive integer and where $d > 1$ is admissible. Another alternative mentioned above is the case of the function $g(x) = |x|$, the absolute value transformation that turns negative real numbers into positive ones (and leaves positive real numbers unchanged). In practice, independence implies not only the absence of any serial correlation in the *level* of returns—i.e., in the first power of returns, $\hat{Q}_k \simeq 0$ for all $k \geq 1$—but it is equivalent to the absence of any serial correlations in all possible functions of returns, $g(R_t)$.

[7] We shall explain later the exact meaning of denoting portfolio returns as R_t. Enders (2004) informs us that Portmanteau (translated as "a little brown suitcase") tests are generally defined as residual-based tests that do not have a specfic alternative hypothesis. In our case, this means that although tests based on $\hat{Q}_k$ will be able to detect any non-zero autocorrelations, this will not extend to inform us on the nature of such autocorrelations (e.g., which specific ARMA model may successfully fit the data).

[8] It is not surprising that the distribution of the test statistic ($\hat{Q}_k$) is derived assuming the null hypothesis holds: the goal is indeed to find sample evidence in the data to reject such a null hypothesis. Therefore the logical background is: are the data providing evidence inconsistent with the statistical properties that $\hat{Q}_k$ should possess under the null?

[9] Technically, one could even state that $Cov[g(R_t), h(R_{t+\tau})] = 0$ for any choice of sufficiently "smooth" functions $g(\cdot)$ and $h(\cdot)$ and $\forall \tau \neq 0$.

The high dependence in series of square and absolute returns proves that the returns process is not made up of IID random variables: appropriate functions of past returns do give information on appropriate functions of current and future returns. . The evidence implies that large squared returns are more likely to be followed by large squared returns than small squared returns are. The fact that past *squared* returns predict subsequent *squared* return does not imply that past returns may predict subsequent returns: clearly, This relates to a phenomenon that we have already commented in chapter 1: at (relatively) high frequencies, it is possible that higher-order moments—in this case, the second—may be strongly predictable even when the level of asset returns is not, so that they are well approximated by the simple model

$$R_{t+1} = \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{D}(0,1),$$

where the fact that σ_{t+1} changes over time captures the predictability in squared returns that we have just illustrated.

At this point we face two challenges. First, and this is a challenge we are not about to pursue, one wonders what type of economic phenomenon may cause the predictability in squares (or more generally, in higher-order moments, as parameterized by a choice of $m \geq 3$ in $g(x) = x^m$), commonly referred to as *volatility clustering*, the fact that periods of high (low) squared returns tend to be followed by other periods of high (low) squared returns. In short, the general consensus in asset pricing theory is that changes in the speed of flow of relevant—concerning either the exposures to risks or their prices—information to the market causes changes in price volatility that creates clusters of high and low volatility. However, this just moves the question of what may trigger such changes in the speed of information flows elsewhere. Although a range of explanations have been proposed (among them, the effects of transaction costs when trading securities, the fact that investors must learn the process of the fundamentals underlying asset prices in a complex and uncertain world, special features of investors' preferences such as habit formation and loss aversion, etc.) we will drop the issue for the time being. Second, given this evidence of volatility clustering, one feels a need to develop models in which volatility follows a stochastic process where today's volatility is positively correlated with the volatility of subsequent returns. This is what GARCH models are for.

2.2. *Testing and Measuring Deviations from Normality*

Another potentially interesting feature of high frequency data is deviation form the normal distribution. The relevance of these deviation is immediataley illustrated by th fact that in Figure 1 we observe outliers that are several standard deviations away from the zero mean of the daily returns. To evaluate the importance of deviations from normality, we develop statistical

tools to perform tests of non-normality applied to an empirical density (of either returns or standardized residuals). We also provide a quick primer to methods of estimation of empirical densities, to try and "quantify" any such deviations from a Gaussian benchmark.

The key tool to perform statistical tests of normality is Jarque and Bera's (1980) test.[10] The test has a very intuitive structure and is based on a simple fact: if $X_t \sim \mathcal{N}(\mu, \sigma^2)$, then the distribution of X_t is symmetric—therefore it has zero skewness—and it has a kurtosis of 3.[11] In particular, if we define the unconditional mean $\mu \equiv E[X_t]$ and the variance $\sigma^2 \equiv Var[X_t]$, then skewness is

$$Skew[X_t] \equiv \frac{E[(X_t - \mu)^3]}{(Var[X_t])^{3/2}} = \frac{E[(X_t - \mu)^3]}{\sigma^3},$$

while kurtosis is[12]

$$Kurt[X_t] \equiv \frac{E[(X_t - \mu)^4]}{(Var[X_t])^2} = \frac{E[(X_t - \mu)^4]}{\sigma^4} \geq 0.$$

Clearly, skewness is the scaled third central moment, while kurtosis is the scaled fourth central moment.[13] When skewness is positive (negative), then $E[(X_t - \mu)^3] > 0$ (< 0) and this means that there is a larger probability mass below (above) the mean μ than there is above (below). Because a normal distribution implies perfect symmetry around the mean and therefore the same probability below and above μ, then $Skew[X_t] = 0$ when $X_t \sim \mathcal{N}(\mu, \sigma^2)$. We also call *excess* kurtosis the quantity $Kurt[X_t] - 3$, which derives from the fact that $Kurt[X_t] = 3$ when $X_t \sim \mathcal{N}(\mu, \sigma^2)$. A positive (negative) excess kurtosis implies that X_t has fatter (thinner) tails than a normal distribution. Because $Kurt[X_t] \geq 0$, then excess kurtosis may at most be equal to -3.

Jarque and Bera's test is based on *sample* estimates of skewness and (excess kurtosis) from the data, here either raw asset returns or standardized residuals from an earlier estimation of some dynamic econometric model. Denoting with a "hat" sample estimates of central moments obtained from the data, under the null hypothesis of normally distributed errors, Jarque and

[10] This is not the only test available, but it is certainly the most widely used in applied finance.

[11] Here X_t is any generic time series. In this chapter, we shall be interested in two cases: when $X_t = R_{PF,t}$ and when $X_t = \hat{z}_t$ from some model. In the second case, when we deal with standardized residuals, we shall ignore the fact that $\hat{z}_t$ depends on some vector of estimated parameters, $\hat{\boldsymbol{\theta}}$; to take that into account would introduce considerable complications because it would make each $\hat{z}_t$ a function of the entire data sample, $\{\hat{z}_t\}_{t=1}^T$. This occurs because the entire data set $\{\hat{z}_t\}_{t=1}^T$ has been presumably used to estimate $\hat{\boldsymbol{\theta}}$.

[12] Later skewness will also be called ζ_1 and excess kurtosis ζ_2.

[13] A central moment is defined as $\mu_k \equiv E[(X_t - \mu)^k]$ where k is an integer number. Skewness and kurtosis are scaled central moments because they are divided by σ^k. This derives from the desire to express skewness and kurtosis as pure numbers, which is obtained by dividing them by another central moment (here the second), raised to the appropriate power so that the unit of measurement at the numerator and denominator (e.g., percentage) exactly cancel out. The fact that skewness and kurtosis are pure numbers means that these can be compared across different series, different periods, etc. Because kurtosis is the ratio of two (powers) of positive central moments, then it can only be non-negative.

Bera's test statistic is:

$$\widehat{JB} \equiv \frac{T}{6}\left\{\widehat{Skew}[X_t]\right\}^2 + \frac{T}{24}\left\{\widehat{Kurt}[X_t] - 3\right\}^2 \overset{a}{\sim} \chi^2_2,$$

where T is sample size, and the pedix 2 in χ^2_2 indicates that the critical value needs to be found under a chi-square distribution with 2 degrees of freedom. As usual, large values of this statistic—exceeding some critical value under the χ^2_2 selected for a given size (i.e., probability of a type I error) of the test—will indicate departures from normality. Note that $\widehat{JB}$ is a function of *excess* kurtosis and not of kurtosis only. This result derives from the fact that $\widehat{JB}$ is the sum of the squares of two random variables (technically, sample statistics) that have each a normal asymptotic distribution,[14]

$$\sqrt{T}\widehat{Skew}[X_t] \overset{a}{\sim} \mathcal{N}(0,6)$$
$$\sqrt{T}\left\{\widehat{Kurt}[X_t] - 3\right\} \overset{a}{\sim} \mathcal{N}(0,24),$$

are also asymptotically independently distributed.

For instance, daily returns on our portfolio on the sample the Jarque-Bera statistic in this case is huge: 2925.9398 which is well above any critical values under a χ^2_2 (e.g., these are 5.99 for $p = 5\%$; 9.21 for $p = 1\%$; 13.82 for $p = 0.1\%$)! Clearly, the null hypothesis of returns being normally distributed can be rejected at any significance level; in fact, the p-value associated with such a large value of $\widehat{JB}$ is essentially zero.

To see how this is reflected in the distribution of return we can use a kernel density estimator of our variable and compare it with the normal distribution. A kernel density estimator is an empirical density "smoother" based on the choice of two objects: (i) the *kernel function* $K(x)$, and (ii) the *bandwidth parameter*, h. The kernel function is defined as some smooth function (read, continuous and sometimes also differentiable) that integrates to 1:

$$\int_{-\infty}^{+\infty} K(x)\, dx = 1.$$

For instance, a typical kernel function is the Gaussian one,

$$K^{Gauss}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \tag{1}$$

which also corresponds to the probability density function of a $\mathcal{N}(0,1)$ variate (right?). Here x represents any possible value that the generic random variable X_t may take.[15] The bandwidth

[14] It is well known that if $Z_j \sim \mathcal{N}(\mu_j, \sigma^2_j)$ $j = 1, 2, ..., k$ and are independent, then $Z^2_1 + Z^2_2 + ...Z^2_j + ...Z^2_k \overset{a}{\sim} \chi^2_k$. The notation $\overset{a}{\sim} \mathcal{D}$ means that asymptotically, as $T \to \infty$, the distribution of the statistic under examination is $\mathcal{D}$.

[15] Generic, because we are still trying to deal with both the case of asset or portfolio returns, $X_t = R_{PF,t}$ and with $X_t = \hat{z}_t$ from some model.

parameter is instead used to allocate weight to values of x_i in the support of X_t that differ from a given x. This last claim can be understood only by inspecting the general definition of a *kernel density estimator*:

$$\hat{f}_X^{\text{ker}}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2}$$

where n is the number of points over which the estimation is based, usually the size of the sample at hand (in this case, $n = T$). Two aspects need to be adequately emphasized. First, in (2) we are estimating not a parameter of the population but the entire *density* of such a population. This means that $\hat{f}_X^{\text{ker}}(x)$ represents an estimator of the true but unknown $f_X(x)$.[16] Second, the mechanics of (2) is easy to understand: for each x_i in your data set, you compute $\hat{f}_X^{\text{ker}}(x)$ for any arbitrary value x in the support of X_t, by running through your entire sample, computing for each x_i the kernel "scores" $K\left((x - x_i)/h\right)$ and summing them. Note that because you have n observations in your sample and the differences $(x - x_i)$ are re-weighted by the bandwidth h, the total sum is scaled by the factor nh. In this sense, note that a large (small) h tends to strongly (weakly) shrink any $(x - x_i) \neq 0$, which justifies our claim that the bandwidth parameter allocates weight to values of x_i in the support of X_t that differ from a given x.

As esoteric as this may sound, the truth is that since the early ages you have been implicitly trained to compute and use kernel density estimators all the time. As it often occurs however, you have also been educated to use a very poor—in a statistical sense—kernel density estimator, the so-called "histogram estimator" that is obtained from the general formula in (2) when $h = 1$ (as we shall see, $h = 1$ is hardly optimal) and the kernel function is Dirac (usually denoted as $\delta(x)$), i.e., a sort of indicator function:

$$K_{hist}\left(x - x_i\right) = \delta(x_i) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}.$$

[16] Yes, it is possible. In case you are asking yourselves what is the point of spending years studying how to estimate parameters of such a population density while one may actually attack the problem by estimating the density itself, don't. The branch of statistics that deals with the second task is called *nonparametric* statistics (econometrics). Although its goals are as general as ambitious, these do not solve all the problems that applied finance people usually face. For instance, in finance we care a lot for not only fitting/modelling objects of interest, but also in understanding their dynamics over time (because we would like to predict them). Nonparametric econometrics becomes very problematic when it is employed in view of this second type of objective. Hence parametric econometrics remains a crucial subject and most work in applied finance and economics is still organized around parametric methods.

As a result, every time you build a histogram, you are using:[17]

$$\hat{f}_X^{hist}(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(x = x_i\right) = \text{Fraction of your data equal to } x.$$

Of course, there is no good reason to set $K\left(x - x_i\right) = \delta(x)$ or $h = 1$. On the contrary, after the naive histogram estimator, the most common type of kernel function used in applied finance is the Gaussian kernel in (1). A $K(x)$ with optimal (in a Mean-Squared Error sense) properties is instead Epanechnikov's:

$$K_{Epan}(x) = \frac{3}{4\sqrt{5}}\left(1 - 0.2x^2\right) I(-\sqrt{5} \leq x \leq \sqrt{5}). \tag{3}$$

Other popular kernels are the triangular and box kernels:

$$K_{Box}(x) = \frac{1}{2} I(|x| < 1) \qquad K_{Triang}(x) = (1 - |x|) I(|x| < 1). \tag{4}$$

Figure 3 shows the kernels in (3) and (4)

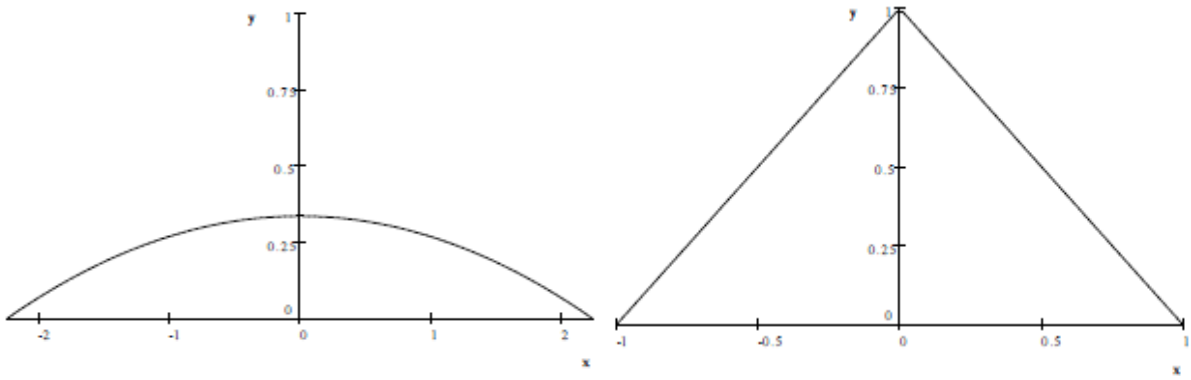


Figure 3: The Epanechnikov (left) and Triangular (right) kernels

The fact that Epanechnikov's kernel is optimal—because it minimizes the average squared deviations $[f_X(x) - \hat{f}_X^{\text{ker}}(x)]^2$—while the Gaussian is not, illustrates one general point, that to minimize the integrated MSE,

$$E\int_{-\infty}^{+\infty} [f_X(x) - \hat{f}_X^{\text{ker}}(x)]^2 dx,$$

kernel functions that are truncated and do not extend to the infinite right and left tails tend to display superior properties when compared to kernels that do. However, the histogram kernel over-does it in this dimension and seems to excessively truncate, because it prevents that any

[17] Usually, what we do to present smarter-looking results, is to organize the possible values of x_i in buckets (intervals) and estimate the probability of that interval as the percentage of your sample that falls in that bucket. However, the nature of the resulting density estimator is the same, alas. In the following formula, note that $I\left(x = x_i\right)$ and $I_{\{x=x_i\}}$ have the same meaning.

$x_i \neq x$ may bring any information useful to the estimation of $f_X(x)$. Finally, the bandwidth parameter h is usually chosen according to the rule (n here is again the sample size):

$$h = 0.9 \cdot \hat{\sigma} \cdot n^{-1/5},$$

which minimizes the integrated MSE across kernels.

How does one use kernel density estimators and do different choices of $K(x)$ make a big difference when it comes to assess deviations from normality? The first question has a trivial answer: here we are in the notoriously difficult (and silly) "eyeballing domain" and—as we did above in our comments—every time one notices large departures of the kernel density estimates from a given benchmark (for us, the normal distribution), you have legitimation to debate the issue, and especially how and why the deviation occurs. However, it is doubtful that the choice of optimal vs. sub-optimal kernel density estimators may make a first-order differences for our ability to assess whether data are normal or not. For instance, in the following Figure 4, it seems that financial returns (in this case, value-weighted U.S. stock returns) are easily assessed to be leptokurtic, i.e., they have fat tails and highly peaked densities around the mean, independently of the specific kernel density estimator that is employed.

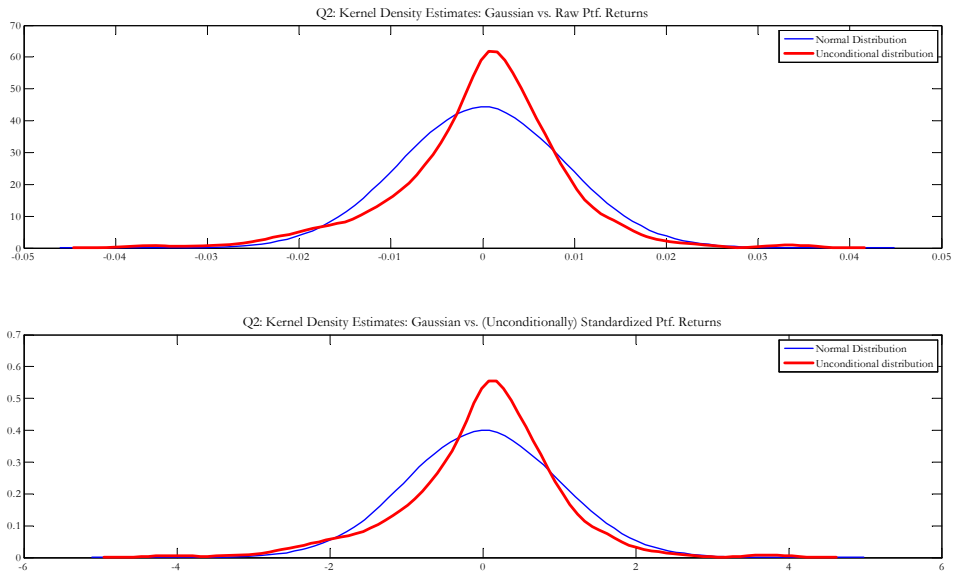


Figure 4: The non-normality of daily. returns using kernel density estimators

If you are ready to work with visual tools instead of performing formal inference on the null hypothesis of normally distributed returns or standardized residuals, another informal and yet powerful method to visualize non-normalities consists of *quantile-quantile* (Q-Q) *plots*. The idea is to plot in a standard Cartesian reference graph:

- the quantiles of the series under consideration, X_t, either raw returns or standardized residuals from the earlier fit of some conditional econometric model;
- against the quantiles of the normal distribution.

If the returns were truly normal, then the graph should look like a straight line with a 45-degree angle. The reason is that if the theoretical (in this case, normal) and empirical quantiles are exactly identical, then they must fall on the 45-degree line. Systematic deviations from the 45-degree line signal that the returns are not well described by the normal distribution and give ground to rejection of the null of normality. The recipe to build a Q-Q plot is simple: first, sort all (standardized) returns in ascending order, and call the ith sorted value x_i; second, compute the empirical probability of getting a value below the actual as $(i-0.5)/T$, where T is number of observations available in the sample.[18] Finally, we calculate the standard normal quantiles as $\Phi^{-1}\left((i-0.5)/T\right)$, where $\Phi^{-1}\left(\cdot\right)$ denotes the inverse of a standard normal density. At this point, we can represent on a scatter plot the (standardized) returns and sort the data on the Y-axis against the standard normal quantiles on the X-axis. Figure 5 shows an examples of Q-Q

[18]The subtraction of 0.5 is an adjustment allowing for the fact that we are using a finite sample and a discrete density estimator to estimate a continuous distribution.

plots applied to the our portfolio returns.

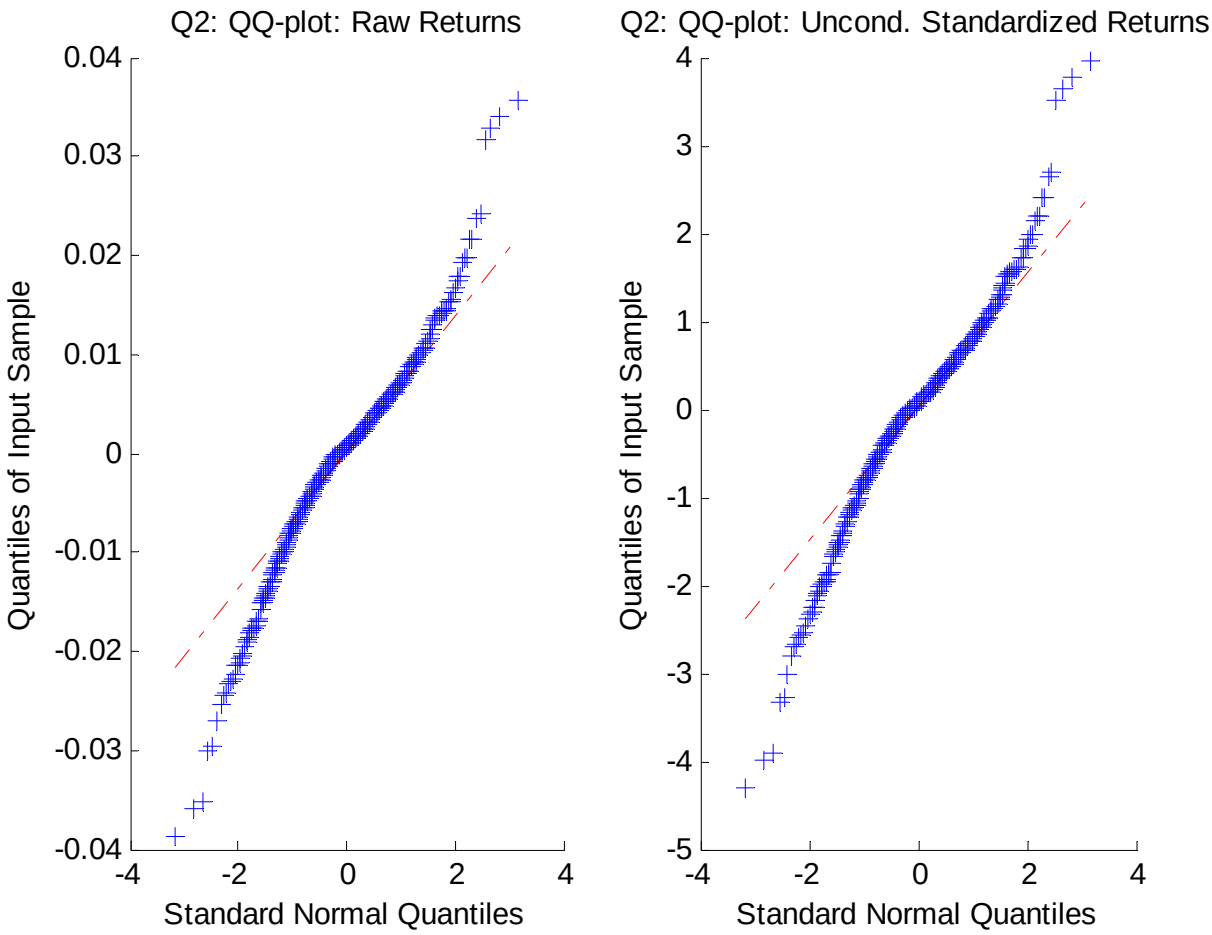


Figure 5: Q-Q plots of raw vs. standardized returns

In Figure 5, both plots reject normality. The deviations from the 45-degree line are obvious and massive in both tails. In particular, the empirical quantiles in the *left* tail are all smaller—i.e., the point in the return distribution below which a given percentage of the sample lies occurs for a return level that is smaller, i.e., more negative—than the theoretical quantiles that one obtains under a theoretical normal distribution that has the same mean and the same variance as the sample of raw returns. This means that the left tail of the empirical distribution of returns is thicker/fatter than the normal tail: in reality, extreme negative market declines have a higher probability than in a Gaussian world. On the contrary, the empirical quantiles in the *right* tail are all larger—i.e., the point in the empirical support above which a given percentage of the sample lies occurs for a return level that is larger—than the theoretical quantiles that one obtains under a theoretical normal distribution that has the same mean and the same variance as the sample data. This means that the right tail of the empirical distribution of S&P 500 returns is *thicker* than the normal tail: in reality, extreme, positive market outcomes have a

lower probability than in a Gaussian world.

3. Generalized Autoregressive Conditional Heteroskedastic (GARCH) Variance Models

A parismonious model is capable of capturing all the features of high-frquency returns described in the previous section:

$$\begin{aligned} R_{t+1} &= \mu + \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0,1), \\ \sigma_{t+1}^2 &= \omega + \alpha \left(R_t - \mu\right)^2 + \beta \sigma_t^2 \\ \alpha + \beta &< 1 \end{aligned}$$

where returns have a constant mean (that is usually zero) and a time varying GARCH(1,1) structure.

In a model like this the innovation $\epsilon_t \equiv \sigma_t z_t$ has zero mean and is serially uncorrelated at all lags $j \geqslant 1$.

R_{t+1} has a finite unconditional long-run variance of $\frac{\omega}{1-\alpha-\beta}$

$$\begin{aligned} \sigma^2 &= E\left(\sigma_{t+1}^2\right) = \omega + \alpha E\left(R_t - \mu\right)^2 + \beta \sigma^2 \\ &= \omega + \alpha \sigma^2 + \beta \sigma^2 \\ &= \frac{\omega}{1 - \alpha - \beta} \end{aligned}$$

Substituting ω out of the GARCH expression:

$$\begin{aligned} \sigma_{t+1}^2 &= \left(1 - \alpha - \beta\right) \sigma^2 + \alpha R_t^2 + \beta \sigma_t^2 \\ &= \sigma^2 + \alpha \left(\left(R_t - \mu\right)^2 - \sigma^2\right) + \beta \left(\sigma_t^2 - \sigma^2\right) \end{aligned}$$

which illustrates the relation between predicted variance and long-run variance in a GARCH model . Under a GARCH(1,1), the forecast of tomorrow's variance is the long-run average variance, adjusted by:

- adding (subtracting) a term that measures whether today's demeaned squared return is above (below) its long-run average, and
- adding (subtracting) a term that measures whether today's variance is above (below) its long-run average.

A GARCH(1,1) model can be considered as the equivalent of an ARMA(1,1) model for the variance, More generally, in the ARMA(q,p) case, we have:

$$\sigma_{t+1}^2 = \omega + \sum_{i=1}^{q} \alpha_i \left(R_{t+1-i}^2 - \mu\right)^2 + \sum_{j=1}^{p} \beta_j \sigma_{t+1-j}^2. \tag{5}$$

setting $\sigma^2 \equiv E[\sigma_{t+1}^2]$:[19]

$$\begin{aligned}
\bar{\sigma}^2 &= E[\sigma_{t+1}^2] = \omega + \sum_{i=1}^{q} \alpha_i E[\left(R_{t+1-i}^2 - \mu\right)^2] + \sum_{j=1}^{p} \beta_j E[\sigma_{t+1-j}^2] = \omega + \sum_{i=1}^{q} \alpha_i \bar{\sigma}^2 + \sum_{j=1}^{p} \beta_j \bar{\sigma}^2 \\
&= \omega + \bar{\sigma}^2 \left(\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j \right) \Longrightarrow \bar{\sigma}^2 = \frac{\omega}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j}.
\end{aligned}$$

Because unconditional variance exists only if $\bar{\sigma}^2 > 0$, the equation above implies that when $\omega > 0$, the condition

$$1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j > 0 \Longrightarrow \sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$$

must hold. When the long-run (i.e. ergodic) variance of a GARCH process exists, because in a GARCH model the only source of time-variation in conditional moments comes from the variance, we say that the GARCH process is stationary and we also refer to the condition $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$ as a stationarity condition. Moreover, because also existence of conditional variances requires that $\sigma_{t+1}^2 > 0$, the additional restrictions that $\omega > 0$, α_1, α_2, ..., $\alpha_q > 0$, β_1, β_2, ..., $\beta_p > 0$ are usually added both in theoretical work and in applied estimation. Of course in the $q = p = 1$ case, such restrictions are simply $\omega > 0$, $\alpha > 0$, $\beta > 0$ and $\alpha + \beta < 1$.

Even though they are straightforward logical extensions of GARCH(1,1), rich GARCH(q,p) models with q and p exceeding one are rarely encountered in practice. This occurs not only because most data sets do not seem to strongly need the specification of higher-order lags q and p in GARCH models, but also because in practical estimation so many constraints have to be imposed to ensure that variance is positive and the process stationary, that numerical optimization may often be problematic. It is natural to ask why can it be that a simple GARCH(1,1) is so popular and successful? Conside, for simplicity, the case in which $\mu = 0$, notice that by

[19] The following derivation exploits the fact that $\bar{\sigma}^2 = E[\sigma_{t+j}^2]$ $\forall j \geq 0$. This is true of any stationary process: its properties do not depend on the exact indexing of the time series under investigation.

recursive substitution,

$$\begin{aligned}
\sigma_{t+1}^2 &= \omega + \alpha R_t^2 + \beta\sigma_t^2 = \omega + \alpha R_t^2 + \beta[\underbrace{\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2}_{\sigma_t^2}] = \omega(1+\beta) + \alpha(1+\beta)R_{t-1}^2 + \beta^2\sigma_{t-1}^2 \\
&= \omega(1+\beta) + \alpha(1+\beta)R_{t-1}^2 + \beta^2[\underbrace{\omega + \alpha R_{t-2}^2 + \beta\sigma_{t-2}^2}_{\sigma_{t-1}^2}] \\
&= \omega(1+\beta+\beta^2) + \alpha R_t^2 + \alpha\beta R_{t-1}^2 + \alpha\beta^2 R_{t-2}^2 + \beta^3\sigma_{t-2}^2 \\
&= ... = \omega\sum_{j=0}^{\infty}\beta^j + \alpha\sum_{j=0}^{\infty}\beta^j R_{t-j}^2 + \lim_{j\to+\infty}\beta^j\sigma_{t-j}^2. \qquad (6)
\end{aligned}$$

If the return series had started in the sufficiently "distant" past or, equivalently, when $t \to \infty$, so that (6) is a parsimonious parameterization in which todays variance depends on an infinite sum with a particular structure of decaying power weights, given by $\alpha\sum_{j=0}^{\infty}\beta^j$. This is called a an ARCH(∞) model. Because $0 < \beta < 1$ implies that

$$\omega\sum_{j=0}^{\infty}\beta^j = \frac{\omega}{1-\beta},$$

(6) is then equivalent to

$$\sigma_{t+1}^2 = \frac{\omega}{1-\beta} + \text{ARCH}(\infty).$$

$$\lim_{j\to+\infty}\beta^j\sigma_{t-j}^2 = 0$$

which is implied by $\alpha + \beta < 1$ or $\beta < 1 - \alpha < 1$ (as $\alpha > 0$). Therefore, the GARCH(1,1) is a stylish way of parameterizing parsimoniously a ARCH(∞), and its empirical power should be a little less than surprising.

3.1. *A formal (G)ARCH test*

A more formal (Lagrange multiplier) test for (G)ARCH in returns/disturbances has been proposed by Engle (1982). The methodology involves the following two steps: First, use simple OLS to estimate the most appropriate regression equation or ARMA model on asset returns and let $\{\hat{z}_t^2\}$ denote the squares of the standardized returns (residuals), for instance coming from a homoskedastic model, $\hat{z}_t^2 = R_t^2/\hat{\sigma}^2$; Second, regress these squared residuals on a constant and on q lagged values $\hat{z}_{t-1}^2$, $\hat{z}_{t+2}^2$, ..., $\hat{z}_{t-q}^2$ (e_t is a white noise shock):

$$\hat{z}_t^2 = \xi_0 + \xi_1\hat{z}_{t-1}^2 + \xi_2\hat{z}_{t-2}^2 + ... + \xi_q\hat{z}_{t-q}^2 + e_t. \qquad (7)$$

If there are no ARCH effects, the estimated values of ξ_1 through ξ_q should be zero, $\xi_1 = \xi_2 = ... = \xi_q$. Hence, this regression will have little explanatory power so that the coefficient of determination (i.e., the usual R^2) will be quite low. Using a sample of T standardized returns,

under the null hypothesis of no ARCH errors, the test statistic TR^2 converges to a χ^2_q. If TR^2 is sufficiently large, rejection of the null hypothesis that ξ_1 through ξ_q are jointly equal to zero is equivalent to rejection of the null hypothesis of no ARCH errors. On the other hand, if TR^2 is sufficiently low, it is possible to conclude that there are no ARCH effects.[20]

A straightforward extension of (7) can also be used to test alternative specifications of (G)ARCH models. For instance, to test for ARCH(q_1) against ARCH(q_2), with $q_2 > q_1$, you simply estimate (7) by regressing the standardized squared residuals from the ARCH(q_1) model on q_2 lags of the same squared residuals and then use an F-test for the null hypothesis that $\xi_{q_1} = \xi_{q_1+1} = ... = \xi_{q_2}$ in:

$$\hat{z}_t^2 = \xi_0 + \xi_{q_1}\hat{z}_{t-q_1-1}^2 + \xi_{q_1+1}\hat{z}_{t-q_1-2}^2 + ... + \xi_{q_2}\hat{z}_{t-q_2}^2 + e_t.$$

Note that these tests will be valid in small samples only if all the competing ARCH models have been estimated on the same data sets, in the sense that the total number of observations should be identical even though $q_2 > q_1$.

It is also possible to specifically test for GARCH effects by performing a Lagrange multiplier regression-based test. For instance, if one has initially estimated a ARCH(q) model and wants to test for p generalized ARCH terms, then the needed auxiliary regression is:

$$\hat{z}_t^2 = \varsigma_0 + \varsigma_1\hat{\sigma}_{t-1}^{2,ARCH(q)} + \varsigma_2\hat{\sigma}_{t-2}^{2,ARCH(q)} + ... + \varsigma_p\hat{\sigma}_{t-p}^{2,ARCH(q)} + e_t,$$

where $\hat{\sigma}_t^{2,ARCH(q)}$ is the time series of filtered, in-sample ARCH(q) conditional variances obtained in the first-stage estimation. Also in this case, if there are no GARCH effects, the estimated values of ς_1 through ς_p should be zero, $\varsigma_1 = \varsigma_2 = ... = \varsigma_p$. Hence, this regression will have little explanatory power so that the coefficient of determination (i.e., the usual R^2) will be quite low. Using a sample of T standardized returns, under the null hypothesis of no ARCH errors, the test statistic TR^2 converges to a χ^2_q. As before, in small samples, an F test may have superior power.

3.2. *Forecasting with GARCH models*

We have emphasized on several occasions that the point of GARCH models is more proposing forecasts of subsequent future variance than telling or supporting some economic story for why variance may be time-varying. It is therefore natural to ask how does one forecast conditional

[20] With the small samples typically used in applied work, an F-test for the null hypothesis $\xi_1 = \xi_2 = ... = \xi_q$ has been shown to be superior to a χ^2_q test. In this case, we compare the sample value of F to the values in an F-table with q degrees of freedom in the numerator and $T - q$ degrees of freedom in the denominator.

variance with a GARCH model.[21] At one level, the answer is very simple because the one-step (one-day) ahead forecast of variance, $\sigma^2_{t+1|t}$, is given directly by the model in (**??**):

$$\sigma^2_{t+1|t} = \omega + \alpha R^2_t + \beta \sigma^2_t,$$

where the notation $\sigma^2_{t+1|t} \equiv E_t[\sigma^2_{t+1}]$ now stresses that such a prediction for time $t+1$ is obtained on the basis of information up to time t, i.e., that $\sigma^2_{t+1|t}$ is a short-hand for $Var[R_t|\mathcal{F}_t] = E[R^2_t|\mathcal{F}_t]$, where the equality derives from the fact that we have assumed $\mu_{t+1} = 0$.

However we are rarely interested in just forecasting one-step ahead. Consider a generic forecast horizon, $H \geq 1$. In this case, it is easy to show that from (**??**),

$$\begin{aligned}
\sigma^2_{t+H|t} - \bar{\sigma}^2 &= E_t[\sigma^2_{t+H}] - \bar{\sigma}^2 = \alpha E_t[R^2_{t+H-1} - \bar{\sigma}^2] + \beta E_t[\sigma^2_{t+H-1} - \bar{\sigma}^2] \\
&= \alpha(E_t[R^2_{t+H-1}] - \bar{\sigma}^2) + \beta(E_t[\sigma^2_{t+H-1}] - \bar{\sigma}^2) \\
&= \alpha(\sigma^2_{t+H-1|t} - \bar{\sigma}^2) + \beta(\sigma^2_{t+H-1|t} - \bar{\sigma}^2) = (\alpha+\beta)(\sigma^2_{t+H-1|t} - \bar{\sigma}^2).
\end{aligned}$$

This establishes a recursive relationship: the predicted deviations of $t+H$ forecasts from the unconditional, long-run variance on the left-hand side equal $(\alpha+\beta) < 1$ times the predicted deviations of $t+H-1$ forecasts from the unconditional, long-run variance. All the forecasts are computed conditioning on time t information. However, we know from the recursion that $\sigma^2_{t+H-1|t} - \bar{\sigma}^2 = (\alpha+\beta)(\sigma^2_{t+H-2|t} - \bar{\sigma}^2)$, and

$$\sigma^2_{t+H|t} - \bar{\sigma}^2 = (\alpha+\beta)\left[\underbrace{(\alpha+\beta)(\sigma^2_{t+H-2|t} - \bar{\sigma}^2)}_{\sigma^2_{t+H-1|t} - \bar{\sigma}^2}\right] = (\alpha+\beta)^2(\sigma^2_{t+H-2|t} - \bar{\sigma}^2).$$

Working backwards this way $H-1$ times, it is easy to see that

$$\sigma^2_{t+H|t} - \bar{\sigma}^2 = (\alpha+\beta)^{H-1}(\sigma^2_{t+1|t} - \bar{\sigma}^2) \tag{8}$$

or

$$\sigma^2_{t+H|t} = \bar{\sigma}^2 + (\alpha+\beta)^{H-1}(\sigma^2_{t+1} - \bar{\sigma}^2) = \bar{\sigma}^2 + (\alpha+\beta)^{H-1}[\alpha(R^2_t - \bar{\sigma}^2) + \beta(\sigma^2_t - \bar{\sigma}^2)].$$

This expression implies that as the forecast horizon H grows, because for $(\alpha+\beta) < 1$ the limit of $(\alpha+\beta)^{H-1}$ is 0, we obtain

$$\lim_{H\to\infty} \sigma^2_{t+H|t} = \bar{\sigma}^2,$$

[21] For concreteness, in what follows we focus on the case of a simple GARCH(1,1) model. All these results, at the cost of tedious algebra, may be generalized to the GARCH(q,p) case. This may represent a useful (possibly, boring) exercise.

i.e., the very long horizon forecast from a stationary GARCH(1,1) model is the long-run variance itself. Practically, this means that because stationary GARCH models are mean-reverting, any long-run forecast will simply exploit this fact, i.e., use $\bar{\sigma}^2$ as the prediction. Of course, for finite but large H it is easy to see that when $(\alpha+\beta)$ is relatively small, then $\sigma^2_{t+H|t}$ will be close to $\bar{\sigma}^2$ for relatively modest values of H; when $(\alpha+\beta)$ is instead close to 1, $\sigma^2_{t+H|t}$ will depart from $\bar{\sigma}^2$ even for large values of H. (8) has another key implication: because in a GARCH we also restrict both α and β to be positive, $(\alpha+\beta) \in (0,1)$ implies that $(\alpha+\beta)^{H-1} > 0$ for all values of the horizon $H \geq 1$. Therefore it is clear that $\sigma^2_{t+H|t} > \bar{\sigma}^2$ when $\sigma^2_{t+1|t} > \bar{\sigma}^2$, and vice-versa. This means that H-step ahead forecasts of the variance will exceed long-run variance if 1-step ahead forecasts exceed long-run variance, and vice-versa. As you have understood at this point, the coefficient sum $(\alpha+\beta)$ plays a crucial role in all matters concerning forecasting with GARCH models and is commonly called the *persistence level/index* of the model: a high persistence, $(\alpha+\beta)$ close to 1, implies that shocks which push variance away from its long-run average will persist for a long time, even though eventually the long-horizon forecast will be the long-run average variance, $\bar{\sigma}^2$.

In asset allocation problems, we sometimes care for the variance of long-horizon returns,

$$R_{t+1:t+H} \equiv \sum_{h=1}^{H} R_{t+h}.$$

Chapter 1 has already extensively discussed the properties of long-horizon returns, emphasizing how simple sums make sense in the case of continuously compounded returns.[22] Here we specifically investigate conditional forecasts (expectations) of the variance of long-horizon returns.The simple model $R_{t+1} = \sigma_{t+1} z_{t+1}$, $z_{t+1} \sim$IID $\mathcal{N}(0,1)$, implies that financial returns have zero autocorrelations, the variance of the cumulative H-day returns is:

$$\begin{aligned} \sigma^2_{t+1:t+H} &\equiv Var_t\left[\sum_{h=1}^{H} R_{t+h}\right] = E_t\left[\left(\sum_{h=1}^{H} R_{t+h}\right)^2\right] = E_t\left[\sum_{h=1}^{H} R^2_{t+h}\right] \\ &= \sum_{h=1}^{H} E_t[R^2_{t+h}] = \sum_{h=1}^{H} \sigma^2_{t+h|t}. \end{aligned}$$

Note that $Var_t\left[\sum_{h=1}^{H} R_{t+h}\right] = E_t\left[\sum_{h=1}^{H} R^2_{t+h}\right]$ because $E_t\left[\sum_{h=1}^{H} R_{t+h}\right] = \sum_{h=1}^{H} E_t[R_{t+h}] = 0$. Moreover, $E_t[(\sum_{h=1}^{H} R_{t+h})^2] = E_t\left[\sum_{h=1}^{H} R^2_{t+h}\right]$ because the absence of autocorrelation in returns leads to all the conditional expectations of the cross-products, $E_t\left[R_{t+\tau} R^2_{t+\tau+k}\right]$ $(k \neq 0)$

[22] The notation $R_{t+1:t+H}$ may be new, but it is also rather self-evident.

to vanish by construction. Solving in the GARCH(1,1) case, we have:

$$\begin{aligned}\sigma^2_{t+1:t+H} &= \sum_{h=1}^{H}\bar{\sigma}^2 + \sum_{h=1}^{H}(\alpha+\beta)^{h-1}(\sigma^2_{t+1|t}-\bar{\sigma}^2) \\ &= H\bar{\sigma}^2 + \sum_{h=1}^{H}(\alpha+\beta)^{h-1}(\sigma^2_{t+1|t}-\bar{\sigma}^2) \neq H\bar{\sigma}^2.\end{aligned}$$

In particular, $\sigma^2_{t+1:t+H} \gtrless H\bar{\sigma}^2$ when $\sum_{h=1}^{H}(\alpha+\beta)^{h-1}(\sigma^2_{t+1|t}-\bar{\sigma}^2)$, which requires that $\sigma^2_{t+1|t} \gtrless \bar{\sigma}^2$. More importantly, note that the variance of the (log-) long horizon returns is not simply H times their unconditional, long-run variance: the term $H\bar{\sigma}^2$ needs to be adjusted to take into account transitory effects, concerning each of the R_{t+h} contributing to $R_{t+1:t+H}$. Note that given the result above, the per-period variance from a GARCH model is:

$$\frac{\sigma^2_{t+1:t+H}}{H} = \bar{\sigma}^2 + \frac{1}{H}\sum_{h=1}^{H}(\alpha+\beta)^{h-1}(\sigma^2_{t+1|t}-\bar{\sigma}^2)$$

so that

$$\lim_{H\to\infty}\bar{\sigma}^2 + \frac{1}{H}\sum_{h=1}^{H}(\alpha+\beta)^{h-1}(\sigma^2_{t+1|t}-\bar{\sigma}^2) = \bar{\sigma}^2.$$

4. **Maximum Likelihood Estimation of GARCH Models**

MLE is based on knowledge of the likelihood function of the sample of data, which is affine (i.e., it is not always identical, but for all practical purposes, it is) to the joint probability density function (PDF) of the same data. In general, models that are estimated by maximum likelihood must be fully specified *parametric* models, in the sense that once the parameter values are known, all necessary information is available to simulate the (dependent) variable(s) of interest; yet, if one can simulate the process of returns, this means that their PDF must be known, both for each observation as a scalar random variable, and for the full sample as a vector random variable. The intuition of ML estimation has been already illustrated above: to look for a unique $\hat{\boldsymbol{\theta}} \in \Theta$ (Θ is the space of possible values of the parameters, to accommodate any restrictions or constraints) such that the joint, total probability that the observed data sample has been generated by the assumed stochastic process parameterized by $\boldsymbol{\theta}$ is maximized when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. In what follows, for concreteness, we refer to the MLE for a standard GARCH(1,1) model, when $\boldsymbol{\theta} \equiv [\omega \ \alpha \ \beta]'$. However, it will be clear that these concepts easily generalize to all conditional heteroskedastic models covered in this chapter and therefore to any possible structure for $\boldsymbol{\theta} \in \Theta$.

The assumption of IID normal shocks (z_t),

$$R_{t+1} = \sigma_{t+1}z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0,1),$$

implies (from normality and identical distribution of z_{t+1}) that the density of the time t observation is:

$$l_t \equiv \Pr(R_t;\boldsymbol{\theta}) = \frac{1}{\sigma_t(\boldsymbol{\theta})\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{R_t^2}{\sigma_t^2(\boldsymbol{\theta})}\right),$$

where the notation $\sigma_t^2(\boldsymbol{\theta})$ emphasizes that conditional variance depends on $\boldsymbol{\theta} \in \Theta$. Because each shock is independent of the others (from independence over time of z_{t+1}), the total probability density function (PDF) of the entire sample is then the product of T such densities:

$$L(R_1, R_2, ..., R_T;\boldsymbol{\theta}) \equiv \prod_{t=1}^{T} l_t = \prod_{t=1}^{T} \frac{1}{\sigma_t(\boldsymbol{\theta})\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{R_t^2}{\sigma_t^2(\boldsymbol{\theta})}\right). \tag{9}$$

This is called the *likelihood function*. However, because it is more convenient to work—especially when we are about to take the derivatives required by first-order conditions, and also to avoid numerical problems when computers are involved—with sums than with products, we usually consider the natural logarithm of the likelihood function,

$$\mathcal{L}(R_1, R_2, ..., R_T;\boldsymbol{\theta}) \equiv \log L(R_1, R_2, ..., R_T;\boldsymbol{\theta}) = \log \prod_{t=1}^{T} l_t = \sum_{t=1}^{T} \log l_t$$

$$\begin{aligned} &= \sum_{t=1}^{T} \left[-\log \sigma_t(\boldsymbol{\theta}) - \log\sqrt{2\pi} - \frac{1}{2}\frac{R_t^2}{\sigma_t^2(\boldsymbol{\theta})} \right] \\ &= -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T} \log \sigma_t^2(\boldsymbol{\theta}) - \frac{1}{2}\sum_{t=1}^{T} \frac{R_t^2}{\sigma_t^2(\boldsymbol{\theta})}, \end{aligned} \tag{10}$$

where we have used several obvious properties of natural logarithms, including the fact that $\log\sqrt{x} = \log x^{1/2} = 0.5 \log x$ and $\log \sigma_t(\boldsymbol{\theta}) = \log\sqrt{\sigma_t^2(\boldsymbol{\theta})} = 0.5 \log \sigma_t^2(\boldsymbol{\theta})$. $\mathcal{L}(R_1, R_2, ..., R_T;\boldsymbol{\theta})$ is also called *log-likelihood function* and the notation employed emphasizes that it is the log joint probability of the sample of data, given a choice for the parameter vector $\boldsymbol{\theta} \in \Theta$. However, nothing prevents you from seeing the log-likelihood as a function that simply depends on the unknown parameters in (say) $\boldsymbol{\theta} \equiv [\omega \ \alpha \ \beta]'$. Note that whatever value of $\boldsymbol{\theta} \in \Theta$ maximizes(10) will also maximize the likelihood function (9), because $\mathcal{L}(R_1, R_2, ..., R_T;\boldsymbol{\theta})$ is just a monotonic transformation of $L(R_1, R_2, ..., R_T;\boldsymbol{\theta})$. Therefore MLE is simply based on the idea that once the functional form of (10) has been written down, for instance

$$\mathcal{L}(R_1, R_2, ..., R_T;\boldsymbol{\theta}) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T} \log\left[\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2\right] - \frac{1}{2}\sum_{t=1}^{T} \frac{R_t^2}{\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2},$$

and initialized at

$$\sigma_0^2 = \frac{\omega}{1-\alpha-\beta},$$

simply maximizing the log-likelihood to select the unknown parameters,

$$\max_{\boldsymbol{\theta}\in\Theta}\left\{-\frac{T}{2}\log 2\pi-\frac{1}{2}\sum_{t=1}^{T}\log\left[\omega+\alpha R_{t-1}^2+\beta\sigma_{t-1}^2\right]-\frac{1}{2}\sum_{t=1}^{T}\frac{R_t^2}{\omega+\alpha R_{t-1}^2+\beta\sigma_{t-1}^2}\right\}$$

will deliver the MLE, denoted as $\hat{\boldsymbol{\theta}}_T^{ML}$, or

$$\hat{\boldsymbol{\theta}}_T^{ML}=\arg\max_{\boldsymbol{\theta}\in\Theta}\left\{-\frac{T}{2}\log 2\pi-\frac{1}{2}\sum_{t=1}^{T}\log\left[\omega+\alpha R_{t-1}^2+\beta\sigma_{t-1}^2\right]-\frac{1}{2}\sum_{t=1}^{T}\frac{R_t^2}{\omega+\alpha R_{t-1}^2+\beta\sigma_{t-1}^2}\right\}.$$

Here the reference to some need to "initialize" σ_0^2 refers to the fact that the log-likelihood function has a clear recursive structure: given σ_0^2, $\omega+\beta\sigma_0^2$ can be evaluated and therefore the $t=1$ term of $\mathcal{L}$ can be numerically assessed for a given choice of ω and α;[23] at this point, given the value of σ_1^2, $\omega+\alpha R_1^2+\beta\sigma_1^2$ can be evaluated and therefore the $t=2$ term of $\mathcal{L}$ can be numerically assessed for a given choice of ω, α, and β. The algorithm proceeds now iteratively until time T, when given the value of σ_{T-1}^2, $\omega+\alpha R_{T-1}^2+\beta\sigma_{T-1}^2$ can be evaluated and therefore the $t=T$ term of $\mathcal{L}$ can be numerically assessed for a given choice of ω α, and β.

Another aspect needs some care: note that $\hat{\boldsymbol{\theta}}_T^{ML}$ is the maximizer of the log-likelihood function for $\boldsymbol{\theta}\in\Theta$. As already mentioned, this is a compact way to state that ML estimation may be performed subject to a number of constraints, such as positivity restrictions on the parameters and the stationarity condition by which $\alpha+\beta<1$. How do you do all this amazing amount of calculations? Surely enough, not using paper and pencil. Note that even in our short description of the recursive structure of the log-likelihood function calculation, that was done only for a given choice of the parameters $\boldsymbol{\theta}\in\Theta$: infinite such choices remain possible. Therefore, at least in principle, to maximize $\mathcal{L}$ you will then need to repeat this operation an infinite number of times, to span all the vectors of parameters in Θ. Needless to say, it takes an infinite amount time to span all of Θ. Therefore, appropriate methods of numerical, constrained optimization need to be implemented: this is what packages such as Matlab®, Gauss or Stata are for.[24]

What about the desired good properties of the estimator? ML estimators have very strong theoretical properties:

[23] R_0^2 does not appear because it is not available and it is implicitly set to zero, which in this corresponds to the unconditional mean of the process. You know from your ML estimation theory for AR(q) models, that this is not an innocent choice. However, asymptotically, for $T\to\infty$ as it is frequently assumed in finance, such a short-cut will not matter.

[24] For instance, Newton's method makes use of the Hessian, which is a $K\times K$ matrix $\mathcal{H}(\boldsymbol{\theta})\equiv\partial^2\mathcal{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ that collects second partial derivatives of the log-likelihood function with respect to each of the parameters in $\boldsymbol{\theta}$. Similarly the gradient $\partial\mathcal{L}(\hat{\boldsymbol{\theta}}_j)/\partial\boldsymbol{\theta}$ collects the first partial derivatives of the log-likelihood function with respect to each of the elements in $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_j$ denote the value of the vector of estimates at step j of the algorithm, and let $\partial\mathcal{L}(\hat{\boldsymbol{\theta}}_j)/\partial\boldsymbol{\theta}$ and $\mathcal{H}(\hat{\boldsymbol{\theta}}_j)$ denote, respectively, the gradient and the Hessian evaluated at $\hat{\boldsymbol{\theta}}_j$. Then the fundamental equation for Newton's Method is $\hat{\boldsymbol{\theta}}_{j+1}=\hat{\boldsymbol{\theta}}_j-\mathcal{H}^{-1}(\hat{\boldsymbol{\theta}}_j)[\partial\mathcal{L}(\hat{\boldsymbol{\theta}}_j)/\partial\boldsymbol{\theta}]$. Because the log-likelihood function is to be maximized, the Hessian should be negative definite, at least when $\hat{\boldsymbol{\theta}}_j$ is sufficiently near $\hat{\boldsymbol{\theta}}_T$. This ensures that this step is in an uphill direction.

- They are *consistent* estimators: this means that as the sample size $T \to \infty$, the probability that the estimator $\hat{\boldsymbol{\theta}}_T^{ML}$ (in repeated samples) shows a large divergence from the true (unfortunately unknown) parameter values $\boldsymbol{\theta}$, goes to 0.

- They are the *most efficient* estimators (i.e., those that give estimates with the smallest standard errors, in repeated samples) among all the (asymptotically) unbiased estimators.[25]

Asymptotic normality of the MLE and QMLE has been proved in the univariate case under low level assumptions, one of which is the existence of moments of order four or higher of the innovations (see Lee and Hansen, 1994). However, Hall and Yao (2003) show that the asymptotic distribution of the QMLE in the univariate GARCH(p, q) model is not normal, but is a multivariate stable distribution (with fatter tails than the normal) if the innovations are in the domain of attraction of a stable law with exponent smaller than two (implying nonexisting fourth moments).

The concept of efficiency begs the question of how does one compute standard errors for ML estimates, in particular with reference to GARCH estimation. If the econometric model is correctly specified, such an operation is based on the concept of *information matrix*, that under correct model specification is given by:

$$\mathcal{I}(\boldsymbol{\theta}) = \lim_{T\to\infty} -E\left[\frac{1}{T}\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]. \tag{11}$$

Correct specification means that the conditional mean and variance functions (i.e., μ_{t+1} and σ_{t+1}^2) should be correct and that the parametric distribution of the shocks (here, so far it was $z_{t+1} \sim$ IID $\mathcal{N}(0,1)$) is also correct. Visibly, the information matrix is based on the Hessian of the MLE problem.[26] In fact, under the assumption of correct specification, the result in (11)

[25] What does asymptotically unbiased mean? Something related to consistency (not exactly the same, but the same for most cases) and for the time being, you may ignore the details of the technical differences between the two concepts. One indirect but equivalent way to state that the MLE is the most efficient estimator is to state that "it achieves the Crámer-Rao lower bound" for the variance of the estimator. Such famous bound represents the least possible covariance matrix among all possible estimators, $\hat{\boldsymbol{\theta}}$.

[26] Wow, big words flying here... The Hessian is simply the matrix of second partial derivatives of the objective function—here the log-likelihood function—and the vector of parameters $\boldsymbol{\theta} \in \Theta$. Let's quickly review it with one example: given the function $\mathcal{L}(\theta_1, \theta_2)$, the Hessian is:

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial^2 \mathcal{L}(\theta_1,\theta_2)}{\partial\theta_1^2} & \frac{\partial^2 \mathcal{L}(\theta_1,\theta_2)}{\partial\theta_1\partial\theta_2} \\ \frac{\partial^2 \mathcal{L}(\theta_1,\theta_2)}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 \mathcal{L}(\theta_1,\theta_2)}{\partial\theta_2^2} \end{bmatrix}.$$

Clearly, the Hessian is a symmetric matrix because $\frac{\partial^2 \mathcal{L}(\theta_1,\theta_2)}{\partial\theta_1\partial\theta_2} = \frac{\partial^2 \mathcal{L}(\theta_1,\theta_2)}{\partial\theta_2\partial\theta_1}$. Also note that the main diagonal of the Hessian collects second partial derivatives vs. the same variable (here, parameter), while the off-diagonal elements collect the cross-partial derivatives.

is called *information matrix equality* (to the Hessian). In particular, it is the inverse of the information matrix, $\mathcal{I}^{-1}(\boldsymbol{\theta})$ that will provide the asymptotic covariance of the estimates:

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T^{ML} - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta})\right),$$

where $\xrightarrow{D}$ denotes convergence in distribution. Obviously, this result implies that $\hat{\boldsymbol{\theta}}_T^{ML} \xrightarrow{a} \boldsymbol{\theta}$.[27] Consistent estimates of the information matrix may be calculated from T sample observations as:[28]

$$\mathcal{I}_T(\hat{\boldsymbol{\theta}}_T^{ML}) = -\frac{1}{T}\sum_{t=1}^{T}\left[\left.\frac{\partial^2 \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right],$$

where, for instance, in the GARCH(1,1) case the log-likelihood contribution $\mathcal{L}(R_t;\boldsymbol{\theta})$ is:

$$\mathcal{L}(R_t;\boldsymbol{\theta}) \equiv -\log 2\pi - \frac{1}{2}\log\left[\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2\right] - \frac{1}{2}\frac{R_t^2}{\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2}.$$

The information matrix measures the average amount of information about the parameters that is contained in the observations of the sample. As $T \to \infty$, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_T^{ML}$ allows us to approximate its variance as:

$$Var[\hat{\boldsymbol{\theta}}_T^{ML}] \simeq \left\{-\frac{1}{T}\sum_{t=1}^{T}\left[\left.\frac{\partial^2 \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right\}^{-1}. \tag{12}$$

The inverse of this matrix can be used for hypothesis testing by constructing the usual z-ratio statistic. As usual, asymptotically valid tests of hypothesis are built as z ratios that have a structure similar to t-ratios, although their normal distribution obtains only asymptotically, as $T \to \infty$. For instance, consider testing the null hypothesis that the parameter $\alpha = \alpha^*$ (α^* is not necessarily zero, but $\alpha^* = 0$ is very common) from a GARCH(1,1), i.e., $H_0 : \alpha = \alpha^*$. The first step is to find the MLE estimate $\hat{\alpha}_T^{ML}$. Second, we compute an estimate of the covariance matrix, i.e.

$$\mathbf{e}_2'\left\{-\frac{1}{T}\sum_{t=1}^{T}\left[\left.\frac{\partial^2 \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right\}^{-1}\mathbf{e}_2,$$

where $\mathbf{e}_2 = [0\ 1\ 0]'$ (because α is the second element in $\boldsymbol{\theta} \equiv [\omega\ \alpha\ \beta]' \in \mathcal{R}_+^3$). Third, we define the ratio

$$z(\hat{\alpha}_T^{ML};\alpha^*) \equiv \frac{\hat{\alpha}_T^{ML} - \alpha^*}{\mathbf{e}_2'\left\{-\frac{1}{T}\sum_{t=1}^{T}\left[\left.\frac{\partial^2 \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right\}^{-1}\mathbf{e}_2}$$

[27] Technically, under adequate assumptions, this may be stated as $\hat{\boldsymbol{\theta}}_T^{ML}$ converging to $\boldsymbol{\theta}$ almost surely (a.s.), meaning that the event in which asymptotically $\hat{\boldsymbol{\theta}}_T^{ML} \neq \boldsymbol{\theta}$ has probability zero.

[28] Probably you are wondering about the origin of the negative sign in the definition of the Hessian. Just think about it: heuristically, you are maximizing the log-likelihood function, which is a function from $\Theta \subseteq \mathcal{R}^K$ into $\mathcal{R}$, $K \geq 1$; at any (also local) maximum a function that is being maximized will be concave; hence, in correspondence to $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, the second derivative should be negative; but for a function from $\Theta \subseteq \mathcal{R}^K$ into $\mathcal{R}$ such a second derivative is in fact the Hessian; hence the Hessian is expected to be negative at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$; only taking the opposite of the negative definite Hessian, one obtaines a positive definite covariance matrix, and we know that covariance matrix ought to be positive definite by construction.

and compare it with a chosen critical value under a $\mathcal{N}(0,1)$, assuming α^* belongs to the feasible set, $\Theta \subseteq \mathcal{R}^K$.[29]

4.1. *Quasi maximum likelihood (QML) estimation*

One key aspect needs to be further emphasized: although the idea of trying and finding a unique $\hat{\boldsymbol{\theta}}_T^{ML} \in \Theta$ that maximizes the joint probability that the sample of data actually came from the process parameterized by $\boldsymbol{\theta} \in \Theta$ is highly intuitive—it answers the layman question "let's rig the assumed model (e.g., a GARCH) to make it as consistent as possible to what we see out there in real life and real financial markets"—one detail should not go unnoticed: the fact that MLE requires knowledge of

$$R_{t+1} = \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0,1). \tag{13}$$

In fact, as we have seen, both the IID nature of z_{t+1} and the fact that $z_{t+1} \sim \mathcal{N}(0,1)$ has been repeatedly exploited in building the log-likelihood function. What if you are not positive about the fact that (13) actually adequately describes the data? For instance, what if all you can say is that

$$R_{t+1} = \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{D}(0,1),$$

but it looks rather unlikely that $\mathcal{D}(0,1)$ may actually turn out to be a $\mathcal{N}(0,1)$?[30] Can we still somehow do what we have described above and enjoy *some* of the good properties of MLE? The answer is a qualified—i.e., that will hold subject to specific but possibly verifiable conditions—"yes" and the resulting estimator is called a *quasi (or pseudo) maximum likelihood estimator* (QMLE). Interestingly, the corresponding statistical result is one of the most useful and frequently exploited finding in modern econometrics—in a way, as close to "magic" as econometrics can go.

The key finding concerning the QMLE estimator is that even though the conditional distribution of the shocks z_t is **not** normal (i.e., $z_{t+1} \sim$IID $\mathcal{D}(0,1)$ and $\mathcal{D}$ does not reduce to a $\mathcal{N}$), *under some conditions*, an application of MLE based on $z_{t+1} \sim$IID $\mathcal{N}(0,1)$ will yield estimators of the mean and variance parameters which converge to the true parameters as the sample gets

[29] For instance, if the test is based on a type I error of 5%, then if $|z(\hat{\alpha}_T^{ML}; \alpha^*)| \gtrapprox 1.96$, the null hypothesis of $\hat{\alpha}_T^{ML} = \alpha^*$ is rejected; if instead $|z(\hat{\alpha}_T^{ML}; \alpha^*)| < 1.96$, the null cannot be rejected. $\mathbf{e}_2' \left\{ -\frac{1}{T} \sum_{t=1}^{T} \left[\frac{\partial^2 \mathcal{L}(R_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right] \right\}^{-1} \mathbf{e}_2$ is simply the matrix algebra operation that selects the second element on the diagonal of the approximate covariance matrix of $\hat{\boldsymbol{\theta}}$. You may find quicker ways to refer to this element of the main diagonal of the covariance matrix.

[30] For instance, you may feel that in fact $z_{t+1} \sim$IID t-student$(0,1)$ may be more sensible. We will deal with this case extensively in the next chapter.

infinitely large, i.e. that are *consistent*.[31] What are the conditions mentioned above? You will need that:

- The conditional variance function, σ^2_{t+1} seen as a function of the information at time t, $\mathcal{F}_t$, must be correctly specified.
- The conditional mean function, μ_{t+1} seen as a function of the information at time t, $\mathcal{F}_t$, must be correctly specified.

Two issues need to be clarified. First, "correctly specified" means that the mathematical, functional specification of the models for the conditional mean and variance are "right". In practice, most of this chapter may be taken as a survey of alternative and increasingly complex conditional variance functions. One example of what it means to mis-specify a model will help understanding what correct specification means. Suppose the world as we know it, is actually ruled—as far conditional variance of the market portfolio (say)— by a EGARCH(1,1) process:

$$\log \sigma^2_{t+1} = \omega + \beta \log \sigma^2_t + g\left(z_t\right) \qquad g\left(z_t\right) = \theta z_t + \alpha(|z_t| - E|z_t|).$$

However, you do not know it and just out of sheer laziness you proceed to estimate a plain-vanilla, off-the-shelf GARCH(1,1) model,

$$\sigma^2_{t+1} = \omega + \alpha R^2_t + \beta \sigma^2_t.$$

Therefore the very functional form that you use, not to mention the fact that you should be paying attention to 4 parameters (ω, β, θ, and α in the EGARCH) and not 3 (ω, β, and α in the GARCH) will be a source of a violation of the needed assumptions to operationalize the QMLE. How would you know in practice that you are making a mistake and using the wrong model for the conditional variance? It is not easy and we shall return to this point, but one useful experiment would be: simulate a long time series of returns from (13) under some EGARCH(1,1). Instead of estimating such a EGARCH(1,1) model on the simulated data, estimate mistakenly a GARCH(1,1) model and look at the resulting standardized residuals, $\hat{z}_{t+1} = R_{t+1}/\hat{\sigma}^{GARCH}_{t+1}$, where the hat alludes to the fact that the GARCH standard deviations have been computed (filtered) under the estimated GARCH model. Because the data came from (13), you know that in a long sample you should never reject the (joint) null hypothesis that $\hat{z}_{t+1} \sim$IID $\mathcal{N}(0,1)$. Trust me: if you performed this experiment, because you have incorrectly estimated a GARCH in place of a EGARCH, $\hat{z}_{t+1} \sim$IID $\mathcal{N}(0,1)$ will be instead rejected in most long samples of

[31] Such conditions and technical details are presented in Bollerslev and Wooldridge (1992).

data.[32] Second, note that the set of assumptions needed for the properties of QMLE to obtain include the correct specification of the conditional mean function, μ_{t+1}. Although technically this necessary and sufficient for the key QMLE result to obtain, clearly in this chapter this is not strictly relevant because we have assumed that $\mu_{t+1} = 0$. However, more generally, also the assumption that μ_{t+1} has been correctly specified will have to be tested.[33]

This may feel as the classical case of "Too good to be true", and you would be right in your instincts: QMLE methods do imply a precise cost, in a statistical sense as they will in general be less efficient than ML estimators are. By using QMLE, we trade-off theoretical asymptotic parameter efficiency for practicality.[34]

In short, the QMLE result says that we can still use MLE estimation *based on normality assumptions* even when the shocks are not normally distributed, if our choices of conditional mean and variance function are defendable, at least in empirical terms. However, because the maintained model still has that $R_{t+1} = \sigma_{t+1} z_{t+1}$ with $z_{t+1} \sim$ IID $\mathcal{D}(0,1)$, the shocks will have to be anyway IID: you can just do without normality, but the convenience of $z_{t+1} \sim$ IID $\mathcal{D}(0,1)$ needs to be preserved. In practice, QMLE buys us the freedom to worry about the conditional distribution later on, and we shall, in the next chapter.

In this case, you will have to take our world for good, but it can be shown that although QMLE yields an estimator that is as consistent as the true MLE one (i.e., they both converge to the same, true $\boldsymbol{\theta} \in \Theta$), the covariance estimator of the QMLE needs to be adjusted with respect to (12). In the QMLE, the optimal estimator of $Var[\hat{\boldsymbol{\theta}}_T^{QML}]$ becomes:

$$\begin{aligned} Var[\hat{\boldsymbol{\theta}}_T^{QML}] \simeq & \left\{ -\frac{1}{T} \sum_{t=1}^{T} \left[\frac{\partial^2 \mathcal{L}(R_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \right\}^{-1} \left\{ -\frac{1}{T} \sum_{t=1}^{T} \left[\frac{\partial \mathcal{L}(R_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \right\} \times \\ & \left\{ -\frac{1}{T} \sum_{t=1}^{T} \left[\frac{\partial \mathcal{L}(R_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \right\}' \left\{ -\frac{1}{T} \sum_{t=1}^{T} \left[\frac{\partial^2 \mathcal{L}(R_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \right\}^{-1}, \end{aligned}$$

where the $K \times 1$ vector $-\frac{1}{T} \sum_{t=1}^{T} \left[\frac{\partial \mathcal{L}(R_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right]$ is called the sum of the sample gradients of the log-likelihood function, i.e., the first-partial derivative of the log-likelihood evaluated in correspondence to $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{QML}$. Such a vector is also called the sample *score* of the log-likelihood

[32] One good reason for that is that the data are simulated to include asymmetric effects that you would be instead completely ignoring under a simpler, incorrect GARCH. Therefore $\hat{z}_{t+1} \sim$IID $\mathcal{N}(0,1)$ will be rejected because the filtered standard residuals will have an asymmetric distribution, which is inconsistent with the null of $\mathcal{N}(0,1)$.

[33] Notice that all misspecification tests that you have encountered in your econometrics sequence so far concerned indeed tests of the correct specification of the conditional mean function, for instance when μ_{t+1} was a simple regression.

[34] Equivalently, a QMLE fails to "achieve the Crámer-Rao lower bound" for the variance among all possible estimators. Such lower bound is in fact attained by the MLE, which however requires that you can both correctly specify the joint density of the data and that shocks are IID.

function.[35]

4.2. *Sequential estimators as QMLEs*

There is one special case in which we may indulge into QMLE estimation even though our key problem is not really the correct specification of the joint density of the shocks to returns, i.e., we may need to invoke the QMLE result even though (13) actually holds. This occurs when estimation of some vector of parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathcal{R}^K$ is conveniently—this is only reason why we would do that, because we now understand that QMLE implies costs—split up in a number of *sequential estimation* stages. For instance, if $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2']' \in \Theta$, the idea is that one would first estimate by full MLE $\boldsymbol{\theta}_1$ and then, conditional on the $\hat{\boldsymbol{\theta}}_1$ obtained during the first stage, estimate—again, at least in principle by full MLE—$\boldsymbol{\theta}_2$. Why would we do that? Sometimes because of practicality, because estimation would be otherwise much harder; in other occasions, to avoid numerical optimization.

The problem with sequential estimation is simply defined: successive waves of (seemingly) partial MLE that may even, at least on the surface, fully exploit (13) will not deliver the optimal statistical properties and characterization of the MLE. On the contrary, a sequential ML-based estimator may be characterized as a QMLE and as such it will be subject to the same limitations as all QMLEs are: loss of asymptotic efficiency. Intuitively, this is due to the fact that when we split $\boldsymbol{\theta}$ down into $[\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2']'$ to separately estimate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, this very separation in a sequential estimator will imply that for all $\hat{\theta}_{1i} \in \hat{\boldsymbol{\theta}}_1$ and $\hat{\theta}_{2j} \in \hat{\boldsymbol{\theta}}_2$, $Cov[\hat{\theta}_{1i}, \hat{\theta}_{2j}] = 0$ even though empirically there is no presumption that this should or might be the case. A few examples will help to clarify this point but also to appreciate the potential advantages from sequential estimation.

4.2.1. **Example 1 (OLS estimation of ARCH models)**

Let's go back to our AR(1)-ARCH(1) example. We know what the right estimation approach is: MLE applied to full log-likelihood function, that in this case will take the form

$$\mathcal{L}(R_1, R_2, ..., R_T; \phi_0, \phi_1, \omega, \alpha) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log\left[\omega + \alpha\epsilon_{t-1}^2\right] - \frac{1}{2}\sum_{t=1}^{T}\frac{(R_t - \phi_0 - \phi_1 R_{t-1})^2}{\omega + \alpha\epsilon_{t-1}^2}, \tag{14}$$

where $\epsilon_{t-1} \equiv R_{t-1} - \phi_0 - \phi_1 R_{t-2}$. Note that $\mathcal{L}(R_1, R_2, ..., R_T; \phi_0, \phi_1, \omega, \alpha, \beta)$ jointly and simultaneously depends on all the 4 parameters that characterize our AR(1)-ARCH(1) model. Yet,

[35] The elements of such a vector are K because $\boldsymbol{\theta}$ has K elements and therefore the same holds for $\partial\mathcal{L}(R_t;\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Moreover,

$$\left\{-\frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial\mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right\}\left\{-\frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial\mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right\}'$$

is a $K \times K$ square, symmetric matrix.

many of you have been subject to a temptation : why not obtain the estimated OLS residuals from a simple regression as

$$\hat{\epsilon}_t = R_t - \hat{\phi}_0 - \hat{\phi}_1 R_{t-1}$$

(which incidentally already gives estimates for ϕ_0 and ϕ_1) and then separately estimate ω and α from maximization of

$$\mathcal{L}_2(\hat{\epsilon}_1, \hat{\epsilon}_2, ..., \hat{\epsilon}_T; \omega, \alpha) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log\left[\omega + \alpha\hat{\epsilon}_{t-1}^2\right] - \frac{1}{2}\sum_{t=1}^{T}\frac{\hat{\epsilon}_t^2}{\omega + \alpha\hat{\epsilon}_{t-1}^2},$$

where the $\{\hat{\epsilon}_t\}_{t=1}^T$ are considered as if they were data even though these are obtained conditional on the OLS estimates of $\hat{\phi}_0$ and $\hat{\phi}_1$. In this case, given $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2']'$, we have $\boldsymbol{\theta}_1 \equiv [\phi_0 \ \phi_1]'$ and $\boldsymbol{\theta}_2 \equiv [\omega \ \alpha]'$. Clearly, there is no illusion: this is a QMLE and the loss of efficiency vs. maximization of (14) may be dramatic. In fact, you even suspect that the very estimation of ϕ_0 and ϕ_1 by OLS in the first stage may be problematic, as in the case of an AR(q) process, MLE does not correspond to OLS. In short, OLS estimation of GARCH models should be avoided in favor of MLE.

4.2.2. **Example 2 (variance targeting)**

This is another common example of sequential estimation that frequently appears in practice. Because we know that the long-run (ergodic) variance from a GARCH(1,1) is $\bar{\sigma}^2 = \omega/(1-\alpha-\beta)$, instead of jointly estimating ω, α, and β, you simply set

$$\tilde{\omega} = (1 - \alpha - \beta)\left[\frac{1}{T}\sum_{t=1}^{T} R_t^2\right]$$

for whatever values of α and β, where the term in square brackets is simply the sample variance of financial returns to be estimate beforehand, on the basis of the data. In this case, given $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_1' \ \theta_2]'$, we have $\boldsymbol{\theta}_1 \equiv [\alpha \ \beta]'$ and $\theta_2 \equiv \omega$. Here the sample variance estimator for $\bar{\sigma}^2$, $\hat{S}^2 \equiv T^{-1}\sum_{t=1}^{T} R_t^2$, is itself a first-step MLE. Of course, the fact that a pre-MLE run of estimation concerning the sample variance to scale down the dimension of $\boldsymbol{\theta}$ makes the resulting estimates of $\hat{\boldsymbol{\theta}}_T$ a QMLE. There are, as usual, two obvious advantages from this approach: (i) you impose the long-run variance estimate on the GARCH model directly and avoid that the model may yield nonsensical estimates;[36] (ii) you have reduced the number of parameters to be estimated in the model by one. These benefits must be carefully contrasted with the well-known costs, the loss of efficiency caused by QMLE.

[36]Note that MLE is not set up to match the sample moments of the data: this means that once $\hat{\boldsymbol{\theta}}_T^{ML}$ is obtained, if the implied moments of the process—for instance, mean and variance—were computed, this may differ from those in the data because of the structure of the log-likelihood function that in general weighs means and variances in a highly non-linear fashion. We shall return on this distinction between MLE and method-of-moment estimators in the next chapter.

5. **Evaluating Conditional Variance Models**

Let's now move where the money is (or not): how can you tell whether a (univariate) volatility model works in practice? A number of methods—called diagnostic or misspecification checks—exist. In this concluding section, we simply discuss four among the many possible methods, even though a few more ideas on how to test whether conditional variance models are correctly specified will emerge in later chapters.

The first, rather simple (and already mentioned, to some extent) method consists of applying standard *univariate tests of normality*, that aim at checking whether data from a given stochastic process $\{X_t\}$ may have been generated by a normal random variable. In practice, if you have estimated the parameters of a conditional volatility model by MLE and exploited the assumption that $z_{t+1} \sim$ IID $\mathcal{N}(0,1)$ in (13), then this implies that the standardized model residuals defined as $\hat{z}_{t+1} \equiv R_{t+1}/\hat{\sigma}_{t+1}$ should have a normal distribution with zero mean and unit variance, where $\hat{\sigma}_{t+1}$ denotes the time series of filtered standard deviations derived from the estimated volatility model. Moreover, because a standard normal distribution is symmetric around 0 and the thickness of its tails are used as benchmarks to measure tail thickness of all distributions (i.e., the excess kurtosis of a normal is set to 0 by construction), the empirical (unconditional, overall) distribution of $\hat{z}_{t+1}$ should be characterized by zero skewness and zero excess kurtosis. At this point, a typical approach consists of using Jarque and Bera's (JB) test : JB proposed a test that measures departures from normality in terms of the skewness and kurtosis of standardized residuals. Under the null hypothesis of normally distributed errors, the JB statistic has a known asymptotic distribution:[37]

$$\widehat{JB}(z) \equiv \frac{T}{6}\left[\underbrace{\widehat{Skew}(z)}_{=0 \text{ under } N(0,1)}\right]^2 + \frac{T}{24}\left[\underbrace{\widehat{Kurt}(z)-3}_{=0 \text{ under } N(0,1)}\right]^2 \overset{a}{\sim} \chi^2_2,$$

where "hats" denote samples estimates of the moments under investigation. Clearly, $\widehat{JB}(z) = 0$ under the null of normality; a large value of $\widehat{JB}(z)$ denotes a departure from normality, and JB tests will formally reject the null of normality when $\widehat{JB}(z)$ exceeds the critical value under a χ^2_2. This means that when the null of normality is rejected, then there is evidence against $z_{t+1} \sim$ IID $\mathcal{N}(0,1)$, which is an indication of model misspecification.

A second method echoes our earlier tests of time series independence of z_{t+1}: this derives

[37] In the expression that follows, we define:

$$\widehat{Skew}(z) \equiv \frac{\sum_{t=1}^T \hat{z}_t^3}{\left(\sum_{t=1}^T \hat{z}_t^2\right)^{3/2}} \quad \widehat{Kurt}(z) \equiv \frac{\sum_{t=1}^T \hat{z}_t^4}{\left(\sum_{t=1}^T \hat{z}_t^2\right)^2}.$$

The intuition behind these scaled unconditional sample moments will be further explained in the next chapter.

from the fact that even though normality has not been assumed (this is the case of QMLE) so that the assumed model for returns is $z_{t+1} \sim$ IID $\mathcal{D}(0,1)$ and $\mathcal{D}(0,1)$ is not $\mathcal{N}(0,1)$, a correctly specified anyway implies

$$z_{t+1} \sim \text{IID}.$$

As we know, independence implies that $\hat{Q}_k^g(z) \simeq 0$ for all $k \geq 1$ where

$$\hat{Q}_k^g(z) \equiv T \sum_{\tau=1}^{k} (\hat{\rho}_\tau^g)^2 \overset{a}{\sim} \chi_k^2 \qquad \hat{\rho}_\tau^g \equiv \frac{\sum_{t=1}^{T-\tau} (g(z_t) - \overline{g(z_t)})(g(z_{t+\tau}) - \overline{g(z_t)})}{\sum_{t=1}^{T-\tau} (g(z_t) - \overline{g(z_t)})^2}$$

and $g(\cdot)$ is any (measurable) function. Because we are testing the correct specification of a conditional volatility model, it is typical to set $g(x) = x^2$, i.e., we test whether the squared standardized residuals, $\hat{z}_{t+1}^2 \equiv R_{t+1}^2 / \hat{\sigma}_{t+1}^2$, display any systematic autocorrelation patterns. As it is now clear, one often simply uses sample autocorrelations to test the null of IID standardized residuals, possibly with tests based on the Bartlett's asymptotic standard errors.

However, the more informative way in which conditional volatility models are typically tested for misspecification is by a smart use of so-called "variance regressions". The idea is simply to regress squared returns computed over a forecast period on the forecasts derived from the conditional variance model under examination:[38]

$$R_{t+1}^2 = b_0 + b_1 \hat{\sigma}_{t+1}^2 + e_{t+1},$$

where e_{t+1} follows a white noise process, i.e., $e_{t+1} \sim \mathcal{D}(0,1)$. Estimation may be simply performed using OLS, no sweat. Let's first state how one proceeds to use such a regression to test whether the conditional variance forecasts obtained from the model, $\hat{\sigma}_{t+1}^2$, are consistent with the null hypothesis of correct specification: in this case, $b_0 = 0$ and $b_1 = 1$. When $b_0 = 0$, we say that the variance model yields *unbiased* forecasts; $b_1 = 1$ implies that the variance model is *efficient*. Our goal is then to use standard OLS inference (as you have learned it from the first part of the Financial Econometrics sequence) to test whether $b_0 = 0$ and $b_1 = 1$. The reason for why correct specification is equivalent to $b_0 = 0$ and $b_1 = 1$ is that under these restrictions

$$R_{t+1}^2 = \hat{\sigma}_{t+1}^2 + e_{t+1} \iff E_t[R_{t+1}^2] = \hat{\sigma}_{t+1}^2, \tag{15}$$

which is indeed what we expect of an unbiased and efficient forecast.

This variance forecast regression has however one problem: the squared returns are used as a proxy (technically, estimator) for the true but unobserved variance in period $t+1$; one wonders, whether this proxy for squared returns is any good. On the one hand, in principle we

[38] It just occured to me: R_{t+1}^2 has nothing to do with the OLS coefficient of determination, R^2, often also called "R-square"!

are fine because from our model $R_{t+1} = \sigma_{t+1} z_{t+1}$ with $z_{t+1} \sim$ IID $\mathcal{D}(0,1)$, so that $E_t[R_{t+1}^2] = \sigma_{t+1}^2 E_t[z_{t+1}^2] = \sigma_{t+1}^2$ because $Var_t[z_{t+1}^2] = 1 = E_t[z_{t+1}^2] - \{E_t[z_{t+1}]\}^2 = E_t[z_{t+1}^2]$ by assumption. $E_t[R_{t+1}^2] = \sigma_{t+1}^2$ means that R_{t+1}^2 is an unbiased estimator of conditional variance. On the other hand, you know better than assessing estimators just on the basis of their being unbiased: the optimal estimator ought to be also the most efficient one. Therefore one wonders what the variance of R_{t+1}^2 as an estimator of σ_{t+1}^2 is:

$$\begin{aligned} Var_t[R_{t+1}^2] &= E_t[(R_{t+1}^2 - E_t[R_{t+1}^2])^2] = E_t[(R_{t+1}^2 - \sigma_{t+1}^2)^2] = E_t[(\sigma_{t+1}^2 z_{t+1}^2 - \sigma_{t+1}^2)^2] \\ &= E_t[\sigma_{t+1}^2 (z_{t+1}^2 - 1)^2] = \sigma_{t+1}^4 E_t[z_{t+1}^4 - 2z_{t+1}^2 + 1] \\ &= \sigma_{t+1}^4 \{\underbrace{E_t[z_{t+1}^4]}_{=\kappa} - 2\underbrace{E_t[z_{t+1}^2]}_{=1} + 1\} = \sigma_{t+1}^4 (\kappa - 1), \end{aligned}$$

where κ is the kurtosis coefficient of z_{t+1}.[39] Because κ for typical (especially, daily) empirical standardized residuals tends to be much higher than 3, the variance of the square return proxy for realized variance is often very poor (i.e., imprecisely estimated), in the sense that $Var_t[R_{t+1}^2]$ in excess of 10 times σ_{t+1}^4 emerges not infrequently. More generally, if we take the coefficient of variation (defined as $E[\hat{\theta}]/\sqrt{Var[\hat{\theta}]}$) as a measure of the variability of an estimator, then

$$\frac{E_t[R_{t+1}^2]}{Var_t[R_{t+1}^2]} = \frac{\sigma_{t+1}^2}{\sqrt{\sigma_{t+1}^4(\kappa - 1)}} = \frac{1}{\sqrt{\kappa - 1}}$$

and this coefficient declines as κ increases. Due to the high degree of noise in squared financial returns, the fit of the variance regression as measured by the regression R^2 (coefficient of determination) is typically very low, typically around 5 to 10%, even if the variance model used to forecast is indeed the correctly specified one. Thus obtaining a low R^2 in such regressions should not lead one to reject a variance model even though the fact that variance regressions lead to a poor fit is certainly not something that can be completely dismissed. What can be done about the fact that (15) is based on an estimator of realized variance, R_{t+1}^2, that is extremely inefficient? Simply enough, to replace the estimator with a better estimator. How can that be done, will be analyzed in later chapters.

Finally, alternative conditional heteroskedastic models can also be compared using *penalized measures of fit* which trade-off in-sample fit with parsimony, i.e., whose value increases as the fit to the data improves but also decreases as the number of estimated parameters increase. Since

[39] Note that there is no contradiction between $E_t[z_{t+1}^4] = \kappa$ and our general assumptions that $R_{t+1} = \sigma_{t+1} z_{t+1}$ with $z_{t+1} \sim$ IID $\mathcal{D}(0,1)$. Naturally, when $R_{t+1} = \sigma_{t+1} z_{t+1}$ with $z_{t+1} \sim$ IID $\mathcal{N}(0,1)$, then $E_t[z_{t+1}^4] = 3$ and $Var_t[R_{t+1}^2] = 2\sigma_{t+1}^4$. As for the fact that $E_t[z_{t+1}^4] = \kappa$ is the kurtosis coefficient, note that

$$Kurt_t(z_{t+1}) \equiv \frac{E_t[z_{t+1}^4]}{\{E_t[z_{t+1}^2]\}^2} = \frac{E_t[z_{t+1}^4]}{\{1\}^2} = E_t[z_{t+1}^4].$$

your early age you have been familiar with one such measure, the adjusted R^2 (often denoted as $\bar{R}^2$) which, indeed, penalizes the standard R^2 with a measure of the parameter vector dimension to prevent that big models have an unfair advantage over smaller, tightly parameterized ones. Why do we value parsimony? Because in general terms the forecasting performance of a model improves as the number of parameters used to fit the data in sample declines—i.e., smaller models tend to perform better than bigger ones do. For instance, the general empirical finding is that, given an identical in-sample fit, e.g., a GARCH(1,1) model will perform better than a GARCH(2,2) when it comes to actual, out-of-sample volatility prediction because the latter implies two additional parameters to be estimated. This is of course the forecasting analog of Occam's razor. In a maximum likelihood set up, the traditional concept of $\bar{R}^2$ is generalized to *information criteria*: in the same way in which the $\bar{R}^2$ is based on the application of penalties to the classical coefficient of determination (R^2), information criteria are based on the concept of applying additional penalty terms to the maximized log-likelihood. Their general structure is:

$$-(\text{Maximized Log-Lik}) + l(dim(\hat{\boldsymbol{\theta}})),$$

where $l(\cdot)$ is a penalty function, and $dim(\hat{\boldsymbol{\theta}})$ is the notation for a counter of the number of different parameters in to be estimated in $\boldsymbol{\theta} \in \Theta$ (this was K in our early treatment). You may wonder way the maximized log-likelihood function enters information criteria with a negative sign: this is due to the fact that, as we have seen, most numerical optimization software actually minimize the negative of the log-likelihood function. Because the maximized log-likelihood is multiplied by -1 while the penalty has been added, it is clear that empirically we shall select models that actually *minimize* information criteria, not maximize them. Three information criteria are widely employed:

- The Bayes-Schwartz information criterion (BIC): $-2\mathcal{L}(\hat{\boldsymbol{\theta}}) + (dim(\hat{\boldsymbol{\theta}}) ln(T)/T)$; this criterion is known to select rather parsimonious models and it appears to be very popular in the applied literature.

- The Akaike information criterion (AIC): $-2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2(dim(\hat{\boldsymbol{\theta}})/T)$; this criterion is also popular because it has optimal asymptotic properties (it is consistent, according to an appropriate definition), although it is also known to select too large non-linear models in small samples (GARCH are non-linear models).

- The Hannan-Quinn information criterion (H-Q):$-2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2[dim(\hat{\boldsymbol{\theta}}) \log(\log(T))/T]$; this criterion has been shown to perform very strongly in small samples and for non-linear models; numerically, it can be shown that it represents a compromise between BIC and AIC.

6. GARCH Specification, Estimation and Forecasting in MATLAB

A number of MATLAB routines are available to specify, estimate and forecast with GARCH models. To illustrate them let us go back to the daily returns of our equally weighted portfolio.

The following commands specify a GARCH(1,1) and structure called spec and estimate it over the sample ss-se (2006-2008 in our case) to display results

```
spec=garchset('P',1,'Q',1); %garchset sets the ARMAX/GARCH model specification
parameters
[coeff, errors,llf,innovation,sigma,summary]=garchfit(spec,Port(ss:se,:));
 garchdisp(coeff,errors);
```

The following output is obtained:

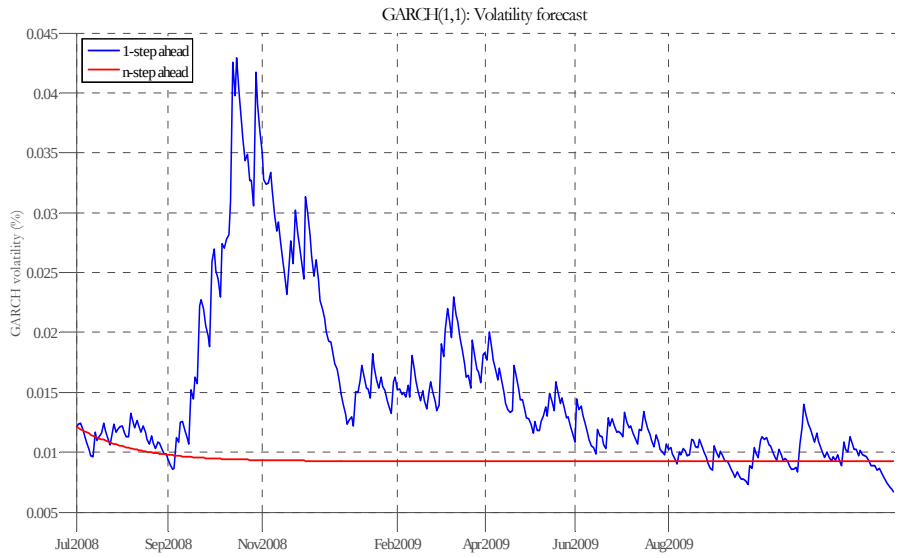
```
Mean:  ARMAX(0,0,0); Variance:  GARCH(1,1)
 Conditional Probability Distribution:  Gaussian
 Number of Model Parameters Estimated:  4
 Standard T
 Parameter Value         Error  Statistic
 ----------- ----------- ------------ -----------
 C                 0.00047595 0.00034739     1.3701
 K                 3.4241e-06 1.009e-06     3.3935
 GARCH(1)     0.86484 0.032627     26.5065
 ARCH(1)     0.095188 0.022505         4.2297
```

Given estimation recursive 1-step ahead and n-step ahead forecast for a given forecasting horizon (from July 2008 to December 2009 in our case) are computed:

```
% Compute GARCH(1,1) 1-step and n-step ahead forecasts using Garchpred
spec_pred=garchset('C',coeff.C,'K',coeff.K,'ARCH',coeff.ARCH,'GARCH',coeff.GARCH);
% This produces a new garchset specification starting from earlier
 % estimates
garch_1s_for=NaN(se_os-se,1); % garchpred forecasts ARMAX/GARCH model responses

for i=1:(se_os-se)
 [SigmaForecast,MeanForecast,SigmaTotal,MeanRMSE] = garchpred(spec_pred,Port(ss:se+i-1),

 garch_1s_for(i)=SigmaForecast(1);
end
```

```
[SigmaForecast,MeanForecast,SigmaTotal,MeanRMSE] = garchpred(spec_pred,Port(ss:se),Tfor)

garch_for=SigmaForecast;
```

The forecast can then be plotted:



7. From GARCH to VaR

After estimation of the GARCH model, the specification can be used to construct the distribution of future returns and compute value at risk. An econometrician that has estimated on the sample 1-T a GARCH(1,1) can use the following model to construct the distribution of returns at time t+1:

$$\begin{aligned} R_{t+1} &= \mu + \sigma_{t+1} z_{t+1} \\ \sigma_{t+1}^2 &= \omega + \alpha \left(R_t^2\right) + \beta \sigma_t^2 \end{aligned}$$

In practice Monte-Carlo simulation and bootstrap can be adopted. In the case Monte-Carlo simulation is chosen, some assumption is made on the distribution of z_{t+1}, for example $z_{t+1} \sim i.i.d.N\left(0,1\right)$, then an artificial sample of size p is drawn form that distribution, given the drawn z_{t+1} and the simulated σ_{t+1} from the model an artificial sample of size p for $R_{i,t+1}^{mc}$ is then generated and VaR can be computed as follows:

$$VaR_{t+1}^{p,mc} = -Percentile\left\{\left\{R_{i,t+1}\right\}_{i=1}^{mc}, 100p\right\}$$

Filtered Historical Simulation (Bootstrap) proceeds exactly as Monte-Carlo simulation with the only difference that the artificial sample of residuls is obtained by drawing with replacement from the empirical distribution of within sample residuals that can be constructed on the basis of the GARCH estimates,

$$\hat{z}_{t+1-\tau} = \frac{R_{t+1-\tau}}{\sigma_{t+1-\tau}}$$

and draw, with replacement, from this empirical distribution. By drawing N times we can get an empirical distribution and use this. After having drawn an artificial sample of size p we can compute VaR as follows:

$$VaR_{t+1}^{p,bs} = -Percentile\left\{\left\{R_{i,t+1}\right\}_{i=1}^{bs}, 100p\right\}$$

An interesting benchmark for model based VaR can be offered by VaR obtained by Historical Simulation. In this case, given the time-series of observations on portfolio returns R_t, the value at risk with coverage rate p is then simply calculated as $100p^{th}$ percentile of the sequence of past portfolo returns:

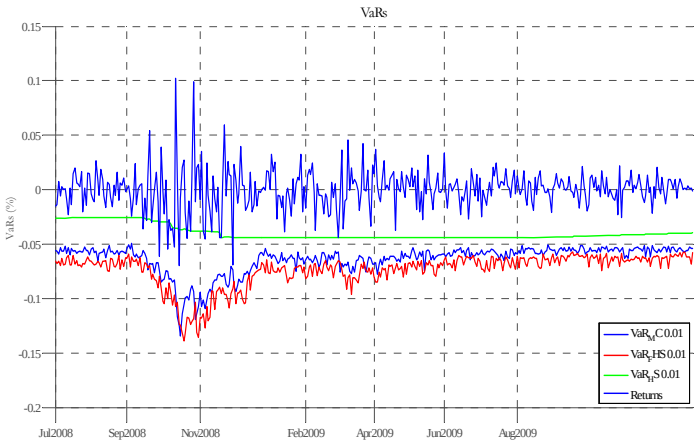
$$VaR_{t+1}^{p,hs} = -Percentile\left\{\left\{RP_{t+1-\tau}\right\}_{\tau=1}^{m}, 100p\right\}$$

The following lines of MATLAB programme compute the three alternative value at risk at each data point from the first day of July 2008 to the end of 2009 and plot them against realized returns:

```
%% VaR with GARCH
alpha=0.01;
Reps=1000;
spec=garchset('C',1,'P',1,'Q',1);
VaR_MC=NaN(rows(Port),1);
VaR_FHS=NaN(rows(Port),1);
VaR_HS=NaN(rows(Port),1);
[coeff, errors,llf,innovation,sigma,summary]=garchfit(spec,Port(ss:se,:));
model = garch('Constant',coeff.C,'GARCH',coeff.GARCH,'ARCH',coeff.ARCH);
sim=garchset('C',1,'P',1,'Q',1,'FixC',1,'FixGARCH',1,...
...'FixARCH',1,'C',coeff.C,'GARCH',coeff.GARCH,'ARCH',coeff.ARCH);
for i=1:(se_os-se)
[coeff1, errors,llf,innovation,sigma,summary]=garchfit(sim,Port(ss:se+i-1,:));
%Monte-Carlo simulation
[v,e] = simulate(model,1,'NumPaths',Reps,'E0',innovation(se+i-ss),'V0',sigma(se+i-ss)^2)
```

```
ret_mc=coeff.C+e;
VaR_MC(se+i,:)=prctile(ret_mc,alpha*100);
%Bootstrap aka Filtered Historical Simulation
Z=innovation./sigma;
Z_boot=randsample(Z,Reps,'true');
[V,E] = filter(model,Z_boot','Z0',0,'V0',innovation(se+i-ss)^2);
ret_FHS=coeff.C+E;
VaR_FHS(se+i,:)=prctile(ret_FHS,alpha*100);
%Historical Simulation
VaR_HS(se+i,:)=prctile(Port(ss:se+i),alpha*100);
end
```

When plotting the results against realized returns we obtain:

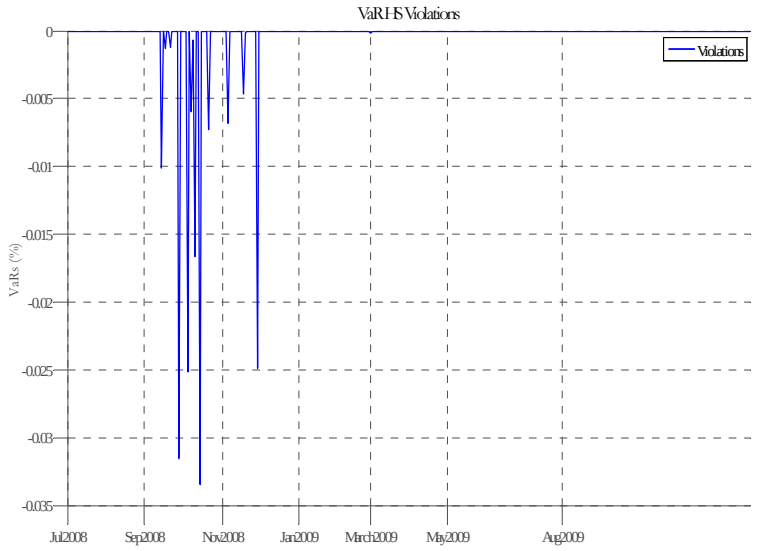


8. **Backtesting VaRs**

How do we test the validity of a VaR model ? The relevant evidence to judge a VaR model are violations. After we have reconstructed the behaviour of a VaR modeli in rela time for a given sample we can observe violations. Consider, for example, the case of the VaR obtained by Historical Simulation methods in the previous section we can genrate violations by constructing the following series:

$$Min(RP_{t+1} - VaR_{t+1}^{p,hs}, 0)$$

the following figure emerges when the generated series is plotted:



Violations are interesting in that:

(a) A good VaR model should not feature neither too few nor too many violations.

(b) We have to few violations when a VaR at the confidence level of alpha shows less then 100*alpha violations in a sample of 100 observations. In this case the VaR model is too conservative. (Note that this happens to be case for the FHS and the Monte Carlo based VaR models in our sample).

(c) when we have violations there are two interesting aspects of that: their number and their timing. A five per cent VaR that feature 5 violations in five successive periods cannot be taken as a valid VaR model as violations are not independent. Clustering of violations is a problem that should lead to reject specific VaR models.

A statistical test for the validity of a VaR model can be derived by analyzing violations and their timing. The testing procedure will concentrate first on the numbero of violations, and then on their timing

8.1. *Unconditional Coverage Testing*

Given a time-series of VaR and observed returns the "hit sequence" of VaR violations is defined as follows:

$$\begin{aligned} I_{t+1} &= 1, if\ RP_{t+1} > VaR_{t+1}^{p,hs} \\ I_{t+1} &= 0, if\ RP_{t+1} > VaR_{t+1}^{p,hs} \end{aligned}$$

In the null hypothesis that the VaR is a valid model violations should not be predictable: the forecast of the probability of a VaR violation should be p every day. The hit sequence in this case should be distributed over time as a Bernoulli variable that takes the value 1 with probability p and the value 0 with probability $1-p$. So

$$\begin{aligned} H_0 &: I_{t+1} \sim i.i.d.\ \text{Bernoulli}(p) \\ f(I_{t+1}, p) &= (1-p)^{1-I_{t+1}} p^{I_{t+1}} \end{aligned}$$

The first test of validity of a VaR is therefore constructed as follows. Take a Bernoulli distribution (I_{t+1}, x) for the that the number of violations, derive a maximum likelihood estimator $\hat{x}$ of x, and test using a likelihood ratio test that $\hat{x}$ is not statistically different from p.

$$\begin{aligned} L(I_{t+1}, x) &= \prod_{i=1}^{T} (1-x)^{1-I_{t+1}} x^{I_{t+1}} \\ &= (1-x)^{T_0} x^{T_1} \end{aligned}$$

where T_1 is the number of violations of the VaR observed in the sample, and $T_0 = T - T_1$. The maximum likelihood estimator $\hat{x} = \frac{T_1}{T}$.

A likelihood ratio test of the null hypothesis $\hat{x} = p$, can then be constructed as follows:

$$LR_{uc} = -2 \ln \left[\frac{L(p)}{L\left(\hat{x}\right)} \right]$$

which is distributed as a χ^2 with one degree of freedom and it is labelled uc as it is a test of unconditional coverage of the VaR, i.e. the number of violations and not the timing of violations is the relevant criterion to test the validity of the VaR model.

Note that usually the number of violations and the number of observations available will not be large, so rather than relying upon the χ^2 distribution, it is adviceble to use Monte-Carlo simulations to build the relevant distribution to conduct the test. In this case the simulated P-values would be obtained by drawing an artificial sample of the relevant size from the null, and using as a P-value the share of simulated test that are larger than the observed one.

8.2. *Independence Testing*

We concentrate now on a test able to reject a VaR with clustered violations. In this case the hit sequence is dependent over time and its evolution over time can be described by a so called

Markov sequence where the transition from the relevant states (violation and no violation) can be described by the following transition probability matrix

$$X_1 = \begin{bmatrix} x_{00} & 1 - x_{00} \\ 1 - x_{11} & x_{11} \end{bmatrix}$$

where:

$$\begin{aligned} x_{00} &= \Pr\left(I_{t+1} = 0 \mid I_t = 0\right) \\ 1 - x_{00} &= \Pr\left(I_{t+1} = 1 \mid I_t = 0\right) \\ x_{11} &= \Pr\left(I_{t+1} = 1 \mid I_t = 1\right) \\ 1 - x_{11} &= \Pr\left(I_{t+1} = 0 \mid I_t = 1\right) \end{aligned}$$

If we observe a sample of T observations the likelihood function of a first order Markov process can be written as follows:

$$L\left(X_1, I_{t+1}\right) = x_{00}^{T_{00}} \left(1 - x_{00}\right)^{T_{01}} \left(1 - x_{11}\right)^{T_{10}} x_{11}^{T_{11}}$$

The maximum likelihood estimates of the relevant parameters are then

$$\begin{aligned} \hat{x}_{00} &= \frac{T_{00}}{T_{00} + T_{01}} \\ \hat{x}_{11} &= \frac{T_{11}}{T_{10} + T_{11}} \end{aligned}$$

and so

$$\hat{X}_1 = \begin{bmatrix} \frac{T_{00}}{T_{00}+T_{01}} & \frac{T_{01}}{T_{00}+T_{01}} \\ \frac{T_{10}}{T_{10}+T_{11}} & \frac{T_{11}}{T_{10}+T_{11}} \end{bmatrix}$$

Under independence

$$\hat{X}_1^{id} = \begin{bmatrix} 1 - \hat{x} & \hat{x} \\ 1 - \hat{x} & \hat{x} \end{bmatrix}$$

and therefore the independence hypothesis $\left(1 - \hat{x}_{00}\right) = \hat{x}_{11}$can be tested using a likelihood ratio test

$$LR_{ind} = -2 \ln \left[\frac{L\left(\hat{X}_1^{id}\right)}{L\left(\hat{X}_1\right)} \right] \sim \chi_1^2$$

As for the unconditional coverage test small sample problems can be fixed by Monte Carlo simulation of the critical values, moreover samples in which $T_{11} = 0$ are often observed. In this cases the likelihood function is computed as

$$L(X_1, I_{t+1}) = x_{00}^{T_{00}} (1 - x_{00})^{T_{01}}$$

8.3. *Conditional Coverage Testing*

Having constructed the test for independence we can test jointly the hypothesis of conditional coverage and independence via the following likelihood ratio test:

$$LR_{cc} = -2 \ln \left[\frac{L(p)}{L\left(\hat{X}_1\right)} \right] \sim \chi_2^2$$

note that

$$LR_{cc} = LR_{uc} + LR_{ind}$$