5 Single-Factor Conditionally Heteroskedastic Models, ARCH and GARCH – On-Line Supplementary Material

On-Line Ex.5A. We use daily data on US aggregate excess market returns (x_{t+1}) obtained as the difference between CRSP value-weighted stock returns (concerning all NYSE, NASDAQ, and AMEX listed stocks, over the relevant periods) for the long sample period Jan. 2, 1963 – Dec. 31, 2016, for a total of 13,594 observations. In particular, we specify a simple Gaussian AR(1) model for the conditional mean function and a Riskmetrics model for the conditional variance function:

$$x_{t+1} = \phi_0 + \phi_1 x_t + \varepsilon_{t+1} \qquad \varepsilon_{t+1} \text{ IID } N(0, \sigma^2_{t+1|t}(\lambda))$$
$$\sigma^2_{t+1|t} = (1-\lambda)\varepsilon_t^2 + \lambda \sigma^2_{t|t-1},$$

The conditional mean function is $\mu_t(\phi_0, \phi_1) = \phi_0 + \phi_1 x_t$. Using E-Views, we have estimated by ML the model obtaining the following estimates (p-values are in parentheses underneath the corresponding coefficient):

$$x_{t+1} = \underset{(0.000)}{0.035} + \underset{(0.000)}{0.123} x_t + \varepsilon_{t+1} \qquad \varepsilon_{t+1} \text{ IID } N(0, \sigma^2_{t+1|t})$$
$$\sigma^2_{t+1|t} = \underset{(0.000)}{0.064}\, \varepsilon_t^2 + \underset{(0.000)}{0.936}\, \sigma^2_{t|t-1},$$

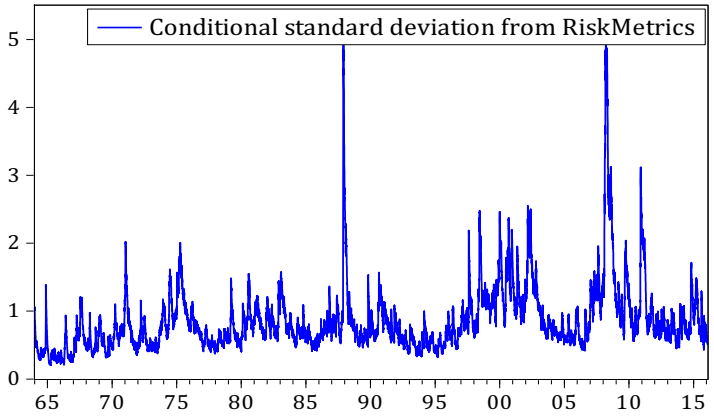


Figure 5A.1 – Plot of One-Day RiskMetrics Volatility Forecasts for US Excess Aggregate Stock Returns

We use the estimated model to forecast the conditional standard deviation of the excess stock return process, since

$$Var_t[x_{t+1}] = \phi_0 + \phi_1 Var_t[x_t] + Var_t[\varepsilon_{t+1}] = \sigma^2_{t+1|t},$$

so that it is natural to use $\sigma_{t+1|t} \equiv \sqrt{\sigma^2_{t+1|t}}$ as a forecast of volatility. Figure 5A.1 shows such forecasts.

Once more, the "law of the 0.94" estimate strikes: almost 30 years later, we find that $\hat{\lambda} = 0.936$, which is close to 0.94 indeed. We move one step further and test this law on a different series of equity-related returns, those on the SMB ("Small-minus-Big") portfolio that goes long in the lowest quintile of the CRSP universe stocks in terms of market value and finances that position by shorting the highest quintile of CRSP stocks when sorted by their total market value. ML estimates are (*p*-values are in parentheses):

$$x_{t+1} = \underset{(0.000)}{0.010} + \underset{(0.000)}{0.111} x_t + \varepsilon_{t+1} \qquad \varepsilon_{t+1} \text{ IID } N(0, \sigma^2_{t+1|t})$$
$$\sigma^2_{t+1|t} = \underset{(0.000)}{0.065}\, \varepsilon_t^2 + \underset{(0.000)}{0.935}\, \sigma^2_{t|t-1},$$

Strikingly, even though the portfolio is very different (just think this is a long-short portfolio that in principle has no or small net beta exposure on the aggregate market portfolio), we obtain similar parameter estimates and also in this case $\hat{\lambda} = 0.935$ falls very close to the 0.94 often recommended by the RiskMetrics experts. We use the estimated model to forecast the conditional standard deviation of the excess stock return process, as shown below.

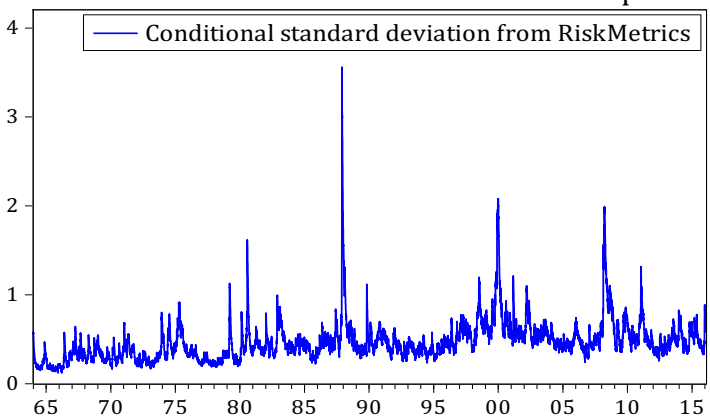


Figure 5A.2 – One-Day RiskMetrics Volatility Forecasts for SMB Returns

On-Line Ex. 5B. Here we would like to compare the different degree of exposure to asymmetries of different asset classes. To this purpose, we use daily *excess* CRSP equity returns, weekly (negative) differences in 10-year US Treasury rates, and monthly US equity *total* returns from Bloomberg. It is interesting to compare the results from the first and the third series to investigate both the effects of the frequency at which a series is sampled and, at least possibly, the impact of subtracting short-term rates from equity returns. The following table conducts a step-by-step model specification search for each of the three series within the threshold-GARCH(p, d, q) class. In the light of earlier evidence, we start off with the case of both p and q being positive, i.e., we rule out simpler ARCH(p) models.

Conditional Mean Model	Conditional Variance Model	p	d	q	Maximized Log-Lik	BIC	Hannan-Quinn	AIC
Daily 1963-2016 Excess Equity Returns from CRSP								
MA(1)	Homeskedastic	0	0	0	-19069.6	2.8070	2.8063	2.8059
MA(1)	GARCH(1,1)	1	0	1	-16255.1	2.3950	2.3932	2.3922
MA(1)	T-GARCH(1,1,1)	1	1	1	-16103.9	2.3735	2.3713	2.3702
MA(1)	T-GARCH(2,1,1)	2	1	1	-16102.3	2.3739	2.3714	2.3701
MA(1)	*T-GARCH(2,2,1)*	*2*	*2*	*1*	*-16091.2*	*2.3730*	*2.3700*	*2.3686*
MA(1)	T-GARCH(2,2,2)**	2	2	2	-16022.1	2.3635	2.3602	2.3586
MA(1)	T-GARCH(2,1,2)	2	1	2	-16099.2	2.3742	2.3712	2.3698
MA(1)	T-GARCH(3,2,2)**	3	2	2	-16035.8	2.3662	2.3626	2.3607
Weekly 1982-2016 10-year Treasury Yield Changes								
AR(1)	Homeskedastic	0	0	0	1257.2	-1.3744	-1.3763	-1.3774
AR(1)	*GARCH(1,1)*	*1*	*0*	*1*	*1424.7*	*-1.5457*	*-1.5536*	*-1.5578*
AR(1)	T-GARCH(1,1,1)	1	1	1	1425.1	-1.5420	-1.5515	-1.5571
AR(1)	T-GARCH(2,1,1)	2	1	1	1426.7	-1.5397	-1.5511	-1.5578
AR(1)	T-GARCH(2,2,1)**	2	2	1	1430.0	-1.5391	-1.5525	-1.5603
AR(1)	T-GARCH(2,2,2)**	2	2	2	1430.8	-1.5359	-1.5512	-1.5601
AR(1)	T-GARCH(2,1,2)	2	1	2	1427.0	-1.5358	-1.5492	-1.5570
AR(1)	T-GARCH(3,2,2)	3	2	2	1430.0	-1.5349	-1.5502	-1.5591
Monthly 1977-2016 Equity Returns								
CER	Homeskedastic	0	0	0	-1394.4	5.8228	5.8175	5.8141
CER	GARCH(1,1)	1	0	1	-1379.6	5.8000	5.7786	5.7649
CER	T-GARCH(1,1,1)	1	1	1	-1376.5	5.7998	5.7734	5.7563
CER	T-GARCH(2,1,1)**	2	1	1	-1375.0	5.8065	5.7748	5.7543
CER	T-GARCH(2,2,1)**	2	2	1	-1365.1	5.7782	5.7412	5.7173
CER	*T-GARCH(2,2,2)*	*2*	*2*	*2*	*-1365.2*	*5.7911*	*5.7488*	*5.7215*
CER	T-GARCH(2,1,2)	2	1	2	-1373.4	5.8125	5.7755	5.7516
CER	T-GARCH(3,2,2)**	3	2	2	-1364.7	5.8019	5.7544	5.7236

** = some of the ML estimates of GARCH coefficients turned out to be negative

Table 5.B1 –Information Criteria-Based Model Selection for Different Data Sets and Frequencies

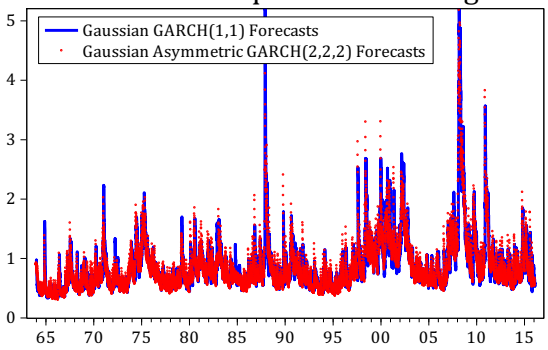
The results are in some ways expected: stock return data contain strong evidence of a need to incorporate asymmetries in a GARCH model; bond returns data do not, and economically it would be more complex to find any justification for such asymmetries. On the opposite, the frequency of the data does not seem to be crucial: stock index returns data contain leverage both at daily and monthly frequency and over two quite different time periods (the former is considerably longer, also including 1963-1976). Interestingly, for both equity data sets, the same, rich GARCH(2,2,2) model prevails. Technically, slightly more parsimonious GARCH(2,2,1) models prevail in terms of minimizing the information criteria, but these models are characterized by a few negative coefficients that make them unsuitable to practical uses. This tendency of relatively large GARCH models to be selected by (all) standard information criteria is something relatively novel when compared to the empirical finance literature, possibly due to the fact that we are performing these estimations 30 years later the birth of the original GARCH model, so that much longer time series of data have become available.

To give an idea of typical estimates, in Figure 5.B1, we report ML estimates and plot the forecasts of conditional volatility they imply. In the case of stock index returns, we also compare such forecasts with those that a symmetric GARCH(1,1) would obtain. *P*-values are in parentheses below ML estimates (obtained assuming normality); we also computed and report long-run ergodic volatilities obtained as the square root of long-run variances.

GARCH(2,2,2):

$$\sigma^2_{t+1|t} = \underset{(0.000)}{0.0003} + \underset{(0.340)}{0.003}\varepsilon_t^2 + \underset{(0.000)}{0.180} I_{\{\varepsilon_t<0\}}\varepsilon_t^2 + \underset{(0.909)}{0.000}\varepsilon_{t-1}^2 - \underset{(0.000)}{0.176} I_{\{\varepsilon_{t-1}<0\}}\varepsilon_{t-1}^2 + \underset{(0.000)}{0.794}\sigma^2_{t|t-1} + \underset{(0.000)}{0.202}\sigma^2_{t-1|t-2}$$

$$\bar{\sigma} = 0.958$$

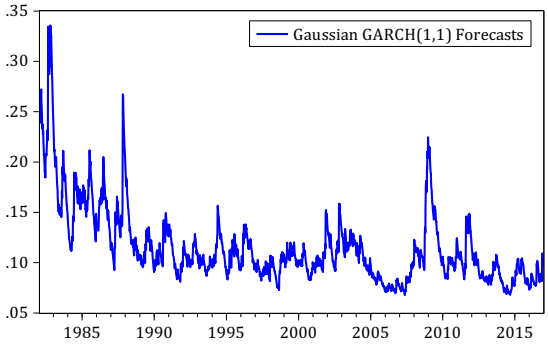


Panel (a) Daily US Equity Excess Returns

GARCH(1,0,1):

$$\sigma^2_{t+1|t} = \underset{(0.003)}{0.0002} + \underset{(0.000)}{0.073}\varepsilon_t^2 + \underset{(0.000)}{0.910}\sigma^2_{t|t-1}$$

$$\bar{\sigma} = 0.117$$

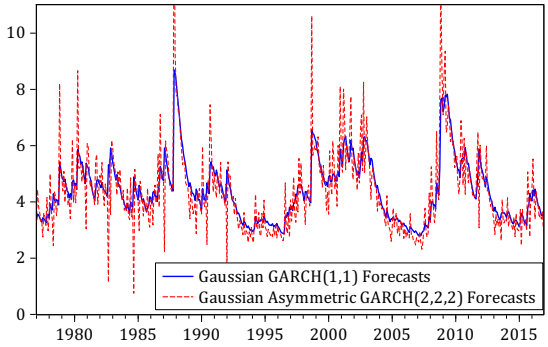


Panel (b) Weekly 10-year Treasury Returns

GARCH(2,2,2)

$$\sigma^2_{t+1|t} = \underset{(0.130)}{0.749} + \underset{(0.000)}{0.103}\varepsilon_t^2 + \underset{(0.000)}{0.463} I_{\{\varepsilon_t<0\}}\varepsilon_t^2 + \underset{(0.000)}{0.235}\varepsilon_{t-1}^2 - \underset{(0.000)}{0.431} I_{\{\varepsilon_{t-1}<0\}}\varepsilon_{t-1}^2 + \underset{(0.000)}{0.778}\sigma^2_{t|t-1} + \underset{(0.797)}{0.049}\sigma^2_{t-1|t-2}$$

$$\bar{\sigma} = 5.353$$



Panel (c) Monthly US Equity Returns

Figure 5.B1 –Estimates and In-Sample Volatility Forecasts from Alternative GARCH Models

While in panel (a), the differences between plain vanilla GARCH(1,0,1) and a richer GARCH(2,2,2) with two lags of leverage effects are modest, in panel (c), at a monthly frequency, these become much more visible. Indeed, in panel (a) the correlation between the two sets of forecasts is 0.96 vs. 0.83 in panel (c). The threshold GARCH forecasts seem to jiggle around the symmetric GARCH one, and therefore appear "spikier" in both directions, in the sense there are a few months in the 1980s characterized by predicted volatility well below 2% per month that could not be forecast from the smoother GARCH(1,0,1). Such differences are presumably due to the fact that while in panel (a) both the standard ARCH coefficients are small and fail to be significant while the two asymmetric ARCH coefficients almost exactly cancel out, this does not occur in panel (c).

On-Line Ex. 5C. Consider a t-Student GARCH(1,1) model with fixed known parameters, controlled by a persistence index $\Delta \equiv \alpha+\beta$:

$$\sigma^2_{t+1|t} = 2 + \frac{\Delta}{9}\varepsilon^2_t + \frac{8\Delta}{9}\sigma^2_{t|t-1} \quad \varepsilon_{t+1} \text{ IID t}(0,\sigma^2_{t+1|t};\nu), \quad \nu > 2$$

Here, we have parameterized the GARCH process in such a way that 1/9 of the total persistence comes from the lagged squared shock, ARCH-type ε^2_t term, and 8/9 come from lagged conditional variance. When Δ=0.9, this corresponds to α=0.1 and β=0.8 that are fairly typical values in the literature. The other parameter that we shall be experimenting with is, of course, ν. We now compute unconditional and conditional *excess* kurtosis

$$ExKurt(R_{t+1}) \equiv \frac{E[\sigma^4_{t+1|t} z^4_{t+1}]}{(Var[R_{t+1}])^2} - 3 \quad \text{and} \quad ExKurt_t(R_{t+1}) = \frac{6}{\nu - 4},$$

for a few representative values of Δ and ν, reporting results. These are summarized at the bottom of the example, for your convenience. When Δ=0 (no GARCH) and ν=4.01, we have $ExKurt(R_{t+1}) = 300$ and $ExKurt_t(R_{t+1}) = 300$ because excess kurtosis can only come from the t-Student shocks. In general, when Δ=0, $ExKurt(R_{t+1}) = ExKurt_t(R_{t+1}) = \frac{6}{\nu-4}$ will always obtain independently of ν>4. When Δ=0.99 (very persistent GARCH) and ν=4.01, we have $ExKurt(R_{t+1}) \simeq 2{,}565$ and $ExKurt_t(R_{t+1}) = 300$: in this case, the sources of excess kurtosis, volatility clustering and fat-tailed shocks compound to give massively thick tails. The approximate equality is used here because the unconditional kurtosis is computed using simulation methods with 100,000 independent trials (because of the notorious difficulty in estimating excess kurtosis with any precision). When Δ=0.8 (persistent GARCH) and ν=9 (which seems typical of many financial series), we have $ExKurt(R_{t+1}) \simeq 1.80$ and $ExKurt_t(R_{t+1}) = 1.2$, which means that a persistent GARCH contributes a 1/3 increase in excess kurtosis on top of a fat-tailed t-Student. Finally, when Δ=0.99 (an extremely persistent GARCH) and ν=20 (where the tails of the t-Student stop being significantly different from a normal distribution), we have $ExKurt(R_{t+1}) \simeq 390$ and $ExKurt_t(R_{t+1}) = 0.375$; this configuration gives powerful evidence of the interaction effects between the thick tails generated by GARCH and the tails of the marginal density that characterizes the assumed t-Student shocks. In fact, $\Delta = 0.99$ by itself but under a Gaussian distribution (say, assuming ν=9999) does not generate such a massive excess kurtosis: $ExKurt(R_{t+1}) \simeq 23.2$. Therefore, it is the interaction between t-Student shocks and GARCH persistence that captures empirically relevant excess kurtosis.

Here the danger is that by incorrectly assuming Gaussian shocks in all circumstances, a researcher may force her GARCH model to express too high a persistence just because implicitly the parameter ν is forced to diverge to infinity while the data would often "prefer" some ν<20. For instance, data generated by a mildly persistent process with mildly fat tails (say, Δ=0.6 and ν=15) would imply $ExKurt(R_{t+1}) \simeq 0.62$ and $ExKurt_t(R_{t+1}) = 0.55$. If we forced the shocks to be drawn from a Gaussian distribution, instead we would have to resort

to a much higher persistence of Δ≅0.94, to obtain $ExKurt(R_{t+1}) \simeq 0.61$ while $ExKurt_t(R_{t+1}) = 0$. In other words, persistence has to be inflated from 0.6 to 0.94 simply to match the fourth moment of the data, regardless of the false view of the nature of the process that this implies.

Δ	ν	$ExKurt(R_{t+1})$	$ExKurt_t(R_{t+1})$
0	4.01	300	300
0.99	4.01	2,565	300
0.8	9	1.80	1.20
0.99	20	390	0.375
0.99	∞	23.2	0
0.6	15	0.62	0.55
0.94	∞	0.61	0

On-Line Ex. 5D. Among the large number of predetermined variables that have been proposed in the empirical literature, one (family) has recently acquired considerable importance in exercises aimed at forecasting variance: **option implied volatilities**, and in particular the (square of the) CBOE's (Chicago Board Options Exchange) VXO and VIX indices as well as other functions and transformations of the same.

The VXO represents a weighted, average implied volatility (IV) computed on S&P 100 index options and offers a longer time series vs. VIX, that instead concerns implied volatilities computed on S&P 500 index options. As discussed in Poon and Granger (2005), IV tends to be more accurate than GARCH and related models at predicting future variance, even though this is surprising because IV is normally based on a larger and timelier information set that is by construction forward looking. However, options are written on a limited number of assets and indices: for instance, emerging market equity and bond indices and small stocks are important building blocks of optimal portfolios but there are no options written on them. So, the time-series models covered in this book, although inferior to option-implied models, will continue to play an important role going forward.

In general, models that use explanatory variables to capture time-variation in variance are represented as:

$$\sigma^2_{t+1|t} = \omega + g(\mathbf{X}_t) + \alpha\sigma^2_{t|t-1}z^2_t + \beta\sigma^2_{t-1|t}, \qquad (5D.1)$$

which is one more case of augmented GARCH and in which $\mathbf{X}_t$ is a vector of predetermined variables that may as well include implied volatility. Note that because this volatility model is not written in log-exponential form, it is important to ensure that the model always generates a positive variance forecast, which will require that restrictions—either of an economic type or at least in the form of mathematical constraints to be numerically imposed during estimation—must be satisfied, to ensure that $g(\mathbf{X}_t) > 0$ for all possible values of $\mathbf{X}_t$, besides the usual ω, α, β≥ 0 (with one positive).

When $\mathbf{X}_t$ consists of implied variance (say VXO for concreteness),

$$R_{t+1} = \mu + \sigma_{t+1|t} z_{t+1} \text{ with } z_{t+1} \sim \text{IID } N(0,1)$$
$$\sigma^2_{t+1|t} = \omega + \alpha \varepsilon_t^2 + \beta \sigma^2_{t|t-1} + \lambda VXO_t \quad (5D.2)$$

then there are interesting implications to explore. Assume that *VXO* follows a stationary autoregressive first-order process, $VXO_t = \delta_0 + \delta_1 VXO_{t-1} + \zeta_t$ with ζ_t white noise. The expression for the unconditional variance remains easy to derive but it will be influenced by the fact that over the long-run, on average, also *VXO* can be taken to represent a predictor of variance. If the process for *VXO* is stationary, we know that $0<|\delta_1|<1$ and from

$$E[VXO_t] = \delta_0 + \delta_1 E[VXO_{t-1}] \Rightarrow E[VXO_t] = E[VXO_{t-1}] = \frac{\delta_0}{1-\delta_1}, \quad (5D.3)$$

which is finite because $|\delta_1|<1$, we have:

$$E[\sigma^2_{t+1|t}] = \omega + \alpha E[\varepsilon_t^2] + \beta E[\sigma^2_{t+1|t}] + \lambda E[VXO_t]$$
$$= \omega + (\alpha+\beta) E[\sigma^2_{t+1|t}] + \lambda \frac{\delta_0}{1-\delta_1} \Rightarrow E[\sigma^2_{t+1|t}] = \frac{\omega + \lambda \frac{\delta_0}{1-\delta_1}}{1-\alpha-\beta}. \quad (5D.4)$$

One may actually make more progress by imposing economic restrictions. For instance, taking into account that, if the options markets are efficient, then $E[VXO_t] = E[\sigma^2_{t+1|t}]$ obtains, one can establish a further connection between the parameters δ_0 and δ_1 on one side, and ω, α, and:[1]

$$E[\sigma^2_{t+1|t}] = \omega + \alpha E[\varepsilon_t^2] + \beta E[\sigma^2_{t+1|t}] + \lambda E[VIX_t]$$
$$= \omega + (\alpha+\beta) E[\sigma^2_{t+1|t}] + \lambda E[\sigma^2_{t+1|t}] \Rightarrow E[\sigma^2_{t+1|t}] = \frac{\omega}{1-\alpha-\beta-\lambda}. \quad (5D.5)$$

Because $E[\sigma^2_{t+1|t}] = \delta_0/(1-\delta_1)$ but also $E[\sigma^2_{t+1|t}] = \omega/(1-\alpha-\beta-\lambda)$, we derive the restriction that

$$\delta_0/(1-\delta_1) = \frac{\omega}{(1-\alpha-\beta-\lambda)} \quad (5D.6)$$

should hold, which is an interesting and testable restriction.

We start by regressing log-gross monthly US value-weighted returns on the log of VXO divided by 1200. The latter transformation is required by the fact that since 1986, CBOE has been reporting VXO as a percentage annualized volatility, while here we need a monthly series comparable to the log of gross returns. The resulting series of log-monthly VXO data contains however a unit root on the basis of an augmented Dickey-Fuller test and a simple regression of log-returns on long-monthly VXO would represent an unbalanced regression that, when estimated by OLS on a 1986-2016 sample, gives:

$$\ln(1+R_{t+1}) = \underset{(0.000)}{-0.123} \ \underset{(0.000)}{-0.046} \ln\left(\frac{VXO_t}{1200}\right) + e_{t+1},$$

with an R-square of 6.93%. Using the formula of Example 5.19, the resulting RMSPE for squared monthly returns is 38.01. For comparison, a GARCH(1,1) model, gives a RMSPE of 39.45.

In any event, we also proceed to estimate a balanced predictive regression:

$$\ln(1+R_{t+1}) = \underset{(0.000)}{0.008} \ \underset{(0.000)}{-0.149} \Delta \ln\left(\frac{VXO_t}{1200}\right) + e_{t+1},$$

with a striking R-square of 41.1%: it is not really a high implied volatility that forecasts

[1] For the asset pricing buffs, $E[VXO_t] = E[\sigma^2_{t+1|t}]$ may present some problems, as VXO (and VIX) is normally calculated under the risk-neutral measure while $E[\sigma^2_{t+1|t}]$ refers to the physical measure. Strictly speaking, the equality only holds assuming (at least, local) risk-neutrality.

negative returns, it is an increasing implied volatility that does so. The resulting RMSPE is interesting, only 34.0.

An interesting alternative is to use VXO not in alternative to GARCH but along with GARCH:

$$R_{t+1} = \underset{(0.000)}{1.725} + \varepsilon_{t+1} \qquad \varepsilon_{t+1} \text{ IID } N(0,\sigma^2_{t+1|t})$$

$$\sigma^2_{t+1|t} = \underset{(0.194)}{0.078} + \underset{(0.108)}{0.025}\varepsilon_t^2 + \underset{(0.000)}{0.976}\sigma^2_{t|t-1} + \underset{(0.000)}{0.465}\Delta\left(\frac{VXO_t^2}{12}\right)$$

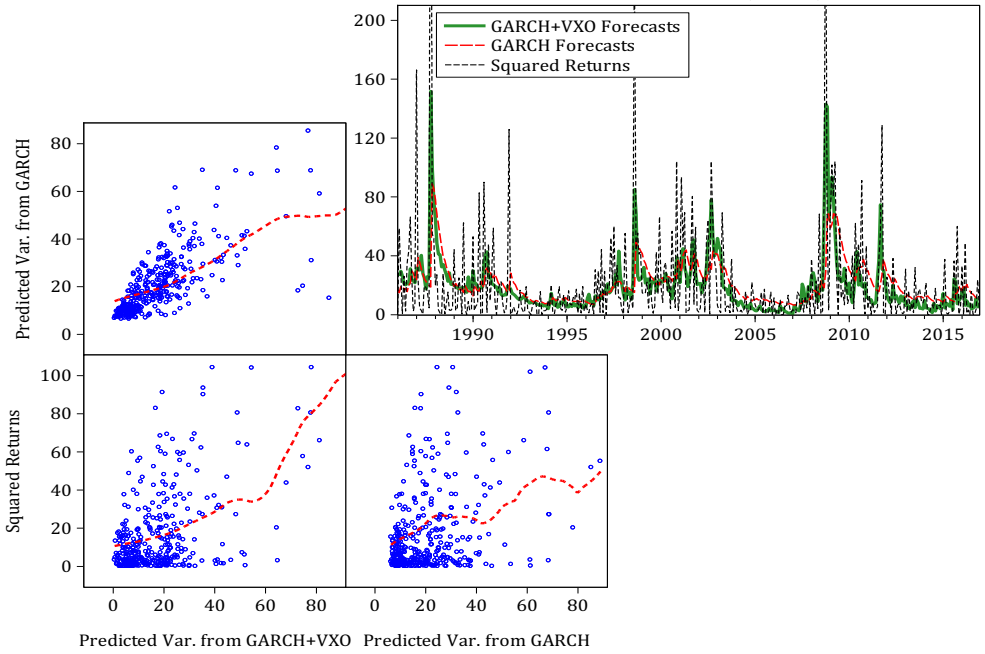


Figure 5.D1 –Improvement in In-Sample Forecasting Accuracy from Using Implied Volatility Predictions

Even though the GARCH model features a small and imprecisely estimated alpha coefficient, the resulting RMSPE is low, 30.99 only, reducing by almost 25% the RMSPE of a plain-vanilla GARCH model. This much is the value of using implied volatility in addition to time series methods in variance forecasting. Figure 5.D1 shows why augmenting a standard GARCH(1,1) model with past VXO values yields more accurate forecast of one-month ahead squared realized returns. Moreover, the scatter plots show that while the plain vanilla and VXO-augmented predictions are very similar for squared values between 0 and 40, for more extreme values the GARCH model faces limitations in predicting spikes, and when the VXO is also employed, this occurs less. In fact, above 40, GARCH forecasts stop reacting to information, while VXO can still yield predictions in excess of 70, which was very useful to forecast variance during the Great Financial Crisis and then again in 2011.

On-Line Ex. 5E. The most famous NIC functional forms are derived by simply extending the GARCH NIC, $NIC(z_t \mid \sigma_t^2 = \sigma^2) = A + \alpha\sigma^2 z_t^2$ (where $A \equiv \omega + \beta\sigma^2 > 0$) to a family of volatility models parameterized by θ_1, θ_2, and θ_3 that can be written as follows:

$$NIC(z_t) = [|z_t - \theta_1| - \theta_2(z_t - \theta_1)]^{2\theta_3}. \qquad (5E.1)$$

The objective is then to estimate the parameters (θ_1, θ_2, θ_3, and β) of models with structure:

$$\sigma^2_{t+1|t} = NIC(z_t) + \beta\sigma^2_{t|t-1} = \omega + [|z_t - \theta_1| - \theta_2(z_t - \theta_1)]^{2\theta_3} + \beta\sigma^2_{t|t-1} \qquad (5E.2)$$

One can retrieve a standard, plain vanilla GARCH(1,1) by setting $\theta_1 = 0$, $\theta_2 = 0$, and $\theta_3 = 1$. Another important case that we have already encountered in Section 5.2.6, the NA-GARCH(1,1) model, that can be obtained from (5E.1) by setting $\theta_2 = 0$ and $\theta_3 = 1$. Under

these restrictions, the NIC becomes $NIC(z_t)=(|z_t-\theta_1|)^2=(z_t-\theta_1)^2$ (squaring an absolute value makes the absolute value operator irrelevant, i.e., $|f(x)|^2=(f(x))^2$) and an asymmetry derives from the fact that when $\theta_1>0$,

$$(z_t-\theta_1)^2=\begin{cases}(z_t-\theta_1)^2<z_t^2 & \text{if } z_t\geq 0\\ (z_t-\theta_1)^2>z_t^2 & \text{if } z_t<0\end{cases}, \quad (5E.3)$$

in words, while positive standardized errors are reduced by $\theta_1>0$, negative news are magnified in their impact on subsequent variance.

On-Line Ex. 5F. Extending Example 5.22, we also compute the per-week cumulative variance forecasts of approximate 5-year Treasury note returns at the end of the sample for horizons H that vary between 1 week and the end of 2027. We also estimate a IGARCH(1) model with t-Student innovations and proceed to compute the same per-week cumulative variance forecasts. The estimated RiskMetrics model is:

$$R_{t+1}=\underset{(0.572)}{-0.002}+\underset{(0.000)}{0.254}R_t+\underset{(0.193)}{0.374}\sigma^2_{t|t-1}+\varepsilon_{t+1} \quad \varepsilon_{t+1}\ \text{IID}\ \text{t}(0,\sigma^2_{t+1|t};\ \underset{(0.000)}{10.37})$$

$$\sigma^2_{t+1|t}=\underset{(0.000)}{0.050}\varepsilon_t^2+\underset{(0.000)}{0.950}\ \sigma^2_{t|t-1},$$

Figure 5F.1 compares the variance forecasts of cumulative returns from the two models. As expected, as the horizon grows, forecasts from a stationary GARCH(1,1) stabilize in correspondence to $\bar{\sigma}^2$, while the average, per-period forecast from RiskMetrics is "stuck" in correspondence to the most recent forecast, $\sigma^{2,RiskMcs}_{t+1|t}$.

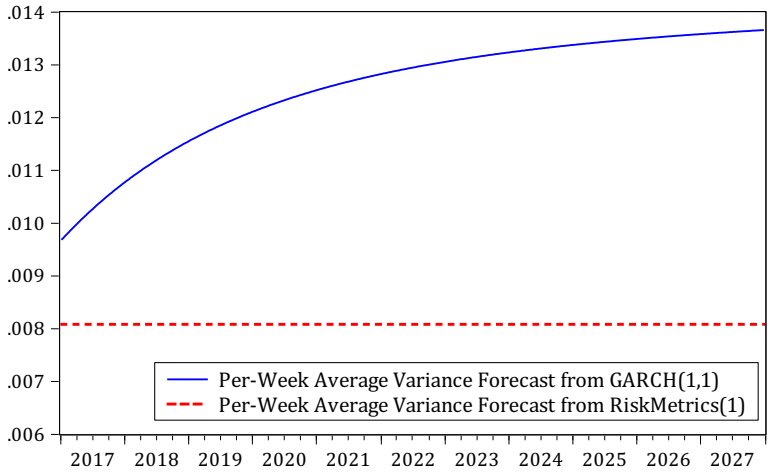


Figure 5F.1 –Comparing Per-Week Average Cumulative Variance Forecasts from t-Student Stationary GARCH(1,1) and RiskMetrics(1)

Of course, if were to repeat this experiment in correspondence to a forecast origin at which $\sigma^2_{t+1|t}>\bar{\sigma}^2$ we would find a monotone declining shape of the GARCH(1,1) prediction as a function of *H*.

On-Line Ex. 5G. Consider the four Fama-French-Carhart factor portfolios often used in asset pricing and asset management. For each of the value-weighted excess market, SMB, HML, and momentum (these were all previously defined) portfolio returns, we have the following sample moments and implied MM estimates of μ, σ, and ν in constant mean, homoskedastic process $R_{t+1}=\mu+\sigma z_{t+1}$, z_{t+1} IID t(0,1;ν):

Ptf.	Sample Mean	Sample Vol.	Sample Ex. Kurt.	μ	σ	ν
Market	0.025	0.986	15.717	0.025	0.727	4.382
SMB	0.007	0.522	24.457	0.007	0.402	4.245
HML	0.019	0.499	10.856	0.019	0.374	4.553
Mom	0.030	0.702	14.799	0.030	0.519	4.405

Interestingly, all portfolios display massive excess kurtosis, hence they cannot be modeled as having a marginal normal distribution and as a result all the estimated MM values for ν fall between 4.25 and 4.55.

On-Line Supp. 5H. How do we proceed to maximize the log-likelihood function of a sample by selecting the optimizing parameters, subject to $\theta \in \Theta$? Appropriate **methods of numerical, constrained optimization** need to be implemented: this is what packages such as Matlab, Gauss, E-Views, or Stata are for. For instance (i.e., other, better but more complex methods are feasible), **Newton's method** makes use of the **Hessian**, which is a $K \times K$ matrix $H(\boldsymbol{\theta}) \equiv \partial^2 (\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$ that collects second partial derivatives of the log-likelihood function with respect to each of the parameters in $\boldsymbol{\Theta}$. Similarly the $K \times 1$ **gradient** $\partial (\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ 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 $\partial (\hat{\boldsymbol{\theta}}_j) / \partial \boldsymbol{\theta}$ and $\mathrm{H}(\hat{\boldsymbol{\theta}}_j)$ denote, respectively, the gradient and the Hessian evaluated at $\hat{\boldsymbol{\theta}}_j$. Then the fundamental equation to update the estimates according to Newton's algorithm is:

$$\hat{\boldsymbol{\theta}}_{j+1} = \hat{\boldsymbol{\theta}}_j - \mathrm{H}^{-1}(\hat{\boldsymbol{\theta}}_j)[\partial (\hat{\boldsymbol{\theta}}_j) / \partial \boldsymbol{\theta}] \qquad (5\mathrm{H}.1)$$

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. The maximization process therefore proceeds through the following steps:

- Set an initial vector of parameters, $\hat{\boldsymbol{\theta}}_0$, and compute $\mathrm{H}^{-1}(\hat{\boldsymbol{\theta}}_0)$ and $\partial (\hat{\boldsymbol{\theta}}_0) / \partial \boldsymbol{\theta}$.
- Compute the new vector of estimated parameters $\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\theta}}_0 - \mathrm{H}^{-1}(\hat{\boldsymbol{\theta}}_0)[\partial (\hat{\boldsymbol{\theta}}_0) / \partial \boldsymbol{\theta}]$ and therefore $\mathrm{H}^{-1}(\hat{\boldsymbol{\theta}}_1)$ and $\partial (\hat{\boldsymbol{\theta}}_1) / \partial \boldsymbol{\theta}$; check that the **Euclidean norm** $|| \hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0 ||$ (in words, this is the square root of the sum of all squared differences between the elements of $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_0$) is not inferior to some small threshold parameter (typically, 10^{-5}).
- Update the vector of parameter estimates to $\hat{\boldsymbol{\theta}}_2 = \hat{\boldsymbol{\theta}}_1 - \mathrm{H}^{-1}(\hat{\boldsymbol{\theta}}_1)[\partial (\hat{\boldsymbol{\theta}}_1) / \partial \boldsymbol{\theta}]$ and check that the norm $|| \hat{\boldsymbol{\theta}}_2 - \hat{\boldsymbol{\theta}}_1 ||$ is not inferior to the threshold parameter.
- Continue (unless a maximum number of iteration has been exceeded, but with fast computers often thousands of iterations are affordable in the space of a few minutes only) until $\hat{\boldsymbol{\theta}}_j = \hat{\boldsymbol{\theta}}_{j-1} - \mathrm{H}^{-1}(\hat{\boldsymbol{\theta}}_{j-1})[\partial (\hat{\boldsymbol{\theta}}_{j-1}) / \partial \boldsymbol{\theta}]$ is such that $|| \hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}_{j-1} ||$ falls below the

fixed convergence threshold, that signals that the optimizing vector has stopped changing.

- Set $\hat{\boldsymbol{\theta}}_T^{ML} = \hat{\boldsymbol{\theta}}_j$.

Numerical optimization is a very sensitive business; a myriad of choices are considered to be crucial to obtain "reliable" results, such as the initial value $\hat{\boldsymbol{\theta}}_0$, the convergence tolerance criterion, and often how much the algorithm is supposed to "travel" in the direction indicated by the inverse Hessian matrix, i.e., the coefficient τ in the iteration in (5.H1), generalized to read as $\hat{\boldsymbol{\theta}}_{j+1} = \hat{\boldsymbol{\theta}}_j - \tau \mathrm{H}^{-1}(\hat{\boldsymbol{\theta}}_j)[\partial\, (\hat{\boldsymbol{\theta}}_j)/\partial\boldsymbol{\theta}]$, where $\tau > 0$ (clearly a $\tau < 1$ "dims" the step taken in direction $[\partial\, (\hat{\boldsymbol{\theta}}_j)/\partial\boldsymbol{\theta}]$, while a $\tau > 1$ acts as a multiplier). Reliability here is often evidence or even taken to offer some guarantee that $\hat{\boldsymbol{\theta}}_T^{ML} = \hat{\boldsymbol{\theta}}_j$ truly represents a **global** (as opposed to local) **maximizer** of the log-likelihood function and as such it is unique, as assumed. For instance, just to get hard evidence on this aspect, it is often advised to start off the maximization algorithm in correspondence of a range of alternative starting values and then retain, for the true and often lengthy iterative Newton-style search, the most promising one(s).

Other numerical optimization methods are of course possible. A few of them are faster than Newton's method because they replace the Hessian matrix with cheaper to compute negative definite $K \times K$ matrices, for instance $OPG(\boldsymbol{\theta}) \equiv -[\partial\, (\boldsymbol{\theta})/\partial\boldsymbol{\theta}][\partial\, (\boldsymbol{\theta})/\partial\boldsymbol{\theta}]'$, which is negative definite by construction, unless $\partial\, (\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$, which would instead show that a stationary point has been reached. The advantage of this expression is that it only requires calculation (often numerically) of first-order derivatives. Moreover, our simplified illustration of Newton's method ignores the role played by constraints, that may interfere with setting $\hat{\boldsymbol{\theta}}_j = \hat{\boldsymbol{\theta}}_{j-1} - \mathrm{H}^{-1}(\hat{\boldsymbol{\theta}}_{j-1})\,[\partial\, (\hat{\boldsymbol{\theta}}_{j-1})/\partial\boldsymbol{\theta}]$, when the constraints are violated.

On-Line Ex. 5I. As an example of calculations of confidence intervals based on the last set of ML estimates in Example 5.24, we have:

Coefficient	99% CI	95% CI	90% CI	Coefficient	90% CI	95% CI	99% CI
$\kappa 0$	0.022	0.026	0.028	0.039	0.049	0.051	0.055
$\kappa 1$	0.099	0.105	0.108	0.122	0.137	0.140	0.145
ω	0.004	0.005	0.005	0.007	0.008	0.008	0.009
α	0.012	0.014	0.016	0.024	0.032	0.033	0.036
δ	0.092	0.096	0.099	0.111	0.124	0.126	0.131
β	0.903	0.906	0.907	0.914	0.922	0.923	0.926
ν	6.712	6.978	7.114	7.823	8.532	8.668	8.933

Table 5I.1 –90%, 95%, and 99% Asymptotic Confidence Intervals from t-Student MA(1), Threshold GARCH(1,1,1) for US Excess Stock Returns

Clearly, none of the intervals contains zero as a lower bound, and this derives from the fact that all coefficients had been found to be significant with *p*-values lower than 1% early on (note that here the widest interval reported is computed at confidence of 1–*p*=99%). It is also useful to look at the **joint, pairwise confidence ellipses** of the estimated ML parameters considered in pairs, as we do in Figure 5I.1 for the GARCH related parameters α, δ, β, and ν (the remaining parameters have been dropped just to keep the figure readable).

To familiarize with a (1–*p*)% confidence ellipse, consider the top, rightmost panel of Figure 5I.1, where we have the 95% ellipse involving the persistence ($\alpha + 0.5\delta + \beta$) index of the *t*-Student threshold GARCH(1,1,1) model and the "number of degrees-of-freedom"

parameter, ν. The region inside the circular ellipse includes the infinite combinations of the persistence index and of ν that can be assigned a confidence of 0.95, i.e., in a frequentist sense, that shall characterize the parameter estimates in 95% of all samples of a size identical to the sample that has been used to obtain parameter estimates in this example. For instance, a combination of a persistence index of 0.998 and of a value for ν that equals 9, should occur rather infrequently. The dot at the center of the ellipse—in this case at approximately a persistence index of 0.994 and a value of ν equal to 7.9—corresponds to the point estimates reported in Table 5I.1. The four dotted lines represent for each pair, the same 95% confidence intervals in Table 5I.1, for instance ν falls with probability 95% between 6.9 and 8.7; interestingly, the persistence index is 95% of the time between 0.989 and 0.998, i.e., it is estimated to be high and in excess of 0.99 with considerable precision. Finally, perfectly round ellipses indicate that there is approximate zero correlation between ML estimates in pair; ellipses that are "slanted" to the left (right), like the case of the persistence index and ν, indicate the existence of negative (positive) correlation.

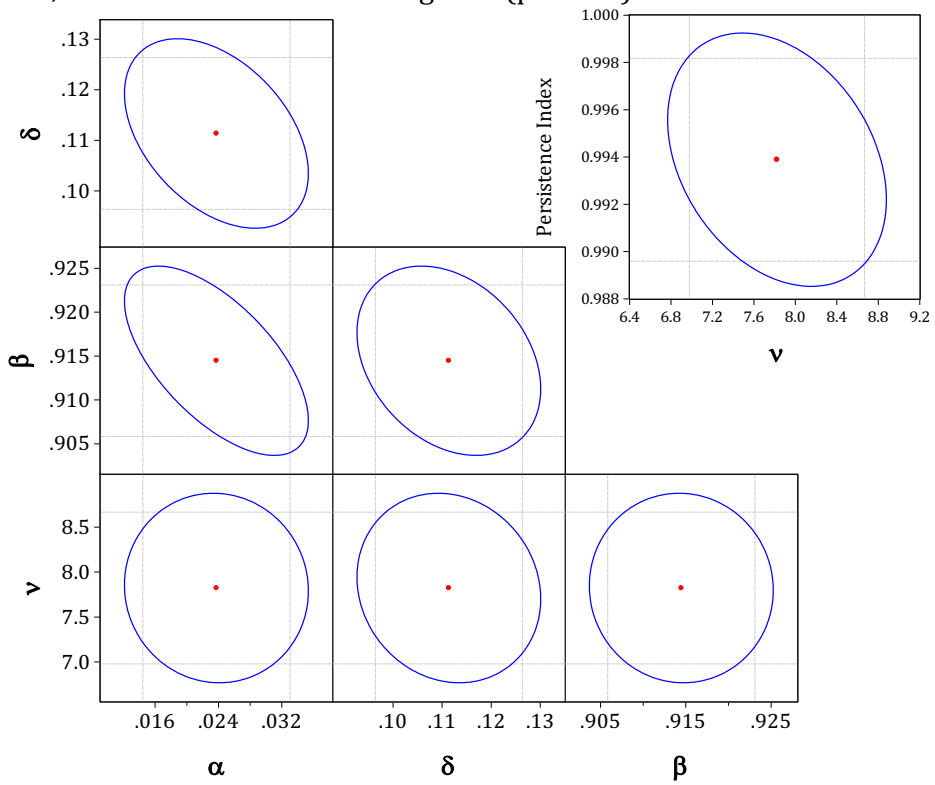


Figure 5I.1 – 95% Confidence Ellipses for Pairs of ML Parameter Estimates from t-Student MA(1), Threshold GARCH(1,1,1) for US Excess Stock Returns

The Figure ends up showing the existence of a rather strong and negative correlation between ML estimates of α and δ (which is expected, given their similar meaning and interpretation), and also α and β. As already remarked, fatter-tailed distributions for the errors tend to lead to higher and not lower estimates of the GARCH persistence, that is contrary to what sometimes thought.

On-Line Ex. 5L. Consider again the simple Gaussian MA(1)-threshold GARCH(1,1) estimated in Example 5I above on daily US excess stock returns:

$$x_{t+1} = \underset{(0.000)}{0.026} + \underset{(0.000)}{0.129}\,\varepsilon_t + \sigma_{t+1|t}^2 z_{t+1} \qquad z_{t+1} \text{ IID N}(0,1)$$

$$\sigma_{t+1|t}^2 = \underset{(0.000)}{0.009} + \underset{(0.000)}{0.024}\,\varepsilon_t^2 + \underset{(0.000)}{0.112}\,I_{\{\varepsilon_t<0\}}\varepsilon_t^2 + \underset{(0.000)}{0.911}\,\sigma_{t|t-1}^2,$$

The Jarque-Bera test on the standardized residuals from this model is $JB(z) = 4095.1$ which commands a p-value of zero. The model as such is rejected and the assumption

$z_{t+1} \sim \text{IID } N(0,1)$ can at most be retained in a QMLE framework, i.e., only as an approximation. The following histogram and nonparametric (kernel) density estimator let us appreciate what the key source of non-normality is in the case of these data: even though the excess kurtosis ends up exceeding 2.5, the sample skewness of -0.40 appears to be too large for normality to fare well.

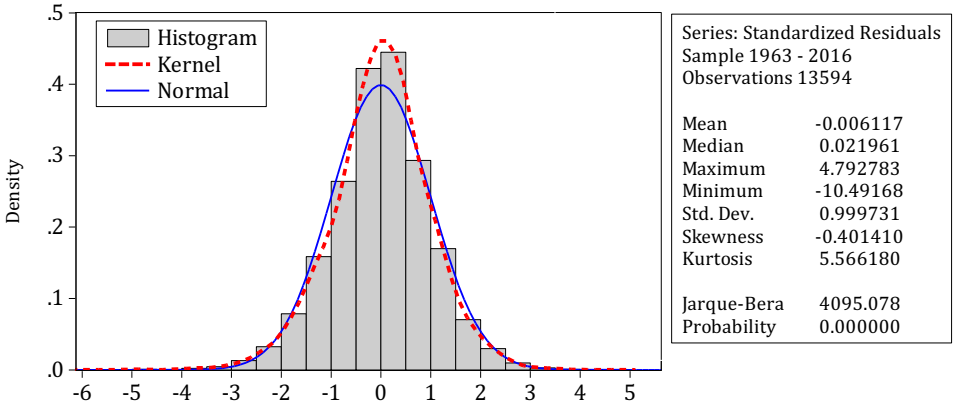


Figure 5L.1 – Histogram and Key Summary Statistics for Standardized Residuals Obtained from a Gaussian threshold GARCH(1,1,1)

In case you are wondering, we have estimated and analyzed the normality of standardized residuals from the entire range of models featured in the examples that appear in Chapter 5 with references to this particular series, finding results that are not qualitatively different from Figure 5L.1: it is hard for textbook GARCH models to capture the thick tails of the data, especially the left tail that therefore commands a large and statistically significant negative skewness.