Today’s Agenda

1. The importance of an economic model
2. Spurious regressions
3. Time-Varying Variances: ARCH/GARCH, Realized Volatility, MIDAS Volatility, Stochastic Volatility
4. GMM—Formal Treatment
   – Asset Pricing Models
5. Maximum Likelihood Estimation
   – ARCH estimation
   – GARCH estimation
1 The Importance of an Economic Model: The Gordon-Growth Model in the Risk-Return Relation

• Suppose we have an estimate of \( \sigma_t \).

• We can run the regression

\[
    r_t = \alpha + \beta \sigma_{t-1} + \varepsilon_t
\]

• Under the null hypothesis that there is a risk-return trade-off, we expect \( \beta > 0 \).

• If the relationship does not hold, there are several possibilities:
  – The relationship is not well specified (perhaps see VAR)
  – The volatility is not well estimated
  – There are other variables that must enter into a VAR (investment opportunity set is time-varying).
  – A combination of the above

• Surprisingly, thus far, the evidence for \( \hat{\beta} > 0 \) is not conclusive.

• Several papers find \( \hat{\beta} < 0 \).
Recall the definition of a return:

\[ R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t} - 1 \]

which can be re-written, if we assume that \( E_t R_{t+1} = R \), as:

\[
P_t = E_t \left[ \frac{P_{t+1} + D_{t+1}}{1 + R} \right]
\]

\[
= E_t \left[ \sum_{i=1}^{K} \left( \frac{1}{1 + R} \right)^i D_{t+i} \right] + E_t \left[ \left( \frac{1}{1 + R} \right)^K P_{t+K} \right]
\]

\[
\simeq E_t \left[ \sum_{i=1}^{\infty} \left( \frac{1}{1 + R} \right)^i D_{t+i} \right]
\]

Note:

\[ E_t P_{t+1} = (1 + R) P_t - E_t D_{t+1} \]

1. A constant expected return (as assumed) does not imply that \( P_t \) would follow a martingale.

2. Recall: To obtain a martingale, we must construct a portfolio for which all dividend payments are re-invested in the stock. The value of this portfolio is a martingale (Campbell et al. p. 257)!
Now, suppose that $D_{t+1} = (1 + G) D_t + \varepsilon_t$, where $\varepsilon_t$ is iid. Then $E_t D_{t+i} = (1 + G)^i D_t$ and

$$P_t = E_t \left[ \sum_{i=1}^{\infty} \left( \frac{1}{1+R} \right)^i D_{t+i} \right]$$

$$= \frac{(1 + G) D_t}{R - G} = \frac{E_t D_{t+1}}{R - G}$$

This is called the Gordon (constant) growth model.

Note the unrealistic assumptions: $E_t R_{t+1} = R$ and $G$ are constant.

This is how people think about asset prices.

But the definition seems circuitous: Prices are determined by the discount rate, what is the discount rate determined by?
• But it must be the case that (I am cheating a bit...changing assumptions):

\[ E_t R_{t+1} = a + \gamma \sigma_t^2 \]

• Hence:

\[ P_t \approx \frac{E_t D_{t+1}}{E_t R_{t+1} - G} \]

• Q: Why is this useful?
• A: Suppose $\alpha = 0.08$ and $\gamma = 2.5$, and $\sigma_t^2 = 0.15^2$. Then $E_t R_{t+1} = 0.1362$.

• Suppose that $E_t D_{t+1} = 1$, and $G = 0$. Then:
  
  $$P_t \approx \frac{1}{E_t R_{t+1}} = \frac{1}{0.1362} = 7.34$$

• Now, suppose that volatility increases at $t + 1$, or $\sigma_{t+1}^2 = 0.25^2$. Hence, $E_{t+1} R_{t+2} = 0.2363$. This implies that:
  
  $$P_{t+1} \approx \frac{1}{0.2363} = 4.23$$

• What happened to $R_{t+1}$ (no dividend got paid)?
  
  $$R_{t+1} = \frac{P_{t+1} - P_t}{P_t} = \frac{4.23 - 7.34}{7.34} < 0$$

• MORAL: $COV \left( R_{t+1}, \sigma_{t+1}^2 \right) < 0$.

• BUT: $COV( R_{t+1}, \sigma_t^2 ) > 0$, on average.

• ALSO: $COV \left( R_{t+1}, \Delta \sigma_{t+1}^2 \right) < 0$. 


• BTW, this conclusion holds for any state variable
  – Any variable that fluctuates with expected returns:
    \[ E_t R_{t+1} = a + \gamma X_t \]
  – We can show that
    * \( COV(R_{t+1}, X_{t+1}) < 0 \) (contemporaneous relation)
    * \( COV(R_{t+1}, X_t) > 0 \) (predictive regression)

• Insight: If the sign of the model and the regression are in disagreement:
  – Check your model
  – Check the lag and the timing of the data
2 Spurious Regressions and Non-Stationary Time Series

• Thus far, we have been looking at stationary time series. We have focused on $r_t$, $\sigma^2_t$, assuming that they are stationary. For some series, this assumption is more tenable than for others.

• But suppose you want to work with non-stationary time-series, i.e. prices, volume, number of investors in a particular fund, number of funds, etc. Those processes are inherently non-stationary.

• Let $p_t$ be the log-price. We know that
  \[ p_t = p_{t-1} + \varepsilon_t \]
  or $p_t$ is an AR(1) process with an unit-root.

• This process is non-stationary. We cannot apply the CLT.

• But we are still interested in testing the null $\phi = 1$ versus $\phi < 1$.

• Problem. Under the null, the process is non-stationary.

• Under the alternative, the process is stationary.
• It turns out that (Functional CLT)

\[
\frac{1}{\sqrt{T}}p_t = \frac{1}{\sqrt{T}} \sum_{s=1}^{t} \varepsilon_s
\]

\[
= \frac{1}{\sqrt{T}} \sum_{s=1}^{[rT]} \varepsilon_s \Rightarrow W(r)
\]

where \( W(r) \) is a Brownian motion on \([0, 1]\). I.e., \( W(r) \sim N(0, r), 0 < r < 1 \).

• Note that this result is asymptotic. We don’t have to assume that the \( \varepsilon' \)s are normal (hence the Functional CLT, or FCLT).

• Q: Can’t we standardize the non-stationary processes by a power of \( T \) in order for them to converge.

• A: Yes.

• Let’s get a “flavor” of how things work:
Recall that
\[
\hat{\phi} = \frac{\sum p_t p_{t-1}}{\sum p_{t-1}^2} = \frac{\sum p_{t-1} (\phi p_{t-1} + \varepsilon_t)}{\sum p_{t-1}^2} = \phi + \frac{\sum p_{t-1} \varepsilon_t}{\sum p_{t-1}^2}
\]

If \( \phi < 1 \), we had
\[
\hat{\phi} = \phi + \frac{1}{T} \sum \varepsilon_t p_{t-1} \rightarrow^p \phi
\]
\[
\sqrt{T} \left( \hat{\phi} - \phi \right) \sim N \left( 0, \sigma_{\hat{\phi}}^2 \right)
\]

But if \( \phi = 1 \), the results do not hold. But
\[
\hat{\phi} = \phi + \frac{\sum \varepsilon_t p_{t-1}}{\sum p_{t-1}^2}
\]
\[
T \left( \hat{\phi} - \phi \right) \Rightarrow O_p (1)
\]

In other words, \( \hat{\phi} \) is super-consistent.

Q: But since we don’t know the distribution of \( \hat{\phi} \), can we use this result for testing?

A: Yes, if we simulate the distribution.
• Dickey-Fuller (DF) Test:
  \[ H_o : \phi = 1 \]
  \[ H_a : \phi < 1 \]
• The test is: \[ t = \frac{\hat{\phi} - 1}{se(\hat{\phi})} \]
• Suppose that \( \varepsilon_t \) follows an AR(p) process. The distribution of the DF test is influenced by those parameters. Not good.
• To get rid of those parameters, we run the following regression:
  \[ p_t = \phi p_{t-1} + \zeta_1 \Delta p_{t-1} + \zeta_2 \Delta p_{t-2} + \ldots + \zeta_k \Delta p_{t-k} + u_t \]
• Then, \[ t = \frac{\hat{\phi} - 1}{se(\hat{\phi})} \].
• This is called the Augmented DF, or ADF test.
Summary of ADF test–Testing for a unit root:
- Regress $p_t$ on $p_{t-1}, \Delta p_{t-1}, \Delta p_{t-2}, \ldots, \Delta p_{t=k}$
- Form: $t = \frac{\hat{\phi}^{-1}}{se(\hat{\phi})}$
- Get the critical value from simulations.
• So, is working with non-stationary variables that easy?
• NO:
• Suppose $p^1_t$ and $p^2_t$ are the prices of the same asset traded on two markets. Then, it must be the case that

$$p^1_t = p^2_t$$

• Empirically, this is almost true. We find

$$p^1_t - p^2_t = \varepsilon_t$$

where $\varepsilon_t$ is almost iid, and $E(\varepsilon_t) = 0$.

• How do we take advantage of this?
• Regress:

$$p^1_t = \gamma p^2_t + \varepsilon_t$$

• If

$$\varepsilon_t > 0$$

$$p^1_t > \gamma p^2_t$$

• The asset is “too expensive” in market 1.
• Easy, right?
Wrong!

Suppose we have two assets, with prices $p_t$ and $q_t$. One might be tempted to look for arbitrage strategies as in

$$p_t = \gamma q_t + \varepsilon_t$$

If $\gamma > 0$, there is a relationship, and we can trade.

No.

Suppose

$$p_t = p_{t-1} + u_t$$
$$q_t = q_{t-1} + v_t$$
$$\text{cov}(u_tv_t) = 0$$

Note: The two log-prices represent two independent discrete Brownian motions.
• But:

\[
\hat{\gamma} = \frac{\sum_{t=1}^{T} p_t q_t}{\sum_{t=1}^{T} q_t^2} = \frac{\sum_{t=1}^{T} \left( \sum_{s=1}^{t} v_s \right) \left( \sum_{s=1}^{t} u_s \right)}{\sum_{t=1}^{T} \left( \sum_{s=1}^{t} u_s \right)^2} = \frac{1}{T^2} \sum_{t=1}^{T} \left( \sum_{s=1}^{t} v_s \right) \left( \sum_{s=1}^{t} u_s \right) \frac{1}{\sum_{t=1}^{T} \left( \sum_{s=1}^{t} u_s \right)^2} = O_p(1)
\]

• Similarly:

\[
t = \frac{\hat{\gamma}}{se(\hat{\gamma})}
\]

is not consistent.

• The \(R^2\) does not converge to 0, as \(T \rightarrow \infty\).

• Illustration: spurious.m
Co-integration:

- Suppose that $p_t$ and $q_t$ are unit root (integrated) processes, but there is a linear combination of $p_t$ and $q_t$ that is stationary. That is, there exists a vector $\gamma = \begin{bmatrix} 1 & -\gamma_1 \end{bmatrix}$ such that
  \[ p_t - \gamma q_t = \varepsilon_t \]
  and $\varepsilon_t$ is a stationary process. Then, $p_t$ and $q_t$ are said to be cointegrated.

- There are formal tests for cointegration, but they have low power against the alternative. Why? (Think spurious correlation).

- Cointegration is only between contemporaneous variables. I.e. $p_t - \gamma q_t$ is a cointegrating vector. But $\Delta p_t$ is not. However, the latter is also a way of stationarizing the process.

- Cointegration occurs “naturally” in economics. It is dictated by theory.
• Examples:
  – Prices $p^1_t$ and $p^2_t$ of the same asset traded on two markets.
  – Dividend price ratio is cointegrated, or:
    $$d_t - p_t$$
    must be stationary
  – The long and the short rate must be cointegrated
  – Consumption and GDP must be cointegrated, etc.
  – Lettau & Ludvigson (2001), the “cay” ratio:
    $$c_t - \gamma_1 a_t - \gamma_2 y_t$$
3 Time-Varying Variances

• Suppose we have the returns process \( \{r_t\}_{t=1}^T \).

• First, we model the conditional mean, for example as

\[
    r_t = \mu + \phi r_{t-1} + u_t
\]

• We know that the unconditional first moments are

\[
    E(r_t) = \frac{\mu}{1-\phi} \quad \text{and} \quad Var(r_t) = \frac{\sigma_u^2}{1-\phi^2}
\]

• We also know that the conditional first moment

\[
    E_{t-1}(r_t) = \mu + \phi r_t \quad \text{is time varying, even though the unconditional moment is not!}
\]

• Q: Can we also have the same situation for the second conditional moment, i.e. to have a time-varying conditional second moment, although the unconditional second moment constant over time?

• A: Yes.

• Note: The unconditional second moment of \( u_t \) is \( \sigma_u^2 \).

• Last class, we saw a few models of conditional volatility, \( E_{t-1}(u_t^2) \)

\[
    u_t = \sqrt{h_t v_t}
\]
3.1 ARCH/GARCH Models

- In the specification
  \[ u_t = \sqrt{h_t} v_t \]

- ARCH(1) is:
  \[ h_t = \zeta + au_{t-1}^2 \]
  where \( v_t \) is iid with mean 0, and \( E(v_t^2) = 1 \).

- GARCH(1,1): Suppose, we specify \( h_t \) as
  \[ h_t = \zeta + \delta h_{t-1} + au_{t-1}^2 \]

- GARCH(p,q):
  \[ u_t^2 = \zeta + \delta_1 h_{t-1} + \delta_2 h_{t-2} + \ldots + \delta_p h_{t-p} + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \ldots + \alpha_q u_{t-q}^2 + \ldots \]
• For most practical purposes a GARCH(1,1) is GREAT.
  – There is a trade-off. You introduce more parameters to capture the accurate dynamics, but there are more parameters to estimate
  – Those parameters have restrictions. The estimation is tricky.
  – Bottom line, for 99% of the applications, GARCH(1,1) does a great job.

• GARCH is successful, because it can capture the persistence in $\sigma_t^2$, which is the most significant feature that needs to be captured.

• See Hansen and Lunde (2001) article.
3.2 Simple Introduction to GMM (Last Lecture):

- Recall that any variable $x_t$ has a distribution $F_x(x)$. If $x$ has moments $E(x^j), j = 1, \ldots$ then those moments can be used to retrieve $F_x(x)$.

- Caution: Some variables do not have moments (Cauchy distribution case).

- Suppose we have random variables $x_t, y_t, z_t$, and a function $g(.)$.
  - A population moment of those variables is $E[ g(x_t, y_t, z_t | \beta) ]$
  - A sample moment of those variables is $\frac{1}{T} \sum_{t=1}^{T} g(x_t, y_t, z_t | \beta)$
  - By the ergodicity theorem (or the LLN in cross section, we know that)
    $$\frac{1}{T} \sum_{t=1}^{T} g(x_t, y_t, z_t | \beta) \rightarrow^{p} E[ g(x_t, y_t, z_t | \beta) ]$$
    under some mild conditions on the function $g(\cdot | \beta)$. 
• We must have at least as many conditions as parameters.

• What if we have more conditions than parameters?

• If there are more conditions than parameters (dimension of \( g(\cdot|\beta) \) is higher than dimension of \( \beta \)), then how do we proceed?

• We can construct a quadratic function, as:

\[
E \left( g(x|\beta)'Wg(x|\beta) \right) = 0
\]

for some symmetric positive definite matrix \( W \) and \( k \) is the dimensionality of \( g(\cdot|\beta) \).

• Now, we have the information into the three equations, weighted by the elements of the matrix \( W \).

• Problem: What matrix \( W \) to choose?

• A: Any symmetric positive matrix will give us consistent estimates (i.e. \( \hat{\beta}_W \rightarrow^p \beta \)), but we are concerned with efficiency, or smallest possible standard errors around \( \hat{\beta}_W \).
3.3 Estimation of ARCH using GMM

- Recall the model:
  \[ r_t = \mu + \phi r_{t-1} + u_t \]
  \[ u_t^2 = h_t + w_t = \zeta + \alpha u_{t-1}^2 + w_t \]
  \[ E_{t-1}(u_t^2) = \zeta + \alpha u_{t-1}^2 \]
  - First equation: Conditional Mean
  - Second equation: Conditional Variance

- The four moment conditions are:
  \[ E(u_t) = 0 \]  \[ E(u_tr_{t-1}) = E((r_t - \mu - \phi r_{t-1})r_{t-1}) = 0 \]
  \[ E \left( u_t^2 - \frac{\zeta}{1 - \alpha} \right) = 0 \]
  \[ E(w_tr_{t-1}) = E \left( (u_t^2 - \zeta - \alpha u_{t-1}^2) r_{t-1} \right) = 0 \]

- Note: four moments and four parameters (\(\mu, \phi, \zeta, \alpha\)).

- Replace those conditions by their sample analogues:
  \[
  \frac{1}{T} \sum_{t=1}^{T} \begin{bmatrix}
  (r_t - \mu - \phi r_{t-1}) \\
  ((r_t - \mu - \phi r_{t-1})r_{t-1}) \\
  (u_t^2 - \frac{\zeta}{1 - \alpha}) \\
  ((u_t^2 - \zeta - \alpha u_{t-1}^2) r_{t-1})
  \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}
  \]

- Solve for (\(\mu, \phi, \zeta, \alpha\)).
• We need $u_{t-1}$ for the last equation.
• A better way to estimate ARCH/GARCH: maximum likelihood
3.4 Specification and GMM Estimation of Non-Autoregressive Models of Volatility

- Chan et al. (1992): Estimate a “very” general process for the short rate.

- Q: Why?

- A: Because, many models in finance have to specify a short rate. Usually they specify the short rate process that is most convenient, i.e. the short rate process that allows them to derive nice, simple, closed-form results.

- But one has to realize that the results from those models are conditional upon having the right short-rate process.
The proposed (general) process for the short rate is:

\[ dr = (\alpha + \beta r) \, dt + \sigma r^\gamma dZ \]

The short rate is written in continuous time to relate it to most finance models.

The estimation is done in discrete time.

The discretization is done by taking that \( dt \) is one month.

No other adjustments are made.

Discretization

\[
\begin{align*}
    r_{t+1} - r_t &= \alpha + \beta r_t + \sigma r_t^\gamma v_t \\
    r_{t+1} &= \alpha + (1 + \beta) r_t + \sigma r_t^\gamma v_t
\end{align*}
\]
Implementation:


- The short rate is:
  \[ \Delta r_{t+1} = \alpha + \beta r_t + u_{t+1} \]
  or \[ r_{t+1} = \alpha + (1 + \beta) r_t + u_{t+1} \]
  \[ E \left( u_{t+1}^2 \right) = \sigma^2 r_t^{2\gamma} \]

- Note that while the continuous time processes assumed that \( dZ_t \) is normally distributed, this is purely for tractability.

- The shock \( u_{t+1} \) can have any distribution!

- The processes have to be stationary!–This is the only restriction.

- The parameters to be estimated are: \( \alpha, \beta, \gamma, \) and \( \sigma^2. \)
• Of course, this setup is very similar to our first (ARCH) setup, but:

\[ r_t = \mu + \phi r_{t-1} + u_t \]
\[ u_t = \sigma r_{t-1}^{\gamma} v_t \]
\[ E_{t-1}(u_t^2) = \sigma^2 r_{t-1}^{2\gamma} \]

- First equation: Conditional Mean
  * We can take the first difference of returns and write (Chan et al., 1992)

\[ r_{t+1} - r_t = \mu + (\phi - 1) r_t + \sigma r_{t-1}^{\gamma} v_t \]

- Second equation: Conditional Variance

- Note: The variance is not auto-regressive—not suitable for forecasting

• The four moment conditions are:

\[ E(u_t) = 0 \] \hspace{1cm} (3) \]
\[ E(u_t r_{t-1}) = E((r_t - \mu - \phi r_{t-1}) r_{t-1}) = 0 \] \hspace{1cm} (4) \]
\[ E\left(u_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) = 0 \]
\[ E(w_t r_{t-1}) = E\left(\left(u_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) r_{t-1}\right) = 0 \]

• Note: four moments and four parameters ($\mu, \phi, \sigma, \gamma$).

• Replace those conditions by their sample ana-
\[
\frac{1}{T} \sum_{t=1}^{T} \begin{bmatrix}
(r_t - \mu - \phi r_{t-1}) \\
((r_t - \mu - \phi r_{t-1}) r_{t-1}) \\
(u_t^2 - \sigma^2 r_{t-1}^{2\gamma}) \\
\left(\left(u_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) r_{t-1}\right)
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}
\]

- Solve for \((\mu, \phi, \sigma, \gamma)\).
- The same algorithm applies for all problems that can be written as a moment condition.
- Example: stock_vol_illustration2.m
We could also proceed in the following way:

- Regress $r_t$ on $r_{t-1}$ using OLS. Get estimates $\hat{\alpha}, \hat{\beta},$ and $\hat{\varepsilon}_t$.

- Estimate $\sigma^2$ and $\gamma$ using non-linear least squares:

$$\min_{\sigma^2, \gamma} \sum_{t=1}^{T} \left( \hat{\varepsilon}_{t+1}^2 - \sigma^2 r_t^{2\gamma} \right)^2$$

- Would we obtain the same answer?
- What method is preferable in theory?
- What method is preferable in practice?
• A good exercise: Redo the Chan et al. (1992) paper with data up to 2001 and with monthly and daily data. Compare the results and compare them to the GARCH homework.
3.5 GARCH in Mean or GARCH-M models

(Engle, Lilien, and Robins (1987))

\[ r_t = a + bx_{t-1} + c\sigma_t^2 + \varepsilon_t \]
\[ \varepsilon_t = z_t\sigma_t \]
\[ \sigma_t^2 = \kappa + \beta (L) \sigma_{t-1}^2 + \alpha (L) \varepsilon_{t-1}^2 \]

- The difference from the previous models ARCH/GARCH models is that the volatility enters also in the mean of the return.
- This is exactly what Merton’s (1973, 1980) ICAPM produces—risk-return tradeoff.
- It must be the case that \( b > 0 \).
- The GARCH-M is estimated with ML or QML (to come!)
3.6 MIDAS Estimators–Mixed Data Sampling Estimators

(Ghysels, Santa-Clara, Valkanov (2006a,b))

- Idea: Use data at different frequencies to estimate the risk-return tradeoff

\[ R_{t+1} = \alpha + \beta \left( \sum_{d=1}^{D} w_d r_{t-d}^2 \right) + \varepsilon_t \]

- where \( R_t \) is at monthly frequency and \( r_t \) is at daily.
- The weights \( w_d \) sum up to one.
- Given that vol is persistent, there might be many weights to estimate, which would result in inefficient estimators.
- Hence, we parameterize \( w_d(\theta) \) and estimate the shape of the weights.
- There are several advantages:
  - Higher frequency data, i.e. better estimates of vol.
  - Joint estimation of \( \theta, \alpha, \beta \)
  - Flexibility of weights
  - Easy to implement other variables, asymmetries.
3.7 Stochastic volatility models

\[ r_t = a + br_{t-1} + \varepsilon_t \]
\[ \varepsilon_t = z_t \sigma_t \]
\[ \sigma_t = \kappa + \beta \sigma_{t-1} + \nu_t \]

- The difference here is that the shocks that govern the volatility are not necessarily \( \varepsilon_t^2 \)’s.
- This is really a discretization of a continuous-time model, where the mean and the variance follow two OU processes.
- Constrast with Chan et al. (1992) and ARCH/GARCH
- Stochastic volatility models can be estimated by MLE or other methods.
• We can also think of modelling the entire variance covariance matrix.
  – Bollerslev (1990) provides a particularly elegant model with constant correlations, but time-varying covariances.

• Here is a new and good way of modelling the entire variance-covariance matrix
  – To keep it simple, we focus on 2 assets ($R_{1,t}$ and $R_{2,t}$) and 1 exogenous variable, $X_{t-1}$ (think D/P ratio). First, we model the conditional means as:

$$R_{1,t+1} = \mu_1 + k_{11}X_t + Y_{1,t+1}$$
$$R_{2,t+1} = \mu_2 + k_{22}X_t + Y_{2,t+1}$$

Let $Y_{t+1} = (Y_{1,t+1} Y_{2,t+1})$, and $\Sigma_t = E(Y_{t+1}' Y_{t+1} | X_t) = E\left(\begin{array}{c}
Y_{1,t+1}^2 & Y_{1,t+1}Y_{2,t+1} \\
Y_{1,t+1}Y_{2,t+1} & Y_{2,t+1}^2
\end{array}\right) | X_t$.

• NOTE: Usually, we handle conditional expectations with projections (regressions), but we cannot regress $Y_{1,t+1}^2$, $Y_{1,t+1}Y_{2,t+1}$, and $Y_{2,t+1}^2$ on $X_t$ because $\Sigma_t$ must be positive semi-definite.

LET:

$$\Sigma_t = \hat{\Sigma}_t + \begin{pmatrix}
\varepsilon_{11,t+1} & \varepsilon_{12,t+1} \\
\varepsilon_{21,t+1} & \varepsilon_{22,t+1}
\end{pmatrix}$$

Then, we can use any triangular decomposition
(say Cholesky) to write:
\[ \hat{\Sigma}_t = U_t'U_t \]

where
\[
U_t = \begin{pmatrix}
U_{11,t} & U_{12,t} \\
0 & U_{22,t}
\end{pmatrix} = \\
\begin{pmatrix}
\alpha_{11} + \beta_{11}X_t + \gamma_{11}U_{11,t-1} & \alpha_{12} + \beta_{12}X_t + \gamma_{12}U_{12,t-1} \\
0 & \alpha_{22} + \beta_{22}X_t + \gamma_{22}U_{22,t-1}
\end{pmatrix}
\]

Then, we can write
\[ Y_{1,t+1}^2 = \hat{\Sigma}_{11,t} + \varepsilon_{11,t+1} = U_{11,t}^2 + \varepsilon_{11,t+1} \]
\[ Y_{1,t+1}Y_{2,t+1} = \hat{\Sigma}_{12,t} + \varepsilon_{12,t+1} = U_{11}U_{12} + \varepsilon_{12,t+1} \]
\[ Y_{2,t+1}^2 = \hat{\Sigma}_{22,t} + \varepsilon_{22,t+1} = U_{12,t}^2 + U_{22,t}^2 + \varepsilon_{22,t+1} \]

– NOTE: The positive definiteness restrictions are insured by the Cholesky decomposition.

– NOTE: Estimating the \( \alpha' \)'s, \( \beta' \)'s, and \( \gamma' \)'s is done with non-linear least squares.

• NOTE: 2 assets, 5 explanatory variables–5 seconds.
• Comments:
  – Easy to generalize to N assets and M exogenous variables.

• Note that we can write:

\[ U_{11,t} = \kappa_{11} + \beta_{11} \sum_{k=0}^{\infty} \gamma_{11}^k X_{t-k-1} \]
We start the estimation from an “orthogonality” condition:

\[ E(h(w_t; \theta_0)) = 0 \]

where

- \( h(w_t; \theta) \) is a \( r \) dimensional vector of moment conditions, which depends on the data on some unknown parameters to be estimated.
- The parameters are collected in vector \( \theta \) of dimension \( a \), where \( a \leq r \). The true value of \( \theta \) is denoted by \( \theta_0 \).
- Note that \( h(., .) \) is a random variable.

The “Method of Moments” principle states that we can estimate parameters by working with sample moments instead of population moments (Why?).

Therefore, instead of working with

\[ E(h(w_t; \theta_0)) = 0 \]

which we cannot evaluate (Why?), we work with its sample analogue:

\[ g(w_t; \theta) = \frac{1}{T} \sum_{t=1}^{T} h(w_t; \theta) \]
Example: OLS \( y_t = \beta x_t + \varepsilon_t \)

\[
E(x_t\varepsilon_t) = 0 \\
E(x_t^2\varepsilon) = 0 \\
E(x_t^3\varepsilon_t) = 0
\]

- Here, we have three moment conditions \((r = 3)\), and one parameter to estimate \((a = 1)\).

- You can think of \( h(w_t; \theta) = \begin{bmatrix} x_t(y_t - \beta x_t) \\ x_t^2(y_t - \beta x_t) \\ x_t^3(y_t - \beta x_t) \end{bmatrix} \), \( w_t = (y_t, x_t) \) and \( \theta = \beta \).

- We will work with the sample analogues

\[
g(w_t; \theta) = \frac{1}{T} \sum_{t=1}^{T} \begin{bmatrix} x_t(y_t - \beta x_t) \\ x_t^2(y_t - \beta x_t) \\ x_t^3(y_t - \beta x_t) \end{bmatrix}
\]

- Note, that from the e... theorem, we have

\[
g(w_t; \theta) \xrightarrow{p} E(h(w_t; \theta))
\]
Since there might be more moment conditions than parameters to estimate, we will work with the quadratic

\[ Q = g (w_t; \theta)' W_T g (w_t; \theta) \]

where \( W_T \) is a positive definite matrix that depends on the data.

The above quadratic can be minimized with respect to \( \theta \) using analytic or numerical methods (depending on the complexity of \( h \)).

It would be “logical” to put more weight on moments whose variance is smaller. Therefore, we want the matrix \( W_T \) to be inversely related to \( Var (h (w_t; \theta)) \), or \( W_T = Var (h (w_t; \theta))^{-1} \).

Before we pose the problem, we note that the weighing matrix \( Var (h (w_t; \theta))^{-1} \) does not take into account the dependence in the data. Therefore, we will work with

\[ \Gamma_j = E (h (w_t; \theta) h (w_{t-j}; \theta)) \]

\[ S = \sum_{j=0}^{\infty} \Gamma_j \]

The matrix \( S \) takes into account the dependence in the data.
– Long-run variance
– $2\pi$ spectrum at frequency zero.
• It turns out that we can prove (CLT with serially dependent data)

\[ \sqrt{T} \left( g(w_t; \theta_0) \right) \sim^a N(0, S) \]

• Note that if \( \Gamma_j = 0, j \geq 1 \) (serially independent data), then \( S = Var(h(w_t; \theta)) = E(h(w_t; \theta) h(w_t; \theta)) \).

• Finally we will let \( W_T = S_T^{-1} \)

• Therefore, the problem is:

\[ Q = g(w_t; \theta)' S_T^{-1} g(w_t; \theta) \]

• The FOC is:

\[ \left\{ \frac{\partial g}{\partial \theta} \Big|_{\theta = \hat{\theta}} \right\}_a^r x_r S_T^{-1} g(w_t; \hat{\theta}) = 0 \]

• So, what are the properties of \( \hat{\theta} \)?
• Denote

\[ D'_T = \begin{bmatrix}
\frac{\partial g_1(w_i;\hat{\theta})}{\partial \theta'} \\
\frac{\partial g_r(w_i;\hat{\theta})}{\partial \theta'}
\end{bmatrix} \]

• We will show that

\[ \sqrt{T} \left( \hat{\theta} - \theta_0 \right) \sim^a N \left( 0, \left( DS^{-1}D' \right)^{-1} \right) \]
The “proof” follows a few very simple steps

– Use the Mean-Value theorem, to write

\[ g \left( w_t; \hat{\theta} \right) = g \left( w_t; \theta_0 \right) + D'_T \left( \hat{\theta} - \theta_0 \right) \]

– Pre-multiply both sides by \( \left\{ \frac{\partial g}{\partial \theta'} |_{\theta = \hat{\theta}} \right\}' S_T^{-1} \) to get

\[ \left\{ \frac{\partial g}{\partial \theta'} |_{\theta = \hat{\theta}} \right\}' S_T^{-1} g \left( w_t; \hat{\theta} \right) \]

\[ = \left\{ \frac{\partial g}{\partial \theta'} |_{\theta = \hat{\theta}} \right\}' S_T^{-1} g \left( w_t; \theta_0 \right) + \left\{ \frac{\partial g}{\partial \theta'} |_{\theta = \hat{\theta}} \right\}' S_T^{-1} D'_T \left( \hat{\theta} - \theta_0 \right) \]

– But, we know by definition that

\[ \left\{ \frac{\partial g}{\partial \theta'} |_{\theta = \hat{\theta}} \right\}' S_T^{-1} g \left( w_t; \hat{\theta} \right) = 0 \]
– Or
\[
\begin{align*}
\left\{ \frac{\partial g}{\partial \theta} \right\}_{\theta = \hat{\theta}} \right. & \quad S_{T}^{-1} g (w_t; \theta_0) = - \left\{ \left. \frac{\partial g}{\partial \theta} \right\}_{\theta = \hat{\theta}} \right. \\
& \quad S_{T}^{-1} D_{T} \left( \hat{\theta} - \theta_0 \right)
\end{align*}
\]
– Rearranging, we get
\[
\left( \hat{\theta} - \theta_0 \right) = D_{T}^{-1} g (w_t; \theta_0)
\]
– Then,
\[
\sqrt{T} \left( \hat{\theta} - \theta_0 \right) = D_{T}^{-1} \sqrt{T} g (w_t; \theta_0)
\]
– But, recall that
\[
\sqrt{T} (g (w_t; \theta_0)) \sim a N(0, S)
\]
– Hence,
\[
\sqrt{T} \left( \hat{\theta} - \theta_0 \right) \sim a N(0, D_{T}^{-1} S D_{T}^{-1'})
\]
\[
\sim a N \left( 0, \left( D_{T} S^{-1} D_{T}' \right)^{-1} \right)
\]
Final Result: The GMM estimates are asymptotically normally distributed with a variance-covariance matrix equal to \((D_T S^{-1} D_T')^{-1}/T\).

This is a huge result. All we needed was a set of moment conditions, nothing else!

We also need the data \(w_t\) to be stationary.

Many “standard” problems can be written as GMM.

The real power of GMM is that one framework can handle a lot of interesting problems.
It should be immediately obvious that the number of orthogonality conditions and the conditioning information matter

- In practice, the “type” of conditioning information will have a great impact on the estimates $\hat{\theta}$. Think instrumental variables.
- The question is: Which moments to choose?
- This is quite discomforting. If slight variations in our problem yields widely different estimates of $\theta$, what can we conclude?
• Also: Estimating the matrix $S$ makes a huge difference. Recall that
\[ \Gamma_j = \mathbb{E} \left( h(w_t; \theta) h(w_{t-j}; \theta) \right) \]
\[ S = \sum_{j=0}^{\infty} \Gamma_j \]

• Using sample analogues to obtain $\hat{\Gamma}_j$ and $\hat{S}$ is not the right way. Newey and West (1987) have proposed a “corrected” way, which is:
\[ \hat{S} = \hat{\Gamma}_0 + \sum_{v=1}^{q} \left( 1 - \frac{v}{q+1} \right) \left( \hat{\Gamma}_j + \hat{\Gamma}'_j \right) \]

• Even the Newey-West method does not yield good results when the dimension of the system is large.

• Moreover, the truncation point, $q$, introduces another source of error.
- People have shown that small perturbations in $\hat{S}$ results in big differences in the estimates $\hat{\theta}$. In other words, suppose we use a matrix 
  $$\hat{S} + P$$
  where $P$ has small values on its diagonals (perturbing the variances only). This results in widely different estimates. So, small differences in estimating $\hat{S}$ matter a lot.

- The mechanics of why this is so reside in taking inverses...

- Since we only need the optimal weighing matrix $S$ for efficiency (smallest variance), is it possible to find a matrix that, although not yielding efficient estimates, yields robust estimates?

- In practice: The best (most robust) results are obtained with $I$, the identity matrix.

- Empirical rule of thumb: Try the identity matrix first. Then, try the optimal weighing matrix, $\hat{S}$. If the results are widely different, stick with $I$. 
In his 1982 seminal paper, L.P. Hansen argued that the multiplicity of the moments, or the over-identifications, are an advantage, rather than a disadvantage.

- Even though we cannot have \( g(\hat{\theta}, w_t) = 0 \), it must be the case that at, and close to, the true value \( \theta_0 \), \( g() \) will be close to zero.

- Note that, since 
  \[ \sqrt{T} (g(w_t; \theta_0)) \sim^a \text{N}(0, S) \]
  then, it must be the case that 
  \[ Tg(w_t; \theta_0)' S^{-1} g(w_t; \theta_0) \sim^a ??? \]

- It might be conjectured that the same would hold true for \( \hat{\theta} \), or that 
  \[ Tg(w_t; \hat{\theta})' S^{-1} g(w_t; \hat{\theta}) \sim^a \chi^2(r) \]

- However, this intuition is false, because it is not necessarily the same linear combination of \( g(w_t; \hat{\theta}) \) that would be close to zero. Instead, it can be shown that 
  \[ Tg(w_t; \hat{\theta})' S^{-1} g(w_t; \hat{\theta}) \sim^a \chi^2(r - a) \]

- Note: This statistic is trivial to estimate. Plug \( \hat{\theta} \) into \( g(.) \), etc.
The test

\[ J = T g \left( w_t; \hat{\theta} \right)' S^{-1} g \left( w_t; \hat{\theta} \right) \]

called the rank test, the test for over-identifying restrictions, Hansen’s J test, etc. has been used extensively in finance.

Indeed, people have relied exclusively on this test to judge the fit of their models.

Problems:

- As discussed above, the over-identifying restrictions are subject to the “which moments” critique.
- The J test also depends crucially on \( S \), which cannot be estimated accurately.

Not surprisingly, the J test rejects a lot of models. But, people are now aware of its deficiencies.
The GMM framework is rich enough that we can think of many other ways of testing the hypotheses of interest. As a starting point, we can break the orthogonality restrictions into those that identify and those that over-identify the parameters

\[ E \begin{pmatrix} h_1(w_t; \theta_0) \\ h_2(w_t; \theta_0) \end{pmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \]

Some people have suggested to see how the estimates would change as we add more restrictions to \( h_2 \) (say, starting from no over-identifying restrictions, and adding progressively).

This set-up has also yielded insights into the stability properties of the moments and (or versus) the estimates.
• Example (Hansen, EMA 1982; Hansen and Singleton, EMA 1982):

\[ p_t = E_t \left( m(zt|\theta) x_{t+1} \right) \]

- \( m(zt|\theta) \) is the stochastic discount factor that depends on some state variables and parameters.
- \( x_{t+1} \) is the payoff from the asset.

• Example from last lecture:

\[ p_t = E_t \left[ \beta \left( \frac{ct+1}{ct} \right)^{-\gamma} x_{t+1} \right] \]

• We don’t need to log-linearize anything with GMM.
• Non-linearities are not a problem.
• Robustness is a real issue.
• But this is not what people use. Why?
We can re-write the pricing equation as:

\[ 1 = E_t \left( m(z_t|\theta) \frac{x_{t+1}}{p_t} \right) \]

\[ E_t \left( m(z_t|\theta) R_{t+1} \right) - 1 = 0 \]

where \( R_{t+1} = \frac{x_{t+1}}{p_t} \). In the case of stocks, think \( x_{t+1} = p_{t+1} + d_{t+1} \).

- In this case, returns are stationary.
- If \( z_t \) in the stochastic discount factor is also stationary, then the ergoticity theorem conditions will be satisfied (under some other mild conditions).

- Example from last lecture:

\[ E_t \left[ \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} (1 + R_{t+1}) \right] - 1 = 0 \]

- We have the data \( \{c_t, R_t\} \).
- Have to estimate one parameter, \( \gamma \) (assuming \( \beta \) is 0.995).
5 Maximum Likelihood Estimation

(Preliminaries for Kalman Filtering)

- Suppose we have the series \( \{Y_1, Y_2, \ldots, Y_T\} \) with a joint density \( f_{Y_T, \ldots, Y_1}(\theta) \) that depends on some parameters \( \theta \) (such as means, variances, etc.)

- We observe a realization of \( Y_t \).

- If we make some functional assumptions on \( f \), we can think of \( f \) as the probability of having observed this particular sample, given the parameters \( \theta \).

- The maximum likelihood estimate (MLE) of \( \theta \) is the value of the parameters \( \theta \) for which this sample is most likely to have been observed.

- In other words, \( \hat{\theta}^{MLE} \) is the value that maximizes \( f_{Y_T, \ldots, Y_1}(\theta) \).
Q: But, how do we know what $f$—the true density of the data—is?

A: We don’t.

Usually, we assume that $f$ is normal, but this is strictly for simplicity. The fact that we have to make distributional assumptions limits the use of MLE in many financial applications.

Recall that if $Y_t$ are independent over time, then

$$f_{Y_T...Y_1}(\theta) = f_{Y_T}(\theta_T)f_{Y_{T-1}}(\theta_{T-1})...f_{Y_1}(\theta_1) = \prod_{i=1}^{T} f_{Y_i}(\theta_i)$$

Sometimes it is more convenient to take the log of the likelihood function, then

$$\Lambda(\theta) = \log f_{Y_T...Y_1}(\theta) = \sum_{i=1}^{T} \log f_{Y_i}(\theta)$$
• However, in most time series applications, the independence assumption is untenable. Instead, we use a conditioning trick.

• Recall that

\[ f_{Y_2Y_1} = f_{Y_2|Y_1}f_{Y_1} \]

• In a similar fashion, we can write

\[ f_{Y_T....Y_1}(\theta) = f_{Y_T|Y_{T-1}....Y_1}(\theta)f_{Y_{T-1}|Y_{T-2}....Y_1}(\theta)…f_{Y_1}(\theta) \]

• The log likelihood can be expressed as

\[ \Lambda(\theta) = \log f_{Y_T....Y_1}(\theta) = \sum_{i=1}^{T} \log f_{Y_i|Y_{i-1},...,Y_1}(\theta_i) \]
Example: The log-likelihood of an AR(1) process

\[ Y_t = c + \phi Y_{t-1} + \varepsilon_t \]

Suppose that \( \varepsilon_t \) is iid \( N(0, \sigma^2) \)

Recall that \( E(Y_t) = \frac{c}{1-\phi} \) and \( Var(Y_t) = \frac{\sigma^2}{1-\phi^2} \)

Since \( Y_t \) is a linear function of the \( \varepsilon_t \)'s, then it is also Normal (sum of normals is a normal).

Therefore, the density (unconditional) of \( Y_t \) is Normal.

Result: If \( Y_1 \) and \( Y_2 \) are jointly Normal, then the marginals are also normal.

Therefore,

\[ f_{Y_2|Y_1} \text{ is } N \left( (c + \phi y_1) , \sigma^2 \right) \]

or

\[ f_{Y_2|Y_1} = \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left[ -\frac{(y_2 - c - \phi y_1)^2}{2\sigma^2} \right] \]
Similarly,

\[
f_{Y_3 | Y_2} \text{ is } N \left( (c + \phi y_2), \sigma^2 \right)
\]

or

\[
f_{Y_3 | Y_2} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_3 - c - \phi y_2)^2}{2\sigma^2} \right]
\]
Then, the log likelihood can be written as

$$\Lambda(\theta) = \log f_Y + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}$$

$$= \frac{1}{2} \log (2\pi) - \frac{1}{2} \log \left(\sigma^2 / (1 - \phi^2)\right)$$

$$- \frac{\{y_1 - (c/ (1 - \phi))\}^2}{2\sigma^2 / (1 - \phi^2)}$$

$$- \frac{(T - 1)}{2} \log (2\pi) - \frac{(T - 1)}{2} \log (\sigma^2)$$

$$- \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}$$

- The unknown parameters are collected in $\theta = (c, \phi, \sigma)$

- We can maximize $\Lambda(\theta)$ with respect to all those parameters and find the estimates that maximize the probability of having observed such a sample.

$$\max_{\theta} \Lambda(\theta)$$

- Sometimes, we can even put constraints (such as $|\phi| < 1$)

- Q: Is it necessary to put the constraint $\sigma^2 > 0$?
• Note: If we forget the first observation, then we can write (setting $c = 0$) the FOC:

$$- \sum_{t=2}^{T} \frac{\partial}{\partial \phi} \left( \frac{(y_t - \phi y_{t-1})^2}{2\sigma^2} \right) = 0$$

$$\sum_{t=2}^{T} y_{t-1} (y_t - \phi y_{t-1}) = 0$$

$$\hat{\phi} = \frac{\sum_{t=2}^{T} y_{t-1} y_t}{\sum_{t=2}^{T} y_{t-1}^2}$$

• RESULT: In the univariate linear regression case, OLS, GMM, MLE are equivalent!!!
To summarize the maximum likelihood principle:
(a) Make a distributional assumption about the data
(b) Use the conditioning to write the joint likelihood function
(c) For convenience, we work with the log-likelihood function
(d) Maximize the likelihood function with respect to the parameters

There are some subtle points.
- We had to specify the unconditional distribution of the first observation
- We had to make an assumption about the dependence in the series

But sometimes, MLE is the only way to go.
MLE is particularly appealing if we know the distribution of the series. Most other deficiencies can be circumvented.
• Now, you will ask: What are the properties of $\hat{\theta}^{MLE}$? More specifically, is it consistent? What is its distribution, where
\[ \hat{\theta}^{MLE} = \arg \max \Lambda (\theta) \]
• Yes, $\hat{\theta}^{MLE}$ is a consistent estimator of $\theta$.
• As you probably expect the asymptotic distribution of $\hat{\theta}^{MLE}$ is normal.
• Result:
\[ T^{1/2} \left( \hat{\theta}^{MLE} - \theta \right) \sim aN (0, V) \]
\[ V = \left[ -\frac{\partial^2 \Lambda (\theta)}{\partial \theta \partial \theta'} \bigg|_{\hat{\theta}^{MLE}} \right]^{-1} \]
\text{or}
\[ V = \sum_{t=1}^{T} l \left( \hat{\theta}^{MLE}, y \right) l \left( \hat{\theta}^{MLE}, y \right) \]
\[ l \left( \hat{\theta}^{MLE}, y \right) = \frac{\partial f}{\partial \theta} (\hat{\theta}^{MLE}, y) \]
• But we will not dwell on proving those properties.
Another Example: The log-likelihood of an AR(1)+ARCH(1) process

\[ Y_t = c + \phi Y_{t-1} + u_t \]

- where,
  \[ u_t = \sqrt{h_t} v_t \]

- ARCH(1) is:
  \[ h_t = \zeta + au_{t-1}^2 \]
  where \( v_t \) is iid with mean 0, and \( E(v_t^2) = 1 \).

- GARCH(1,1): Suppose, we specify \( h_t \) as
  \[ h_t = \zeta + \delta h_{t-1} + au_{t-1}^2 \]

- Recall that \( E(Y_t) = \frac{c}{1-\phi} \) and \( \text{Var}(Y_t) = \frac{\sigma^2}{1-\phi^2} \)

- Since \( Y_t \) is a linear function of the \( \varepsilon_t \)'s, then it is also Normal (sum of normals is a normal).

- Therefore, the density (unconditional) of \( Y_t \) is Normal.

- Result: If \( Y_1 \) and \( Y_2 \) are jointly Normal, then the marginals are also normal.

- Therefore,
  \[ f_{Y_2|Y_1} \text{ is } N((c + \phi y_1), h_2) \]

  or for the ARCH(1)

\[
 f_{Y_2|Y_1} = \frac{1}{\sqrt{2\pi (\zeta + au_1^2)}} \exp \left[ \frac{-(y_2 - c - \phi y_1)^2}{2(\zeta + au_1^2)} \right]
\]
• Similarly,

\[ f_{Y_3|Y_2} \text{ is } N((c + \phi y_2), h_3) \]

or

\[ f_{Y_3|Y_2} = \frac{1}{\sqrt{2\pi (\zeta + au_2^2)}} \exp \left[ \frac{-(y_3 - c - \phi y_2)^2}{2(\zeta + au_2^2)} \right] \]
Then, the conditional log likelihood can be written as

\[
\Lambda(\theta|y_1) = \sum_{t=2}^{T} \log f_{Y_i|Y_{t-1}}
\]

\[
= -\frac{(T - 1)}{2} \log (2\pi) - \frac{1}{2} \sum_{t=2}^{T} \log (\zeta + au_{t-1}^2)
\]

\[
- \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2(\zeta + \alpha u_{t-1}^2)}
\]

- The unknown parameters are collected in \( \theta = (c, \phi, \zeta, \alpha) \)

- We can maximize \( \Lambda(\theta) \) with respect to all those parameters and find the estimates that maximize the probability of having observed such a sample.

\[
\max_{\theta} \Lambda(\theta)
\]

- Example: mle_arch.m
Similarly for GARCH(1,1):

$$\Lambda (\theta | y_1) = \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}$$

$$= -\frac{(T - 1)}{2} \log (2\pi) - \frac{1}{2} \sum_{t=2}^{T} \log (h_t)$$

$$- \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2 (h_t)}$$

where

$$h_t = \zeta + \delta h_{t-1} + \alpha u_{t-1}^2$$