Lecture 12
Heteroscedasticity

Two-Step Estimation of the GR Model: Review

- Use the GLS estimator with an estimate of $\Omega$
  1. $\Omega$ is parameterized by a few estimable parameters, $\Omega = \Omega(\theta)$.
     Example: Harvey’s heteroscedastic model.
  2. Iterative estimation procedure:
     (a) Use OLS residuals to estimate the variance function.
     (b) Use the estimated $\Omega$ in GLS - Feasible GLS, or FGLS.

- True GLS estimator
  $b_{\text{GLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ (converges in probability to $\beta$.)

- We seek a vector which converges to the same thing that this does.
  Call it FGLS, based on $[X'\Omega^{-1}X]^{-1}X'\Omega^{-1}y$
Two-Step Estimation of the GR Model: Review

The object is to find a set of parameters such that

\[ [X'\Omega^{-1}X]X'\Omega^{-1}y - [X'\tilde{\Omega}^{-1}X]X'\tilde{\Omega}^{-1}y \to 0 \]

This does not require us to find estimates such that \( \tilde{\Omega} - \Omega \to 0 \). This does not make sense, actually, because \( \Omega \) is non-n, and is growing in size with the sample size. So, you could not say that \( \tilde{\Omega} \) converges to anything. What is required is that the two \( K \times K \) matrices, 

\[ [X'\Omega^{-1}X] \] and \( [X'\tilde{\Omega}^{-1}X] \) converge to the same thing. Note that it is meaningful to speak of “convergence” of these matrices, whereas it is not to speak of convergence of \( \tilde{\Omega} \). We'll examine Harvey's model in detail. For asymptotic properties, we require that

\[ (1/n) [X'\Omega^{-1}e \cdot X'\tilde{\Omega}^{-1}e] \to 0 \]

This is the difference of two \( K \times 1 \) vectors. The logic is the same.

Two-Step Estimation of the GR Model: FGLS

- Feasible GLS is based on finding an estimator which has the same properties as the true GLS.

Example: \( \text{Var}[e] = \sigma^2 \exp(\gamma'z_i) \).

True GLS: Regress \( y_i/\sigma \exp((1/2)\gamma'z_i) \) on \( x_i/\sigma \exp((1/2)\gamma'z_i) \)

FGLS: With a consistent estimator of \( [\sigma, \gamma] \), say \( [\hat{s}, \hat{c}] \), we do the same computation with our estimates.

Note: If plim \( [s, c] = [\sigma, \gamma] \), then, FGLS is as good as true GLS.

- Remark: To achieve full efficiency, we do not need an efficient estimate of the parameters in \( \Omega \), only a consistent one.
Heteroscedasticity

- Assumption (A3) is violated in a particular way: $\varepsilon$ has unequal variances, but $\varepsilon_i$ and $\varepsilon_j$ are still not correlated with each other. Some observations (lower variance) are more informative than others (higher variance).

$$E(y|X) = b_0 + b_1x$$

Now, we have the CLM regression with hetero-(different) scedastic (variance) disturbances.

(A1) DGP: $y = X \beta + \varepsilon$ is correctly specified.

(A2) $E[\varepsilon|X] = 0$

(A3') $\text{Var}[\varepsilon_i] = \sigma^2 \omega_i, \quad \omega_i > 0. \quad (\text{CLM} \Rightarrow \omega_i = 1, \text{for all } i.)$

(A4) $X$ has full column rank -- $\text{rank}(X)=k$, where $T \geq k$.

- Popular normalization: $\Sigma_i \omega_i = 1$. (A scaling, absorbed into $\sigma^2$.)

- A characterization of the heteroscedasticity: Well defined estimators and methods for testing hypotheses will be obtainable if the heteroscedasticity is “well behaved” in the sense that
  $$\omega_i / \Sigma_i \omega_i \to 0 \text{ as } T \to \infty. \quad \text{-i.e., no single observation becomes dominant.}$$
  $$\frac{1}{T} \Sigma_i \omega_i \to \text{some stable constant.} \quad \text{(Not a plim!)}$$
GR Model and Testing

• Implications for conventional OLS and hypothesis testing:
  1. $b$ is still unbiased.
  2. Consistent? We need the more general proof. Not difficult.
  3. If \( \text{plim} \ b = \beta \), then \( \text{plim} \ s^2 = \sigma^2 \) (with the normalization).
  4. Under usual assumptions, we have asymptotic normality.

• Two main problems with OLS estimation under heteroscedasticity:
  1. The usual standard errors are not correct. (They are biased!)
  2. OLS is not BLUE.

• Since the standard errors are biased, we cannot use the usual $t$-statistics or $F$-statistics or $LM$ statistics for drawing inferences. This is a serious issue.

Heteroscedasticity: Inference Based on OLS

• Q: But, what happens if we still use $s^2(X'X)^{-1}$?
  A: It depends on $X'\Omega X - X'X$. If they are nearly the same, the OLS covariance matrix will give OK inferences.

But, when will $X'\Omega X - X'X$ be nearly the same? The answer is based on a property of weighted averages. Suppose $\omega_i$ is randomly drawn from a distribution with $E[\omega_i] = 1$. Then,

\[
\frac{1}{T} \sum \omega_i x_i^2 \rightarrow E[x^2]
\]

--just like \( \frac{1}{T} \sum x_i^2 \).

• Remark: For the heteroscedasticity to be a significant issue for estimation and inference by OLS, the weights must be correlated with $x$ and/or $x^2$. The higher correlation, heteroscedasticity becomes more important ($b$ is more inefficient).
There are several theoretical reasons why the $\omega_i$ may be related to $x$ and/or $x_i^2$:

1. Following the error-learning models, as people learn, their errors of behavior become smaller over time. Then, $\sigma_i^2$ is expected to decrease.
2. As data collecting techniques improve, $\sigma_i^2$ is likely to decrease. Companies with sophisticated data processing techniques are likely to commit fewer errors in forecasting customer’s orders.
3. As incomes grow, people have more discretionary income and, thus, more choice about how to spend their income. Hence, $\sigma_i^2$ is likely to increase with income.
4. Similarly, companies with larger profits are expected to show greater variability in their dividend/buyback policies than companies with lower profits.

Finding Heteroscedasticity

- Heteroscedasticity can also be the result of model misspecification.
- It can arise as a result of the presence of outliers (either very small or very large). The inclusion/exclusion of an outlier, especially if $T$ is small, can affect the results of regressions.
- Violations of (A1) - model is correctly specified-, can produce heteroscedasticity, due to omitted variables from the model.
- Skewness in the distribution of one or more regressors included in the model can induce heteroscedasticity. Examples are economic variables such as income, wealth, and education.
- David Hendry notes that heteroscedasticity can also arise because of
  - (1) incorrect data transformation (e.g., ratio or first difference transformations).
  - (2) incorrect functional form (e.g., linear vs log-linear models).
Finding Heteroscedasticity

- Heteroscedasticity is usually modeled using one of the following specifications:
  - $H_1$: $\sigma_t^2$ is a function of past $\varepsilon_t^2$ and past $\sigma_t^2$ (GARCH model).
  - $H_2$: $\sigma_t^2$ increases monotonically with one (or several) exogenous variable(s) ($x_1, \ldots, x_T$).
  - $H_3$: $\sigma_t^2$ increases monotonically with $E(y_t)$.
  - $H_4$: $\sigma_t^2$ is the same within $p$ subsets of the data but differs across the subsets (grouped heteroscedasticity). This specification allows for structural breaks.

- These are the usual alternatives hypothesis in the heteroscedasticity tests.

Finding Heteroscedasticity

- **Visual test**
  In a plot of residuals against dependent variable or other variable will often produce a fan shape.
Testing for Heteroscedasticity

• Usual strategy when heteroscedasticity is suspected: Use OLS along the White estimator. This will give us consistent inferences.

• Q: Why do we want to test for heteroscedasticity?
  A: OLS is no longer efficient. There is an estimator with lower asymptotic variance (the GLS/FGLS estimator).

• We want to test: $H_0: E(\varepsilon^2|x_1, x_2, \ldots, x_k) = E(\varepsilon^2) = \sigma^2$

• The key is whether $E[\varepsilon^2] = \sigma^2\omega_i$ is related to $x$ and/or $x_i^2$. Suppose we suspect a particular independent variable, say $X_1$, is driving $\omega_i$.

• Then, a simple test: Check the RSS for large values of $X_1$, and the RSS for small values of $X_1$. This is the Goldfeld-Quandt test.

Testing for Heteroscedasticity

• The Goldfeld-Quandt test
  - Step 1. Arrange the data from small to large values of the independent variable suspected of causing heteroscedasticity, $X_j$.
  - Step 2. Run two separate regressions, one for small values of $X_j$ and one for large values of $X_j$, omitting $d$ middle observations ($\approx 20\%$). Get the RSS for each regression: $RSS_1$ for small values of $X_j$ and $RSS_2$ for large $X_j$'s.
  - Step 3. Calculate the F ratio
    $GQ = RSS_2/RSS_1$, $\sim F_{df,df}$ with $df = [(T - d) - 2(k+1)]/2$ ($A5$ holds).

If ($A5$) does not hold, we have $GQ$ is asymptotically $\chi^2$. 
Testing for Heteroscedasticity

• The Goldfeld-Quandt test

Note: When we suspect more than one variable is driving the $\omega_i$'s, this test is not very useful.

• But, the GQ test is a popular to test for structural breaks (two regimes) in variance. For these tests, we rewrite step 3 to allow for different size in the sub-samples 1 and 2.

- Step 3. Calculate the F-test ratio
\[ GQ = \frac{RSS_2/ (T_2 - k)}{RSS_1/ (T_1 - k)} \]

Testing for Heteroscedasticity: LR Test

• The Likelihood Ratio Test

Let’s define the likelihood function, assuming normality, for a general case, where we have $g$ different variances:

\[ \ln L = -\frac{T}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{g} \frac{1}{\sigma_i^2} \ln(\sigma_i^2) - \frac{1}{2} \sum_{i=1}^{g} \frac{1}{\sigma_i^2} (y_i - X_i\beta)'(y_i - X_i\beta) \]

We have two models:

(R) Restricted under $H_0$: $\sigma_i^2 = \sigma^2$. From this model, we calculate $\ln L_R$:

\[ \ln L_R = -\frac{T}{2} [\ln(2\pi) + 1] - \frac{T}{2} \ln(\hat{\sigma}^2) \]

(U) Unrestricted. From this model, we calculate the log likelihood.

\[ \ln L_U = -\frac{T}{2} [\ln(2\pi) + 1] - \frac{1}{2} \sum_{i=1}^{g} \frac{T}{2} \ln(\hat{\sigma}_i^2); \quad \hat{\sigma}_i^2 = \frac{1}{T_i}(y_i - X_i\beta)'(y_i - X_i\beta) \]
Testing for Heteroscedasticity: LR Test

• Now, we can estimate the Likelihood Ratio (LR) test:

\[ LR = 2 \ln L_U - \ln L_R = T \ln \hat{\sigma}^2 - \sum_{i=1}^{g} T_i \ln \hat{\sigma}_i^2 \stackrel{a}{\longrightarrow} \chi^2_{g-1} \]

Under the usual regularity conditions, LR is approximated by a \( \chi^2_{g-1} \).

• Using specific functions for \( \sigma_i^2 \), this test has been used by Rutemiller and Bowers (1968) and in Harvey’s (1976) groupwise heteroscedasticity paper.

Testing for Heteroscedasticity

• **Score LM tests**

• We want to develop tests of \( H_0: \text{E}(\varepsilon^2 | x_1, x_2, \ldots, x_k) = \sigma^2 \) against an \( H_1 \) with a general functional form.

• Recall the central issue is whether \( \text{E}[\varepsilon^2] = \sigma^2 \omega_i \) is related to \( x \) and/or \( x_i^2 \). Then, a simple strategy is to use OLS residuals to estimate disturbances and look for relationships between \( \varepsilon_i^2 \) and \( x_i \) and/or \( x_i^2 \).

• Suppose that the relationship between \( \varepsilon^2 \) and \( X \) is linear:

\[ \varepsilon^2 = X \alpha + v \]

Then, we test: \( H_0: \alpha = 0 \) against \( H_1: \alpha \neq 0 \).

• We can base the test on how the squared OLS residuals \( \varepsilon \) correlate with \( X \).
Testing for Heteroscedasticity

- Popular heteroscedasticity LM tests:
  - Breusch and Pagan (1979)’s LM test (BP).
  - White (1980)’s general test.

- Both tests are based on OLS residuals. That is, calculated under $H_0$: No heteroscedasticity.

- The BP test is an LM test, based on the score of the log likelihood function, calculated under normality. It is a general tests designed to detect any linear forms of heteroskedasticity.

- The White test is an asymptotic Wald-type test, normality is not needed. It allows for nonlinearities by using squares and crossproducts of all the $x$’s in the auxiliary regression.

Testing for Heteroscedasticity: BP Test

- Let’s start with a general form of heteroscedasticity:
  \[ h_1(\alpha_0 + z_{i1}' \alpha_1 + \ldots + z_{im}' \alpha_m) = \sigma_i^2 \]

- We want to test: $H_0$: $E(\varepsilon_i^2 | z_1, z_2, \ldots, z_k) = h_1(z_i' \alpha) = \sigma^2$
  or $H_0$: $\alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$ (m restrictions)

- Assume normality. That is, the log likelihood function is:
  \[ \log L = \text{constant} + \frac{1}{2} \sum \log \sigma_i^2 - \frac{1}{2} \sum \frac{\varepsilon_i^2}{\sigma_i^2} \]

Then, construct an LM test:

\[ \text{LM} = S(\theta_R)' I(\theta_R)^{-1} S(\theta_R) \quad \theta = (\beta, \alpha) \]

\[ S(\theta) = \frac{\partial \log L}{\partial \theta} = [\Sigma \sigma_i^{-2} X' \varepsilon_i ; -\frac{1}{2} \Sigma (\partial h / \partial \alpha) z_i \sigma_i^{-2} + \frac{1}{2} \Sigma \sigma_i^{-4} \varepsilon_i^2 (\partial h / \partial \alpha) z_i] \]

\[ I(\theta) = E[- \partial^2 \log L / \partial \theta \partial \theta'] \]

- We have block diagonality, we can rewrite the LM test, under $H_0$:
  \[ \text{LM} = S(\alpha_0, 0)' [I_{22} - I_{21} I_{11} I_{21}]^{-1} S(\alpha_0, 0) \]
Testing for Heteroscedasticity: BP Test

- We have block diagonality, we can rewrite the LM test, under H0:

\[
LM = S(α₀, 0) \left[ I_{22} - I_{21} \right]^{-1} S(α₀, 0)
\]

\[
S(α₀, 0) = -\frac{1}{2} \sum_i \left( \frac{∂h}{∂α} \mid α₀, R, 0 \right) z_i^2 \sigma_R^{-2} + \frac{1}{2} \sum_i \sigma_R^{-2} c_i^2 \left( \frac{∂h}{∂α} \mid α₀, R, 0 \right) z_i^2
\]

\[
= \frac{1}{2} \sigma_R^{-2} \left( \frac{∂h}{∂α} \mid α₀, R, 0 \right) \sum_i z_i \left( c_i^2 / \sigma_R^{-2} - 1 \right)
\]

\[
b = c_i^2 / \sigma_R^{-2} - 1 = g_i - 1
\]

\[
I_{22}(α₀, 0) = E[− \nabla^2 \log L / \nabla α α'] = \frac{1}{2} \left[ \sigma_R^{-2} \left( \frac{∂h}{∂α} \mid α₀, R, 0 \right) \right]^2 \sum_i z_i z_i'
\]

\[
I_{21}(α₀, 0) = 0
\]

\[
\sigma_R^{-2} = (1/T) \sum_i c_i^2
\]

(MLE of σ under H0).

Then,

\[
LM = \frac{1}{2} \left( \sum_i z_i \omega_i \right)^2 \sum_i z_i \omega_i = \frac{1}{2} W'Z (Z'Z)^{-1} Z'W \sim \chi^2_m
\]

Note: Recall R² = \left[ y'X (X'X)^{-1} X'y - TV^2 \right] / \left[ y'y - TV^2 \right] = ESS / TSS

Also note that under H₀: E[ω_i] = 0, E[ω_i^2] = 1.

Testing for Heteroscedasticity: BP Test

- LM = \frac{1}{2} W'Z (Z'Z)^{-1} Z'W = \frac{1}{2} ESS

ESS = Explained SS in regression of ω_i (= c_i^2 / σ_R^2 - 1) against z_i.

- Under the usual regularity condition, and under H₀,

\[
\sqrt{T} (α_ML - α) \xrightarrow{d} N(0, 2 σ^4 (Z'Z/T)^{-1})
\]

Then,

\[
LM-BP = (2 σ_R^{-4})^{-1} ESS \xrightarrow{d} \chi^2_m
\]

ESS_e = ESS in regression of c_i^2 (= g_i σ_R^2) against z_i.

Since σ_R^{-4} \xrightarrow{d} σ^4 \Rightarrow LM-BP \xrightarrow{d} \chi^2_m

Note: Recall R² = \left[ y'X (X'X)^{-1} X'y - TV^2 \right] / \left[ y'y - TV^2 \right]

Under H₀: E[ω_i] = 0, E[ω_i^2] = 1, then, the LM test can be shown to be asymptotically equivalent to a T R². (Think of \bar{y}=0 and y'y/T=1 above).
Testing for Heteroscedasticity: BP Test

- The LM test is asymptotically equivalent to a $T \ R^2$ test, where $R^2$ is calculated from a regression of $e_i^2/\sigma R^2$ on the variables $Z$.

- Usual calculation of the Breusch-Pagan test
  - Step 1. (Auxiliary Regression). Run the regression of $e_i^2$ on all the explanatory variables, $z$. In our example,
    \[ e_i^2 = \alpha_0 + z_i,1' \alpha_1 + \ldots + z_i,m' \alpha_m + v_i \]
  - Step 2. Keep the $R^2$ from this regression. Let’s call it $R_{e2}^2$. Calculate either
    \begin{align*}
    (a) \quad F &= \frac{R_{e2}^2/m}{(1-R_{e2}^2)/(T-(m+1))}, \text{ which follows a } F_{m,(T-(m+1))} \\
    (b) \quad LM &= T \ R_{e2}^2, \text{ which follows } \chi^2_{m}.
    \end{align*}

Testing for Heteroscedasticity: BP Test

- Variations:
  1. Glesjer (1969) test. Use absolute values instead of $e_i^2$ to estimate the varying second moment. Following our previous example,
     \[ |e_i| = \alpha_0 + z_i,1' \alpha_1 + \ldots + z_i,m' \alpha_m + v_i \]
  2. Harvey-Godfrey (1978) test. Use $\ln(e_i^2)$. Then, the implied model for $\sigma_i^2$ is an exponential model.
     \[ \ln(e_i^2) = \alpha_0 + z_i,1' \alpha_1 + \ldots + z_i,m' \alpha_m + v_i \]

Note: Implied model for $\sigma_i^2 = \exp\{\alpha_0 + z_i,1' \alpha_1 + \ldots + z_i,m' \alpha_m + v_i\}$. 
Testing for Heteroscedasticity: BP Test

- Variations:
  
  (3) Koenker’s (1981) studentized LM test. A usual problem with statistic LM is that it crucially depends on the assumption that $\varepsilon$ is normal. Koenker (1981) proposed studentizing the statistic $LM-BP$ by

  \[
  LM-S = \left(2\sigma_R^4\right)LM-BP / \left[\Sigma (\hat{\varepsilon}_i^2 - \sigma_R^2)^2 / T\right] \rightarrow \chi^2_m
  \]

Testing for Heteroscedasticity: White Test

- Based on the difference between OLS and true OLS variances:

  \[
  \sigma^2(X'\Omega X - XX') = X'\Sigma X - \sigma^2XX' = \Sigma_i (E[\varepsilon_i^2] - \sigma^2)x_i'x_i
  \]

- Empirical counterpart: $(1/T) \Sigma_i (\varepsilon_i^2 - \bar{\varepsilon}^2)x_i'x_i$

- We can express each element of the $k(k+1)$ matrix as:

  \[
  (1/T) \Sigma_i (\varepsilon_i^2 - \bar{\varepsilon}^2) = (\psi_1, \psi_2, \ldots, \psi_m)'
  \]

  \[
  \psi_i = (\psi_{1i}, \psi_{2i}, \ldots, \psi_{mi}), \quad \psi_{ij} = \psi_{pi} \psi_{qi}, \quad p \geq q, \quad p, q = 1, 2, \ldots, k
  \]

- White heteroscedasticity test:

  \[
  W = [(1/T) \Sigma_i (\varepsilon_i^2 - \bar{\varepsilon}^2)]' D_T^{-1} [(1/T) \Sigma_i (\varepsilon_i^2 - \bar{\varepsilon}^2)] \rightarrow \chi^2_m
  \]

  where

  \[
  D_T = \text{Var} [(1/T) \Sigma_i (\varepsilon_i^2 - \bar{\varepsilon}^2)]
  \]

  Note: $W$ is asymptotically equivalent to a $T^2$ test, where $R^2$ is calculated from a regression of $\varepsilon_i^2/\sigma_R^2$ on the $\phi_i$'s.
Testing for Heteroscedasticity: White Test

- Usual calculation of the White test
  - Step 1. (Auxiliary Regression). Regress $e^2$ on all the explanatory variables ($X_j$), their squares ($X_j^2$), and all their cross products. For example, when the model contains $k = 2$ explanatory variables, the test is based on:

  \[ e_i^2 = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i}^2 + \beta_4 x_{2,i}^2 + \beta_5 x_1 x_2,i + v_i \]

  Let $m$ be the number of regressors in auxiliary regression. Keep $R^2$, say $R_{e2}^2$.
  - Step 2. Compute the statistic
    \[ LM = T R_{e2}^2, \text{ which follows a } \chi^2_m \]

Testing for Heteroscedasticity: Remarks

- Drawbacks of the Breusch-Pagan test:
  - It has been shown to be sensitive to violations of the normality assumption.
  - Three other popular LM tests: the Glejser test; the Harvey-Godfrey test, and the Park test, are also sensitive to such violations.

- Drawbacks of the White test
  - If a model has several regressors, the test can consume a lot of df’s.
  - In cases where the White test statistic is statistically significant, heteroscedasticity may not necessarily be the cause, but model specification errors.
  - It is general. It does not give us a clue about how to model heteroscedasticity to do FGLS. The BP test points us in a direction.
Testing for Heteroscedasticity: Remarks

- Drawbacks of the White test (continuation)
  - In simulations, it does not perform well relative to others, especially, for time-varying heteroscedasticity, typical of financial time series.
  - The White test does not depend on normality; but the Koenker’s test is also not very sensitive to normality. In simulations, Koenker’s test seems to have more power – see, Lyon and Tsai (1996) for a Monte Carlo study of the heteroscedasticity tests presented here.

Testing for Heteroscedasticity: Remarks

- General problems with heteroscedasticity tests:
  - The tests rely on the first four assumptions of the CLM being true.
  - In particular, (A2) violations. That is, if the zero conditional mean assumption, then a test for heteroskedasticity may reject the null hypothesis even if \( \text{Var}(y | X) \) is constant.
  - This is true if our functional form is specified incorrectly (omitted variables or specifying a log instead of a level). Recall David Hendry’s comment.

- Knowing the true source (functional form) of heteroscedasticity may be difficult. A practical solution is to avoid modeling heteroscedasticity altogether and use OLS along the White heteroskedasticity-robust standard errors.
Estimation: WLS form of GLS

• While it is always possible to estimate robust standard errors for OLS estimates, if we know the specific form of the heteroskedasticity, we can obtain more efficient estimates than OLS: GLS.

• GLS basic idea: Efficient estimation through the transform the model into one that has homoskedastic errors – called WLS.

• Suppose the heteroskedasticity can be modeled as:
  \[ \text{Var}(\varepsilon \mid x) = \sigma^2 h(x) \]

• The key is to figure out what \( h(x) \) looks like. Suppose that we know \( h_i \). For example, \( h_i(x) = x_j^2 \). (make sure \( h_i \) is always positive.)

• Then, use \( 1 / \sqrt{x_j^2} \) to transform the model.

Estimation: WLS form of GLS

• Suppose that we know \( h_i(x) = x_j^2 \). Then, use \( 1 / \sqrt{x_j^2} \) to transform the model:
  \[ \text{Var}(\varepsilon / \sqrt{h_i} \mid x) = \sigma^2 \]

• Thus, if we divide our whole equation by \( \sqrt{h_i} \) we get a (transformed) model where the error is homoskedastic.

• Assuming weights are known, we have a two-step GLS estimation:
  - Step 1: Use OLS, then the residuals to estimate the weights.
  - Step 2: Weighted least squares using the estimated weights.

• Greene has a proof based on our asymptotic theory for the asymptotic equivalence of the second step to true GLS.
Estimation: FGLS

• More typical is the situation where we do not know the form of the heteroskedasticity. In this case, we need to estimate \( h(x) \).

• Typically, we start by assuming a fairly flexible model, such as

\[
\text{Var}(\varepsilon | x) = \sigma^2 \exp(X\delta) \quad \text{--make sure Var}(\varepsilon_i | x) > 0.)
\]

Since we don’t know \( \delta \), it must be estimated. Our assumption implies that

\[
\varepsilon^2 = \sigma^2 \exp(X\delta) \nu \quad \text{with E}(\nu | X) = 1.
\]

Then, if E(\( \nu \)) = 1

\[
\ln(\varepsilon^2) = X\delta + u \quad (*)
\]

where E(\( u \)) = 0 and u is independent of X.

Now, we know that e is an estimate of \( \varepsilon \), so we can estimate (*) by OLS.

Estimation: FGLS

• Now, an estimate of \( b \) is obtained as \( \hat{b} = \exp(\hat{\delta}) \), and the inverse of this is our weight. Now, we can do GLS as usual.

• Summary of FGLS

1. Run the original OLS model, save the residuals, e. Get ln(e^2).
2. Regress ln(e^2) on all of the independent variables. Get fitted values, \( \hat{\delta} \).
3. Do WLS using 1/exp(\( \hat{\delta} \)) as the weight.
4. Iterate to gain efficiency.

• Remark: We are using WLS just for efficiency – OLS is still unbiased and consistent. Sandwich estimator gives us consistent inferences.
Estimation: MLE

- ML estimates all the parameters simultaneously. To construct the likelihood, we assume a distribution for $\varepsilon$. Under normality (A5):

  $$\ln L = -\frac{T}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{T} \ln \sigma_i^2 - \frac{1}{2} \sum_{i=1}^{T} \frac{1}{\sigma_i^2} (y_i - X_i \beta)'(y_i - X_i \beta)$$

- Suppose $\sigma_i^2 = \exp(\alpha_0 + z_{i1} \alpha_1 + \ldots + z_{im} \alpha_m) = \exp(z_i' \alpha)$

- Then, the first derivatives of the log likelihood wrt $\theta = (\beta, \alpha)$ are:

  $$\frac{\partial \ln L}{\partial \beta} = -\sum_{i=1}^{T} x_i e_i / \sigma_i^2 = X' \Sigma^{-1} e$$

  $$\frac{\partial \ln L}{\partial \alpha_i} = -\frac{1}{2} \sum_{i=1}^{T} 1/\sigma_i^2 \exp(z_i' \alpha) z_i - \left(-\frac{1}{2}\right) \sum_{i=1}^{T} e_i^2 / \sigma_i^4 \exp(z_i' \alpha) z_i - \frac{1}{2} \sum_{i=1}^{T} z_i (e_i^2 / \sigma_i^2 - 1)$$

- Then, we get the f.o.c. We get a non-linear system of equations.

Estimation: MLE

- We take second derivatives to calculate the information matrix:

  $$\frac{\partial \ln L^2}{\partial \beta \partial \beta'} = -\sum_{i=1}^{T} x_i x_i' / \sigma_i^2 = X' \Sigma^{-1} X$$

  $$\frac{\partial \ln L}{\partial \beta \alpha_i} = -\frac{1}{2} \sum_{i=1}^{T} x_i z_i' e_i / \sigma_i^2$$

  $$\frac{\partial \ln L}{\partial \alpha_i \alpha_i'} = -\frac{1}{2} \sum_{i=1}^{T} z_i z_i' e_i^2 / \sigma_i^2$$

- Then,

  $$I(\theta) = E\left[ -\frac{\partial \ln L}{\partial \theta \partial \theta'} \right] = \begin{bmatrix} X' \Sigma^{-1} X & 0 \\ 0 & \frac{1}{2} Z' Z \end{bmatrix}$$

- We can estimate the model using Gauss-Newton:

  $$\theta_{t+1} = \theta_t - H_t^{-1} g_t$$

  $$g_t = \partial \log L / \partial \theta'$$
Estimation: MLE

- We estimate the model using Gauss-Newton:

\[ \theta_{j+1} = \theta_j - H_j^{-1} g_j \quad g_j = \frac{\partial \log L_j}{\partial \theta}' \]

Since \( H_t \) is block diagonal,

\[ \beta_{j+1} = \beta_j - (X' \Sigma_j^{-1} X)^{-1} X' \Sigma_j^{-1} e_j \]

\[ \alpha_{j+1} = \alpha_j - \left( \frac{1}{2} Z'Z \right)^{-1} \left[ \frac{1}{2} \Sigma_i z_i \left( \frac{\epsilon_i^2}{\sigma_i^2} - 1 \right) \right] = \alpha_j - (Z'Z)^{-1} Z'v, \]

where

\[ v = \left( \frac{\epsilon_i^2}{\sigma_i^2} - 1 \right) \]

Convergence will be achieved when \( g_j = \frac{\partial \log L_j}{\partial \theta}' \) is close to zero.

- We have an iterative algorithm \( \Rightarrow \) Iterative FGLS = MLE!

Heteroscedasticity: Log Transformations

- A log transformation of the data, can eliminate (or reduce) a certain type of heteroskedasticity.

- Assume

\[ \mu_t = \text{E}[Z_t] \]

\[ \text{Var}[Z_t] = \delta \mu_t^2 \]

(Variance proportional to the squared mean)

- We log transformed the data: \( \log(Z_t) \). Then, we use the delta method to approximate the variance of the transformed variable. Recall: \( \text{Var}[f(X)] \) using delta method:

\[ \text{Var}[f(X)] \approx f'(\theta)^2 \text{Var}[X] \]

- Then, the variance of \( \log(Z_t) \) is roughly constant:

\[ \text{Var} [\log(Z_t)] \approx \left( \frac{1}{\mu_t} \right)^2 \text{Var} [Z_t] = \delta \]
ARCH Models

• Until the early 1980s econometrics had focused almost solely on modeling the means of series - i.e., their actual values. 
\[ y_t = E_t[y_t|x] + \varepsilon_t, \quad \varepsilon_t \sim D(0, \sigma^2) \]
For an AR(1) process:
\[ E_{t+1}[y_t|x] = E_t[y_t] = a + \beta y_{t-1} \]

Note: The unconditional mean and variance are:
\[ E[y_t] = \frac{a}{(1-\beta)}, \quad \text{Var}_t[y_t] = \frac{\sigma^2}{(1-\beta^2)} \]

The conditional mean is time varying; the unconditional mean is not!

Key distinction: Conditional vs. Unconditional moments.

• Similar idea for the variance
Unconditional variance: \( \text{Var}_t[y_t] = E[(y_t - E[y_t])^2] = \frac{\sigma^2}{(1-\beta^2)} \)
Conditional variance: \( \text{Var}_{t-1}[y_t] = E_{t-1}[(y_t - E_{t-1}[y_t])^2] = E_{t-1}[\varepsilon_t^2] \)

ARCH Models

• The unconditional variance measures the overall uncertainty. In the AR(1) example, time \( t-1 \) information plays no role: \( \text{Var}_t[y_t] = \frac{\sigma^2}{(1-\beta^2)} \)
• The conditional variance, \( \text{Var}_{t-1}[y_t] \), is a better measure of uncertainty at time \( t-1 \). It is a function of information at time \( t-1 \).
ARCH Models: Stylized Facts of Asset Returns

- **Thick tails** - Mandelbrot (1963): leptokurtic (thicker than Normal)

- **Volatility clustering** - Mandelbrot (1963): “large changes tend to be followed by large changes of either sign.”

- **Leverage Effects** – Black (1976), Christie (1982): Tendency for changes in stock prices to be negatively correlated with changes in volatility.

- **Non-trading Effects, Weekend Effects** – Fama (1965), French and Roll (1986): When a market is closed information accumulates at a different rate to when it is open – for example, the weekend effect, where stock price volatility on Monday is not three times the volatility on Friday.

ARCH Models: Stylized Facts of Asset Returns

- **Expected events** – Cornell (1978), Patell and Wolfson (1979), etc: Volatility is high at regular times such as news announcements or other expected events, or even at certain times of day – for example, less volatile in the early afternoon.

- **Volatility and serial correlation** – LeBaron (1992): Inverse relationship between the two.

- **Co-movements in volatility** – Ramchand and Susmel (1998): Volatility is positively correlated across markets/assets.

• We need a model that accommodates all these facts.
ARCH Models: Stylized Facts of Asset Returns

- Easy to check leptokurtosis (Stylized Fact #1)

Figure: Descriptive Statistics and Distribution for Monthly S&P500 Returns

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (%)</td>
<td>0.626332</td>
</tr>
<tr>
<td>(p-value: 0.0004)</td>
<td></td>
</tr>
<tr>
<td>Standard Dev (%)</td>
<td>4.37721</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.43764</td>
</tr>
<tr>
<td>Excess Kurtosis</td>
<td>2.29395</td>
</tr>
<tr>
<td>Jarque-Bera</td>
<td>145.72</td>
</tr>
<tr>
<td>(p-value: &lt;0.0001)</td>
<td></td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.0258</td>
</tr>
<tr>
<td>(p-value: 0.5249)</td>
<td></td>
</tr>
</tbody>
</table>

ARCH Models: Stylized Facts of Asset Returns

- Easy to check Volatility Clustering (Stylized Fact #2)

Figure: Monthly S&P500 Returns (1964:1-2014:9)

- We start with assumptions (A1) to (A5), but with a specific (A3): 
  \[ Y_t = \beta X_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_t^2) \]
  \[ \sigma_t^2 = \text{Var}_{t-1}(\varepsilon_t) = E_{t-1}(\varepsilon_t^2) = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 = \omega + \alpha(L)\varepsilon_t^2 \]
  define \( \nu_t = \varepsilon_t^2 - \sigma_t^2 \)
  \[ \varepsilon_t^2 = \omega + \alpha(L)\varepsilon_t^2 + \nu_t \]

- This is an AR(q) model for squared innovations. That is, we have an ARCH model: Auto-Regressive Conditional Heteroskedasticity

This model estimates the unobservable (latent) variance.

Note: We are dealing with a variance, we usually impose \( \omega > 0 \) and \( \alpha_i > 0 \) for all \( i \).


- The unconditional variance is determined by:
  \[ \sigma^2 = E[\sigma_t^2] = \omega + \sum_{i=1}^{q} \alpha_i E[\varepsilon_{t-i}^2] = \omega + \sum_{i=1}^{q} \alpha_i \sigma^2 \]
  That is, \( \sigma^2 = \frac{\omega}{1 - \sum_{i=1}^{q} \alpha_i} \)
  To obtain a positive \( \sigma^2 \), we impose another restriction: \( 1 - \sum \alpha_i > 0 \).

- Example: ARCH(1)
  \[ Y_t = \beta X_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_t^2) \]
  \[ \sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 \]

- We need to impose restrictions: \( \alpha > 0 \) & \( 1 - \alpha > 0 \).

• Even though the errors may be serially uncorrelated they are not independent: There will be volatility clustering and fat tails. Let’s define standardized errors:
  \[ z_t = \frac{\epsilon_t}{\sigma_t} \]

• They have conditional mean zero and a time invariant conditional variance equal to 1. That is, \( z_t \sim D(0,1) \). If \( z_t \) is assumed to follow a N(0,1), with a finite fourth moment (use Jensen’s inequality). Then:

\[
E(\epsilon_t^4) = E(z_t^4)E(\sigma_t^4) \geq E(z_t^4)E(\epsilon_t^2)^2 = E(z_t^4)E(\epsilon_t^2)^2 = 3E(\epsilon_t^2)^2
\]

\[
\kappa(\epsilon_t) = E(\epsilon_t^4) / E(\epsilon_t^2)^2 \geq 3.
\]

• For an ARCH(1), the 4th moment for an ARCH(1):

\[
\kappa(\epsilon_t) = 3(1 - \alpha^2) / (1 - 3\alpha^2) \quad \text{if } 3\alpha^2 < 1.
\]


• More convenient, but less intuitive, presentation of the ARCH(1) model:

\[ \epsilon_t = \sqrt{\sigma_t^2} \nu_t \]

where \( \nu_t \) is i.i.d. with mean 0, and Var[\( \nu_t \)] = 1. Since \( \nu_t \) is i.i.d., then:

\[
E_{t-1}[\epsilon_t^2] = E_{t-1}[\sigma_t^2 \nu_t^2] = E_{t-1}[\sigma_t^2]E_{t-1}[\nu_t^2] = \omega + \alpha_t \epsilon_{t-1}^2
\]

• It turns out that \( \sigma_t^2 \) is a very persistent process. Such a process can be captured with an ARCH(q), where q is large. This is not efficient.
GARCH: Bollerslev (1986)

- In practice, q is often large. A more parsimonious representation is the Generalized ARCH model or GARCH(q,p):

\[
\sigma^2_t = \omega + \sum_{i=1}^q \alpha_i \varepsilon^2_{t-i} + \sum_{j=1}^p \beta_j \sigma^2_{t-j}
\]

\[
= \omega + \alpha (L) \varepsilon^2 + \beta (L) \sigma^2
\]

define \( \nu_t \equiv \varepsilon^2_t - \sigma^2_t \)

\[
\varepsilon^2_t = \omega + (\alpha (L) + \beta (L)) \varepsilon^2 + \beta (L) \nu_t + \nu_t
\]

which is an ARMA(max(p,q),p) model for the squared innovations.

- Popular GARCH model: GARCH(1,1):

\[
\sigma^2_{t+1} = \omega + \alpha_1 \varepsilon^2_t + \beta_1 \sigma^2_t
\]

GARCH: Bollerslev (1986)

- Technical details: This is covariance stationary if all the roots of \( \alpha (L) + \beta (L) = 1 \) lie outside the unit circle. For the GARCH(1,1) this amounts to \( \alpha_1 + \beta_1 < 1 \).

- Bollerslev (1986) showed that if \( 3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1 \), the second and 4th moments of \( \varepsilon_t \) exist:

\[
E[\varepsilon^2_t] = \frac{\omega}{(1 - \alpha_1 - \beta_1)}
\]

\[
E[\varepsilon^4_t] = \frac{3\omega^2 (1 + \alpha_1 + \beta_1)}{(1 - \alpha_1 - \beta_1) (1 - \beta_1^2 - 2\alpha_1\beta_1 - 3\alpha_1^2)} \quad \text{if } (1 - \beta_1^2 - 2\alpha_1\beta_1 - 3\alpha_1^2) \neq 0
\]
GARCH: Forecasting and Persistence

- Consider the forecast in a GARCH(1,1) model

\[ \sigma_{t+1}^2 = \omega + \alpha_i \epsilon_t^2 + \beta_i \sigma_t^2 = \omega + \sigma_t^2 (\alpha_i \epsilon_t^2 + \beta_t) \quad (\epsilon_t^2 = \sigma_t^2 \epsilon_t^2) \]

Taking expectation at time t

\[ E_i[\sigma_{t+1}^2] = \omega + \sigma_t^2 (\alpha_i 1 + \beta_t) \]

Then, by repeated substitutions:

\[ E_i[\sigma_{t+j}^2] = \omega \left[ \sum_{i=0}^{j-1} (\alpha_i + \beta_t)^i \right] + \sigma_t^2 (\alpha_i + \beta_t)^j \]

As \( j \to \infty \), the forecast reverts to the unconditional variance:

\[ \omega / (1 - \alpha_i - \beta_i). \]

- When \( \alpha_i + \beta_i = 1 \), today’s volatility affect future forecasts forever:

\[ E_i[\sigma_{t+j}^2] = \sigma_t^2 + j \omega \]

GARCH-X

- In the GARCH-X model, exogenous variables are added to the conditional variance equation.

Consider the GARCH(1,1)-X model:

\[ \sigma_i^2 = \omega + \alpha_i \epsilon_{i-1}^2 + \beta_i \sigma_{i-1}^2 + \delta f(X_{i-1}, \theta) \]

where \( f(X_{i-1}, \theta) \) is strictly positive for all \( t \). Usually, \( X_t \) is an observed economic variable or indicator, say liquidity index, and \( f(.) \) is a non-linear transformation, which is non-negative.

Recall the technical detail: The standard GARCH model:
\[ \sigma_j^2 = \omega + \alpha(L)\epsilon_t^2 + \beta(L)\sigma_t^2 \]
is covariance stationary if \( \alpha(1) + \beta(1) < 1 \).

But strict stationarity does not require such a stringent restriction (That is, that the unconditional variance does not depend on \( t \)).

In the GARCH(1,1) model, if \( \alpha_1 + \beta_1 = 1 \), we have the Integrated GARCH (IGARCH) model.

In the IGARCH model, the autoregressive polynomial in the ARMA representation has a unit root: a shock to the conditional variance is “persistent.”

Variance forecasts are generated with: 
\[ E_{t-1}(\sigma_j^2) = \sigma_j^2 + j\omega \]
That is, today’s variance remains important for future forecasts of all horizons.

Nelson (1990) establishes that, as this satisfies the requirement for strict stationarity, it is a well defined model.

In practice, it is often found that \( \alpha_1 + \beta_1 \) are close to 1.

It is often argued that IGARCH is a product of omitted variables; For example, structural breaks. See Lamoreux and Lastrapes (1989), Hamilton and Susmel (1994), & Mikosch and Starica (2004).

**GARCH: Variations**

  It models an exponential function for the time-varying variance:
  \[
  \log(\sigma_i^2) = \omega + \sum_{i=1}^{q} \alpha_i (z_{i,j} + \gamma |z_{i,j}| - E[z_{i,j}]) + \sum_{j=1}^{p} \beta_j \log(\sigma_{i-j}^2)
  \]

- **GJR-GARCH model** -- Glosten Jagannathan and Runkle (JF, 1993):
  \[
  \sigma_i^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{i-i}^2 + \sum_{i=1}^{q} \alpha_i \epsilon_{i-i}^2 * I_{i-i} + \sum_{j=1}^{q} \beta_j \sigma_{i-j}^2
  \]
  where \(I_{i-i} = 1\) if \(\epsilon_{i-i} < 0\); 0 otherwise.

- **Remark:** Both models capture sign (asymmetric) effects in volatility: Negative news (\(\epsilon_{t-i} < 0\)) increase the conditional volatility (leverage effect).

---

**GARCH: Variations**


  These models apply the Box-Cox-type transformation to the conditional variance:
  \[
  \sigma_i^2 = \omega + \sum_{i=1}^{q} \alpha_i |\epsilon_{i-i}|^{\gamma} + \sum_{j=1}^{p} \beta_j \sigma_{i-j}^{\gamma}
  \]
  Special case: \(\gamma = 2\) (standard GARCH model).

  **Note:** The variance depends on both the size and the sign of the variance which helps to capture leverage type (asymmetric) effects.
GARCH: Variations

• Threshold ARCH (TARCH) -- Rabemananjara, R. and J.M. Zakoian (JAE, 1993).

Large events to have an effect but no effect from small events

\[ \sigma_t^2 = \omega + \sum_{i=1}^{q} (\alpha^+ I(\epsilon_{t-i} > \kappa) + \alpha^- I(\epsilon_{t-i} \leq \kappa)) \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \]

There are two variances:

\[ \sigma_{t-1}^2 = \omega + \sum_{i=1}^{q} \alpha^+ \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \quad \text{if } (\epsilon_{t-i} > \kappa) \]

\[ \sigma_{t-1}^2 = \omega + \sum_{i=1}^{q} \alpha^- \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \quad \text{if } (\epsilon_{t-i} \leq \kappa) \]

Many other versions are possible by adding minor asymmetries or non-linearities in a variety of ways.

GARCH: Variations

• Switching ARCH (SWARCH) - Hamilton and Susmel (JE, 1994).

Intuition: \( \sigma_t^2 \) depends on the state of the economy -- regime. It’s based on Hamilton’s (1989) time series models with changes of regime:

\[ \sigma_t^2 = \omega_{\gamma_t} + \sum_{i=1}^{q} \alpha_i \gamma_{t-i} \epsilon_{t-i}^2 \]

The key is to select a parsimonious representation:

\[ \frac{\sigma_t^2}{\gamma_{t}} = \omega + \sum_{i=1}^{q} \alpha_i \frac{\epsilon_{t-i}^2}{\gamma_{t-i}} \]

For a SWARCH(1) with 2 states (1 and 2) we have 4 possible \( \sigma_t^2 \):

\[ \sigma_t^2 = \omega \gamma_1 + \alpha_1 \epsilon_{t-1}^2 / \gamma_1, \quad s_t = 1, s_{t-1} = 1 \]
\[ \sigma_t^2 = \omega \gamma_1 + \alpha_1 \epsilon_{t-1}^2 / \gamma_2, \quad s_t = 1, s_{t-1} = 2 \]
\[ \sigma_t^2 = \omega \gamma_2 + \alpha_1 \epsilon_{t-1}^2 / \gamma_1, \quad s_t = 2, s_{t-1} = 1 \]
\[ \sigma_t^2 = \omega \gamma_2 + \alpha_1 \epsilon_{t-1}^2 / \gamma_2, \quad s_t = 2, s_{t-1} = 2 \]
ARCH Estimation: MLE

• All of these models can be estimated by maximum likelihood. First we need to construct the sample likelihood.

• Since we are dealing with dependent variables, we use the conditioning trick to get the joint distribution:

\[ f(y_{1}, y_{2}, \ldots, y_{T}; \theta) = f(y_{1} \mid x_{1}; \theta) f(y_{2} \mid y_{1}, x_{2}; \theta) \ldots \]
\[ \ldots f(y_{T} \mid y_{T-1}, \ldots, y_{1}, x_{T-1}, \ldots, x_{1}; \theta). \]

Taking logs
\[ L = \log(f(y_{1}, y_{2}, \ldots, y_{T}; \theta)) = \log(f(y_{1} \mid x_{1}; \theta)) + \log(f(y_{2} \mid y_{1}, x_{2}; \theta)) + \ldots + \]
\[ \ldots + \log(f(y_{T} \mid y_{T-1}, \ldots, y_{1}, x_{T-1}, \ldots, x_{1}; \theta)) \]
\[ = \sum_{t=1}^{T} \log(f(y_{t} \mid y_{t-1}, X_{t}; \theta)) \]

ARCH Estimation: MLE

• Note that the \( \delta L / \delta \gamma = 0 \) (k f.o.c.’s) will give us GLS.

• Denote \( \delta L / \delta \theta = S(y_{t}, \theta) = 0 \) (\( S(.) \) is the score vector)

- We have a \((k+2)(k+2)\) system. But, it is a non-linear system. We will need to use numerical optimization. Gauss-Newton or BHHH (also approximates \( H \) by the product of \( S(y_{t}, \theta) \)’s) can be easily implemented.

- Given the AR structure, we will need to make assumptions about \( \sigma_{0} \) (and \( \varepsilon_{0}, \varepsilon_{1}, \ldots, \varepsilon_{p} \) if we assume an AR(p) process for the mean).

- Alternatively, we can take \( \sigma_{0} \) (and \( \varepsilon_{0}, \varepsilon_{1}, \ldots, \varepsilon_{p} \)) as parameters to be estimated (it can be computationally more intensive and estimation can lose power.)
ARCH Estimation: MLE

- If the conditional density is well specified and \(\theta_0\) belongs to \(\Omega\), then

\[
T^{1/2}(\hat{\theta} - \theta_0) \to N(0, A_0^{-1}), \quad \text{where } A_0^{-1} = T^{-1}\sum_{t=1}^T \frac{\partial S_t(y_t, \theta_0)}{\partial \theta}
\]

- Under the correct specification assumption, \(A_0 = B_0\), where

\[
B_0 = T^{-1}\sum_{t=1}^T E[S_t(y_t, \theta_0), S_t(y_t, \theta_0)']
\]

We estimate \(A_0\) and \(B_0\) by replacing \(\theta_0\) by its estimated MLE value.

- The estimator \(B_0\) has a computational advantage over \(A_0\): Only first derivatives are needed. But \(A_0 = B_0\) only if the distribution is correctly specified. This is very difficult to know in practice.

- Common practice in empirical studies: Assume the necessary regularity conditions are satisfied.

ARCH Estimation: MLE - Example

- Assuming normality, we maximize with respect to \(\theta\) the function:

\[
L = \sum_{i=1}^T \log(f_i) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^T \left(-\log(\sigma_i^2) - \frac{\varepsilon_i^2}{\sigma_i^2}\right)
\]

Example: ARCH(1) model.

\[
L = \sum_{i=1}^T \log(f_i) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^T \log(\omega + \alpha_i \varepsilon_{i-1}^2) - \frac{1}{2} \sum_{i=1}^T \varepsilon_i^2 / (\omega + \alpha_i \varepsilon_{i-1}^2)
\]

Taking derivatives with respect to \(\theta = (\omega, \alpha, \gamma)\), where \(\gamma = k\) mean pars:

\[
\frac{\partial L}{\partial \omega} = -\sum_{i=1}^T 1 / (\omega + \alpha_i \varepsilon_{i-1}^2) - (-1/2) \sum_{i=1}^T \varepsilon_i^2 / (\omega + \alpha_i \varepsilon_{i-1}^2)^2
\]

\[
\frac{\partial L}{\partial \alpha_i} = -\sum_{i=1}^T \varepsilon_{i-1}^2 / (\omega + \alpha_i \varepsilon_{i-1}^2) - (-1/2) \sum_{i=1}^T \varepsilon_i^2 \varepsilon_{i-1}^2 / (\omega + \alpha_i \varepsilon_{i-1}^2)^2
\]

\[
\frac{\partial L}{\partial \gamma} = -\sum_{i=1}^T x_i \varepsilon_i / \sigma_i^2
\]
ARCH Estimation: MLE

• Then, we set the f.o.c. \( \frac{\delta L}{\delta \theta} = 0. \)

• We have a \((k+2)\) system. It is a non-linear system. We will use
numerical optimization; usually, Gauss-Newton or BHHH.

• Again, note that \( \frac{\delta L}{\delta \gamma} = 0 \) \((k\ f.o.c.’s)\) will give us GLS.

\[
\frac{\partial L}{\partial \gamma} = -\sum_{i=1}^{T} x_i e_i (\gamma_{MLE}) / \sigma_i^2 (\omega_{MLE}, \alpha_{MLE}, \cdot) = 0
\]

ARCH Estimation: MLE – Example (in R)

• Log likelihood of GARCH(1,1) Model:

```r
log_lik_garch11 <- function(theta, data) {
  mu <- theta[1]; alpha0 <- abs(theta[2]); alpha1 <- abs(theta[3]); beta1 <- abs(theta[4]);
  chk0 <- (1 - alpha1 - beta1)
  r <- ts(data)
  n <- length(r)
  u <- vector(length=n); u <- ts(u)
  for (t in 2:n) {
    u[t] = r[t] - mu # this setup allows for ARMA in mean
  }
  h <- vector(length=n); h <- ts(h)
  h[1] = alpha0/chk0 # set initial value for h[t] series
  if (chk0==0) {h[1]=.000001} # check to avoid dividing by 0
  for (t in 2:n) {
    h[t] = abs(alpha0 + alpha1*(u[t-1]^2)+ beta1*h[t-1])
    if (h[t]==0) {h[t]=.00001} #check to avoid log(0)
  }
  return(-sum(-0.5*log(2*pi) - 0.5*log(abs(h[2:n])) - 0.5*(u[2:n]^2)/abs(h[2:n])))
}
```
ARCH Estimation: MLE – Example (in R)

• To maximize the likelihood we use optim (mln can also be used):

```r
dat_xy <- read.csv("C:/IFM/datastream-K-DIS.csv",head=TRUE,sep="","")
summary(dat_xy)
names(dat_xy)

z <- dat_xy$SP500 # S&P 500 90-2016 monthly data
theta0 = c(0.01, -0.1, -0.001, 0.2, 0.7) # initial values
ml_2 <- optim(theta0, log_lik_garch11, data=z, method="BFGS", hessian=TRUE)
I_Var_m2 <- ml_2$hessian
eigen(I_Var_m2) # check if Hessian is pd.
sqrt(diag(solve(I_Var_m2))) # parameters SE
```

GARCH(1,1): Example – USD/CHF

• We estimate an AR(1)-GARCH(1,1) for $S_t$ - USD/CHF:

\[
\begin{align*}
    s_t &= [\log(S_t) - \log(S_{t-1})] \times 100 = a_0 + a_1 s_{t-1} + \varepsilon_t, \\
    \sigma_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2.
\end{align*}
\]


The estimated model for $s_t$ is given by:

\[
\begin{align*}
    s_t &= -0.055 + 0.059 s_{t-1}, \\
    \sigma_t^2 &= 0.140 + 0.079 \varepsilon_{t-1}^2 + 0.876 \sigma_{t-1}^2.
\end{align*}
\]

Note: $\alpha_1 + \beta_1 = 0.955 < 1$. (Very persistent!)
ARCH Estimation: MLE – Regularity Conditions

Note: The appeal of MLE is the optimal properties of the resulting estimators under ideal conditions.

• Crowder (1976) gives one set of sufficient regularity conditions for the MLE in models with dependent observations to be consistent and asymptotically normally distributed.

• Verifying these regularity conditions is very difficult for general ARCH models - proof for special cases like GARCH(1,1) exists.

For example, for the GARCH(1,1) model: if $E[\ln(\alpha_1 z_t^2 + \beta_0)] < 0$, the model is strictly stationary and ergodic. See Nelson (1990) & Lumsdaine (1996).
ARCH Estimation: MLE – Regularity Conditions

- Block-diagonality
  In many applications of ARCH, the parameters can be partitioned into mean parameters, $\theta_1$, and variance parameters, $\theta_2$.

Then, $\delta \mu_t(\theta)/\delta \theta_2=0$ and, although, $\delta \sigma_t(\theta)/\delta \theta_1 \neq 0$, the Information matrix is block-diagonal (under general symmetric distributions for $z_t$ and for particular ARCH specifications).

Not a bad result:
- Regression can be consistently done with OLS.
- Asymptotically efficient estimates for the ARCH parameters can be obtained on the basis of the OLS residuals.

ARCH Estimation: MLE – Remarks

- But, block diagonality cannot buy everything:
  - Conventional OLS standard errors could be terrible.
  - When testing for serial correlation, in the presence of ARCH, the conventional Bartlett s.e. $- T^{1/2} -$ could seriously underestimate the true standard errors.
ARCH Estimation: QMLE

- The assumption of conditional normality is difficult to justify in many empirical applications. But, it is convenient.

- The MLE based on the normal density may be given a quasi-maximum likelihood (QMLE) interpretation.

- If the conditional mean and variance functions are correctly specified, the normal quasi-score evaluated at $\theta_0$ has a martingale difference property:

$$E\{\delta L/\delta \theta \mid y_t, \theta_0\} = 0$$

Since this equation holds for any value of the true parameters, the QMLE, say $\hat{\theta}_{QMLE}$ is Fisher-consistent — i.e., $E[S(y_T, y_{T-1}, \ldots y_1 ; \theta)] = 0$ for any $\theta \in \Omega$.

ARCH Estimation: QMLE

- The asymptotic distribution for the QMLE takes the form:

$$T^{1/2}(\hat{\theta}_{QMLE} - \theta_0) \rightarrow N(0, A_0^{-1} B_0 A_0^{-1})$$

The covariance matrix $(A_0^{-1} B_0 A_0^{-1})$ is called “robust.” Robust to departures from “normality.”

- Bollerslev and Wooldridge (1992) study the finite sample distribution of the QMLE and the Wald statistics based on the robust covariance matrix estimator:

For symmetric departures from conditional normality, the QMLE is generally close to the exact MLE.

For non-symmetric conditional distributions both the asymptotic and the finite sample loss in efficiency may be large.
ARCH Estimation: Non-Normality

• The basic GARCH model allows a certain amount of leptokurtosis. It is often insufficient to explain real world data.

Solution: Assume a distribution other than the normal which help to allow for the fat tails in the distribution.

• t Distribution - Bollerslev (1987)
The t distribution has a degrees of freedom parameter which allows greater kurtosis. The t likelihood function is

\[ l_t = \ln(\Gamma(0.5(v+1)))\Gamma(0.5v)^{-1}(v-2)^{-1/2}(1+z_t^2)^{-v/2} - 0.5\ln(\sigma_t^2) \]

where \( \Gamma \) is the gamma function and \( v \) is the degrees of freedom. As \( v \to \infty \), this tends to the normal distribution.

• GED Distribution - Nelson (1991)

ARCH Estimation: GMM

• Suppose we have an ARCH(q). We need moment conditions:

\[ (1) - E[m_1] = E[\gamma(y_i - x_i\gamma)] = 0 \]
\[ (2) - E[m_2] = E[\epsilon_i^2(\epsilon_i^2 - \sigma_i^2)] = 0 \]
\[ (3) - E[m_3] = E[\epsilon_i^3 - \omega/1 - \alpha_1 - \ldots - \alpha_q)] = 0 \]

Note: (1) refers to the conditional mean, (2) refers to the conditional variance, and (3) to the unconditional mean.

• GMM objective function:

\[ Q(X, y; \theta) = \hat{E}[m(\theta; X, y)]' W \hat{E}[m(\theta; X, y)] \]

where

\[ \hat{E}[m(\theta; X, y)] = [\hat{E}[m_1], \hat{E}[m_2], \ldots, \hat{E}[m_q]]' \]
ARCH Estimation: GMM

• \( \gamma \) has \( k \) free parameters; \( \alpha \) has \( q \) free parameters. Then, we have \( r = k + q + 1 \) parameters.

\[ m(\theta;X,y) \text{ has } r = k + q + 1 \text{ equations.} \]

**Dimensions:** \( Q \) is \( 1 \times 1 \); \( E[m(\theta;X,y)] \) is \( r \times 1 \); \( W \) is \( r \times r \).

• Problem is *over-identified*: more equations than parameters so cannot solve \( E[m(\theta;X,y)]=0 \), exactly.

• Choose a weighting matrix \( W \) for objective function and minimize using numerical optimization.

• Optimal weighting matrix: \( W = [E[m(\theta;X,y)]E[m(\theta;X,y)]]^{-1} \).

\[ \text{Var}(\theta) = (1/T)[DW^{-1}D']^{-1}, \]

where \( D = \delta E[m(\theta;X,y)]/\delta \theta' \) —expressions evaluated at \( \theta_{GMM} \).

ARCH Estimation: Testing

• Standard BP test, with auxiliary regression given by:

\[ c_t^2 = \alpha_0 + \alpha_1 c_{t-1}^2 + \ldots + \alpha_m c_{t-q}^2 + v_t \]

\( H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_q = 0 \) (No ARCH). It is not possible to do GARCH test, since we are using the same lagged squared residuals.

Then, the LM test is \( (T-q)R^2 \overset{d}{\rightarrow} \chi^2_q \) —Engle’s (1982).

• In ARCH Models, testing as usual: LR, Wald, and LM tests.

Reliable inference from the LM, Wald and LR test statistics generally does require moderately large sample sizes of at least two hundred or more observations.
ARCH Estimation: Testing

- Issues:
  - Non-negative constraints must be imposed. $\theta_0$ is often on the boundary of $\Omega$. (Two sided tests may be conservative)
  - Lack of identification of certain parameters under $H_0$ creates a singularity of the Information matrix under $H_0$. For example, under $H_0$: $\gamma_1 = 0$ (No ARCH), in the GARCH(1,1), $\omega$ and $\beta_1$ are not jointly identified. See Davies (1977).

- Ignoring ARCH
  - You suspect $y_t$ has an AR structure: $y_t = \gamma_0 + \gamma_1 y_{t-1} + \varepsilon_t$
  Hamilton (2008) finds that OLS t-test with no correction for ARCH spuriously reject $H_0$: $\gamma_1 = 0$ with arbitrarily high probability for sufficiently large $T$. White’s (1980) SE help. NW SE help less.

ARCH Estimation: Testing

**Figure.** From Hamilton (2008), Fraction of samples in which OLS t-test leads to rejection of $H_0$: $\gamma_1 = 0$ as a function of $T$ for regression with Gaussian errors (solid line) and Student’s t errors (dashed line). Note: $H_0$ is actually true and test has nominal size of 5%.
ARCH: Which Model to Use

• Questions
  1) Lots of ARCH models. Which one to use?
  2) Choice of $p$ and $q$. How many lags to use?

• Hansen and Lunde (2004) compared lots of ARCH models:
  - It turns out that the GARCH(1,1) is a great starting model.
  - Add a leverage effect for financial series and it’s even better.
  - A t-distribution is also a good addition.

Realized Volatility (RV) Models

• French, Schwert and Stambaugh’s (1987) use higher frequency to estimate the variance as:
  \[ s_t^2 = \frac{1}{k} \sum_{i=1}^{k} r_{t+i}^2 \]
  where $r_t$ is realized returns in days, and we estimate monthly variance.

• Model-free measure –i.e., no need for ARCH-family specifications. It is very popular for intra-daily data, called high frequency (HF) data. The measure is called realized volatility (RV).

• Very popular to calculate intra-day or daily volatility. For example, based on TAQ data, say, 1’ or 10’ realized returns ($r_{ij}$ is $j^{th}$ interval return on day $t$), we can calculate the daily variance, or $RV_t$:
  \[ RV_t = \sum_{j=1}^{M} r_{t,j}^2, \quad t = 1, 2, \ldots, T \]
Realized Volatility (RV) Models

\[ RV_t = \sum_{j=1}^{\nu} r_{t,j}^2, \quad t = 1, 2, \ldots, T \]

where \( r_{t,j} \) is \( j \)th interval return on day \( t \). That is, RV is defined as the sum of squared intraday returns.

- We can use time series models –say an ARIMA- for \( RV_t \) to forecast daily volatility.

- RV is affected by microstructure effects: bid-ask bounce, infrequent trading, calendar effects, etc.. For example, the bid-ask bounce induces serial correlation in intra-day returns, which biases \( RV_t \). (Big problem!)

- Proposed Solution: Filter the intra-day returns using MA or AR models before constructing RV measures.
Realized Volatility (RV) Models - Properties

- Under some conditions (bounded kurtosis and autocorrelation of squared returns less than 1), RV_t is consistent and m.s. convergent.
- Realized volatility is a measure. It has a distribution.
- For returns, the distribution of RV is non-normal (as expected). It tends to be skewed right and leptokurtic. For log returns, the distribution is approximately normal.
- Daily returns standardized by RV measures are nearly Gaussian.
- RV is highly persistent.
- The key problem is the choice of sampling frequency (or number of observations per day).

— Bandi and Russell (2003) propose a data-based method for choosing frequency that minimizes the MSE of the measurement error.
— Simulations and empirical examples suggest optimal sampling is around 1-3 minutes for equity returns.

RV Models - Variation

- Another method: AR model for volatility:
  \[ |\varepsilon_t| = \alpha + \gamma |\varepsilon_{t-1}| + \nu_t \]

The \( \varepsilon_t \) are estimated from a first step procedure -i.e., a regression. Asymmetric/Leverage effects can also be introduced.

OLS estimation possible. Make sure that the variance estimates are positive.
Other Models - Parkinson’s (1980) estimator

• The Parkinson’s (1980) estimator:
  
  \[ s_t^2 = \{ \sum_t [\ln(H_t) - \ln(L_t)]^2 / (4\ln(2)) \}, \]

  where \( H_t \) is the highest price and \( L_t \) is the lowest price.

• There is an RV counterpart, using HF data: Realized Range (RR):
  
  \[ RR_t = \{ \sum_j [100 \times (\ln(H_{t,j}) - \ln(L_{t,j}))]^2 / (4\ln(2)) \}, \]

  where \( H_{t,j} \) and \( L_{t,j} \) are the highest and lowest price in the \( j \)th interval.

• These “range estimators are very good and very efficient.


Stochastic volatility (SV/SVOL) models

• Now, instead of a known volatility at time \( t \), like ARCH models, we allow for a stochastic shock to \( \sigma_t, \eta_t; \)
  
  \[ \sigma_t = \omega + \beta \sigma_{t-1} + \eta_t; \quad \eta_t \sim N(0, \sigma^2) \]

  Or using logs:
  
  \[ \log \sigma_t = \omega + \beta \log \sigma_{t-1} + \nu_t; \quad \nu_t \sim N(0, \sigma^2) \]

• The difference with ARCH models: The shocks that govern the volatility are not necessarily \( \varepsilon_t \)’s.

• Usually, the standard model centers log volatility around \( \omega; \)
  
  \[ \log \sigma_t = \omega + \beta (\log \sigma_{t-1} - \omega) + \nu_t \]

Then,

\[ E[\log(\sigma_t)] = \omega \]
\[ \text{Var}[\log(\sigma_t)] = \kappa^2 = \sigma^2/(1-\beta^2). \]

\[ \Rightarrow \text{Unconditional distribution: } \log(\sigma_t) \sim N(\omega, \kappa^2) \]
### Stochastic volatility (SV/SVOL) models

- Like ARCH models, SV models produce returns with kurtosis > 3 (and, also, positive autocorrelations between squared excess returns):

\[
\begin{align*}
\text{Var}[r_t] &= E[(r_t - E[r_t])^2] = E[\sigma_t^2 z_t^2] = E[\sigma_t^2] E[z_t^2] \\
&= E[\sigma_t^2] = \exp(2\omega + 2\kappa^2) \quad \text{(property of log normal)}
\end{align*}
\]

\[
\begin{align*}
\text{kurt}[r_t] &= E[(r_t - E[r_t])^4] / \{E[(r_t - E[r_t])^2]^2 \} \\
&= E[\sigma_t^4] E[z_t^4] / \{E[\sigma_t^2]^2 \} E[z_t^2]^2 \\
&= 3 \exp(4\omega + 8\kappa^2) / \exp(4\omega + 4\kappa^2) = 3 \exp(4\kappa^2) > 3!
\end{align*}
\]

- We have 3 SVOL parameters to estimate: \( \psi = (\omega, \beta, \sigma) \).

- Estimation:
  - Bayesian: Using MCMC methods (mainly, Gibbs sampling). Modern approach.

### Stochastic volatility (SV/SVOL) models

- The Bayesian approach takes advantage of the idea of hierarchical structure:
  - \( f(y | h) \) (distribution of the data given the volatilities)
  - \( f(h | \psi) \) (distribution of the volatilities given the parameters)
  - \( f(\psi) \) (distribution of the parameters)

Algorithm: MCMC (JPR (1994).)

Augment the parameter space to include \( h_t \).

Using a proper prior for \( f(h_t | \psi) \) MCMC methods provide inference about the joint posterior \( f(h_t, \psi | y) \). We'll go over this topic in Lecture 17.