Asymptotic Theory for ARCH Models: Estimation and Testing

Andrew A. Weiss


Stable URL:
http://links.jstor.org/sici?sici=0266-4666%28198604%292%3A1%3C107%3AATFAME%3E2.0.CO%3B2-T

*Econometric Theory* is currently published by Cambridge University Press.

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/cup.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.
ASYMPTOTIC THEORY FOR
ARCH MODELS:
ESTIMATION AND TESTING

ANDREW A. WEISS

University of Southern California at Los Angeles

In the context of a linear dynamic model with moving average errors, we consider a heteroscedastic model which represents an extension of the ARCH model introduced by Engle [4]. We discuss the properties of maximum likelihood and least squares estimates of the parameters of both the regression and ARCH equations, and also the properties of various tests of the model that are available. We do not assume that the errors are normally distributed.

1. INTRODUCTION

We begin with the situation in which a researcher wishes to model the heteroscedasticity in a time series regression. For this, Engle [4] has introduced the concept of autoregressive conditional heteroscedasticity (ARCH). This is seen as an extension of time series behavior in the mean, allowing the variance of the errors to change if the process takes into account past experience but assumes it constant if this experience is not known. In a process with stochastic regressors, which is the case in most time series processes, this corresponds to the usual properties of the mean of the output from the regression model. Hence it is more appealing than the common assumption of unconditional heteroscedasticity, i.e., where the unconditional variance changes through time, which also implies, for example, that the stochastic variables in the process cannot affect both the mean and variance of the dependent variable.

The heteroscedasticity model we analyze extends that in Engle [4] to a more general form of conditional heteroscedasticity. The conditional variance of the errors is given as a function of lagged errors, lagged dependent variables, a forecast of the dependent variable and exogenous variables. Furthermore, we do not assume that the conditional distribution of the errors

I would like to thank the referees and co-editor for their very helpful comments.

© 1986 Cambridge University Press
is normal. Hence, as Engle and Kraft [6] note, this allows the analysis of reduced form errors via ARCH when the errors in a linear structural equation system are ARCH. In particular, even if the errors in the structural equations are conditionally normal, those in the reduced form may not be. Examples of the use of these models may be found in Pagan, Hall, and Trivedi [15], Weiss [19], Engle [4] and Engle and Kraft [6].

In this paper we consider the estimation of the model, the asymptotic properties of the estimates and sufficient conditions for these to hold. Given the specification of equations for both the mean and variance, the basic estimation technique is maximum likelihood (ML) and the likelihood function (LF) is derived as though the errors are, in fact, conditionally normal. As we shall see, the LF is still maximized at the true parameters and hence the estimates have the usual asymptotic properties. However, because the LF is not correct, the form of the covariance matrix of the (quasi-) ML estimates is more complex than the usual inverse of the information matrix. We also consider the least squares (LS) estimation of the mean and variance equations and tests associated with the model. In the LS estimation, the covariance matrices are affected by the heteroscedasticity.

The rest of the paper begins, in Section 2, by specifying the model, stating some preliminary assumptions, and summarizing the results. ML and LS estimation are considered in Sections 3 and 4, respectively, while Section 5 contains some concluding comments.

2. THE MODEL, PRELIMINARY ASSUMPTIONS, AND SUMMARY

The regression equation is given by a linear dynamic model with moving average errors, i.e.,

$$a(B)(y_t - \mu) = \beta'x_t + b(B)e_t,$$

where $y_t$ is a scalar output with mean $\mu$, $a(B)$ and $b(B)$ are the usual polynomials in the lag operator $B$ of lengths $p$ and $q$, respectively, i.e.,

$$a(B) = 1 - a_1 B - \cdots - a_p B^p$$

and

$$b(B) = 1 + b_1 B + \cdots + b_q B^q,$$

$x_t$ contains $k$ strongly exogenous variables, $\beta$ is a $(k \times 1)$ vector of fixed parameters and the error $e_t$ is a random variable following an ARCH process. We assume that the roots of $a(z) = 0$ and $b(z) = 0$ lie outside the unit circle.
and that the process generating the \( x_t \) is second order stationary. Since \( \mu = E(y_t) \), this implies that \( x_t \) is measured from its mean and that any deterministic components in \( x_t \), e.g., seasonal dummies, have been removed and so do not feed into the level (or variance) of \( y_t \).\(^1\) To help ensure identification, we also assume, as usual, that \( \mu \) and the elements of \( x_t \) are linearly independent.

To specify the ARCH equation, assume

\[
E(\varepsilon_t | I_{t-1}) = 0, \tag{2}
\]

where \( I_{t-1} \) is the information set containing information about the process up to and including time \( t - 1 \). Then the conditional variance of \( \varepsilon_t \) is given by

\[
h_t = E(\varepsilon_t^2 | I_{t-1})
= \alpha_0 + \sum_{i=1}^{R} \alpha_i \varepsilon_{t-i}^2 + \delta_0 (y_t^* - \mu)^2 + \sum_{i=1}^{S} \delta_i (y_{t-i} - \mu)^2 + w_t' P w_t, \tag{3}
\]

where \( w_t' = (x_t^2, x_{t-1}^2, \ldots, x_{t-u+1}^2) \) and \( P \) is diagonal. \( y_t^* \) is the forecast of \( y_t \) from time \( t - 1 \) which, under a least squares criterion, is given by

\[
y_t^* = E(y_t | I_{t-1})
= a_0 y_{t-1} + \cdots + a_p y_{t-p} + \beta' \hat{x}_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q}, \tag{4}
\]

where \( \hat{x}_t = E(x_t | I_{t-1}) \). We do not specify the conditional distribution of \( \varepsilon_t \) further, other than assuming that the distribution is continuous.

Equation (3) represents an extension of the ARCH process in Engle [4] since this included only the lagged squared errors. The further inclusion of cross products of \( \varepsilon_t, y_t, \) and \( x_t \) in the ARCH equation is straightforward. We shall continue to refer to equation (3) as ARCH and the process with \( \delta_i = 0, i = 0, \ldots, S, \) and \( P = 0 \) will be “Engle ARCH.”

If \( x_t \) contains contemporaneous variables, then associated with equations (3) and (4) will be a forecast equation for \( x_t \). To keep within a single equation context, and following Granger [9], we assume \( x_t \) contains only lagged variables, i.e., \( \hat{x}_t = x_t \). Multivariate ARCH processes have been considered by, for example, Granger and Robins [10], who study an application of a particular bivariate ARCH process and Engle and Kraft [6] who extend Engle [4] to multivariate Engle ARCH. These studies are either less general in their use of lagged dependent variables or are empirical in nature, and hence relative to the process considered here can afford the additional variables. Presumably, of course, the results we give can be extended to more general systems, but this remains a topic for future research.

For the identification of the parameters in the ARCH equation, we require that the right-hand side variables in equation (3) are linearly independent.
Notice that linear combinations of these variables being equal to zero implies a quadratic in $e_{t-1}$ and hence two solutions. We therefore rule this out, and assume that $x_t$ and the squared elements of $w_t$ are linearly independent and that conditional of $I_{t-1}$ and future $x_t$, $e_t$ cannot take on only two values almost surely. This is formalized in Lemma 3.2.

Finally, because of equation (3), $e_t$ and $x_t$ are not independent. We require instead, that in addition to equation (2) and given $I_{t-1}$,

$$E(e_t|x_{t+1}) = 0, \quad i > 0. \quad (5)$$

Note however that this condition, and the previous one on $e_t$, are satisfied by assuming that conditional on $I_{t-1}$, $e_t$ and future $x_t$ are independent. But this follows because $x_t$ is strongly exogenous and so $e_t$ ($y_t$) does not Granger cause future $x_t$, given $x_t, x_{t-1}, \ldots$ (see Engle, Hendry, and Richard [5]).

We define the vector of parameters $\theta = (v':m')'$, where $v' = (\alpha_0, \ldots, \alpha_\kappa \delta_0, \ldots, \delta_s P_{11} \ldots P_{kk_k})$ are the parameters in the ARCH or variance equation and $m' = (a_1, \ldots, a_\mu \beta' b_1 \ldots b_4)$ are those in the mean equation. In $v$, $P_{ii}$ is the $i$th diagonal element of $P$. We assume $\theta \in \Theta$, where $\Theta$ is a compact subspace of Euclidean space. $\Theta$ will be determined such that for each $\theta \in \Theta$, the conditions and assumptions on the process required for the theory are satisfied, and hence its definition is largely left implicit. As a minimum, for each $\theta \in \Theta$, the roots of $a(z) = 0$ and $b(z) = 0$ must lie outside the unit circle and condition (6) below must be satisfied. Also, since $\Theta$ is compact, it is closed and bounded. This implies, for example, that if $r_j, j = 1, \ldots, p$ are the roots of $a(z) = 0$, then there exists a $v > 0$ such that $|r_j| \geq 1 + v, j = 1, \ldots, p$. $\tau$ is the total number of parameters and $\theta_0 = (v_0':m_0')$ is the true parameter vector.

The most basic assumptions on the process are those needed for the process to be well defined and for $e_t$ and hence $y_t$ to have finite and constant second moments, i.e., $\alpha_0 > 0, \alpha_i \geq 0, i > 0, \delta_i \geq 0, i > 0, P_{ii} = 0, i > 0$, and stability of the difference equation in $E(h_i)$ implied by substituting expressions for $(y_{t-i}-\mu)^2$ and $(y_{t-i}^2-\mu^2)$, obtained from equation (1), into equation (3). From Weiss [18], this difference equation is stable when

$$\sum_{i=1}^{R} \alpha_i + \sum_{i=0}^{S} \delta_i \left( \frac{\text{var}(y_i)}{\text{var}(e_t)} - c_1 \right) - \delta_0 < 1, \quad (6)$$

where

$$c_1 = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} (\psi_i \otimes \psi_j) \text{vec} E(x_{t-i}x_{t-j})$$

and

$$\sum_{i=0}^{\infty} \psi_i z^i = a^{-1}(z)\beta'.\]
Although the moment assumptions below extend past the second, we do not obtain conditions for moments higher than the second to be finite. This would require knowledge of the evolution of moments higher than the second and since we have not specified the conditional distribution of the \( \varepsilon_t \) or equations for their higher moments, we do not have such knowledge. Similarly, we shall simply assume the invertibility of certain matrices, the strict stationarity and ergodicity of the process and the existence of higher-order moments of \( x_t \).

As noted in the introduction, the LF is based on the normal distribution. For a sample of \( T \) observations, this gives

\[
L_T(\theta) = -\frac{1}{2T} \sum_{t=1}^{T} \log h_t - \frac{1}{2T} \sum_{t=1}^{T} \varepsilon_t^2 h_t^{-1}
\]

where \( \varepsilon_t \) and \( h_t \) are now treated as functions of \( \theta \), although we note that \( \varepsilon_t \) is a function of \( m \) only. The true errors will be denoted \( \varepsilon_{0t} \) and \( h_t \) evaluated as \( \theta_0 \) is \( h_{0t} \). The vector of ML estimates, \( \hat{\theta}_T \), is that \( \theta \in \Theta \) which maximizes \( L_T(\theta) \).

In Section 3, we prove that the ML estimates are consistent and asymptotically normal, i.e.,

\[
\hat{\theta}_T \xrightarrow{p} \theta_0,
\]

and

\[
B_0^{-1/2} A_0 T^{1/2}(\hat{\theta}_T - \theta_0) \overset{d}{\sim} N(0, I),
\]

where

\[
A_0 = -E \left( \frac{\partial^2 L_T(\theta_0)}{\partial \theta \partial \theta'} \right)
\]

and

\[
B_0 = E \left( T \frac{\partial L_T(\theta_0)}{\partial \theta} \frac{\partial L_T(\theta_0)}{\partial \theta'} \right).
\]

The notation here means the derivatives are evaluated at \( \theta_0 \). Note that if the distribution of the errors is truly normal, \( A_0 = B_0 \). Also, from the asymptotic normality, we can easily show that the LS estimates under the null of no ARCH are consistent and asymptotically normal and hence that the usual LM test for heteroscedasticity is asymptotically \( \chi^2 \). Other tests associated with the ML estimation include the information matrix test derived from comparing \( A_0 \) and \( B_0 \).
Since the unconditional second moments are finite, we can derive LS estimates of the parameters in the regression equation, denoted $m_{LS}$, by minimizing

$$Q_T(m) = T^{-1} \sum_{t=1}^{T} e_t^2.$$ 

The heteroscedasticity affects the covariance matrix of these, and

$$B_{LS}^{-1/2} A_{LS} T^{1/2} (m_{LS} - m_0) \sim N(0, I)$$

where

$$A_{LS} = 2E \left( \frac{\partial e_t}{\partial m} \frac{\partial e_t}{\partial m'} \right)$$

and

$$B_{LS} = 4E \left( e_{0t}^2 \frac{\partial e_t}{\partial m} \frac{\partial e_t}{\partial m'} \right),$$

both evaluated at $m_0$. Expressions for the various derivatives are given in the mathematical appendix.

Finally, least squares estimates of the parameters in the ARCH equation, denoted $v_{LS}$, are obtained by running the artificial regression based on equation (3), i.e.,

$$\hat{\epsilon}_t^2 = \alpha_0 + \sum_{i=1}^{R} \alpha_i \hat{\epsilon}_{t-i}^2 + \delta_0 (\hat{y}_t^* - \mu_{LS})^2 + \sum_{i=1}^{S} \delta_i (y_{t-i} - \mu_{LS})^2 + w_t^t P w_t,$$  \(8\)

where the $\hat{\epsilon}_t$ are the LS residuals, $\hat{y}_t^*$ is $y_t^*$ evaluated at $m_{LS}$, and $\mu_{LS}$ is the LS estimate of $\mu$. This regression also forms the basis of the calculation of the LM test for ARCH. The complication in the regression (8) is the existence of generated regressors on the right-hand side. The influence of the generation of these does not disappear and as a result, the form of $F$, the covariance matrix of $v_{LS}$, is complex.

3. MAXIMUM LIKELIHOOD ESTIMATION

We begin by concentrating the LF with respect to $\alpha_0$. Since $\alpha_0 > 0$, this gives

$$\bar{L}_T(\theta^*) = -\frac{1}{2T} \sum \log \hat{h}_t - \frac{1}{2} \log \hat{\alpha}_0$$ \(9\)
where
\[ \hat{h}_t = h_t / x_0, \]
\[ \hat{\alpha}_0 = T^{-1} \sum \tilde{e}_t \tilde{h}_t^{-1} \]
and \( \theta^* \) contains the variance parameters excluding \( x_0 \) and divided by \( x_0 \), and \( m, \theta^*_m \) will be \( \theta^* \) evaluated at \( \theta_0 \). Also, the summations run from \( t = 1 \) to \( t = T \), and we shall continue to drop the indices if they are obvious.

Concentrating the LF not only has the advantage of reducing by one the number of parameters over which the maximization must be done, but the form (9) is convenient for studying the shape of the LF. In particular, we show in Theorem 3.1 and Corollary 3.1 that the LF is still maximized at the true parameters despite the fact that it isn't necessarily correct. The consistency (Theorem 3.2) and asymptotic normality (Theorem 3.3) of the ML estimates can then be verified. With asymptotic normality, we can then discuss the tests associated with the ML estimation.

The first formal result concerns the properties of \( E[(\partial \varepsilon_t / \partial m)(\partial \varepsilon_t / \partial m')] \). We require that this matrix is well defined and positive definite, the latter corresponding to the usual identification condition (on \( "X'X" \)) in the general linear model and ensuring that the least squares estimates exist for \( T \) large enough. We have, with all proofs in the Appendix:

**Lemma 3.1.** For all \( \theta \in \Theta \), there exists \( M < \infty \) not depending on \( \theta \) such that

\[ E \left[ \frac{\partial \varepsilon_t}{\partial m} \frac{\partial \varepsilon_t}{\partial m'} \right] < M. \]

For all \( \theta \in \Theta \)

\[ \det E \left[ \frac{\partial \varepsilon_t}{\partial m} \frac{\partial \varepsilon_t}{\partial m'} \right] > 0. \]

The equivalent result for \( h_t \) is:

**Lemma 3.2.** Assume that \( E(\varepsilon_{i0}') < \infty \). Then for all \( \theta \in \Theta \), there exists \( M_1 < \infty \) not depending on \( \theta \) such that

\[ E \left[ \frac{\partial h_t}{\partial v} \frac{\partial h_t}{\partial v'} \right] < M_1. \]

Further, for all \( \theta \in \Theta \),

\[ \det E \left[ \frac{\partial h_t}{\partial v} \frac{\partial h_t}{\partial v'} \right] > 0. \]
Expressions for \( \partial \varepsilon_i / \partial \eta \) and \( \partial h_i / \partial v \) are given in the proofs of Lemmas 3.1 and 3.2. The expression for \( \partial h_i / \partial v \) shows that the right-hand side variables in the ARCH regression (8) are just the elements of \( \partial h_i / \partial v \). Therefore Lemma 3.2 implies that the LS estimates of equation (8) exist for \( T \) large enough. Together Lemmas 3.1 and 3.2 ensure the identification of \( \theta_0 \).

**THEOREM 3.1.** For the ARCH model (1)–(5) with \( \theta \in \Theta \) and under the same conditions as Lemma 3.2, \( \lim_{T \to \infty} L_T(\theta^*) \) exists a.s. for all \( \theta \in \Theta \) and the limit \( L(\theta^*) \) is uniquely maximized at \( \theta^*_0 \).

**COROLLARY 3.1.** Under the same conditions as Theorem 3.1, \( L(\theta) = \lim_{T \to \infty} L_T(\theta) \) exists a.s. for all \( \theta \in \Theta \) and is uniquely maximized at \( \theta_0 \).

That is, in the limit, the LF is uniquely maximized at \( \theta_0 \), even though it is not necessarily correct. This in turn suggests that the parameters that maximize the LF for any \( T \) (i.e., the ML estimates \( \hat{\theta}_T \)) will converge to the parameters that maximize the LF as \( T \to \infty \) (i.e., the true parameters). We formalize this in Theorem 3.2.

Lemmas 3.1 and 3.2 also imply that minus the expected value of the matrix of second derivatives of the LF is positive definite. In particular, we have defined

\[
A_0 = -E[V^2 L_T(\theta_0)] \\
= \frac{1}{2} E(h_0^{-2} \nabla h \nabla^2 h) + E(h_0^{-1} \nabla \varepsilon \nabla^2 \varepsilon),
\]

where we have used the notation \( \nabla = \partial / \partial \theta \), \( \nabla' = \partial / \partial \theta' \), and \( \nabla^2 = \partial^2 / \partial \theta \partial \theta' \), and the derivatives are evaluated at \( \theta_0 \). The second expression for \( A_0 \) is derived in Lemma 3.3. Similarly, for any \( \theta \in \Theta \), let

\[
A = \frac{1}{2} E(h_i^{-2} \nabla h_i \nabla^2 h_i) + E(h_i^{-1} \nabla \varepsilon_i \nabla^2 \varepsilon_i).
\]

Then we have

**LEMMA 3.3.** Under the same conditions as Lemma 3.2, and \( M_2 < \infty \) not depending on \( \theta \), for all \( \theta \in \Theta \)

\[
A < M_2
\]

and

\[
det A > 0.
\]

In particular, \( A_0 \) is positive definite. These results provide the basis for the consistency theorem.
THEOREM 3.2 (consistency). Under the same conditions as Lemma 3.2, plus \( \theta_0 \) interior to \( \Theta \), the ML estimate \( \hat{\theta}_T \) is consistent for \( \theta_0 \).

The condition that \( \theta_0 \) is interior to \( \Theta \) ensures that, for \( T \) large enough, the first derivatives of \( L_T(\theta) \) are "well-behaved" at \( \theta_0 \). The following corollary implies that consistent estimates for all the terms in the stability condition (6) are available. Let \( \hat{\epsilon}_t \) be the residuals evaluated at the ML estimates, let the estimate of \( \sigma^2 \) be

\[
\hat{\sigma}^2 = \frac{1}{T} \sum_{i=0}^{T-1} (\hat{\psi}_i \otimes \hat{\psi}_j) \text{vec} \left( \frac{1}{T} \sum_{i=1}^{T} (x_{i-j} x'_{i-j}) \right),
\]

where \( \hat{\psi}_i \) is the obvious estimate of \( \psi_i \) and \( \gamma < 1 - \ln 2/\ln T \), and let \( \hat{\mu} = T^{-1} \sum y_t \).

COROLLARY 3.2. Under the same conditions as Theorem 3.2,

\[
\frac{1}{T} \sum (y_t - \hat{\mu})^2 \xrightarrow{p} \text{var}(y_t)
\]

\[
\frac{1}{T} \sum \hat{\epsilon}_t^2 \xrightarrow{p} \text{E}(\epsilon^2)
\]

and

\[
\hat{c}_1 \xrightarrow{p} c_1.
\]

Hence, following estimation, we can check whether this fundamental condition on the process is satisfied.

As an example of the restrictions on \( \Theta \) implied by the fourth moment condition \( \text{E}(\epsilon^4) < \infty \), consider the Engle ARCH process

\[
h_{0t} = \alpha_0 + \alpha_1 \epsilon^2_{0t-1}.
\]

(10)

\( \alpha_1 < 1 \) is sufficient for \( \text{E}(h_{0t}) < \infty \), but for \( \text{E}(h^2_{0t}) \) and hence \( \text{E}(\epsilon^4_{0t}) < \infty \), we must have \( \alpha_1^2 < \frac{1}{3} \), or \( \alpha_1 < 0.58 \), if the conditional distribution of \( \epsilon_{0t} \) is normal. Alternatively, if \( \epsilon_{0t} h_{0t}^{-1/2} \) has (say) a \( t_5 \) distribution, standardized such that \( \text{E}(\epsilon^2_{0t} h_{0t}^{-1}) = 1 \), then for \( \text{E}(\epsilon^4_{0t}) < \infty \), the condition is \( \alpha_1 < \frac{3}{5} \). For other examples in the Engle ARCH model, see Milhoj [13].

Next, consider the asymptotic distribution of the ML estimates. For this, we have defined \( A_0 \), and \( B_0 = E[T^2 \text{VL}_T(\theta_0) \text{VL}_T(\theta_0)] \). The expression for \( \text{VL}_T(\theta) \) is given in the proof of Theorem 3.2. Both \( A_0 \) and \( B_0 \) appear in the covariance matrix of the asymptotic distribution of the ML estimates and hence are required to be invertible.
THEOREM 3.3 (asymptotic normality). Under the same conditions as Theorem 3.2, plus det $B_0 > 0$,

$$B_0^{-1/2}A_0 T^{1/2}(\hat{\theta}_T - \theta_0) \overset{d}{\approx} N(0, I).$$

Further, consistent estimates of $A_0$ and $B_0$ are given by

$$\hat{A}_T = \frac{1}{2T} \sum h_t^{-2} \nabla h_t \nabla' h_t + \frac{1}{T} \sum h_t^{-1} \nabla e_t \nabla' e_t$$

and

$$\hat{B}_T = \frac{1}{T} \sum \nabla l_t \nabla' l_t,$$

respectively, where $1_t = -\frac{1}{2} \log h_t - \frac{1}{2} e_t^2 h_t^{-1}$ and the various terms are evaluated at $\hat{\theta}_T$.

The asymptotic covariance matrix of $\hat{\theta}_T$ takes the form $A_0^{-1}B_0 A_0^{-1}$ because the LF may not be correct. When the conditional distribution of $\varepsilon_{0_t}$ is truly normal, $A_0 = B_0$ and var $(\hat{\theta}_T) = A_0^{-1}$, the usual form. Theorem 3.3, and White [21], also give the Wald and LM tests as asymptotically $\chi^2$, with $\hat{A}_T^{-1} \hat{B}_T \hat{A}_T^{-1}$ the consistent estimate of the covariance matrix of the ML estimates of the parameters. Under the null hypothesis, these tests are asymptotically equivalent to each other, and, under normality of the $\varepsilon_{0_t}$, are also equivalent to the likelihood ratio (LR) test. As White [21] notes, the LR test is not $\chi^2$ without normality.

The LM test we are primarily concerned with is the well-known LM test for ARCH. The null is of course $h_t = \alpha_0$ for all $t$, i.e., the familiar ARMAX model. In this case, ML estimation reduces to LS and with the appropriate changes to the LF, Theorem 3.2 gives the consistency of the estimates, and from Theorem 3.3,

$$C_0^{1/2} T^{1/2} (m_{LS} - m_0) \overset{d}{\approx} N(0, I_{\tau_1}),$$

where $m_{LS}$ contains the LS estimates of $m_0$, $C_0 = \alpha_0^{-1} E[(\partial \varepsilon_t/\partial m)(\partial \varepsilon_t/\partial m')]$, where the derivatives are evaluated at $m_0$, and $\tau_1 = 1 + p + q + k$. $C_0$ is invertible and is consistently estimated by

$$\hat{C}_0 = \hat{\sigma}_e^{-2} T^{-1} \sum \frac{\partial \hat{\varepsilon}_t}{\partial m} \frac{\partial \hat{\varepsilon}_t}{\partial m'},$$

where the $\hat{\varepsilon}_t$ are the LS residuals, $\hat{\sigma}_e^2 = T^{-1} \sum \hat{\varepsilon}_t^2$ is the consistent estimate of $\alpha_0$ and the derivatives are evaluated at $m_{LS}$. 
Since $B_0$ is block-diagonal under the null, the LM test has the form 
\[
\xi_{LM} \sim \chi^2_{\tau_2}
\]
where $\tau_2 = 1 + R + S + ku$ and

\[
\xi_{LM} = T \frac{\partial L_T(\theta)}{\partial v'} B_{vv}^{-1} \frac{\partial L_T(\theta)}{\partial v}.
\]

$B_{vv}$ is the block of $B(\theta)$ corresponding to $v$ and $\theta$ is the vector of parameters under the null. $\xi_{LM}$ is equivalent to

\[
\xi_{LM} = \frac{Tf_0'(z'z)^{-1}z'f_0}{f_0'f_0},
\]

where $f_0$ is the $(T \times 1)$ vector with $t$th element $(\hat{\varepsilon}_t^2/\hat{\sigma}_t^2 - 1)$, $z$ is the $(T \times (\tau_2 + 1))$ matrix with $t$th row $\hat{\varepsilon}_t^2(1\hat{\varepsilon}_{t-1}^2 \ldots \hat{\varepsilon}_{t-R}^2(y_{t-1} - \mu_{LS})^2(y_{t-S} - \mu_{LS})^2 \ldots (y_{t-1} - \mu_{LS})^2 w_i)$ and the various terms are evaluated under the null, i.e., at $\varepsilon_i = 0$, $i = 1, \ldots, R$, $\delta_i = 0$, $i = 0, \ldots, S$, $P = 0$ and $m = m_{LS}$. A simpler estimate is $TR^2$ from the artificial regression (8).

Notice also that the ARCH null implies $\theta_0$ is on the boundary of the parameter space. Thus neither the Wald or (under normality) the likelihood ratio test is necessarily asymptotically $\chi^2$. As in Gouriéroux, Holly, and Monfort [8], we may expect their distribution to be more concentrated towards the origin than a $\chi^2$.

The other main test associated with the ML estimation is the information matrix test (White [21]). This is based on the fact that under the null hypothesis of both normality of the conditional distribution of $\varepsilon_0t$ and the correct specification of the model, $A_0 = B_0$. Hence a test for this null hypothesis may be based upon the difference between $\hat{A}_T$ and $\hat{B}_T$. Define $T(\tau + 1)/2$ vectors $d_k(\theta)$ and $D_T(\theta)$ such that $d_k(\theta)$ has $k$th element

\[
d_{ik}(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} + \frac{\partial l(\theta)}{\partial \theta_i} \frac{\partial l(\theta)}{\partial \theta_j}
\]

$k = 1, \ldots, \frac{1}{2}(\tau + 1)$

$i, j = 1, \ldots, \tau$

$i \leq j$

and

\[
D_T(\theta) = \frac{1}{T} \sum d_i(\theta);
\]

and let

\[
V(\theta) = E[d(\theta)d(\theta)']
\]

Then under the various assumptions we have made, plus that $V(\theta)$ is invertible and for cases when $\varepsilon_{0t}$ is not normal, $E(\varepsilon_t^8h_t^{-4}|I_{t-1}) < \infty$ for all
\(\theta \in \Theta\), the conditions for White [21] Theorem 4.1 are satisfied and it follows that

\[TD_T(\hat{\theta}_T)V_T(\hat{\theta}_T)^{-1}D_T(\hat{\theta}_T) \sim \chi_{p+1}^2,\]  

where \(V_T(\hat{\theta}_T)\) is the obvious estimate of \(V(\theta)\). Alternatively, we may use elements of \(A_T\) and \(\hat{B}_T\) in place of the derivatives of the LF or test only some subset of the entries of \(d_t(\theta)\), in both cases giving results analogous to equation (11). When the test does not reject the null, \(A_0^{-1}\) may be used as the variance covariance matrix of the ML estimates. Further evidence for normality may be obtained by applying normality tests such as those in Lomnicki [12] to the standardized residuals \(e_t h_t^{-1/2}\). Similarly, alternative specification tests include the Hausman tests comparing the ML estimates to the LS estimates considered in Section 4.

The ML estimates are found in practice using a technique such as scoring. At iteration \(i + 1\),

\[\theta^{i+1} = \theta^i + \lambda_i \hat{A}_T^{-1}VL_T(\theta^i),\]  

where \(\hat{A}_T\) and \(VL_T(\theta)\) are evaluated at \(\theta^i\), the value of \(\theta\) at the \(i\)th iteration, and \(\lambda_i\) is a step length. Note that the algorithm only requires first derivatives and following Dhrymes and Taylor [3], one step of equation (12) from consistent starting values (such as the LS estimates discussed in Section 4) also gives estimates with the same asymptotic distribution as the ML estimates. An alternative to scoring is to replace \(\hat{A}_T\) by \(\hat{B}_T\), giving the Berndt, Hall, Hall, and Hausman [2] algorithm. If the LF is not correct, however, then the estimates from one step of this no longer have the same asymptotic distribution as the ML estimates.

Finally, in this section, we note that this and the ensuing theory is also valid, under the relevant conditions, if certain of the assumptions are relaxed. First, as an alternative to the strict stationarity, we may assume the process is mixing (White and Domowitz [22]). Second, provided the various moment conditions on the model and the LF and its derivatives can be satisfied, the linearity assumption can be dropped. For a general nonlinear equation, i.e., \(y_t = f(x_t, \theta) + \epsilon_t\), moment assumptions analogous to those in White and Domowitz [22] would be sufficient. In specific models where the form of \(f\) is given, sufficient conditions for stability and the existence of moments may be more difficult to obtain than in the essentially linear model we are considering.

4. LEAST SQUARES ESTIMATION

In practice, the ARCH model would be constructed by first estimating the regression equation using LS and then testing for and estimating the ARCH equation. Finally, the combined model would be estimated via ML and tests
based on this performed. For pure ARMA models, this model-building process is discussed in Weiss [19]. At the first stage, we cannot distinguish between LS under the null of no ARCH or the alternative. Hence we use the same notation in both cases, i.e., \( m_{LS} \) is the LS estimate, \( \hat{\epsilon}_t \) and \( \epsilon_{o_t} \) are the residuals at \( m_{LS} \) and \( m_0 \), respectively, and \( \epsilon_t \) refers to the residuals at any point \( m \).

We first give the equivalent of Corollary 3.1 and Theorem 3.2 describing the shape of the criterion function \( Q_T(m) = T^{-1}\sum_{i=1}^{T} \hat{\epsilon}_t^2 \) as \( T \to \infty \), and the consistency theorem for the LS estimates.

**THEOREM 4.1.** Under the same conditions as Lemma 3.1, \( \lim_{T \to \infty} T^{-1}\sum_{i=1}^{T} \hat{\epsilon}_t^2 \) exists a.s. for all \( \theta \in \Theta \) and the limit, \( E(\epsilon_t^2) \), is uniquely minimized at \( m_0 \).

The proof of Theorem 4.1 is obvious from that of Theorem 3.1 and so is omitted. Similarly, the proofs of the consistency and asymptotic normality results, Theorems 4.2 and 4.3, respectively, are now standard and are omitted. See, for example, Rissanen and Caines [16] or Nicholls and Quinn [14].

**THEOREM 4.2 (consistency).** Under the same conditions as Theorem 4.1 plus \( \theta_0 \) interior to \( \Theta \), \( m_{LS} \) is consistent for \( m_0 \). Furthermore, \( \hat{\sigma}_t^2 = T^{-1}\sum_{i=1}^{T} \hat{\epsilon}_t^2 \) is consistent for \( E(\epsilon_{o_t}^2) \).

As noted in Section 2, the covariance matrix of \( m_{LS} \) depends on two matrices \( A_{LS} \) and \( B_{LS} \). The existence of \( A_{LS} \) was proved in Lemma 3.1, while for \( B_{LS} \) we note that

\[
B_{LS,ij} = 4E \left( \epsilon_{o_t}^2 \frac{\partial \epsilon_t}{\partial m_i} \frac{\partial \epsilon_t}{\partial m_j} \right)
\leq 4[E(\epsilon_{o_t}^2)]^{1/2} \left[ E \left( \left( \frac{\partial \epsilon_t}{\partial m_i} \right)^2 \left( \frac{\partial \epsilon_t}{\partial m_j} \right)^2 \right) \right]^{1/2}
\]

(by the Cauchy–Schwartz inequality), where \( B_{LS,ij} \) is the \( ij \)th element of \( B_{LS} \) and the derivatives are evaluated at \( \theta_0 \). The right-hand side of the inequality is finite provided \( E(\epsilon_{o_t}^2) < \infty \). As with \( A_0 \) and \( B_0 \), \( A_{LS} \) and \( B_{LS} \) must also be invertible. For \( A_{LS} \), we again appeal to Lemma 3.1 (giving the identification condition) while for \( B_{LS} \) we have

\[
B_{LS} = 4E \left[ E \left( \epsilon_{o_t}^2 \frac{\partial \epsilon_t}{\partial m} \frac{\partial \epsilon_t}{\partial m} \right) I_{t-1} \right]
= 4E \left( h_{o_t} \frac{\partial \epsilon_t}{\partial m} \frac{\partial \epsilon_t}{\partial m} \right)
\geq 4\lambda_0 E \left( \frac{\partial \epsilon_t}{\partial m} \frac{\partial \epsilon_t}{\partial m} \right) > 0.
\]
THEOREM 4.3 (asymptotic normality). Under the same conditions as Theorem 4.2, plus $E(e_{0i}^4) < \infty$,

$$B_{LS}^{-1/2} A_{LS} T^{1/2}(m_{LS} - m_0) \overset{d}{\sim} N(0, I_r).$$

Further, consistent estimates of $A_{LS}$ and $B_{LS}$ are given by

$$\hat{A}_{LS} = \frac{2}{T} \sum \frac{\partial \varepsilon_t}{\partial m} \frac{\partial \varepsilon_t}{\partial m'}$$

and

$$\hat{B}_{LS} = \frac{4}{T} \sum \varepsilon_t^2 \frac{\partial \varepsilon_t}{\partial m} \frac{\partial \varepsilon_t}{\partial m'}$$

respectively, where the derivatives are evaluated at $m_{LS}$.

Theorems 4.2 and 4.3 also correspond to the consistency and asymptotic normality theorems in White and Domowitz [22], Theorems 3.1 and 3.2, respectively, with $\hat{C}_{LS} = \hat{A}_{LS}^{-1} \hat{B}_{LS} A_{LS}^{-1}$ the heteroscedasticity consistent estimate of the covariance matrix of the LS estimates. Note that this matrix is different from the covariance matrix under the null of no ARCH and hence provides a possible method, other than the LM test, of distinguishing between the models under the null and alternative. Such a general heteroscedasticity test was suggested by White and Domowitz [22], although presumably the LM test is asymptotically more powerful against the ARCH alternative. As White and Domowitz also note, in the linear model ($m(t) = 1$) $\partial \varepsilon_t / \partial m$ is just “$X_t$,” and

$$\hat{C}_{LS} = \left( \frac{X'X}{T} \right)^{-1} \left( \frac{X'\hat{\Sigma}X}{T} \right) \left( \frac{X'X}{T} \right)^{-1},$$

where, with $\hat{\Sigma}_{ij}$ as the $ij$th element of $\hat{\Sigma}$, $\hat{\Sigma}_{ii} = \hat{\varepsilon}_i^2$ and $\hat{\Sigma}_{ij} = 0$.

The consistency and asymptotic normality theorems also imply that if, in the specification of the model, a Type II error is made with respect to ARCH, i.e., it is wrongly believed ARCH is not present, the LS estimates will still be consistent and asymptotic normal although the implicit use of $C_0$ in the asymptotic distribution implies their covariance matrix will be incorrectly estimated. Any subsequent tests will also have incorrect size asymptotically.

Least squares estimates of the ARCH equation may be obtained by running the ARCH artificial regression, equation (8). Define $z_t$ as $\partial h_t / \partial \varepsilon$ evaluated at $m_{LS}$, i.e., $\hat{\varepsilon}_t^2$ times the $t$th row of $z$ defined in the previous section, and $\hat{z}_t$ as the row vector formed from $z_t$ by replacing $\hat{\varepsilon}_t$ by $\varepsilon_{0t}$, $\mu_{LS}$ by $\mu_0$, and $\hat{\gamma}_t$ by
ASYMPTOTIC THEORY FOR ARCH MODELS

Also define $\eta_t = \delta^2_t - z_tv$ and $\eta_t = \epsilon^2_0 - \tilde{z}_tv_0$. The LS estimate is

$$v_{LS} = (\sum z_t z_t)^{-1}(\sum z_t^2).$$

With $\eta_{vt} = \delta^2_t - z_tv$, i.e., $\eta_t$ evaluated at $v_0$, we have

$$(v_{LS} - v_0) = (T^{-1}\sum z_t z_t)^{-1}T^{-1}\sum z_t^2$$

and

$$T^{1/2}(v_{LS} - v_0) = (T^{-1}\sum z_t z_t)^{-1}T^{-1/2}\sum z_t^2.$$

Hence the asymptotic distribution of $v_{LS}$ depends upon that of $T^{-1/2}\sum z_t^2$. A mean value expansion of $T^{-1/2}\sum z_t^2$ about $m_0$ gives

$$T^{-1/2}\sum z_t^2 = T^{-1/2}\sum z_t^2 + T^{-1}\sum \left( \frac{\partial(z_t^2)}{\partial m} \right)_{m_0} T^{1/2}(m_{LS} - m_0),$$

where $m_t$ lies between $m_{LS}$ and $m_0$. Now

$$T^{-1}\sum \frac{\partial(z_t^2)}{\partial m} = T^{-1}\sum \frac{\partial z_t^2}{\partial m} \eta_t + 2T^{-1}\sum \delta_t \frac{\partial \epsilon_t}{\partial m} z_t - T^{-1}\sum \frac{\partial z_t^2}{\partial m} v_0 z_t,$$

where the various terms are evaluated at $m_t$, and while the first two terms on the right-hand side converge to zero as $m_{LS} \to m_0$, in general the third does not. These arguments are formalized in Theorem 4.4, but it can be seen by substituting equation (15) into equation (14) and letting $T \to \infty$, that this leaves terms involving $m_{LS}$ in the asymptotic distribution of $v_{LS}$ and thus represents a contrast from similar regressions in, for example, Nicholls and Quinn [14], Pagan, Hall, and Trivedi [15], and White and Domowitz [22]. The second term converging to zero suggests that the regression is not asymptotically affected by using $\delta_t^2$ rather than $\epsilon_0^2$ as the dependent variable. Similarly, in special cases of the artificial regression where $z_t$ contains only lagged $y_t^2$ (Nicholls and Quinn [14] and Pagan, Hall, and Trivedi [15]) or exogenous variables (Pagan, Hall, and Trivedi [15]), $\partial z_t^2/\partial m = 0$ and the influence of $m_{LS}$ is not present. White and Domowitz [22] use a regression of this type in their test for unconditional heteroscedasticity which has a null hypothesis of no heteroscedasticity and explicitly excludes ARCH. Under this null, all but the first element of $v_0$ is zero and hence, since the first column of $\partial z_t^2/\partial m$ is also zero, $(\partial z_t^2/\partial m)v_0 = 0$.

To formalize the properties of $v_{LS}$, we begin by noting that the identification condition, $\det E(z_t^2) > 0$, is just Lemma 3.2. Next, we define the matrix
ANDREW A. WEISS

$F$, which in Theorem 4.4 is shown to be the covariance matrix of the asymptotic distribution of $v_{1S}$.

$$F = E(\bar{z}_t^2) -1\left[ E(\bar{z}_t^2) + E\left( \frac{\partial z'_t}{\partial m} v_0 \bar{z}_t \right) A_{LS}^{-1} B_{LS} A_{LS}^{-1} E\left( \frac{\partial z'_t}{\partial m} v_0 \bar{z}_t \right) 
+ 2E\left( \frac{\partial z'_t}{\partial m} e_{0t} \right) A_{LS}^{-1} E\left( \frac{\partial z'_t}{\partial m} v_0 \bar{z}_t \right) 
+ 2E\left( \frac{\partial z'_t}{\partial m} v_0 \bar{z}_t \right) A_{LS}^{-1} E\left( \frac{\partial e_t}{\partial m} e_{0t} \bar{z}_t \right) \right] E(\bar{z}_t^2) -1,$$

(16)

where the derivatives are evaluated at $m_0$. Also, from equation (14), since the asymptotic distribution of $T^{-1/2}\sum \bar{z}_t \bar{z}_t$ enters into that of $T^{1/2}(v_{LS} - v_0)$, we require that its covariance matrix, $E(\bar{z}_t^2 \bar{z}_t)$, be invertible. But det $E(\bar{z}_t^2 \bar{z}_t)$ = 0 implies there exists a $\lambda \neq 0$ such that $E(\lambda \bar{z}_t^2 \bar{z}_t) = 0$, or $\lambda \bar{z}_t^2 = 0$ a.s. This is not possible since $\bar{z}_t \neq 0$ a.s. and $\lambda \neq 0$ a.s., the latter holding because $E(\bar{z}_t^2 \bar{z}_t)$ is positive definite.

**THEOREM 4.4 (consistency and asymptotic normality).** For the model given by (1)–(5), under the same conditions as Theorem 4.3, the LS estimates of the ARCH parameters are consistent. Further, if

(i) $E(e_{0t}^8) < \infty$,

and

(ii) det $F > 0$,

in

$$F^{-1/2}T^{1/2}(v_{LS} - v_0) \overset{d}{\sim} N(0, I_{t+1}).$$

A consistent estimate of $F$ is given by replacing the expectations in equation (16) by their sample means evaluated at $v_{LS}$.

The eighth moment condition is used as a sufficient condition for the existence of $F$ and is typical of the moment conditions needed for this type of regression. Equivalent conditions may be found in Nicholls and Quinn [14], White and Domowitz [22], and Weiss [17]. If the LS estimates are to be used only as initial estimates in the scoring algorithm to obtain the ML estimates, then only fourth moments are required. Presumably the small sample properties of the estimate of $F$ may also be poor, especially in the presence of outliers, so that paying undue attention to the standard errors derived from the estimate of $F$ may be unwise. In any case, in the simple Engle ARCH process (10), we now require $\varkappa_1 < \frac{1}{105}$, or $\varkappa_1 < 0.32$ (assuming normality). Note also that if $E(e_{0t}^8| I_{t-1}) = 0$, then $E(\bar{z}_t^2 v_0 \bar{z}_t) | I_{t-1} = 0$. Further, if $\gamma^*$ is not included in the variance equation, then $E[(\partial \bar{z}_t' / \partial m) v_0 \bar{z}_t] = 0$, and $F$ reduces to $E(\bar{z}_t^2 \bar{z}_t)^{-1} E(\bar{z}_t^2 \bar{z}_t) E(\bar{z}_t^2 \bar{z}_t)^{-1}$. Condition (ii) in the Theorem 4.4 is then not needed.
5. CONCLUDING COMMENTS

In this paper we have considered the asymptotic properties of the estimates of the ARCH model, and noted a number of hypothesis tests. A variety of assumptions have been made to facilitate the analysis and these also, of course, give the obvious avenues for further research. Many have concerned the existence of higher (unconditional) moments, but because of the relative complexity of the model, sufficient conditions for these to hold have not been derived. However, given our results and (say) equations for higher conditional moments, the general approach is straightforward.

The various hypothesis tests suggest an important research direction. That is, analyzing the results from a large number of tests, presumably all having different power against specific alternatives both within and outside the class of ARCH models. Other important areas include removing the linearity assumption and investigation of the small sample properties of the estimators and tests. Finally, the multivariate case has also received little attention.

NOTES

1. Different specifications for $x_t$ are possible under the alternative (mixing) assumptions noted at the end of Section 3. However, for constant unconditional variance some modification of the ARCH equation (3) may then be necessary.

2. We are implicitly conditioning on the pre-sample values of $y_t$ and setting pre-sample $e_t$ to zero. As in Godfrey [7], this will not affect the asymptotic results.

3. A more detailed version of the appendix is available in Weiss [20]. The theorems there give strong, rather than weak, consistency, and in many cases also utilize weaker assumptions. This is at the expense of more complex proofs.

REFERENCES


APPENDIX

Proof of Lemma 3.1. We begin by giving expressions for $\partial e_i / \partial m$. From equation (1),

$$b(B) \frac{\partial e_i}{\partial \mu} = -a(1),$$

$$b(B) \frac{\partial e_i}{\partial a_i} = -y_{t-i},$$

$$b(B) \frac{\partial e_i}{\partial b_i} = -e_{t-i},$$

and

$$b(B) \frac{\partial e_i}{\partial \beta_i} = -x_{ti},$$

where $\beta_i$ and $x_{ti}$ are the $i$th elements of $\beta$ and $x_t$, respectively. Next, for constant vectors $\lambda$,

$$W'_i = (-a(1) - y_{t-1} \cdots - y_{t-p} - e_{t-1} \cdots - e_{t-q} - x_t).$$

That is, $\lambda' \frac{\partial e_i}{\partial m}$ is given by a difference equation that has input with finite second moments and is stable due to the invertibility condition on $b(B)$. Hence $E(\lambda' \frac{\partial e_i}{\partial m} \frac{\partial e_i}{\partial m'}) < \infty$. Further, because $\Theta$ is compact, $h(z)$ is invertible and the roots of a polynomial are continuous functions of its parameters, there exists $\rho_1 < 1$ and $K_1, K_2 > 0$, independent of $\theta$, such that

$$|e_t| \leq K_1 \sum_{i=0}^\infty \rho_1^i |y_{t-i}| + K_2 \sum_{i=0}^\infty \sum_{j=1}^K \rho_1^j |x_{t-i,j}|.$$ 

Hence there exists a constant $M_1 < \infty$ such that $E(e_t^2) < M_1$ for all $\theta \in \Theta$. Applying the same method to $\partial e_i / \partial m$ implies $E(\partial e_i / \partial m \; \partial e_i / \partial m') < M < \infty$ for all $\theta \in \Theta$.

Next, $\det E(\partial e_i / \partial m \; \partial e_i / \partial m') > 0$ is equivalent to $E(\lambda' \frac{\partial e_i}{\partial m} \frac{\partial e_i}{\partial m'} \lambda) > 0$ for all $\lambda \neq 0$. We proceed by contradiction. Assume there exists $\lambda \neq 0$ such that $E(\lambda' \frac{\partial e_i}{\partial m} \frac{\partial e_i}{\partial m'} \lambda) = 0$. Then $\lambda' \frac{\partial e_i}{\partial m} = 0$ a.s. for all $t$, which, writing

$$\lambda' \frac{\partial e_i}{\partial m} = -\sum_{i=1}^q b_i \lambda' \frac{\partial e_{t-i}}{\partial m} + \lambda' W_t,$$

implies $\lambda' W_t = 0$ a.s. for all $t$. But this, together with

$$y_t = e_{ot} + \hat{y}_t,$$
and (using a mean value expansion)

\[ \epsilon_t = \epsilon_{0t} + \frac{\partial \epsilon_t}{\partial m'} (m - m_0), \]

(A.1)

where the derivative is evaluated at \( m' \), which lies between \( m \) and \( m_0 \), implies \( \epsilon_{0t-1} \) is a function of \( I_{t-2} \) and \( x_t \). But then

\[ E(\epsilon_{0t-1} \mid I_{t-2}, x_t) = \epsilon_{0t-1} \]

while from equations (2) and (5),

\[ E(\epsilon_{0t-1} \mid I_{t-2}, x_t) = 0. \]

This implies \( \epsilon_{0t-1} = 0 \) a.s. for all \( t \), which contradicts the fact that \( E(\epsilon_{0t}^2) = \sigma_t^2 > 0 \). Therefore, no such \( \lambda \) exists and \( E(\lambda' \partial e_t/\partial m \partial e_t/\partial m' \lambda) > 0 \) a.s. as required. Since 

\[ E(\epsilon_{0t} \mid x_{t+1}) = 0, i > 0, \]

we may use this approach unless \( \lambda_i = 0, i = 2, \ldots, p + q + 1 \), where \( \lambda' = (\lambda_1, \ldots, \lambda_{p+q+1}, \lambda^\ast) \). In that case, we have \( \lambda^\ast(1: x'_t) = 0 \) a.s. for all \( t \), which is not allowed.

Proof of Lemma 3.2. Differentiating equation (3) and using \( \partial e_t/\partial y = 0 \) gives

\[ \frac{\partial h_t}{\partial \pi_0} = 1, \quad \frac{\partial h_t}{\partial \pi_i} = \epsilon_i^2 - i, \quad \frac{\partial h_t}{\partial \pi_0} = (y_t^\ast - \mu)^2, \]

\[ \frac{\partial h_t}{\partial \beta_{ij}} = (y_t - \mu)^2 \quad \text{and} \quad \frac{\partial h_t}{\partial P_{ii}} = w_{ii}^2, \]

where \( w_{ii} \) is the \( i \)th element of \( w_t \). Since \( E(\epsilon_{0t}^2) < \infty \) (and \( E(w_{ii}^2) < \infty \)), \( E(y_t - \mu)^4 < \infty \) and \( E(\epsilon_t^4) < M_2 < \infty \) for all \( \theta \in \Theta \). Thus the first part of the lemma is obvious.

Next, applying the method of the second part of Lemma 3.1 to \( \lambda' \partial h_t/\partial y = 0 \) and using

\[ \epsilon_t = \epsilon_{0t} + \frac{\partial \epsilon_t}{\partial \theta} (\theta - \theta_0), \]

\[ y_t = \mu + \sum_{i=0}^\infty \pi_i x_{t-i} + \sum_{i=0}^\infty \psi_i x_{t-i} + (\mu_0 - \mu), \]

and

\[ y_t^\ast = \mu + \sum_{i=0}^\infty \mu_i x_{t-i} + \sum_{i=0}^\infty v_i x_{t-i} + (\mu_0 - \mu), \]

for suitable constants \( \pi_i, \psi_i, \mu_i, \) and \( v_i \), implies a quadratic function in \( \epsilon_{0t} \), which may be solved for \( \epsilon_{0t} = f_1(t) \) or \( \epsilon_{0t} = f_2(t) \) a.s. But \( f_1(t) \) and \( f_2(t) \), the two solutions, are functions of \( I_{t-1} \) and \( x_{t+1} \), and \( \epsilon_{0t} \) having only two values conditional on \( I_{t-1} \) and \( x_{t+1} \) was excluded. Therefore, no such \( \lambda \) exists and \( \det E(\partial h_t/\partial y \partial h_t/\partial y') > 0 \).
Unless \( \lambda_i = 0, i = 2, \ldots, R + S + 2 \), where \( \lambda' = (\lambda_1 \ldots \lambda_{R+S+2}) \), we use this approach. If these \( \lambda_i \) are all zero, we write

\[
w'_t D w_t = -\lambda_1 \quad \text{a.s. for all } t,
\]

where \( D \) is the diagonal matrix with the elements of \( \lambda^* \) on its diagonal, which was also excluded by assuming that \( x_0 \) and the elements of \( w_t \) are not linearly dependent.

\[
w'_t D w_t = -\lambda_1 \quad \text{a.s. for all } t.
\]

Proof of Theorem 3.1. From the ergodic theorem (see, for example, Hannan [11], p. 201), for any \( \theta \in \Theta \),

\[
\tilde{L}(\theta^*) = \lim_{T \to \infty} \tilde{L}_T(\theta^*) = -\frac{1}{2} E(\log \hat{h}_0) - \frac{1}{2} \log E(\varepsilon_t^2 \hat{h}_t^{-1}) \quad \text{a.s.}
\]

if the expectations exist. Now, since \( \log E(X) \geq E \log(X) \) for all positive random variables \( X \) with equality only when \( X \) is a constant a.s., \( \log E(h_t) \geq E(\log h_t) \). But \( E(h_t) < \infty \), so \( E(\log h_t) < \infty \). Also, from Lemma 3.1, \( E(\varepsilon_t^2) < \infty \) for all \( \theta \in \Theta \). Therefore, since \( E(e_t^2 \hat{h}_t^{-1}) \leq E(e_t^2), E(e_t^2 \hat{h}_t^{-1}) < \infty \).

Next,

\[
E(e_t^2 \hat{h}_t^{-1}) = E(\hat{h}_t^{-1}(e_t + \varepsilon_{0t} - \varepsilon_{00})^2)
\]

\[
= E(\hat{h}_t^{-1} \varepsilon_{0t}^2) + E(\hat{h}_t^{-1}(e_t - \varepsilon_{00})^2),
\]

since \( \hat{h}_t^{-1}(e_t - \varepsilon_{00}) \) depends only on \( I_{t-1} \) and \( E(\varepsilon_{0t} | I_{t-1}) = 0 \). Thus \( E(e_t^2 \hat{h}_t^{-1}) \geq E(e_{00}^2 \hat{h}_t^{-1}) \) with equality when \( e_t = \varepsilon_{0t} \) for all \( t \) a.s. But from equation (A.1) and Lemma 3.1,

\[
E(\varepsilon_t^2) = E(\varepsilon_{00}^2) + (m - m_0) E \left( \frac{\partial \varepsilon_t}{\partial m} \frac{\partial \varepsilon_t}{\partial m} \right) (m - m_0)
\]

\[
\geq E(\varepsilon_{00}^2)
\]

with equality only at \( m = m_0 \). Hence \( e_t = \varepsilon_{0t} \) for all \( t \) a.s. only when \( m = m_0 \). Next, with \( x_{00} \), the true value of \( x_0 \),

\[
\tilde{L}(\theta^*) = -\frac{1}{2} E(\log \hat{h}_0) - \frac{1}{2} \log E(\varepsilon_t^2 \hat{h}_t^{-1})
\]

\[
= -\frac{1}{2} \log x_0 - \frac{1}{2} E(\log \hat{h}_0) - \frac{1}{2} E(\varepsilon_t^2 \hat{h}_t^{-1})
\]

with equality only at \( \theta^* = \theta_0^* \). The second inequality follows because \( \log E(e_{00}^2 \hat{h}_t^{-1}) = E(\varepsilon_{00}^2) + E(\hat{h}_0 \hat{h}_t^{-1}) \) and \( E(\hat{h}_0 \hat{h}_t^{-1}) \geq E(\hat{h}_0 \hat{h}_t^{-1}) \) with equality only when \( \hat{h}_0 = \hat{h}_t \). But because of Lemma 3.2, this only occurs at \( \theta^* = \theta_0^* \).

Proof of Corollary 3.1.

\[
\lim_{T \to \infty} L_T(\theta) = -\frac{1}{2} E(\log h_t) - \frac{1}{2} E(e_t^2 h_t^{-1})
\]

\[
= \frac{1}{2} \log x_0 - \frac{1}{2} E(\log \hat{h}_0) - \frac{1}{2} E(\varepsilon_t^2 \hat{h}_t^{-1}),
\]
which exists by Theorem 3.1. Maximizing with respect to $\alpha_0$ gives

$$\hat{\alpha}_0 = E(\varepsilon_i^2 \hat{h}_i^{-1}).$$

and by Theorem 3.1, the concentrated LF, $-\frac{1}{2} \log \alpha_0 - \frac{1}{2} E(\log \hat{h}_i)$ is maximized at $\theta_0^*$. The value of $\alpha_0$ which maximizes the LF is thus

$$\hat{\alpha}_0 = E(v_{0i}^2 \hat{h}_{0i}^{-1}) = \alpha_{00},$$

where $\hat{h}_{0i} = \hat{h}_i(\theta_0^*)$.

Proof of Lemma 3.3. Differentiating equation (7) gives

$$V_{i}^{2}L_{r}(\theta) = T^{-1} \sum V_{i}^{2}l(\theta)$$

where $V_{ij} \equiv \partial^{2}/\partial \theta_{i} \partial \theta_{j}$, $\theta_{i}$ is the $i$th element of $\theta$, $l_{t} = - \frac{1}{2} \log h_{t} - \frac{1}{2} \varepsilon_{t}^{2} h_{t}^{-1}$ (i.e., $L_{r}(\theta) = T^{-1} \sum l(\theta)$), and

$$V_{i}^{2}l_{t}(\theta) = - h_{t}^{-2} \nabla h_{i} \nabla h_{t} \varepsilon_{t}^{2} h_{t}^{-1} + h_{t}^{-2} \nabla h_{i} \varepsilon_{t} \nabla h_{t} \varepsilon_{t} - \nabla \varepsilon_{t} \nabla h_{t} h_{t}^{-1} - \varepsilon_{t} \nabla h_{t}^{2} h_{i}^{-1}$$

$$+ h_{t}^{-2} \nabla \varepsilon_{t} \nabla h_{t} + \frac{1}{2} h_{t}^{-2} \nabla h_{i} \nabla h_{t} + \frac{1}{2} h_{t}^{-2} \nabla h_{t}^{2} h_{i}^{-1} - 1),$$

(A.2)

where $V_{i} \equiv \partial / \partial \theta_{i}$. Expressions for $\nabla \varepsilon_{t}$ and $\partial h_{i} / \partial \varepsilon_{t}$ are given in Lemmas 3.1 and 3.2, respectively, while for $\partial h_{i} / \partial \varepsilon_{t}$ we have

$$\frac{\partial h_{i}}{\partial \varepsilon_{t}} = 2 \sum_{i=1}^{k} \frac{\partial \varepsilon_{t-i}}{\partial \varepsilon_{t}} + 2 \delta_{0}(y_{t}^{*} - \mu) \left( \frac{\partial y_{t}^{*}}{\partial \varepsilon_{t}} - \delta_{\mu} \right) - 2 \delta_{0}(y_{t}^{*} - \mu) \delta_{\mu},$$

where $m_{j}$ is the $j$th element of $m$ and

$$\delta_{\mu} = 1 \quad \text{if } m_{j} = \mu \quad \text{and} \quad 0 \quad \text{otherwise}.$$

Since $y_{t}^{*} = y_{t} - \varepsilon_{t}$, $\partial y_{t}^{*} / \partial m_{j} = - \partial \varepsilon / \partial m_{j}$. Clearly, the $\nabla h_{i}$ have finite, bounded, second moments. Equations for the second derivatives are also easily obtained, and we note that from the expressions for $\nabla \varepsilon_{t}$ and $\nabla \varepsilon_{t}^{2}$, these derivatives have bounded fourth moments. Similarly, since every term is $\nabla h_{i}$ and $\nabla h_{i}^{2}$, also appears in $h_{t}$ itself, the $h_{t}^{-1} \nabla h_{i}$ and $h_{t}^{-1} \nabla h_{i}^{2}$ are uniformly bounded from above. Hence, evaluating equation (A.2) at $\theta_{0}$ implies the second, fourth, fifth, and last terms are zero,

$$E(h_{0i}^{2} \nabla h_{i} \nabla h_{t} \varepsilon_{t}^{2} h_{0i}^{-1}) = E(h_{0i}^{2} \nabla h_{i} \nabla h_{t}) < \infty,$$

and

$$E(\nabla \varepsilon_{t} \nabla h_{t} h_{0i}^{-1}) \leq \alpha_{00} E(\nabla \varepsilon_{t} \nabla \varepsilon_{t}) < \infty.$$

Hence,

$$A_{0} = \frac{1}{2} E \left( h_{0i}^{2} \frac{\partial h_{i}}{\partial \theta} \frac{\partial']}{\partial \theta'} \right) + E \left( h_{0i}^{2} \frac{\partial \varepsilon_{t}}{\partial \theta} \frac{\partial \varepsilon_{t}}{\partial \theta'} \right) < \infty,$$

where the derivatives are evaluated at $\theta_{0}$. 

More generally, for constants $M, M_1 < \infty$ not depending on $\theta$,

$$E(h_t^{-2}\nabla h_t \nabla h_t \varepsilon_t^2 \varepsilon_t^{-1}) \leq ME(\varepsilon_t^2) < \infty$$

and

$$E(\nabla \varepsilon_t \nabla \varepsilon_t h_t^{-1}) \leq z_0^{-1} E(\nabla \varepsilon_t \nabla \varepsilon_t) < M_1 < \infty,$$

where $z_0$ is the lower bound for $z_0$ (recall that $\Theta$ is compact).

Next, for $\lambda \neq 0$,

$$\lambda' \Lambda \lambda = \frac{1}{2} E(\lambda' h_t^{-2} \nabla h_t \nabla h_t \lambda) + E(\lambda' h_t^{-1} \nabla \varepsilon_t \nabla \varepsilon_t \lambda).$$

Both terms are non-negative and since $\partial \varepsilon_t / \partial \varepsilon = 0$, the second term is equal to $E(\lambda' h_t^{-1} \partial \varepsilon_t / \partial \theta \partial \varepsilon_t / \partial \theta')$, where $\lambda' = (\lambda_1 : \lambda_2)$ is partitioned to conform with $\theta' = (\theta' : m')$. From Lemma 3.1, this is positive unless $\lambda_2 = 0$. In this case, $\lambda_1 \neq 0$ and the first term becomes $\frac{1}{2} E(\lambda_1 h_t^{-2} \partial h_t / \partial \theta \partial h_t / \partial \theta' \lambda_1)$, which is positive because of Lemma 3.2.

Proof of Theorem 3.2. Basawa, Feigin, and Heyde [1] have analyzed the consistency and asymptotic normality of ML estimators in processes with dependent observations and have provided a set of sufficient conditions that can be checked. Adapted to our problem, these results imply that there exists a consistent root of the equation $\partial L_T(\theta) / \partial \theta = 0$ if

(i) $T^{-1} \sum \nabla l_t(\theta_0) \overset{p}{\to} 0$.
(ii) There exists a nonrandom matrix $M(\theta_0) > 0$ such that for all $\varepsilon > 0$

$$P\{ -T^{-1} \sum \nabla^2 l_t(\theta_0) \geq M(\theta_0) \} > 1 - \varepsilon$$

for all $T > T_0(\varepsilon)$.
(iii) There exists a constant $M < \infty$ such that

$$E|\nabla^3 l_t(\theta)| < M$$

for all $\theta \in \Theta$ and where $\nabla^3_{ijk} \equiv \partial^3 / \partial \theta_i \partial \theta_j \partial \theta_k$.

Corollary 3.1 then implies that $\hat{\theta}_T \overset{p}{\to} \theta_0$. The method of the proof is to show that these three conditions are satisfied.

(i) From equation (7),

$$\nabla l_t(\theta_0) = \frac{1}{2} h_{0t}^{-1} \nabla h_t (\varepsilon_{0t}^2 h_{0t}^{-1} - 1) - \varepsilon_{0t} \nabla \varepsilon_t h_{0t}^{-1},$$

where the derivatives are evaluated at $\theta_0$. Therefore, $E[\nabla l_t(\theta_0)] = 0$ since $E(\varepsilon_{0t} | I_{t-1}) = 0$ and $E(\varepsilon_{0t}^2 | I_{t-1}) = h_{0t}$. The ergodic theorem then implies that $T^{-1} \sum \nabla l_t(\theta_0) \overset{p}{\to} 0$. 

By the ergodic theorem, for any constant vector $\lambda \neq 0$,

$$T^{-1}\sum_{i} \bar{\lambda} V_{i}^{2} l_{i}(\theta_{0}) \lambda \rightarrow E[\bar{\lambda} V_{1}^{2} l_{1}(\theta_{0}) \lambda] \quad \text{a.s.}$$

$$= -\lambda^{'} A_{0} \lambda.$$

For given $\lambda$, let $0 < \delta(\lambda) < -\frac{1}{2} \lambda^{'} E[V_{1}^{2} l_{1}(\theta_{0})] \lambda$. Then, for all $\varepsilon > 0$, there exists $T_{1} = T_{1}(\varepsilon)$ such that

$$P\{|T^{-1}\sum_{i} \bar{\lambda} V_{i}^{2} l_{i}(\theta_{0}) \lambda - E(\bar{\lambda} V_{1}^{2} l_{1}(\theta_{0}) \lambda)| < \delta\} > 1 - \varepsilon$$

for all $T > T_{1}$. That is, setting $M(\theta_{0}) = -\frac{1}{2} E[V_{1}^{2} l_{1}(\theta_{0})]$,

$$P\{-T^{-1}\sum_{i} \bar{\lambda} V_{i}^{2} l_{i}(\theta_{0}) \lambda > \lambda^{'} M(\theta_{0}) \lambda\} > 1 - \varepsilon$$

for all $T > T_{1}$.

(iii) By differentiation of equation (A.2), it can be seen that the only terms that cannot be dealt with as above are those involving $V_{i,h}$, $V_{i,E}$, and $V_{i,E,h}$. Extension of the analysis for first and second derivatives implies that the $h_{t}^{-1} V_{i,h}^{2} h_{t}$ are bounded and the $V_{i,h}^{2}$ have bounded second moments.

Proof of Corollary 3.2.

$$|T^{-1}\sum_{i} (y_{i} - \hat{\mu})^{2} - \text{var}(y_{i})| \leq |T^{-1}\sum_{i} (y_{i} - \mu)^{2} - \text{var}(y_{i})| + (\mu - \hat{\mu})^{2},$$

which converges to zero a.s. by the ergodic theorem. Similarly,

$$|T^{-1}\sum_{i} e_{i}^{2} - E(e_{i}^{2})| \leq |T^{-1}\sum_{i} e_{i}^{2} - E(e_{i}^{2})| + |E(e_{i}^{2}) - E(e_{0}^{2})|.$$ 

Now

$$T^{-1}\sum_{i} e_{i}(\theta_{1})^{2} - T^{-1}\sum_{i} e_{i}(\theta_{2})^{2} = \frac{2}{T} \sum_{i} e_{i}(\theta_{0}) \frac{\partial \hat{c}_{i}}{\partial \theta} (\theta_{1} - \theta_{2}), \quad (A.3)$$

where the derivative is evaluated at $\theta_{0}$, which lies between $\theta_{1}$ and $\theta_{2}$. Since $\Theta$ is compact, $T^{-1}\sum_{i} e_{i}^{2} \rightarrow E(e_{i}^{2})$ uniformly a.s. if $T^{-1}\sum_{i} e_{i}^{2}$ is equicontinuous, i.e., if $|T^{-1}\sum_{i} e_{i}^{2} \partial \hat{c}_{i}/\partial \theta|$ is a.s. uniformly bounded (see, for example, Rissanen and Caines [16]). But this follows from the ergodic theorem. Taking expectations in equation (A.3) shows that $E(e_{i}^{2})$ is continuous in $\theta$ and hence since $\hat{\theta}_{T} \Rightarrow \theta_{0}$,

$$T^{-1}\sum_{i} e_{i}^{2} \Rightarrow E(e_{0}^{2}).$$

Similarly, since $\hat{\theta}_{T} \Rightarrow \theta_{0}$, $T_{y} \rightarrow \infty$, $\hat{\psi}_{i} \Rightarrow \psi_{i0}$ and $(T - i - j)^{-1}\sum_{i} x_{i} x_{i-j} \Rightarrow E(x_{i} x_{i-j})$ a.s. as $T \rightarrow \infty$, it follows that $\hat{c}_{i} \Rightarrow c_{i}$. 

Proof of Theorem 3.3. We again verify the requirements set out in Basawa, Feigin, and Heyde [1]:

(i) $T^{-1/2} \sum l_{i}(\theta_{0}) \overset{d}{\rightarrow} N(0, B_{0})$ for nonrandom $B_{0} > 0$.

(ii) $T^{-1} \sum l_{i}(\theta_{0}) \overset{p}{\rightarrow} A_{0}$ for nonrandom $A_{0} > 0$.

(iii) Condition (iii) of Theorem 3.2.
i) First, we have assumed that $B_0 > 0$. Next, from part (i) of the proof of Theorem 3.2,

$$E(\nabla l_1(\theta_0)|I_{t-1}) = 0.$$ 

Hence we may apply a Martingale central limit theorem (CLT) (see, for example, Theorem A.1.4 of Nicholls and Quinn [14]) if $E[\nabla l_1(\theta_0)\nabla' l_1(\theta_0)] < \infty$ and $E[T\nabla v'(\theta_0)\nabla' L_1(\theta_0)] = B_0 < \infty$. Since $E(\nabla l_1(\theta_0)|I_{t-1}) = 0$, these conditions are equivalent. To show that $E[\nabla l_1(\theta_0)\nabla' l_1(\theta_0)] < \infty$, we use the method in Lemma 3.3 used to show that $A < \infty$. Note also that $B_0 = A_0$ if the conditional distribution of $\varepsilon_{0t}$ is normal.

ii) Since $A_0 = -E[\nabla^2 l_1(\theta_0)]$, see Lemma 3.3.

Proof of Theorem 4.4. From equation (13), $v_{tS} \overset{p}{\to} v_0$ if $T^{-1}\sum\hat{z}_t z_t \overset{p}{\to} T^{-1}\sum\hat{z}^2_t$, and $T^{-1}\sum\hat{z}_t z_t \overset{p}{\to} 0$ (as $m_{LS} \overset{p}{\to} m_0$). But this follows using the method applied to $T^{-1}\sum\varepsilon_t^2$ in Corollary 3.2.

Next, from equation (14), the asymptotic distribution of $T^{1/2}(v_{tS} - v_0)$ depends upon $T^{-1/2}\sum\hat{z}_t \hat{z}_t$ and $T^{-1}\sum \hat{z}_t \eta_t / \hat{m}' T^{1/2}(m_{LS} - m_0)$. But $T^{-1}\sum \hat{z}_t \eta_t / \hat{m}$ converges to $E(\hat{z}_t \eta_t^{1/2} / \hat{m} v_0 \hat{z}_t)$, $T^{1/2}(m_{LS} - m_0)$ is asymptotically equivalent to $-2A_{LS}^{-1}T^{1/2}\sum \varepsilon_t / \hat{m}$ and

$$E(\hat{z}_t \eta_t | I_{t-1}) = 0.$$ 

Also, for any nonzero constant vector $\lambda < \infty$,

$$E(\lambda' \hat{z}_t \eta_t)^2 = E(\lambda' \hat{z}_t \eta_t^2 \hat{z}_t \lambda) < \infty$$ 

since $E(\varepsilon_{0t}^2) < \infty$ (and $E(w_{0t}^2) < \infty$) and hence $E(\varepsilon_t^2) < \infty$. Therefore, appealing to the Martingale CLT as before,

$$F^{-1/2}T^{1/2}(v_{tS} - v_0) \overset{d}{\to} N(0, I_{t^2 + 1}),$$ 

where $F$ is given in equation (16). The existence of $F$ is ensured when $E(\varepsilon_{0t}^2) < \infty$ and estimates of the expectations in equation (16) have the usual form, given by replacing the expectations by sample means evaluated at $v_{tS}$ (and $m_{LS}$). For example, $E(\hat{z}_t \hat{z}_t)$ is estimated by $T^{-1}\sum\hat{z}_t^2$. The consistency of this term was noted earlier, and the consistency of the others is proved in a similar fashion. The estimate of $v_0$ in $F$ is the obvious one, $v_{tS}$. 

\[\blacksquare\]