1 The Linear model with Serially Correlated Data

1.1 Background: asymptotics for serially correlated processes (Review)

• Ergodicity
A process is ergodic if its elements are asymptotically independent – that is, if random variables variance that are far apart in the sequence are essentially statistical independent of one another (see Hayashi, page 101 and Hamilton, pages 46-47). Ergodicity is important because, together with stationarity it leads to a SLLN:

– Ergodic Theorem
Suppose \( \{ z_t \} \) is stationary and ergodic with \( E(z_t) = \mu \). Then

\[
T^{-1} \sum_{t=1}^{T} z_t \xrightarrow{a.s.} \mu.
\]

This is a generalization of the SLLN for week 1’s lecture notes. (For a proof of the theorem and a more detailed discussion see Karlin and Taylor (1975).)

– If \( z_t \) is stationary and ergodic, then so is \( x_t = f(z_t) \) for arbitrary function \( f \).

• CLT for martingale difference sequences (MDS).
Let \( \{ g_t \} \) be a (possibly vector-valued) mds that is stationary and ergodic with \( E(g_t g'_t) = \Sigma_{gg} \). Then

\[
\sqrt{Tg} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t \Rightarrow N(0, \Sigma_{gg})
\]

See Hayashi (p. 106).

Notes:
– While \( \{ g_t \} \) is serially uncorrelated, it may be serially dependent (through higher order moments).
– \( E(g_t g'_t) = \Sigma \) concerns the “unconditional” variance. The conditional variance may be non-constant.

1.2 Linear Model with Serially Correlated Regressors (Review)

Consider the linear model (which defines the parameter \( \beta \)):
(1) \( y_t = x_t' \beta + \varepsilon_t \)

Assume:

(2) \( \{y_t, x_t\} \) is a stationary and ergodic process

(3) \( E(\varepsilon_t x_t) = 0 \), or letting \( g_t = \varepsilon_t x_t \) then \( E(g_t) = 0 \)

(4) \( E(x_t x_t') = \Sigma_{xx} \) which is non-singular

(5) \( \{g_t\} \) is a mds with \( E(g_t g_t') = \Sigma_{gg} \).

Notes:

- We will relax (2’) later in these notes when we discuss integrated processes.
- (3) is weaker than \( E(\varepsilon_t | X) = 0 \) and \( E(\varepsilon_t | x_t) = 0 \). It is sometimes (as in Hamilton) call the assumption of predetermined regressors. We will return to the difference between these two assumptions when we discuss GLS in the context of serially correlated errors.
- Assumption (5) allows the errors to be heteroskedastic conditional on the regressors – that is, it does not require that \( \Sigma_{gg} = \sigma^2 \Sigma_{xx} \).
- (5) is a “high-level” assumption concerning the cross product of \( \varepsilon_t \) and \( x_t \). At a more primitive level it is implied by
  \[
  E(\varepsilon_t | \{\varepsilon_i, x_i\}_{i=1}^{t-l}, x_t) = 0
  \]

Let \( \hat{\beta} \) denote the OLS estimator of \( \beta \).

- Properties of \( \hat{\beta} \)
  1. \( \hat{\beta} \overset{P}{\rightarrow} \beta \) (consistency)
     Proof:
     \[
     \hat{\beta} - \beta = \left( \frac{1}{T} \sum x_t x_t' \right)^{-1} \left( \frac{1}{T} \sum g_t \right)
     \]
     and \( \{x_t x_t'\} \) is stationary and ergodic, with \( E\{x_t x_t'\} = \Sigma_{xx} \), so that
     \[
     \frac{1}{T} \sum x_t x_t' \overset{P}{\rightarrow} \Sigma_{xx}
     \]
     which is non-singular. Also \( \{g_t\} \) is stationary and ergodic with \( E(g_t) = 0 \). Thus
     \[
     \frac{1}{T} \sum g_t \overset{P}{\rightarrow} 0
     \]
     and the result follows from Slutsky’s theorem.
2. $\sqrt{T}(\hat{\beta} - \beta) \overset{d}{\rightarrow} N(0, V_\beta)$, where $V_\beta = \Sigma_{xx}^{-1}\Sigma_{gg}\Sigma_{xx}^{-1}$.

Proof:

$$\sqrt{T}(\hat{\beta} - \beta) = \left(\frac{1}{T} \sum x_t x_t'\right)^{-1} \left(\frac{1}{\sqrt{T}} \sum g_t\right)$$

and

$$\frac{1}{T} \sum x_t x_t' \xrightarrow{p} \Sigma_{xx} \text{ (nonsingular)}$$

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T g_t \xrightarrow{d} N(0, \Sigma_{gg})$$

which follows by the CLT for mds. The result then follows by Slutsky’s theorem.

3. Let $\hat{\Sigma}_{gg}$ be a consistent estimator of $\Sigma_{gg}$. Then

$$\hat{V}_\beta = S_{xx}^{-1}\hat{\Sigma}_{gg}S_{xx}^{-1} \xrightarrow{p} V_\beta$$

where $S_{xx} = \frac{1}{T} \sum x_t x_t'$.

Proof: (Students to fill in – use Slutky’s Theorem)

4. $t_j = \frac{\sqrt{T}(\hat{\beta}_j - \beta_j)}{\sqrt{(\hat{V}_\beta)_{jj}}} \xrightarrow{d} N(0, 1)$

Proof: (Student’s to fill in)

5. Let the Wald Statistic be written as

$$\xi_W = T(R\hat{\beta} - R\beta)'(R\hat{V}_\beta R)'^{-1}(R\hat{\beta} - \beta)$$

then

$$\xi_W \xrightarrow{d} \chi^2_m$$

and

$$\frac{\xi_W}{m} \xrightarrow{d} F_{m, \infty}$$

where $\text{rank}(R) = m$.

Proof: (Student’s to fill in – use from Slutsky’s theorem and the continuous mapping theorem.)

6. Suppose that in addition to (1), (2)-(5) that $\mathbb{E}[(x_{t,i}, x_{t,j})^2]$ is finite for all $i$ and $j$. Let $\hat{\varepsilon}_t = y_t - x_t\hat{\beta}$, $\tilde{g}_t = \hat{\varepsilon}_t x_t$, and $S_{gg} = \frac{1}{T} \sum \tilde{g}_t^2 = \frac{1}{T} \sum \hat{\varepsilon}_t^2 x_t x_t'$. Then

$$S_{gg} \xrightarrow{p} \Sigma_{gg}$$

Proof: (in the model with $k = 1$, for simplicity)

$$\hat{\varepsilon}_t = y_t - x_t\beta - x_t(\hat{\beta} - \beta) = \varepsilon_t - x_t(\hat{\beta} - \beta)$$
so that

\[
S_{\hat{g}\hat{g}} = \frac{1}{T} \sum \varepsilon^2_t x^2_t \\
+ (\hat{\beta} - \beta)^2 \frac{1}{T} \sum x^4_t \\
- 2(\hat{\beta} - \beta) \frac{1}{T} \sum \varepsilon_t x^3_t
\]

Now

(a) \(\frac{1}{T} \sum \varepsilon^2_t x^2_t = \frac{1}{T} \sum g^2_t \to \Sigma_{gg}\) by the ergodic theorem.

(b) \((\hat{\beta} - \beta) \frac{1}{T} \sum x^4_t \to E(x^4_t)
\]

Thus

\[
(\hat{\beta} - \beta)^2 \frac{1}{T} \sum x^4_t \to 0
\]

(c) Note \(E(\varepsilon_t x^3_t)\) is finite since

\[
|E(\varepsilon_t x^3_t)| \leq [E(\varepsilon^2_t x^2_t) E(x^4_t)]^{\frac{1}{2}} \quad \text{(Cauchy-Schwartz)}
\]

and thus

\[
\frac{1}{T} \sum \varepsilon_t x^3_t \to E(\varepsilon_t x^3_t)
\]

so that

\[
(\hat{\beta} - \beta) \frac{1}{T} \sum \varepsilon_t x^3_t \to 0
\]

and the result follows by (a)-(c)

• Application to the AR\((p)\) model

Some preliminaries:

– Suppose \(y_t\) follows the MA process

\[
y_t = \theta(L)\varepsilon_t = \sum_{i=0}^{\infty} \theta_i \varepsilon_{t-i}
\]

where \(\varepsilon_t\) is iid(0,\(\sigma^2\)). Let

\[
\lambda_i = \mathbb{E}(y_t y_{t-i})
\]

denote the \(i'th\) autocovariance of \(\{y_t\}\). If

\[
\sum_{i=0}^{\infty} |\theta_i| < \infty
\]

then

\[
\sum_{i=0}^{\infty} |\lambda_i| < \infty
\]
and the process is stationary and ergodic.

Proof: Hamilton page 69-70 and Dhrymes page 370.

- Suppose $\phi(L)y_t = \varepsilon_t$ where $\varepsilon_t$ is iid$(0, \sigma^2)$, $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ with roots outside the unit circle. Then

$$y_t = \theta(L)\varepsilon_t$$

with $\sum |\theta_i| < \infty$, which can be verified using the notes presented above for AR models.

Now, consider the AR model. Maintain the assumptions that $\varepsilon_t$ is iid$(0, \sigma^2)$ and that the roots of $\phi(z)$ are outside the unit circle. Write the model as

$$y_t = x_t' \beta + \varepsilon_t$$

where $x_t = (y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ and $\beta = (\phi_1, \phi_2, \ldots, \phi_p)$.

Then

$$\hat{\beta} \xrightarrow{p} \beta$$

and

$$\sqrt{T}(\hat{\beta} - \beta) \Rightarrow N(0, V_\beta)$$

where

$$V_\beta = \sigma^2 \Sigma_{xx}^{-1}$$

with $\Sigma_{xx} = E(x_t x_t')$.

Proof: Key Points

- $\{y_t, x_t\}$ is stationary and ergodic
- $[E(x_t x_t')]_{ij} = E(y_{t-i} y_{t-j}) = \lambda_{|i-j|}$
- $g_t = \varepsilon_t x_t$, $\varepsilon_t$ is independent of $x_{t-i}$ for $i \leq 0$, independent of $\varepsilon_{t-i}$ for $i < 0$ and $E(\varepsilon_t) = 0$. Thus $E(\varepsilon_t | \varepsilon_i, x_i)_{t-i} = 0$ and so $g_t$ is a mds.
- $E(g_t g_t') = E(\varepsilon_t^2 x_t x_t') = E[E(\varepsilon_t^2 | x_t) x_t] = \sigma^2 \Sigma_{xx}$.

And the results follow from the general results given above.

AR(1) Example

$$y_t = \beta x_t + \varepsilon_t$$

with $\beta = \phi$ and $x_t = y_{t-1}$. Then

$$\Sigma_{xx} = var(y_{t-1}) = var(y_t) = \frac{\sigma^2}{1 - \phi^2}$$
and

\[ \sqrt{T}(\hat{\phi} - \phi) \Rightarrow N(0, V_{\hat{\phi}}) \]

with

\[ V_{\hat{\phi}} = \sigma^2 \Sigma_{xx}^{-1} = (1 - \phi^2) \]

so that

\[ \hat{\phi} \sim N(\phi, \frac{1}{T}(1 - \phi^2)) \]

and an approximate 95% confidence interval for \( \phi \) is given by

\[ \hat{\phi} \pm 1.96 \left[ \frac{1}{T}(1 - \hat{\phi}^2) \right]^{\frac{1}{2}} \]

### 1.3 Dropping the Assumption that \( \{g_t\} \) is a mds

Reference: Hayashi Chapter 6. Now we drop the assumption that \( \{g_t\} \) is a mds. To do this we need a central limit that we can use for \( \bar{g} \) when \( \{g_t\} \) is not a martingale difference sequence. That is, we need some sufficient conditions that imply

\[ \frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t \Rightarrow N(0, \Omega) \]

Before stating a set of assumptions that gives this result, it is useful to work out an expression for \( \Omega \).

Suppose \( g_t \) is stationary with autocovariances \( \lambda_i \). Then

\[
\text{var}(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t) = \frac{1}{T} \left\{ T\lambda_0 + (T - 1)(\lambda_1 + \lambda_{-1}) + (T - 2)(\lambda_2 + \lambda_{-2}) + ... + (T - r + 1)(\lambda_{r-1} + \lambda_{-T}) \right\} \\
= \sum_{j=-T+1}^{T-1} \lambda_j - \frac{1}{T} \sum_{j=1}^{T-1} j(\lambda_j + \lambda_{-j})
\]

If the autocovariances are “1-summable” so that \( \sum |\lambda_j| < \infty \) then

\[
\text{var}(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t) \to \sum_{j=-\infty}^{\infty} \lambda_j
\]

Recall the autocovariance generating function for \( g_t \) is \( \lambda(z) = \sum_{i=-\infty}^{\infty} \lambda_i \). Thus \( \Omega \) is recognizes as the ACGF evaluated at \( z = 1 \), that is \( \Omega = \lambda(1) \).

Now, a few results:

1. CLT for the \( MA(\infty) \) model

Let

\[ y_t = \mu + \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-1-j} \]
where $\varepsilon_t$ is iid$(0, \sigma^2)$ and $\sum_{j} |\theta_j| < \infty$. Then

$$
\sqrt{T}(\bar{y} - \mu) \Rightarrow N(0, \sum_{j=-\infty}^{\infty} \lambda_j)
$$


- Example: MA(1)
  Suppose $y_t = \mu + \varepsilon_t - \theta \varepsilon_{t-1}$, then $\sqrt{T}(\bar{y} - \mu) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t - \theta \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} \varepsilon_t = (1-\theta) \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t + \theta \frac{1}{\sqrt{T}} (\varepsilon_T - \varepsilon_0)$. Notice $(1-\theta) \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t \overset{L}{\rightarrow} N(0, \sigma^2(1-\theta)^2)$ and $\theta \frac{1}{\sqrt{T}} (\varepsilon_T - \varepsilon_0) \overset{p}{\rightarrow} 0$. Finally, $\sum_{j=-\infty}^{\infty} \lambda_j = \sigma^2 \theta^2(1-\theta)^2$.

2. CLT for ergodic and stationary processes
Suppose that $\{y_t\}$ is stationary and ergodic with finite variance. Then under a set of “dependence” assumptions analogous to absolutely summable MA coefficients

$$
\sqrt{T}(\bar{y} - \mu) \Rightarrow N(0, \sum_{j=-\infty}^{\infty} \lambda_j)
$$

The additional assumptions are given in White (1984), Theorem 5.15.

### 1.4 OLS With Serially Correlated Errors

Maintain all of the OLS assumptions except replace the mds assumption for $g_t$, and let $\lambda_j = E(g_t g'_{t-j})$. Assume

$$
\sqrt{T} \bar{y} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t \Rightarrow N(0, \Omega).
$$

Then all of the OLS results given above continue to hold with $\Omega$ replacing $\Sigma_{gg}$. That is

$$
\sqrt{T}(\hat{\beta} - \beta) \Rightarrow N(0, V_{\hat{\beta}}) \text{ with } V_{\hat{\beta}} = \Sigma_{XX}^{-1} \Omega \Sigma_{XX}^{-1}.
$$

Results for GMM generalize analogously.

### 1.5 HAC and HAR inference

Let $\hat{V}_{\hat{\beta}} = S_{XX}^{-1} \hat{\Omega} S_{XX}^{-1}$ and $\xi_W = T(\hat{\beta} - \beta)' \hat{V}_{\hat{\beta}}^{-1}(\hat{\beta} - \beta)$.

If $\hat{\Omega} \overset{p}{\rightarrow} \Omega$, then $\hat{V}_{\hat{\beta}} \overset{p}{\rightarrow} V_{\beta}$, and $\xi_W \Rightarrow \chi^2_k$.

Here, let’s focus on estimators of $\Omega$ and investigate how this affects inference. The most straightforward example uses $x_t = 1$ and $\beta = 0$, so inference is about the mean of the random variable $y$ when the true mean is zero.
In this example \( g_t = y_t \) are the data, \( E(g_t) = 0 \),

\[
\hat{\beta} - \beta = \bar{g} = T^{-1} \sum_{t=1}^{T} g_t,
\]

so that \( \sqrt{T} \bar{g} \Rightarrow N(0, \Omega) \),

\[
\xi_W(\Omega) = T \bar{g} \Omega^{-1} \bar{g} \Rightarrow \chi^2_1
\]

and \( \xi_W(\hat{\Omega}) = T \hat{g} \hat{\Omega}^{-1} \hat{g} \).

### 1.5.1 Estimators for \( \Omega \)

The goal is estimate \( \Omega = \sum_{j=-\infty}^{\infty} \lambda_j \). With a finite sample of data it is impossible to consistently estimate \( \Omega \) for all possible sequences \( \{\lambda_j\} \). But for special sequences, consistent estimation is possible. The estimators are referred to as ‘heteroskeastic-autocorrelation-consistent’ estimators, abbreviated as HAC.

Here are two examples:

**Example 1:**

Suppose \( |\lambda_j| = 0 \) for \( |j| > 1 \) (so \( g_t \) follows an MA(1) process). In this model, one needs only estimate the variance and first auto-covariance of the process. These can be estimated consistently. Thus

\[ \hat{\Omega} = \sum_{j=-1}^{1} \hat{\lambda}_j \] is consistent.

**Example 2:**

Suppose \( g_t \sim \text{AR}(1) \). In this case \( \lambda_j = \sigma^2 \phi^j / (1 - \phi^2) \), and

\[
\Omega = \sum_{j=-\infty}^{\infty} \lambda_j = \frac{\sigma^2}{(1 - \phi^2)} \sum_{j=-\infty}^{\infty} \phi^j = \frac{\sigma^2}{(1 - \phi)^2}.
\]

(Note, this result can also be deduced from the ACGF for an AR(1), which is \( \lambda(z) = \sigma^2 / [(1 - \phi z)(1 - \phi z^{-1})] \).)

In this case, \( \Omega \) can be consistently estimated by estimating the two parameters characterizing the AR(1) process, \( \sigma^2 \) and \( \phi \) and yields, yielding the estimator:

\[
\hat{\Omega} = \hat{\sigma}^2 / (1 - \hat{\phi})^2.
\]

These two examples are easily generalized.

The logic of example 1 accommodates \( \lambda_{|j|} = 0 \) for \( |j| > k \), where \( k \) is finite.

The logic of example 2 accommodates any (vector) finite order ARMA model.
And even these can be generalized, if they hold “approximately”. There is a large literature on this.

**Truncated estimators:** \( \hat{\Omega} = \sum_{j=-k}^{k} \hat{\lambda}_j \) with \( \hat{\lambda}_j = T^{-1} \sum_{t=1}^{T-j} g_t g'_{t+j} \).

A calculation shows that truncated estimators are not PSD. (They can generate values of \( \hat{\Omega} \) that are not PSD – so in a scalar model \( \hat{\Omega} < 0 \) is possible).

**Weighted truncated estimators:** \( \hat{\Omega}(w) = \sum_{j=-k}^{k} w_j \hat{\lambda}_j \) where \( w_j \) are weights.

Carefully chosen weights ensure PSD estimators. The most widely used estimator is the “Newey-West” estimator:

\[
\hat{\Omega}^{NW} = \sum_{j=-k}^{k} w_j \hat{\lambda}_j
\]

where \( w_{|j|} = (k + 1 - j)/(k + 1) \). These are sometimes called 'Bartlett' weights. The truncation parameter \( k \) is typically chosen as a function of \( T \). (The Stock-Watson UG textbook suggests \( k = 0.75 T^{1/3} \)). \( \hat{\Omega}^{NW} \) is PSD.

Notice that there is a bias-variance tradeoff when choosing \( k \): Small values of \( k \) will lead to estimators \( \hat{\Omega}^{NW} \) with small variance but potentially large bias. Large values of \( k \) will lead to estimators with small bias but large variance.

With Bartlett (i.e., Newey-West) weights, Andrews (1991, ECTA) shows:

- \( \hat{\Omega}^{NW} \xrightarrow{p} \Omega \) if \( k = k(T) \) with \( k(T) \to \infty \) and \( k(T)/T \to 0 \).
- \( \text{MSE}(\hat{\Omega}^{NW}) \) is minimized with \( k(T) \sim O(T^{1/3}) \).

With \( g_t \sim \text{AR}(1) \) with coefficient \( \phi \), Andrews derives the following formula for the MSE minimizing value of \( k \):

\[
\phi \quad k^* \quad T
\]

<table>
<thead>
<tr>
<th>( \phi )</th>
<th>( k^* )</th>
<th>100</th>
<th>400</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.25</td>
<td>0.72 \times T^{1/3}</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>0.50</td>
<td>1.15 \times T^{1/3}</td>
<td>5</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>0.75</td>
<td>1.50 \times T^{1/3}</td>
<td>6</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>0.90</td>
<td>1.70 \times T^{1/3}</td>
<td>7</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>0.95</td>
<td>1.76 \times T^{1/3}</td>
<td>8</td>
<td>12</td>
<td>17</td>
</tr>
</tbody>
</table>

Using rules such as these: \( \xi_W(\hat{\Omega}^{NW}) - \xi_W(\Omega) \xrightarrow{p} 0 \)
so $\xi_W(\hat{\Omega}^{NW}) \Rightarrow \chi^2$

This is all fine, ‘asymptotically,’ but it turns out that size distortions can be large using $\chi^2$ critical values for $\xi_W(\hat{\Omega}^{NW})$.

Example: $\varrho_t$ is a Gaussian AR(1) with coefficient $\phi$. The table below shows the size of 10% tests of $\mu_\varrho = 0$ when $T = 250$ using the Andrews rule given above:

<table>
<thead>
<tr>
<th>$\phi$</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>0.25</td>
<td>0.13</td>
</tr>
<tr>
<td>0.50</td>
<td>0.15</td>
</tr>
<tr>
<td>0.75</td>
<td>0.21</td>
</tr>
<tr>
<td>0.90</td>
<td>0.33</td>
</tr>
<tr>
<td>0.95</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Going beyond ‘first-order asymptotics’: (a calculation in Phillips and Sun (2008, *ECTA*)�

Consider our scalar example where $\xi(\Omega) = T g \Omega^{-1} g$ and $\xi(\hat{\Omega}) = T g \hat{\Omega}^{-1} g$. In this case:

$$P\left(\xi(\hat{\Omega}) > c\right) = P\left(\xi(\Omega) > \frac{\hat{\Omega}}{\Omega} c\right).$$

Let $P\left(\xi(\Omega) > \frac{\hat{\Omega}}{\Omega} c\right) = F(c, \hat{\Omega})$. Then

$$P\left(\xi(\hat{\Omega}) > c\right) = \int P\left(\xi(\Omega) > \frac{\hat{\Omega}}{\Omega} c\bigg| \Omega\right) f(\Omega)d\Omega$$

$$= \int F(c, \hat{\Omega}) f(\hat{\Omega})d\hat{\Omega} = E_{\hat{\Omega}}(F(c, \hat{\Omega}))$$

Write

$$F(c, \hat{\Omega}) = F(c, \Omega) + (\hat{\Omega} - \Omega)F'(c, \Omega) + \frac{1}{2}(\hat{\Omega} - \Omega)^2F''(c, \Omega) + ...$$

then

$$P\left(\xi(\hat{\Omega}) > c\right) \approx F(c, \Omega) + E(\hat{\Omega} - \Omega)F'(c, \Omega) + \frac{1}{2}E(\hat{\Omega} - \Omega)^2F''(c, \Omega)$$

or

$$P\left(\xi(\hat{\Omega}) > c\right) \approx F(c, \Omega) + Bias(\hat{\Omega})F'(c, \Omega) + \frac{1}{2}MSE(\hat{\Omega})F''(c, \Omega)$$

If $\sqrt{T}\bar{g}$ is independent of $\hat{\Omega}$ (as it would be, for example when the data are Gaussian and $\hat{\Omega}$ is computed using 2nd moments of $g$), $F(c, \Omega) = P(\xi(\Omega) > c)$ which can be computed (exactly) from the $\chi^2$ when the data are Gaussian. This is the first-order asymptotic term.
The Bias and MSE are higher-order terms.

Evidently, bias is important (above and beyond its role in MSE).

This suggests that $k$ should be larger than the value chosen to minimize MSE.

1.5.2 HAR estimators for $\Omega$

Heteroskedastic and Autocorrelation Robust (HAR) estimators refers to estimators (and inference methods using $\xi_W(\hat{\Omega})$) for which $\hat{\Omega}$ is not necessarily consistent, and sampling error in $\hat{\Omega}$ is accounted for when conducting inference. Examples include Newey-West estimators with large values of $k$ and ‘projection-based’ estimators. This remains an active area of research. Two recent papers are

- Lazarus, Lewis, Stock (2021, ECTA) who discuss size distortion versus power loss tradeoff. For $\hat{\Omega}^{NW}$, they suggest using $k = 1.3 T^{1/2}$ along with modified (non-$\chi^2$) critical values for $\xi_W(\hat{\Omega})$.
  - With $T = 400$, The Stock-Watson textbook rule of thumb is $k = 0.75 T^{1/3} \approx 6$. The Lazarus, et al rule is much larger: $k = 1.3 T^{1/2} \approx 26$.

1.6 When should you use OLS and HAC standard errors instead of GLS?

1.6.1 Digression: review of GLS

Consider the regression model $Y = X\beta + u$ where $Y$ is $n \times 1$, $X$ is $n \times k$ and so forth.

Suppose that $E(u|X) = 0$ and $Var(u|X) = \Lambda$.

If $\Lambda = \sigma^2 I$, then the OLS estimator of $\beta$, say $\hat{\beta}^{OLS}$ is the best linear unbiased estimator of $\beta$ conditional on $X$ (the Gauss-Markov theorem). Moreover, if the errors have a conditional Gaussian distribution, $\hat{\beta}^{OLS}$ is the MLE and achieves the CR lower bound and therefore in the minimum variance unbiased estimator conditional on $X$.

When $\Lambda \neq \sigma^2 I$, $\hat{\beta}^{OLS}$ does not (in general) have these efficiency properties. Another estimator, the “generalized least squares estimator” does. This estimator is

$$\hat{\beta}^{GLS} = (X'\Lambda^{-1}X)^{-1}X'\Lambda^{-1}Y.$$

To motivate this estimator, write $\Lambda^{-1} = \Lambda^{-1/2} \Lambda^{-1/2}$, so that $\Lambda = \Lambda^{1/2} \Lambda^{1/2}$ and $\Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I$.

Multiplying the regression relation by $\Lambda^{-1/2}$ yields
\[ \Lambda^{-1/2}Y = \Lambda^{-1/2}X\beta + \Lambda^{-1/2}u \] or \[ \tilde{Y} = \tilde{X}\beta + \tilde{u}, \]

where \( \tilde{Y} = \Lambda^{-1/2}Y, \tilde{X} = \Lambda^{-1/2}X, \) and \( \tilde{u} = \Lambda^{-1/2}u. \)

Note that \( \mathbb{E}(\tilde{u}|\tilde{X}) = 0 \) and \( \text{var}(\tilde{u}|\tilde{X}) = I. \)

Because \( \tilde{Y} \) and \( \tilde{X} \) are nonsingular transformations of \( Y \) and \( X \), the best linear unbiased estimator of \( \beta \) conditional on \( X \) is (from the Gauss-Markov Theorem):

\[ \hat{\beta}_{\text{GLS}} = \left( \tilde{X}'\tilde{X} \right)^{-1} \tilde{X}'\tilde{Y} = \left( X'\Lambda^{-1}X \right)^{-1} X'\Lambda^{-1}Y \]

where the final equality follows from the definitions of \( \tilde{Y} \) and \( \tilde{X}. \)

An alternative way to derive the estimator is to write the Gaussian conditional likelihood function: \( Y|X \sim N(X\beta, \Lambda) \), so that

\[ f(Y|X) = (2\pi)^{-n/2}|\Lambda|^{-1/2} \exp \left\{ -\frac{1}{2} (Y - X\beta)' \Lambda^{-1} (Y - X\beta) \right\}, \]

and the MLE solves

\[ \min_b \left( Y - Xb \right)' \Lambda^{-1} \left( Y - Xb \right) \]

Carrying out this minimization yields \( \hat{\beta}_{\text{GLS}} \)

**Examples:**

**Weighted least squares:** Suppose that \( \Lambda = \text{diag}(\sigma^2_i) \). Then \( \Lambda^{-1/2} = \text{diag}\left( \frac{1}{\sigma_i} \right) \),

\[ \hat{y}_i = \frac{y_i}{\sigma_i} \text{ and } \hat{x}_i = \frac{x_i}{\sigma_i}. \]

The GLS estimator can be constructed as the OLS regression of \( \hat{y}_i \) onto \( \hat{x}_i. \)

Notice that this estimator is OLS after re-weighting the observations, where the weight applied to the \( i \)'th observation is \( 1/\sigma_i \) (so that observations corresponding to \( u_i \) with a low variance receive more weight).

**GLS in time series models:**

Suppose that \( u_t = c(L)e_t, \) where \( e_t \sim iid(0,\sigma^2). \)

Then (ignoring effects associated with initial conditions) \( \tilde{u}_t = c(L)^{-1}u_t = e_t, \)

so that the GLS estimator can be constructed by regressing \( c(L)^{-1}y_t \) onto \( c(L)^{-1}x_t \) via OLS.

As an example: suppose that \( (1-\rho L)u_t = \varepsilon_t, \) so that \( c(L) = (1-\rho L)^{-1}. \)

Then the GLS estimator is formed by regressing...
(1−ρL)y_t = y_t−ρy_{t−1}
onumber

onto

(1−ρL)x_t = x_t−ρx_{t−1}.

Notice that one observation is “lost” using this transformation.

A calculation shows that the initial observation should be constructed as \( \tilde{y}_1 = (1 − ρ^2)^{1/2}y_1 \) and \( \tilde{x}_1 = (1 − ρ^2)^{1/2}x_1 \).

Because the first observation is asymptotically negligible relative to the information in the other \( T−1 \) observations, the first observation is often dropped from the analysis.

**Feasible GLS:**

To construct the GLS estimator, you need to know \( Λ \).

Suppose that it is unknown, but depends on a small number of parameters, say \( θ \), so that \( Λ = Λ(θ) \). This suggests that \( Λ \) can be estimated as \( \hat{Λ} = Λ(\hat{θ}) \).

The “Feasible” GLS is

\[
\hat{β}_{FGLS} = (X'\hat{Λ}^{-1}X)^{-1}X'\hat{Λ}^{-1}Y.
\]

In many models it is possible to show that \( \sqrt{n}(\hat{β}_{FGLS} − β_{GLS}) \rightarrow 0 \), so that, in large samples, \( \hat{β}_{FGLS} \) shares the same efficiency properties as \( β_{GLS} \).

1.7 OLS (with HAC inference) or GLS?

Consider the regression model

\[ y_t = x_t'β + u_t \]

where \( u_t \) is the regression error. Suppose that \( E(u_t|x_t) = 0 \), so that \( E(u_tX_t) = 0 \). Given the other assumptions discussed above, the OLS estimator of \( β \) is consistent and asymptotically normal.

Now suppose that \( u_t = ρu_{t−1} + ε_t \) where \( ε_t \) is i.i.d. \( (0, σ^2) \). From above, the GLS regresses \( \tilde{y}_t \) (= \( y_t − ρy_{t−1} \)) onto \( \tilde{x}_t \) using OLS. The error in this regression is \( ε_t = u_t − ρu_{t−1} \).

The key assumption underlying the consistency of the GLS estimator is that \( E(ε_t\tilde{x}) = 0 \). Alternatively, this can be written as

\[
E[(u_t − ρu_{t−1})(x_t − ρx_{t−1})] = E(u_tX_t) + ρ^2E(u_{t−1}x_{t−1}) − ρE(u_tx_{t−1}) − ρE(u_{t−1}x_t) = 0
\]
for this to be true for all values of $\rho$ it must be the case that

\[
E(u_t x_t) = 0 \quad \text{(Term 1)}
\]
\[
E(u_{t-1} x_{t-1}) = 0 \quad \text{(Term 2)}
\]
\[
E(u_t x_{t-1}) = 0 \quad \text{(Term 3)}
\]
\[
E(u_{t-1} x_t) = 0 \quad \text{(Term 4)}
\]

The first two of these are implied by $E(u_t | x_t) = 0$, the same assumption used for the consistency of OLS.

The last two restrictions are not. These two restrictions are implied be stronger assumptions. They are commonly stated as

- $E(u_t | x_t, x_{t-1}, ... ) = 0$. When this assumption holds the regressors are said to be *exogenous* or *predetermined*.

- $E(u_t | ... x_{t+1}, x_t, x_{t-1}, ... ) = 0$. When this assumption holds the regressors are said to be *strictly exogenous*.

Evidently, GLS requires the assumption of strict exogeneity.

## 2 The Functional Central Limit

In the discussion in the last section the data were serially correlated, but stationary, and with autocovariances that died out relatively quickly. When the stochastic processes under study are more persistent, alternative methods are used to characterize and/or approximate the distribution of statistics. An important tool for this analysis is the Functional Central Limit Theorem.

To begin, let’s review three things from Weeks 1 and 2

### 2.1 Review of 3 things from Weeks 1 and 2

1. **Convergence in distribution or “weak convergence”:** $\xi_T, T = 1, 2, ...$ is a sequence of random variables.

   $\xi_T \Rightarrow \xi$ (or $\xi_T \xrightarrow{d}\xi$) means that the probability distribution function (PDF) of $\xi_T$ converges to the PDF of $\xi$.

   (Equivalently, $E(g(\xi_T)) \to E(g(\xi))$ for any continuous bounded function $g$.)

   As a practical matter this means that we can approximate the PDF of $\xi_T$ using the PDF of $\xi$ when $T$ is large.

2. **Central Limit Theorem:** Let $\varepsilon_t$ be a martingale difference sequence with mean 0, variance $\sigma_\varepsilon^2$ and with $2+\delta$ moments. Let
\[ \xi_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t. \] Then \( \xi_T \Rightarrow \xi \sim N(0, \sigma^2). \)

3. **Continuous mapping theorem:** Let \( g \) be a continuous function and \[ \xi_T \Rightarrow \xi, \] then \( g(\xi_T) \Rightarrow g(\xi). \)

Example: \( \xi_T \) is the usual \( t \)-statistic, and \( \xi_T \Rightarrow \xi \sim N(0, 1) \), then \( \xi_T^2 \Rightarrow \xi^2 \sim \chi^2_1. \)

### 2.2 Extensions to random functions:

These ideas can be extended to random functions:

A particular random function: The Wiener Process, a continuous-time stochastic process sometimes called Standard Brownian Motion that will play the role of a “standard normal” in the relevant function space.

Denote the process by \( W(s) \) defined on \( s \in [0,1] \) with the following properties

1. \( W(0) = 0 \)

2. For any dates \( 0 \leq t_1 < t_2 < \ldots < t_k \leq 1 \), \( W(t_2) - W(t_1), W(t_3) - W(t_2), \ldots, W(t_k) - W(t_{k-1}) \) are independent normally distributed random variables with \( W(t_i) - W(t_{i-1}) \sim N(0, t_i - t_{i-1}). \)

3. Realizations of \( W(s) \) are continuous w.p. 1.

From properties (1) and (2), note that \( W(1) \sim N(0, 1). \)

Another Random Function: Suppose \( \varepsilon_t \sim iidN(0,1) \) for \( t = 1, \ldots, T \), and let \( \xi_T(s) \) denote the function that linearly interpolates between the points \( \xi_T(t/T) = \frac{1}{\sqrt{T}} \sum_{i=1}^{T} \varepsilon_i. \)

Can we use \( W \) to approximate the probability law of \( \xi_T(s) \) if \( T \) is large?

More generally, we want to know whether the probability distribution of a random function can be well approximated by the PDF of another (perhaps simpler, maybe Gaussian) function when \( T \) is large. Formally, we want to study weak convergence on function spaces.

Useful References: Hall and Heyde (1980)\(^1\), Davidson (1994)\(^2\), Andrews (1994)\(^3\).

Suppose we limit our attention to continuous functions on \( s \in [0,1] \) (the space of such functions is denoted \( C[0,1] \), and define the distance between two functions, say \( x \) and \( y \) as \( d(x,y) = \sup_{0 \leq s \leq 1} |x(s) - y(s)|. \)

---


Three important theorems (Hall and Heyde (1980) and Davidson (1994, part VI):

Important Theorem 1: (Hall and Heyde Theorem A.2) Weak Convergence of random functions on $C[0, 1]$.

Weak convergence follows from (i) and (ii), where

(i) Let $0 \leq s_1 < s_2 \ldots < s_k \leq 1$, denote a set of $k$ points. Suppose that $(\xi_T(s_1), \xi_T(s_2), ..., \xi_T(s_k)) \Rightarrow (\xi(s_1), \xi(s_2), ..., \xi(s_k))$ for any set of $k$ points, $\{s_i\}$.

(ii) The function $\xi_T(s)$ is “tight” (or more generally satisfies “stochastic equicontinuity” as discussed in Andrews (1994)), meaning

(a) For each $\epsilon > 0$, $P(\sup_{|s-t|<\delta} |\xi_T(s) - \xi_T(t)| > \epsilon) \to 0$ as $\delta \to 0$ uniformly in $T$. (This says that the function $\xi_T$ does not get too “wild” as $T$ grows.)

(b) $P(\xi_T(0) > \delta) \to 0$ as $\delta \to \infty$ uniformly in $T$. (This says the function $\xi_T$ can’t get too crazy at the origin as $T$ grows.)

Important Theorem 2: (Hall on Heyde Theorem A.3) Continuous Mapping Theorem

Let $g : C[0, 1] \to \mathbb{R}$ be a continuous function and suppose $\xi_T(.) \Rightarrow \xi(.)$.

Then $g(\xi_T) \Rightarrow g(\xi)$.

Important Theorem 3: (Hall and Heyde) Functional Central Limit Theorem:

Suppose $\varepsilon_i \sim$ mds with variance $\sigma_\varepsilon^2$ and bounded $2+\delta$ moments for some $\delta > 0$.

(a) Let $\xi_T(s)$ denote the function that linearly interpolates between the points $\xi(t/T) = \frac{1}{\sqrt{T}} \sum_{i=1}^{t} \varepsilon_i(t/T)$. Then $\xi_T \Rightarrow \sigma_\varepsilon W$, where $W$ is a Wiener process (standard Brownian motion).

(b) The results can be extended to $\xi_T(s) = \frac{1}{\sqrt{T}} \sum_{i=1}^{[sT]} \varepsilon_i$, the step-function interpolation, where $\lfloor . \rfloor$ is the “less than or equal to integer function” (so that $[3.1] = 3$, $[3.0] = 3$, $[3.9999] = 3$, and so forth).

See Davidson Ch. 29 for extensions.

An Example:

Let $x_t = \sum_{i=1}^{t} \varepsilon_i$, where $\varepsilon_i$ is mds$(0, \sigma_\varepsilon^2)$, and let $\xi_T(s) = \frac{1}{\sqrt{T}} \sum_{i=1}^{[sT]} \varepsilon_i = \frac{1}{\sqrt{T}} x_{[sT]}$ be a step function approximation of $W(s)$.

Then

$$v_T = \frac{1}{T^{3/2}} \sum_{t=1}^{T} x_t = \frac{1}{T} \sum_{t=1}^{T} \left[ \frac{1}{T^{1/2}} \sum_{i=1}^{t} \varepsilon_i \right] = \sigma_\varepsilon \int_{0}^{1} \xi_T(s)ds \Rightarrow \sigma_\varepsilon \int_{0}^{1} W(s)ds = v$$
What does this all mean?

Suppose I want to approximate the 95\(^{th}\) quantile of the distribution of, say, 
\[ \nu_T = \frac{1}{T^{3/2}} \sum_{t=1}^{T} x_t. \] Because \( \nu_T \Rightarrow v = \sigma \int_0^1 W(s)ds \), I can use the 95\(^{th}\) quantile of \( v \) are the approximator.

How do I find (or approximate) the 95\(^{th}\) quantile of \( v \)?

Use Monte Carlo draws of \( \sigma \epsilon N^{-3/2} \sum_{i=1}^{N} \sum_{i=1}^{t} z_i \) where \( z_i \sim \text{iidN}(0,1) \) and \( N \) is very large.

This approximation works well when \( T \) is reasonably large, and does not require knowledge of the distribution of \( x \).

### 2.3 Application: Testing for a “Break” in a regression function

Model: \( y_t = \beta_t + \epsilon_t \), where \( \epsilon_t \sim \text{iid (0, } \sigma^2_\epsilon) \)

\[ \beta_t = \begin{cases} 
\beta & \text{for } t \leq \tau \\
\beta + \delta & \text{for } t > \tau 
\end{cases} \]

Null and alternative: \( H_0 : \delta = 0 \) vs. \( H_a : \delta \neq 0 \)

Tests for \( H_0 \) versus \( H_a \) depend on whether \( \tau \) is known or unknown.

#### 2.3.1 Chow\(^4\) tests (known break date)

The least squares estimator of \( \delta \) is \( \hat{\delta} = \bar{Y}_2 - \bar{Y}_1 \)

where \( \bar{Y}_1 = \frac{1}{\tau} \sum_{i=1}^{\tau} y_t \) and \( \bar{Y}_2 = \frac{1}{T-\tau} \sum_{i=\tau+1}^{T} y_t \).

Note that \( \bar{Y}_1 \sim N(\beta, \sigma^2_\epsilon) \), \( \bar{Y}_2 \sim N(\beta + \delta, \sigma^2_\epsilon \frac{1}{T-\tau}) \), and they are independent. Thus \( \hat{\delta} \sim N \left( \delta, \sigma^2_\epsilon \left( \frac{1}{\tau} + \frac{1}{T-\tau} \right) \right) \).

The Wald statistic is therefore: \( \xi_W = \frac{1}{\hat{\sigma}^2} \frac{\hat{\delta}^2}{1 + \frac{1}{\tau} + \frac{1}{T-\tau}} \Rightarrow \xi \sim \chi^2_1 \)

Under \( H_0 \) \( \xi_W \) is distributed as a \( \chi^2_1 \) random variable in large \( (\tau \text{ and } T-\tau) \) samples. Thus, critical values for the test can be determined from the \( \chi^2 \) distribution.

2.3.2 Quandt Tests (Sup Wald or QLR) (unknown break date)

Quandt (1960) suggested computing the Chow statistic for a large number of possible values of $\tau$ and using the largest of these as the test statistics.

QLR statistic: $\xi_Q = \max_{\tau_1 \leq \tau \leq \tau_2} \xi_W(\tau)$

where the Chow statistic $\xi_W$ is now indexed by the break date.

The problem is then to find the distribution of $\xi_Q$ under the null (it will not be $\chi^2$), so that the critical value for the test can be determined.

Let $s = \tau/T$. Under the null $\delta = 0$, and (now using $s$ as the index), we can then write $\xi_W$ as

$$\xi_{W,T}(s) = \frac{1}{\hat{\sigma}^2} \left[ \frac{1}{(sT)} \sum_{t=1}^{sT} y_t - \frac{1}{((1-s)T)} \sum_{t=sT+1}^{T} y_t \right]^2$$

$$= \frac{1}{\hat{\sigma}^2} \frac{1}{(sT)} \sum_{t=1}^{sT} \varepsilon_t - \frac{1}{((1-s)T)} \sum_{t=sT+1}^{T} \varepsilon_t^2$$

$$= \frac{1}{\hat{\sigma}^2} \frac{1}{s} \frac{1}{\sqrt{T}} \sum_{t=1}^{sT} \varepsilon_t - \frac{1}{(1-s)} \frac{1}{\sqrt{T}} \sum_{t=sT+1}^{T} \varepsilon_t^2$$

$$= \frac{1}{\hat{\sigma}^2} \left[ \frac{1}{s} W_T(s) - \frac{1}{(1-s)} (W_T(1) - W_T(s)) \right]^2$$

$$= \frac{[\frac{1}{s} W_T(s) - sW_T(1)]^2}{s(1-s)}$$

where $W_T(s) = \frac{1}{\hat{\sigma} \sqrt{T}} \sum_{t=1}^{sT} \varepsilon_t$, and the last equality follows from multiplying the numerator and denominator by $s^2(1-s)^2$ and simplifying.

Thus, using FCLT, $\xi_T \Rightarrow \xi$ where $\xi(s) = \frac{[W(s) - sW(1)]}{s(1-s)}$.

Suppose that $\tau_1$ is chosen as $\lambda T$ and $\tau_2$ is chosen as $((1-\lambda) T)$, where

$0 < \lambda < 0.5$. Then

$$\xi_Q = \sup_{\lambda \leq s \leq (1-\lambda)} \xi_{W,T}(s), \text{ and } \xi_Q \Rightarrow \sup_{\lambda \leq s \leq (1-\lambda)} \xi(s) \text{ (using the continuous mapping theorem, noting that } '\sup' \text{ is a continuous function).}$$

It has become standard practice to use a value of $\lambda = 0.15$.

The results have been derived here for the case of a single constant regressor. Extensions to the case of multiple (non-constant) regressors can be found in Andrews (1993)\textsuperscript{6} (Critical values for the test statistic are also given in Andrews (1993) with corrections in Andrews (2003), reprinted in Stock and Watson UG textbook.)

<table>
<thead>
<tr>
<th>Number of Restrictions ($q$)</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.12</td>
<td>8.68</td>
<td>12.16</td>
</tr>
<tr>
<td>2</td>
<td>5.00</td>
<td>5.86</td>
<td>7.78</td>
</tr>
<tr>
<td>3</td>
<td>4.09</td>
<td>4.71</td>
<td>6.02</td>
</tr>
<tr>
<td>4</td>
<td>3.59</td>
<td>4.09</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>3.26</td>
<td>3.66</td>
<td>4.53</td>
</tr>
<tr>
<td>6</td>
<td>3.02</td>
<td>3.37</td>
<td>4.12</td>
</tr>
<tr>
<td>7</td>
<td>2.84</td>
<td>3.15</td>
<td>3.82</td>
</tr>
<tr>
<td>8</td>
<td>2.69</td>
<td>2.98</td>
<td>3.57</td>
</tr>
<tr>
<td>9</td>
<td>2.58</td>
<td>2.84</td>
<td>3.38</td>
</tr>
<tr>
<td>10</td>
<td>2.48</td>
<td>2.71</td>
<td>3.23</td>
</tr>
<tr>
<td>11</td>
<td>2.40</td>
<td>2.62</td>
<td>3.09</td>
</tr>
<tr>
<td>12</td>
<td>2.33</td>
<td>2.54</td>
<td>2.97</td>
</tr>
<tr>
<td>13</td>
<td>2.27</td>
<td>2.46</td>
<td>2.87</td>
</tr>
<tr>
<td>14</td>
<td>2.21</td>
<td>2.40</td>
<td>2.78</td>
</tr>
<tr>
<td>15</td>
<td>2.16</td>
<td>2.34</td>
<td>2.71</td>
</tr>
<tr>
<td>16</td>
<td>2.12</td>
<td>2.29</td>
<td>2.64</td>
</tr>
<tr>
<td>17</td>
<td>2.08</td>
<td>2.25</td>
<td>2.58</td>
</tr>
<tr>
<td>18</td>
<td>2.05</td>
<td>2.20</td>
<td>2.53</td>
</tr>
<tr>
<td>19</td>
<td>2.01</td>
<td>2.17</td>
<td>2.48</td>
</tr>
<tr>
<td>20</td>
<td>1.99</td>
<td>2.13</td>
<td>2.43</td>
</tr>
</tbody>
</table>

Note: These critical values apply when $\tau_0 = 0.15T$ and $\tau_1 = 0.85T$ (rounded to the nearest integer), so the $F$-statistic is computed for all potential break dates in the central 70% of the sample. The number of restrictions $q$ is the number of restrictions tested by each individual $F$-statistic. Critical values for other trimming percentages are given in Andrews (2003).

### 2.4 Application: Unit root AR(1) model

Consider the Gaussian AR(1) model

$$y_t = \phi y_{t-1} + \varepsilon_t$$

---

with $y_0 = 0$ and where $\varepsilon_t \sim \text{Niid}(0, 1)$. Let

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

denote the least squares estimator or $\phi$.

In the stationary model ($|\phi| < 1$), we derived several results including:

1. $\hat{\phi} \to \phi$ at rate $T^{\frac{1}{2}}$
2. The limiting distribution of $\hat{\phi}$ is normal with $T^{1/2}(\hat{\phi} - \phi) \Rightarrow N(0, (1 - \phi^2))$.

and that Hayashi-style GMM inference methods can be used.

From (2) note that the limiting variance converges to 0 as $\phi \to 1$ suggesting the normal approximation is poor for $\phi$ close to 1. Indeed when $\phi = 1$ these result change in an important way. Specifically:

1. $\hat{\phi} \to \phi$ at rate $T$
2. The limiting distribution of $\hat{\phi}$ is Non-normal

We will work out some results for a model in which $\phi = 1$. This makes it easy to contrast limiting results to the stationary model.

To begin, as usual we have the identity

$$\hat{\phi} - \phi = \frac{\sum y_{t-1} \varepsilon_t}{\sum y_{t-1}^2}$$

so that

$$T(\hat{\phi} - \phi) = \frac{1}{T} \frac{\sum y_{t-1} \varepsilon_t}{\sum y_{t-1}^2}$$

(Note the non-standard standardization factors involving $T$. ) We will analyze the numerator and denominator separately. We begin with the numerator.

Since $\phi = 1$,

$$y_t = y_{t-1} + \varepsilon_t$$

$$\sum y_t^2 = \sum y_{t-1}^2 + 2 \sum y_{t-1} \varepsilon_t + \sum \varepsilon_t^2$$

so that

$$\frac{1}{T} \sum y_{t-1} \varepsilon_t = \frac{1}{2} \left[ \frac{1}{T} \sum y_t^2 - \frac{1}{T} \sum y_{t-1}^2 - \frac{1}{T} \sum \varepsilon_t^2 \right]$$

$$= \frac{1}{2} \left[ \frac{1}{T} \sum y_t^2 - \frac{1}{T} \sum \varepsilon_t^2 \right]$$

Also

$$y_T = \sum_{t=1}^{T} \varepsilon_t$$
so that
\[ \frac{1}{T} y_T^2 = \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_t \right)^2 \sim [N(0,1)]^2 \sim \chi_1^2 \]
and
\[ \frac{1}{T} \sum \varepsilon_t^2 \overset{p}{\rightarrow} 1 \]
Putting these together
\[ \frac{1}{T} \sum y_{t-1} \varepsilon_t \overset{d}{\rightarrow} \frac{1}{2} [\chi_1^2 - 1] \]
And thus the numerator is non-normal. From this expression we can learn about the median bias of \( \hat{\phi} \). Note that \( \text{sign}(\hat{\phi} - \phi) = \text{sign}(\frac{1}{T} \sum y_{t-1} \varepsilon_t) \), so that \( \mathbb{P}(\hat{\phi} < \phi) = \mathbb{P}(\chi_1^2 < 1) = 0.68 \).

The denominator can be analyzed using the FCLT. Write
\[ \frac{1}{T^2} \sum_{t=1}^T y_t^2 = \frac{1}{T} \sum_{t=1}^T \left( \frac{1}{\sqrt{T}} y_t \right)^2 \]
and let \( \xi_T(s) = \frac{1}{\sqrt{T}} y_{[sT]} \) be a step function approximation of \( W(s) \). Then
\[ \frac{1}{T} \sum_{t=1}^T \left( \frac{1}{\sqrt{T}} y_t \right)^2 = \int_0^1 \xi_T(s)^2 ds \Rightarrow \int_0^1 W(s)^2 ds \]
Putting this together with the expression for the numerator, we have
\[ T(\hat{\phi} - 1) \Rightarrow \frac{1}{2} \frac{[\chi_1^2 - 1]}{\int_0^1 W(s)^2 ds} \]
There is another limiting representations of the random variable that make the dependence of the numerator and denominator clearer:
\[ \frac{1}{T} \sum y_{t-1} \varepsilon_t \Rightarrow \int_0^1 W(s)dW(s) \sim \frac{1}{2} [\chi_1^2 - 1] \]
But, the important thing to see, however, is the limiting random variable is non-normal and so standard inference procedures cannot be used.

The non-standard distribution carries over to the t-statistic for testing \( H_0 : \phi = 1 \). A calculation like the one carried above shows
\[ t - \text{stat} = \frac{\int_0^1 W(s)dW(s)}{\left[ \int_0^1 W(s)^2 ds \right]^{1/2}}. \]
Tests of \( H_0 : \phi = 1 \) versus \( H_a : \phi < 1 \) are called ‘unit root’ tests. These tests typically use a one-sided t-test using critical values that are appropriate for the (non-Standard Normal) large-sample distribution of the t-statistic under the null. (The preceding paragraph contains an example.) As it turns out, when \( \phi = 1 \), the distribution of the t-statistic depends on whether or not the estimating equation includes an intercept (a constant regressor) or a time trend or other deterministic regressors. You can see details in in Hamilton, Hayashi, the Stock-Watson undergraduate textbook, or many other places.