Stationarity and Autocorrelation

TOPICS IN FINANCIAL ECONOMETRICS: CHAPTER 1

September 6, 2016

1 Stationarity

Time series is a collection of random variables over time $X_1, \cdots, X_t, \cdots, X_T$. Stationarity is a common assumption used in a time series model. A time series is stationary if its first two moments are constant over time (does not depend on $t$). That is

$$EX_t = \mu$$
$$\text{Cov}(X_t, X_{t-k}) = \gamma_k \text{ (lag-$k$ autocovariance)}.$$

The lag zero autocovariance $\gamma_0$ is just the variance of the time series.

$\gamma_k$’s are of central importance in time series analysis as they characterize the serial dependence over observations (i.e. over time). We usually do not have iid data in time series contexts.

A white noise series is stationary. A white noise (WN) series is defined as an uncorrelated series with $EX_t = 0$ and $EX_t^2 = \sigma^2$.

A trend model is not stationary. Let $X_t = \alpha + \beta t + \varepsilon_t$, where $\varepsilon_t$ is white noise.

A random walk is not stationary either. Let $X_t$ be such that $X_t = X_{t-1} + \varepsilon_t$ (where $\varepsilon_t$ is WN). Assuming $X_0 = 0$ (the initial value), $X_t = \varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_1$. Although $EX_t = 0$, $\text{Var}(X_t) = t\sigma^2$ depends on $t$.

2 Autocorrelations

Autocorrelations can sometime be more convenient than autocovariances, as they always take value between $-1$ and $1$. Autocorrelations are defined as

$$\rho_k = \text{Cor}(X_t, X_{t-k}) = \gamma_k / \gamma_0.$$

Under stationarity, $\rho_k$ is also constant over time. $\rho_k$ is also called the autocorrelation function (ACF) of $x_t$. Plotting $\rho_k$ against $k$ is a useful way to check serial correlations (called ACF plots).

3 An example: MA model

Moving-average (MA) model is an extension of the white noise. MA(1) model

$$X_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1},$$
where $\varepsilon_t$ is WN$(0, \sigma^2)$. For example, if $X_t$ is price change, then $\varepsilon_t$ represents the effects of unexpected news. The impact may not be fully absorbed by the market; say, it takes 2 days. It leads to an MA(1) model of $X_t$.

MA(1) model is stationary. First and second moments are

\[
\begin{align*}
\mathbb{E}X_t &= \mu \\
\gamma_0 &= \text{Var}(X_t) = (1 + \theta^2)\sigma^2 \\
\gamma_1 &= \theta \sigma^2 \\
\gamma_2 &= \gamma_3 = \ldots = 0 \\
\rho_1 &= \theta/(1 + \theta^2) \\
\rho_2 &= \rho_3 = \ldots = 0.
\end{align*}
\]

In general, MA($q$) model:

\[X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q}.\]

MA($q$) model can only capture the first $q$-order correlation (i.e. finite memory).

4 How to estimate autocovariances?

By sample autocovariances. They are

\[
\hat{\gamma}_k = \frac{1}{T} \sum_{t=k+1}^{T} (X_t - \overline{X})(X_{t-k} - \overline{X}), \quad 0 \leq k \leq T - 1,
\]

where $\overline{X} = T^{-1} \sum_{t=1}^{T} X_t$. Some software uses the divisor $T - k$ (instead of $T$ above), but the difference is very small asymptotically (i.e. when $T$ is large).

Similarly, $\rho_k$ is estimated by the sample autocorrelation $\hat{\rho}_k$:

\[
\hat{\rho}_1 = \frac{\hat{\gamma}_1}{\hat{\gamma}_0} = \frac{\sum_{t=2}^{T} (X_t - \overline{X})(X_{t-1} - \overline{X})}{\sum_{t=1}^{T} (X_t - \overline{X})^2},
\]

\[
\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=k+1}^{T} (X_t - \overline{X})(X_{t-k} - \overline{X})}{\sum_{t=1}^{T} (X_t - \overline{X})^2}, \quad 0 \leq k \leq T - 1.
\]

5 The distribution of autocorrelations

We will establish two important results in (1) and (2). They tell you how large autocorrelations are deemed as "large".

The first one is concerned about the individual $\hat{\rho}_k$. If $X_t$ is iid $(0, \sigma^2)$,

\[
\sqrt{T}\hat{\rho}_k \overset{d}{\to} \mathcal{N}(0,1).
\]  

for any $k \geq 1$. If your data do not support (1), they are not likely to be iid.

The second is about the joint behavior of $K$ correlations, $\{\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_K\}$. If $X_t$ is iid $(0, \sigma^2)$, then

\[
Q(K) = T \sum_{k=1}^{K} \hat{\rho}_k^2 \overset{d}{\to} \chi^2(K).
\]
$Q(K)$ is called Box-Pierce statistic. These two results provide theoretical basis for two most commonly used methods to check serial correlation: ACF plots and Box-Pierce tests. For example, to test the hypothesis $\rho_1 = \rho_2 = \ldots = \rho_{10} = 0$ (the first ten autocorrelations are all zeros), we reject (at 5% level) if $Q(10)$ is larger than the 95% quantile of $\chi^2(10)$. We accept otherwise.

6 Multivariate time series

The concepts above can be extended to a multivariate time series. Consider a bivariate series $X_t = (X_{1t}, X_{2t})'$. $X_t$ is stationary if its first moment (now a 2-dim vector) and second moments (2 x 2 matrices) are time-invariant. Define

$$
\begin{align*}
\mu & = \mathbb{E}X_t \\
\Gamma_k & = \mathbb{E}(X_t - \mu)(X_{t-k} - \mu)'
\end{align*}
$$

where $k \in \{-2, -1, 0, 1, 2, \ldots\}$. $\mu$ and $\Gamma_k$ do not rely on $t$.

$\Gamma_k$ is called the lag-$k$ autocovariance matrix. The lag-0 autocovariance matrix, $\Gamma_0 = \mathbb{E}(X_t - \mu)(X_t - \mu)'$, is the variance matrix of $X_t$.

7 Autocorrelation matrices

Suppose $X_t$ is a $m$-dim time series. The concurrent correlation matrix of $X_t$, $\rho_0$, is such that its $(i,j)$-th element is

$$
\rho_{0,ij} = \frac{\text{cor}(X_{it}, X_{jt})}{\sqrt{\Gamma_{0,ii}} \sqrt{\Gamma_{0,jj}}},
$$

where $i, j = 1, \ldots, m$. How the matrices $\rho_0$ and $\Gamma_0$ are connected?:

$$
\rho_0 = D^{-1}\Gamma_0 D^{-1}.
$$

where $D = \text{diag}(\sqrt{\Gamma_{0,11}}, \ldots, \sqrt{\Gamma_{0,mm}})$. (Please verify.) The matrix $\rho_0$ is symmetric.

$\rho_0$ is also called the lag-0 autocorrelation matrix. In general, the lag-$k$ autocorrelation matrix

$$
\rho_k = D^{-1}\Gamma_k D^{-1} = (\rho_{k,ij} = \text{cor}(X_{it}, X_{j,t-k}))_{i,j=1,\ldots,m}.
$$

$^1$The 95% quantile of the $\chi^2(10)$ distribution is the number $q$ such that

$$
P(\text{a } \chi^2(10) \text{ random variable } < q) = 0.95.$$
In other words, $\rho_{k,ij}$ is the lag-$k$ autocorrelation between $X_i$ and $X_j$ (series $i$ leads series $j$ by $k$ periods).

$\Gamma_k$ and $\rho_k$ are in general not symmetric for $k \neq 0$, meaning that $\Gamma_k \neq \Gamma'_k$ and $\rho_k \neq \rho'_k$. However, we have

$$
\Gamma_k = \Gamma'_{-k}, \\
\rho_k = \rho'_{-k},
$$

e.g. $\rho_{k,12} = \rho_{-k,21}$.

To understand the asymmetry of $\rho_k$, look at the following graph. The correlations in two pairs ($\rho_{1,21}$ and $\rho_{1,12}$) are not necessarily equal.

For example, index returns may predict returns of an individual stock. But the reverse is not true. (Put it differently, the response of $X_1$ to previous movements in $X_2$ could be completely different from the response of $X_2$ to previous movements in $X_1$.)

This should not be confused with stationarity. Stationarity means the correlations in two pairs (in the following graph) are the same.

$$
\begin{array}{ccc}
X_{1,t-1} & X_{1,t} & X_{1,t+1} \\
\times & \times & \times \\
X_{2,t-1} & X_{2,t} & X_{2,t+1}
\end{array}
$$

8 Verification of (1) and (2)

This verification (which does not serve as a rigorous proof) is a good exercise to get familiar with time series concepts. (1) and (2) hold from (3) in the following.

If $X_t$ is iid $(0, \sigma^2)$,

$$
\sqrt{T} \left( \begin{array}{c} \hat{\rho}_1 \\ \vdots \\ \hat{\rho}_K \end{array} \right) \overset{d}{\rightarrow} N(0, I_K),
$$

where $I_K$ is the $K \times K$ identity matrix.

We consider a simple case in which we know the mean is zero. Extension to non-zero mean is a matter of more involved algebra. Then we can use

$$
\hat{\rho}_k = \frac{\sum_{t=k+1}^{T} X_t X_{t-k}}{\sum_{t=1}^{T} X_t^2},
$$

and we will show (3) holds in this simple case. We need to show

(i). $\sqrt{T}\hat{\rho}_k \overset{d}{\rightarrow} N(0, 1)$.
(ii). $TE\hat{\rho}_k \hat{\rho}_{k'} \rightarrow 0$ (for $k \neq k'$).

Since

$$
\sqrt{T}\hat{\rho}_k = \frac{T^{-1/2} \sum_{t=k+1}^{T} X_t X_{t-k}}{T^{-1} \sum_{t=1}^{T} X_t^2},
$$
and $T^{-1} \sum_{t=1}^{T} X_t^2 \rightarrow \sigma^2$, we need to show

$$T^{-1/2} \sum_{t=k+1}^{T} X_t X_{t-k} \xrightarrow{d} N(0, \sigma^4). \quad (4)$$

Note that $E(T^{-1/2} \sum_{t=k+1}^{T} X_t X_{t-k}) = 0$ and $\text{Var}(T^{-1/2} \sum_{t=k+1}^{T} X_t X_{t-k}) = E(T^{-1/2} \sum_{t=k+1}^{T} X_t X_{t-k})^2 = \sigma^4$ (since $X_t$ is iid $(0, \sigma^2)$). The normality in (4) follows from the central limit theorem. So (i) holds.

Consider

$$T \hat{p}_1 \hat{p}_2 = \frac{T^{-1} \sum_{t=2}^{T} X_t X_{t-1} \sum_{t=3}^{T} X_t X_{t-2}}{(T^{-1} \sum_{t=1}^{T} X_t^2)^2}.$$ 

Since $(T^{-1} \sum_{t=1}^{T} X_t^2)^2 \rightarrow \sigma^4$ and $E(T^{-1} \sum_{t=2}^{T} X_t X_{t-1} \sum_{t=3}^{T} X_t X_{t-2} = 0$, so (ii) is verified for $k = 1$ and $k' = 1$. The general case follows similarly. Thus (3) holds.
Autoregression Models

TOPICS IN FINANCIAL ECONOMETRICS: CHAPTER 2

September 13, 2016

1 AR(p) model

If we have detected that the time series $x_t$ is not likely to be serially uncorrelated (e.g. through ACF plots or Box-Pierce test), autoregression (AR) model is the simplest model to capture its serial dependence.

We say that $x_t$ follows an AR($p$) model if

$$x_t = a_0 + a_1 x_{t-1} + \ldots + a_p x_{t-p} + \varepsilon_t;$$

(1)

where $\varepsilon_t$ is the error term (white noise, or WN(0,$\sigma^2$)). We can write it differently as

$$\begin{align*}
(1-a_1 L - \ldots - a_p L^p) x_t &= a_0 + \varepsilon_t, \\
 a(L)x_t &= a_0 + \varepsilon_t;
\end{align*}$$

(2)

where $L$ is the lag operator ($L x_t = x_{t-1}$, $L^2 x_t = x_{t-2}$, etc.), and $a(L)$ is a polynomial of $L$.

$a_0, a_1, \ldots, a_p$ (slopes) and $\sigma^2$ (the error variance) are parameters to be estimated by the data. The most commonly used estimator is by the OLS (ordinary least squares), i.e. running a regression of $x_t$ on an intercept, and $p$ regressors $x_{t-1}, \ldots, x_{t-p}$. If $x_t$ is actually serially uncorrelated, $a_1, \ldots, a_p$ should be all close to zero.

You will also need to decide on $p$, i.e. how many lags you include in your model.

Consider the AR(1) model

$$x_t = a_0 + a_1 x_{t-1} + \varepsilon_t.$$

If $x_t$ follows this model, it is not always stationary. For stationarity, we need $-1 < a_1 < 1$ (with no constraints on $a_0$). If $a_1 = 1$, $x_t$ is called to follow a random walk (which is not stationary).

2 Autocovariance-generating function

We may ask what autocovariances look like if $x_t$ follows an AR(p) model. We have asked the same question in Topic 1 when $x_t$ is a moving average (MA) model. One reason we are interested in this type of question is to see what the model limitations are. For example, MA models can not capture autocorrelation beyond a certain lag. AR models can do better in this sense but have other limitations. Another reason to answer this question is, we can calculate all the autocovariances ($\gamma_j$’s) and autocorrelations ($\rho_j$’s) when the model parameters ($a_0, a_1, \ldots, a_p, \sigma^2$) have been decided. This way of estimating $\gamma_j$’s and $\rho_j$’s can be more accurate (than we did in Topic 1) if the model is correct.
The following concept will be useful. The autocovariance-generating function (AGF) of $x_t$ is defined as

$$g_x(z) = \sum_{j=-\infty}^{\infty} \gamma_j z^j$$

$$= \gamma_0 + \sum_{j=1}^{\infty} \gamma_j (z^j + z^{-j}).$$

If $x_t$ is WN(0,$\sigma^2$), then $g_x(z) = \sigma^2$.

3 A powerful result

**Lemma:** Suppose $x_t$ is stationary and $\sum_{j=0}^{\infty} |\gamma_j| < \infty$. If

$$x_t = h(L)e_t = (1 + h_1L + h_2L^2 + ...)e_t$$

$$= e_t + h_1e_{t-1} + h_2e_{t-2} + ...$$

($\sum_{j=0}^{\infty} |h_j| < \infty$), where $e_t$ is WN(0,$\sigma^2$), then

$$g_x(z) = \sigma^2 h(z)h(z^{-1}). \quad (3)$$

(3) The result will be useful later when we calculate an important quantity, the long-run variance.) We now apply this result to MA and AR models.

3.1 MA model

Consider MA(2) model, which follows

$$x_t = \mu + e_t + b_1e_{t-1} + b_2e_{t-2}$$

$$= \mu + (1 + b_1L + b_2L^2)e_t,$$

with $e_t$ being WN. Using the Lemma,

$$g_x(z) = \sigma^2 (1 + b_1z + b_2z^2)(1 + b_1z^{-1} + b_2z^{-2})$$

$$= \sigma^2 [1 + b_1z + b_2z^2 + b_1(z^{-1} + b_1z^{-1} + b_2z^{-1} + b_2)]$$

So collecting the coefficients on terms $1, z, z^2$ respectively gives

$$\gamma_0 = \sigma^2 (1 + b_1^2 + b_2^2)$$

$$\gamma_1 = \sigma^2 (b_1 + b_1b_2)$$

$$\gamma_2 = \sigma^2 b_2.$$

(What are $\gamma_3$ and $\gamma_4$?). These confirm what we obtained in Topic 1.
3.2 AR model

Consider the AR\((p)\) model (as defined in (2)). Assume \(a_0 = 0\). We can equivalently write

\[ x_t = a(L)^{-1} e_t \]

Using the Lemma,

\[ g_x(z) = \frac{\sigma^2}{a(z)a(z^{-1})}. \]

For AR\((1)\) model \((x_t = a_1x_{t-1} + e_t)\),

\[ g_x(z) = \frac{\sigma^2}{(1 - a_1z)(1 - a_1/z)} = \sigma^2 (1 + a_1z + a_1^2z^2 + \ldots)(1 + a_1z^{-1} + a_1^2z^{-2} + \ldots). \]

Collecting coefficients in the polynomial above,

\[ \gamma_0 = \sigma^2(1 + a_1 + a_1^2 + \ldots) = \sigma^2/(1 - a_1^2) \]
\[ \gamma_1 = \sigma^2(a_1 + a_1^3 + a_1^5 + \ldots) = \sigma^2a_1/(1 - a_1^2) \]
\[ \gamma_2 = \sigma^2(a_1^2 + a_1^4 + a_1^6 + \ldots) = \sigma^2a_1^2/(1 - a_1^2). \]

(What are \(\gamma_3\) and \(\gamma_4^2\)).

4 Vector Autoregression

A \(k\)-dimensional time series follows a vector autoregression (VAR) model if

\[ x_t = \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_p \end{bmatrix} x_{t-1} + \begin{bmatrix} \vdots \\ \vdots \end{bmatrix} + \begin{bmatrix} A_p x_{t-p} + e_t \\ \vdots \end{bmatrix}, \quad (4) \]

where \(e_t\) is WN\((0,\Phi)\). Note that the variance-covariance matrix \(\Phi\) is \(k \times k\).

The coefficient matrices \(A_0, A_1, \ldots, A_p\) can be estimated by using OLS to each equation in (4). (How many regressors for each equation?) Note that there are \(k\) equations in (4). This is called the equation-by-equation OLS procedure.

The autocovariance-generating function (AGF) of the \(k\)-dim time series \(x_t\) is defined as

\[ g_x(z) = \sum_{j=-\infty}^{\infty} \Gamma_j z^j = \Gamma_0 + \sum_{j=1}^{\infty} (\Gamma_j z^j + \Gamma_j' z^{-j}), \]

since \(\Gamma_{-j} = \Gamma_j'\).

The multivariate extension of the result (3) is the following. Suppose \(x_t\) is stationary and \(\sum_{j=0}^{\infty} |\Gamma_j| < \infty\). If

\[ x_t = \begin{bmatrix} H(L) e_t \\ \vdots \\ \vdots \end{bmatrix}, \quad (\sum_{j=0}^{\infty} |H_j| < \infty), \text{ with } e_t \text{ being WN}(0,\Phi), \text{ then} \]

\[ g_x(z) = H(z)\Phi H(z^{-1})'. \]
5 Different definitions of stock returns

5.1 Simple returns and log returns

The return, defined as a rate, is the profit of holding an asset over a time period (quoted as a proportion of the original asset value). Let the asset price at time $t$ be $P_t$. The simple gross return is defined as $P_t/P_{t-1}$. The simple net return (or simple return) is defined as a percentage change of the price:

$$R_t = P_t/P_{t-1} - 1.$$  

The continuously compounded return (or log return) is defined as

$$r_t = \log(P_t/P_{t-1}).$$

In other words, $r_t$ is the return such that if it is received continuously over the whole period, the value would increase from $P_{t-1}$ to $P_t$. Suppose that the return is received in $m$ sub-periods (between $t-1$ and $t$) (each period receives $r_t/m$), then $P_{t-1}(1 + r_t/m)^m = P_t$. That is,

$$P_t/P_{t-1} = (1 + r_t/m)^m.$$  

Letting $m \to \infty$ (the return is received continuously), we have $P_t/P_{t-1} = \exp(r_t)$, which gives the definition of the continuously compounded return $r_t$. For the simple return $R_t$, compounding takes place only at the end of the period, i.e. $m=1$, (unlike $r_t$).

For example, the S&P 500 index this month (Sep 1, 2016) is 2128, and the index last month is 2170. Then $R_t = -1.935\%$ and $r_t = -1.954\%$ (which are pretty close).

In general, $R_t$ and $r_t$ are close if $R_t$ is near zero. This is because $\log(1 + x) \approx x$ when $x$ is close to zero.  

1 (Almost always $R_t > r_t$).

Figure 1 shows $R_t$ and $r_t$ for annual S&P 500 index data from 1872-2015. When $|R_t|$ is large, there is some non-negligible difference between $R_t$ and $r_t$. Otherwise they are very close.

5.2 Other returns

For the return of holding for $k$ periods, from $t$ to $t+k-1$, the continuously compounded multi-period return (denoted as $r_t[k]$) solves

$$P_{t+k-1}/P_{t-1} = \exp(r_t[k]),$$

by definition. So $r_t[k] = \log(P_{t+k-1}/P_{t-1})$. Note that

$$r_t[k] = r_{t+k-1} + r_{t+k-2} + \ldots + r_t,$$

which is the sum of returns for single periods. This is a convenient property. Multi-period returns using $R_t$ do not enjoy this property.

So far we assume the asset does not pay dividends. If it pays dividends, the simple return is defined as $R_t = (P_t + D_t)/P_{t-1} - 1$. The continuously compounded return (including dividends) is defined as $r_t = \log(P_t + D_t) - \log P_{t-1}$.

Paying dividend can make a difference. The S&P 500 indices at the end of years 2015 and 2014 are $P_{2015} = 2092$ and $P_{2014} = 2105$. The dividend paid in the year of 2015 is 44.14.  

Thus $r_{2015} = \log(2092) - \log(2105) = -0.62\%$ (without dividends) and $r_{2015} = \log(2092 + 44.14) - \log(2105) = 1.47\%$ (including dividends).

---

1 Use the approximation $\log(1 + x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots$ for $|x| < 1$.

2 Thus the dividend yield is 44.14/2092 = 2.11\%
Figure 1: How close are $R_t$ and $r_t$?

Excess return $z_t$ is the asset return over a risk-free return $i_t$:

$$z_t = r_t - i_t.$$  

Treasure bill rates are often used for $i_t$. (e.g. one-month T-bill rate for monthly data, and three-month T-bill rate for quarterly data).
1 The long run variance

Before considering the regression methods, we introduce one important quantity. For a time series \( w_t \) (denoting its mean and variance to be \( \mu \) and \( \sigma^2 \)), its long run variance (LRV) is defined as

\[
\Omega \triangleq \lim_{T \to \infty} \text{Var} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} w_t \right).
\] (1)

Why might we be interested in \( \Omega \)? It gives the asymptotic variance for the sample mean \( \bar{w} = T^{-1} \sum_{t=1}^{T} w_t \) when observations \( w_t \) are correlated. More precisely,

\[
\bar{w} \xrightarrow{d} \mathcal{N}(\mu, \Omega/T), \tag{2}
\]

where \( \xrightarrow{d} \) is read as "is distributed approximately as (when \( T \) is large)". If \( w_t \) is iid (as if generated by random sampling in basic econometrics), then \( \Omega = \sigma^2 \), and (2) thus reduces to the elementary result

\[
\bar{w} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/T). \tag{3}
\]

In a time series setting, (3) is generally incorrect (since it ignores autocovariances of \( w_t \)).\(^1\) We will rely on (2) instead.

An equivalent definition of LRV is

\[
\Omega = \gamma_0 + 2 \sum_{j=1}^{\infty} \gamma_j. \tag{4}
\]

We will justify the equivalence below.

---

\(^1\)...except if \( w_t \) is a white noise.
Denote the autocovariances of $w_t$ as $\{\gamma_0, \gamma_1, \ldots\}$. An expansion shows\(^2\)

\[
T^{-1} \operatorname{Var} \left( \sum_{t=1}^{T} w_t \right) = \frac{1}{T} [T \gamma_0 + 2(T-1)\gamma_1 + 2(T-2)\gamma_2 + \cdots + 2\gamma_{T-1}]
\]

\[
= \gamma_0 + 2 \sum_{j=1}^{T-1} (1 - \frac{j}{T}) \gamma_j.
\]

(5) is an exact formula, meaning it is true for all $T$. When $T \to \infty$,

\[
T^{-1} \operatorname{Var} \left( \sum_{t=1}^{T} w_t \right) = \gamma_0 + 2 \sum_{j=1}^{T-1} \left(1 - \frac{j}{T}\right) \gamma_j
\]

\[
= \gamma_0 + 2 \sum_{j=1}^{T-1} \gamma_j - \frac{2}{T} \sum_{j=1}^{T-1} j \gamma_j
\]

\[
\to \gamma_0 + 2 \sum_{j=1}^{\infty} \gamma_j,
\]

if

\[
T^{-1} \sum_{j=1}^{T-1} j \gamma_j \to 0
\]

(6)

($\gamma_j$ shrinks sufficiently fast). Thus (4) holds.

(6) is true quite broadly. For instance, it holds if (called one-summability)

\[
\sum_{j=1}^{\infty} j |\gamma_j| < \infty.
\]

(7)

We can easily verify that (7) is reasonable for the two models we have seen, MA model and stationary AR model. For MA model, $\sum_{j=1}^{\infty} j |\gamma_j|$ is a sum of finite terms so (7) obviously holds. It is less obvious for AR model (recall what are their autocovariances).

2 How to estimate LRV?

One may start with (4). The most autocovariances we can compute are $\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_{T-1}$ (sample autocovariances of $w_t$). So a natural estimator of $\Omega$ is

\[
\hat{\gamma}_0 + 2 \sum_{j=1}^{T-1} \hat{\gamma}_j.
\]

\(^2\)If (5) does not seem clear, you can verify it by using $T = 2$ and 3. For example,

\[
\begin{align*}
\operatorname{Var}(w_1 + w_2 + w_3) & = 3 \operatorname{Var}(w_1) + 2 \operatorname{Cov}(w_1, w_2) + 2 \operatorname{Cov}(w_2, w_3) + 2 \operatorname{Cov}(w_1, w_3) \\
& = 3\gamma_0 + 2\gamma_1 + 2\gamma_1 + 2\gamma_2
\end{align*}
\]
It is not satisfactory since we know that $\hat{\gamma}_j$’s are reasonably good estimators only when $j$ is small (relatively to $T$).

How about we only use $\hat{\gamma}_j$’s in (8) for small $j$, say $j = 1, 2, \ldots, M$? That is, estimate $\Omega$ by

$$\hat{\gamma}_0 + 2 \sum_{j=1}^M \hat{\gamma}_j.$$  \hfill (9)

This is the idea of truncation, using only $M$ autocovariances (instead of $T - 1$ autocovariances). The rationale behind is that $\gamma_j$ is smaller and smaller (thus is safer to be ignored) as $j$ increases (i.e. covariance between two observations gets smaller when they are farther apart). The estimator (9) itself is not satisfactory either as people realize that (9) may produce a negative number (which is not reasonable for a variance estimator).

A well accepted estimator (called Newey-West estimator) is based on a modification of (9), defined as

$$\hat{\Omega} = \hat{\gamma}_0 + 2 \sum_{j=1}^M (1 - \frac{j}{M+1}) \hat{\gamma}_j.$$  \hfill (10)

It resembles (5) as if we compute $\text{Var}\left(\frac{1}{M} \sum_{t=1}^M w_t\right)$ (i.e. setting $T = M + 1$). The weights applied to each $\hat{\gamma}_j$ in (10) is also called Bartlett weights.

In practice we need to specify the truncation parameter $M$. A simple choice is $M = \lfloor T^{1/3} \rfloor$. Another choice of $M$ is the following

$$M = 1.145 \left( \frac{4T\hat{\alpha}^2}{(1-\hat{\alpha}^2)^2} \right)^{1/3},$$

where $\hat{\alpha}$ is the AR(1) coefficient estimate for $w_t$. If $\hat{\alpha}$ is close to one (i.e. $w_t$ is very persistent), $M$ tends to be large (which makes sense since we need more autocovariances to be included in $\hat{\Omega}$ in this circumstance).

If $T = 400$, and using the simple choice you pick $M = \lfloor 400^{1/3} \rfloor = 7$, then (10) gives you

$$\hat{\Omega} = \hat{\gamma}_0 + 2 \left( \frac{7}{8} \hat{\gamma}_1 + \frac{6}{8} \hat{\gamma}_2 + \frac{5}{8} \hat{\gamma}_3 + \frac{4}{8} \hat{\gamma}_4 + \frac{3}{8} \hat{\gamma}_5 + \frac{2}{8} \hat{\gamma}_6 + \frac{1}{8} \hat{\gamma}_7 \right).$$

3 The regression model

We now consider the usual regression model

$$y_t = x_t' \beta + \epsilon_t,$$  \hfill (11)

where the standard exogeneity assumption is maintained, $E(x_t \epsilon_t) = 0$. The regressor $x_t$ is $K$-dimensional, and the first element of $x_t$ is one.

What is interesting here (in particular, in terms of standard errors) is that we allow the data (for $y$ and $x$) to be serially correlated, i.e. $\text{Cov}(x_t, x_s) \neq 0$ and $E(\epsilon_t \epsilon_s) \neq 0$ for $t \neq s$. If they are independent (and iid), the standard regression toolbox can be used.

We will still use the OLS estimator

$$\hat{\beta}_K = \left( \sum_{t=1}^T x_t x_t' \right)^{-1} \left( \sum_{t=1}^T x_t y_t \right).$$  \hfill (12)
To obtain its standard error, we have the following approximation (see at the end of this chapter for derivation)
\[
\sqrt{T}(\hat{\beta} - \beta) \overset{d}{\to} \mathcal{N}(0, V_{rob}), \quad (13)
\]
where
\[
V_{rob} = \frac{1}{K^2} E \left( x_t x_t' \right)^{-1} \Omega (x_t x_t')^{-1}
\]
and \( \Omega \) is the long-run variance matrix of \( x_t e_t \)
\[
\Omega = \lim_{T \to \infty} \frac{1}{T} E \left( \sum_{t=1}^{T} x_t e_t \right) \left( \sum_{t=1}^{T} x_t' e_t \right).
\]

\( V_{rob} \) is called the robust asymptotic variance. It has a "sandwich" form. The LRV \( \Omega \) is the meat part.

4 Robust standard errors

The non-trivial part of estimating \( V_{rob} \) is to deal with \( \Omega \). It is done by using (10) (with an extension to the vector case). Define
\[
\hat{\Omega} = \hat{\Gamma}_0 + \sum_{j=1}^{M} (1 - \frac{j}{M+1}) \left( \hat{\Gamma}_j + \hat{\Gamma}_j' \right), \quad (14)
\]
where \( \hat{\Gamma}_j \) is the sample covariance matrix of \( x_t e_t \). Then \( V_{rob} \) is estimated by
\[
\hat{V}_{rob} = (T^{-1} \sum_{t=1}^{T} x_t x_t')^{-1} \hat{\Omega} (T^{-1} \sum_{t=1}^{T} x_t x_t')^{-1}.
\]

It is called the Newey-West asymptotic variance estimator, \( (\hat{V}_{rob, kk}/T)^{1/2} \) is the Newey-West standard error for \( \beta_k \), where \( \hat{V}_{rob, kk} \) is the \((k,k)\)-element of \( \hat{V}_{rob} \).

The t-statistic for a particular coefficient \( \beta_j \) has \( \mathcal{N}(0, 1) \) distribution:
\[
\sqrt{T}\hat{V}_{rob, kk}^{1/2} (\hat{\beta}_k - \beta_k) \overset{d}{\to} \mathcal{N}(0, 1).
\]

A 95% confidence interval for \( \beta_k \) is \[ \hat{\beta}_k - 1.96 \hat{V}_{rob, kk}^{1/2} \sqrt{T}, \hat{\beta}_k + 1.96 \hat{V}_{rob, kk}^{1/2} / \sqrt{T} \].

5 A parametric VAR approach

The Newey-West method of estimating the LRV \( \Omega \) is nonparametric, meaning that it does not rely on any particular (parameterized) time series model. A more traditional way is based on a time series model of \( x_t e_t \) to calculate the model-implied LRV.

\[ \hat{\Gamma}_j \] is the estimator of \( \Gamma_j \). Using the definition in Topic 1, they are
\[
\begin{align*}
\Gamma_j &= \text{Cov}(x_t e_t, x_{t-j} e_{t-j}) = E x_t e_t (x_{t-j} e_{t-j})' = E x_t' x_{t-j} e_{t-j}, \\
\hat{\Gamma}_j &= T^{-1} \sum_{t=j+1}^{T} x_t x_{t-j} e_t e_{t-j},
\end{align*}
\]
since \( E x_t e_t = 0 \) (exogeneity).
A natural model (to capture the serial correlation in \(w_t = x_t e_t\)) is VAR(1):

\[
\begin{align*}
\begin{bmatrix} w_{t} \\ K \times 1 
\end{bmatrix} &= \begin{bmatrix} \Phi \\ K \times K \end{bmatrix} \begin{bmatrix} w_{t-1} \\ K \times 1 
\end{bmatrix} + \begin{bmatrix} v_t \end{bmatrix}, \\
\text{(15)}
\end{align*}
\]

where \(v_t\) is WN(0,\(\Sigma_v\)). Then \(\Omega = (1 - \Phi)^{-1} \Sigma_v (1 - \Phi^\prime)^{-1}\). Thus

\[
\hat{\Omega} = (1 - \Phi)^{-1} \hat{\Sigma}_v (1 - \Phi^\prime)^{-1},
\]

where \(\hat{\Sigma}_v = T^{-1} \sum_{t=1}^{T} \hat{v}_t \hat{v}_t^\prime\) with \(\hat{v}_t\) being the residuals in the regression (15).

6 An alternative

An alternative (more recent) way of estimating \(\Omega\) is to use \(M = T - 1\) in (14), i.e. using no truncation. Be careful not to use \(M = T\), since \(\Gamma_T\) is undefined (\(\Gamma_{T-1}\) is the largest autocovariance the data can estimate). This "estimator" is defined as

\[
\begin{align*}
\hat{\Omega}_{M=T-1} &= \hat{\Gamma}_0 + \sum_{j=1}^{T-1} (1 - j\frac{T}{}) \hat{\Gamma}_j + \hat{\Gamma}_j'. \\
\text{(16)}
\end{align*}
\]

An identical way to compute \(\hat{\Omega}_{M=T-1}\) is to use (22) (at the end of the chapter).

While \(\hat{\Omega} \rightarrow \Omega\) when truncation is used, \(\hat{\Omega}_{M=T-1} \rightarrow \Omega\) since more autocovariance estimators introduce too much noise (for \(\hat{\Omega}_{M=T-1}\) to converge to \(\Omega\)). This is consistent with our previous argument that \(\hat{\Gamma}_j\) is not of good quality when \(j\) is large.

The asymptotic variance estimator of \(\beta\) is then

\[
\hat{V}_{rob,M=T-1} = (T^{-1} \sum_{t=1}^{T} x_t x_t^\prime)^{-1} \hat{\Omega}_{M=T-1} (T^{-1} \sum_{t=1}^{T} x_t x_t^\prime)^{-1}.
\]

The t-statistic for a particular coefficient \(\beta_k\) has a limit distribution (called \(D\), which is not standard normal):

\[
\sqrt{T} \hat{V}_{rob,M=T-1,kk}^{1/2} (\hat{\beta}_k - \beta_k) \xrightarrow{d} D.
\]

7 Predictive regression example

We examine the predictive relationship between index returns and dividend price ratios. The data are quarterly from 1926:1-2015:4, having 360 observations (from Welch and Goyal); see Figure 1 for time plots. We OLS-fit the following two equations

\[
\begin{align*}
\text{returns}_t &= 0.136 + 0.0316 \cdot (d/p)_{t-1} + \text{error}_t, \\
\text{(18)} \\
(d/p)_t &= -0.097 + 0.972 \cdot (d/p)_{t-1} + \text{error}_t. \\
\text{(19)}
\end{align*}
\]

\text{This follows from the Lemma in Topic 2 on the autocovariance generating function of } w_t.
Here we use \((d/p)_{t-1}\) as regressor (instead of \((d/p)_t\)), so (18) is called the *predictive regression*. Regression (19) is an autoregression (AR(1)). The estimated slope in (19) (which is very close to one) shows that \((d/p)_t\) is very persistent.

The standard errors are reported in the following two tables. Robust standard errors are generally larger.

![Time plots](image)

**Figure 1: Time plots**

An econometrics haiku:

"T-stat looks too good
Try robust standard errors-
Significance gone."

<table>
<thead>
<tr>
<th>Parameter</th>
<th>usual</th>
<th>VAR(1)</th>
<th>NW</th>
<th>KVB</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\beta}) = 0.0316</td>
<td>0.01278</td>
<td>0.01760</td>
<td>0.01765</td>
<td>0.00759</td>
</tr>
<tr>
<td>se ((\beta=0))</td>
<td>2.468</td>
<td>1.793</td>
<td>1.793</td>
<td>4.156</td>
</tr>
<tr>
<td>p-value ((\beta=0))</td>
<td>1.4%</td>
<td>7.3%</td>
<td>7.3%</td>
<td>[5%, 10%]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>usual</th>
<th>VAR(1)</th>
<th>NW</th>
<th>KVB</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\rho}) = 0.972</td>
<td>0.0125</td>
<td>0.0159</td>
<td>0.0160</td>
<td>0.0098</td>
</tr>
<tr>
<td>se |90% Conf. Int.|</td>
<td>[0.951, 0.993]</td>
<td>[0.946, 0.998]</td>
<td>[0.946, 0.998]</td>
<td>[0.935, 1.009]</td>
</tr>
<tr>
<td>length</td>
<td>0.041</td>
<td>0.052</td>
<td>0.053</td>
<td>0.074</td>
</tr>
</tbody>
</table>
8 Taylor rule example

In this example we estimate the Taylor rule in macroeconomics. Taylor rule is a monetary-policy rule that stipulates how much the central bank should change the nominal interest rate in response to changes in inflation, output, or other economic conditions. It is formulated in equation (20). Figure 2 shows the time series for the dependent variable and two regressors (data are from Ang, Boivin, Dong and Loo-Kung, 2011).

ACF plots in Figure 3 shows the output gap series is less persistent than the other two. ACF plots (Figure 4) for the series $x_t e_t$ (3-dim) show (cross-)serial correlation that can not be overlooked. These (cross-)serial correlations are much reduced by a simple VAR(1) model (Figure 5 shows ACFs for residuals).

The NW standard errors are much larger (about 3 or 4 times) larger than the usual ones, thus yielding wider confidence intervals. The KVB intervals are slightly tighter than the NW ones.

$$\text{interest}_t - \text{rate}_t = 0.005 + 0.025 \cdot \text{output}_t \cdot \text{gap}_t + 0.906 \cdot \text{inflation}_t + \text{error}_t$$  \hspace{1cm} (20)

<table>
<thead>
<tr>
<th></th>
<th>usual</th>
<th>NW</th>
<th>KVB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.025</td>
<td>0.059</td>
<td>0.154</td>
</tr>
<tr>
<td>se</td>
<td></td>
<td>0.196</td>
<td>0.506</td>
</tr>
<tr>
<td>90% Conf. Int.</td>
<td>[-0.073, 0.123]</td>
<td>[-0.228, 0.278]</td>
<td>[-0.218, 0.268]</td>
</tr>
<tr>
<td>length</td>
<td></td>
<td>0.196</td>
<td>0.506</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>usual</th>
<th>NW</th>
<th>KVB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_2$</td>
<td>0.906</td>
<td>0.063</td>
<td>0.226</td>
</tr>
<tr>
<td>se</td>
<td></td>
<td>0.209</td>
<td>0.744</td>
</tr>
<tr>
<td>90% Conf. Int.</td>
<td>[0.801, 1.010]</td>
<td>[0.534, 1.278]</td>
<td>[0.651, 1.161]</td>
</tr>
<tr>
<td>length</td>
<td></td>
<td>0.209</td>
<td>0.744</td>
</tr>
</tbody>
</table>
Figure 3: Taylor rule: ACF of $y_t$ and $x_{1t}, x_{2t}$. (AR coefficients are 0.937, 0.931, 0.982 respectively.)

Figure 4: Taylor rule: ACF of $x_t e_t$. 

9 A few steps to show (17)

Let $v_t = x_{t-1}e_t$. A key step to show (17) is to show

$$
\hat{\Omega}_{M=T-1} = T^{-1} \sum_{i=1}^{T} \sum_{j=1}^{T} (1 - \frac{|i-j|}{T}) v_i v'_j \tag{21}
$$

$$
= T^{-2} \sum_{i=1}^{T} \sum_{j=1}^{T} S_i S'_j, \tag{22}
$$

where $S_i = \sum_{t=1}^{i} v_t$. We here only show (21).

We can write

$$
\hat{\Omega}_{M=T-1} = T^{-1} \sum_{i=1}^{T} \sum_{j=1}^{T} (1 - \frac{|i-j|}{T}) v_i v'_j
$$

$$
= T^{-1} \sum_{k=0}^{T-1} \sum_{j=1}^{T-k} (1 - \frac{|k|}{T}) v_j v'_{j+k} + T^{-1} \sum_{k=-(T-1)}^{1} \sum_{j=T+k+1}^{T} (1 - \frac{|k|}{T}) v_j v'_{j+k}
$$

$$
\underbrace{\sum_{k=0}^{T-1} (1 - \frac{|k|}{T}) \hat{\Gamma}_{-k}} + \underbrace{\sum_{k=-(T-1)}^{1} (1 - \frac{|k|}{T}) \hat{\Gamma}_{-k}}
$$

$$
= T^{-1} \sum_{k=-(T-1)}^{1} (1 - \frac{|k|}{T}) \hat{\Gamma}_{k},
$$

Figure 5: Taylor rule: ACF of VAR(1)- residuals of $x_t e_t$. 
which shows that (16) and (21) are equivalent expressions. Proofs of (22) and (17) can be found in Kiefer and Vogelsang (2002).

10 The derivation of (13)

From (11) and (12),

\[ \hat{\beta} - \beta = \left( \frac{T}{T} \sum_{t=1}^{T} x_t x_t' \right)^{-1} \left( \frac{T}{T} \sum_{t=1}^{T} x_t e_t \right). \]

Thus

\[ \sqrt{T} (\hat{\beta} - \beta) = \left( \frac{1}{T} \sum_{t=1}^{T} x_t x_t' \right)^{-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t e_t \right) \]

\[ \overset{d}{\to} E(x_t x_t')^{-1} \cdot T \left( 0, \text{Var}\left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t e_t \right) \right) \]

\[ = E(x_t x_t')^{-1} \cdot T(0, \Omega) \]

\[ = T(0, E(x_t x_t')^{-1} \Omega E(x_t x_t')^{-1}). \]