

Lecture notes

Undergraduate econometrics

- [Review of calculus](#)
- [Importing data in EasyReg International](#)
- [The two-variable linear regression model](#)
- [Multivariate linear regression](#)
- [Specification of econometric models](#)
- [The Logit model: estimation, testing and interpretation](#)
- [Comparison of the Probit and Logit distributions](#)
- [Modeling fractions](#)
- [Forecasting](#)

Graduate econometrics

Cross-section and panel data

- [The classical linear regression model](#)
- [Multicollinearity](#)
- [Tests of normality of regression errors](#)
- [Method of moments](#)
- [The uniform weak law of large numbers and the consistency of M-estimators](#)

Time series econometrics

- [The Wold decomposition](#)
- [ARMA models](#)
- [Information criteria and model selection](#)
- [Forecasting](#)
- [Vector time series and innovation response analysis](#)
- [Unit roots](#)
- [Spurious regression](#)
- [Cointegration](#)
- [Weak convergence to the matrix stochastic integral \$\int B dB'\$ in the Gaussian case, with application to likelihood-based cointegration analysis](#)

Semi-nonparametric econometrics

- [Introduction to Hilbert spaces](#)
- [Hilbert space theory](#) and its applications to semi-nonparametric modeling and inference
- [Semi-nonparametric identification of the right censored mixed proportional hazard model](#)

Consistent model specification tests

- [The integrated conditional moment test \(Seminar slides\)](#)
- [Review of the integrated conditional moment test](#) and its implementation in [EasyReg International](#)

Miscellaneous

- [The inverse of a partitioned matrix](#)
- [The matrix norm \$\|A\|\$](#)



[Back to my home \(page\)](#)

REVIEW OF CALCULUS

Herman J. Bierens

Pennsylvania State University

(January 28, 2004)

1. *Summation*

Let x_1, x_2, \dots, x_n be a sequence of numbers. The sum of these numbers is usually denoted by

$$x_1 + x_2 + \dots + x_n = \sum_{j=1}^n x_j, \quad \text{or} \quad x_1 + x_2 + \dots + x_n = \sum_{j=1}^n x_j.$$

The index “ j ” may be replaced by any other variable name. Thus, $x_1 + x_2 + \dots + x_n = \sum_{i=1}^n x_i$ as well.

Next, let y_1, y_2, \dots, y_m be another sequence of numbers. Then we can write

$$\begin{aligned} (x_1 + x_2 + \dots + x_n)(y_1 + y_2 + \dots + y_m) &= (x_1 + x_2 + \dots + x_n) \left(\sum_{j=1}^m y_j \right) \\ &= x_1 \left(\sum_{j=1}^m y_j \right) + x_2 \left(\sum_{j=1}^m y_j \right) + \dots + x_n \left(\sum_{j=1}^m y_j \right) \\ &= \sum_{j=1}^m x_1 y_j + \sum_{j=1}^m x_2 y_j + \dots + \sum_{j=1}^m x_n y_j = \sum_{i=1}^n \sum_{j=1}^m x_i y_j. \end{aligned} \tag{1}$$

Of course, the order of the summation does not matter:

$$\sum_{i=1}^n \sum_{j=1}^m x_i y_j = \sum_{j=1}^m \sum_{i=1}^n x_i y_j.$$

Moreover, replacing y_1, y_2, \dots, y_m in (1) by x_1, x_2, \dots, x_n it follows that

$$\left(\sum_{j=1}^n x_j \right)^2 = (x_1 + x_2 + \dots + x_n)^2 = \sum_{i=1}^n \sum_{j=1}^n x_i x_j. \tag{2}$$

Note that the reason for using different indices i and j is that the summation in (2) is done in two steps. First, for each index i we sum up $x_i x_j$ for $j = 1, \dots, n$, and then we sum up $x_i \sum_{j=1}^n x_j$ for $i =$

$1, \dots, n$. The same applies to (1).

The average of the numbers x_1, x_2, \dots, x_n is usually denoted by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{j=1}^n x_j.$$

In particular, we have

$$\begin{aligned} \sum_{j=1}^n (x_j - \bar{x}) &= \sum_{j=1}^n x_j - \sum_{j=1}^n \bar{x} = \sum_{j=1}^n x_j - n\bar{x} \\ &= \sum_{j=1}^n x_j - n \cdot \frac{1}{n} \sum_{j=1}^n x_j = \sum_{j=1}^n x_j - \sum_{j=1}^n x_j = 0. \end{aligned} \tag{3}$$

This easy result will prove useful in regression analysis.

2. Special sums

Consider the sum $1 + 2 + 3 + \dots + n = \sum_{j=1}^n j$. There is an easy formula for this sum:

$$\sum_{j=1}^n j = n(n+1)/2. \tag{4}$$

Rather than memorizing this formula, it is better to memorize how it is derived. Since the order of the summation does not matter, we can write $\sum_{j=1}^n j$ in two ways:

$$\begin{aligned} \sum_{j=1}^n j &= 1 + 2 + 3 + \dots + n, \text{ and} \\ \sum_{j=1}^n j &= n + n-1 + n-2 + \dots + 1 \end{aligned} \tag{5}$$

Adding up the left and right-hand sides of the two equations in (5) yields

$$2 \sum_{j=1}^n j = (n+1) + (n+1) + (n+1) + \dots + (n+1) = n(n+1). \tag{6}$$

Dividing (6) by 2, the result (4) follows.

Next, consider the sum

$$1 + x + x^2 + x^3 + \dots + x^n = x^0 + x^1 + x^2 + x^3 + \dots + x^n = \sum_{j=0}^n x^j,$$

where x is any number. The formula for this sum is:

$$\sum_{j=0}^n x^j = \frac{1 - x^{n+1}}{1 - x}. \quad (7)$$

Again, it is easier to memorize how this formula is derived than to memorize the formula itself. First, observe that $\sum_{j=0}^n x^j = 1 + \sum_{j=1}^n x^j$. Next, replace the index j by $i-1$. Then we can write:

$$\begin{aligned} \sum_{j=0}^n x^j &= 1 + \sum_{j=1}^n x^j = 1 + \sum_{i=0}^{n-1} x^{i+1} = 1 + x \cdot \sum_{i=0}^{n-1} x^i = 1 + x \left(\sum_{j=0}^{n-1} x^j \right) \\ &= 1 + x \left(\sum_{j=0}^n x^j - x^n \right) = 1 - x^{n+1} + x \cdot \sum_{j=0}^n x^j. \end{aligned} \quad (8)$$

Solving equation (8) for $\sum_{j=0}^n x^j$ yields the formula (7).

3. Limits

The limit of a sequence y_n , $n = 1, 2, 3, \dots$, of numbers is a number y such that y_n approaches y if n increases to infinity. The formal definition of a limit is:

A sequence y_n , $n = 1, 2, 3, \dots$, of numbers has a limit y , say, denoted by $y = \lim_{n \rightarrow \infty} y_n$, if and only if for every number $\epsilon > 0$ there exists an index n_0 (which may depend on ϵ) such that $|y_n - y| < \epsilon$ for all $n \geq n_0$.

For example, if $y_n = 1/n$ then $\lim_{n \rightarrow \infty} y_n = 0$. To see this, pick an arbitrary $\epsilon > 0$. Then $|y_n| = 1/n < \epsilon$ if $n > 1/\epsilon$. Thus in this case the index n_0 is the smallest natural number $\geq 1/\epsilon$. Another example is the case $y_n = x^n$, where $|x| < 1$. Given such a number x and an arbitrary number $\epsilon > 0$, it is possible to find a positive integer n_0 such that $|x|^{n_0} < \epsilon$, and then $|y_n| = |x^n| = |x|^n < \epsilon$ for all $n \geq n_0$. Hence

$$\lim_{n \rightarrow \infty} x^n = 0 \text{ if } |x| < 1. \quad (9)$$

Consequently, it follows from (7) and (9) that:

$$\lim_{n \rightarrow \infty} \sum_{j=0}^n x^j = \frac{1 - \lim_{n \rightarrow \infty} x^{n+1}}{1 - x} = \frac{1 - x \cdot \lim_{n \rightarrow \infty} x^n}{1 - x} = \frac{1}{1 - x} \text{ if } |x| < 1. \quad (10)$$

The left-hand side of this equation is usually denoted by $\sum_{j=0}^{\infty} x^j$. Thus,

$$\sum_{j=0}^{\infty} x^j = \frac{1}{1 - x} \text{ if } |x| < 1. \quad (11)$$

More generally, if for a sequence $x_0, x_1, x_2, \dots, x_n, \dots$ of numbers, $\lim_{n \rightarrow \infty} \sum_{j=0}^n x_j$ exists, then this limit is denoted by $\sum_{j=0}^{\infty} x_j$. Moreover, it is left as an exercise to verify that

$$\text{If } \lim_{n \rightarrow \infty} \sum_{j=0}^n x_j \text{ exists then } \lim_{m \rightarrow \infty} \sum_{j=m}^{\infty} x_j = 0. \quad (12)$$

Not every sequence has a limit, though. For example, if $y_n = (-1)^n$ then the limit does not exist, because then $|y_n| = |-1|^n = 1^n = 1$ for $n = 1, 2, 3, \dots$, so that for $0 < \epsilon < 1$ and all positive natural numbers n , $|y_n| > \epsilon$.

4. Functions

4.1 Linear and quadratic functions and their roots

A real function $f(x)$ assigns a real number $y = f(x)$ to x . Important classes of functions are the linear functions:

$$f(x) = \alpha + \beta \cdot x \quad (13)$$

and the quadratic functions:

$$f(x) = \alpha + \beta \cdot x + \gamma \cdot x^2, \quad (14)$$

where α, β and γ are given constants.

The roots of a function $f(x)$ are the values of x for which $f(x) = 0$. In the linear case (13) there is only one root, namely $x = -\alpha/\beta$, provided that $\beta \neq 0$. In the quadratic case (14) the number of roots is either 0, 1 or 2, depending on what α, β and γ are. In order to derive these roots, observe

first that $(x + a)^2 = x^2 + 2ax + a^2$. Next, assume that $\gamma \neq 0$. Then

$$\begin{aligned} a + \beta x + \gamma x^2 = 0 &\Rightarrow x^2 + \frac{\beta}{\gamma}x = -\frac{a}{\gamma} \Rightarrow x^2 + 2 \cdot \frac{\beta}{2\gamma}x = -\frac{a}{\gamma} \\ &\Rightarrow x^2 + 2 \cdot \frac{\beta}{2\gamma}x + \left(\frac{\beta}{2\gamma}\right)^2 = -\frac{a}{\gamma} + \left(\frac{\beta}{2\gamma}\right)^2 \\ &\Rightarrow \left(x + \frac{\beta}{2\gamma}\right)^2 = \frac{\beta^2 - 4\alpha\gamma}{4\gamma^2}. \end{aligned} \quad (15)$$

If $\beta^2 - 4\alpha\gamma > 0$ then the last equality in (15) implies that the quadratic function (14) has two roots:

$$x_1 = \frac{-\beta - \sqrt{\beta^2 - 4\alpha\gamma}}{2\gamma}, \quad x_2 = \frac{-\beta + \sqrt{\beta^2 - 4\alpha\gamma}}{2\gamma}. \quad (16)$$

If $\beta^2 - 4\alpha = 0$ then the quadratic function (14) has only one root:

$$x = \frac{-\beta}{2\gamma}, \quad (17)$$

and if $\beta^2 - 4\alpha < 0$ the quadratic function (14) has no real roots.¹

4.2 The $\exp(\cdot)$ and $\ln(\cdot)$ functions

The exponential function $\exp(x)$ is defined as

$$\exp(x) = e^x, \text{ where } e \approx 2.7182818285 \quad (18)$$

It can be shown² that

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}, \quad (19)$$

¹ However, the roots are then complex-valued:

$$x_1 = \frac{-\beta - i\sqrt{4\alpha\gamma - \beta^2}}{2\gamma}, \quad x_2 = \frac{-\beta + i\sqrt{4\alpha\gamma - \beta^2}}{2\gamma}, \text{ where } i = \sqrt{-1}.$$

² But that requires advanced calculus!

where $k!$ (read: k factorial) is the product of the natural numbers 1 to k : $k! = 1 \times 2 \times 3 \times \dots \times k$, and $0!$ is defined as 1: $0! = 1$.

The numerical value of the number e in (18) is only an approximation, though. The true number e has an infinite number of decimal digits without a repeating pattern, and is therefore irrational. We can only express e exactly as a limit:

$$e = \sum_{k=0}^{\infty} \frac{1}{k!} = \lim_{n \rightarrow \infty} \sum_{k=0}^n \frac{1}{k!}.$$

The natural logarithm is the inverse of the exponential function:

The natural logarithm, $\ln(x)$, is for each $x > 0$ a number y such that $\exp(y) = x$.

Since $\exp(y) > 0$, the function $\ln(x)$ is only defined for $x > 0$. Important properties of the function $\ln(x)$ are:

$$\begin{aligned} \ln(x) &< 0 \text{ for } 0 < x < 1, \\ \ln(x) &= 0 \text{ for } x = 1, \\ \ln(x) &> 0 \text{ for } x > 1, \\ \ln(x \cdot y) &= \ln(x) + \ln(y) \text{ for } x > 0 \text{ and } y > 0, \\ \ln(x/y) &= \ln(x) - \ln(y) \text{ for } x > 0 \text{ and } y > 0, \\ \ln(x^a) &= a \cdot \ln(x) \text{ for } x > 0 \text{ and any } a. \end{aligned} \tag{20}$$

It is left as an exercise to verify these properties from the definition of $\ln(x)$.

4.3 Continuity

Continuity of a function $f(x)$ in a point x_0 can be defined in two (equivalent) ways. The first way is via limits:

A function $f(x)$ is continuous in x_0 if and only if $f(x_0) = \lim_{n \rightarrow \infty} f(x_0 + 1/n)$ and $f(x_0) = \lim_{n \rightarrow \infty} f(x_0 - 1/n)$.

This definition is equivalent to the following official definition:

A function $f(x)$ is continuous in x_0 if and only if for each $\epsilon > 0$ there exists a $\delta > 0$ (possibly depending on x_0) such that $|f(x) - f(x_0)| < \epsilon$ if $|x - x_0| < \delta$.

It is not hard to show that the second definition implies the first one. The proof that the other way around is also true is harder and therefore omitted.

The second definition gives rise to another definition of a limit:

$\lim_{x \rightarrow x_0} f(x) = y$ if and only if for each $\epsilon > 0$ there exists a $\delta > 0$ such that $|f(x) - y| < \epsilon$ if $|x - x_0| < \delta$.

Thus, $f(x)$ is continuous in x_0 if $\lim_{x \rightarrow x_0} f(x) = f(x_0)$.

For example, the linear function (13) is continuous in all x , and so is the quadratic function (14). The latter can be shown as follows. Let x_0 be arbitrary and fixed, and let $|x - x_0| < 1$. Then

$$\begin{aligned} |f(x) - f(x_0)| &= |\beta(x - x_0) + \gamma(x^2 - x_0^2)| = |\beta(x - x_0) + \gamma(x - x_0)(x + x_0)| \\ &= |\beta(x - x_0) + \gamma(x - x_0)^2 + 2x_0\gamma(x - x_0)| \\ &\leq |\beta||x - x_0| + |\gamma||x - x_0|^2 + 2|x_0\gamma||x - x_0| < (|\beta| + |\gamma| + 2|x_0\gamma|)|x - x_0|, \end{aligned} \tag{21}$$

where the last inequality in (21) follows from the fact that $|x - x_0|^2 < |x - x_0|$ if $|x - x_0| < 1$. Next, choose an arbitrary $\epsilon > 0$, and let $\delta = \min[1, \epsilon/(|\beta| + |\gamma| + 2|x_0\gamma|)]$. Then it follows from (21) that $|f(x) - f(x_0)| < \epsilon$ if $|x - x_0| < \delta$.

Also the exponential function (18) is continuous in all x , and the $\ln(x)$ function is continuous in all $x > 0$.

5. Derivatives

5.1 What is a derivative?

The derivative of a function $f(x)$ is denoted by $f'(x)$, and is defined by

$$f'(x) = \lim_{\delta \rightarrow 0} \frac{f(x + \delta) - f(x)}{\delta}. \quad (22)$$

An alternative notation for a derivative is $df(x)/dx$. In order for a derivative to exist, the limit in (22) must exist and be the same regardless whether $\delta > 0$ (so that $\delta \downarrow 0$) or $\delta < 0$ (so that $\delta \uparrow 0$). If so, the function involved is said to be *differentiable* in x .

The derivation of $f'(x)$ is illustrated in the following figure.

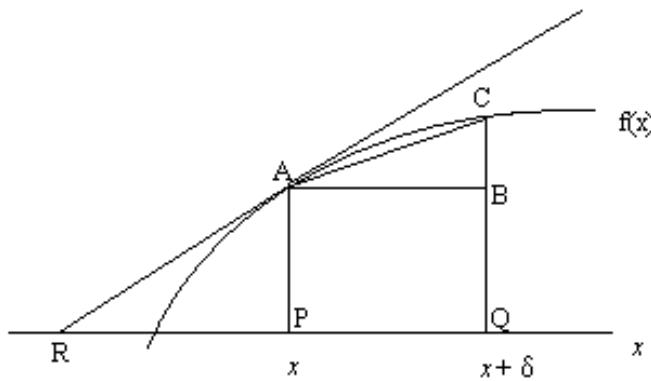


Figure 1: Derivative

The curved line in Figure 1 represents the function $f(x)$. The length of the line piece between the points P and A, $\overline{P \rightarrow A}$ say, is the value of the function $f(x)$ in x at point P, and $\overline{Q \rightarrow C}$ is equal to $f(x + \delta)$. Moreover, $\overline{P \rightarrow Q} = \overline{A \rightarrow B} = \delta$. Thus,

$$\frac{f(x + \delta) - f(x)}{\delta} = \frac{\overline{C \rightarrow B}}{\overline{A \rightarrow B}},$$

which is the tangents³ of the angle between the legs $A \rightarrow B$ and $A \rightarrow C$ of the triangle ABC. Now if we let $\delta \rightarrow 0$ then this angle approaches the angle of the line through the points R and A with the

³

The tangents of an angle φ is defined as: $\tan(\varphi) = \sin(\varphi)/\cos(\varphi)$.

horizontal axis. This line is called the *tangent line* of $f(x)$ in point A. Thus, the derivative of $f(x)$ in x at point P is:

$$f'(x) = \frac{\overline{A \rightarrow P}}{\overline{R \rightarrow P}}.$$

This ratio is called the *slope* of the function $f(x)$ in x at point P.

The quadratic function (14) is differentiable in every x :

$$f(x) = \alpha + \beta \cdot x + \gamma \cdot x^2 \Rightarrow f'(x) = \beta + 2\gamma \cdot x. \quad (23)$$

To see this, observe that in the case (14),

$$\begin{aligned} \lim_{\delta \rightarrow 0} \frac{f(x + \delta) - f(x)}{\delta} &= \lim_{\delta \rightarrow 0} \frac{\beta \cdot \delta + \gamma \cdot (x + \delta)^2 - \gamma \cdot x^2}{\delta} \\ &= \lim_{\delta \rightarrow 0} \frac{\beta \cdot \delta + 2\gamma \cdot \delta x + \gamma \delta^2}{\delta} = \lim_{\delta \rightarrow 0} (\beta + 2\gamma x + \gamma \delta) = \beta + 2\gamma x. \end{aligned} \quad (24)$$

Of course, the derivative of the linear function (13) follows from (23) by setting $\gamma = 0$.

The exponential function has itself as derivative:

$$f(x) = e^x \Rightarrow f'(x) = e^x, \quad (25)$$

because it follows from (19) that

$$\begin{aligned} \lim_{\delta \rightarrow 0} \frac{f(x+\delta) - f(x)}{\delta} &= \lim_{\delta \rightarrow 0} \frac{e^{x+\delta} - e^x}{\delta} = e^x \cdot \lim_{\delta \rightarrow 0} \frac{e^\delta - 1}{\delta} = e^x \cdot \lim_{\delta \rightarrow 0} \frac{\sum_{k=0}^{\infty} \delta^k / k! - 1}{\delta} \\ &= e^x \cdot \lim_{\delta \rightarrow 0} \frac{1 + \sum_{k=1}^{\infty} \delta^k / k! - 1}{\delta} = e^x \cdot \lim_{\delta \rightarrow 0} \sum_{k=1}^{\infty} \delta^{k-1} / k! = e^x \cdot \lim_{\delta \rightarrow 0} \sum_{k=0}^{\infty} \delta^k / (k+1)! \\ &= e^x \cdot \left(1 + \lim_{\delta \rightarrow 0} \sum_{k=1}^{\infty} \delta^k / (k+1)! \right) = e^x. \end{aligned} \quad (26)$$

Moreover, the derivative of $\ln(x)$ is $1/x$:

$$f(x) = \ln(x) \Rightarrow f'(x) = 1/x, \text{ for all } x > 0, \quad (27)$$

To prove (27), let $y = \ln(x)$. Then $x = \exp(y)$, hence it follows from (25) that

$$dx/dy = d\exp(y)/dy = \exp(y) = x. \quad (28)$$

Taking the reciprocal of (28) it follows that

$$dy/dx = d\ln(x)/dx = 1/x. \quad (29)$$

It should be noted that there are many function that are not differentiable in some points, even functions that are continuous in every point. For example, the absolute value function $f(x) = |x|$ is continuous in every x , but is not differentiable in $x = 0$.

5.2 The chain rule

Next, consider two differentiable functions $f(x)$ and $g(x)$, and let $h(x) = f(g(x))$. The question is: Given the derivatives $f'(x)$ and $g'(x)$, what is the derivative $h'(x) = df(g(x))/dx$? The answer is the *chain rule*:

$$h'(x) = \frac{df(g(x))}{dx} = \frac{df(g(x))}{dg(x)} \times \frac{dg(x)}{dx} = f'(g(x)).g'(x). \quad (30)$$

For example, let $f(x) = \ln(x)$ and $g(x) = x^2 + 1$. Then $h(x) = f(g(x)) = \ln(x^2 + 1)$, $f'(x) = 1/x$, and $g'(x) = 2x$, hence it follows from (30) that

$$h'(x) = \frac{2x}{x^2 + 1}.$$

The proof of (30) is left as an exercise.

6. Integrals

6.1 What is an integral?

The concept of an integral is illustrated in Figure 2:



Figure 2: The integral $\int_a^b f(x)dx = \text{grey area.}$

The integral of a function $f(x)$ over an interval $[a,b]$, denoted by $\int_a^b f(x)dx$, is the grey area in Figure 2. This area can be (roughly) approximated by a sum⁴ of rectangle areas with height $f(x)$ and width dx , as illustrated in Figure 3:

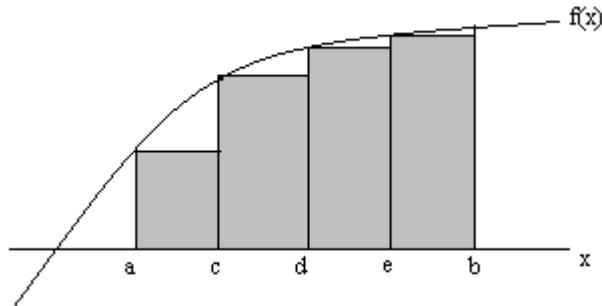


Figure 3: Approximation of $\int_a^b f(x)dx$

The first rectangle has area $f(a)\times(c-a)$, the second has area $f(c)\times(d-c)$, the third has area

⁴ Therefore, the integral symbol \int is actually a stylized version of the letter S in "Sum".

$f(d) \times (e-d)$, and the last one has area $f(e) \times (d-b)$. Assuming that the intervals $[a,c]$, $[c,d]$, $[d,e]$ and $[e,b]$ have equal length $(b-a)/4 = dx$, say, so that

$$c = a + (b-a)/4, \quad d = a + 2(b-a)/4, \quad e = 3(b-a)/4,$$

the total grey area in Figure 3 is:

$$\sum_{k=0}^{n-1} f(a + k(b-a)/n) \times (b-a)/n = \sum_{\substack{x=a+k(b-a)/n \\ dx=(b-a)/n \\ k=0,1,\dots,n-1}} f(x)dx, \text{ for } n = 4.$$

Letting $n \rightarrow \infty$, the limit of this sum is the grey area in Figure 2: $\int_a^b f(x)dx$.

Note that if $f(x) < 0$ on $[a,b]$ then $\int_a^b f(x)dx < 0$, as follows by flipping the pictures in Figures 2 and 3 vertically 180 degrees. Therefore, the integral of $f(x)$ on $[a,b]$ in Figure 4 below is the *difference* of the darker grey area above the horizontal axis, and the lighter grey area below the horizontal axis.

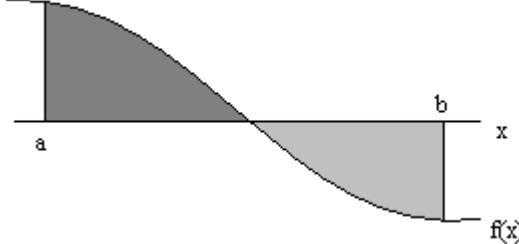


Figure 4: Integral $\int_a^b f(x)dx$ if $f(x)$ flips sign

Moreover,

$$\int_a^b f(x)dx = -\int_b^a f(x)dx. \quad (31)$$

because the latter integral should be interpreted as the grey area in Figure 1 looking from the *back* (the negative side) of the picture: If we flip Figure 1 horizontally 180 degrees, then point b will be at the left of point a , and the new viewpoint is now *behind* Figure 1:

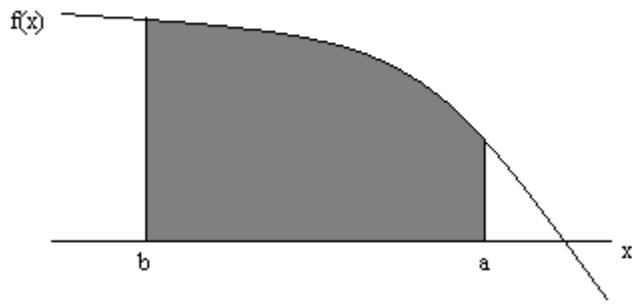


Figure 5: Back side of Figure 1

6.2 Derivative of an integral

Let point e in Figure 3 be $b - \delta$, where $\delta > 0$ is very small. Then the approximation

$$\int_{b-\delta}^b f(x)dx \approx f(b-\delta)\delta \approx f(b)\delta$$

will be close, and so will be

$$\frac{\int_{b-\delta}^b f(x)dx}{\delta} \approx f(b).$$

Therefore,

$$\lim_{\delta \downarrow 0} \frac{\int_{b-\delta}^b f(x)dx}{\delta} = f(b).$$

Similarly it follows that

$$\lim_{\delta \downarrow 0} \frac{\int_b^{b+\delta} f(x)dx}{\delta} = f(b). \quad (32)$$

More generally, we have:

$$\text{Let } F(x) = \int_a^x f(u)du, \text{ where } x > a. \text{ Then } F'(x) = f(x). \quad (33)$$

In order to verify this, observe that

$$F'(x) = \lim_{\delta \rightarrow 0} \frac{F(x+\delta) - F(x)}{\delta} = \lim_{\delta \rightarrow 0} \frac{\int_a^{x+\delta} f(u) du - \int_a^x f(u) du}{\delta} = \lim_{\delta \rightarrow 0} \frac{\int_x^{x+\delta} f(u) du}{\delta} = f(x), \quad (34)$$

where the last equality follows from (32).

Moreover, it follows from (31):

$$\text{Let } F(x) = \int_x^b f(u) du, \text{ where } x < b. \text{ Then } F'(x) = -f(x). \quad (35)$$

The proof of (35) is left as an exercise.

7. Functions of two variables, and their partial derivatives

A real function $f(x,y)$ with two arguments, x and y , assigns a real number $z = f(x,y)$ to a pair (x,y) . These functions are called *bivariate* functions. For example, consider the bivariate quadratic function

$$f(x,y) = x^2 + y^2. \quad (36)$$

The shape of this function is a hyperbola:

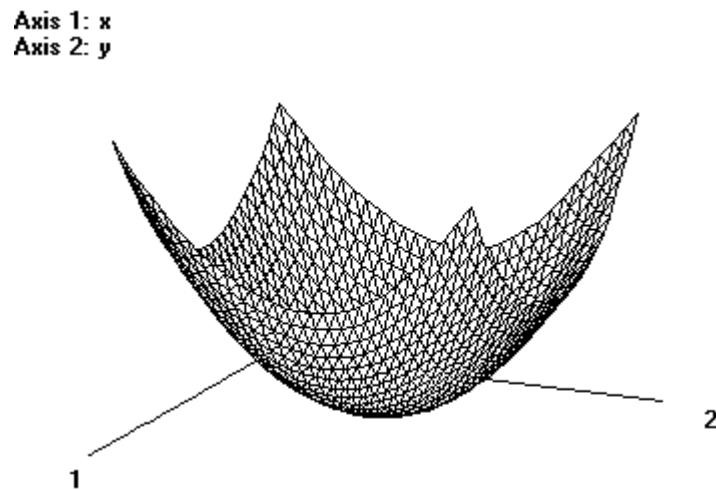


Figure 6: The function $f(x,y) = x^2 + y^2$ on the square $-1 < x < 1, -1 < y < 1$.

If the function $f(x,y)$ is differentiable in both arguments, then we can take the derivative of $f(x,y)$ to x , *treating y as a constant*. This derivative is called the *partial derivative of f(x,y) to x*, and is denoted by $\partial f(x,y)/\partial x$ or $\frac{\partial f(x,y)}{\partial x}$:

$$\frac{\partial f(x,y)}{\partial x} = \lim_{\delta \rightarrow 0} \frac{f(x+\delta, y) - f(x, y)}{\delta}. \quad (37)$$

Similarly, we can also take the derivative of $f(x,y)$ to y , *treating x as a constant*. This derivative is called the *partial derivative of f(x,y) to y*, and is denoted by $\partial f(x,y)/\partial y$ or $\frac{\partial f(x,y)}{\partial y}$:

$$\frac{\partial f(x,y)}{\partial y} = \lim_{\delta \rightarrow 0} \frac{f(x, y+\delta) - f(x, y)}{\delta}. \quad (38)$$

For example, in the case (36) we have $\partial f(x,y)/\partial x = 2x$, $\partial f(x,y)/\partial y = 2y$.

The general bivariate quadratic function takes the form

$$f(x,y) = \gamma_1(x - \beta_1 y - \alpha_1)^2 + \gamma_2(y - \beta_2 x - \alpha_2)^2, \quad (39)$$

where $\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2$ are constants. Then

$$\begin{aligned} \frac{\partial f(x,y)}{\partial x} &= 2\gamma_1(x - \beta_1 y - \alpha_1) - 2\gamma_2\beta_2(y - \beta_2 x - \alpha_2) \\ \frac{\partial f(x,y)}{\partial y} &= 2\gamma_2(y - \beta_2 x - \alpha_2) - 2\gamma_1\beta_1(x - \beta_1 y - \alpha_1) \end{aligned} \quad (40)$$

8. The minimum or maximum of a bivariate quadratic function

Clearly, the function (36) is minimal zero for $x = 0$ and $y = 0$. In this point the partial derivatives involved are zero. This can easily be verified directly, but also from Figure 5: In the point $(0,0)$ the hyperbola touches the horizontal plane at zero level, and is above zero level for any other point (x,y) . Thus, the point (x,y) for which the function $f(x,y) = x^2 + y^2$ is minimal can be found by solving the so-called *first-order conditions*:

$$\begin{aligned}\partial f(x,y)/\partial x &= 2x = 0, \\ \partial f(x,y)/\partial y &= 2y = 0.\end{aligned}$$

Next, let us have a look at the general bivariate quadratic function (39). If $\gamma_1 > 0$ and $\gamma_2 > 0$ then (39) can be written as

$$f(x,y) = \left(\sqrt{\gamma_1}x - \beta_1\sqrt{\gamma_1}y - \alpha_1\sqrt{\gamma_1}\right)^2 + \left(\sqrt{\gamma_2}y - \beta_2\sqrt{\gamma_2}x - \alpha_2\sqrt{\gamma_2}\right)^2. \quad (41)$$

The shape of this function is similar to Figure 5, except that the hyperbola involved will be shifted, squeezed and/or turned horizontally. Also, it is clear that this function is minimal if x and y are such that

$$\begin{aligned}\sqrt{\gamma_1}x - \beta_1\sqrt{\gamma_1}y - \alpha_1\sqrt{\gamma_1} &= 0 \Rightarrow x - \beta_1y - \alpha_1 = 0 \\ \sqrt{\gamma_2}y - \beta_2\sqrt{\gamma_2}x - \alpha_2\sqrt{\gamma_2} &= 0 \Rightarrow y - \beta_2x - \alpha_2 = 0\end{aligned} \quad (42)$$

However, the same conditions can be derived by setting the partial derivatives (40) equal to zero:

$$\begin{aligned}\frac{\partial f(x,y)}{\partial x} &= 2\gamma_1(x - \beta_1y - \alpha_1) - 2\gamma_2\beta_2(y - \beta_2x - \alpha_2) = 0 \\ \frac{\partial f(x,y)}{\partial y} &= 2\gamma_2(y - \beta_2x - \alpha_2) - 2\gamma_1\beta_1(x - \beta_1y - \alpha_1) = 0\end{aligned}\left. \begin{array}{l} \\ \end{array} \right\} \quad (43)$$

$$\Rightarrow \begin{cases} x - \beta_1y - \alpha_1 = 0 \\ y - \beta_2x - \alpha_2 = 0 \end{cases} \Rightarrow \begin{cases} x = (\alpha_1 + \alpha_2\beta_1)/(1 - \beta_1\beta_2) \\ y = (\alpha_2 + \alpha_1\beta_2)/(1 - \beta_1\beta_2) \end{cases}$$

provided that $\beta_1\beta_2 \neq 1$.

If $\gamma_1 < 0$ and $\gamma_2 < 0$ then (39) can be written as

$$f(x,y) = -\left(\sqrt{-\gamma_1}x - \beta_1\sqrt{-\gamma_1}y - \alpha_1\sqrt{-\gamma_1}\right)^2 - \left(\sqrt{-\gamma_2}y - \beta_2\sqrt{-\gamma_2}x - \alpha_2\sqrt{-\gamma_2}\right)^2. \quad (44)$$

It is easy to verify that now the bivariate quadratic function (39) takes a maximum, and that the point (x,y) where it is maximal can be obtained by solving the first-order conditions (43), provided that $\beta_1\beta_2 \neq 1$.

9. *Exercises*

1. Prove (12).
2. Prove (20).
3. Prove that $\exp(x)$ is continuous in all x .
4. Prove that $\ln(x)$ is continuous in all $x > 0$, using one or more of the properties (20).
5. Prove the chain rule (30). Hint: Write $g(x+\delta) = g(x) + (g(x+\delta) - g(x))$, and use the fact that by the continuity of $g(x)$ in x , $\lim_{\delta \rightarrow 0} (g(x+\delta) - g(x)) = 0$.
6. Prove (35) by modifying (34) to this case.
7. Determine the set of points (x,y) where the bivariate quadratic function (39) is minimal or maximal, for the case $\beta_1\beta_2 = 1$ and $\gamma_1 > \gamma_2/\beta_1$.

Importing Data in EasyReg International

Herman J. Bierens

Pennsylvania State University

October 3, 2009

1. *About EasyReg International*

EasyReg (**Easy Regression**) International is a free econometrics software package, which can be downloaded from web page <http://econ.la.psu.edu/~hbierens/EASYREG.HTM>

EasyReg opens with the following window, containing the menus:



Picture 1: EasyReg front window

The date displayed in Picture 1 is the compilation date, i.e. the date of the installation files. However, the displayed date of April 1, 2007, is not the current date. The most recent version of EasyReg is dated October 1, 2009, which has many improvements over previous versions.

It is possible that some upgrades have not yet been included in the current version. Therefore, after installing EasyReg, it is recommended that you upgrade it, via menu item “WWW > Upgrade EasyReg to the latest version.”. At the bottom of the EasyReg web page you will find a link to a web page which lists the upgrades not yet included in the current version.

EasyReg conducts a wide variety of econometric estimation, testing and forecasting tasks on all 32 bit Windows platforms, simply by point-and-click. EasyReg is designed for use in empirical research (including my own), and for teaching econometrics. In the latter case the user can choose his or her own econometrics level.

EasyReg is called International because it accepts dots and/or commas as decimal delimiters, regardless of the local number setting of Windows. Therefore, EasyReg runs everywhere in the world without need to adjust the setting of Windows.

However, if your Windows uses a language with a non-roman alphabet such as Japanese and Chinese, and EasyReg seems to be frozen, the problem is likely due to the language setting for non-Unicode programs. If so, open Control Panel, select "Regional and Language Options", and on the "Advanced" tab, under "Language for non-Unicode programs", select English. This will fix the problem without affecting the language of Windows itself.

When you run EasyReg for the first time, most menu items are disabled, because they need data to work. Therefore, in this paper I will show now how to import data in EasyReg and how to transform them.

2. *EasyReg data files and how to import them*

EasyReg can import two types of data files, Microsoft Excel files in CSV format, and EasyReg data files in space delimited text format. I will discuss the latter type first.

2.1 *EasyReg data files in space delimited text format*

This type of file is a (Notepad or Wordpad) text file with the following format. The first line contains two numbers, the number of variables ($= k$), and the missing value code ($= m$), separated by at least one space. The next k lines contain the names of the k variables involved. The file names may contain any

character, including spaces, commas, etc., and there is no restriction on their length. However, it is recommended to keep the variable names short, say no more than ten characters. The rest of the file contains the data entries $x(j,i)$ for observation j on variable i , where $i = 1, \dots, k$, $j = 1, \dots, n$, separated by one or more spaces, where n is the number of observations. Data entries with value m are interpreted as missing values, except if $m = 0$, which indicates that there are no missing values. Therefore, the missing value code should be chosen such that the real data entries will never take the value m , for example by choosing $m = -99999.99$.

You may use a dot or a comma as decimal delimiter for the numerical data entries. For example, $m = -99999,99$ is also a valid missing value code. EasyReg will automatically replace a decimal comma with a decimal dot when it reads numbers. However, do not use a grouping symbol, because otherwise EasyReg cannot interpret the number involved. For example, do not use numbers with a grouping symbol like 34.123,75 (European style) or 34,123.75 (US style), but instead write them as 34123,75 or 34123.75, respectively.

Thus, the structure of this data file is:

k	m		
Name of variable 1			
⋮			
Name of variable k			
$x(1,1)$	$x(1,2)$	…	$x(1,k)$
⋮	⋮	⋮	⋮
$x(n,1)$	$x(n,2)$	…	$x(n,k)$

For example, consider the following data file in Table 1 below, containing quarterly data on real consumption, GDP and investment in the US, from the first quarter of 1985 to the first quarter of 1995, together with the quarters themselves. The last two quarters contain missing values, indicated by the missing value code -99999.99 .

The inclusion of the time “variable” quarter is optional. As we will see, once you import a data file, EasyReg will ask for the type of data, either cross-section, annual time series, quarterly time series, monthly time series, or other time series, and in the time series cases the year and eventually quarter, month or other frequency (for example bi-annual) of the first observation. Anyhow, if you include a time variable, or observation number in the case of cross-section data, this variable should be a valid numerical data entry.

Table 1: Example of a space delimited EasyReg data file

```

4 -99999.99
Quarter
Real US Consumption
Real US GDP
Real US Investment
1985.1 2824.90 4221.80 717.80
1985.2 2849.70 4254.80 724.60
1985.3 2893.30 4309.00 719.90
1985.4 2895.30 4333.50 732.90
1986.1 2922.40 4390.50 728.20
1986.2 2947.90 4387.70 728.10
1986.3 2993.70 4412.60 723.80
1986.4 3012.50 4427.10 725.90
1987.1 3011.50 4460.00 706.80
1987.2 3046.80 4515.30 718.30
1987.3 3075.80 4559.30 733.00
1987.4 3074.70 4625.50 733.90
1988.1 3128.20 4655.30 737.70
1988.2 3147.80 4704.80 753.30
1988.3 3170.60 4734.50 758.60
1988.4 3202.90 4779.70 764.10
1989.1 3203.60 4817.60 761.70
1989.2 3212.20 4839.00 757.50
1989.3 3235.30 4839.00 753.10
1989.4 3242.00 4856.70 744.60
1990.1 3264.40 4898.30 761.80
1990.2 3271.60 4917.10 745.80
1990.3 3288.40 4906.50 740.10
1990.4 3265.90 4867.20 716.60
1991.1 3242.90 4842.00 686.40
1991.2 3259.50 4867.90 683.40
1991.3 3269.80 4879.90 685.60
1991.4 3265.30 4880.80 684.40
1992.1 3311.40 4918.50 693.50
1992.2 3325.40 4947.50 721.30
1992.3 3357.60 4990.50 728.10
1992.4 3403.40 5060.70 748.60
1993.1 3417.20 5075.30 770.70
1993.2 3439.20 5105.40 787.30
1993.3 3472.20 5139.40 808.80
1993.4 3506.20 5218.00 851.70
1994.1 3546.30 5261.10 873.40
1994.2 3557.80 5314.10 891.70
1994.3 3584.70 5367.00 910.20
1994.4 -99999.99 -99999.99 -99999.99
1995.1 -99999.99 -99999.99 -99999.99

```

Although the data entries in Table 1 are nicely lined-up, there is no need for that. As long as the data entries are separated by a space, EasyReg can read them.

The current version of EasyReg can also read this file if, instead of spaces, tabs are used as data entry separators, provided these tabs are of the type compatible with Notepad or Wordpad tabs. EasyReg will automatically convert these tabs to spaces.

2.2 Excel files in CSV format

CSV stands for Comma-Separated Values. It is one of the formats for saving an Excel file. A CSV file is a text file (and can therefore be imported in Notepad and Wordpad) where the data entries are separated by commas. The data in Table 1 in CSV format is displayed in Table 2.

Table 2: The data in Table 1 in Excel CSV format

Quarter	Real US Consumption	Real US GDP	Real US Investment
1985.1	2824.9	4221.8	717.8
1985.2	2849.7	4254.8	724.6
1985.3	2893.3	4309	719.9
1985.4	2895.3	4333.5	732.9
1986.1	2922.4	4390.5	728.2
1986.2	2947.9	4387.7	728.1
1986.3	2993.7	4412.6	723.8
1986.4	3012.5	4427.1	725.9
1987.1	3011.5	4460	706.8
1987.2	3046.8	4515.3	718.3
1987.3	3075.8	4559.3	733
1987.4	3074.7	4625.5	733.9
1988.1	3128.2	4655.3	737.7
1988.2	3147.8	4704.8	753.3
1988.3	3170.6	4734.5	758.6
1988.4	3202.9	4779.7	764.1
1989.1	3203.6	4817.6	761.7
1989.2	3212.2	4839	757.5
1989.3	3235.3	4839	753.1
1989.4	3242	4856.7	744.6
1990.1	3264.4	4898.3	761.8
1990.2	3271.6	4917.1	745.8
1990.3	3288.4	4906.5	740.1
1990.4	3265.9	4867.2	716.6
1991.1	3242.9	4842	686.4
1991.2	3259.5	4867.9	683.4
1991.3	3269.8	4879.9	685.6
1991.4	3265.3	4880.8	684.4
1992.1	3311.4	4918.5	693.5
1992.2	3325.4	4947.5	721.3
1992.3	3357.6	4990.5	728.1
1992.4	3403.4	5060.7	748.6
1993.1	3417.2	5075.3	770.7
1993.2	3439.2	5105.4	787.3
1993.3	3472.2	5139.4	808.8
1993.4	3506.2	5218	851.7
1994.1	3546.3	5261.1	873.4
1994.2	3557.8	5314.1	891.7
1994.3	3584.7	5367	910.2
1994.4	,,		
1995.1	,,		

The first record contains the variable names, separated by commas. If a variable name contains a comma itself, the name involved has to be enclosed between quotation marks (“”). Older versions of Excel also enclose variable names between quotation marks if the name contains a space. EasyReg will read these files

as well.

Missing values in a CSV file are indicated by an empty data entry. Thus, two adjacent commas (,,) indicate that the data entry corresponding to the position after the first comma is a missing value. Similarly, a comma at the end of a data record indicates that the next data entry is a missing value, and a comma as first character of a data record indicates that the first data entry is a missing value.

The CSV format in Table 2 applies if Windows uses a dot as decimal delimiter, as in the US. If Windows uses a comma as decimal delimiter, as in continental Europe, the commas are replaced by semi-colons (;), and the decimal dots by commas. The CSV file then looks like this:

Table 3: Excel CSV format if the decimal delimiter is a comma

Quarter	Real US Consumption	Real US GDP	Real US Investment
1985,1;2824,9;4221,8;717,8			
1985,2;2849,7;4254,8;724,6			
1985,3;2893,3;4309;719,9			
1985,4;2895,3;4333,5;732,9			
1986,1;2922,4;4390,5;728,2			
1986,2;2947,9;4387,7;728,1			
1986,3;2993,7;4412,6;723,8			
1986,4;3012,5;4427,1;725,9			
1987,1;3011,5;4460;706,8			
1987,2;3046,8;4515,3;718,3			
1987,3;3075,8;4559,3;733			
1987,4;3074,7;4625,5;733,9			
1988,1;3128,2;4655,3;737,7			
1988,2;3147,8;4704,8;753,3			
1988,3;3170,6;4734,5;758,6			
1988,4;3202,9;4779,7;764,1			
1989,1;3203,6;4817,6;761,7			
1989,2;3212,2;4839;757,5			
1989,3;3235,3;4839;753,1			
1989,4;3242;4856,7;744,6			
1990,1;3264,4;4898,3;761,8			
1990,2;3271,6;4917,1;745,8			
1990,3;3288,4;4906,5;740,1			
1990,4;3265,9;4867,2;716,6			
1991,1;3242,9;4842;686,4			
1991,2;3259,5;4867,9;683,4			
1991,3;3269,8;4879,9;685,6			
1991,4;3265,3;4880,8;684,4			
1992,1;3311,4;4918,5;693,5			
1992,2;3325,4;4947,5;721,3			
1992,3;3357,6;4990,5;728,1			
1992,4;3403,4;5060,7;748,6			
1993,1;3417,2;5075,3;770,7			
1993,2;3439,2;5105,4;787,3			
1993,3;3472,2;5139,4;808,8			
1993,4;3506,2;5218;851,7			
1994,1;3546,3;5261,1;873,4			
1994,2;3557,8;5314,1;891,7			
1994,3;3584,7;5367;910,2			
1994,4;;;			
1995,1;;;			

EasyReg automatically checks whether Windows uses a dot or a comma as decimal delimiter, and assumes that the format of the CSV file is either that in Table 2 or in Table 3, respectively.

2.3 Importing data files in EasyReg

To import a data file, open “File > Get data”:

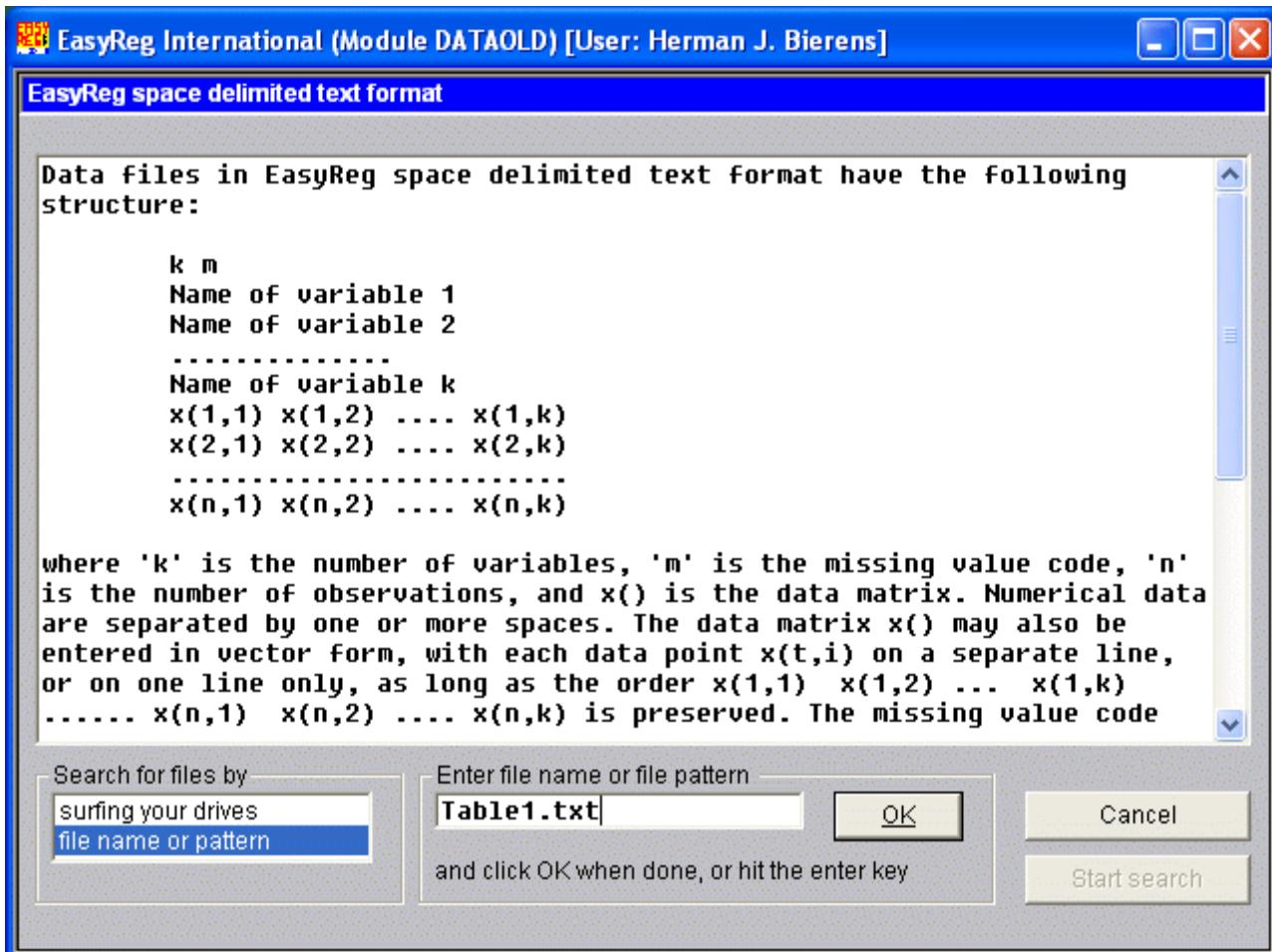


Picture 2: The File menu

2.3.1 Importing an EasyReg data file in space delimited text format

I will now show how to import the data in Table 1 in EasyReg. The data file name is Table1.txt, which is located in folder C:\LectureNotes\UnderGraduate\EasyReg\Data_1.

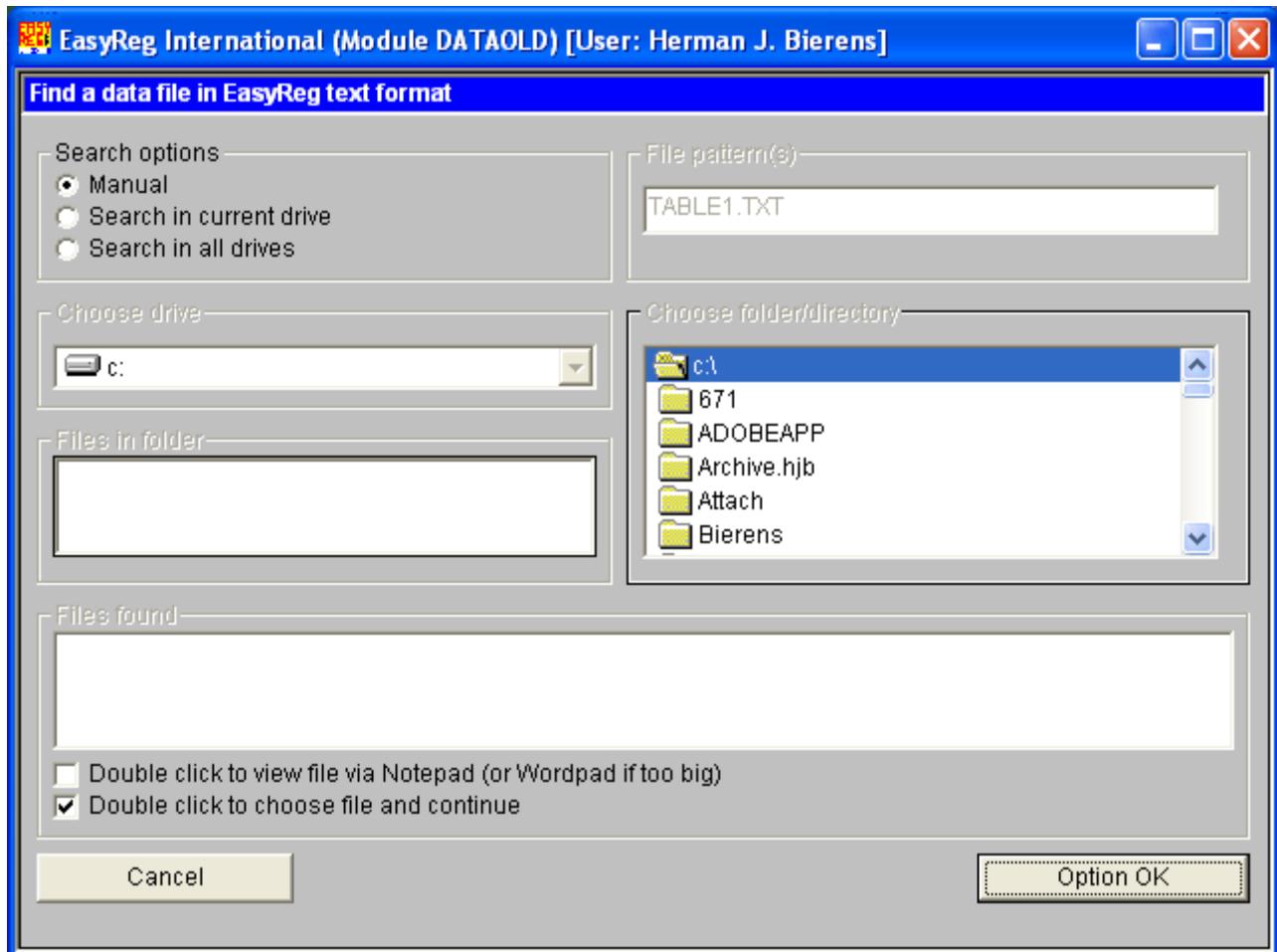
Open “File > Get data > Choose a former default EasyReg input file” (see Picture 2). Then the following window appears:



Picture 3

There are two ways to find a data file in EasyReg space delimited text format, by surfing your drives and folders, and by file name or file pattern. If you know the file name, choose the latter option, enter the file pattern or file name, click OK (or hit the enter key: Buttons with underlined text respond to the enter key as well), and then click “Start search”. Then you can scan all your drives automatically for files of this type, or navigate to the folder where the input file is located.

Because we know where to look for the file, the fastest way to find it is by choosing the option “surfing your drives”, and then click “Start search”. Then EasyReg opens the window displayed in Picture 4 below.



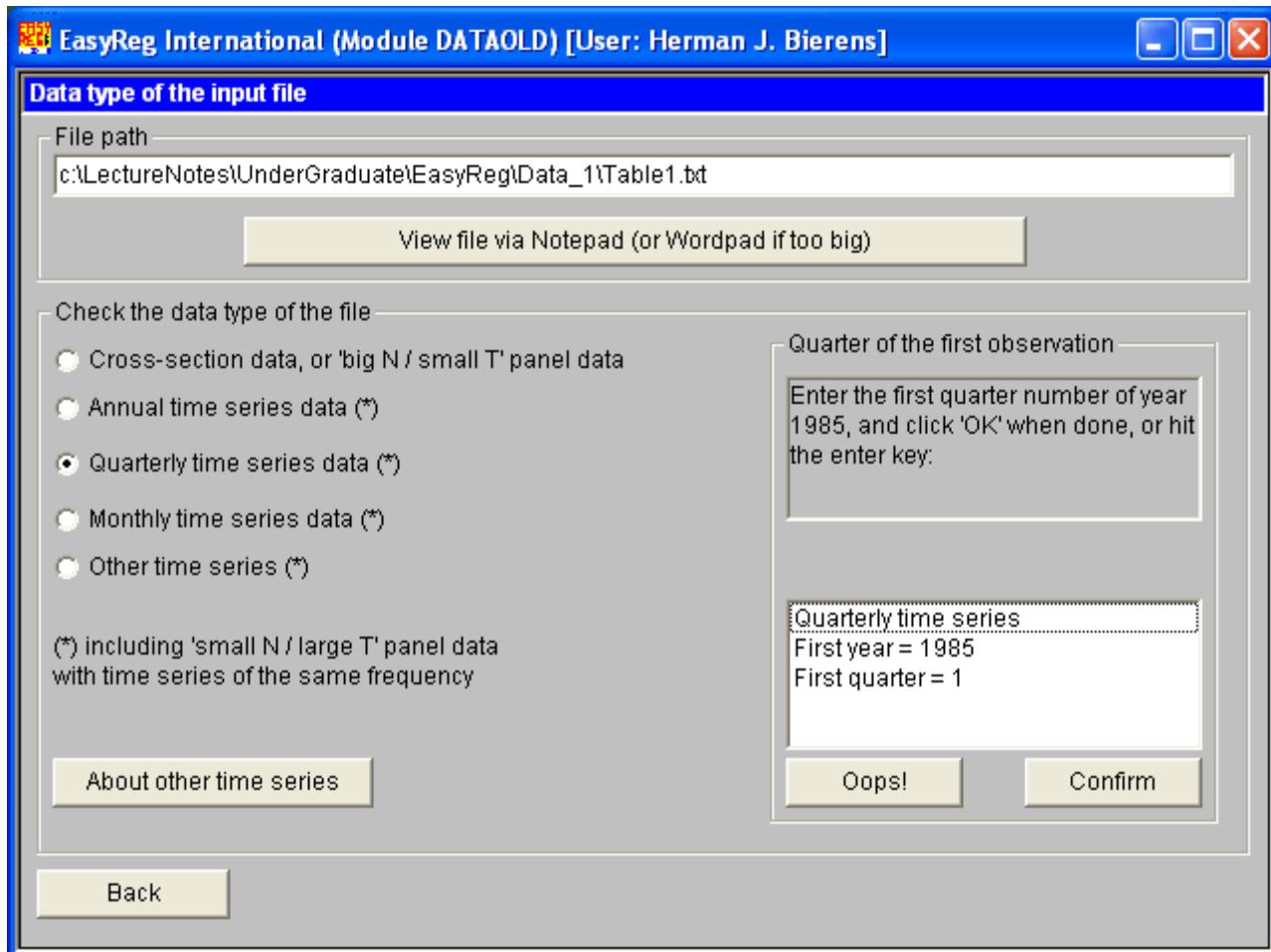
Picture 4

The default search option is “Manually”. Leave it that way, and click “Option OK”. Next, navigate to the folder where the file Table1.txt is located. This is not shown but the steps involved are self-explanatory. Then the file path

C:\LectureNotes\UnderGraduate\EasyReg\ Data_1\Table1.txt

will appear in the list box “Files found” (the last box) .

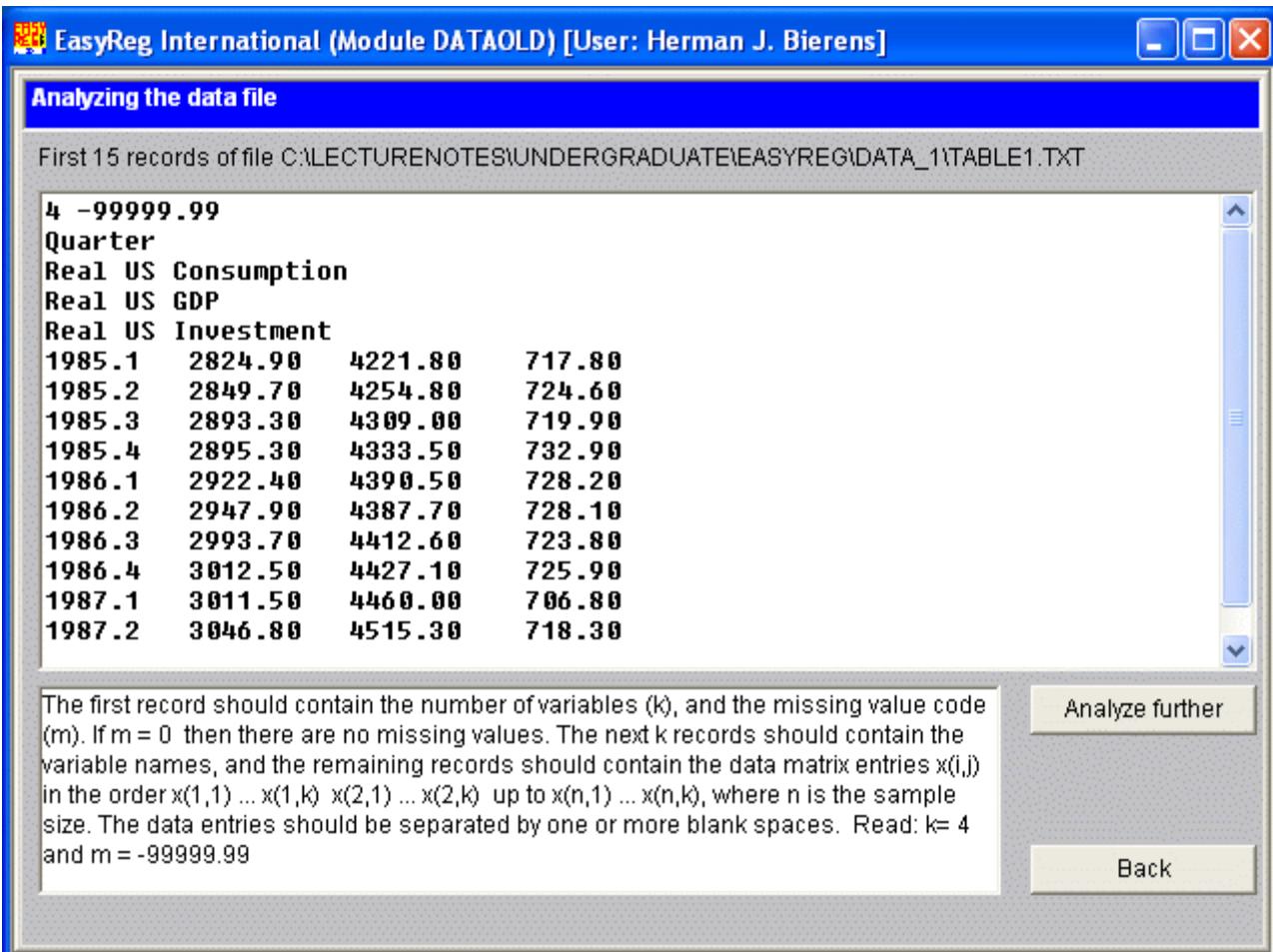
Once you double-click this file path, EasyReg will jump to the next window, which will ask you to indicate the data type, and in the time series cases the year, and for seasonal time series the quarter, month or other frequency, of the first observation. In case of the data in Table 1, the data type is quarterly time series data, starting from quarter 1 of year 1985:



Picture 5

The “Confirm” button opens a window (not shown) in which you can type any comments or information regarding this data file. This information can be retrieved later via “File > Current data” (See Picture 2). You may also skip this option. Now EasyReg will display the variable names and the next ten records of file Table1.txt. See Picture 6 below.

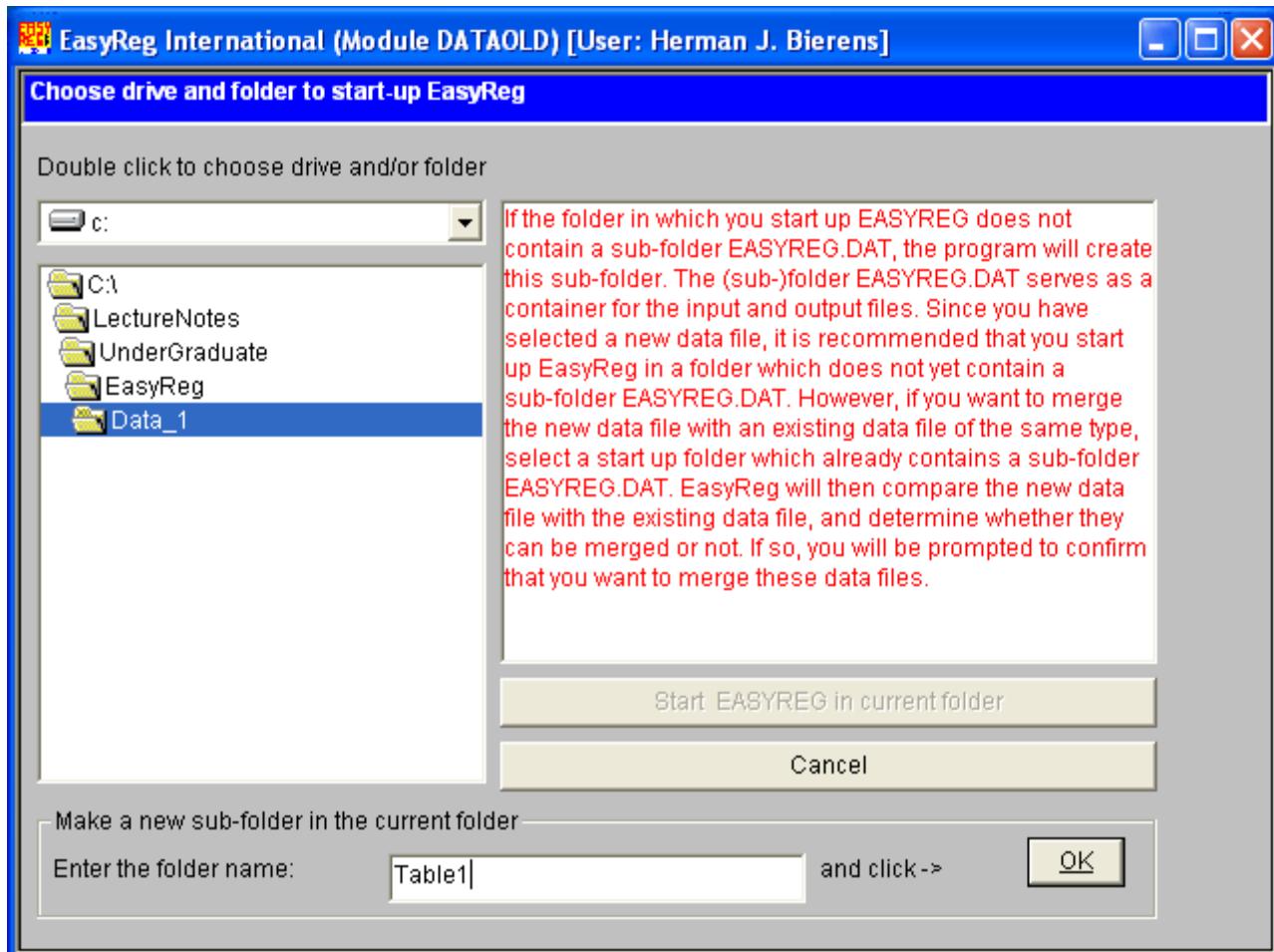
The next step is to analyze the file (via “Analyze further”), and check whether the length of the file corresponds with the number of variables. In particular, the number of numerical data entries, including missing values, divided by the number of variables, should be a natural number, n , being the sample size or the length of the time series. EasyReg also checks whether the numerical data entries are valid numbers.



Picture 6

Once the data file is analyzed and found OK, the data will be copied to a pair of random access binary files, INPUT1.RAN and INPUT2.RAN, because retrieving the data from these files is much faster than reading them from a text file. Also, the data is converted to a CSV file, INPUT.CSV, which you can open to check out whether the data file has been read correctly. The window involved is self-explanatory and therefore not shown.

In first instance the files INPUT1.RAN, INPUT2.RAN and INPUT.CSV are stored in a temporary folder. To complete the data import, you have to choose a folder where you want to store these files. EasyReg will then create a sub-folder EASYREG.DAT in that folder, which serves as a container for the input files, and later on also for the output files. The folder I will choose is a sub-folder Table1 of the folder where I had stored the original data file Table1.txt:



Picture 7

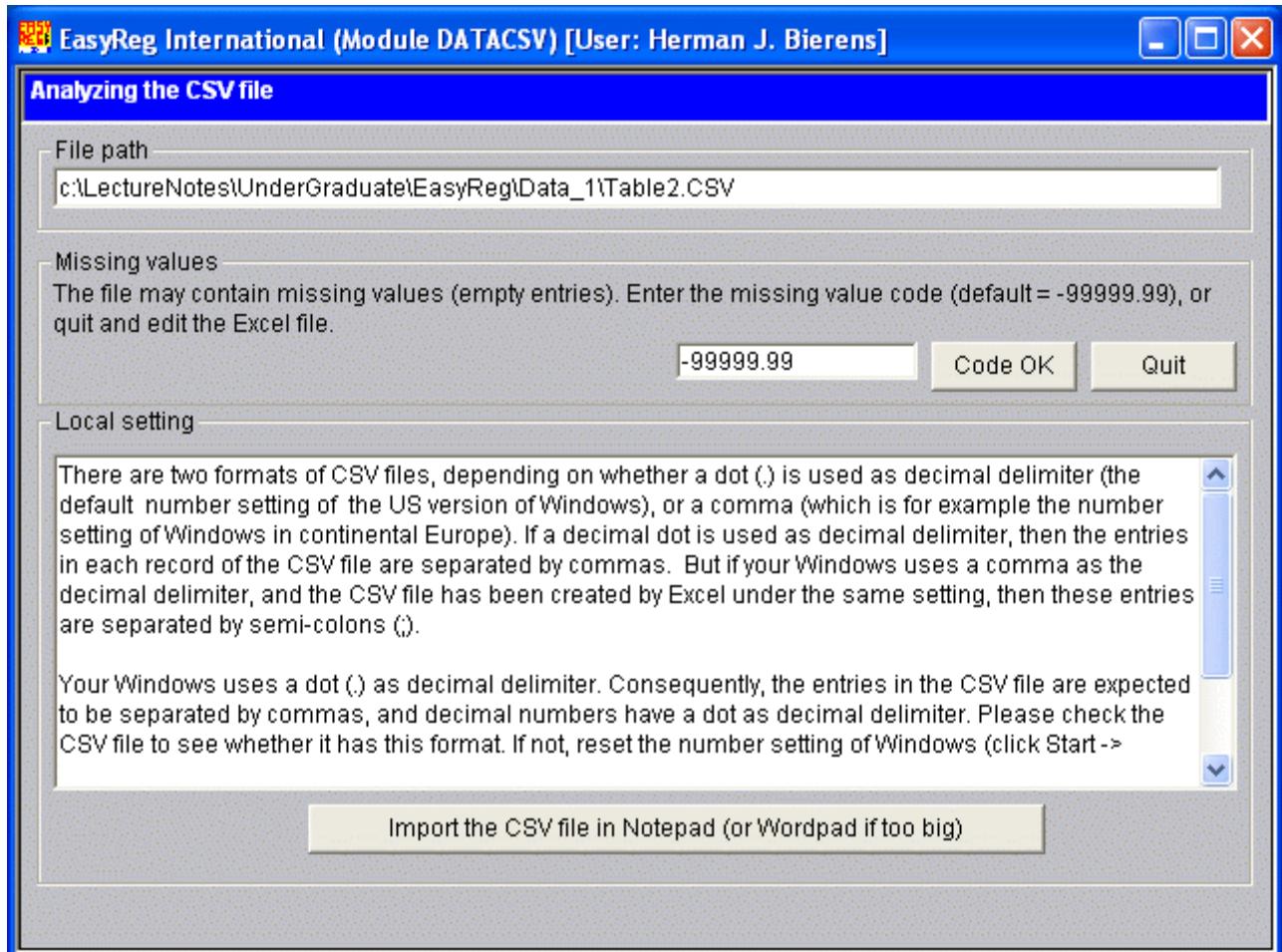
The sub-folder Table1 does not exist yet, but EasyReg can create it. Thus, type in the new sub-folder name, click OK or hit the enter key, and click the “Start EASYREG in current folder” button. Then the new input files INPUT1.RAN, INPUT2.RAN and INPUT.CSV are copied to folder

C:\LectureNotes\UnderGraduate\EasyReg\ Data_1\Table1\EASYREG.DAT,
and you are done with the data import.

2.3.2 Importing a CSV file

The file name of the data in Table 2 is Table2.CSV, which is located in the same folder as Table1.txt: C:\LectureNotes\UnderGraduate\EasyReg\Data_1. To import this file, open “File > Get data > Choose an Excel file in CSV format”. Then EasyReg opens with a window explaining what CSV files are and how they look like. The next window is the same as Picture 4. You have to navigate to the folder

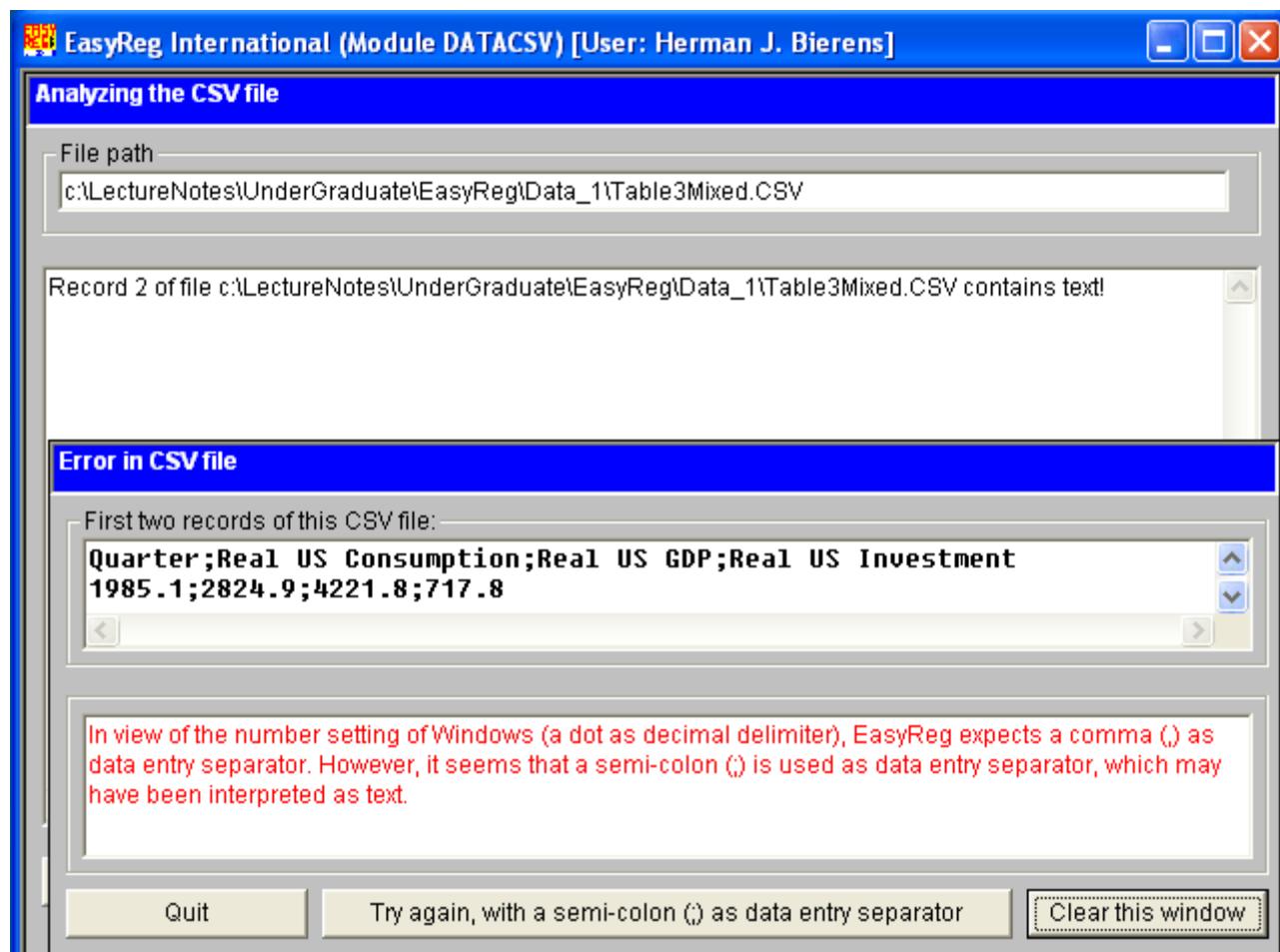
where your CSV file is located, or scan your drives for all CSV files. Once you have reached the folder the file will be displayed in the box “Files found.” The procedure is now the same as in the previous case. However, instead of Picture 6 you will get:



Picture 8

Recall that missing values in a CSV file are represented by empty data entries. Therefore, you have to specify the missing value code with which these empty entries have to be filled. I recommend to adopt the default value. Once you click Code OK, the remaining import procedures are the same as before, provided that the CSV file has the format that is expected by EasyReg. As to the latter, recall that the expected CSV format is either data entries separated by a comma (,) if Windows uses a dot (.) as decimal delimiter, as in Table 2, or data entries separated by a semi-colon (;) if Windows uses a comma as decimal delimiter, as in Table 3. However, some versions of Excel save CSV files in a mixed format: A dot as

decimal delimiter, a semi-colon as data entry separator. If the number setting of Windows is a dot as decimal delimiter, and you import a CSV file of this mixed type, you will get an error message window:



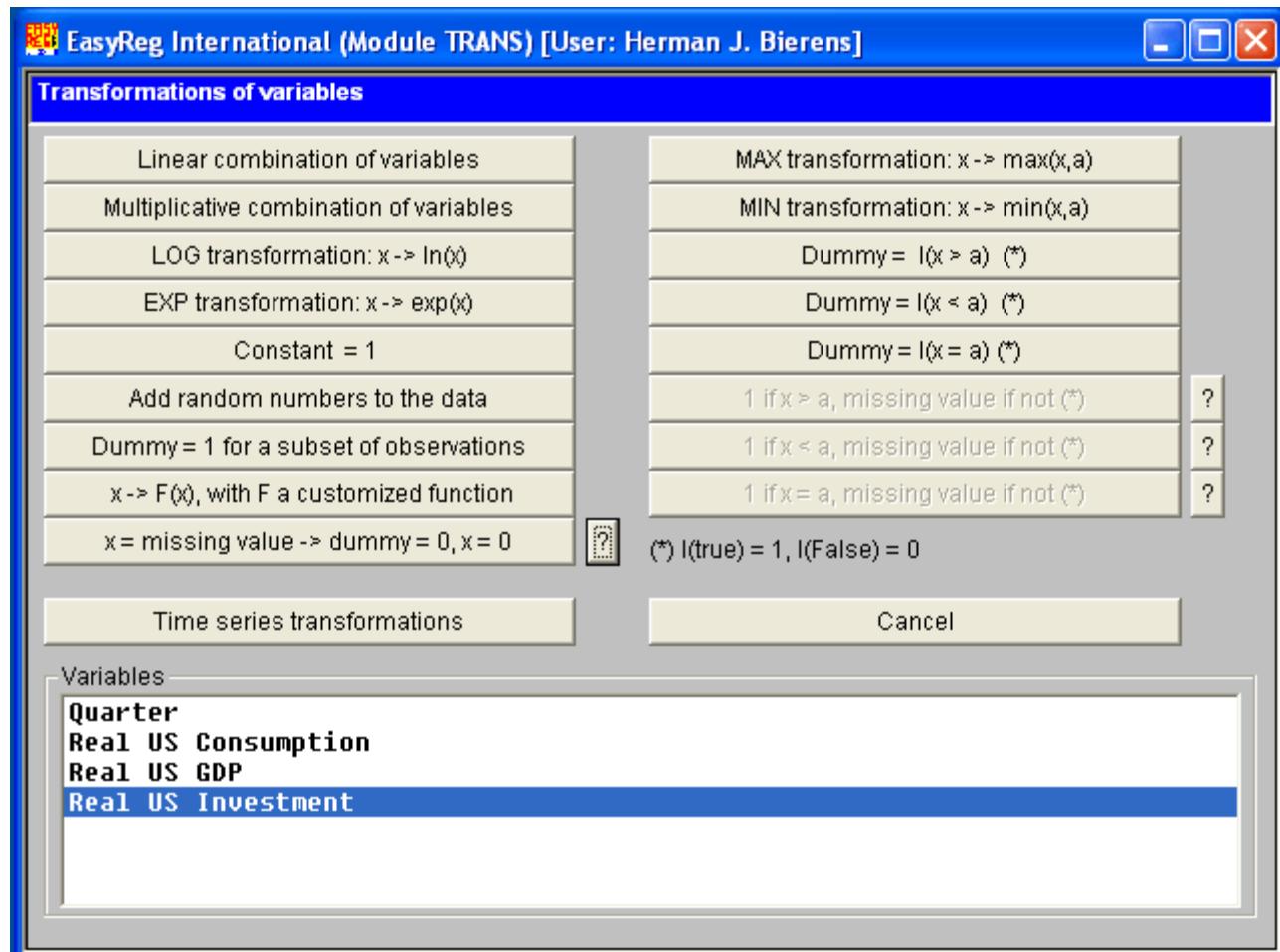
Picture 9

In this case you can read the CSV file as well, after clicking the “Try again,...” button. This will also work if the number setting of Windows is a dot as decimal delimiter and you import a European type CSV file with semi-colons as data separators and a commas as decimal delimiters. The same applies for US type CSV files if Windows uses a comma as decimal delimiter.

Finally, note that EasyReg does not need Excel itself to work with CSV files, because EasyReg reads these files as text files.

3. Data transformations

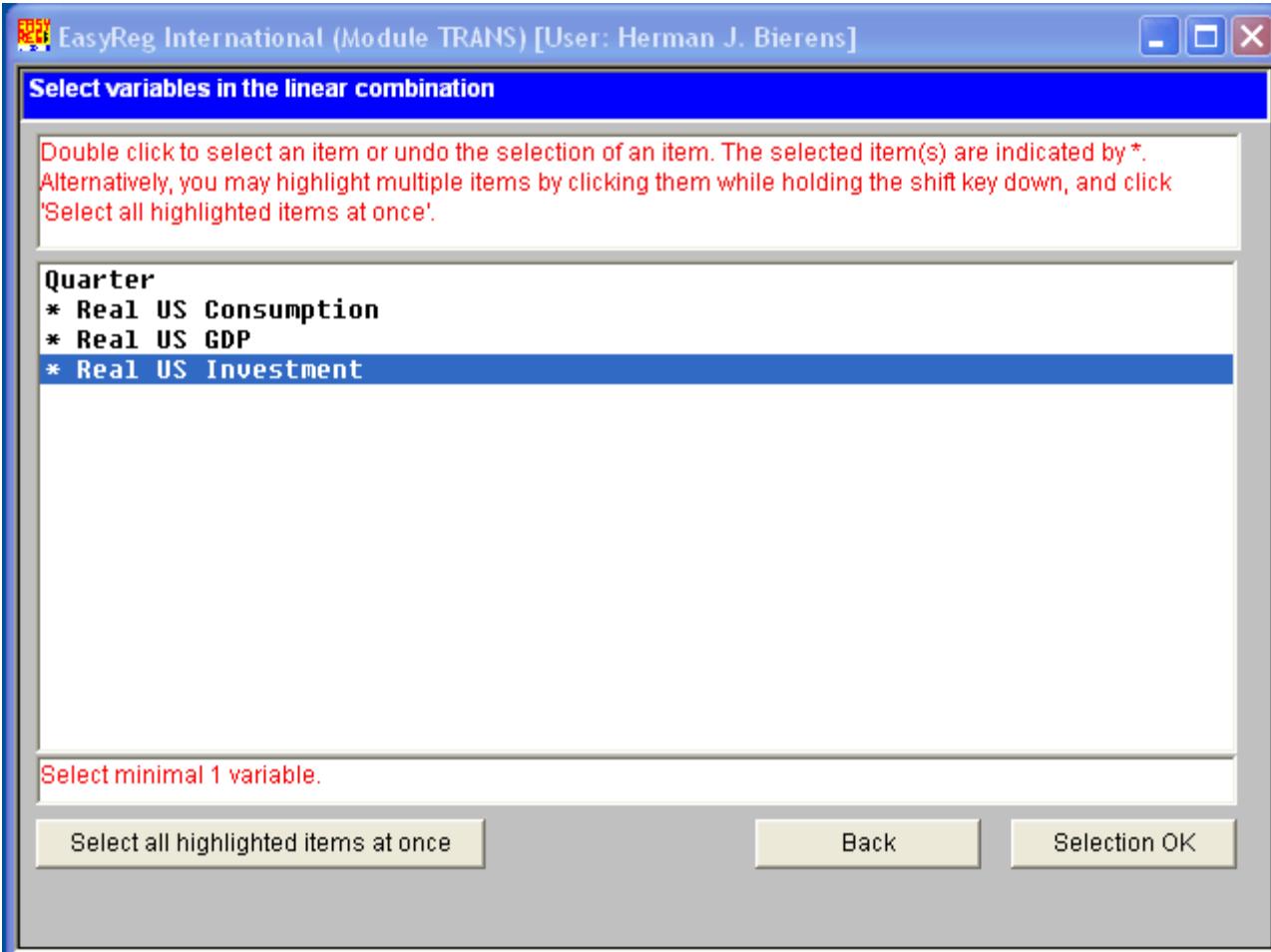
Often you have to transform your data before they can be used. If so, open “Menu > Input > Transform variables”:



Picture 10

For example, let us create a new variable, Real US GDP – Real US Consumption – Real US Investment. This is a linear combination, with coefficients 1, -1, -1. Thus, click the “Linear combination” button. Then the window displayed in Picture 11 below appears.

Double-click on Real US GDP, Real US Consumption and Real US Investment, and click “Selection OK”. EasyReg will then prompt you to enter the coefficients of each of these variables in the linear combination. Once done, you can enter a name for this linear combination. Let us call this linear combination “Other real expenditures”. Then this new variable will be added to the data.

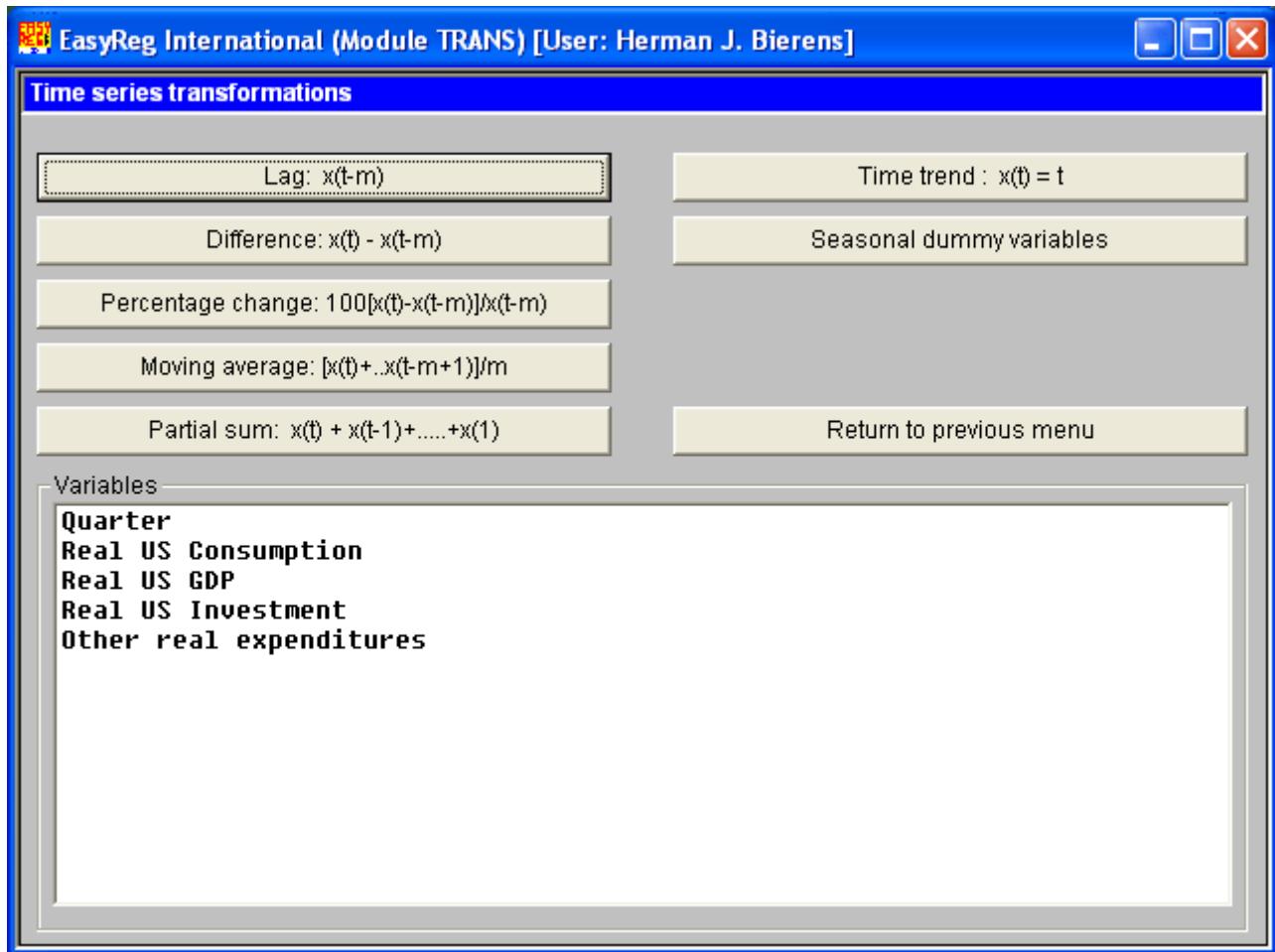


A multiplicative combination is of the form $x_1^{c_1}x_2^{c_2}\dots x_k^{c_k}$, where x_1, x_2, \dots, x_k are the variables involved and c_1, c_2, \dots, c_k are the corresponding powers. The procedure is similar to the linear combination transformation.

The LOG and EXP transformations are self-explanatory. Adding a constant 1 to the data is useful only if you want to add a constant to a variable, by taking a linear combination of 1 and a variable. You don't need it to run a linear regression with an intercept, because the OLS module itself provides the option to include an intercept.

Creating a dummy variable for a subset of observation can be useful if you want to test whether parameters are different in a particular period or not. The other transformation options, except the option "Time series transformations", are only useful for particular advanced econometric tasks, and will therefore not be discussed here.

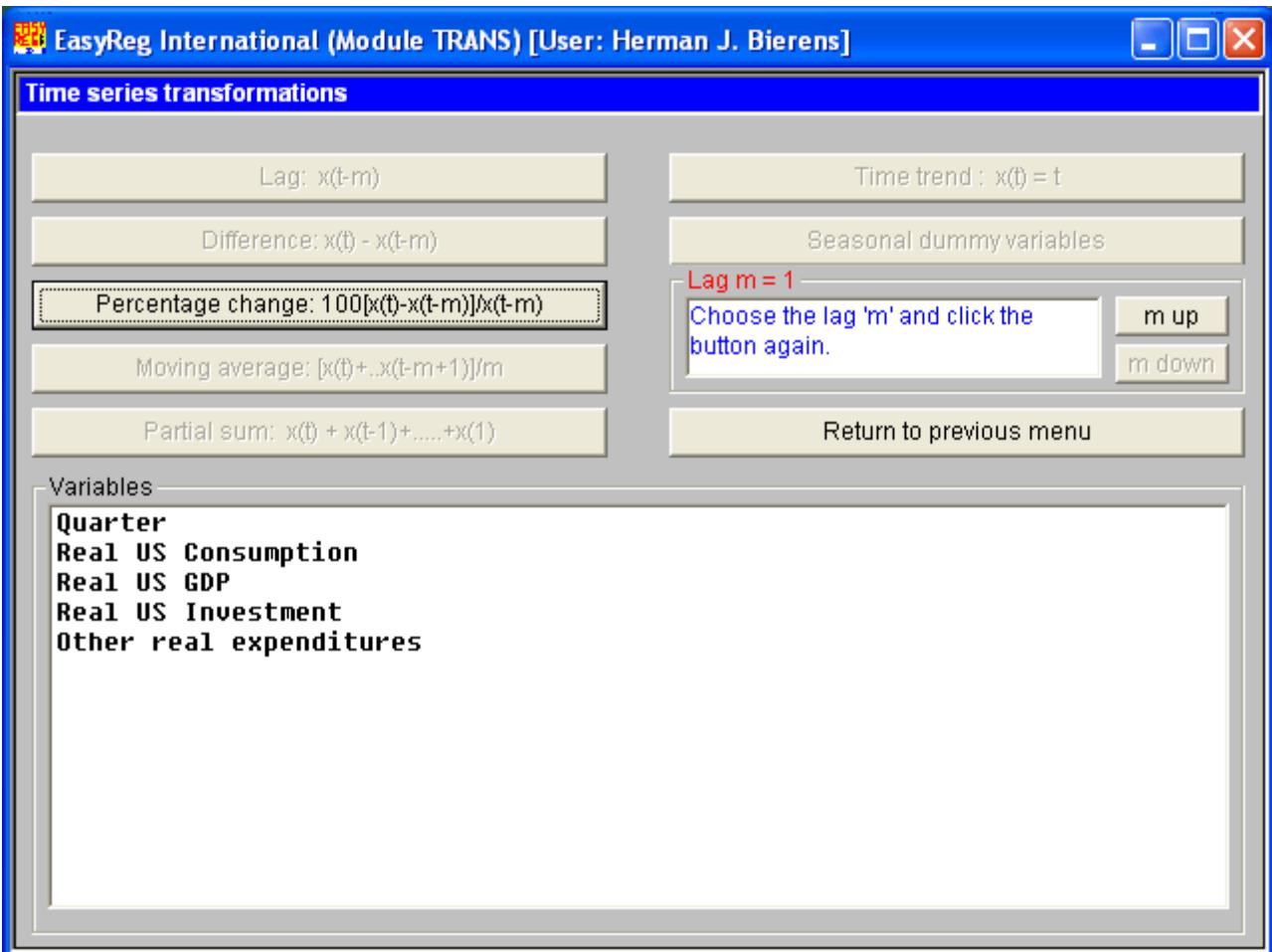
Next, click “Time series transformations”:



Picture 12

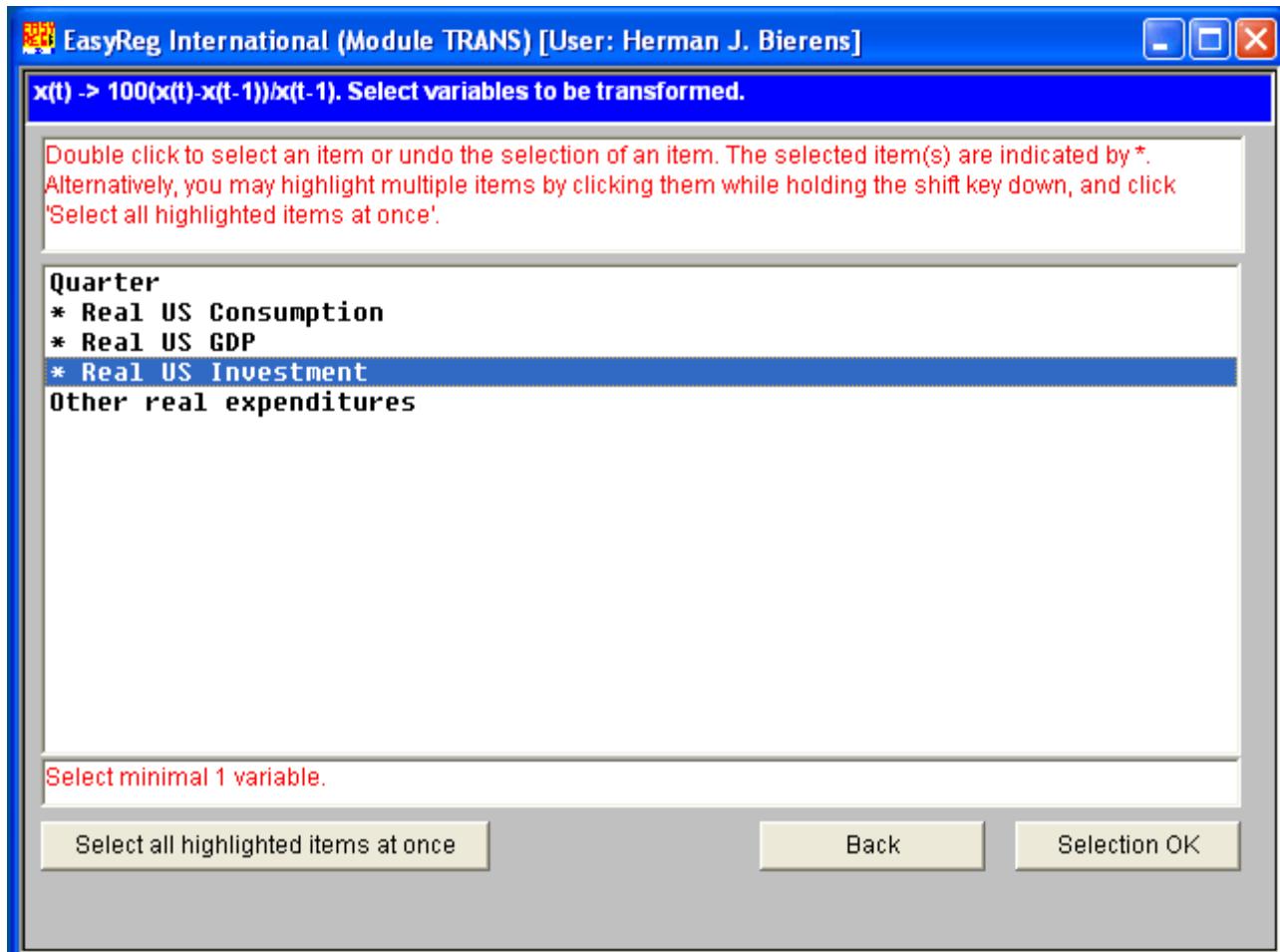
In general you don't need to add a time trend or seasonal dummy variables to the data, because the OLS module in EasyReg provides the option to include a time trend and/or seasonal dummy variables in a regression model. The same applies to the “Lag” transformation: The OLS module provides the option to include lagged dependent and/or independent variables in a regression model. The other options, except the “Difference” and “Percentage change” options, are intended for advanced econometric tasks, and will therefore not be discussed.

The most important time series transformations are the “Difference” and the “Percentage change” transformations, which are used to make the time series stationary. Once you click one of these buttons, you can specify the lag m . To demonstrate this, click “Percentage change”:



Picture 13

Here I will choose $m = 1$, and transform the variables Real US GDP, Real US Consumption and Real US Investment. Thus, click the “Percentage change” button (now with underlined text) again, select the variables Real US GDP, Real US Consumption and Real US Investment in the next window by double-clicking them, and click the “Selection OK” button. See Picture 14 below. Then, upon confirmation, three new variables will be added to the data set: %DIF1[Real US GDP], %DIF1[Real US Consumption] and %DIF1[Real US Investment], being the quarterly percentage changes of the variables involved.

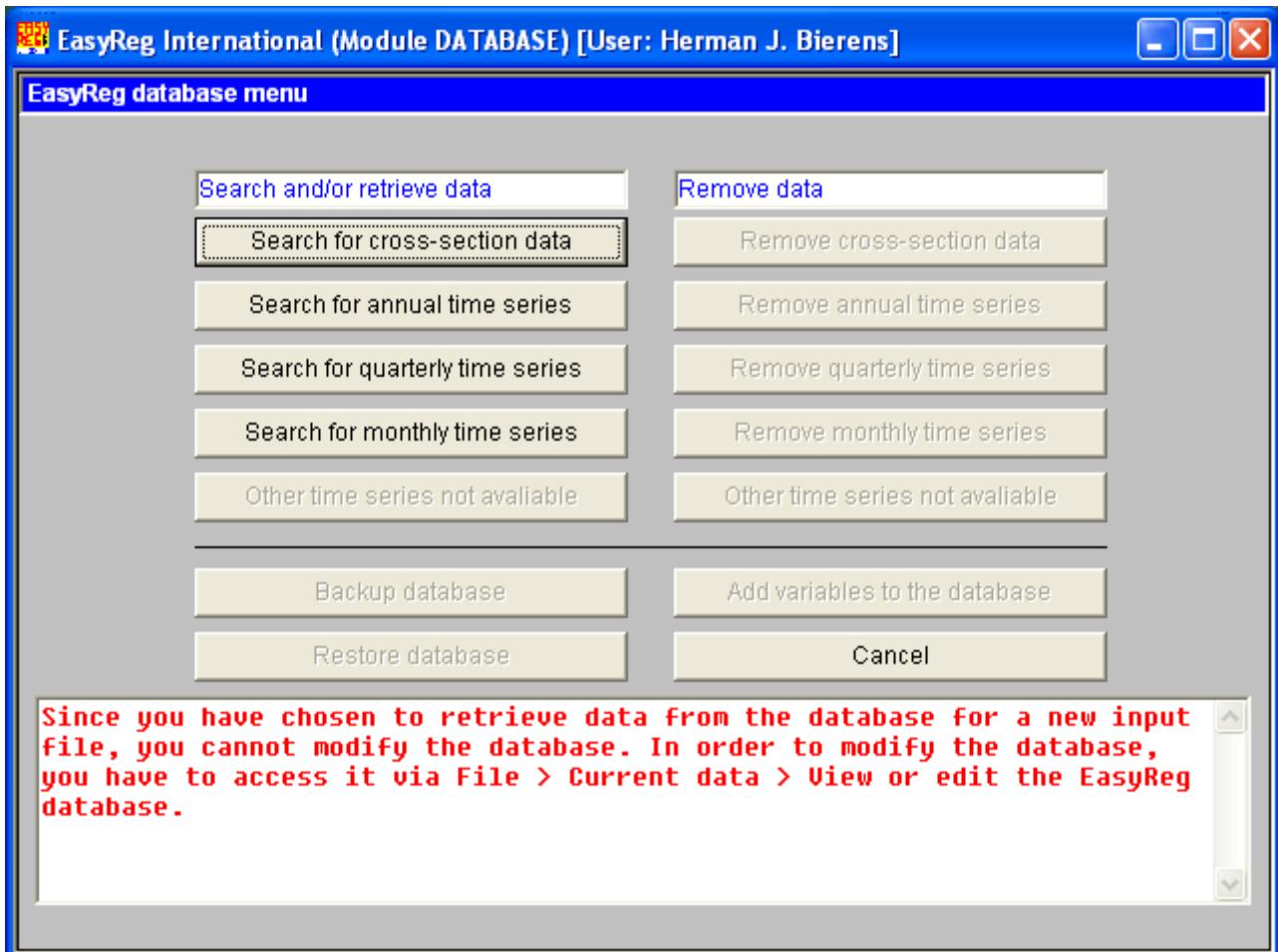


Picture 14

4. Retrieving data from the EasyReg database

EasyReg comes with its own database, in which you can store variables from the current input file, via “File> Current data > View or edit the EasyReg database”, or retrieve variables from for a new input file, via “File > Get data > Choose data from the EasyReg database.” In the latter case the window displayed in Picture 15 below appears.

Actually, the data in Tables 1-2 came from this database, although the time series were shortened in order to be able to display them on one page. The corresponding time series in the database are much longer, starting from quarter 1 of 1947.

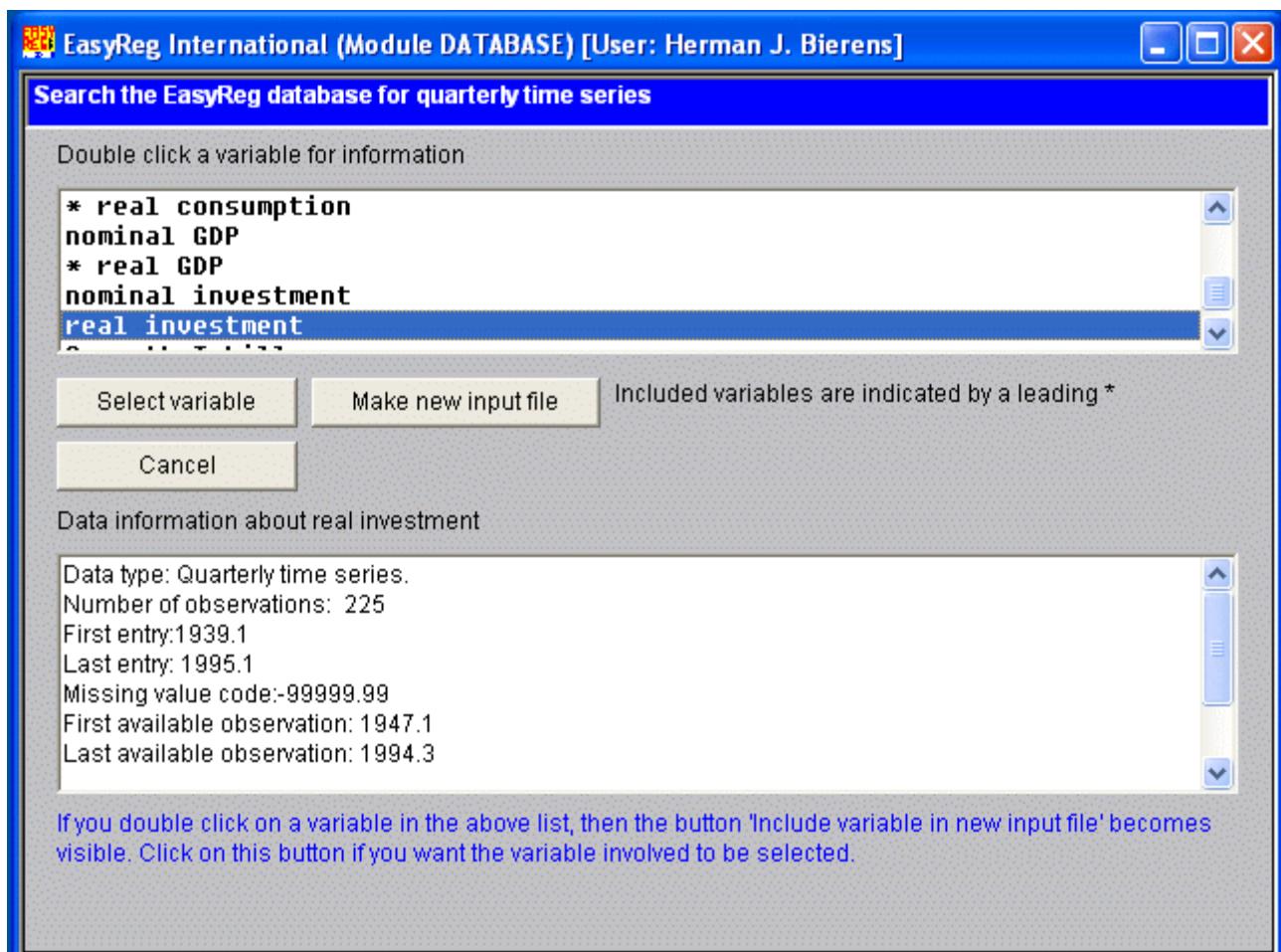


Picture 15

Let us retrieve these longer time series. Thus, click “Search for quarterly time series”. Then the window displayed in Picture 16 below appears.

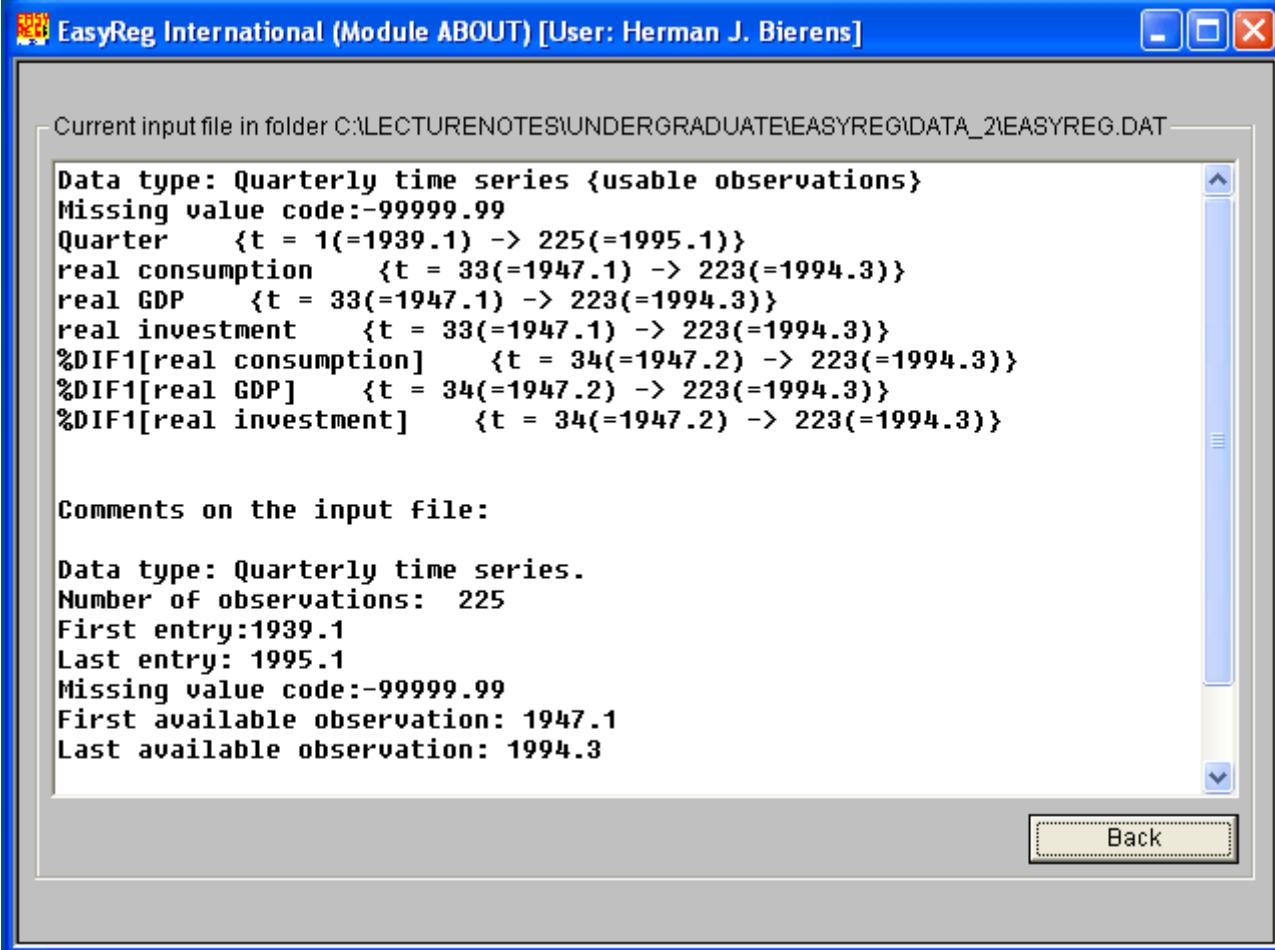
The time series involved are “real consumption”, “real GDP”, and “real investment”. Click a variable that you want to select and then click “Select variable”. Repeat this for each variable that you want to select. The selected variables are then displayed with a leading asterix (*).

EasyReg will in first instance assume that you want to store this data in the same folder as the current data set, namely C:\LectureNotes\UnderGraduate\EasyReg\ Data_1\Table1. If so, you can either overwrite the current data or merge it with the new data set, provided that in the latter case the data are of the same type. However, I will store the new data in folder C:\LectureNotes\UnderGraduate\EasyReg\Data_2. Also, I will transform the three time series by taking quarterly percentage changes, similar as before.



Picture 16

Finally, to check which is the current data set, open “File > Current input file > Show current input file”. Then the window in Picture 17 below appears. EasyReg is now ready to conduct a wide variety of econometrics tasks, for example estimation and testing a dynamic regression model. However, a discussion of these tasks and how to conduct them is beyond the scope of this paper. They are explained in detail in the various guided tours that are downloadable from the EasyReg download web page.



Picture 17: File > Current data > Show current input file

THE TWO-VARIABLE LINEAR REGRESSION MODEL

Herman J. Bierens

Pennsylvania State University

April 30, 2012

1. *Introduction*

Suppose you are an economics or business major in a college close to the beach in the southern part of the US, for example southern California¹, where the weather is almost always nice the whole year around. In order to support yourself through college, you have started your own (weekend) business: an ice cream parlor on the beach. You have experienced that on hot weekends you usually sell more ice cream than on cold weekends. Also, you have recorded the average temperature and the sales of ice cream during eight weekends. Let Y_j be the sales of ice cream on weekend j , measured in \$100, and let X_j be the average temperature on weekend j , measured in units of 10 degrees Fahrenheit:

Table 1: *Ice cream data*

Sales (unit = \$100)	Temperature (unit = 10 degrees)
$Y_1 = 8$	$X_1 = 5$
$Y_2 = 10$	$X_2 = 7$
$Y_3 = 8$	$X_3 = 6$
$Y_4 = 13$	$X_4 = 8$
$Y_5 = 15$	$X_5 = 10$
$Y_6 = 14$	$X_6 = 9$
$Y_7 = 11$	$X_7 = 7$
$Y_8 = 9$	$X_8 = 8$

You want to use this information to forecast next weekend's sales of ice cream, given a good forecast of next weekend's temperature. Such a forecast of the sales will enable you to

¹ These lecture notes are based on lecture notes that I wrote while teaching at the University of California, San Diego, in the winter of 1987.

reduce your cost by adjusting your purchase of ice cream to the expected demand, because the ice cream you don't sell has to be thrown away.

Let your forecasting scheme be

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X,$$

i.e., given the temperature of X times 10 degrees and given the values of $\hat{\alpha}$ and $\hat{\beta}$, \hat{Y} times \$100 will be your forecast of the sales of ice cream. This forecasting scheme together with the points (X_j, Y_j) , $j = 1, 2, \dots, 8$, is plotted in Figure 1:

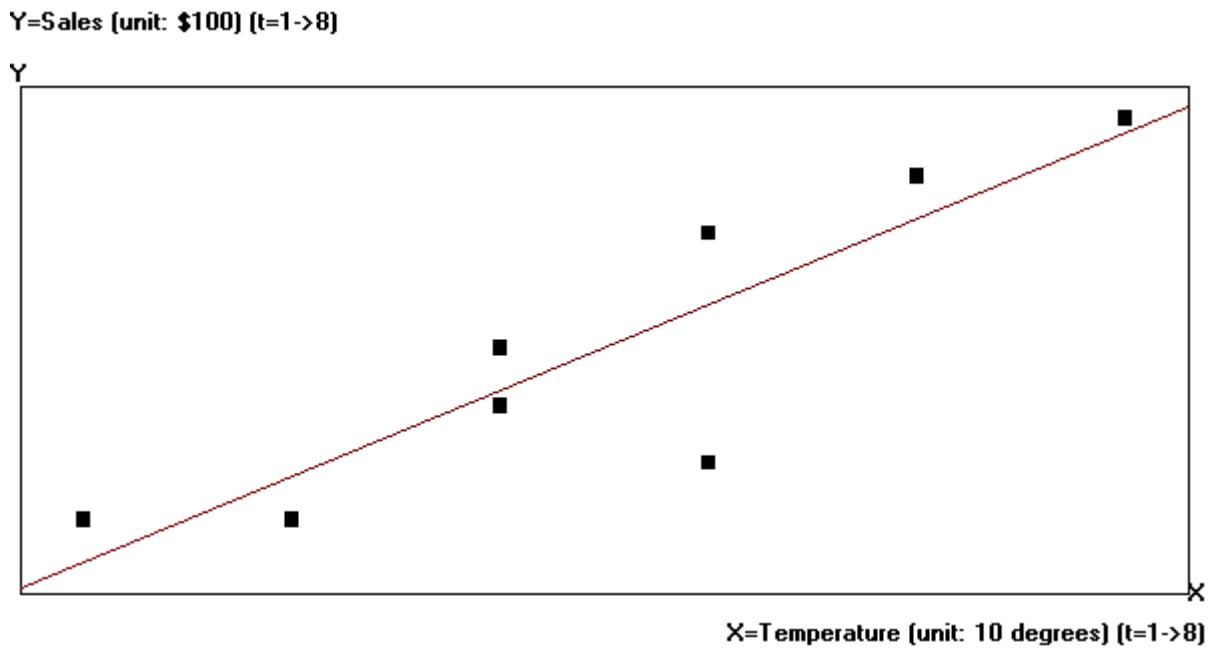


Figure 1 Scatter plot of (X_j, Y_j) , $j = 1, 2, \dots, 8$, together with the line $\hat{Y} = \hat{\alpha} + \hat{\beta}X$.

The best values for $\hat{\alpha}$ and $\hat{\beta}$ are those for which the forecast error (= actual sales minus forecasted sales) is minimal. However, you do not know yet the actual sales in the next weekend, but you do know the actual sales in the eight weekends for which you have recorded your sales and the corresponding temperature. So what you could do is to forecast the sales of ice cream on each of these eight weekends and to determine $\hat{\alpha}$ and $\hat{\beta}$ such that the forecast errors are minimal. Because forecast errors can be positive and negative, as can be seen from Figure 1, the sum of the forecast errors is not a good measure of the performance of your forecasting

scheme, because large positive errors can be offset by large negative errors. Therefore, use the sum of squared errors as your measure of the accuracy of your forecasts:

$$Q(\hat{\alpha}, \hat{\beta}) = \sum_{j=1}^n (Y_j - \hat{Y}_j)^2 = \sum_{j=1}^n (Y_j - \hat{\alpha} - \hat{\beta}X_j)^2,$$

where n is the sample size ($n = 8$ in our example), and minimize $Q(\hat{\alpha}, \hat{\beta})$ to $\hat{\alpha}$ and $\hat{\beta}$. It can be shown (see the Appendix) that $Q(\hat{\alpha}, \hat{\beta})$ is minimal for

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^n (X_j - \bar{X})^2} = \frac{\sum_{j=1}^n (X_j - \bar{X})Y_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \\ \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X}, \text{ where } \bar{X} = (1/n)\sum_{j=1}^n X_j \text{ and } \bar{Y} = (1/n)\sum_{j=1}^n Y_j.\end{aligned}\tag{1}$$

In the ice cream parlor case we have

$$\begin{aligned}n &= 8, \bar{X} = 7.5, \bar{Y} = 11, \sum_{j=1}^n X_j^2 = 468, \sum_{j=1}^n X_j Y_j = 687, \\ \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}) &= \sum_{j=1}^n X_j Y_j - n\bar{X}\bar{Y} = 27, \\ \sum_{j=1}^n (X_j - \bar{X})^2 &= \sum_{j=1}^n X_j^2 - n\bar{X}^2 = 18,\end{aligned}$$

so that

$$\hat{\beta} = 1.5, \quad \hat{\alpha} = -0.25.$$

Thus, our best forecasting scheme is $\hat{Y} = -0.25 + 1.5X$. This is the straight line in Figure 1.

Now suppose that the forecast of next weekend's temperature is 75 degrees. Then $X = 7.5$, hence $\hat{Y} = -0.25 + 1.5(7.5) = 11$. Therefore, the best forecast of next weekend's sales is:
 $\hat{Y} \times \$100 = \$1,100$.

2. The two-variable linear regression model.

In order to answer the question how good this forecast is, we have to make assumptions about the true relationship between the *dependent variable* Y_j and the *independent variable* X_j , (also called the *explanatory variable*). The true relationship we are going to assume is the two-variable linear regression model:

$$Y_j = \alpha + \beta \cdot X_j + U_j, \quad j = 1, 2, \dots, n.\tag{2}$$

The U_j 's are random error variables, called *error terms*, for which we assume:

Assumption I: *The U_j 's are independent and identically distributed (i.i.d) random variables.*

Assumption II: *The mathematical expectation of U_j equals zero: $E(U_j) = 0$ for $j = 1, 2, \dots, n$.*

Assumption III: *The variance $\sigma^2 = \text{var}(U_j) = E[(U_j - E(U_j))^2] = E[U_j^2]$ of the U_j 's is constant and finite.*

Regarding the explanatory variables X_j we shall assume for the time being that

Assumption IV: *The independent variables X_j are non-random.*

This assumption is not strictly necessary, and is actually quite unrealistic in economics, but will be made for the sake of convenience, as it will ease the argument. Finally, we will assume that the errors are normally distributed:

Assumption V: *The errors U_j 's are $N(0, \sigma^2)$ distributed.*

In particular, we shall need the latter assumption in order to say something about the reliability of the forecast. These assumptions will be relaxed later on.

3. *The properties of $\hat{\alpha}$ and $\hat{\beta}$.*

Although we have motivated model (2) by the need to forecast out-of-sample values of the dependent variables Y_j , a linear regression model is more often used for testing economic hypotheses. For example, let Y_j be the hourly wage of wage earner j in a random sample of size n of wage earners, and let X_j be a gender indicator, say $X_j = 1$ if person j is a female, and $X_j = 0$ if person j is a male. If you suspect gender discrimination in the workplace, you can test this suspicion by testing the null hypothesis that $\beta = 0$ (no gender discrimination) against one of

three possible alternative hypotheses:

- (a) $\beta \neq 0$: women are paid different hourly wages than men, either higher or lower;
- (b) $\beta > 0$: women are paid higher hourly wages than men;
- (c) $\beta < 0$: women are paid lower hourly wages than men.

The last hypothesis is usually what is meant by "gender discrimination." A test for the null hypothesis $\beta = 0$ against one of these alternative hypotheses can be based on the estimate $\hat{\beta}$ of β , provided that we know how $\hat{\beta}$ is related to β .

It will be shown below that $\hat{\alpha}$ and $\hat{\beta}$ are indeed reasonable approximations of α and β , respectively, possessing particular desirable properties.

In general an *estimator* of an unknown parameter is a function of the data that serves as an approximation of the parameter involved. It follows from (1) that $\hat{\alpha}$ and $\hat{\beta}$ are functions of the data, $(Y_1, X_1), \dots, (Y_n, X_n)$. Because $\hat{\alpha}$ and $\hat{\beta}$ will be used as approximations of α and β , respectively, and were obtained by minimizing the squared errors, we will call $\hat{\alpha}$ and $\hat{\beta}$ the Ordinary² Least Squares (OLS) estimators of α and β , respectively.

3.1 Unbiasedness

The first property of $\hat{\alpha}$ and $\hat{\beta}$ is that they are *unbiased* estimators of α and β :

Proposition 1. *Under Assumptions II and IV the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are unbiased, which means that $E[\hat{\alpha}] = \alpha$ and $E[\hat{\beta}] = \beta$.*

This result follows from the fact that we can write

$$\hat{\alpha} = \alpha + \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j, \quad \hat{\beta} = \beta + \frac{\sum_{j=1}^n (X_j - \bar{X}) U_j}{\sum_{i=1}^n (X_i - \bar{X})^2}. \quad (3)$$

See the Appendix.

² The estimators $\hat{\alpha}$ and $\hat{\beta}$ are called "Ordinary" least squares estimators to distinguish them from "Nonlinear" least squares estimators.

3.2 The variances of $\hat{\alpha}$ and $\hat{\beta}$.

Our next issue concerns the variances of $\hat{\alpha}$ and $\hat{\beta}$. For deriving these variances the following two lemmas are convenient.

Lemma 1. Let U_1, U_2, \dots, U_n be independent random variables with zero mathematical expectation (thus $E(U_j) = 0$) and variance σ^2 . (Thus $E[(U_j - E(U_j))^2] = E(U_j^2) = \sigma^2$). Let v_1, v_2, \dots, v_n and w_1, w_2, \dots, w_n be given constants. Then $E[(\sum_{j=1}^n v_j U_j)(\sum_{j=1}^n w_j U_j)] = \sigma^2 \sum_{j=1}^n v_j w_j$.

Proof. See the Appendix.

Note that if we choose $v_j = w_j$ for $j = 1, 2, \dots, n$ in Lemma 1 then it reads:

Lemma 2. Let U_1, U_2, \dots, U_n be independent random variables with zero mathematical expectation and variance σ^2 . Let w_1, w_2, \dots, w_n be given constants. Then $E[(\sum_{j=1}^n w_j U_j)^2] = \sigma^2 \sum_{j=1}^n w_j^2$.

Using (3) and Lemmas 1 and 2 it can be shown that

Proposition 2. Under the assumptions I - IV,

$$\begin{aligned} \text{var}(\hat{\alpha}) &= \frac{\sigma^2 \sum_{j=1}^n X_j^2}{n \sum_{j=1}^n (X_j - \bar{X})^2} = \sigma_{\hat{\alpha}}^2, \text{ say, } \text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{j=1}^n (X_j - \bar{X})^2} = \sigma_{\hat{\beta}}^2, \text{ say, and} \\ \text{cov}(\hat{\alpha}, \hat{\beta}) &= \frac{-\sigma^2 \bar{X}}{\sum_{j=1}^n (X_j - \bar{X})^2}. \end{aligned} \tag{4}$$

Proof. See the Appendix

3.3 Normality of $\hat{\alpha}$ and $\hat{\beta}$.

If we also assume normality of the error terms U_j then $\hat{\alpha}$ and $\hat{\beta}$ are also normally distributed. This result follows from the following lemma.

Lemma 3. Let Z_1, Z_2, \dots, Z_m be independent $N(\mu, \sigma^2)$ distributed random variables and let w_1, \dots, w_m be constants. Then $\sum_{j=1}^m w_j Z_j$ is distributed $N[(\sum_{j=1}^m w_j)\mu, (\sum_{j=1}^m w_j^2)\sigma^2]$.

The proof of this lemma requires advanced probability theory and is therefore omitted.

It follows now straightforwardly from Proposition 2, Lemma 3, and (3) that:

Proposition 3. Under the assumptions I - V,

$$\hat{\alpha} - \alpha \sim N\left[0, \frac{\sigma^2 \sum_{j=1}^n X_j^2}{n \sum_{j=1}^n (X_j - \bar{X})^2}\right], \quad \hat{\beta} - \beta \sim N\left[0, \frac{\sigma^2}{\sum_{j=1}^n (X_j - \bar{X})^2}\right], \quad (5)$$

where “~” is the symbol for “is distributed as.”

Moreover, applying Lemma 3 again for $m = 1$ it follows from (5) (Exercise: Why?) that

Proposition 4. Under the assumptions I - V,

$$\frac{(\hat{\alpha} - \alpha)\sqrt{n \sum_{j=1}^n (X_j - \bar{X})^2}}{\sigma \cdot \sqrt{\sum_{j=1}^n X_j^2}} \sim N[0, 1], \quad \frac{(\hat{\beta} - \beta)\sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}{\sigma} \sim N[0, 1]. \quad (6)$$

These results play a key-role in testing hypotheses about α and β . The only problem that prevents us from using these results for testing is that σ is unknown. This problem will be addressed in the next section.

4. How to estimate the error variance σ^2 ?

If α and β were known then we could estimate σ^2 by

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (Y_j - \alpha - \beta \cdot X_j)^2 = \frac{1}{n} \sum_{j=1}^n U_j^2. \quad (7)$$

However, α and β are not known, but we do have OLS estimators of α and β . This suggests to

replace α and β in (7) by their OLS estimators:

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (Y_j - \hat{\alpha} - \hat{\beta} \cdot X_j)^2 = \frac{1}{n} \sum_{j=1}^n \hat{U}_j^2, \quad (8)$$

where

$$\hat{U}_j = Y_j - \hat{\alpha} - \hat{\beta} \cdot X_j \quad (9)$$

is called the regression *residual*. However, the estimator (8) is biased, due to the fact that

Proposition 5. Under the assumptions I - V, $E[\sum_{j=1}^n \hat{U}_j^2] = (n - 2)\sigma^2$.

Proof: See the Appendix.

This result suggests to use

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{j=1}^n \hat{U}_j^2 \quad (10)$$

as an estimator of σ^2 instead of (8), because then by Proposition 5, $\hat{\sigma}^2$ is an unbiased estimator of σ^2 :

$$E[\hat{\sigma}^2] = \sigma^2. \quad (11)$$

The sum $\sum_{j=1}^n \hat{U}_j^2$ is called the Sum of Squares Residuals, shortly *SSR*, or also called the Residual Sum of Squares (RSS), and $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ is called the Standard Error of the Residuals, shortly *SER*. Thus,

$$SSR = \sum_{j=1}^n \hat{U}_j^2, \quad SER = \sqrt{\frac{\sum_{j=1}^n \hat{U}_j^2}{n-2}} = \sqrt{\frac{SSR}{n-2}} (= \hat{\sigma}). \quad (12)$$

Finally, note that the sum of squared residuals can be computed as follows:

$$SSR = \sum_{j=1}^n (Y_j - \bar{Y})^2 - \hat{\beta}^2 \sum_{j=1}^n (X_j - \bar{X})^2. \quad (13)$$

See the Appendix.

5. *Standard errors, t-values and p-values of the OLS estimators*

The variances of $\hat{\alpha}$ and $\hat{\beta}$ can now be estimated by replacing σ^2 in (4) by $\hat{\sigma}^2$:

$$\begin{aligned} \text{Estimated var}(\hat{\alpha}) &= \frac{\hat{\sigma}^2 \sum_{j=1}^n X_j^2}{n \sum_{j=1}^n (X_j - \bar{X})^2} = \hat{\sigma}_{\hat{\alpha}}^2, \text{ say,} \\ \text{Estimated var}(\hat{\beta}) &= \frac{\hat{\sigma}^2}{\sum_{j=1}^n (X_j - \bar{X})^2} = \hat{\sigma}_{\hat{\beta}}^2, \text{ say.} \end{aligned} \quad (14)$$

Then $\hat{\sigma}_{\hat{\alpha}} = \sqrt{\hat{\sigma}_{\hat{\alpha}}^2}$ is called the standard error of $\hat{\alpha}$, also denoted by $SE(\hat{\alpha})$, and $\hat{\sigma}_{\hat{\beta}} = \sqrt{\hat{\sigma}_{\hat{\beta}}^2}$ is called the standard error of $\hat{\beta}$, also denoted by $SE(\hat{\beta})$.

If we replace σ in Proposition 4 by the SER, $\hat{\sigma}$, the standard normality results involved change:

Proposition 6. *Under the assumptions I - V,*

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}} = \frac{(\hat{\alpha} - \alpha)\sqrt{n \sum_{j=1}^n (X_j - \bar{X})^2}}{\hat{\sigma} \cdot \sqrt{\sum_{j=1}^n X_j^2}} \sim t_{n-2}, \quad \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} = \frac{(\hat{\beta} - \beta)\sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}{\hat{\sigma}} \sim t_{n-2}. \quad (15)$$

The proof of Proposition 6 is based on the fact that under these assumptions, SSR/σ^2 is distributed χ_{n-2}^2 and is independent of $\hat{\alpha}$ and $\hat{\beta}$, but the proof involved requires advanced probability theory and is therefore omitted.

Because for large degrees of freedom the t distribution is approximately equal to the standard normal distribution, and due to the central limit theorem, Proposition 4 holds if n is large and the errors are not normally distributed, we also have:

Proposition 7. *If the sample size n is large then under the assumptions I - IV we have approximately,*

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}} = \frac{(\hat{\alpha} - \alpha)\sqrt{n \sum_{j=1}^n (X_j - \bar{X})^2}}{\hat{\sigma} \cdot \sqrt{\sum_{j=1}^n X_j^2}} \sim N(0,1), \quad \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} = \frac{(\hat{\beta} - \beta)\sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}{\hat{\sigma}} \sim N(0,1). \quad (16)$$

The results in Proposition 6 now enable us to test hypotheses about α and β . In particular the null hypothesis that $\beta = 0$ is of importance, because this hypothesis implies that X has no effect on Y . The test statistic for testing this hypothesis is the t-value (or t-statistic) of $\hat{\beta}$:

$$\hat{t}_{\hat{\beta}} \text{ } (= \text{ } t\text{-value of } \hat{\beta}) \stackrel{\text{def.}}{=} \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta} \sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}{\hat{\sigma}} \sim t_{n-2} \text{ if } \beta = 0. \quad (17)$$

If $\beta > 0$ and $n \rightarrow \infty$ then the t-value of $\hat{\beta}$ converges in probability to $+\infty$, and if $\beta < 0$ and $n \rightarrow \infty$ then the t-value of $\hat{\beta}$ converges in probability to $-\infty$. Moreover, if the sample size n is large then by Proposition 7 we may use the standard normal distribution instead of the t distribution to find critical values of the test.

Similarly,

$$\hat{t}_{\hat{\alpha}} \text{ } (= \text{ } t\text{-value of } \hat{\alpha}) \stackrel{\text{def.}}{=} \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}} \sim t_{n-2} \text{ if } \alpha = 0. \quad (18)$$

However, the hypothesis $\alpha = 0$ is often of no interest.

In the ice cream example,

$$\begin{aligned} \sum_{j=1}^n (X_j - \bar{X})^2 &= 18 \Rightarrow \sqrt{\sum_{j=1}^n (X_j - \bar{X})^2} = \sqrt{18} \approx 4.24264, \\ \sum_{j=1}^n (Y_j - \bar{Y})^2 &= \sum_{j=1}^n Y_j^2 - n \bar{Y}^2 = 1020 - 8 \times 11^2 = 52 \end{aligned}$$

and by (13),

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{j=1}^n \hat{U}_j^2 = \frac{1}{n-2} \sum_{j=1}^n (Y_j - \bar{Y})^2 - \hat{\beta}^2 \frac{1}{n-2} \sum_{j=1}^n (X_j - \bar{X})^2 \\ &= \frac{52 - (1.5)^2 \cdot 18}{8-2} = \frac{11.5}{6} \approx 1.916667 \Rightarrow \hat{\sigma} \approx 1.384437 \end{aligned}$$

Hence,

$$\hat{t}_{\hat{\beta}} = \frac{\hat{\beta} \sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}{\hat{\sigma}} = \frac{1.5 \times 4.24264}{1.384437} \approx 4.597 \quad (19)$$

Assuming that the conditions of Proposition 6 hold, the null hypothesis $H_0: \beta = 0$ can be tested

against the alternative hypothesis $H_1: \beta \neq 0$ using the two-sided t-test at say the 5% significance level, as follows. Under the null hypothesis, (19) is a random drawing from the t distribution with $n-2 = 6$ degrees of freedom. Look up in the table of the t distribution the value t_* such that for $T \sim t_6$, $P[|T| > t_*] = 0.05$. This value is $t_* = 2.447$. Then accept the null hypothesis if $-t_* = -2.447 \leq \hat{t}_\beta \leq 2.447 = t_*$, and reject the null hypothesis in favor of the alternative hypothesis if $|\hat{t}_\beta| > t_* = 2.447$. Thus, in the ice cream example we reject the null hypothesis $H_0: \beta = 0$ because $|\hat{t}_\beta| = 4.597 > 2.447 = t_*$.

This test is illustrated in Figure 2 below. The curved line in Figure 2 is the density of the t distribution with 6 degrees of freedom. The grey areas are each 0.025, so that the total grey area is 0.05.

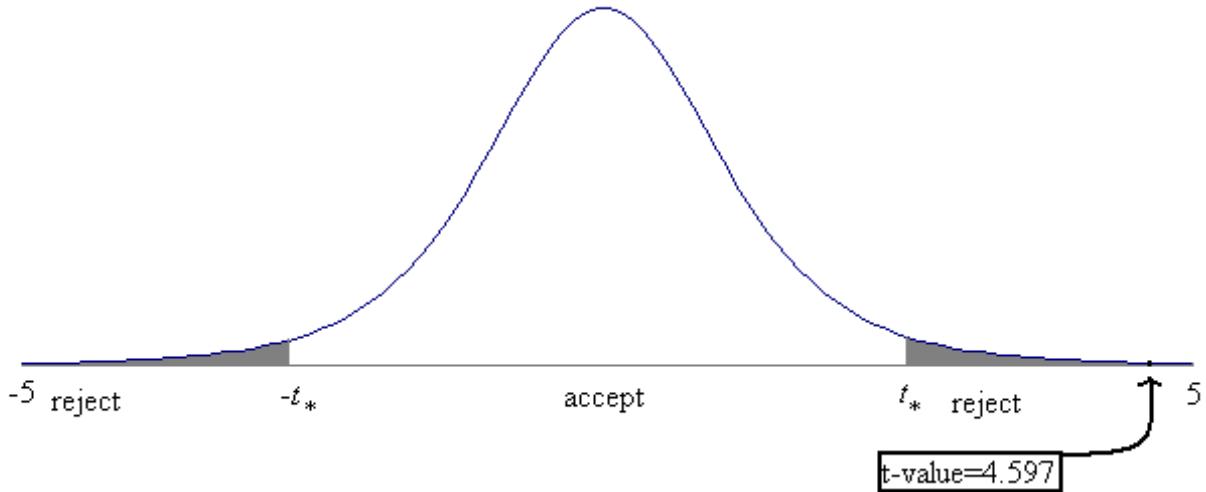


Figure 2 Two-sided t-test of $H_0: \beta = 0$ against the alternative hypothesis $H_1: \beta \neq 0$.

The null hypothesis $H_0: \beta = 0$ can be tested against the alternative hypothesis $H_1: \beta > 0$ at the 5% significance level by the right-sided t-test. Now look up in the table of the t distribution the value t_* such that for $T \sim t_6$, $P[T > t_*] = 0.05$. This value corresponds to the critical value of the two-sided t-test at the 10% significance level: $t_* = 1.943$. Then accept the null hypothesis if $\hat{t}_\beta \leq t_* = 1.943$, and reject the null hypothesis in favor of the alternative hypothesis if $\hat{t}_\beta > t_* = 1.943$. Thus, in the ice cream case we reject the null hypothesis

$H_0: \beta = 0$ in favor of the alternative hypothesis $H_1: \beta > 0$.

This right-sided t-test is illustrated in Figure 3 below. Again, the curved line in Figure 3 is the density of the t distribution with 6 degrees of freedom, and the grey area is 0.05.

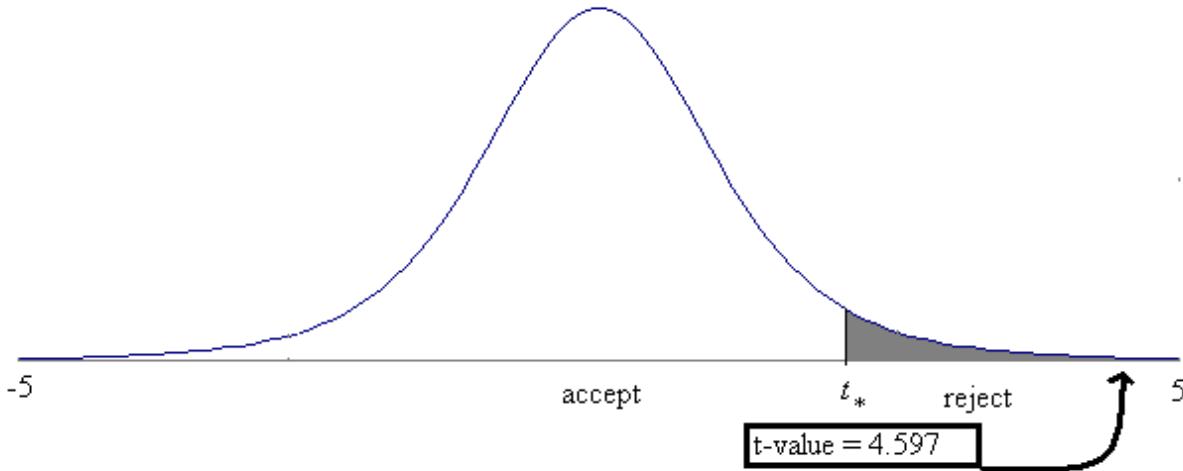


Figure 3 Right-sided t-test of $H_0: \beta = 0$ against the alternative hypothesis $H_1: \beta > 0$.

If the sample size n is large, so that $\hat{t}_\beta \sim N(0,1)$ if $\beta = 0$, then an alternative way of testing the null hypothesis $\beta = 0$ against the alternative hypothesis $\beta \neq 0$ is to use the (two-sided) p-value:

$$\hat{p}_\beta \text{ } (= \text{ } p\text{-value of } \hat{t}_\beta) \stackrel{\text{def.}}{=} P[|U| > |\hat{t}_\beta|], \text{ where } U \sim N(0,1). \quad (20)$$

For example, if $\hat{p}_\beta < 0.05$ we reject the null hypothesis $\beta = 0$ in favor of the alternative hypothesis $\beta \neq 0$ at the 5% significance level, and if $\hat{p}_\beta \geq 0.05$ we accept the null hypothesis $\beta = 0$. The p-value for $\hat{\alpha}$ is defined and used similarly.

Although a t-value is a test statistics of the null hypothesis that the corresponding coefficient in the regression model is zero, it is quite easy to rebuild the t-value for testing other null hypotheses, as follows. Suppose you want to test the null hypothesis that $\beta = \beta_0$, where β_0 is a given number, for example $\beta_0 = 1$. Then

$$\frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} - \frac{\beta_0}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} - \frac{\beta_0 \hat{\beta}}{\hat{\beta} \hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \left(1 - \frac{\beta_0}{\hat{\beta}} \right) = \frac{\hat{\beta} - \beta_0}{\hat{\beta}} \cdot \hat{t}_{\hat{\beta}}, \quad (21)$$

so that by Proposition 5,

$$\hat{t}_{\hat{\beta}, \beta=\beta_0} = \frac{\hat{\beta} - \beta_0}{\hat{\beta}} \cdot \hat{t}_{\hat{\beta}} \sim t_{n-2}. \quad (22)$$

For example, suppose that in the ice cream case we want to test the null hypothesis $H_0: \beta = 1$.

Then

$$\hat{t}_{\hat{\beta}, \beta=1} = \frac{\hat{\beta} - 1}{\hat{\beta}} \cdot \hat{t}_{\hat{\beta}} = \frac{1.5 - 1}{1.5} \times 4.597 \approx 1.532, \quad (23)$$

which under the null hypothesis $H_0: \beta = 1$ is a random drawing from the t distribution with 6 degrees of freedom. Note that the value of this test statistic is in the acceptance regions in Figures 2 and 3.

This trick is useful if the econometric software you are using only reports the t-values but not the standard errors. If the standard errors are reported, you can compute $\hat{t}_{\hat{\beta}, \beta=\beta_0}$ directly as $\hat{t}_{\hat{\beta}, \beta=\beta_0} = (\hat{\beta} - \beta_0)/\hat{\sigma}_{\hat{\beta}}$. Of course, if only the standard errors are reported and not the t-values you can compute the t-value of $\hat{\beta}$ as $\hat{t}_{\hat{\beta}} = \hat{\beta}/\hat{\sigma}_{\hat{\beta}}$.

6. The R^2

The R^2 of a regression model compares the sum of squared residuals (SSR) of the model with the SSR of a “regression model” without regressors:

$$Y_j = \alpha + U_j, \quad j = 1, 2, \dots, n. \quad (24)$$

It is easy to verify that the OLS estimator $\tilde{\alpha}$ of α is just the sample mean of the Y_j 's:

$$\tilde{\alpha} = \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j.$$

Therefore, the SSR of “regression model” (24) is $\sum_{j=1}^n (Y_j - \bar{Y})^2$, which is called the Total Sum of Squares (TSS), is

$$TSS = \sum_{j=1}^n (Y_j - \bar{Y})^2. \quad (26)$$

The R^2 is now defined as:

$$R^2 \stackrel{\text{def.}}{=} 1 - \frac{SSR}{TSS}. \quad (27)$$

The R^2 is always between zero and one, because $SSR \leq TSS$. (*Exercise: Why?*) If $SSR = TSS$, so that $R^2 = 0$, then model (24) explains the dependent variable Y_j ‘s equally well as model (2). In other words, the explanatory variables X_j in (2) do not matter: $\beta = 0$. The other extreme case is where $R^2 = 1$, which corresponds to $SSR = 0$. Then the dependent variable Y_j in model (2) is completely explained by X_j , without error: $Y_j \equiv \alpha + \beta X_j$. Thus, the R^2 measures how well the explanatory variables X_j are able to explain the corresponding dependent variables Y_j . For example, in the ice cream case, $SSR = 11.5$ and $TSS = 52$, hence $R^2 = 0.778846$. Loosely speaking, this means that about 78% of the variation of ice cream sales can be explained by the variation in temperature.

7. *Presenting regression results*

When you need to report regression results you should include, next to the OLS estimates of course, either the corresponding t-values or the standard errors, the sample size n , the standard error of the residuals (SER), and the R^2 , because this information will enable the reader to judge your results. For example, our ice cream estimation results should be displayed as either

$$\begin{aligned} Sales &= -0.25 + 1.5 \text{Temp.}, \quad n = 8, \quad SER = 1.384437, \quad R^2 = 0.778846 \\ &\quad (-0.100) \quad (4.597) \end{aligned}$$

(*t-values between brackets*)

or

$$Sales = -0.25 + 1.5 Temp., \quad n = 8, \quad SER = 1.384437, \quad R^2 = 0.778846$$

(2.49583) (0.32632)

(standard errors between brackets)

It is helpful to the reader if you would indicate whether you have displayed the t-values between brackets or the standard errors, but you only need to mention this once.

8. Out-of-sample forecasting

The linear regression model was introduced as a forecasting scheme. The question we now address is: How reliable is an out-of-sample forecast?

Consider the linear regression model (2), and suppose we observe X_{n+1} . Then the forecast of Y_{n+1} is $\hat{Y}_{n+1} = \hat{\alpha} + \hat{\beta} \cdot X_{n+1}$, where the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are computed on the basis of the observations for $j = 1, 2, \dots, n$. The actual but unknown value of Y_{n+1} is

$$Y_{n+1} = \alpha + \beta \cdot X_{n+1} + U_{n+1},$$

so that the forecast error is:

$$Y_{n+1} - \hat{Y}_{n+1} = U_{n+1} - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) \cdot X_{n+1} = U_{n+1} - \sum_{j=1}^n \left(\frac{1}{n} + \frac{(X_{n+1} - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \cdot U_j. \quad (28)$$

See the Appendix for the latter equality. It follows now from Lemma 3 that under Assumptions I through V, $Y_{n+1} - \hat{Y}_{n+1} \sim N[0, \sigma_{Y_{n+1} - \hat{Y}_{n+1}}^2]$, where

$$\sigma_{Y_{n+1} - \hat{Y}_{n+1}}^2 = \sigma^2 \left(\frac{n+1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right). \quad (29)$$

See the Appendix. Denoting,

$$\hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}}^2 = \hat{\sigma}^2 \left(\frac{n+1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right), \quad (30)$$

it follows now similar to Proposition 6 that

Proposition 8. Under assumptions I - V, $(Y_{n+1} - \hat{Y}_{n+1})/\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}} \sim t_{n-2}$.

This result can be used to construct a 95% confidence interval, say, of Y_{n+1} . Look up in the table of the t distribution the critical value t_* of the two-sided t-test with $n-2$ degrees of freedom. Then it follows from Proposition 7 that

$$\begin{aligned} 0.95 &= P[-t_* \leq (Y_{n+1} - \hat{Y}_{n+1})/\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}} \leq t_*] \\ &= P[-t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}} \leq Y_{n+1} - \hat{Y}_{n+1} \leq t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}] \\ &= P[\hat{Y}_{n+1} - t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}} \leq Y_{n+1} \leq \hat{Y}_{n+1} + t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}] \end{aligned} \quad (31)$$

Thus, the 95% confidence interval of Y_{n+1} is $[\hat{Y}_{n+1} - t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}, \hat{Y}_{n+1} + t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}]$.

Observe from (30) that $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}$ increases with $(X_{n+1} - \bar{X})^2$, and so does the width of the confidence interval. Thus, the farther X_{n+1} is away from \bar{X} , the more unreliable the forecast \hat{Y}_{n+1} of Y_{n+1} becomes. Also observe from (30) that $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}} \geq \hat{\sigma}$, and that $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}$ gets close to $\hat{\sigma}$ if n is large because $\lim_{n \rightarrow \infty} \sum_{j=1}^n (X_j - \bar{X})^2 = \infty$.

9. Relaxing the non-random regressor assumption

As said before, the assumption that the regressors X_j are non-random is too strong an assumption in economics. Therefore, we now assume that the X_j 's are random variables. This requires the following modifications of the Assumptions I-V:

Assumption I*: The pairs (X_j, Y_j) , $j = 1, 2, 3, \dots, n$, are independent and identically distributed.

Assumption II*: The conditional expectations $E[U_j | X_j]$ are equal to zero: $E[U_j | X_j] \equiv 0$.

Assumption III*: The conditional expectations $E[U_j^2 | X_j]$ do not depend on the X_j 's and are finite, constant and equal: $E[U_j^2 | X_j] \equiv \sigma^2 < \infty$. (This is called the **homoscedasticity assumption**.)

Assumption IV*: *Conditional on X_j , U_j is $N(0, \sigma^2)$ distributed.*

The Assumptions I* and II* imply that for $j = 1, \dots, n$,

$$E[U_j | X_1, X_2, \dots, X_n] \equiv 0, \quad (32)$$

and similarly the Assumptions I* and III* imply that for $j = 1, \dots, n$,

$$E[U_j^2 | X_1, X_2, \dots, X_n] \equiv \sigma^2. \quad (33)$$

Because (loosely speaking) conditioning on X_1, X_2, \dots, X_n is effectively the same as treating them as given constants, most of the previous propositions carry over:

Proposition 9. *Under Assumptions I*-IV*, Propositions 1 and 4 through 7 carry over, and the results in Propositions 2 and 3 now hold conditional on X_1, X_2, \dots, X_n .*

However, without Assumption IV* we need an additional condition in Proposition 6 in order to use the central limit theorem, namely:

Proposition 10. *If the sample size n is large then under the assumptions I* - III* and the additional condition $E[X_j^2] < \infty$ the approximate normality results in Proposition 7 carry over.*

Moreover, without Assumption IV* the Propositions 6 and 8 are no longer true. As to Proposition 6, this not a big deal, as in large samples we can still use Proposition 7, but without Assumption IV* we can no longer derive confidence intervals for the forecasts, as these confidence intervals are based on Proposition 8. It is therefore important to test the normality assumption.

10. Testing the normality assumption

For a normal random variable U with zero expectation and variance σ^2 it can be shown that

$$\begin{aligned}
\text{Kurtosis} & \stackrel{\text{def.}}{=} E[U^4]/\sigma^4 - 3 = 0, \\
\text{Skewness} & \stackrel{\text{def.}}{=} E[U^3] = 0
\end{aligned} \tag{34}$$

Therefore, the normality condition can be tested by testing whether the kurtosis and the skewness of the model errors are zero, using the residuals. This is the idea behind the Jarque-Bera³ and Kiefer-Salmon⁴ tests. Under the null hypothesis (34) the test statistic involved has a χ_2^2 distribution

11. Heteroscedasticity⁵

We say that the errors U_j of regression model (2) are heteroskedastic if assumption III* does not hold:

$$E[U_j^2|X_j] = \Psi(X_j) \text{ for some function } \Psi(.). \tag{35}$$

Heteroscedasticity often occurs in practice. It is actually the rule rather than the exception. The main consequence of heteroscedasticity is that the conditional variance formulas in Propositions 2 and 3 do no longer hold, although the unbiasedness result in Proposition 1 is not affected by heteroscedasticity. Therefore, the Propositions 4-8 are no longer valid as well. In particular, the conditional variance of $\hat{\beta}$ [see (60)] under heteroscedasticity takes the form

$$\text{var}(\hat{\beta}|X_1, \dots, X_n) = E[(\hat{\beta} - \beta)^2|X_1, \dots, X_n] = \frac{\sum_{j=1}^n (X_j - \bar{X})^2 \Psi(X_j)}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2}. \tag{36}$$

A cure for the heteroscedasticity problem is to replace the standard error of $\hat{\beta}$ by

³ Jarque, C.M. and A.K. Bera, (1980), "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals". *Economics Letters* 6, 255--259.

⁴ Kiefer, N. and M. Salmon (1983), "Testing Normality in Econometric Models", *Economic Letters* 11, 123-127.

⁵ Also spelled as "Heteroskedasticity."

$$\tilde{\sigma}_{\hat{\beta}} = \sqrt{\left(\frac{n}{n-2}\right) \frac{\sum_{j=1}^n (X_j - \bar{X})^2 \hat{U}_j^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2}}. \quad (37)$$

This is known as the Heteroscedasticity Consistent (H.C.) standard error. The H.C. t-value then becomes $\tilde{t}_{\hat{\beta}} = \hat{\beta}/\tilde{\sigma}_{\hat{\beta}}$. Under the null hypothesis $\beta = 0$ this t-value is no longer t distributed, but the standard normal approximation remains valid if the sample size n is large.

A popular test for heteroscedasticity is the Breusch-Pagan⁶ test. Given that

$$E[U_j^2 | X_j] = g(\gamma_0 + \gamma_1 X_j) \text{ for some unknown function } g(.). \quad (38)$$

the Breusch-Pagan test tests the null hypothesis

$$H_0: \gamma_1 = 0 \Leftrightarrow E[U_j^2 | X_j] = g(\gamma_0) = \sigma^2, \text{ say} \quad (39)$$

against the alternative hypothesis

$$H_0: \gamma_1 \neq 0 \Leftrightarrow E[U_j^2 | X_j] = g(\gamma_0 + \gamma_1 X_j) = \psi(X_j), \text{ say}. \quad (40)$$

Under the null hypothesis (39) of homoskedasticity the test statistic of the Breusch-Pagan test has a χ^2_1 distribution⁷, and the test is conducted right-sided.

12. How close are OLS estimators?

The ice cream data in Table 1 is not based on any actual observations on sales and temperature; I have picked the numbers for X_j and Y_j quite arbitrarily. Therefore, there is no way to find out how close the OLS estimates $\hat{\alpha} = -0.25$, $\hat{\beta} = 1.5$ are to the unknown parameters α and β . Actually, we do not know either whether the linear regression model (2) and its assumptions are applicable to this artificial data.

In order to show how well OLS estimators approximate the corresponding parameters I

⁶ Breusch, T. and A. Pagan (1979), "A Simple Test for Heteroscedasticity and Random Coefficient Variation", *Econometrica* 47, 1287-1294.

⁷ In the multiple regression case the degrees of freedom is equal to the number of parameters minus 1 for the intercept.

have generated random samples⁸ $(Y_1, X_1), \dots, (Y_n, X_n)$ for three sample sizes: $n = 10$, $n = 100$ and $n = 1000$, as follows. The explanatory variables X_j have been drawn independently from the χ^2_1 distribution, the regression errors U_j have been drawn independently from the $N(0,1)$ distribution, and the Y_j 's have been generated by

$$Y_j = 1 + X_j + U_j, \quad j = 1, 2, \dots, n. \quad (41)$$

Thus, in this case the parameters α and β in model (2) are $\alpha = 1$ and $\beta = 1$, and the standard error of U_j is $\sigma = 1$. Moreover, note that the Assumptions I*-IV* hold for model (41).

The true R^2 can be defined by

$$R_0^2 = 1 - \frac{E[SSR]}{E[TSS]} = 1 - \frac{(n-2)\sigma^2}{\sum_{j=1}^n E[(Y_j - \bar{Y})^2]}.$$

In the case (41), $\sigma^2 = 1$, $\mu_Y = E(Y_j) = 1 + E(X_j) = 2$,

$$\sum_{j=1}^n E[(Y_j - \bar{Y})^2] = E\left[\sum_{j=1}^n ((Y_j - \mu_Y) - (\bar{Y} - \mu_Y))^2\right] = E\left[\sum_{j=1}^n (Y_j - \mu_Y)^2 - n(\bar{Y} - \mu_Y)^2\right] = (n-1)\text{var}(Y_j)$$

and

$$\text{var}(Y_j) = E[(X_j - 1 + U_j)^2] = E[(X_j - 1)^2] + E[U_j^2] = E[(X_j - 1)^2] + 1 = 3,$$

because X_j is χ^2_1 distributed and therefore has the same distribution as U_j^2 , and it can be shown that for standard normal random variables U_j , $E[(U_j^2 - 1)^2] = 2$. Thus, the true R^2 in this case is

$$R_0^2 = 1 - \frac{n-2}{3(n-1)} = \frac{2n-1}{3n-3} \approx \begin{cases} 0.7037 & \text{for } n = 10 \\ 0.6700 & \text{for } n = 100 \\ 0.6670 & \text{for } n = 1000 \end{cases}$$

The estimation results involved are given in Table 2:

⁸ Via the *EasyReg International* menus File → Choose an input file → Create artificial data. Rather than generating one random sample of size $n = 1000$ and then using subsamples of sizes $n = 10$ and $n = 100$, these samples have been generated separately for $n = 10$, $n = 100$ and $n = 1000$.

Table 2: Artificial regression estimation results

	$\hat{\beta}$	$\hat{\alpha}$	SER (= $\hat{\sigma}$)	R^2	n
estimate:	1.11748	0.55912	0.919045	0.8842	10
(t-value):	(7.817)	(1.675)			
estimate:	1.03309	0.96028	0.992502	0.8284	100
(t-value):	(21.753)	(8.237)			
estimate:	1.02360	0.98518	0.983608	0.6899	1000
(t-value):	(47.124)	(26.037)			

Even for a sample size of $n = 10$ the OLS estimator $\hat{\beta}$ is already pretty close to its true value 1, and the same applies to $\hat{\sigma}$, but $\hat{\alpha}$ is too far away from the true value $\alpha = 1$. However, for $n = 100$ the OLS estimators $\hat{\beta}$ and $\hat{\alpha}$ deviate only about $\pm 4\%$ from their true values $\alpha = \beta = 1$, and $\hat{\sigma}$ deviates about -1% from its true value 1. In the case $n = 1000$ these deviations reduce to about $\pm 2\%$. The R^2 's are too high, and only for $n = 1000$ is the R^2 reasonably close to its true value. However, the R^2 is only a descriptive statistic; it does not play a role in hypotheses testing, so that the unreliability of the R^2 in small samples is harmless.

Notice the quite dramatic increase of the t-values. Recall that these t-values are the test statistics of the null hypotheses that the corresponding parameters are zero. Because the true parameters are equal to 1, what you see in Table 2 is the increase of the power of the t-test with the sample size.

APPENDIX

Proof of (1):

The first-order conditions for a minimum of $Q(\hat{\alpha}, \hat{\beta}) = \sum_{j=1}^n (Y_j - \hat{\alpha} - \hat{\beta}X_j)^2$ are:

$$\begin{aligned}
dQ(\hat{\alpha}, \hat{\beta})/d\hat{\alpha} = 0 &\Leftrightarrow \sum_{j=1}^n 2(Y_j - \hat{\alpha} - \hat{\beta}X_j)(-1) = 0 \\
&\Leftrightarrow \sum_{j=1}^n (Y_j - \hat{\alpha} - \hat{\beta}X_j) = 0 \\
&\Leftrightarrow \sum_{j=1}^n Y_j - \sum_{j=1}^n \hat{\alpha} - \sum_{j=1}^n (\hat{\beta}X_j) = 0 \\
&\Leftrightarrow \sum_{j=1}^n Y_j = n\hat{\alpha} + \hat{\beta}\sum_{j=1}^n X_j = 0 \\
&\Leftrightarrow \bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X},
\end{aligned} \tag{42}$$

and

$$\begin{aligned}
dQ(\hat{\alpha}, \hat{\beta})/d\hat{\beta} = 0 &\Leftrightarrow \sum_{j=1}^n 2(Y_j - \hat{\alpha} - \hat{\beta}X_j)(-X_j) = 0 \\
&\Leftrightarrow \sum_{j=1}^n (Y_j X_j - \hat{\alpha}X_j - \hat{\beta}X_j^2) = 0 \\
&\Leftrightarrow \sum_{j=1}^n X_j Y_j - \hat{\alpha} \sum_{j=1}^n X_j - \hat{\beta} \sum_{j=1}^n X_j^2 = 0 \\
&\Leftrightarrow \sum_{j=1}^n X_j Y_j = \hat{\alpha} \sum_{j=1}^n X_j + \hat{\beta} \sum_{j=1}^n X_j^2 \\
&\Leftrightarrow \frac{1}{n} \sum_{j=1}^n X_j Y_j = \hat{\alpha} \bar{X} + \hat{\beta} \frac{1}{n} \sum_{j=1}^n X_j^2
\end{aligned} \tag{43}$$

where $\bar{X} = (1/n)\sum_{j=1}^n X_j$ and $\bar{Y} = (1/n)\sum_{j=1}^n Y_j$ are the sample means of the X_j 's and Y_j 's, respectively. The last equations in (42) and (43) are called the *normal equations*:

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}, \tag{44}$$

$$\frac{1}{n} \sum_{j=1}^n X_j Y_j = \hat{\alpha} \bar{X} + \hat{\beta} \frac{1}{n} \sum_{j=1}^n X_j^2. \tag{45}$$

To solve these normal equations, substitute $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ in (45). Then we get

$$\begin{aligned}
\frac{1}{n} \sum_{j=1}^n X_j Y_j &= (\bar{Y} - \hat{\beta} \bar{X}) \frac{1}{n} \sum_{j=1}^n X_j + \hat{\beta} \frac{1}{n} \sum_{j=1}^n X_j^2 \\
&= \bar{Y} \cdot \bar{X} - \hat{\beta} \bar{X}^2 + \hat{\beta} \frac{1}{n} \sum_{j=1}^n X_j^2 \\
&= \bar{X} \cdot \bar{Y} + \hat{\beta} \left(\frac{1}{n} \sum_{j=1}^n X_j^2 - \bar{X}^2 \right)
\end{aligned}$$

hence

$$\frac{1}{n} \sum_{j=1}^n X_j Y_j - \bar{X} \cdot \bar{Y} = \hat{\beta} \left(\frac{1}{n} \sum_{j=1}^n X_j^2 - \bar{X}^2 \right). \quad (46)$$

Equation (46) can also be written as

$$\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}) = \hat{\beta} \left(\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 \right), \quad (47)$$

because

$$\begin{aligned}
\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}) &= \frac{1}{n} \sum_{j=1}^n (X_j Y_j - \bar{X} \cdot Y_j - X_j \cdot \bar{Y} + \bar{X} \cdot \bar{Y}) \\
&= \frac{1}{n} \sum_{j=1}^n X_j Y_j - \bar{X} \cdot \frac{1}{n} \sum_{j=1}^n Y_j - \bar{Y} \cdot \frac{1}{n} \sum_{j=1}^n X_j + \bar{X} \cdot \bar{Y} \\
&= \frac{1}{n} \sum_{j=1}^n X_j Y_j - \bar{X} \cdot \bar{Y}
\end{aligned} \quad (48)$$

and similarly

$$\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 = \frac{1}{n} \sum_{j=1}^n X_j^2 - \bar{X}^2. \quad (49)$$

Moreover,

$$\begin{aligned}
\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}) &= \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}) Y_j - \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}) \bar{Y} \\
&= \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}) Y_j - (\bar{X} - \bar{X}) \bar{Y} \\
&= \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}) Y_j
\end{aligned} \quad (50)$$

The result (1) now follows from (44) and (46) through (50).

Proof of Proposition 1.

Recall from (1) that

$$\hat{\beta} = \frac{\sum_{j=1}^n (X_j - \bar{X}) Y_j}{\sum_{j=1}^n (X_j - \bar{X})^2}. \quad (51)$$

Substitute model (2) in (51). Then

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{j=1}^n (X_j - \bar{X})(\alpha + \beta X_j + U_j)}{\sum_{j=1}^n (X_j - \bar{X})^2} \\ &= \frac{\alpha \sum_{j=1}^n (X_j - \bar{X}) + \beta \sum_{j=1}^n (X_j - \bar{X}) X_j + \sum_{j=1}^n (X_j - \bar{X}) U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \\ &= \beta \cdot \frac{\sum_{j=1}^n (X_j - \bar{X}) X_j}{\sum_{j=1}^n (X_j - \bar{X})^2} + \frac{\sum_{j=1}^n (X_j - \bar{X}) U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \\ &= \beta + \frac{\sum_{j=1}^n (X_j - \bar{X}) U_j}{\sum_{j=1}^n (X_j - \bar{X})^2}, \end{aligned} \quad (52)$$

where the last step follows from the fact that similar to (50),

$$\sum_{j=1}^n (X_j - \bar{X})^2 = \sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X}) = \sum_{j=1}^n (X_j - \bar{X}) X_j. \quad (53)$$

Now take the mathematical expectation at both sides of (52). Then,

$$E[\hat{\beta}] = \beta + E\left(\frac{\sum_{j=1}^n (X_j - \bar{X}) U_j}{\sum_{j=1}^n (X_j - \bar{X})^2}\right) = \beta + \frac{\sum_{j=1}^n (X_j - \bar{X}) E(U_j)}{\sum_{j=1}^n (X_j - \bar{X})^2} = \beta, \quad (54)$$

because taking the mathematical expectation of a constant (β) does not effect that constant, and taking the mathematical expectation of a linear function of random variables is equal to taking the linear function of the mathematical expectation of these random variables. The last conclusion in (54) follows from assumption II, and the second step in (54) can be taken because

we have assumed that the X_j 's are non-random (assumption IV).

Next consider $\hat{\alpha}$. We have already established that $\hat{\alpha} = \bar{Y} - \hat{\beta} \cdot \bar{X}$. Substituting the right-hand side of (52) for $\hat{\beta}$ in this equation yields

$$\hat{\alpha} = \bar{Y} - \left(\beta + \frac{\sum_{j=1}^n (X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \right) \cdot \bar{X} = \bar{Y} - \beta \cdot \bar{X} - \frac{\sum_{j=1}^n \bar{X}(X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2}. \quad (55)$$

Substituting

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j = \frac{1}{n} \sum_{j=1}^n (\alpha + \beta X_j + U_j) = \alpha + \beta \cdot \bar{X} + \frac{1}{n} \sum_{j=1}^n U_j$$

in (55) yields

$$\hat{\alpha} = \alpha + \frac{1}{n} \sum_{j=1}^n U_j - \frac{\sum_{j=1}^n \bar{X}(X_j - \bar{X})U_j}{\sum_{i=1}^n (X_i - \bar{X})^2} = \alpha + \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j. \quad (56)$$

Similar as for $\hat{\beta}$ we therefore have:

$$E[\hat{\alpha}] = \alpha + \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) E[U_j] = \alpha. \quad (57)$$

This completes the proof of Proposition 1.

Proof of Lemma 1:

We have

$$\begin{aligned} E\left[\left(\sum_{j=1}^n v_j U_j\right)\left(\sum_{j=1}^n w_j U_j\right)\right] &= E\left[\sum_{i=1}^n \sum_{j=1}^n v_i w_j U_i U_j\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n v_i w_j E(U_i U_j) \\ &= \sum_{j=1}^n v_j w_j \sigma^2, \end{aligned} \quad (58)$$

where the last equality in (58) follows from

$$\begin{aligned} E(U_i U_j) &= E(U_i) E(U_j) = 0 \text{ if } i \neq j, \\ &= E(U_i^2) = \sigma^2 \text{ if } i = j. \end{aligned} \quad (59)$$

Proof of Proposition 2:

It follows from formula (52) and Lemma 2 that

$$\begin{aligned}
 \text{var}(\hat{\beta}) &= E[(\hat{\beta} - \beta)^2] \\
 &= E\left[\left(\sum_{j=1}^n \left(\frac{X_j - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j \right)^2 \right] = \sigma^2 \sum_{j=1}^n \left(\frac{X_j - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)^2 \\
 &= \sigma^2 \frac{\sum_{j=1}^n (X_j - \bar{X})^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} = \sigma^2 \frac{\sum_{j=1}^n (X_j - \bar{X})^2}{\left(\sum_{j=1}^n (X_j - \bar{X})^2 \right)^2} = \frac{\sigma^2}{\sum_{j=1}^n (X_j - \bar{X})^2}.
 \end{aligned} \tag{60}$$

Similarly, it follows from formula (56) and Lemma 2 that

$$\begin{aligned}
 \text{var}(\hat{\alpha}) &= E[(\hat{\alpha} - \alpha)^2] \\
 &= E\left[\left(\sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j \right)^2 \right] = \sigma^2 \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)^2 \\
 &= \sigma^2 \sum_{j=1}^n \left(\frac{1}{n^2} - \frac{\frac{2}{n}\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{\bar{X}^2(X_j - \bar{X})^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} \right) \\
 &= \sigma^2 \left(\frac{1}{n} - \frac{2\bar{X}(1/n)\sum_{j=1}^n (X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{\bar{X}^2 \sum_{j=1}^n (X_j - \bar{X})^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} \right) \\
 &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right) \\
 &= \sigma^2 \left(\frac{(1/n)\sum_{j=1}^n (X_j - \bar{X})^2 + \bar{X}^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right) = \frac{\sigma^2 \sum_{j=1}^n X_j^2}{n \sum_{j=1}^n (X_j - \bar{X})^2},
 \end{aligned} \tag{61}$$

where the last equality follows from the fact that $(1/n)\sum_{j=1}^n (X_j - \bar{X})^2 = (1/n)\sum_{j=1}^n X_j^2 - \bar{X}^2$.

Finally, it follows from Lemma 1 and the formulas (52) and (56) that

$$\begin{aligned}
\text{cov}(\hat{\alpha}, \hat{\beta}) &= E[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)] = E\left[\left(\sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) U_j\right) \left(\sum_{j=1}^n \left(\frac{X_j - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) U_j\right)\right] \\
&= \sigma^2 \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \left(\frac{(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)
\end{aligned} \tag{62}$$

which can be rewritten as

$$\text{cov}(\hat{\alpha}, \hat{\beta}) = \sigma^2 \left(\frac{(1/n)\sum_{j=1}^n (X_j - \bar{X}) - \bar{X} \sum_{j=1}^n (X_j - \bar{X})^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2} \right) = \frac{-\sigma^2 \bar{X}}{\sum_{j=1}^n (X_j - \bar{X})^2}. \tag{63}$$

Proof of Proposition 5.

Observe first from (44) and (9) that

$$\frac{1}{n} \sum_{j=1}^n \hat{U}_j = \bar{Y} - \hat{\alpha} - \hat{\beta} \cdot \bar{X} = 0 \tag{64}$$

so that we can write

$$\hat{U}_j = \hat{U}_j - \frac{1}{n} \sum_{i=1}^n \hat{U}_i = (Y_j - \bar{Y}) - \hat{\beta} \cdot (X_j - \bar{X}). \tag{65}$$

Next, observe from (2) that $Y_j - \bar{Y} = U_j - \bar{U} + \beta \cdot (X_j - \bar{X})$, where $\bar{U} = (1/n)\sum_{j=1}^n U_j$.

Substituting the former equation in (65) yields

$$\hat{U}_j = (U_j - \bar{U}) - (\hat{\beta} - \beta)(X_j - \bar{X}), \tag{66}$$

hence

$$\begin{aligned}
\sum_{j=1}^n \hat{U}_j^2 &= \sum_{j=1}^n \left((U_j - \bar{U}) - (\hat{\beta} - \beta)(X_j - \bar{X}) \right)^2 \\
&= \sum_{j=1}^n (U_j - \bar{U})^2 - 2(\hat{\beta} - \beta) \sum_{j=1}^n (X_j - \bar{X})(U_j - \bar{U}) + (\hat{\beta} - \beta)^2 \sum_{j=1}^n (X_j - \bar{X})^2 \\
&= \sum_{j=1}^n (U_j - \bar{U})^2 - 2(\hat{\beta} - \beta) \sum_{j=1}^n (X_j - \bar{X})U_j + (\hat{\beta} - \beta)^2 \sum_{j=1}^n (X_j - \bar{X})^2,
\end{aligned} \tag{67}$$

where the last equality follows from the fact that $\sum_{j=1}^n (X_j - \bar{X})\bar{U} = 0$. It follows from (52), (67) and the equality $\sum_{j=1}^n (U_j - \bar{U})^2 = \sum_{j=1}^n U_j^2 - n\bar{U}^2$ that

$$\begin{aligned}\sum_{j=1}^n \hat{U}_j^2 &= \sum_{j=1}^n (U_j - \bar{U})^2 - (\hat{\beta} - \beta)^2 \sum_{j=1}^n (X_j - \bar{X})^2 = \sum_{j=1}^n U_j^2 - n\bar{U}^2 - (\hat{\beta} - \beta)^2 \sum_{j=1}^n (X_j - \bar{X})^2. \\ &= \sum_{j=1}^n U_j^2 - \frac{1}{n} \left(\sum_{i=1}^n U_i \right)^2 - (\hat{\beta} - \beta)^2 \sum_{j=1}^n (X_j - \bar{X})^2.\end{aligned}\tag{68}$$

Taking expectations and using Lemma 2 and Proposition 2 it follows now from (68) that

$$\begin{aligned}E[\sum_{j=1}^n \hat{U}_j^2] &= \sum_{j=1}^n E[U_j^2] - \frac{1}{n} E\left[\left(\sum_{i=1}^n U_i\right)^2\right] - (E(\hat{\beta} - \beta)^2) \sum_{j=1}^n (X_j - \bar{X})^2 \\ &= n\sigma^2 - \sigma^2 - \sigma^2 = (n-2)\sigma^2.\end{aligned}\tag{69}$$

Proof of (13):

$$\begin{aligned}SSR &= \sum_{j=1}^n \hat{U}_j^2 = \sum_{j=1}^n (Y_j - \hat{\alpha} - \hat{\beta} \cdot X_j)^2 = \sum_{j=1}^n (Y_j - (\bar{Y} - \hat{\beta} \cdot \bar{X}) - \hat{\beta} \cdot X_j)^2 \\ &= \sum_{j=1}^n ((Y_j - \bar{Y}) - \hat{\beta} \cdot (X_j - \bar{X}))^2 \\ &= \sum_{j=1}^n (Y_j - \bar{Y})^2 - 2\hat{\beta} \sum_{j=1}^n (Y_j - \bar{Y})(X_j - \bar{X}) + \hat{\beta}^2 \sum_{j=1}^n (X_j - \bar{X})^2 \\ &= \sum_{j=1}^n (Y_j - \bar{Y})^2 - \hat{\beta}^2 \sum_{j=1}^n (X_j - \bar{X})^2.\end{aligned}\tag{70}$$

Proof of (28):

It follows from (3) that

$$\begin{aligned}Y_{n+1} - \hat{Y}_{n+1} &= U_{n+1} - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) \cdot X_{n+1} \\ &= U_{n+1} - \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \cdot U_j - \sum_{j=1}^n \left(\frac{X_{n+1}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j \\ &= U_{n+1} - \sum_{j=1}^n \left(\frac{1}{n} + \frac{(X_{n+1} - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \cdot U_j.\end{aligned}\tag{71}$$

Proof of (29):

It follows from (28) and Lemma 3 that

$$\begin{aligned}
 \sigma_{Y_{n+1} - \hat{Y}_{n+1}}^2 &= \sigma^2 + \sum_{j=1}^n \left(\frac{1}{n} + \frac{(X_{n+1} - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)^2 \cdot \sigma^2 \\
 &= \sigma^2 \left(1 + \frac{1}{n} + \frac{2}{n} \cdot \frac{(X_{n+1} - \bar{X}) \sum_{j=1}^n (X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{(X_{n+1} - \bar{X})^2 \sum_{j=1}^n (X_j - \bar{X})^2}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \right) \quad (72) \\
 &= \sigma^2 \left(\frac{n+1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right).
 \end{aligned}$$

MULTIVARIATE LINEAR REGRESSION

Herman J. Bierens

Pennsylvania State University

February 13, 2010

1. *Missing variables*

Suppose you assume that the relationship between a dependent variable Y_j and an explanatory variable X_j for observations $j = 1, \dots, n$ is given by

$$Y_j = \alpha + \beta \cdot X_j + U_j, \quad j = 1, 2, \dots, n, \quad (1)$$

whereas in reality

$$Y_j = \alpha + \beta \cdot X_j + \gamma \cdot Z_j + U_j, \quad j = 1, 2, \dots, n, \quad (2)$$

where Z_j is a missing explanatory variable, and U_j is the error term. Recall that the OLS estimator of β , assuming that model (1) is correct, is

$$\hat{\beta} = \frac{\sum_{j=1}^n (X_j - \bar{X}) Y_j}{\sum_{j=1}^n (X_j - \bar{X})^2}. \quad (3)$$

Substituting (2) in (3), it can be shown (see the Appendix) that

$$\hat{\beta} = \beta + \gamma \frac{\sum_{j=1}^n (X_j - \bar{X})(Z_j - \bar{Z})}{\sum_{j=1}^n (X_j - \bar{X})^2} + \frac{\sum_{j=1}^n (X_j - \bar{X}) U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \quad (4)$$

where \bar{Z} is the sample mean of the Z_j 's. Assuming that the variables involved are independent across the observations j and that model (2) is correct, in the sense that $E[U_j | X_j, Z_j] = 0$, it follows from (4) that

$$E[\hat{\beta}] = \beta + \gamma \cdot E \left[\frac{\sum_{j=1}^n (X_j - \bar{X})(Z_j - \bar{Z})}{\sum_{j=1}^n (X_j - \bar{X})^2} \right] \neq \beta. \quad (5)$$

In other words, the OLS estimator $\hat{\beta}$ of β in model (1) is no longer unbiased, due to the missing explanatory variable Z_j , except if the sample covariance $(1/n) \sum_{j=1}^n (X_j - \bar{X})(Z_j - \bar{Z})$ is exactly zero.

To demonstrate the effect of missing explanatory variables, I have drawn X_j for $j = 1, \dots, n$

$= 500$ independently from the $N(0,1)$ distribution, then generated the Z_j 's by $Z_j = X_j + V_j$ for $j = 1, \dots, 500$, where the V_j 's have been drawn independently from the $N(0,1)$ distribution, and next I have generated the Y_j 's by $Y_j = 1 + X_j + Z_j + U_j$ for $j = 1, \dots, 500$, where the U_j 's have been drawn independently from the $N(0,1)$ distribution. Thus, model (2) is applicable to this artificial data set, with $\alpha = \beta = \gamma = 1$.

The EasyReg output of the regression according to model (1) is:

Dependent variable:

$Y (= 1+X+Z+U)$

X variables:

$X(1) = X$

$X(2) = 1$

Model:

$Y = b(1)X(1) + b(2)X(2) + U$, where U is the error term satisfying
 $E[U|X(1), X(2)] = 0$.

OLS estimation results

Parameters	Estimate	t-value	H.C. t-value
		(S.E.)	(H.C. S.E.)
		[p-value]	[H.C. p-value]
$b(1)$	2.08383	32.704	34.040
		(0.06372)	(0.06122)
		[0.000000]	[0.000000]
$b(2)$	0.97896	15.475	15.496
		(0.06326)	(0.06317)
		[0.000000]	[0.000000]

Notes:

1: S.E. = Standard error

2: H.C. = Heteroskedasticity Consistent. These t-values and standard errors are based on White's heteroskedasticity consistent variance matrix.

3: The two-sided p-values are based on the normal approximation.

Effective sample size (n): 500

Variance of the residuals: 1.996027

Standard error of the residuals (SER) :	1.412808
Residual sum of squares (RSS) :	994.021241
(Also called SSR = Sum of Squared Residuals)	
Total sum of squares (TSS) :	3128.841499
R-square:	0.682304

Breusch-Pagan test = 0.048472
 Null hypothesis: The errors are homoskedastic
 Null distribution: Chi-square(1)
 p-value = 0.82574
 Significance levels: 10% 5%
 Critical values: 2.71 3.84
 Conclusions: accept accept

Thus, the OLS estimator of β in model (1) is $\hat{\beta} = 2.08383$, which is more than 100% larger than the true value $\beta = 1$.

Note that in this case

$$Y_j = 1 + X_j + Z_j + U_j = 1 + 2X_j + V_j + U_j = 1 + 2X_j + U_j^*, \quad (6)$$

say, where $U_j^* = V_j + U_j$ is the new error term, which satisfies $E[U_j^*|X_j] = 0$ and $\text{var}(U_j^*|X_j) = \text{var}(V_j|X_j) + \text{var}(U_j|X_j) = 1 + 1 = 2$.

Thus, all relevant explanatory variables should be included in the regression model, as otherwise the OLS estimators of the parameters of interest may be biased, and will then stay away from the true values even if the sample size n grows to infinity.

2. The multiple regression model, and least squares estimation

The multiple regression model with an intercept takes the form

$$Y_j = \beta_1 X_{1,j} + \beta_2 X_{2,j} + \dots + \beta_{k-1} X_{k-1,j} + \beta_k + U_j, \quad j = 1, 2, \dots, n, \quad (7)$$

where $X_{i,j}$, $i = 1, 2, \dots, k-1$, are the explanatory variables, U_j is the error term, and β_k is the intercept. We can write this model more compactly as

$$Y_j = \sum_{i=1}^k \beta_i X_{i,j} + U_j, \quad j = 1, 2, \dots, n, \quad (8)$$

where $X_{k,j} \equiv 1$.

The OLS estimators $\hat{\beta}_i$, $i = 1, 2, \dots, k$, of the corresponding parameters β_i can be obtained by minimizing the sum of squared residuals:

$$\min_{\hat{\beta}_1, \dots, \hat{\beta}_k} \sum_{j=1}^n (Y_j - \sum_{i=1}^k \hat{\beta}_i X_{i,j})^2. \quad (9)$$

The first-order conditions for this problem take the form

$$\begin{aligned} \left(\sum_{j=1}^n X_{1,j} X_{1,j} \right) \hat{\beta}_1 + \dots + \left(\sum_{j=1}^n X_{1,j} X_{k,j} \right) \hat{\beta}_k &= \sum_{j=1}^n X_{1,j} Y_j \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \left(\sum_{j=1}^n X_{k,j} X_{1,j} \right) \hat{\beta}_1 + \dots + \left(\sum_{j=1}^n X_{k,j} X_{k,j} \right) \hat{\beta}_k &= \sum_{j=1}^n X_{k,j} Y_j \end{aligned} \quad (10)$$

This is a system of k equations (called the normal equations) in k unknowns $\hat{\beta}_1, \dots, \hat{\beta}_k$. If

Assumption 1: *The normal equations (10) can be solved uniquely for $\hat{\beta}_1, \dots, \hat{\beta}_k$,*

then the solutions involved take the form

$$\begin{aligned} \hat{\beta}_1 &= c_{1,1} \sum_{j=1}^n X_{1,j} Y_j + \dots + c_{1,k} \sum_{j=1}^n X_{k,j} Y_j \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \hat{\beta}_k &= c_{k,1} \sum_{j=1}^n X_{1,j} Y_j + \dots + c_{k,k} \sum_{j=1}^n X_{k,j} Y_j \end{aligned} \quad (11)$$

where the coefficients $c_{m,i}$, $m, i = 1, 2, \dots, k$, are (complicated) functions of the explanatory variables $X_{i,j}$, $i = 1, 2, \dots, k$, $j = 1, \dots, n$.

To determine how the solutions (11) are related to the corresponding true parameter values β_i and the error terms U_j , substitute model (2) in (11). Then

$$\begin{aligned} \hat{\beta}_1 - \beta_1 &= c_{1,1} \sum_{j=1}^n X_{1,j} U_j + \dots + c_{1,k} \sum_{j=1}^n X_{k,j} U_j \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \hat{\beta}_k - \beta_k &= c_{k,1} \sum_{j=1}^n X_{1,j} U_j + \dots + c_{k,k} \sum_{j=1}^n X_{k,j} U_j \end{aligned} \quad (12)$$

where the coefficients $c_{m,i}$, $m, i = 1, 2, \dots, k$, are the same as before. Moreover, denoting

$$w_{i,j} = \sum_{m=1}^k c_{i,m} X_{m,j} \quad (13)$$

we can write the solutions (12) as

$$\begin{aligned}
\hat{\beta}_1 - \beta_1 &= \sum_{j=1}^n w_{1,j} U_j \\
&\vdots \quad \vdots \quad \vdots \\
\hat{\beta}_k - \beta_k &= \sum_{j=1}^n w_{k,j} U_j
\end{aligned} \tag{14}$$

Now let us for the time being assume that

Assumption 2: *The explanatory variables $X_{i,j}$ are non-random,*

and that

Assumption 3: *The error terms U_j are independently $N(0, \sigma^2)$ distributed.*

Because the coefficients (13) are functions of the explanatory variables $X_{i,j}$ only, they are nonrandom too. Consequently, it follows from (14) and Assumptions 2-3 that the OLS estimators $\hat{\beta}_i$ are unbiased:

$$\begin{aligned}
E[\hat{\beta}_1] &= \beta_1 + \sum_{j=1}^n w_{1,j} E[U_j] = \beta_1 \\
&\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
E[\hat{\beta}_k] &= \beta_k + \sum_{j=1}^n w_{k,j} E[U_j] = \beta_k
\end{aligned} \tag{15}$$

Moreover, the variances involved are

$$\begin{aligned}
\text{var}[\hat{\beta}_1] &= \sigma^2 \sum_{j=1}^n w_{1,j}^2 \\
&\vdots \quad \vdots \quad \vdots \\
\text{var}[\hat{\beta}_k] &= \sigma^2 \sum_{j=1}^n w_{k,j}^2
\end{aligned} \tag{16}$$

hence,

$$\begin{aligned}
\hat{\beta}_1 &\sim N[\beta_1, \sigma^2 \sum_{j=1}^n w_{1,j}^2] \\
&\vdots \quad \vdots \quad \vdots \\
\hat{\beta}_k &\sim N[\beta_k, \sigma^2 \sum_{j=1}^n w_{k,j}^2]
\end{aligned} \tag{17}$$

because the error terms U_j are independently $N(0, \sigma^2)$ distributed and the $\hat{\beta}_i$'s are linear combinations of the U_j 's and therefore normally distributed themselves. It follows now from (17) that under Assumptions 1-3,

$$\begin{aligned} \frac{\hat{\beta}_1 - \beta_1}{\sigma \sqrt{\sum_{j=1}^n w_{1,j}^2}} &\sim N[0, 1] \\ &\vdots \quad \vdots \quad \vdots \\ \frac{\hat{\beta}_k - \beta_k}{\sigma \sqrt{\sum_{j=1}^n w_{k,j}^2}} &\sim N[0, 1] \end{aligned} \tag{18}$$

The error variance σ^2 can be estimated similar to the case of the two-variable linear regression model, namely using the sum of squared residuals

$$SSR = \sum_{j=1}^n \hat{U}_j^2, \tag{19}$$

where

$$\hat{U}_j = Y_j - \sum_{i=1}^k \hat{\beta}_i X_{i,j} \tag{20}$$

is the OLS residual.

It can be shown that under Assumptions 1-3,

$$\frac{\sum_{j=1}^n \hat{U}_j^2}{\sigma^2} \sim \chi_{n-k}^2. \tag{21}$$

Since the expected value of a χ_{n-k}^2 distributed random variable is $n-k$, the result (21) suggests to estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{j=1}^n \hat{U}_j^2. \tag{22}$$

Due to (21), this estimator is unbiased: $E[\hat{\sigma}^2] = \sigma^2$. Moreover, it can be shown that under Assumptions 1-3, $\sum_{j=1}^n \hat{U}_j^2$ is independent of the $\hat{\beta}_i$'s, hence it follows from (18) and (21) and the definition of the t distribution that under Assumptions 1 and 2,

$$\begin{aligned}
\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} \sqrt{\sum_{j=1}^n w_{1,j}^2}} &\sim t_{n-k} \\
&\vdots \quad \vdots \quad \vdots \\
\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma} \sqrt{\sum_{j=1}^n w_{k,j}^2}} &\sim t_{n-k}
\end{aligned} \tag{23}$$

The denominators involved are the standard errors of the corresponding OLS estimators:

$$\hat{\sigma}_i = \hat{\sigma} \sqrt{\sum_{j=1}^n w_{i,j}^2} \text{ (= standard error of } \hat{\beta}_i\text{).} \tag{24}$$

The results (18), (21) and (23) do not hinge on the assumption that the explanatory variables $X_{i,j}$ are nonrandom, though. They also hold if we replace Assumption 2 by

Assumption 2*: *The model variables $Y_j, X_{1,j}, \dots, X_{k-1,j}$ are independent and identically distributed across the observations $j = 1, \dots, n$,*

and if we replace Assumption 3 by

Assumption 3*: *Conditionally on $X_{1,j}, \dots, X_{k-1,j}$ the errors U_j are $N(0, \sigma^2)$ distributed.*

Proposition 1: *Under Assumptions 1, 2* and 3* the results (18), (21) and (23) carry over.*

Furthermore, if instead of Assumption 3*,

Assumption 3**: $E[U_j | X_{1,j}, \dots, X_{k-1,j}] = 0$, $E[U_j^2 | X_{1,j}, \dots, X_{k-1,j}] = \sigma^2 < \infty$ and $E[X_{i,j}^2] < \infty$ for $i = 1, \dots, k-1$,

then it can be shown that

Proposition 2: Under Assumptions 1, 2* and 3** ,

$$\begin{aligned} \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}\sqrt{\sum_{j=1}^n w_{1,j}^2}} &\sim N(0,1) \\ \vdots &\quad \vdots \quad \vdots \\ \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{\sum_{j=1}^n w_{k,j}^2}} &\sim N(0,1) \end{aligned} \tag{25}$$

provided that n is large.

3. Testing parameter hypotheses

The results (23) and (25) can be used to test whether a particular coefficient β_i is zero or not, similar to the case of the two-variable linear regression model. The test statistic involved is the corresponding t-value,

$$\hat{t}_i = \frac{\hat{\beta}_i}{\hat{\sigma}_i} = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{\sum_{j=1}^n w_{i,j}^2}}. \tag{26}$$

Proposition 3: Under the null hypothesis $\beta_i = 0$ and the conditions of Proposition 1, $\hat{t}_i \sim t_{n-k}$, and under the null hypothesis involved and the conditions of Proposition 2, $\hat{t}_i \sim N(0,1)$.

Moreover, if $\beta_i > 0$ then \hat{t}_i converges in probability to ∞ if $n \rightarrow \infty$ ¹, and if $\beta_i < 0$ then \hat{t}_i converges in probability to $-\infty$ if $n \rightarrow \infty$ ².

The test can now be conducted in the same way as in the case of the two-variable linear regression model, either left-sided, right-sided or two-sided. The only difference is the degrees of freedom, which is $n-k$ instead of $n-2$ in the two-variable linear regression case.

Now suppose you want to test the joint hypothesis that, for example, $\beta_i = 0$ for $i = 1, \dots, m$,

¹ This means that for any constant $K > 0$, $\lim_{n \rightarrow \infty} P(\hat{t}_i > K) = 1$.

² This means that for any constant $K > 0$, $\lim_{n \rightarrow \infty} P(\hat{t}_i < -K) = 1$.

where $m \leq k - 1$, against the alternative hypothesis that the null hypothesis is false: $\beta_i \neq 0$ for at least one index $i \leq m$. One possible way of testing this hypothesis is to conduct m separate two-sided t tests for $i = 1, \dots, m$. However, the problem is that the left-hand side random variables in (23) are in general not independent, hence under the null hypothesis $\beta_1 = \dots = \beta_m = 0$ the test statistics $\hat{t}_1, \dots, \hat{t}_m$ are in general not independent. In particular, it is impossible to select a critical value t_* such that for a given significance level $\alpha \times 100\%$, $P[|\hat{t}_1| > t_*, |\hat{t}_2| > t_*, \dots, |\hat{t}_m| > t_*] = \alpha$, because we do not know the joint distribution of $\hat{t}_1, \dots, \hat{t}_m$.

The solution of this problem is the following. Consider the restricted regression model

$$Y_j = \beta_{m+1} X_{m+1,j} + \beta_{m+2} X_{m+2,j} + \dots + \beta_{k-1} X_{k-1,j} + \beta_k + U_j, \quad j = 1, 2, \dots, n. \quad (27)$$

Then it can be shown that:

Proposition 4: Under the null hypothesis $\beta_1 = \dots = \beta_m = 0$ and the conditions of Proposition 2,

$$\hat{F} = \frac{(SSR_0 - SSR)/m}{SSR/(n-k)} \sim F_{m,n-k}, \quad (28)$$

and under the conditions of Proposition 2,

$$\hat{W} = m \cdot \hat{F} = \frac{SSR_0 - SSR}{SSR/(n-k)} \sim \chi_m^2, \quad (29)$$

where SSR is the sum of squared residuals of the unrestricted model (7) and SSR_0 is the sum of squared residuals of the restricted model (27). Moreover, under the alternative hypothesis that for at least one index $i \leq m$, $\beta_i \neq 0$, the test statistics \hat{F} and \hat{W} converge in probability to ∞ as $n \rightarrow \infty$ ³.

The test based on \hat{F} is called, for obvious reasons, the F test, and the test based on \hat{W} is called the Wald test, named after the statistician with that name who proposed this test. The tests involved are conducted right-sided. In particular in the case of the Wald test the null hypothesis involved is rejected at say the 5% significance level if $\hat{W} > c$, where the critical value c is chosen

³ Again, this means that for any constant $K > 0$, $\lim_{n \rightarrow \infty} P(\hat{F} > K) = 1$ and $\lim_{n \rightarrow \infty} P(\hat{W} > K) = 1$.

such that for a χ^2_m distributed random variable W , $P[W > c] = 0.05$.

If $m = k - 1$ then the restricted model (27) takes the form

$$Y_j = \beta_k + U_j, \quad j = 1, 2, \dots, n. \quad (30)$$

The sum of squares residuals of this model, SSR_0 , is then equal to the total sum of squares of model (7),

$$TSS = \sum_{j=1}^n (Y_j - \bar{Y})^2, \quad (31)$$

where $\bar{Y} = (1/n) \sum_{j=1}^n Y_j$. The F test involved then has test statistic

$$\tilde{F} = \frac{(TSS - SSR)/(k-1)}{SSR/(n-k)}, \quad (32)$$

which has an $F_{k-1, n-k}$ distribution under the null hypothesis that $\beta_1 = \dots = \beta_{k-1} = 0$ and the conditions of Proposition 2. This test is called the overall F test. Its null hypothesis amounts to the hypothesis that none of the explanatory variables $X_{i,j}$, $i = 1, \dots, k-1$, have an effect on the dependent variable Y_j .

4. The adjusted R^2

The R^2 in the multiple regression case is defined the same as in the two-variable regression case:

$$R^2 = 1 - \frac{SSR}{TSS}. \quad (33)$$

The problem with the R^2 is that it can be inflated towards 1 by including more explanatory variables in the model, because

$$\min_{\hat{\beta}_1, \dots, \hat{\beta}_k} \sum_{j=1}^n (Y_j - \sum_{i=1}^k \hat{\beta}_i X_{i,j})^2 > \min_{\hat{\beta}_1, \dots, \hat{\beta}_k, \hat{\beta}_{k+1}} \sum_{j=1}^n (Y_j - \sum_{i=1}^{k+1} \hat{\beta}_i X_{i,j})^2. \quad (34)$$

The extreme case is where $k = n$. Then $SSR = 0$, hence $R^2 = 1$. To penalize this, the R^2 is adjusted as:

$$\bar{R}^2 = 1 - \frac{SSR/(n-k)}{TSS/(n-1)}. \quad (35)$$

The reason for this particular adjustment is that under the conditions of Proposition 1, $SSR \sim \chi^2_{n-k}$, whereas under the null hypothesis $\beta_1 = \dots = \beta_{k-1} = 0$, $TSS \sim \chi^2_{n-1}$.

5. Multicollinearity

Multicollinearity is the phenomenon that (some of) the explanatory variables are highly correlated. The effect of multicollinearity is that the t-values are deflated. To demonstrate this, I have generated artificial data $Y_j, X_{1,j}, X_{2,j}$ for $j = 1, \dots, n = 500$ as follows. The explanatory variables $X_{1,j}$ have been drawn independently from the $N(0,1)$ distribution. Next, I have drawn random variables V_j independently from the $N(0,1)$ distribution, and have set $X_{2,j} = X_{1,j} + 0.01V_j$. Due to this construction the explanatory variables $X_{1,j}$ and $X_{2,j}$ are highly correlated. In particular, the R^2 of the regression of $X_{2,j}$ on $X_{1,j}$ is 0.999892. Finally, I have drawn the errors U_j independently from the $N(0,1)$ distribution, and have generated the dependent variables by

$$Y_j = X_{1,j} + X_{2,j} + 1 + U_j, \quad j = 1, \dots, n = 500. \quad (36)$$

This is model (7) with $k = 3$ and $\beta_1 = \beta_2 = \beta_3 = 1$.

The EasyReg output involved is below.

OLS estimation results

Parameters	Estimate	t-value	H.C. t-value
		(S.E.)	(H.C. S.E.)
		[p-value]	[H.C. p-value]
b (1)	0.45258	0.101	0.104
		(4.48811)	(4.35703)
		[0.91968]	[0.91727]
b (2)	1.57385	0.351	0.362
		(4.48634)	(4.35350)
		[0.72573]	[0.71772]
b (3)	0.95879	21.386	21.410
		(0.04483)	(0.04478)
		[0.00000]	[0.00000]

Notes:

1: S.E. = Standard error

2: H.C. = Heteroskedasticity Consistent. These t-values and standard errors are based on White's heteroskedasticity consistent variance matrix.

3: The two-sided p-values are based on the normal approximation.

Effective sample size (n) :	500
Variance of the residuals:	1.00464
Standard error of the residuals (SER) :	1.002317
Residual sum of squares (RSS) :	499.305951
(Also called SSR = Sum of Squared Residuals)	
Total sum of squares (TSS) :	2396.64478
R-square:	0.791665
Adjusted R-square:	0.790826

Overall F test: F(2,497) = 944.29

p-value = 0.00000

Significance levels:	10%	5%
----------------------	-----	----

Critical values:	2.31	3.01
------------------	------	------

Conclusions:	reject	reject
--------------	--------	--------

Breusch-Pagan test = 3.286972

Null hypothesis: The errors are homoskedastic

Null distribution: Chi-square(2)

p-value = 0.19331

Significance levels:	10%	5%
----------------------	-----	----

Critical values:	4.61	5.99
------------------	------	------

Conclusions:	accept	accept
--------------	--------	--------

Note that $b(1)$, $b(2)$ and $b(3)$ are the OLS estimators of β_1 , β_2 and β_3 , respectively.

Observe that not only the OLS estimators of β_1 and β_2 are way off from the true value 1, but that also the corresponding t-values are deflated towards levels where the null hypotheses $\beta_1 = 0$ and $\beta_2 = 0$ cannot be rejected by the separate t tests at any reasonable significance level. On the other hand, the overall F test strongly rejects the joint null hypothesis that $\beta_1 = \beta_2 = 0$. These contradictory results are due to multicollinearity.

There is no cure for multicollinearity. The only thing you can do is be aware of it, and always test joint hypotheses using the F or Wald test rather than using separate t tests.

Assumption 1 does not rule out multicollinearity, but only its extreme form, where one explanatory variable is an exact linear function of other explanatory variables. For example, consider model (2), and suppose that $Z_j = \eta + \delta X_j$ without error. Then model (2) becomes

$$\begin{aligned}
Y_j &= \alpha + \beta \cdot X_j + \gamma \cdot Z_j + U_j = \alpha + \beta \cdot X_j + \gamma \cdot (\eta + \delta X_j) + U_j \\
&= (\alpha + \gamma \cdot \eta) + (\beta + \gamma \cdot \delta) \cdot X_j + U_j, \quad j = 1, 2, \dots, n,
\end{aligned} \tag{37}$$

Clearly, only $\alpha + \gamma \cdot \eta$ and $\beta + \gamma \cdot \delta$ can be estimated by OLS, but not α , β and γ separately without knowing η and δ . This case is ruled out by Assumption 1.

6. Heteroscedasticity

Recall that the errors U_j of a regression model are heteroskedastic if the conditional variance of U_j given the explanatory variables is not constant, but a function of the explanatory variables. In particular, the error terms in model (7) are heteroskedastic if there exists a non-constant function ψ such that

$$E[U_j^2 | X_{1,j}, X_{2,j}, \dots, X_{k-1,j}] = \psi(X_{1,j}, X_{2,j}, \dots, X_{k-1,j}). \tag{38}$$

Heteroscedasticity often occurs in practice. It is actually the rule rather than the exception. One of the problems of heteroscedasticity is that the standard errors and t-values of the OLS parameter estimators are no longer valid. However, this problem can easily be cured by replacing the standard errors (24) with the heteroscedasticity consistent (H.C.) standard errors:

$$\tilde{\sigma}_i = \sqrt{\sum_{j=1}^n w_{i,j}^2 \hat{U}_j^2} \quad (= \text{H.C. standard error of } \hat{\beta}_i). \tag{39}$$

and the t-values with the heteroscedasticity consistent (H.C.) t-values

$$\tilde{t}_i = \frac{\hat{\beta}_i}{\tilde{\sigma}_i} \quad (= \text{H.C. t-value of } \hat{\beta}_i). \tag{40}$$

The F and Wald tests in Proposition 4 are also no longer valid under heteroscedasticity, but the cure for this is difficult to explain at the undergraduate level. To test joint hypotheses under heteroscedasticity with EasyReg you have to increase the econometrics level to “Intermediate”. Then after running your regression you will get the option to conduct Wald tests of linear parameter restrictions. This option gives you two versions of the Wald test, one for the homoscedastic case and one for the heteroskedastic case. See the guided tour on OLS estimation.

To decide whether the errors of model (7) are homoscedastic or heteroskedastic, use the Breusch-Pagan⁴ test. Given that

$$E[U_j^2 | X_{1,j}, X_{2,j}, \dots, X_{k-1,j}] = g\left(\sum_{i=1}^{k-1} \gamma_i X_{i,j} + \gamma_k\right) \text{ for some unknown function } g(\cdot). \quad (41)$$

the Breusch-Pagan test tests the null hypothesis

$$H_0: \gamma_1 = \dots = \gamma_{k-1} = 0 \Leftrightarrow E[U_j^2 | X_{1,j}, X_{2,j}, \dots, X_{k-1,j}] = g(\gamma_k) = \sigma^2, \quad (42)$$

against the alternative hypothesis

$$H_0: E[U_j^2 | X_{1,j}, X_{2,j}, \dots, X_{k-1,j}] = g\left(\sum_{i=1}^{k-1} \gamma_i X_{i,j} + \gamma_k\right) \neq g(\gamma_k) \quad (43)$$

Under the null hypothesis (42) of homoskedasticity the test statistic of the Breusch-Pagan test has a χ_{k-1}^2 distribution, and the test is conducted right-sided.

APPENDIX

Proof of (14):

Substituting (2) in (3) yields

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{j=1}^n (X_j - \bar{X})(\alpha + \beta \cdot X_j + \gamma \cdot Z_j + U_j)}{\sum_{j=1}^n (X_j - \bar{X})^2} \\ &= \alpha \frac{\sum_{j=1}^n (X_j - \bar{X})}{\sum_{j=1}^n (X_j - \bar{X})^2} + \beta \frac{\sum_{j=1}^n (X_j - \bar{X}) X_j}{\sum_{j=1}^n (X_j - \bar{X})^2} + \gamma \frac{\sum_{j=1}^n (X_j - \bar{X}) Z_j}{\sum_{j=1}^n (X_j - \bar{X})^2} + \frac{\sum_{j=1}^n (X_j - \bar{X}) U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \\ &= \beta + \gamma \frac{\sum_{j=1}^n (X_j - \bar{X})(Z_j - \bar{Z})}{\sum_{j=1}^n (X_j - \bar{X})^2} + \frac{\sum_{j=1}^n (X_j - \bar{X}) U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \end{aligned} \quad (44)$$

where \bar{Z} is the sample mean of the Z_j 's. Note that the last equality in (4) follows from the fact that $\sum_{j=1}^n (X_j - \bar{X}) = 0$.

⁴ Breusch, T. and A. Pagan (1979), "A Simple Test for Heteroscedasticity and Random Coefficient Variation", *Econometrica* 47, 1287-1294.

SPECIFICATION OF ECONOMETRIC MODELS

Herman J. Bierens

Pennsylvania State University

March 26, 2009

1 *Introduction*

Most econometric models link an observable dependent variable Y to observable explanatory variables X_1, \dots, X_m , an unobservable variable U (the error term if the model is a linear regression), and parameters β_1, \dots, β_k , via some function f :

$$Y = f(X_1, \dots, X_m, U, \beta_1, \dots, \beta_k). \quad (1)$$

For example, in the case of a linear regression model with intercept this function f is specified as

$$f(X_1, \dots, X_m, U, \beta_1, \dots, \beta_k) = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{k-1} X_{k-1} + \beta_k + U, \quad m = k-1. \quad (2)$$

The question is: “How should we specify this function f ?” There is no single answer to this question, of course, but there are a few rules to narrow down the set of possible choices for f .

2 *Feasibility rule*

The first rule is the feasibility rule:

Rule 1: (Feasibility) *The function f should be specified such that the equality (1) can hold for all possible values of the dependent variable Y , the explanatory variables X_1, \dots, X_m and the error term U .*

As an example of a violation of this rule, consider the linear regression model

$$Y = \beta_1 X_1 + \beta_2 + U, \quad E[U|X_1] = 0. \quad (3)$$

where Y is the demand for a particular good or service, and X_1 is the price of that good or service. For instance, let Y be the demand for cruises to the Bahamas during the peak season, measured as the total number of people taking cruises to the Bahamas during the peak season times the average duration of a cruise, and let X_1 be the cost of a cruise for one person per day. The law of

demand predicts that if the price X_1 goes up then the demand Y goes down. Therefore, $\beta_1 < 0$. But then $\beta_1 X_1 + \beta_2 < 0$ if $X_1 > -\beta_2/\beta_1$, so that $E[Y|X_1 > -\beta_2/\beta_1] = \beta_1 X_1 + \beta_2 < 0$. However, this is impossible because demand Y cannot be negative.

At first sight one could think of fixing this problem by specifying the demand model as

$$\ln(Y) = \beta_1 X_1 + \beta_2 + U, \quad E[U|X_1] = 0.$$

so that

$$Y = f(X_1, U, \beta_1, \beta_2) = \exp(\beta_1 X_1 + \beta_2) \exp(U).$$

Assume for sake of the argument that U and X_1 are independent and that $U \sim N(0, \sigma^2)$. Then it can be shown (but that is beyond the level of this course) that $E[\exp(U)|X_1] = \exp(\sigma^2/2)$, hence

$$E[Y|X_1] = \exp(\beta_1 X_1 + \beta_2) \exp(\sigma^2/2).$$

Again, the law of demand predicts that $\beta_1 < 0$. But now this model predicts that if the price X_1 decreases towards zero the log of demand Y increases towards an upper bound. In particular,

$$E[Y|X_1] \uparrow \exp(\beta_2) \exp(\sigma^2/2) < \infty \text{ if } X_1 \downarrow 0.$$

This means that even if the good or service involved is completely free the expected demand will be bounded. This may be reasonable for some goods or services, but not for cruises to the Bahamas!

Therefore, it is better to specify this demand model as

$$\ln(Y) = \beta_1 \ln(X_1) + \beta_2 + U, \quad E[U|X_1] = 0. \quad (4)$$

Because $\beta_1 < 0$ we now have that $Y \uparrow \infty$ if $X_1 \downarrow 0$ and $Y \downarrow 0$ if $X_1 \uparrow \infty$. Note that the parameter β_1 in model (4) can be interpreted as the elasticity of demand:

$$\beta_1 = \frac{d\ln(Y)}{d\ln(X_1)} = \frac{(dY)/Y}{(dX_1)/X_1},$$

where the second equality follows from the fact that $d\ln(x)/dx = 1/x$, hence $d\ln(x) = (dx)/x$.

Another violation of rule 1 is the so-called linear probability model, which is discussed in most undergraduate econometrics textbooks. Consider the case where the dependent variable Y is a binary variable: it takes only two values, 0 or 1. For example, let $Y = 1$ if a household owns the home it lives in, and $Y = 0$ if not, and let X_1 be household income. The linear probability model is then the regression model (3) for this case.

Because $E[Y|X_1] = 0 \times P[Y = 0|X_1] + 1 \times P[Y = 1|X_1] = P[Y = 1|X_1]$ the linear

probability model states that

$$P[Y = 1|X_1] = \beta_1 X_1 + \beta_2. \quad (5)$$

If we would believe this to be true, then we should expect $\beta_1 > 0$, because the higher the household income X_1 the more likely the household will own the home it lives in. But then

$$P[Y = 1|X_1] > 1 \text{ if } X_1 > (1 - \beta_2)/\beta_1,$$

which is impossible. The only case where the linear probability model is valid is the case where the explanatory variable X_1 is a binary variable itself: $X_1 = 0$ or $X_1 = 1$, and this explanatory variable is the only one in the model. Then it follows from (5) that

$$P[Y=1|X_1=0] = \beta_2, \quad P[Y=1|X_1=1] = \beta_1 + \beta_2,$$

which is possible if $0 < \beta_2 < 1$ and $0 < \beta_1 + \beta_2 < 0$.

The problem of how to specify the function f in (1) in the case of a binary dependent variable Y will be addressed in a separate lecture note.

The problem with the linear probability model also applies to the case where the dependent variable Y is a fraction, so that $0 < Y < 1$. For example, let Y be the share of the expenditures on food and clothing in total expenditures of a household, and let X_1 be household income or the log of household income. The linear regression model (3) is not appropriate in this case, because it will be impossible to force $\beta_1 X_1 + \beta_2$ between zero and one for all possible values of X_1 if $\beta_1 \neq 0$. However, in this case there is an easy solution, namely, let

$$Y = f(X_1, U, \beta_1, \beta_2) = \frac{\exp(\beta_1 X_1 + \beta_2 + U)}{1 + \exp(\beta_1 X_1 + \beta_2 + U)}.$$

Then $0 < Y < 1$. It is easy to verify that this model can be rewritten as

$$\ln\left(\frac{Y}{1-Y}\right) = \beta_1 X_1 + \beta_2 + U, \quad (6)$$

which is a valid regression model if $E[U|X_1] = 0$.

If in model (6) Y is the share of the expenditures on food and clothing in total expenditures of a household and X_1 is household income, one may expect that $\beta_1 < 0$ because the higher income, the more the household will spend on other items than food and clothing

relative to total expenditures. Then Y is maximal for $X_1 = 0$, but this maximum is less than 1 if X_1 is household income itself rather than the log of income, because

$$\max Y = \lim_{X_1 \rightarrow 0} \frac{\exp(\beta_1 X_1 + \beta_2 + U)}{1 + \exp(\beta_1 X_1 + \beta_2 + U)} = \frac{\exp(\beta_2 + U)}{1 + \exp(\beta_2 + U)} < 1.$$

This is not realistic. However, this problem can easily be cured by replacing X_1 with the log of X_1 :

$$\ln\left(\frac{Y}{1-Y}\right) = \beta_1 \ln(X_1) + \beta_2 + U. \quad (7)$$

Then

$$\max Y = \lim_{X_1 \rightarrow 0} \frac{\exp(\beta_1 \ln(X_1) + \beta_2 + U)}{1 + \exp(\beta_1 \ln(X_1) + \beta_2 + U)} = \lim_{z \rightarrow \infty} \frac{\exp(z)}{1 + \exp(z)} = \lim_{z \rightarrow \infty} \frac{1}{1 + \exp(-z)} = 1,$$

which is more plausible.

3 Scale invariance rule

In January 2002, twelve members of the European union (Belgium, Germany, Greece, Spain, France, Ireland, Italy, Luxembourg, the Netherlands, Austria, Portugal and Finland) changed their local currencies to a common currency, the Euro. Each local currency was exchanged for Euros at a fixed exchange rate. For example, the exchange rate in the Netherlands was 2.20371 Dutch guilders for one Euro.

Consider the model $Y = f(X_1, U, \beta_1, \beta_2)$, where again Y is the demand for a good or service, measured in some non-monetary unit, and X_1 is the price of this good or service, in the Netherlands before the introduction of the Euro. Thus, the price X_1 was initially measured in Dutch guilders. After the introduction of the Euro the price became $X_1/2.20371$ Euros. Did this conversion to the Euro have an effect on demand Y ? In January 2002 and a few months thereafter it had! All prices in Euros were suddenly about 45% of what they were before in Dutch guilders, and it took a while for the public to realize that their income was also about 45% less in Euros than what it was before in Dutch guilders. But once the public got used to the new Euro

the effect on demand of the conversion to the Euro vanished.

For the demand model $Y = f(X_1, U, \beta_1, \beta_2)$ to be valid before and after the introduction of the Euro, it must be possible to adjust the parameters β_1 and β_2 to β_1^* and β_2^* such that $Y = f(X_1, U, \beta_1, \beta_2) = f(X_1/2.20371, U, \beta_1^*, \beta_2^*)$. In other words, changes in the unit of measurement of the explanatory variables should not affect the functional form f of the model, because this functional form represents the behavior of economic agents that should not be affected by units of measurements. This leads to our second rule:

Rule 2: (Scale invariance) *The function f in (1) should be specified such that changes in the unit of measurements of the explanatory variables X_1, \dots, X_m can be compensated by changes in the parameters β_1, \dots, β_k : If $Y = f(X_1, \dots, X_m, U, \beta_1, \dots, \beta_k)$ then for arbitrary positive numbers $\lambda_1, \dots, \lambda_m$ there exist parameters $\beta_1^*, \dots, \beta_k^*$ such that $Y = f(\lambda_1 X_1, \dots, \lambda_m X_m, U, \beta_1^*, \dots, \beta_k^*)$.*

Clearly this rule holds for the linear regression model (2):

$$f(\lambda_1 X_1, \dots, \lambda_m X_m, U, \beta_1^*, \dots, \beta_k^*) = \beta_1^* \lambda_1 X_1 + \beta_2^* \lambda_2 X_2 + \dots + \beta_{k-1}^* \lambda_{k-1} X_{k-1} + \beta_k^* + U, \quad m = k-1.$$

where $\beta_i^* = \beta_i / \lambda_i$ for $i = 1, \dots, k-1$ and $\beta_k^* = \beta_k$. It also holds for a log-linear model, provided that there is an intercept. For example, in the case (4),

$$\ln(Y) = \beta_1 \ln(X_1) + \beta_2 + U = \beta_1 \ln(\lambda_1 X_1) + \beta_2 - \beta_1 \ln(\lambda_1) + U,$$

so that $\beta_1^* = \beta_1$ and $\beta_2^* = \beta_2 - \beta_1 \ln(\lambda_1)$. Therefore, if one or more variables enter the model in log form you have to include an intercept, as otherwise rule 2 will be violated.

4 Use economic theory

In the example of the demand for cruises to the Bahamas we have combined rule 1 with the law of demand. Thus, we have already applied the following third specification rule:

Rule 3: (Use economic theory) *Base your model specification on economic theory, if possible.*

However, often economic theory tells you more about which variables to select, and the direction

of the effect of the independent variables on the dependent variable, than about the functional form of the model. Nevertheless, there are a few cases where the functional form can be derived from economic theory. An example is the Mincer wage equation.

The basic idea of Mincer's¹ theory is that wages increase with experience on the job up to a certain point, after which wages decrease with experience. The underlying economic theory is human capital theory: The productivity of a worker increases with on-the-job training, hence it is advantageous for firms to invest in on-the job training of their workers. The training may not be a formal training. Just by having more experience in doing a particular job one can do the job better and faster. However, human capital depreciates over time. For example, a particular job may become obsolete due to change in technology. Anyhow, the human capital theory predicts that the marginal product of a worker first increases with experience on the job but after a certain point will decrease. Consequently, wages follow the same pattern because the optimal² wage of a worker is equal to his or her marginal product.

A functional form of the wage-experience relationship that can mimic this pattern is a quadratic function:

$$Y = \alpha + \beta X + \gamma X^2 + U, \quad (8)$$

where Y is some measurement of wage (to be discussed below), X is experience on the job and U is an error term. The quadratic function involved is maximal at say X_0 years of experience if the slope $\beta + 2\gamma X$ is positive for $X < X_0$ and negative for $X > X_0$. A necessary condition for this is that $\beta > 0$ and $\gamma < 0$. Then $X_0 = -0.5\beta/\gamma > 0$.

As to the dependent variable Y , we cannot choose for Y the wage itself, because the right-hand side of (8) can become negative for large X , whereas the wage cannot be negative, so that then rule 1 will be violated. Therefore, let $Y = \ln(Wage)$. Then the basic Mincer wage equation takes the form

$$\begin{aligned} \ln(Wage) &= \alpha + \beta \cdot Experience + \gamma \cdot Experience^2 + U, \\ \beta &> 0, \gamma < 0. \end{aligned} \quad (9)$$

¹ Mincer, J. (1974), *Schooling, Experience and Earnings*, New York: Columbia University Press.

² Optimal from the point of view of the firm.

Often this model is augmented with a variety of other explanatory variables, such as gender and race indicators, years or level of schooling, location variables, etc.

5 Testing for misspecification of functional form

5.1 Ramsey's RESET test³

Consider the general linear regression model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{k-1} X_{k-1} + \beta_k + U. \quad (10)$$

Recall that the crucial condition for correctness of this model is that the conditional expectation of the error term U given the explanatory variables is zero:

$$E[U|X_1, X_2, \dots, X_{k-1}] = 0 \quad (11)$$

because then

$$E[Y|X_1, X_2, \dots, X_{k-1}] = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{k-1} X_{k-1} + \beta_k. \quad (12)$$

The conditions (11) and (12) are equivalent: If (12) holds then condition (11) holds for the error term U in model (10), and vice versa.

The question is how to test the null hypothesis (11) (or equivalently, the null hypothesis (12)). Before we can answer this question, we need to formulate an alternative hypothesis. The most general alternative hypothesis is that (12) is not true, but for practical purposes we have to narrow down this alternative, as follows. Assume that for some function g and parameters

$$\beta_1, \dots, \beta_{k-1}, \beta_k,$$

$$E[Y|X_1, X_2, \dots, X_{k-1}] = g(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{k-1} X_{k-1} + \beta_k). \quad (13)$$

Then the null hypothesis (12) boils down to the hypothesis that $g(x) = x$.

If the function g in (13) is p times differentiable we can approximate it by a polynomial of order p :

$$g(x) \approx \gamma_0 + \gamma_1 x + \dots + \gamma_p x^p. \quad (14)$$

Often this approximation is already pretty close for $p = 2$, so let us focus on that case. Thus, consider the alternative hypothesis

³ Ramsey, J. (1974), "Classical Model Selection Through Specification Error Tests", in P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press.

$$\begin{aligned}
E[Y|X_1, X_2, \dots, X_{k-1}] &= \gamma_0 + \gamma_1(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{k-1} X_{k-1} + \beta_k) \\
&\quad + \gamma_2(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{k-1} X_{k-1} + \beta_k)^2 \\
&= \delta_1 X_1 + \delta_2 X_2 + \dots + \delta_{k-1} X_{k-1} + \delta_k + \gamma_2(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{k-1} X_{k-1} + \beta_k)^2,
\end{aligned} \tag{15}$$

where $\delta_k = \gamma_0 + \beta_k$ and $\delta_i = \gamma_1 \beta_i$ for $i = 1, \dots, k-1$. Now the null hypothesis (12) corresponds to the null hypothesis that $\gamma_2 = 0$.

This suggests the following testing procedure. First, estimate model (10) by OLS, and compute

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_{k-1} X_{k-1} + \hat{\beta}_k, \tag{16}$$

where the $\hat{\beta}_i$'s are the OLS estimates of the β_i 's. You can do that in EasyReg in two steps. Once your estimation results are displayed in the “What to do next?” window, open “Options” and click “Write residuals to the input file”. Then the OLS residuals

$$\hat{U} = Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \dots - \hat{\beta}_{k-1} X_{k-1} - \hat{\beta}_k = Y - \hat{Y}$$

are added to the input file, as new variable OLS Residual of “Y”, where “Y” is the actual name of the dependent variable Y . Next, open “Menu > Input > Transform variables” in the EasyReg main window, click “Linear combination of variables”, select the variables Y and \hat{U} (= OLS Residual of “Y”), and make the linear combination $Y - \hat{U} = \hat{Y}$.

Second, include the variable \hat{Y}^2 in the augmented regression model:

$$Y = \gamma_2 \hat{Y}^2 + \delta_1 X_1 + \delta_2 X_2 + \dots + \delta_{k-1} X_{k-1} + \delta_k + U, \tag{17}$$

and re-estimate it by OLS. Of course, you have to make the variable \hat{Y}^2 first, which can be done in EasyReg as follows. Open “Menu > Input > Transform variables” in the EasyReg main window again, click “Multiplicative combination of variables”, select the new variable \hat{Y} and choose power 2. This creates the variable \hat{Y}^2 .

Finally, test the null hypothesis $\gamma_2 = 0$ against the alternative hypothesis $\gamma_2 \neq 0$, using the t-test.

A more general test is to estimate the augmented regression model

$$Y = \gamma_2 \hat{Y}^2 + \dots + \gamma_p \hat{Y}^p + \delta_1 X_1 + \delta_2 X_2 + \dots + \delta_{k-1} X_{k-1} + \delta_k + U, \tag{18}$$

by OLS and test the joint null hypothesis $\gamma_2 = 0, \dots, \gamma_p = 0$ against the alternative that $\gamma_i \neq 0$ for some index i ($=2, \dots, p$), using the Wald test. This test is known as Ramsey's

Regression Specification Error Test (RESET).

5.2 An application

The data set for this application of the RESET test is random sample⁴ of size $n = 2000$ from the Dutch Wage Structure Survey 1997, containing the following variables:

- ln(Wage), where Wage is the hourly wage in Dutch guilders, times 100
- Gender (Female = 1, Male = 0)
- College (1 if education level ≥ 5 , 0 if not)⁵
- Experience (Experience with the present employer, in years)
- Age (in years)

The initial model is model (9), augmented with the variables Gender, College, and Age. The EasyReg output involved is:

```

Y = LN[Wage]
X variables:
X(1) = experience
X(2) = experience^2
X(3) = Gender
X(4) = College
X(5) = age
X(6) = 1

Model: Y = b(1)X(1) +....+ b(6)X(6) + U, where U is the error term,
satisfying E[U|X(1),...,X(6)] = 0.

OLS estimation results
Parameters Estimate      t-value      H.C. t-value
                  (S.E.)      (H.C. S.E.)
                  [p-value]  [H.C. p-value]
b(1)        0.01772      7.779      7.921
                  (0.00228)  (0.00224)
                  [0.00000]  [0.00000]
b(2)       -0.00045     -7.139     -7.626
                  (0.00006)  (0.00006)
                  [0.00000]  [0.00000]

```

⁴ See Bierens, H.J. and J. Hartog (1988), "Non-Linear Regression with Discrete Explanatory Variables", *Journal of Econometrics* 38, 269-299, for a description of this sample. This data set is included in the EasyReg database.

⁵ The original data set contains the variable "level of education", ranging from 1 to 7. Level 5 is "higher general" which is comparable with a BA degree.

b (3)	- 0.16947	- 10.322 (0.01642) [0.00000]	- 11.420 (0.01484) [0.00000]
b (4)	0.46154	28.765 (0.01605) [0.00000]	24.445 (0.01888) [0.00000]
b (5)	0.01024	14.493 (0.00071) [0.00000]	12.659 (0.00081) [0.00000]
b (6)	6.85378	289.287 (0.02369) [0.00000]	270.929 (0.02530) [0.00000]

Notes:

- 1: S.E. = Standard error
- 2: H.C. = Heteroskedasticity Consistent. These t-values and standard errors are based on White's heteroskedasticity consistent variance matrix.
- 3: The two-sided p-values are based on the normal approximation.

Effective sample size (n):	2000
Variance of the residuals:	0.073569
Standard error of the residuals (SER):	0.271236
Residual sum of squares (RSS):	146.697047
(Also called SSR = Sum of Squared Residuals)	
Total sum of squares (TSS):	290.506519
R-square:	0.4950
Adjusted R-square:	0.4938

Breusch-Pagan test = 143.749687
Null hypothesis: The errors are homoskedastic
Null distribution: Chi-square(5)
p-value = 0.00000
Significance levels: 10% 5%
Critical values: 9.24 11.07
Conclusions: reject reject

Next, augment this model with $\hat{Y}^2 = (\text{LN}[Wage] - \text{OLS Residual of LN}[Wage])^2$. Then the results become:

X variables:
X(1) = experience
X(2) = experience²
X(3) = Gender
X(4) = College
X(5) = age
X(6) = (LN[Wage] - OLS Residual of LN[Wage])²
X(7) = 1

Model: $Y = b(1)X(1) + \dots + b(7)X(7) + U$, where U is the error term,
satisfying $E[U|X(1), \dots, X(7)] = 0$.

OLS estimation results

Parameters	Estimate	t-value (S.E.)	H.C. t-value (H.C. S.E.)
		[p-value]	[H.C. p-value]
b (1)	-0.06223	-2.770 (0.02247) [0.00561]	-2.392 (0.02602) [0.01677]
b (2)	0.00159	2.766 (0.00057) [0.00567]	2.390 (0.00066) [0.01684]
b (3)	0.57380	2.753 (0.20843) [0.00591]	2.373 (0.24179) [0.01764]
b (4)	-1.66944	-2.801 (0.59594) [0.00509]	-2.411 (0.69237) [0.01590]
b (5)	-0.03613	-2.783 (0.01298) [0.00539]	-2.399 (0.01506) [0.01644]
b (6)	0.30645	3.577 (0.08567) [0.00035]	3.068 (0.09988) [0.00215]
b (7)	-7.46569	-1.865 (4.00313) [0.06219]	-1.599 (4.67026) [0.10992]

Notes:

1: S.E. = Standard error

2: H.C. = Heteroskedasticity Consistent. These t-values and standard errors are based on White's heteroskedasticity consistent variance matrix.

3: The two-sided p-values are based on the normal approximation.

Effective sample size (n):	2000
Variance of the residuals:	0.073137
Standard error of the residuals (SER):	0.270438
Residual sum of squares (RSS):	145.761204
(Also called SSR = Sum of Squared Residuals)	
Total sum of squares (TSS):	290.506519
R-square:	0.4983
Adjusted R-square:	0.4967

Breusch-Pagan test = 129.195866

Null hypothesis: The errors are homoskedastic

Null distribution: Chi-square(6)

p-value = 0.000000

Significance levels: 10% 5%

Critical values: 10.64 12.59

Conclusions: reject reject

Note that the parameter γ_2 in (17) corresponds to b(6). Thus, the test statistic of the RESET test is the t-value corresponding to b(6). Because there is strong evidence of heteroskedasticity, the appropriate t-value is the H.C. t-value (3.068). Under the null hypothesis $\gamma_2 = 0$ this t-value is

a random drawing from the standard normal distribution (because the sample size n is large). Moreover, recall that the 5% critical value of the two-sided standard normal test is 1.96. Clearly, the null hypothesis is rejected at the 5% significance level, and in view of the p-value involved even at the 0.1% significance level!

But now the problem arises how to fix the misspecification. Sometimes you can do that by allowing for interactions between variables. In this case it is conceivable that the effect of experience and age on the log of the wage is different for females and males, and college graduates and non-college graduates. Also, maybe we need higher powers of experience as well.

To test this, regress LN(Wage) on

```
X(1) = experience
X(2) = experience^2
X(3) = experience^3
X(4) = Gender
X(5) = College
X(6) = age
X(7) = Gender x College
X(8) = Gender x experience
X(9) = Gender x experience^2
X(10) = Gender x experience^3
X(11) = College x experience
X(12) = College x experience^2
X(13) = College x experience^3
X(14) = Gender x age
X(15) = College x age
X(16) = College x Gender x experience
X(17) = College x Gender x experience^2
X(18) = College x Gender x experience^3
X(19) = College x Gender x age
X(20) = 1
```

Model:

$Y = b(1)X(1) + \dots + b(20)X(20) + U$, where U is the error term, satisfying $E[U|X(1), \dots, X(20)] = 0$.

OLS estimation results

Parameters	Estimate	t-value	H.C. t-value
		(S.E.)	(H.C. S.E.)
		[p-value]	[H.C. p-value]
b(1)	0.02369	4.792	4.457
		(0.00494)	(0.00531)
		[0.000000]	[0.00001]
b(2)	-0.00079	-2.579	-2.397
		(0.00030)	(0.00033)
		[0.00991]	[0.01652]
b(3)	0.00001	1.315	1.222
		(0.00001)	(0.00001)
		[0.18843]	[0.22153]

b (4)	- 0.41656	- 6.512 (0.06397) [0.00000]	- 6.033 (0.06904) [0.00000]
b (5)	- 0.22960	- 2.816 (0.08154) [0.00487]	- 2.238 (0.10259) [0.02522]
b (6)	0.00650	8.260 (0.00079) [0.00000]	7.454 (0.00087) [0.00000]
b (7)	- 0.00628	- 0.031 (0.20386) [0.97541]	- 0.030 (0.21032) [0.97617]
b (8)	0.02803	1.573 (0.01782) [0.11581]	1.873 (0.01497) [0.06106]
b (9)	- 0.00279	- 1.565 (0.00179) [0.11750]	- 1.835 (0.00152) [0.06651]
b (10)	0.00005	1.108 (0.00005) [0.26781]	1.357 (0.00004) [0.17492]
b (11)	- 0.00814	- 0.578 (0.01407) [0.56305]	- 0.560 (0.01454) [0.57564]
b (12)	- 0.00034	- 0.344 (0.00099) [0.73082]	- 0.329 (0.00103) [0.74189]
b (13)	0.00001	0.378 (0.00002) [0.70525]	0.388 (0.00002) [0.69776]
b (14)	0.00776	3.592 (0.00216) [0.00033]	3.135 (0.00247) [0.00172]
b (15)	0.02142	9.964 (0.00215) [0.00000]	7.651 (0.00280) [0.00000]
b (16)	0.07313	1.147 (0.06376) [0.25141]	1.423 (0.05141) [0.15488]
b (17)	- 0.01026	- 1.394 (0.00736) [0.16323]	- 1.512 (0.00678) [0.13056]
b (18)	0.00037	1.595 (0.00023) [0.11069]	1.639 (0.00023) [0.10125]
b (19)	- 0.00844	- 1.235 (0.00683) [0.21683]	- 1.097 (0.00769) [0.27246]
b (20)	6.95949	243.701 (0.02856) [0.00000]	216.759 (0.03211) [0.00000]

Notes:

1: S.E. = Standard error

2: H.C. = Heteroskedasticity Consistent. These t-values and standard errors are based on White's heteroskedasticity consistent variance matrix.

3: The two-sided p-values are based on the normal approximation.

Effective sample size (n) :	2000
Variance of the residuals:	0.067767
Standard error of the residuals (SER) :	0.260321
Residual sum of squares (RSS) :	134.178399
(Also called SSR = Sum of Squared Residuals)	
Total sum of squares (TSS) :	290.506519
R-square:	0.5381
Adjusted R-square:	0.5337

Breusch-Pagan test = 144.599224
 Null hypothesis: The errors are homoskedastic
 Null distribution: Chi-square(19)
 p-value = 0.00000
 Significance levels: 10% 5%
 Critical values: 27.2 30.14
 Conclusions: reject reject

Indeed, quite a few interaction variables are significant. None of the interaction variables involving Gender \times College are significant, but they are jointly significant. However, if we augment this model with $\hat{Y}^2 = (\text{LN}[Wage] - \text{OLS Residual of LN}[Wage])^2$, the coefficient involved is still significant at the 5% level. Thus, despite the interaction variables the model is still misspecified. As shown by Bierens and Hartog (1988) [see footnote 4], the non-linearity of the model involved is much more complicated than can be captured by interaction variables and higher powers. Thus, *at the undergraduate econometrics level* the model is beyond repair.

The Logit Model: Estimation, Testing and Interpretation

Herman J. Bierens

October 25, 2008

1 Introduction to maximum likelihood estimation

1.1 The likelihood function

Consider a random sample Y_1, \dots, Y_n from the Bernoulli distribution:

$$\begin{aligned}\Pr[Y_j = 1] &= p_0 \\ \Pr[Y_j = 0] &= 1 - p_0,\end{aligned}$$

where p_0 is unknown. For example, toss n times a coin for which you suspect that it is unfair: $p_0 \neq 0.5$, and for each tossing j assign $Y_j = 1$ if the outcome is heads and $Y_j = 0$ if the outcome is tails. The question is how to estimate p_0 and how to test the null hypothesis that the coin is fair: $p_0 = 0.5$.

The probability function involved can be written as

$$\begin{aligned}f(y|p_0) &= \Pr[Y_j = y] \\ &= p_0^y (1 - p_0)^{1-y} = \begin{cases} p_0 & \text{if } y = 1, \\ 1 - p_0 & \text{if } y = 0. \end{cases}\end{aligned}$$

Next, let y_1, \dots, y_n be a given sequence of zeros and ones. Thus, each y_j is either 0 or 1. The joint probability function of the random sample Y_1, Y_2, \dots, Y_n is defined as

$$f_n(y_1, \dots, y_n|p_0) = \Pr[Y_1 = y_1 \text{ and } Y_2 = y_2 \dots \text{ and } Y_n = y_n].$$

Because the random variables Y_1, Y_2, \dots, Y_n are independent, we can write

$$\begin{aligned}\Pr[Y_1 &= y_1 \text{ and } Y_2 = y_2 \dots \text{ and } Y_n = y_n] \\ &= \Pr[Y_1 = y_1] \times \Pr[Y_2 = y_2] \times \dots \times \Pr[Y_n = y_n] \\ &= f(y_1|p_0) \times f(y_2|p_0) \times \dots \times f(y_n|p_0) \\ &= \prod_{j=1}^n f(y_j|p_0),\end{aligned}$$

hence

$$\begin{aligned}f_n(y_1, \dots, y_n|p_0) &= \prod_{j=1}^n p_0^{y_j} (1-p_0)^{1-y_j} \\ &= \left(\prod_{j=1}^n p_0^{y_j} \right) \left(\prod_{j=1}^n (1-p_0)^{1-y_j} \right) \\ &= p_0^{\sum_{j=1}^n y_j} (1-p_0)^{n-\sum_{j=1}^n y_j}.\end{aligned}$$

Replacing the given non-random sequence y_1, \dots, y_n by the random sample Y_1, Y_2, \dots, Y_n and the unknown probability p_0 by a variable p in the interval $(0, 1)$ yields the likelihood function

$$L_n(p) = f_n(Y_1, \dots, Y_n|p) = p^{\sum_{j=1}^n Y_j} (1-p)^{n-\sum_{j=1}^n Y_j}$$

For the case $p = p_0$ the likelihood function can be interpreted as the joint probability that we draw a particular sample Y_1, \dots, Y_n .

1.2 Maximum likelihood estimation

The idea of maximum likelihood (ML) estimation is now to choose p such that $L_n(p)$ is maximal. In other words, choose p such that the probability of drawing this particular sample Y_1, \dots, Y_n is maximal.

Note that maximizing $L_n(p)$ is equivalent to maximizing $\ln(L_n(p))$, i.e.,

$$\begin{aligned}\ln(L_n(p)) &= \left(\sum_{j=1}^n Y_j \right) \ln(p) + \left(n - \sum_{j=1}^n Y_j \right) \ln(1-p) \\ &= n \left(\bar{Y} \ln(p) + (1-\bar{Y}) \ln(1-p) \right),\end{aligned}$$

where

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$$

is the sample mean. Therefore, the ML estimator \hat{p} in this case can be obtained from the first-order condition for a maximum of $\ln(L_n(p))$ in $p = \hat{p}$:

$$\begin{aligned} 0 &= \frac{d \ln(L_n(\hat{p}))}{d\hat{p}} = n \left(\bar{Y} \frac{d \ln(\hat{p})}{d\hat{p}} + (1 - \bar{Y}) \frac{d \ln(1 - \hat{p})}{d\hat{p}} \right) \\ &= n \left(\bar{Y} \frac{d \ln(\hat{p})}{d\hat{p}} + (1 - \bar{Y}) \frac{d \ln(1 - \hat{p})}{d(1 - \hat{p})} \times \frac{d(1 - \hat{p})}{d\hat{p}} \right) \\ &= n \left(\bar{Y} \frac{1}{\hat{p}} + (1 - \bar{Y}) \frac{1}{1 - \hat{p}} \times (-1) \right) \\ &= n \left(\frac{\bar{Y}}{\hat{p}} - \frac{1 - \bar{Y}}{1 - \hat{p}} \right) = n \left(\frac{\bar{Y}(1 - \hat{p}) - \hat{p}(1 - \bar{Y})}{\hat{p}(1 - \hat{p})} \right) \\ &= n \left(\frac{\bar{Y} - \hat{p}}{\hat{p}(1 - \hat{p})} \right) \end{aligned}$$

where we have used the fact that $d \ln(x)/dx = 1/x$. Thus, in this case the ML estimator \hat{p} of p_0 is the sample mean:

$$\hat{p} = \bar{Y}.$$

Note that this is an unbiased estimator: $E(\hat{p}) = \frac{1}{n} \sum_{j=1}^n E(Y_j) = p_0$.

1.3 Large sample statistical inference

It can be shown (but this requires advanced probability theory) that if the sample size n is large then $\sqrt{n}(\hat{p} - p_0)$ is approximately normally distributed, i.e.,

$$\sqrt{n}(\hat{p} - p_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^n (Y_j - p_0) \sim N[0, \sigma_0^2],$$

where

$$\begin{aligned} \sigma_0^2 &= \text{var}(Y_j) = E[(Y_j - p_0)^2] \\ &= (1 - p_0)^2 p_0 + (-p_0)^2 (1 - p_0) \\ &= p_0(1 - p_0). \end{aligned}$$

Thus, for large sample size n ,

$$\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1 - p_0)}} \sim N[0, 1]. \quad (1)$$

This result can be used to test hypotheses about p_0 . In particular, under the null hypothesis that the coin is fair, $p_0 = 0.5$, we have

$$2\sqrt{n}(\hat{p} - 0.5) = \frac{\sqrt{n}(\hat{p} - 0.5)}{\sqrt{0.5 \times 0.5}} \sim N[0, 1],$$

Therefore, $2\sqrt{n}(\hat{p} - 0.5)$ can be used as the test statistic of the standard normal test of the null hypothesis $p_0 = 1/2$, as follows. Recall that for a standard normal random variable U , $\Pr [|U| > 1.96] = 0.05$. Thus, under the null hypothesis $p_0 = 1/2$ one would expect that

$$\begin{aligned} \Pr [|2\sqrt{n}(\hat{p} - 0.5)| > 1.96] &= 0.05 \\ \Pr [|2\sqrt{n}(\hat{p} - 0.5)| \leq 1.96] &= 0.95 \end{aligned}$$

If $|2\sqrt{n}(\hat{p} - 0.5)| > 1.96$ then we reject the null hypothesis $p_0 = 1/2$ at the 5% significance level, because this is not what one would expect if the null hypothesis is true, and if $|2\sqrt{n}(\hat{p} - 0.5)| \leq 1.96$ then we accept this null hypothesis, as this result is then in accordance with the null hypothesis $p_0 = 1/2$.

The result (1) can also be used to endow the unknown probability p_0 with a confidence interval, for example the 95% confidence interval, as follows. The result (1) implies

$$\Pr \left[\left| \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1 - p_0)}} \right| \leq 1.96 \right] = 0.95,$$

which, after some straightforward calculations, can be shown to be equivalent to

$$\Pr [p_n \leq p_0 \leq \bar{p}_n] = 0.95$$

where

$$\begin{aligned} p_n &= \frac{n.\hat{p} + (1.96)^2 / 2 - 1.96\sqrt{n.\hat{p}(1 - \hat{p}) + (1.96)^2 / 4}}{n + (1.96)^2} \\ \bar{p}_n &= \frac{n.\hat{p} + (1.96)^2 / 2 + 1.96\sqrt{n.\hat{p}(1 - \hat{p}) + (1.96)^2 / 4}}{n + (1.96)^2} \end{aligned}$$

The interval $[\underline{p}_n, \bar{p}_n]$ is now the 95% confidence interval for p_0 .

1.4 An application election polls

Consider a presidential election with two candidates, candidate A and candidate B , and let p_0 be the fraction of likely voters who favor candidate A , just before the election is held. To predict the outcome of the election, a polling agency draws a random sample of size $n = 3000$, for example, from the population of likely voters.¹ Suppose that 1800 of the respondents express a preference for candidate A . Thus, the fraction of respondents favoring candidate A is $\hat{p} = 0.6$. Substituting $n = 3000$ and $\hat{p} = 0.6$ in the formulas for \underline{p}_n and \bar{p}_n yields

$$\underline{p}_n = 0.58, \bar{p}_n = 0.62$$

Thus, the 95% confidence interval of $100 \times p_0$ is $[58, 62]$. The polling results are therefore stated as: 60% of the likely voters will vote for candidate A , with a margin of error of ± 2 points.

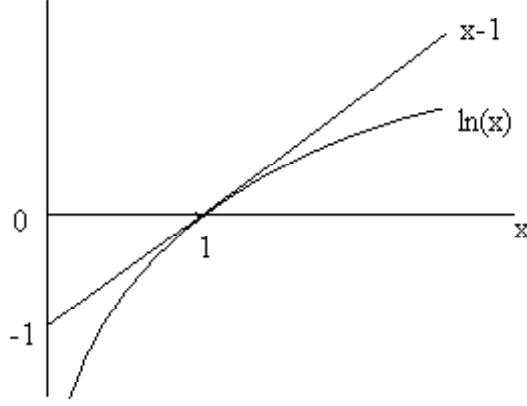
2 Motivation for maximum likelihood estimation

A more formal motivation for ML estimation is based on the fact that for $0 < x < 1$ and $x > 1$,

$$\ln(x) < x - 1.$$

This is illustrated in the following picture:

¹ How to draw such a sample is beyond the scope of this lecture note.



$$\ln(x) \leq x - 1.$$

The inequality $\ln(x) < x - 1$ is strict for $x \neq 1$, and $\ln(1) = 0$. Consequently, taking $x = f(Y_j|p)/f(Y_j|p_0)$, we have the inequality

$$\ln\left(\frac{f(Y_j|p)}{f(Y_j|p_0)}\right) \leq \frac{f(Y_j|p)}{f(Y_j|p_0)} - 1.$$

Taking expectations, it follows that

$$\begin{aligned} E\left[\ln\left(\frac{f(Y_j|p)}{f(Y_j|p_0)}\right)\right] &\leq E\left[\frac{f(Y_j|p)}{f(Y_j|p_0)}\right] - 1 \\ &= \frac{f(1|p)}{f(1|p_0)} \Pr[Y_j = 1] + \frac{f(0|p)}{f(0|p_0)} \Pr[Y_j = 0] - 1 \\ &= \frac{p}{p_0} p_0 + \frac{1-p}{1-p_0} (1-p_0) - 1 \\ &= p + 1 - p - 1 = 0, \end{aligned} \tag{2}$$

hence

$$E[\ln(f(Y_j|p))] - E[\ln(f(Y_j|p_0))] = E\left[\ln\left(\frac{f(Y_j|p)}{f(Y_j|p_0)}\right)\right] \leq 0,$$

and therefore,

$$E[\ln(L_n(p))] \leq E[\ln(L_n(p_0))]. \tag{3}$$

Thus, $E[\ln(L_n(p))]$ is maximal for $p = p_0$, and it can be shown that this maximum is unique.

3 Maximum likelihood estimation of the Logit model

3.1 The Logit model with one explanatory variable

Next, let $(Y_1, X_1), \dots, (Y_n, X_n)$ be a random sample from the conditional Logit distribution:

$$\begin{aligned}\Pr[Y_j = 1|X_j] &= \frac{1}{1 + \exp(-\alpha_0 - \beta_0 X_j)}, \\ \Pr[Y_j = 0|X_j] &= 1 - \Pr[Y_j = 1|X_j] \\ &= \frac{\exp(-\alpha_0 - \beta_0 X_j)}{1 + \exp(-\alpha_0 - \beta_0 X_j)}\end{aligned}\tag{4}$$

where the X_j 's are the explanatory variables and α_0 and β_0 are unknown parameters to be estimated. This model is called a Logit model, because

$$\Pr[Y_j = 1|X_j] = F(\alpha_0 + \beta_0 X_j) \tag{5}$$

where

$$F(x) = \frac{1}{1 + \exp(-x)} \tag{6}$$

is the distribution function of the logistic (Logit) distribution.

The conditional probability function involved is

$$\begin{aligned}f(y|X_j, \alpha_0, \beta_0) &= \Pr[Y_j = y|X_j] \\ &= F(\alpha_0 + \beta_0 X_j)^y (1 - F(\alpha_0 + \beta_0 X_j))^{1-y} \\ &= \begin{cases} F(\alpha_0 + \beta_0 X_j) & \text{if } y = 1, \\ 1 - F(\alpha_0 + \beta_0 X_j) & \text{if } y = 0. \end{cases}\end{aligned}$$

Now the conditional log-likelihood function is

$$\begin{aligned}\ln(L_n(\alpha, \beta)) &= \sum_{j=1}^n \ln(f(Y_j|X_j, \alpha, \beta)) \\ &= \sum_{j=1}^n Y_j \ln(F(\alpha + \beta X_j)) + \sum_{j=1}^n (1 - Y_j) \ln(1 - F(\alpha + \beta X_j)) \\ &= - \sum_{j=1}^n (1 - Y_j)(\alpha + \beta X_j) - \sum_{j=1}^n \ln(1 + \exp(-\alpha - \beta X_j)).\end{aligned}\tag{7}$$

Similar to (3) we have

$$E [\ln (L_n(\alpha, \beta)) | X_1, \dots, X_n] \leq E [\ln (L_n(\alpha_0, \beta_0)) | X_1, \dots, X_n].$$

Again, this result motivates to estimate α_0 and β_0 by maximizing $\ln (L_n(\alpha, \beta))$ to α and β :

$$\ln (L_n(\hat{\alpha}, \hat{\beta})) = \max_{\alpha, \beta} \ln (L_n(\alpha, \beta)).$$

However, there is no longer an explicit solution for $\hat{\alpha}$ and $\hat{\beta}$. These ML estimators have to be solved numerically. Your econometrics software will do that for you.

3.2 Pseudo t-values

It can be shown that if the sample size n is large then

$$\sqrt{n} (\hat{\alpha} - \alpha_0) \sim N(0, \sigma_\alpha^2), \quad \sqrt{n} (\hat{\beta} - \beta_0) \sim N(0, \sigma_\beta^2).$$

Given consistent estimators $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\beta^2$ of the unknown variances σ_α^2 and σ_β^2 , respectively (which are computed by your econometrics software), we then have

$$\frac{\sqrt{n} (\hat{\alpha} - \alpha_0)}{\hat{\sigma}_\alpha} \sim N(0, 1), \quad \frac{\sqrt{n} (\hat{\beta} - \beta_0)}{\hat{\sigma}_\beta} \sim N(0, 1).$$

These results can be used to test whether the coefficients α_0 and β_0 are zero or not. In particular the null hypothesis $\beta_0 = 0$ is of interest, because this hypothesis implies that the conditional probability $\Pr[Y_j = 1 | X_j]$ does not depend on X_j . Under the null hypothesis $\beta_0 = 0$ we have

$$\hat{t}_\beta = \frac{\sqrt{n} \hat{\beta}}{\hat{\sigma}_\beta} \sim N(0, 1).$$

Recall that the 5% critical value of the two-sided standard normal test is 1.96. Thus, for example, the null hypothesis $\beta_0 = 0$ is rejected at the 5% significance level in favor of the alternative hypothesis $\beta_0 \neq 0$ if $|\hat{t}_\beta| > 1.96$, and accepted if $|\hat{t}_\beta| \leq 1.96$.

The statistic \hat{t}_β is called the *pseudo t-value* of $\hat{\beta}$ because it is used in the same way as the t-value in linear regression, and $\hat{\sigma}_\beta$ is called the standard error of $\hat{\beta}$. Your econometric software will report the ML estimators together with their corresponding pseudo t-values and/or standard errors.

3.3 The general Logit model

The general Logit model takes the form

$$\begin{aligned}\Pr[Y_j = 1 | X_{1j}, \dots, X_{kj}] &= \frac{1}{1 + \exp(-\beta_1^0 X_{1j} - \dots - \beta_k^0 X_{kj})} \\ &= \frac{1}{1 + \exp\left(-\sum_{i=1}^k \beta_i^0 X_{ij}\right)},\end{aligned}\tag{8}$$

where one of the X_{ij} equals 1 for the constant term, for example, let $X_{kj} = 1$, and the β_i^0 's are the true parameter values. This model can be estimated by ML in the same way as before. Thus, the log-likelihood function is

$$\ln(L_n(\beta_1, \dots, \beta_k)) = -\sum_{j=1}^n (1 - Y_j) \sum_{i=1}^k \beta_i X_{ij} - \sum_{j=1}^n \ln\left(1 + \exp\left(-\sum_{i=1}^k \beta_i X_{ij}\right)\right),\tag{9}$$

and the ML estimators $\hat{\beta}_1, \dots, \hat{\beta}_k$ are obtained by maximizing $\ln(L_n(\beta_1, \dots, \beta_k))$:

$$\ln(L_n(\hat{\beta}_1, \dots, \hat{\beta}_k)) = \max_{\beta_1, \dots, \beta_k} \ln(L_n(\beta_1, \dots, \beta_k)).$$

Again, it can be shown that if n is large then for $i = 1, \dots, k$,

$$\sqrt{n}(\hat{\beta}_i - \beta_i^0) \sim N[0, \sigma_i^2].$$

Given consistent estimators $\hat{\sigma}_i^2$ of the variances σ_i^2 , it follows then that

$$\frac{\sqrt{n}(\hat{\beta}_i - \beta_i^0)}{\hat{\sigma}_i} \sim N[0, 1]$$

for $i = 1, \dots, k$. Your econometrics software will report the ML estimators $\hat{\beta}_i$ together with their corresponding pseudo t-values $\hat{t}_i = \sqrt{n}\hat{\beta}_i/\hat{\sigma}_i$ and/or standard errors $\hat{\sigma}_i$.

3.4 Testing joint significance

Now suppose you want to test the joint null hypothesis

$$H_0: \beta_1^0 = 0, \beta_2^0 = 0, \dots, \beta_m^0 = 0,\tag{10}$$

where $m < k$.

There are two ways to do that. One way is akin to the F test in linear regression: Re-estimate the Logit model under the null hypothesis:

$$\ln(L_n(0, \dots, 0, \tilde{\beta}_{m+1}, \dots, \tilde{\beta}_k)) = \max_{\beta_{m+1}, \dots, \beta_k} \ln(L_n(0, \dots, 0, \beta_{m+1}, \dots, \beta_k)).$$

and compare the log-likelihoods². It can be shown that under the null hypothesis (10) and for large samples,

$$LR_m = -2 \ln \left(\frac{L_n(0, \dots, 0, \tilde{\beta}_{m+1}, \dots, \tilde{\beta}_k)}{L_n(\hat{\beta}_1, \dots, \hat{\beta}_k)} \right) \sim \chi_m^2,$$

where the degrees of freedom m corresponds to the number of restrictions imposed under the null hypothesis. This is the so-called likelihood ratio test, which is conducted right-sided. For example, choose the 5% significance level, look up in the table of the χ^2 distribution the critical value c such that for a χ_m^2 distributed random variable Z_m , $\Pr[Z_m > c] = 0.05$. Then the null hypothesis (10) is rejected at the 5% significance level if $LR_m > c$ and accepted if $LR_m \leq c$.

An alternative test of the null hypothesis (10) is the Wald test, which is conducted in the same way as for linear regression models.³ Under the null hypothesis (10) the Wald test statistic has also a χ_m^2 distribution.

4 Interpretation of the coefficients of the Logit model

4.1 Marginal effects

Consider the Logit model (5). If $\beta_0 > 0$ then $\Pr[Y_j = 1 | X_j] = F(\alpha_0 + \beta_0 X_j)$ is an increasing function of X_j :

$$\frac{dP[Y_j = 1 | X_j]}{dX_j} = \beta_0 \cdot F'(\alpha_0 + \beta_0 X_j),$$

where F' is the derivative of (6):

²Your econometric software will report the log-likelihood function value.

³In *EasyReg International* the Wald test can be conducted simply by point-and-click.

$$\begin{aligned}
F'(x) &= \frac{\exp(-x)}{(1 + \exp(-x))^2} = \frac{1 + \exp(-x)}{(1 + \exp(-x))^2} - \frac{1}{(1 + \exp(-x))^2} \\
&= \frac{1}{1 + \exp(-x)} - \frac{1}{(1 + \exp(-x))^2} = F(x) - F(x)^2 \\
&= F(x)(1 - F(x)).
\end{aligned}$$

Therefore, the marginal effect of X_j on $\Pr[Y_j = 1|X_j]$ depends on X_j :

$$\frac{dP[Y_j = 1|X_j]}{dX_j} = \beta_0 \cdot F(\alpha_0 + \beta_0 X_j) (1 - F(\alpha_0 + \beta_0 X_j)),$$

which renders the interpretation of β_0 difficult.

However, the coefficient β_0 can be interpreted in terms of relative changes in odds.

4.2 Odds and odds ratios

The odds is the ratio of the probability that something is true divided by the probability that it is not true. Thus, in the Logit case (4),

$$\text{Odds}(X) = \frac{\Pr[Y_j = 1|X_j]}{\Pr[Y_j = 0|X_j]} = \frac{F(\alpha_0 + \beta_0 X_j)}{1 - F(\alpha_0 + \beta_0 X_j)} = \exp(\alpha_0 + \beta_0 X_j). \quad (11)$$

The odds ratio is the ratio of two odds for different values of X_j , say $X_j = x$ and $X_j = x + \Delta x$:

$$\frac{\text{Odds}(x + \Delta x)}{\text{Odds}(x)} = \frac{\exp(\alpha + \beta x + \beta \Delta x)}{\exp(\alpha + \beta x)} = \exp(\beta \Delta x),$$

where Δx is a small change in x . Then

$$\begin{aligned}
&\lim_{\Delta x \rightarrow 0} \frac{1}{\Delta x} \left(\frac{\text{Odds}(x + \Delta x) - \text{Odds}(x)}{\text{Odds}(x)} \right) = \lim_{\Delta x \rightarrow 0} \frac{\exp(\beta_0 \Delta x) - 1}{\Delta x} \\
&= \beta_0 \lim_{\beta_0 \Delta x \rightarrow 0} \frac{\exp(\beta_0 \Delta x) - 1}{\beta_0 \Delta x} = \beta_0 \times \frac{d \exp(u)}{du} \Big|_{u=0} = \beta_0 \exp(0) = \beta_0.
\end{aligned}$$

Thus, β_0 may be interpreted as the *relative* change in the odds due to a small change Δx in X_j :

$$\frac{\text{Odds}(x + \Delta x) - \text{Odds}(x)}{\text{Odds}(x)} = \frac{\text{Odds}(x + \Delta x)}{\text{Odds}(x)} - 1 \approx \beta_0 \Delta x \quad (12)$$

If X_j is a binary variable itself, $X_j = 0$ or $X_j = 1$, then the only reasonable choices for $x + \Delta x$ and x are 1 and 0, respectively, so that then

$$\frac{\text{Odds}(1)}{\text{Odds}(0)} - 1 = \frac{\text{Odds}(1) - \text{Odds}(0)}{\text{Odds}(0)} = \exp(\beta_0) - 1.$$

Only if β_0 is small we may then use the approximation $\exp(\beta_0) - 1 \approx \beta_0$. If not, one has to interpret β_0 in terms of the log of the odds ratio involved:

$$\ln\left(\frac{\text{Odds}(1)}{\text{Odds}(0)}\right) = \beta_0.$$

The interpretation of the coefficients $\beta_i^0, i = 1, \dots, k - 1$ in the general Logit model (8) is similar as in the case (12):

$$\frac{\text{Odds}(X_{1j}, \dots, X_{i-1,j}, X_{i,j} + \Delta X_{i,j}, X_{i+1,j}, \dots, X_{k,j})}{\text{Odds}(X_{1j}, \dots, X_{i-1,j}, X_{i,j}, X_{i+1,j}, \dots, X_{k,j})} - 1 \approx \beta_i^0 \Delta X_{i,j}$$

if $\Delta X_{i,j}$ is small. For example, β_i^0 may be interpreted as the percentage change in $\text{Odds}(X_{1j}, \dots, X_{k,j})$ due to a small percentage change $100 \times \Delta X_{i,j} = 1$ in $X_{i,j}$.

Comparison of Probit and Logit Analysis

The following figure compares the standard normal density $f(x)$ with the density $g(x)$ of the **rescaled** Logit distribution

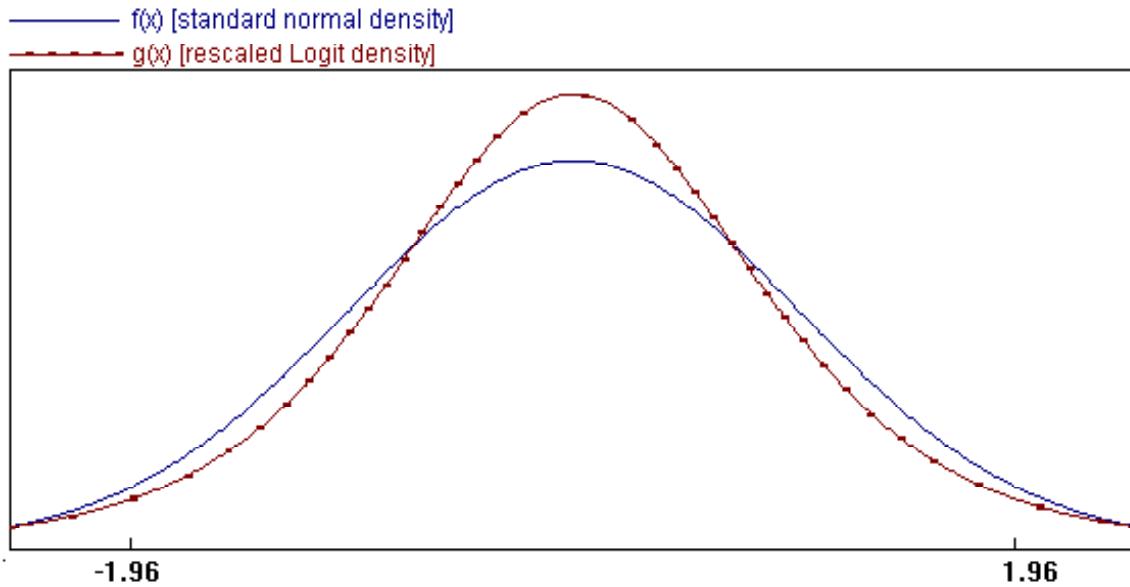
$$G(x) = \frac{1}{1 + \exp(-x/\sigma)},$$

i.e.,

$$g(x) = \frac{1}{\sigma} G(x) (1 - G(x)),$$

where σ is chosen such that $G(1.96) = 0.975$ as for the standard normal distribution. This is the case for

$$\sigma = 0.5349985$$



We see that $f(x)$ is somewhat flatter than $g(x)$. Nevertheless we may expect that Probit and Logit analyses for the same data yield similar result, taking

into account the rescaling. To check this, I have generated data according to a standard Probit model

$$\Pr [Y_j = 1|X_j] = F(\alpha + \beta X_j), \quad F(x) = \int_{-\infty}^x \frac{\exp(-u^2/2)}{\sqrt{2\pi}} du,$$

for $j = 1, \dots, n = 500$, with $\alpha = \beta = 1$, and the X_j 's drawn from the standard normal distribution. If we estimate this model as a standard Logit model, we may expect that the Logit estimates $\hat{\alpha}_L, \hat{\beta}_L$ are related to the Probit estimates $\hat{\alpha}_P, \hat{\beta}_P$ as follows

$$\hat{\alpha}_P \approx \sigma \hat{\alpha}_L, \quad \hat{\beta}_P \approx \sigma \hat{\beta}_L$$

The Probit estimation results are

$$\hat{\alpha}_P = 1.117529, \quad \hat{\beta}_P = 1.147434$$

and the Logit estimation results are

$$\hat{\alpha}_L = 1.928988, \quad \hat{\beta}_L = 2.019920$$

Thus,

$$\hat{\alpha}_P/\hat{\alpha}_L = 0.579334, \quad \hat{\beta}_P/\hat{\beta}_L = 0.568059$$

which are reasonably close to $\sigma = 0.5349985$. Therefore, the Probit parameter estimates are between 50% and 60% smaller in absolute value than the corresponding Logit parameter estimates.

Modeling fractions

Herman J. Bierens

December 12, 2007

1 Modeling a single fraction

Let Y be a dependent variable which is bounded between zero and one, $Y \in (0, 1)$, for example if Y is a fraction. A possible way to model the distribution of Y conditional on a vector X of predetermined variables, including 1 for the constant term, is to assume that

$$Y = \frac{\exp(\beta'X + U)}{1 + \exp(\beta'X + U)} = \frac{1}{1 + \exp(-\beta'X - U)}, \quad (1)$$

where U is an unobserved error term. Then

$$\ln[Y/(1 - Y)] = \beta'X + U, \quad (2)$$

which, under standard assumptions on the error term U , can be estimated by OLS.

Note that model (1) is of the form

$$Y = F(\beta'X + U),$$

where $F(x) = (1 + \exp(-x))^{-1}$ is the logistic distribution function. This distribution function is convenient because its inverse F^{-1} has a closed form: If $y = F(x)$ then $x = F^{-1}(y) = \ln[y/(1 - y)]$.

Of course, if Y is a percentage rather than a fraction, i.e., $Y \in (0, 100)$, then (2) has to be modified to

$$\ln[Y/(100 - Y)] = \beta'X + U.$$

As to the interpretation of the coefficients in the vector β , let $\beta = (\beta_1, \dots, \beta_m)'$ and $X = (X_1, \dots, X_m)'$, where $X_m = 1$ for the constant term. It follows from (1) that for $i = 1, \dots, m - 1$,

$$\frac{\partial Y}{\partial X_i} = \frac{\exp(-\beta' X - U)\beta_i}{(1 + \exp(-\beta' X - U))^2} = Y^2 \exp(-\beta' X - U)\beta_i = \frac{\partial Y}{\partial U} \cdot \beta_i \quad (3)$$

Thus, β_i measures the marginal effect of X_i on Y relative to the marginal effect of the error term U on Y :

$$\beta_i = \frac{\partial Y / \partial X_i}{\partial Y / \partial U}.$$

2 Dynamic fraction models

In the presence of a lagged dependent variables model (2) can be modified to

$$\ln[Y_t/(1 - Y_t)] = \alpha \ln[Y_{t-1}/(1 - Y_{t-1})] + \beta' X_t + U_t, \quad (4)$$

or

$$\ln[Y_t/(1 - Y_t)] = \gamma Y_{t-1} + \beta' X_t + U_t, \quad (5)$$

for example.

Because

$$\frac{d \ln[Y_{t-1}/(1 - Y_{t-1})]}{d Y_{t-1}} = \frac{1}{Y_{t-1}} - \frac{1}{1 - Y_{t-1}}$$

it follows from (3) that in the case of model (4),

$$\begin{aligned} \frac{\partial Y_t}{\partial Y_{t-1}} &= \frac{\partial Y_t}{\partial \ln[Y_{t-1}/(1 - Y_{t-1})]} \times \frac{d \ln[Y_{t-1}/(1 - Y_{t-1})]}{d Y_{t-1}} \\ &= \left(\frac{1}{Y_{t-1}} - \frac{1}{1 - Y_{t-1}} \right) \frac{\partial Y_t}{\partial U_t} \cdot \alpha \end{aligned}$$

so that the previous interpretation of α in terms of relative marginal effects no longer applies. On the other hand, in the case of model (5) we have

$$\gamma = \frac{\partial Y_t / \partial Y_{t-1}}{\partial Y_t / \partial U_t}.$$

Nevertheless, I would prefer model (4) over model (5) because the dynamic properties of model (4) are standard. In particular, if $|\alpha| < 1$ then under some further regularity conditions it follows by backwards substitution that

$$\ln[Y_t/(1 - Y_t)] = \sum_{j=0}^{\infty} \alpha^j \beta' X_{t-j} + \sum_{j=0}^{\infty} \alpha^j U_{t-j}.$$

Consequently, with $X_{i,t-j}$ component i of X_{t-j} ,

$$\begin{aligned} \frac{\partial Y_t}{\partial X_{i,t-j}} &= \frac{\exp\left(-\sum_{j=0}^{\infty} \alpha^j \beta' X_{t-j} - \sum_{j=0}^{\infty} \alpha^j U_{t-j}\right) \alpha^j \beta_i}{\left(1 + \exp\left(-\sum_{j=0}^{\infty} \alpha^j \beta' X_{t-j} - \sum_{j=0}^{\infty} \alpha^j U_{t-j}\right)\right)^2} \\ \frac{\partial Y_t}{\partial U_{t-j}} &= \frac{\exp\left(-\sum_{j=0}^{\infty} \alpha^j \beta' X_{t-j} - \sum_{j=0}^{\infty} \alpha^j U_{t-j}\right) \alpha^j}{\left(1 + \exp\left(-\sum_{j=0}^{\infty} \alpha^j \beta' X_{t-j} - \sum_{j=0}^{\infty} \alpha^j U_{t-j}\right)\right)^2} \end{aligned}$$

Hence

$$\beta_i = \frac{\partial Y_t / \partial X_{i,t-j}}{\partial Y_t / \partial U_{t-j}}, \quad \alpha = \frac{\partial Y_t / \partial U_{t-j-1}}{\partial Y_t / \partial U_{t-j}}$$

for $j = 0, 1, 2, \dots$

3 Modeling multiple fractions

Now suppose that we have multiple fractions Y_0, \dots, Y_k as dependent variables, i.e., each $Y_i \in (0, 1)$, and $\sum_{i=0}^k Y_i = 1$. For example, let Y_i be the fraction of people smoking a particular brand $i \in \{1, \dots, k\}$ of cigarettes, with $Y_0 = 1 - \sum_{i=1}^k Y_i$ the fraction of all other people. The question now is how to model the joint distribution of Y_0, \dots, Y_k conditional on a vector X of predetermined variables such that the conditions $Y_i \in (0, 1)$, and $\sum_{i=0}^k Y_i = 1$ hold, and how to estimate the parameters of this model. A possible way is the following.

Let for $i = 1, \dots, k$,

$$Y_i = \frac{\exp(\beta'_i X + U_i)}{1 + \sum_{j=1}^k \exp(\beta'_j X + U_j)},$$

and

$$Y_0 = \frac{1}{1 + \sum_{j=1}^k \exp(\beta'_j X + U_j)},$$

where the U_i 's are unobserved error terms. Then

$$\ln(Y_i/Y_0) = \beta'_i X + U_i, \quad i = 1, \dots, k, \quad (6)$$

which is a system of seemly unrelated regressions (SUR). If the vector X is common and there are no restrictions on the parameters β_i , SUR estimation is equivalent to estimating each equation separately by OLS. Otherwise one has to conduct SUR estimation in the usual (textbook) way. The latter can be done in EasyReg via the GMM module.

Note that model (6) also applies if Y_0, \dots, Y_k are percentages which add up to 100%.

FORECASTING
Herman J. Bierens
Pennsylvania State University

April 30, 2012

1. *Conditional expectations as best forecasting schemes*

Consider a pair of random variables, X and Y , for which you know the joint distribution. Suppose that Y is not yet observed, and that you want to forecast Y , given that you have observed X . The question is: what is the best forecasting scheme?

Any forecasting scheme for Y is a function of X , for example

$$\hat{Y} = \varphi(X). \quad (1)$$

Using (1) as a forecast of Y , the mean-square forecast error involved is defined as

$$E[(Y - \hat{Y})^2] = E[(Y - \varphi(X))^2]. \quad (2)$$

Now the best forecasting scheme is the function (1) that minimizes (2). So now the question arises: for which function φ is the mean-square forecast error (2) minimal?

The answer is: the conditional expectation function: $\varphi(X) = E[Y|X]$, as will be shown in the Appendix, section A.1. Then the forecast error is $U = Y - \varphi(X) = Y - E[Y|X]$, which satisfies $E[U|X] = E[Y|X] - E[\varphi(X)|X] = \varphi(X) - \varphi(X) = 0$. Thus, we can always write

$$Y = \varphi(X) + U, \text{ where } E[U|X] = 0. \quad (3)$$

which is a general form of a regression model, with U the error term and $\varphi(X)$ the regression function. The property $E[U|X] = 0$ is the usual assumption about the error term of a regression model. Thus, this assumption implies that the regression function $\varphi(X)$ is equal to the conditional expectation of Y given X : $\varphi(X) = E[Y|X]$. Therefore, for a regression model to be correctly specified, the regression function has to represent the conditional expectation of the dependent variable given the regressors.

2. *Out-of-sample forecasting with a linear regression model*

Consider the linear regression model

$$Y_j = \alpha + \beta \cdot X_j + U_j, \quad j = 1, 2, \dots, n, \quad (4)$$

where the unobserved error terms are assumed to be independent normally distributed:

$$U_j \sim i.i.d. N(0, \sigma^2), \quad (5)$$

and each U_j is independent of all the explanatory variables $X_1, X_2, \dots, X_n, \dots$. Let $\hat{\beta}$ and $\hat{\alpha}$ be the OLS estimators of β and α in model (4) on the basis of the observations (Y_j, X_j) for $j = 1, \dots, n$.

Next, suppose we observe X_{n+1} . Then the one-step ahead out-of-sample forecast of Y_{n+1} is

$$\hat{Y}_{n+1} = \hat{\alpha} + \hat{\beta} \cdot X_{n+1}, \quad (6)$$

where the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are computed on the basis of the observations for $j = 1, 2, \dots, n$.

The actual but unknown value of Y_{n+1} is $Y_{n+1} = \alpha + \beta \cdot X_{n+1} + U_{n+1}$, so that the forecast error is $Y_{n+1} - \hat{Y}_{n+1}$. In the Appendix it will be shown that

$$Y_{n+1} - \hat{Y}_{n+1} \sim N[0, \sigma_{Y_{n+1} - \hat{Y}_{n+1}}^2], \text{ where } \sigma_{Y_{n+1} - \hat{Y}_{n+1}}^2 = \sigma^2 \left(\frac{n+1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right). \quad (7)$$

Moreover, denoting,

$$\hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}}^2 = \hat{\sigma}^2 \left(\frac{n+1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right), \quad (8)$$

where $\hat{\sigma}^2$ is the (unbiased) OLS estimator of σ^2 :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{j=1}^n (Y_j - \hat{\alpha} - \hat{\beta} \cdot X_j)^2, \quad (9)$$

it can be shown that

Proposition 1. *In the case of model (4), $(Y_{n+1} - \hat{Y}_{n+1})/\hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}} \sim t_{n-2}$.*

This result can be used to construct a 95% confidence interval, say, of Y_{n+1} . Look up in

the table of the t distribution the critical value t_* of the two-sided t-test at the 5% significance level with $n-2$ degrees of freedom. Note that if $n-2 > 30$ then t_* is approximately equal to the critical value of the two-sided standard normal test at the 5% significance level: $t_* \approx 1.96$. Then it follows from Proposition 1 that

$$\begin{aligned} P[-t_* \leq (Y_{n+1} - \hat{Y}_{n+1})/\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}} \leq t_*] &= P[-t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}} \leq Y_{n+1} - \hat{Y}_{n+1} \leq t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}] \\ &= P[\hat{Y}_{n+1} - t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}} \leq Y_{n+1} \leq \hat{Y}_{n+1} + t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}] = 0.95 \end{aligned} \quad (10)$$

Thus, the 95% confidence interval of Y_{n+1} is $[\hat{Y}_{n+1} - t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}, \hat{Y}_{n+1} + t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}]$.

Observe from (8) that $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}^2$ increases with $(X_{n+1} - \bar{X})^2$, and so does the width of the confidence interval. Thus, the farther X_{n+1} is away from \bar{X} , the more unreliable the forecast \hat{Y}_{n+1} of Y_{n+1} becomes. Also observe from (8) that $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}^2 \geq \hat{\sigma}^2$, and that $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}^2$ gets close to $\hat{\sigma}^2$ if n is large because $\lim_{n \rightarrow \infty} \sum_{j=1}^n (X_j - \bar{X})^2 = \infty$.

These results apply to general regression models as well, with only a minor modification. Consider the general linear regression model

$$Y_j = \beta_0 + \beta_1 \cdot X_{1,j} + \dots + \beta_k \cdot X_{k,j} + U_j, \quad j = 1, 2, \dots, n, \quad (11)$$

where again the error terms U_j satisfy the normality condition (5) and are independent of all the regressors $X_{i,t}$ for $i = 1, \dots, k$ and all observation indices t . If the regressors $X_{i,n+1}$ are observed then we can forecast Y_{n+1} by

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_{1,n+1} + \dots + \hat{\beta}_k \cdot X_{k,n+1}, \quad (12)$$

where the $\hat{\beta}_i$'s are the OLS estimators of the corresponding parameters β_i on the basis of the observations for $j = 1, \dots, n$. It is no longer possible to give an explicit expression for the forecast standard error $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}$ without using matrix calculus, but your econometrics software will compute $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}$ or $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}^2$. Then

Proposition 2. *In the case of model (12), $(Y_{n+1} - \hat{Y}_{n+1})/\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}} \sim t_{n-k-1}$.*

The difference with the result in Proposition 1 are the degrees of freedom of the t distributions involved. In Proposition 1 the degrees of freedom is the sample size n minus the number of parameters ($= 2$) in model (4), whereas in Proposition 2 the degrees of freedom is the sample size

n minus the number of parameters ($= k+1$) in model (12).

Also in this case $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}^2$ increases with $(X_{i,n+1} - \bar{X}_i)^2$, where $\bar{X}_i = (1/n)\sum_{j=1}^n X_{i,j}$.

The construction of the 95% confidence interval of Y_{n+1} can be done in the same way as before. Look up in the table of the t distribution the critical value t_* of the two-sided t-test at the 5% significance level with $n-k-1$ degrees of freedom (instead of $n-2$ degrees of freedom!). Note again that if $n-k-1 > 30$ then t_* is approximately equal to the critical value of the two-sided standard normal test at the 5% significance level: $t_* \approx 1.96$. Then it follows from

Proposition 2 that $[\hat{Y}_{n+1} - t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}, \hat{Y}_{n+1} + t_* \hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}]$ is the 95% confidence interval of Y_{n+1} .

3. Forecasting stationary time series

Let Y_t be a time series process, and suppose that we have observed Y_t for all time periods $t \leq n$. The problem is how to predict the unknown future value Y_{n+1} , given that we have observed Y_t for all time periods $t \leq n$. In practice of course we only observe Y_t from some initial time period onwards, say $t = 1$, but I will deal with that problem later.

The answer is similar to the simple case considered in section 1, namely the best predictor \hat{Y}_{n+1} of Y_{n+1} is the conditional expectation of Y_{n+1} given the whole past of the time series up to the last period $t = n$:

$$\hat{Y}_{n+1} = E[Y_{n+1} | Y_n, Y_{n-1}, Y_{n-2}, Y_{n-3}, \dots]. \quad (13)$$

Also in this case the right-hand side of (13) is a function of the conditioning variables, but this is in general a function with an infinite number of arguments, and its shape may depend on n as well:

$$E[Y_{n+1} | Y_n, Y_{n-1}, Y_{n-2}, Y_{n-3}, \dots] = g_n(Y_n, Y_{n-1}, Y_{n-2}, Y_{n-3}, \dots) \quad (14)$$

If indeed the shape of $g_n(\cdot)$ changes with n , there is no way to determine its shape, so that (14) is not a feasible predictor. For the shape of $g_n(\cdot)$ to be constant, $g_n(\cdot) = g(\cdot)$, we need to require that the time series Y_t is **stationary**:

Definition 1: A time series process Y_t is said to be strictly stationary if for arbitrary integers $m_1 < m_2 < \dots < m_n$ the joint distribution of $Y_{t-m_1}, \dots, Y_{t-m_n}$ does not depend on the time index t .

A weaker version of stationarity is **covariance stationarity**, which requires that the expectations and covariances are well-defined and do not depend on the time index t .

Definition 2: A time series process Y_t is said to be covariance stationary (or weakly stationary) if for all integers t and m , the expectations $E[Y_t]$ and the covariances $E[(Y_t - \mu)(Y_{t-m} - \mu)]$ are finite and do not depend on the time index t : $E[Y_t] = \mu$ and $E[(Y_t - \mu)(Y_{t-m} - \mu)] = \gamma(m)$. The function $\gamma(m)$ is called the covariance function.

Note that a strictly stationary time series process Y_t is covariance stationary if $E[Y_t^2] < \infty$, because then $E[Y_t]$ and $E[(Y_t - \mu)(Y_{t-m} - \mu)]$ are finite and do not depend on t .

Definition 3: A covariance stationary process Y_t is said to be Gaussian if for any finite sequence a_i , $i = 0, 1, \dots, m$, of coefficients, $X_t = \sum_{i=0}^m a_i Y_{t-i}$ is normally distributed.

It can now be shown:

Proposition 3. If Y_t strictly stationary and $E[Y_t^2] < \infty$, then the shape of the function $g_n(\cdot)$ in (14) does not change with n : There exists a function $g(\cdot)$ such that for all t ,

$$E[Y_t | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots] = g(Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots). \quad (15)$$

If Y_t is covariance stationary and Gaussian then Y_t is strictly stationary, and the conditional expectation function g is linear:

$$g(Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots) = \beta_0 + \sum_{j=1}^{\infty} \beta_j Y_{t-j}. \quad (16)$$

Moreover, in that case the random variables

$$U_t = Y_t - E[Y_t | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots] \quad (17)$$

are independent normally distributed with zero expectation and constant variance, and for each t , U_t is independent of $Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots$

Thus, a stationary Gaussian process Y_t can be written as

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + \dots + U_t \\ &= \beta_0 + \sum_{j=1}^{\infty} \beta_j Y_{t-j} + U_t, \text{ where } U_t \sim i.i.d. N(0, \sigma^2). \end{aligned} \quad (18)$$

Model (18) is the general linear autoregressive model. Almost all time series models for univariate stationary time series¹ are special cases of (18).

Note that similar to (3),

$$E[U_t | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots] = 0, \quad (19)$$

regardless whether Y_t is Gaussian or not.

In order for (18) to be usable for forecasting, we need to estimate the parameters β_j for $j = 0, 1, 2, \dots$. In general this will not be possible without imposing restrictions on the β_j 's, because it is impossible to estimate an infinite number of parameters.

If we assume that $\beta_j = 0$ for all $j > p$, where p is a given natural number, we get an Auto-Regression of order p , shortly an AR(p) model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + U_t. \quad (20)$$

If the order p in (20) is known and not too large, this model can be estimated by OLS. In practice we do not know p , but there are various ways to determine p , as I will show below.

4. The AR(1) model

The simplest AR(p) model is the one for the case $p = 1$:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + U_t. \quad (21)$$

In the Appendix it will be shown, under the assumption that the errors U_t are independent $N(0, \sigma^2)$ distributed, that a necessary condition for the stationarity of the process (21) is that

$$|\beta_1| < 1. \quad (22)$$

and that then, by repeated backwards substitution of (21), we can write Y_t as

$$Y_t = \frac{\beta_0}{1 - \beta_1} + \sum_{k=0}^{\infty} \beta_1^k U_{t-k}. \quad (23)$$

¹ The exceptions are ARCH and GARCH models, which will be discussed in the last section.

This is called the **infinite-order moving average** [MA(∞)] representation of the AR(1) process (21).

From this expression we can derive the covariance function of the AR(1) process, as follows. Denoting $\mu = \beta_0/(1-\beta_1)$, it follows from (23) that

$$\begin{aligned}
\gamma(m) &= E[(Y_t - \mu)(Y_{t-m} - \mu)] = E\left[\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-k}\right)\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)\right] \\
&= E\left[\left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k} + \sum_{k=m}^{\infty} \beta_1^k U_{t-k}\right)\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)\right] \\
&= E\left[\left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k} + \beta_1^m \sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)\right] \\
&= E\left[\left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k}\right)\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)\right] + \beta_1^m E\left[\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)^2\right] \\
&= \beta_1^m E\left[\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)^2\right] = \beta_1^m \sigma^2 \sum_{k=0}^{\infty} \beta_1^{2k} = \sigma^2 \beta_1^m / (1 - \beta_1^2).
\end{aligned} \tag{24}$$

6. Lag operators

A lag operator L is the instruction to shift the time back with one period: $LY_t = Y_{t-1}$. If we apply the lag operator again we get $L^2Y_t = L(LY_t) = LY_{t-1} = Y_{t-2}$, and more generally

$$L^m Y_t \stackrel{\text{def.}}{=} Y_{t-m}, \quad m = 0, 1, 2, \dots \tag{25}$$

Using the lag operator, the AR(1) model (21) can be written as

$$(1 - \beta_1 L)Y_t = \beta_0 + U_t \tag{26}$$

In the previous section we have in several places used the equality

$$\sum_{k=0}^{\infty} z^k = \frac{1}{1-z}, \quad \text{provided that } |z| < 1, \tag{27}$$

which follows from the equalities

$$\sum_{k=0}^{\infty} z^k = 1 + \sum_{k=1}^{\infty} z^k = 1 + \sum_{k=0}^{\infty} z^{k+1} = 1 + z \cdot \sum_{k=0}^{\infty} z^k.$$

Now suppose that we may treat $\beta_1 L$ as the variable z in (27). If so, it follows from (27) that

$$\frac{1}{1 - \beta_1 L} = \sum_{k=0}^{\infty} (\beta_1 L)^k = \sum_{k=0}^{\infty} \beta_1^k L^k. \tag{28}$$

Applying this lag polynomial to both sides of (26) then yields

$$\begin{aligned}
Y_t &= \frac{1}{1 - \beta_1 L}(1 - \beta_1 L)Y_t = \sum_{k=0}^{\infty} \beta_1^k L^k \beta_0 + \sum_{k=0}^{\infty} \beta_1^k L^k U_t = \sum_{k=0}^{\infty} \beta_1^k \beta_0 + \sum_{k=0}^{\infty} \beta_1^k U_{t-k} \\
&\quad 7 \\
&= \frac{\beta_0}{1 - \beta_1} + \sum_{k=0}^{\infty} \beta_1^k U_{t-k},
\end{aligned} \tag{29}$$

which is exactly the moving average representation (23). Note that in the second equality in (29) I have used the fact that the lag operator has no effect on a constant: $L\beta_0 = \beta_0$, hence $L^k\beta_0 = \beta_0$. Thus, the equality (28) holds if $|\beta_1| < 1$:

Proposition 4. *The lag polynomial $\sum_{k=0}^{\infty} \beta^k L^k$ may be treated as $1/(1-\beta L)$, in the sense that $(1-\beta L)\sum_{k=0}^{\infty} \beta^k L^k = 1$, provided that $|\beta| < 1$.*

7. The AR(2) model

Consider the AR(2) process

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + U_t, \quad (30)$$

where the errors U_t have the same properties as before. Similar to (26) we can write this model in lag-polynomial form as

$$(1 - \beta_1 L - \beta_2 L^2)Y_t = \beta_0 + U_t. \quad (31)$$

We can always write

$$1 - \beta_1 L - \beta_2 L^2 = (1 - \alpha_1 L)(1 - \alpha_2 L), \quad (32)$$

by solving the equations $\alpha_1 + \alpha_2 = \beta_1$, $\alpha_1 \alpha_2 = -\beta_2$,² so that (31) can be written as

$$(1 - \alpha_1 L)(1 - \alpha_2 L)Y_t = \beta_0 + U_t. \quad (33)$$

Now if $|\alpha_1| < 1$ and $|\alpha_2| < 1$ then it follows from Proposition 4 that

²

Although the solutions involved may be complex valued.

$$\begin{aligned}
Y_t &= \frac{1}{(1-\alpha_1 L)(1-\alpha_2 L)} (\beta_0 + U_t) = \left(\sum_{k=0}^{\infty} \alpha_1^k L^{-k} \right) \left(\sum_{m=0}^{\infty} \alpha_2^m L^{-m} \right) \beta_0 + \left(\sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \alpha_1^k \alpha_2^m L^{-k-m} \right) U_t \\
&= \left(\sum_{k=0}^{\infty} \alpha_1^k \right) \left(\sum_{m=0}^{\infty} \alpha_2^m \right) \beta_0 + \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \alpha_1^k \alpha_2^m U_{t-k-m} \\
&= \frac{\beta_0}{(1-\alpha_1)(1-\alpha_2)} + \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \alpha_1^k \alpha_2^m U_{t-k-m} = \frac{\beta_0}{1-\beta_1-\beta_2} + \sum_{j=0}^{\infty} \left(\sum_{m=0}^j \alpha_1^{j-m} \alpha_2^m \right) U_{t-j}
\end{aligned} \tag{34}$$

Note that under the assumption that the errors U_t are independent $N(0, \sigma^2)$ distributed, Y_t in (34) is normally distributed, with expectation $\mu = \beta_0/(1-\beta_1-\beta_2)$ and variance $\sigma^2 \sum_{j=0}^{\infty} \left(\sum_{m=0}^j \alpha_1^{j-m} \alpha_2^m \right)^2$.

Consequently, the necessary conditions for the covariance stationarity of the AR(2) process (30) is that the errors U_t are covariance stationary and that the solutions $1/\alpha_1$ and $1/\alpha_2$ of the equation $0 = 1 - \beta_1 z - \beta_2 z^2 = (1 - \alpha_1 z)(1 - \alpha_2 z)$ are larger than one in absolute value.

Similar conditions apply to general AR(p) processes:

Proposition 5. *The necessary conditions for the covariance stationarity of the AR(p) process (20) are that the errors U_t are covariance stationary and the solutions z_1, \dots, z_p of the equation $0 = 1 - \beta_1 z - \beta_2 z^2 - \dots - \beta_p z^p$ are all greater than one in absolute value: $|z_j| > 1$ for $j = 1, \dots, p$.*

8. How to determine the order p of an AR(p) process

8.1 The partial autocorrelation function.

If the correct order of an AR process is p_0 but you estimate the AR(p) model (20) with $p > p_0$ by OLS, then the OLS estimates of the coefficients $\beta_{p_0+1}, \dots, \beta_p$ will be small and insignificant, because these coefficients are then all zero: $\beta_{p_0+1} = \beta_{p_0+2} = \dots = \beta_p = 0$. This suggests the following procedure for selecting p . Estimate the AR(p) model for $p = 1, 2, \dots, \bar{p}$, where $\bar{p} > p_0$.

$$\hat{Y}_t = \hat{\beta}_{p,0} + \hat{\beta}_{p,1} Y_{t-1} + \hat{\beta}_{p,2} Y_{t-2} + \dots + \hat{\beta}_{p,p} Y_{t-p}, \tag{35}$$

where the $\hat{\beta}_{p,j}$'s are OLS estimates. Then the (estimated) partial autocorrelation function, PAC(p), is defined by

$$PAC(p) \stackrel{def.}{=} \hat{\beta}_{p,p}, p = 1, 2, 3, \dots, PAC(0) = 1. \quad (36)$$

For example, suppose that an AR(p) model $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + U_t$ has been fitted for $p = 1, 2, 3, 4, 5$ to 500 observation of a time series Y_t , with the following estimation results:

p	β_0	β_1	β_2	β_3	β_4	β_5
1	0.04070	0.02760				
	(0.06481)	(0.04470)				
2	0.06902	0.05358	-0.71257			
	(0.04609)	(0.03189)	(0.03221)			
3	0.06068	0.10470	-0.72159	0.07156		
	(0.04607)	(0.04481)	(0.03231)	(0.04525)		
4	0.06264	0.10524	-0.76661	0.07215	-0.06511	
	(0.04612)	(0.04500)	(0.04548)	(0.04577)	(0.04534)	
5	0.06283	0.10032	-0.76805	0.04648	-0.06783	-0.03274
	(0.04624)	(0.04516)	(0.04575)	(0.05715)	(0.04592)	(0.04544)

The entries that are not enclosed in brackets are the OLS estimates of the AR parameters, and the entries in brackets are the standard errors of the corresponding OLS estimates. Then

p	$PAC(p)$	(s.e.)
1	0.02760	(0.04470)
2	-0.71257	(0.03221)
3	0.07156	(0.04525)
4	-0.06511	(0.04534)
5	-0.03274	(0.04544)

In EasyReg the PAC function can be computed automatically, via Menu > Data analysis > Auto/Cross correlation, and the results will then be displayed as a plot. For example, the $PAC(p)$ for the AR(2) model

$$Y_t = 1.144123Y_{t-1} - 0.5Y_{t-2} + U_t, U_t \text{ i.i.d. } N(0,1), t = 1, \dots, 500, \quad (37)$$

is displayed in Figure 1 below. The dots are the lower and upper bound of the one and two times the standard error bands, which correspond to the 68% and 95% confidence intervals of $\hat{\beta}_{p,p}$,

respectively. The value $\text{PAC}(0) = 1$ is arbitrary, and is chosen because $\text{PAC}(p) < 1$ for $p \geq 1$.

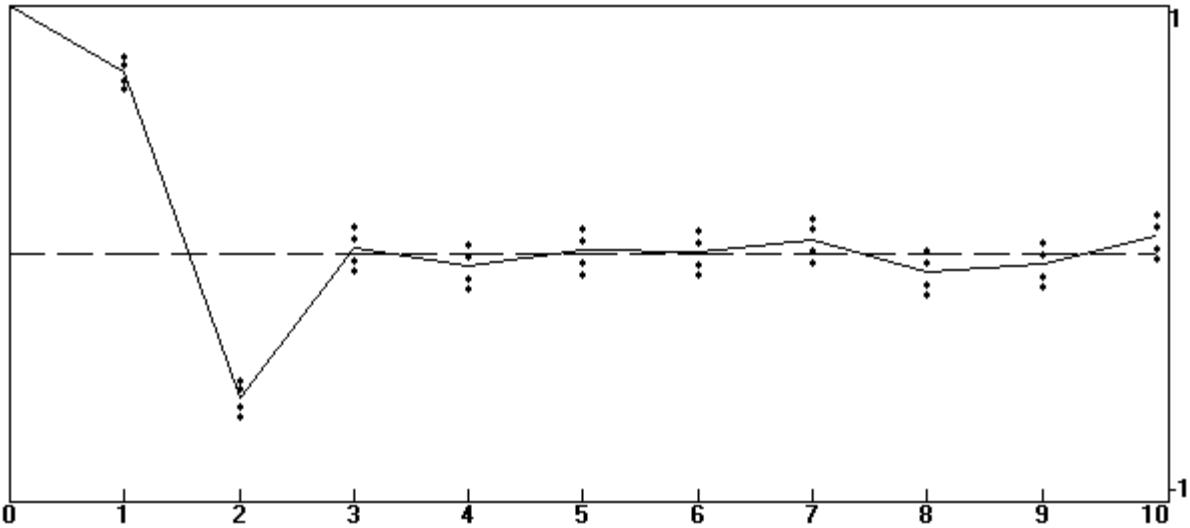


Figure 1: Partial autocorrelation function, $\text{PAC}(m)$, of the AR(2) process (37)

In Figure 1, at $p = 3$, the zero level is contained in the smaller 68% confidence interval, and at $p = 4$ the zero level is contained in the larger 95% confidence interval. From $p = 3$ onwards the zero level is contained in either the 68% and/or 95% confidence intervals, which indicates that the true value of p is $p_0 = 2$.

8.2 Information criteria

An alternative approach to determine the order p of the AR(p) model (20) is to use the Akaike (1974, 1976), Hannan-Quinn (1979), or Schwarz (1978) information criteria:

$$\begin{aligned} \text{Akaike: } c_n^{AR}(p) &= \ln(\hat{\sigma}_p^2) + 2(1+p)/n, \\ \text{Hannan-Quinn: } c_n^{AR}(p) &= \ln(\hat{\sigma}_p^2) + 2(1+p)\ln(\ln(n))/n, \\ \text{Schwarz:}^3 \quad c_n^{AR}(p) &= \ln(\hat{\sigma}_p^2) + (1+p)\ln(n)/n, \end{aligned}$$

³ The Schwarz information criterion is also known as the Bayesian Information Criterion (BIC).

where n is the effective sample size of the regression (20) (so that Y_t is observed for $t = 1-p, \dots, n$), and $\hat{\sigma}_p^2$ is the OLS estimator of the error variance $\sigma^2 = E[U_t^2]$. Denoting by \hat{p} the value of p for which $c_n^{AR}(p)$ is minimal:

$$c_n^{AR}(\hat{p}) = \min\{c_n^{AR}(1), \dots, c_n^{AR}(\bar{p})\},$$

where $\bar{p} > p_0$, with p_0 the true value of p , we have in the Hannan-Quinn and Schwarz cases:

$\lim_{n \rightarrow \infty} P[\hat{p} = p_0] = 1$, and in the Akaike case $\lim_{n \rightarrow \infty} P[\hat{p} \geq p_0] = 1$ but $\lim_{n \rightarrow \infty} P[\hat{p} = p_0] < 1$.

These results are explained in the Appendix. Thus, the Akaike criterion may “overshoot” the true value. Due to the latter, it is recommended to use the Hannan-Quinn or Schwarz criterion instead of the Akaike criterion. Note however that in small samples the Hannan-Quinn and Schwarz criteria may give different results for \hat{p} .

For example, for the same data on which Figure 1 was based, namely the AR(2) model $Y_t = 1.144123Y_{t-1} - 0.5Y_{t-2} + U_t$, $t = 1, \dots, 500$, with independent $N(0,1)$ distributed errors U_t , and upper bound $\bar{p} = 4$, we get

p	Akaike	Hannan-Quinn	Schwarz
1	5.14474E-01	5.21089E-01	5.31332E-01
2	1.08788E-01	1.18711E-01	1.34076E-01
3	1.12462E-01	1.25692E-01	1.46179E-01
4	1.13783E-01	1.30321E-01	1.55929E-01

All three criteria are minimal for $p = 2$, hence $\hat{p} = 2$, which is equal to the true value $p_0 = 2$.

8.3 The Wald test

A third way to determine the correct order p_0 of the AR(p) model (20) is the following. Determine an upper bound $\bar{p} > p_0$ on the basis of the PAC function and the information criteria, estimate the model (20) for $p = \bar{p}$ and test whether p can be reduced, using the Wald test, via Options > Wald test of linear parameter restrictions > Test joint significance, in the “What to do next?” module of EasyReg. For example, for the same data as in the previous section, and $\bar{p} = 4$, we get the OLS results

<i>Parameters</i>	<i>OLS estimate</i>	<i>t-value</i>
β_0	0.06910	1.449
β_1	1.17283	26.033
β_2	-0.63395	-9.134
β_3	0.07841	1.130
β_4	-0.05090	-1.130

The t-value of β_4 is well within the range -1.96, +1.96, hence the null hypothesis that $\beta_4 = 0$ cannot be rejected at the 5% significance level. To test whether $\beta_3 = 0$ as well, you need to test the joint null hypothesis $\beta_3 = \beta_4 = 0$, using the Wald test. In this case the test result involved is:

```

Wald test:                      1.45
Asymptotic null distribution: Chi-square(2)
p-value = 0.48398
Significance levels:          10%      5%
Critical values:              4.61      5.99
Conclusions:                  accept    accept

```

Thus, the null hypothesis $\beta_3 = \beta_4 = 0$ cannot be rejected, hence we may reduce p to 2.

Since β_2 is strongly significant, there is no need to test the null hypothesis $\beta_2 = \beta_3 = \beta_4 = 0$, but if we do the null hypothesis will be rejected:

```

Wald test:                      253.39
Asymptotic null distribution: Chi-square(3)
p-value = 0.00000
Significance levels:          10%      5%
Critical values:              6.25      7.81
Conclusions:                  reject   reject

```

Thus, the test results involved lead to the same conclusion as the one on the basis of the PAC function and the information criteria, namely that $p_0 = 2$.

9. Moving average processes

A moving average process of order q , denoted by $\text{MA}(q)$, takes the form

$$Y_t = \mu + U_t - \theta_1 U_{t-1} - \dots - \theta_q U_{t-q}, \quad (38)$$

where $\mu = E[Y_t]$. Under regularity conditions an MA process has an infinite order AR representation, as I will demonstrate for the case $q = 1$.

Consider the $\text{MA}(1)$ process

$$Y_t = \mu + U_t - \theta U_{t-1}. \quad (39)$$

Using the lag operator, we can write this $\text{MA}(1)$ model as

$$Y_t = \mu + (1 - \theta L)U_t. \quad (40)$$

Now it follows from Proposition 4 and (40) that if $|\theta| < 1$ then

$$\sum_{j=0}^{\infty} \theta^j L^j Y_t = (1 - \theta L)^{-1} Y_t = (1 - \theta L)^{-1} \mu + U_t = \frac{\mu}{1-\theta} + U_t, \quad (41)$$

hence

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + U_t = \beta_0 + \sum_{j=1}^{\infty} \beta_j Y_{t-j} + U_t, \quad (42)$$

where $\beta_0 = \mu/(1-\theta)$, $\beta_j = -\theta^j$ for $j = 1, 2, 3, \dots$

More generally we have:

Proposition 6. If the solutions z_1, \dots, z_q of the equation $0 = 1 - \theta_1 z - \theta_2 z^2 - \dots - \theta_q z^q$ are all greater than one in absolute value: $|z_j| > 1$ for $j = 1, \dots, q$, then the $\text{MA}(q)$ process (38) can be written as an infinite order AR process: $Y_t = \beta_0 + \sum_{j=1}^{\infty} \beta_j Y_{t-j} + U_t$, where $\beta_0 = \mu/(1 - \theta_1 - \theta_2 - \dots - \theta_q)$ and $1 - \sum_{j=1}^{\infty} \beta_j L^j = 1/(1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_q L^q)$.

This result is the main reasons for working with MA models, because the best one-step ahead linear forecast of Y_t takes the form $\beta_0 + \sum_{j=1}^{\infty} \beta_j Y_{t-j}$, which can be approximated quite well by an $\text{MA}(q)$ model.

10. *How to determine the order q of a MA(q) process*

10.1 *The (regular) autocorrelation function*

The autocorrelation function of a time series process Y_t is defined by

$$\rho(m) = \frac{\text{cov}(Y_t, Y_{t-m})}{\text{var}(Y_t)}, m = 0, 1, 2, \dots \dots \quad (43)$$

Proposition 7. *For an MA(q) process, $\rho(m) = 0$ for $m > q$, and $\rho(q) \neq 0$.*

To see this, consider first the MA(1) process (39). For this process,

$$\begin{aligned} \text{cov}(Y_t, Y_{t-1}) &\stackrel{\text{def.}}{=} E[(Y_t - \mu)(Y_{t-1} - \mu)] = E[(U_t - \theta U_{t-1})(U_{t-1} - \theta U_{t-2})] \\ &= E[U_t U_{t-1}] - \theta \cdot E[U_{t-1}^2] - \theta \cdot E[U_t U_{t-2}] + \theta^2 E[U_{t-1} U_{t-2}] = -\theta \cdot E[U_{t-1}^2], \end{aligned} \quad (44)$$

where the last equality follows from the conditions that the errors U_t are uncorrelated and have zero expectation: $E[U_t U_{t-k}] = E[U_t]E[U_{t-k}]$ for $k \neq 0$. Similarly we have for $m > 1$,

$$\begin{aligned} \text{cov}(Y_t, Y_{t-m}) &\stackrel{\text{def.}}{=} E[(Y_t - \mu)(Y_{t-m} - \mu)] = E[(U_t - \theta U_{t-1})(U_{t-m} - \theta U_{t-m-1})] \\ &= E[U_t U_{t-m}] - \theta \cdot E[U_{t-1} U_{t-m}] - \theta \cdot E[U_t U_{t-m-1}] + \theta^2 E[U_{t-1} U_{t-m-1}] = 0 \end{aligned} \quad (45)$$

and

$$\begin{aligned} \text{var}(Y_t) &\stackrel{\text{def.}}{=} E[(Y_t - \mu)^2] = E[(U_t - \theta U_{t-1})^2] \\ &= E[U_t^2] - 2\theta \cdot E[U_t U_{t-1}] + \theta^2 E[U_{t-1}^2] = E[U_t^2] + \theta^2 E[U_{t-1}^2] = (1 + \theta^2)E[U_t^2], \end{aligned} \quad (46)$$

where the last equality follows from the stationarity of U_t . Thus in the MA(1) case,

$$\rho(1) = -\frac{\theta}{1 + \theta^2} \neq 0, \rho(m) = 0 \text{ for } m = 2, 3, \dots \dots \quad (47)$$

Along similar lines it can be shown that Proposition 7 holds.

The actual autocorrelation function cannot be calculated, but it can be estimated in various ways. EasyReg estimates $\rho(m)$ by

$$\hat{\rho}(m) = \frac{(1/(n-m))\sum_{t=m+1}^n(Y_t - \bar{Y})(Y_{t-m} - \bar{Y})}{\sqrt{(1/(n-m))\sum_{t=1}^{n-m}(Y_t - \bar{Y})^2}\sqrt{(1/(n-m))\sum_{t=m+1}^n(Y_{t-m} - \bar{Y})^2}}, \quad m = 0, 1, 2, \dots \quad (48)$$

where $\bar{Y} = (1/n)\sum_{t=1}^n Y_t$.

For an AR(p) process the autocorrelation function does not provide information about p . To see this, consider again the AR(1) process (21) satisfying condition (22). Then it follows from (24) that $cov(Y_t, Y_{t-m}) = \gamma(m) = \sigma^2 \beta_1^m / (1 - \beta_1^2)$ and $var(Y_t) = \gamma(0) = \sigma^2 / (1 - \beta_1^2)$, hence in the AR(1) case, $\rho(m) = \beta_1^m$. Therefore, in this case the autocorrelation function will not drop sharply to zero for $m > 1$. The same applies to more general AR processes.

For example, for the same data on which Figure 1 was based we have:

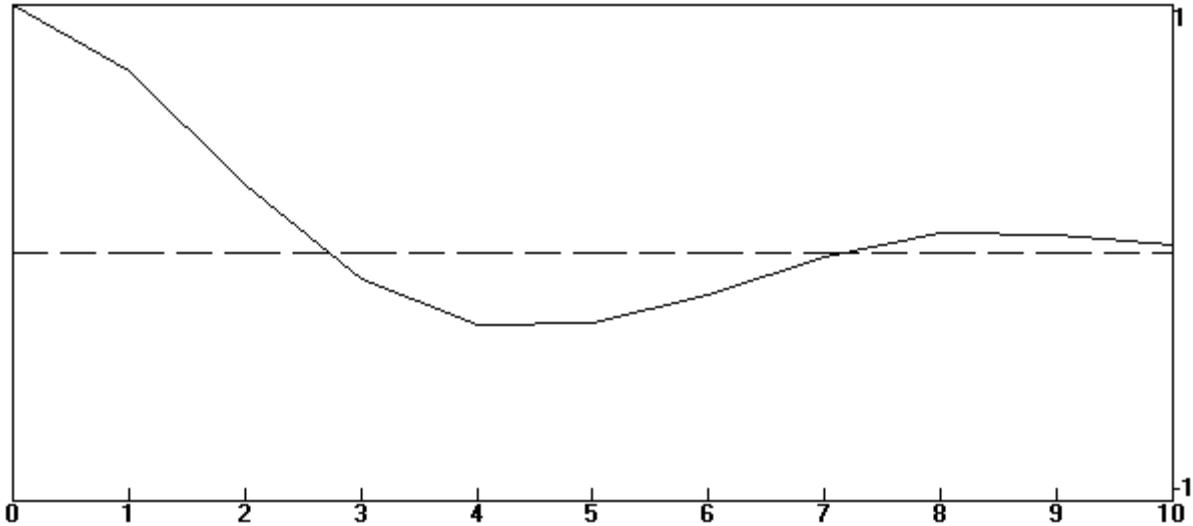


Figure 2: Estimated autocorrelation function $\hat{\rho}(m)$ of the AR(2) process (37).

To demonstrate how to use the estimated autocorrelation function to determine the order q of an MA(q) process, I have generated 500 observations according to the model

$$Y_t = U_t - 1.4U_{t-1} + 0.5U_{t-2}, \quad U_t \sim i.i.d. N(0,1), \quad t = 1, 2, \dots, 500 \quad (49)$$

The estimated autocorrelation function $\hat{\rho}(m)$ involved is displayed in Figure 3 below, for $m = 0, 1, \dots, 10$.

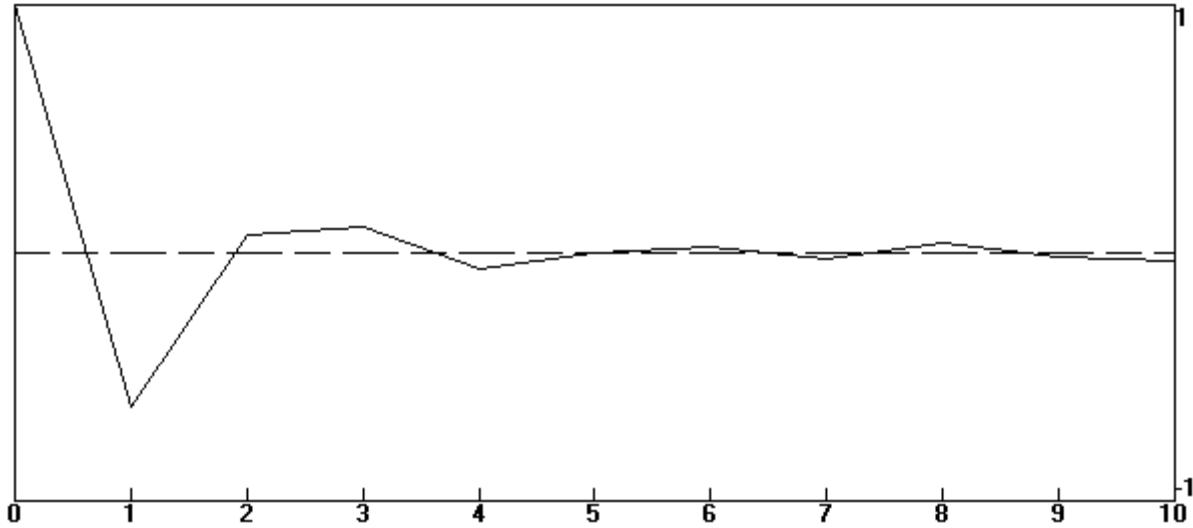


Figure 3: Estimated autocorrelation function $\hat{\rho}(m)$ of the MA(2) process (49)

Because $\hat{\rho}(m)$ is not endowed with standard error bands, it is not obvious at which value of m the true autocorrelation function $\rho(m)$ becomes zero. But at least we can determine an upper bound \bar{q} of q from Figure 3: It seems that $\hat{\rho}(m)$ is approximately zero for $m \geq 5$, indicating that $q \leq 4$. Thus, let $\bar{q} = 4$.

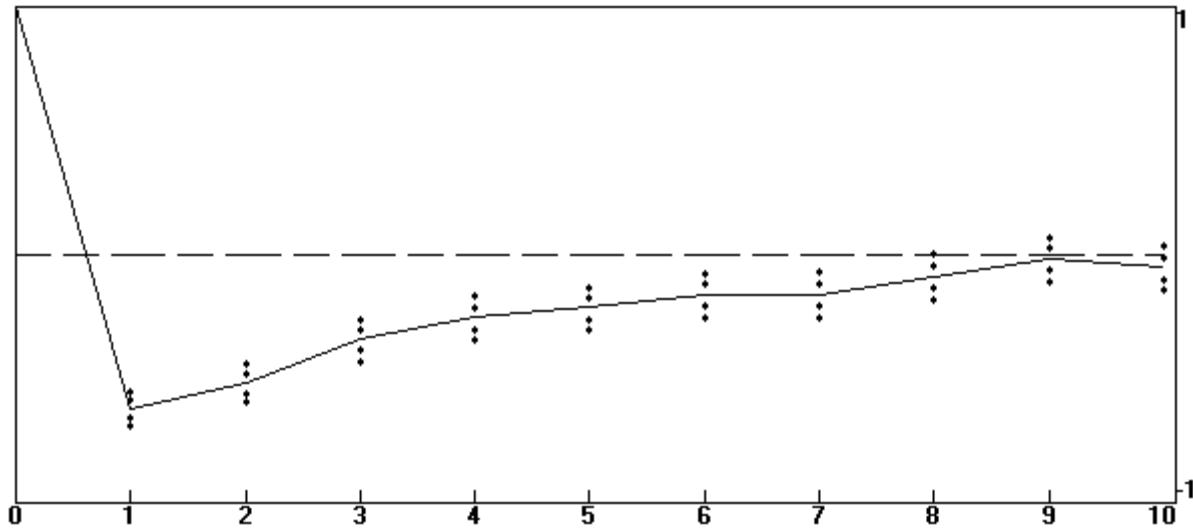


Figure 4: Partial autocorrelation function, $\text{PAC}(m)$, of the MA(2) process (49)

The partial autocorrelation function of an MA(q) process is of no use for determining q or an upper bound of q , because of the AR(∞) representation of an MA(q) process. See Proposition 6. For example, the PAC(m) of the MA(2) process (49) does not drop sharply to zero for $m > 2$, as is demonstrated in Figure 4.

10.2 Information criteria

The three information criteria, Akaike, Hannan-Quinn and Schwarz also apply to MA processes. Therefore, estimate the MA(q) model (38) for $q = 1,2,3,4 (= \bar{q})$, and compare the information criteria:

q	Akaike	Hannan-Quinn	Schwarz
1	$1.91941E-01$	$1.98556E-01$	$2.08799E-01$
2	$1.24771E-02$	$2.23999E-02$	$3.77647E-02$
3	$1.63628E-02$	$2.95933E-02$	$5.00797E-02$
4	$1.68145E-02$	$3.33526E-02$	$5.89606E-02$

All three criteria are minimal for $q = 2$, which is the true value.

10.3 Wald test

As a double check, estimate the MA model (38) for $q = 4$ (in EasyReg via Menu > Single equation models > ARIMA estimation and forecasting), and test whether $\theta_3 = \theta_4 = 0$, using the Wald test:

	Parameters	Estimate	t-value
	μ	0.000234	0.050
	θ_1	1.348470	29.979
	θ_2	-0.427742	-5.637
	θ_3	-0.074225	-0.978
	θ_4	0.050943	1.127

Wald test:		1.29
Asymptotic null distribution:	Chi-square(2)	
p-value = 0.52588		
Significance levels:	10%	5%
Critical values:	4.61	5.99
Conclusions:	accept	accept

Thus, the null hypothesis $\theta_3 = \theta_4 = 0$ cannot be rejected, hence we may reduce q from 4 to $q = 2$. Since θ_2 is strongly significant, we cannot reduce q further.

Re-estimating model (38) for $q = 2$ yields:

Parameters	Estimate	t-value
μ	0.000187	0.039
θ_1	1.348684	33.682
θ_2	-0.456211	-11.349

which is reasonably close to the true values of the parameters: $\mu = 0$, $\theta_1 = 1.4$, $\theta_2 = -0.5$.

11. ARMA models

11.1 Introduction

As I have shown before, both AR(p) models and MA(q) models are parsimonious⁴ approximations of more general AR(∞) processes, $Y_t = \beta_0 + \sum_{j=1}^{\infty} \beta_j Y_{t-j} + U_t$, and the best one-step ahead linear forecast of Y_t given all past values of Y_t takes the form $\hat{Y}_t = \beta_0 + \sum_{j=1}^{\infty} \beta_j Y_{t-j}$. This suggests that even closer approximations can be achieved by combining the two types of models:

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + V_t, \\ V_t &= U_t - \theta_1 U_{t-1} - \dots - \theta_q U_{t-q}. \end{aligned} \quad (50)$$

This is called an ARMA(p,q) model. Denoting

$$\begin{aligned} \varphi_p(L) &= 1 - \beta_1 L - \beta_2 L^2 - \dots - \beta_p L^p, \\ \psi_q(L) &= 1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_q L^q, \end{aligned} \quad (51)$$

⁴

In the sense that only a limited number of parameters are used.

an ARMA(p,q) model can be written more compactly as

$$\varphi_p(L)Y_t = \beta_0 + \psi_q(L)U_t. \quad (52)$$

Proposition 8. If $\varphi_p(z_1) = 0$ implies $|z_1| > 1$ and $\psi_q(z_2) = 0$ implies $|z_2| > 1$ then the ARMA(p,q) process (52) is stationary, with AR(∞) representation $\psi_q(L)^{-1}\varphi_p(L)Y_t = \beta_0/\psi_q(1) + U_t$ and MA(∞) representation $Y_t = \beta_0/\varphi_p(1) + \varphi_p(L)^{-1}\psi_q(L)U_t$.

The last result implies that $E[Y_t] = \beta_0/\varphi_p(1) + \varphi_p(L)^{-1}\psi_q(L)E[U_t] = \beta_0/\varphi_p(1)$.

I will demonstrate Proposition 8 for the case $p = q = 1$:

$$(1 - \beta_1 L)Y_t = \beta_0 + (1 - \theta_1 L)U_t. \quad (53)$$

The condition that $\varphi_1(z) = 0$ implies $|z| > 1$ is equivalent to $|\beta_1| < 1$, because $\varphi_1(z) = 1 - \beta_1 z = 0$ implies that $z = 1/\beta_1$. Similarly, the condition that $\psi_1(z) = 1 - \theta_1 z = 0$ implies $|z| > 1$ is equivalent to $|\theta_1| < 1$. It follows now from Proposition 4 that $\psi_1(L)^{-1} = (1 - \theta_1 L)^{-1} = \sum_{j=0}^{\infty} \theta_1^j L^j$, hence

$$\begin{aligned} \sum_{j=0}^{\infty} \theta_1^j L^j (1 - \beta_1 L)Y_t &= \psi_1(L)^{-1}(1 - \beta_1 L)Y_t = \sum_{j=0}^{\infty} \theta_1^j L^j \beta_0 + \sum_{j=0}^{\infty} \theta_1^j L^j (1 - \theta_1 L)U_t \\ &= \sum_{j=0}^{\infty} \theta_1^j \beta_0 + U_t = \beta_0/(1 - \theta_1) + U_t \end{aligned} \quad (54)$$

and

$$\begin{aligned} \sum_{j=0}^{\infty} \theta_1^j L^j (1 - \beta_1 L)Y_t &= \sum_{j=0}^{\infty} \theta_1^j L^j Y_t - \beta_1 \sum_{j=0}^{\infty} \theta_1^j L^{j+1} Y_t \\ &= Y_t + \sum_{j=1}^{\infty} \theta_1^j Y_{t-j} - \beta_1 \sum_{j=0}^{\infty} \theta_1^j Y_{t-j-1} = Y_t + \theta_1 \sum_{j=0}^{\infty} \theta_1^j Y_{t-1-j} - \beta_1 \sum_{j=0}^{\infty} \theta_1^j Y_{t-1-j} \\ &= Y_t - (\beta_1 - \theta_1) \sum_{j=0}^{\infty} \theta_1^j Y_{t-1-j}. \end{aligned} \quad (55)$$

Combining these results yields

$$Y_t = \beta_0/(1 - \theta_1) + (\beta_1 - \theta_1) \sum_{j=0}^{\infty} \theta_1^j Y_{t-1-j} + U_t, \quad (56)$$

which is the AR(∞) representation of the ARMA(1,1) process under review.

Similarly, it follows from Proposition 4 that $\varphi_1(L)^{-1} = (1 - \beta_1 L)^{-1} = \sum_{j=0}^{\infty} \beta_1^j L^j$, hence

$$\begin{aligned}
Y_t &= (1-\beta_1 L)^{-1}(1-\beta_1 L)Y_t = (1-\beta_1 L)^{-1}\beta_0 + (1-\beta_1 L)^{-1}(1-\theta_1 L)U_t \\
&= \beta_0/(1-\beta_1) + \sum_{j=0}^{\infty} \beta_1^j L^j (1-\theta_1 L)U_t \\
&= \beta_0/(1-\beta_1) + \sum_{j=0}^{\infty} \beta_1^j L^j U_t - \theta_1 \sum_{j=0}^{\infty} \beta_1^j L^{j+1} U_t \\
&= \beta_0/(1-\beta_1) + U_t - (\theta_1 - \beta_1) \sum_{j=0}^{\infty} \beta_1^j U_{t-1-j},
\end{aligned} \tag{57}$$

which is the MA(∞) representation of the ARMA(1,1) process under review.

11.2 Common roots

Observe from (56) and (57) that if $\beta_1 = \theta_1$ then $Y_t = \beta_0/(1-\beta_1) + U_t$, which is an ARMA(0,0) process (also called a *white noise* process). This is the **common roots** problem:

Proposition 9. *Let the conditions in Proposition 8 be satisfied. If there exists a $\delta \neq 0$ such that*

$\varphi_p(1/\delta) = \psi_q(1/\delta) = 0$ *then we can write the lag polynomials in ARMA(p,q) model (52) as*
 $\varphi_p(L) = (1-\delta L)\varphi_{p-1}^*(L)$ *and* $\psi_q(L) = (1-\delta L)\psi_{q-1}^*(L)$, *where* $\varphi_{p-1}^*(L)$ *and* $\psi_{q-1}^*(L)$ *are lag polynomials of order p-1 and q-1, respectively, satisfying the conditions in Proposition 8. The ARMA(p,q) process (52) is then equivalent to the ARMA(p-1,q-1) process*
 $\varphi_{p-1}^*(L)Y_t = \beta_0^* + \psi_{q-1}^*(L)U_t$, *where* $\beta_0^* = \varphi_{p-1}^*(1)E[Y_t]$.

Because the value of δ does not matter, the parameters in the lag polynomials $\varphi_p(L)$ and $\psi_q(L)$ are no longer identified. The same applies to the constant β_0 in model (52) because $\beta_0 = (1-\delta)\beta_0^*$ for arbitrary δ . For example, let for $p = q = 2$,

$$\begin{aligned}
\varphi_2(L) &= (1-\delta L)(1-\beta L) = 1-2(\delta+\beta)L + \delta.\beta L^2 = 1-\beta_1 L-\beta_2 L^2 \\
\psi_2(L) &= (1-\delta L)(1-\theta L) = 1-2(\delta+\theta)L + \delta.\theta L^2 = 1-\theta_1 L-\theta_2 L^2
\end{aligned} \tag{58}$$

where $|\beta| < 1$, $|\theta| < 1$, and $|\delta| < 1$, and let $E[Y_t] = 0$. Then the ARMA(2,2) model $\varphi_2(L)Y_t = \psi_2(L)U_t$ is equivalent to the ARMA(1,1) model $(1-\beta L)Y_t = (1-\theta L)U_t$ for all values of δ . Hence, given β and θ , $\beta_1 = 2(\delta+\beta)$, $\beta_2 = -\delta.\beta$, $\theta_1 = 2(\delta+\theta)$, $\theta_2 = -\delta.\theta$ for arbitrary δ . As a consequence, the estimates of the parameters β_1 , β_2 , θ_1 , θ_2 are no longer consistent, and the t-test and Wald test for testing the (joint) significance of the parameters are no longer valid. In particular, in the ARMA(2,2) case under review the Wald test of the null hypothesis $\beta_2 = \theta_2 = 0$

is no longer valid. Therefore, we should not use the Wald test to test whether the AR and MA orders p and q can be reduced to $p-1$ and $q-1$.

The problem of common roots in ARMA models is similar to the multicollinearity problem in linear regression. As in the latter case, the t values of the parameters will be deflated towards zero. Therefore, if all the t values of the ARMA parameters are insignificant this may indicate that the AR and MA lag polynomials have a common root.

Although we should not use the Wald test to test for common roots, we can still use the information criteria to determine whether the AR and MA orders p and q can be reduced to $p-1$ and $q-1$. In the case of a common root, the variance σ^2 of the errors U_t of the ARMA(p,q) model in Proposition 9 is the same as the variance of the errors U_t in the ARMA($p-1,q-1$) model

$\Phi_{p-1}^*(L)Y_t = \beta_0^* + \psi_{q-1}^*(L)U_t$. Therefore, the estimate $\hat{\sigma}_{p,q}^2$ of the errors U_t of the ARMA(p,q) model involved will be close to the estimate $\hat{\sigma}_{p-1,q-1}^2$ of the errors of the equivalent ARMA($p-1,q-1$) model, and asymptotically they will be equal:

$$\text{plim}_{n \rightarrow \infty} \hat{\sigma}_{p,q}^2 = \text{plim}_{n \rightarrow \infty} \hat{\sigma}_{p-1,q-1}^2 = \sigma^2. \quad (59)$$

In the ARMA case the three information criteria take the form

$$\begin{aligned} \text{Akaike: } c_n^{ARMA}(p,q) &= \ln(\hat{\sigma}_{p,q}^2) + 2(1+p+q)/n, \\ \text{Hannan-Quinn: } c_n^{ARMA}(p,q) &= \ln(\hat{\sigma}_{p,q}^2) + 2(1+p+q)\ln(\ln(n))/n, \\ \text{Schwarz: } c_n^{ARMA}(p,q) &= \ln(\hat{\sigma}_{p,q}^2) + (1+p+q)\ln(n)/n, \end{aligned}$$

Therefore, in the case of a common root, $c_n^{ARMA}(p-1,q-1) < c_n^{ARMA}(p,q)$ if n is large enough, due to (59).

To demonstrate the common roots phenomenon, I have generated a time series Y_t , $t = 1, \dots, 500$, according to the ARMA(1,1) model

$$Y_t = 0.3 + 0.7Y_{t-1} + U_t + 0.5U_{t-1}, \quad U_t \sim i.i.d N(0,1), \quad (60)$$

and estimated this model as an ARMA(2,2) model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + U_t - \theta_1 U_{t-1} - \theta_2 U_{t-2}. \quad (61)$$

The EasyReg estimation results involved are:

<i>Parameters</i>	<i>Estimate</i>	<i>t-value</i>
$\mu = \beta_0/(1-\beta_1-\beta_2)$	0.776272	3.273
β_1	1.087430	0.170
β_2	-0.267512	-0.058
θ_1	-0.166275	-0.026
θ_2	0.189439	0.055
σ	1.008897	

Information criteria:

Akaike:	2.76651E-02
Hannan-Quinn:	4.42031E-02
Schwarz:	6.98112E-02

Apart from the estimate of $\mu = E[Y_t]$, the AR and MA parameters are insignificant, due to a common root in the AR and MA lag polynomials.

Next, I have estimated the model as an ARMA(1,1) model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + U_t - \theta_1 U_{t-1}. \quad (62)$$

The EasyReg estimation results are:

<i>Parameters</i>	<i>Estimate</i>	<i>t-value</i>
$\mu = \beta_0/(1-\beta_1)$	0.775967	3.249
β_1	0.720705	20.784
θ_1	-0.530183	-12.597
σ	1.006879	

Information criteria:

Akaike:	1.96937E-02
Hannan-Quinn:	2.96166E-02
Schwarz:	4.49814E-02

Indeed, the information criteria for the latter model are substantial lower (and thus better) than for the previous ARMA(2,2) model. Moreover, observe that in the latter case the estimates of β_1 , θ_1 and σ are close to the true values $\beta_1 = 0.7$, $\theta_1 = -0.5$ and $\sigma = 1$, respectively,

although at first sight the estimate $\hat{\mu} = 0.775967$ of $\mu = E[Y_t]$ seems quite different from the true value $\mu = 0.3/(1-0.7) = 1$. However, it can be shown that $\hat{\mu}$ is not significantly different from 1.

11.3 How to distinguish an ARMA process from an AR process

The AR(∞) representation (56) of the ARMA(1,1) process (60) is

$$\begin{aligned}
 Y_t &= \beta_0/(1-\theta_1) + (\beta_1 - \theta_1)\sum_{j=0}^{\infty}\theta_1^jY_{t-1-j} + U_t \\
 &= 0.3/(1+0.5) + 1.2\sum_{j=0}^{\infty}(-0.5)^jY_{t-1-j} + U_t \\
 &= 0.2 + 1.2\sum_{j=0}^{\infty}(-0.5)^jY_{t-1-j} + U_t \\
 &= 0.2 + 1.2Y_{t-1} - 0.6Y_{t-2} + 0.3Y_{t-3} - 0.15Y_{t-4} + 0.075Y_{t-5} + \dots + U_t
 \end{aligned} \tag{63}$$

which is close to an AR(4) process. Therefore, the partial autocorrelation function, PAC(m), of this process will look like the PAC(m) of an AR process:

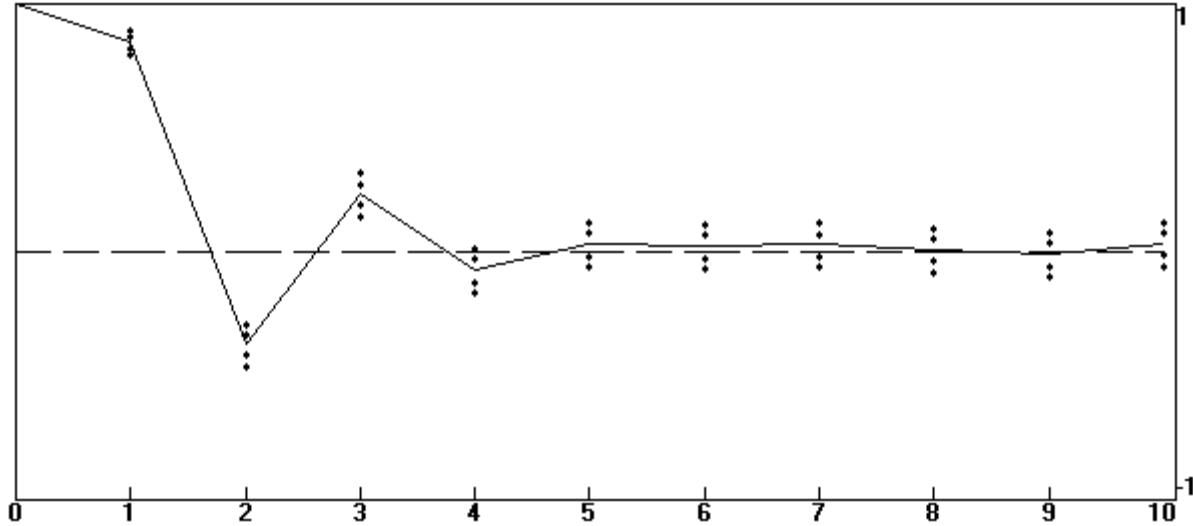


Figure 5: Partial autocorrelation function, PAC(m), of the ARMA(1,1) process (60)

Indeed, on the basis of this plot one may be tempted to conclude (erroneously) that the process is an AR(4) process, and the estimated autocorrelation function would actually corroborates this:

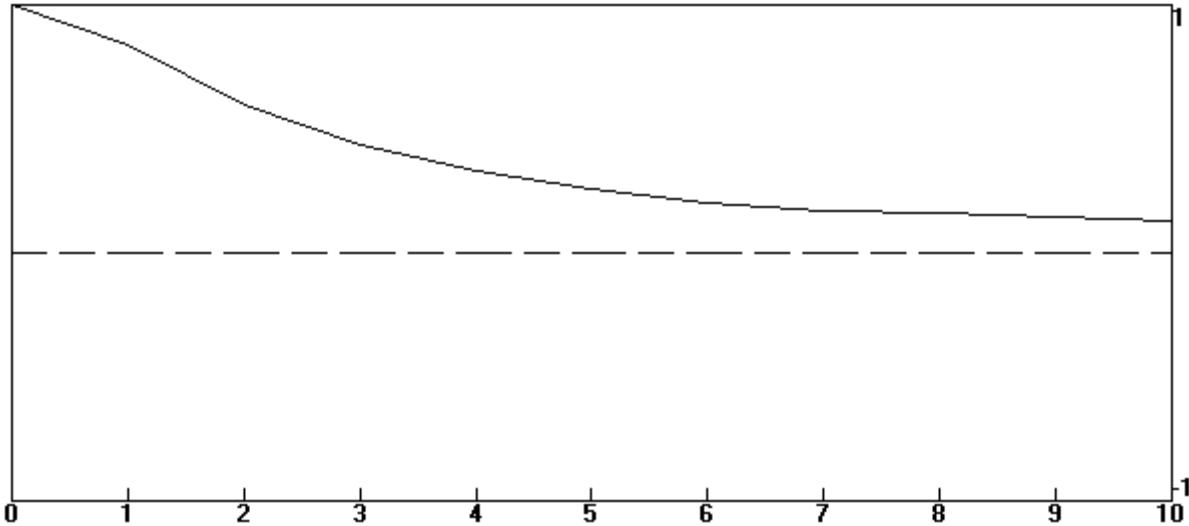


Figure 6: Estimated autocorrelation function $\hat{\rho}(m)$ of the ARMA(1,1) process (60)

Therefore, the partial and regular autocorrelation functions are of no help in distinguishing an ARMA model from an AR model.

So how to proceed? In this particular case I recommend the following: In first instance assume, on the basis of Figure 5, that the model is an AR(4) model, and estimate it by OLS or the ARIMA module. Then try all ARMA(p,q) models with $p + q \leq 4$, and pick the model with the lowest value of one of the information criteria. For example, the Hannan-Quinn information criteria for the model (60) are:

p	$q=0$	$q=1$	$q=2$	$q=3$	$q=4$
4	$3.99778E-02$				
3	$3.79859E-02$	$4.15536E-02$			
2	$8.56998E-02$	$3.61047E-02$	$4.42031E-02$		
1	$22.9013E-02$	$2.96166E-02$	$3.57906E-02$	$3.92521E-02$	
0	$145.019E-02$	$54.6481E-02$	$19.5530E-02$	$10.7008E-02$	$7.73787E-02$

The smallest value of the Hannan-Quinn information criterion is $2.96166E-02$ for $p = q = 1$, hence the conclusion is that the process involved is an ARMA(1,1) process.

This model selection procedure can be conducted automatically in EasyReg, via Menu > Single equation models > ARIMA model selection via information criteria. The only thing you

have to do is to specify an initial, possibly over-parametrized, ARMA model. EasyReg will then compute the Akaike, Hannan-Quinn and Schwarz information criteria for this model and all sub-models, and indicate which model is optimal.

11.4 Forecasting with an ARMA model

In EasyReg the AR(∞) representation of an ARMA model is used as forecasting scheme, because it represents the conditional expectation function. For example, in the ARMA(1,1) case the forecasting scheme for Y_{n+1} given its past up to time n is

$$\hat{Y}_{n+1} = \beta_0/(1-\theta_1) + (\beta_1-\theta_1)\sum_{j=0}^{\infty}\theta_1^j Y_{n-j}, \quad (64)$$

where n is the last observed time period. Compare (56). In practice we have to replace the parameters involved by estimates. Moreover, usually we do not observe all values of Y_{n-j} , but only for $n-j \geq 1$, say. Therefore, replace Y_t for $t < 1$ in (64) by its sample mean $\bar{Y} = (1/n)\sum_{t=1}^n Y_t$. Thus, the actual forecast of Y_{n+1} is:

$$\begin{aligned} \tilde{Y}_{n+1|n} &= \hat{\beta}_0/(1-\hat{\theta}_1) + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=0}^{n-1}\hat{\theta}_1^j Y_{n-j} + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=n}^{\infty}\hat{\theta}_1^j \bar{Y} \\ &= \frac{\hat{\beta}_0}{1-\hat{\theta}_1} + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=0}^{n-1}\hat{\theta}_1^j Y_{n-j} + \frac{(\hat{\beta}_1-\hat{\theta}_1)\hat{\theta}_1^n \bar{Y}}{1-\hat{\theta}_1}, \end{aligned} \quad (65)$$

where $\hat{\beta}_1$ and $\hat{\theta}_1$ are the estimates of β_1 and θ_1 , respectively, based on the data up to time n .

To forecast Y_{n+2} given its past up to time n , replace n in (65) by $n+1$, and the unobserved Y_{n+1} by its forecast:

$$\tilde{Y}_{n+2|n} = \hat{\beta}_0/(1-\hat{\theta}_1) + (\hat{\beta}_1-\hat{\theta}_1)\tilde{Y}_{n+1|n} + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=1}^n\hat{\theta}_1^j Y_{n+1-j} + \frac{(\hat{\beta}_1-\hat{\theta}_1)\hat{\theta}_1^{n+1} \bar{Y}}{1-\hat{\theta}_1}. \quad (66)$$

This procedure is called *recursive forecasting*. More generally, the h step ahead recursive forecast of Y_{n+h} given its past up to time n is

$$\tilde{Y}_{n+h|n} = \frac{\hat{\beta}_0}{1-\hat{\theta}_1} + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=0}^{h-2}\hat{\theta}_1^j \tilde{Y}_{n+h-1-j|n} + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=h+1}^{n+h-2}\hat{\theta}_1^j Y_{n+h-1-j} + \frac{(\hat{\beta}_1-\hat{\theta}_1)\hat{\theta}_1^{n+h-1} \bar{Y}}{1-\hat{\theta}_1}. \quad (67)$$

Note however, that in this case

$$\begin{aligned}\lim_{h \rightarrow \infty} \tilde{Y}_{n+h|n} &= \hat{\beta}_0 / (1 - \hat{\theta}_1) + (\hat{\beta}_1 - \hat{\theta}_1) \lim_{h \rightarrow \infty} \sum_{j=0}^{h-2} \hat{\theta}_1^j \tilde{Y}_{n+h-1-j|n} \\ &= \hat{\beta}_0 / (1 - \hat{\theta}_1) + (\hat{\beta}_1 - \hat{\theta}_1) \sum_{j=0}^{\infty} \hat{\theta}_1^j \lim_{h \rightarrow \infty} \tilde{Y}_{n+h|n} = \hat{\beta}_0 / (1 - \hat{\theta}_1) + ((\hat{\beta}_1 - \hat{\theta}_1) / (1 - \hat{\theta}_1)) \lim_{h \rightarrow \infty} \tilde{Y}_{n+h|n}.\end{aligned}\quad (68)$$

Solving this equality for $\lim_{h \rightarrow \infty} \tilde{Y}_{n+h|n}$ yields

$$\lim_{h \rightarrow \infty} \tilde{Y}_{n+h|n} = \hat{\beta}_0 / (1 - \hat{\beta}_1), \quad (69)$$

which is just the estimate of $\mu = E[Y_t]$. Compare (57). Thus, if we choose the forecast horizon h too large, the recursive forecast $\tilde{Y}_{n+h|n}$ will be close to the expectation $\mu = E[Y_t]$.

12. ARMA models for seasonal time series

12.1 Seasonal dummy variables

The effect of seasonality may manifest itself through seasonal varying expectations as well as seasonal patterns in the AR and/or MA lag polynomials. As to the former, time varying expectations can easily be modeled using seasonal dummy variables. For example, if Y_t is a quarterly time series, $E[Y_t]$ can be modeled either by

$$E[Y_t] = \mu_0 + \mu_1 Q_{1,t} + \mu_2 Q_{2,t} + \mu_3 Q_{3,t} \quad (70)$$

or

$$E[Y_t] = \mu_1^* Q_{1,t} + \mu_2^* Q_{2,t} + \mu_3^* Q_{3,t} + \mu_4^* Q_{4,t}, \quad (71)$$

where the $Q_{s,t}$'s are seasonal dummy variables:

$$Q_{s,t} = 1 \text{ if the quarter of } t \text{ is } s, Q_{s,t} = 0 \text{ if not.} \quad (72)$$

The equivalence of (70) and (71) follows from the fact that $\sum_{s=1}^4 Q_{s,t} = 1$, so that

$$\begin{aligned}E[Y_t] &= \mu_1^* Q_{1,t} + \mu_2^* Q_{2,t} + \mu_3^* Q_{3,t} + \mu_4^* (1 - Q_{1,t} - Q_{2,t} - Q_{3,t}) \\ &= \mu_4^* + (\mu_1^* - \mu_4^*) Q_{1,t} + (\mu_2^* - \mu_4^*) Q_{2,t} + (\mu_3^* - \mu_4^*) Q_{3,t},\end{aligned}\quad (73)$$

hence $\mu_0 = \mu_4^*$, $\mu_1 = \mu_1^* - \mu_4^*$, $\mu_2 = \mu_2^* - \mu_4^*$, $\mu_3 = \mu_3^* - \mu_4^*$.

Note that if we had defined (71) as $E[Y_t] = \mu_0^* + \mu_1^* Q_{1,t} + \mu_2^* Q_{2,t} + \mu_3^* Q_{3,t} + \mu_4^* Q_{4,t}$, the parameters involved are no longer identified, because then (73) becomes

$$E[Y_t] = (\mu_0^* + \mu_4^*) + (\mu_1^* - \mu_4^*) Q_{1,t} + (\mu_2^* - \mu_4^*) Q_{2,t} + (\mu_3^* - \mu_4^*) Q_{3,t}, \quad (74)$$

which is also equivalent to (70). Hence

$$\mu_0 = \mu_0^* + \mu_4^*, \mu_1 = \mu_1^* - \mu_4^*, \mu_2 = \mu_2^* - \mu_4^*, \mu_3 = \mu_3^* - \mu_4^*, \quad (75)$$

which is a system of four equations in five unknowns.

The presence of seasonally varying expectations can be observed from the autocorrelation function. For example, let Y_t be a quarterly time series satisfying

$$Y_t = \mu_0 + \mu_1 Q_{1,t} + \mu_2 Q_{2,t} + \mu_3 Q_{3,t} + X_t \quad (76)$$

where X_t is zero-mean covariance stationary with covariance function $\gamma_x(m) = E(X_t X_{t-m})$. The sample average of Y_t is

$$\begin{aligned} \bar{Y}_n &= \mu_0 + \mu_1(1/n)\sum_{t=1}^n Q_{1,t} + \mu_2(1/n)\sum_{t=1}^n Q_{2,t} + \mu_3(1/n)\sum_{t=1}^n Q_{3,t} + (1/n)\sum_{t=1}^n X_t \\ &\approx \mu_0 + 0.25\mu_1 + 0.25\mu_2 + 0.25\mu_3 \end{aligned} \quad (77)$$

if n is large, because for each $s = 1, 2, 3$ the fraction of values of $Q_{s,t}$ for $t = 1, \dots, n$ that are equal to 1 tends towards 0.25 if $n \rightarrow \infty$, and $\text{plim}_{n \rightarrow \infty}(1/n)\sum_{t=1}^n X_t = E[X_t] = 0$ by the law of large numbers. Then it can be shown⁵ that there exists constants c_s , $s = 1, 2, 3, 4$, such that for $n \rightarrow \infty$,

$$\frac{1}{n-m} \sum_{t=m+1}^n (Y_t - \bar{Y}_n)(Y_{t-m} - \bar{Y}_n) \rightarrow \begin{cases} \gamma_x(m) + c_1 \text{ for } m = 0, 4, 8, 12, \dots \\ \gamma_x(m) + c_2 \text{ for } m = 1, 5, 9, 13, \dots \\ \gamma_x(m) + c_3 \text{ for } m = 2, 6, 10, 14, \dots \\ \gamma_x(m) + c_4 \text{ for } m = 3, 7, 11, 15, \dots \end{cases} \quad (78)$$

in probability. It follows now from (48) and (78) that the estimated autocorrelation function $\hat{\rho}(m)$ will have spikes at distances of four lags, and will not die out to zero:

$$\hat{\rho}(m) \rightarrow \rho(m), \text{ where } \rho(m) = \begin{cases} (\gamma_x(m) + c_1)/(\gamma_x(0) + c_1) \text{ for } m = 0, 4, 8, 12, \dots \\ (\gamma_x(m) + c_2)/(\gamma_x(0) + c_1) \text{ for } m = 1, 5, 9, 13, \dots \\ (\gamma_x(m) + c_3)/(\gamma_x(0) + c_1) \text{ for } m = 2, 6, 10, 14, \dots \\ (\gamma_x(m) + c_4)/(\gamma_x(0) + c_1) \text{ for } m = 3, 7, 11, 15, \dots \end{cases} \quad (79)$$

in probability as $n \rightarrow \infty$.

For example consider the quarterly process

⁵

But the derivation involved is too tedious and therefore omitted.

$$Y_t = 1 + 2Q_{1,t} - Q_{2,t} - 2Q_{3,t} + X_t, \text{ where } X_t \sim \text{i.i.d. } N(0,1), \quad (80)$$

for $t = 1, 2, \dots, 225$. The estimated autocorrelation function $\hat{\rho}(m)$ of this process is displayed in Figure 7.

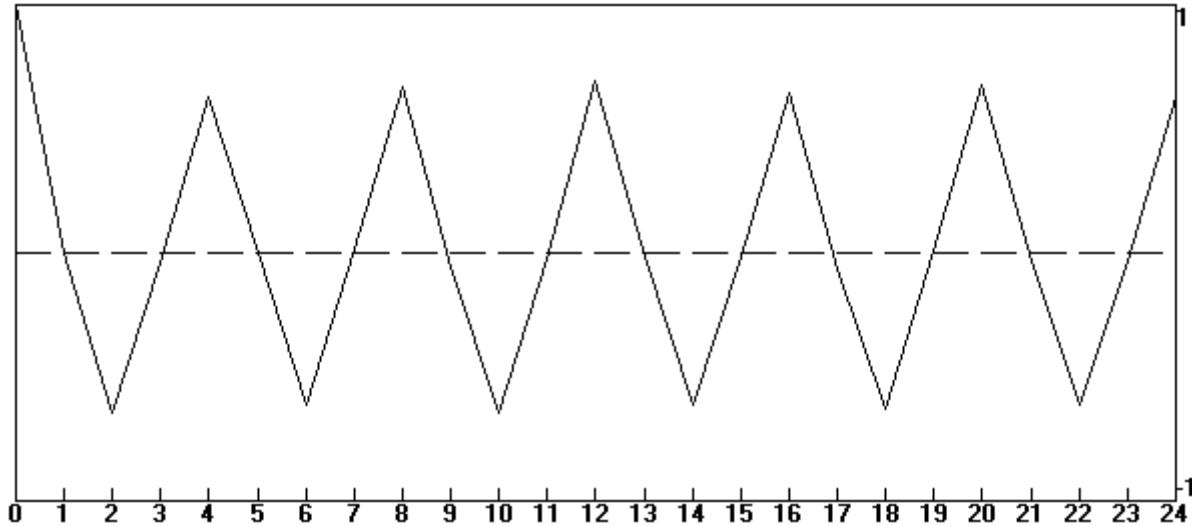


Figure 7: Estimated autocorrelation function $\hat{\rho}(m)$ of quarterly process (80)

12.2 Seasonal lag polynomials

Seasonality may also occur in the process X_t in (76) itself. For example, let X_t be a quarterly ARMA process

$$\varphi_p(L)\lambda_r(L^4)X_t = \beta_0 + \psi_q(L)\eta_s(L^4)U_t, \quad (81)$$

where $\varphi_p(L)$ and $\psi_q(L)$ are the non-seasonal AR and MA lag polynomials of orders p and q , respectively, defined before, and $\lambda_r(z)$ and $\eta_s(z)$ are the seasonal AR and MA polynomials of orders r and s , respectively.

In EasyReg these polynomials are specified via the window displayed in Figure 8 below. The coefficients $a(1,i)$, $i = 1, \dots, p$, are the coefficients of the non-seasonal AR polynomial $\varphi_p(L)$, the coefficients $a(2,i)$, $i = 1, \dots, q$, are the coefficients of the non-seasonal MA polynomial $\psi_q(L)$, the coefficients $c(1,i)$, $i = 1, \dots, r$, are the coefficients of the seasonal AR polynomial $\lambda_r(L^4)$, and the coefficients $c(2,i)$, $i = 1, \dots, s$, are the coefficients of the seasonal MA polynomial $\eta_s(L^4)$. The displayed specification is for $p = q = r = s = 2$.

The specification procedure for p , q , r and s is similar to the non-seasonal ARMA case:

First, specify upper bounds of p , q , r and s , and then use the information criteria to select the correct p , q , r and s , via Menu > Single equation models > ARIMA model selection via information criteria.

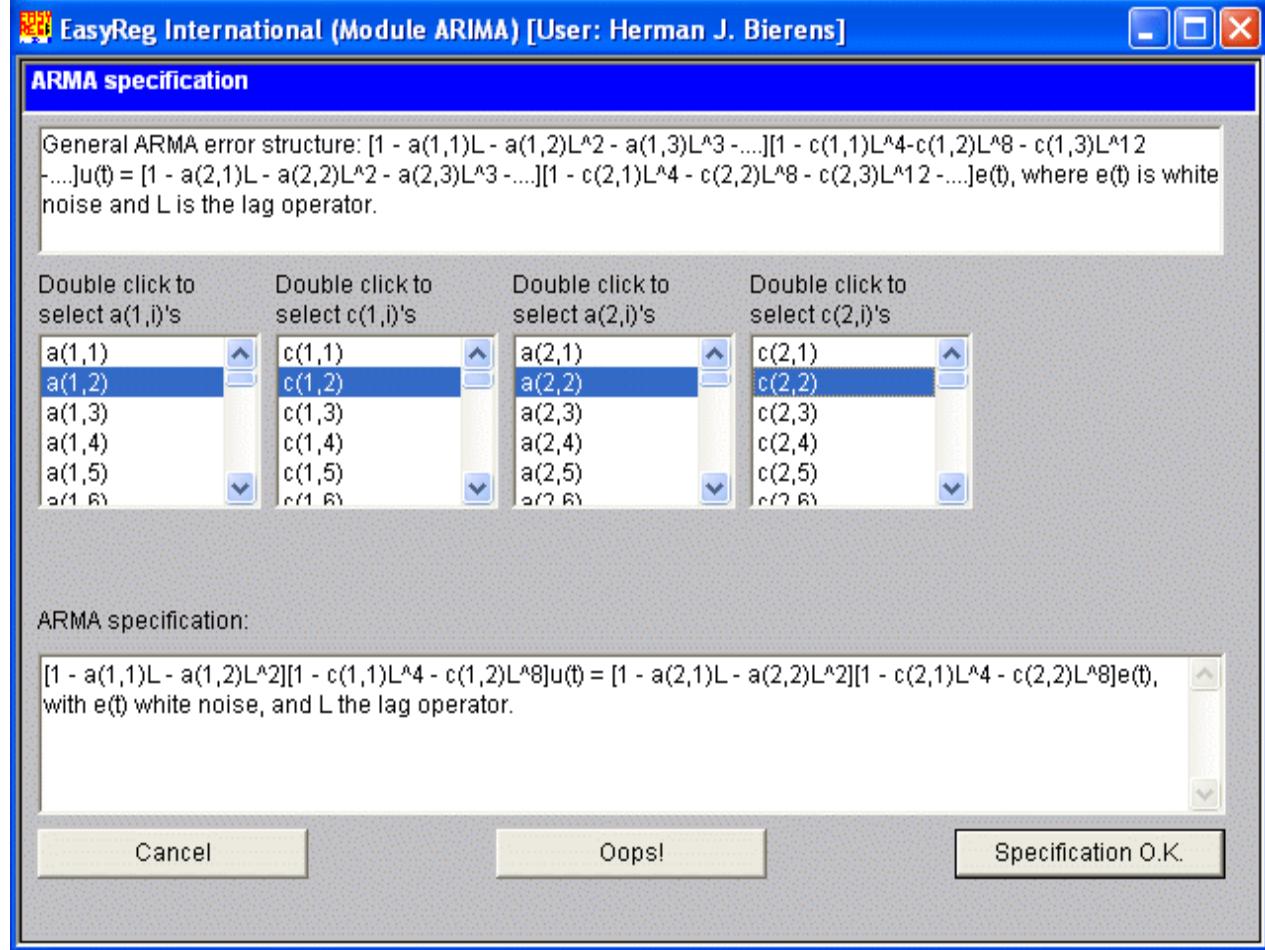


Figure 8: Specification of a seasonal ARMA model in EasyReg.

13. Unit roots

13.1 What is a unit root process?

Consider the AR(p) process (20), written as

$$\varphi_p(L)Y_t = \beta_0 + U_t, \text{ where } \varphi_p(L) = 1 - \beta_1L - \beta_2L^2 - \dots - \beta_pL^p. \quad (82)$$

If $\varphi_p(1) = 0$, this process is called a unit root process. If so, $\varphi_p(L) = (1-L)\varphi_{p-1}^*(L)$, where

$\varphi_{p-1}^*(L)$ is a lag polynomial of order $p-1$. Moreover, the process (82) is then no longer stationary, because the condition in Proposition 5 is no longer satisfied. But if $\varphi_{p-1}^*(L)$ satisfies the condition in Proposition 5 then the first difference of Y_t , $\Delta Y_t = Y_t - Y_{t-1}$, is a stationary AR($p-1$) process: $\varphi_{p-1}^*(L)\Delta Y_t = \beta_0 + U_t$.

For example, consider the case $p = 1$, $\beta_0 = 0$:

$$Y_t = Y_{t-1} + U_t, \text{ where } U_t \sim \text{i.i.d. } N(0, \sigma^2). \quad (83)$$

This process is called a random walk. By t times backwards substitution of (83) we get

$$Y_t = Y_0 + \sum_{j=1}^t U_j. \quad (84)$$

Because $\text{var}(\sum_{j=1}^t U_j) = \sum_{j=1}^t \text{var}(U_j) = \sigma^2 t$, it is clear that Y_t is not stationary.

13.2 The Augmented Dickey-Fuller (ADF) tests

Now the question arises how to distinguish the case (83) from a stationary AR(1) process

$Y_t = \beta Y_{t-1} + U_t$, $|\beta| < 1$. The latter process can be written as

$$\Delta Y_t = Y_t - Y_{t-1} = (\beta - 1)Y_{t-1} + U_t = \alpha Y_{t-1} + U_t, \quad (85)$$

say, where $\alpha = \beta - 1$. Now the non-stationary case (83) corresponds to $\alpha = 0$, and the stationary case ($|\beta| < 1$) corresponds to $-2 < \alpha < 0$. This suggests to estimate the model (85) by OLS, and to use the t value \hat{t}_α of α to test the null hypothesis $\alpha = 0$ against the alternative hypothesis $\alpha < 0$.

The problem, however, is that under the null hypothesis $\alpha = 0$ the test statistic \hat{t}_α has no longer a standard normal distribution, as is demonstrated in Figure 9.

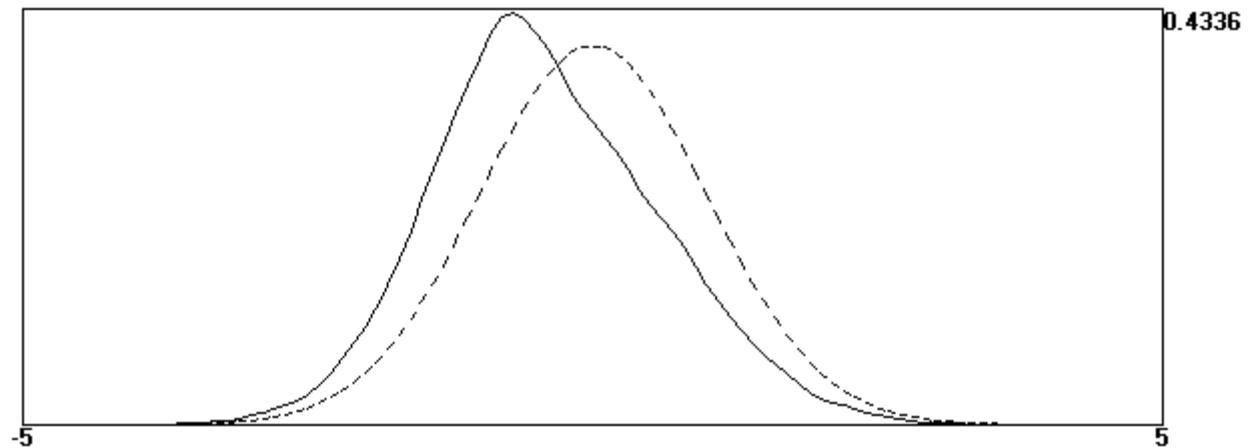


Figure 9: Density of \hat{t}_α under the null hypothesis $\alpha = 0$ (solid curve) compared with the standard normal density (dashed curve)

Therefore, the asymptotic critical values of the left-sided standard normal test are not valid.

Instead we have under the null hypothesis $\alpha = 0$ and large sample size n ,

$$P(\hat{t}_\alpha \leq -1.95) = 0.05, \quad P(\hat{t}_\alpha \leq -1.62) = 0.10. \quad (86)$$

Thus, the null hypothesis $\alpha = 0$ is rejected at the 5% significance level in favor of the alternative hypothesis $\alpha < 0$ if $\hat{t}_\alpha \leq -1.95$, and the null hypothesis $\alpha = 0$ is accepted at the 5% significance level if $\hat{t}_\alpha > -1.95$.

The assumption that the U_t 's in (83) are independent is much too restrictive, though. To relax this assumption, consider an AR(2) process without intercept:

$$(1-\delta L)(1-\beta L)Y_t = U_t, \text{ where } |\beta| < 1, \text{ and either } \delta = 1 \text{ or } |\delta| < 1. \quad (87)$$

The unit root hypothesis corresponds to the case $\delta = 1$, and the stationarity hypothesis corresponds to the case $|\delta| < 1$. Using the easy equality $(1-\delta L)(1-\beta L) = 1 - (\delta + \beta)L + \delta\beta L^2$ this model can be written as

$$\begin{aligned} \Delta Y_t &= Y_t - Y_{t-1} = (\delta + \beta)Y_{t-1} - Y_{t-1} - \delta\beta Y_{t-2} + U_t \\ &= (\delta + \beta)Y_{t-1} - Y_{t-1} - \delta\beta Y_{t-1} + \delta\beta Y_{t-1} - \delta\beta Y_{t-2} + U_t \\ &= (\delta + \beta - 1 - \delta\beta)Y_{t-1} + \delta\beta(Y_{t-1} - Y_{t-2}) + U_t \\ &= \alpha Y_{t-1} + \gamma \Delta Y_{t-1} + U_t, \end{aligned} \quad (88)$$

where

$$\begin{aligned} \alpha &= \delta + \beta - 1 - \delta\beta = -(1-\delta)(1-\beta), \\ \gamma &= \delta\beta. \end{aligned} \quad (89)$$

Now the unit root hypothesis $\delta = 1$ corresponds to $\alpha = 0$, and the stationarity hypothesis $|\delta| < 1$ corresponds to $\alpha < 0$ (and $\alpha > -4$).

More generally we have:

Proposition 10. An AR(p) process Y_t with intercept can always be written as

$$\Delta Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \sum_{j=1}^{p-1} \gamma_j \Delta Y_{t-j} + U_t. \quad (90)$$

If the process Y_t is stationary then $\alpha_1 < 0$, and if Y_t is a unit root process then $\alpha_1 = 0$. If model (90) is estimated without an intercept (thus $\alpha_0 = 0$) then under the unit root hypothesis $\alpha_1 = 0$ the t value \hat{t}_{α_1} of α_1 converges in distribution to a non-normal distribution with density displayed in Figure 9. If model (90) is estimated with an intercept then under the unit root hypothesis

$\alpha_1 = 0$ the t value \hat{t}_{α_1} of α_1 converges in distribution to a non-normal distribution with density displayed in Figure 10 below.

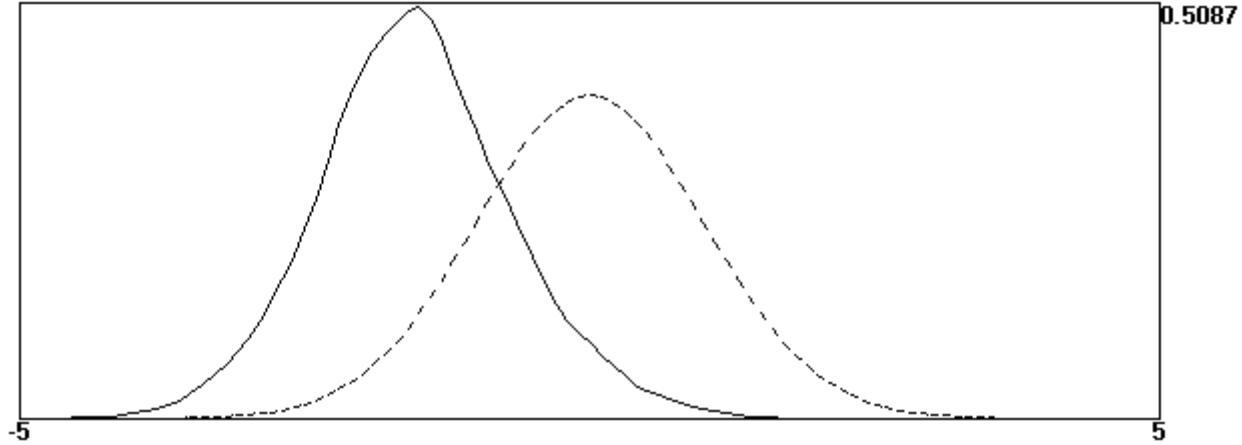


Figure 10: Density of \hat{t}_{α_1} for model (90) with intercept under the null hypothesis $\alpha_1 = 0$ (solid curve) compared with the standard normal density (dashed curve)

Observe that the density of \hat{t}_{α_1} in Figure 10 is shifted even more farther to the left of the standard normal density than in Figure 9, hence the left-sided standard normal test would result in a dramatically higher type 1 error than in the case without an intercept. In particular, we now have that under the null hypothesis $\alpha_1 = 0$ and for sufficiently long time series,

$$P(\hat{t}_{\alpha_1} \leq -2.86) = 0.05, \quad P(\hat{t}_{\alpha_1} \leq -2.57) = 0.1. \quad (91)$$

Thus, the null hypothesis $\alpha_1 = 0$ is rejected at the 5% significance level in favor of the alternative hypothesis $\alpha_1 < 0$ if $\hat{t}_{\alpha_1} \leq -2.86$, and the null hypothesis $\alpha_1 = 0$ is accepted at the 5% significance level if $\hat{t}_{\alpha_1} > -2.86$.

The left-sided test of the unit root (null) hypothesis $\alpha_1 = 0$ against the (alternative) stationarity hypothesis $\alpha_1 < 0$ based on the t value \hat{t}_{α_1} of α_1 in model (90) is known as the Augmented Dickey-Fuller (ADF) test. If model (90) is estimated without an intercept the alternative hypothesis is zero-mean stationarity: $E[Y_t] = 0$. This is ADF test 1 in EasyReg. However, zero-mean stationarity is very rare for economic time series. Therefore, it is recommended to always include an intercept in model (90), so that the alternative hypothesis is

stationarity about a constant. This is ADF test 2 in EasyReg.

13.3 Unit root processes with drift

Consider the random walk with a constant:

$$Y_t = Y_{t-1} + \tau + U_t. \quad (92)$$

This is a random walk with **drift**, with τ the drift parameter. By t times backwards substitution of (92) we get

$$Y_t = Y_0 + \tau \cdot t + \sum_{j=1}^t U_j. \quad (93)$$

Due to the deterministic time trend $\tau \cdot t$ this process moves upwards if $\tau > 0$ and downwards if $\tau < 0$. An example of a unit root process with drift is the log of the nominal GDP of the U.S., which is displayed in Figure 11 for the years 1909-1988.

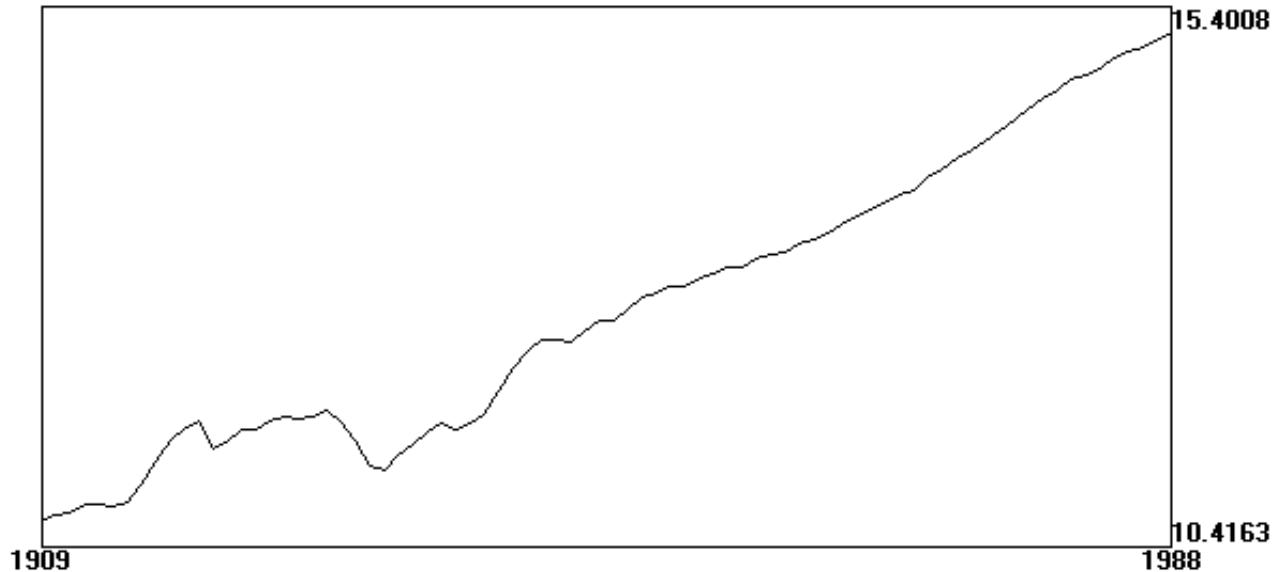


Figure 11: Log of nominal GDP of the U.S.

In this case the proper alternative to the unit root with drift hypothesis is linear trend stationarity:

$$Y_t = \tau_0 + \tau_1 \cdot t + X_t, \text{ where } X_t \text{ is zero-mean stationary.} \quad (94)$$

In particular, if X_t is an AR(p) process, then Y_t can be written as (90), but now including a time trend as well.

Proposition 11. An AR(p) process Y_t with intercept and time trend can always be written as

$$\Delta Y_t = \alpha_0 + \alpha_1 t + \alpha_2 Y_{t-1} + \sum_{j=1}^{p-1} \gamma_j \Delta Y_{t-j} + U_t. \quad (95)$$

If the process Y_t is trend stationary then $\alpha_2 < 0$, and if Y_t is a unit root process with drift process then $\alpha_2 = 0$. In the latter case the t value \hat{t}_{α_2} of α_2 converges in distribution to a non-normal distribution with density displayed in Figure 12.

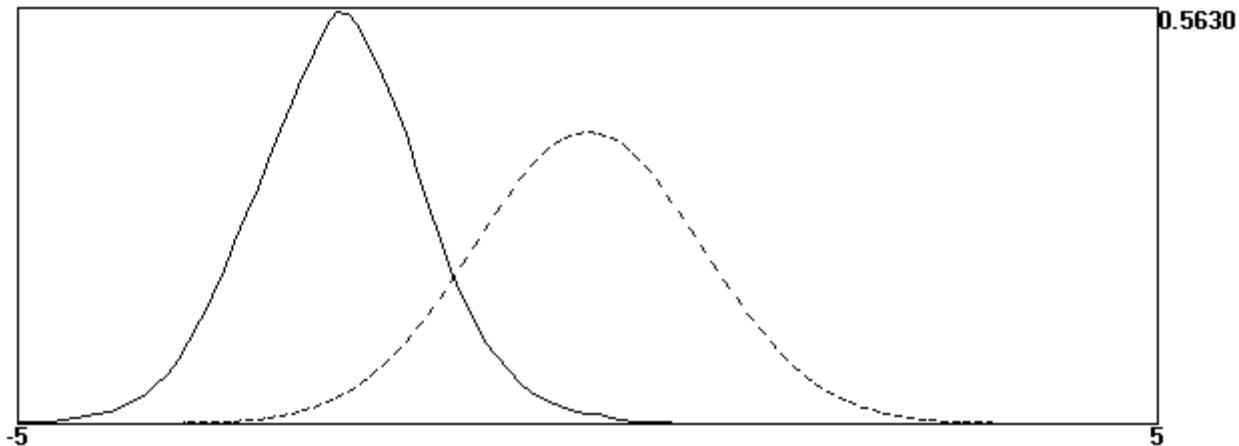


Figure 12: Density of \hat{t}_{α_2} for model (95) under the null hypothesis $\alpha_2 = 0$ (solid curve) compared with the standard normal density (dashed curve)

Note that in this case the asymptotic density of \hat{t}_{α_2} under the null hypothesis is even farther to the left of the standard normal density than in the previous two cases.

The asymptotic 5% and 10% quantiles are now

$$P(\hat{t}_{\alpha_2} \leq -3.41) = 0.05, \quad P(\hat{t}_{\alpha_2} \leq -3.13) = 0.1. \quad (96)$$

Thus, the null hypothesis $\alpha_2 = 0$ is rejected at the 5% significance level in favor of the alternative hypothesis $\alpha_2 < 0$ if $\hat{t}_{\alpha_2} \leq -3.41$, and the null hypothesis $\alpha_2 = 0$ is accepted at the 5% significance level if $\hat{t}_{\alpha_2} > -3.41$.

13.4 Choices you have to make

In conducting the ADF test we have to make three decisions: First, we have to determine whether to transform the time series to make it unbounded, because a bounded time series cannot

be a unit root process. For example, the time series Y_t = “Percentage unemployment rate” takes only values between 0 and 100 (%). If Y_t were a unit root process without drift then

$Y_t - Y_{t-1} = X_t$, where X_t is a zero-mean stationary process, and $Y_t = Y_0 + \sum_{j=1}^t X_j$. It follows from the central limit theorem that for $t \rightarrow \infty$, $(1/\sqrt{t})\sum_{j=1}^t X_j$ converges in distribution to the normal distribution with zero expectation, hence

$$\lim_{t \rightarrow \infty} P[Y_t \leq 0] = \lim_{t \rightarrow \infty} P[Y_t/\sqrt{t} \leq 0] = \lim_{t \rightarrow \infty} P[Y_0/\sqrt{t} + (1/\sqrt{t})\sum_{j=1}^t X_j \leq 0] = 1/2,$$

which is impossible. Therefore, we have to make Y_t unbounded. For the unemployment rate involved we can do that by transforming Y_t to $Y_t^* = \ln(Y_t/(100-Y_t))$, and then test whether Y_t^* is a unit root process.

Now suppose that the time series Y_t is positive valued and not bounded from above, and that $Y_t - Y_{t-1} = X_t$, where X_t is a stationary process. Because $Y_t > 0$ we must have that for all t , $X_t > -Y_{t-1}$. However, this condition implies that X_t depends on the non-stationary process Y_{t-1} , and therefore X_t is non-stationary itself! Only if $X_t > 0$ for all t is it possible that $Y_t - Y_{t-1} = X_t$ with X_t is a stationary process, but then it is not possible that $Y_t < Y_{t-1}$, which is rare for economic time series. Therefore, also in the case where Y_t is positive valued one should make it unbounded, by taking the log transformation, before testing for a unit root.

Second, we have to determine what the appropriate alternative hypothesis is: Either stationarity about a constant (ADF test 2), or trend stationarity (ADF test 3). As to this choice, always plot the time series first and see whether there is a trend pattern in the time series, like in Figure 11. However, there are cases where the time series plot shows a trend but trend stationarity is impossible. For example, let us have a look at the time series $Y_t^* = \ln(Y_t/(100-Y_t))$, where Y_t is the monthly percentage unemployment rate in the US from month 1948.01 to month 1995.09.

At first sight one would conclude from Figure 13 below that Y_t^* is either a unit root with drift process or a trend stationary process, with positive trend slope. However, in the latter case $\lim_{t \rightarrow \infty} Y_t^* = \infty$, hence $\lim_{t \rightarrow \infty} Y_t = 100$. Thus, if the trend stationarity hypothesis were true the unemployment rate in the far future would becomes 100%, which seems quite unlikely. Consequently, despite its appearance, Y_t^* cannot be trend stationary, hence the only plausible alternative hypothesis is that Y_t^* is stationary about a constant.

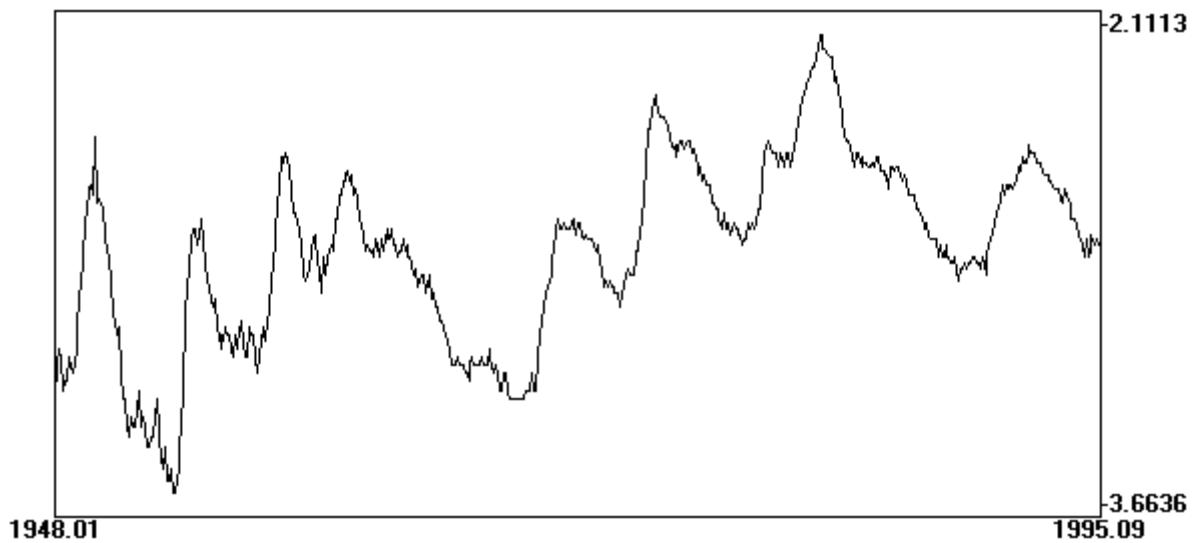


Figure 13: Plot of $\ln(Y_t/(100-Y_t))$, where Y_t is the monthly percentage unemployment rate in the US

Third, we have to choose p . There are various ways to do that. Given an initial choice of p , EasyReg will compute the Akaike, Hannan-Quinn and Schwarz information criteria, up to the initial choice of p . EasyReg will also automatically conduct a sequence of Wald tests to test whether the initial p can be reduced. Moreover, you can also fit an AR model to the first differences of the time series involved, and determine the appropriate order of the AR model involved as discussed earlier, because the choice of p is only critical under the null hypothesis.

13.5 The Breitung tests

The disadvantage of the ADF tests is that these tests only apply to $\text{AR}(p)$ processes with a unit root, and that seasonal variation is excluded. An alternative test for a unit root that also applies to ARMA processes with a unit root in the AR lag polynomial is the Breitung test, which only requires to specify the alternative hypothesis.

Given a time series Y_1, Y_2, \dots, Y_n , let

$$X_t = Y_t \text{ if the alternative hypothesis is zero-mean stationarity,} \quad (97)$$

$$X_t = Y_t - \bar{Y} \text{ if the alternative hypothesis is stationarity about a constant,} \quad (98)$$

where $\bar{Y} = (1/n)\sum_{t=1}^n Y_t$, or

$$X_t = Y_t - \hat{\alpha} - \hat{\beta}t \text{ if the alternative hypothesis is trend stationarity,} \quad (99)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the OLS estimates of the intercept and the slope parameter, respectively, of the linear regression of Y_t on the constant 1 and time t . Thus, in the latter case X_t is the OLS residual of the trend regression involved.

Next, let $S_t = \sum_{j=1}^t X_j$. Then the test statistic of the Breitung test is B_n/n , where

$$B_n = \frac{\sum_{t=1}^n S_t^2}{n \cdot \sum_{t=1}^n X_t^2}. \quad (100)$$

Under the unit root hypothesis B_n/n converges in distribution, where the limiting distribution is non-normal and different for each of the three cases (97), (98) and (99). On the other hand, under one of the alternatives (97), (98) or (99), B_n itself converges in distribution, hence B_n/n converges in probability to zero. Thus, the Breitung test is a left-sided test: The unit root (with drift) hypothesis is rejected if B_n/n is smaller than a critical value. The critical values are different for the cases (97), (98) and (99).

In the case of seasonal data, the Breitung test should be conducted on seasonal moving averages in order to eliminate possible seasonal variation in the time series involved. For example, if Y_t is a quarterly time series satisfying (76), the Breitung test should be conducted with Y_t replaced by $\bar{Y}_t = (1/4)\sum_{i=0}^3 Y_{t-i}$, because then the seasonal dummy variables in (76) are wiped out.

14. ARIMA models

A non-seasonal ARIMA(p,r,q) model for a time series Y_t takes the form

$$X_t = (1-L)^r Y_t, \quad \varphi_p(L)X_t = \beta_0 + \psi_q(L)U_t, \quad (101)$$

where X_t is a stationary ARMA(p,q) process. Compare (52) and Proposition 8. The parameter r indicates how many time we need to difference Y_t to get a stationary ARMA(p,q) process X_t . For economic time series r is usually either 0 or 1, and can be determined on the basis of the Breitung test. Once you have determined r and transformed Y_t to X_t , the ARMA model for X_t can be specified and estimated as before, and be used for forecasting out-of-sample values of X_t and Y_t .

15. *ARCH and GARCH models, and forecasting volatility*

15.1 *ARCH errors*

ARCH stands for Auto-Regressive Conditional Heteroskedasticity, and relates to the conditional variance of model errors. Let U_t be the error in a (conditional expectation) model for Y_t . As we have seen before, the errors of a correctly specified model for a time series Y_t should satisfy $E[U_t | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots] = 0$. See (19). However, in general this condition does not imply that the conditional variance of U_t ,

$$\sigma_t^2 = E[U_t^2 | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots], \quad (102)$$

is constant. Only if Y_t is Gaussian it is guaranteed that σ_t^2 is constant.

In the case of ARCH(p) errors, U_t is specified as

$$U_t = e_t \sqrt{a_0 + a_1 U_{t-1}^2 + a_2 U_{t-2}^2 + \dots + a_p U_{t-p}^2} = e_t \sqrt{a_0 + \sum_{j=1}^p a_j U_{t-j}^2}, \quad (103)$$

where $e_t \sim i.i.d. N(0,1)$, $a_0 > 0$, and $a_j \geq 0$, $j = 1, 2, \dots, p$.

The restrictions on the parameters involved are needed to guarantee that the expression over which the square root is taken is always positive. It follows from (103), and the conditional expectation property (120) [with $V = e_t$ and $X = \sqrt{a_0 + \sum_{j=1}^p a_j U_{t-j}^2}$], that

$$\begin{aligned} E[U_t | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots] &= E[e_t | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots] \cdot \sqrt{a_0 + \sum_{j=1}^p a_j U_{t-j}^2} \\ &= E[e_t] \cdot \sqrt{a_0 + \sum_{j=1}^p a_j U_{t-j}^2} = 0. \end{aligned} \quad (104)$$

Moreover, similar to (104) we have

$$\begin{aligned} \sigma_t^2 &= E[U_t^2 | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots] = E[e_t^2 | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots] \cdot (a_0 + \sum_{j=1}^p a_j U_{t-j}^2) \\ &= E[e_t^2] \cdot (a_0 + \sum_{j=1}^p a_j U_{t-j}^2) = a_0 + \sum_{j=1}^p a_j U_{t-j}^2. \end{aligned} \quad (105)$$

Denoting $V_t = U_t^2 - \sigma_t^2$, the result (105) can be written as an AR(p) model for U_t^2 :

$$U_t^2 = a_0 + \sum_{j=1}^p a_j U_{t-j}^2 + V_t, \quad (106)$$

where $E[V_t | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots] = 0$. Denoting

$$\alpha(L) = 1 - a_1 L - a_2 L^2 - \dots - a_p L^p, \quad (107)$$

the process (106) is stationary if $\alpha(z) = 0$ implies $|z| > 1$. The latter condition has to be imposed as well.

15.2 GARCH errors

GARCH stands for Generalized Auto-Regressive Conditional Heteroskedasticity. In the case of GARCH(q,p) errors, U_t is specified as

$U_t = e_t \sigma_t$, where $e_t \sim i.i.d. N(0,1)$ and

$$\sigma_t^2 = \alpha_0 + \alpha_1 U_{t-1}^2 + \alpha_2 U_{t-2}^2 + \dots + \alpha_p U_{t-p}^2 + \theta_1 \sigma_{t-1}^2 + \theta_2 \sigma_{t-2}^2 + \dots + \theta_q \sigma_{t-q}^2 \quad (108)$$

with $\alpha_0 > 0$, $\alpha_j \geq 0$, $j = 1, 2, \dots, p$, $\theta_i \geq 0$, $i = 1, 2, \dots, q$.

Again, the restrictions on the parameters involved are needed to guarantee that $\sigma_t^2 > 0$. Denoting

$$\theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_q L^q \quad (109)$$

and using (107), the model for σ_t^2 in (108) can be written as

$$\theta(L)\sigma_t^2 = \alpha_0 + (1 - \alpha(L))U_t^2. \quad (110)$$

Assuming that $\theta(z) = 0$ implies $|z| > 1$, the lag polynomial $\theta(L)$ is invertible, hence it follows from (110) that

$$\sigma_t^2 = \alpha_0/\theta(1) + \theta(L)^{-1}(1 - \alpha(L))U_t^2, \quad (111)$$

and thus

$$U_t^2 = \alpha_0/\theta(1) + \theta(L)^{-1}(1 - \alpha(L))U_t^2 + V_t, \quad (112)$$

where $V_t = U_t^2 - \sigma_t^2$. Applying the lag polynomial $\theta(L)$ to both sides of equation (112) yields an ARMA(max(p,q), q) model for U_t^2 :

$$(\theta(L) + \alpha(L) - 1)U_t^2 = \left(1 - \sum_{i=1}^q \theta_i L^i - \sum_{j=1}^p \alpha_j L^j\right)U_t^2 = \alpha_0 + \theta(L)V_t, \quad (113)$$

Note that if we would choose $p = 0$, so that $\alpha(L) = 1$, then (113) becomes an ARMA model with common AR and MA lag polynomial $\theta(L)$ and thus with common roots: $\theta(L)U_t^2 = \alpha_0 + \theta(L)V_t$, which is equivalent to $U_t^2 = \alpha_0/\theta(1) + V_t$. Consequently, $\sigma_t^2 = \alpha_0/\theta(1)$ is then constant because $E[V_t | Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, \dots] = 0$. This result follows also directly from (111): If $\alpha(L) = 1$ then (111) implies $\sigma_t^2 = \alpha_0/\theta(1)$. Therefore, in specifying a GARCH model for the errors U_t you have to choose $p \geq 1$.

15.3 Estimating a model with (G)ARCH errors via EasyReg

The option to estimate a model with ARCH or GARCH errors is only available in

EasyReg if you estimate a linear regression model first, via Menu > Single equation models > Linear regression models. The ARIMA module in EasyReg does not have this option. This is not a restriction, though. You can estimate an ARIMA model with GARCH errors as follows.

- (1) First, determine whether you have to difference the time series in order to make it stationary, by testing whether the time series is a unit root process. If so, take the first differences of the time series, via Menu > Input > Transform variables > Time series transformations. Let Y_t be the stationary time series involved.
- (2) Add missing values to the data, via Menu > Input > Prepare data for forecasting, in order to enable out of sample forecasting.
- (3) Use the option “ARIMA section via information criteria” to specify an ARMA model. If Y_t is stationary about a constant, include the constant 1. If Y_t is trend stationary, include the constant 1 and the time t , and if Y_t is a seasonal time series you may include seasonal dummy variables as well.
- (4) Regress Y_t on a constant, and eventually a time trend and/or seasonal dummy variables, via OLS.
- (5) Once you have estimated this model, choose the option “Re-estimate the model with ARMA errors”, under menu item “Options” in the “What to do next?” window.
- (6) When done, choose the option “Re-estimate the model with GARCH errors”. This yields an ARMA model with GARCH errors.

If after that you want to specify a different ARMA model for Y_t , or a different GARCH model for the errors, you have to redo the last three steps. The details of these steps are explained in the EasyReg guided tour on OLS estimation, which you can access via Tours > OLS.HTM.

To illustrate these procedure, I will choose $Y_t = \%DIF1[SP 500]$, which is the percentage change of the monthly SP 500 index in the US, starting at month 1 of 1950. The latter time series is in the EasyReg data base. Moreover, I have added missing values to the time series, to enable out-of-sample forecasting. The time series is stationary about a constant. The automatic ARMA model selection on the basis of the information criteria suggests that Y_t is an MA(1) process.

Regressing Y_t on a constant only:

Dependent variable:
 $Y = \%DIF1[SP500 \text{ index}]$
 First available observation = 2 (=1939.02)
 Last available observation = 673 (=1995.01)
 First chosen observation = 133 (=1950.01)
 Last chosen observation = 673 (=1995.01)
 Number of usable chosen observations: 541
 X variables:
 $X(1) = 1$

Model:
 $Y = b(1)X(1) + U,$
 where U is the error term, satisfying $E[U|X(1)] = 0.$

OLS estimation results

Parameters	Estimate	t-value	H.C.	t-value
		(S.E.)	(H.C.)	S.E.)
		[p-value]	[H.C.]	p-value]
b(1)	0.67191	4.723	4.723	
		(0.14227)	(0.14227)	
		[0.00000]	[0.00000]	

Notes:

- 1: S.E. = Standard error
- 2: H.C. = Heteroskedasticity Consistent. These t-values and standard errors are based on White's heteroskedasticity consistent variance matrix.
- 3: The two-sided p-values are based on the normal approximation.

Effective sample size (n):	541
Variance of the residuals:	10.949943
Standard error of the residuals (SER):	3.30907
Residual sum of squares (RSS):	5912.969032
(Also called SSR = Sum of Squared Residuals)	
Total sum of squares (TSS):	5912.969032

Information criteria:

Akaike:	2.39518E+00
Hannan-Quinn:	2.39828E+00
Schwarz:	2.40312E+00

Next, choose the option “Re-estimate the model with ARMA errors”, and specify an ARMA(0,1) model:

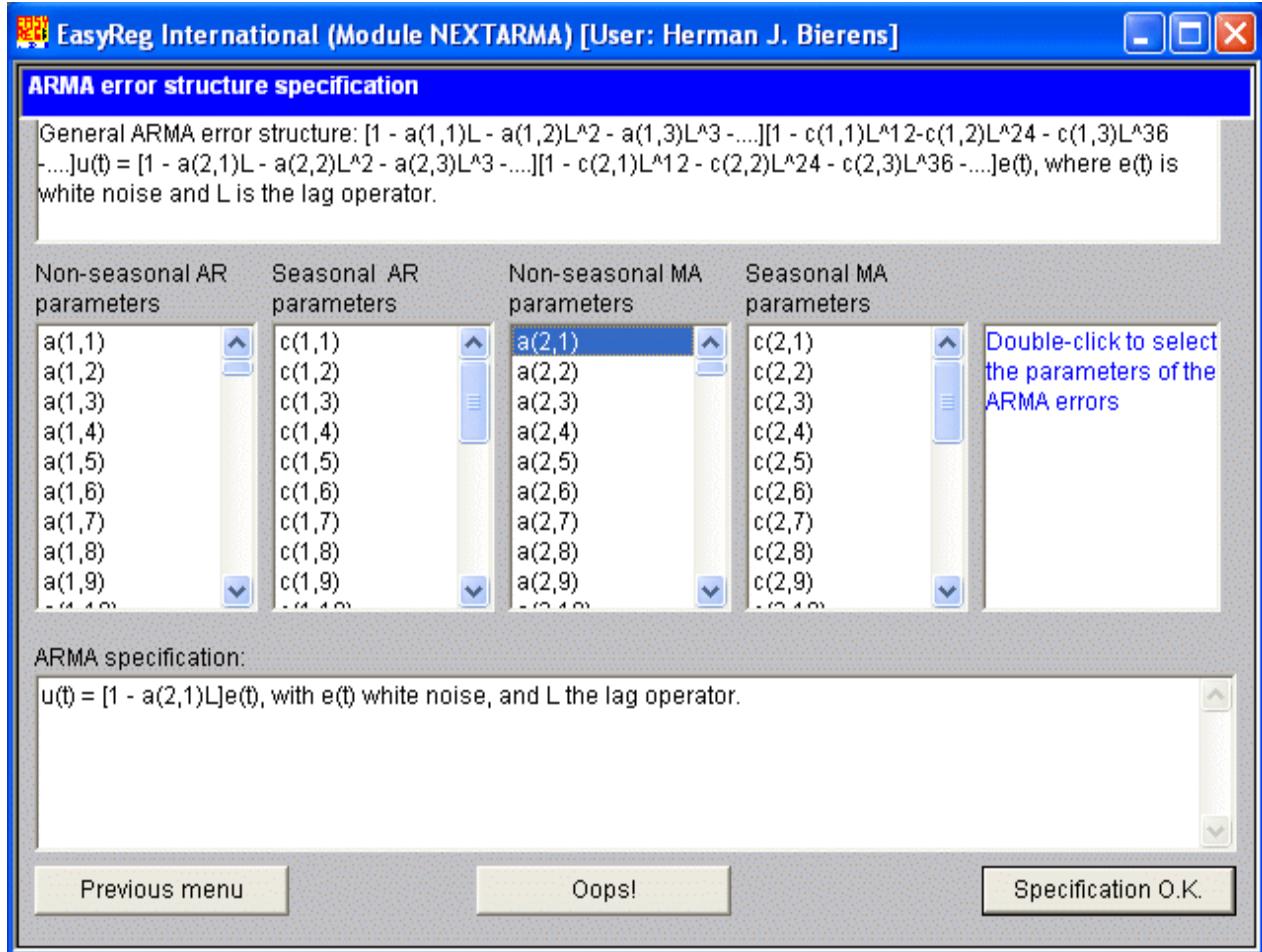


Figure 14: MA(1) error specification

The EasyReg results are:

Error specification: $u(t) = [1 - a(2,1)L]e(t)$, with $e(t)$ white noise and L the lag operator.

```

Parameters   estimate t-value [p-value] HC t-value) [HC p-value]
b(1)        0.672896  3.862 [0.00011]      3.872 [0.00011]
a(2,1)      -0.265969 -6.404 [0.00000]     -5.288 [0.00000]
RSS = 55.271441184E+02
s.e. = 32.022557328E-01
R-square = 0.0653
n = 541

```

Information criteria:
 Akaike: 2.33140E+00
 Hannan-Quinn: 2.33761E+00
 Schwarz: 2.34727E+00

Next, re-estimate the MA(1) model with GARCH errors. I will specify an GARCH(1,1) model.

See Figure 15.

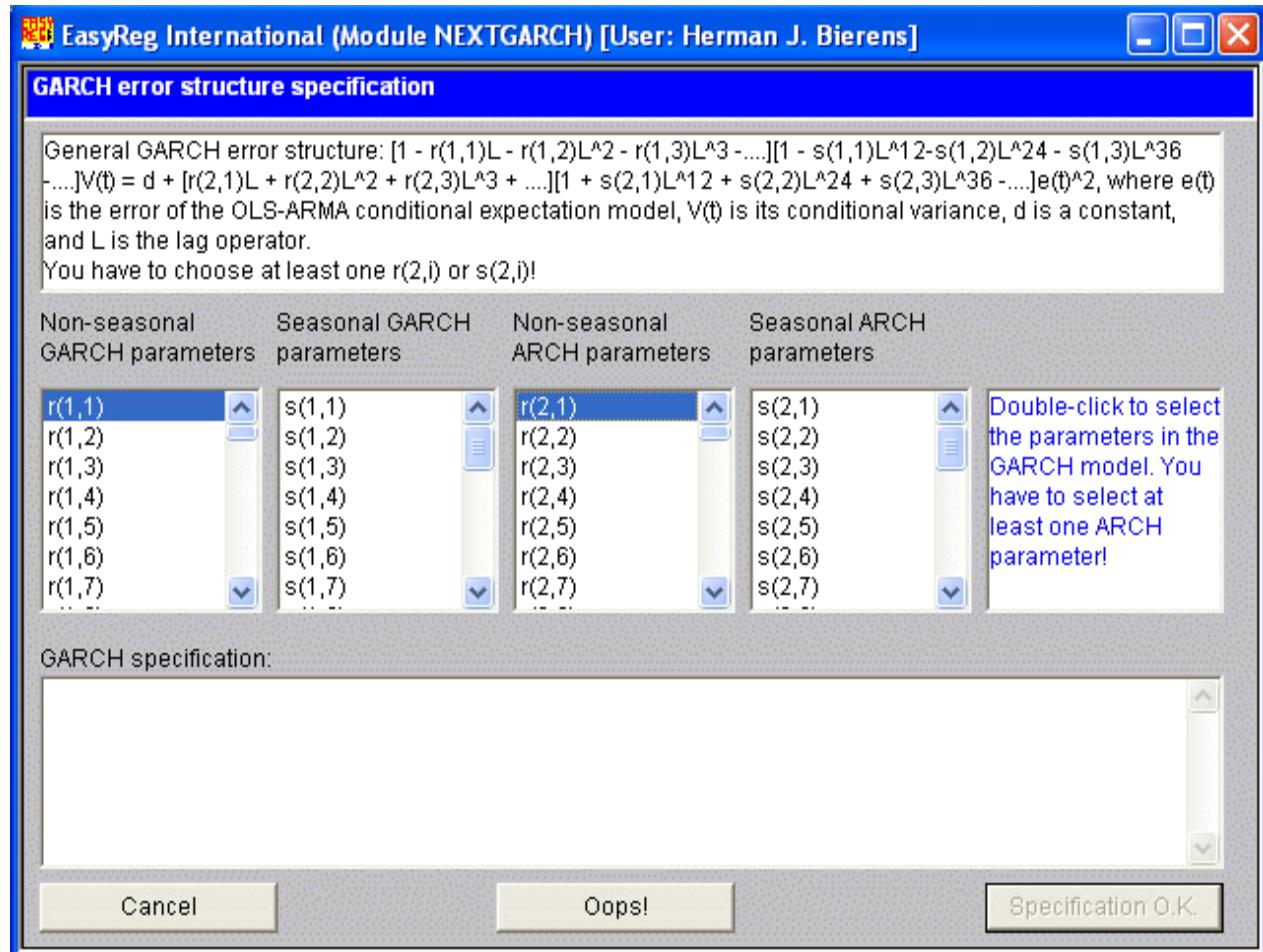


Figure 15

The EasyReg results are:

GARCH specification:
 $[1 - r(1,1)L]V(t) = d + [r(2,1)L]e(t)^2,$
 where $e(t)$ is the error of the OLS-ARMA conditional expectation model, $V(t)$ is its conditional variance, d is a constant, and L

```

is the lag operator.
Maximum likelihood estimation results:
Parameters   ML estimate t-value [p-value]
b(1)          0.764270  4.669 [0.00000]
a(2,1)        -0.121729 -4.916 [0.00000]
d             8.656218  3.076 [0.00210]
r(1,1)        0.001671  0.006 [0.99546]
r(2,1)        0.149719  2.983 [0.00286]
[The two-sided p-values are based on the normal approximation]

Log-Likelihood = -13.890927734E+02
RSS            = 56.461249666E+02
s.e.           = 32.455839592E-01
R-square       = 0.0451
n              = 541
Information criteria:
    Akaike:          2.315885497
    Hannan-Quinn:    2.331403141
    Schwarz:         2.355565897

```

The parameter $r(1,1)$ is not significant, indicating the model is an MA(1)-ARCH(1) model. To re-estimate the model as an MA(1)-ARCH(1) model we have to start all over again. The results are:

```

Maximum likelihood estimation results:
Parameters   ML estimate t-value [p-value]
b(1)          0.768396  4.703 [0.00000]
a(2,1)        -0.119612 -4.851 [0.00000]
d             8.685394  15.866 [0.00000]
r(2,1)        0.147329  3.106 [0.00190]
[The two-sided p-values are based on the normal approximation]

Log-Likelihood = -13.890972900E+02
RSS            = 56.498197461E+02
s.e.           = 32.436213715E-01
R-square       = 0.0445
n              = 541
Information criteria:
    Akaike:          2.312205337
    Hannan-Quinn:    2.324619452
    Schwarz:         2.343949657

```

EasyReg also provides the option to plot the GARCH variances:

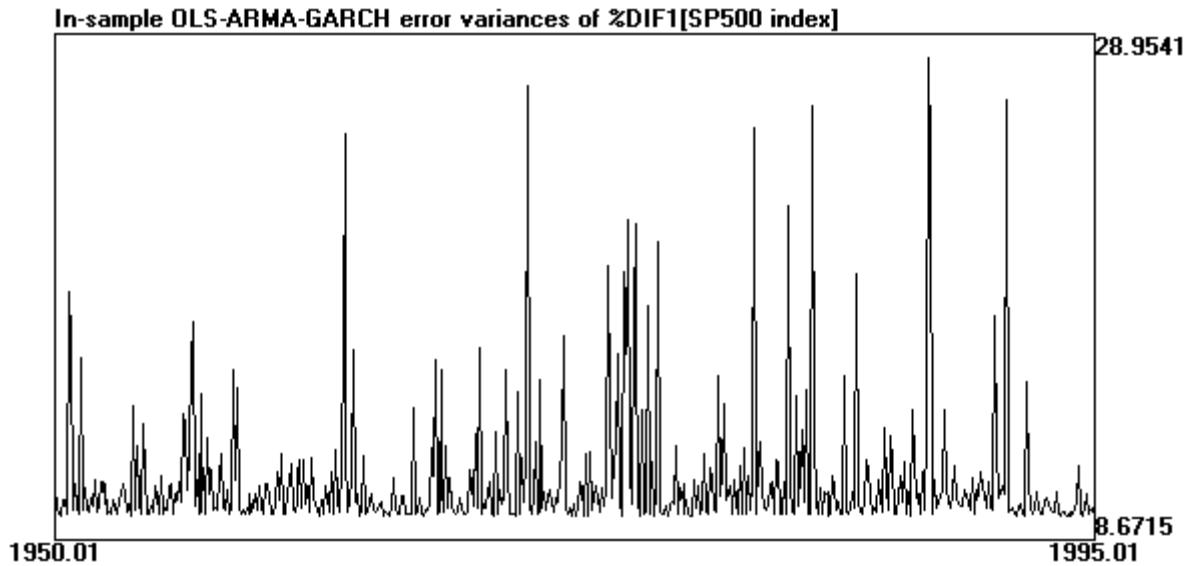


Figure 16

15.4 Forecasting volatility

In finance the (G)ARCH variance σ_t^2 is called the **volatility** of a financial time series. EasyReg does not allow you to recursively forecast the conditional variance σ_t^2 directly, so that you have to use the following indirect approach.

First, add at least one missing value to the data, via Menu > Input > Prepare time series for forecasting, in order to enable the forecast option.

After completion of step (6) above, open “Options > Write the residuals to the input file”.

Next, take the square of the residuals \hat{U}_t , via Menu > Input > Transform variables > Multiplicative transformation, with power 2. Then \hat{U}_t^2 is added to the data file.

Now estimate an ARMA model for \hat{U}_t^2 , similar to the ARMA(max(p,q), q) model (113) for GARCH(p,q) errors, or an AR(p) model for \hat{U}_t^2 in the ARCH(p) case.

Finally, choose the recursive forecast option. Then you get the recursive out-of-sample forecasts of σ_t^2 .

Technical Appendix

This appendix contains some tedious derivations together with advanced material. None is part of the course: You may skip it if you wish. I will not ask questions on the exam about the contents of this appendix, nor is understanding of this material essential for doing well on the exam.

A.1 *Optimality of conditional expectations for forecasting*

To show that the conditional expectation function $E[Y|X]$ is the best forecast of Y given X , recall that the conditional expectation $E[Y|X]$ is a function of X , for example

$$E[Y|X] = g(X). \quad (114)$$

In particular, if $f(y,x)$ is the joint density of Y and X , then $f_x(x) = \int_{-\infty}^{\infty} f(y,x) dy$ is the marginal density of X , $f(y|x) = f(y,x)/f_x(x)$ is the conditional density of Y given $X = x$, and

$$E[Y|X=x] \stackrel{\text{def.}}{=} \int_{-\infty}^{\infty} y f(y|x) dy = g(x), \quad (115)$$

is the conditional expectation of Y given $X = x$. Plugging in X for x , we get (114).

Define

$$U = Y - E[Y|X] = Y - g(X). \quad (116)$$

Substituting (116) (in the form $Y = g(X) + U$) in (2) now yields:

$$\begin{aligned} E[(Y - \hat{Y})^2] &= E[(U + g(X) - \phi(X))^2] \\ &= E[U^2 + 2.U(g(X) - \phi(X)) + (g(X) - \phi(X))^2] \\ &= E[U^2] + 2.E[U(g(X) - \phi(X))] + E[(g(X) - \phi(X))^2]. \end{aligned} \quad (117)$$

Next, recall the following properties of conditional expectations: For any function ψ of X , and random variable V and Z ,

$$E[\psi(X)|X] = \psi(X), \quad (118)$$

$$E[E[Z|X]] = E[Z], \quad (119)$$

$$E[V.\psi(X)|X] = \psi(X).E[V|X]. \quad (120)$$

Property (119) is known as the law of iterated expectations.

It follows from (114), (116) and (118), with $\psi(x) = g(x)$ in (118), that

$$E[U|X] = E[Y|X] - E[g(X)|X] = E[Y|X] - g(X) = g(X) - g(X) = 0. \quad (121)$$

Moreover it follows from (120) and (121), with $V = U$ and $\psi(X) = g(X) - \varphi(X)$ in (120), that

$$E[U(g(X) - \varphi(X))|X] = (g(X) - \varphi(X))E[U|X] = 0, \quad (122)$$

hence it follows from (119), with $Z = U(g(X) - \varphi(X))$, and (122) that

$$E[U(g(X) - \varphi(X))] = E[E(U(g(X) - \varphi(X))|X)] = E[(g(X) - \varphi(X))E(U|X)] = 0. \quad (123)$$

The latter result implies that (117) can be written as

$$E[(Y - \hat{Y})^2] = E[U^2] + E[(g(X) - \varphi(X))^2]. \quad (124)$$

Clearly, (124) is minimal for $\varphi(X) = g(X)$. Thus the conditional expectation $g(X) = E[Y|X]$ is the best forecasting scheme for Y given X . This argument can easily be generalized to the case with multiple explanatory variables.

A.2 The distribution of the forecast error

Recall that the OLS estimators of β and α in model (4) on the basis of the observations

(Y_j, X_j) for $j = 1, \dots, n$ are

$$\hat{\beta} = \frac{\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^n (X_j - \bar{X})^2} = \frac{\sum_{j=1}^n (X_j - \bar{X})Y_j}{\sum_{j=1}^n (X_j - \bar{X})^2} = \beta + \frac{\sum_{j=1}^n (X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \quad (125)$$

and

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \cdot \bar{X} = \alpha + \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j, \quad (126)$$

respectively, where $\bar{X} = (1/n)\sum_{j=1}^n X_j$ and $\bar{Y} = (1/n)\sum_{j=1}^n Y_j$. Then the forecast error $Y_{n+1} - \hat{Y}_{n+1} = \alpha + \beta X_{n+1} + U_{n+1} - \hat{\alpha} - \hat{\beta} X_{n+1}$ can be written as

$$\begin{aligned} Y_{n+1} - \hat{Y}_{n+1} &= U_{n+1} - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) \cdot X_{n+1} = U_{n+1} - \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j \\ &\quad - \sum_{j=1}^n \left(\frac{X_{n+1}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j = U_{n+1} - \sum_{j=1}^n \left(\frac{1}{n} + \frac{(X_{n+1} - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j. \end{aligned} \quad (127)$$

where the second equality in (127) follows from (125) and (126). Moreover, due to the independence of the explanatory variables X_j and the model errors U_j the variance of the forecast error can be computed as follows.

$$\begin{aligned}
\sigma_{Y_{n+1} - \hat{Y}_{n+1}}^2 &= \sigma^2 + \sum_{j=1}^n \left(\frac{1}{n} + \frac{(X_{n+1} - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)^2 \cdot \sigma^2 \\
&= \sigma^2 \left(1 + \frac{1}{n} + \frac{2}{n} \cdot \frac{(X_{n+1} - \bar{X}) \sum_{j=1}^n (X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{(X_{n+1} - \bar{X})^2 \sum_{j=1}^n (X_j - \bar{X})^2}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \right) \\
&= \sigma^2 \left(\frac{n+1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right).
\end{aligned} \tag{128}$$

The result (7) follows now from the fact that linear combinations of normally distributed random variables are normally distributed themselves.

A.3 Stationarity condition and moving average representation for an AR(1) process

By repeated backwards substitution of (21) we can write

$$\begin{aligned}
Y_t &= \beta_0 + \beta_1(\beta_0 + \beta_1 Y_{t-2} + U_{t-1}) + U_t \\
&= \beta_0 + \beta_1 \beta_0 + \beta_1^2 Y_{t-2} + U_t + \beta_1 U_{t-1} \\
&= \beta_0 + \beta_1 \beta_0 + \beta_1^2 (\beta_0 + \beta_1 Y_{t-3} + U_{t-2}) + U_t + \beta_1 U_{t-1} \\
&= \beta_0 + \beta_1 \beta_0 + \beta_1^2 \beta_0 + \beta_1^3 Y_{t-3} + U_t + \beta_1 U_{t-1} + \beta_1^2 U_{t-2} \\
&\quad \dots \\
&= \beta_0 + \beta_1 \beta_0 + \beta_1^2 \beta_0 + \dots + \beta_1^{m-1} \beta_0 + \beta_1^m Y_{t-m} \\
&\quad + U_t + \beta_1 U_{t-1} + \beta_1^2 U_{t-2} + \dots + \beta_1^{m-1} U_{t-m+1} \\
&= \beta_0 \sum_{k=0}^{m-1} \beta_1^k + \sum_{k=0}^{m-1} \beta_1^k U_{t-k} + \beta_1^m Y_{t-m}.
\end{aligned} \tag{129}$$

If Y_t is covariance stationary then $E[Y_t] = \mu$ and $E[(Y_t - \mu)(Y_{t-m} - \mu)] = \gamma(m)$ for all t . Then it follows from (129) that

$$\mu = \beta_0 \sum_{k=0}^{m-1} \beta_1^k + \sum_{k=0}^{m-1} \beta_1^k E[U_{t-k}] + \beta_1^m \mu = \beta_0 \sum_{k=0}^{m-1} \beta_1^k + \beta_1^m \mu. \tag{130}$$

Next, subtract (130) from (129),

$$Y_t - \mu = \sum_{k=0}^{m-1} \beta_1^k U_{t-k} + \beta_1^m (Y_{t-m} - \mu), \tag{131}$$

take the square of both sides,

$$(Y_t - \mu)^2 = \left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k} \right)^2 + 2\beta_1^m \left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k} \right) (Y_{t-m} - \mu) + \beta_1^{2m} (Y_{t-m} - \mu)^2 \quad (132)$$

and take expectations,

$$\begin{aligned} E[(Y_t - \mu)^2] &= E \left[\left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k} \right)^2 \right] + 2\beta_1^m \sum_{k=0}^{m-1} \beta_1^k E[U_{t-k}(Y_{t-m} - \mu)] + \beta_1^{2m} E(Y_{t-m} - \mu)^2 \\ &= \sigma^2 \sum_{k=0}^{m-1} \beta_1^{2k} + \beta_1^{2m} E(Y_{t-m} - \mu)^2. \end{aligned} \quad (133)$$

The last equality in (133) follows from (19) and property (120), and the fact that under the assumption that the errors U_t are independent $N(0, \sigma^2)$ distributed,

$$E \left[\left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k} \right)^2 \right] = \sigma^2 \sum_{k=0}^{m-1} \beta_1^{2k}. \quad (134)$$

If Y_t is covariance stationary then $E[(Y_{t-m} - \mu)^2] = E[(Y_t - \mu)^2] = \gamma(0)$, so that (133) reads:

$$\gamma(0) = \sigma^2 \sum_{k=0}^{m-1} \beta_1^{2k} + \beta_1^{2m} \gamma(0). \quad (135)$$

However, if $|\beta_1| \geq 1$ then the right-hand side of (135) converges to ∞ if we let $m \rightarrow \infty$, which contradicts the condition that $\gamma(0) < \infty$. On the other hand, if $|\beta_1| < 1$ then $\beta_1^{2m} E[Y_{t-m}^2] = \beta_1^{2m} E[Y_1^2] \rightarrow 0$ as $m \rightarrow \infty$, hence it follows from (129) by letting $m \rightarrow \infty$ that

$$Y_t = \beta_0 \sum_{k=0}^{\infty} \beta_1^k + \sum_{k=0}^{\infty} \beta_1^k U_{t-k} = \frac{\beta_0}{1-\beta_1} + \sum_{k=0}^{\infty} \beta_1^k U_{t-k}. \quad (136)$$

Note that under the assumption that the errors U_t are independent $N(0, \sigma^2)$ distributed, Y_t in (136) is normally distributed, with expectation $\mu = \beta_0/(1-\beta_1)$ and variance $E[(\sum_{k=0}^{\infty} \beta_1^k U_{t-k})^2] = \sigma^2 \sum_{k=0}^{\infty} \beta_1^{2k} = \sigma^2 / (1 - \beta_1^2)$.

The expression at the right-hand side of (136) is called the Moving Average (MA) representation of a covariance stationary time series.

A.4 AR order selection via information criteria

Recall that for an AR(p) model the Akaike, Hannan-Quinn, and Schwarz information criteria take the form

$$\begin{aligned}\text{Akaike:} \quad c_n^{AR}(p) &= \ln(\hat{\sigma}_p^2) + 2(1+p)/n, \\ \text{Hannan-Quinn:} \quad c_n^{AR}(p) &= \ln(\hat{\sigma}_p^2) + 2(1+p)\ln(\ln(n))/n, \\ \text{Schwarz:} \quad c_n^{AR}(p) &= \ln(\hat{\sigma}_p^2) + (1+p)\ln(n)/n,\end{aligned}$$

where n is the effective sample size and $\hat{\sigma}_p^2$ is the OLS estimator of the error variance $\sigma^2 = E[U_t^2]$. Denoting by \hat{p} the value of p for which $c_n^{AR}(p)$ is minimal:

$$c_n^{AR}(\hat{p}) = \min\{c_n^{AR}(1), \dots, c_n^{AR}(\bar{p})\},$$

where $\bar{p} > p_0$, with p_0 the true value of p , we have in the Hannan-Quinn and Schwarz cases: $\lim_{n \rightarrow \infty} P[\hat{p} = p_0] = 1$, and in the Akaike case $\lim_{n \rightarrow \infty} P[\hat{p} \geq p_0] = 1$ but $\lim_{n \rightarrow \infty} P[\hat{p} = p_0] < 1$. Thus, the Akaike criterion may “overshoot” the true value.

These results are based on the following facts. If $p < p_0$ then $\text{plim}_{n \rightarrow \infty} \hat{\sigma}_p^2 > \text{plim}_{n \rightarrow \infty} \hat{\sigma}_{p_0}^2$, hence in all three cases, $\lim_{n \rightarrow \infty} P[c_n^{AR}(p_0) < c_n^{AR}(p)] = 1$, whereas for $p > p_0$,

$$n \left(\ln(\hat{\sigma}_{p_0}^2) - \ln(\hat{\sigma}_p^2) \right) \xrightarrow{} \chi_{p-p_0}^2 \tag{137}$$

in distribution if $n \rightarrow \infty$. The result (137) is due to the so-called likelihood-ratio test.⁶ Then in the Akaike case,

$$n \left(c_n^{AR}(p_0) - c_n^{AR}(p) \right) = n \left(\ln(\hat{\sigma}_{p_0}^2) - \ln(\hat{\sigma}_p^2) \right) - 2(p-p_0) \xrightarrow{} X_{p-p_0} - 2(p-p_0) \tag{138}$$

in distribution if $n \rightarrow \infty$, where $X_{p-p_0} \sim \chi_{p-p_0}^2$, hence

$$\lim_{n \rightarrow \infty} P[c_n^{AR}(p_0) > c_n^{AR}(p)] = P[X_{p-p_0} > 2(p-p_0)] > 0. \tag{139}$$

Consequently, in the Akaike case we have $\lim_{n \rightarrow \infty} P[\hat{p} \geq p_0] = 1$, but $\lim_{n \rightarrow \infty} P[\hat{p} > p_0] > 0$. Therefore, the Akaike criterion may asymptotically overshoot the correct number of parameters.

Since (137) implies $\text{plim}_{n \rightarrow \infty} n(\ln(\hat{\sigma}_{p_0}^2) - \ln(\hat{\sigma}_p^2))/\ln(\ln(n)) = 0$ and $\text{plim}_{n \rightarrow \infty} n(\ln(\hat{\sigma}_{p_0}^2) -$

⁶ Which, however, is beyond the undergraduate econometrics level, so you have to believe this.

$\ln(\hat{\sigma}_p^2)/\ln(n) = 0$ it follows that in the Hannan-Quinn case,

$$\text{plim}_{n \rightarrow \infty} n(c_n^{AR}(p_0) - c_n^{AR}(p))/\ln(\ln(n)) = 2(p - p_0) \geq 2$$

and in the Schwarz case,

$$\text{plim}_{n \rightarrow \infty} n(c_n^{AR}(p_0) - c_n^{AR}(p))/\ln(n) = p - p_0 \geq 1,$$

so that in both cases $\lim_{n \rightarrow \infty} P[c_n^{AR}(p_0) > c_n^{AR}(p)] = 0$. Hence, $\lim_{n \rightarrow \infty} P[\hat{p} = p_0] = 1$.

A.5 Motivation for the Dickey-Fuller test

Consider an AR(1) model without a constant

$$Y_t = \beta Y_{t-1} + U_t, \text{ where } U_t \sim \text{i.i.d. } N(0,1), \quad (140)$$

which can be written as

$$(1 - \beta L)Y_t = U_t. \quad (141)$$

If $\beta = 1$ then the AR lag polynomial involved has a unit root: $1 - z = 0$ implies $z = 1$. The model then becomes a random walk:

$$Y_t = Y_{t-1} + U_t. \quad (142)$$

This is a nonstationary process. To see this, assume for convenience that

$$Y_t = 0 \text{ for } t \leq 0. \quad (143)$$

Then it follows by backwards substitution of (142) that

$$Y_t = U_1 + U_2 + \dots + U_t = \sum_{j=1}^t U_j, \quad (144)$$

so that under the normality assumption in (140), $Y_t \sim N(0, t)$. Thus, in this case the variance of Y_t blows up to infinity with t .

The model (140) can be rewritten as

$$\Delta Y_t = Y_t - Y_{t-1} = (\beta - 1)Y_{t-1} + U_t = \alpha Y_{t-1} + U_t, \text{ where } \alpha = \beta - 1. \quad (145)$$

Then the unit root hypothesis $\beta = 1$ corresponds to $\alpha = 0$, and the stationarity hypothesis $|\beta| < 1$ corresponds to $-2 < \alpha < 0$. This suggests to test the null hypothesis $\alpha = 0$ by estimating model (145) by OLS and using the t-value \hat{t}_α of α for a left-sided t test. However, the problem is that under the null hypothesis $\alpha = 0$ the t-value \hat{t}_α has no longer an asymptotic standard normal distribution.

To derive the distribution of \hat{t}_α , let us first derive the asymptotic distribution of the OLS estimator $\hat{\alpha}$ of α in the case $\alpha = 0$:

$$\hat{\alpha} = \frac{\sum_{t=1}^n Y_{t-1} \Delta Y_t}{\sum_{t=1}^n Y_{t-1}^2} = \frac{\sum_{t=1}^n U_t Y_{t-1}}{\sum_{t=1}^n Y_{t-1}^2}. \quad (146)$$

Note that by (142), $Y_t^2 - Y_{t-1}^2 = (U_t + Y_{t-1})^2 - Y_{t-1}^2 = U_t^2 + 2U_t Y_{t-1}$, hence

$$2 \sum_{t=1}^n U_t Y_{t-1} = \sum_{t=1}^n (Y_t^2 - Y_{t-1}^2) - \sum_{t=1}^n U_t^2 = Y_n^2 - Y_0^2 - \sum_{t=1}^n U_t^2 = Y_n^2 - \sum_{t=1}^n U_t^2, \quad (147)$$

where the latter equality follows from (143). Therefore we can write

$$n \cdot \hat{\alpha} = \frac{\frac{1}{2} \left((Y_n/\sqrt{n})^2 - (1/n) \sum_{t=1}^n U_t^2 \right)}{(1/n) \sum_{t=1}^n (Y_{t-1}/\sqrt{n})^2}. \quad (148)$$

Note that by (144) and the normality assumption in (140),

$$Y_n/\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{t=1}^n U_t \sim N(0,1), \quad (149)$$

so that $(Y_n/\sqrt{n})^2$ has a χ_1^2 distribution. Moreover, it follows from the law of large numbers that $(1/n) \sum_{t=1}^n U_t^2$ converges in probability to the expectation $E[U_t^2] = 1$, so that

$$(1/n) \sum_{t=1}^n U_t^2 = 1 + r_n, \text{ where } \text{plim}_{n \rightarrow \infty} r_n = 0. \quad (150)$$

To determine the distribution of $(1/n) \sum_{t=1}^n (Y_{t-1}/\sqrt{n})^2$, denote

$$W_n(x) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[x.n]} U_t \text{ if } 1/n \leq x \leq 1, \quad (151)$$

$$W_n(x) = 0 \text{ if } 0 \leq x < 1/n,$$

where $[x.n]$ means the largest natural number $\leq x.n$. For example, $[2.5] = 2$, $[5] = 5$, $[6.999] = 6$.

Then $Y_{t-1}/\sqrt{n} = W_n((t-1)/n)$, hence

$$\begin{aligned} (1/n) \sum_{t=1}^n (Y_{t-1}/\sqrt{n})^2 &= (1/n) \sum_{t=1}^n (W_n((t-1)/n))^2 = (1/n) \sum_{t=1}^n (W_n((t-1)/n))^2 \int_{t-1}^t dz \\ &= (1/n) \sum_{t=1}^n \int_{t-1}^t (W_n((z-1)/n))^2 dz = (1/n) \sum_{t=1}^n \int_{t-1}^t (W_n(z/n))^2 dz = (1/n) \int_0^n (W_n(z/n))^2 dz \\ &= \int_0^1 (W_n(x))^2 dx, \end{aligned} \quad (152)$$

where the fourth equality follows from the fact that for $t-1 \leq z < t$, $[z] = t-1$, hence, $W_n(z/n) = W_n((t-1)/n)$, and the last equality follows by replacing z by $n.x$. Moreover, it follows from (150) and (151) that

$$\left(Y_n/\sqrt{n}\right)^2 - (1/n)\sum_{t=1}^n U_t^2 = \left(W_n(1)\right)^2 - 1 - r_n \quad (153)$$

Combining (148), (152) and (153) it follows now that

$$n.\hat{\alpha} = \frac{\frac{1}{2}\left(\left(W_n(1)\right)^2 - 1 - r_n\right)}{\int_0^1 \left(W_n(x)\right)^2 dx}. \quad (154)$$

Now let us have a closer look at the function $W_n(x)$. First, it follows from (151) and the normality assumption in (140) that for $0 < x \leq 1$,

$$W_n(x) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[x.n]} U_t \sim N\left(0, \frac{[x.n]}{n}\right) \rightarrow N(0, x) \text{ in distribution if } n \rightarrow \infty \quad (155)$$

Moreover, for $0 < x < y \leq 1$,

$$W_n(y) - W_n(x) = \frac{1}{\sqrt{n}} \sum_{t=[x.n]+1}^{[y.n]} U_t \sim N\left(0, \frac{[y.n]-[x.n]}{n}\right) \rightarrow N(0, y-x) \text{ in distribution if } n \rightarrow \infty, \text{ and} \quad (156)$$

$W_n(y) - W_n(x)$ and $W_n(x)$ are independent.

This suggests the existence of a random function $W(x)$ on $[0,1]$, called a Brownian motion or Wiener process, with the following properties:

$$\begin{aligned} &\text{for } 0 < x \leq 1, W(x) \sim N(0, x), \\ &\text{for } 0 < x < y \leq 1, W(y) - W(x) \sim N(0, y-x), \text{ and} \\ &W(x) \text{ and } W(y) - W(x) \text{ are independent,} \end{aligned} \quad (157)$$

such that $W_n(x) \rightarrow W(x)$ in distribution. This result, together with $\text{plim}_{n \rightarrow \infty} r_n = 0$, implies that

$$n.\hat{\alpha} = \frac{\frac{1}{2}\left(\left(W_n(1)\right)^2 - 1 - r_n\right)}{\int_0^1 \left(W_n(x)\right)^2 dx} \rightarrow \frac{\frac{1}{2}\left(\left(W(1)\right)^2 - 1\right)}{\int_0^1 \left(W(x)\right)^2 dx} \text{ in distribution if } n \rightarrow \infty. \quad (158)$$

The t value \hat{t}_α of $\hat{\alpha}$ is defined by

$$\hat{t}_\alpha = \frac{\sqrt{n} \hat{\alpha} \cdot \sqrt{(1/n) \sum_{t=1}^n Y_{t-1}^2}}{\hat{\sigma}}, \quad (159)$$

which can be rewritten as

$$\hat{t}_\alpha = \frac{n \hat{\alpha} \cdot \sqrt{(1/n^2) \sum_{t=1}^n Y_{t-1}^2}}{\hat{\sigma}} = \frac{n \hat{\alpha} \cdot \sqrt{\int_0^1 (W_n(x))^2 dx}}{\hat{\sigma}}, \quad (160)$$

where

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{t=1}^n (\Delta Y_t - \hat{\alpha} Y_{t-1})^2. \quad (161)$$

Under the null hypothesis, and using the previous results, it follows that

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-1} \sum_{t=1}^n (U_t - \hat{\alpha} Y_{t-1})^2 = \frac{1}{n-1} \sum_{t=1}^n U_t^2 - \hat{\alpha}^2 \frac{1}{n-1} \sum_{t=1}^n Y_{t-1}^2 \\ &= \frac{n}{n-1} \left((1/n) \sum_{t=1}^n U_t^2 \right) - \frac{1}{n-1} (n \cdot \hat{\alpha})^2 (1/n^2) \sum_{t=1}^n Y_{t-1}^2 \\ &= \frac{n}{n-1} (1 + r_n) - \frac{1}{n-1} (n \cdot \hat{\alpha})^2 \int_0^1 (W_n(x))^2 dx \rightarrow 1 \text{ in probability if } n \rightarrow \infty. \end{aligned} \quad (162)$$

hence

$$\hat{t}_\alpha \rightarrow \frac{\frac{1}{2} ((W(1))^2 - 1)}{\sqrt{\int_0^1 (W(x))^2 dx}} \text{ in distribution if } n \rightarrow \infty. \quad (163)$$

The density function of the limiting random variable involved is displayed in Figure 9.

A.6 Motivation for the Breitung test

In order to explain the Breitung unit root test, suppose first that

$$Y_t = Y_{t-1} + U_t, \text{ where } U_t \sim \text{i.i.d. } N(0,1) \text{ and } Y_t = 0 \text{ for } t \leq 0, \quad (164)$$

so that

$$Y_t = U_1 + U_2 + \dots + U_t = \sum_{j=1}^t U_j = \sqrt{n} \cdot W_n(t/n), \quad (165)$$

where again $W_n(x)$ is defined by (151). Then similar to (152),

$$\begin{aligned} (1/n^2) \sum_{t=1}^n Y_t^2 &= (1/n^2) \sum_{t=0}^n Y_t^2 = (1/n) \cdot \sum_{t=0}^n (W_n(t/n))^2 = W_n(1)/n + \int_0^1 (W_n(x))^2 \\ &\approx \int_0^1 (W_n(x))^2, \end{aligned} \quad (166)$$

where the latter approximation follows from the fact that $W_n(1)/n \sim N(0, 1/n^2) \rightarrow 0$ if $n \rightarrow \infty$.

Moreover,

$$\begin{aligned} S_t &= \sum_{j=1}^t Y_j = \sqrt{n} \cdot \sum_{j=1}^t W_n(j/n) = \sqrt{n} \cdot \sum_{j=1}^t \int_j^{j+1} W_n(z/n) dz \\ &= \sqrt{n} \cdot \int_1^{t+1} W_n(z/n) dz = n\sqrt{n} \cdot \int_1^{t+1} W_n(z/n) d(z/n) = n\sqrt{n} \cdot \int_0^{t+1} W_n(z/n) d(z/n) \\ &= n\sqrt{n} \cdot \int_0^{(t+1)/n} W_n(x) dx. \end{aligned} \quad (167)$$

so that

$$\begin{aligned} (1/n^4) \sum_{t=1}^n S_t^2 &= (1/n) \sum_{t=1}^n \left(\int_0^{t/n+1/n} W_n(x) dx \right)^2 \approx \int_0^n \left(\int_0^{z/n+1/n} W_n(x) dx \right)^2 d(z/n) \\ &= \int_0^1 \left(\int_0^{y+1/n} W_n(x) dx \right)^2 dy \approx \int_0^1 \left(\int_0^y W_n(x) dx \right)^2 dy \end{aligned}$$

Thus, with B_n defined by (100),

$$B_n/n = \frac{\sum_{t=1}^n S_t^2}{n^2 \cdot \sum_{t=1}^n Y_t^2} = \frac{(1/n^4) \sum_{t=1}^n S_t^2}{(1/n^2) \sum_{t=1}^n Y_t^2} \approx \frac{\int_0^1 \left(\int_0^y W_n(x) dx \right)^2 dy}{\int_0^1 (W_n(x))^2} \rightarrow \frac{\int_0^1 \left(\int_0^y W(x) dx \right)^2 dy}{\int_0^1 (W(x))^2} \quad (169)$$

in distribution if $n \rightarrow \infty$.

On the other hand, if

$$Y_t = U_t, \text{ where } U_t \sim \text{i.i.d. } N(0, 1), \quad (170)$$

then $S_t = \sqrt{n} W_n(t/n)$, hence

$$\begin{aligned}
B_n &= \frac{\sum_{t=1}^n S_t^2}{n \sum_{t=1}^n Y_t^2} = \frac{n \sum_{t=1}^n (W_n(t/n))^2}{n \sum_{t=1}^n U_t^2} = \frac{(1/n) \sum_{t=1}^n (W_n(t/n))^2}{(1/n) \sum_{t=1}^n U_t^2} \approx \int_0^1 (W_n(x))^2 dx \\
&\rightarrow \int_0^1 (W(x))^2 dx
\end{aligned} \tag{171}$$

in distribution if $n \rightarrow \infty$. Consequently, $B_n/n \rightarrow 0$ in probability if $n \rightarrow \infty$.

THE CLASSICAL LINEAR REGRESSION MODEL

Herman J. Bierens

Pennsylvania State University

September 1, 2002

1. Introduction

The classical linear regression model takes the form

$$y_j = \theta_1 x_{1,j} + \dots + \theta_k x_{k,j} + u_j, \quad j = 1, \dots, n, \quad (1)$$

where y_j is the dependent variable, the $x_{i,j}$'s are the independent (or explanatory) variables, the u_j 's are unobservable error terms, n is the sample size, and the θ_i 's are the model parameters. If the model contains an intercept, then one of the $x_{i,j}$'s is equal to 1, say $x_{1,j}$.

Denoting

$$x_j = \begin{pmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{k,j} \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix}, \quad (2)$$

we can write model (1) more compactly as

$$y_j = x_j^T \theta + u_j, \quad j = 1, \dots, n \quad (3)$$

For the time being we assume:

ASSUMPTION 1: *The $x_{i,j}$'s are non-stochastic.*

ASSUMPTION 2: *The u_j 's are independent $N(0, \sigma^2)$ distributed.*

In particular Assumption 1 is very unrealistic for economic data, but we impose it for pedagogical reasons only. As will appear later on, we may without loss of generality assume that the $x_{i,j}$'s are

random (except for the one corresponding to the intercept). For example, we may replace Assumption 1 by the assumption that for each $t = 1, \dots, n$, u_t is independent of all the $x_{i,j}$'s.

Denoting:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{1,1} & \dots & x_{k,1} \\ x_{1,2} & \dots & x_{k,2} \\ \vdots & \dots & \vdots \\ x_{1,n} & \dots & x_{k,n} \end{pmatrix}, u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \quad (4)$$

we can cast model (1) in vector/matrix form as:

$$y = X\theta + u, \quad u \sim N_n(0, \sigma^2 I_n), \quad (5)$$

hence,

$$y \sim N_n(X\theta, \sigma^2 I_n), \quad (6)$$

because by Assumption 1, $X\theta$ is non-random.

2. Least squares estimation

For an arbitrary vector $\theta_* \in \mathbb{R}^k$ we have:

$$\begin{aligned} E((y - X\theta_*)^T(y - X\theta_*)) &= E((u + X\theta - X\theta_*)^T(u + X\theta - X\theta_*)) \\ &= E(u^T u - u^T X(\theta_* - \theta) - (\theta_* - \theta)^T X^T u + (\theta_* - \theta)^T X^T X(\theta_* - \theta)) \\ &= n\sigma^2 + (\theta_* - \theta)^T X^T X(\theta_* - \theta), \end{aligned} \quad (7)$$

which is minimal for $\theta_* = \theta$. This solution is unique if:

ASSUMPTION 3: The matrix $X^T X$ is nonsingular,

because then $X^T X$ is positive definite. This result motivates the least squares estimation of θ . *In general an estimator is a function of the data which serves as an approximation of a parameter or a parameter vector.* In particular, the least squares estimator $\hat{\theta}$ of θ is the solution of the minimization problem:

$$\begin{aligned} \min_{\hat{\theta}} (y - X\hat{\theta})^T(y - X\hat{\theta}) &= \min_{\hat{\theta}} (y^T y - \hat{\theta}^T X^T y - y^T X \hat{\theta} + \hat{\theta}^T X^T X \hat{\theta}) \\ &= \min_{\hat{\theta}} (y^T y - 2\hat{\theta}^T X^T y + \hat{\theta}^T X^T X \hat{\theta}). \end{aligned} \quad (8)$$

The first-order condition for the minimum involved is:

$$\frac{\partial (y^T y - 2\hat{\theta}^T X^T y + \hat{\theta}^T X^T X \hat{\theta})}{\partial \hat{\theta}^T} = -2X^T y + 2X^T X \hat{\theta} = 0, \quad (9)$$

hence:

$$\hat{\theta} = (X^T X)^{-1} X^T y. \quad (10)$$

Remarks: In (9) we have used the notation:

$$\frac{\partial f(x)}{\partial x^T} = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}, \text{ where } x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \text{ and } f(x) \text{ is a function of } x. \quad (11)$$

In particular, if $f(x) = a + x^T b + x^T C x$ where

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \quad C = \begin{pmatrix} c_{1,1} & \dots & c_{1,n} \\ \vdots & \dots & \vdots \\ c_{n,1} & \dots & c_{n,n} \end{pmatrix}, \text{ with } c_{ij} = c_{ji} \text{ (thus } C = C^T), \quad (12)$$

then

$$\begin{aligned}
\frac{\partial f(x)/\partial x_k}{\partial x_k} &= \frac{\partial \left(a + \sum_{i=1}^n b_i x_i + \sum_{i=1}^n \sum_{j=1}^n x_i c_{ij} x_j \right)}{\partial x_k} \\
&= \sum_{i=1}^n b_i \frac{\partial x_i}{\partial x_k} + \sum_{i=1}^n \sum_{j=1}^n \frac{\partial x_i c_{ij} x_j}{\partial x_k} = b_k + 2c_{k,k}x_k + \sum_{\substack{i=1 \\ i \neq k}}^n x_i c_{i,k} + \sum_{\substack{j=1 \\ j \neq k}}^n c_{k,j} x_j \\
&= b_k + 2 \sum_{j=1}^n c_{k,j} x_j, \quad k = 1, \dots, n,
\end{aligned} \tag{13}$$

hence

$$\frac{\partial f(x)}{\partial x^T} = b + 2Cx. \tag{14}$$

If C is not symmetric, we may without loss of generality replace C in the quadratic function $f(x)$ by the symmetric matrix $(C + C^T)/2$ (because $x^T C x = (x^T C x)^T = x^T C^T x$), so that then

$$\frac{\partial f(x)}{\partial x^T} = b + Cx + C^T x. \tag{15}$$

3. Properties of the least squares estimator for fixed sample size

Substituting (5) in (10) yields:

$$\hat{\theta} = (X^T X)^{-1} X^T (X\theta + u) = \theta + (X^T X)^{-1} X^T u. \tag{16}$$

Since u is multivariate normally distributed and X is assumed to be nonstochastic, it follows that $\hat{\theta}$ is k -variate normally distributed with expectation

$$E(\hat{\theta}) = \theta + E((X^T X)^{-1} X^T u) = \theta + (X^T X)^{-1} X^T E(u) = \theta \tag{17}$$

(hence $\hat{\theta}$ is an *unbiased* estimator) and variance matrix

$$\begin{aligned}
E((\hat{\theta} - \theta)(\hat{\theta} - \theta)^T) &= E((X^T X)^{-1} X^T u u^T X (X^T X)^{-1}) = (X^T X)^{-1} X^T (E(u u^T)) X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}.
\end{aligned} \tag{18}$$

Moreover, the least squares estimator is the *best linear unbiased estimator* (BLUE), in the sense that for all estimators of the form $\hat{\theta}_* = Cy$, where C is a $k \times n$ matrix such that $E(\hat{\theta}_*) = \theta$, we have that

$$Var(\hat{\theta}_*) = E[(\hat{\theta}_* - \theta)(\hat{\theta}_* - \theta)^T] = \sigma^2 (X^T X)^{-1} + D, \text{ where } D \text{ is a positive semi-definite matrix.}$$

The proof of this proposition is quite easy. First, observe that the unbiasedness condition implies that $CX = I_k$, hence $\hat{\theta}_* = C(X\theta + u) = \theta + Cu$, and thus $Var(\hat{\theta}_*) = \sigma^2 CC^T$. Now

$$\begin{aligned}
D &= \sigma^2 [CC^T - (X^T X)^{-1}] = \sigma^2 [CC^T - CX(X^T X)^{-1} X^T C^T] \\
&= \sigma^2 C[I_k - X(X^T X)^{-1} X^T]C^T = \sigma^2 CMC^T,
\end{aligned} \tag{19}$$

say, where the second equality follows from the unbiasedness condition $CX = I_k$. The matrix

$$M = I_n - X(X^T X)^{-1} X^T \tag{20}$$

is idempotent:

$$\begin{aligned}
M^2 &= (I_n - X(X^T X)^{-1} X^T)(I_n - X(X^T X)^{-1} X^T) = I_n - 2X(X^T X)^{-1} X^T \\
&\quad + X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = I_n - X(X^T X)^{-1} X^T = M,
\end{aligned} \tag{21}$$

hence its eigenvalues are either 1 or 0. Since all the eigenvalues are non-negative, M is positive semi-definite, and so is CMC^T . Thus we have:

THEOREM 1: Under Assumptions 1-3, $\hat{\theta} - \theta \sim N_k(0, \sigma^2 (X^T X)^{-1})$, and $\hat{\theta}$ is BLUE.

The latter result is known as the Gauss-Markov theorem.

4. *Estimation of the error variance*

Since by (1),

$$\min_{\theta_*} \frac{1}{n} E \left((y - X\theta_*)^T (y - X\theta_*) \right) = \sigma^2, \quad (22)$$

it seems at first sight a good idea (but not at second sight as will appear) to estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\theta})^T (y - X\hat{\theta}) = \frac{1}{n} \sum_{j=1}^n (y_j - x_j^T \hat{\theta})^2 = \frac{1}{n} \sum_{j=1}^n \hat{u}_j^2, \quad (23)$$

where $\hat{u}_j = y_j - x_j^T \hat{\theta}$ is called the residual of y_j (i.e., the part of y_j that is left over after accounting for

the effect of x_j). The quadratic form involved is called the Residual Sum of Squares (RSS):

$$RSS = (y - X\hat{\theta})^T (y - X\hat{\theta}) = \sum_{j=1}^n \hat{u}_j^2. \quad (24)$$

This is sometimes also referred to as the Sum of Squared Residuals (SSR).

Note that the RSS can be computed without computing each of the \hat{u}_j 's separately, as follows:

$$RSS = y^T y - y^T X \hat{\theta} - \hat{\theta}^T X^T y + \hat{\theta}^T X^T X \hat{\theta} = y^T y - \hat{\theta}^T X^T y = y^T y - \hat{\theta}^T X^T X \hat{\theta}. \quad (25)$$

(See Exercise 1) Substituting (5) and (16) in (24) yield:

$$\begin{aligned} RSS &= (u + X\theta - X\hat{\theta})^T (u + X\theta - X\hat{\theta}) = (u - X(\hat{\theta} - \theta))^T (u - X(\hat{\theta} - \theta)) \\ &= (u - X(X^T X)^{-1} X^T u)^T (u - (X^T X)^{-1} X^T u) = u^T M^2 u, \end{aligned} \quad (26)$$

where M is defined by (20). As shown in (21), M is idempotent, hence

$$\begin{aligned} \text{rank}(M) &= \text{trace}(M) = \text{trace}(I_n - X(X^T X)^{-1} X^T) = \text{trace}(I_n) - \text{trace}(X(X^T X)^{-1} X) \\ &= \text{trace}(I_n) - \text{trace}((X^T X)^{-1} X^T X) = \text{trace}(I_n) - \text{trace}(I_k) = n - k. \end{aligned} \quad (27)$$

Thus:

$$RSS = u^T M u \quad (28)$$

and consequently, by one of the results for the multivariate normal distribution, hereafter indicated as “an *MND* result” (which one?),

THEOREM 2: Under Assumptions 1-3, $RSS/\sigma^2 \sim \chi_{n-k}^2$.

This result implies that

$$E[\hat{\sigma}^2] = (n-k)\sigma^2/n, \quad (29)$$

hence $\hat{\sigma}^2$ is a *biased* estimator. However, the following correction yields an *unbiased* estimator:

$$s^2 = \frac{1}{n-k}(y - X\hat{\theta})^T(y - X\hat{\theta}). \quad (30)$$

Next we show that s^2 and $\hat{\theta}$ are independent, by showing that $(X^T X)^{-1} X^T u$ and $u^T M u$ are independent. A necessary and sufficient condition for the latter is that

$$(X^T X)^{-1} X^T M = O. \quad (31)$$

Condition (31) follows from:

$$\begin{aligned} (X^T X)^{-1} X^T (I - X(X^T X)^{-1} X^T) &= (X^T X)^{-1} X^T - (X^T X)^{-1} X^T X (X^T X)^{-1} X^T \\ &= (X^T X)^{-1} X^T - (X^T X)^{-1} X^T = O. \end{aligned} \quad (32)$$

Summarizing, we have shown that:

THEOREM 3: Under Assumptions 1-3, $E(s^2) = \sigma^2$, $(n-k)s^2/\sigma^2 \sim \chi_{n-k}^2$, and s^2 and $\hat{\theta}$

are independent.

5. The t-test.

Suppose we want to test the null hypothesis that the i -th component θ_i of θ equals zero:

$$H_0: \theta_i = 0, \quad (33)$$

which amounts to the hypothesis that the corresponding variable $x_{i,j}$ can be deleted from model (1), against the alternative hypothesis

$$H_1: \theta_i \neq 0. \quad (34)$$

The general procedure for testing statistical hypothesis is to construct a function of the data, called test statistic, that has under the null hypothesis a particular distribution, and under the alternative hypothesis a distribution that differs from the one under the null hypothesis.

In order to construct a test statistic for the null hypothesis (33), we first isolate θ_i and its least squares estimator $\hat{\theta}_i$ from θ and $\hat{\theta}$, respectively, by taking the linear transformations

$$\theta_i = e_i^T \theta, \quad \hat{\theta}_i = e_i^T \hat{\theta} \quad (35)$$

where e_i is the i -th column of the $k \times k$ unit matrix I_k . Thus, e_i is a k -vector of zeros, except for the i -the component, which equals 1. Then it follows from Theorem 1, together with an MND result (*Exercise: Which one?*), that

$$\hat{\theta}_i - \theta_i = e_i^T (\hat{\theta} - \theta) \sim N(0, \sigma^2 e_i^T (X^T X)^{-1} e_i), \quad (36)$$

hence

$$\frac{\hat{\theta}_i - \theta_i}{\sigma \sqrt{e_i^T (X^T X)^{-1} e_i}} \sim N(0, 1). \quad (37)$$

Note that $e_i^T (X^T X)^{-1} e_i$ is just the i -th diagonal element of $(X^T X)^{-1}$. Using the definition of the t -distribution, and the result in Theorem 3, it is now easy to verify that

THEOREM 4: Under Assumptions 1-3,

$$\frac{\hat{\theta}_i - \theta_i}{s\sqrt{e_i^T(X^T X)^{-1}e_i}} \sim t_{n-k}. \quad (38)$$

Proof: Exercise.

The denominator in the left-hand side expression involved is called the *standard error* (*se*) of $\hat{\theta}_i$:

$$se(\hat{\theta}_i) = s\sqrt{e_i^T(X^T X)^{-1}e_i}. \quad (39)$$

Now under the null hypothesis (33) we have

$$\hat{t}_i = \frac{\hat{\theta}_i}{s\sqrt{e_i^T(X^T X)^{-1}e_i}} \sim t_{n-k}. \quad (40)$$

The statistic \hat{t}_i is called the *t-value* of $\hat{\theta}_i$, and will be our test statistic. Most econometric software packages report either the standard error, or the t-value, or both, for each least squares estimate $\hat{\theta}_i$.

Before we turn to the actual testing procedure, we need to pay attention to what happens with the t-value if the null hypothesis is false, i.e., if (34) is true. Then

$$\hat{t}_i = \frac{\hat{\theta}_i - \theta_i}{s\sqrt{e_i^T(X^T X)^{-1}e_i}} + \frac{\theta_i}{s\sqrt{e_i^T(X^T X)^{-1}e_i}}. \quad (41)$$

The first term in this expression is *t* distributed with $n - k$ degrees of freedom, which converges in distribution to the $N(0,1)$ distribution if n increases to infinity:

$$\frac{\hat{\theta}_i - \theta_i}{s\sqrt{e_i^T(X^T X)^{-1}e_i}} \rightarrow N(0,1) \text{ in distr. if } n \rightarrow \infty. \quad (42)$$

Moreover,

$$\operatorname{plim}_{n \rightarrow \infty} s^2 = \sigma^2, \quad (43)$$

hence

$$\operatorname{plim}_{n \rightarrow \infty} s = \sigma. \quad (44)$$

Next, assume that

ASSUMPTION 4: $\lim_{n \rightarrow \infty} (1/n) \sum_{j=1}^n x_j x_j^T = Q$, where Q is a finite positive definite matrix.

Then $\lim_{n \rightarrow \infty} X^T X / n = Q$, hence $\lim_{n \rightarrow \infty} n(X^T X)^{-1} = Q^{-1}$ and thus

$$\lim_{n \rightarrow \infty} \sqrt{n} \sqrt{e_i^T (X^T X)^{-1} e_i} = \sqrt{e_i^T Q^{-1} e_i} > 0. \quad (45)$$

It follows now from (40), (42), (44) and (45) that

THEOREM 5: Under Assumptions 1-3 and the null hypothesis (33), $\hat{t}_i \sim t_{n-k}$, whereas under Assumptions 1-4 and the alternative hypothesis (34), $\operatorname{plim}_{n \rightarrow \infty} \hat{t}_i / \sqrt{n} = \theta_i / \sqrt{\sigma^2 e_i^T Q^{-1} e_i}$.

Proof: Exercise.

Thus, under the alternative (34) we have for an arbitrary large positive number K ,

$$\lim_{n \rightarrow \infty} P(\hat{t}_i > K) = 1 \text{ if } \theta_i > 0, \quad \lim_{n \rightarrow \infty} P(\hat{t}_i < -K) = 1 \text{ if } \theta_i < 0. \quad (46)$$

This result suggests a decision rule where we accept the null hypothesis (33) if for an a priori chosen constant $K > 0$, $|\hat{t}_i| \leq K$, and we reject the null hypothesis (33) in favor of the alternative hypothesis (34) if $|\hat{t}_i| > K$. Of course it is possible that the correct null hypothesis is rejected, with probability

$\alpha = P(|t_{n-k}| > K)$. This probability α is called the *Type I error*, or the *significance level*, and can be controlled by choosing K such that $P(|t_{n-k}| > K) = \alpha$ for a given value of α , using the table of the t distribution. Traditional values for the significance level are $\alpha = 0.05$ (5% significance level) and $\alpha = 0.1$ (10% significance level).

The *power* function of the test is:

$$\phi_n(\theta_i) = P(|\hat{t}_i| > K), \quad (47)$$

where K , called *critical value*, is chosen such that $P(|t_{n-k}| > K) = \alpha$ for a given significance level α . The value of the power function under the alternative hypothesis (34) is called the *power* of the test, and 1 minus the power is called the *Type II error*, which is the probability of not rejecting the null hypothesis if the alternative is true.

The t-test discussed so far is a two-sided test, because under the alternative both $\theta_i > 0$ and $\theta_i < 0$ are possible. If the latter is not possible, and the choice is between the null hypothesis $\theta_i = 0$ against the alternative $\theta_i > 0$, we should conduct a one-sided test: Choose for given significance level α the critical value K such that $P(t_{n-k} > K) = \alpha$. Then accept the null hypothesis $\theta_i = 0$ if $\hat{t}_i \leq K$ and reject the null in favor of the alternative $\theta_i > 0$ if $\hat{t}_i > K$. For the case of the alternative $\theta_i < 0$ we accept the null if $\hat{t}_i \geq -K$ and we reject the null in favor of the alternative if $\hat{t}_i < -K$.

6. The F-test

We now consider testing of a null hypothesis of the form

$$H_0: R\theta = q, \quad (48)$$

where R is a given $r \times k$ matrix with rank r , and q is $r \times 1$ vector of given constants. The alternative we are going to consider is the alternative that this null hypothesis is false.

It follows from Theorem 1, together with an MND result (*Exercise: Which one?*), that

$$R(\hat{\theta} - \theta) \sim N(0, \sigma^2 R(X^T X)^{-1} R^T), \quad (49)$$

hence it follows from an MND result (which one?) that

$$\frac{(\hat{\theta} - \theta)^T R^T (R(X^T X)^{-1} R^T)^{-1} R(\hat{\theta} - \theta)}{\sigma^2} \sim \chi_r^2. \quad (50)$$

Combining this result with the results of Theorem 3, it follows from the definition of the F distribution that

$$\frac{(\hat{\theta} - \theta)^T R^T (R(X^T X)^{-1} R^T)^{-1} R(\hat{\theta} - \theta)/r}{s^2} \sim F_{r, n-k}. \quad (51)$$

Thus under the null hypothesis (48) we have:

$$\hat{F} = \frac{(R\hat{\theta} - q)^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\theta} - q)/r}{s^2} \sim F_{r, n-k}, \quad (52)$$

which is the test statistic of the F-test under review.

If the null hypothesis (48) is false, we have

$$\begin{aligned} \hat{F} &= \frac{(\hat{\theta} - \theta)^T R^T (R(X^T X)^{-1} R^T)^{-1} R(\hat{\theta} - \theta)/r}{s^2} \\ &+ 2 \frac{(R\theta - q)^T (R(X^T X)^{-1} R^T)^{-1} R(\hat{\theta} - \theta)/r}{s^2} \\ &+ \frac{(R\theta - q)^T (R(X^T X)^{-1} R^T)^{-1} (R\theta - q)/r}{s^2}, \end{aligned} \quad (53)$$

hence:

THEOREM 6: Under Assumptions 1-3 and the null hypothesis (48), $\hat{F} \sim F_{r, n-k}$. If the null hypothesis (48) is false then under Assumptions 1-4,

$$\operatorname{plim}_{n \rightarrow \infty} \hat{F}/n = \frac{(R\theta - q)^T \{RQ^{-1}R^T\}^{-1}(R\theta - q)/r}{\sigma^2} > 0. \quad (54)$$

Proof: Exercise.

This result suggests the following one-sided test: Given a significance level α , look up in the table of the F distribution the critical value K for which $P(F_{r,n-k} > K) = \alpha$. Then accept the null hypothesis (48) if $\hat{F} \leq K$, and reject it if $\hat{F} > K$.

In practice the F-test is conducted differently, as follows: Implement the null hypothesis (48), and re-estimate the regression model (5) with the parameter vector θ reparametrized such that the condition $R\theta = q$ holds. This can be done by augmenting the matrix R with a $(k-r) \times k$ matrix R_* such that

$$\begin{pmatrix} R_* \\ R \end{pmatrix} \theta = \begin{pmatrix} \beta \\ q \end{pmatrix}, \text{ where } \begin{pmatrix} R_* \\ R \end{pmatrix} \text{ is nonsingular.} \quad (55)$$

Then

$$\theta = \begin{pmatrix} R_* \\ R \end{pmatrix}^{-1} \begin{pmatrix} \beta \\ q \end{pmatrix} = R_1 \beta + R_2 q, \quad (56)$$

say. Substituting (56) in model (3) yields the restricted model

$$y_j - x_j^T R_2 q = x_j^T R_1 \beta + u_j, \quad j = 1, \dots, n. \quad (57)$$

Now estimate the parameter vector β by least squares, and compute the Residual Sum of Squares RSS_0 involved. Then

THEOREM 7: The test statistic \hat{F} defined by (52) is equal to

$$\hat{F} = \frac{(RSS_0 - RSS)/r}{RSS/(n-k)}. \quad (58)$$

Proof: Note that

$$RSS_0 = \min_{R\hat{\theta}_0 = q} (y - X\hat{\theta}_0)^T(y - X\hat{\theta}_0). \quad (59)$$

The Lagrange function for this minimization problem is:

$$\mathcal{L} = y^T y - 2\hat{\theta}_0^T X^T y + \hat{\theta}_0^T X^T X \hat{\theta}_0 - 2(R\hat{\theta}_0 - q)^T \lambda. \quad (60)$$

The first-order conditions for the minimum are:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{\theta}_0^T} &= -2X^T y + 2X^T X \hat{\theta}_0 - 2R^T \lambda = 0 \Rightarrow \hat{\theta}_0 = (X^T X)^{-1} X^T y + (X^T X)^{-1} R^T \lambda, \\ \frac{\partial \mathcal{L}}{\partial \lambda^T} &= -2(R\hat{\theta}_0 - q) = 0 \Rightarrow R\hat{\theta}_0 = q. \end{aligned} \quad (61)$$

Thus,

$$\hat{\theta}_0 = \hat{\theta} + (X^T X)^{-1} R^T \lambda, \quad (62)$$

and

$$q = R\hat{\theta}_0 = R\hat{\theta} + R(X^T X)^{-1} R^T \lambda, \quad (63)$$

hence

$$\lambda = -(R(X^T X)^{-1} R^T)^{-1} (R\hat{\theta} - q). \quad (64)$$

Substituting (64) in (62) yields:

$$\hat{\theta}_0 = \hat{\theta} - (X^T X)^{-1} R^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\theta} - q). \quad (65)$$

Therefore

$$\begin{aligned}
RSS_0 &= \left(y - X\hat{\theta} - X(X^T X)^{-1} R^T \{R(X^T X)^{-1} R^T\}^{-1} (R\hat{\theta} - q) \right)^T \\
&\quad \times \left(y - X\hat{\theta} - X(X^T X)^{-1} R^T \{R(X^T X)^{-1} R^T\}^{-1} (R\hat{\theta} - q) \right) \\
&= (y - X\hat{\theta})^T (y - X\hat{\theta}) - 2(y - X\hat{\theta})^T \{X(X^T X)^{-1} R^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\theta} - q)\} \\
&\quad + \{X(X^T X)^{-1} R^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\theta} - q)\}^T \{X(X^T X)^{-1} R^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\theta} - q)\}
\end{aligned} \tag{66}$$

Since

$$(y - X\hat{\theta})^T X = 0 \tag{67}$$

(why?), this expression simplifies to:

$$RSS_0 = RSS + (R\hat{\theta} - q)^T \{R(X^T X)^{-1} R^T\}^{-1} (R\hat{\theta} - q). \tag{68}$$

Using the fact that $s^2 = RSS/(n-k)$, the theorem follows. Q.E.D.

Some econometric software packages automatically report the F-statistic. This statistic is the test statistic of the F-test that all the slope coefficients are zero, in a linear regression model with an intercept. Thus let the model be as in (1), with $x_{1,j} = 1$. Then the null hypothesis involved is that

$$H_0: \theta_2 = \theta_3 = \dots = \theta_k = 0. \tag{69}$$

This null hypothesis is equivalent to the hypothesis that the model can be simplified to

$$y_j = \theta_1 + u_j. \tag{70}$$

As is easy to verify, the least squares estimator $\hat{\theta}_1$ of the parameter θ_1 in model (70) is just the sample mean of the y_j 's:

$$\hat{\theta}_1 = \bar{y} = (1/n) \sum_{j=1}^n y_j \tag{71}$$

and the RSS of model (70) is also called the Total Sum of Squares (TSS):

$$TSS = \sum_{j=1}^n (y_j - \bar{y})^2. \quad (72)$$

Thus, the F-test involved is:

$$\hat{F} = \frac{(TSS - RSS)/(k-1)}{RSS/(n-k)}, \quad (73)$$

which under the null hypothesis (69) has an $F_{k-1,n-k}$ distribution.

7. The R-square

The R^2 statistic compares the RSS of the model (1) with intercept (thus $x_{1,j} = 1$) with the RSS (= TSS) of model (70):

$$R^2 = 1 - \frac{RSS}{TSS}. \quad (74)$$

It is easy to verify that $0 \leq R^2 \leq 1$, where the value $R^2 = 0$ corresponds to $RSS = TSS$, hence the explanatory variables $x_{i,j}$, $i = 2, \dots, k$, in model (1) do not contribute anything at all to the explanation of y_j . The case $R^2 = 1$ corresponds to $RSS = 0$, hence $\text{Var}(u_j) = 0$. The model then fits without error. The R^2 is related to the F statistic (73):

$$\hat{F} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}. \quad (75)$$

8 Relaxing the assumption of non-stochastic regressors

As said before, Assumption 1 is too restrictive for economic data. However, we may replace Assumptions 1-4 by the following assumptions:

ASSUMPTION 1^(*): Conditionally on the random variables in the matrix X , the error vector u in model (5) satisfies $u \sim N_n(0, I_n)$.

ASSUMPTION 2^(*): $P(\det(X^T X) > 0) = 1$.

ASSUMPTION 3^(*): $\text{plim}_{n \rightarrow \infty} X^T X / n = Q$, where Q is a positive definite matrix.

Then:

THEOREM 8: With Assumptions 1-3 replaced by Assumptions 1^(*) and 2^(*), and Assumption 4 by Assumption 3^(*), Theorem 1 now holds conditionally on the random variables in the matrix X , and Theorems 2-7 hold unconditionally.

Proof: Exercise.

Hints: It is easy to verify that all the previous results go through conditionally on X . In order to prove that s^2 and $\hat{\theta}$ are still independent (cf. Theorem 3), observe that the joint density of s^2 and $\hat{\theta}$ conditional on X is now the product of the conditional density of s^2 and the conditional density of $\hat{\theta}$, given X . But the conditional density of s^2 does not depend on X , because $(n-k)s^2/\sigma^2 \sim \chi_{n-k}^2$ conditionally on X and therefore also unconditionally (why?). Integrating X out in the joint conditional density of s^2 and $\hat{\theta}$ then yields the product of the unconditional density of s^2 and the unconditional density of $\hat{\theta}$, which proves the result involved. The rest of the conclusion of Theorem 8 can be proved by a similar argument.

A more general version of the above argument is stated in the following lemma:

LEMMA 1: Let x , y and z be random vector or variables such that y and z are

conditionally independent, relative to x , i.e., the joint conditional distribution function of y and z , given x , is the product of the conditional distribution function of y and the conditional distribution function of z , given x . If z and x are independent then y and z are independent.

Proof: For convenience assume that the joint distribution of x , y and z is continuous, with marginal densities $f_x(x)$, $f_y(y)$ and $f_z(z)$, and conditional densities $f_{yz}(y,z|x)$, $f_y(y|x)$ and $f_z(z|x)$. Since z and x are independent we have $f_z(z|x) = f_z(z)$. Now

$$\begin{aligned} f_{yz}(y,z) &= \int f_{yz}(y,z|x)f_x(x)dx = \int f_y(y|x)f_z(z|x)f_x(x)dx = \int f_y(y|x)f_x(x)dx f_z(z) \\ &= f_y(y)f_z(z). \end{aligned} \tag{76}$$

This result, however, also holds without the assumption that x , y and z are continuously distributed.

9. Large sample theory without the normality assumption

If our sample size n is large, we may even get rid of the normality assumption on the errors u_j in model (1). Assume that the source of the data $\{(y_j, x_j), j = 1, \dots, n\}$ is a random sample

ASSUMPTION 1^():** *The random vectors $(y_j, x_j^T)^T$, $j = 1, \dots, n, \dots$, (or the sub-vectors of random variables if one of the components of x_j equals 1) are i.i.d. Moreover, $E(y_j|x_j) = x_j^T\theta$, hence $E(u_j|x_j) = 0$. Furthermore, $Var(u_j|x_j) = \sigma^2 < \infty$, and $E(x_j x_j^T) = Q$, where Q is the same positive definite matrix as in Assumption 3^(*).*

Then:

THEOREM 9: *Under Assumptions 1^(**) and 2^(*) we have:*

$$E(\hat{\theta}) = \theta, \quad E(s^2) = \sigma^2. \tag{77}$$

Conditional on X , the least squares estimator $\hat{\theta}$ is BLUE. Moreover, under the additional

Assumption 3^()* we have:

$$\operatorname{plim}_{n \rightarrow \infty} \hat{\theta} = \theta, \quad \operatorname{plim}_{n \rightarrow \infty} s^2 = \sigma^2, \quad (78)$$

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N_k(0, \sigma^2 Q^{-1}) \text{ in distr. as } n \rightarrow \infty, \quad (79)$$

the *t*-value \hat{t}_i of the *i*-th component of $\hat{\theta}$ satisfies

- (a) if $\theta_i = 0$ then $\hat{t}_i \rightarrow N(0,1)$ in distr as $n \rightarrow \infty$,
- (b) if $\theta_i \neq 0$ then $\operatorname{plim}_{n \rightarrow \infty} \hat{t}_i / \sqrt{n} = \theta_i / \sqrt{\sigma^2 e_i^T Q^{-1} e_i} \neq 0$,

and the *F*-statistic \hat{F} for testing the null hypothesis $R\theta = q$, where R is a $r \times k$ matrix with rank r , satisfies

- (a) if $R\theta = q$ then $r\hat{F} \rightarrow \chi_r^2$ in distr as $n \rightarrow \infty$,
- (b) if $R\theta \neq q$ then $\operatorname{plim}_{n \rightarrow \infty} r\hat{F}/n = (R\theta - q)^T [RQ^{-1}R^T]^{-1}(R\theta - q)/\sigma^2 > 0$.

Proof: Assumption 1^(**) implies that $E(u|X) = 0$ and $E(uu^T|X) = \sigma^2 I_n$, which in their turn

imply the unbiasedness of the two estimators (*Proof:* Exercise). Moreover, Assumption 1^(**) and the law of large numbers imply that $\operatorname{plim}_{n \rightarrow \infty} (1/n)X^T u = 0$. Together with Assumptions 2^(*) and 3^(*) this

result implies (78). Furthermore, Assumption 1^(**) and the central limit theorem imply that

$(1/\sqrt{n})X^T u \rightarrow N_k(0, \sigma^2 Q)$ in distribution (*Proof:* Exercise). Together with Assumption 3^(*) this result

implies (79). The rest of Theorem 9 is easy to prove.

10. Tests of structural change: The Chow tests

The classical linear regression model (1) assumes that the parameters are the same for all observations. In order to test this crucial hypothesis, the sample is split in say m subsamples of sizes

n_1, \dots, n_m , respectively, with $n = \sum_{i=1}^m n_i$ and $n_i > k$, where the parameters are allowed to be different across subsamples:

$$\begin{aligned}
y_j &= \theta_{1,1}x_{1,j} + \dots + \theta_{1,k}x_{k,j} + u_j, \quad j = 1, \dots, n_1, \\
y_j &= \theta_{2,1}x_{1,j} + \dots + \theta_{2,k}x_{k,j} + u_j, \quad j = n_1 + 1, \dots, n_1 + n_2 \\
&\dots \\
y_j &= \theta_{m,1}x_{1,j} + \dots + \theta_{m,k}x_{k,j} + u_j, \quad j = \sum_{i=1}^{m-1} n_i + 1, \dots, n.
\end{aligned} \tag{82}$$

For each subsample i we can write the model in matrix form as

$$y^{(i)} = X^{(i)}\theta^{(i)} + u^{(i)}, \quad i = 1, \dots, m, \tag{83}$$

and stacking all these equations yields the *unrestricted model*:

$$\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix} = \begin{pmatrix} X^{(1)} & O & .. & O \\ O & X^{(2)} & .. & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & .. & X^{(m)} \end{pmatrix} \begin{pmatrix} \theta^{(1)} \\ \theta^{(2)} \\ \vdots \\ \theta^{(m)} \end{pmatrix} + \begin{pmatrix} u^{(1)} \\ u^{(2)} \\ \vdots \\ u^{(m)} \end{pmatrix}. \tag{84}$$

Note that the total number of parameters is $k \times m$. It is not hard to verify that the least squares estimators $\hat{\theta}^{(i)}$ of the parameter vectors $\theta^{(i)}$ are:

$$\hat{\theta}^{(i)} = (X^{(i)T}X^{(i)})^{-1}X^{(i)T}y^{(i)} \tag{85}$$

and that the RSS_u of model (84) is just the sum of the residual sums of squares RSS_i of the regressions (83):

$$RSS_u = \sum_{i=1}^m RSS_i \tag{86}$$

The null hypothesis that model (3) is correct corresponds to the set of hypotheses

$$\begin{aligned}
& \theta^{(1)} = \theta^{(2)} \\
H_0: & \quad \theta^{(2)} = \theta^{(3)} \\
& \quad \vdots \\
& \theta^{(m-1)} = \theta^{(m)}
\end{aligned} \tag{87}$$

hence the total number of restrictions is $(m-1) \times k$. Therefore:

THEOREM 10: (Chow test) The F test of the hypothesis (87) is:

$$\hat{F} = \frac{(RSS - \sum_{i=1}^m RSS_i)/((m-1)k)}{(\sum_{i=1}^m RSS_i)/(n-km)} \tag{88}$$

where RSS is the residual sum of squares of the restricted model (3). Under the null hypothesis involved this F statistic is $F_{(m-1)k, n-km}$ distributed.

If one of the subsample sizes n_i is less or equal to the number k of parameters, it is possible to fit the corresponding regression (83) without any residuals, so that $RSS_i = 0$ (why?). This case, with $m = 2$, may occur for example in a situation where the model is used for prediction and we want to test whether the predicted dependent variable y_n is governed by the same model as for the observations $j \leq n-1$. In that case the degrees of freedom of the F test will be different, as we will demonstrate now for the case $m = 2$. Thus, we want to test the validity of the null model (1) against the alternative model

$$\begin{aligned}
y_j &= \theta_{1,1}x_{1,j} + \dots + \theta_{1,k}x_{k,j} + u_j, \quad j = 1, \dots, n_1, \\
y_j &= \theta_{2,1}x_{1,j} + \dots + \theta_{2,k}x_{k,j} + u_j, \quad j = n_1 + 1, \dots, n,
\end{aligned} \tag{89}$$

where

$$n_2 = n - n_1 \leq k. \tag{90}$$

Since in this case $RSS_2 = 0$, the F test now compares the RSS of the restricted model (1) for the full sample of size n , with the RSS_1 of the regression for the larger subsample of size n_1 :

THEOREM 11: (*Chow predictive test*) The F test of the null hypothesis that model (1) applies to all observations against the alternative that model (89) with (90) applies, takes the form

$$\hat{F} = \frac{(RSS - RSS_1)/n_2}{RSS_1/(n_1 - k)}, \quad (91)$$

which under the null hypothesis has an $F_{n_2, n_1 - k}$ distribution.

Proof: Adopting the matrix notation (83) of the models for the two subsamples, we can write

$$\begin{aligned} RSS &= u^T \left(I - X(X^T X)^{-1} X^T \right) u \\ &= u^T \left(\begin{pmatrix} I_{n_1} & O \\ O & I_{n_2} \end{pmatrix} - \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} (X^T X)^{-1} (X^{(1)T}, X^{(2)T}) \right) u \\ &= u^T \left(\begin{pmatrix} I_{n_1} & O \\ O & I_{n_2} \end{pmatrix} - \begin{pmatrix} X^{(1)} (X^T X)^{-1} X^{(1)T} & X^{(1)} (X^T X)^{-1} X^{(2)T} \\ X^{(2)} (X^T X)^{-1} X^{(1)T} & X^{(2)} (X^T X)^{-1} X^{(2)T} \end{pmatrix} \right) u = u^T M u, \end{aligned} \quad (92)$$

say, and

$$\begin{aligned} RSS_1 &= u^{(1)T} \left(I_{n_1} - X^{(1)} (X^{(1)T} X^{(1)})^{-1} X^{(1)T} \right) u^{(1)} \\ &= u^T \left(\begin{pmatrix} I_{n_1} & O \\ O & I_{n_2} \end{pmatrix} - \begin{pmatrix} X^{(1)} (X^{(1)T} X^{(1)})^{-1} X^{(1)T} & O \\ O & I_{n_2} \end{pmatrix} \right) u = u^T M_1 u, \end{aligned} \quad (93)$$

say, hence

$$\begin{aligned}
& RSS - RSS_1 \\
&= u^T \begin{pmatrix} X^{(1)}(X^{(1)T}X^{(1)})^{-1}X^{(1)T} & O \\ O & I_{n_2} \end{pmatrix} u - u^T \begin{pmatrix} X^{(1)}(X^TX)^{-1}X^{(1)T} & X^{(1)}(X^TX)^{-1}X^{(2)T} \\ X^{(2)}(X^TX)^{-1}X^{(1)T} & X^{(2)}(X^TX)^{-1}X^{(2)T} \end{pmatrix} u \quad (94) \\
&= u^T M_2 u,
\end{aligned}$$

say. Clearly, the matrix M_1 is idempotent, with $\text{rank}(M_1) = \text{trace}(M_1) = n_1 - k$. Also M_2 is idempotent, because

$$\begin{aligned}
& \begin{pmatrix} X^{(1)}(X^{(1)T}X^{(1)})^{-1}X^{(1)T} & O \\ O & I_{n_2} \end{pmatrix} \begin{pmatrix} X^{(1)}(X^TX)^{-1}X^{(1)T} & X^{(1)}(X^TX)^{-1}X^{(2)T} \\ X^{(2)}(X^TX)^{-1}X^{(1)T} & X^{(2)}(X^TX)^{-1}X^{(2)T} \end{pmatrix} \\
&= \begin{pmatrix} X^{(1)}(X^TX)^{-1}X^{(1)T} & X^{(1)}(X^TX)^{-1}X^{(2)T} \\ X^{(2)}(X^TX)^{-1}X^{(1)T} & X^{(2)}(X^TX)^{-1}X^{(2)T} \end{pmatrix} \quad (95)
\end{aligned}$$

hence $M_1 M = M M_1 = M_1$ and thus $(M - M_1)^2 = M - M_1$. The rank of M_2 is: $\text{rank}(M_2) = \text{trace}(M_2) = \text{trace}(M) - \text{trace}(M_1) = (n - k) - (n_1 - k) = n - n_1 = n_2$. Using an MND result (which one?), it now follows that under the null hypothesis

$$(RSS - RSS_1)/\sigma^2 \sim \chi^2_{n_2}, \quad RSS_1/\sigma^2 \sim \chi^2_{n_1-k}. \quad (96)$$

Finally it remains to show that the two random variables involved are independent. This follows from the fact that $M_1 M_2 = M_1(M - M_1) = M_1 M - M_1 M_1 = M_1 - M_1 = O$, and an MND result (which one?). Q.E.D.

11. Tests of partial structural change

If some but not all of the parameters are allowed to be different across subsamples, the F tests in section 10 no longer apply. We then have to merge the different models by using dummy variables. To illustrate this, consider the bivariate regressions

$$\begin{aligned} y_j &= \theta_{1,1} + \theta_{2,1}x_j + u_j, \quad j = 1, \dots, n_1 \\ y_j &= \theta_{1,2} + \theta_{2,2}x_j + u_j, \quad j = n_1 + 1, \dots, n \end{aligned} \tag{97}$$

Consider first the case with maintained hypothesis

Case 1: $\theta_{2,1} = \theta_{2,2} = \theta_2$

(A maintained hypothesis is a hypothesis that is assumed to hold under both the null and alternative hypothesis). Rewriting $\theta_{1,1} = \theta_1$, $\theta_{1,2} = \theta_1 + \theta_3$, and creating the dummy variable

$$d_j = 0 \text{ for } j = 1, \dots, n_1, \quad d_j = 1 \text{ for } j = n_1 + 1, \dots, n, \tag{98}$$

we can now merge the two models involved into a single reparametrized model:

$$y_j = \theta_1 + \theta_2x_j + \theta_3d_j + u_j, \quad j = 1, \dots, n, \tag{99}$$

and the null hypothesis that the two models are the same is now equivalent to the hypothesis $\theta_3 = 0$,

which can be tested by the t-test.

The case with maintained hypothesis

Case 2: $\theta_{1,1} = \theta_{1,2} = \theta_1$

is similar. Rewriting $\theta_{2,1} = \theta_2$, $\theta_{2,2} = \theta_2 + \theta_3$, we can merge the two models involved into a single model:

$$y_j = \theta_1 + \theta_2x_j + \theta_3(d_jx_j) + u_j, \quad j = 1, \dots, n \tag{10}$$

and the null hypothesis that the two models are the same is again equivalent to the hypothesis $\theta_3 = 0$.

The extension of this approach to multivariate models is straightforward and left to the reader.

Exercises:

1. Prove **(25)**.
2. Which MND¹ result is used in the proof of Theorem 2?
3. Prove **(29)**.
4. Which MND results are used to show that s^2 and $\hat{\theta}$ are independent (c.f. Theorem 3)?
5. Which MND results are used in **(36)** and **(37)**?
6. Give the details of the proof of Theorem 4. In particular, indicate which MND results are used.
7. Prove **(42)**.
8. Prove **(43)**.
9. Give the details of the proof of Theorem 5. In particular, indicate which MND results are used.
10. Which MND results are used in **(49)** and **(50)**?
11. Give the details of the proof of Theorem 6. In particular, indicate which MND results are used.
12. Why is **(67)** true?
13. Give the details of the proof of Theorem 8.
14. Give the details of the proof of Theorem 9.
15. Which MND results are used in **(96)**?

¹ Recall that MND stands for Multivariate Normal Distribution.

MULTICOLLINEARITY

Herman J. Bierens

Pennsylvania State University

Revised: April 5, 2007

1. *Introduction*

Consider the linear regression model

$$y_j = \theta_1 x_{1,j} + \dots + \theta_k x_{k,j} + u_j, \quad j = 1, \dots, n, \quad (1)$$

where y_j is the dependent variable, the $x_{i,j}$'s are the independent (or explanatory) variables, the u_j 's are unobservable error terms, n is the sample size, and the θ_i 's are the model parameters. If the model contains an intercept, then one of the $x_{i,j}$'s is equal to 1, say $x_{1,j}$. Throughout we assume that the errors are normally distributed:

$$u_j \sim N(0, \sigma^2), \quad (2)$$

Also, for the sake of the argument we assume that the x -variables are nonstochastic. This assumption is of course not very realistic, but it allow the discussion below to stay focused, and is harmless in that everything we are going to derive also holds conditional on the x variables if they are stochastic.

Denoting

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{1,1} & \dots & x_{k,1} \\ x_{1,2} & \dots & x_{k,2} \\ \vdots & \dots & \vdots \\ x_{1,n} & \dots & x_{k,n} \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \quad (3)$$

we can cast model (1) in vector/matrix form as:

$$y = X\theta + u, \quad u \sim N_n(0, \sigma^2 I_n), \quad (4)$$

Exact multicollinearity occurs if the columns of the matrix X are linearly dependent: at least one of the columns can be written as a linear combination of the other columns. This case is always due to model specification errors, or to transformation errors. For example, if there are dummy variables among the x variables like seasonal dummies that add up to 1, and if also an intercept is included, these dummy variables will be exactly multicollinear with the column on 1-s in the matrix X . The

solution is to delete either one of the dummy variables or the intercept itself. See Greene. Another common error is where you want to specify a quadratic model, for example $y_j = \theta_0 + \theta_1 x_j + \theta_2 x_j^2 + u_j$, and instead of making the variable x_j^2 by using the option "multiplicative transformation" you use the option "linear transformation", with coefficient 2. Then the transformed variable will be $2x_j$ rather than x_j^2 , and obviously x_j and $2x_j$ are exactly multicollinear.

Note that in the case of exact multicollinearity the matrix $X^T X$ is singular, hence we cannot compute the least squares estimator $\hat{\theta} = (X^T X)^{-1} X^T y$.

If the matrix $X^T X$ is nonsingular but the smallest eigenvalue is very small, we have a case of near-multicollinearity.

2. *The effect of near-multicollinearity on the t-values*

Denoting by e_i the i -th unit vector of length k , i.e., e_i is the i -th column of the unit matrix I_k , we can write the variance of the i -th component $\hat{\theta}_i$ of the least squares estimator $\hat{\theta}$ as

$$Var(\hat{\theta}_i) = \sigma^2 e_i^T (X^T X)^{-1} e_i. \quad (5)$$

Since we can write

$$X^T X = Q \Lambda Q^T, \quad (6)$$

where Λ is the diagonal matrix of eigenvalues of $X^T X$ and Q is the orthogonal matrix of corresponding eigenvectors, we can write this variance also as

$$Var(\hat{\theta}_i) = \sigma^2 e_i^T Q \Lambda^{-1} Q^T e_i = \sigma^2 \sum_{m=1}^k \frac{q_{m,i}^2}{\lambda_m}, \quad (7)$$

where $q_{m,i}$ is the (m,i) -th element of the matrix Q and the λ_m 's are the eigenvalues (λ_m is the m -th diagonal element of Λ). Now if λ_1 is the smallest eigenvalue, and $q_{1,i} \neq 0$, then the smaller λ_1 , the larger the variance involved. Therefore, near-multicollinearity may inflate all the variances and consequently deflate all the t-values. But how can we distinguish true insignificance from near-multicollinearity?

If the insignificant parameters are really zero, then the F-test of the joint hypothesis involved

should not reject. If it does, we have an indication that the low t-values are due to near-multicollinearity. In order to illustrate this, suppose that all the t-values are insignificant due to near-multicollinearity. Then the F-test of the (incorrect) null hypothesis $\theta = 0$ takes the form

$$\hat{F} = \frac{\hat{\theta}^T(X^T X)^{-1}\hat{\theta}/k}{s^2}, \quad (8)$$

where s^2 is the well-known unbiased estimator of σ^2 . It is also well-known that $\hat{\theta}$ is unbiased and independent of s^2 . Denoting $\gamma = (\gamma_1, \dots, \gamma_k)^T = Q^T\theta$ we therefore have:

$$E(\hat{F}|s^2) = \frac{\theta^T(X^T X)^{-1}\theta/k}{s^2} = \frac{\theta^T Q \Lambda^{-1} Q^T \theta/k}{s^2} = \frac{\gamma^T \Lambda^{-1} \gamma}{ks^2} = \frac{1}{ks^2} \sum_{i=1}^k \gamma_i^2 \lambda_i^{-1}. \quad (9)$$

Since $\gamma \neq 0$ as otherwise $\theta = 0$, we see that a very small λ_1 inflates the F-statistic rather than deflating it.

3. A cure for near-multicollinearity

The only cure for near-multicollinearity is to reduce the number of explanatory variables by imposing restrictions on the parameters. The best way of doing this is to impose restrictions that are prescribed by economic theory. Take for example the translog production function

$$\ln(Y) = \beta_0 + \beta_1 \ln(L) + \beta_2 \ln(K) + \beta_3 \frac{(\ln(L))^2}{2} + \beta_4 \ln(L) \ln(K) + \beta_5 \frac{(\ln(K))^2}{2} + u, \quad (10)$$

where Y is output, L is labor, and K is capital. If you suspect that $\ln(L)$ and $\ln(K)$ are near-multicollinear, then imposing the restriction of constant return to scale, which amounts to the restriction $\beta_1 + \beta_2 = 1$ and $\beta_3 = \beta_5 = -\beta_4$, will cure the problem, because then the model becomes

$$\ln(Y) - \ln(K) = \beta_0 + \beta_1 (\ln(L) - \ln(K)) + \beta_3 \frac{(\ln(L) - \ln(K))^2}{2} + u. \quad (11)$$

If economic theory does not provide guidelines for parameter restrictions, the only other option is to delete the variables from the model that cause the problem. As we have seen, the t-values cannot be used for testing which set of variables should be deleted, as they are "polluted" by the near-singularity of the $X^T X$ matrix. The solution I propose is to orthogonalize the columns of the matrix

X similarly to the Gramm-Schmidt orthogonalization of a basis of a vector space, as follows.

First, determine which of the insignificant x -variables is the least important for your economic theory, because you don't want to start off throwing away the key-variables. Suppose it is the last one. Then we can partition the matrix X in:

$$X = (X_{k-1}, x_k), \quad (12)$$

where X_{k-1} is the matrix of the first $k-1$ columns of X and x_k is the last column. Next, regress x_k on X_{k-1} , and let the vector of residuals be z_k :

$$z_k = x_k - X_{k-1}\alpha_{k-1}, \text{ where } \alpha_{k-1} = (X_{k-1}^T X_{k-1})^{-1} X_{k-1}^T x_k. \quad (13)$$

Then

$$X_{k-1}^T z_k = 0 \quad (14)$$

(exercise: why?). Next, replace x_k in X by z_k . Partitioning the parameter vector θ as

$$\theta = \begin{pmatrix} \theta_{k-1}^* \\ \theta_k \end{pmatrix} \quad (15)$$

the new model is related to the old one as follows:

$$y = X\theta + u = X_{k-1}\theta_{k-1}^* + x_k\theta_k + u = X_{k-1}(\theta_{k-1}^* + \alpha_{k-1}\theta_k) + z_k\theta_k + u \quad (16)$$

hence

$$y = X_{k-1}\beta_{k-1} + z_k\theta_k + u, \quad (17)$$

where $\beta_{k-1} = \theta_{k-1}^* + \alpha_{k-1}\theta_k$. Thus we only have reparametrized the model, but now z_k is orthogonal to the columns in X_{k-1} .

THEOREM 1: *The least square estimator and the t-value of the parameter θ_k in the reparametrized model (17) are the same as in the original model (4).*

Proof: We have

$$\begin{aligned} (X^T X)^{-1} &= \begin{pmatrix} X_{k-1}^T X_{k-1} & X_{k-1}^T x_k \\ x_k^T X_{k-1} & x_k^T x_k \end{pmatrix}^{-1} = \begin{pmatrix} (X_{k-1}^T X_{k-1})^{-1} + \frac{\alpha_{k-1} \alpha_{k-1}^T}{z_k^T z_k} & -\frac{\alpha_{k-1}}{z_k^T z_k} \\ -\frac{\alpha_{k-1}^T}{z_k^T z_k} & \frac{1}{z_k^T z_k} \end{pmatrix} \end{aligned} \quad (18)$$

(exercise: Verify this), and

$$X^T y = \begin{pmatrix} X_{k-1}^T y \\ x_k^T y \end{pmatrix} \quad (19)$$

hence

$$\hat{\theta}_k = \frac{x_k^T y - \alpha_{k-1}^T X_{k-1}^T y}{z_k^T z_k} = \frac{z_k^T y}{z_k^T z_k} \quad (20)$$

Moreover, the least squares estimator of the parameter vector $(\beta_k^T, \theta_k)^T$ in model (17) is

$$\begin{aligned} (X_{k-1}, z_k)^T (X_{k-1}, z_k)^{-1} (X_{k-1}, z_k)^T y &= \begin{pmatrix} (X_{k-1}^T X_{k-1})^{-1} & 0 \\ 0 & \frac{1}{z_k^T z_k} \end{pmatrix} \begin{pmatrix} X_{k-1}^T y \\ z_k^T y \end{pmatrix} \\ &= \begin{pmatrix} X_{k-1}^T X_{k-1}^{-1} X_{k-1}^T y \\ \frac{z_k^T y}{z_k^T z_k} \end{pmatrix}. \end{aligned} \quad (21)$$

Comparing the last two results, it follows that the least squares estimator of θ_k in both models is the same. Replacing y in (20) by the right-hand side of (17), and using (14), it easily follows that

$$\hat{\theta}_k - \theta_k = \frac{z_k^T u}{z_k^T z_k} \sim N\left(0, \frac{\sigma^2}{z_k^T z_k}\right). \quad (22)$$

From this result it follows that also the t-values are the same, because the sum of squared residuals of both models is the same, and so is the estimator s^2 of σ^2 . Q.E.D.

The t-values of the x -variables in the matrix X_{k-1} , however, will change if there is near-multicollinearity. If all the t-values of these variables are now significant, and some of them were not before, then you are done, in the sense that you now may blame the near-multicollinearity on x_k , and solve the problem by removing x_k from the model. If some t-values of variables in X_{k-1} are still insignificant, we may repeat the procedure, by selecting among the insignificant variables the one that is of the least importance (but do not choose the intercept! The intercept is important for other reasons.), and replace it by the residual of the regression of the variable involved on the remaining variables in X_{k-1} . If the last column, x_{k-1} , in the matrix X_{k-1} is this variable, and partitioning $X_{k-1} = (X_{k-2}, x_{k-1})$, then we replace x_{k-1} by the vector z_{k-1} of residuals of the regression of x_{k-1} on X_{k-2} , and run the regression

$$y = X_{k-2}\beta_{k-2} + z_{k-1}\beta_{k-1,k-1} + z_k\theta_k + u, \quad (23)$$

where $\beta_{k-1,k-1}$ is the last component of β_{k-1} . Again the t-value of $\beta_{k-1,k-1}$ will be the same as in model (17), as follows easily from Theorem 1. Repeating this procedure until all the t-values of the remaining variables are significant, we end up with a model of the form

$$y = X_{k-m}\beta_{k-m} + \sum_{i=1}^{m-1} z_{k-i}\beta_{k-i,k-i} + z_k\theta_k + u, \quad (24)$$

where the t-values of the parameters in β_{k-m} are significant, except perhaps the intercept, and the coefficients of the residuals z_i are all insignificant. This model is still equivalent to the original model (4), but we now have isolated the source of the problem, and since the residuals z_i are all insignificant, it is now clear which variables to remove from the model: the variables corresponding to the residuals in model (24), so that the final model becomes

$$y = X_{k-m}\beta_{k-m} + u. \quad (25)$$

This approach has the advantage that in the selection process one can also weigh the theoretical importance of each x -variable.

TESTS OF NORMALITY OF REGRESSION ERRORS

Herman J. Bierens

Pennsylvania State University

February 13, 2010

In this lecture note I will derive the Jarque-Bera (1980) and-Kiefer- Salmon (1983) tests of the normality of the regression errors. Some of the steps in the derivations follow from standard results, in particular Chapters 5 and 6 in Bierens (2004) and are therefore left as exercises, where the readers have to figure out for themselves which results in Bierens (2004) are applicable.

Consider the linear regression with an intercept

$$y_j = \theta^T x_j + u_j, \quad x_j^T = (1, x_{2,j}, \dots, x_{k,j}), \quad j = 1, \dots, n, \quad (1)$$

where

ASSUMPTION 1: *The vectors $(u_j, x_{2,j}, \dots, x_{k,j})^T$ are independent random drawings from a k-variate distribution, and $E(u_j | x_{2,j}, \dots, x_{k,j}) = 0$ with probability 1.*

We want to test the null hypothesis that (conditionally on the x variables),

$$H_0: u_j | x_j \sim N(0, \sigma^2). \quad (2)$$

Under the null hypotheses we have:

$$E(u_j^3) = 0, \quad E(u_j^4) = 3\sigma^4, \quad (3)$$

and it will be more feasible to test the latter implication of the null hypothesis than the null hypothesis (2) itself.

If the errors u_j would be observable and the variance σ^2 known, we could test the hypothesis (3) on the basis of the random vector

$$\frac{1}{\sqrt{n}} \begin{pmatrix} \sum_{j=1}^n (u_j/\sigma)^3 \\ \sum_{j=1}^n ((u_j/\sigma)^4 - 3) \end{pmatrix}, \quad (4)$$

which under the null hypothesis converges in distribution to the bivariate normal distribution with zero mean vector and variance matrix

$$\begin{pmatrix} E(u_1/\sigma)^6 & E(u_1/\sigma)^7 - 3E(u_1/\sigma)^3 \\ E(u_1/\sigma)^7 - 3E(u_1/\sigma)^3 & E(u_1/\sigma)^8 - 6E(u_1/\sigma)^4 + 9 \end{pmatrix} = \begin{pmatrix} 15 & 0 \\ 0 & 96 \end{pmatrix}, \quad (5)$$

where the equality involved follows from the fact that under the hypothesis (2),

$$\begin{aligned} E(u_j^{2m-1}) &= 0 \quad \text{for } m = 1, 2, 3, \dots \\ E(u_j^2) &= \sigma^2 \\ E(u_j^4) &= 3\sigma^4 \\ E(u_j^6) &= 15\sigma^6 \\ E(u_j^8) &= 105\sigma^8 \end{aligned} \quad (6)$$

The results in (6) can be derived using the moment generating function of the standard normal distribution. (*Exercise:* Try it!) Thus under the null hypothesis (2) we have

$$\frac{1}{\sqrt{n}} \begin{pmatrix} \sum_{j=1}^n (u_j/\sigma)^3 \\ \sum_{j=1}^n ((u_j/\sigma)^4 - 3) \end{pmatrix} \rightarrow N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 15 & 0 \\ 0 & 96 \end{pmatrix} \right] \text{ in distr.} \quad (7)$$

and consequently,

$$\frac{\left(\sum_{j=1}^n (u_j/\sigma)^3 \right)^2}{15n} + \frac{\left(\sum_{j=1}^n ((u_j/\sigma)^4 - 3) \right)^2}{96n} \rightarrow \chi_2^2 \text{ in distr.} \quad (8)$$

(*Exercise:* Why?) However, we do not observe the u_j 's and the variance σ^2 , so that this test is not feasible.

The idea behind the Jarque-Bera (1980) and Kiefer-Salmon (1983) tests is to replace in (7)

the u_j 's by the least squares residuals

$$\hat{u}_j = y_j - \hat{\theta}^T x_j = u_j - (\hat{\theta} - \theta)^T x_j, \quad (9)$$

where $\hat{\theta}$ is the OLS estimator of θ , and σ^2 by the estimator

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n \hat{u}_j^2, \quad (10)$$

and to prove that the asymptotic normality result (7) goes through, although with a different variance matrix.

Note that, due to the presence of an intercept,

$$\sum_{j=1}^n \hat{u}_j = 0. \quad (11)$$

(Exercise: Why?). Therefore, we can also write

$$\hat{u}_j = u_j - \bar{u} - (\hat{\theta} - \theta)^T (x_j - \bar{x}), \quad (12)$$

where $\bar{u} = (1/n) \sum_{j=1}^n u_j$ and $\bar{x} = (1/n) \sum_{j=1}^n x_j$. This will be convenient in proving Lemmas 2 and 3 below.

Next assume:

ASSUMPTION 2: $Q = E(x_j x_j^T)$ is nonsingular, and $E(\|x_j\|^4) < \infty$,

where $\|x\| = \sqrt{x^T x}$ is the Euclidean norm. Then:

LEMMA 1: Under Assumptions 1-2 and the null hypothesis (2) we have

$$\text{plim}_{n \rightarrow \infty} [(1/\sqrt{n}) \sum_{j=1}^n \hat{u}_j^2 - (1/\sqrt{n}) \sum_{j=1}^n u_j^2] = 0.$$

Proof: It follows from (9) that

$$\begin{aligned}
\sum_{j=1}^n \hat{u}_j^2 &= \sum_{j=1}^n (u_j - (\hat{\theta} - \theta)^T x_j)^2 = \sum_{j=1}^n u_j^2 - 2(\hat{\theta} - \theta)^T \sum_{j=1}^n u_j x_j + (\hat{\theta} - \theta)^T \sum_{j=1}^n x_j x_j^T (\hat{\theta} - \theta) \\
&= \sum_{j=1}^n u_j^2 - \sqrt{n}(\hat{\theta} - \theta)^T \left(\frac{1}{n} \sum_{j=1}^n x_j x_j^T \right) \sqrt{n}(\hat{\theta} - \theta).
\end{aligned} \tag{13}$$

(Exercise: Prove the last step). Moreover, under the conditions of the lemma we have

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N_k(0, \sigma^2 Q^{-1}) \text{ in distr.} \tag{14}$$

and by the law of large numbers,

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n x_j x_j^T = Q. \tag{15}$$

(Exercise: Verify the conditions of the law of large numbers). These two results imply that

$$\sum_{j=1}^n u_j^2 - \sum_{j=1}^n \hat{u}_j^2 = \sqrt{n}(\hat{\theta} - \theta)^T \left(\frac{1}{n} \sum_{j=1}^n x_j x_j^T \right) \sqrt{n}(\hat{\theta} - \theta) \rightarrow \chi_k^2 \text{ in distr.} \tag{16}$$

(Exercise: Which result in Bierens (2004) has been applied?). The lemma follows now from (16)

(Exercise: Why?).

LEMMA 2: Under Assumptions 1-2 and the null hypothesis (2) we have

$$\text{plim}_{n \rightarrow \infty} [(1/\sqrt{n}) \sum_{j=1}^n \hat{u}_j^3 - (1/\sqrt{n}) \sum_{j=1}^n (u_j^3 - 3\sigma^2 u_j)] = 0.$$

Proof: Using (12) and the well-known equality $(a-b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$, we have

$$\begin{aligned}
\sum_{j=1}^n \hat{u}_j^3 &= \sum_{j=1}^n (u_j - \bar{u} - (\hat{\theta} - \theta)^T (x_j - \bar{x}))^3 = \sum_{j=1}^n (u_j - \bar{u})^3 - 3(\hat{\theta} - \theta)^T \sum_{j=1}^n (u_j - \bar{u})^2 (x_j - \bar{x}) \\
&\quad + 3(\hat{\theta} - \theta)^T \sum_{j=1}^n (u_j - \bar{u}) (x_j - \bar{x}) (x_j - \bar{x})^T (\hat{\theta} - \theta) - \sum_{j=1}^n ((\hat{\theta} - \theta)^T (x_j - \bar{x}))^3
\end{aligned} \tag{17}$$

Since under the conditions of the lemma,

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n (u_j - \bar{u})^2 (x_j - \bar{x}) = 0 \tag{18}$$

and

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n (u_j - \bar{u}) (x_j - \bar{x}) (x_j - \bar{x})^T = O \tag{19}$$

(exercise: why?), it follow that

$$\text{plim}_{n \rightarrow \infty} \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n \hat{u}_j^3 - \frac{1}{\sqrt{n}} \sum_{j=1}^n (u_j - \bar{u})^3 \right) = 0. \quad (20)$$

Next, we have

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{j=1}^n (u_j - \bar{u})^3 &= \frac{1}{\sqrt{n}} \sum_{j=1}^n u_j^3 - 3\sqrt{n}\bar{u} \frac{1}{n} \sum_{j=1}^n u_j^2 + 2\sqrt{n}\bar{u}^3 \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^n (u_j^3 - 3\sigma^2 u_j) - 3\sqrt{n}\bar{u} \frac{1}{n} \sum_{j=1}^n (u_j^2 - \sigma^2) + 2\sqrt{n}\bar{u}^3, \end{aligned} \quad (21)$$

hence

$$\text{plim}_{n \rightarrow \infty} \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n (u_j - \bar{u})^3 - \frac{1}{\sqrt{n}} \sum_{j=1}^n (u_j^3 - 3\sigma^2 u_j) \right) = 0. \quad (22)$$

(Exercise: Why?). Combining (20) and (22), the lemma follows.

LEMMA 3: Under Assumptions 1-2 and the null hypothesis (2) we have

$$\text{plim}_{n \rightarrow \infty} \left((1/\sqrt{n}) \sum_{j=1}^n \hat{u}_j^4 - (1/\sqrt{n}) \sum_{j=1}^n u_j^4 \right) = 0.$$

Proof: Again using (12), and the well-known equality $(a-b)^4 = a^4 - 4a^3b + 6a^2b^2 - 4ab^3 + b^4$, we have

$$\begin{aligned} \sum_{j=1}^n \hat{u}_j^4 &= \sum_{j=1}^n (u_j - \bar{u} - (\hat{\theta} - \theta)^T(x_j - \bar{x}))^4 \\ &= \sum_{j=1}^n (u_j - \bar{u})^4 - 4(\hat{\theta} - \theta)^T \sum_{j=1}^n (u_j - \bar{u})^3 (x_j - \bar{x}) \\ &\quad + 6(\hat{\theta} - \theta)^T \sum_{j=1}^n (u_j - \bar{u})^2 (x_j - \bar{x}) (x_j - \bar{x})^T (\hat{\theta} - \theta) \\ &\quad - 4 \sum_{j=1}^n (u_j - \bar{u}) (\hat{\theta} - \theta)^T (x_j - \bar{x})^3 + \sum_{j=1}^n ((\hat{\theta} - \theta)^T (x_j - \bar{x}))^4 \end{aligned} \quad (23)$$

Therefore, similar to the proof of Lemmas 2 we have:

$$\text{plim}_{n \rightarrow \infty} \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n \hat{u}_j^4 - \frac{1}{\sqrt{n}} \sum_{j=1}^n (u_j - \bar{u})^4 \right) = 0. \quad (24)$$

(Exercise: Prove this). Moreover,

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n (u_j - \bar{u})^4 = \frac{1}{\sqrt{n}} \sum_{j=1}^n u_j^4 - 4\sqrt{n}\bar{u} \frac{1}{n} \sum_{j=1}^n u_j^3 + 6\frac{1}{\sqrt{n}} (\sqrt{n}\bar{u})^2 \frac{1}{n} \sum_{j=1}^n u_j^2 - 3\frac{1}{n\sqrt{n}} (\sqrt{n}\bar{u})^4. \quad (25)$$

Since $\sqrt{n}\bar{u} \sim N(0, \sigma^2)$, $\text{plim}_{n \rightarrow \infty} (1/n) \sum_{j=1}^n u_j^3 = 0$ and $\text{plim}_{n \rightarrow \infty} (1/n) \sum_{j=1}^n u_j^2 = \sigma^2$, it follows now that

$$\text{plim}_{n \rightarrow \infty} \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n (u_j - \bar{u})^4 - \frac{1}{\sqrt{n}} \sum_{j=1}^n u_j^4 \right) = 0. \quad (26)$$

(Exercise: Why?). Combining (24) and (26), Lemma 3 follows.

LEMMA 4: Under Assumptions 1-2 and the null hypothesis (2) we have

$$\text{plim}_{n \rightarrow \infty} [(1/\sqrt{n}) \sum_{j=1}^n (\hat{u}_j^4 - 3\hat{\sigma}^4) - (1/\sqrt{n}) \sum_{j=1}^n (u_j^4 - 6\sigma^2 u_j^2 + 3\sigma^4)] = 0.$$

Proof: It follows from (10), Lemma 1, and the law of large numbers and the central limit theorem that

$$\text{plim}_{n \rightarrow \infty} \hat{\sigma}^2 = \sigma^2, \quad \sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{D} N(0, 2\sigma^4) \text{ in distr}, \quad (27)$$

(Exercise: Why?) and thus

$$\text{plim}_{n \rightarrow \infty} \left(\sqrt{n}(\hat{\sigma}^4 - \sigma^4) - 2\sigma^2 \frac{1}{\sqrt{n}} \sum_{j=1}^n (u_j^2 - \sigma^2) \right) = 0 \quad (28)$$

(Exercise: Why?). Combining the latter result with Lemma 3, Lemma 4 follows.

Finally, it follows from (6) that

$$E \begin{pmatrix} u_j^3 - 3\sigma^2 u_j \\ u_j^4 - 6\sigma^2 u_j^2 + 3\sigma^4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{Var} \begin{pmatrix} u_j^3 - 3\sigma^2 u_j \\ u_j^4 - 6\sigma^2 u_j^2 + 3\sigma^4 \end{pmatrix} = \begin{pmatrix} 6\sigma^6 & 0 \\ 0 & 24\sigma^8 \end{pmatrix} \quad (29)$$

hence, using the central limit theorem, it follows that

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n \begin{pmatrix} u_j^3 - 3\sigma^2 u_j \\ u_j^4 - 6\sigma^2 u_j + 3\sigma^4 \end{pmatrix} \rightarrow N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 6\sigma^6 & 0 \\ 0 & 24\sigma^8 \end{pmatrix} \right] \text{ in distr.} \quad (30)$$

Using this result and those of Lemmas 1-4, it follows that

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n \begin{pmatrix} (\hat{u}_j/\hat{\sigma})^3 \\ (\hat{u}_j/\hat{\sigma})^4 - 3 \end{pmatrix} \rightarrow N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix} \right] \text{ in distr.} \quad (31)$$

hence

THEOREM 1: Under Assumptions 1-2 and the null hypothesis (2) we have

$$\hat{T}_N = n \left(\frac{[(1/n) \sum_{j=1}^n (\hat{u}_j/\hat{\sigma})^3]^2}{6} + \frac{[(1/n) \sum_{j=1}^n (\hat{u}_j/\hat{\sigma})^4 - 3]^2}{24} \right) \rightarrow \chi^2_2 \text{ in distr.} \quad (32)$$

Proof: Exercise.

The statistic \hat{T}_N is the test statistic of the Kiefer-Salmon (1983) test. The Jarque-Bera (1980) test is derived in a slightly different way, but is essentially the same.

REFERENCES

- Bierens, H.J. (2004), *Introductions to the Mathematical and Statistical Foundations of Econometrics*, Cambridge University Press.
- Jarque, C.M. and A.K. Bera, (1980), "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals". *Economics Letters* 6, 255–259.
- Kiefer, N. and M. Salmon (1983), "Testing Normality in Econometric Models", *Economic Letters* 11, 123-127.

METHOD OF MOMENTS

Herman J. Bierens

Pennsylvania State University

September 19, 2005

1. Linear method of moments

1.1. The model

Consider a system of k linear equations,

$$y_{i,t} = x_{i,t}^T \theta_i + u_{i,t}, \quad t = 1, \dots, n, \quad i = 1, \dots, k, \quad \theta_i \in \mathbb{R}^{p_i}, \quad (1)$$

where the $x_{i,t}$ vectors possibly contain some of the dependent variables $y_{j,t}$, and the errors $u_{i,t}$ have zero expectation but are contemporaneously dependent. Consequently, the usual regression assumption $E[u_{i,t}|x_{i,t}] = 0$ may not apply. However, suppose we have q_i -vectors $z_{i,t}$ of instrumental variables such that $E[u_{i,t}z_{i,t}] = 0$, hence

$$\left(E[z_{i,t}x_{i,t}^T] \right) \theta_i = E[z_{i,t}y_{i,t}] \quad (2)$$

is a system of q_i linear equations in p_i unknown elements of θ_i . If $q_i \geq p_i$, and the rank of the matrix $E[z_{i,t}x_{i,t}^T]$ is p_i or larger, the parameter vector θ_i is identified by the moment conditions (2).

Note that this case is only one of the many cases for which the method of moment estimation approach is applicable. For example, least squares and two-stage least squares estimation are special cases of method of moment estimation techniques.

Denoting

$$p = \sum_{i=1}^k p_i, \quad q = \sum_{i=1}^k q_i, \quad (3)$$

we can write this model in vector form as

$$y_t = X_t^T \theta_0 + u_t, \quad E[Z_t u_t] = 0, \quad (4)$$

where

$$y_t = \begin{pmatrix} y_{1,t} \\ \vdots \\ y_{k,t} \end{pmatrix}, \quad X_t = \begin{pmatrix} x_{1,t} & 0 & \dots & 0 \\ 0 & x_{2,t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_{k,t} \end{pmatrix} \quad (p \times k), \quad \theta_0 = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}, \quad u_t = \begin{pmatrix} u_{1,t} \\ \vdots \\ u_{k,t} \end{pmatrix}, \quad (5)$$

and

$$Z_t = \begin{pmatrix} z_{1,t} & 0 & \dots & 0 \\ 0 & z_{2,t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & z_{k,t} \end{pmatrix} \quad (q \times k). \quad (6)$$

The moment conditions (2) now take the form

$$E[Z_t X_t^T] \theta_0 = E[Z_t y_t]. \quad (7)$$

The implicit assumption that the parameter vectors θ_i in model (1) are different is not essential. If some or all of the components of the parameter vectors θ_i are common, we may augment the $x_{i,t}$ vectors with zeros corresponding to the parameters that are not part of the equation involved, and write model (1) as

$$y_{i,t} = x_{i,t}^T \theta_0 + u_{i,t}, \quad t = 1, \dots, n, \quad i = 1, \dots, k, \quad \theta_0 \in \mathbb{R}^p. \quad (8)$$

The only difference is then that the matrix X_t in (5) becomes

$$X_t = (x_{1,t}, \dots, x_{k,t}) \quad (p \times k). \quad (9)$$

The moment conditions (7) suggest to estimate θ_0 by minimizing the quadratic form

$$Q_n(\theta) = M_n(\theta)^T W_n M_n(\theta), \quad (10)$$

where

$$M_n(\theta) = \frac{1}{n} \sum_{t=1}^n Z_t (y_t - X_t^T \theta) = \frac{1}{n} \sum_{t=1}^n Z_t y_t - \left(\frac{1}{n} \sum_{t=1}^n Z_t X_t^T \right) \theta. \quad (11)$$

and W_n is a positive definite $q \times q$ matrix, to be determined later. Thus, the *method of moment* estimator involved is:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} Q_n(\theta). \quad (12)$$

If $q = p$, the solution is the same as the solution of $M_n(\theta) = 0$, namely

$$\hat{\theta} = \left(\frac{1}{n} \sum_{t=1}^n Z_t X_t^T \right)^{-1} \left(\frac{1}{n} \sum_{t=1}^n Z_t y_t \right), \quad (13)$$

provided that the inverted matrix is nonsingular, which is known as the just identified case, but the overidentified case $q > p$ is more interesting and challenging.

The first-order condition for a minimum of $Q_n(\theta)$ is:

$$\frac{\partial Q_n(\theta)}{\partial \theta^T} = 2 \left(\frac{\partial M_n(\theta)^T}{\partial \theta^T} \right) W_n M_n(\theta) = -2 \left(\frac{1}{n} \sum_{t=1}^n X_t Z_t^T \right) W_n \left(\frac{1}{n} \sum_{t=1}^n Z_t y_t - \frac{1}{n} \sum_{t=1}^n Z_t X_t^T \theta \right) = 0 \quad (14)$$

hence the solution is:

$$\begin{aligned} \hat{\theta} &= \left[\left(\frac{1}{n} \sum_{t=1}^n X_t Z_t^T \right) W_n \left(\frac{1}{n} \sum_{t=1}^n Z_t X_t^T \right) \right]^{-1} \left(\frac{1}{n} \sum_{t=1}^n X_t Z_t^T \right) W_n \left(\frac{1}{n} \sum_{t=1}^n Z_t y_t \right) \\ &= \theta_0 + \left[\left(\frac{1}{n} \sum_{t=1}^n X_t Z_t^T \right) W_n \left(\frac{1}{n} \sum_{t=1}^n Z_t X_t^T \right) \right]^{-1} \cdot \left(\frac{1}{n} \sum_{t=1}^n X_t Z_t^T \right) W_n \left(\frac{1}{n} \sum_{t=1}^n Z_t u_t \right) \end{aligned} \quad (15)$$

Now assume that

Assumption 1: $\frac{1}{\sqrt{n}} \sum_{t=1}^n Z_t u_t \rightarrow N_k(0, A)$ in distr., where $A = \operatorname{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n Z_t u_t u_t^T Z_t^T$.

This condition is satisfied if for example $Z_t u_t$ is i.i.d. and the variance matrix A of $Z_t u_t$ is finite.

Moreover, assume that

Assumption 2: $B = \text{plim}_{n \rightarrow \infty} (1/n) \sum_{t=1}^n X_t Z_t^T$ exists and is finite,

Assumption 3: $W = \text{plim}_{n \rightarrow \infty} W_n$ is finite and positive definite.

Assumption 4: BWB^T is nonsingular.

Then under Assumptions 1-4,

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N_p[0, (BWB^T)^{-1}(BWA WB^T)(BWB^T)^{-1}] \text{ in distr.} \quad (16)$$

Because the variance matrix involved depends on W , the question now arises:

1.2. What is the best choice for W_n ?

In order to answer this question, consider the linear regression model

$$y = A^{-1/2}B^T\theta + e, \quad e \sim N(0, I). \quad (17)$$

It follows from the Gauss-Markov theorem that the best linear unbiased estimator of θ is the least squares estimator:

$$\hat{\theta} = (BA^{-1}B^T)^{-1}BA^{-1/2}y = \theta + (BA^{-1}B^T)^{-1}BA^{-1/2}e \sim N[\theta, (BA^{-1}B^T)^{-1}] \quad (18)$$

Next, consider the alternative unbiased estimator

$$\begin{aligned} \tilde{\theta} &= [(BWB^T)^{-1}BWA^{1/2}]y = \theta + [(BWB^T)^{-1}BWA^{1/2}]e \\ &\sim N[\theta, (BWB^T)^{-1}(BWA WB^T)(BWB^T)^{-1}] \end{aligned} \quad (19)$$

By the Gauss-Markov theorem,

$$D = (BWB^T)^{-1}(BWA WB^T)(BWB^T)^{-1} - (BA^{-1}B^T)^{-1} \quad (20)$$

is a positive semi-definite matrix. The direct proof follows from the fact that we can write

$$D = [(BWB^T)^{-1}BWA^{1/2}]\left(I - A^{-1/2}B^T(BA^{-1}B^T)^{-1}BA^{-1/2}\right)[A^{1/2}WB^T(BWB^T)^{-1}] \quad (21)$$

and that the matrix $I - A^{-1/2}B^T(BA^{-1}B^T)^{-1}BA^{-1/2}$ is idempotent, hence positive semi-definite.

Since $D = O$ if $W = A^{-1}$, the best choice for W_n is therefore such that

$$\operatorname{plim}_{n \rightarrow \infty} W_n = A^{-1}. \quad (22)$$

The *efficient method of moment estimation* procedure is now as follows. First, choose an initial matrix W_n , for example, let $W_n = I_q$. Then compute the first stage method of moment estimator $\hat{\theta}$, and denote

$$\hat{A} = \frac{1}{n} \sum_{t=1}^n Z_t \hat{u}_t \hat{u}_t^T Z_t^T, \quad \text{where } \hat{u}_t = y_t - X_t^T \hat{\theta}. \quad (23)$$

Next, choose

$$W_n = \hat{A}^{-1}, \quad (24)$$

which under Assumptions 1-4 is a consistent estimator of A^{-1} . Using this matrix W_n in the second stage now yields the *efficient method of moment* estimator:

$$\hat{\theta}_{EMM} = \underset{\theta}{\operatorname{argmin}} M_n(\theta)^T \hat{A}^{-1} M_n(\theta), \quad (25)$$

with limiting normal distribution:

$$\sqrt{n}(\hat{\theta}_{EMM} - \theta_0) \rightarrow N_p[0, (BA^{-1}B^T)^{-1}]. \quad (26)$$

Moreover, denoting

$$\hat{B} = \frac{1}{n} \sum_{t=1}^n X_t Z_t^T, \quad (27)$$

it follows from Assumptions 1,2, and 4 that the asymptotic variance matrix in (26) can be estimated consistently by $(\hat{B}\hat{A}^{-1}\hat{B}^T)^{-1}$.

1.3. Testing the adequacy of the instruments

It follows from (15) with (24) and (27) that

$$\sqrt{n}(\hat{\theta}_{EMM} - \theta_0) = (\hat{B}\hat{A}^{-1}\hat{B}^T)^{-1} (\hat{B}\hat{A}^{-1}) \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n Z_t u_t \right), \quad (28)$$

hence it follows from (11) and Assumption 1 that

$$\begin{aligned}
\sqrt{n}\hat{A}^{-1/2}M_n(\hat{\theta}_{EMM}) &= \hat{A}^{-1/2} \frac{1}{\sqrt{n}} \sum_{t=1}^n Z_t u_t - \hat{A}^{-1/2} \hat{B}^T \sqrt{n}(\hat{\theta}_{EMM} - \theta_0) \\
&= \left[\hat{A}^{-1/2} - \hat{A}^{-1/2} \hat{B}^T (\hat{B} \hat{A}^{-1} \hat{B}^T)^{-1} (\hat{B} \hat{A}^{-1}) \right] \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n Z_t u_t \right) \\
&= \left[I_q - \hat{A}^{-1/2} \hat{B}^T (\hat{B} \hat{A}^{-1} \hat{B}^T)^{-1} (\hat{B} \hat{A}^{-1/2}) \right] \left(\hat{A}^{-1/2} \frac{1}{\sqrt{n}} \sum_{t=1}^n Z_t u_t \right) \\
&\rightarrow N_q[0, M] \text{ in distribution,}
\end{aligned} \tag{29}$$

where

$$M = I_q - A^{-1/2} B^T (BA^{-1} B^T)^{-1} BA^{-1/2}. \tag{30}$$

Since M is idempotent (*Exercise: Why?*), it follows now that under Assumptions 1-4,

$$nM_n(\hat{\theta}_{EMM})^T \hat{A}^{-1} M_n(\hat{\theta}_{EMM}) \rightarrow \chi_{q-p}^2 \text{ in distribution.} \tag{31}$$

(*Exercise: Why?*) On the other hand, if

$$\operatorname{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n Z_t u_t = \eta \neq 0, \tag{32}$$

and $q > p$ then under Assumptions 2-4,

$$\operatorname{plim}_{n \rightarrow \infty} \hat{A}^{-1/2} M_n(\hat{\theta}_{EMM}) = \left[A^{-1/2} - A^{-1/2} B^T (BA^{-1} B^T)^{-1} (BA^{-1}) \right] \eta \neq 0 \tag{33}$$

hence

$$\operatorname{plim}_{n \rightarrow \infty} n M_n(\hat{\theta}_{EMM})^T \hat{A}^{-1} M_n(\hat{\theta}_{EMM}) = \infty. \tag{34}$$

Therefore, we can use $n M_n(\hat{\theta}_{EMM})^T \hat{A}^{-1} M_n(\hat{\theta}_{EMM})$ as a test for the adequacy of the instruments.

1.4. Application to static panel data models

A static panel data model takes the form¹

$$y_{i,t}^* = x_{i,t}^{*T} \theta_0 + \alpha_i + \varepsilon_{i,t}, \quad t = 1, \dots, T, \quad i = 1, \dots, N, \quad \theta_0 \in \mathbb{R}^p, \quad (35)$$

where α_i is a fixed or random effect which is constant over time t but varies with the cross-section index i , the $x_{i,t}^*$ are $p \times 1$ vectors of exogenous variables, **none** of which are constant over time, and the $\varepsilon_{i,t}$ are i.i.d. $(0, \sigma^2)$ errors that are independent of the exogenous variables. It will be assumed that the cross-section dimension N is much larger than the time dimension T , so that T may be considered as fixed, whereas all the asymptotic properties follow from letting $N \rightarrow \infty$.

In order to get rid of α_i , we take first differences:

$$\begin{aligned} y_{i,t}^* - y_{i,t-1}^* &= (x_{i,t}^* - x_{i,t-1}^*)^T \theta_0 + \varepsilon_{i,t} - \varepsilon_{i,t-1}, \\ t &= 2, \dots, T, \quad i = 1, \dots, N, \quad \theta_0 \in \mathbb{R}^p, \end{aligned} \quad (36)$$

Since $\varepsilon_{i,t} - \varepsilon_{i,t-1}$ is independent of $x_{i,t}^*$ and $x_{i,t-1}^*$, we can choose either $x_{i,t}^* - x_{i,t-1}^*$ or $x_{i,t}^*$ and $x_{i,t-1}^*$ as instruments. Choosing the latter, we can now write the model in vector form as

$$y_i = X_i^T \theta_0 + u_i, \quad E[Z_i u_i] = 0, \quad i = 1, \dots, N, \quad (37)$$

where

$$y_i = \begin{pmatrix} y_{i,2}^* - y_{i,1}^* \\ \vdots \\ y_{i,T}^* - y_{i,T-1}^* \end{pmatrix}, \quad X_i = (x_{i,2}^* - x_{i,1}^*, \dots, x_{i,T}^* - x_{i,T-1}^*), \quad u_i = \begin{pmatrix} \varepsilon_{i,2}^* - \varepsilon_{i,1}^* \\ \vdots \\ \varepsilon_{i,T}^* - \varepsilon_{i,T-1}^* \end{pmatrix}, \quad (38)$$

and

¹ Note that T now denotes the length of the time series, whereas the superscript T still denotes the "transpose".

$$Z_t = \begin{pmatrix} x_{i,1}^* & 0 & \dots & 0 \\ x_{i,2}^* & 0 & \dots & 0 \\ 0 & x_{i,2}^* & \dots & 0 \\ 0 & x_{i,3}^* & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_{i,T-1}^* \\ 0 & 0 & \dots & x_{i,T}^* \end{pmatrix} \quad (q \times k), \quad q = 2p, \quad k = T-1. \quad (39)$$

Note that

$$Var(u_i) = \sigma^2 \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & 0 & -1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 \end{pmatrix} = \sigma^2 \Omega, \quad say. \quad (40)$$

Hence

$$A = \text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N Z_i u_i u_i^T Z_t^T = \sigma^2 \text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N Z_i \Omega Z_t^T. \quad (41)$$

Therefore, if we choose

$$W_N = \left(\frac{1}{N} \sum_{i=1}^N Z_i \Omega Z_t^T \right)^{-1} \quad (42)$$

as the weight matrix, we obtain the efficient method of moment estimator $\hat{\theta}_{EMM}$ in one step. The only difference with the general case is that $N M_N(\hat{\theta}_{EMM})^T W_N M_N(\hat{\theta}_{EMM})$ needs to be divided by a consistent estimate $\hat{\sigma}^2$ of the variance σ^2 of the errors $\varepsilon_{i,t}$'s in order to be used as a chi-square

test of model correctness. Of course, we also need $\hat{\sigma}^2$ to estimate the asymptotic variance matrix of $\sqrt{N}(\hat{\theta}_{EMM} - \theta_0)$. For example, given the residual vector $\hat{u}_i = y_i - X_i^T \hat{\theta}_{EMM}$, the variance σ^2 can be estimated consistently by

$$\hat{\sigma}^2 = \frac{1}{2N(T-1)} \sum_{i=1}^N \hat{u}_i^T \hat{u}_i. \quad (43)$$

(Exercise: Why?)

1.5. Application to dynamic panel data models

A dynamic panel data model takes the form

$$y_{i,t}^* = \rho_0 y_{i,t-1}^* + x_{i,t}^{*T} \beta_0 + \alpha_i + \varepsilon_{i,t}, \quad t = 2, \dots, T, \quad i = 1, \dots, N, \quad \beta_0 \in \mathbb{R}^p, \quad |\rho_0| < 1, \quad (44)$$

where again α_i is a fixed or random effect which is constant over time t but varies with the cross-section index i , the $x_{i,t}^*$ are $p \times 1$ vectors of exogenous variables which are not constant over time, and the $\varepsilon_{i,t}$ are i.i.d. $(0, \sigma^2)$ errors that are independent of the exogenous variables. Also now it will be assumed that the cross-section dimension N is much larger than the time dimension T , so that T may be considered as fixed, whereas all the asymptotic properties follow from letting $N \rightarrow \infty$.

Taking first differences yields for $t = 3, \dots, T$, $i = 1, \dots, N$,

$$y_{i,t}^* - y_{i,t-1}^* = \rho_0(y_{i,t-1}^* - y_{i,t-2}^*) + (x_{i,t}^* - x_{i,t-1}^*)^T \beta_0 + \varepsilon_{i,t} - \varepsilon_{i,t-1}. \quad (45)$$

Due to the dynamic structure of the model, we now have a much richer choice of instruments, because $\rho_0(y_{i,t-1}^* - y_{i,t-2}^*) + (x_{i,t}^* - x_{i,t-1}^*)^T \beta_0$ depends on $y_{i,t-2-j}^*$ for $j = 0, \dots, t-2$ as well as on $x_{i,t-j}^*$ for $j = 0, \dots, t$. Denoting

$$X_i = (y_{i,2}^* - y_{i,1}^*, x_{i,3}^* - x_{i,2}^*, \dots, y_{i,T-1}^* - y_{i,T-2}^*, x_{i,T}^* - x_{i,T-1}^*) \quad (46)$$

$$\theta_0 = \begin{pmatrix} \rho_0 \\ \beta_0 \end{pmatrix}, \quad y_i = \begin{pmatrix} y_{i,3}^* - y_{i,2}^* \\ \vdots \\ y_{i,T}^* - y_{i,T-1}^* \end{pmatrix}, \quad u_i = \begin{pmatrix} \varepsilon_{i,3}^* - \varepsilon_{i,2}^* \\ \vdots \\ \varepsilon_{i,T}^* - \varepsilon_{i,T-1}^* \end{pmatrix}, \quad (47)$$

$$Z_t = \begin{pmatrix} x_{i,1}^* & x_{i,1}^* & \dots & x_{i,1}^* \\ x_{i,2}^* & x_{i,2}^* & \dots & x_{i,2}^* \\ x_{i,3}^* & x_{i,3}^* & \dots & x_{i,3}^* \\ y_{i,1}^* & x_{i,4}^* & \dots & x_{i,4}^* \\ 0 & y_{i,1}^* & \dots & x_{i,5}^* \\ 0 & y_{i,2}^* & \dots & x_{i,6}^* \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_{i,T}^* \\ 0 & 0 & \dots & y_{i,1}^* \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & y_{i,T-2}^* \end{pmatrix} \quad (q \times k), \quad q = pT+T-1, \quad k = T-2, \quad (48)$$

we can write the model again as (37). Therefore, the same results as in the previous section hold, except that we have to modify (43) to

$$\hat{\sigma}^2 = \frac{1}{2N(T-2)} \sum_{i=1}^N \hat{u}_i^T \hat{u}_i. \quad (49)$$

(Exercise: Why?)

2. Nonlinear method of moments

2.1 The model

Consider now the case where a model for a random vector $X_t \in \mathbb{R}^k$ is implicitly defined by a set of moment conditions:

$$m_t(\theta) = \begin{pmatrix} \mu_1(X_t, \theta) \\ \vdots \\ \mu_q(X_t, \theta) \end{pmatrix}, \quad \theta \in \Theta \subset \mathbb{R}^p, \quad \exists \theta_0 \in \Theta: E[m_t(\theta_0)] = 0, \quad (50)$$

where $q \geq p$, Θ is the parameter space, θ_0 is the parameter vector of interest. The random vectors

X_t are observable for $t = 1, \dots, n$. For convenience of the exposition we will assume that

Assumption A: *the X_t ‘s are i.i.d.,*

but under some mild additional conditions the results below will also hold if the X_t ‘s are realizations of a stationary vector time series process, or are panel data observations.

The following assumptions allow us to apply the central limit theorem and the uniform law of large numbers:

Assumption B: *The functions $\mu_i(x, \theta)$ are twice continuously differentiable in θ , and for each $\theta \in \Theta$ Borel measurable in $x \in \mathbb{R}^k$. The parameter space Θ is compact and convex, and θ_0 is an interior point of Θ .*

Assumption C: *For $i = 1, \dots, k$,*

$$E\left(\sup_{\theta \in \Theta} \mu_i(X_t, \theta)^2\right) < \infty, \quad E\left(\sup_{\theta \in \Theta} \left\| \frac{\partial \mu_i(X_t, \theta)}{\partial \theta^T} \right\| \right) < \infty, \quad E\left(\sup_{\theta \in \Theta} \left\| \frac{\partial^2 \mu_i(X_t, \theta)}{\partial \theta \partial \theta^T} \right\| \right) < \infty.$$

In the latter case one should interpret the matrix norm $\|\cdot\|$ as the maximum of the absolute values of the elements of the matrix involved. Moreover, in order for the parameter vector θ_0 to be identified, we need to ensure that

Assumption D: $\|E[m_t(\theta)]\| = 0$ if and only if $\theta = \theta_0$.

2.2. Strong consistency

Denote

$$M_n(\theta) = \frac{1}{n} \sum_{t=1}^n m_t(\theta), \quad \bar{M}(\theta) = E[m_t(\theta)]. \quad (51)$$

Under Assumptions A-C we have

$$\sup_{\theta \in \Theta} \|M_n(\theta) - \bar{M}(\theta)\| \rightarrow 0 \text{ a.s.}, \quad (52)$$

hence, denoting

$$Q_n(\theta) = M_n(\theta)^T W_n M_n(\theta), \quad \bar{Q}(\theta) = \bar{M}(\theta)^T W \bar{M}(\theta), \quad (53)$$

where

$$W_n \rightarrow W \text{ a.s.}, \quad \text{with } W \text{ a positive definite symmetric matrix}, \quad (54)$$

it follows that

$$\sup_{\theta \in \Theta} |Q_n(\theta) - \bar{Q}(\theta)| \rightarrow 0 \text{ a.s..} \quad (55)$$

This result, together with Assumption D, imply that

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} Q_n(\theta) \rightarrow \theta_0 \text{ a.s.} \quad (56)$$

(Exercise: Why?)

3.3. Asymptotic normality

Assumptions A-C are also sufficient conditions for the application of the central limit theorem:

$$\frac{1}{\sqrt{n}} \sum_{t=1}^n m_t(\theta_0) \rightarrow N_q[0, A] \text{ in distr., where } A = E[m_t(\theta_0)m_t(\theta_0)^T]. \quad (57)$$

The limiting normal distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ can be derived as follows. The first-order conditions for a minimum of $Q_n(\theta)$ are:

$$\frac{\partial Q_n(\theta)}{\partial \theta_i} = 2 \left(\frac{\partial M_n(\theta)^T}{\partial \theta_i} \right) W_n M_n(\theta) = 2 \left(\frac{1}{n} \sum_{t=1}^n \frac{\partial m_t(\theta)^T}{\partial \theta_i} \right) W_n M_n(\theta) = 0 \quad (58)$$

for $i = 1, \dots, p$. If $\hat{\theta}$ is on the border of the parameter space Θ , these first-order conditions may not

hold for $\hat{\theta}$. However, since by Assumption B θ_0 is an interior point of Θ , and $\hat{\theta} \rightarrow \theta_0$ a.s., we have that

$$P\left[\left.\left((1/n)\sum_{t=1}^n(\partial m_t(\theta)^T/\partial\theta_i)\right)W_nM_n(\theta)\right|_{\theta=\hat{\theta}} = 0\right] \rightarrow 1 \quad (59)$$

Next observe that by the mean value theorem, and the convexity of the parameter space Θ , there exists a mean value $\tilde{\theta}_i \in \Theta$, with $\|\tilde{\theta}_i - \theta_0\| \leq \|\hat{\theta} - \theta_0\|$, such that for $i = 1, \dots, p$,

$$\begin{aligned} \sqrt{n}\left.\left((1/n)\sum_{t=1}^n(\partial m_t(\theta)^T/\partial\theta_i)\right)W_nM_n(\theta)\right|_{\theta=\hat{\theta}} &= \sqrt{n}\left.\left((1/n)\sum_{t=1}^n(\partial m_t(\theta)^T/\partial\theta_i)\right)W_nM_n(\theta)\right|_{\theta=\theta_0} \\ &+ \frac{\partial}{\partial\theta}\left.\left((1/n)\sum_{t=1}^n(\partial m_t(\theta)^T/\partial\theta_i)\right)W_nM_n(\theta)\right|_{\theta=\tilde{\theta}_i} \sqrt{n}(\hat{\theta} - \theta_0). \end{aligned} \quad (60)$$

It follows from (59) that the left-hand side of (60) converges in probability to zero. Moreover, since

$$\begin{aligned} \frac{\partial}{\partial\theta_j}\left.\left((1/n)\sum_{t=1}^n(\partial m_t(\theta)^T/\partial\theta_i)\right)W_nM_n(\theta)\right| &= \left((1/n)\sum_{t=1}^n\frac{\partial^2 m_t(\theta)^T}{\partial\theta_i\partial\theta_j}\right)W_nM_n(\theta) \\ &+ \left.\left((1/n)\sum_{t=1}^n(\partial m_t(\theta)^T/\partial\theta_i)\right)W_n\left((1/n)\sum_{t=1}^n(\partial m_t(\theta)/\partial\theta_j)\right)\right| \\ &\rightarrow \left(E\left[\frac{\partial^2 m_t(\theta)^T}{\partial\theta_i\partial\theta_j}\right]\right)W\bar{M}(\theta) + \left(E(\partial m_t(\theta)^T/\partial\theta_i)\right)W\left(E(\partial m_t(\theta)/\partial\theta_j)\right) \text{ a.s., uniformly on } \Theta, \end{aligned} \quad (61)$$

and $\tilde{\theta}_i \rightarrow \theta_0$ a.s., it follow that

$$\tilde{C} = \begin{pmatrix} \frac{\partial}{\partial\theta}\left.\left((1/n)\sum_{t=1}^n(\partial m_t(\theta)^T/\partial\theta_1)\right)W_nM_n(\theta)\right|_{\theta=\tilde{\theta}_1} \\ \vdots \\ \frac{\partial}{\partial\theta}\left.\left((1/n)\sum_{t=1}^n(\partial m_t(\theta)^T/\partial\theta_p)\right)W_nM_n(\theta)\right|_{\theta=\tilde{\theta}_p} \end{pmatrix} \rightarrow BWB^T \text{ a.s.,} \quad (62)$$

(Exercise: Why?) where

$$B = E \left[\frac{\partial m_t(\theta)}{\partial \theta} \Bigg|_{\theta=\theta_0} \right]. \quad (63)$$

Furthermore, it follows from the strong law of large numbers that

$$(1/n) \sum_{t=1}^n (\partial m_t(\theta)^T / \partial \theta) \Bigg|_{\theta=\theta_0} \rightarrow B \text{ a.s.} \quad (64)$$

hence it follows from (54) and (57) that

$$\sqrt{n} \left((1/n) \sum_{t=1}^n (\partial m_t(\theta)^T / \partial \theta) \right) W_n M_n(\theta) \Bigg|_{\theta=\theta_0} \rightarrow N_p(0, BWB^T) \text{ in distr.} \quad (65)$$

(Exercise: Why?) Combining the results (59), (60), (62), and (65), now yield (Exercise: Why?)

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N_p[0, (BWB^T)^{-1}(BWAWB^T)(BWB^T)^{-1}] \text{ in distr.} \quad (66)$$

provided that BWB^T is nonsingular. Again, the variance matrix involved is the smallest for $W = A^{-1}$, provided of course that

Assumption E: *The matrix $BA^{-1}B^T$ is nonsingular.*

Thus, $W_n = \hat{A}^{-1}$ is an optimal choice.

Finally, observe that under Assumptions A-D,

$$\hat{A} = \frac{1}{n} \sum_{t=1}^n m_t(\theta) m_t(\theta)^T \Bigg|_{\theta=\hat{\theta}} \rightarrow A \text{ a.s.}, \quad \hat{B} = \frac{1}{n} \sum_{t=1}^n \frac{\partial m_t(\theta)}{\partial \theta} \Bigg|_{\theta=\hat{\theta}} \rightarrow B \text{ a.s..} \quad (67)$$

(Exercise: Why?) Thus, under Assumptions A-E the *efficient method of moment estimator*

$$\hat{\theta}_{EMM} = \underset{\theta}{\operatorname{argmin}} M_n(\theta)^T \hat{A}^{-1} M_n(\theta), \quad (68)$$

is strongly consistent: $\hat{\theta}_{EMM} \rightarrow \theta_0$ a.s., and has limiting normal distribution:

$$\sqrt{n}(\hat{\theta}_{EMM} - \theta_0) \rightarrow N_p[0, (BA^{-1}B^T)^{-1}]. \quad (69)$$

Moreover,

$$(\hat{B}\hat{A}^{-1}\hat{B}^T)^{-1} \rightarrow (BA^{-1}B^T)^{-1} \text{ a.s.} \quad (70)$$

Finally, observe that similarly to the linear case, under Assumptions A-E and the null hypothesis $H_0: E[m_t(\theta_0)] = 0$ for some $\theta_0 \in \Theta$,

$$nM_n(\hat{\theta}_{EMM})^T \hat{A}^{-1} M_n(\hat{\theta}_{EMM}) \rightarrow \chi_{q-p}^2 \text{ in distr.,} \quad (71)$$

whereas under the alternative hypothesis that the null hypothesis is incorrect, and the maintained Assumptions A, B,C,E,

$$nM_n(\hat{\theta}_{EMM})^T \hat{A}^{-1} M_n(\hat{\theta}_{EMM}) \rightarrow \infty \text{ a.s.} \quad (72)$$

Thus, the left-hand side of (71) is the test statistic of the Wald test that the moment conditions involved are correct.

The Uniform Weak Law of Large Numbers and the Consistency of M-Estimators of Cross-Section and Time Series Models

Herman J. Bierens

Pennsylvania State University

September 16, 2005

1. *The uniform weak law of large numbers*

In econometrics we often have to deal with sample means of random functions. A random function is a function that is a random variable for each fixed value of its argument. In cross-section econometrics random functions usually take the form of a function $g(Z, \theta)$ of a random vector Z and a non-random vector θ . For example, consider a Logit model:

$$P[Y_j = y | X_j] = \frac{y + (1 - y)\exp(-\alpha - \beta^T X_j)}{1 + \exp(-\alpha - \beta^T X_j)}, \quad y = 0,1,$$

where $Y_j \in \{0,1\}$ is the dependent variable and $X_j \in \mathbb{R}^k$ is a vector of explanatory variables. Denoting $Z_j = (Y_j, X_j^T)^T$, and given a random sample $\{Z_1, Z_2, \dots, Z_n\}$, the log-likelihood function involved takes the form $\sum_{j=1}^n g(Z_j, \theta)$, where

$$\begin{aligned} g(Z_j, \theta) &= \ln(Y_j + (1 - Y_j)\exp(-\alpha - \beta^T X_j)) - \ln(1 + \exp(-\alpha - \beta^T X_j)) \\ &= Y_j(\alpha + \beta^T X_j) - \ln(1 + \exp(\alpha + \beta^T X_j)), \text{ where } \theta = (\alpha, \beta^T)^T. \end{aligned} \tag{1}$$

For such functions we can extend the weak law of large numbers for i.i.d. random variables to a Uniform Weak Law of Large Numbers (UWLLN):

Theorem 1: Let $Z_j, j = 1, \dots, n$, be a random sample from a k -variate distribution. Let $g(z, \theta)$ be a Borel measurable function on $\mathbf{Z} \times \Theta$, where $\mathbf{Z} \subset \mathbb{R}^k$ is a Borel set such that $P[Z_t \in \mathbf{Z}] = 1$, and Θ is a compact subset of \mathbb{R}^m , such that for each $z \in \mathbf{Z}$, $g(z, \theta)$ is a continuous function on Θ . Furthermore, let

$$E[\sup_{\theta \in \Theta} |g(Z_j, \theta)|] < \infty. \quad (2)$$

Then $\text{plim}_{n \rightarrow \infty} \sup_{\theta \in \Theta} |(1/n) \sum_{j=1}^n g(Z_j, \theta) - E[g(Z_1, \theta)]| = 0$.

Note that subsets of Euclidean spaces are compact if and only if they are closed and bounded. See, for example, Bierens (2004), Appendix II, Theorem II.2.

The original proof of the stronger result

$$\sup_{\theta \in \Theta} |(1/n) \sum_{j=1}^n g(Z_j, \theta) - E[g(Z_1, \theta)]| \rightarrow 0 \text{ a.s.},$$

was given in the seminal paper of Jennrich (1969). This proof is explained in detail in Bierens (2004, Appendix to Chapter 6).

The condition that the random vectors Z_j are i.i.d. can be relaxed, because the result in Theorem 1 also holds for strictly stationary time series processes with a vanishing memory:

Definition 1: A (vector) time series process $X_t \in \mathbb{R}^k$ is strictly stationary if for arbitrary integers $m_1 < m_2 < \dots < m_n$ the joint distribution of $\begin{pmatrix} X_{t-m_1}^T, \dots, X_{t-m_n}^T \end{pmatrix}^T$ does not depend on the time index t .

Definition 2: A (vector) time series process $X_t \in \mathbb{R}^k$ has a vanishing memory if all the sets in the remote σ -algebra $\mathcal{F}_{-\infty} = \bigcap_t \sigma(\{X_{t-j}\}_{j=0}^\infty)$ have either probability zero or one.

Note that if the X_t 's are independent then by Kolmogorov's zero-one law the time series X_t has a vanishing memory.

It has been shown in Bierens (2004, Theorem 7.4) that

Theorem 2: If $X_t \in \mathbb{R}^k$ is a strictly stationary time series process with vanishing memory, and $E[\|X_t\|] < \infty$, then $\text{plim}_{n \rightarrow \infty} (1/n) \sum_{t=1}^n X_t = E[X_1]$.

I will use this result to prove the following more general version of Theorem 1. To be able to generalize the UWLLN to the time series case where the random functions involved depend on

the entire past of the time series rather than on a finite dimensional vector of variables, I will reformulate and prove Theorem 1 under slightly different moment conditions.

Theorem 3: Let $Z_t \in \mathbb{R}^k$ be a strictly stationary vector time series process with a vanishing memory,¹ defined on a common probability space $\{\Omega, \mathcal{F}, P\}$. Let $g(z, \theta)$ be a Borel measurable real function on $\mathbf{Z} \times \Theta_0$, where $\mathbf{Z} \subset \mathbb{R}^k$ is a Borel set such that $P[Z_t \in \mathbf{Z}] = 1$, and Θ_0 is an open subset of \mathbb{R}^m , such that for each $z \in \mathbf{Z}$, $g(z, \theta)$ is a continuous function on Θ_0 . Furthermore, let Θ be a compact subset of Θ_0 . Finally, assume that for each $\theta_* \in \Theta$ there exists an arbitrary small $\delta > 0$, possibly depending on θ_* , such that

$$E\left[\sup_{\|\theta - \theta_*\| \leq \delta} g(Z_1, \theta)\right] < \infty, \quad E\left[\inf_{\|\theta - \theta_*\| \leq \delta} g(Z_1, \theta)\right] > -\infty. \quad (3)$$

Then $\text{plim}_{n \rightarrow \infty} \sup_{\theta \in \Theta} |(1/n) \sum_{j=1}^n g(Z_j, \theta) - E[g(Z_1, \theta)]| = 0$.

Proof: Observe from condition (3) that for each $\theta \in \Theta$, $E[g(Z_1, \theta)]$ is well-defined. Actually, due to the compactness of Θ , (3) implies (2) [Exercise: Why?], so that the latter is a weaker condition than (3). Moreover, it follows from condition (3), the continuity of $g(z, \theta)$ in θ , and the dominated convergence theorem, that

$$\lim_{\delta \downarrow 0} E\left[\sup_{\|\theta - \theta_*\| \leq \delta} g(Z_1, \theta) - \inf_{\|\theta - \theta_*\| \leq \delta} g(Z_1, \theta)\right] = 0, \quad (4)$$

pointwise in $\theta_* \in \Theta$. Therefore, for an arbitrary $\varepsilon > 0$ and each $\theta_* \in \Theta$ we can choose a positive number $\delta(\theta_*, \varepsilon)$ such that, with

$$N(\theta_* | \varepsilon) = \{\theta \in \Theta_0 : \|\theta - \theta_*\| < \delta(\theta_*, \varepsilon)\}, \quad (5)$$

we have

$$0 \leq E\left[\sup_{\theta \in N(\theta_* | \varepsilon)} g(Z_1, \theta) - \inf_{\theta \in N(\theta_* | \varepsilon)} g(Z_1, \theta)\right] < \varepsilon. \quad (6)$$

Next, observe that the sets (5) are open, so that $\bigcup_{\theta_* \in \Theta} N(\theta_* | \varepsilon)$ is an open covering of Θ . Then by the compactness of Θ there exists a finite sub-covering of Θ :

¹

Which includes the case that the Z_t 's are i.i.d.

$$\Theta \subset \bigcup_{i=1}^K N(\theta_i | \varepsilon), \quad (7)$$

where K and the vectors $\theta_i \in \Theta$ depend on ε .

Using the easy inequality $\sup_x |f(x)| \leq |\sup_x f(x)| + |\inf_x f(x)|$, it is not hard to verify that for each $\theta_i \in \Theta$,

$$\begin{aligned} & \sup_{\theta \in N(\theta_i | \varepsilon)} |(1/n) \sum_{t=1}^n g(Z_t, \theta) - E[g(Z_1, \theta)]| \\ & \leq 2|(1/n) \sum_{t=1}^n \sup_{\theta \in N(\theta_i | \varepsilon)} g(Z_t, \theta) - E[\sup_{\theta \in N(\theta_i | \varepsilon)} g(Z_1, \theta)]| \\ & \quad + 2|(1/n) \sum_{t=1}^n \inf_{\theta \in N(\theta_i | \varepsilon)} g(Z_t, \theta) - E[\inf_{\theta \in N(\theta_i | \varepsilon)} g(Z_1, \theta)]| \\ & \quad + 2(E[\sup_{\theta \in N(\theta_i | \varepsilon)} g(Z_1, \theta)] - E[\inf_{\theta \in N(\theta_i | \varepsilon)} g(Z_1, \theta)]). \end{aligned} \quad (8)$$

It follows from Theorem 2 that the first two terms at the right-hand side of (8) converge in probability to zero, and from (6) that the last term is less than 2ε . Hence,

$$\begin{aligned} & \sup_{\theta \in \Theta} |(1/n) \sum_{t=1}^n g(Z_t, \theta) - E[g(Z_1, \theta)]| \\ & \leq \max_{1 \leq i \leq K} \sup_{\theta \in N(\theta_i | \varepsilon)} |(1/n) \sum_{t=1}^n g(Z_t, \theta) - E[g(Z_1, \theta)]| \\ & \leq R_n(\varepsilon) + 2\varepsilon, \text{ where } \text{plim}_{n \rightarrow \infty} R_n(\varepsilon) = 0. \end{aligned} \quad (9)$$

Theorem 3 follows now straightforwardly from (9). Q.E.D.

In time series econometrics there are quite a few cases where we need a UWLLN for functions $g(., \theta)$ depending on Z_{t-j} for all $j \geq 0$. In that case $g(., \theta)$ takes a more general form as a random function:

Definition 3: Let $\{\Omega, \mathcal{F}, P\}$ be the probability space. A random function $f(\theta)$ on a subset Θ of a Euclidean space is a mapping $f(\omega, \theta): \Omega \times \Theta \rightarrow \mathbb{R}$ such that for each Borel set B in \mathbb{R} and each $\theta \in \Theta$, $\{\omega \in \Omega: f(\omega, \theta) \in B\} \in \mathcal{F}$.

Definition 4: A random function $f(\theta)$ on a subset Θ of a Euclidean space is almost surely continuous on Θ if there exists a set A with probability one such that for each $\omega \in A$, $f(\omega, \theta)$ is continuous in $\theta \in \Theta$.

For example, let $Z_t \in \mathbb{R}$ be a stationary Gaussian moving average process of order 1 [alias an MA(1) process]:

$$Z_t = U_t - \alpha_0 U_{t-1}, \quad |\alpha_0| < 1, \quad U_t \sim i.i.d. \ N(0, \sigma_0^2). \quad (10)$$

Then backwards substitution of $U_t = \alpha_0 U_{t-1} + Z_t$ yields $U_t = \sum_{j=0}^{\infty} \alpha_0^j Z_{t-j}$, hence

$$Z_t = -\sum_{j=1}^{\infty} \alpha_0^j Z_{t-1} + U_t \quad (11)$$

Thus, denoting $\mathcal{F}_t = \sigma(U_t, U_{t-1}, U_{t-2}, \dots)$, the distribution of Z_t conditional on \mathcal{F}_{t-1} is normal with conditional expectation $-\sum_{j=1}^{\infty} \alpha_0^j Z_{t-1}$ and conditional variance σ_0^2 .

If the Z_t 's were observable for all $t \leq n$, a version of the log-likelihood would take the form $\sum_{j=1}^n g_t(\theta)$, where

$$g_t(\theta) = -\frac{1}{2\sigma^2} \left(\sum_{j=0}^{\infty} \alpha_0^j Z_{t-j} \right)^2 - \frac{1}{2} \ln(\sigma^2) - \ln(\sqrt{2\pi}), \quad \theta = (\alpha, \sigma^2)^T, \quad (12)$$

is a random function. In that case we need to reformulate Theorem 3 as follows.

Theorem 4: Let $\mathcal{F}_t = \sigma(U_t, U_{t-1}, U_{t-2}, \dots)$, where U_t is a time series process with vanishing memory. Let $g_t(\theta)$ be a sequence of a.s. continuous random function on an open subset Θ_0 of a Euclidean space, and let Θ be a compact subset of Θ_0 . If for each $\theta_* \in \Theta$ there exists an arbitrarily small $\delta > 0$ such that

- (a) $g_t(\theta_*)$, $\sup_{\|\theta - \theta_*\| \leq \delta} g_t(\theta)$ and $\inf_{\|\theta - \theta_*\| \leq \delta} g_t(\theta)$ are measurable \mathcal{F}_t and strictly stationary,
 - (b) $E[\sup_{\|\theta - \theta_*\| \leq \delta} g_t(\theta)] < \infty$, $E[\inf_{\|\theta - \theta_*\| \leq \delta} g_t(\theta)] > -\infty$,
- then $\text{plim}_{n \rightarrow \infty} \sup_{\theta \in \Theta} |(1/n) \sum_{t=1}^n g_t(\theta) - E[g_t(\theta)]| = 0$.

2. Consistency of M-estimators

Theorems 3 and 4 are important tools for proving consistency of parameter estimators. A large class of estimators are obtained by maximizing or minimizing an objective function of the form $(1/n)\sum_{t=1}^n g_t(\theta)$, for example maximum likelihood estimators or nonlinear least squares estimators. These estimators are called M-estimators (where the M indicates that the estimator is obtained by Maximizing or Minimizing a Mean of random functions).

Suppose that the conditions of Theorem 4 are satisfied, and that the parameter vector of interest is

$$\theta_0 = \operatorname{argmax}_{\theta \in \Theta} E[g_1(\theta)]. \quad (13)$$

Note that "argmax" is a short-hand notation for the argument for which the function involved is maximal. Then it seems a natural choice to use

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} (1/n)\sum_{t=1}^n g_t(\theta) \quad (14)$$

as an estimator of θ_0 . Indeed, under some mild conditions the estimator involved is consistent:

Theorem 5: (Consistency of M-estimators) Let $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \hat{Q}(\theta)$ and $\theta_0 = \operatorname{argmax}_{\theta \in \Theta} \bar{Q}(\theta)$, where $\hat{Q}(\theta) = (1/n)\sum_{t=1}^n g_t(\theta)$ and $\bar{Q}(\theta) = E[\hat{Q}(\theta)] = E[g_1(\theta)]$. If θ_0 is unique then under the conditions of Theorem 4, $\operatorname{plim}_{n \rightarrow \infty} \hat{\theta} = \theta_0$.

Proof. Since a continuous function on a compact set takes its maximum value in this set [see, for example, Bierens (2004, Appendix II)], it follows that $\hat{\theta} \in \Theta$ and $\theta_0 \in \Theta$. Moreover, by the same result it follows from the continuity of $\bar{Q}(\theta)$ and the uniqueness of θ_0 that for every $\varepsilon > 0$ for which the set $\{\theta \in \Theta: \|\theta - \theta_0\| \geq \varepsilon\}$ is non-empty,

$$\bar{Q}(\theta_0) > \sup_{\theta \in \Theta, \|\theta - \theta_0\| \geq \varepsilon} \bar{Q}(\theta) \quad (15)$$

[Exercise: Why?] Now by the definition of θ_0 ,

$$\begin{aligned}
0 &\leq \bar{Q}(\theta_0) - \bar{Q}(\hat{\theta}) = \bar{Q}(\theta_0) - \hat{Q}(\theta_0) + \hat{Q}(\theta_0) - \bar{Q}(\hat{\theta}) \\
&\leq \bar{Q}(\theta_0) - \hat{Q}(\theta_0) + \hat{Q}(\hat{\theta}) - \bar{Q}(\hat{\theta}) \leq 2\sup_{\theta \in \Theta} |\hat{Q}(\theta) - \bar{Q}(\theta)|,
\end{aligned} \tag{16}$$

and it follows from Theorem 4 that the right-hand side of (16) converges in probability to zero.

Thus:

$$\text{plim}_{n \rightarrow \infty} \bar{Q}(\hat{\theta}) = \bar{Q}(\theta_0). \tag{17}$$

Moreover, (15) implies that for arbitrary $\varepsilon > 0$ there exists a $\delta > 0$ such that $\bar{Q}(\theta_0) - \bar{Q}(\hat{\theta}) \geq \delta$ if $\|\hat{\theta} - \theta_0\| \geq \varepsilon$, hence

$$P(\|\hat{\theta} - \theta_0\| > \varepsilon) \leq P(\bar{Q}(\theta_0) - \bar{Q}(\hat{\theta}) \geq \delta). \tag{18}$$

Combining (17) and (18), the theorem under review follows. Q.E.D.

It is easy to verify that Theorem 5 carries over to the "argmin" case.

References

- Bierens, H. J. (2004): *Introduction to the Mathematical and Statistical Foundations of Econometrics*, Cambridge University Press, Cambridge, U.K.
- Jennrich, R. I. (1969): “Asymptotic Properties of Non-Linear Least Squares Estimators”, *Annals of Mathematical Statistics* 40, 633-643.

The Wold Decomposition

Herman J. Bierens*

February 1, 2012

Abstract

In Chapter 7 in Bierens (2004) the Wold decomposition was motivated by claiming that every zero-mean covariance stationary process X_t can be written as $X_t = \sum_{j=1}^{\infty} \beta_j X_{t-j} + U_t$, where $E[U_t X_{t-j}] = 0$ for all $j \geq 1$, and $\sum_{j=1}^{\infty} \beta_j X_{t-j}$ is the projection of X_t on its past. However, in general this claim is incorrect. In this note I will give a more general (and hopefully correct) proof of the Wold decomposition.

1 Projections on spaces spanned by a sequence

The fundamental projection theorem states that:

Theorem 1. *Given a sub-Hilbert space \mathcal{S} of a Hilbert space \mathcal{H} and an element $y \in \mathcal{H}$, there exists a unique element $\hat{y} \in \mathcal{S}$ such that $\|y - \hat{y}\| = \inf_{z \in \mathcal{S}} \|y - z\|$. Moreover the residual $u = y - \hat{y}$ is orthogonal to any $z \in \mathcal{S}$: $\langle u, z \rangle = 0$.*

Proof: See for example Bierens (2004, Th. 7.A.3, p. 202).

This result is the basis for the famous Wold (1938) decomposition for covariance stationary time series, which in its turn is the basis for time series analysis.

*Thanks to Peter Boswijk (University of Amsterdam) for pointing out an error in a previous version of this note. Moreover, the queries of the students in my graduate time series courses have led to substantial improvements of the proof of the Wold decomposition.

The proof of the Wold decomposition in Anderson (1994) is more transparent than the original proof by Wold (1938). However, rather than following Anderson's proof, I will in this note derive first a general Wold decomposition for a regular sequence¹ in a general Hilbert space, and then specialize this result to the Wold decomposition for covariance stationary time series.

First, we need to define sub-Hilbert spaces spanned by a sequence in a Hilbert space, as follows.

Let $\{x_k\}_{k=1}^\infty$ be a sequence of elements of a Hilbert space \mathcal{H} , and let

$$\mathcal{M}_m = \text{span}(\{x_j\}_{j=1}^m)$$

be the space spanned by x_1, \dots, x_m , i.e., \mathcal{M}_m consists of all linear combinations of x_1, \dots, x_m . Then

Lemma 1. \mathcal{M}_m is a Hilbert space.

Proof: Without loss of generality we may assume that the $m \times m$ matrix Σ_m with elements $\langle x_i, x_j \rangle$, $i, j = 1, \dots, m$, is non-singular, as otherwise we can remove one or more x_j 's from the list $\{x_j\}_{j=1}^m$ and still span the same space. For example, suppose that $\text{rank}(\Sigma_m) = m - 1$, and let $c = (c_1, \dots, c_m)'$ be the eigenvector corresponding to the zero eigenvalue. Then $\left\| \sum_{j=1}^m c_j x_j \right\|^2 = c' \Sigma_m c = 0$, hence $\sum_{j=1}^m c_j x_j = 0$ (the latter being the zero element of \mathcal{M}_m). Since at least one component of c is non-zero, for example c_i , we can write

$$x_i = \begin{cases} -\sum_{j=2}^m (c_j/c_1)x_j & \text{if } i = 1, \\ -\sum_{j=1}^{m-1} (c_j/c_m)x_j & \text{if } i = m, \\ -\sum_{j=1}^{i-1} (c_j/c_i)x_j - \sum_{j=i+1}^m (c_j/c_i)x_j & \text{if } 1 < i < m, \end{cases}$$

so that

$$\mathcal{M}_m = \begin{cases} \text{span}(\{x_j\}_{j=2}^m) & \text{if } i = 1, \\ \text{span}(\{x_j\}_{j=1}^{m-1}) & \text{if } i = m, \\ \text{span}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_m) & \text{if } 1 < i < m. \end{cases}$$

Now let $z_n = \sum_{j=1}^m \beta_{j,n} x_j$ be a Cauchy sequence in \mathcal{M}_m , and denote $\beta_n = (\beta_{1,n}, \dots, \beta_{m,n})'$. Then for each j , $\beta_{j,n}$ is a Cauchy sequence in \mathbb{R} because

$$0 = \lim_{\min(n_1, n_2) \rightarrow \infty} \|z_{n_1} - z_{n_2}\|^2 = \lim_{\min(n_1, n_2) \rightarrow \infty} \left\| \sum_{j=1}^m (\beta_{j,n_1} - \beta_{j,n_2}) x_j \right\|^2$$

¹ See Definition 4 below.

$$\begin{aligned}
&= \lim_{\min(n_1, n_2) \rightarrow \infty} (\beta_{n_1} - \beta_{n_2})' \Sigma_m (\beta_{n_1} - \beta_{n_2}) \\
&\geq \lambda_{\min}(\Sigma_m) \cdot \lim_{\min(n_1, n_2) \rightarrow \infty} \|\beta_{n_1} - \beta_{n_2}\|^2,
\end{aligned}$$

where $\lambda_{\min}(\Sigma_m) > 0$ is the smallest eigenvalue of Σ_m . Consequently, $\lim_{n \rightarrow \infty} \|z - z_n\| = 0$, where $z = \sum_{j=1}^m \beta_j x_j \in \mathcal{M}_m$ with $\beta_j = \lim_{n \rightarrow \infty} \beta_{j,n}$. Q.E.D.

Definition 1. The space $\mathcal{M}_\infty = \overline{\cup_{n=1}^\infty \mathcal{M}_n}$ (which is the closure of $\cup_{n=1}^\infty \mathcal{M}_n$) is called the space spanned by $\{x_j\}_{j=1}^\infty$, and is also denoted by $\text{span}(\{x_j\}_{j=1}^\infty)$.

Lemma 2. \mathcal{M}_∞ is a Hilbert space.

Proof: Let z_m be a Cauchy sequence in \mathcal{M}_∞ , with limit $\bar{z} \in \mathcal{H}$. If $\bar{z} \notin \mathcal{M}_\infty$ then, because \mathcal{M}_∞ is closed, there exists an $\varepsilon > 0$ such that the set $\{z \in \mathcal{H} : \|z - \bar{z}\| < \varepsilon\}$ is completely outside \mathcal{M}_∞ : $\{z \in \mathcal{H} : \|z - \bar{z}\| < \varepsilon\} \cap \mathcal{M}_\infty = \emptyset$. But then there exists an m such that $z_m \notin \mathcal{M}_\infty$. Since this is impossible, $\bar{z} \in \mathcal{M}_\infty$. Q.E.D.

Next, let us focus on projections on a space spanned by a sequence in a Hilbert space.

Lemma 3. For $z \in \mathcal{M}_\infty$ let \hat{z}_n be the projection of z on \mathcal{M}_n . Then $\lim_{n \rightarrow \infty} \|z - \hat{z}_n\| = 0$.

Proof: If $z \in \cup_{n=1}^\infty \mathcal{M}_n$ then there exists an n_0 such that $z \in \mathcal{M}_{n_0}$, hence for $n \geq n_0$, $\hat{z}_n = z$ and thus $\lim_{n \rightarrow \infty} \|z - \hat{z}_n\| = 0$. Now let $z \in \mathcal{M}_\infty \setminus (\cup_{n=1}^\infty \mathcal{M}_n)$. Since $\mathcal{M}_\infty = \overline{\cup_{n=1}^\infty \mathcal{M}_n}$ is closed and $\mathcal{M}_n \subset \mathcal{M}_{n+1}$, for each n there exists an $z_n \in \mathcal{M}_n$ such that $\lim_{n \rightarrow \infty} \|z - z_n\|^2 = 0$, hence for $n \rightarrow \infty$, $\|z - \hat{z}_n\|^2 \leq \|z - z_n\|^2 \rightarrow 0$. Q.E.D.

More generally we have:

Theorem 2. For $z \in \mathcal{H}$, let \hat{z} be the projection of z on $\mathcal{M}_\infty = \text{span}(\{x_j\}_{j=1}^\infty)$ and let \hat{z}_n be the projection of z on $\mathcal{M}_n = \text{span}(\{x_j\}_{j=1}^n)$. Then $\lim_{n \rightarrow \infty} \|\hat{z} - \hat{z}_n\| = 0$.

Proof: We may without loss of generality assume that $\hat{z} \in \mathcal{M}_\infty \setminus (\cup_{n=1}^\infty \mathcal{M}_n)$, as otherwise the result of Theorem 2 holds trivially. Since \mathcal{M}_∞ is closed this

assumption implies that for each n we can select a $z_n \in \mathcal{M}_n$ such that

$$\lim_{n \rightarrow \infty} \|\widehat{z} - z_n\| = 0. \quad (1)$$

Let $\|z - \widehat{z}\| = \delta$ and $\|z - \widehat{z}_n\| = \delta_n$, and note that $\delta_n \geq \delta$. Since

$$\begin{aligned} \delta_n^2 &= \|z - \widehat{z}_n\|^2 \leq \|z - z_n\|^2 = \|z - \widehat{z} + \widehat{z} - z_n\|^2 \\ &= \|z - \widehat{z}\|^2 + \|\widehat{z} - z_n\|^2 + 2 \langle z - \widehat{z}, \widehat{z} - z_n \rangle \\ &= \delta^2 + \|\widehat{z} - z_n\|^2 \end{aligned}$$

it follows from (1) that

$$\lim_{n \rightarrow \infty} \delta_n = \delta. \quad (2)$$

Recall that $z = \widehat{z} + u$, where $\langle u, x \rangle = 0$ for all $x \in \mathcal{M}_\infty$. Hence

$$\begin{aligned} \|\widehat{z} - \widehat{z}_n\|^2 &= \|z - \widehat{z}_n - u\|^2 = \|z - \widehat{z}_n\|^2 + \|u\|^2 - 2 \langle z - \widehat{z}_n, u \rangle \\ &= \|z - \widehat{z}_n\|^2 + \|u\|^2 - 2 \langle z, u \rangle = \delta_n^2 - \delta^2 \end{aligned} \quad (3)$$

where the last equality follows from

$$\langle z, u \rangle - \langle u, u \rangle = \langle \widehat{z}, u \rangle = 0 \quad (4)$$

and $\langle u, u \rangle = \|u\|^2 = \delta^2$. The theorem now follows from (2) and (3). Q.E.D.

Remark 1. Although each projection \widehat{z}_n is a linear combination of x_1, \dots, x_n , in general the result of Theorem 2 does **not** imply that there exists a sequence $\{\theta_j\}_{j=1}^\infty$ such that $\widehat{z} = \sum_{j=1}^\infty \theta_j x_j$.

As a counter example, consider the Hilbert space \mathcal{R}_0 of zero-mean random variables with finite second moments, endowed with the inner product $\langle X, Y \rangle = E[X.Y]$ and associated norm and metric. Let

$$X_t = V_t - V_{t-1},$$

where the V_t 's are independent $N(0, 1)$ distributed. This is clearly a zero-mean covariance stationary process, with covariance function $\gamma(0) = 2$, $\gamma(1) = -1$, $\gamma(m) = 0$ for $m \geq 2$. Hence $X_t \in \mathcal{R}_0$ for all t .

For given t , let

$$\mathcal{M}_{-\infty}^{t-1} = \text{span}(\{X_{t-m}\}_{m=1}^\infty), \quad \mathcal{M}_{t-n}^{t-1} = \text{span}(X_{t-1}, \dots, X_{t-n}).$$

The projection $\hat{X}_{t,n}$ of X_t on \mathcal{M}_{t-n}^{t-1} takes the form

$$\hat{X}_{t,n} = \sum_{j=1}^n \theta_{n,j} X_{t-j}$$

where the coefficients $\theta_{n,j}$ are the solutions of the normal equations

$$\gamma(m) = \sum_{k=1}^n \gamma(|k-m|) \theta_{n,k}, \quad m = 1, \dots, n.$$

hence for $n \geq 3$,

$$\begin{aligned} -1 &= 2\theta_{n,1} - \theta_{n,2} \\ 0 &= -\theta_{n,1} + 2\theta_{n,2} - \theta_{n,3} \\ 0 &= -\theta_{n,2} + 2\theta_{n,3} - \theta_{n,4} \\ &\vdots \\ 0 &= -\theta_{n,n-2} + 2\theta_{n,n-1} - \theta_{n,n} \\ 0 &= -\theta_{n,n-1} + 2\theta_{n,n} \end{aligned}$$

The solutions of these normal equations are

$$\theta_{n,j} = \frac{j}{n+1} - 1, \quad j = 1, \dots, n,$$

hence

$$\hat{X}_{t,n} = \sum_{j=1}^n \left(\frac{j}{n+1} - 1 \right) X_{t-j} \quad (5)$$

Next, let \hat{X}_t be the projection of X_t on $\mathcal{M}_{-\infty}^{t-1}$, and suppose that there exists a sequence $\{\theta_j\}_{j=1}^{\infty}$ such that $\hat{X}_t = \sum_{j=1}^{\infty} \theta_j X_{t-j}$. Note that the latter is merely a short-hand notation for

$$\lim_{n \rightarrow \infty} \left\| \hat{X}_t - \sum_{j=1}^n \theta_j X_{t-j} \right\|^2 = \lim_{n \rightarrow \infty} E \left[\left(\hat{X}_t - \sum_{j=1}^n \theta_j X_{t-j} \right)^2 \right] = 0 \quad (6)$$

If so, it follows from Theorem 2 and (5) that

$$0 = \lim_{n \rightarrow \infty} E \left[\left(\sum_{j=1}^n \theta_j X_{t-j} - \sum_{j=1}^n \left(\frac{j}{n+1} - 1 \right) X_{t-j} \right)^2 \right]$$

$$= \lim_{n \rightarrow \infty} E \left[\left(\sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) X_{t-j} \right)^2 \right] \quad (7)$$

But

$$\begin{aligned} \sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) X_{t-j} &= \sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) (V_{t-j} - V_{t-j-1}) \\ &= - \left(\frac{n}{n+1} + \theta_1 \right) V_{t-1} - \sum_{j=1}^{n-1} \left(\theta_{j+1} - \theta_j - \frac{1}{n+1} \right) V_{t-j-1} \\ &\quad + \left(\frac{1}{n+1} + \theta_n \right) V_{t-n-1} \end{aligned}$$

hence

$$\begin{aligned} E \left[\left(\sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) X_{t-j} \right)^2 \right] &= \left(\frac{n}{n+1} + \theta_1 \right)^2 \\ &\quad + \sum_{j=1}^{n-1} \left(\theta_{j+1} - \theta_j - \frac{1}{n+1} \right)^2 + \left(\frac{1}{n+1} + \theta_n \right)^2 \end{aligned} \quad (8)$$

This equality implies that for arbitrary integers $m \geq 1$,

$$\begin{aligned} \liminf_{n \rightarrow \infty} E \left[\left(\sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) X_{t-j} \right)^2 \right] &\geq \liminf_{n \rightarrow \infty} \left(\frac{n}{n+1} + \theta_1 \right)^2 + \liminf_{n \rightarrow \infty} \left(\theta_{m+1} - \theta_m - \frac{1}{n+1} \right)^2 \\ &= (\theta_1 + 1)^2 + (\theta_{m+1} - \theta_m)^2. \end{aligned}$$

Therefore, a necessary condition for (7) is that $\theta_m = -1$ for $m = 1, 2, 3, \dots$. But then it follows from (8) that

$$\lim_{n \rightarrow \infty} E \left[\left(\sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) X_{t-j} \right)^2 \right] = \lim_{n \rightarrow \infty} \left(\frac{1}{n+1} - 1 \right)^2 = 1$$

which contradicts (7). Thus, in this case there does **not** exist a sequence $\{\theta_j\}_{j=1}^\infty$ such that (6) holds.

2 Projections on the span of an orthonormal sequence

On the other hand,

Theorem 3. *If a sequence $\{x_j\}_{j=1}^\infty$ in a Hilbert space \mathcal{H} is orthonormal, i.e.,*

$$\langle x_i, x_j \rangle = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \quad (9)$$

then any projection \hat{z} of $z \in \mathcal{H}$ on $\text{span}(\{x_j\}_{j=1}^\infty)$ takes the form $\hat{z} = \sum_{j=1}^\infty \theta_j x_j$ (in the sense that $\lim_{n \rightarrow \infty} \|\hat{z} - \sum_{j=1}^n \theta_j x_j\| = 0$), where $\theta_j = \langle z, x_j \rangle$ and $\sum_{j=1}^\infty \theta_j^2 < \infty$.

Proof: Let \hat{z}_n be the projection of z on $\text{span}(\{x_j\}_{j=1}^n)$. Then

$$\begin{aligned} \|z - \hat{z}_n\|^2 &= \min_{c_1, \dots, c_n} \left\| z - \sum_{j=1}^n c_j x_j \right\|^2 \\ &= \min_{c_1, \dots, c_n} \left\{ \|z\|^2 - 2 \sum_{j=1}^n c_j \langle z, x_j \rangle + \sum_{i=1}^n \sum_{j=1}^n c_i c_j \langle x_i, x_j \rangle \right\} \\ &= \min_{c_1, \dots, c_n} \left\{ \|z\|^2 - 2 \sum_{j=1}^n c_j \langle z, x_j \rangle + \sum_{j=1}^n c_j^2 \right\}, \end{aligned}$$

hence,

$$\hat{z}_n = \sum_{j=1}^n \theta_j x_j, \text{ where } \theta_j = \langle z, x_j \rangle. \quad (10)$$

Moreover, denoting $u_n = z - \hat{z}_n$, it follows from (9) and (10) that

$$\begin{aligned} \|u_n\|^2 &= \left\| z - \sum_{j=1}^n \theta_j x_j \right\|^2 = \|z\|^2 - 2 \sum_{j=1}^n \theta_j \langle z, x_j \rangle + \sum_{j=1}^n \sum_{i=1}^n \theta_j \theta_i \langle x_j, x_i \rangle \\ &= \|z\|^2 - \sum_{j=1}^n \theta_j^2 \geq 0 \end{aligned} \quad (11)$$

so that $\sum_{j=1}^n \theta_j^2 \leq \|z\|^2$ for all n and thus $\sum_{j=1}^\infty \theta_j^2 < \infty$. Finally, it follows from Theorem 2 that

$$\lim_{n \rightarrow \infty} \left\| \hat{z} - \sum_{j=1}^n \theta_j x_j \right\| = \lim_{n \rightarrow \infty} \|\hat{z} - \hat{z}_n\| = 0.$$

Q.E.D.

3 The general Wold decomposition

3.1 Preliminary definitions and results

Let $\mathcal{S}_1, \mathcal{S}_2$ be a pair of subspaces of a Hilbert space \mathcal{H} . We say that:

Definition 2. \mathcal{S}_1 and \mathcal{S}_2 are orthogonal, denoted by $\mathcal{S}_1 \perp \mathcal{S}_2$, if for each $x_1 \in \mathcal{S}_1$ and each $x_2 \in \mathcal{S}_2$, $\langle x_1, x_2 \rangle = 0$.

Lemma 4. Let \mathcal{S}_1 and \mathcal{S}_2 be sub-Hilbert spaces satisfying $\mathcal{S}_1 \perp \mathcal{S}_2$. Then

$$\text{span}(\mathcal{S}_1, \mathcal{S}_2) = \{y = x_1 + x_2 : x_1 \in \mathcal{S}_1, x_2 \in \mathcal{S}_2\}$$

is a Hilbert space.

Proof: Let y_n be a Cauchy sequence in $\text{span}(\mathcal{S}_1, \mathcal{S}_2)$. Then $y_n = x_{1,n} + x_{2,n}$, where $x_{1,n} \in \mathcal{S}_1$ and $x_{2,n} \in \mathcal{S}_2$. Since $x_{1,n} - x_{1,m} \in \mathcal{S}_1$ and $x_{2,n} - x_{2,m} \in \mathcal{S}_2$ it follows from the orthogonality condition $\mathcal{S}_1 \perp \mathcal{S}_2$ that

$$\begin{aligned} \|y_n - y_m\|^2 &= \|(x_{1,n} - x_{1,m}) + (x_{2,n} - x_{2,m})\|^2 \\ &= \|x_{1,n} - x_{1,m}\|^2 + \|x_{2,n} - x_{2,m}\|^2 \\ &\quad + 2 \langle x_{1,n} - x_{1,m}, x_{2,n} - x_{2,m} \rangle \\ &= \|x_{1,n} - x_{1,m}\|^2 + \|x_{2,n} - x_{2,m}\|^2, \end{aligned}$$

hence $\lim_{\min(n,m) \rightarrow \infty} \|y_n - y_m\| = 0$ implies that $\lim_{\min(n,m) \rightarrow \infty} \|x_{1,n} - x_{1,m}\| = 0$ and $\lim_{\min(n,m) \rightarrow \infty} \|x_{2,n} - x_{2,m}\| = 0$. Because \mathcal{S}_1 and \mathcal{S}_2 are Hilbert spaces there exist an $x_1 \in \mathcal{S}_1$ and an $x_2 \in \mathcal{S}_2$ such that $\lim_{n \rightarrow \infty} \|x_{1,n} - x_1\| = 0$ and $\lim_{n \rightarrow \infty} \|x_{2,n} - x_2\| = 0$, hence $\lim_{n \rightarrow \infty} \|y_n - y\| = 0$, where $y = x_1 + x_2 \in \text{span}(\mathcal{S}_1, \mathcal{S}_2)$. Q.E.D.

We also need the definitions of orthogonal complement and regularity:

Definition 3. *The orthogonal complement of a subspace \mathcal{S} of a Hilbert space \mathcal{H} , denoted by \mathcal{S}^\perp , is the subset of \mathcal{H} defined by*

$$\mathcal{S}^\perp = \{y \in \mathcal{H} : \langle x, y \rangle = 0 \text{ for all } x \in \mathcal{S}\}.$$

Lemma 5. *Orthogonal complements are Hilbert spaces.*

Proof: Let $x \in \mathcal{S}$ be arbitrary and let y_n be a Cauchy sequence in \mathcal{S}^\perp . Then there exists an $y \in \mathcal{H}$ such that $\lim_{n \rightarrow \infty} \|y - y_n\| = 0$. Since $\langle x, y_n \rangle = 0$ we have $\langle x, y \rangle = \langle x, y - y_n \rangle$. It follows now from the Cauchy-Schwarz inequality that $|\langle x, y \rangle| = |\langle x, y - y_n \rangle| \leq \|x\| \cdot \|y - y_n\| \rightarrow 0$. Hence $y \in \mathcal{S}^\perp$. Q.E.D.

Note that the result of Lemma 5 does not require that \mathcal{S} is a Hilbert space, although in the application below it is.

Definition 4. *Let $\{x_k\}_{k=1}^\infty$ be a sequence in a Hilbert space \mathcal{H} . Let \hat{x}_k be the projection of x_k on $\text{span}(\{x_m\}_{m=k+1}^\infty)$, and denote $u_k = x_k - \hat{x}_k$. The sequence $\{x_k\}_{k=1}^\infty$ is called regular if $\|u_k\| > 0$ for all $k \geq 1$.*

Note that the regularity concept is related to the concept of linear independence in Euclidean spaces.

3.2 The Wold decomposition for regular sequences in a Hilbert space

We can now formulate the following general version of the Wold decomposition:

Theorem 4. *Given a regular sequence $\{x_k\}_{k=1}^\infty$ in a Hilbert space, every $x \in \mathcal{M}_\infty = \text{span}(\{x_k\}_{k=1}^\infty)$ can be written as*

$$x = \sum_{k=1}^{\infty} \alpha_k e_k + w, \quad (12)$$

in the sense that $\lim_{n \rightarrow \infty} \|x - w - \sum_{k=1}^n \alpha_k e_k\| = 0$, where $\{e_k\}_{k=1}^\infty$ is an orthonormal sequence in \mathcal{M}_∞ , $\alpha_k = \langle x, e_k \rangle$, $\sum_{k=1}^\infty \alpha_k^2 < \infty$, and

$$w \in \mathcal{S}_\infty \cap \mathcal{U}_\infty^\perp, \quad (13)$$

with $\mathcal{S}_\infty = \cap_{n=1}^\infty \text{span}(\{x_k\}_{k=n}^\infty)$ and \mathcal{U}_∞^\perp the orthogonal complement of $\mathcal{U}_\infty = \text{span}(\{e_k\}_{k=1}^\infty)$. Note that (13) implies that w is orthogonal to all the e_k 's: $\langle e_k, w \rangle = 0$ for $k = 1, 2, 3, \dots$

Proof: Denote

$$\mathcal{S}_n = \text{span}(\{x_k\}_{k=n}^\infty).$$

Note that $\mathcal{M}_\infty = \mathcal{S}_1$. Project each x_k on \mathcal{S}_{k+1} , so that $x_k = \hat{x}_k + u_k$ with projection $\hat{x}_k \in \mathcal{S}_{k+1}$ and residual u_k . Recall that by the regularity condition, $\|u_k\| > 0$, hence $e_k = u_k/\|u_k\|$ is well defined. It is not hard to verify that the residuals u_k are orthogonal, so that the e_k 's are orthonormal, and that $\mathcal{U}_\infty \subset \mathcal{M}_\infty$. It follows now from Theorem 3 that (12) holds with $\alpha_k = \langle x, e_k \rangle$, $\sum_{k=1}^\infty \alpha_k^2 < \infty$, and $w \in \mathcal{U}_\infty^\perp$, where the latter follows from the fact that w is the residual of the projection of x on \mathcal{U}_∞ . Therefore, the actual contents of Theorem 4 is that $w \in \mathcal{S}_\infty$.

The theorem under review will be proved in six steps:

Step 1. As before, let $\mathcal{M}_n = \text{span}(\{x_k\}_{k=1}^n)$. I will show first that

$$\mathcal{M}_n \subset \text{span}(\mathcal{U}_n, \mathcal{U}_n^\perp \cap \mathcal{S}_2) \quad (14)$$

[c.f. Lemma 4], where $\mathcal{U}_n = \text{span}(e_1, \dots, e_n) = \text{span}(u_1, \dots, u_n)$ and \mathcal{U}_n^\perp is the orthogonal complement of \mathcal{U}_n .

Proof. Let $z \in \mathcal{M}_n$ be arbitrary. Recall that z takes the form $z = \sum_{k=1}^n c_k x_k$. Substituting $x_k = \hat{x}_k + u_k = \hat{x}_k + \|u_k\|e_k$ we can write z as

$$\begin{aligned} z &= \sum_{k=1}^n c_k (\hat{x}_k + u_k) = \sum_{k=1}^n c_k u_k + \sum_{k=1}^n c_k \hat{x}_k \\ &= \sum_{k=1}^n c_k \|u_k\| e_k + \sum_{k=1}^n c_k \hat{x}_k \\ &= \sum_{k=1}^n c_k \|u_k\| e_k + z_2 \end{aligned}$$

where

$$z_2 = \sum_{k=1}^n c_k \hat{x}_k$$

Note that

$$z_2 = \sum_{k=1}^n c_k \hat{x}_k \in \mathcal{S}_2 \quad (15)$$

because $\hat{x}_k \in \mathcal{S}_{k+1} \subset \mathcal{S}_2$ for $k = 1, 2, \dots, n$.

Next, project z_2 on \mathcal{U}_n . This projection takes the form

$$\hat{p}_n = \sum_{k=1}^n d_k e_k, \text{ where } d_k = \langle z_2, e_k \rangle,$$

with residual

$$w_{n+1} \in \mathcal{U}_n^\perp. \quad (16)$$

However, e_1 is orthogonal to any element of \mathcal{S}_2 , and $z_2 \in \mathcal{S}_2$. Therefore, $d_1 = \langle z_2, e_1 \rangle = 0$ and thus

$$\hat{p}_n = \sum_{k=2}^n d_k e_k \in \text{span}(\{e_k\}_{k=2}^n) \subset \mathcal{S}_2,$$

where the latter follows from $e_k \in \mathcal{S}_k \subset \mathcal{S}_2$ for $k = 2, 3, \dots, n$. Because $w_{n+1} = z_2 - \hat{p}_n$ where both terms are elements of \mathcal{S}_2 , it follows that

$$w_{n+1} \in \mathcal{S}_2. \quad (17)$$

Combining (16) and (17) now yields

$$w_{n+1} \in \mathcal{U}_n^\perp \cap \mathcal{S}_2.$$

Thus, denoting $\alpha_1 = c_1 \|u_1\|$, $\alpha_k = c_k \|u_k\| + d_k$ for $k = 2, 3, \dots, n$, we can write

$$z = \sum_{k=1}^n \alpha_k e_k + w_{n+1}, \text{ where } w_{n+1} \in \mathcal{U}_n^\perp \cap \mathcal{S}_2,$$

hence $z \in \text{span}(\mathcal{U}_n, \mathcal{U}_n^\perp \cap \mathcal{S}_2)$. This proves (14).

Step 2. Next, it will be shown that

$$\text{span}(\mathcal{U}_n, \mathcal{U}_n^\perp \cap \mathcal{S}_2) = \text{span}(\mathcal{U}_n, \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1}), \quad (18)$$

so that by (14),

$$\mathcal{M}_n \subset \text{span}(\mathcal{U}_n, \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1}). \quad (19)$$

Proof. Denote for $k < m$,

$$\mathcal{S}_{k,m} = \text{span}(\{x_j\}_{j=k}^m).$$

I will show first that for $m > n$,

$$\mathcal{U}_n^\perp \cap \mathcal{S}_{2,m} \subset \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1,m}, \quad (20)$$

as follows. Let $z \in \mathcal{U}_n^\perp \cap \mathcal{S}_{2,m}$ be arbitrary. Since $z \in \mathcal{S}_{2,m}$ there exists constants c_k such that

$$\begin{aligned} z &= \sum_{k=2}^m c_k x_k = \sum_{k=2}^n c_k (\hat{x}_k + u_k) + \sum_{k=n+1}^m c_k x_k \\ &= \sum_{k=2}^n c_k \|u_k\| e_k + \sum_{k=2}^n c_k \hat{x}_k + \sum_{k=n+1}^m c_k x_k. \end{aligned}$$

Moreover, since $z \in \mathcal{U}_n^\perp$ it follows that $\langle z, e_k \rangle = 0$ for $k = 1, \dots, n$. In particular,

$$0 = \langle z, e_2 \rangle = c_2 \|u_2\| + \sum_{k=2}^n c_k \langle \hat{x}_k, e_2 \rangle + \sum_{k=n+1}^m c_k \langle x_k, e_2 \rangle = c_2 \|u_2\|$$

because $\hat{x}_k \in S_{k+1} \subset S_3$ for $k = 2, \dots, n$, $x_k \in S_k \subset S_{n+1}$ for $k = n+1, \dots, m$, and e_2 is orthogonal to S_3 and S_{n+1} . Hence $c_2 = 0$ and thus

$$z = \sum_{k=3}^n c_k \|u_k\| e_k + \sum_{k=3}^n c_k \hat{x}_k + \sum_{k=n+1}^m c_k x_k.$$

It follows now similarly that $c_k = 0$ for $k = 3, \dots, n$, hence

$$z = \sum_{k=n+1}^m c_k x_k \in \mathcal{S}_{n+1,m}.$$

Because $z \in \mathcal{U}_n^\perp$ as well, it follows now that

$$z \in \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1,m},$$

which implies (20).

However, $\mathcal{S}_{n+1,m} \subset \mathcal{S}_{2,m}$ and therefore

$$\mathcal{U}_n^\perp \cap \mathcal{S}_{n+1,m} \subset \mathcal{U}_n^\perp \cap \mathcal{S}_{2,m}. \quad (21)$$

Combining (20) and (21) now yields

$$\mathcal{U}_n^\perp \cap \mathcal{S}_{2,m} = \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1,m} \text{ for } m > n,$$

which in its turn implies that

$$\mathcal{U}_n^\perp \cap (\overline{\cup_{m=n+1}^{\infty} \mathcal{S}_{2,m}}) = \mathcal{U}_n^\perp \cap (\overline{\cup_{m=n+1}^{\infty} \mathcal{S}_{n+1,m}}). \quad (22)$$

Finally, note that $\mathcal{S}_2 = \overline{\cup_{m=n+1}^{\infty} \mathcal{S}_{2,m}}$ and $\mathcal{S}_{n+1} = \overline{\cup_{m=n+1}^{\infty} \mathcal{S}_{n+1,m}}$, hence it follows from (22) that (18) holds.

Step 3. Denote $\mathcal{R}_n = \text{span}(\mathcal{U}_n, \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1})$. Then

$$\mathcal{M}_\infty = \overline{\cup_{n=1}^{\infty} \mathcal{R}_n}. \quad (23)$$

Proof. It follows from (19) that $\mathcal{M}_n \subset \mathcal{R}_n$, hence

$$\mathcal{M}_\infty = \overline{\cup_{n=1}^{\infty} \mathcal{M}_n} \subset \overline{\cup_{n=1}^{\infty} \mathcal{R}_n}. \quad (24)$$

However, we also have $\mathcal{R}_n \subset \mathcal{M}_\infty$, as is not hard to verify, hence

$$\overline{\cup_{n=1}^{\infty} \mathcal{R}_n} \subset \mathcal{M}_\infty. \quad (25)$$

Thus, the result (23) follows from (24) and (25).

Step 4. For an $x \in \mathcal{M}_\infty$, let \hat{x}_n be the projection of x on \mathcal{R}_n . Then

$$\hat{x}_n = \sum_{j=1}^n \alpha_j e_j + w_{n+1} \quad (26)$$

where $\alpha_j = \langle x, e_j \rangle$ and w_{n+1} is the projection of x on $\mathcal{U}_n^\perp \cap \mathcal{S}_{n+1}$. Moreover,

$$\sum_{j=1}^{\infty} \alpha_j^2 < \infty. \quad (27)$$

Furthermore,

$$\lim_{n \rightarrow \infty} \left\| x - \sum_{j=1}^n \alpha_j e_j - w_{n+1} \right\| = 0. \quad (28)$$

Proof. By the definition of \mathcal{R}_n and Lemma 4, $\hat{x}_n = \sum_{j=1}^n \theta_j e_j + w$ for some constants θ_j and a $w \in \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1}$. To determine the θ_j 's and w , note that

$$\begin{aligned} \left\| x - \sum_{j=1}^n \theta_j e_j - w \right\|^2 &= \|x - w\|^2 - 2 \sum_{j=1}^n \theta_j \langle e_j, x \rangle + 2 \sum_{j=1}^n \theta_j \langle e_j, w \rangle \\ &\quad + \left\| \sum_{j=1}^n \theta_j e_j \right\|^2 \\ &= \|x - w\|^2 - 2 \sum_{j=1}^n \theta_j \langle e_j, x \rangle + \sum_{j=1}^n \theta_j^2 \end{aligned}$$

because $w \in \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1} \subset \mathcal{U}_n^\perp$ implies $\langle e_j, w \rangle = 0$ and

$$\left\| \sum_{j=1}^n \theta_j e_j \right\|^2 = \sum_{j=1}^n \sum_{i=1}^n \theta_j \theta_i \langle e_j, e_i \rangle = \sum_{j=1}^n \theta_j^2 \langle e_j, e_j \rangle = \sum_{j=1}^n \theta_j^2.$$

Thus

$$\begin{aligned} \|x - \hat{x}_n\|^2 &= \inf_{\theta_1, \dots, \theta_n, w \in \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1}} \left\| x - \sum_{j=1}^n \theta_j e_j - w \right\|^2 \\ &= \inf_{\theta_1, \dots, \theta_n, w \in \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1}} \left(\|x - w\|^2 - 2 \sum_{j=1}^n \theta_j \langle e_j, x \rangle + \sum_{j=1}^n \theta_j^2 \right) \\ &= \inf_{w \in \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1}} \|x - w\|^2 - \sum_{j=1}^n \alpha_j^2 \\ &= \|x - w_{n+1}\|^2 - \sum_{j=1}^n \alpha_j^2 \end{aligned} \quad (29)$$

where $\alpha_j = \langle x, e_j \rangle$ and w_{n+1} is the projection of x on $\mathcal{U}_n^\perp \cap \mathcal{S}_{n+1}$.

This result implies that for all n ,

$$\sum_{j=1}^n \alpha_j^2 \leq \|x - w_{n+1}\|^2 \leq \|x\|^2 \quad (30)$$

so that (27) holds.

Finally, to prove (28), let \hat{x} be the projection of x on $\overline{\cup_{n=1}^{\infty} \mathcal{R}_n}$. Then it follows from Theorem 2 that $\lim_{n \rightarrow \infty} \|\hat{x}_n - \hat{x}\| = 0$. But (23) implies $\hat{x} \in \mathcal{M}_{\infty}$, hence $x = \hat{x}$, so that $\lim_{n \rightarrow \infty} \|\hat{x}_n - x\| = 0$.

Step 5. Let $z_n = \sum_{j=1}^n \alpha_j e_j$. Then

$$\lim_{n \rightarrow \infty} \|z - z_n\| = 0, \text{ where } z \in \mathcal{U}_{\infty}. \quad (31)$$

Proof. This follows from the fact that z_n is a Cauchy sequence in \mathcal{U}_{∞} because

$$\begin{aligned} \|z_n - z_m\|^2 &= \left\| \sum_{j=\min(m,n)+1}^{\max(m,n)} \alpha_j e_j \right\|^2 = \sum_{j=\min(m,n)+1}^{\max(m,n)} \alpha_j^2 \\ &\leq \sum_{j=\min(m,n)+1}^{\infty} \alpha_j^2 \rightarrow 0 \end{aligned}$$

as $\min(m, n) \rightarrow \infty$, where the latter is due to $\sum_{j=1}^{\infty} \alpha_j^2 < \infty$.

Step 6. There exists a $w \in \mathcal{U}_{\infty}^{\perp} \cap S_{\infty}$ such that

$$\lim_{n \rightarrow \infty} \|w_{n+1} - w\| = 0. \quad (32)$$

Proof. Recall from Step 4 that $w_{n+1} \in \mathcal{U}_n^{\perp} \cap \mathcal{S}_{n+1}$. Moreover, it is easy to verify that $\mathcal{U}_{n+1}^{\perp} \subset \mathcal{U}_n^{\perp}$ whereas it is trivial that $\mathcal{S}_{n+2} \subset \mathcal{S}_{n+1}$, hence $\mathcal{U}_{n+1}^{\perp} \cap \mathcal{S}_{n+2} \subset \mathcal{U}_n^{\perp} \cap \mathcal{S}_{n+1}$. Thus for an arbitrary $k \geq 1$ and all $n \geq k$,

$$w_{n+1} \in \mathcal{U}_k^{\perp} \cap \mathcal{S}_{k+1}$$

Furthermore for $n \geq k$, w_{n+1} is a Cauchy sequence in $\mathcal{U}_k^{\perp} \cap \mathcal{S}_{k+1}$ because

$$\begin{aligned} \|w_{n+1} - w_{m+1}\| &= \|\hat{x}_n - z_n - \hat{x}_m + z_m\| \\ &\leq \|\hat{x}_n - \hat{x}_m\| + \|z_n - z_m\| \\ &\leq \|\hat{x}_n - x\| + \|\hat{x}_m - x\| + \|z_n - z_m\| \\ &\rightarrow 0 \end{aligned}$$

as $\min(m, n) \rightarrow \infty$. Therefore, there exists a $w \in \mathcal{U}_k^\perp \cap \mathcal{S}_{k+1}$ such that (32) holds. Since k was arbitrary we now have $w \in \cap_{k=1}^\infty \mathcal{U}_k^\perp = \mathcal{U}_\infty^\perp$ and $w \in \cap_{k=1}^\infty \mathcal{S}_{k+1} = \mathcal{S}_\infty$, hence

$$w \in \mathcal{U}_\infty^\perp \cap \mathcal{S}_\infty.$$

This completes the proof of Step 6.

The theorem now follows from (27), (31), (32) and the fact that $w \in \mathcal{U}_\infty^\perp \cap \mathcal{S}_\infty \subset \mathcal{U}_\infty^\perp$, which implies that $\langle w, e_k \rangle = 0$ for $k = 1, 2, 3, \dots$. Q.E.D.

4 The Wold decomposition for covariance stationary time series

In the case of the Hilbert space \mathcal{R}_0 of zero-mean random variables with finite second moments, with inner product $\langle X, Y \rangle = E[X.Y]$ and associated norm and metric, the results of Theorem 4 translate as follows:

Theorem 5. *Let X_t be a regular univariate zero-mean covariance stationary time series process. Then X_t can be written as*

$$X_t = \sum_{j=0}^{\infty} \alpha_j U_{t-j} + W_t \text{ a.s.,} \quad (33)$$

where the U_t is a zero-mean uncorrelated process with variance 1,

$$\alpha_j = E[X_t U_{t-j}], \quad \sum_{j=0}^{\infty} \alpha_j^2 < \infty, \quad (34)$$

and W_t is a zero-mean covariance stationary process satisfying

$$W_t \in \mathcal{U}_t^\perp \cap \mathcal{S}_{-\infty}, \quad (35)$$

where $\mathcal{S}_{-\infty} = \cap_n \text{span}(\{X_{n-k}\}_{k=1}^\infty)$ and \mathcal{U}_t^\perp is the orthogonal complement of $\mathcal{U}_t = \text{span}(\{U_{t-k}\}_{k=0}^\infty)$. The result (35) implies that

$$W_t \in \text{span}(\{W_{t-m}\}_{m=1}^\infty), \quad (36)$$

which in its turn implies that W_t is perfectly predictable from its past values $W_{t-1}, W_{t-2}, W_{t-3}, \dots$. In other words, W_t is a deterministic process. Moreover, (35) implies that

$$E[W_t U_{t-m}] = 0 \quad (37)$$

for all leads and lags m .

Proof: Recall that $U_t = \tilde{U}_t / \sqrt{E[\tilde{U}_t^2]}$, where $\tilde{U}_t = X_t - \hat{X}_t$ with \hat{X}_t the projection of X_t on $\text{span}(\{X_{t-j}\}_{j=1}^\infty)$. The uncorrelatedness of the \tilde{U}_t 's follows from Theorem 4, but we still need to show that $E[\tilde{U}_t] = 0$ and $E[\tilde{U}_t^2] = \sigma^2$ for all t .

Proof of $E[\tilde{U}_t] = 0$

Let $\hat{X}_{t,n}$ be the projection of X_t on $\text{span}(\{X_{t-j}\}_{j=1}^n)$. Then $\hat{X}_{t,n}$ takes the form

$$\hat{X}_{t,n} = \sum_{j=1}^n \beta_{j,n} X_{t-j},$$

where the $\beta_{j,n}$'s do not depend on t . The latter follows from the fact that the $\beta_{j,n}$'s are the solutions of the normal equations

$$\sum_{j=1}^n \beta_{j,n} \gamma(i-j) = \gamma(i), \quad i = 1, 2, \dots, n,$$

where $\gamma(i) = E[X_t X_{t-i}]$ is the covariance function of X_t . Hence $E[\hat{X}_{t,n}] = 0$.

It follows from Theorem 2 that

$$\lim_{n \rightarrow \infty} \left\| \hat{X}_{t,n} - \hat{X}_t \right\|^2 = \lim_{n \rightarrow \infty} E \left[\left(\hat{X}_{t,n} - \hat{X}_t \right)^2 \right] = 0 \quad (38)$$

so that by Liapounov's inequality and $E[\hat{X}_{t,n}] = 0$,

$$\begin{aligned} \lim_{n \rightarrow \infty} |E[\hat{X}_t]| &= \lim_{n \rightarrow \infty} |E[\hat{X}_t - \hat{X}_{t,n}]| \leq \lim_{n \rightarrow \infty} E \left[|\hat{X}_t - \hat{X}_{t,n}| \right] \\ &\leq \sqrt{\lim_{n \rightarrow \infty} E \left[\left(\hat{X}_{t,n} - \hat{X}_t \right)^2 \right]} = 0. \end{aligned}$$

Thus $E[\hat{X}_t] = 0$ and therefore $E[\tilde{U}_t] = E[X_t - \hat{X}_t] = 0$.

Proof of $E[\tilde{U}_t^2] = \sigma^2$

Let $\tilde{U}_{t,n} = X_t - \hat{X}_{t,n}$. It follows from (38) that

$$\lim_{n \rightarrow \infty} E \left[(\tilde{U}_t - \tilde{U}_{t,n})^2 \right] = \lim_{n \rightarrow \infty} E \left[(\hat{X}_{t,n} - \hat{X}_t)^2 \right] = 0 \quad (39)$$

Moreover,

$$\begin{aligned} E \left[\tilde{U}_{t,n}^2 \right] &= \|X_t - \hat{X}_{t,n}\|^2 = E \left[\left(X_t - \sum_{j=1}^n \beta_{j,n} X_{t-j} \right)^2 \right] \\ &= \gamma(0) - 2 \sum_{j=1}^n \beta_{j,n} \gamma(j) + \sum_{j=1}^n \sum_{i=1}^n \beta_{j,n} \beta_{i,n} \gamma(i-j) \\ &= \sigma_n^2 \end{aligned}$$

say, which does not depend on t . Furthermore, note that σ_n^2 is non-increasing in n , so that

$$\lim_{n \rightarrow \infty} \sigma_n^2 = \sigma^2$$

exists, and that

$$\begin{aligned} E \left[(\tilde{U}_t - \tilde{U}_{t,n})^2 \right] &= \|\hat{X}_{t,n} - \hat{X}_t\|^2 = \|\hat{X}_{t,n} - X_t + \tilde{U}_t\|^2 \\ &= \|\hat{X}_{t,n} - X_t\|^2 + 2 \langle \hat{X}_{t,n} - X_t, \tilde{U}_t \rangle + \|\tilde{U}_t\|^2 \\ &= \|\tilde{U}_{t,n}\|^2 - 2 \langle X_t, \tilde{U}_t \rangle + \|\tilde{U}_t\|^2 \\ &= \|\tilde{U}_{t,n}\|^2 - 2 \langle \hat{X}_t + \tilde{U}_t, \tilde{U}_t \rangle + \|\tilde{U}_t\|^2 \\ &= \|\tilde{U}_{t,n}\|^2 - 2 \langle \tilde{U}_t, \tilde{U}_t \rangle + \|\tilde{U}_t\|^2 \\ &= \|\tilde{U}_{t,n}\|^2 - \|\tilde{U}_t\|^2 \\ &= E \left[\tilde{U}_{t,n}^2 \right] - E \left[\tilde{U}_t^2 \right]. \end{aligned}$$

Thus,

$$E \left[\tilde{U}_t^2 \right] = \sigma_n^2 - E \left[(\tilde{U}_t - \tilde{U}_{t,n})^2 \right] \rightarrow \sigma^2.$$

Proof of (34), (35) and (37)

The result of Theorem 4 can now be translated as

$$\lim_{n \rightarrow \infty} \left\| X_t - \sum_{j=0}^n \alpha_j U_{t-j} - W_t \right\| = 0, \quad (40)$$

where U_t is a zero-mean uncorrelated covariance stationary process with unit variance, and $\alpha_k = \langle X_t, U_{t-k} \rangle = E[X_t U_{t-k}]$ with $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.

We still need to prove that the α_k 's do not depend on t , as follows. Recall from the proof of $E[\tilde{U}_t^2] = \sigma^2$ that $\tilde{U}_{t,n} = X_t - \sum_{j=1}^n \beta_{j,n} X_{t-j}$, so that

$$E \left[X_{t+k} \tilde{U}_{t,n} \right] = \gamma(k) - \sum_{j=1}^n \beta_{j,n} \gamma(k+j),$$

which does not depend on t . Moreover, by the Cauchy-Schwarz inequality and (39),

$$\lim_{n \rightarrow \infty} \left| E \left[X_{t+k} \left(\tilde{U}_{t,n} - \tilde{U}_t \right) \right] \right|^2 \leq \gamma(0) \lim_{n \rightarrow \infty} E \left[\left(\tilde{U}_{t,n} - \tilde{U}_t \right)^2 \right] = 0.$$

Thus $E \left[X_{t+k} \tilde{U}_t \right] = \lim_{n \rightarrow \infty} E \left[X_{t+k} \tilde{U}_{t,n} \right]$. Since the latter does not depend on t , neither does $\alpha_k = E[X_{t+k} U_t] = E[X_{t+k} \tilde{U}_t / ||\tilde{U}_t||]$.

The results (35) and (37) follow straightforwardly from Theorem 4.

Proof of (33)

The result (40) implies, by Chebyshev's inequality, that

$$X_t = p \lim_{n \rightarrow \infty} \sum_{j=0}^n \alpha_j U_{t-j} + W_t. \quad (41)$$

Recall that convergence in probability for $n \rightarrow \infty$ is equivalent to a.s. convergence along a further subsequence k_m of an arbitrary subsequence of n . See for example Bierens (2004, Theorem 6.B.3, p. 168). Thus for such a subsequence k_m ,

$$\sum_{j=0}^{k_m} \alpha_j U_{t-j} \rightarrow X_t - W_t \text{ a.s.} \quad (42)$$

as $m \rightarrow \infty$, and the same holds for any further subsequence of k_m .

Without loss of generality we may choose $k_0 = 0$. Then for each $n > 0$ we can find an m_n such that

$$k_{m_{n-1}} < n \leq k_{m_n}. \quad (43)$$

Moreover, (42) implies that

$$\sum_{j=0}^{k_{m_n}} \alpha_j U_{t-j} \rightarrow X_t - W_t \text{ a.s. as } n \rightarrow \infty. \quad (44)$$

Due to (43),

$$\begin{aligned} \sum_{n=1}^{\infty} E \left[\left(\sum_{j=0}^{k_{m_n}} \alpha_j U_{t-j} - \sum_{j=0}^n \alpha_j U_{t-j} \right)^2 \right] &= \sum_{n=1}^{\infty} E \left[\left(\sum_{j=n+1}^{k_{m_n}} \alpha_j U_{t-j} \right)^2 \right] \\ &\leq \sum_{n=1}^{\infty} \sum_{j=k_{m_{n-1}}+1}^{k_{m_n}} \alpha_j^2 \leq \sum_{j=0}^{\infty} \alpha_j^2 < \infty, \end{aligned}$$

so that by Chebyshev's inequality, for arbitrary $\varepsilon > 0$,

$$\sum_{n=0}^{\infty} P \left[\left| \sum_{j=0}^{k_{m_n}} \alpha_j U_{t-j} - \sum_{j=0}^n \alpha_j U_{t-j} \right| > \varepsilon \right] < \infty.$$

This result implies, by the Borel-Cantelli lemma,² that

$$\sum_{j=0}^{k_{m_n}} \alpha_j U_{t-j} - \sum_{j=0}^n \alpha_j U_{t-j} \rightarrow 0 \text{ a.s. as } n \rightarrow \infty. \quad (45)$$

Combining (44) and (45) it follows now that

$$\sum_{j=0}^n \alpha_j U_{t-j} \rightarrow X_t - W_t \text{ a.s. as } n \rightarrow \infty. \quad (46)$$

Since $\sum_{j=0}^{\infty} \alpha_j U_{t-j}$ is defined as $\lim_{n \rightarrow \infty} \sum_{j=0}^n \alpha_j U_{t-j}$, (33) is equivalent to (46).

²See for example Bierens (2004, Theorem 6.B.2, p. 168).

The zero-mean covariance stationarity of W_t

It follows now trivially from (33) that $E[W_t] = 0$. Moreover, W_t is covariance stationary because for $m \geq 0$,

$$\begin{aligned} E[W_t W_{t-m}] &= E\left[\left(X_t - \sum_{j=0}^{\infty} \alpha_j U_{t-j}\right)\left(X_{t-m} - \sum_{j=0}^{\infty} \alpha_j U_{t-m-j}\right)\right] \\ &= E[X_t X_{t-m}] - \sum_{j=m}^{\infty} \alpha_j E[U_{t-j} X_{t-m}] \\ &\quad - \sum_{j=0}^{\infty} \alpha_j E[U_{t-m-j} X_t] + \sum_{j=0}^{\infty} \alpha_{j+m} \alpha_j \\ &= \gamma(m) - \sum_{j=0}^{\infty} \alpha_{j+m} \alpha_j. \end{aligned}$$

Proof of (36)

Finally, $W_t \in \cap_n \text{span}(\{X_{n-j}\}_{j=0}^{\infty})$ implies that $W_t \in \text{span}(\{X_{t-j}\}_{j=1}^{\infty})$, hence the projection of W_t on $\text{span}(\{X_{t-j}\}_{j=1}^{\infty})$ is W_t itself. Since by (33),

$$\text{span}(\{X_{t-j}\}_{j=1}^{\infty}) = \text{span}(\text{span}(\{U_{t-j}\}_{j=1}^{\infty}), \text{span}(\{W_{t-j}\}_{j=1}^{\infty}))$$

and the projection of W_t on $\text{span}(\{U_{t-j}\}_{j=1}^{\infty})$ is zero, it follows that the projection of W_t on $\text{span}(\{W_{t-j}\}_{j=1}^{\infty})$ is W_t itself, which proves (36). Q.E.D.

Remark 2. The condition $\text{var}(U_t) = 1$ is not essential as long as X_t is regular. Without loss of generality we may then replace U_t with $\tilde{U}_t = \sigma U_t$, $\sigma > 0$, and α_k with $\tilde{\alpha}_k/\sigma$, where σ can be pinned down by normalizing $\tilde{\alpha}_0 = 1$.

5 Further analysis of the deterministic term

5.1 An example

The conclusion that the term W_t in (33) is deterministic does *not* imply that W_t is non-random. For example, consider the following sequence of random variables:

$$W_t = U \cos(t) + V \sin(t),$$

where U and V are independent standard normal random variables which do not depend on t . Suppose that for a given t , W_{t-1} and W_{t-2} are observed. Then

$$\begin{pmatrix} W_{t-1} \\ W_{t-2} \end{pmatrix} = \begin{pmatrix} \cos(t-1) & \sin(t-1) \\ \cos(t-2) & \sin(t-2) \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix}$$

hence

$$\begin{aligned} \begin{pmatrix} U \\ V \end{pmatrix} &= \frac{1}{\sin(t-2)\cos(t-1) - \cos(t-2)\sin(t-1)} \\ &\quad \times \begin{pmatrix} \sin(t-2) & -\sin(t-1) \\ -\cos(t-2) & \cos(t-1) \end{pmatrix} \begin{pmatrix} W_{t-1} \\ W_{t-2} \end{pmatrix} \\ &= \frac{-1}{\sin(1)} \begin{pmatrix} \sin(t-2) & -\sin(t-1) \\ -\cos(t-2) & \cos(t-1) \end{pmatrix} \begin{pmatrix} W_{t-1} \\ W_{t-2} \end{pmatrix}, \end{aligned}$$

so that

$$\begin{aligned} W_t &= \frac{-1}{\sin(1)} (\cos(t), \sin(t)) \begin{pmatrix} \sin(t-2) & -\sin(t-1) \\ -\cos(t-2) & \cos(t-1) \end{pmatrix} \begin{pmatrix} W_{t-1} \\ W_{t-2} \end{pmatrix} \\ &= \frac{-1}{\sin(1)} (\sin(t-2)\cos(t) - \cos(t-2)\sin(t)) W_{t-1} \\ &\quad - \frac{1}{\sin(1)} (\sin(t)\cos(t-1) - \cos(t)\sin(t-1)) W_{t-2} \\ &= \frac{\sin(2)}{\sin(1)} \cdot W_{t-1} - W_{t-2}. \end{aligned}$$

Moreover, it is trivial that $E[W_t] = 0$ and

$$E[W_t W_{t-m}] = \cos(t)\cos(t-m) + \sin(t).\sin(t-m) = \cos(m),$$

hence W_t is a zero-mean covariance stationary process.

5.2 Measurability

Again, the crux of Theorem 5 is that

$$W_t \in \mathcal{S}_{-\infty} = \cap_n \text{span}(\{X_{n-j}\}_{j=0}^{\infty}), \quad (47)$$

as the other conclusions of Theorem 5 follow straightforwardly from Theorem 3. The question I will now address is how (47) translates in terms of σ -algebras generated by sequences of the type $\{X_{n-j}\}_{j=0}^{\infty}$.

Denote for natural numbers n and m ,

$$\begin{aligned}\mathcal{F}_{t-n-m}^{t-n} &= \sigma(X_{t-n}, X_{t-n-1}, \dots, X_{t-n-m}), \\ \mathcal{S}_{t-n-m}^{t-n} &= \text{span}(X_{t-n}, X_{t-n-1}, \dots, X_{t-n-m})\end{aligned}$$

where $\mathcal{F}_{t-n-m}^{t-n}$ is the σ -algebra generated by $X_{t-n}, X_{t-n-1}, \dots, X_{t-n-m}$. Recall that

$$\mathcal{F}_{-\infty}^{t-n} = \sigma(\cup_{m=1}^{\infty} \mathcal{F}_{t-n-m}^{t-n})$$

is the smallest σ -algebra containing $\cup_{m=1}^{\infty} \mathcal{F}_{t-n-m}^{t-n}$, which is the σ -algebra generated by the sequence $\{X_{t-n-j}\}_{j=0}^{\infty}$, and that

$$\mathcal{F}_{-\infty} = \cap_{n=1}^{\infty} \mathcal{F}_{-\infty}^{t-n} = \cap_t \mathcal{F}_{-\infty}^t$$

is the remote σ -algebra of the sequence X_t , whereas

$$\begin{aligned}\mathcal{S}_{-\infty}^{t-n} &= \overline{\cup_{m=1}^{\infty} \mathcal{S}_{t-n-m}^{t-n}} = \text{span}(\{X_{t-n-j}\}_{j=0}^{\infty}), \\ \mathcal{S}_{-\infty} &= \cap_n \mathcal{S}_{-\infty}^{t-n}.\end{aligned}$$

The similarity between $\mathcal{S}_{-\infty}$ and $\mathcal{F}_{-\infty}$ suggests that, possibly, W_t is measurable with respect to (w.r.t.) the remote σ -algebra $\mathcal{F}_{-\infty}$. In Bierens (2004, p. 182) I have stated the latter as a fact, but in hindsight this is not trivial.

To prove this proposition, we need the following lemma:

Lemma 6. *Let $\{Z_n\}_{n=1}^{\infty}$ be a sequence of random variables [vectors] defined on a common probability space $\{\Omega, \mathcal{F}, P\}$ such that for a set $A \in \mathcal{F}$ with $P(A) = 1$, $\lim_{n \rightarrow \infty} Z_n(\omega) = Z(\omega)$ pointwise in $\omega \in A$. Then Z is a random variable [vector]³ defined on $\{\Omega, \mathcal{F}, P\}$, i.e., Z is measurable with respect to \mathcal{F} .*

Proof: See for example Theorem 2.13 in Bierens (2004, p. 47).

Let $\widehat{W}_{t-n,m}$ be the projection of W_t on $\mathcal{S}_{t-n-m}^{t-n}$, and let \widehat{W}_{t-n} be the projection of W_t on $\mathcal{S}_{-\infty}^{t-n}$. Then it follows from Theorem 2 that

$$\lim_{m \rightarrow \infty} \|\widehat{W}_{t-n,m} - \widehat{W}_{t-n}\|^2 = \lim_{m \rightarrow \infty} E \left[\left(\widehat{W}_{t-n,m} - \widehat{W}_{t-n} \right)^2 \right] = 0,$$

³Possibly extended arbitrarily for $\omega \in \Omega \setminus A$.

hence by Chebyshev inequality, $\text{plim}_{m \rightarrow \infty} \widehat{W}_{t-n,m} = \widehat{W}_{t-n}$. Recall⁴ that the latter implies that along a subsequence m_k ,

$$\widehat{W}_{t-n,m_k} \xrightarrow{\text{a.s.}} \widehat{W}_{t-n} \text{ as } k \rightarrow \infty. \quad (48)$$

Moreover, recall that $\widehat{W}_{t-n,m_k} \in \mathcal{S}_{t-n-m_k}^{t-n}$ takes the form of a linear combination of $X_{t-n}, X_{t-n-1}, \dots, X_{t-n-m_k}$, hence \widehat{W}_{t-n,m_k} is measurable w.r.t. $\mathcal{F}_{t-n-m_k}^{t-n}$ and therefore is also measurable w.r.t. $\mathcal{F}_{-\infty}^{t-n}$. It follows now from (48) and Lemma 6 that \widehat{W}_{t-n} is measurable w.r.t. $\mathcal{F}_{-\infty}^{t-n}$.

Because $W_t \in \mathcal{S}_{-\infty} \subset \mathcal{S}_{-\infty}^{t-n}$ and therefore $\widehat{W}_{t-n} = W_t$ a.s., it follows now that for arbitrary n , W_t is measurable w.r.t. $\mathcal{F}_{-\infty}^{t-n}$ and therefore W_t is also measurable w.r.t. $\cap_{n=1}^{\infty} \mathcal{F}_{-\infty}^{t-n}$. Thus,

Theorem 6. *The deterministic term W_t in the Wold decomposition (33) is measurable with respect to the remote σ -algebra $\mathcal{F}_{-\infty}$ of the time series X_t .*

5.3 When is the deterministic term equal to zero?

Finally, I will address the question: under what condition(s) is $W_t = 0$ a.s.. For this we need the concept of *vanishing memory* introduced in Bierens (2004, Definition 7.3, p.183):

Definition 5. *A time series process X_t is said to have a vanishing memory if all the sets in its remote σ -algebra $\mathcal{F}_{-\infty}$ have either probability zero or one.*

In this case any random variable W that is measurable w.r.t. to such a σ -algebra $\mathcal{F}_{-\infty}$ satisfies

$$E[W|\mathcal{F}_{-\infty}] = E[W] \text{ a.s.}$$

To see this, let $Z = E[W|\mathcal{F}_{-\infty}]$ and $A = \{\omega \in \Omega : Z(\omega) - E[W] > 0\}$. Recall from the definition of conditional expectation relative to a σ -algebra that Z is measurable w.r.t. $\mathcal{F}_{-\infty}$, so that $A \in \mathcal{F}_{-\infty}$, and that $\int_A Z dP = \int_A W dP$. Now suppose that $P(A) = 1$. Then

$$\begin{aligned} \int_A E[W] dP &= E[W].P(A) = E[W] = \int W dP \\ &= \int_A W dP + \int_{\Omega \setminus A} W dP = \int_A W dP \end{aligned}$$

⁴See for example Bierens (2004, Theorem 6.B.3, p. 168).

hence $\int_A (Z - E[W])dP = 0$. But this implies that $P(A) = 0$,⁵ so that $P(A) = 1$ is not possible, and therefore the only alternative is that $P(A) = 0$. The same applies to $A = \{\omega \in \Omega : Z(\omega) - E[W] < 0\}$. Hence, $Z = E[W]$ a.s..

Moreover, because W is measurable w.r.t. to $\mathcal{F}_{-\infty}$ we also have⁶

$$E[W|\mathcal{F}_{-\infty}] = W \text{ a.s.}$$

Because $E[W_t] = 0$ it follows now that

Theorem 7. *If the time series X_t in Theorem 5 has a vanishing memory then the deterministic term W_t in the Wold decomposition (33) is a.s. zero.*

6 The multivariate Wold decomposition

To prove the multivariate version of the Wold decomposition for a k -variate covariance stationary process X_t , consider the Hilbert space \mathcal{R}_k of zero mean random vectors in \mathbb{R}^k with finite second moment matrices, endowed with the inner product $\langle X, Y \rangle = E[X'Y]$ and associated norm and metric. Let \hat{X}_t be the projection of X_t on $\text{span}(\{X_{t-j}\}_{j=1}^\infty)$, with residual vector $V_t = X_t - \hat{X}_t$, and let $\Sigma = E[V_t V_t']$. In this case we need to extend the notion of regularity by requiring that Σ is positive definite rather than only $\|V_t\|^2 = E[V_t' V_t] > 0$, so that we can define $U_t = \Sigma^{-1/2}V_t$. Then the projection \tilde{X}_t of X_t on $\text{span}(\{U_{t-j}\}_{j=0}^n)$ takes the form $\tilde{X}_t = \sum_{j=1}^n A_j U_{t-j}$, where $A_j = E[X_t U_{t-j}']$. It follows now straightforwardly from the proofs of Theorems 4 and 5 that

$$X_t = \sum_{j=1}^{\infty} A_j U_{t-j} + W_t \text{ a.s.,}$$

where the process U_t is uncorrelated with zero expectation vector and variance matrix I_k , and $W_t \in \mathcal{U}_t^\perp \cap \mathcal{S}_{-\infty}$, with \mathcal{U}_t^\perp and $\mathcal{S}_{-\infty}$ defined in Theorem 5.

7 References

Anderson, T. W. (1994), *The Statistical Analysis of Time Series*, Wiley

⁵See for example Bierens (2004, Lemma 3.1, p.71).

⁶See for example Bierens (2004, Theorem 3.4, p. 73).

- Bierens, H. J. (2004). *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Cambridge University Press.
- Wold, H. (1938), *A Study in the Analysis of Stationary Time Series*. Almqvist and Wiksell, Sweden.

ARMA MODELS
Herman J. Bierens
Pennsylvania State University

February 23, 2009

1. *Introduction*

Given a covariance stationary process Y_t with vanishing memory¹ and expectation $\mu = E[Y_t]$, the Wold decomposition states that

$$Y_t - \mu = \sum_{j=0}^{\infty} \alpha_j U_{t-j}, \text{ with } \alpha_0 = 1, \sum_{j=0}^{\infty} \alpha_j^2 < \infty, \quad (1)$$

where U_t is an uncorrelated zero-mean covariance stationary process. If in addition the process Y_t is Gaussian then the U_t 's are i.i.d. $N(0, \sigma_u^2)$ [Why?]. If so then Y_t is strictly stationary [Why?]. In the sequel I will assume that (in the non-seasonal case) Y_t is covariance stationary and Gaussian, with a vanishing memory, so that the U_t 's are i.i.d. $N(0, \sigma_u^2)$.

To approximate the process Y_t by a process that only involves a finite number of parameters, denote $X_t = Y_t - \mu$. Next, project X_t on $X_{t-1}, X_{t-2}, \dots, X_{t-p}$ for some $p \geq 1$. This projection takes the form $\hat{X}_t = \sum_{j=1}^p \beta_j X_{t-j}$. Then we can write

$$X_t = \sum_{k=1}^p \beta_k X_{t-k} + V_t, \quad (2)$$

where

$$\begin{aligned} V_t &= X_t - \sum_{k=1}^p \beta_k X_{t-k} = U_t + \sum_{j=1}^{\infty} \alpha_j U_{t-j} - \sum_{k=1}^p \beta_k \left(U_{t-k} + \sum_{j=1}^{\infty} \alpha_j U_{t-k-j} \right) \\ &= U_t - \sum_{m=1}^{\infty} \theta_m U_{t-m}, \text{ say.} \end{aligned} \quad (3)$$

Since the θ_m 's are functions of the α_j 's for $j \geq m$ and $\sum_{j=0}^{\infty} \alpha_j^2 < \infty$, it follows that $\sum_{m=1}^{\infty} \theta_m^2 < \infty$. Consequently, for arbitrary small $\varepsilon > 0$ we can find a q such that

$$\text{var}\left(V_t - \left(U_t - \sum_{m=1}^q \theta_m U_{t-m}\right)\right) = \sum_{m=q+1}^{\infty} \theta_m^2 < \varepsilon.$$

This motivates to specify V_t in (2) as $U_t - \sum_{m=1}^q \theta_m U_{t-m}$, which gives rise to the ARMA(p, q) model

$$X_t = \sum_{k=1}^p \beta_k X_{t-k} + U_t - \sum_{m=1}^q \theta_m U_{t-m}. \quad (4)$$

Substituting $X_t = Y_t - \mu$ in (4) then yields

¹

See Bierens (2004), Chapter 7.

$$Y_t = \beta_0 + \sum_{k=1}^p \beta_k Y_{t-k} + U_t - \sum_{m=1}^q \theta_m U_{t-m}, \quad (5)$$

where $\beta_0 = (1 - \sum_{k=1}^p \beta_k) \mu$. This is the general ARMA(p, q) model.

All stationary time series models are of the form (5) or are special cases of (5). In particular, the ARMA($p, 0$) case is known as the autoregressive model of order p , shortly an $AR(p)$ model:

$$Y_t = \beta_0 + \sum_{j=1}^p \beta_j Y_{t-j} + U_t, \quad (6)$$

and the ARMA($0, q$) case is known as the moving average model of order q , shortly an $MA(q)$ model:

$$Y_t = \beta_0 + U_t - \sum_{m=1}^q \theta_m U_{t-m}. \quad (7)$$

2. The $AR(1)$ model

The simplest $AR(p)$ model is the one for the case $p = 1$:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + U_t. \quad (8)$$

I will first show that a necessary condition for the stationarity of the process (8) is that

$$|\beta_1| < 1. \quad (9)$$

By repeated backwards substitution of (8) we can write

$$Y_t = \beta_0 \sum_{k=0}^{m-1} \beta_1^k + \sum_{k=0}^{m-1} \beta_1^k U_{t-k} + \beta_1^m Y_{t-m}. \quad (10)$$

Since Y_t is covariance stationary, $E[Y_t] = \mu$ and $E[(Y_t - \mu)(Y_{t-m} - \mu)] = \gamma(m)$ for all t . Then it follows from (10) that

$$\mu = \beta_0 \sum_{k=0}^{m-1} \beta_1^k + \sum_{k=0}^{m-1} \beta_1^k E[U_{t-k}] + \beta_1^m \mu = \beta_0 \sum_{k=0}^{m-1} \beta_1^k + \beta_1^m \mu. \quad (11)$$

Next, subtract (11) from (10),

$$Y_t - \mu = \sum_{k=0}^{m-1} \beta_1^k U_{t-k} + \beta_1^m (Y_{t-m} - \mu), \quad (12)$$

take the square of both sides,

$$(Y_t - \mu)^2 = \left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k} \right)^2 + 2\beta_1^m \left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k} \right) (Y_{t-m} - \mu) + \beta_1^{2m} (Y_{t-m} - \mu)^2 \quad (13)$$

and take expectations,

$$\begin{aligned}
E[(Y_t - \mu)^2] &= E\left[\left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k}\right)^2\right] + 2\beta_1^m \sum_{k=0}^{m-1} \beta_1^k E[U_{t-k}(Y_{t-m} - \mu)] + \beta_1^{2m} E(Y_{t-m} - \mu)^2 \\
&= \sigma_u^2 \sum_{k=0}^{m-1} \beta_1^{2k} + \beta_1^{2m} E(Y_{t-m} - \mu)^2.
\end{aligned} \tag{14}$$

The last equality in (14) follows from the Wold decomposition (1). Because $\gamma(0) = E[(Y_t - \mu)^2] = E[(Y_{t-m} - \mu)^2]$, (14) reads:

$$\gamma(0) = \sigma_u^2 \sum_{k=0}^{m-1} \beta_1^{2k} + \beta_1^{2m} \gamma(0). \tag{15}$$

However, if $|\beta_1| \geq 1$ then the right-hand side of (15) converges to ∞ if $m \rightarrow \infty$, which contradicts the condition that $\gamma(0) < \infty$. On the other hand, if $|\beta_1| < 1$ then $\beta_1^{2m} \gamma(0) \rightarrow 0$ as $m \rightarrow \infty$, hence it follows from (10) by letting $m \rightarrow \infty$ that

$$Y_t = \beta_0 \sum_{k=0}^{\infty} \beta_1^k + \sum_{k=0}^{\infty} \beta_1^k U_{t-k} = \frac{\beta_0}{1-\beta_1} + \sum_{k=0}^{\infty} \beta_1^k U_{t-k}, \tag{16}$$

which is the Wold decomposition (1) :

$$\mu = \frac{\beta_0}{1-\beta_1}, \alpha_k = \beta_1^k, k = 0, 1, 2, 3, \dots \tag{17}$$

The expression at the right-hand side of (16) is also called the Moving Average (MA) representation of a covariance stationary time series. From this expression we can derive the covariance function of the AR(1) process, as follows:

$$\begin{aligned}
\gamma(m) &= E[(Y_t - \mu)(Y_{t-m} - \mu)] = E\left[\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-k}\right)\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)\right] \\
&= E\left[\left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k} + \sum_{k=m}^{\infty} \beta_1^k U_{t-k}\right)\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)\right] \\
&= E\left[\left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k} + \beta_1^m \sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)\right] \\
&= E\left[\left(\sum_{k=0}^{m-1} \beta_1^k U_{t-k}\right)\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)\right] + \beta_1^m E\left[\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)^2\right] \\
&= \beta_1^m E\left[\left(\sum_{k=0}^{\infty} \beta_1^k U_{t-m-k}\right)^2\right] = \beta_1^m \sigma_u^2 \sum_{k=0}^{\infty} \beta_1^{2k} = \sigma_u^2 \beta_1^m / (1 - \beta_1^2), m = 0, 1, 2, 3, \dots
\end{aligned} \tag{18}$$

3. Lag operators

A lag operator L is the instruction to shift the time back with one period: $L.Y_t = Y_{t-1}$. If we apply the lag operator again we get $L^2.Y_t = L(L.Y_t) = L.Y_{t-1} = Y_{t-2}$, and more generally

$$L^m.Y_t \stackrel{\text{def.}}{=} Y_{t-m}, \quad m = 0, 1, 2, 3, \dots \quad (19)$$

Using the lag operator, the AR(1) model (8) can be written as

$$(1 - \beta_1 L)Y_t = \beta_0 + U_t \quad (20)$$

In the previous section we have in several places used the equality

$$\sum_{k=0}^{\infty} z^k = \frac{1}{1-z}, \quad \text{provided that } |z| < 1, \quad (21)$$

which follows from the equalities $\sum_{k=0}^{\infty} z^k = 1 + \sum_{k=1}^{\infty} z^k = 1 + \sum_{k=0}^{\infty} z^{k+1} = 1 + z \cdot \sum_{k=0}^{\infty} z^k$. Now suppose that we may treat $\beta_1 L$ as the variable z in (21). If so, it follows from (21) that

$$\frac{1}{1 - \beta_1 L} = \sum_{k=0}^{\infty} (\beta_1 L)^k = \sum_{k=0}^{\infty} \beta_1^k L^k. \quad (22)$$

Applying this lag function to both sides of (20) then yields

$$\begin{aligned} Y_t &= \frac{1}{1 - \beta_1 L}(1 - \beta_1 L)Y_t = \sum_{k=0}^{\infty} \beta_1^k L^k \beta_0 + \sum_{k=0}^{\infty} \beta_1^k L^k U_t = \sum_{k=0}^{\infty} \beta_1^k \beta_0 + \sum_{k=0}^{\infty} \beta_1^k U_{t-k} \\ &= \frac{\beta_0}{1 - \beta_1} + \sum_{k=0}^{\infty} \beta_1^k U_{t-k}, \end{aligned} \quad (23)$$

which is exactly the moving average representation (16). Note that in the second equality in (23) I have used the fact that the lag operator has no effect on a constant: $L.\beta_0 = \beta_0$, hence $L^k \beta_0 = \beta_0$. Thus, the equality (22) holds if $|\beta_1| < 1$:

Proposition 1. *The lag function $\sum_{k=0}^{\infty} \beta^k L^k$ may be treated as $1/(1 - \beta L)$, in the sense that $(1 - \beta L) \sum_{k=0}^{\infty} \beta^k L^k = 1$, provided that $|\beta| < 1$.*

4. The AR(2) model

Consider the AR(2) process

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + U_t, \quad (24)$$

where the errors U_t have the same properties as before. Similar to (20) we can write this model in lag-polynomial form as

$$(1 - \beta_1 L - \beta_2 L^2)Y_t = \beta_0 + U_t. \quad (25)$$

We can always write

$$1 - \beta_1 L - \beta_2 L^2 = (1 - \alpha_1 L)(1 - \alpha_2 L), \quad (26)$$

by solving the equations $\alpha_1 + \alpha_2 = \beta_1$, $\alpha_1 \alpha_2 = -\beta_2$ ² so that (25) can be written as

$$(1 - \alpha_1 L)(1 - \alpha_2 L)Y_t = \beta_0 + U_t. \quad (27)$$

Now if $|\alpha_1| < 1$ and $|\alpha_2| < 1$ then it follows from Proposition 1 that

$$\begin{aligned} Y_t &= \frac{1}{(1-\alpha_1 L)(1-\alpha_2 L)}(\beta_0 + U_t) = \left(\sum_{k=0}^{\infty} \alpha_1^k L^{-k} \right) \left(\sum_{m=0}^{\infty} \alpha_2^m L^{-m} \right) \beta_0 + \left(\sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \alpha_1^k \alpha_2^m L^{-k-m} \right) U_t \\ &= \left(\sum_{k=0}^{\infty} \alpha_1^k \right) \left(\sum_{m=0}^{\infty} \alpha_2^m \right) \beta_0 + \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \alpha_1^k \alpha_2^m U_{t-k-m} \\ &= \frac{\beta_0}{(1-\alpha_1)(1-\alpha_2)} + \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \alpha_1^k \alpha_2^m U_{t-k-m} = \frac{\beta_0}{1-\beta_1-\beta_2} + \sum_{j=0}^{\infty} \left(\sum_{m=0}^j \alpha_1^{j-m} \alpha_2^m \right) U_{t-j}. \end{aligned} \quad (28)$$

Note that Y_t in (28) has expectation $\mu = \beta_0/(1-\beta_1-\beta_2)$ and variance $\sigma_u^2 \sum_{j=0}^{\infty} \left(\sum_{m=0}^j \alpha_1^{j-m} \alpha_2^m \right)^2$.

Consequently, the necessary conditions for the covariance stationarity of the AR(2) process (24) is that the errors U_t are covariance stationary and that the solutions $1/\alpha_1$ and $1/\alpha_2$ of the equation $0 = 1 - \beta_1 z - \beta_2 z^2 = (1 - \alpha_1 z)(1 - \alpha_2 z)$ are larger than one in absolute value. Similar conditions apply to general AR(p) processes:

² Although the solutions involved may be complex valued. If $\beta_2 \neq 0$ then the solutions are $\alpha_1 = 0.5\beta_1 \pm 0.5\sqrt{\beta_1^2 + 4\beta_2}$, $\alpha_2 = -\beta_2/\alpha_1$. If $\beta_1^2 + 4\beta_2 < 0$ then α_1 and α_2 are complex conjugate: $\alpha_1 = 0.5\beta_1 + i.0.5\sqrt{-\beta_1^2-4\beta_2}$, $\alpha_2 = 0.5\beta_1 - i.0.5\sqrt{-\beta_1^2-4\beta_2}$.

Proposition 2. *The necessary conditions for the covariance stationarity of the AR(p) process (6) are that the errors U_t are covariance stationary and the solutions z_1, \dots, z_p of the equation $0 = 1 - \beta_1 z - \beta_2 z^2 - \dots - \beta_p z^p$ are all greater than one in absolute value: $|z_j| > 1$ for $j = 1, \dots, p$.*

5. How to determine the order p of an AR(p) process

5.1 The partial autocorrelation function.

If the correct order of an AR process is p_0 but you estimate the AR(p) model (6) with $p > p_0$ by OLS, then the OLS estimates of the coefficients $\beta_{p_0+1}, \dots, \beta_p$ will be small and insignificant, because these coefficients are then all zero: $\beta_{p_0+1} = \beta_{p_0+2} = \dots = \beta_p = 0$. This suggests the following procedure for selecting p . Estimate the AR(p) model for $p = 1, 2, \dots, \bar{p}$, where $\bar{p} > p_0$.

$$\hat{Y}_t = \hat{\beta}_{p,0} + \hat{\beta}_{p,1} Y_{t-1} + \hat{\beta}_{p,2} Y_{t-2} + \dots + \hat{\beta}_{p,p} Y_{t-p}, \quad (29)$$

where the $\hat{\beta}_{p,j}$'s are OLS estimates. Then the (estimated) partial autocorrelation function,

PAC(p), is defined by

$$PAC(p) \stackrel{\text{def.}}{=} \hat{\beta}_{p,p}, \quad p = 1, 2, 3, \dots, \quad PAC(0) = 1. \quad (30)$$

For example, suppose that an AR(p) model $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + U_t$ has been fitted for $p = 1, 2, 3, 4, 5$ to 500 observation of a time series Y_t , with the following estimation results:

p	β_0	β_1	β_2	β_3	β_4	β_5
1	0.04070	0.02760				
	(0.06481)	(0.04470)				
2	0.06902	0.05358	-0.71257			
	(0.04609)	(0.03189)	(0.03221)			
3	0.06068	0.10470	-0.72159	0.07156		
	(0.04607)	(0.04481)	(0.03231)	(0.04525)		
4	0.06264	0.10524	-0.76661	0.07215	-0.06511	
	(0.04612)	(0.04500)	(0.04548)	(0.04577)	(0.04534)	
5	0.06283	0.10032	-0.76805	0.04648	-0.06783	-0.03274
	(0.04624)	(0.04516)	(0.04575)	(0.05715)	(0.04592)	(0.04544)

The entries that are not enclosed in brackets are the OLS estimates of the AR parameters, and the entries in brackets are the standard errors of the corresponding OLS estimates. Then

p	$PAC(p)$	(s.e.)
1	0.02760	(0.04470)
2	-0.71257	(0.03221)
3	0.07156	(0.04525)
4	-0.06511	(0.04534)
5	-0.03274	(0.04544)

In EasyReg (see Bierens 2008a) the PAC function can be computed automatically, via Menu > Data analysis > Auto/Cross correlation, and the results will then be displayed as a plot. For example, the $PAC(p)$ for the AR(2) model

$$Y_t = 1.144123Y_{t-1} - 0.5Y_{t-2} + U_t, \quad U_t \text{ i.i.d. } N(0,1), \quad t = 1, \dots, 500, \quad (31)$$

is displayed in Figure 1 below. The dots are the lower and upper bound of the one and two times the standard error bands, which correspond to the 68% and 95% confidence intervals of $\hat{\beta}_{p,p}$, respectively. The value $PAC(0) = 1$ is arbitrary, and is chosen because $PAC(p) < 1$ for $p \geq 1$.

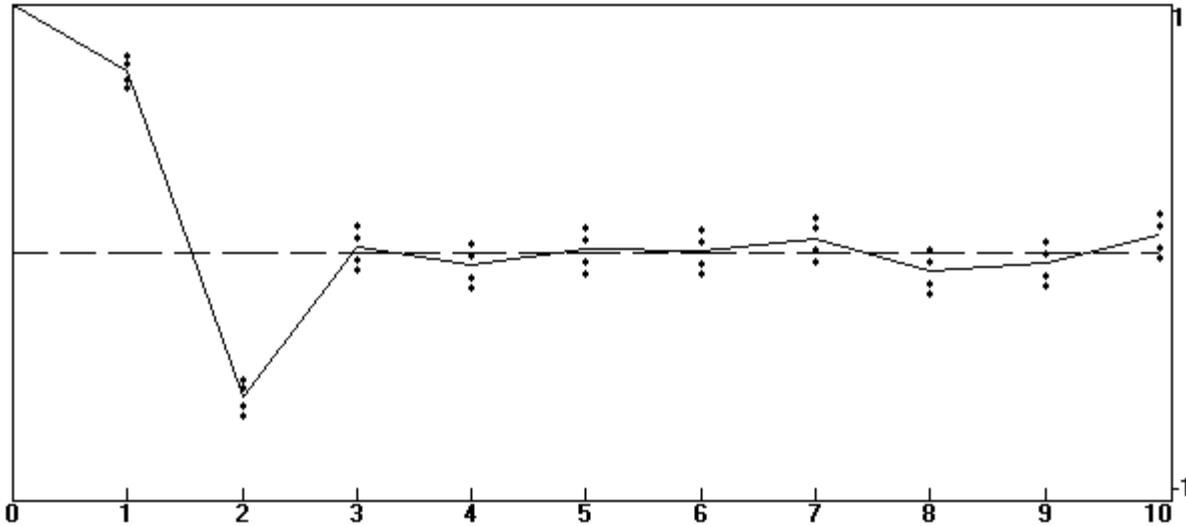


Figure 1: Partial autocorrelation function, $PAC(m)$, of the AR(2) process (31)

In Figure 1, at $p = 3$, the zero level is contained in the smaller 68% confidence interval, and at $p = 4$ the zero level is contained in the larger 95% confidence interval. From $p = 3$

onwards the zero level is contained in either the 68% and/or 95% confidence intervals, which indicates that the true value of p is $p_0 = 2$.

5.2 Information criteria

An alternative approach to determine the order p of the AR(p) model (6) is to use the Akaike (1974, 1976), Hannan-Quinn (1979), or Schwarz (1978) information criteria:

$$\begin{aligned} \text{Akaike: } c_n^{AR}(p) &= \ln(\hat{\sigma}_p^2) + 2(1+p)/n, \\ \text{Hannan-Quinn: } c_n^{AR}(p) &= \ln(\hat{\sigma}_p^2) + 2(1+p)\ln(\ln(n))/n, \\ \text{Schwarz:}^3 \quad c_n^{AR}(p) &= \ln(\hat{\sigma}_p^2) + (1+p)\ln(n)/n, \end{aligned}$$

where n is the effective sample size of the regression (6) (so that Y_t is observed for $t = 1-p, \dots, n$), and $\hat{\sigma}_p^2$ is the OLS estimator of the error variance $\sigma^2 = E[U_t^2]$. Denoting by \hat{p} the value of p for which $c_n^{AR}(p)$ is minimal:

$$c_n^{AR}(\hat{p}) = \min\{c_n^{AR}(1), \dots, c_n^{AR}(\bar{p})\},$$

where $\bar{p} > p_0$, with p_0 the true value of p , we have in the Hannan-Quinn and Schwarz cases: $\lim_{n \rightarrow \infty} P[\hat{p} = p_0] = 1$, and in the Akaike case $\lim_{n \rightarrow \infty} P[\hat{p} \geq p_0] = 1$ but $\lim_{n \rightarrow \infty} P[\hat{p} = p_0] < 1$. Thus, the Akaike criterion may “overshoot” the true value.

These results are based on the following facts. If $p < p_0$ then $\text{plim}_{n \rightarrow \infty} \hat{\sigma}_p^2 > \text{plim}_{n \rightarrow \infty} \hat{\sigma}_{p_0}^2$, hence in all three cases, $\lim_{n \rightarrow \infty} P[c_n^{AR}(p_0) < c_n^{AR}(p)] = 1$, whereas for $p > p_0$,

$$n \left(\ln(\hat{\sigma}_{p_0}^2) - \ln(\hat{\sigma}_p^2) \right) \xrightarrow{d} \chi_{p-p_0}^2, \quad (32)$$

where \xrightarrow{d} indicates convergence in distribution. The result (32) is due to the likelihood-ratio test. Then in the Akaike case,

$$n \left(c_n^{AR}(p_0) - c_n^{AR}(p) \right) = n \left(\ln(\hat{\sigma}_{p_0}^2) - \ln(\hat{\sigma}_p^2) \right) - 2(p-p_0) \xrightarrow{d} X_{p-p_0} - 2(p-p_0),$$

where $X_{p-p_0} \sim \chi_{p-p_0}^2$, hence $\lim_{n \rightarrow \infty} P[c_n^{AR}(p_0) > c_n^{AR}(p)] = P[X_{p-p_0} > 2(p-p_0)] > 0$.

³ The Schwarz information criterion is also known as the Bayesian Information Criterion (BIC).

Consequently, in the Akaike case we have $\lim_{n \rightarrow \infty} P[\hat{p} \geq p_0] = 1$, but $\lim_{n \rightarrow \infty} P[\hat{p} > p_0] > 0$. Therefore, the Akaike criterion may asymptotically overshoot the correct number of parameters.

Since (32) implies $\text{plim}_{n \rightarrow \infty} n(\ln(\hat{\sigma}_{p_0}^2) - \ln(\hat{\sigma}_p^2))/\ln(\ln(n)) = 0$ and $\text{plim}_{n \rightarrow \infty} n(\ln(\hat{\sigma}_{p_0}^2) - \ln(\hat{\sigma}_p^2))/\ln(n) = 0$ it follows that in the Hannan-Quinn case,

$$\text{plim}_{n \rightarrow \infty} n(c_n^{AR}(p_0) - c_n^{AR}(p))/\ln(\ln(n)) = 2(p - p_0) \geq 2$$

and in the Schwarz case,

$$\text{plim}_{n \rightarrow \infty} n(c_n^{AR}(p_0) - c_n^{AR}(p))/\ln(n) = p - p_0 \geq 1,$$

so that in both cases $\lim_{n \rightarrow \infty} P[c_n^{AR}(p_0) > c_n^{AR}(p)] = 0$. Hence, $\lim_{n \rightarrow \infty} P[\hat{p} = p_0] = 1$. Due to the latter, it is recommended to use the Hannan-Quinn or Schwarz criterion instead of the Akaike criterion. Note however that in small samples the Hannan-Quinn and Schwarz criteria may give different results for \hat{p} .

For example, for the same data on which Figure 1 was based, namely the AR(2) model $Y_t = 1.144123Y_{t-1} - 0.5Y_{t-2} + U_t$, $t = 1, \dots, 500$, with independent $N(0,1)$ distributed errors U_t , and upper bound $\bar{p} = 4$, we get

p	Akaike	Hannan-Quinn	Schwarz
1	5.14474E-01	5.21089E-01	5.31332E-01
2	1.08788E-01	1.18711E-01	1.34076E-01
3	1.12462E-01	1.25692E-01	1.46179E-01
4	1.13783E-01	1.30321E-01	1.55929E-01

All three criteria are minimal for $p = 2$, hence $\hat{p} = 2$, which is equal to the true value $p_0 = 2$.

5.3 The Wald test

A third way to determine the correct order p_0 of the AR(p) model (6) is the following. Determine an upper bound $\bar{p} > p_0$ on the basis of the PAC function and the information criteria, estimate the model (6) for $p = \bar{p}$ and test whether p can be reduced, using the Wald test, via Options > Wald test of linear parameter restrictions > Test joint significance, in the “What to do next?” module of EasyReg. For example, for the same data as in the previous section, and $\bar{p} = 4$, we get the OLS results

<i>Parameters</i>	<i>OLS estimate</i>	<i>t-value</i>
β_0	0.06910	1.449
β_1	1.17283	26.033
β_2	-0.63395	-9.134
β_3	0.07841	1.130
β_4	-0.05090	-1.130

The t-value of β_4 is well within the range -1.96, +1.96, hence the null hypothesis that $\beta_4 = 0$ cannot be rejected at the 5% significance level. To test whether $\beta_3 = 0$ as well, you need to test the joint null hypothesis $\beta_3 = \beta_4 = 0$, using the Wald test. In this case the test result involved is:

```

Wald test:                      1.45
Asymptotic null distribution: Chi-square(2)
p-value = 0.48398
Significance levels:          10%      5%
Critical values:              4.61     5.99
Conclusions:                  accept   accept

```

Thus, the null hypothesis $\beta_3 = \beta_4 = 0$ cannot be rejected, hence we may reduce p to 2.

Since β_2 is strongly significant, there is no need to test the null hypothesis $\beta_2 = \beta_3 = \beta_4 = 0$, but if we do so the null hypothesis will be rejected:

```

Wald test:                      253.39
Asymptotic null distribution: Chi-square(3)
p-value = 0.00000
Significance levels:          10%      5%
Critical values:              6.25     7.81
Conclusions:                  reject   reject

```

Thus, the test results involved lead to the same conclusion as the one on the basis of the PAC function and the information criteria, namely that $p_0 = 2$.

6. Moving average processes

Recall that a moving average process of order q , denoted by $\text{MA}(q)$, takes the form

$$Y_t = \mu + U_t - \theta_1 U_{t-1} - \dots - \theta_q U_{t-q}, \quad (33)$$

where $\mu = E[Y_t]$. Under regularity conditions an MA process has an infinite order AR representation, as I will demonstrate for the case $q = 1$.

Consider the $\text{MA}(1)$ process

$$Y_t = \mu + U_t - \theta U_{t-1}. \quad (34)$$

Using the lag operator, we can write this $\text{MA}(1)$ model as

$$Y_t = \mu + (1 - \theta L)U_t. \quad (35)$$

Now it follows from Proposition 1 and (35) that if $|\theta| < 1$ then

$$\sum_{j=0}^{\infty} \theta^j L^j Y_t = (1 - \theta L)^{-1} Y_t = (1 - \theta L)^{-1} \mu + U_t = \frac{\mu}{1-\theta} + U_t, \quad (36)$$

hence

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + U_t = \beta_0 + \sum_{j=1}^{\infty} \beta_j Y_{t-j} + U_t, \quad (37)$$

where $\beta_0 = \mu/(1-\theta)$, $\beta_j = -\theta^j$ for $j = 1, 2, 3, \dots$

More generally we have:

Proposition 3. If the solutions z_1, \dots, z_q of the equation $0 = 1 - \theta_1 z - \theta_2 z^2 - \dots - \theta_q z^q$ are all greater than one in absolute value: $|z_j| > 1$ for $j = 1, \dots, q$, then the $\text{MA}(q)$ process (33) can be written as an infinite order AR process: $Y_t = \beta_0 + \sum_{j=1}^{\infty} \beta_j Y_{t-j} + U_t$, where $\beta_0 = \mu/(1 - \theta_1 - \theta_2 - \dots - \theta_q)$ and $1 - \sum_{j=1}^{\infty} \beta_j L^j = 1/(1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_q L^q)$.

7. How to determine the order q of a $\text{MA}(q)$ process

7.1 The (regular) autocorrelation function

The autocorrelation function of a time series process Y_t is defined by

$$\rho(m) = \frac{\text{cov}(Y_t, Y_{t-m})}{\text{var}(Y_t)}, m = 0, 1, 2, \dots \dots \quad (38)$$

It is trivial that:

Proposition 4. For an MA(q) process, $\rho(m) = 0$ for $m > q$, and $\rho(q) \neq 0$.

The actual autocorrelation function cannot be calculated, but it can be estimated in various ways. EasyReg estimates $\rho(m)$ by

$$\hat{\rho}(m) = \frac{(1/(n-m))\sum_{t=m+1}^n(Y_t - \bar{Y})(Y_{t-m} - \bar{Y})}{\sqrt{(1/(n-m))\sum_{t=1}^{n-m}(Y_t - \bar{Y})^2}\sqrt{(1/(n-m))\sum_{t=m+1}^n(Y_{t-m} - \bar{Y})^2}}, m = 0, 1, 2, \dots \dots \quad (39)$$

where $\bar{Y} = (1/n)\sum_{t=1}^n Y_t$.

For an AR(p) process the autocorrelation function does not provide information about p . To see this, consider again the AR(1) process (8) satisfying condition (9). Then it follows from (18) that $\text{cov}(Y_t, Y_{t-m}) = \gamma(m) = \sigma^2\beta_1^m/(1-\beta_1^2)$ and $\text{var}(Y_t) = \gamma(0) = \sigma^2/(1-\beta_1^2)$, hence in the AR(1) case, $\rho(m) = \beta_1^m$. Therefore, in this case the autocorrelation function will not drop sharply to zero for $m > 1$, as is demonstrated in Figure 2.. The same applies to more general AR processes.

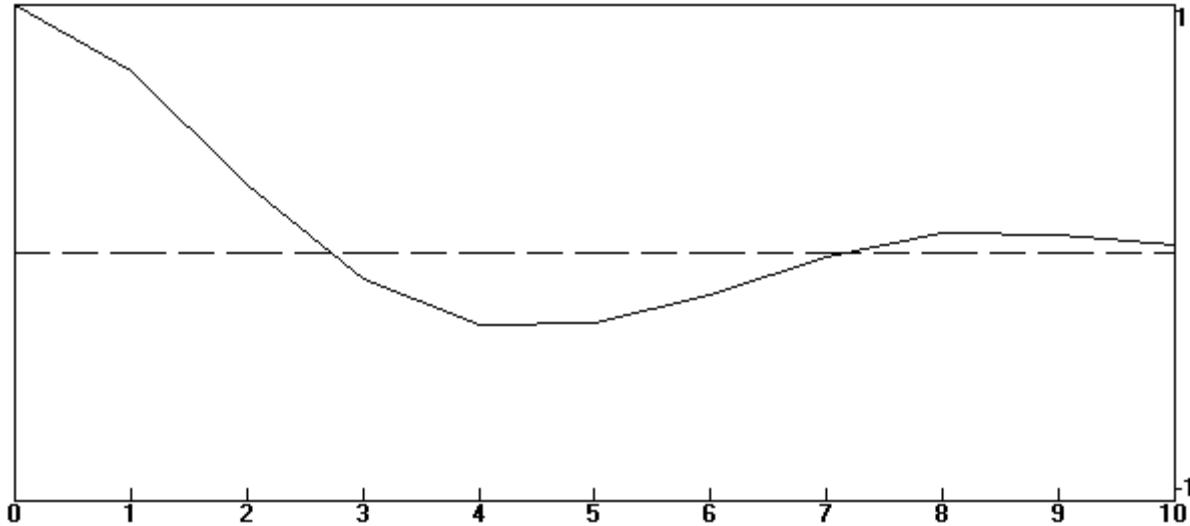


Figure 2: Estimated autocorrelation function $\hat{\rho}(m)$ of the AR(2) process (31).

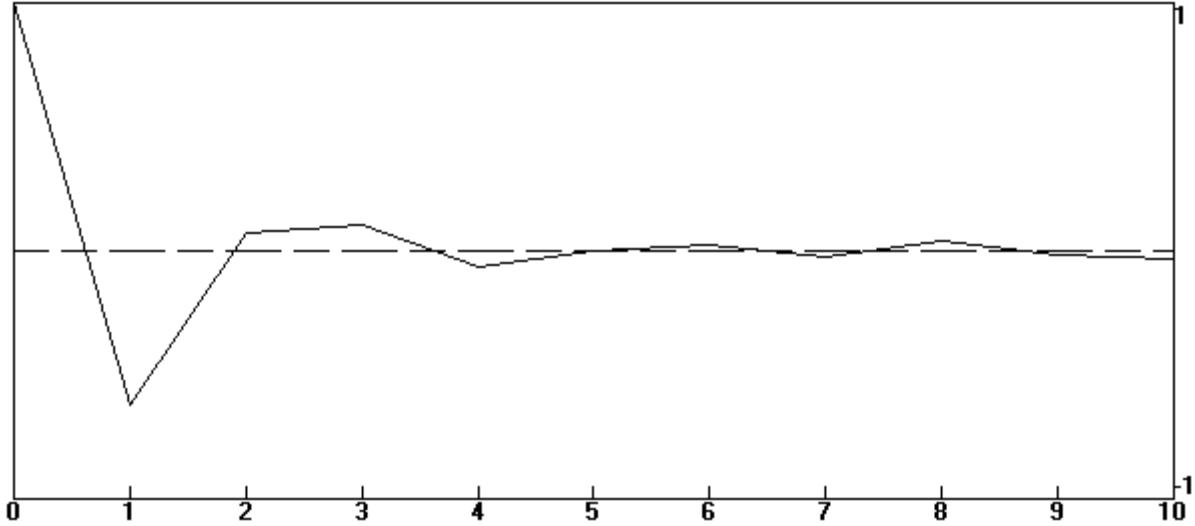


Figure 3: Estimated autocorrelation function $\hat{\rho}(m)$ of the MA(2) process (40)

To demonstrate how to use the estimated autocorrelation function to determine the order q of an MA(q) process, I have generated 500 observations according to the model

$$Y_t = U_t - 1.4U_{t-1} + 0.5U_{t-2}, \quad U_t \sim i.i.d. N(0,1), \quad t = 1, 2, \dots, 500 \quad (40)$$

The estimated autocorrelation function $\hat{\rho}(m)$ involved is displayed in Figure 3, for $m = 0, 1, \dots, 10$.

Because $\hat{\rho}(m)$ is not endowed with standard error bands, it is not obvious at which value of m the true autocorrelation function $\rho(m)$ becomes zero. But at least we can determine an upper bound \bar{q} of q from Figure 3: It seems that $\hat{\rho}(m)$ is approximately zero for $m \geq 5$, indicating that $q \leq 4$. Thus, let $\bar{q} = 4$.

The partial autocorrelation function of an MA(q) process is of no use for determining q or an upper bound of q , because of the AR(∞) representation of an MA(q) process. For example, the PAC(m) of the MA(2) process (40) does not drop sharply to zero for $m > 2$, as is demonstrated in Figure 4.

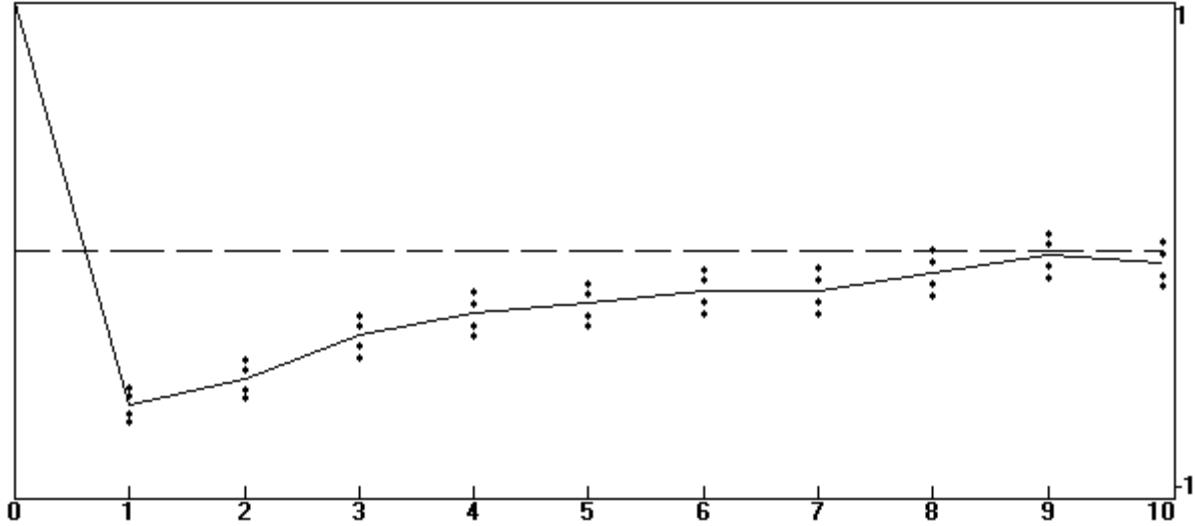


Figure 4: Partial autocorrelation function, $\text{PAC}(m)$, of the $\text{MA}(2)$ process (40)

7.2 Information criteria

The three information criteria, Akaike, Hannan-Quinn and Schwarz also apply to MA processes. Therefore, estimate the $\text{MA}(q)$ model (33) for $q = 1, 2, 3, 4$ ($= \bar{q}$), and compare the information criteria:

q	Akaike	Hannan-Quinn	Schwarz
1	$1.91941E-01$	$1.98556E-01$	$2.08799E-01$
2	$1.24771E-02$	$2.23999E-02$	$3.77647E-02$
3	$1.63628E-02$	$2.95933E-02$	$5.00797E-02$
4	$1.68145E-02$	$3.33526E-02$	$5.89606E-02$

All three criteria are minimal for $q = 2$, which is the true value.

7.3 Wald test

As a double check, estimate the MA model (33) for $q = 4$ (in EasyReg via Menu > Single equation models > ARIMA estimation and forecasting), and test whether $\theta_3 = \theta_4 = 0$, using the Wald test:

<i>Parameters</i>	<i>Estimate</i>	<i>t-value</i>
μ	0.000234	0.050
θ_1	1.348470	29.979
θ_2	-0.427742	-5.637
θ_3	-0.074225	-0.978
θ_4	0.050943	1.127

Wald test: 1.29
 Asymptotic null distribution: Chi-square(2)
 p-value = 0.52588
 Significance levels: 10% 5%
 Critical values: 4.61 5.99
 Conclusions: accept accept

Thus, the null hypothesis $\theta_3 = \theta_4 = 0$ cannot be rejected, hence we may reduce q from 4 to $q = 2$. Since θ_2 is strongly significant, we cannot reduce q further.

Re-estimating model (33) for $q = 2$ yields:

<i>Parameters</i>	<i>Estimate</i>	<i>t-value</i>
μ	0.000187	0.039
θ_1	1.348684	33.682
θ_2	-0.456211	-11.349

which is reasonably close to the true values of the parameters: $\mu = 0$, $\theta_1 = 1.4$, $\theta_2 = -0.5$.

8. ARMA models

8.1 Invertibility conditions

Denote

$$\begin{aligned}\varphi_p(L) &= 1 - \beta_1 L - \beta_2 L^2 - \dots - \beta_p L^p, \\ \psi_q(L) &= 1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_q L^q,\end{aligned}\tag{41}$$

Then the ARMA(p,q) model (5) can be written more compactly as

$$\varphi_p(L)Y_t = \beta_0 + \psi_q(L)U_t. \quad (42)$$

Proposition 5. A necessary condition for the stationarity of ARMA(p,q) process (42) is that

$\varphi_p(z_1) = 0$ implies $|z_1| > 1$. If so, it has the MA(∞) representation (or equivalently, the Wold decomposition) $Y_t = \beta_0/\varphi_p(1) + \varphi_p(L)^{-1}\psi_q(L)U_t$. If in addition $\psi_q(z_2) = 0$ implies $|z_2| > 1$ then the ARMA(p,q) process (42) has an AR(∞) representation: $\psi_q(L)^{-1}\varphi_p(L)Y_t = \beta_0/\psi_q(1) + U_t$.

I will demonstrate Proposition 5 for the case $p = q = 1$:

$$(1 - \beta_1 L)Y_t = \beta_0 + (1 - \theta_1 L)U_t. \quad (43)$$

The condition that $\varphi_1(z) = 0$ implies $|z| > 1$ is equivalent to $|\beta_1| < 1$, because $\varphi_1(z) = 1 - \beta_1 z = 0$ implies that $z = 1/\beta_1$. Similarly, the condition that $\psi_1(z) = 1 - \theta_1 z = 0$ implies $|z| > 1$ is equivalent to $|\theta_1| < 1$. It follows now from Proposition 1 that $\psi_1(L)^{-1} = (1 - \theta_1 L)^{-1} = \sum_{j=0}^{\infty} \theta_1^j L^j$, hence

$$\begin{aligned} \sum_{j=0}^{\infty} \theta_1^j L^j (1 - \beta_1 L)Y_t &= \psi_1(L)^{-1}(1 - \beta_1 L)Y_t = \sum_{j=0}^{\infty} \theta_1^j L^j \beta_0 + \sum_{j=0}^{\infty} \theta_1^j L^j (1 - \theta_1 L)U_t \\ &= \sum_{j=0}^{\infty} \theta_1^j \beta_0 + U_t = \beta_0/(1 - \theta_1) + U_t \end{aligned} \quad (44)$$

and

$$\begin{aligned} \sum_{j=0}^{\infty} \theta_1^j L^j (1 - \beta_1 L)Y_t &= \sum_{j=0}^{\infty} \theta_1^j L^j Y_t - \beta_1 \sum_{j=0}^{\infty} \theta_1^j L^{j+1} Y_t \\ &= Y_t + \sum_{j=1}^{\infty} \theta_1^j Y_{t-j} - \beta_1 \sum_{j=0}^{\infty} \theta_1^j Y_{t-j-1} = Y_t + \theta_1 \sum_{j=0}^{\infty} \theta_1^j Y_{t-1-j} - \beta_1 \sum_{j=0}^{\infty} \theta_1^j Y_{t-1-j} \\ &= Y_t - (\beta_1 - \theta_1) \sum_{j=0}^{\infty} \theta_1^j Y_{t-1-j}. \end{aligned} \quad (45)$$

Combining these results yields

$$Y_t = \beta_0/(1 - \theta_1) + (\beta_1 - \theta_1) \sum_{j=0}^{\infty} \theta_1^j Y_{t-1-j} + U_t, \quad (46)$$

which is the AR(∞) representation of the ARMA(1,1) process under review.

Similarly, it follows from Proposition 1 that $\varphi_1(L)^{-1} = (1 - \beta_1 L)^{-1} = \sum_{j=0}^{\infty} \beta_1^j L^j$, hence

$$\begin{aligned}
Y_t &= (1-\beta_1 L)^{-1}(1-\beta_1 L)Y_t = (1-\beta_1 L)^{-1}\beta_0 + (1-\beta_1 L)^{-1}(1-\theta_1 L)U_t \\
&= \beta_0/(1-\beta_1) + \sum_{j=0}^{\infty} \beta_1^j L^j (1-\theta_1 L)U_t \\
&= \beta_0/(1-\beta_1) + \sum_{j=0}^{\infty} \beta_1^j L^j U_t - \theta_1 \sum_{j=0}^{\infty} \beta_1^j L^{j+1} U_t \\
&= \beta_0/(1-\beta_1) + U_t - (\theta_1 - \beta_1) \sum_{j=0}^{\infty} \beta_1^j U_{t-1-j},
\end{aligned} \tag{47}$$

which is the MA(∞) representation of the ARMA(1,1) process under review.

8.2 Common roots

Observe from (46) and (47) that if $\beta_1 = \theta_1$ then $Y_t = \beta_0/(1-\beta_1) + U_t$, which is an ARMA(0,0) process (also called a *white noise* process). This is the **common roots** problem:

Proposition 6. *Let the conditions in Proposition 5 be satisfied. If there exists a $\delta \neq 0$ such that*

$\varphi_p(1/\delta) = \psi_q(1/\delta) = 0$ *then we can write the lag polynomials in ARMA(p,q) model (42) as*
 $\varphi_p(L) = (1-\delta L)\varphi_{p-1}^*(L)$ *and* $\psi_q(L) = (1-\delta L)\psi_{q-1}^*(L)$, *where* $\varphi_{p-1}^*(L)$ *and* $\psi_{q-1}^*(L)$ *are lag polynomials of order p-1 and q-1, respectively, satisfying the conditions in Proposition 5. The ARMA(p,q) process (42) is then equivalent to the ARMA(p-1,q-1) process*
 $\varphi_{p-1}^*(L)Y_t = \beta_0^* + \psi_{q-1}^*(L)U_t$, *where* $\beta_0^* = \varphi_{p-1}^*(1)E[Y_t]$.

Because the value of δ does not matter, the parameters in the lag polynomials $\varphi_p(L)$ and $\psi_q(L)$ are no longer identified. The same applies to the constant β_0 in model (42) because $\beta_0 = (1-\delta)\beta_0^*$ for arbitrary δ . For example, let for $p = q = 2$,

$$\begin{aligned}
\varphi_2(L) &= (1-\delta L)(1-\beta L) = 1-(\delta+\beta)L + \delta.\beta L^2 = 1-\beta_1 L-\beta_2 L^2 \\
\psi_2(L) &= (1-\delta L)(1-\theta L) = 1-(\delta+\theta)L + \delta.\theta L^2 = 1-\theta_1 L-\theta_2 L^2
\end{aligned} \tag{48}$$

where $|\beta| < 1$, $|\theta| < 1$, and $|\delta| < 1$, and let $E[Y_t] = 0$. Then the ARMA(2,2) model $\varphi_2(L)Y_t = \psi_2(L)U_t$ is equivalent to the ARMA(1,1) model $(1-\beta L)Y_t = (1-\theta L)U_t$ for all values of δ . Hence, given β and θ , $\beta_1 = \delta+\beta$, $\beta_2 = -\delta.\beta$, $\theta_1 = \delta+\theta$, $\theta_2 = -\delta.\theta$ for arbitrary δ . As a consequence, the estimates of the parameters β_1 , β_2 , θ_1 , θ_2 are no longer consistent, and the t-test and Wald test for testing the (joint) significance of the parameters are no longer valid. In particular, in the ARMA(2,2) case under review the Wald test of the null hypothesis $\beta_2 = \theta_2 = 0$

is no longer valid. Therefore, we should not use the Wald test to test whether the AR and MA orders p and q can be reduced to $p-1$ and $q-1$.

The problem of common roots in ARMA models is similar to the multicollinearity problem in linear regression. As in the latter case, the t values of the parameters will be deflated towards zero. Therefore, if all the t values of the ARMA parameters are insignificant this may indicate that the AR and MA lag polynomials have a common root.

Although we should not use the Wald test to test for common roots, we can still use the information criteria to determine whether the AR and MA orders p and q can be reduced to $p-1$ and $q-1$. In the case of a common root, the variance σ^2 of the errors U_t of the ARMA(p,q) model in Proposition 6 is the same as the variance of the errors U_t in the ARMA($p-1,q-1$) model

$\Phi_{p-1}^*(L)Y_t = \beta_0^* + \psi_{q-1}^*(L)U_t$. Therefore, the estimate $\hat{\sigma}_{p,q}^2$ of the errors U_t of the ARMA(p,q) model involved will be close to the estimate $\hat{\sigma}_{p-1,q-1}^2$ of the errors of the equivalent ARMA($p-1,q-1$) model, and asymptotically they will be equal:

$$\text{plim}_{n \rightarrow \infty} \hat{\sigma}_{p,q}^2 = \text{plim}_{n \rightarrow \infty} \hat{\sigma}_{p-1,q-1}^2 = \sigma^2. \quad (49)$$

In the ARMA case the three information criteria take the form

$$\begin{aligned} \text{Akaike: } c_n^{ARMA}(p,q) &= \ln(\hat{\sigma}_{p,q}^2) + 2(1+p+q)/n, \\ \text{Hannan-Quinn: } c_n^{ARMA}(p,q) &= \ln(\hat{\sigma}_{p,q}^2) + 2(1+p+q)\ln(\ln(n))/n, \\ \text{Schwarz: } c_n^{ARMA}(p,q) &= \ln(\hat{\sigma}_{p,q}^2) + (1+p+q)\ln(n)/n, \end{aligned}$$

Therefore, in the case of a common root, $c_n^{ARMA}(p-1,q-1) < c_n^{ARMA}(p,q)$ if n is large enough, due to (49).

To demonstrate the common roots phenomenon, I have generated a time series Y_t , $t = 1, \dots, 500$, according to the ARMA(1,1) model

$$Y_t = 0.3 + 0.7Y_{t-1} + U_t + 0.5U_{t-1}, \quad U_t \sim i.i.d N(0,1), \quad (50)$$

and estimated this model as an ARMA(2,2) model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + U_t - \theta_1 U_{t-1} - \theta_2 U_{t-2}. \quad (51)$$

The EasyReg estimation results involved are:

<i>Parameters</i>	<i>Estimate</i>	<i>t-value</i>
$\mu = \beta_0/(1-\beta_1-\beta_2)$	0.776272	3.273
β_1	1.087430	0.170
β_2	-0.267512	-0.058
θ_1	-0.166275	-0.026
θ_2	0.189439	0.055
σ	1.008897	

Information criteria:

Akaike:	2.76651E-02
Hannan-Quinn:	4.42031E-02
Schwarz:	6.98112E-02

Apart from the estimate of $\mu = E[Y_t]$, the AR and MA parameters are insignificant, due to a common root in the AR and MA lag polynomials.

Next, I have estimated the model as an ARMA(1,1) model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + U_t - \theta_1 U_{t-1}. \quad (52)$$

The EasyReg estimation results are:

<i>Parameters</i>	<i>Estimate</i>	<i>t-value</i>
$\mu = \beta_0/(1-\beta_1)$	0.775967	3.249
β_1	0.720705	20.784
θ_1	-0.530183	-12.597
σ	1.006879	

Information criteria:

Akaike:	1.96937E-02
Hannan-Quinn:	2.96166E-02
Schwarz:	4.49814E-02

Indeed, the information criteria for the latter model are substantial lower (and thus better) than for the previous ARMA(2,2) model. Moreover, observe that in the latter case the estimates of β_1 , θ_1 and σ are close to the true values $\beta_1 = 0.7$, $\theta_1 = -0.5$ and $\sigma = 1$, respectively,

although at first sight the estimate $\hat{\mu} = 0.775967$ of $\mu = E[Y_t]$ seems quite different from the true value $\mu = 0.3/(1-0.7) = 1$. However, it can be shown that $\hat{\mu}$ is not significantly different from 1.

8.3 How to distinguish an ARMA process from an AR process

The AR(∞) representation (46) of the ARMA(1,1) process (50) is

$$\begin{aligned}
 Y_t &= \beta_0/(1-\theta_1) + (\beta_1 - \theta_1)\sum_{j=0}^{\infty}\theta_1^jY_{t-1-j} + U_t \\
 &= 0.3/(1+0.5) + 1.2\sum_{j=0}^{\infty}(-0.5)^jY_{t-1-j} + U_t \\
 &= 0.2 + 1.2\sum_{j=0}^{\infty}(-0.5)^jY_{t-1-j} + U_t \\
 &= 0.2 + 1.2Y_{t-1} - 0.6Y_{t-2} + 0.3Y_{t-3} - 0.15Y_{t-4} + 0.075Y_{t-5} + \dots + U_t
 \end{aligned} \tag{53}$$

which is close to an AR(4) process. Therefore, the partial autocorrelation function, PAC(m), of this process will look like the PAC(m) of an AR process:

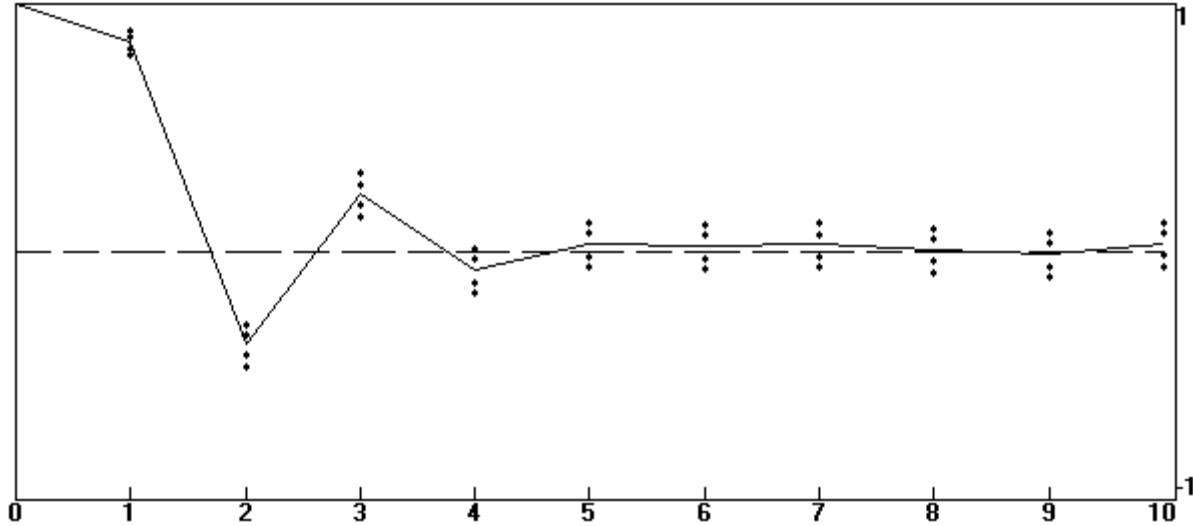


Figure 5: Partial autocorrelation function, PAC(m), of the ARMA(1,1) process (50)

Indeed, on the basis of this plot one may be tempted to conclude (erroneously) that the process is an AR(4) process, and the estimated autocorrelation function would actually corroborate this:

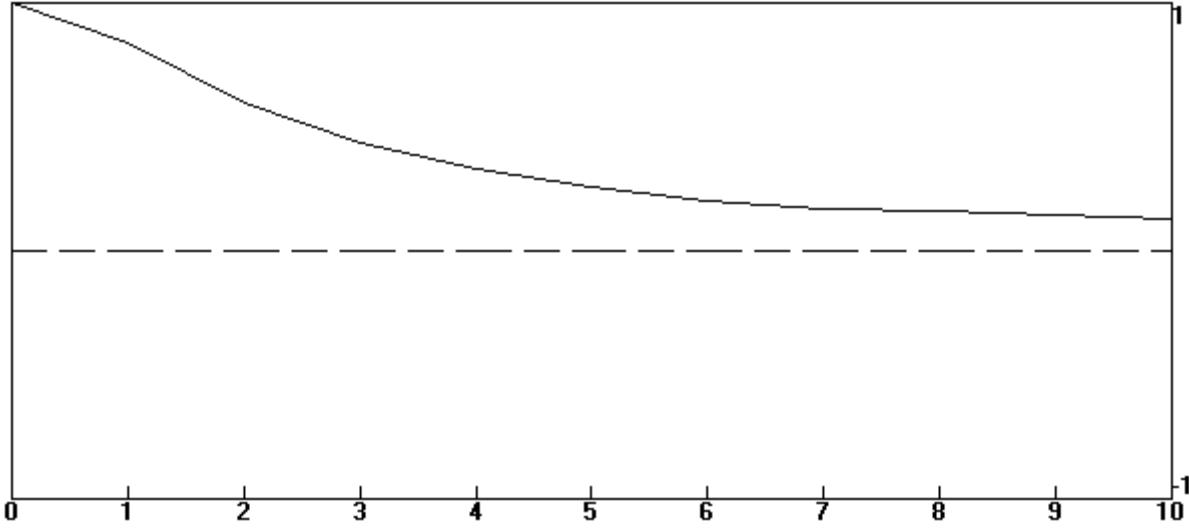


Figure 6: Estimated autocorrelation function $\hat{\rho}(m)$ of the ARMA(1,1) process (50)

Therefore, the partial and regular autocorrelation functions are of no help in distinguishing an ARMA model from an AR model.

So how to proceed? My recommendation is the following: Select an upper bound \bar{p} of p , for example on the basis of the PAC plot, and select an upper bound \bar{q} of q . The latter is a matter of guess work; there is no rule of thumb for this. Then try all ARMA(p,q) models with $p \leq \bar{p}$, $q \leq \bar{q}$, and pick the model with the lowest value of one of the information criteria. This can be done automatically in EasyReg, via Menu > Single equation models > ARIMA model selection via information criteria. For example in the case of (50), with $\bar{p} = \bar{q} = 4$, the Hannan-Quinn information criterion takes its smallest value for $p = q = 1$, which are the true values.

8.4 Forecasting with an ARMA model

In EasyReg the AR(∞) representation of an ARMA model is used as forecasting scheme, because for covariance stationary Gaussian processes it represents the conditional expectation function. For example, in the ARMA(1,1) case the forecasting scheme for Y_{n+1} given its past up to time n is

$$\hat{Y}_{n+1} = \beta_0 / (1 - \theta_1) + (\beta_1 - \theta_1) \sum_{j=0}^{\infty} \theta_1^j Y_{n-j}, \quad (54)$$

where n is the last observed time period. Compare (46). In practice we have to replace the

parameters involved by estimates. Moreover, usually we do not observe all values of Y_{n-j} , but only for $n-j \geq 1$, say. Therefore, replace Y_t for $t < 1$ in (54) by its sample mean $\bar{Y} = (1/n)\sum_{t=1}^n Y_t$. Thus, the actual forecast of Y_{n+1} is:

$$\begin{aligned}\tilde{Y}_{n+1|n} &= \hat{\beta}_0/(1-\hat{\theta}_1) + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=0}^{n-1}\hat{\theta}_1^j Y_{n-j} + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=n}^{\infty}\hat{\theta}_1^j \bar{Y} \\ &= \frac{\hat{\beta}_0}{1-\hat{\theta}_1} + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=0}^{n-1}\hat{\theta}_1^j Y_{n-j} + \frac{(\hat{\beta}_1-\hat{\theta}_1)\hat{\theta}_1^n \bar{Y}}{1-\hat{\theta}_1},\end{aligned}\quad (55)$$

where $\hat{\beta}_1$ and $\hat{\theta}_1$ are the estimates of β_1 and θ_1 , respectively, based on the data up to time n .

To forecast Y_{n+2} given its past up to time n , replace n in (55) by $n+1$, and the unobserved Y_{n+1} by its forecast:

$$\tilde{Y}_{n+2|n} = \hat{\beta}_0/(1-\hat{\theta}_1) + (\hat{\beta}_1-\hat{\theta}_1)\tilde{Y}_{n+1|n} + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=1}^n\hat{\theta}_1^j Y_{n+1-j} + \frac{(\hat{\beta}_1-\hat{\theta}_1)\hat{\theta}_1^{n+1} \bar{Y}}{1-\hat{\theta}_1}. \quad (56)$$

This procedure is called *recursive forecasting*. More generally, the h step ahead recursive forecast of Y_{n+h} given its past up to time n is

$$\tilde{Y}_{n+h|n} = \frac{\hat{\beta}_0}{1-\hat{\theta}_1} + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=0}^{h-2}\hat{\theta}_1^j \tilde{Y}_{n+h-1-j|n} + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=h+1}^{n+h-2}\hat{\theta}_1^j Y_{n+h-1-j} + \frac{(\hat{\beta}_1-\hat{\theta}_1)\hat{\theta}_1^{n+h-1} \bar{Y}}{1-\hat{\theta}_1}. \quad (57)$$

Note however, that in this case

$$\begin{aligned}\lim_{h \rightarrow \infty} \tilde{Y}_{n+h|n} &= \hat{\beta}_0/(1-\hat{\theta}_1) + (\hat{\beta}_1-\hat{\theta}_1)\lim_{h \rightarrow \infty} \sum_{j=0}^{h-2}\hat{\theta}_1^j \tilde{Y}_{n+h-1-j|n} \\ &= \hat{\beta}_0/(1-\hat{\theta}_1) + (\hat{\beta}_1-\hat{\theta}_1)\sum_{j=0}^{\infty}\hat{\theta}_1^j \lim_{h \rightarrow \infty} \tilde{Y}_{n+h|n} = \hat{\beta}_0/(1-\hat{\theta}_1) + \left((\hat{\beta}_1-\hat{\theta}_1)/(1-\hat{\theta}_1)\right) \lim_{h \rightarrow \infty} \tilde{Y}_{n+h|n} \\ &= \hat{\beta}_0/(1-\hat{\beta}_1)\end{aligned}\quad (58)$$

which is just the estimate of $\mu = E[Y_t]$. Compare (47). Thus, if we choose the forecast horizon h too large, the recursive forecast $\tilde{Y}_{n+h|n}$ will be close to the expectation $\mu = E[Y_t]$.

The forecast issue will be treated in more detail separately. See Bierens (2008b).

9. ARMA models for seasonal time series

9.1 Seasonal dummy variables

The effect of seasonality may manifest itself through seasonal varying expectations as well as seasonal patterns in the AR and/or MA lag polynomials. As to the former, time varying expectations can easily be modeled using seasonal dummy variables. For example, if Y_t is a quarterly time series, $E[Y_t]$ can be modeled either by

$$E[Y_t] = \mu_0 + \mu_1 Q_{1,t} + \mu_2 Q_{2,t} + \mu_3 Q_{3,t} \quad (59)$$

or

$$E[Y_t] = \mu_1^* Q_{1,t} + \mu_2^* Q_{2,t} + \mu_3^* Q_{3,t} + \mu_4^* Q_{4,t}, \quad (60)$$

where the $Q_{s,t}$'s are seasonal dummy variables:

$$Q_{s,t} = 1 \text{ if the quarter of } t \text{ is } s, Q_{s,t} = 0 \text{ if not.} \quad (61)$$

The equivalence of (59) and (60) follows from the fact that $\sum_{s=1}^4 Q_{s,t} = 1$, so that

$$\begin{aligned} E[Y_t] &= \mu_1^* Q_{1,t} + \mu_2^* Q_{2,t} + \mu_3^* Q_{3,t} + \mu_4^* (1 - Q_{1,t} - Q_{2,t} - Q_{3,t}) \\ &= \mu_4^* + (\mu_1^* - \mu_4^*) Q_{1,t} + (\mu_2^* - \mu_4^*) Q_{2,t} + (\mu_3^* - \mu_4^*) Q_{3,t}, \end{aligned} \quad (62)$$

hence $\mu_0 = \mu_4^*$, $\mu_1 = \mu_1^* - \mu_4^*$, $\mu_2 = \mu_2^* - \mu_4^*$, $\mu_3 = \mu_3^* - \mu_4^*$.

Note that if we had defined (60) as $E[Y_t] = \mu_0^* + \mu_1^* Q_{1,t} + \mu_2^* Q_{2,t} + \mu_3^* Q_{3,t} + \mu_4^* Q_{4,t}$, the parameters involved are no longer identified, because then (62) becomes

$$E[Y_t] = (\mu_0^* + \mu_4^*) + (\mu_1^* - \mu_4^*) Q_{1,t} + (\mu_2^* - \mu_4^*) Q_{2,t} + (\mu_3^* - \mu_4^*) Q_{3,t}, \quad (63)$$

which is also equivalent to (59). Hence

$$\mu_0 = \mu_0^* + \mu_4^*, \mu_1 = \mu_1^* - \mu_4^*, \mu_2 = \mu_2^* - \mu_4^*, \mu_3 = \mu_3^* - \mu_4^*, \quad (64)$$

which is a system of four equations in five unknowns.

The presence of seasonally varying expectations can be observed from the autocorrelation function. For example, let Y_t be a quarterly time series satisfying

$$Y_t = \mu_0 + \mu_1 Q_{1,t} + \mu_2 Q_{2,t} + \mu_3 Q_{3,t} + X_t \quad (65)$$

where X_t is zero-mean covariance stationary with covariance function $\gamma_x(m) = E(X_t X_{t-m})$. The sample average of Y_t is

$$\begin{aligned} \bar{Y}_n &= \mu_0 + \mu_1 (1/n) \sum_{t=1}^n Q_{1,t} + \mu_2 (1/n) \sum_{t=1}^n Q_{2,t} + \mu_3 (1/n) \sum_{t=1}^n Q_{3,t} + (1/n) \sum_{t=1}^n X_t \\ &\approx \mu_0 + 0.25\mu_1 + 0.25\mu_2 + 0.25\mu_3 \end{aligned} \quad (66)$$

if n is large, because for each $s = 1,2,3$ the fraction of values of $Q_{s,t}$ for $t = 1,\dots,n$ that are equal to 1 tends towards 0.25 if $n \rightarrow \infty$, and $\text{plim}_{n \rightarrow \infty} (1/n) \sum_{t=1}^n X_t = E[X_t] = 0$ by the law of large numbers.

Then it is not hard to show that there exists constants c_s , $s = 1,2,3,4$, such that for $n \rightarrow \infty$,

$$\frac{1}{n-m} \sum_{t=m+1}^n (Y_t - \bar{Y}_n)(Y_{t-m} - \bar{Y}_n) \rightarrow \begin{cases} \gamma_x(m) + c_1 & \text{for } m = 0,4,8,12,\dots \\ \gamma_x(m) + c_2 & \text{for } m = 1,5,9,13,\dots \\ \gamma_x(m) + c_3 & \text{for } m = 2,6,10,14,\dots \\ \gamma_x(m) + c_4 & \text{for } m = 3,7,11,15,\dots \end{cases} \quad (67)$$

in probability. It follows now from (39) and (67) that the estimated autocorrelation function $\hat{\rho}(m)$ will have spikes at distances of four lags, and will not die out to zero. For example consider the quarterly process

$$Y_t = 1 + 2Q_{1,t} - Q_{2,t} - 2Q_{3,t} + X_t, \text{ where } X_t \sim \text{i.i.d. } N(0,1), \quad (68)$$

for $t = 1,2,\dots,225$. The estimated autocorrelation function $\hat{\rho}(m)$ of this process is displayed in Figure 7.

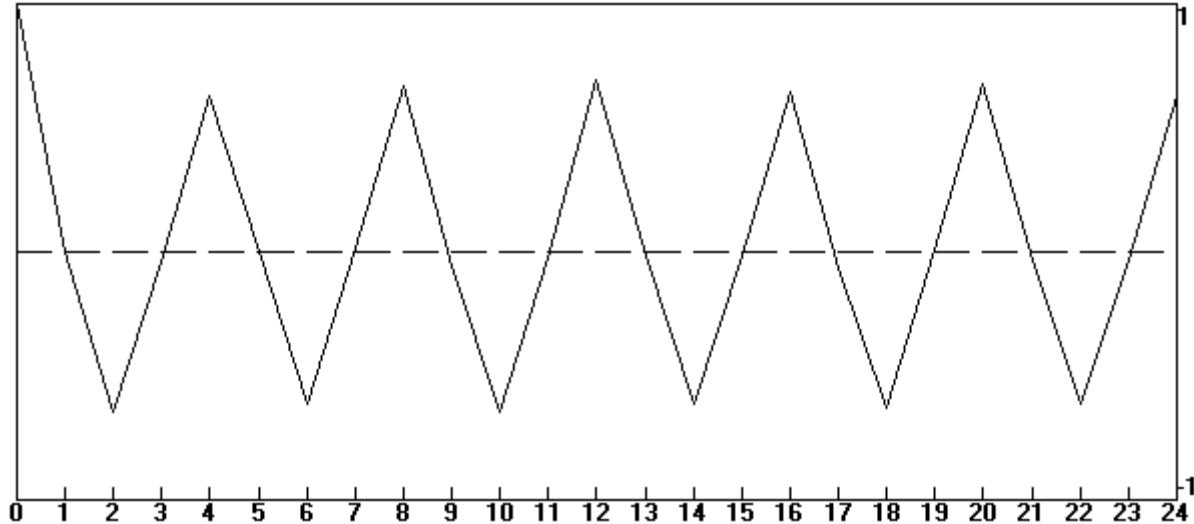


Figure 7: Estimated autocorrelation function $\hat{\rho}(m)$ of quarterly process (68)

9.2 Seasonal lag polynomials

Seasonality may also occur in the process X_t in (65) itself. For example, let X_t be a quarterly ARMA process

$$\varphi_p(L)\lambda_r(L^4)X_t = \psi_q(L)\eta_s(L^4)U_t, \quad (69)$$

where $\varphi_p(L)$ and $\psi_q(L)$ are the non-seasonal AR and MA lag polynomials of orders p and q , respectively, defined before, and $\lambda_r(z)$ and $\eta_s(z)$ are the seasonal AR and MA polynomials of orders r and s , respectively.

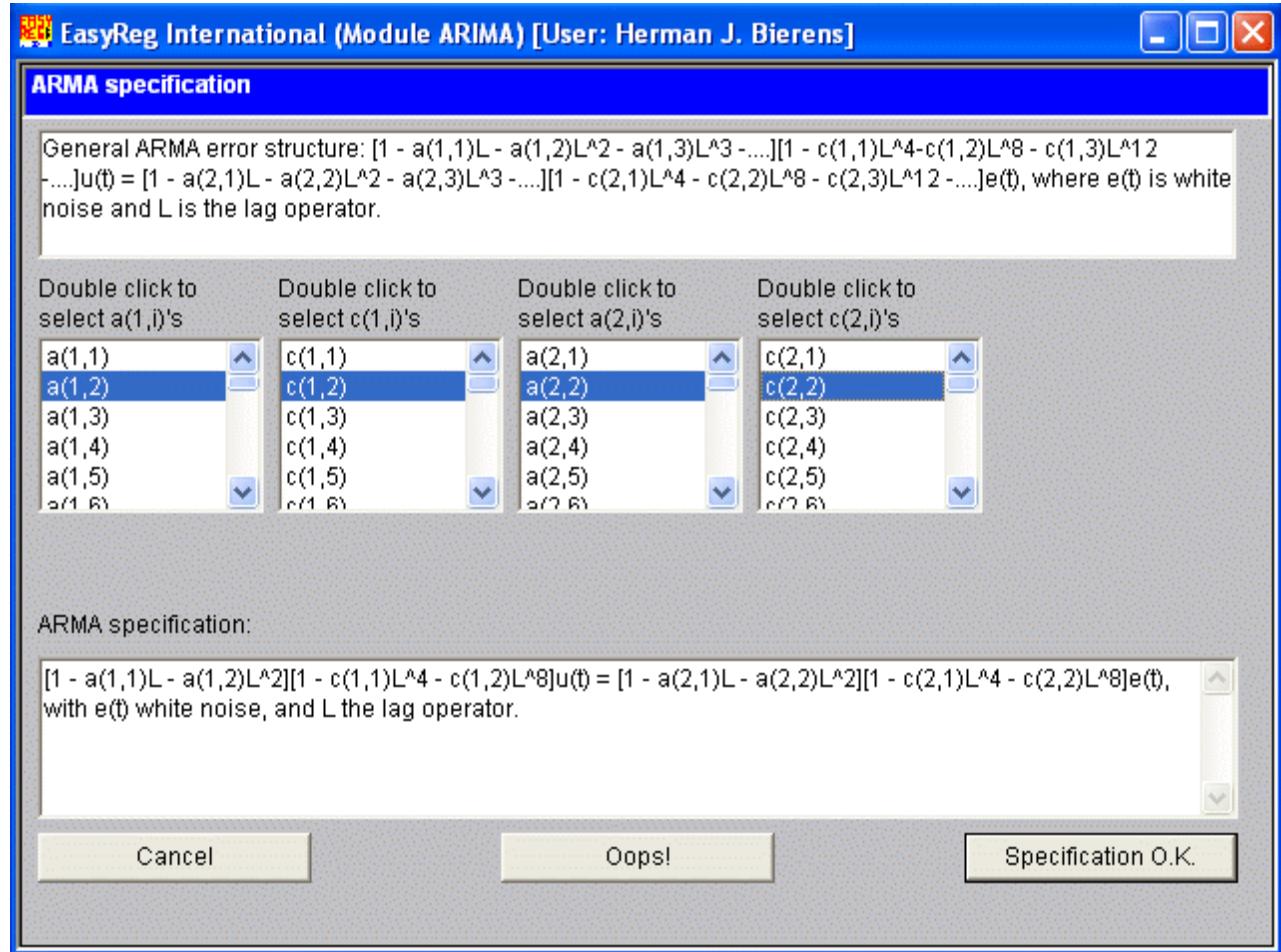


Figure 8: Specification of a seasonal ARMA model in EasyReg.

In EasyReg these polynomials are specified via the window displayed in Figure 8. The coefficients $a(1,i)$, $i = 1, \dots, p$, are the coefficients of the non-seasonal AR polynomial $\varphi_p(L)$, the coefficients $a(2,i)$, $i = 1, \dots, q$, are the coefficients of the non-seasonal MA polynomial $\psi_q(L)$, the coefficients $c(1,i)$, $i = 1, \dots, r$, are the coefficients of the seasonal AR polynomial $\lambda_r(L^4)$, and the coefficients $c(2,i)$, $i = 1, \dots, s$, are the coefficients of the seasonal MA polynomial $\eta_s(L^4)$. The

displayed specification is for $p = q = r = s = 2$.

The specification procedure for p, q, r and s is similar to the non-seasonal ARMA case: First, specify upper bounds of p, q, r and s , and then use the information criteria to select the correct p, q, r and s , via Menu > Single equation models > ARIMA model selection via information criteria.

References

- Akaike, H. (1974): "A New Look at the Statistical Model Identification," *I.E.E.E. Transactions on Automatic Control, AC 19*, 716-723.
- Akaike, H. (1976): "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion," in R. K. Mehra and D. G. Lainotis (eds.), *System Identification: Advances and Case Studies*, Academic Press, New York, 52-107.
- Bierens, H. J. (2004): *Introduction to the Mathematical and Statistical Foundations of Econometrics*, Cambridge University Press, Cambridge, UK.
- Bierens, H. J. (2008a): *EasyReg International*, free econometrics software package, downloadable from <http://econ.la.psu.edu/~hbierens/EASYREG.HTM>
- Bierens, H. J. (2008b): *Forecasting*, Lecture note, downloadable from <http://econ.la.psu.edu/~hbierens/FORECAST.PDF>
- Hannan, E. J., and B. G. Quinn (1979): "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, B*, 41, 190-195.
- Schwarz, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.

Information Criteria and Model Selection

Herman J. Bierens

Pennsylvania State University

March 12, 2006

1. Introduction

Let $L_n(k)$ be the maximum likelihood of a model with k parameters based on a sample of size n , and let k_0 be the correct number of parameters. Suppose that for $k > k_0$ the model with k parameters is nested in the model with k_0 parameters, so that $L_n(k_0)$ is obtained by setting $k - k_0$ parameters in the larger model to constants. Without loss of generality we may assume that these constants are zeros. Thus, denoting the likelihood function of the least parsimonious model by $\hat{L}_n(\theta)$, $\theta \in \Theta \subset \mathbb{R}^m$,

$$L_n(k) = \max_{\theta \in \Theta_k} \hat{L}_n(\theta), \text{ where } \Theta_k = \left\{ \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \in \Theta : \theta_2 = 0 \in \mathbb{R}^{m-k} \right\} \quad (1)$$

for $k \leq m$. Thus, the models with $k < k_0$ parameters are misspecified, and the models with $k > k_0$ parameters are correctly specified but over-parametrized.

The Akaike (1974, 1976), Hannan-Quinn (1979), and Schwarz (1978) information criteria for selecting the most parsimonious correct model are

$$\text{Akaike: } c_n(k) = -2.\ln(L_n(k))/n + 2k/n,$$

$$\text{Hannan-Quinn: } c_n(k) = -2.\ln(L_n(k))/n + 2k.\ln(\ln(n))/n,$$

$$\text{Schwarz: } c_n(k) = -2.\ln(L_n(k))/n + k.\ln(n)/n,$$

respectively. Since the Schwarz information criterion is derived using Bayesian arguments, this criterion is also known as the Bayesian Information Criterion (BIC).

These criteria take the general form

$$c_n(k) = -2.\ln(L_n(k))/n + k.\varphi(n)/n, \quad (2)$$

where $\varphi(n) = 2$ in the Akaike case, $\varphi(n) = 2.\ln(\ln(n))$ in the Hannan-Quinn case, and $\varphi(n) = \ln(n)$ in the Schwarz case. Using these criteria, the model is selected that corresponds to

$$\hat{k} = \operatorname{argmin}_{k \leq m} c_n(k). \quad (3)$$

2. Consistency

If $k < k_0$ then the model with k parameters is misspecified, so that

$$\text{plim}_{n \rightarrow \infty} \ln(L_n(k))/n < \text{plim}_{n \rightarrow \infty} \ln(L_n(k_0))/n. \quad (4)$$

Hence, it follows from (2), (4) and $\lim_{n \rightarrow \infty} \varphi(n)/n = 0$ that in all three cases

$$\begin{aligned} \lim_{n \rightarrow \infty} P[c_n(k_0) \geq c_n(k)] \\ = \lim_{n \rightarrow \infty} P[-2\ln(L_n(k_0))/n + k_0 \cdot \varphi(n)/n \geq -2\ln(L_n(k))/n + k \cdot \varphi(n)/n] \\ = \lim_{n \rightarrow \infty} P[\ln(L_n(k_0))/n - \ln(L_n(k))/n \leq 0.5(k_0 - k) \cdot \varphi(n)/n] = 0, \end{aligned} \quad (5)$$

so that

$$\begin{aligned} \lim_{n \rightarrow \infty} P[\hat{k} < k_0] &\leq \lim_{n \rightarrow \infty} P[c_n(k_0) \geq c_n(k) \text{ for some } k < k_0] \\ &\leq \sum_{k < k_0} \lim_{n \rightarrow \infty} P[c_n(k_0) \geq c_n(k)] = 0 \end{aligned} \quad (6)$$

For $k > k_0$ it follows from the likelihood ratio test that

$$2(\ln(L_n(k)) - \ln(L_n(k_0))) \xrightarrow{d} X_{k-k_0} \sim \chi^2_{k-k_0}, \quad (7)$$

where \xrightarrow{d} indicates convergence in distribution. Then in the Akaike case,

$$n(c_n(k_0) - c_n(k)) = 2(\ln(L_n(k)) - \ln(L_n(k_0))) - 2(k - k_0) \xrightarrow{d} X_{k-k_0} - 2(k - k_0),$$

hence

$$\lim_{n \rightarrow \infty} P[c_n(k_0) > c_n(k)] = P[X_{k-k_0} > 2(k - k_0)] > 0.$$

Therefore, the Akaike criterion may asymptotically overshoot the correct number of parameters:

$$\lim_{n \rightarrow \infty} P[\hat{k} \geq k_0] = 1, \text{ but } \lim_{n \rightarrow \infty} P[\hat{k} > k_0] > 0,$$

Since in the Hannan-Quinn and Schwarz cases, $\lim_{n \rightarrow \infty} \varphi(n) = \infty$, (7) implies that in these two cases

$$\text{plim}_{n \rightarrow \infty} -2(\ln(L_n(k_0)) - \ln(L_n(k)))/\varphi(n) = 0$$

hence

$$\text{plim}_{n \rightarrow \infty} n(c_n(k_0) - c_n(k))/\varphi(n) = \text{plim}_{n \rightarrow \infty} -2(\ln(L_n(k_0)) - \ln(L_n(k)))/\varphi(n) + k_0 - k = k_0 - k \leq -1$$

so that

$$\lim_{n \rightarrow \infty} P[c_n(k_0) \geq c_n(k)] = 0.$$

This implies, similar to (6), that $\lim_{n \rightarrow \infty} P[\hat{k} > k_0] = 0$. Thus, in the Hannan-Quinn and Schwarz cases,

$$\lim_{n \rightarrow \infty} P[\hat{k} = k_0] = 1. \quad (8)$$

Note that the consistency result (8) holds for any criterion of the type (2) with

$$\lim_{n \rightarrow \infty} \varphi(n)/n = 0 \text{ and } \lim_{n \rightarrow \infty} \varphi(n) = \infty, \quad (9)$$

for example, let $\varphi(n) = \sqrt{n}$.

3. Applications

3.1 VAR and AR model selection

Let $L_n(k)$ be the maximum likelihood of a d -variate Gaussian VAR(p) model,

$$Y_t = a_0 + \sum_{j=1}^p A_j Y_{t-j} + U_t, \quad U_t \sim i.i.d. N_d[0, \Sigma],$$

where $Y_t \in \mathbb{R}^d$ is observed for $t = 1-p, \dots, n$. Then $k = d + d^2 p$ and

$$\ln(L_n(k)) = -\frac{1}{2}n.d - \frac{1}{2}n.d.\ln(2\pi) - \frac{1}{2}n.\ln(\det(\hat{\Sigma}_p)),$$

where $\hat{\Sigma}_p$ is the maximum likelihood estimator of the error variance Σ . Hence,

$$-2.\ln(L_n(k))/n = \ln(\det(\hat{\Sigma}_p)) + d.(1 + \ln(2\pi)). \quad (10)$$

The second term does not depend on p . Therefore, the model is selected that corresponds to

$$\hat{p} = \operatorname{argmin}_p c_n^{VAR}(p), \text{ where}$$

Akaike:	$c_n^{VAR}(p) = \ln(\det(\hat{\Sigma}_p)) + 2(d+d^2p)/n,$
Hannan-Quinn:	$c_n^{VAR}(p) = \ln(\det(\hat{\Sigma}_p)) + 2(d+d^2p)\ln(\ln(n))/n,$
Schwarz:	$c_n^{VAR}(p) = \ln(\det(\hat{\Sigma}_p)) + (d+d^2p)\ln(n)/n.$

Similarly, these criteria can also be used to determine the order p of an AR(p) model:

$$Y_t = a_0 + \sum_{j=1}^p \alpha_j Y_{t-j} + U_t, \quad U_t \sim i.i.d. N[0, \sigma^2],$$

where again $Y_t \in \mathbb{R}$ is observed for $t = 1-p, \dots, n$, simply by replacing d with 1 and $\det(\hat{\Sigma}_p)$ with the ML estimator $\hat{\sigma}_p^2$ of the error variance σ^2 :

Akaike:	$c_n^{AR}(p) = \ln(\hat{\sigma}_p^2) + 2(1+p)/n,$
Hannan-Quinn:	$c_n^{AR}(p) = \ln(\hat{\sigma}_p^2) + 2(1+p)\ln(\ln(n))/n,$
Schwarz:	$c_n^{AR}(p) = \ln(\hat{\sigma}_p^2) + (1+p)\ln(n)/n.$

3.2 ARMA model specification

Similarly, in the ARMA(p,q) case

$$Y_t = a_0 + \sum_{j=1}^p \alpha_j Y_{t-j} + U_t - \sum_{j=1}^q \beta_j U_{t-j}, \quad U_t \sim i.i.d. N[0, \sigma^2],$$

these criteria become

$$\begin{aligned} \text{Akaike: } c_n^{ARMA}(p,q) &= \ln(\hat{\sigma}_{p,q}^2) + 2(1+p+q)/n, \\ \text{Hannan-Quinn: } c_n^{ARMA}(p,q) &= \ln(\hat{\sigma}_{p,q}^2) + 2(1+p+q)\ln(\ln(n))/n, \\ \text{Schwarz: } c_n^{ARMA}(p,q) &= \ln(\hat{\sigma}_{p,q}^2) + (1+p+q)\ln(n)/n, \end{aligned}$$

where now $\hat{\sigma}_{p,q}^2$ is the ML estimator of the error variance σ^2 and n is the number of observations used in the ML estimation.

It can be shown [see Hannan (1980)] that in the case of common roots in the AR and MA polynomials the Hannan-Quinn and Schwarz criteria still select the correct orders p and q consistently: Given upper bounds $\bar{p} \geq p_0$ and $\bar{q} \geq q_0$, where p_0 and q_0 are the correct orders of an ARMA(p,q) process, we have $\lim_{n \rightarrow \infty} P[\hat{p} = p_0, \hat{q} = q_0] = 1$, where

$$(\hat{p}, \hat{q}) = \operatorname{argmin}_{0 \leq p \leq \bar{p}, 0 \leq q \leq \bar{q}} c_n^{ARMA}(p,q).$$

3.3 ARCH and GARCH models

If a model is extended to include ARCH or GARCH errors, it is recommended to subtract the term $1 + \ln(2\pi)$ from $-2.\ln(L_n(k))/n$ [see (10)] in the formula for the information criteria, in order to make these criteria comparable with those for the model without (G)ARCH errors. Thus,

$$\begin{aligned} \text{Akaike: } c_n^{(G)ARCH}(k) &= -2.\ln(L_n(k))/n + 2k/n - 1 - \ln(2\pi), \\ \text{Hannan-Quinn: } c_n^{(G)ARCH}(k) &= -2.\ln(L_n(k))/n + 2k.\ln(\ln(n))/n - 1 - \ln(2\pi), \\ \text{Schwarz: } c_n^{(G)ARCH}(k) &= -2.\ln(L_n(k))/n + k.\ln(n)/n - 1 - \ln(2\pi), \end{aligned}$$

where again k is the number of parameters, including the (G)ARCH parameters.

References

- Akaike, H. (1974): "A New Look at the Statistical Model Identification," *I.E.E.E. Transactions on Automatic Control*, AC 19, 716-723.
- Akaike, H. (1976): "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion," in R. K. Mehra and D. G. Lainotis (eds.), *System Identification: Advances and Case Studies*, Academic Press, New York, 52-107.

Hannan, E. J. (1980): "The Estimation of the Order of an ARMA Process", *Annals of Statistics*, 8, 1071-1081.

Hannan, E. J., and B. G. Quinn (1979): "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, B*, 41, 190-195.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.

VECTOR TIME SERIES AND INNOVATION RESPONSE ANALYSIS

Herman J. Bierens

Pennsylvania State University

(Revised: October 15, 2007)

1. *The Wold decomposition theorem, and the vector auto-regressive (VAR) and vector moving average (VMA) representations*

Let X_t be a k -variate covariance stationary time series process: $X_t \in \mathbb{R}^k$, $E(X_t) = \mu$, $E[(X_t - \mu)(X_{t-m} - \mu)^T] = \Xi(m)$, where the expectation vector μ and the covariance matrices $\Xi(m)$ do not depend on the time index t . The multivariate Wold decomposition theorem states that the process X_t has the Wold representation

$$X_t = \sum_{s=0}^{\infty} \Gamma_s U_{t-s} + W_t, \quad (1)$$

where the $k \times k$ matrices Γ_s are such that

$$\Gamma_0 = I_k, \quad \sum_{s=1}^{\infty} \Gamma_s \Gamma_s^T \text{ converges}, \quad (2)$$

the process U_t is k -variate white noise:

$$U_t \in \mathbb{R}^k, \quad E(U_t) = 0, \quad E(U_t U_t^T) = \Sigma, \quad E(U_t U_{t-m}^T) = O \text{ for } m \neq 0, \quad (3)$$

and $W_t \in \mathbb{R}^k$ is a linear deterministic process: there exists a k -vector c_0 and $k \times k$ matrices C_s

such that without error,

$$W_t = c_0 + \sum_{s=1}^{\infty} C_s W_{t-s}, \quad \text{and } E[U_t W_{t-m}^T] = O \text{ for } m = 0, \pm 1, \pm 2, \dots \quad (4)$$

Assuming that the deterministic process W_t is constant, it follows from (1) that W_t must be equal to the expectation of X_t : $W_t = E(X_t) = \mu$. Then the Wold decomposition theorem implies that the covariance stationary process X_t is a Vector Moving Average (VMA) process, possibly of infinite order:

$$X_t = \mu + \sum_{m=0}^{\infty} \Gamma_m U_{t-m} = \mu + \Gamma(L)U_t, \text{ where } \Gamma_0 = I_k, \Gamma(L) = I_k + \sum_{m=1}^{\infty} \Gamma_m L^m. \quad (5)$$

Now suppose that there exists a matrix valued lag polynomial $C(L)$ such that $C(L)\Gamma(L) = I_k$. Note that $C(0) = I_k$ because $\Gamma(0) = I_k$, so that we can write

$$\Gamma(L)^{-1} = C(L) = I_k - \sum_{m=1}^{\infty} C_m L^m. \quad (6)$$

Then

$$X_t = \eta + \sum_{s=1}^{\infty} C_s X_{t-s} + U_t, \quad (7)$$

where $\eta = C(1)\mu = \Gamma(0)^{-1}\mu$. This is the Vector Auto-Regressive (VAR) representation.

Next, assume in addition to the covariance stationarity condition that X_t is a Gaussian process: for arbitrary $m \geq 1$ and arbitrary indices $t_1 < t_2 < \dots < t_m$, the vector $(X_{t_1}^T, \dots, X_{t_m}^T)^T$ is jointly normally distributed. Then U_t is a Gaussian process, and since the U_t 's are serially uncorrelated, they are independent [Exercise: Why?] and normally distributed: U_t is i.i.d. $N_k[0, \Sigma]$. Consequently, X_t is then strictly stationary: the distribution of $(X_{t_1}^T, \dots, X_{t_m}^T)^T$ only depends on the differences of the indices t_j , and not on their levels, and moreover,

$$E(X_t | X_{t-1}, X_{t-2}, \dots) = \eta + \sum_{s=1}^{\infty} C_s X_{t-s}, \quad (8)$$

because $E(U_t) = 0$ and U_t is independent of X_{t-m} for $m > 0$, so that

$$E(U_t | X_{t-1}, X_{t-2}, \dots) = E(U_t) = 0. \quad (9)$$

2. VAR(p) models

All linear vector time series models are approximations of model (7). In particular, the assumption of the VAR(p) model is that in model (7), $C_s = O$ for $s > p$:

$$X_t = \eta + \sum_{s=1}^p C_s X_{t-s} + U_t. \quad (10)$$

A necessary condition for the strict stationarity of the $VAR(p)$ model is that the error process U_t is strictly stationary, and the lag polynomial

$$C(L) = I_k - C_1 L - \dots - C_p L^p \quad (11)$$

can be inverted: $C(L)^{-1} = \Gamma(L)$, where $\Gamma(L)$ is the same as before, because then $X_t = \mu + \Gamma(L)U_t$, where the right-hand side is a moving average of a stationary process and therefore stationary itself.

As is well-known, the inverse of $C(L)$ is:

$$C(L)^{-1} = [c^{ij}(L)] = \frac{1}{\det C(L)} [\text{cof}_{j,i}\{C(L)\}]. \quad (12)$$

Thus, denoting

$$C^*(L) = [\text{cof}_{j,i}\{C(L)\}] \quad (13)$$

we have

$$C^*(L)C(L) = \det(C(L))I \quad (14)$$

hence:

$$\det(C(L))X_t = C^*(1)\eta + C^*(L)U_t. \quad (15)$$

Note that $C^*(L)$ consists of finite-order lag polynomials, and that $\det(C(L))$ is a finite-order lag polynomial. Now it follows from the condition for stationarity of univariate $AR(p)$ processes that:

PROPOSITION 1: *A necessary condition for the (strict) stationarity of the $VAR(p)$ process (10) is that all the roots of $\det(C(L))$ are located outside the complex unit circle.*

3. Granger causality

Consider the strictly stationary bi-variate vector time series process $Z_t = (x_t, y_t)^T$. As is well-known (or it ought to be), the best one-step ahead forecast of Z_t given the whole past of the

process Z_t is the conditional expectation

$$E[Z_t | Z_{t-1}, Z_{t-2}, \dots] = \begin{pmatrix} E[x_t | Z_{t-1}, Z_{t-2}, \dots] \\ E[y_t | Z_{t-1}, Z_{t-2}, \dots] \end{pmatrix}. \quad (16)$$

Now if the past of the y_t process does not contribute to the best forecast of x_t , one says that y_t does *not* Granger cause x_t :

DEFINITION 1: y_t does not Granger-cause x_t if

$$E[x_t - E(x_t | x_{t-1}, y_{t-1}, x_{t-2}, y_{t-2}, \dots)]^2 = E[x_t - E(x_t | x_{t-1}, x_{t-2}, \dots)]^2. \quad (17)$$

If y_t Granger-causes x_t , then one can predict x_t better using the whole past of the x_t and y_t processes than using only the past of x_t :

$$E[x_t - E(x_t | x_{t-1}, y_{t-1}, x_{t-2}, y_{t-2}, \dots)]^2 < E[x_t - E(x_t | x_{t-1}, x_{t-2}, \dots)]^2. \quad (18)$$

Suppose that Z_t is a Gaussian VAR(p) process:

$$Z_t = \eta + C_1 Z_{t-1} + \dots + C_p Z_{t-p} + U_t, \quad U_t \text{ is i.i.d. } N_2(0, \Sigma). \quad (19)$$

Then

$$E(Z_t | Z_{t-1}, Z_{t-2}, \dots) = \eta + \sum_{s=1}^p C_s Z_{t-s} \quad (20)$$

hence

$$E(x_t | Z_{t-1}, Z_{t-2}, \dots) = \eta_1 + \sum_{s=1}^p (c_{1,1,s} x_{t-s} + c_{1,2,s} y_{t-s}), \quad (21)$$

where η_1 is the first component of η and

$$C_s = \begin{pmatrix} c_{1,1,s} & c_{1,2,s} \\ c_{2,1,s} & c_{2,2,s} \end{pmatrix}. \quad (22)$$

Now y_t does *not* Granger cause x_t if $c_{1,2,s} = 0$ for $s = 1, \dots, p$, so that then the matrices C_s are lower-triangular, as then the lagged y_t 's disappear from the right-hand side of (21).

Finally, the above argument easily extends to the case where y_t and/or x_t are vectors themselves.

4. Sims' nonstructural VAR innovation response analysis

In his seminal paper "Macroeconomics and Reality", Sims (1980) works with a vector X_t of macro-economics variables (dimension $k = 6$). He assumes for X_t a stationary VAR(p) process:

$$C(L)X_t = X_t - C_1X_{t-1} - \dots - C_pX_{t-p} = \eta + U_t, \text{ with } U_t \text{ i.i.d. } N(0, \Sigma). \quad (23)$$

Since by stationarity all the roots of $\det(C(L))$ are located outside the unit circle, the lag polynomial $C(L)$ is invertible: $C(L)^{-1} = \Gamma(L)$. We can now write the VAR(p) process as a VMA(∞) process:

$$X_t = \mu + \sum_{s=0}^{\infty} \Gamma_s U_{t-s}, \quad \Gamma_0 = I, \quad (24)$$

which is just the Wold representation.

Since the innovations U_t are i.i.d. $N(0, \Sigma)$ we have for $m \geq 0$

$$E(X_{t+m}|U_t) = \Gamma_m U_t + \mu. \quad (25)$$

The expected "impact" of the first component $u_{1,t}$ of U_t on X_{t+m} is:

$$E(X_{t+m}|u_{1,t}) = E[E(X_{t+m}|U_t)|u_{1,t}] = \Gamma_m E(U_t|u_{1,t}) + \mu. \quad (26)$$

The *innovation response* of $u_{1,t}$ on X_{t+m} is now defined as

$$E(X_{t+m}|u_{1,t}) - E(X_{t+m}) = \Gamma_m E(U_t|u_{1,t}) = \Gamma_m \begin{pmatrix} u_{1,t} \\ E(u_{2,t}|u_{1,t}) \\ \vdots \\ E(u_{k,t}|u_{1,t}) \end{pmatrix}. \quad (27)$$

We see that the innovation $u_{1,t}$ has two effects on X_{t+m} : a direct effect and an indirect effect via $E(u_{j,t}|u_{1,t})$ for $j > 1$.

Sims (1980, 1982) proposes to calculate $E(U_t|u_{1,t})$ as follows. Observe that $\text{Var}(U_t) = \Sigma$. Since Σ is positive definite we can write $\Sigma = \Delta\Delta^T$, where Δ is a lower triangular matrix:

$$\Delta = \begin{pmatrix} \delta_{1,1} & 0 & 0 & \dots & 0 \\ \delta_{2,1} & \delta_{2,2} & 0 & \dots & 0 \\ \delta_{3,1} & \delta_{3,2} & \delta_{3,3} & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \delta_{k,1} & \delta_{k,2} & \delta_{k,3} & \dots & \delta_{k,k} \end{pmatrix} \quad (28)$$

with positive diagonal elements. Let $e_t = \Delta^{-1}U_t$. Then e_t is i.i.d. $N(0, I)$. Note that $u_{1,t} = \delta_{1,1}e_{1,t}$ with $e_{1,t}$ the first element of e_t . We can now write:

$$E(U_t | u_{1,t}) = \Delta E(e_t | e_{1,t}) = \Delta \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} e_{1,t} = \delta_1 e_{1,t} \quad (29)$$

with δ_1 the first column of Δ . Thus the response of $u_{1,t}$ on X_{t+m} is $\Gamma_m \delta_1 e_{1,t}$. Sims replaces in his analysis $e_{1,t}$ by its standard error (=1). He calls it a unit shock. Thus the *innovation response* of a unit shock in the first component of X_t on X_{t+m} is given by $\Gamma_m \delta_1$, for $m = 0, 1, 2, \dots$

We have seen that U_t can be written as $U_t = \Delta e_t$, with Δ a lower triangular matrix and e_t i.i.d. $N(0, I)$. Thus:

$$U_t = \begin{pmatrix} u_{1,t} \\ u_{2,t} \\ \vdots \\ u_{i,t} \\ \vdots \\ u_{k,t} \end{pmatrix} = \begin{pmatrix} \delta_{1,1} e_{1,t} \\ \delta_{2,1} e_{1,t} + \delta_{2,2} e_{2,t} \\ \vdots \\ \sum_{j=1}^i \delta_{i,j} e_{j,t} \\ \vdots \\ \sum_{j=1}^k \delta_{k,j} e_{j,t} \end{pmatrix}, \quad (30)$$

hence

$$u_{i,t} = \sum_{j=1}^i \delta_{i,j} e_{j,t} \quad (i = 1, \dots, k). \quad (31)$$

Rather than identifying the innovation response of a shock in variable 2 on the future values of X_t

by $E(X_{t+m}|u_{2,t})$, Sims considers the *net* effect of $u_{2,t}$ on X_{t+m} :

$$E(X_{t+m}|u_{1,t}, u_{2,t}) - E(X_{t+m}|u_{1,t}) = E(X_{t+m}|e_{1,t}, e_{2,t}) - E(X_{t+m}|e_{1,t}) \quad (32)$$

The equality follows from the fact

$$\begin{pmatrix} u_{1,t} \\ u_{2,t} \end{pmatrix} = \begin{pmatrix} \delta_{1,1} & 0 \\ \delta_{2,1} & \delta_{2,2} \end{pmatrix} \begin{pmatrix} e_{1,t} \\ e_{2,t} \end{pmatrix} \quad (33)$$

is a one-to-one mapping:

$$\begin{pmatrix} e_{1,t} \\ e_{2,t} \end{pmatrix} = \frac{1}{\delta_{1,1}\delta_{2,2}} \begin{pmatrix} \delta_{2,2} & 0 \\ -\delta_{2,1} & \delta_{2,2} \end{pmatrix} \begin{pmatrix} u_{1,t} \\ u_{2,t} \end{pmatrix}. \quad (34)$$

Since

$$E(X_{t+m}|e_{1,t}, e_{2,t}) = \Gamma_m(\delta_1 e_{1,t} + \delta_2 e_{2,t}), \quad E(X_{t+m}|e_{1,t}) = \Gamma_m \delta_1 e_{1,t} \quad (35)$$

with δ_1 and δ_2 the first two columns of Δ , the *net* effect of $u_{2,t}$ on X_{t+m} is:

$$E(X_{t+m}|u_{1,t}, u_{2,t}) - E(X_{t+m}|u_{1,t}) = \Gamma_m \delta_2 e_{2,t} \quad (36)$$

Again, Sims replaces $e_{2,t}$ by its standard error 1, so that $\Gamma_m \delta_2$ is the innovation response of a unit shock in variable 2 on X_{t+m} .

In general we have for $i > 1$:

$$\begin{aligned} & E(X_{t+m}|u_{1,t}, \dots, u_{i,t}) - E(X_{t+m}|u_{1,t}, \dots, u_{i-1,t}) \\ &= E(X_{t+m}|e_{1,t}, \dots, e_{i,t}) - E(X_{t+m}|e_{1,t}, \dots, e_{i-1,t}) = \Gamma_m \delta_i e_{it} \end{aligned} \quad (37)$$

with δ_i the i -th column of Δ . Again, Sims replaces $e_{i,t}$ by its standard error 1. Thus the innovation response of a unit shock in the i -th component of X_t on X_{t+m} is $\Gamma_m \delta_i$.

Summarizing:

DEFINITION 2: Denoting $R_m = (r_{ij}(m)) = \Gamma_m \Delta$, the innovation response of a unit shock in variable j on variable i is given by $r_{ij}(m)$, $m = 0, 1, 2, \dots$, where $r_{ij}(m)$ is the element of R_m in the i -th row and j -th column.

The graphs in Sims' (1980) article are the plots of the functions $r_{ij}(m)$ for $m = 0, 1, 2, \dots$

Remark 1. Since the innovations e_t are the errors of the VAR model for $\Delta^{-1}X_t$, it is in general

not possible to measure the components of e_t in the unit of measurement of the corresponding components of X_t . For example, let X_t be bivariate, with component 1 measured in US dollars and component 2 measured in euros. Then the same applies to the corresponding VAR errors $u_{1,t}$ and $u_{2,t}$ in (33). Hence it follows from (34) that the unit of measurement of $e_{1,t}$ is the same as for $u_{1,t}$ (US \$), but the unit of measurement of $e_{2,t}$ is a combination of US dollars and euros. On the other hand, the units of measurement of the innovation responses are the same as for the variables involved.

Note that the matrix Δ is not unique. It depends on the *order* of the components of X_t . Therefore, in conducting innovation response analysis one has to determine the order of the shocks to the system in advance.

The matrixes C_1, \dots, C_p and the vector η of intercepts can be estimated consistently by OLS. Let $\hat{C}_1, \dots, \hat{C}_p, \hat{\eta}$ be these OLS estimators and let

$$\hat{U}_t = X_t - \hat{\eta} - \sum_{i=1}^p \hat{C}_i X_{t-i} \quad (t = p+1, \dots, n). \quad (38)$$

be the vectors of OLS residuals. Then

$$\hat{\Sigma} = \frac{1}{n-p} \sum_{t=p+1}^n \hat{U}_t \hat{U}_t^T \quad (39)$$

is a consistent estimator of Σ (the variance matrix of U_t). Under the normality assumption, these estimators are just the maximum likelihood estimators.

Given the order of the variables in X_t we can calculate the lower triangular matrix $\hat{\Delta}$ such that $\hat{\Sigma} = \hat{\Delta} \hat{\Delta}^T$. Thus the main problem is how to estimate the matrices Γ_m for $m = 0, 1, 2, \dots$. This can be done as follows.

First, note that $\Gamma_0 = I$. Next, observe that the matrices Γ_m can be obtained from backwards substitution of the recursive relation $\Gamma_m = C_1 \Gamma_{m-1} + \dots + C_p \Gamma_{m-p} + E_m$, where $\Gamma_j = O$ for $j < 0$, $\Gamma_0 = I$, $E_j = O$ for $j \neq 0$, and $E_0 = I$. Replacing C_j by its OLS estimate \hat{C}_j , we then get consistent estimates $\hat{\Gamma}_m$ of Γ_m . Denoting $\hat{R}_m = (\hat{r}_{i,j}(m)) = \hat{\Gamma}_m \hat{\Delta}$, the estimated innovation response of a unit shock in variable j on variable i is $\hat{r}_{i,j}(m)$.

For each m , $\hat{r}_{i,j}(m)$ is a nonlinear but continuously differentiable function of the elements of the matrices \hat{C}_i , $i = 1, \dots, p$, and $\hat{\Sigma}$, and the vector $\hat{\beta}$ of these stacked elements has asymptotically a multivariate normal distribution:

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, \Omega) \text{ in distr.}, \quad (40)$$

where β is the vector of corresponding elements of the matrices C_i , $i = 1, \dots, p$, and Σ . Therefore, it follows by the so-called delta method that for each i , j , and m ,

$$\sqrt{n}(\hat{r}_{i,j}(m) - r_{i,j}(m)) \rightarrow N(0, \xi_{i,j,m}^T \Omega \xi_{i,j,m}) \text{ in distr.}, \quad (41)$$

where

$$\xi_{i,j,m} = \partial r_{i,j}(m) / \partial \beta^T. \quad (42)$$

Hence

$$\frac{\sqrt{n}(\hat{r}_{i,j}(m) - r_{i,j}(m))}{\sqrt{\hat{\xi}_{i,j,m}^T \hat{\Omega} \hat{\xi}_{i,j,m}}} \rightarrow N(0, 1) \text{ in distr.}, \quad (43)$$

where

$$\hat{\xi}_{i,j,m} = \partial \hat{r}_{i,j}(m) / \partial \hat{\beta}^T, \quad (44)$$

and $\hat{\Omega}$ is a consistent estimator of Ω . [Exercise: Assuming normality, how would you determine $\hat{\Omega}$?] On the basis of this result it is possible to endow each of the estimated innovation responses $\hat{r}_{i,j}(m)$ with asymptotic confidence intervals. See Baillie (1987).

5. Bernanke-Sims' structural VAR innovation response analysis

A disadvantage of Sims' (1980, 1982) nonstructural VAR approach is that the only way to incorporate economic theory into the VAR model is through the order of the variables in X_t . Therefore, Bernanke (1986) and Sims (1986) propose the following structural VAR(p) model:

$$BX_t = a_0 + \sum_{s=1}^p A_s X_{t-s} + e_t, \quad e_t \sim i.i.d. N_k(0, I). \quad (45)$$

The matrix B is the matrix of structural coefficients. This matrix represents the contemporaneous interaction between the variables in X_t , similar to the classical structural equations model.

Assuming that the matrix B is nonsingular, this structural model is related to the non-structural $\text{Var}(p)$ model (10):

$$X_t = B^{-1}a_0 + \sum_{s=1}^p B^{-1}A_s X_{t-s} + B^{-1}e_t = \eta + \sum_{s=1}^p C_s X_{t-s} + U_t. \quad (46)$$

Therefore, the matrix B links the nonstructural errors U_t to the innovations e_t :

$$BU_t = e_t, \text{ where } e_t \sim N_k(0, I). \quad (47)$$

Note that there are many matrices B for which this is true, for example let $B = \Sigma^{-1/2}$.

Given that B is specified such that B is invertible, (47) reads

$$U_t = B^{-1}e_t, \quad (48)$$

hence B^{-1} now takes over the role of the lower triangular matrix Δ in non-structural VAR innovation response analysis. Thus,

DEFINITION 3: Denoting $R_m^{(s)} = (r_{i,j}^{(s)}(m)) = \Gamma_m B^{-1}$, the (structural) innovation response of a unit shock in variable j on variable i is given by $r_{i,j}^{(s)}(m)$, $m = 0, 1, 2, \dots$, where $r_{i,j}^{(s)}(m)$ is the element of $R_m^{(s)}$ in the i -th row and j -th column.

Remark 2. Note that Remark 1 applies to the structural case as well.

Since

$$\Sigma = B^{-1}(B^T)^{-1} = (B^T B)^{-1}, \text{ hence } B^T B = \Sigma^{-1}, \quad (49)$$

and Σ contains only $k + (k^2 - k)/2$ different elements, we can identify no more than $k + (k^2 - k)/2$ elements of B . Therefore, we have to set at least $(k^2 - k)/2$ elements of B equal to given constants (usually zeros). This is where economic theory comes in the picture.

However, even if the matrix B contain $k + (k^2 - k)/2$ or less non-zero entries, identification is not guaranteed. For example, consider the case

$$B = \begin{pmatrix} b_1 & 0 & b_5 \\ 0 & b_2 & b_4 \\ b_6 & 0 & b_3 \end{pmatrix}. \quad (50)$$

Then

$$B^T B = \begin{pmatrix} b_1 & 0 & b_6 \\ 0 & b_2 & 0 \\ b_5 & b_4 & b_3 \end{pmatrix} \begin{pmatrix} b_1 & 0 & b_5 \\ 0 & b_2 & b_4 \\ b_6 & 0 & b_3 \end{pmatrix} = \begin{pmatrix} b_1^2 + b_6^2 & 0 & b_1 b_5 + b_3 b_6 \\ 0 & b_2^2 & b_2 b_4 \\ b_1 b_5 + b_3 b_6 & b_2 b_4 & b_3^2 + b_5^2 \end{pmatrix} = \Sigma^{-1}, \quad (51)$$

which effectively consists of five different nonlinear equations in six unknowns.

Given appropriate restrictions on B , we can estimate B by solving the nonlinear system of equations

$$\hat{B}^T \hat{B} = \hat{\Sigma}^{-1} \quad (52)$$

in the just-identified case, or by maximum likelihood in the over-identified case, where there are less unknown elements of B than $k + (k^2 - k)/2$. Replacing the lower triangular matrices Δ and $\hat{\Delta}$ in Sims' approach by B^{-1} and \hat{B}^{-1} , respectively, now yields the structural innovation responses.

REFERENCES

- Baillie, R. T. (1987): "Inference in Dynamic Models Containing 'Surprise' Variables", *Journal of Econometrics* 35, 101-117.
- Bernanke, B.S. (1986): "Alternative Explanations of the Money-Income Correlation", *Carnegie-Rochester Conference Series on Public Policy* 25, 49-100
- Sims, C.A. (1980): "Macroeconomics and Reality", *Econometrica* 48, 1-48
- Sims, C.A. (1982): "Policy Analysis with Econometric Models", *Brookings Papers on Economics Activity* 1, 107-152
- Sims, C.A. (1986): "Are Forecasting Models Usable for Policy Analysis?", *Federal Reserve Bank of Minneapolis Quarterly Review*, 1-16.

UNIT ROOTS

Herman J. Bierens¹

Pennsylvania State University

(October 30, 2007)

1. Introduction

In this chapter I will explain the two most frequently applied types of unit root tests, namely the Augmented Dickey-Fuller tests [see Fuller (1996), Dickey and Fuller (1979, 1981)], and the Phillips-Perron tests [see Phillips (1987) and Phillips and Perron (1988)]. The statistics and econometrics levels required for understanding the material below are Hogg and Craig (1978) or a similar level for statistics, and Green (1997) or a similar level for econometrics. The functional central limit theorem [see Billingsley (1968)], which plays a key-role in the derivations involved, will be explained in this chapter by showing its analogy with the concept of convergence in distribution of random variables, and by confining the discussion to Gaussian unit root processes.

This chapter is not a review of the vast literature on unit roots. Such a review would entail a long list of descriptions of the many different recipes for unit root testing proposed in the literature, and would leave no space for motivation, let alone proofs. I have chosen for depth rather than breadth, by focusing on the most influential papers on unit root testing, and discussing them in detail, without assuming that the reader has any previous knowledge about this topic.

As an introduction of the concept of a unit root and its consequences, consider the Gaussian AR(1) process $y_t = \beta_0 + \beta_1 y_{t-1} + u_t$, or equivalently $(1 - \beta_1 L)y_t = \beta_0 + u_t$, where L is the lag operator: $Ly_t = y_{t-1}$, and the u_t 's are i.i.d. $N(0, \sigma^2)$. The lag polynomial $1 - \beta_1 L$ has root equal to $1/\beta_1$. If $|\beta_1| < 1$, then by backwards substitution we can write $y_t = \beta_0/(1-\beta_1) + \sum_{j=0}^{\infty} \beta_1^j u_{t-j}$, so that y_t is strictly stationary, i.e., for arbitrary natural numbers $m_1 < m_2 < \dots < m_{k-1}$ the joint distribution of $y_t, y_{t-m_1}, y_{t-m_2}, \dots, y_{t-m_{k-1}}$ does not depend on t , but only on the lags or leads m_1, m_2, \dots, m_{k-1} . Moreover, the distribution of y_t , $t > 0$, conditional on $y_0, y_{-1}, y_{-2}, \dots$, then converges to the marginal distribution

¹ This is a slightly revised version of a chapter in Badi Baltagi (Ed.), *Companion in Theoretical Econometrics*, Blackwell Publishers. The useful comments of three referees are gratefully acknowledged.

of y_t if $t \rightarrow \infty$. In other words, y_t has a vanishing memory: y_t becomes independent of its past, $y_0, y_{-1}, y_{-2}, \dots$, if $t \rightarrow \infty$.

If $\beta_1 = 1$, so that the lag polynomial $1 - \beta_1 L$ has a unit root, then y_t is called a unit root process. In this case the AR(1) process under review becomes $y_t = y_{t-1} + \beta_0 + u_t$, which by backwards substitution yields for $t > 0$, $y_t = y_0 + \beta_0 t + \sum_{j=1}^t u_j$. Thus now the distribution of y_t , $t > 0$, conditional on $y_0, y_{-1}, y_{-2}, \dots$, is $N(y_0 + \beta_0 t, \sigma^2 t)$, so that y_t has no longer a vanishing memory: a shock in y_0 will have a persistent effect on y_t . The former intercept β_0 now becomes the *drift* parameter of the unit root process involved.

It is important to distinguish stationary processes from unit root processes, for the following reasons:

1. Regressions involving unit root processes may give spurious results. If y_t and x_t are mutually independent unit root processes, i.e. y_t is independent of x_{t-j} for all t and j , then the OLS regression of y_t on x_t for $t = 1, \dots, n$, with or without an intercept, will yield a significant estimate of the slope parameter if n is large: the absolute value of the t-value of the slope converges in probability to ∞ if $n \rightarrow \infty$. We then might conclude that y_t depends on x_t , while in reality the y_t 's are independent of the x_t 's. This phenomenon is called *spurious regression*.² One should therefore be very cautious when conducting standard econometric analysis using time series. If the time series involved are unit root processes, naive application of regression analysis may yield nonsense results.

2. For two or more unit root processes there may exist linear combinations which are stationary, and these linear combinations may be interpreted as long-run relationships. This phenomenon is called *cointegration*³, and plays a dominant role in modern empirical macroeconomic research.

² See the chapter on spurious regression in Badi Baltagi (Ed.), *Companion in Theoretical Econometrics*, Blackwell Publishers. This phenomenon can easily be demonstrated by using my free software package *EasyReg*, which is downloadable from website <http://econ.la.psu.edu/~hbierens/EASYREG.HTM> (Click on "Tools", and then on "Teaching tools").

³ See the chapter on cointegration in Badi Baltagi (Ed.), *Companion in Theoretical Econometrics*, Blackwell Publishers..

3. Tests of parameter restrictions in (auto)regressions involving unit root processes have in general different null distributions than in the case of stationary processes. In particular, if one would test the null hypothesis $\beta_1 = 1$ in the above AR(1) model using the usual t-test, the null distribution involved is non-normal. Therefore, naive application of classical inference may give incorrectly results. We will demonstrate the latter first, and in the process derive the Dickey-Fuller test [see Fuller (1996), Dickey and Fuller (1979, 1981)], by rewriting the AR(1) model as

$$\Delta y_t = y_t - y_{t-1} = \beta_0 + (\beta_1 - 1)y_{t-1} + u_t = \alpha_0 + \alpha_1 y_{t-1} + u_t, \quad (1)$$

say, estimating the parameter α_1 by OLS on the basis of observations y_0, y_1, \dots, y_n , and then testing the unit root hypothesis $\alpha_1 = 0$ against the stationarity hypothesis $-2 < \alpha_1 < 0$, using the t-value of α_1 . In Section 2 we consider the case where $\alpha_0 = 0$ under both the unit root hypothesis and the stationarity hypothesis. In Section 3 we consider the case where $\alpha_0 = 0$ under the unit root hypothesis but not under the stationarity hypothesis.

The assumption that the error process u_t is independent is quite unrealistic for macroeconomic time series. Therefore, in Sections 4 and 5 this assumption will be relaxed, and two types of appropriate unit root tests will be discussed: the Augmented Dickey-Fuller (ADF) tests, and the Phillips-Perron (PP) tests.

In Section 6 we consider the unit root *with* drift case, and we discuss the ADF and PP tests of the unit root with drift hypothesis, against the alternative of trend stationarity.

Finally, Section 7 contains some concluding remarks.

2. The Gaussian AR(1) case without intercept: Part 1

2.1 Introduction

Consider the AR(1) model without intercept, rewritten as⁴

$$\Delta y_t = \alpha_0 y_{t-1} + u_t, \text{ where } u_t \text{ is i.i.d. } N(0, \sigma^2), \quad (2)$$

and y_t is observed for $t = 1, 2, \dots, n$. For convenience I will assume that

⁴ The reason for changing the subscript of α from 1 in (1) to 0 is to indicate the number of other parameters at the right-hand side of the equation. See also (39).

$$y_t = 0 \text{ for } t \leq 0. \quad (3)$$

This assumption is, of course, quite unrealistic, but is made for the sake of transparency of the argument, and will appear to be innocent.

The OLS estimator of α_0 is:

$$\hat{\alpha}_0 = \frac{\sum_{t=1}^n y_{t-1} \Delta y_t}{\sum_{t=1}^n y_{t-1}^2} = \alpha_0 + \frac{\sum_{t=1}^n y_{t-1} u_t}{\sum_{t=1}^n y_{t-1}^2}. \quad (4)$$

If $-2 < \alpha_0 < 0$, so that y_t is stationary, then it is a standard exercise to verify that $\sqrt{n}(\hat{\alpha}_0 - \alpha_0) \rightarrow N(0, 1 - (1 + \alpha_0)^2)$ in distribution. On the other hand, if $\alpha_0 = 0$, so that y_t is a unit root process, this result reads: $\sqrt{n}\hat{\alpha}_0 \rightarrow N(0, 0)$ in distr., hence $\text{plim}_{n \rightarrow \infty} \sqrt{n}\hat{\alpha}_0 = 0$. However, we show now that a much stronger result holds, namely that $\hat{\rho}_0 \equiv n\hat{\alpha}_0$ converges in distribution, but the limiting distribution involved is non-normal. Thus, the presence of a unit root is actually advantageous for the efficiency of the OLS estimator $\hat{\alpha}_0$. The main problem is that the t-test of the null hypothesis that $\alpha_0 = 0$ has no longer a standard normal asymptotic null distribution, so that we cannot test for a unit root using standard methods. The same applies to more general unit root processes.

In the unit root case under review we have $y_t = y_{t-1} + u_t = y_0 + \sum_{j=1}^t u_j = \sum_{j=1}^t u_j$ for $t > 0$, where the last equality involved is due to assumption (3). Denoting

$$S_t = 0 \text{ for } t \leq 0, \quad S_t = \sum_{j=1}^t u_j \text{ for } t \geq 1. \quad (5)$$

and $\hat{\sigma}^2 = (1/n)\sum_{t=1}^n u_t^2$, it follows that

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n u_t y_{t-1} &= \frac{1}{2n} \sum_{t=1}^n \left((u_t + y_{t-1})^2 - y_{t-1}^2 - u_t^2 \right) = \frac{1}{2} \left(\frac{1}{n} \sum_{t=1}^n y_t^2 - \frac{1}{n} \sum_{t=1}^n y_{t-1}^2 - \frac{1}{n} \sum_{t=1}^n u_t^2 \right) \\ &= \frac{1}{2} \left(y_n^2/n - y_0^2/n - \hat{\sigma}^2 \right) = \frac{1}{2} (S_n^2/n - \hat{\sigma}^2), \end{aligned} \quad (6)$$

and similarly,

$$\frac{1}{n^2} \sum_{t=1}^n y_{t-1}^2 = \frac{1}{n} \sum_{t=1}^n (S_{t-1}/\sqrt{n})^2. \quad (7)$$

Next, let

$$W_n(x) = S_{[nx]} / (\sigma\sqrt{n}) \quad \text{for } x \in [0,1], \quad (8)$$

where $[z]$ means truncation to the nearest integer $\leq z$. Then we have⁵:

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n u_t y_{t-1} &= \frac{1}{2}(\sigma^2 W_n(1)^2 - \hat{\sigma}^2) \\ &= \frac{1}{2}(\sigma^2 W_n(1)^2 - \sigma^2 - O_p(1/\sqrt{n})) = \sigma^2 \frac{1}{2}(W_n(1)^2 - 1) + o_p(1), \end{aligned} \quad (9)$$

and

$$\frac{1}{n^2} \sum_{t=1}^n y_{t-1}^2 = \frac{1}{n} \sum_{t=1}^n \sigma^2 W_n((t-1)/n)^2 = \sigma^2 \int W_n(x)^2 dx, \quad (10)$$

where the integral in (10) and below, unless otherwise indicated, is taken over the unit interval $[0,1]$. The last equality in (9) follows from the law of large numbers, by which $\hat{\sigma}^2 = \sigma^2 + O_p(1/\sqrt{n})$. The last equality in (10) follows from the fact that for any power m ,

$$\begin{aligned} \int W_n(x)^m dx &= \int_0^1 W_n(x)^m dx = \frac{1}{n} \int_0^n W_n(z/n)^m dz = \frac{1}{n} \sum_{t=1}^n \int_{t-1}^t W_n(z/n)^m dz \\ &= \frac{1}{n^{1+m/2}} \sum_{t=1}^n \int_{t-1}^t (S_{[z]}/\sigma)^m dz = \frac{1}{n^{1+m/2}} \sum_{t=1}^n (S_{t-1}/\sigma)^m. \end{aligned} \quad (11)$$

Moreover, observe from (11), with $m = 1$, that $\int W_n(x) dx$ is a linear combination of i.i.d. standard normal random variables, and therefore normal itself, with zero mean and variance

$$E\left(\int W_n(x) dx\right)^2 = \iint E(W_n(x) W_n(y)) dx dy = \iint \frac{\min([nx], [ny])}{n} dx dy \rightarrow \iint \min(x, y) dx dy = \frac{1}{3}. \quad (12)$$

Thus, $\int W_n(x) dx \rightarrow N(0, 1/3)$ in distribution. Since $\int W_n(x)^2 dx \geq (\int W_n(x) dx)^2$, it follows therefore that $\int W_n(x)^2 dx$ is bounded away from zero:

⁵ Recall that the notation $o_p(a_n)$, with a_n a deterministic sequence, stands for a sequence of random variables or vectors x_n , say, such that $\text{plim}_{n \rightarrow \infty} x_n/a_n = 0$, and that the notation $O_p(a_n)$ stands for a sequence of random variables or vectors x_n such that x_n/a_n is stochastically bounded: $\forall \varepsilon \in (0,1) \exists M \in (0,\infty) : \sup_{n \geq 1} P(|x_n/a_n| > M) < \varepsilon$. Also, recall that convergence in distribution implies stochastic boundedness.

$$\left(\int W_n(x)^2 dx \right)^{-1} = O_p(1). \quad (13)$$

Combining (9), (10), and (13), we now have:

$$\hat{\rho}_0 \equiv n\hat{a}_0 = \frac{(1/n)\sum_{t=1}^n u_t y_{t-1}}{(1/n^2)\sum_{t=1}^n y_{t-1}^2} = \frac{(1/2)(W_n(1)^2 - 1) + o_p(1)}{\int W_n(x)^2 dx} = \frac{1}{2} \left(\frac{W_n(1)^2 - 1}{\int W_n(x)^2 dx} \right) + o_p(1). \quad (14)$$

This result does not depend on assumption (3).

2.2 Weak convergence of random functions

In order to establish the limiting distribution of (14), and other asymptotic results, we need to extend the well-known concept of convergence in distribution of random variables to convergence in distribution of a sequence of random functions. Recall that for random variables $X_n, X, X_n \rightarrow X$ in distribution if the distribution function $F_n(x)$ of X_n converges pointwise to the distribution function $F(x)$ of X in the continuity points of $F(x)$. Moreover, recall that distribution functions are uniquely associated to probability measures on the Borel sets⁶, i.e., there exists one and only one probability measure $\mu_n(B)$ on the Borel sets B such that $F_n(x) = \mu_n((-\infty, x])$, and similarly, $F(x)$ is uniquely associated to a probability measure μ on the Borel sets, such that $F(x) = \mu((-\infty, x])$. The statement $X_n \rightarrow X$ in distribution can now be expressed in terms of the probability measures μ_n and μ : $\mu_n(B) \rightarrow \mu(B)$ for all Borel sets B with boundary δB satisfying $\mu(\delta B) = 0$.

In order to extend the latter to random functions, we need to define Borel sets of functions. For our purpose it suffices to define Borel sets of continuous functions on $[0,1]$. Let $C[0,1]$ be the set of all continuous functions on the unit interval $[0,1]$. Define the distance between two functions f and g in $C[0,1]$ by the sup-norm: $\rho(f,g) = \sup_{0 \leq x \leq 1} |f(x) - g(x)|$. Endowed with this norm, the set

⁶ The Borel sets in \mathbb{R} are the members of the smallest σ -algebra containing the collection \mathfrak{C} , say, of all half-open intervals $(-\infty, x]$, $x \in \mathbb{R}$. Equivalently, we may also define the Borel sets as the members of the smallest σ -algebra containing the collection of open subsets of \mathbb{R} . A collection \mathcal{F} of subsets of a set Ω is called a σ -algebra if the following three conditions hold: $\Omega \in \mathcal{F}$; $A \in \mathcal{F}$ implies that its complement also belongs to \mathcal{F} : $\Omega \setminus A \in \mathcal{F}$ (hence, the empty set \emptyset belongs to \mathcal{F}); $A_n \in \mathcal{F}$, $n = 1, 2, 3, \dots$, implies $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$. The smallest σ -algebra containing a collection \mathfrak{C} of sets is the intersection of all σ -algebras containing the collection \mathfrak{C} .

$C[0,1]$ becomes a metric space, for which we can define open subsets, similarly to the concept of an open subset of \mathbb{R} : A set B in $C[0,1]$ is open if for each function f in B we can find an $\varepsilon > 0$ such that $\{g \in C[0,1] : \rho(g, f) < \varepsilon\} \subset B$. Now the smallest σ -algebra of subsets of $C[0,1]$ containing the collection of all open subsets of $C[0,1]$ is just the collection of Borel sets of functions in $C[0,1]$.

A random element of $C[0,1]$ is a random function $W(x)$, say, on $[0,1]$, which is continuous with probability 1. For such a random element W , say, we can define a probability measure μ on the Borel sets B in $C[0,1]$ by $\mu(B) = P(W \in B)$. Now a sequence W_n^* of random elements of $C[0,1]$, with corresponding probability measures μ_n , is said to converge weakly to a random element W of $C[0,1]$, with corresponding probability measure μ , if for each Borel set B in $C[0,1]$ with boundary δB satisfying $\mu(\delta B) = 0$, we have $\mu_n(B) \rightarrow \mu(B)$. This is usually denoted by: $W_n^* \Rightarrow W$ (on $[0,1]$). Thus, weak convergence is the extension to random functions of the concept of convergence in distribution.

In order to verify that $W_n^* \Rightarrow W$ on $[0,1]$, we have to verify two conditions. See Billingsley (1968). First, we have to verify that the finite distributions of W_n^* converge to the corresponding finite distributions of W , i.e., for arbitrary points x_1, \dots, x_m in $[0,1]$, $(W_n^*(x_1), \dots, W_n^*(x_m)) \Rightarrow (W(x_1), \dots, W(x_m))$ in distribution. Second, we have to verify that W_n^* is tight. Tightness is the extension of the concept of stochastic boundedness to random functions: for each ε in $(0,1)$ there exists a compact (Borel) set K in $C[0,1]$ such that $\mu_n(K) > 1 - \varepsilon$ for $n = 1, 2, \dots$. Since convergence in distribution implies stochastic boundedness, we cannot have convergence in distribution without stochastic boundedness, and the same applies to weak convergence: tightness is a necessary condition for weak convergence.

As is well-known, if $X_n \rightarrow X$ in distribution, and Φ is a continuous mapping from the support of X into a Euclidean space, then by Slutsky's theorem, $\Phi(X_n) \rightarrow \Phi(X)$ in distribution. A similar result holds for weak convergence, which is known as the continuous mapping theorem: If Φ is a continuous mapping from $C[0,1]$ into a Euclidean space, then $W_n^* \Rightarrow W$ implies $\Phi(W_n^*) \rightarrow \Phi(W)$ in distribution. For example, the integral $\Phi(f) = \int f(x)^2 dx$ with $f \in C[0,1]$ is a continuous mapping from $C[0,1]$ into the real line, hence $W_n^* \Rightarrow W$ implies that $\int W_n^*(x)^2 dx \rightarrow \int W(x)^2 dx$ in distribution.

The random function W_n defined by (8) is a step function on $[0,1]$, and therefore not a random element of $C[0,1]$. However, the steps involved can be smoothed by piecewise linear interpolation, yielding a random element W_n^* of $C[0,1]$ such that $\sup_{0 \leq x \leq 1} |W_n^*(x) - W_n(x)| = o_p(1)$. The finite

distributions of W_n^* are therefore asymptotically the same as the finite distributions of W_n . In order to analyze the latter, redefine W_n as

$$W_n(x) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nx]} e_t \text{ for } x \in [n^{-1}, 1], \quad W_n(x) = 0 \text{ for } x \in [0, n^{-1}), \quad e_t \text{ is i.i.d. } N(0, 1). \quad (15)$$

(Thus, $e_t = u_t/\sigma$, and let

$$\begin{aligned} W_n^*(x) &= W_n\left(\frac{t-1}{n}\right) + (nx - (t-1)) \left(W_n\left(\frac{t}{n}\right) - W_n\left(\frac{t-1}{n}\right) \right) \\ &= W_n(x) + (nx - (t-1)) \frac{e_t}{\sqrt{n}} \text{ for } x \in \left[\frac{t-1}{n}, \frac{t}{n}\right], \quad t = 1, \dots, n, \quad W_n^*(0) = 0. \end{aligned} \quad (16)$$

Then

$$\sup_{0 \leq x \leq 1} |W_n^*(x) - W_n(x)| \leq \frac{\max_{1 \leq t \leq n} |e_t|}{\sqrt{n}} = o_p(1). \quad (17)$$

The latter conclusion is not too hard an exercise.⁷

It is easy to verify that for *fixed* $0 \leq x < y \leq 1$ we have

$$\begin{aligned} \begin{pmatrix} W_n(x) \\ W_n(y) - W_n(x) \end{pmatrix} &= \frac{1}{\sqrt{n}} \begin{pmatrix} \sum_{t=1}^{[nx]} e_t \\ \sum_{t=[nx]+1}^{[ny]} e_t \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{[nx]}{n} & 0 \\ 0 & \frac{[ny]-[nx]}{n} \end{pmatrix} \right) \\ &\rightarrow \begin{pmatrix} W(x) \\ W(y)-W(x) \end{pmatrix} \text{ in distr.}, \end{aligned} \quad (18)$$

where $W(x)$ is a random function on $[0, 1]$ such that for $0 \leq x < y \leq 1$,

$$\begin{pmatrix} W(x) \\ W(y) - W(x) \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} x & 0 \\ 0 & y-x \end{pmatrix} \right). \quad (19)$$

⁷ Under the assumption that e_t is i.i.d. $N(0, 1)$,

$$P\left(\max_{1 \leq t \leq n} |e_t| \leq \epsilon \sqrt{n}\right) = \left(1 - 2 \int_{\epsilon \sqrt{n}}^{\infty} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx\right)^n \rightarrow 1$$

for arbitrary $\epsilon > 0$.

This random function $W(x)$ is called a standard Wiener process, or Brownian motion. Similarly, for arbitrary fixed x, y in $[0,1]$,

$$\begin{pmatrix} W_n(x) \\ W_n(y) \end{pmatrix} \rightarrow \begin{pmatrix} W(x) \\ W(y) \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} x & \min(x,y) \\ \min(x,y) & y \end{pmatrix}\right) \text{ in distr.} \quad (20)$$

and it follows from (17) that the same applies to W_n^* . Therefore, the finite distributions of W_n^* converge to the corresponding finite distributions of W . Also, it can be shown that W_n^* is tight [see Billingsley (1968)]. Hence, $W_n^* \Rightarrow W$, and by the continuous mapping theorem,

$$(W_n^*(1), \int W_n^*(x) dx, \int W_n^*(x)^2 dx, \int x W_n^*(x) dx)^T \rightarrow (W(1), \int W(x) dx, \int W(x)^2 dx, \int x W(x) dx)^T \quad (21)$$

in distr. This result, together with (17), implies that:

LEMMA 1. For W_n defined by (15), $(W_n(1), \int W_n(x) dx, \int W_n(x)^2 dx, \int x W_n(x) dx)^T$ converges jointly in distribution to $(W(1), \int W(x) dx, \int W(x)^2 dx, \int x W(x) dx)^T$.

2.3 Asymptotic null distributions

Using Lemma 1, it follows now straightforwardly from (14) that:

$$\hat{\rho}_0 \equiv n\hat{\alpha}_0 \rightarrow \rho_0 \equiv \frac{1}{2} \left(\frac{W(1)^2 - 1}{\int W(x)^2 dx} \right) \text{ in distr.} \quad (22)$$

The density⁸ of the distribution of ρ_0 is displayed in Figure 1, which clearly shows that the distribution involved is non-normal and asymmetric, with a fat left tail.

⁸ This density is actually a kernel estimate of the density of $\hat{\rho}_0$ on the basis of 10,000 replications of a Gaussian random walk $y_t = y_{t-1} + e_t$, $t = 0, 1, \dots, 1000$, $y_0 = 0$ for $t < 0$. The kernel involved is the standard normal density, and the bandwidth $h = c.s.10000^{-1/5}$, where s is the sample standard error, and $c = 1$. The scale factor c has been chosen by experimenting with various values. The value $c = 1$ is about the smallest one for which the kernel estimate remains a smooth curve; for smaller values of c the kernel estimate becomes wobbly. The densities of ρ_1 , τ_1 , ρ_2 , and τ_2 in Figures 2-6 have been constructed in the same way, with $c = 1$.

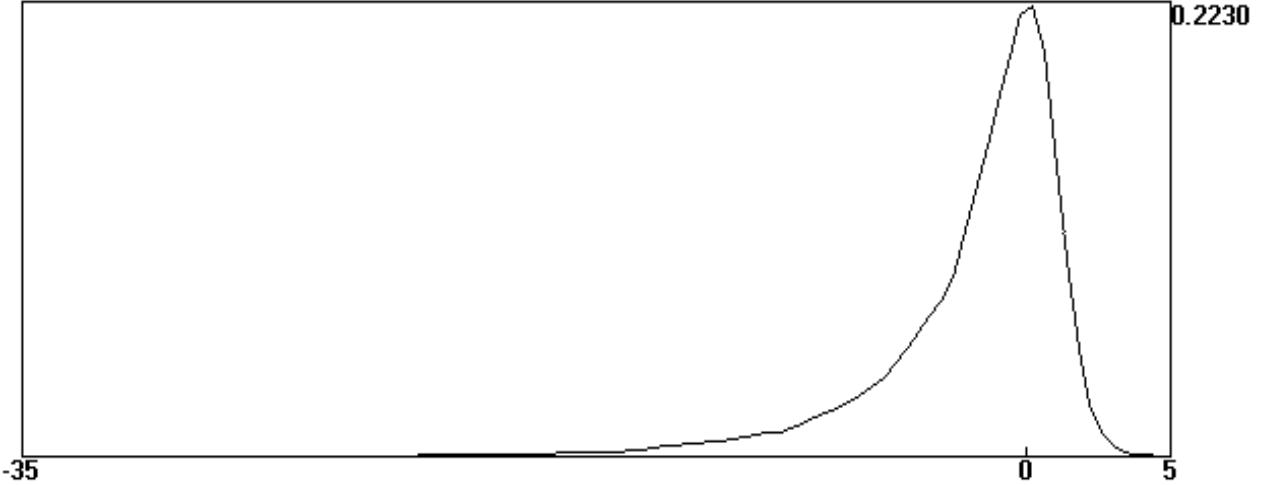


Figure 1: Density of ρ_0

Also the limiting distribution of the usual t-test statistic of the null hypothesis $\alpha_0 = 0$ is non-normal. First, observe that due to (10), (22), and Lemma 1, the residual sum of squares (RSS) of the regression (2) under the unit root hypothesis is:

$$RSS = \sum_{t=1}^n (\Delta y_t - \hat{\alpha}_0 y_{t-1})^2 = \sum_{t=1}^n u_t^2 - (n\hat{\alpha}_0)^2(1/n^2)\sum_{t=1}^n y_{t-1}^2 = \sum_{t=1}^n u_t^2 + O_p(1). \quad (23)$$

Hence $RSS/(n-1) = \sigma^2 + O_p(1/n)$. Therefore, similarly to (14) and (22), the Dickey-Fuller t-statistic $\hat{\tau}_0$ involved satisfies:

$$\hat{\tau}_0 \equiv n\hat{\alpha}_0 \frac{\sqrt{(1/n^2)\sum_{t=1}^n y_{t-1}^2}}{\sqrt{RSS/(n-1)}} = \frac{(W_n(1)^2 - 1)/2}{\sqrt{\int W_n(x)^2 dx}} + o_p(1) \rightarrow \tau_0 \equiv \frac{(W(1)^2 - 1)/2}{\sqrt{\int W(x)^2 dx}} \text{ in distr.} \quad (24)$$

Note that the unit root tests based on the statistics $\hat{\rho}_0 \equiv n\hat{\alpha}_0$ and $\hat{\tau}_0$ are left-sided: under the alternative of stationarity, $-2 < \alpha_0 < 0$, we have $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_0 = \alpha_0 < 0$, hence $\hat{\rho}_0 \rightarrow -\infty$ in probability at rate n , and $\hat{\tau}_0 \rightarrow -\infty$ in probability at rate \sqrt{n} .

The non-normality of the limiting distributions ρ_0 and τ_0 is no problem, though, as long one is aware of it. The distributions involved are free of nuisance parameters, and asymptotic critical values of the unit root tests $\hat{\rho}_0$ and $\hat{\tau}_0$ can easily be tabulated, using Monte Carlo simulation. In particular,

$$P(\tau_0 \leq -1.95) = 0.05, \quad P(\tau_0 \leq -1.62) = 0.10, \quad (25)$$

(see Fuller 1996, p. 642), whereas for a standard normal random variable e ,

$$P(e \leq -1.64) = 0.05, \quad P(e \leq -1.28) = 0.10 \quad (26)$$

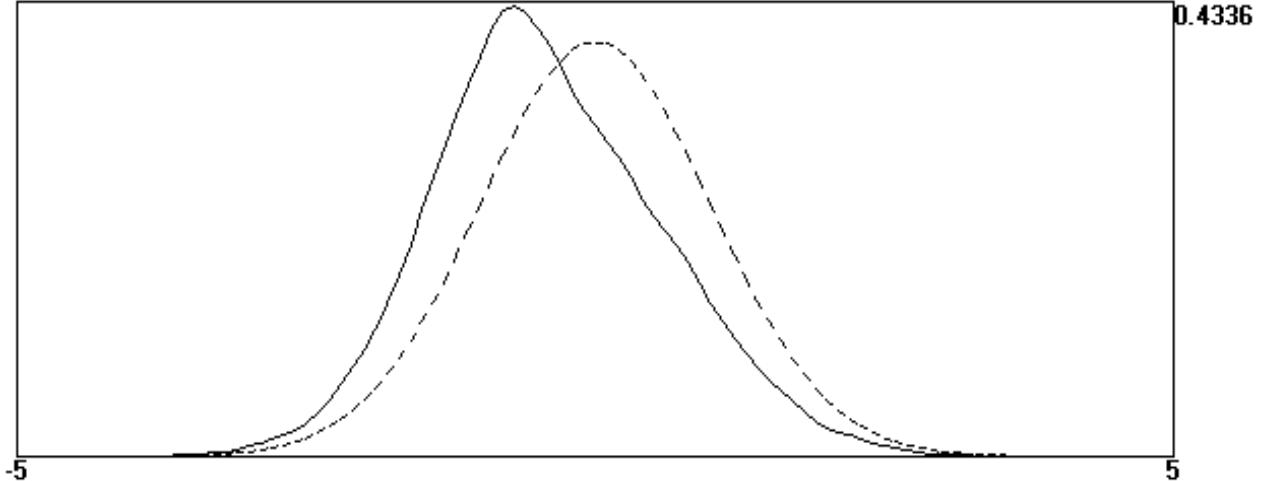


Figure 2: Density of τ_0 compared with the standard normal density (dashed curve)

In Figure 2 the density of τ_0 is compared with the standard normal density. We see that the density of τ_0 is shifted to left of the standard normal density, which causes the difference between (25) and (26). Using the left-sided standard normal test would result in a type 1 error of about twice the size: compare (26) with

$$P(\tau_0 \leq -1.64) \approx 0.09, \quad P(\tau_0 \leq -1.28) \approx 0.18 \quad (27)$$

3. The Gaussian AR(1) case with intercept under the alternative of stationarity

If under the stationarity hypothesis the AR(1) process has an intercept, but not under the unit root hypothesis, the AR(1) model that covers both the null and the alternative is:

$$\Delta y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t, \quad \text{where } \alpha_0 = -c\alpha_1. \quad (28)$$

If $-2 < \alpha_1 < 0$, then the process y_t is stationary around the constant c :

$$y_t = -c\alpha_1 + (1+\alpha_1)y_{t-1} + u_t = \sum_{j=0}^{\infty} (1+\alpha_1)^j (-c\alpha_1 + u_{t-j}) = c + \sum_{j=0}^{\infty} (1+\alpha_1)^j u_{t-j}, \quad (29)$$

hence $E(y_t^2) = c^2 + (1-(1+\alpha_1)^2)^{-1}\sigma^2$, $E(y_t y_{t-1}) = c^2 + (1+\alpha_1)(1-(1+\alpha_1)^2)^{-1}\sigma^2$, and the OLS estimator (4) of α_0 in model (2) satisfies

$$\operatorname{plim}_{n \rightarrow \infty} \hat{\alpha}_0 = \frac{E(y_t y_{t-1})}{E(y_{t-1}^2)} - 1 = \frac{\alpha_1}{1 + (c/\sigma)^2(1-(1+\alpha_1)^2)}, \quad (30)$$

which approaches zero if $c^2/\sigma^2 \rightarrow \infty$. Therefore, the power of the test $\hat{\rho}_0$ will be low if the variance of u_t is small relative to $[E(y_t)]^2$. The same applies to the t-test $\hat{\tau}_0$. We should therefore use the OLS estimator of α_1 and the corresponding t-value in the regression of Δy_t on y_{t-1} with intercept.

Denoting $\bar{y}_{-1} = (1/n)\sum_{t=1}^n y_{t-1}$, $\bar{u} = (1/n)\sum_{t=1}^n u_t$, the OLS estimator of α_1 is:

$$\hat{\alpha}_1 = \alpha_1 + \frac{\sum_{t=1}^n u_t y_{t-1} - n\bar{u}\bar{y}_{-1}}{\sum_{t=1}^n y_{t-1}^2 - n\bar{y}_{-1}^2}. \quad (31)$$

Since by (8), $\sqrt{n}\bar{u} = \sigma W_n(1)$, and under the null hypothesis $\alpha_1 = 0$ and the maintained hypothesis (3),

$$\bar{y}_{-1}/\sqrt{n} = \frac{1}{n\sqrt{n}} \sum_{t=1}^n S_{t-1} = \sigma \int W_n(x) dx, \quad (32)$$

where the last equality follows from (11) with $m = 1$, it follows from Lemma 1, similarly to (14) and (22) that

$$\begin{aligned} \hat{\rho}_1 &\equiv n\hat{\alpha}_1 = \frac{(1/2)(W_n(1)^2 - 1) - W_n(1) \int W_n(x) dx}{\int W_n(x)^2 dx - (\int W_n(x) dx)^2} + o_p(1) \\ &\rightarrow \rho_1 \equiv \frac{(1/2)(W(1)^2 - 1) - W(1) \int W(x) dx}{\int W(x)^2 dx - (\int W(x) dx)^2} \text{ in distr.} \end{aligned} \quad (33)$$

The density of ρ_1 is displayed in Figure 3. Comparing Figures 1 and 3, we see that the density of ρ_1 is farther left of zero than the density of ρ_0 , and has a fatter left tail.

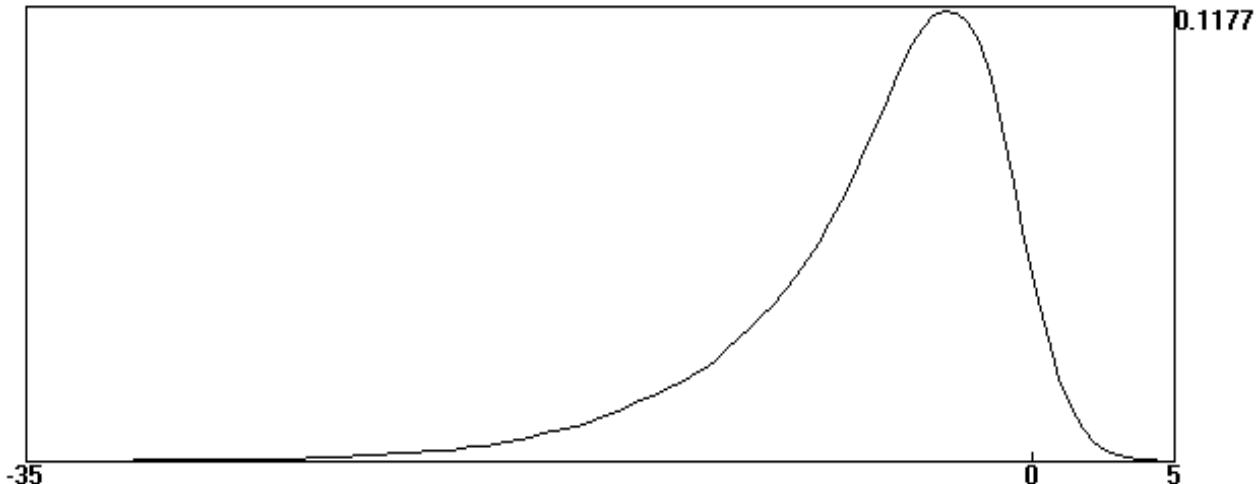


Figure 3: Density of ρ_1

As to the t-value $\hat{\tau}_1$ of α_1 in this case, it follows similarly to (24) and (33) that under the unit root hypothesis,

$$\hat{\tau}_1 \rightarrow \tau_1 \equiv \frac{(1/2)(W(1)^2 - 1) - W(1) \int W(x)dx}{\sqrt{\int W(x)^2 dx - (\int W(x)dx)^2}} \text{ in distr.} \quad (34)$$

Again, the results (33) and (34) do not hinge on assumption (3).

The distribution of τ_1 is even farther away from the normal distribution than the distribution of τ_0 , as follows from comparison of (26) with

$$P(\tau_1 \leq -2.86) = 0.05, \quad P(\tau_1 \leq -2.57) = 0.1 \quad (35)$$

See again Fuller (1996, p. 642). This is corroborated by Figure 4, where the density of τ_1 is compared with the standard normal density.

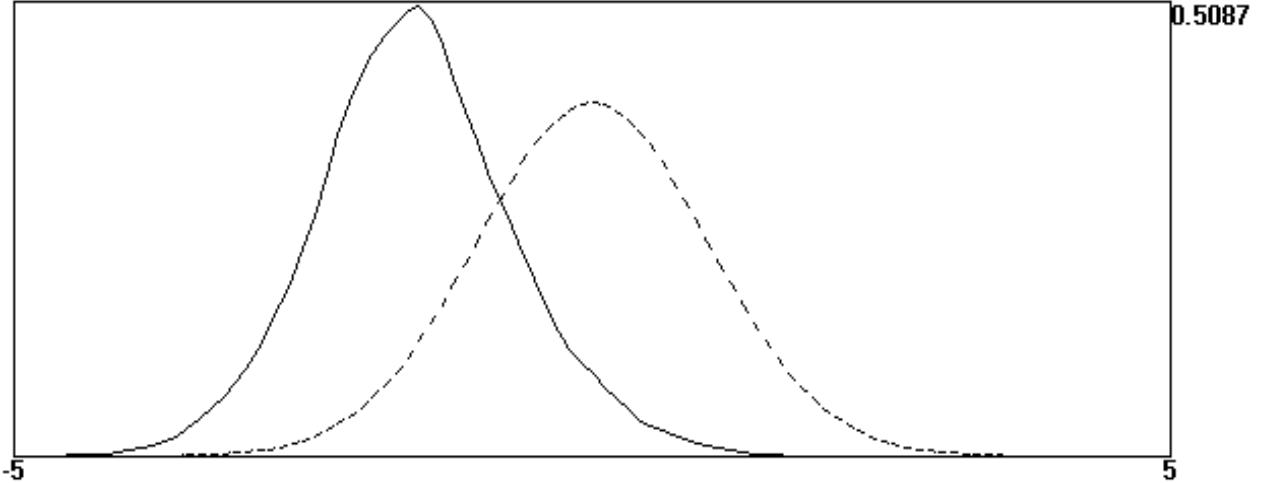


Figure 4: Density of τ_1 compared with the standard normal density (dashed curve)

We see that the density of τ_1 is shifted even more to the left of the standard normal density than in Figure 2, hence the left-sided standard normal test would result in a dramatically higher type 1 error than in the case without an intercept: compare

$$P(\tau_1 \leq -1.64) \approx 0.46, \quad P(\tau_1 \leq -1.28) \approx 0.64 \quad (36)$$

with (26) and (27).

4. General AR processes with a unit root, and the Augmented Dickey-Fuller test

The assumption made in Sections 2 and 3 that the data-generating process is an AR(1) process, is not very realistic for macroeconomic time series, because even after differencing most of these time series will still display a fair amount of dependence. Therefore we now consider an AR(p) process:

$$y_t = \beta_0 + \sum_{j=1}^p \beta_j y_{t-j} + u_t, \quad u_t \sim i.i.d. N(0, \sigma^2) \quad (37)$$

By recursively replacing y_{t-j} by $\Delta y_{t-j} + y_{t-p+j}$ for $j = 0, 1, \dots, p-1$, this model can be written as

$$\Delta y_t = \alpha_0 + \sum_{j=1}^{p-1} \alpha_j \Delta y_{t-j} + \alpha_p y_{t-p} + u_t, \quad u_t \sim i.i.d. N(0, \sigma^2), \quad (38)$$

where $\alpha_0 = \beta_0$, $\alpha_j = \sum_{i=1}^j \beta_i - 1$, $j = 1, \dots, p$. Alternatively and equivalently, by recursively replacing y_{t-p+j} by $y_{t-p+j+1} - \Delta y_{t-p+j+1}$ for $j = 0, 1, \dots, p-1$, model (37) can also be written as

$$\Delta y_t = \alpha_0 + \sum_{j=1}^{p-1} \alpha_j \Delta y_{t-j} + \alpha_p y_{t-1} + u_t, \quad u_t \sim i.i.d. N(0, \sigma^2), \quad (39)$$

where now $\alpha_j = -\sum_{i=1}^j \beta_i$, $j = 1, \dots, p-1$, $\alpha_p = \sum_{i=1}^p \beta_i - 1$.

If the AP(p) process (37) has a unit root, then clearly $\alpha_p = 0$ in (38) and (39). If the process (37) is stationary, i.e., all the roots of the lag polynomial $1 - \sum_{j=1}^p \beta_j L^j$ lie outside the complex unit circle, then $\alpha_p = \sum_{j=1}^p \beta_j - 1 < 0$ in (38) and (39).⁹ The unit root hypothesis can therefore be tested by testing the null hypothesis $\alpha_p = 0$ against the alternative hypothesis $\alpha_p < 0$, using the t-value \hat{t}_p of α_p in model (38) or model (39). This test is known as the Augmented Dickey-Fuller (ADF) tests.

We will show now for the case $p = 2$, with intercept under the alternative, i.e.,

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \alpha_2 y_{t-1} + u_t, \quad u_t \sim i.i.d. N(0, \sigma^2), \quad t = 1, \dots, n. \quad (40)$$

that under the unit root (without drift¹⁰) hypothesis the limiting distribution of $n\hat{\alpha}_p$ is proportional to the limiting distribution in (33), and the limiting distribution of \hat{t}_p is the same as in (34).

Under the unit root hypothesis, i.e., $\alpha_0 = \alpha_2 = 0$, $|\alpha_1| < 1$, we have

$$\begin{aligned} \Delta y_t &= \alpha_1 \Delta y_{t-1} + u_t = (1 - \alpha_1 L)^{-1} u_t = (1 - \alpha_1)^{-1} u_t + [(1 - \alpha_1 L)^{-1} - (1 - \alpha_1)^{-1}] u_t \\ &= (1 - \alpha_1)^{-1} u_t - \alpha_1 (1 - \alpha_1)^{-1} (1 - \alpha_1 L)^{-1} (1 - L) u_t = (1 - \alpha_1)^{-1} u_t + v_t - v_{t-1}, \end{aligned} \quad (41)$$

say, where $v_t = -\alpha_1 (1 - \alpha_1)^{-1} (1 - \alpha_1 L)^{-1} u_t = -\alpha_1 (1 - \alpha_1)^{-1} \sum_{j=0}^{\infty} \alpha_1^j u_{t-j}$ is a stationary process. Hence:

$$\begin{aligned} y_t / \sqrt{n} &= y_0 / \sqrt{n} + v_t / \sqrt{n} - v_0 / \sqrt{n} + (1 - \alpha_1)^{-1} (1 / \sqrt{n}) \sum_{j=1}^t u_j \\ &= y_0 / \sqrt{n} + v_t / \sqrt{n} - v_0 / \sqrt{n} + \sigma (1 - \alpha_1)^{-1} W_n(t/n) \end{aligned} \quad (42)$$

and therefore, similarly to (6), (7), and (32), it follows that

⁹ To see this, write $1 - \sum_{j=1}^p \beta_j L^j = \prod_{j=1}^p (1 - \rho_j L)$, so that $1 - \sum_{j=1}^p \beta_j = \prod_{j=1}^p (1 - \rho_j)$, where the $1/\rho_j$'s are the roots of the lag polynomial involved. If root $1/\rho_j$ is real valued, then the stationarity condition implies $-1 < \rho_j < 1$, so that $1 - \rho_j > 0$. If some roots are complex-valued, then these roots come in complex-conjugate pairs, say $1/\rho_1 = a+i.b$ and $1/\rho_2 = a-i.b$, hence $(1 - \rho_1)(1 - \rho_2) = (1/\rho_1 - 1)(1/\rho_2 - 1)\rho_1\rho_2 = ((a-1)^2 + b^2)/(a^2 + b^2) > 0$.

¹⁰ In the sequel we shall suppress the statement "without drift". A unit root process is from now on by default a unit root without drift process, except if otherwise indicated.

$$(1/n) \sum_{t=1}^n y_{t-1} / \sqrt{n} = \sigma(1-\alpha_1)^{-1} \int W_n(x) dx + o_p(1), \quad (43)$$

$$(1/n^2) \sum_{t=1}^n y_{t-1}^2 = \sigma^2(1-\alpha_1)^{-2} \int W_n(x)^2 dx + o_p(1), \quad (44)$$

$$\begin{aligned} (1/n) \sum_{t=1}^n u_t y_{t-1} &= (1/n) \sum_{t=1}^n u_t \left((1-\alpha_1)^{-1} \sum_{j=1}^{t-1} u_j + y_0 + v_{t-1} - v_0 \right) \\ &= (1-\alpha_1)^{-1} (1/n) \sum_{t=1}^n u_t \sum_{j=1}^{t-1} u_j + (y_0 - v_0) (1/n) \sum_{t=1}^n u_t + (1/n) \sum_{t=1}^n u_t v_{t-1} \\ &= \frac{(1-\alpha_1)^{-1} \sigma^2}{2} \left(W_n(1)^2 - 1 \right) + o_p(1) \end{aligned} \quad (45)$$

Moreover,

$$\operatorname{plim}_{n \rightarrow \infty} (1/n) \sum_{t=1}^n \Delta y_{t-1} = E(\Delta y_t) = 0, \quad \operatorname{plim}_{n \rightarrow \infty} (1/n) \sum_{t=1}^n (\Delta y_{t-1})^2 = E(\Delta y_t)^2 = \sigma^2/(1-\alpha_1^2) \quad (46)$$

and

$$\begin{aligned} (1/n) \sum_{t=1}^n y_{t-1} \Delta y_{t-1} &= (1/n) \sum_{t=1}^n (\Delta y_{t-1})^2 + (1/n) \sum_{t=1}^n y_{t-2} \Delta y_{t-1} \\ &= (1/n) \sum_{t=1}^n (\Delta y_{t-1})^2 + \frac{1}{2} \left((1/n) \sum_{t=1}^n y_{t-1}^2 - (1/n) \sum_{t=1}^n y_{t-2}^2 - (1/n) \sum_{t=1}^n (\Delta y_{t-1})^2 \right) \\ &= \frac{1}{2} \left((1/n) \sum_{t=1}^n (\Delta y_{t-1})^2 + y_{n-1}^2/n - y_{-1}^2/n \right) = \frac{1}{2} \left(\sigma^2/(1-\alpha_1^2) + \sigma^2(1-\alpha_1)^{-2} W_n(1)^2 \right) + o_p(1) \end{aligned} \quad (47)$$

hence

$$(1/n) \sum_{t=1}^n y_{t-1} \Delta y_{t-1} / \sqrt{n} = O_p(1/\sqrt{n}). \quad (48)$$

Next, let $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)^T$ be the OLS estimator of $\alpha = (\alpha_0, \alpha_1, \alpha_2)^T$. Under the unit root hypothesis we have

$$\begin{pmatrix} \sqrt{n} \hat{\alpha}_0 \\ \sqrt{n} (\hat{\alpha}_1 - \alpha_1) \\ n \hat{\alpha}_2 \end{pmatrix} = \sqrt{n} D_n \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xu} = \left(D_n^{-1} \hat{\Sigma}_{xx} D_n^{-1} \right)^{-1} \sqrt{n} D_n^{-1} \hat{\Sigma}_{xu}, \quad (49)$$

where

$$D_n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sqrt{n} \end{pmatrix}, \quad (50)$$

$$\hat{\Sigma}_{xx} = \begin{pmatrix} 1 & (1/n)\sum_{t=1}^n \Delta y_{t-1} & (1/n)\sum_{t=1}^n y_{t-1} \\ (1/n)\sum_{t=1}^n \Delta y_{t-1} & (1/n)\sum_{t=1}^n (\Delta y_{t-1})^2 & (1/n)\sum_{t=1}^n y_{t-1} \Delta y_{t-1} \\ (1/n)\sum_{t=1}^n y_{t-1} & (1/n)\sum_{t=1}^n y_{t-1} \Delta y_{t-1} & (1/n)\sum_{t=1}^n y_{t-1}^2 \end{pmatrix}, \quad (51)$$

and

$$\hat{\Sigma}_{xu} = \begin{pmatrix} (1/n)\sum_{t=1}^n u_t \\ (1/n)\sum_{t=1}^n u_t \Delta y_{t-1} \\ (1/n)\sum_{t=1}^n u_t y_{t-1} \end{pmatrix}. \quad (52)$$

It follows from (43) through (48) that

$$D_n^{-1} \hat{\Sigma}_{xx} D_n^{-1} = \begin{pmatrix} 1 & 0 & \sigma(1-\alpha_1)^{-1} \int W_n(x) dx \\ 0 & \sigma^2/(1-\alpha_1^2) & 0 \\ \sigma(1-\alpha_1)^{-1} \int W_n(x) dx & 0 & \sigma^2(1-\alpha_1)^{-2} \int W_n(x)^2 dx \end{pmatrix} + o_p(1), \quad (53)$$

hence, using the easy equality

$$\begin{pmatrix} 1 & 0 & a \\ 0 & b & 0 \\ a & 0 & c \end{pmatrix}^{-1} = \frac{1}{c-a^2} \begin{pmatrix} c & 0 & -a \\ 0 & b^{-1}(c-a^2) & 0 \\ -a & 0 & 1 \end{pmatrix},$$

it follows that

$$\begin{aligned} & \left(D_n^{-1} \hat{\Sigma}_{xx} D_n^{-1} \right)^{-1} = \frac{\sigma^{-2}(1-\alpha_1)^2}{\int W_n(x)^2 dx - (\int W_n(x) dx)^2} \\ & \times \begin{pmatrix} \sigma^2(1-\alpha_1)^{-2} \int W_n(x)^2 dx & 0 & -\sigma(1-\alpha_1)^{-1} \int W_n(x) dx \\ 0 & \frac{\int W_n(x)^2 dx - (\int W_n(x) dx)}{(1-\alpha_1^2)(1-\alpha_1)^2} & 0 \\ -\sigma(1-\alpha_1)^{-1} \int W_n(x) dx & 0 & 1 \end{pmatrix} + o_p(1). \end{aligned} \quad (54)$$

Moreover, it follows from (8) and (45) that

$$\sqrt{n} D_n^{-1} \hat{\Sigma}_{xu} = \begin{pmatrix} \sigma W_n(1) \\ (1/\sqrt{n}) \sum_{t=1}^n u_t \Delta y_{t-1} \\ \sigma^2(1-\alpha_1)^{-2} (W_n(1)^2 - 1)/2 \end{pmatrix} + o_p(1). \quad (55)$$

Combining (49), (54) and (55), and using Lemma 1, it follows now easily that

$$\frac{n \hat{\alpha}_2}{1-\alpha_1} = \frac{\frac{1}{2} (W_n(1)^2 - 1) - W_n(1) \int W_n(x) dx}{\int W_n(x)^2 dx - (\int W_n(x) dx)^2} + o_p(1) \rightarrow \rho_1 \text{ in distr.}, \quad (56)$$

where ρ_1 is defined in (33). Along the same lines it can be shown:

THEOREM 1. Let y_t be generated by (39), and let $\hat{\alpha}_p$ be the OLS estimator of α_p . Under the unit root hypothesis, i.e., $\alpha_p = 0$ and $\alpha_0 = 0$, the following hold: If model (39) is estimated without intercept, then $n \hat{\alpha}_p \rightarrow (1 - \sum_{j=1}^{p-1} \alpha_j) \rho_0$ in distr., where ρ_0 is defined in (22). If model (39) is estimated with intercept, then $n \hat{\alpha}_p \rightarrow (1 - \sum_{j=1}^{p-1} \alpha_j) \rho_1$ in distr., where ρ_1 is defined in (33). Moreover, under the stationarity hypothesis, $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_p = \alpha_p < 0$, hence $\text{plim}_{n \rightarrow \infty} n \hat{\alpha}_p = -\infty$, provided that in the case where the model is estimated without intercept this intercept, α_0 , is indeed zero.

Due to the factor $1 - \sum_{j=1}^{p-1} \alpha_j$ in the limiting distribution of $n \hat{\alpha}_p$ under the unit root

hypothesis, we cannot use $n\hat{\alpha}_p$ directly as a unit root test. However, it can be shown that under the unit root hypothesis this factor can be consistently estimated by $1 - \sum_{j=1}^{p-1} \hat{\alpha}_j$, hence we can use $n\hat{\alpha}_p / |1 - \sum_{j=1}^{p-1} \hat{\alpha}_j|$ as a unit root test statistic, with limiting distribution given by (22) or (33). The reason for the absolute value is that under the alternative of stationarity the probability limit of $1 - \sum_{j=1}^{p-1} \hat{\alpha}_j$ may be negative¹¹.

The actual ADF test is based on the t-value of α_p , because the factor $1 - \sum_{j=1}^{p-1} \alpha_j$ will cancel out in the limiting distribution involved. We will show this for the AR(2) case.

First, it is not too hard to verify from (43) through (48), and (54), that the residual sum of squares RSS of the regression (40) satisfies:

$$RSS = \sum_{t=1}^n u_t^2 + O_p(1). \quad (57)$$

This result carries over to the general AR(p) case, and also holds under the stationarity hypothesis. Moreover, under the unit root hypothesis it follows easily from (54) and (57) that the OLS standard error, s_2 , say, of $\hat{\alpha}_2$ in model (40) satisfies:

$$ns_2 = \sqrt{\frac{(RSS/(n-3))\sigma^{-2}(1-\alpha_1)^2}{\int W_n(x)^2 dx - (\int W_n(x)dx)^2}} + o_p(1) = \frac{1-\alpha_1}{\sqrt{\int W_n(x)^2 dx - (\int W_n(x)dx)^2}} + o_p(1), \quad (58)$$

hence it follows from (56) that the t-value \hat{t}_2 of $\hat{\alpha}_2$ in model (40) satisfies (34). Again, this result carries over to the general AR(p) case:

THEOREM 2. Let y_t be generated by (39), and let \hat{t}_p be t-value of the OLS estimator of α_p . Under the unit root hypothesis, i.e., $\alpha_p = 0$ and $\alpha_0 = 0$, the following hold: If model (39) is estimated without intercept, then $\hat{t}_p \rightarrow \tau_0$ in distr., where τ_0 is defined in (24). If model (39) is estimated with intercept, then $\hat{t}_p \rightarrow \tau_1$ in distr., where τ_1 is defined in (34). Moreover, under the stationarity

¹¹ For example, let $p = 2$ in (37) and (39). Then $\alpha_1 = -\beta_1$, hence if $\beta_1 < -1$ then $1 - \alpha_1 < 0$. In order to show that $\beta_1 < -1$ can be compatible with stationarity, assume that $\beta_1^2 = 4\beta_2$, so that the lag polynomial $1 - \beta_1 L - \beta_2 L^2$ has two common roots $-2/|\beta_1|$. Then the AR(2) process involved is stationary for $-2 < \beta_1 < -1$.

hypothesis, $\text{plim}_{n \rightarrow \infty} \hat{t}_p / \sqrt{n} < 0$, hence $\text{plim}_{n \rightarrow \infty} \hat{t}_p = -\infty$, provided that in the case where the model is estimated without intercept this intercept, α_0 , is indeed zero.

5. ARIMA processes, and the Phillips-Perron test

The ADF test requires that the order p of the AR model involved is finite, and correctly specified, i.e., the specified order should not be smaller than the actual order. In order to analyze what happens if p is misspecified, suppose that the actual data-generating process is given by (39) with $\alpha_0 = \alpha_2 = 0$ and $p > 1$, and that the unit root hypothesis is tested on the basis of the assumption that $p = 1$. Denoting $e_t = u_t / \sigma$, model (39) with $\alpha_0 = \alpha_2 = 0$ can be rewritten as

$$\Delta y_t = (\sum_{j=0}^{\infty} \gamma_j L^j) e_t = \gamma(L) e_t, \quad e_t \sim \text{i.i.d. } N(0,1), \quad (59)$$

where $\gamma(L) = \alpha(L)^{-1}$, with $\alpha(L) = 1 - \sum_{j=1}^{p-1} \alpha_j L^j$. This data-generating process can be nested in the auxiliary model

$$\Delta y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t, \quad u_t = \gamma(L) e_t, \quad e_t \sim \text{i.i.d. } N(0,1). \quad (60)$$

We will now determine the limiting distribution of the OLS estimate $\hat{\alpha}_1$ and corresponding t-value \hat{t}_1 of the parameter α_1 in the regression (60), derived under the assumption that the u_t 's are independent, while in reality (59) holds.

Similarly to (41) we can write $\Delta y_t = \gamma(1) e_t + v_t - v_{t-1}$, where $v_t = [(\gamma(L) - \gamma(1)) / (1 - L)] e_t$ is a stationary process. The latter follows from the fact that by construction the lag polynomial $\gamma(L) - \gamma(1)$ has a unit root, and therefore contains a factor $1 - L$. Next, redefining $W_n(x)$ as

$$W_n(x) = (1/\sqrt{n}) \sum_{t=1}^{[nx]} e_t \quad \text{if } x \in [n^{-1}, 1], \quad W_n(x) = 0 \quad \text{if } x \in [0, n^{-1}), \quad (61)$$

it follows similarly to (42) that

$$y_t / \sqrt{n} = y_0 / \sqrt{n} + v_t / \sqrt{n} - v_0 / \sqrt{n} + \gamma(1) W_n(t/n), \quad (62)$$

hence

$$y_n / \sqrt{n} = \gamma(1) W_n(1) + O_p(1/\sqrt{n}), \quad (63)$$

and similarly to (43) and (44) that

$$\bar{y}_{-1}/\sqrt{n} = \frac{1}{n} \sum_{t=1}^n y_{t-1}/\sqrt{n} = \gamma(1) \int W_n(x) dx + o_p(1), \quad (64)$$

and

$$\frac{1}{n^2} \sum_{t=1}^n y_{t-1}^2 = \gamma(1)^2 \int W_n(x)^2 dx + o_p(1). \quad (65)$$

Moreover, similarly to (6) we have

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n (\Delta y_t) y_{t-1} &= \frac{1}{2} \left(y_n^2/n - y_0^2/n - \frac{1}{n} \sum_{t=1}^n (\Delta y_t)^2 \right) \\ &= \frac{1}{2} \left(\gamma(1)^2 W_n(1)^2 - \frac{1}{n} \sum_{t=1}^n (\gamma(L)e_t)^2 \right) + o_p(1) = \gamma(1)^2 \frac{1}{2} (W_n(1) - \lambda) + o_p(1), \end{aligned} \quad (66)$$

where

$$\lambda = \frac{E(\gamma(L)e_t)^2}{\gamma(1)^2} = \frac{\sum_{j=0}^{\infty} \gamma_j^2}{(\sum_{j=0}^{\infty} \gamma_j)^2}. \quad (67)$$

Therefore, (33) now becomes:

$$n\hat{\alpha}_1 = \frac{(1/2)(W_n(1)^2 - \lambda) - W_n(1) \int W_n(x) dx}{\int W_n(x)^2 dx - (\int W_n(x) dx)^2} + o_p(1) \rightarrow \rho_1 + \frac{0.5(1-\lambda)}{\int W(x)^2 dx - (\int W(x) dx)^2} \quad (68)$$

in distr., and (34) becomes:

$$\hat{t}_1 = \frac{(1/2)(W_n(1)^2 - \lambda) - W_n(1) \int W_n(x) dx}{\sqrt{\int W_n(x)^2 dx - (\int W_n(x) dx)^2}} + o_p(1) \rightarrow \tau_1 + \frac{0.5(1-\lambda)}{\sqrt{\int W(x)^2 dx - (\int W(x) dx)^2}} \quad (69)$$

in distr. These results carry straightforwardly over to the case where the actual data-generating process is an ARIMA process $\alpha(L)\Delta y_t = \beta(L)e_t$, simply by redefining $\gamma(L) = \beta(L)/\alpha(L)$.

The parameter $\gamma(1)^2$ is known as the long-run variance of $u_t = \gamma(L)e_t$:

$$\sigma_L^2 = \lim_{n \rightarrow \infty} \text{var}\left[(1/\sqrt{n}) \sum_{t=1}^n u_t\right] = \gamma(1)^2 \quad (70)$$

which in general is different from the variance of u_t itself:

$$\sigma_u^2 = \text{var}(u_t) = E(u_t^2) = E(\sum_{j=0}^{\infty} \gamma_j e_{t-j})^2 = \sum_{j=0}^{\infty} \gamma_j^2. \quad (71)$$

If we would know σ_L^2 and σ_u^2 , and thus $\lambda = \sigma_u^2/\sigma_L^2$, then it follows from (64), (65), and Lemma 1, that

$$\frac{\sigma_L^2 - \sigma_u^2}{(1/n^2)\sum_{t=1}^n(y_{t-1} - \bar{y}_{-1})^2} \rightarrow \frac{1 - \lambda}{\int W(x)^2 dx - (\int W(x)dx)^2} \text{ in distr.} \quad (72)$$

It is an easy exercise to verify that this result also holds if we replace y_{t-1} by y_t and \bar{y}_{-1} by $\bar{y} = (1/n)\sum_{t=1}^n y_t$. Therefore it follows from (68) and (72) that,

THEOREM 3. (*Phillips-Perron test I*) *Under the unit root hypothesis, and given consistent estimators $\hat{\sigma}_L^2$ and $\hat{\sigma}_u^2$ of σ_L^2 and σ_u^2 , respectively, we have*

$$\hat{Z}_1 = n \left(\hat{\alpha}_1 - \frac{(\hat{\sigma}_L^2 - \hat{\sigma}_u^2)/2}{(1/n)\sum_{t=1}^n(y_t - \bar{y})^2} \right) \rightarrow \rho_1 \text{ in distr.} \quad (73)$$

This correction of (68) has been proposed by Phillips and Perron (1988) for particular estimators $\hat{\sigma}_L^2$ and $\hat{\sigma}_u^2$, following the approach of Phillips (1987) for the case where the intercept α_0 in (60) is assumed to be zero.

It is desirable to choose the estimators $\hat{\sigma}_L^2$ and $\hat{\sigma}_u^2$ such that under the stationarity alternative, $\text{plim}_{n \rightarrow \infty} \hat{Z}_1 = -\infty$. We show now that this is the case if we choose

$$\hat{\sigma}_u^2 = \frac{1}{n} \sum_{t=1}^n \hat{u}_t^2, \quad \text{where } \hat{u}_t = \Delta y_t - \hat{\alpha}_0 - \hat{\alpha}_1 y_{t-1}, \quad (74)$$

and $\hat{\sigma}_L^2$ such that $\bar{\sigma}_L^2 = \text{plim}_{n \rightarrow \infty} \hat{\sigma}_L^2 \geq 0$ under the alternative of stationarity.

First, it is easy to verify that $\hat{\sigma}_u^2$ is consistent under the null hypothesis, by verifying that (57) still holds. Under stationarity we have $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_1 = \text{cov}(y_t, y_{t-1})/\text{var}(y_t) - 1 = \alpha_1^*$, say, $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_0 = -\alpha_1^* E(y_t) = \alpha_0^*$, say, and $\text{plim}_{n \rightarrow \infty} \hat{\sigma}_u^2 = (1 - (\alpha_1^* + 1)^2)\text{var}(y_t) = \sigma_*^2$, say. Therefore,

$$\text{plim}_{n \rightarrow \infty} \hat{Z}_1/n = -0.5(\alpha_1^{*2} + \bar{\sigma}_L^2/\text{var}(y_t)) < 0. \quad (75)$$

Phillips and Perron (1988) propose to estimate the long-run variance by the Newey-West (1987) estimator

$$\hat{\sigma}_L^2 = \hat{\sigma}_u^2 + 2 \sum_{i=1}^m [1 - i/(m+1)](1/n) \sum_{t=i+1}^n \hat{u}_t \hat{u}_{t-i}, \quad (76)$$

where \hat{u}_t is defined in (74), and m converges to infinity with n at rate $o(n^{1/4})$. Andrews (1991) has shown (and we will show it again along the lines in Bierens (1994)) that the rate $o(n^{1/4})$ can be relaxed to $o(n^{1/2})$. The weights $1 - j/(m+1)$ guarantee that this estimator is always positive. The reason for the latter is the following. Let $u_t^* = u_t$ for $t = 1, \dots, n$, and $u_t^* = 0$ for $t < 1$ and $t > n$. Then,

$$\begin{aligned} \hat{\sigma}_L^{*2} &\equiv \frac{1}{n} \sum_{t=1}^{n+m} \left(\frac{1}{\sqrt{m+1}} \sum_{j=0}^m u_{t-j}^* \right)^2 = \frac{1}{m+1} \sum_{j=0}^m \frac{1}{n} \sum_{t=1}^{n+m} u_{t-j}^{*2} + 2 \frac{1}{m+1} \sum_{j=0}^{m-1} \sum_{i=1}^{m-j} \frac{1}{n} \sum_{t=1}^{n+m} u_{t-j}^* u_{t-j-i}^* \\ &= \frac{1}{m+1} \sum_{j=0}^m \frac{1}{n} \sum_{t=1-j}^{n+m-j} u_t^{*2} + 2 \frac{1}{m+1} \sum_{j=0}^{m-1} \sum_{i=1}^{m-j} \frac{1}{n} \sum_{t=1-j}^{n+m-j} u_t^* u_{t-i}^* \\ &= \frac{1}{n} \sum_{t=1}^n u_t^2 + 2 \frac{1}{m+1} \sum_{j=0}^{m-1} \sum_{i=1}^{m-j} \frac{1}{n} \sum_{t=i+1}^n u_t u_{t-i} = \frac{1}{n} \sum_{t=1}^n u_t^2 + 2 \frac{1}{m+1} \sum_{i=1}^m (m+1-i) \frac{1}{n} \sum_{t=i+1}^n u_t u_{t-i} \end{aligned} \quad (77)$$

is positive, and so is $\hat{\sigma}_L^2$. Next, observe from (62) and (74) that

$$\hat{u}_t = u_t - \sqrt{n} \hat{\alpha}_1 \gamma(1) W_n(t/n) - \hat{\alpha}_1 v_t + \hat{\alpha}_1 (v_0 - y_0) - \hat{\alpha}_0. \quad (78)$$

Since

$$E \left| (1/n) \sum_{t=1+i}^n u_t W_n((t-i)/n) \right| \leq \sqrt{(1/n) \sum_{t=1+i}^n E(u_t^2)} \sqrt{(1/n) \sum_{t=1+i}^n E(W_n((t-i)/n)^2)} = O(1),$$

it follows that $(1/n) \sum_{t=1+i}^n u_t W_n((t-i)/n) = O_p(1)$. Similarly, $(1/n) \sum_{t=1+i}^n u_{t-i} W_n(t/n) = O_p(1)$. Moreover, $\hat{\alpha}_1 = O_p(1/n)$, and similarly, it can be shown that $\hat{\alpha}_0 = O_p(1/\sqrt{n})$. Therefore, it follows from (77) and (78) that

$$\hat{\sigma}_L^2 - \hat{\sigma}_L^{*2} = O_p(1/n) + O_p \left(\sum_{i=1}^m [1 - i/(m+1)] / \sqrt{n} \right) = O_p(1/n) + O_p(m/\sqrt{n}). \quad (79)$$

A similar result holds under the stationarity hypothesis. Moreover, substituting $u_t = \sigma_L^2 e_t + v_t - v_{t-1}$, and denoting $e_t^* = e_t$, $v_t^* = v_t$ for $t = 1, \dots, n$, $v_t^* = e_t^* = 0$ for $t < 1$ and $t > n$, it is easy to verify that under the unit root hypothesis,

$$\begin{aligned}
\hat{\sigma}_L^{*2} &= \frac{1}{n} \sum_{t=1}^{n+m} \left(\sigma_L \frac{1}{\sqrt{m+1}} \sum_{j=0}^m e_{t-j}^* + \frac{v_t^* - v_{t-m}^*}{\sqrt{m+1}} \right)^2 \\
&= \sigma_L^2 \frac{1}{n} \sum_{t=1}^{n+m} \left(\frac{1}{\sqrt{m+1}} \sum_{j=0}^m e_{t-j}^* \right)^2 + 2\sigma_L \frac{1}{n} \sum_{t=1}^{n+m} \left(\frac{1}{\sqrt{m+1}} \sum_{j=0}^m e_{t-j}^* \right) \left(\frac{v_t^* - v_{t-m}^*}{\sqrt{m+1}} \right) \\
&\quad + \frac{1}{n} \sum_{t=1}^{n+m} \left(\frac{v_t^* - v_{t-m}^*}{\sqrt{m+1}} \right)^2 = \sigma_L^2 + O_p(\sqrt{m/n}) + O_p(1/\sqrt{m}) + O_p(1/m).
\end{aligned} \tag{80}$$

A similar result holds under the stationarity hypothesis. Thus:

THEOREM 4. Let m increase with n to infinity at rate $o(n^{1/2})$. Then under both the unit root and stationarity hypothesis, $\text{plim}_{n \rightarrow \infty}(\hat{\sigma}_L^2 - \hat{\sigma}_L^{*2}) = 0$. Moreover, under the unit root hypothesis, $\text{plim}_{n \rightarrow \infty} \hat{\sigma}_L^{*2} = \sigma_L^2$, and under the stationarity hypothesis, $\text{plim}_{n \rightarrow \infty} \hat{\sigma}_L^{*2} > 0$. Consequently, under stationarity, the Phillips-Perron test satisfies $\text{plim}_{n \rightarrow \infty} \hat{Z}_1/n < 0$.

Finally, note that the advantage of the PP test is that there is no need to specify the ARIMA process under the null hypothesis. It is in essence a nonparametric test. Of course, we still have to specify the Newey-West truncation lag m as a function of n , but as long as $m = o(\sqrt{n})$, this specification is asymptotically not critical.

6. Unit root with drift versus trend stationarity

Most macroeconomic time series in (log) levels have an upwards sloping pattern. Therefore, if they are (covariance) stationary, then they are stationary around a deterministic trend. If we would conduct the ADF and PP tests in Sections 4 and 5 to a linear trend stationary process, we will likely accept the unit root hypothesis, due to the following. Suppose we conduct the ADF test under the hypothesis $p = 1$ to the trend stationary process $y_t = \beta_0 + \beta_1 t + u_t$, where the u_t 's are i.i.d. $N(0, \sigma^2)$. It is a standard exercise to verify that then $\text{plim}_{n \rightarrow \infty} n \hat{\alpha}_1 = 0$, hence the ADF and PP tests in sections 4 and 5 have no power against linear trend stationarity!

Therefore, if one wishes to test the unit root hypothesis against linear trend stationarity, then a trend term should be included in the auxiliary regressions (39) in the ADF case, and in (60) in the

PP case: Thus the ADF regression (39) now becomes

$$\Delta y_t = \alpha_0 + \sum_{j=1}^{p-1} \alpha_j \Delta y_{t-j} + \alpha_p y_{t-1} + \alpha_{p+1} t + u_t, \quad u_t \sim i.i.d. N(0, \sigma_2^2) \quad (81)$$

where the null hypothesis of a unit root with drift corresponds to the hypothesis $\alpha_p = \alpha_{p+1} = 0$, and the PP regression becomes:

$$\Delta y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 t + u_t, \quad u_t = \gamma(L)e_t, \quad e_t \sim i.i.d. N(0, 1). \quad (82)$$

The asymptotic null distributions of the ADF and PP tests for the case with drift are quite similar to the ADF test without an intercept. The difference is that the Wiener process $W(x)$ is replaced by the detrended Wiener process:

$$W^{**}(x) = W(x) - 4 \int W(z) dz + 6 \int z W(z) dz + 6x \left(\int W(z) dz - 2 \int z W(z) dz \right) x$$

After some tedious but not too difficult calculations it can be shown that effectively the statistics $n\hat{\alpha}_p/(1-\sum_{j=1}^{p-1}\alpha_j)$ and \hat{t}_p are asymptotically equivalent to the Dickey-Fuller tests statistics $\hat{\rho}_0$ and $\hat{\tau}_0$, respectively, applied to detrended time series.

THEOREM 5. Let y_t be generated by (81), and let $\hat{\alpha}_p$ and \hat{t}_p be the OLS estimator and corresponding t-value of α_p . Under the unit root with drift hypothesis, i.e., $\alpha_p = \alpha_{p+1} = 0$, we have $n\hat{\alpha}_p \rightarrow (1-\sum_{j=1}^{p-1}\alpha_j)\rho_2$ and $\hat{t}_p \rightarrow \tau_2$ in distr., where

$$\rho_2 = \frac{1}{2} \left(\frac{W^{**}(1) - 1}{\int W^{**}(x)^2 dx} \right), \quad \tau_2 = \frac{1}{2} \left(\frac{W^{**}(1) - 1}{\sqrt{\int W^{**}(x)^2 dx}} \right).$$

Under the trend stationarity hypothesis, $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_p = \alpha_p < 0$, hence $\text{plim}_{n \rightarrow \infty} \hat{t}_p / \sqrt{n} < 0$.

The densities of ρ_2 and τ_2 (the latter compared with the standard normal density), are displayed in Figures 5 and 6, respectively. Again, these densities are farther to the left, and heavier left-tailed, than the corresponding densities displayed in Figures 1-4. The asymptotic 5% and 10% critical values of the Dickey-Fuller t-test are:

$$P(\tau_2 < -3.41) = 0.05, \quad P(\tau_2 < -3.13) = 0.10$$

Moreover, comparing (26) with

$$P(\tau_2 \leq -1.64) \approx 0.77, \quad P(\tau_2 \leq -1.28) \approx 0.89,$$

we see that the standard normal tests at the 5% and 10% significance level would reject the correct unit root with drift hypothesis with probabilities of about 0.77 and 0.89, respectively!

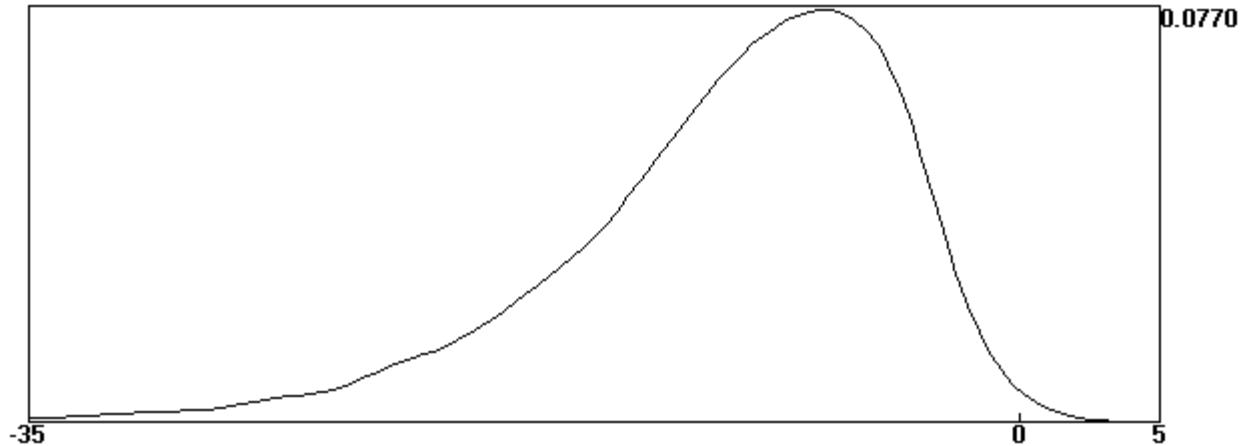


Figure 5: Density of ρ_2

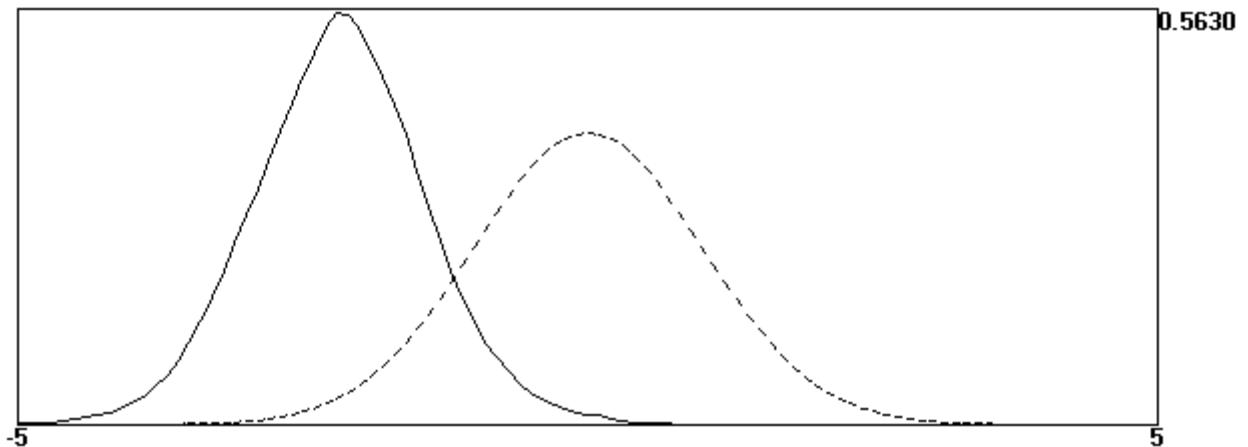


Figure 6: Density of τ_2 compared with the standard normal density (dashed curve)

A similar result as in Theorem 5 can be derived for the PP test, on the basis of the OLS estimator of α_1 in, and the residuals \hat{u}_t of, the auxiliary regression (82):

THEOREM 6. (*Phillips-Perron test 2*) Let \hat{r}_t be the residuals of the OLS regression of y_t on t and a constant, and let $\hat{\sigma}_u^2$ and $\hat{\sigma}_L^2$ be as before, with the \hat{u}_t 's the OLS residuals of the auxiliary

regression (82). Under the unit root with drift hypothesis,

$$\hat{Z}_2 = n \left(\hat{\alpha}_1 - \frac{(\hat{\sigma}_L^2 - \hat{\sigma}_u^2)/2}{(1/n)\sum_{t=1}^n \hat{r}_t^2} \right) \rightarrow \rho_2 \text{ in distr.}, \quad (83)$$

whereas under trend stationarity $\text{plim}_{n \rightarrow \infty} \hat{Z}_2/n < 0$.

7. Concluding remarks

In the discussion of the ADF test we have assumed that the lag length p of the auxiliary regression (81) is fixed. It should be noted that we may choose p as a function of the length n of the time series involved, similarly to the truncation width of the Newey-West estimator of the long-run variance in the Phillips-Perron test. See Said and Dickey (1984).

We have seen that the ADF and Phillips-Perron tests for a unit root against stationarity around a constant have almost no power if the correct alternative is linear trend stationarity. However, the same may apply to the tests discussed in section 6 if the alternative is trend stationarity with a broken trend. See Perron (1988, 1989, 1990), Perron and Vogelsang (1992), and Zivot and Andrews (1992), among others.

All the tests discussed so far have the unit root as the null hypothesis, and (trend) stationarity as the alternative. However, it is also possible to test the other way around. See Bierens and Guo (1993), and Kwiatkowski et.al. (1992). The latter test is known as the KPSS test.

Finally, note that the ADF and Phillips-Perron tests can easily be conducted by various econometric software packages, for example the commercial software packages TSP, EViews, RATS, and my freeware *EasyReg International*.¹²

¹² *EasyReg International* can be downloaded from URL
<http://econ.la.psu.edu/~hbierens/EASYREG.HTM>.

EasyReg International also contains my own unit root tests, Bierens (1993, 1997), Bierens and Guo (1993), and the KPSS test.

References

- Andrews, D.W.K., 1991, Heteroskedasticity and autocorrelation consistent covariance matrix estimators, *Econometrica* 59, 817-858.
- Bierens, H.J., 1993, Higher order sample autocorrelations and the unit root hypothesis, *Journal of Econometrics* 57, 137-160.
- Bierens, H.J., 1997, Testing the unit root hypothesis against nonlinear trend stationarity, with an application to the price level and interest rate in the U.S, *Journal of Econometrics* 81, 29-64.
- Bierens, H.J., 1994, *Topics in advanced econometrics: estimation, testing and specification of cross-section and time series models* (Cambridge University Press, Cambridge, U.K.).
- Bierens, H.J. and S. Guo, 1993, Testing stationarity and trend stationarity against the unit root hypothesis, *Econometric Reviews* 12, 1-32.
- Billingsley (1968), P., 1968, *Convergence of probability measures* (John Wiley, New York).
- Dickey, D.A. and W.A. Fuller, 1979, Distribution of the estimators for autoregressive times series with a unit root, *Journal of the American Statistical Association* 74, 427-431.
- Dickey, D.A. and W.A. Fuller, 1981, Likelihood ratio statistics for autoregressive time series with a unit root, *Econometrica* 49, 1057-1072.
- Fuller, W.A., 1996, *Introduction to statistical time series* (John Wiley, New York).
- Green, W., 1997, *Econometric analysis* (Prentice Hall, Upper Saddle River, NJ).
- Hogg, R.V. and A.T. Craig, 1978, *Introduction to mathematical statistics* (Macmillan, London).
- Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin, 1992, Testing the null of stationarity against the alternative of a unit root, *Journal of Econometrics* 54, 159-178.
- Newey, W.K. and K.D. West, 1987, A simple positive definite heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703-708.
- Perron, P., 1988, Trends and random walks in macroeconomic time series: further evidence from a new approach, *Journal of Economic Dynamics and Control* 12, 297-332.
- Perron, P., 1989, The great crash, the oil price shock and the unit root hypothesis, *Econometrica* 57, 1361-1402.

Perron, P., 1990, Testing the unit root in a time series with a changing mean, *Journal of Business and Economic Statistics* 8, 153-162.

Perron, P. and T.J. Vogelsang, 1992, Nonstationarity and level shifts with an application to purchasing power parity, *Journal of Business and Economic Statistics* 10, 301-320.

Phillips, P.C.B., 1987, Time series regression with a unit root, *Econometrica* 55, 277-301.

Phillips, P.C.B. and P. Perron, 1988, Testing for a unit root in time series regression, *Biometrika* 75, 335-346.

Said, S.E. and D.A. Dickey, 1984, Testing for unit roots in autoregressive-moving average of unknown order, *Biometrika* 71, 599-607.

Zivot, E. and D.W.K. Andrews, 1992, Further evidence on the great crash, the oil price shock, and the unit root hypothesis, *Journal of Business and Economic Statistics* 10, 251-270.

FORECASTING
Herman J. Bierens
Pennsylvania State University

February 26, 2009

1. Recursive best linear forecasting

Let Y_t be a covariance stationary time series process, with $E[Y_t] = 0$. The best linear h -step ahead forecast of Y_{t+h} , $h = 1, 2, 3, \dots$, given the observations on $Y_t, Y_{t-1}, Y_{t-2}, \dots, Y_{t-m}$ is a linear function of Y_{t-j} , $j = 0, 1, \dots, m$, say:

$$\hat{Y}_{t+h|t,m} = \sum_{j=0}^m \gamma_{h,m,j} Y_{t-j}, \quad (1)$$

such that the mean-square forecast error

$$E(Y_{t+h} - \hat{Y}_{t+h|t,m})^2 = E\left(Y_{t+h} - \sum_{j=0}^m \gamma_{h,m,j} Y_{t-j}\right)^2 \quad (2)$$

is minimal. Therefore, the coefficients $\gamma_{h,m,j}$ are such that the first-order conditions

$$E\left(Y_{t+h} - \sum_{j=0}^m \gamma_{h,m,j} Y_{t-j}\right) Y_{t-k} = 0 \text{ for } k = 0, 1, 2, \dots, m \quad (3)$$

are satisfied.

Note that we can write (2) and (3) in terms of the covariance function

$$f(k) = \text{cov}(Y_t, Y_{t-k}) = E[Y_t Y_{t-k}] \quad (4)$$

(the last equality follows from the assumption that $E[Y_t] = 0$):

$$E(Y_{t+h} - \hat{Y}_{t+h|t,m})^2 = f(0) - 2 \sum_{j=0}^m \gamma_{h,m,j} f(h+j) + \sum_{i=0}^m \sum_{j=0}^m \gamma_{h,m,i} \gamma_{h,m,j} f(|i-j|) \quad (5)$$

with first-order conditions:

$$f(h+k) = \sum_{j=0}^m \gamma_{h,m,j} f(|k-j|), \quad k = 0, 1, 2, \dots, m \quad (6)$$

Given the covariance function $f()$, we can in general solve the coefficients $\gamma_{h,m,j}$ uniquely from (6).

Obviously, the mean-square error (2) is non-increasing in m , hence

$$\lim_{m \rightarrow \infty} E(Y_{t+h} - \hat{Y}_{t+h|t,m})^2 = \lim_{m \rightarrow \infty} E\left(Y_{t+h} - \sum_{j=0}^m \gamma_{h,m,j} Y_{t-j}\right)^2 \quad (7)$$

exists. However, in general this result does not imply that $\lim_{m \rightarrow \infty} \hat{Y}_{t+h|t,m}$ exists,¹ but it does imply that there exists a random variable $\hat{Y}_{t+h|t}$ measurable with respect to the σ -algebra $\mathcal{F}_{-\infty}^t$ generated by $Y_t, Y_{t-1}, Y_{t-2}, \dots$ such that

$$\lim_{m \rightarrow \infty} E(\hat{Y}_{t+h|t} - \hat{Y}_{t+h|t,m})^2 = \lim_{m \rightarrow \infty} E\left(Y_{t+h|t} - \sum_{j=0}^m \gamma_{h,m,j} Y_{t-j}\right)^2 = 0 \quad (8)$$

This random variable $\hat{Y}_{t+h|t}$ is called the best linear h -step ahead linear forecast of Y_{t+h} . On the other hand, if Y_t is a covariance stationary ARMA process with invertible MA lag polynomial² then $\hat{Y}_{t+h|t} = \lim_{m \rightarrow \infty} \hat{Y}_{t+h|t,m}$ exists and takes the form

$$\hat{Y}_{t+h|t} = \sum_{j=0}^{\infty} \gamma_{h,j} Y_{t-j} \quad (9)$$

with coefficients determined by

$$f(h+k) = \sum_{j=0}^{\infty} \gamma_{h,j} f(|k-j|), \quad k = 0, 1, 2, 3, \dots \quad (10)$$

In the rest of this lecture note I will focus on this case only.

Now consider the best linear one-step ahead forecast of Y_{t+2} :

$$\hat{Y}_{t+2|t+1} = \sum_{j=0}^{\infty} \gamma_{1,j} Y_{t+1-j} = \gamma_{1,0} Y_{t+1} + \sum_{j=0}^{\infty} \gamma_{1,j+1} Y_{t-j}. \quad (11)$$

This expression can be rewritten as

¹ See for example: Bierens, H.J. (2009), "The space spanned by a countable infinite sequence in an Hilbert space, with application to the Wold decomposition", lecture note downloadable from URL http://econ.la.psu.edu/~hbierens/HILBERT_SPAN.PDF

² I.e, the roots of the MA lag polynomial involved are located outside the complex unit circle, and for the ARMA process to be covariance stationary the same must hold for the AR lag polynomial.

$$\hat{Y}_{t+2|t+1} = \gamma_{1,0}(Y_{t+1} - \hat{Y}_{t+1|t}) + \gamma_{1,0}\hat{Y}_{t+1|t} + \sum_{j=0}^{\infty} \gamma_{1,j+1}Y_{t-j}. \quad (12)$$

It follows from the first-order conditions that for $k = 0, 1, 2, \dots$

$$\begin{aligned} 0 &= E(Y_{t+2} - \hat{Y}_{t+2|t+1})Y_{t-k} = E\left[-\gamma_{1,0}(Y_{t+1} - \hat{Y}_{t+1|t}) + Y_{t+2} - \gamma_{1,0}\hat{Y}_{t+1|t} - \sum_{j=0}^{\infty} \gamma_{1,j+1}Y_{t-j}\right]Y_{t-k} \\ &= -\gamma_{1,0}E(Y_{t+1} - \hat{Y}_{t+1|t})Y_{t-k} + E\left[Y_{t+2} - \gamma_{1,0}\hat{Y}_{t+1|t} - \sum_{j=0}^{\infty} \gamma_{1,j+1}Y_{t-j}\right]Y_{t-k} \quad (13) \\ &= E\left[Y_{t+2} - \gamma_{1,0}\hat{Y}_{t+1|t} - \sum_{j=0}^{\infty} \gamma_{1,j+1}Y_{t-j}\right]Y_{t-k}, \end{aligned}$$

hence:

$$\hat{Y}_{t+2|t} = \gamma_{1,0}\hat{Y}_{t+1|t} + \sum_{j=0}^{\infty} \gamma_{1,j+1}Y_{t-j}. \quad (14)$$

More generally we have:

THEOREM 1. Let Y_t be a covariance stationary ARMA process with $E[Y_t] = 0$ and invertible MA lag polynomial. Let Y_{t-j} be observable for $j = 0, 1, 2, \dots$. Replacing in the expression for the best linear one-step ahead forecast $\hat{Y}_{t+h|t+h-1}$ of Y_{t+h} , i.e.,

$$\hat{Y}_{t+h|t+h-1} = \sum_{j=0}^{\infty} \gamma_{1,j}Y_{t+h-1-j} = \sum_{j=0}^{h-2} \gamma_{1,j}Y_{t+h-1-j} + \sum_{j=h-1}^{\infty} \gamma_{1,j}Y_{t+h-1-j} \quad (15)$$

the unobserved $Y_{t+h-1-j}$, $j = 0, \dots, h-2$, by best linear forecasts $\hat{Y}_{t+h-1-j|t}$, respectively, yields the best linear h -step ahead forecast of Y_{t+h} :

$$\hat{Y}_{t+h|t} = \sum_{j=0}^{h-2} \gamma_{1,j}\hat{Y}_{t+h-1-j|t} + \sum_{j=h-1}^{\infty} \gamma_{1,j}Y_{t+h-1-j}. \quad (16)$$

If $E[Y_t] = \mu \neq 0$, the best linear h -step ahead forecast takes the form

$$\hat{Y}_{t+h|t} = \delta_h + \sum_{j=0}^{\infty} \gamma_{h,j}Y_{t-j}. \quad (17)$$

Exercise 1: Show that

$$\delta_h = \left(1 - \sum_{j=0}^{\infty} \gamma_{h,j} \right) \mu \quad (18)$$

with the coefficients $\gamma_{h,j}$ determined by (10), so that

$$\hat{Y}_{t+h|t} = \mu + \sum_{j=0}^{\infty} \gamma_{h,j} (Y_{t-j} - \mu). \quad (19)$$

The practical implication of this result is that in forecasting Y_{t+h} we may first forecast $Y_{t+h} - \mu$, using the result of Theorem 1, and then add μ to the forecast involved.

Exercise 2: Prove that:

THEOREM 2: For the case $E[Y_t] = \mu \neq 0$ the result of Theorem 1 becomes

$$\hat{Y}_{t+h|t} = \sum_{j=0}^{h-2} \gamma_{1,j} \hat{Y}_{t+h-1-j|t} + \sum_{j=h-1}^{\infty} \gamma_{1,j} Y_{t+h-1-j} + \left(1 - \sum_{j=0}^{\infty} \gamma_{1,j} \right) \mu. \quad (20)$$

2. Forecasting with an ARMA(p,q) model

Consider the ARMA(p,q) process

$$Y_t = \mu + u_t, \quad \alpha(L)u_t = \beta(L)e_t,$$

where

$$\begin{aligned} \alpha(L) &= 1 - \sum_{j=1}^p \alpha_j L^j, \quad \beta(L) = 1 - \sum_{j=1}^q \beta_j L^j, \\ \alpha(z) &= 0 \Rightarrow |z| > 1, \quad \beta(z) = 0 \Rightarrow |z| > 1, \\ e_t \text{ is white noise: } E(e_t) &= 0, \quad E(e_t^2) = \sigma^2 < \infty, \quad E(e_t e_{t-j}) = 0 \text{ for } j \neq 0. \end{aligned} \quad (21)$$

Moreover, we have to assume that the lag polynomials $\alpha(L)$ and $\beta(L)$ do not have common roots (*Exercise 3: Why?*). Since the lag polynomial $\beta(L)$ is invertible, because all its roots are outside the unit circle, we can write this process as an AR(∞) process:

$$\gamma(L)(Y_t - \mu) = e_t, \text{ where } \gamma(L) = \beta(L)^{-1}\alpha(L) = 1 - \sum_{j=0}^{\infty} \gamma_j L^{j+1}, \quad (22)$$

say. Note that $\gamma(z) = 0 \Rightarrow |z| > 1$. Thus:

$$Y_{t+1} = \delta + \sum_{j=0}^{\infty} \gamma_j Y_{t-j} + e_{t+1}, \text{ where } \delta = \left(1 - \sum_{j=0}^{\infty} \gamma_j \right) \mu. \quad (23)$$

Consequently, the best linear one-step ahead forecast of Y_{t+1} is:

$$\hat{Y}_{t+1|t} = \delta + \sum_{j=0}^{\infty} \gamma_j Y_{t-j} = \mu + \sum_{j=0}^{\infty} \gamma_j (Y_{t-j} - \mu). \quad (24)$$

(Exercise 4: Why?) Using Theorem 2, we can recursively find the best linear h -step ahead forecast of Y_{t+h} by

$$\hat{Y}_{t+h|t} = \sum_{j=0}^{h-2} \gamma_j \hat{Y}_{t+h-1-j|t} + \sum_{j=h-1}^{\infty} \gamma_j Y_{t+h-1-j} + \left(1 - \sum_{j=0}^{\infty} \gamma_j \right) \mu. \quad (25)$$

The practical problem with the above approach is three-fold: First, we usually do not observe the whole process Y_t , but only a finite number of Y_t 's, say for $t = 1, \dots, n$. Second, p and q are unknown. We will address that problem later. Third, we do not observe the coefficients α_i , β_j directly. These coefficients have to be estimated. The latter can be done by maximum likelihood, but that requires further assumptions on the distribution of the white noise errors e_t .

An alternative approach is nonlinear least squares estimation, together with the assumption that $e_t = 0$ for $t < 1$, hence $u_t = 0$ for $t < 1$ and $Y_t = \mu$ for $t < 1$. The assumption $e_t = 0$ for $t < 1$ is asymptotically innocent: it does not affect the consistency or asymptotic normality of the parameter estimates. The least squares problem involved is:

$$\begin{aligned} & \min_{\theta} \sum_{t=1}^n e_t(\theta)^2, \\ & \text{subject to} \\ & e_t(\theta) = \sum_{j=1}^q \beta_j I(t-j>0) e_{t-j}(\theta) + Y_t - \mu - \sum_{j=1}^p \alpha_j I(t-j>0) (Y_{t-j} - \mu), \quad t = 1, \dots, n, \\ & \text{where } \theta = (\mu, \alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_q)^T, \end{aligned} \quad (26)$$

with $I(\cdot)$ the indicator function: $I(\text{true}) = 1$, $I(\text{false}) = 0$. Under some regularity conditions, in particular the condition that p and q are correctly specified, and the condition that the errors e_t are martingale differences: $E[e_t | e_{t-1}, e_{t-2}, e_{t-3}, \dots] = 0$, it can be shown that the nonlinear least squares estimator $\hat{\theta} = (\hat{\mu}, \hat{\alpha}_1, \dots, \hat{\alpha}_p, \hat{\beta}_1, \dots, \hat{\beta}_q)^T$ is consistent and asymptotically normally distributed:

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N_{p+q+1}[0, \Omega_1^{-1} \Omega_2 \Omega_1^{-1}] \text{ in distribution,}$$

where

$$\begin{aligned}\Omega_1 &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n E \left[\left(\frac{\partial e_t(\theta)}{\partial \theta^T} \right) \left(\frac{\partial e_t(\theta)}{\partial \theta} \right) \right], \\ \Omega_2 &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n E \left[e_t(\theta)^2 \left(\frac{\partial e_t(\theta)}{\partial \theta^T} \right) \left(\frac{\partial e_t(\theta)}{\partial \theta} \right) \right].\end{aligned}\tag{27}$$

Moreover, these two matrices can be consistently estimated by

$$\hat{\Omega}_1 = \frac{1}{n} \sum_{t=1}^n \left(\frac{\partial e_t(\theta)}{\partial \theta^T} \right) \left(\frac{\partial e_t(\theta)}{\partial \theta} \right) \Bigg|_{\theta=\hat{\theta}}, \quad \hat{\Omega}_2 = \frac{1}{n} \sum_{t=1}^n e_t(\theta)^2 \left(\frac{\partial e_t(\theta)}{\partial \theta^T} \right) \left(\frac{\partial e_t(\theta)}{\partial \theta} \right) \Bigg|_{\theta=\hat{\theta}},\tag{28}$$

respectively.

Once we have estimated the parameters α_i , β_j , we can compute the γ_j 's recursively, as follows. Observe from (22) that

$$e_t = u_t - \sum_{j=0}^{\infty} \gamma_j u_{t-1-j}.\tag{29}$$

If we set in (29) $u_t = -1$ for $t = -1$, $u_t = 0$ for $t \neq -1$, then

$$\begin{aligned}e_j &= 0 \text{ for } j < -1 \\ e_{-1} &= u_{-1} = -1 \\ e_0 &= u_0 - \gamma_0 u_{-1} = \gamma_0 \\ e_1 &= u_1 - \gamma_0 u_0 - \gamma_1 u_{-1} = \gamma_1 \\ e_2 &= u_2 - \gamma_0 u_2 - \gamma_1 u_0 - \gamma_2 u_{-1} = \gamma_2 \\ &\dots \\ e_t &= \gamma_t, \quad t \geq 0.\end{aligned}\tag{30}$$

But it follows from (21) that also

$$e_t = \sum_{j=1}^q \beta_j e_{t-j} + u_t - \sum_{j=1}^p \alpha_j u_{t-j}. \quad (31)$$

Thus if we set in (31), $u_t = -1$ for $t = -1$, $u_t = 0$ for $t \neq -1$, $e_j = 0$ for $j < -1$, then $e_{-1} = -1$ and $e_t = \gamma_t$ for $t = 0, 1, 2, \dots$. Therefore, the γ_j 's can be solved recursively, on the basis of the nonlinear least squares estimation results, by:

$$\begin{aligned} \hat{\gamma}_{-1-j} &= 0 \text{ for } j > 0, \\ \hat{\gamma}_{-1} &= -1, \\ \hat{\gamma}_0 &= \hat{\alpha}_1 - \hat{\beta}_1, \\ \hat{\gamma}_j &= \sum_{i=1}^q \hat{\beta}_i \hat{\gamma}_{j-i} + \hat{\alpha}_{j+1} \text{ for } j = 2, \dots, p-1, \\ \hat{\gamma}_j &= \sum_{i=1}^q \hat{\beta}_i \hat{\gamma}_{j-i} \text{ for } j \geq p. \end{aligned} \quad (32)$$

Replacing in (24) the Y_{t-j} for $j \geq t$ by $\hat{\mu}$ and the other parameters by their estimates, yields the feasible best linear one-step ahead forecast

$$\tilde{Y}_{t+1|t} = \hat{\mu} + \sum_{j=0}^{t-1} \hat{\gamma}_j (Y_{t-j} - \hat{\mu}),$$

and replacing in (25) $Y_{t+h-1-j}$ for $j \geq t+h-1$ by $\hat{\mu}$ and the other parameters by their estimates, yields the recursive formula for the feasible best linear h -step ahead forecast:

$$\tilde{Y}_{t+h|t} = \hat{\mu} + \sum_{j=0}^{h-2} \hat{\gamma}_j \tilde{Y}_{t+h-1-j|t} + \sum_{j=h-1}^{t+h-2} \hat{\gamma}_j (Y_{t+h-1-j} - \hat{\mu}). \quad (34)$$

As to the choice of p and q , there are a few model selection tools on the market such as the Akaike, Hannan-Quinn and Schwarz information criteria.³ However, if forecasting is the goal, then the out-of-sample forecasting performance may be a better criterion. Thus, select a sub-sample Y_1, \dots, Y_m , $m < n$, and estimate the parameters of the ARMA(p, q) model using the sub-sample only. Then choose p and q such that the sum of squared out-of-sample forecast errors,

$$\sum_{h=1}^{n-m} (Y_{m+h} - \tilde{Y}_{m+h|m})^2$$

³ See for example Bierens, H.J. (2006), "Information Criteria and Model Selection", lecture note, downloadable from <http://econ.la.psu.edu/~hbierens/INFORMATIONCRIT.PDF>

is minimal. Once you have determined p and q , re-estimate the parameters using the whole sample Y_1, \dots, Y_n , and forecast Y_{n+h} by $\tilde{Y}_{m+h|n}$.

SPURIOUS REGRESSION

Herman J. Bierens

Pennsylvania State University

Consider two independent unit root processes, $\Delta y_t = u_t$ and $\Delta x_t = v_t$, where the u_t 's and the v_t 's are independent, say:

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} \sim i.i.d. N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right). \quad (1)$$

Then y_t , with all leads and lags, is independent of x_t , with all leads and lags. If we regress y_t on x_t for $t = 1, \dots, n$, one would therefore expect the slope coefficient to be insignificant, but as we show now, that is not true. First, consider the OLS regression of y_t on x_t without an intercept. The OLS estimate of the slope is:

$$\begin{aligned} \hat{\gamma}_0 &= \frac{\sum_{t=1}^n x_t y_t}{\sum_{t=1}^n x_t^2} = \frac{(1/n) \sum_{t=1}^n \left(x_0/\sqrt{n} + (1/\sqrt{n}) \sum_{i=1}^t v_i \right) \left(y_0/\sqrt{n} + (1/\sqrt{n}) \sum_{j=1}^t u_j \right)}{(1/n) \sum_{t=1}^n \left(x_0/\sqrt{n} + (1/\sqrt{n}) \sum_{i=1}^t v_i \right)^2} \\ &= \frac{(1/n) \sum_{t=1}^n \left(x_0/\sqrt{n} + W_{x,n}(t/n) \right) \left(y_0/\sqrt{n} + W_{y,n}(t/n) \right)}{(1/n) \sum_{t=1}^n \left(x_0/\sqrt{n} + W_{x,n}(t/n) \right)^2}, \end{aligned} \quad (2)$$

where

$$W_{x,n}(r) = (1/\sqrt{n}) \sum_{j=1}^{[nr]} v_j \text{ if } r \in [n^{-1}, 1], \quad W_{x,n}(r) = 0 \text{ if } r \in [0, n^{-1}), \quad (3)$$

$$W_{y,n}(r) = (1/\sqrt{n}) \sum_{j=1}^{[nr]} u_j \text{ if } r \in [n^{-1}, 1], \quad W_{y,n}(r) = 0 \text{ if } r \in [0, n^{-1}), \quad (4)$$

with $[rn]$ the largest natural number $\leq rn$. Since these functions are step functions, and both $W_{x,n}(1)$ and $W_{y,n}(1)$ are standard normally distributed, hence $W_{x,n}(1) = O_p(1)$ and $W_{y,n}(1) = O_p(1)$, we have

$$\begin{aligned}
(1/n) \sum_{t=1}^n W_{x,n}(t/n)^2 &= (1/n) \sum_{t=0}^{n-1} \int_t^{t+1} W_{x,n}(z/n)^2 dz + W_{x,n}(1)^2/n \\
&= (1/n) \int_0^n W_{x,n}(z/n)^2 dz + W_{x,n}(1)^2/n = \int W_{x,n}(r)^2 dr + O_p(1/n),
\end{aligned} \tag{5}$$

and similarly

$$(1/n) \sum_{t=1}^n W_{y,n}(t/n)^2 = \int W_{y,n}(r)^2 dr + O_p(1/n), \tag{6}$$

$$(1/n) \sum_{t=1}^n W_{x,n}(t/n) W_{y,n}(t/n) = \int W_{x,n}(r) W_{y,n}(r) dr + O_p(1/n). \tag{7}$$

The integrals in (5) through (7), and in the sequel, are taken over the unit interval [0,1], unless otherwise indicated.

It can be shown that the integrals in (5) through (7) converge jointly in distribution, due to the functional central limit theorem and the continuous mapping theorem:

LEMMA 1. $\left(\int W_{x,n}(r)^2 dr, \int W_{y,n}(r)^2 dr, \int W_{x,n}(r) W_{y,n}(r) dr \right)^T$ converges in distribution to

$\left(\int W_x(r)^2 dr, \int W_y(r)^2 dr, \int W_x(r) W_y(r) dr \right)^T$, where $W_x(r)$ and $W_y(r)$ are independent standard Wiener processes.

For the proof of Lemma 1, see Billingsley (1968)¹.

Using Lemma 1, we now have

$$\hat{\gamma}_0 = \frac{\int W_{y,n}(r) W_{x,n}(r) dr}{\int W_{x,n}(r)^2 dr} + o_p(1) \rightarrow \gamma_0 = \frac{\int W_y(r) W_x(r) dr}{\int W_x(r)^2 dr} \tag{8}$$

¹ Billingsley, Patric: *Convergence of Probability Measures*, John Wiley, New York, 1968

in distribution. Note that the limiting random variable γ_0 is continuously distributed, and in particular, $P(\gamma_0 = 0) = 0$.

Along the same lines, and using (8), it follows that the residual sum of squares, RSS_0 , of the regression involved, divided by n^2 , satisfies

$$\begin{aligned} \frac{RSS_0}{n^2} &= \frac{1}{n^2} \sum_{t=1}^n y_t^2 - \hat{\gamma}_0^2 \frac{1}{n^2} \sum_{t=1}^n x_t^2 = \int W_{y,n}(r)^2 dr - \hat{\gamma}_0^2 \int W_{x,n}(r)^2 dr + o_p(1) \\ &\rightarrow \int W_y(r)^2 dr - \gamma_0^2 \int W_x(r)^2 dr \text{ in distr.} \end{aligned} \quad (9)$$

hence the t-value \hat{t}_0 , say, of the slope, divided by \sqrt{n} , satisfies:

$$\begin{aligned} \frac{\hat{t}_0}{\sqrt{n}} &= \frac{\hat{\gamma}_0 \sqrt{(1/n) \sum_{t=1}^n x_t^2}}{\sqrt{RSS_0/(n-1)}} = \sqrt{n/(n-1)} \frac{\hat{\gamma}_0 \sqrt{(1/n^2) \sum_{t=1}^n x_t^2}}{\sqrt{RSS_0/n^2}} \\ &= \frac{\hat{\gamma}_0 \sqrt{\int W_{x,n}(r)^2 dr}}{\sqrt{\int W_{y,n}(r)^2 dr - \hat{\gamma}_0^2 \int W_{x,n}(r)^2 dr}} + o_p(1) \rightarrow \frac{\gamma_0 \sqrt{\int W_x(r)^2 dr}}{\sqrt{\int W_y(r)^2 dr - \gamma_0^2 \int W_x(r)^2 dr}} \end{aligned} \quad (10)$$

in distribution. This result, together with $P(\gamma_0 = 0) = 0$, implies that $\text{plim}_{n \rightarrow \infty} |\hat{t}_0| = \infty$.

Similar results as in (8) and (10) hold for slope parameter $\hat{\gamma}_1$ and corresponding t-value \hat{t}_1 of the regression of y_t on x_t with intercept:

$$\hat{\gamma}_1 \rightarrow \gamma_1 = \frac{\int W_y(r) W_x(r) dr - \left(\int W_y(r) dr \right) \left(\int W_x(r) dr \right)}{\int W_x(r)^2 dr - \left(\int W_x(r) dr \right)^2} \quad (11)$$

and

$$\frac{\hat{t}_1}{\sqrt{n}} \rightarrow \gamma_1 \frac{\sqrt{\int W_x(r)^2 dr - \left(\int W_x(r) dr \right)^2}}{\sqrt{\int W_y(r)^2 dr - \left(\int W_y(r) dr \right)^2 - \gamma_1^2 \int W_x(r)^2 dr + \gamma_1^2 \left(\int W_x(r) dr \right)^2}} \quad (12)$$

in distribution. Moreover, the R^2 of the regression involved satisfies:

$$R^2 \rightarrow \gamma_1^2 \frac{\int W_x(r)^2 dr - \left(\int W_x(r) dr \right)^2}{\int W_y(r)^2 dr - \left(\int W_y(r) dr \right)^2} \quad (13)$$

in distribution

The conclusion from these results is that one should be very cautious when conducting standard econometric analysis using time series. If the time series involved are unit root processes, naive application of regression analysis may yield nonsense results.

COINTEGRATION ANALYSIS

Herman J. Bierens¹

Pennsylvania State University

April 7, 2010

1. *Introduction*

1.1 *What is cointegration?*

The basic idea behind cointegration is that if all the components of a vector time series process z_t have a unit root, or in other words, if z_t is a multivariate $I(1)$ process, there may exist linear combinations $\beta^T z_t$ without a unit root. These linear combinations may then be interpreted as long term relations between the components of z_t , or in economic terms as static equilibrium relations.

For bivariate economic $I(1)$ processes, cointegration often manifests itself by more or less parallel shapes of the plots of the two series involved. Figure 1 displays a typical example of such a pair of cointegrated economic time series, namely the log of nominal income (upper curve) and the log of nominal consumption (lower curve) in Sweden² from 1861 to 1988.

According to Friedman's (1957) permanent income theory, the long run marginal propensity to consume from permanent income should be close to one. With the logs of consumption and income being unit root with drift processes, the modern interpretation of the permanent income hypothesis therefore is that the difference of the logs of consumption and income is stationary: $\beta = (1, -1)^T$. However, income in Friedman's theory is net income rather than gross income, so that the long run marginal propensity to consume from gross income might be less than one. Anyhow, Friedman's theory predicts that the logs of consumption and income are cointegrated. The time series displayed in Figure 1 will be used in an empirical application in section 6.

¹ This paper is an updated and extended version of Bierens (1997b).

² I like to thank Philip Hans Franses for providing me with this data set. The original sources of these time series are Krantz and Nilson (1975) and Melander, Vredin and Warne (1992).

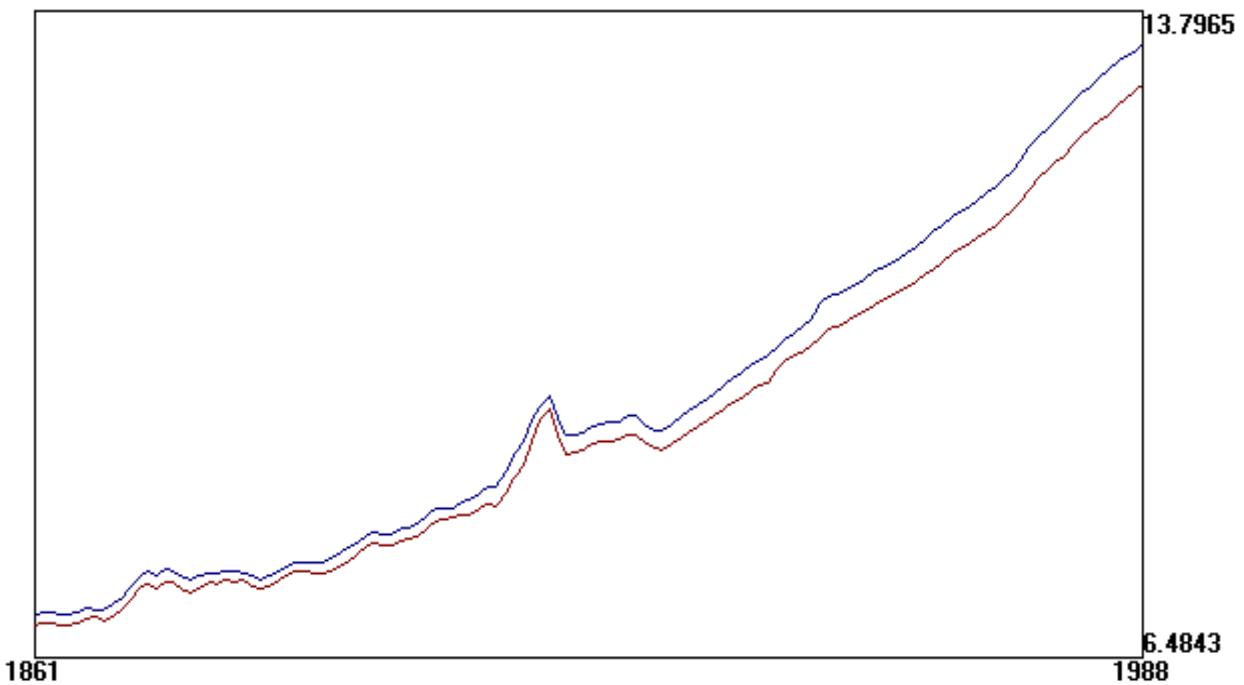


Figure 1: *Logs of Income and Consumption in Sweden*

1.2 The literature on cointegration

The concept of cointegration was first introduced by Granger (1981) and elaborated further by Engle and Granger (1987), Engle and Yoo (1987, 1991), Phillips and Ouliaris (1990), Stock and Watson (1988), Phillips (1991) and Johansen (1988, 1991, 1994), among others.

Working in the context of a bivariate system with at most one cointegrating vector, Engle and Granger (1987) propose to estimate the cointegrating vector $\beta = (1, \beta_2)^T$ by regressing the first component $z_{1,t}$ of z_t on the second component $z_{2,t}$, using OLS (which is called the cointegrating regression), and then testing whether the OLS residuals of this regression have a unit root, using the Augmented Dickey-Fuller (ADF) test. See Fuller (1976), Dickey and Fuller (1979, 1981) and Said and Dickey (1984) for the latter. However, since the ADF test is conducted on estimated residuals, the tables of the critical values of this test in Fuller (1976) do not apply anymore. The correct critical values involved can be found in Engle and Yoo (1987). Phillips and Ouliaris' (1990) tests are also based on these residuals, but instead of using the ADF test for testing the presence of a unit root they use further elaborations of the Phillips (1987) and Phillips-Perron

(1988) unit root tests. Both types of tests have absence of cointegration as the null hypothesis. Park (1990) proposes a test for unit root and cointegration using the variable addition approach, by regressing the OLS residuals of the cointegrating regression on powers of time and testing whether the coefficients involved are jointly zero. The same idea has been used by Bierens and Guo (1993) to test (trend) stationarity against the unit root hypothesis. However, also Park's approach requires consistent estimation of the long-run variance of the errors of the true cointegrating regression by a Newey-West (1987) type estimator, which sacrifices a substantial amount of asymptotic power of the test. Cf. Bierens and Guo (1993) for the latter. Also the tests of Hansen (1992) and Park (1992) are based on a single cointegrating regression, and both tests employ variants of the instrumental variables estimation method of Phillips and Hansen (1990). Finally, Boswijk (1994, 1995) links the single-equation and system approaches by using structural single-equations as a basis for cointegration analysis.

The above approaches test the null or alternative hypothesis of absence of cointegration, but if the tests indicate the presence of cointegration in systems with three variables or more we still don't know how many linear independent cointegrating vectors there are. In such cases one may use the approach of Stock and Watson (1988), which is a multivariate extension of the Engle-Granger and Phillips-Ouliaris tests. The basic idea is to linearly transform the q -variate cointegrated process z_t with say r linear independent cointegrating vectors such that the first r components of the transformed z_t are stationary and the last $q-r$ components, stacked in a vector w_t , say, are integrated. The transformation matrix involved can be consistently estimated using principal components of z_t . Then test whether w_t is a $q-r$ variate unit root process, using a multivariate version of the ADF test or the Phillips (1987) test. The critical values of this test differ according to whether the initial value z_0 is non-zero or not and whether the unit root process z_t has drift or not.

In a series of influential papers, Johansen (1988, 1991) and Johansen and Juselius (1990) propose an ingenious and practical full maximum likelihood estimation and testing approach, based on the following *Vector Error Correction Model* (hereafter indicted by ECM) for the q -variate unit root process z_t :

$$\Delta z_t = \Pi_0 d_t + \sum_{j=1}^{p-1} \Pi_j \Delta z_{t-j} + \alpha \beta^T z_{t-p} + e_t. \quad (1)$$

Here $\Delta z_t = z_t - z_{t-1}$, d_t is a vector of deterministic variables, such as a constant and seasonal dummy variables, the $\Pi_j, j \geq 0$, are $q \times q$ and β and α are $q \times r$ parameter matrices, where β and α are of full column rank, with r the number of linear independent cointegrating vectors (the columns of β), the e_t are i.i.d. $N_q(0, \Sigma)$ errors, and $\det(I - \sum_{j=1}^{p-1} \Pi_j L^j)$ has all its roots outside the complex unit circle. Note that if $r = q$, so that then the matrix $\alpha\beta^T$ is of full rank, and if $d_t = 1$, then model (1) generates a stationary AR(p) process z_t .

The VECM (1) is based on the Engle-Granger (1987) error correction representation theorem for cointegrated systems, and the asymptotic inference involved is related to the work of Sims, Stock and Watson (1990). By step-wise concentrating all the parameter matrices in the likelihood function out, except the matrix β , Johansen shows that the maximum likelihood estimator of β can be derived as the solution of a generalized eigenvalue problem. Likelihood ratio tests of hypotheses about the number of cointegrating vectors can then be based on these eigenvalues. Moreover, Johansen (1988) also proposes likelihood ratio tests for linear restrictions on the cointegrating vectors.

Initially, Johansen (1988) considered the case where d_t is absent. Later on, Johansen (1991) extended his approach to the case where d_t contains an intercept and seasonal dummy variables, and in Johansen (1994) also a time trend in d_t (but no seasonal dummy variables) is allowed. These three cases lead to different null distributions of the likelihood ratio tests of the number of cointegrating vectors. Moreover, also possible restrictions on the vector of intercepts or the vector of trend coefficients may lead to different null distributions. Thus, application of Johansen's tests actually requires some a priori knowledge about the true parameters of the VECM (1).

Phillips' (1991) efficient error correction modeling approach differs from that of Johansen (1988) in that Phillips specifies the VECM directly on the basis of the cointegrating relations $z_{1,t} = Bz_{2,t} + u_t$, with u_t a stationary zero mean Gaussian process, leading to an VECM of the form

$$\Delta z_t = \begin{pmatrix} I_r & -B \\ O & O \end{pmatrix} z_{t-t} + v_t, \quad (2)$$

where r is the number of cointegrating relations and v_t is a stationary Gaussian process with long

run variance matrix $\Omega = \lim_{n \rightarrow \infty} \text{Var}[(1/\sqrt{n})\sum_{t=1}^n v_t]$. Phillips shows that under the i.i.d. assumption on v_t , the maximum likelihood estimator of B is efficient, and that this efficiency carries over to the case with dependent errors v_t if B is estimated by maximum likelihood on the basis of model (2) with i.i.d. $N(0, \Omega)$ errors v_t , provided Ω is replaced by a consistent estimator. In contrast with Johansen's maximum likelihood method, however, Phillips' efficient maximum likelihood approach has not yet been widely applied in empirical research, probably due to the fact that the limiting distribution of the maximum likelihood estimator of the matrix B depends on the long run variance matrix Ω .

The Stock and Watson (1988), Phillips (1991) and Johansen (1988, 1991, 1994) approaches require consistent estimation of nuisance and/or structural parameters. In Bierens (1997a) I have proposed consistent cointegration tests that do not need specification of the data generating process, apart from some mild regularity conditions, or estimation of (nuisance) parameters. Thus these tests are completely nonparametric. My tests are conducted analogously to Johansen's tests, inclusive the test for parametric restrictions on the cointegrating vectors, namely on the basis of the ordered solutions of a generalized eigenvalue problem. Moreover, similarly to Johansen's approach one can consistently estimate a basis of the space of cointegrating vectors, using the eigenvectors of the generalized eigenvalue problem involved. However, the two matrices involved are constructed independently of the data generating process on the basis of weighted means of z_t and Δz_t , respectively, where the weights involved are Chebishev time polynomials [cf. Hamming (1973)] of even order.

1.3 *Contents*

In these lecture notes I will review some new developments in cointegration analysis, in particular Johansen's (1988, 1991, 1994) maximum likelihood approach on the basis of the VECM (1), and my nonparametric cointegration approach. First, in section 2, I will explain in more detail what cointegration is about, and in section 3 I will discuss (in an informal way) the Granger representation theorem that gives rise to the VECM specification (1). In sections 4 and 5 I will review Johansen's and my nonparametric approach, respectively. The main reason for focusing on Johansen's approach is that it is presently the most popular one in empirical

macroeconomic cointegration research, due to its own merits as well as the fact that Johansen's approach is now available in most time series oriented econometric software packages. Finally, in section 6 I will apply both the Johansen approach and my nonparametric approach to the Swedish data on the logs of consumption and income displayed in Figure 1.

2. *Introduction to cointegration*

Consider the q -variate unit root process $z_t = z_{t-1} + u_t$, where u_t is a zero mean stationary process, and let z_t be observable for $t = 0, 1, 2, \dots, n$. Due to the Wold decomposition theorem, we can write (under some mild regularity conditions), $u_t = C(L)v_t$, where v_t is a q -variate stationary white noise process with unit variance, i.e.,

$$E[v_t] = 0, E[v_t v_t^T] = I_q, E[v_t v_{t-j}^T] = O \text{ for } j \neq 0, \quad (3)$$

and $C(L)$ is a $q \times q$ matrix of lag series: $C(L) = \sum_{k=0}^{\infty} C_k L^k$, where L is the lag operator. Since by construction the lag polynomial $C(L) - C(1)$ is zero at $L = 1$, we can write

$$C(L)v_t = C(1)v_t + (C(L) - C(1))v_t = C(1)v_t + (1-L)D(L)v_t, \quad (4)$$

where

$$D(L) = \sum_{k=0}^{\infty} D_k L^k = (C(L) - C(1))/(1-L). \quad (5)$$

Denoting $w_t = D(L)v_t$ we now have $u_t = C(1)v_t + w_t - w_{t-1}$, hence

$$z_t = z_0 - w_0 + w_t + C(1)\sum_{j=1}^t v_j. \quad (6)$$

If v_t is a Gaussian process then by the white noise assumption (3) the v_t 's are i.i.d. $N_q(0, I_q)$. Since Johansen's approach is based on Gaussian maximum likelihood theory, this normality condition will be assumed. Moreover, we need regularity conditions that ensure that u_t and w_t are stationary processes. Therefore, it will be assumed that:

Assumption 1: *The process u_t can be written as $u_t = C(L)v_t$, where v_t is i.i.d. $N_q(0, I_q)$, $C(L) = C_1(L)^{-1}C_2(L)$, with $C_1(L)$ and $C_2(L)$ finite-order lag polynomials, and $\det(C_1(L))$ has all its roots outside the complex unit circle.*

Note that this condition on $C(L)$ implies that $\sum C_k$, $\sum C_k C_k^T$, $\sum D_k$, and $\sum D_k D_k^T$ converge, so that together with the normality condition, it follows that u_t and w_t are stationary Gaussian processes.

Cf. Engle and Yoo (1991). Moreover, Assumption 1 excludes the usual condition that also $\det(C_2(L))$ has roots all outside the unit circle. This is necessary because for cointegration we need to allow the matrix $C(1)$ to be singular.

As far as the nonparametric cointegration approach is concerned, Assumption 1 is more restrictive than necessary, but it will keep the argument below transparent, and focused on the main issues. See Phillips and Solo (1992) for weaker conditions in the case of linear processes. Also, in the nonparametric cointegration case we could assume instead of Assumption 1 that u_t is stationary and ergodic, so that we can write $u_t = \varepsilon_t + w_t - w_{t-1}$, where ε_t is a martingale difference process with variance matrix comparable with $C(1)C(1)^T$. Cf. Hall and Heyde (1980, p.136).

Now if $\text{rank}(C(1)) = q - r < q$ then the process z_t is cointegrated: there exist r linear independent cointegrating vectors β_j , $j = 1, \dots, r$, say, such that $\beta_j^T C(1) = 0^T$, hence it follows from (6) that $\beta_j^T z_t = \beta_j^T (z_0 - w_0) + \beta_j^T w_t$, $j = 1, \dots, r$. Thus the $\beta_j^T z_t$'s are now asymptotically stationary processes, in the sense that the stochastic intercept $\beta_j^T (z_0 - w_0)$ becomes independent of w_t if t approaches infinity, so that we then may condition on $z_0 - w_0$ and treat it as a constant.

The factorization (4) can be applied to the matrix $D(L)$ as well, so that similarly to (4), $C(L)v_t = C(1)v_t + D(1)(1-L)v_t + (1-L)^2 G(L)v_t$, where $G(L) = (D(L) - D(1))/(1-L)$. However, if there would exist a cointegrating vector β such that $\beta^T D(1) = 0^T$ then, with $\varepsilon_t = G(L)v_t$ a stationary process, we would have $\beta^T u_t = \Delta^2 \beta^T \varepsilon_t$, hence $\sum_{j=1}^r \beta_j^T z_t = \beta^T (z_0 - \varepsilon_0 + \varepsilon_{-1})t + \beta^T \varepsilon_t$ is trend stationary. As we will see in the next section, this would violate one of the conditions for the existence of an autoregressive error correction representation of a cointegrated system. Also, we need to exclude this case for the nonparametric cointegration approach. Therefore I assume:

Assumption 2: Let R_r be the matrix of eigenvectors of $C(1)C(1)^T$ corresponding to the r zero eigenvalues. Then the matrix $R_r^T D(1) D(1)^T R_r$ is nonsingular.

3. The error correction form of a cointegrated system

Following the approach of Engle and Yoo (1991) I will show now that under some regularity conditions a cointegrated process can be modelled as an VECM of the type (1). This result is due to Granger. Cf. Engle and Granger (1987). For convenience the discussion will be

confined to the bivariate case ($q = 2$) with one cointegrating vector.

Let β be the cointegrating vector. Without loss of generality we may normalize $\beta = (1, \beta_2)^T$. Consider the matrices

$$\begin{aligned}\Phi &= \begin{pmatrix} 1 & \beta_2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \beta^T \\ \varphi_2^T \end{pmatrix}, \quad V^{-1}(L) = \begin{pmatrix} \beta^T D(L) \\ \varphi_2^T C(L) \end{pmatrix}, \\ M(L) &= \begin{pmatrix} 1-L & 0 \\ 0 & 1 \end{pmatrix}, \quad M^*(L) = \begin{pmatrix} 1 & 0 \\ 0 & 1-L \end{pmatrix}.\end{aligned}\tag{7}$$

Then

$$\Phi \Delta z_t = \begin{pmatrix} (1-L)\beta^T D(L) \\ \varphi_2^T C(L) \end{pmatrix} v_t = M(L)V^{-1}(L)v_t.\tag{8}$$

and $M^*(L)M(L) = (1-L)I_2$. Next, assume that $V^{-1}(L)$ is invertible with inverse $V(L)$. This assumption is related to Assumption 2: if Assumption 2 does not hold then $V^{-1}(1)$ is singular so that $V^{-1}(L)$ is not invertible. Furthermore, denote $A(L) = V(L)M^*(L)\Phi$. Then $(1-L)A(L)z_t = (1-L)v_t$, which yields the AR form of the model:

$$A(L)z_t = \mu_0 + v_t,\tag{9}$$

where $\mu_0 = A(L)z_0 - v_0$. Now observe that

$$A(1) = V(1)M^*(1)\Phi = V(1)\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & \beta_2 \\ 0 & 1 \end{pmatrix} = V(1)\begin{pmatrix} 1 & \beta_2 \\ 0 & 0 \end{pmatrix} = \gamma_1 \beta^T,\tag{10}$$

where γ_1 is the first column of $V(1)$. Moreover, similarly to (4) we can write

$$A(L) = A(1)L + (1-L)B(L).\tag{11}$$

Combining (9), (10) and (11) we get the VECM $B(L)\Delta z_t = \mu_0 - \gamma_1 \beta^T z_{t-1} + v_t$. Finally, assume that $B(0)$ is invertible and that $B(L)$ is a $(p-1)$ -order lag polynomial, so that we may write

$\Pi(L) = B(0)^{-1}B(L) = I - \Pi_1 L - \dots - \Pi_{p-1} L^{p-1}$. Denoting $\pi_0 = B(0)^{-1}\mu_0$, $\alpha = -B(0)^{-1}\gamma_1$, and $e_t = B(0)^{-1}v_t$, we get the VECM

$$\Delta z_t = \pi_0 + \sum_{j=1}^{p-1} \Pi_j \Delta z_{t-j} + \alpha \beta^T z_{t-1} + e_t.\tag{12}$$

Note that the lag of the level variable z_{t-1} does not matter. Without loss of generality we may replace $\Pi(L)$ by the lag polynomial $\Pi_*(L) = \Pi(L) - \sum_{j=1}^{p-1} \alpha \beta^T L^j$, which yields an VECM of the form (1) with $d_t = 1$.

4. Johansen's maximum likelihood approach

4.1 Introduction

Johansen's cointegration testing approach is based on maximum likelihood estimation and likelihood ratio testing of the VECM (1), by step-wise concentrating the parameters out (i.e., maximizing the likelihood function over a subset of parameters, treating the other parameters as known), given the number r of cointegrating vectors, where the matrix β is the last to be concentrated out. Denoting the concentrated likelihood, given r and β , by $\hat{L}(r, \beta)$, and the maximum likelihood estimator of β given r by $\hat{\beta}_r$, where β and its maximum likelihood estimate are interpreted as zero vectors if $r = 0$, Johansen proposes two tests for the number of cointegrated vectors, namely the likelihood ratio test $-2 \ln(\hat{L}(r, \hat{\beta}_r)/\hat{L}(r+1, \hat{\beta}_{r+1}))$ of the null hypothesis that there are r cointegrated vectors (for $r = 0, 1, \dots, q-1$) against the alternative that there are $r+1$ cointegrating vectors, and the likelihood ratio test $-2 \ln(\hat{L}(r, \hat{\beta}_r)/\hat{L}(q, \hat{\beta}_q))$ of the same null hypothesis against the alternative that there are q cointegrated vectors. The latter alternative corresponds to the case where β is square and of full rank, which in its turn corresponds to the case that z_t is stationary rather than a multivariate unit root process. Since the usual regularity condition for maximum likelihood estimation do not apply in this case, the likelihood ratio tests involved have nonstandard limiting null distributions. Moreover, given the number r of cointegrating vectors, Johansen also proposes a likelihood ratio test of parametric restrictions on β of the form $\beta = H\varphi$, where H is a given $q \times s$ matrix of rank $s \leq r$ and φ is an unrestricted $s \times r$ matrix. For example, in the case $r = 1, q = 2$, one might wish to test whether β^T is proportional to $(1, -1) = H^T$. The likelihood ratio test statistic

$$-2 \ln[\sup_{\varphi} \hat{L}(r, H\varphi)/\hat{L}(r, \hat{\beta}_r)] \quad (13)$$

involved has a limiting χ^2 null distribution with $r(q-s)$ degrees of freedom.

4.2 The lambda-max and trace tests

I will now illustrate how Johansen's cointegration tests are conducted for the case where the data generating process is a Gaussian VECM of the form (1) with $d_t = 1$ and $p = 2$, where z_t is observable for $t = -1, 0, \dots, n$:

$$\Delta z_t = \pi_0 + \Pi_1 \Delta z_{t-1} + \alpha \beta^T z_{t-2} + e_t, \quad e_t \sim i.i.d. N_q(0, \Sigma). \quad (14)$$

Given β , α and Σ , the maximum likelihood estimates of π_0 and Π_1 can be obtained simply by regressing $\Delta z_t - \alpha \beta^T z_{t-2}$ on an intercept 1 and Δz_{t-1} , using OLS. The residuals of this regression are $\hat{R}_{1,t} - \alpha \beta^T \hat{R}_{2,t}$, where $\hat{R}_{1,t}$ is the residual of the regression of Δz_t on 1 and Δz_{t-1} , and $\hat{R}_{2,t}$ is the residual of the regression of z_{t-2} on 1 and Δz_{t-1} . Now the log-likelihood function with π_0 and Π_1 concentrated out is of the form

$$- 0.5n \ln(\det \Sigma) - 0.5 \sum_{t=1}^n (\hat{R}_{1,t} - \alpha \beta^T \hat{R}_{2,t})^T \Sigma^{-1} (\hat{R}_{1,t} - \alpha \beta^T \hat{R}_{2,t}) + \text{rest}, \quad (15)$$

where "rest" stands for the terms that do not depend on parameters. Similarly, we can concentrate α out, given β and Σ , by regressing $\hat{R}_{1,t}$ on $\beta^T \hat{R}_{2,t}$, which yields the estimate

$$\hat{\alpha}(\beta) = \hat{S}_{1,2} \beta [\beta^T \hat{S}_{2,2} \beta]^{-1}, \quad (16)$$

where $\hat{S}_{i,j} = (1/n) \sum_{t=1}^n \hat{R}_{i,t} \hat{R}_{j,t}^T$, $i, j = 1, 2$. Next, concentrate Σ out, given β , by substituting the well-known maximum likelihood estimator of the variance matrix of a normal distribution with zero mean vector:

$$\hat{\Sigma}(\beta) = (1/n) \sum_{t=1}^n (\hat{R}_{1,t} - \hat{\alpha}(\beta) \beta^T \hat{R}_{2,t}) (\hat{R}_{1,t} - \hat{\alpha}(\beta) \beta^T \hat{R}_{2,t})^T = \hat{S}_{1,1} - \hat{S}_{1,2} \beta [\beta^T \hat{S}_{2,2} \beta]^{-1} \beta^T \hat{S}_{2,1}. \quad (17)$$

Thus, the concentrated log-likelihood now becomes

$$\ln(\hat{L}(r, \beta)) = -0.5n \ln(\det \hat{\Sigma}(\beta)) + \text{rest}, \quad (18)$$

hence the maximum likelihood estimator of β is found by solving the minimization problem

$$\min \det(\hat{S}_{1,1} - \hat{S}_{1,2} \beta [\beta^T \hat{S}_{2,2} \beta]^{-1} \beta^T \hat{S}_{2,1}), \quad (19)$$

where the minimum is taken over all $q \times r$ matrices β . Using the matrix equalities

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} = \begin{pmatrix} A & O \\ B^T & I \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ O & C - B^T A^{-1}B \end{pmatrix} = \begin{pmatrix} I & B \\ O & C \end{pmatrix} \begin{pmatrix} A - BC^{-1}B^T & O \\ C^{-1}B^T & I \end{pmatrix},$$

where A and C are nonsingular square matrices, it is a standard exercise to verify that the minimization problem (19) is equivalent to

$$\min \det(\beta^T \hat{S}_{2,2} \beta - \beta^T \hat{S}_{2,1} \hat{S}_{1,1}^{-1} \hat{S}_{1,2} \beta) \det(\hat{S}_{1,1}) / \det(\beta^T \hat{S}_{2,2} \beta). \quad (20)$$

Note that the solution involved is not unique, as we may freely multiply β by a conformable nonsingular matrix. It is now quite easy to recognize the minimization problem (20) as a generalized eigenvalue problem: let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_q$ be the ordered solutions of the generalized eigenvalue problem $\det(\lambda \hat{S}_{2,2} - \hat{S}_{2,1}\hat{S}_{1,1}^{-1}\hat{S}_{1,2}) = 0$, let $\hat{B} = (\hat{b}_1, \dots, \hat{b}_q)$ be the matrix of corresponding eigenvectors, normalized such that $\hat{B}^T \hat{S}_{2,2} \hat{B} = I_q$, and choose $\beta = \hat{B}\xi$, where ξ is a $q \times r$ matrix normalized such that $\xi^T \xi = I_r$. Then the minimization problem (20) becomes

$$\begin{aligned}
& \min_{\substack{\xi^T \xi = I_r}} \det\left(\xi^T \hat{B}^T \hat{S}_{2,2} \hat{B} \xi - \xi^T \hat{B}^T \hat{S}_{2,1} \hat{S}_{1,1}^{-1} \hat{S}_{1,2} \hat{B} \xi\right) \det(\hat{S}_{1,1}) / \det(\xi^T \xi) \\
&= \min_{\substack{\xi^T \xi = I_r}} \det\left(\xi^T \xi - \xi^T \hat{B}^T \hat{S}_{2,1} \hat{S}_{1,1}^{-1} \hat{S}_{1,2} \hat{B} \xi\right) \det(\hat{S}_{1,1}) / \det(\xi^T \xi) \\
&= \min_{\substack{\xi^T \xi = I_r}} \det\left(I_r - \xi^T \hat{B}^T \hat{S}_{2,1} \hat{S}_{1,1}^{-1} \hat{S}_{1,2} \hat{B} \xi\right) \det(\hat{S}_{1,1}) / \det(I_r) \\
&= \min_{\substack{\xi^T \xi = I_r}} \det\left(I_r - \xi^T \hat{\Lambda} \xi\right) \det(\hat{S}_{1,1})
\end{aligned} \tag{21}$$

where $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_q)$. Clearly, the solution of (21) is $\xi^T = (I_r, O)$, hence the maximum likelihood estimator $\hat{\beta}_r$ of β , given the number r of cointegrating vectors, is equal to the matrix of the first r columns of \hat{B} : $\hat{\beta}_r = (\hat{b}_1, \dots, \hat{b}_r)$. Moreover, $\det[\hat{\Sigma}(\hat{\beta}_r)] = \det(I_r - \hat{\Lambda}_r) \det(\hat{S}_{1,1})$, where $\hat{\Lambda}_r = \text{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_r)$, so that the maximum log-likelihood given r becomes:

$$\ln(\hat{L}(r, \hat{\beta}_r)) = -0.5n \sum_{i=1}^r \ln(1 - \hat{\lambda}_i) - .5n \ln[\det(\hat{S}_{1,1})] + \text{rest.} \tag{22}$$

Thus, the likelihood ratio test $-2 \ln(\hat{L}(r, \hat{\beta}_r) / \hat{L}(r+1, \hat{\beta}_{r+1}))$ of the null hypothesis that there are r cointegrated vectors against the alternative that there are $r+1$ cointegrating vectors becomes

$-n \ln(1 - \hat{\lambda}_{r+1}) \approx n \hat{\lambda}_{r+1}$, and the likelihood ratio test $-2 \ln(\hat{L}(r, \hat{\beta}_r) / \hat{L}(q, \hat{\beta}_q))$ of the same null hypothesis against the alternative that there are q cointegrated vectors becomes

$-n \sum_{i=r+1}^q \ln(1 - \hat{\lambda}_i) \approx n \sum_{i=r+1}^q \hat{\lambda}_i$. Johansen (1988, 1991) proves that under the null of r cointegrating vectors, $(\hat{\lambda}_1, \dots, \hat{\lambda}_r)^T$ converges in probability to a vector of constants between zero and one, and $n(\hat{\lambda}_{r+1}, \dots, \hat{\lambda}_q)^T$ converges in distribution to the vector of ordered eigenvalues $\lambda_1 \geq \dots \geq \lambda_{q-r}$ of a stochastic a.s. positive definite $(q-r) \times (q-r)$ matrix which components are functionals of a $q-r$ -variate standard Brownian motion. Therefore, the likelihood ratio test

$-2 \ln(\hat{L}(r, \hat{\beta}_r) / \hat{L}(r+1, \hat{\beta}_{r+1})) \approx n \hat{\lambda}_{r+1}$ is called the *lambda-max* test, and the likelihood ratio test

$-2 \ln(\hat{L}(r, \hat{\beta}_r) / \hat{L}(q, \hat{\beta}_q)) \approx n \sum_{i=r+1}^q \hat{\lambda}_i$ is called the *trace test*.

4.3 Testing parametric restrictions on the cointegrating vectors

Similarly to (22) it can be shown that under the null hypothesis $\beta = H\phi$, where H is a given $q \times s$ matrix of rank $s \leq r$ with r given, and ϕ an unrestricted $s \times r$ matrix, the log-likelihood is $-.5n \sum_{i=1}^s \ln[1 - \tilde{\lambda}_i] - .5n \ln[\det(H^T \hat{S}_{1,1} H)] + \text{rest}$, where $\tilde{\lambda}_1 \geq \dots \geq \tilde{\lambda}_s$ are the solutions of the generalized eigenvalue problem $\det(\lambda H^T \hat{S}_{2,2} H - H^T \hat{S}_{2,1} \hat{S}_{1,1}^{-1} \hat{S}_{1,2} H) = 0$ and the rest term is the same as in (22). Thus, the likelihood ratio test statistic involved is:

$$-2\ln(LR) = n \sum_{i=1}^s \ln(1 - \tilde{\lambda}_i) - n \sum_{i=1}^r \ln(1 - \hat{\lambda}_i) + n \ln[\det(H^T \hat{S}_{1,1} H)] - n \ln[\det(\hat{S}_{1,1})].$$

Johansen (1988, 1991) proved that this likelihood ratio test has a χ^2 null distribution with $r(q-s)$ degrees of freedom.

4.4 Cointegrating restrictions on the intercept parameters

The null distributions of the lambda-max and trace tests in the above case depends on whether the vector π_0 of intercept parameters in model (14) can be written as

$$\pi_0 = \alpha\delta, \quad \text{with } \delta \in \mathbb{R}^r, \quad (23)$$

or not. If so, the VECM (14) becomes $\Delta z_t = \Pi_1 \Delta z_{t-1} + \alpha[\delta + \beta^T z_{t-2}] + e_t$.

Proposition 1: Under the restriction (23), $\delta + \beta^T z_{t-2}$ is a zero-mean stationary process, hence Δz_t is then a zero-mean stationary process, so that z_t itself is a multivariate unit root process **without drift**.

Proof: Recall that the general VECM

$$\Delta z_t = \pi_0 + \sum_{j=1}^{p-1} \Pi_j \Delta z_{t-j} + \alpha \beta^T z_{t-p} + e_t \quad (24)$$

is derived from

$$\Delta z_t = \mu + C(L)v_t = \mu + C(1)v_t + w_t - w_{t-1}, \quad (25)$$

where v_t is a bivariate zero-mean white noise process with unit variance matrix, $w_t = D(L)v_t$ with $D(L) = (1-L)^{-1}(C(L)-C(1))$, and μ is the non-zero vector of drift parameters. Moreover, $\beta^T C(1) = 0^T$ and $\beta^T D(1) \neq 0^T$. As to the latter, I will impose the stronger condition that

$$\det(D(1)) \neq 0. \quad (26)$$

Backwards substitution of (25) yields

$$z_t = \mu \cdot t + C(1) \sum_{j=1}^t v_j + w_t + z_0 - w_0, \quad (27)$$

hence

$$\beta^T z_t = \beta^T \mu \cdot t + \beta^T w_t + \beta^T (z_0 - w_0). \quad (28)$$

The last term in (28) acts as constant, and because $\beta^T z_t$ is stationary around a constant we must have that

$$\beta^T \mu = 0. \quad (29)$$

Thus,

$$\beta^T z_{t-p} = \beta^T w_{t-p} + \beta^T (z_0 - w_0) = \beta^T L^p D(L) v_t + c, \text{ where } c = \beta^T (z_0 - w_0). \quad (30)$$

Substituting (25) and (30) in (24) yields:

$$\left(I - \sum_{j=1}^{p-1} \Pi_j \right) \mu + \left(I - \sum_{j=1}^{p-1} \Pi_j L^j \right) C(L) v_t = \pi_0 + \alpha \beta^T L^p D(L) v_t + \alpha c + e_t. \quad (31)$$

This result implies that

$$e_t = C(0) v_t \quad (32)$$

(why?), where

$$\det(C(0)) \neq 0, \quad (33)$$

$$\mu = \left(I - \sum_{j=1}^{p-1} \Pi_j \right)^{-1} (\pi_0 + \alpha c) \quad (34)$$

and

$$\left(I - \sum_{j=1}^{p-1} \Pi_j L^j \right) C(L) = \alpha \beta^T L^p D(L) + C(0) \quad (35)$$

The latter result implies that

$$C(1) = \left(I - \sum_{j=1}^{p-1} \Pi_j \right)^{-1} \alpha \beta^T D(1) + \left(I - \sum_{j=1}^{p-1} \Pi_j \right)^{-1} C(0) \quad (36)$$

hence

$$\beta^T C(1) = \beta^T \left(I - \sum_{j=1}^{p-1} \Pi_j \right)^{-1} \alpha \beta^T D(1) + \beta^T \left(I - \sum_{j=1}^{p-1} \Pi_j \right)^{-1} C(0) = 0 \quad (37)$$

and thus

$$\beta^T \left(I - \sum_{j=1}^{p-1} \Pi_j \right)^{-1} \alpha = \frac{-1}{\beta^T D(1) D(1)^T \beta} \beta^T \left(I - \sum_{j=1}^{p-1} \Pi_j \right)^{-1} C(0) D(1)^T \beta \quad (38)$$

Due to (26) and (33), the right-hand side of (38) is non-zero:

$$\beta^T \left(I - \sum_{j=1}^{p-1} \Pi_j \right)^{-1} \alpha \neq 0. \quad (39)$$

Now suppose that (23) is true. Then (34) becomes

$$\mu = (\delta + c) \left(I - \sum_{j=1}^{p-1} \Pi_j \right)^{-1} \alpha. \quad (40)$$

But it follows now from (29) that

$$0 = \beta^T \mu = (\delta + c) \beta^T \left(I - \sum_{j=1}^{p-1} \Pi_j \right)^{-1} \alpha, \quad (41)$$

which by (39) implies that $c = -\delta$. Substituting this solution in (40) yields $\mu = 0$. Therefore, imposing the cointegrating restriction (23) in VECM (24) removes the drift! Q.E.D.

Consequently, the restriction (23) should only be imposed if the time series involved seem to run parallel, without drift, like in Figure 2:

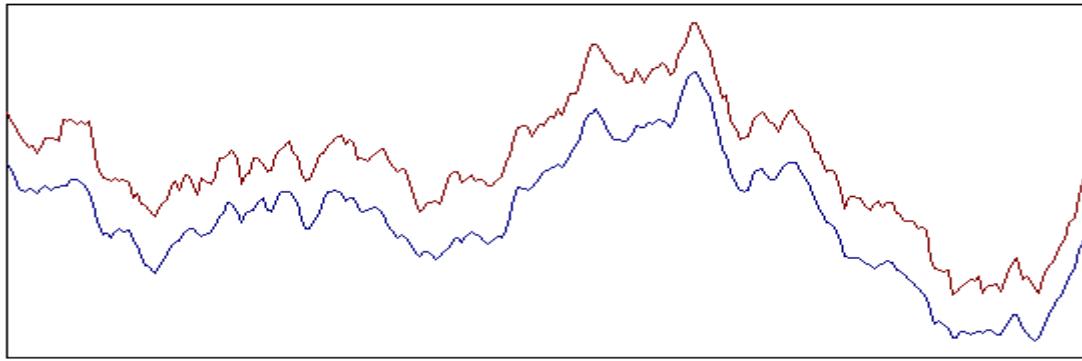


Figure 2 VECM with cointegrating restrictions on the intercept parameters

Without the restriction (23), the vector π_0 of intercept parameters produces common drift in the time series, like in Figure 3:

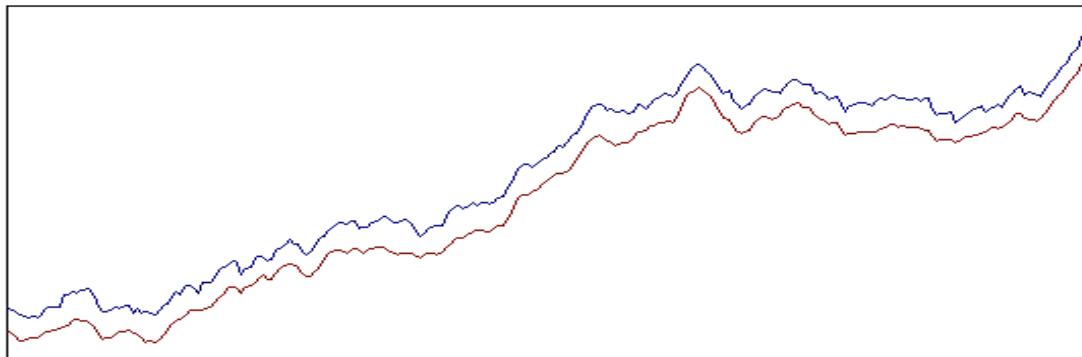


Figure 3 VECM without cointegrating restrictions on the intercept parameters

If we believe that the cointegrating restriction (23) on the intercept parameters holds, we can impose it as follows. First, concentrate Π_1 out, by regressing $\Delta z_t - \alpha(\delta + \beta^T z_{t-2})$ on Δz_{t-1} . The residuals of this regression are $\hat{R}_{1,t} - \alpha(\delta, \beta^T) \hat{R}_{2,t}$, where $\hat{R}_{1,t}$ is now the residual of the regression of Δz_t on Δz_{t-1} alone, and $\hat{R}_{2,t}$ is now the residual of the regression of $(1, z_{t-2})^T$ on Δz_{t-1} . Then proceed as before, with β replaced by $\beta_* = (\delta, \beta^T)^T$. Note that in this case the size of the matrix $\hat{S}_{2,2}$ is now $(q+1) \times (q+1)$, and the sizes of the matrices $\hat{S}_{1,2}$ and $\hat{S}_{2,1}$ are now $q \times (q+1)$ and $(q+1) \times q$, respectively. The limiting null distributions of the lambda-max and trace tests however are different from the ones before. Thus, there are three cases with different null distributions:

- (i) The cointegrating restrictions (23) on the intercept parameters do not hold and are not imposed;
- (ii) The cointegrating restrictions (23) on the intercept parameters hold but are not imposed;
- (iii) The cointegrating restrictions (23) on the intercept parameters hold and are imposed.

Recall that case (i) corresponds to Figure 3, and the cases (ii) and (iii) correspond to Figure 2.

The problem with testing parametric restrictions on the cointegrating vectors in case (iii) is that we cannot confine our attention to restrictions of the form $\beta = H\phi$ only, but that we have to include δ as well. Thus, we can only test restrictions of the form $\beta_* = (\delta, \beta^T)^T = H\phi$, where H is now a given $(q+1) \times s$ matrix with rank $s \leq r$, and ϕ is a conformable matrix of free parameters. However, the parameter vector δ is in general of no (economic) interest, so that one has to re-estimate the model without imposing the restriction (23) in order to test restrictions on β only.

4.5 Further extensions

Along the same lines as above one may include seasonal dummy variables in the VECM, provided they are taken in deviation from their sample means so that they become orthogonal to the intercept, without affecting the null distributions of the lambda-max and trace tests.

Moreover, recently Johansen (1994) considered also the case where a time trend is included in the VECM, i.e.,

$$\Delta z_t = \pi_{0,0} + \pi_{0,1} t + \sum_{j=1}^{p-1} \Pi_j \Delta z_{t-j} + \alpha \beta^T z_{t-p} + e_t. \quad (42)$$

In this case cointegrating restrictions on the trend parameters take the form

$$\pi_{0,1} = \alpha\gamma, \quad (43)$$

so that then

$$\Delta z_t = \pi_{0,0} + \sum_{j=1}^{p-1} \Pi_j \Delta z_{t-j} + \alpha[\gamma.t + \beta^T z_{t-p}] + e_t. \quad (44)$$

Similar to Proposition 1 it can be shown that under the conditions (43), $\gamma.t + \beta^T z_{t-p}$ is zero-mean stationary, hence $\beta^T z_t$ is trend stationary, and z_t is a multivariate unit root **with drift** process. In this case the time series pattern is as in Figure 4:

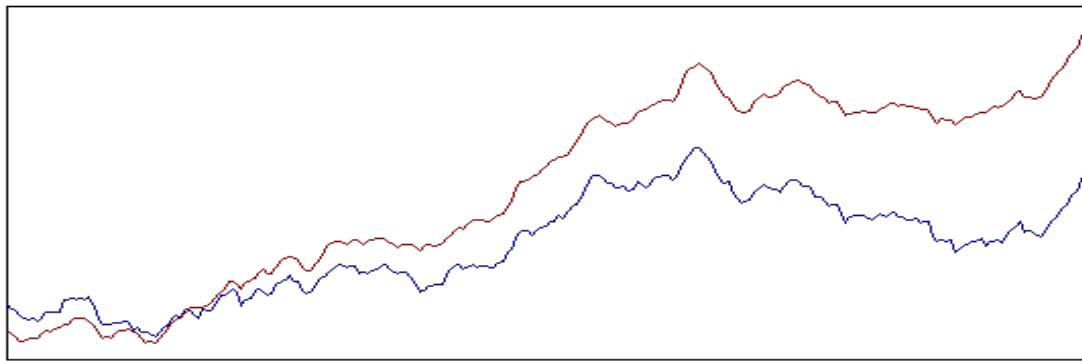


Figure 4 VECM with cointegrating restrictions on the trend parameters imposed

Without the restriction (43) z_t has linear drift and thus a quadratic time trend, which is unlikely in practice. Therefore, if the time series have drift and veer apart, VECM (44) is the appropriate model. Note that the time series in Figure 1 veer apart, which corresponds to VECM (44).

Again, the null distributions of the lambda-max and trace tests differ from the cases with an intercept only, and between the cases where cointegrating restrictions on the trend parameters are imposed or not.

5. Nonparametric cointegration analysis

5.1 Nonparametric tests of the number of cointegrating vectors

The basic ideas behind my nonparametric cointegration approach is that the difference in asymptotic behavior of certain weighted means of z_t and Δz_t under cointegration can be exploited to construct cointegration tests. In particular, these weighted means can be used to construct two random matrices such that cointegration tests can be based on their generalized eigenvalues,

similarly to Johansen's approach. I will only outline the main ideas; for the details and the proofs I refer to Bierens (1997a) and the separate appendix to that paper.

Denote the partial sums associated to z_t and Δz_t by $S_n^z(x) = 0$ if $x \in [0, n^{-1}]$, $S_n^z(x) = \sum_{t=1}^{[xn]} z_t$ if $x \in (n^{-1}, 1]$, and $S_n^{\Delta z}(x) = 0$ if $x \in [0, n^{-1}]$, $S_n^{\Delta z}(x) = \sum_{t=1}^{[xn]} \Delta z_t$ if $x \in (n^{-1}, 1]$, respectively. Then it is not hard to prove that under Assumption 1,

$$\begin{pmatrix} S_n^z(x)/(n\sqrt{n}) \\ S_n^{\Delta z}(x)/\sqrt{n} \end{pmatrix} \Rightarrow \begin{pmatrix} C(1) \int_0^x W(y) dy \\ C(1)W(x) \end{pmatrix}, \quad (45)$$

where $W(\cdot)$ is a q -variate standard Wiener process, and " \Rightarrow " means weak convergence. Cf. Billingsley (1968). The latter symbol will also be used to indicate convergence in distribution and convergence in probability, as these concepts are special cases of weak convergence.

Next, consider the following class of weighted means of z_t and Δz_t :

$$M_n^z(F) = \frac{1}{n} \sum_{t=1}^n F(t/n) z_t, \quad M_n^{\Delta z}(F) = \frac{1}{n} \sum_{t=1}^n F(t/n) \Delta z_t, \quad (46)$$

where F is a continuously differentiable function on the unit interval $[0, 1]$ with derivative f . Then it is pretty straightforward to verify from (45) and Lemma 9.6.3 in Bierens (1994, p.200) that

$$\begin{aligned} \begin{pmatrix} M_n^z(F)/\sqrt{n} \\ M_n^{\Delta z}(F)\sqrt{n} \end{pmatrix} &= F(1) \begin{pmatrix} S_n^z(1)/(n\sqrt{n}) \\ S_n^{\Delta z}(1)/\sqrt{n} \end{pmatrix} - \int f(x) \begin{pmatrix} S_n^z(x)/(n\sqrt{n}) \\ S_n^{\Delta z}(x)/\sqrt{n} \end{pmatrix} dx \Rightarrow \\ &\begin{pmatrix} C(1) \int f(x) W(x) dx \\ C(1)(F(1)W(1) - \int f(x) W(x) dx) \end{pmatrix} \sim N_{2q}(0, (C(1)C(1)^T) \otimes \Sigma_F), \end{aligned} \quad (47)$$

where

$$\Sigma_F = \begin{pmatrix} \int \int F(x) F(y) \min(x, y) dx dy & \frac{1}{2} \left(\int F(x) dx \right)^2 \\ \frac{1}{2} \left(\int F(x) dx \right)^2 & \int F(x)^2 dx \end{pmatrix}. \quad (48)$$

(The integrals in (47), Σ_F and below are taken over the unit interval, unless otherwise indicated). Note that if we choose F such that

$$\int F(x)dx = 0 \quad (49)$$

then Σ_F becomes a diagonal matrix, so that then the two components on the right-hand side of (47) are independent normally distributed:

Lemma 1: *Under Assumption 1 and condition (49),*

$$\begin{pmatrix} M_n^z(F)/\sqrt{n} \\ M_n^{\Delta z}(F)\sqrt{n} \end{pmatrix} \rightarrow \begin{pmatrix} C(1)X_F \sqrt{\int f(x)F(y)\min(x,y)dx dy} \\ C(1)Y_F \sqrt{\int f(x)^2 dx} \end{pmatrix}, \quad (50)$$

where X_F and Y_F are independent q -variate standard normally distributed random vectors depending on F in the following way:

$$X_F = \frac{\int F(x)W(x)dx}{\sqrt{\int f(x)F(y)\min(x,y)dx dy}}, \quad Y_F = \frac{F(1)W(1) - \int f(x)W(x)dx}{\sqrt{\int f(x)^2 dx}}. \quad (51)$$

Note that in the case of cointegration the matrix $C(1)C(1)^T$ is singular, so that the limiting normal distribution at the right-hand side of (50) is singular, hence for any cointegrating vector ξ we have $\xi^T M_n^z(F)/\sqrt{n} \Rightarrow 0$ and $\xi^T M_n^{\Delta z}(F)\sqrt{n} \Rightarrow 0$. This suggests that for cointegrating vectors ξ the rates of convergence of $\xi^T M_n^z(F)$ and $\xi^T M_n^{\Delta z}(F)$ will be different from the case in Lemma 1:

Lemma 2: *Let Assumption 1 and condition (49) hold. If z_t is cointegrated then for each matrix $\Xi = (\xi_1, \dots, \xi_r)$ of cointegrating vectors ξ_i ,*

$$\begin{pmatrix} \Xi^T M_n^z(F)\sqrt{n} \\ \Xi^T M_n^{\Delta z}(F)n \end{pmatrix} \Rightarrow \begin{pmatrix} \Xi^T D(1)Y_F \sqrt{\int f(x)^2 dx} \\ F(1)\Xi^T D_* Z \end{pmatrix}, \quad (52)$$

where Y_F and Z are independent q -variate standard normally distributed, with Y_F defined by (51) and $D_* = [\sum_{j=0}^n D_j D_j^T]^{1/2}$. [c.f. (5)].

Comparing Lemmas 1 and 2 we see that the asymptotic behavior, in particular the absolute and relative rates of convergence, of the statistics (46) differ substantially according to whether z_t is cointegrated or not. These differences can now be exploited in constructing

nonparametric cointegration tests, as follows.

Choose a sequence F_k , $k = 1, 2, \dots, m$, with $m \geq q$, of continuously differentiable real functions on $[0,1]$ with derivatives f_k satisfying condition (49), i.e., $\int F_k(x)dx = 0$ for $k = 1, \dots, m$, so that the random vectors

$$X_k = \frac{\int F_k(x)W(x)dx}{\sqrt{\int \int F_k(x)F_k(y)\min(x,y)dxdy}}, \quad Y_k = \frac{F_k(1)W(1) - \int f_k(x)W(x)dx}{\sqrt{\int f_k(x)^2dx}} \quad (53)$$

[cf. (51)] are mutually independent, together with conditions ensuring that these random vectors are also independent for $k = 1, 2, \dots$. Such functions F_k do exist. For example, let

$$F_k(x) = \cos(2k\pi x), \quad k = 1, 2, 3, \dots \quad (54)$$

Actually, this choice of F_k is "optimal" in the sense that it maximizes a lower bound of the power function of the nonparametric cointegration test.

Next, construct the random matrices $\hat{A}_m = \sum_{k=1}^m a_{n,k} a_{n,k}^T$ and $\hat{B}_m = \sum_{k=1}^m b_{n,k} b_{n,k}^T$, where

$$a_{n,k} = \frac{M_n^z(F_k(\cdot))}{\sqrt{\int \int F_k(x)F_k(y)\min(x,y)dxdy}}, \quad b_{n,k} = \frac{\sqrt{n} M_n^{\Delta z}(F_k(\cdot))}{\sqrt{\int f_k(x)^2dx}}. \quad (55)$$

Moreover, denote

$$\gamma_k = \frac{\sqrt{\int f_k(x)^2dx}}{\sqrt{\int \int F_k(x)F_k(y)\min(x,y)dxdy}}, \quad \delta_k = \frac{F_k(1)}{\sqrt{\int f_k(x)^2dx}}. \quad (56)$$

Then it follows from Lemmas 1-2:

Lemma 3: Let $\text{rank } C(1) = q - r$, let R_{q-r} be the matrix of orthonormal eigenvectors of $C(1)C(1)^T$ corresponding to the $q-r$ positive eigenvalues, let R_r be the matrix of orthonormal eigenvectors corresponding to the r zero eigenvalues, and denote $R = (R_{q-r}, R_r)$. Then under Assumption 1:

$$\begin{pmatrix} I_{q-r} & O \\ O & nI_r \end{pmatrix} R^T \hat{A}_m^T R \begin{pmatrix} I_{q-r} & O \\ O & nI_r \end{pmatrix} = \begin{pmatrix} R_{q-r}^T \hat{A}_m R_{q-r} & nR_{q-r}^T \hat{A}_m R_r \\ nR_r^T \hat{A}_m R_{q-r} & n^2 R_r^T \hat{A}_m R_r \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} R_{q-r}^T C(1) \sum_{k=1}^m X_k X_k^T C(1)^T R_{q-r} & R_{q-r}^T C(1) \sum_{k=1}^m \gamma_k X_k Y_k^T D(1)^T R_r \\ R_r^T D(1) \sum_{k=1}^m \gamma_k Y_k X_k^T C(1)^T R_{q-r} & R_r^T D(1) \sum_{k=1}^m \gamma_k^2 Y_k Y_k^T D(1)^T R_r \end{pmatrix} \quad (57)$$

and

$$\begin{aligned} \begin{pmatrix} I_{q-r} & O \\ O & \sqrt{n}I_r \end{pmatrix} R^T \hat{B}_m R \begin{pmatrix} I_{q-r} & O \\ O & \sqrt{n}I_r \end{pmatrix} &= \begin{pmatrix} R_{q-r}^T \hat{B}_m R_{q-r} & \sqrt{n}R_{q-r}^T \hat{B}_m R_r \\ \sqrt{n}R_r^T \hat{B}_m R_{q-r} & nR_r^T \hat{B}_m R_r \end{pmatrix} \\ \Rightarrow \begin{pmatrix} R_{q-r}^T C(1) \sum_{k=1}^m Y_k Y_k^T C(1)^T R_{q-r} & R_{q-r}^T C(1) \sum_{k=1}^m \delta_k Y_k Z^T D_*^T R_r \\ R_r^T D_* \sum_{k=1}^m \delta_k Z Y_k^T C(1)^T R_{q-r} & R_r^T D_* \sum_{k=1}^m \delta_k^2 Z Z^T D_*^T R_r \end{pmatrix}, \end{aligned} \quad (58)$$

where the X_i 's, the Y_j 's and Z are independent q -variate standard normally distributed.

Now at first sight one might think of employing these results for constructing cointegration tests by using the solutions of the generalized eigenvalue problem $\det(\hat{A}_m - \lambda \hat{B}_m) = 0$, similarly to Johansen's approach. However, the problem is that under cointegration both matrices converge in distribution to singular matrices. In deriving the limiting distribution of the generalized eigenvalues, Johansen used a result of Andersen, Brøns and Jensen (1983) saying that the ordered solutions of the generalized eigenvalue problem $\det(P_n - \lambda Q_n) = 0$, where P_n and Q_n are stochastic matrices converging jointly in distribution to P_* and Q_* , say, converge in distribution to the ordered solutions of the generalized eigenvalue problem $\det(P_* - \lambda Q_*) = 0$, provided Q_* is a.s. nonsingular. Due to the latter condition, this result cannot be used to derive the limiting distribution of the ordered solutions of the generalized eigenvalue problem $\det(\hat{A}_m - \lambda \hat{B}_m) = 0$. However, the following trick will cure the problem.

Observe that part (57) of Lemma 3 implies

$$\frac{R^T \hat{A}_m^{-1} R}{n^2} \Rightarrow \begin{pmatrix} O & O \\ O & V_{r,m}^{-1} \end{pmatrix}, \quad (59)$$

where

$$\begin{aligned}
V_{r,m} &= R_r^T D(1) \sum_{k=1}^m \gamma_k^2 Y_k Y_k^T D(1)^T R_r - \left(R_r^T D(1) \sum_{k=1}^m \gamma_k Y_k X_k^T C(1)^T R_{q-r} \right) \\
&\quad \times \left(R_{q-r}^T C(1) \sum_{k=1}^m X_k X_k^T C(1)^T R_{q-r} \right)^{-1} \left(R_{q-r}^T C(1) \sum_{k=1}^m \gamma_k X_k Y_k^T D(1)^T R_r \right). \tag{60}
\end{aligned}$$

Note that by Assumption 2, this matrix is a.s. nonsingular. Hence $R^T(\hat{B}_m + n^{-2}\hat{A}_m^{-1})R$ converges in distribution to a nonsingular block-diagonal matrix. Now using the result of Andersen, Brøns and Jensen (1983) it follows straightforwardly:

Theorem 1: Let $\lambda_{1,m} \geq \dots \geq \lambda_{q,m}$ be the ordered solutions of the generalized eigenvalue problem

$$\det(\hat{A}_m - \lambda(\hat{B}_m + n^{-2}\hat{A}_m^{-1})) = 0, \tag{61}$$

and let $\lambda_{1,m} \geq \dots \geq \lambda_{q-r,m}$ be the ordered solution of the generalized eigenvalue problem

$\det(\sum_{k=1}^m X_k^{*T} X_k^* - \lambda \sum_{k=1}^m Y_k^{*T} Y_k^*) = 0$, where the X_i^* 's and Y_j^* 's are i.i.d. $N_{q-r}(0, I_{q-r})$. If z_t is

cointegrated with r linear independent cointegrating vectors then under Assumptions 1-2,

$$(\lambda_{1,m}, \dots, \lambda_{q,m}) \Rightarrow (\lambda_{1,m}, \dots, \lambda_{q-r,m}, 0, \dots, 0).$$

This result suggests to use $\lambda_{q-r,m}$ as a test statistic for testing the null hypothesis that there are r cointegrating vectors against the alternative that there are $r+1$ cointegrating vectors. The test involved is a left-sided test: the null is rejected if $\lambda_{q-r,m}$ is smaller than a critical value. See Bierens (1997, Table 2) for the critical values involved.

The power of the test involved depends on the choice of m as well as on the choice of the functions F_k . As mentioned before, the choice (54) is "optimal" in that it maximizes a lower bound of the power function of the test. However, the asymptotically equivalent functions $F_k(x) = \cos(2k\pi(x - 0.5/n))$ will do an even better job because then the test becomes invariant for drift in the multivariate unit root process z_t , i.e. the case where $z_t = z_{t-1} + c + u_t$, where c is a vector of drift parameters and u_t is a zero mean stationary process satisfying Assumption 1. The matrices \hat{A}_m and \hat{B}_m then become

$$\hat{A}_m = \frac{8\pi^2}{n} \sum_{k=1}^m k^2 \left(\frac{1}{n} \sum_{t=1}^n \cos(2k\pi(t-0.5)/n) z_t \right) \left(\frac{1}{n} \sum_{t=1}^n \cos(2k\pi(t-0.5)/n) z_t \right)^T \tag{62}$$

$$\hat{B}_m = 2n \sum_{k=1}^m \left(\frac{1}{n} \sum_{t=1}^n \cos(2k\pi(t-0.5)/n) \Delta z_t \right) \left(\frac{1}{n} \sum_{t=1}^n \cos(2k\pi(t-0.5)/n) \Delta z_t \right)^T. \quad (63)$$

The same lower bound of the power function of the test mentioned before depends on m , hence maximizing this lower bound w.r.t. m would yield a sensible choice for m . The resulting values for m for significance levels $s \times 5\%$, $s = 1, 2$, and $0 \leq r \leq 4$, $1 \leq q \leq 5$, can be expressed by the formula:

$$m = q + I(q \geq s + 1)I(r = 0), \quad (64)$$

where $I(\cdot)$ is the indicator function.

5.2 Testing linear restrictions on the cointegrating vectors

Once the number r of cointegrating vectors is established, and $0 < r < q$, one may wish to verify whether there exist cointegrating vectors β satisfying the linear restriction

$H_0: \beta = H\varphi$, $\varphi \in \mathbb{R}^s$, where H is a given $q \times s$ matrix with full column rank $s \leq r$ and φ is arbitrary. Thus, the null hypothesis is that the space spanned by the columns of the matrix H is contained in the space of cointegrating vectors. For example, in the case $q = 3$ we may wish to test whether there exists a cointegrating vector $\beta = (\beta_1, \beta_2, \beta_3)^T$ such that $\beta_1 + \beta_2 = 0$ and $\beta_3 = 0$, so that then $H = (1, -1, 0)^T$.

At first sight one might think of mimicking Johansen's tests for these restrictions, on the basis of the matrices \hat{A}_m and $\hat{B}_m + n^{-2}\hat{A}_m^{-1}$. However, that leads to a case-dependent null distribution. Therefore I propose two test, the trace test and the lambda-max test, on the basis of the matrix \hat{A}_m only. The recipe for the lambda-max test is as follows. Choose $m = 2q$. The lambda-max test is based on the maximum solution, say $\lambda_{\max}(H)$, of the generalized eigenvalue problem

$$\det[H^T \hat{A}_m H - \lambda H^T (\hat{A}_m + n^{-2}\hat{A}_m^{-1})^{-1} H] = 0. \quad (65)$$

The test statistic involved is $n^2\lambda_{\max}(H)$, and we reject the null hypothesis if $n^2\lambda_{\max}(H)$ is larger than a critical value. See Bierens (1997, Table 4). The trace test statistic is n^2 times the sum of the solutions of (65), and the critical values involved are given in Bierens (1997, Table 3).

The choice of $m = 2q$ is somewhat heuristic: a lower bound of the power function is

monotonic increasing in m , but too large an m may mess up the size of the test. Since this lower bound of the power function is almost flat for $m > 2q$, I recommend the "rule of thumb" $m = 2q$.

5.3 Consistent estimation of a basis of the space of cointegrating vectors

Given that there are r linear independent (but unknown) cointegrating vectors ξ_1, \dots, ξ_r , one can consistently estimate a basis of the space of cointegrated vectors as follows. Choose again $m = 2q$, and let H be the matrix of eigenvectors corresponding to the r smallest eigenvalues of the generalized eigenvalue problem

$$\det[\hat{A}_m - \lambda(\hat{A}_m + n^{-2}\hat{A}_m^{-1})^{-1}] = 0, \quad (66)$$

where H is standardized such that

$$\hat{H}^T(\hat{A}_m + n^{-2}\hat{A}_m^{-1})^{-1}\hat{H} = I_r. \quad (67)$$

Then $\hat{H} = (\xi_1, \dots, \xi_r)\hat{\Gamma}_r + O_p(1/n)$, where $\hat{\Gamma}_r$ is $r \times r$ with $\text{rank}(\hat{\Gamma}_r) = r$. Since the cointegrating vectors ξ_1, \dots, ξ_r can be chosen orthonormal, we can interpret this result also in terms of projections: The distances between the columns of H and their corresponding projections on the space of cointegrating vectors vanish at order $O_p(1/n)$.

5.4 Seasonal drift

The above results apply to multivariate unit root processes with constant drift, but not to processes with seasonal drift. In the latter case one should replace z_t in the matrices A_m and B_m by seasonal moving averages $\bar{z}_t = (1/s)\sum_{\tau=0}^{s-1} z_{t-\tau}$, where s is the number of seasons. With this modification, the nonparametric approach is applicable to time series with seasonal drift.

5.5 Concluding remarks

My nonparametric cointegration approach has some clear advances over Johansen's maximum likelihood approach, in particular that it does not require to specify a lag length p and the deterministic variables d_t of the VECM (1), and that the critical values are case independent. This will become more clear in the empirical example in section 6. However, there is also a disadvantage, namely that the nonparametric tests are not invariant for scale: Replacing z_t by $z_t^* = Qz_t$, where Q is a nonsingular matrix, the generalized eigenvalue problem (61) becomes

$\det[\hat{A}_m - \lambda(\hat{B}_m + n^{-2}(Q^T Q)^{-1}\hat{A}_m^{-1}(Q^T Q)^{-1})] = 0$, and similarly the matrix \hat{A}_m^{-1} in (65), (66) and (67) changes accordingly to $(Q^T Q)^{-1}\hat{A}_m^{-1}(Q^T Q)^{-1}$. Of course, asymptotically this does not matter, but in small samples it clearly will. On the other hand, due to the fact that for $k = 1, 2, 3, \dots$,

$$\sum_{t=1}^n \cos(2\pi k(t - 0.5)/n) = 0, \quad \sum_{t=1}^n t \cos(2\pi k(t - 0.5)/n) = 0, \quad (68)$$

the tests are invariant for location shifts in Δz_t . Therefore, if all the variables in z_t are in logs the units of measurement of the original variables do not matter, due to (68), but one should be cautious in conducting the nonparametric tests to vector time series processes z_t if not all components are in logs.

6. An empirical example

I will now apply my nonparametric and Johansen's likelihood ratio cointegration tests to the annual data on the logs of consumption and income in Sweden from 1861 to 1988. However, before conducting cointegration analysis, one should test first whether the time series involved are unit root processes or not. From Figure 1 it is obvious that the appropriate hypotheses to be tested are the unit root *with drift* hypothesis against *trend* stationarity. Therefore, I have conducted the Augmented Dickey-Fuller (ADF) t-test of the null hypothesis $\alpha = 0$ in the auxiliary regression

$$\Delta y_t = \alpha y_{t-1} + \sum_{j=1}^p \beta_j \Delta y_{t-j} + \gamma_0 + \gamma_1 t + \varepsilon_t$$

with p depending on the sample size n [see Said and Dickey (1984)], and the Phillips-Perron (1988) test Z_α of the unit root with drift hypothesis against the trend stationarity hypothesis. The truncation lag p of the Newey-West (1987) estimator of the long run variance of Δy_t employed by the Phillips-Perron test, as well as the ADF lag length p have been chosen: $p = [5n^{1/4}] = 16$ for $n = 128$. The result is that for both time series the unit root with drift hypothesis can not be rejected at the 10% significance level. Also, I have conducted the Bierens-Guo (1993) tests of the trend stationarity hypothesis against the unit root with drift hypothesis, and for both time series the null hypothesis of linear trend stationarity is rejected at the 5% significance level.

The results of the nonparametric cointegration tests, conducted at the 10% significance level, are:

H_0	H_1	Test statistic	10% critical region	Conclusion
$r = 0$	$r = 1$	0.00005	(0, 0.005)	Reject H_0
$r = 1$	$r = 2$	24.33266	(0, 0.111)	Accept H_0

Thus the conclusion is that there is one cointegrating vector: $r = 1$. The estimate $\hat{\beta}$ of the parameter β in the cointegrating vector $(1, -\beta)^T$ is: $\hat{\beta} = 0.9444$, and the null hypothesis $\beta = 1$ is not rejected at the 10% significance level. The latter hypothesis corresponds to the hypothesis that the long run marginal propensity to consume from income equals 1.

Next, I have conducted Johansen's tests on the basis of VECM (1) for $p = 1, \dots, 6$, for the following five cases w.r.t the deterministic part $\Pi_0 d_t$:

1. $\Pi_0 d_t = \pi_0$, where π_0 is not proportional to α .
2. $\Pi_0 d_t = \pi_0$, where π_0 is proportional to α but this restriction is not imposed.
3. $\Pi_0 d_t = \pi_0$, where π_0 is proportional to α and this restriction is imposed.
4. $\Pi_0 d_t = \pi_0 + \pi_1 t$, where π_1 is proportional to α but this restriction is not imposed.
5. $\Pi_0 d_t = \pi_0 + \pi_1 t$, where π_1 is proportional to α and this restriction is imposed.

In view of Figure 1, the options 1, 2 and 3 are not suitable because they imply that the time series run parallel, whereas the two time series involved veer apart. Therefore, only the options 4 or 5 are applicable. Nevertheless, to demonstrate the sensitivity of the Johansen approach for the specification of the deterministic part of the VECM, I will try all five options. In the cases 3 and 5 with test result $r = 1$ I have also tested whether the imposed cointegrating restriction holds. Moreover, for the cases with test result $r = 1$ I have tested the hypothesis that the cointegrating vector $(1, -\beta)^T$ is equal to $(1, -1)^T$, so that $\beta = 1$. All tests are conducted at the 10% significance level. The results are presented in Tables 1-3.

Table 1: Johansen's test results for the number r of cointegrating vectors

Case 1		Case 2		Case 3		Case 4		Case 5	
p	r	p	r	p	r	p	r	p	r
1	2	1	1	1	2	1	1	1	1
2	2	2	1	2	2	2	1	2	1
3	2	3	1	3	2	3	1	3	1
4	0 or 2	4	0 or 1	4	2	4	0	4	0
5	0	5	0	5	1	5	0	5	0
6	0 or 2	6	0	6	1	6	0	6	0

Table 2: $\hat{\beta}$ and test of $H_0: \beta = 1$ for $r = 1$

Case 2			Case 3			Case 4			Case 5		
p	$\hat{\beta}$	H_0									
1	0.9448	reject	5	0.9420	accept	1	0.9245	reject	1	0.9245	reject
2	0.9367	reject	6	0.9442	accept	2	0.9115	reject	2	0.9115	reject
3	0.9397	reject				3	0.9165	reject	3	0.9165	reject

Table 3: Test of $H_0: \pi_i = a_i \gamma$ for $r = 1$

Case 3		Case 5	
p	H_0	p	H_0
5	reject	1	reject
6	reject	2	accept
		3	reject

The results in Table 1 where r takes two possible values are due to the fact that the lambda-max and trace tests gave different test results. The test result $r = 2$ would imply that both series are stationary, but the unit root and trend stationarity tests conducted on the single series indicate that they are unit root processes. For case 3 with $p = 5$ and 6 the imposed cointegrating restriction on π_0 is rejected, so that the result $r = 1$ in the cases 2 and 3 should be ignored. In case 5 with $p = 1$ and 3 the cointegrating restriction on the trend parameter vector π_1 is rejected, which would

imply the presence of a linear time trend in the drift. Since this is implausible, because then the growth rates of consumption and income have a linear trend and therefore grow to infinity themselves, there is only one case with $r = 1$ left that make sense, namely case 5 with $p = 2$. For this case the estimated cointegrating vector is $(1, -\hat{\beta})^T$, where $\hat{\beta} = 0.9115$, and the hypothesis $\beta = 1$ is rejected. The latter result is probably more accurate than the corresponding result of the nonparametric test, because the nonparametric test of restrictions on the cointegrating vector seems less powerful than the corresponding Johansen test. See Bierens (1997a).

References

- Anderson, S.A., H.K. Brons and S.T. Jensen (1983): "Distribution of Eigenvalues in Multivariate Statistical Analysis" *Annals of Statistics* **11**, pp. 392-415.
- Bierens, H.J. (1994): *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*. Cambridge: Cambridge University Press.
- Bierens, H.J. (1997a): "Nonparametric Cointegration Analysis", *Journal of Econometrics* **77**, 379-404.
- Bierens, H.J. (1997b): "Cointegration Analysis", in C. Heij, J.M. Schumacher, B. Hanzon and C. Praagman (Eds.), *System Dynamics in Economic and Financial Models*, John Wiley, 217-246.
- Bierens, H.J. and S.Guo (1993): "Testing Stationarity and Trend Stationarity Against the Unit Root Hypothesis", *Econometric Reviews* **12**, pp. 1-32.
- Billingsley, P. (1968): *Convergence of Probability Measures*. New York: John Wiley.
- Boswijk, H.P. (1994): "Testing for an Unstable Root in Conditional and Structural Error Correction Models", *Journal of Econometrics* **63**, pp. 37-60.
- Boswijk, H.P. (1995): "Efficient Inference on Cointegrating Parameters in Structural Error Correction Models", *Journal of Econometrics* **69**, pp. 133-158.
- Dickey, D.A. and W.A. Fuller (1979), "Distribution of the Estimators for Auto-regressive Times Series with a Unit Root", *Journal of the American Statistical Association* **74**, pp. 427-431.
- Dickey, D.A. and W.A. Fuller (1981), "Likelihood Ratio Statistics for Auto-regressive Time Series with a Unit Root", *Econometrica* **49**, pp. 1057-1072.
- Engle, R.F. and C.W.J.Granger (1987), "Cointegration and Error Correction: Representation, Estimation, and Testing", *Econometrica* **55**, pp. 251-276.
- Engle, R.F. and S.B. Yoo (1987), "Forecasting and Testing in Cointegrated Systems", *Journal of Econometrics* **35**, pp. 143-159
- Engle, R.F. and S.B. Yoo (1991), "Cointegrated Economic Time Series: An Overview with New Results", in R.F. Engle and C.W.J. Granger (eds): *Long-Run Economic Relationships*, Oxford: Oxford University Press, pp. 237-266.
- Friedman, M. (1957), *A Theory of the Consumption Function*. Princeton: Princeton University Press.

- Fuller, W.A. (1976), *Introduction to Statistical Time Series*. New York: John Wiley.
- Granger, C.W.J. (1981), "Some Properties of Time Series and Their Use in Econometric Model Specification", *Journal of Econometrics* **16**, 121-130.
- Hamming, R.W. (1973), *Numerical Methods for Scientists and Engineers*. New York: Dover Publications.
- Hall, P. and C.C. Heyde (1980), *Martingale Limit Theory and Its Applications*. San Diego: Academic Press.
- Hansen, B.E. (1992), "Tests for Parameter Instability in Regressions with I(1) Processes", *Journal of Business and Economic Statistics* **10**, pp. 321-335.
- Johansen, S. (1988), "Statistical Analysis of Cointegrated Vectors", *Journal of Economic Dynamics and Control* **12**, pp. 231-254.
- Johansen, S. (1991), "Estimation and Hypothesis Testing of Cointegrated Vectors in Gaussian Vector Autoregressive Models", *Econometrica* **59**, pp. 1551-1580.
- Johansen, S. (1994), "The Role of the Constant and Linear Terms in Cointegration Analysis of Nonstationary Variables", *Econometric Reviews* **13**, pp. 205-229.
- Johansen, S. and K. Juselius (1990), "Maximum Likelihood Estimation and Inference on Cointegration: With Applications to the Demand for Money", *Oxford Bulletin of Economics and Statistics* **52**, pp. 169-210
- Krantz, O. and L. Nilson (1975), *Swedish National Product 1861-1970: New Aspects on Methods and Measurements*, Lund: G.W.K. Gleerup/Liber Läromedel.
- Melander, E., A. Vredin and A. Warne (1992), "Stochastic Trends and Economic Fluctuations in a Small Open Economy", *Journal of Applied Econometrics* **7**, pp. 369-394.
- Newey, W.K. and K.D. West (1987), "A Simple Positive Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica* **55**, pp. 703-708.
- Park, J.Y (1990), "Testing for Unit Root and Cointegration by Variable Addition", *Advances in Econometrics* **8**, pp. 107-133.
- Perron, P. (1988): "Trends and Random Walks in Macroeconomic Time Series: Further Evidence from a New Approach", *Journal of Economic Dynamics and Control* **12**, pp. 297-332.
- Perron, P. (1989): "The Great Crash, the Oil Price Shock and the Unit Root Hypothesis", *Econometrica* **57**, pp. 1361-1402.
- Perron, P. (1990): "Testing the Unit Root in a Time Series with a Changing Mean", *Journal of Business and Economic Statistics* **8**, pp. 153-162.
- Phillips, P.C.B. (1987), "Time Series Regression With Unit Roots", *Econometrica* **55**, pp. 277-302.
- Phillips, P.C.B. (1991), "Optimal Inference in Cointegrated Systems", *Econometrica* **59**, pp. 283-306.
- Phillips, P.C.B. and B.Hansen (1990), "Statistical Inference in Instrumental Variables Regression with I(1) Processes", *Review of Economic Studies* **57**, pp. 99-125
- Phillips, P.C.B. and S.Ouliaris (1990): "Asymptotic Properties of Residual Based Tests for Cointegration", *Econometrica* **58**, pp. 165-193.
- Phillips, P.C.B. and P.Perron (1988), "Testing for a Unit Roots in Time Series Regression", *Biometrika* **75**, pp. 335-346.

Phillips, P.C.B. and V.Solo (1992), "Asymptotics for Linear Processes", *Annals of Statistics* **20**, pp. 971-1001.

Said, S.E and D.A. Dickey (1984), "Testing for Unit Roots in Autoregressive-Moving Average of Unknown Order", *Biometrika* **71**, pp. 599-607.

Stock, J.H. and M.W. Watson (1988), "Testing for Common Trends", *Journal of the American Statistical Association* **83**, pp. 1097-1107.

Sims,C.A., J.H.Stock and M.W.Watson (1990), "Inference in Linear Time Series Models with Some Unit Roots", *Econometrica* **58**, pp. 113-144.

Weak Convergence to the Matrix Stochastic Integral $\int_0^1 B dB'$ in the Gaussian Case, with Application to Likelihood-Based Cointegration Analysis

Herman J. Bierens
Pennsylvania State University

April 22, 2010

Abstract

Phillips (1988) has set forth conditions on a k -variate time series process x_t such that, with $S_t = \sum_{j=1}^t x_j$, $\frac{1}{T} \sum_{t=1}^T S_{t-1} x'_t$ converges in distribution to the stochastic matrix $\Sigma^{1/2} \left(\int_0^1 B dB' \right) \Sigma^{1/2} + \Sigma'_1$, where B is a k -variate standard Brownian motion on the unit interval, $\Sigma = \lim_{T \rightarrow \infty} E \left[\frac{1}{T} S_T S'_T \right]$, and $\Sigma'_1 = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E[S_{t-1} x'_t]$. However, Phillips' derivation of this result is very complicated. Therefore I will give an alternative proof for the case that x_t is a covariance stationary Gaussian process, employing only fairly standard probability theory. The result will be used to explain Johansen's (1988, 1991, 1995) likelihood-based cointegration analysis, in vector error correction models without and with intercepts.

This is a lecture note rather than a research paper. Therefore, the results are not mine; they are all due to Clive Granger, Soren Johansen and Peter Phillips.

1 Introduction

Let $u_t \in \mathbb{R}^k$ be an i.i.d. white noise vector time series process with $E[u_t u'_t] = I_k$, and let $S_t = \sum_{j=1}^t u_j$, $t \geq 1$, $S_0 = 0$. Moreover, denote for $x \in [0, 1]$, $B_T(x) = \frac{1}{\sqrt{T}} S_{[x.T]}$. Under more general conditions Phillips (1988) has shown that the random matrix $M_T = \frac{1}{T} \sum_{t=1}^T u_t S'_{t-1}$ converges in distribution to a random matrix M represented by the integral $M = \int_0^1 (dB) B'$, where B is a k -variate standard Brownian motion on $[0, 1]$ (also called a k -variate standard Wiener process).

More generally, Phillips (1988) has set forth conditions on a k -variate time series process x_t such that, with $S_t = \sum_{j=1}^t x_j$, $\frac{1}{T} \sum_{t=1}^T S_{t-1} x'_t$ converges in distribution to the stochastic matrix $\Sigma^{1/2} \left(\int_0^1 B dB' \right) \Sigma^{1/2} + \Sigma'_1$, where again B is a k -variate standard Brownian motion on the unit interval, with $\Sigma = \lim_{T \rightarrow \infty} E \left[\frac{1}{T} S_T S'_T \right]$, which is known as the long-run variance of x_t , and $\Sigma'_1 = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E[S_{t-1} x'_t]$.

This result and its generalizations in Section 2 are especially important for cointegration theory. However, Phillips' derivation of this result is very complicated. Therefore, in this lecture note I will give an alternative proof for the case that x_t is a covariance stationary Gaussian process, employing only fairly standard probability theory. After discussing in Section 3 the cointegration phenomenon and the Granger representation theorem, the results will then be used in Section 4 to explain Johansen's (1988, 1991, 1995) likelihood-based cointegration analysis, in particular the asymptotic theory of Johansen's lambda-max and trace tests for the cointegrating rank.

2 Weak convergence to the matrix stochastic integral $\int_0^1 B dB'$

2.1 About the notation $\int_0^1 B dB'$

The reason for denoting the limiting distribution $M = \int_0^1 B dB'$ of $M_T = \frac{1}{T} \sum_{t=1}^T S_{t-1} u'_t$ as an integral is that we can write

$$M'_T = \frac{1}{T} \sum_{t=1}^T u_t S'_{t-1} = \frac{1}{T} \sum_{t=1}^T (S_t - S_{t-1}) S'_{t-1}$$

$$\begin{aligned}
&= \sum_{t=1}^T (B_T(t/T) - B_T((t-1)/T)) B_T((t-1)/T)' \\
&= \sum_{t=1}^T \int_{t-1}^t (B_T(\tau/T + 1/T) - B_T(\tau/T)) B_T(\tau/T)' d\tau \\
&= \int_0^T (B_T(\tau/T + 1/T) - B_T(\tau/T)) B_T(\tau/T)' d\tau \\
&= T \int_0^1 (B_T(x + 1/T) - B_T(x)) B_T(x)' dx \\
&= \int_0^1 \left(\frac{B_T(x + 1/T) - B_T(x)}{1/T} \right) B_T(x)' dx \\
&= \int_0^1 (dB_T(x)/dx) B_T(x)' dx = \int_0^1 (dB_T) B'_T,
\end{aligned}$$

say, where $dB_T(x)/dx$ is defined as

$$dB_T(x)/dx = \frac{B_T(x + 1/T) - B_T(x)}{1/T}.$$

Note, however, that

$$dB_T(x)/dx = \sqrt{T} (S_{[(x+1/T)T]} - S_{[x:T]}) = \sqrt{T} u_{[xT+1]},$$

which does not converge weakly. Therefore, we cannot conclude from the continuous mapping theorem that

$$M_T = \frac{1}{T} \sum_{t=1}^T u_t S'_{t-1} = \int_0^1 (dB_T) B'_T \xrightarrow{d} \int_0^1 (dB) B' = M. \quad (1)$$

Nevertheless, the convergence result (1) holds for some random matrix M .

In the case $k = 1$ the meaning of M follows from the results in Phillips (1986): Note that $S_t^2 = (u_t + S_{t-1})^2 = u_t^2 + 2u_t S_{t-1} + S_{t-1}^2$, so that

$$\begin{aligned}
M_T &= \frac{1}{T} \sum_{t=1}^T u_t S_{t-1} = \frac{1}{2T} \sum_{t=1}^T (S_t^2 - S_{t-1}^2 - u_t^2) \\
&= \frac{1}{2} \left(\left(S_T / \sqrt{T} \right)^2 - \frac{1}{T} \sum_{t=1}^T u_t^2 \right) \\
&= \frac{1}{2} (B_T(1)^2 - 1) - \frac{1}{T} \sum_{t=1}^T (u_t^2 - 1) \xrightarrow{d} \frac{1}{2} (B(1)^2 - 1).
\end{aligned} \quad (2)$$

The convergence result involved follows from the central limit theorem, i.e., $B_T(1) = S_T/\sqrt{T} = \left(1/\sqrt{T}\right) \sum_{t=1}^T u_t \xrightarrow{d} N(0, 1) \sim B(1)$, and the law of large numbers: $p \lim_{T \rightarrow \infty} (1/T) \sum_{t=1}^T (u_t^2 - 1) = 0$. Thus, in the case $k = 1$,

$$M = \int_0^1 (dB) B = \int_0^1 BdB = \frac{1}{2} (B(1)^2 - 1).$$

Along the same lines we find that in the multivariate case,

$$\begin{aligned} M_T + M'_T &= \frac{1}{T} \sum_{t=1}^T S_{t-1} u'_t + \frac{1}{T} \sum_{t=1}^T u_t S'_{t-1} \\ &= \frac{1}{T} S_T S'_T - \frac{1}{T} \sum_{t=1}^T u_t u'_t \\ &= B_T(1) B_T(1)' - I_k - \left(\frac{1}{T} \sum_{t=1}^T u_t u'_t - I_k \right) \\ &\xrightarrow{d} B(1) B(1)' - I_k = M + M'. \end{aligned} \tag{3}$$

2.2 The bivariate Gaussian white noise case

Consider the Gaussian white noise case $k = 2$, i.e., $x_t = u_t$, where

$$\begin{aligned} u_t &= \begin{pmatrix} u_{1,t} \\ u_{2,t} \end{pmatrix} \sim i.i.d. N_2 [0, I_2], \\ S_t &= \begin{pmatrix} S_{1,t} \\ S_{2,t} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^t u_{1,j} \\ \sum_{j=1}^t u_{2,j} \end{pmatrix}, \\ B_T(x) &= \begin{pmatrix} B_{1,T}(x) \\ B_{2,T}(x) \end{pmatrix} = \begin{pmatrix} S_{1,[xT]}/\sqrt{T} \\ S_{2,[xT]}/\sqrt{T} \end{pmatrix}. \end{aligned}$$

It follows from (2) that

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T u_{1,t} S_{1,t-1} &= \frac{1}{2} (B_{1,T}(1)^2 - 1) - \frac{1}{2} \frac{1}{T} \sum_{t=1}^T (u_{1,t}^2 - 1) \\ \frac{1}{T} \sum_{t=1}^T u_{2,t} S_{2,t-1} &= \frac{1}{2} (B_{2,T}(1)^2 - 1) - \frac{1}{2} \frac{1}{T} \sum_{t=1}^T (u_{2,t}^2 - 1) \end{aligned}$$

and

$$\begin{aligned}\frac{1}{T} \sum_{t=1}^T u_{2,t} S_{1,t-1} &= B_{1,T}(1) B_{2,T}(1) - \frac{1}{T} \sum_{t=1}^T u_{1,t} S_{2,t-1} - \frac{1}{T} \sum_{t=1}^T u_{1,t} u_{2,t} \\ &= \frac{1}{\sqrt{T}} \sum_{t=1}^T u_{1,t} \left(B_{2,T}(1) - \frac{S_{2,t-1}}{\sqrt{T}} \right) - \frac{1}{T} \sum_{t=1}^T u_{1,t} u_{2,t}\end{aligned}$$

Thus

$$\begin{aligned}\frac{1}{T} \sum_{t=1}^T u_t S'_{t-1} &= \begin{pmatrix} \frac{1}{T} \sum_{t=1}^T u_{1,t} S_{1,t-1} & \frac{1}{T} \sum_{t=1}^T u_{1,t} S_{2,t-1} \\ \frac{1}{T} \sum_{t=1}^T u_{2,t} S_{1,t-1} & \frac{1}{T} \sum_{t=1}^T u_{2,t} S_{2,t-1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2} (B_{t,T}(1)^2 - 1) & \frac{1}{T} \sum_{t=1}^T u_{1,t} S_{2,t-1} \\ B_{1,T}(1) B_{2,T}(1) - \frac{1}{T} \sum_{t=1}^T u_{1,t} S_{2,t-1} & \frac{1}{2} (B_{2,T}(1)^2 - 1) \end{pmatrix} \quad (4)\end{aligned}$$

$$-\frac{1}{2} \begin{pmatrix} \frac{1}{T} \sum_{t=1}^T (u_{1,t}^2 - 1) & 0 \\ 2 \frac{1}{T} \sum_{t=1}^T u_{1,t} u_{2,t} & \frac{1}{T} \sum_{t=1}^T (u_{2,t}^2 - 1) \end{pmatrix}. \quad (5)$$

The elements of the matrix (4) are functions of

$$Z_T = \begin{pmatrix} \frac{1}{\sqrt{T}} \sum_{t=1}^T u_{1,t} \\ \frac{1}{T} \sum_{t=1}^T u_{1,t} S_{2,t-1} \\ \frac{1}{\sqrt{T}} \sum_{t=1}^T u_{2,t} \end{pmatrix} \quad (6)$$

and the matrix (5) converges in probability to a zero matrix. Therefore, we only need to show that (6) converges in distribution, as follows. Write

$$Z_T = \begin{pmatrix} 0 \\ 0 \\ B_{2,T}(1) \end{pmatrix} + \frac{1}{\sqrt{T}} \sum_{t=1}^T u_{1,t} \begin{pmatrix} 1 \\ S_{2,t-1}/\sqrt{T} \\ 0 \end{pmatrix}.$$

Then conditionally on $u_{2,1}, \dots, u_{2,T}$, Z_T is normally distributed with expectation vector

$$\mu_T = \begin{pmatrix} 0 \\ 0 \\ B_{2,T}(1) \end{pmatrix}$$

and singular variance matrix

$$\Sigma_T = \begin{pmatrix} \Phi_{2,T} & 0 \\ 0' & 0 \end{pmatrix}$$

where

$$\Phi_{2,T} = \begin{pmatrix} 1 & \int_0^1 B_{2,T}(x)dx \\ \int_0^1 B_{2,T}(x)dx & \int_0^1 B_{2,T}(x)^2 dx \end{pmatrix}.$$

Therefore, the characteristic function of Z_T conditional on $u_{2,1}, \dots, u_{2,T}$ is

$$\begin{aligned} E[\exp(i.\xi' Z_T) | u_{2,1}, \dots, u_{2,T}] &= \exp(i.\xi' \mu_T) \exp\left(-\frac{1}{2}\xi' \Sigma_T \xi\right) \\ &= \exp(i.\xi_3 B_{2,T}(1)) \exp\left(-\frac{1}{2}(\xi_1, \xi_2) \Phi_{2,T} (\xi_1, \xi_2)'\right) \\ &= \exp\left(-\frac{1}{2}\xi_1^2\right) \exp\left(\xi_1 \xi_2 \int_0^1 B_{2,T}(x)dx - \frac{1}{2}\xi_2^2 \int_0^1 B_{2,T}(x)^2 dx\right) \\ &\quad \times \exp(i.\xi_3 B_{2,T}(1)), \end{aligned}$$

where $\xi' = (\xi_1, \xi_2, \xi_3)$. Because $B_{2,T}(1)$ and $\Phi_{2,T}$ converges jointly in distribution to $B_2(1)$ and

$$\Phi_2 = \begin{pmatrix} 1 & \int_0^1 B_2(x)dx \\ \int_0^1 B_2(x)dx & \int_0^1 B_2(x)^2 dx \end{pmatrix},$$

respectively, it follows from the continuous mapping theorem that pointwise in ξ ,

$$\begin{aligned} &\exp\left(-\frac{1}{2}\xi_1^2\right) \exp(i.\xi_3 B_{2,T}(1)) \\ &\times \exp\left(\xi_1 \xi_2 \int_0^1 B_{2,T}(x)dx - \frac{1}{2}\xi_2^2 \int_0^1 B_{2,T}(x)^2 dx\right) \\ &\xrightarrow{d} \exp\left(-\frac{1}{2}\xi_1^2\right) \exp(i.\xi_3 B_2(1)) \\ &\times \exp\left(\xi_1 \xi_2 \int_0^1 B_2(x)dx - \frac{1}{2}\xi_2^2 \int_0^1 B_2(x)^2 dx\right) \end{aligned}$$

This result implies that

$$\begin{aligned} \lim_{T \rightarrow \infty} E[\exp(i.\xi' Z_T)] &= \exp\left(-\frac{1}{2}\xi_1^2\right) \\ &\times E\left[\exp(i.\xi_3 B_2(1)) \exp\left(\xi_1 \xi_2 \int_0^1 B_2(x)dx - \frac{1}{2}\xi_2^2 \int_0^1 B_2(x)^2 dx\right)\right], \end{aligned} \tag{7}$$

because convergence in distribution of bounded random variables implies convergence of their expectations. Consequently, $Z_T \xrightarrow{d} Z = (Z_1, Z_2, Z_3)'$, where Z is a random vector with characteristic function (7). Hence,

$$\begin{aligned} M &= \int_0^1 (dB)B' = \begin{pmatrix} \frac{1}{2}(Z_1^2 - 1) & Z_2 \\ Z_1Z_3 - Z_2 & \frac{1}{2}(Z_3^2 - 1) \end{pmatrix}, \\ M' &= \int_0^1 BdB' = \begin{pmatrix} \frac{1}{2}(Z_1^2 - 1) & Z_1Z_3 - Z_2 \\ Z_2 & \frac{1}{2}(Z_3^2 - 1) \end{pmatrix}. \end{aligned}$$

2.3 The tri-variate Gaussian white noise case

Now suppose that $u_{2,t} \in \mathbb{R}^2$. Then

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T u_t S'_{t-1} &= \begin{pmatrix} \frac{1}{T} \sum_{t=1}^T u_{1,t} S_{1,t-1} & \frac{1}{T} \sum_{t=1}^T u_{1,t} S'_{2,t-1} \\ \frac{1}{T} \sum_{t=1}^T u_{2,t} S_{1,t-1} & \frac{1}{T} \sum_{t=1}^T u_{2,t} S'_{2,t-1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2}(B_{1,T}(1)^2 - 1) & \frac{1}{T} \sum_{t=1}^T u_{1,t} S'_{2,t-1} \\ B_{1,T}(1)B_{2,T}(1)' - \frac{1}{T} \sum_{t=1}^T u_{1,t} S'_{2,t-1} & \frac{1}{T} \sum_{t=1}^T u_{2,t} S'_{2,t-1} \end{pmatrix} \\ &\quad - \frac{1}{2} \begin{pmatrix} \frac{1}{T} \sum_{t=1}^T (u_{1,t}^2 - 1) & 0' \\ 2\frac{1}{T} \sum_{t=1}^T u_{1,t} u'_{2,t} & O \end{pmatrix}. \end{aligned}$$

Stack the four elements of $\frac{1}{T} \sum_{t=1}^T u_{2,t} S'_{2,t-1}$ in a vector ψ_T , and let

$$Z_T = \begin{pmatrix} B_{1,T}(1) \\ \frac{1}{T} \sum_{t=1}^T u_{1,t} S_{2,t-1} \\ \psi_T \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{T}} \sum_{t=1}^T u_{1,t} \begin{pmatrix} 1 \\ \frac{1}{\sqrt{T}} S_{2,t-1} \end{pmatrix} \\ \psi_T \end{pmatrix}$$

Then

$$\begin{aligned} E [\exp(i \cdot \xi' Z_T)] | u_{2,1}, \dots, u_{2,T} &= \exp(i \cdot (\xi'_3 \psi_T)) \\ &\times \exp \left(-\frac{1}{2} (\xi_1, \xi'_2) \begin{pmatrix} 1 & \int_0^1 B_{2,T}(x)' dx \\ \int_0^1 B_{2,T}(x) dx & \int_0^1 B_{2,T}(x) B_{2,T}(x)' dx \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi'_2 \end{pmatrix} \right), \end{aligned}$$

where $\xi = (\xi_1, \xi'_2, \xi'_3)'$ has been partitioned conformably with Z_T . But by the previous argument, all the random elements involved converge jointly in distribution, hence $E [\exp(i \cdot \xi' Z_T)]$ converges to a characteristic function.

The k -variate Gaussian case follows now by induction.

Summarizing, the following result has been shown.

Theorem 1. Let u_t be i.i.d. $N[0, I_k]$ and let $S_t = \sum_{j=1}^t u_j$. Then the random matrix $(1/T) \sum_{t=1}^T u_t S'_{t-1}$ converges in distribution to a random matrix, denoted by $\int_0^1 (dB) B'$, where $B(\cdot)$ is a k -variate standard Brownian motion.

2.4 The dependent case

The independence assumption is not essential for the above results, provided that some adjustments are made:

Assumption 1. Let

$$y_t = y_{t-1} + x_t, \quad (8)$$

where x_t is a k -variate zero mean covariance stationary Gaussian vector time series process with Wold decomposition $x_t = \sum_{m=0}^{\infty} C_m u_{t-m}$, with u_t i.i.d. $N_k(0, I_k)$. For $m \rightarrow \infty$ the elements $c_{i,j,m}$ of the $k \times k$ matrices C_m converge to zero at an exponential rate: there exists a $\rho \in (0, 1)$ and a constant $K \in (0, \infty)$ such that $\max_{1 \leq i \leq k, 1 \leq j \leq k} |c_{i,j,m}| < K\rho^m$.

Denote $C(L) = \sum_{m=0}^{\infty} C_m L^m$, with L the lag operator, so that

$$x_t = C(L)u_t. \quad (9)$$

We can always write

$$C(L) = C(1) + \left(\frac{C(L) - C(1)}{1 - L} \right) (1 - L) = C(1) + D(L) (1 - L), \quad (10)$$

say, because all the elements of $C(L) - C(1)$ have root 1, hence these elements are proportional to $1 - L$. This construction is known as the Beveridge-Nelson decomposition. Then,

$$x_t = C(1)u_t + D(L)(1 - L)u_t = C(1)u_t + v_t - v_{t-1}, \quad (11)$$

and

$$y_t = \sum_{j=1}^t x_j + y_0 = C(1) \sum_{j=1}^t u_j + v_t - v_0 + y_0, \quad (12)$$

where

$$v_t = D(L)u_t = \sum_{m=0}^{\infty} D_m u_{t-m}. \quad (13)$$

Assumption 1 implies that v_t is a zero-mean covariance stationary Gaussian process, and that for $m \rightarrow \infty$, $D_m \rightarrow O_{k,k}$ exponentially.

Let us derive first the limiting distribution of $\frac{1}{T} \sum_{t=1}^T u_t y'_{t-1}$. Using the decomposition (12), we have

$$\frac{1}{T} \sum_{t=1}^T u_t y'_{t-1} = \frac{1}{T} \sum_{t=1}^T u_t \sum_{j=1}^{t-1} u_j C(1)' + \frac{1}{T} \sum_{t=1}^T u_t v'_{t-1} + \frac{1}{T} \sum_{t=1}^T u_t (y_0 - v_0)'. \quad (14)$$

Since u_t and v_{t-1} are independent, it follows that $\frac{1}{T} \sum_{t=1}^T u_t v'_{t-1} = O_p\left(1/\sqrt{T}\right)$ and $\frac{1}{T} \sum_{t=1}^T u_t = O_p\left(1/\sqrt{T}\right)$. It follows therefore from Theorem 1 and (14) that

Theorem 2. Under Assumption 1, $\frac{1}{T} \sum_{t=1}^T u_t y'_{t-1} \xrightarrow{d} \int_0^1 (dB) B'C(1)'$.

Next, consider the case $\frac{1}{T} \sum_{t=1}^T x_t y'_{t-1}$. Note that by Assumption 1,

$$\Sigma_{XX'_m} = E[x_t x'_{t+m}] = \begin{cases} \sum_{j=0}^{\infty} C_j C'_{j+m} & \text{if } m \geq 0 \\ \sum_{j=0}^{\infty} C_{j-m} C'_j & \text{if } m < 0 \end{cases} \quad (15)$$

is finite, and that

Lemma 1. For $|m| \rightarrow \infty$, $\Sigma_{XX'_{-m}} \rightarrow O_{k,k}$ exponentially, and for fixed m , $p \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T x_t x'_{t+m} = \Sigma_{XX'_m}$.

The latter result is not hard to prove. See for example Bierens (2004, Ch. 7).

Since $\Sigma_{XX'_{-m}} \rightarrow O_{k,k}$ exponentially it follows that

$$\begin{aligned} \Sigma_{XY'} &= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E[x_t (y_{t-1} - y_0)'] \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{j=1}^{t-1} E[x_t x'_{t-j}] = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{j=1}^{t-1} \Sigma'_{XX'_j} \end{aligned} \quad (16)$$

$$= \sum_{j=1}^{\infty} \Sigma'_{XX'_j}$$

exists and is finite. This matrix will play a role in the generalization of Theorem 2.

Similar to Lemma 1 it follows that

$$p \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T v_t v'_{t+m} = \Sigma_{VV'_m} \quad (17)$$

$$p \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T u_t v'_{t+m} = \Sigma_{UV'_m} \quad (18)$$

where

$$\begin{aligned} \Sigma_{VV'_m} &= E[v_t v'_{t+m}] = \begin{cases} \sum_{j=0}^{\infty} D_j D'_{j+m} & \text{if } m \geq 0 \\ \sum_{j=0}^{\infty} D_{j-m} D'_j & \text{if } m < 0 \end{cases} \\ \Sigma_{UV'_m} &= E[u_t v'_{t+m}] = \begin{cases} D'_m & \text{if } m \geq 0 \\ O_{k,k} & \text{if } m < 0 \end{cases} \end{aligned}$$

We can now write

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T x_t y'_{t-1} &= \frac{1}{T} \sum_{t=1}^T x_t \left(\sum_{j=1}^{t-1} u'_j C(1)' + v_{t-1} \right) + \frac{1}{T} \sum_{t=1}^T x_t (y_0 - v_0)' \\ &= \frac{1}{T} \sum_{t=1}^T (C(1)u_t + v_t - v_{t-1}) \left(\sum_{j=1}^{t-1} u'_j C(1)' + v_{t-1} \right) + o_p(1) \\ &= C(1) \left(\frac{1}{T} \sum_{t=1}^T u_t \sum_{j=1}^{t-1} u'_j \right) C(1)' + \frac{1}{T} \sum_{t=1}^T (v_t - v_{t-1}) \sum_{j=1}^{t-1} u'_j C(1)' \\ &\quad + \frac{1}{T} \sum_{t=1}^T v_t v'_{t-1} - \frac{1}{T} \sum_{t=1}^T v_{t-1} v'_{t-1} + C(1) \frac{1}{T} \sum_{t=1}^T u_t v'_{t-1} + o_p(1) \\ &= C(1) \left(\frac{1}{T} \sum_{t=1}^T u_t \sum_{j=1}^{t-1} u'_j \right) C(1)' + \Sigma_{VV'_{-1}} - \Sigma_{VV'_0} - \Sigma'_{UV'_0} + o_p(1) \quad (19) \end{aligned}$$

The last $o_p(1)$ term follows from (17), (18) and

$$\frac{1}{T} \sum_{t=1}^T (v_t - v_{t-1}) \sum_{j=1}^{t-1} u'_j$$

$$\begin{aligned}
&= \frac{1}{T} \sum_{t=1}^T v_t \sum_{j=1}^{t-1} u'_j - \frac{1}{T} \sum_{t=1}^T v_{t-1} \sum_{j=1}^{t-2} u'_j - \frac{1}{T} \sum_{t=1}^T v_{t-1} u'_{t-1} \\
&= v_T \frac{1}{T} \sum_{j=1}^{T-1} u'_j - \frac{1}{T} \sum_{t=1}^T v_{t-1} u'_{t-1} = -\frac{1}{T} \sum_{t=1}^T v_{t-1} u'_{t-1} + O_p(1/\sqrt{T}) \\
&= -\Sigma'_{UV'_0} + o_p(1)
\end{aligned}$$

Moreover, it follows similar to (19) that

$$\Sigma_{XY'} = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E[x_t (y_{t-1} - y_0)'] = \Sigma_{VV'_{-1}} - \Sigma_{VV'_0} - \Sigma'_{UV'_0}.$$

Thus by Theorem 1,

$$\frac{1}{T} \sum_{t=1}^T x_t y'_{t-1} \xrightarrow{d} C(1) \left(\int_0^1 (dB) B' \right) C(1)' + \Sigma_{XY'} \quad (20)$$

Note that

$$C(1)B(\cdot) \sim (C(1)C(1)')^{1/2} B_*(\cdot),$$

where B_* is also a k -variate standard Brownian motion. Moreover, the matrix $C(1)C(1)'$ is known as the long-run variance matrix of x_t :

$$\Sigma = C(1)C(1)' = \lim_{T \rightarrow \infty} E \left[\left(\frac{1}{\sqrt{T}} \sum_{t=1}^T x_t \right) \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T x_t \right)' \right]. \quad (21)$$

Thus, (20) also reads as

$$\frac{1}{T} \sum_{t=1}^T x_t y'_{t-1} \xrightarrow{d} \Sigma^{1/2} \left(\int_0^1 (dB_*) B'_* \right) \Sigma^{1/2} + \Sigma_{XY'} \quad (22)$$

2.5 A further generalization

In cointegration analysis we will encounter stochastic matrices of the type $\frac{1}{T} \sum_{t=1}^T x_{t-j} y'_{t-1}$, where $j \geq 0$. We have already considered the case $j = 0$, so let us focus on the case $j \geq 1$.

The limit distribution involved can easily be derived from the equality

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^T x_{t-j} y'_{t-1} &= \frac{1}{T} \sum_{t=1-j}^{T-j} x_t y'_{t-1+j} \\
&= \frac{1}{T} \sum_{t=1}^T x_t y'_{t-1+j} + \frac{1}{T} \sum_{t=1-j}^0 x_t y'_{t-1+j} - \frac{1}{T} \sum_{t=T-j+1}^T x_t y'_{t-1+j} \\
&= \frac{1}{T} \sum_{t=1}^T x_t y'_{t-1} + \frac{1}{T} \sum_{t=1}^T x_t (y_{t-1+j} - y_{t-1})' + O_p(1/T) \\
&\quad + O_p(1/\sqrt{T}) \\
&= \frac{1}{T} \sum_{t=1}^T x_t y'_{t-1} + \sum_{i=0}^{j-1} \frac{1}{T} \sum_{t=1}^T x_t x'_{t+i} + o_p(1)
\end{aligned} \tag{23}$$

where the O_p terms follow from

$$\begin{aligned}
\sum_{t=1-j}^0 x_t y'_{t-1+j} &= O_p(1), \\
\left\| \frac{1}{T} \sum_{t=T-j+1}^T x_t y'_{t-1+j} \right\| &\leq \frac{1}{T} \sum_{t=T-j+1}^T \|x_t\| \cdot \|y_{t-1+j}\| \\
&\leq \max_{0 \leq x \leq 1} \left\| y_{[xT]} / \sqrt{T} \right\| \frac{1}{\sqrt{T}} \sum_{t=T-j+1}^T \|x_t\|
\end{aligned}$$

where for a matrix $\|\cdot\|$ denotes the maximum of the absolute values of its elements, and the fact¹ $y_{[xT]} / \sqrt{T} \Rightarrow C(1)B(x)$ implies that

$$\max_{0 \leq x \leq 1} \left\| y_{[xT]} / \sqrt{T} \right\| \xrightarrow{d} \max_{0 \leq x \leq 1} \|B(x)\| = O_p(1).$$

Moreover, it follows from Lemma 1 that

$$p \lim_{T \rightarrow \infty} \sum_{i=0}^{j-1} \frac{1}{T} \sum_{t=1}^T x_t x'_{t+i} = \sum_{i=0}^{j-1} \Sigma_{XX'_i}$$

¹In the sequel, " \Rightarrow " denotes weak convergence.

Finally, note that

$$\Sigma_{XY'} + \sum_{i=0}^{j-1} \Sigma_{XX'_i} = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E [x_{t-j} (y_{t-1} - y_0)'] ,$$

hence the general result is:

Theorem 3. *Under Assumption 1 and for $m \geq 0$,*

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T x_{t-m} y'_{t-1} &\xrightarrow{d} C(1) \left(\int_0^1 (dB) B' \right) C(1)' + \Sigma_{XY'} + \sum_{i=0}^{m-1} \Sigma_{XX'_i} \quad (24) \\ &\sim \Sigma^{1/2} \left(\int_0^1 (dB_*) B'_* \right) \Sigma^{1/2} + \Sigma_{XY'} + \sum_{i=0}^{m-1} \Sigma_{XX'_i} \end{aligned}$$

where B and B_* are k -variate standard Brownian motions and

$$\begin{aligned} \Sigma_{XX'_i} &= E[x_t x'_{t+i}], \\ \Sigma_{XY'} &= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E [x_t (y_{t-1} - y_0)'] = \sum_{j=1}^{\infty} \Sigma'_{XX'_j}, \\ \Sigma &= \lim_{T \rightarrow \infty} \text{Var} \left[T^{-1/2} \sum_{t=1}^T x_t \right]. \end{aligned}$$

3 Cointegration

3.1 Vector error correction model (VECM) representation

If the time series process y_t defined by (8) in Assumption 1 is cointegrated, there exist a $k \times r$ matrix β with rank $r < k$ such that

$$\beta' C(1) = O. \quad (25)$$

Note that (25) implies that the matrix $C(1)$ is singular, with rank $k - r$. Then it follows from (12) that

$$\beta' y_t = \beta' v_t + \beta' y_0 - \beta' v_0. \quad (26)$$

Engle and Granger (1987) have set forth conditions such that then y_t can be written as an vector error correction model of order p , shortly VECM(p):

$$\Delta y_t = \pi_0 + \alpha\beta'y_{t-1} + \sum_{j=1}^{p-1} \Pi_j \Delta y_{t-j} + e_t, \quad (27)$$

where $e_t \sim \text{i.i.d. } N_k [0, \Omega]$, Δ is the difference operator $1 - L$, and the lag polynomial matrix $\Pi(L) = I_k - \sum_{j=1}^{p-1} \Pi_j L^j$ is invertible: $\det(\Pi(z)) = 0$ implies $|z| > 1$. Note that (27) is actually a VAR(p) model in y_t , but with unit roots in the determinant of the VAR lag polynomial involved.

The error term e_t in (27) is of course related to u_t in (9). To see how, substitute $\Delta y_t = C(L)u_t$ and

$$\begin{aligned} y_t &= C(1) \sum_{j=1}^t u_j + D(L)u_t - v_0 + y_0 \\ &= C(1) \left(\sum_{j=1}^t L^{j-1} \right) u_t + D(L)u_t - v_0 + y_0 \end{aligned}$$

in (27). Then

$$\begin{aligned} C(L)u_t &= C_0 u_t + \sum_{j=1}^{\infty} C_j u_{t-1} = C_0 u_t + (C(L) - C_0) u_t \\ &= \pi_0 - \alpha\beta' (v_0 - y_0) + \alpha\beta' C(1)L \left(\sum_{j=1}^t L^{j-1} \right) u_t \\ &\quad + \alpha\beta' L \cdot D(L)u_t + \sum_{j=1}^{p-1} \Pi_j L^j C(L)u_t + e_t, \end{aligned}$$

Since the only terms in this equation that relate to time t are e_t and the leading term $C_0 u_t$ of $C(L)u_t$, we must have

$$e_t = C_0 u_t, \quad (28)$$

hence

$$\Omega = C_0 C_0' \quad (29)$$

Thus, with $x_t = \Delta y_t$ we can write the VECM(p) as

$$x_t = \pi_0 + \alpha\beta'y_{t-1} + \sum_{j=1}^{p-1} \Pi_j x_{t-j} + C_0 u_t. \quad (30)$$

3.2 Granger's representation theorem

To see how (30) is related to (9), let ϕ be a $k \times (k - r)$ matrix with rank $k - r$ such that the matrix polynomial

$$P(L) = \begin{pmatrix} \beta' D(L) \\ \phi' C(L) \end{pmatrix}$$

is invertible, and the matrix

$$\Phi = \begin{pmatrix} \beta' \\ \phi' \end{pmatrix}$$

is nonsingular. Then

$$\begin{aligned} \Phi x_t &= \begin{pmatrix} \beta' (C(1) + (1 - L)D(L)) \\ \phi' C(L) \end{pmatrix} u_t \\ &= \begin{pmatrix} ((1 - L)\beta' D(L)) \\ \phi' C(L) \end{pmatrix} u_t \\ &= \begin{pmatrix} (1 - L)I_r & O \\ O & I_{k-r} \end{pmatrix} P(L)u_t \end{aligned}$$

hence

$$P(L)^{-1} \begin{pmatrix} I_r & O \\ O & (1 - L)I_{k-r} \end{pmatrix} \Phi \Delta y_t = \Delta u_t.$$

Applying the lag operator $1 + \sum_{j=1}^t L^j$ to both sides of this equation yields

$$P(L)^{-1} \begin{pmatrix} I_r & O \\ O & (1 - L)I_{k-r} \end{pmatrix} \Phi (y_t - y_0) = u_t - u_0,$$

hence

$$\begin{aligned} &P(L)^{-1} \begin{pmatrix} I_r & O \\ O & (1 - L)I_{k-r} \end{pmatrix} \Phi y_t \\ &= u_t + P(L)^{-1} \begin{pmatrix} I_r & O \\ O & (1 - L)I_{k-r} \end{pmatrix} \Phi y_0 - u_0 \\ &= u_t + P(1)^{-1} \begin{pmatrix} I_r & O \\ O & O \end{pmatrix} \Phi y_0 - u_0. \end{aligned}$$

Thus, denoting

$$\begin{aligned} A(L) &= P(L)^{-1} \begin{pmatrix} I_r & O \\ O & (1-L)I_{k-r} \end{pmatrix} \Phi, \\ \mu &= P(1)^{-1} \begin{pmatrix} I_r & O \\ O & O \end{pmatrix} \Phi y_0 - u_0 = P(1)^{-1} \begin{pmatrix} \beta' \\ O \end{pmatrix} y_0 - u_0. \end{aligned}$$

we have

$$A(L)y_t = \mu + u_t.$$

This is the VAR representation of y_t . Note however that $A(L)$ is not invertible.

Similar to (10) we can write

$$\begin{aligned} A(L) &= A(1)L + (1-L) \frac{A(L) - A(1)L}{1-L} \\ &= A(1)L + (1-L)\Psi(L), \end{aligned}$$

say, where

$$\Psi(L) = \frac{A(L) - A(1)L}{1-L}.$$

Moreover, note that

$$\begin{aligned} A(1) &= P(1)^{-1} \begin{pmatrix} I_r & O \\ O & O \end{pmatrix} \Phi = P(1)^{-1} \begin{pmatrix} I_r & O \\ O & O \end{pmatrix} \begin{pmatrix} \beta' \\ \phi' \end{pmatrix} \\ &= P(1)^{-1} \begin{pmatrix} \beta' \\ O \end{pmatrix}. \end{aligned}$$

Thus

$$\begin{aligned} A(L)y_t &= A(1)y_{t-1} + \Psi(L)\Delta y_t & (31) \\ &= P(1)^{-1} \begin{pmatrix} \beta' \\ O \end{pmatrix} y_{t-1} + \Psi(L)\Delta y_t = \mu + u_t \end{aligned}$$

Multiplying (31) by $\Phi^{-1}P(0)$ and denoting

$$\Pi(L) = \Phi^{-1}P(0)\Psi(L)$$

it follows that

$$\begin{aligned} \Pi(L)\Delta y_t &= \Phi^{-1}P(0)\Psi(L)\Delta y_t = -\Phi^{-1}P(0)P(1)^{-1} \begin{pmatrix} \beta' \\ O \end{pmatrix} y_{t-1} & (32) \\ &\quad + \Phi^{-1}P(0)\mu + \Phi^{-1}P(0)u_t. \end{aligned}$$

Note that

$$\Phi^{-1}P(0) = \Phi^{-1} \begin{pmatrix} \beta'D_0 \\ \phi'C_0 \end{pmatrix} = \Phi^{-1}\Phi C(0) = C_0 \quad (33)$$

and

$$\Phi^{-1}P(0)\mu = C_0P(1)^{-1} \begin{pmatrix} \beta' \\ O \end{pmatrix} y_0 - C_0u_0$$

Moreover, observe that the leading term in $\Pi(L)$ is

$$\begin{aligned} \Pi(0) &= \Phi^{-1}P(0)\Psi(0) = \Phi^{-1}P(0)A(0) \\ &= \Phi^{-1}P(0)P(0)^{-1}\Phi = I_k. \end{aligned}$$

Finally, assume that

$$\Pi(L) = I_k - \sum_{j=1}^{p-1} \Pi_j L^j,$$

denote

$$\begin{aligned} (\alpha, \alpha_*) &= -\Phi^{-1}P(0)P(1)^{-1} = -C_0P(1)^{-1} \\ &= -C_0 \begin{pmatrix} \beta'D(1) \\ \phi'C(1) \end{pmatrix}^{-1} \end{aligned} \quad (34)$$

where α is a $k \times r$ matrix and let

$$\begin{aligned} \pi_0 &= \Phi^{-1}P(0)\mu = \Phi^{-1}P(0)P(1)^{-1} \begin{pmatrix} \beta' \\ O \end{pmatrix} y_0 - \Phi^{-1}P(0)u_0 \\ &= -\alpha\beta'y_0 - C_0u_0. \end{aligned} \quad (35)$$

Then the VECM(p) model (30) follows.

An interesting special case is where π_0 takes the form

$$\pi_0 = -\alpha\phi \text{ for a vector } \phi \in \mathbb{R}^r, \quad (36)$$

because then the (30) becomes

$$x_t = \alpha\beta'(y_{t-1} - \phi) + \sum_{j=1}^{p-1} \Pi_j x_{t-j} + C_0 u_t. \quad (37)$$

This implies that $\beta'(y_{t-1} - \phi)$ is zero-mean stationary, because $E[x_t] = 0$. Hence, $\beta'y_{t-1}$ is stationary about a "constant" vector $\beta'\phi$. The case (36) is known as "cointegrating restrictions on the intercept parameters." If so, it follows from (35) that $\phi = \beta'y_0 + (\alpha'\alpha)^{-1}\alpha'C_0u_0$, which is actually a random vector.

A by-product of the above argument is:

Lemma 2. *Let α_\perp be an orthogonal complement² of α in VECM(p) model (30). There exists an orthogonal complement β_\perp of β such that $\beta'_\perp C(1) = (\alpha'_\perp \Omega \alpha_\perp)^{-1/2} \alpha'_\perp C_0$.*

Proof: Observe from (34) that

$$(\alpha, \alpha_*) \begin{pmatrix} \beta'D(1) \\ \phi'C(1) \end{pmatrix} = \alpha\beta'D(1) + \alpha_*\phi'C(1) = -C_0,$$

hence $\alpha'_\perp \alpha_* \phi' C(1) = \alpha'_\perp \alpha_* \phi' \beta_\perp \delta = -\alpha'_\perp C_0$. Thus, if we choose

$$\gamma' = -(\alpha'_\perp \Omega \alpha_\perp)^{-1/2} \alpha'_\perp \alpha_* \phi'$$

then

$$\gamma' C(1) = (\alpha'_\perp \Omega \alpha_\perp)^{-1/2} \alpha'_\perp C_0.$$

Next, observe that for any orthogonal complement $\bar{\beta}_\perp$ of β ,

$$\bar{\beta}_\perp \left(\bar{\beta}'_\perp \bar{\beta}_\perp \right)^{-1} \bar{\beta}'_\perp + \beta (\beta' \beta)^{-1} \beta' = I_k,$$

because the left-hand side matrix is idempotent with rank k , so that

$$\gamma = \bar{\beta}_\perp \left(\bar{\beta}'_\perp \bar{\beta}_\perp \right)^{-1} \bar{\beta}'_\perp \gamma + \beta (\beta' \beta)^{-1} \beta' \gamma.$$

Then

$$(\gamma' \bar{\beta}_\perp) \left(\bar{\beta}'_\perp \bar{\beta}_\perp \right)^{-1} \bar{\beta}'_\perp C(1) = (\alpha'_\perp \Omega \alpha_\perp)^{-1/2} \alpha'_\perp C_0.$$

Taking

$$\beta_\perp = \bar{\beta}_\perp \left(\bar{\beta}'_\perp \bar{\beta}_\perp \right)^{-1} \bar{\beta}'_\perp \gamma,$$

Lemma 2 follows. Q.E.D.

This result will play a key-role in the next sections.

²Given a $k \times r$ matrix ξ , where $1 \leq r < k$, an orthogonal complement ξ_\perp of ξ is $k \times (k-r)$ matrix with rank $k-r$ such that $\xi'_\perp \xi = O_{k-r,r}$. Note that ξ_\perp is not unique, because for any nonsingular $(k-r) \times (k-r)$ matrix R_{k-r} , $\xi_\perp R_{k-r}$ is also an orthogonal complement of ξ .

4 Likelihood-based cointegration analysis

4.1 The VECM(p) case without intercepts

To demonstrate the Johansen (1988) approach, assume in first instance that $u_t = 0$ for $t < 1$, so that $y_0 = 0$ and thus by (35), $\pi_0 = 0$, so that

$$x_t = \alpha\beta'y_{t-1} + \sum_{j=1}^{p-1} \Pi_j x_{t-j} + C_0 u_t. \quad (38)$$

where $x_t = \Delta y_t$. Next, denote

$$X_{t-1} = (x'_{t-1}, \dots, x'_{t-p+1})', \quad \Pi = (\Pi_1, \dots, \Pi_{p-1}), \quad (39)$$

Then (38) can be written as

$$x_t = \alpha\beta'y_{t-1} + \Pi X_{t-1} + C_0 u_t. \quad (40)$$

Note that, given β , the identification of α and Π requires that

Assumption 2. $\text{Var}\left[(y'_{t-1}\beta, X'_{t-1})'\right]$ is nonsingular.

The log-likelihood involved takes the form

$$\begin{aligned} & \ln L_T(\alpha, \beta, \Pi, \Omega) \\ &= -\frac{1}{2} \sum_{t=1}^{T+1} (x_t - \alpha\beta'y_{t-1} - \Pi X_{t-1})' \Omega^{-1} (x_t - \alpha\beta'y_{t-1} - \Pi X_{t-1}) \\ & \quad - \frac{1}{2} T \ln(\det \Omega) - T.k.\ln\left(\sqrt{2\pi}\right), \end{aligned}$$

where Ω is defined by (29).

Given α , β , and Ω , the matrix Π can be concentrated out by regressing $x_t - \alpha\beta'y_{t-1}$ on X_{t-1} :

$$\begin{aligned} \widehat{\Pi}(\alpha, \beta) &= \frac{1}{T} \sum_{t=1}^T (x_t X'_{t-1} - \alpha\beta'y_{t-1} X'_{t-1}) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} \\ &= \left(\frac{1}{T} \sum_{t=1}^T x_t X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} \\ & \quad - \alpha\beta' \left(\frac{1}{T} \sum_{t=1}^T y_{t-1} X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1}. \end{aligned}$$

Hence, denoting

$$R_{0,t} = x_t - \left(\frac{1}{T} \sum_{j=1}^T x_j X'_{j-1} \right) \left(\frac{1}{T} \sum_{j=1}^T X_{j-1} X'_{j-1} \right)^{-1} X_{t-1}, \quad (41)$$

$$R_{1,t} = y_{t-1} - \left(\frac{1}{T} \sum_{j=1}^T y_{j-1} X'_{j-1} \right) \left(\frac{1}{T} \sum_{j=1}^T X_{j-1} X'_{j-1} \right)^{-1} X_{t-1}, \quad (42)$$

the concentrated log-likelihood becomes

$$\begin{aligned} \ln L_T(\alpha, \beta, \Omega) &= \max_{\Pi} \ln L_T(\alpha, \beta, \Pi, \Omega) \\ &= -\frac{1}{2} \sum_{t=1}^T (R_{0,t} - \alpha \beta' R_{1,t})' \Omega^{-1} (R_{0,t} - \alpha \beta' R_{1,t}) \\ &\quad - \frac{1}{2} T \cdot \ln(\det \Omega) - T \cdot k \ln(\sqrt{2\pi}). \end{aligned} \quad (43)$$

Next, given β and Ω , α can be concentrated out by replacing $R_{0,t} - \alpha \beta' R_{1,t}$ with the OLS residual of the regression of $R_{0,t}$ on $\beta' R_{0,t}$, which yields

$$\hat{\alpha}(\beta) = \hat{S}_{0,1} \beta \left(\beta' \hat{S}_{1,1} \beta \right)^{-1}$$

where

$$\begin{aligned} \hat{S}_{0,1} &= \frac{1}{T} \sum_{t=1}^T R_{0,t} R'_{1,t} = \frac{1}{T} \sum_{t=1}^T x_t y'_{t-1} \\ &\quad - \left(\frac{1}{T} \sum_{t=1}^T x_t X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \right), \end{aligned} \quad (44)$$

$$\begin{aligned} \hat{S}_{1,1} &= \frac{1}{T} \sum_{t=1}^T R_{1,t} R'_{1,t} = \frac{1}{T} \sum_{t=1}^T y_{t-1} y'_{t-1} \\ &\quad - \left(\frac{1}{T} \sum_{t=1}^T y_{t-1} X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \right). \end{aligned} \quad (45)$$

Hence

$$\ln L_T(\beta, \Omega) = \max_{\Pi, \alpha} \ln L_T(\alpha, \beta, \Pi, \Omega)$$

$$\begin{aligned}
&= -\frac{1}{2} \sum_{t=1}^T (R_{0,t} - \hat{\alpha}(\beta) \beta' R_{1,t})' \Omega^{-1} (R_{0,t} - \hat{\alpha}(\beta) \beta' R_{1,t}) \\
&\quad - \frac{1}{2} T \ln(\det \Omega) - T \cdot k \ln(\sqrt{2\pi}).
\end{aligned}$$

As is well known the ML estimator of Ω given β now takes the form

$$\hat{\Omega}(\beta) = \frac{1}{T} \sum_{t=1}^T (R_{0,t} - \hat{\alpha}(\beta) \beta' R_{1,t}) (R_{0,t} - \hat{\alpha}(\beta) \beta' R_{1,t})'$$

which can be further elaborated as follows:

$$\begin{aligned}
\hat{\Omega}(\beta) &= \frac{1}{T} \sum_{t=1}^T R_{0,t} R'_{0,t} - \hat{\alpha}(\beta) \beta' \frac{1}{T} \sum_{t=1}^T R_{1,t} R'_{1,t} \beta \hat{\alpha}(\beta)' \\
&= \hat{S}_{0,0} - \hat{\alpha}(\beta) (\beta' \hat{S}_{1,1} \beta) \hat{\alpha}(\beta)' \\
&= \hat{S}_{0,0} - \hat{S}_{0,1} \beta (\beta' \hat{S}_{1,1} \beta)^{-1} \beta' \hat{S}_{1,0}
\end{aligned}$$

where

$$\hat{S}_{0,0} = \frac{1}{T} \sum_{t=1}^T R_{0,t} R'_{0,t} = \frac{1}{T} \sum_{t=1}^T x_t x'_t \quad (46)$$

$$\begin{aligned}
&\quad - \left(\frac{1}{T} \sum_{t=1}^T x_t X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} x'_t \right), \\
\hat{S}_{1,0} &= \hat{S}'_{0,1} \quad (47)
\end{aligned}$$

Then

$$\begin{aligned}
&\frac{1}{T} \sum_{t=1}^T (R_{0,t} - \hat{\alpha}(\beta) \beta' R_{1,t})' \left(\hat{\Omega}(\beta) \right)^{-1} (R_{0,t} - \hat{\alpha}(\beta) \beta' R_{1,t}) \\
&= \text{trace} \left(\left(\hat{\Omega}(\beta) \right)^{-1} \hat{\Omega}(\beta) \right) = \text{trace}(I_k) = k,
\end{aligned}$$

hence

$$\begin{aligned}
\ln L_T(\beta) &= \max_{\Pi, \alpha, \Omega} \ln L_T(\alpha, \beta, \Pi, \Omega) \\
&= -\frac{1}{2} T \cdot \ln \left(\det \left(\hat{S}_{0,0} - \hat{S}_{0,1} \beta (\beta' \hat{S}_{1,1} \beta)^{-1} \beta' \hat{S}_{1,0} \right) \right) \\
&\quad - T \cdot k \ln(\sqrt{2\pi}) - kT.
\end{aligned}$$

Thus, the maximum likelihood estimator $\hat{\beta}$ of β can be obtained by minimizing

$$\det \left(\hat{S}_{0,0} - \hat{S}_{0,1}\beta \left(\beta' \hat{S}_{1,1}\beta \right)^{-1} \beta' \hat{S}_{1,0} \right)$$

to β .

To simplify this problem, consider the easy matrix equalities

$$\begin{aligned} \begin{pmatrix} A & B \\ B' & C \end{pmatrix} &= \begin{pmatrix} A & O \\ B' & I \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ O & C - B'A^{-1}B \end{pmatrix} \\ &= \begin{pmatrix} I & B \\ O & C \end{pmatrix} \begin{pmatrix} A - BC^{-1}B' & O \\ C^{-1}B' & I \end{pmatrix}, \end{aligned}$$

where A and C are nonsingular square matrices. These equalities imply that

$$\det(A) \det(C - B'A^{-1}B) = \det(C) \det(A - BC^{-1}B').$$

Taking $A = \beta' \hat{S}_{1,1}\beta$, $B = \beta' \hat{S}_{1,0}$, $C = \hat{S}_{0,0}$ it follows now that

$$\begin{aligned} &\det \left(\hat{S}_{0,0} - \hat{S}_{0,1}\beta \left(\beta' \hat{S}_{1,1}\beta \right)^{-1} \beta' \hat{S}_{1,0} \right) \\ &= \frac{\det(\hat{S}_{0,0}) \det(\beta' \hat{S}_{1,1}\beta - \beta' \hat{S}_{1,0} \hat{S}_{0,0}^{-1} \hat{S}_{0,1}\beta)}{\det(\beta' \hat{S}_{1,1}\beta)}. \end{aligned} \quad (48)$$

Thus, the ML estimator $\hat{\beta}$ of β minimizes (48). However, if $\hat{\beta}$ is a solution then so is $c\hat{\beta}$ for any $c \neq 0$ in the case $r = 1$, and in the case $2 < r < k$, we may replace $\hat{\beta}$ by $\hat{\beta}C_{r,r}$ with $C_{r,r}$ an arbitrary nonsingular $r \times r$ matrix. Thus, we need to normalize $\hat{\beta}$ somehow. How to normalize $\hat{\beta}$ will be explained below.

Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_k$ be the ordered solutions of the generalized eigenvalue problem

$$\det \left(\lambda \hat{S}_{1,1} - \hat{S}_{1,0} \hat{S}_{0,0}^{-1} \hat{S}_{0,1} \right) = 0, \quad (49)$$

and let $\hat{q}_1, \hat{q}_2, \dots, \hat{q}_k$ be the corresponding generalized eigenvectors. Then for $j = 1, \dots, k$,

$$\hat{S}_{1,1} \hat{q}_j \hat{\lambda}_j = \hat{S}_{1,0} \hat{S}_{0,0}^{-1} \hat{S}_{0,1} \hat{q}_j.$$

Hence, denoting

$$\hat{\Lambda} = \text{diag} \left(\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_k \right), \quad \hat{Q} = (\hat{q}_1, \hat{q}_2, \dots, \hat{q}_k),$$

we have

$$\widehat{S}_{1,1}\widehat{Q}\widehat{\Lambda} = \widehat{S}_{1,0}\widehat{S}_{0,0}^{-1}\widehat{S}_{0,1}\widehat{Q}. \quad (50)$$

The eigenvalues $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \dots \geq \widehat{\lambda}_k$ can be obtained by solving the standard eigenvalue problem

$$\det\left(\lambda I_k - \widehat{S}_{1,1}^{-1/2}\widehat{S}_{1,0}\widehat{S}_{0,0}^{-1}\widehat{S}_{0,1}\widehat{S}_{1,1}^{-1/2}\right) = 0, \quad (51)$$

with corresponding orthonormal eigenvector $\widehat{q}_1^*, \widehat{q}_2^*, \dots, \widehat{q}_k^*$. Thus, denoting $\widehat{Q}^* = (\widehat{q}_1^*, \widehat{q}_2^*, \dots, \widehat{q}_k^*)$, we have

$$\widehat{Q}^*\widehat{\Lambda} = \widehat{S}_{1,1}^{-1/2}\widehat{S}_{1,0}\widehat{S}_{0,0}^{-1}\widehat{S}_{0,1}\widehat{S}_{1,1}^{-1/2}\widehat{Q}^*. \quad (52)$$

Comparing (51) and (52) we see that we may choose $\widehat{Q} = \widehat{S}_{1,1}^{-1/2}\widehat{Q}^*$, so that by the orthogonality of \widehat{Q}^* ,

$$\widehat{Q}'\widehat{S}_{1,1}\widehat{Q} = I_k.$$

Next, let

$$\widehat{\beta} = \widehat{Q}\xi,$$

where ξ is normalized such that

$$\xi'\xi = I_r.$$

Because $\widehat{\beta} = \widehat{Q}\xi = \widehat{S}_{1,1}^{-1/2}\widehat{Q}^*\xi$ and \widehat{Q}^* is orthogonal, this normalization implies that

$$\widehat{\beta}'\widehat{S}_{1,1}\widehat{\beta} = I_r.$$

Then

$$\begin{aligned} \frac{\det\left(\beta'\widehat{S}_{1,1}\beta - \beta'\widehat{S}_{1,0}\widehat{S}_{0,0}^{-1}\widehat{S}_{0,1}\beta\right)}{\det\left(\beta'\widehat{S}_{1,1}\beta\right)} &= \det\left(I_r - \xi'\widehat{Q}'\widehat{S}_{1,0}\widehat{S}_{0,0}^{-1}\widehat{S}_{0,1}\widehat{Q}\xi\right) \\ &= \det\left(I_r - \xi'\widehat{\Lambda}\xi\right). \end{aligned}$$

In the case $r = 1$, $\xi = (\xi_1, \dots, \xi_k)' \in \mathbb{R}^k$ and $\det\left(I_r - \xi'\widehat{\Lambda}\xi\right) = 1 - \xi'\widehat{\Lambda}\xi = 1 - \sum_{j=1}^k \widehat{\lambda}_j \xi_j^2$, which is minimal subject to $\sum_{j=1}^k \xi_j^2 = 1$ for $\xi = (1, 0, \dots, 0)'$. In the case $r > 1$ it is not hard to show that the solution is

$$\xi = \begin{pmatrix} I_r \\ O \end{pmatrix}.$$

Consequently,

$$\widehat{\beta} = (\widehat{q}_1, \widehat{q}_2, \dots, \widehat{q}_r)$$

so that

$$\widehat{\lambda}_i \widehat{S}_{1,1} \widehat{q}_i = \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \widehat{q}_i, \quad i = 1, 2, \dots, r, \quad (53)$$

and

$$\begin{aligned} \ln L_T(r) &= \max_{\Pi, \alpha, \beta, \Omega} \ln L_T(\alpha, \beta, \Pi, \Omega) \\ &= -\frac{1}{2} T \cdot \ln \left(\det \left(\widehat{S}_{0,0} \right) \right) - \frac{1}{2} T \cdot \sum_{j=1}^r \ln \left(1 - \widehat{\lambda}_j \right) - T \cdot k \ln \left(\sqrt{2\pi} \right) - kT. \end{aligned} \quad (54)$$

4.2 Tests for the cointegrating rank

Consequently, the likelihood-ratio (LR) test of the null-hypothesis that the cointegrating rank is $r < k$ against the alternative hypothesis that the cointegrating rank is $r + 1$ takes the form

$$\begin{aligned} LR_T(r|r+1) &= -2 (\ln L_T(r) - L_T(r+1)) = -T \cdot \ln \left(1 - \widehat{\lambda}_{r+1} \right) \\ &= T \cdot \widehat{\lambda}_{r+1} + o_p(1), \end{aligned} \quad (55)$$

where the approximation is due to the Taylor expansion of $\ln \left(1 - \widehat{\lambda}_{r+1} \right)$, provided that $T \cdot \widehat{\lambda}_{r+1}$ converges to a distribution under the null hypothesis. The latter will be shown below. Since $\widehat{\lambda}_{r+1}$ is the largest of the $k - r$ smallest solutions of the generalized eigenvalue problem (49), the LR test (55) is called by Johansen (1988) the *lambda-max test*.

Another test proposed by Johansen (1988) is the LR test

$$\begin{aligned} LR_T(r|k) &= -2 (\ln L_T(r) - L_T(k)) = -T \cdot \sum_{i=r+1}^k \ln \left(1 - \widehat{\lambda}_i \right) \\ &= \sum_{i=r+1}^k T \cdot \widehat{\lambda}_i + o_p(1), \end{aligned} \quad (56)$$

which has the same null hypothesis as before, but as alternative hypothesis that the cointegrating rank is k . This alternative hypothesis implies that β is a nonsingular $k \times k$ matrix, which in its turn implies that y_t is stationary,

because if $\beta'y_t$ is stationary and β' is nonsingular then $y_t = (\beta')^{-1}\beta'y_t$ is stationary. This test is called by Johansen (1988) the *trace test*, because its null distribution takes the form of the trace of a random matrix.

Although these tests are designed on the basis of a specific alternative hypothesis, they have power against the more general alternative that the cointegrating rank is larger than r .

4.3 Limiting distributions

Because the log-likelihood (54) is a function of the solutions of the generalized eigenvalue problem (49), which in its turn depend on the matrices $\widehat{S}_{0,0}$, $\widehat{S}_{0,1}$ and $\widehat{S}_{1,1}$, we need to derive the limiting distributions or probability limits of these matrices.

First note that

Lemma 3. *Under Assumption 1 the probability limits*

$$\begin{aligned}\Sigma_{\beta\beta} &= p \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} y_{t-1}' \beta, \\ \Sigma_{\beta X} &= p \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} X_{t-1}', \quad \Sigma_{X\beta} = \Sigma_{\beta X}', \\ \Sigma_{XX} &= p \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_{t-1} X_{t-1}'\end{aligned}$$

exist. Moreover, under Assumption 2, Σ_{XX} is non-singular, and the matrix

$$\Sigma_{\beta\beta}^* = \Sigma_{\beta\beta} - \Sigma_{\beta X} \Sigma_{XX}^{-1} \Sigma_{X\beta}$$

is nonsingular

Proof: Only the last part is not obvious. To show this part, observe that $\Sigma_{\beta\beta}^*$ is the variance matrix of the error in the projection of $\beta'y_{t-1}$ on X_{t-1} : $\beta'y_{t-1} = \Gamma X_{t-1} + \eta_t$, where $\Gamma = \Sigma_{\beta X} \Sigma_{XX}^{-1}$, and $\Sigma_{\beta\beta}^* = \text{Var}(\eta_t)$. If $\Sigma_{\beta\beta}^*$ is singular then there exist vectors $\omega \in \mathbb{R}^r$, $v \in \mathbb{R}^k$ such that $\omega' \beta'y_{t-1} = v' \Gamma X_{t-1}$, which however violates Assumption 2. Q.E.D.

4.3.1 The matrix $\widehat{S}_{0,0}$

Replacing x_t in (46) by the right-hand side of (40) yields

$$\begin{aligned}
\widehat{S}_{0,0} &= \frac{1}{T} \sum_{t=1}^T (\alpha \beta' y_{t-1} + \Pi X_{t-1} + C_0 u_t) (y'_{t-1} \beta \alpha' + X'_{t-1} \Pi' + u'_t C'_0) \\
&\quad - \left(\alpha \frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} X'_{t-1} + \Pi \frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} + C_0 \frac{1}{T} \sum_{t=1}^T u_t X'_{t-1} \right) \\
&\quad \times \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} \\
&\quad \times \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \beta \alpha' + \frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \Pi' + \frac{1}{T} \sum_{t=1}^T X_{t-1} u'_t C'_0 \right) \\
&= \alpha \left(\frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} y'_{t-1} \beta \right) \alpha' + C_0 C'_0 \\
&\quad - \alpha \left(\frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \beta \right) \alpha' \\
&\quad + o_p(1) \\
&= \Omega + \alpha \Sigma_{\beta \beta}^* \alpha' + o_p(1),
\end{aligned}$$

where $\Omega = C_0 C'_0$, and $\Sigma_{\beta \beta}^*$ is defined in Lemma 3.

Next, choose φ such that

$$\alpha'_\perp \Omega \varphi = O. \quad (57)$$

Then

$$\alpha'_\perp \widehat{S}_{0,0} \alpha_\perp = \alpha'_\perp \Omega \alpha_\perp + o_p(1), \quad (58)$$

$$\alpha'_\perp \widehat{S}_{0,0} \varphi = \alpha'_\perp \Omega \varphi + o_p(1) = o_p(1), \quad (59)$$

$$\varphi' \widehat{S}_{0,0} \varphi = \varphi' \Omega \varphi + \varphi' \alpha \Sigma_{\beta \beta}^* \alpha' \varphi + o_p(1) \quad (60)$$

Hence it follows from (58), (59) and (60) that

$$\begin{aligned}
p \lim_{T \rightarrow \infty} \left(\begin{pmatrix} \alpha'_\perp \\ \varphi' \end{pmatrix} \widehat{S}_{0,0} (\alpha_\perp, \varphi) \right)^{-1} &= (\alpha_\perp, \varphi)^{-1} \left(p \lim_{T \rightarrow \infty} \widehat{S}_{0,0}^{-1} \right) \begin{pmatrix} \alpha'_\perp \\ \varphi' \end{pmatrix}^{-1} \\
&= \begin{pmatrix} (\alpha'_\perp \Omega \alpha_\perp)^{-1} & O \\ O & \Psi_0 \end{pmatrix} \quad (61)
\end{aligned}$$

where

$$\Psi_0 = (\varphi' \Omega \varphi + \varphi' \alpha \Sigma_{\beta\beta}^* \alpha' \varphi)^{-1}$$

Finally, note that any $k \times r$ matrix φ with rank r satisfying (57) takes the form

$$\varphi = \Omega^{-1} \alpha . R, \quad (62)$$

where R is a non-singular $r \times r$ matrix, hence

$$\begin{aligned} \Psi_0 &= (R' \alpha' \Omega^{-1} \alpha . R + R' \alpha' \Omega^{-1} \alpha \Sigma_{\beta\beta}^* \alpha' \Omega^{-1} \alpha . R)^{-1} \\ &= \left(R' \alpha' \Omega^{-1} \alpha \left((\alpha' \Omega^{-1} \alpha)^{-1} + \Sigma_{\beta\beta}^* \right) \alpha' \Omega^{-1} \alpha . R \right)^{-1} \end{aligned}$$

Summarizing, it has been shown that:

Lemma 4. *Under VECM (38) and Assumptions 1-2,*

$$\widehat{S}_{0,0} = \Omega + \alpha \Sigma_{\beta\beta}^* \alpha' + o_p(1),$$

and

$$\begin{aligned} p \lim_{T \rightarrow \infty} \widehat{S}_{0,0}^{-1} &= \alpha_\perp (\alpha'_\perp \Omega \alpha_\perp)^{-1} \alpha'_\perp \\ &+ \Omega^{-1} \alpha (\alpha' \Omega^{-1} \alpha)^{-1} \left((\alpha' \Omega^{-1} \alpha)^{-1} + \Sigma_{\beta\beta}^* \right)^{-1} (\alpha' \Omega^{-1} \alpha)^{-1} \alpha' \Omega^{-1}. \end{aligned}$$

where $\Sigma_{\beta\beta}^*$ is defined in Lemma 3.

4.3.2 The matrix $\widehat{S}_{0,1}$

Recall that

$$\begin{aligned} \widehat{S}_{0,1} &= \frac{1}{T} \sum_{t=1}^T x_t y'_{t-1} \\ &- \left(\frac{1}{T} \sum_{t=1}^T x_t X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \right). \end{aligned} \quad (63)$$

Replacing x_t in (63) by the right-hand side of (40) yields

$$\begin{aligned}\widehat{S}_{0,1} &= \alpha \left(\frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} y'_{t-1} \right) \\ &\quad - \alpha \left(\frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \right) \\ &\quad + C_0 \left(\frac{1}{T} \sum_{t=1}^T u_t y'_{t-1} \right) \\ &\quad - C_0 \left(\frac{1}{T} \sum_{t=1}^T u_t X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \right).\end{aligned}$$

Due to (39) and Theorem 3, there exists a $k(p-1) \times k$ random matrix M_* such that

$$\frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \xrightarrow{d} M_*, \quad (64)$$

hence $\frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} = O_p(1)$. Moreover, it is easy to verify that under Assumption 1,

$$p \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T u_t X'_{t-1} = O. \quad (65)$$

Hence,

$$\begin{aligned}\widehat{S}_{0,1} &= \alpha \left(\frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} y'_{t-1} \right) - \alpha \Sigma_{\beta X} \Sigma_{XX}^{-1} \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \right) \\ &\quad + C_0 \left(\frac{1}{T} \sum_{t=1}^T u_t y'_{t-1} \right) + o_p(1)\end{aligned}$$

and therefore

$$\alpha'_\perp \widehat{S}_{0,1} = \alpha'_\perp C_0 \left(\frac{1}{T} \sum_{t=1}^T u_t y'_{t-1} \right) + o_p(1), \quad (66)$$

$$\alpha'_\perp \widehat{S}_{0,1} \beta = \alpha'_\perp C_0 \left(\frac{1}{T} \sum_{t=1}^T u_t y'_{t-1} \beta \right) + o_p(1) = o_p(1) \quad (67)$$

$$\widehat{S}_{0,1} \beta = \alpha \Sigma_{\beta \beta}^* + o_p(1) \quad (68)$$

where the latter follows from Lemma 3.

Let φ be given by (62) and let β_\perp be the $k \times (k - r)$ matrix in Lemma 2. It follows now from (66), (67), (68), Theorem 2 and Lemma 2 that

$$\begin{aligned} \begin{pmatrix} \alpha'_\perp \\ \varphi' \end{pmatrix} \widehat{S}_{0,1}(\beta_\perp, \beta) &= \begin{pmatrix} \alpha'_\perp \widehat{S}_{0,1} \beta_\perp & \alpha'_\perp \widehat{S}_{0,1} \beta \\ \varphi' \widehat{S}_{0,1} \beta_\perp & \varphi' \widehat{S}_{0,1} \beta \end{pmatrix} \\ &\xrightarrow{d} \begin{pmatrix} \alpha'_\perp C_0 \int_0^1 (dB) B' \beta_\perp & O_{k-r,r} \\ \varphi' \widetilde{S}_{0,1} \beta_\perp & \varphi' \alpha (\Sigma_{\beta\beta} - \Sigma_{\beta X} \Sigma_{XX}^{-1} \Sigma_{X\beta}) \end{pmatrix} \end{aligned} \quad (69)$$

where $\widetilde{S}_{0,1}$ is a random matrix such that

$$\widehat{S}_{0,1} \xrightarrow{d} \widetilde{S}_{0,1}.$$

The latter follows from Theorem 3.

To make the right-hand side matrix in (69) block-diagonal, post-multiply (69) by

$$\Theta = \begin{pmatrix} I_{k-r} & O_{k-r,r} \\ \Theta_{21} & I_r \end{pmatrix} \quad (70)$$

where

$$\begin{aligned} \Theta_{21} &= -(\varphi' \alpha \Sigma_{\beta\beta}^*)^{-1} \varphi' \widetilde{S}_{0,1} \beta_\perp \\ &= -(R' \alpha' \Omega^{-1} \alpha \Sigma_{\beta\beta}^*)^{-1} R' \alpha' \Omega^{-1} \alpha \widetilde{S}_{0,1} \beta_\perp \\ &= (\Sigma_{\beta\beta}^*)^{-1} \widetilde{S}_{0,1} \beta_\perp. \end{aligned} \quad (71)$$

Note that the second equality follows from (62). Then

$$\begin{aligned} \begin{pmatrix} \alpha'_\perp \\ \varphi' \end{pmatrix} \widehat{S}_{0,1}(\beta_\perp, \beta) \Theta &\\ &\xrightarrow{d} \begin{pmatrix} \alpha'_\perp C_0 \int_0^1 (dB) B' C(1)' \beta_\perp & O \\ O & \varphi' \alpha \Sigma_{\beta\beta}^* \end{pmatrix} \end{aligned} \quad (72)$$

Combining (72) and (61) it follows now that

Lemma 5. *Under VECM (38) and Assumptions 1-2,*

$$\Theta' \begin{pmatrix} \beta'_\perp \\ \beta' \end{pmatrix} \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1}(\beta_\perp, \beta) \Theta \xrightarrow{d} \begin{pmatrix} \int_0^1 B_{k-r} dB'_{k-r} \int_0^1 (dB_{k-r}) B'_{k-r} & O \\ O & \Psi_* \end{pmatrix} \quad (73)$$

where β_\perp is the $k \times (k - r)$ matrix in Lemma 2, Θ is defined by (70) and (71),

$$B_{k-r} = (\alpha'_\perp \Omega \alpha_\perp)^{-1/2} \alpha'_\perp C_0 B \quad (74)$$

is a $k - r$ variate standard Brownian motion, and

$$\Psi_* = \Sigma_{\beta\beta}^* \left((\alpha' \Omega^{-1} \alpha)^{-1} + \Sigma_{\beta\beta}^* \right)^{-1} \Sigma_{\beta\beta}^*, \quad (75)$$

with $\Sigma_{\beta\beta}^*$ defined in Lemma 3. Moreover, $\widehat{S}_{0,1}\beta = \alpha \Sigma_{\beta\beta}^* + o_p(1)$.³

4.3.3 The matrix $\widehat{S}_{1,1}$

Since by Theorem 3, $\frac{1}{T} \sum_{j=1}^T X_{t-1} y'_{t-1}$ converges in distribution and therefore is of order $O_p(1)$, it follows from (45) and Lemma 3 that

$$\begin{aligned} \widehat{S}_{1,1} &= \frac{1}{T} \sum_{t=1}^T y_{t-1} y'_{t-1} \\ &\quad - \left(\frac{1}{T} \sum_{j=1}^T y_{t-1} X'_{t-1} \right) \Sigma_{XX}^{-1} \left(\frac{1}{T} \sum_{j=1}^T X_{t-1} y'_{t-1} \right) + o_p(1) \\ &= \frac{1}{T} \sum_{t=1}^T y_{t-1} y'_{t-1} + O_p(1) \end{aligned} \quad (76)$$

Moreover,

$$\frac{1}{T^2} \sum_{t=1}^T y_{t-1} y'_{t-1} \xrightarrow{d} C(1) \left(\int_0^1 B(x) B(x)' dx \right) C(1)'$$

due to the functional central limit theorem, i.e.,

$$y_{[x,T]} / \sqrt{T} \Rightarrow C(1) B(x),$$

and the continuous mapping theorem. Hence

$$\frac{1}{T} \widehat{S}_{1,1} \xrightarrow{d} C(1) \left(\int_0^1 B(x) B(x)' dx \right) C(1)'. \quad (77)$$

³This is result (66), which is included in Lemma 5 for later reference.

Thus, with β_\perp defined in Lemma 2

$$\frac{1}{T} \beta'_\perp \widehat{S}_{1,1} \beta_\perp \xrightarrow{d} \beta'_\perp C(1) \left(\int_0^1 BB' \right) C(1)' \beta_\perp = \int_0^1 B_{k-r} B'_{k-r} \quad (78)$$

where B_{k-r} is defined in (74) and $\int_0^1 B_{k-r} B'_{k-r}$ is a short-hand notation for $\int_0^1 B_{k-r}(x) B_{k-r}(x)' dx$. Moreover, it follows from Lemma 3 that

$$p \lim_{T \rightarrow \infty} \beta' \widehat{S}_{1,1} \beta = \Sigma_{\beta\beta}^*$$

Furthermore, it follows from (30) than

$$\beta' y_{t-1} = (\alpha' \alpha)^{-1} \alpha' x_t - (\alpha' \alpha)^{-1} \alpha' \Pi X_{t-1} - (\alpha' \alpha)^{-1} \alpha' C_0 u_t,$$

hence by Theorems 2-3 and Lemma 3,

$$\begin{aligned} \beta' \widehat{S}_{1,1} &= \frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} y'_{t-1} - \Sigma_{\beta X} \Sigma_{XX}^{-1} \left(\frac{1}{T} \sum_{j=1}^T x_j y'_{t-1} \right) + o_p(1) \quad (79) \\ &= \frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} y'_{t-1} + O_p(1) \\ &= (\alpha' \alpha)^{-1} \alpha' \frac{1}{T} \sum_{t=1}^T x_t y'_{t-1} - (\alpha' \alpha)^{-1} \alpha' \Pi \frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \\ &\quad - (\alpha' \alpha)^{-1} \alpha' C_0 \frac{1}{T} \sum_{t=1}^T u_t y'_{t-1} + O_p(1) \\ &= O_p(1) \end{aligned}$$

Thus, for the $k \times (k-r)$ matrix β_\perp in Lemma 2,

$$\begin{pmatrix} T^{-1/2} \beta'_\perp \\ \beta' \end{pmatrix} \widehat{S}_{1,1} \left(T^{-1/2} \beta_\perp, \beta \right) \xrightarrow{d} \begin{pmatrix} \int_0^1 B_{k-r} B'_{k-r} & O \\ O & \Sigma_{\beta\beta}^* \end{pmatrix} \quad (80)$$

To make this result comparable with Lemma 5, observe from (70) that

$$\begin{aligned} \Theta \begin{pmatrix} T^{-1/2} I_{k-r} & O_{k-r,r} \\ O_{r,k-r} & I_r \end{pmatrix} &= \begin{pmatrix} I_{k-r} & O_{k-r,r} \\ \Theta_{21} & I_r \end{pmatrix} \begin{pmatrix} T^{-1/2} I_{k-r} & O_{k-r,r} \\ O_{r,k-r} & I_r \end{pmatrix} \\ &= \begin{pmatrix} T^{-1/2} I_{k-r} & O_{k-r,r} \\ T^{-1/2} \Theta_{21} & I_r \end{pmatrix} \end{aligned}$$

hence it follows from (80) that

Lemma 6. *Under VECM (38) and Assumptions 1-2,*

$$\begin{aligned} & \left(\begin{array}{cc} T^{-1/2}I_{k-r} & O_{k-r,r} \\ O_{r,k-r} & I_r \end{array} \right) \Theta' \left(\begin{array}{c} \beta'_\perp \\ \beta' \end{array} \right) \widehat{S}_{1,1}(\beta_\perp, \beta) \Theta \left(\begin{array}{cc} T^{-1/2}I_{k-r} & O_{k-r,r} \\ O_{r,k-r} & I_r \end{array} \right) \\ &= \left(\begin{array}{c} T^{-1/2}\beta'_\perp + T^{-1/2}\Theta'_{21}\beta' \\ \beta' \end{array} \right) \widehat{S}_{1,1}(T^{-1/2}\beta_\perp + T^{-1/2}\beta\Theta_{21}, \beta) \\ &\xrightarrow{d} \left(\begin{array}{cc} \int_0^1 B_{k-r}B'_{k-r} & O \\ O & \Sigma_{\beta\beta}^* \end{array} \right). \end{aligned} \quad (81)$$

where β_\perp is the $k \times (k-r)$ matrix in Lemma 2, Θ is defined by (70) and (71), and $\Sigma_{\beta\beta}^*$ is defined in Lemma 3.

4.3.4 Limiting distributions of the general eigenvalues and the LR test statistics

Let $\Xi = (\beta_\perp, \beta)\Theta$. Since Ξ is non-singular, the generalized eigenvalue problem (49) is equivalent to finding $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \dots \geq \widehat{\lambda}_k$ such that for $i = 1, \dots, k$,

$$\begin{aligned} 0 &= \det \left(\widehat{\lambda}_i \left(\begin{array}{cc} T^{-1/2}I_{k-r} & O_{k-r,r} \\ O_{r,k-r} & I_r \end{array} \right) \Xi' \widehat{S}_{1,1} \Xi \left(\begin{array}{cc} T^{-1/2}I_{k-r} & O_{k-r,r} \\ O & I_r \end{array} \right) \right. \\ &\quad \left. - \left(\begin{array}{cc} T^{-1/2}I_{k-r} & O_{k-r,r} \\ O_{r,k-r} & I_r \end{array} \right) \Xi' \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \Xi \left(\begin{array}{cc} T^{-1/2}I_{k-r} & O_{k-r,r} \\ O_{r,k-r} & I_r \end{array} \right) \right) \end{aligned} \quad (82)$$

or equivalently,

$$0 = \det \left(T\widehat{\lambda}_i T^{-1} \Xi' \widehat{S}_{1,1} \Xi - \Xi' \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \Xi \right), \quad (83)$$

It follows from Anderson, Brons and Jensen (1983, Lemma 2)⁴ that the solutions $\widehat{\lambda}_i$ of (82) converge in distribution to the corresponding solutions of the generalized eigenvalue problem

$$0 = \det \left(\lambda \left(\begin{array}{cc} \int_0^1 B_{k-r}B'_{k-r} & O \\ O & \Psi_{**} \end{array} \right) - \left(\begin{array}{cc} O & O \\ O & \Psi_* \end{array} \right) \right)$$

⁴See Lemma A.1 in the Appendix.

$$\begin{aligned}
&= \det \left(\lambda \int_0^1 B_{k-r} B'_{k-r} \right) \det (\lambda \Sigma_{\beta\beta}^* - \Psi_*) \\
&= \lambda^{k-r} \det \left(\lambda \Sigma_{\beta\beta}^* - \Sigma_{\beta\beta}^* \left((\alpha' \Omega^{-1} \alpha)^{-1} + \Sigma_{\beta\beta}^* \right)^{-1} \Sigma_{\beta\beta}^* \right) \\
&\quad \times \det \left(\int_0^1 B_{k-r} B'_{k-r} \right)
\end{aligned}$$

Note that $\det \left(\lambda \Sigma_{\beta\beta}^* - \Sigma_{\beta\beta}^* \left((\alpha' \Omega^{-1} \alpha)^{-1} + \Sigma_{\beta\beta}^* \right)^{-1} \Sigma_{\beta\beta}^* \right) = 0$ is equivalent to

$$\det \left(\lambda I_r - (\bar{\Psi} + I_r)^{-1} \right) = 0 \quad (84)$$

where

$$\bar{\Psi} = (\Sigma_{\beta\beta}^*)^{-1/2} (\alpha' \Omega^{-1} \alpha) (\Sigma_{\beta\beta}^*)^{-1/2} \quad (85)$$

Consequently, the solutions of (84) are all between 0 and 1. Thus, the r largest ordered solutions of (82) converge in distribution to the correspondingly ordered solutions of (84). Since the latter are non-random, and convergence in distribution to a constant implies convergence in probability to that constant, it follows that $p \lim_{T \rightarrow \infty} (\hat{\lambda}_1, \dots, \hat{\lambda}_k)' = (\bar{\lambda}_1, \dots, \bar{\lambda}_r, 0, 0, \dots, 0)'$, where $1 > \bar{\lambda}_1 \geq \dots \geq \bar{\lambda}_r > 0$ are the ordered solutions of (84).

In the case of (83) it follows from Lemmas 5 and 6 that

$$\Xi' \hat{S}_{1,0} \hat{S}_{0,0}^{-1} \hat{S}_{0,1} \Xi \xrightarrow{d} \begin{pmatrix} \int_0^1 B_{k-r} dB'_{k-r} \int_0^1 (dB_{k-r}) B'_{k-r} & O \\ O & \Psi_* \end{pmatrix} \quad (86)$$

$$T^{-1} \Xi' \hat{S}_{1,1} \Xi \xrightarrow{d} \begin{pmatrix} \int_0^1 B_{k-r} B'_{k-r} & O \\ O & O \end{pmatrix}, \quad (87)$$

Because the right-hand side matrix in (87) is singular, we cannot conclude directly from Anderson, Brons and Jensen (1983, Lemma 2) that $(T \hat{\lambda}_{r+1}, \dots, T \hat{\lambda}_k)'$ converges in distribution to the corresponding solutions of

$$\begin{aligned}
0 &= \det \left[\lambda \begin{pmatrix} \int_0^1 B_{k-r} B'_{k-r} & O \\ O & O \end{pmatrix} \right. \\
&\quad \left. - \left(\int_0^1 B_{k-r} dB'_{k-r} \int_0^1 (dB_{k-r}) B'_{k-r} & O \\ O & \Psi_* \end{pmatrix} \right) \right] \\
&= \det \left(\lambda \int_0^1 B_{k-r} B'_{k-r} - \int_0^1 B_{k-r} dB'_{k-r} \int_0^1 (dB_{k-r}) B'_{k-r} \right) \det (-\Psi_*) .
\end{aligned} \quad (88)$$

However, generalized eigenvalue problem (83) is equivalent to

$$0 = \det \left(T^{-1} \Xi' \widehat{S}_{1,1} \Xi - \rho \Xi' \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \Xi \right), \text{ where } \rho = \frac{1}{T\lambda} \quad (89)$$

In this form the result of Anderson, Brons and Jensen (1983, Lemma 2) is applicable, because the right-hand side matrix in (86) is non-singular. Thus, the ordered solutions $\widehat{\rho}_1 \leq \widehat{\rho}_2 \leq \dots \leq \widehat{\rho}_k$ of (89) converge in distribution to the ordered solutions of

$$0 = \det \left(\begin{pmatrix} \int_0^1 B_{k-r} B'_{k-r} & O \\ O & O \end{pmatrix} - \rho \begin{pmatrix} \int_0^1 B_{k-r} dB'_{k-r} \int_0^1 (dB_{k-r}) B'_{k-r} & O \\ O & \Psi_* \end{pmatrix} \right).$$

It is now easy to verify that $(T\widehat{\lambda}_{r+1}, \dots, T\widehat{\lambda}_k)'$ converges in distribution to $(\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_{k-r})'$, where $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \dots \geq \widetilde{\lambda}_{k-r}$ are the solutions of (88).

Summarizing, the following result has been shown:

Theorem 4. *Under Assumptions 1-2 and VECM (38), the ordered solutions $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \dots \geq \widehat{\lambda}_k$ of the generalized eigenvalue problem (49) satisfy*

$$p \lim_{T \rightarrow \infty} (\widehat{\lambda}_1, \dots, \widehat{\lambda}_k)' = (\bar{\lambda}_1, \dots, \bar{\lambda}_r, 0, 0, \dots, 0)',$$

where $1 > \bar{\lambda}_1 \geq \dots \geq \bar{\lambda}_r > 0$ are constants⁵, and

$$(T\widehat{\lambda}_{r+1}, \dots, T\widehat{\lambda}_k)' \xrightarrow{d} (\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_{k-r})', \quad (90)$$

where $\widetilde{\lambda}_1 \geq \dots \geq \widetilde{\lambda}_{k-r}$ are the ordered solutions of the generalized eigenvalue problem

$$\det \left(\lambda \int_0^1 B_{k-r} B'_{k-r} - \int_0^1 B_{k-r} dB'_{k-r} \int_0^1 (dB_{k-r}) B'_{k-r} \right) = 0. \quad (91)$$

Consequently, under the null hypothesis $H_0(r)$ that the cointegrating rank is $r < k$ the LR statistic $LR_T(r|r+1)$ [see (55)] converges in distribution to $\widetilde{\lambda}_1$, whereas under the alternative hypothesis $H_1(r_1 > r)$ that the actual cointegrating rank r_1 is larger than r ,

$$p \lim_{T \rightarrow \infty} LR_T(r|r+1)/T = -\ln(1 - \bar{\lambda}_{r+1}) > 0.$$

⁵Recall that $\bar{\lambda}_1, \dots, \bar{\lambda}_r$ are the solutions of eigenvalue problem (84).

Moreover, under $H_0(r)$ the trace test statistic $LR_T(r|k)$ [see (56)] converges in distribution to

$$\sum_{i=1}^{k-r} \tilde{\lambda}_i = \text{trace} \left(\int_0^1 (dB_{k-r}) B'_{k-r} \left(\int_0^1 B_{k-r} B'_{k-r} \right)^{-1} \int_0^1 B_{k-r} dB'_{k-r} \right), \quad (92)$$

whereas under $H_1(r_1 > r)$,

$$p \lim_{T \rightarrow \infty} LR_T(r|k)/T = - \sum_{i=r+1}^{r_1} \ln(1 - \bar{\lambda}_i) > 0.$$

Remark: The proof of (90) in this lecture note is different from the original proof by Johansen. Johansen (1995, page 159) uses the fact that

$$\begin{aligned} \det [(\beta, \beta_\perp)' S(\rho) (\beta, \beta_\perp)] &= \det \begin{pmatrix} \beta' S(\rho) \beta & \beta' S(\rho) \beta_\perp \\ \beta'_\perp S(\rho) \beta & \beta'_\perp S(\rho) \beta_\perp \end{pmatrix} \\ &= \det (\beta' S(\rho) \beta) \det \left(\beta'_\perp \left(S(\rho) - S(\rho) \beta (\beta' S(\rho) \beta)^{-1} \beta' S(\rho) \right) \beta_\perp \right) \\ &= \det \left(-\beta' \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta + o_p(1) \right) \det \left(\rho \cdot \beta_\perp T^{-1} \widehat{S}_{1,1} \beta_\perp - \beta'_\perp \widehat{S}_{1,0} \widehat{N} \widehat{S}_{0,1} \beta_\perp \right) \end{aligned}$$

where

$$S(\rho) = \rho T^{-1} \widehat{S}_{1,1} - \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1}, \quad \rho = T\lambda = O_p(1),$$

and

$$\begin{aligned} \widehat{N} &= \widehat{S}_{0,0}^{-1} - \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta \left(\beta' \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta \right)^{-1} \beta' \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \\ &= \alpha_\perp (\alpha'_\perp \Omega \alpha_\perp)^{-1} \alpha'_\perp + o_p(1) \end{aligned}$$

Since by Lemmas 5 and 6,

$$\begin{aligned} \beta'_\perp T^{-1} \widehat{S}_{1,1} \beta_\perp &\xrightarrow{d} \int_0^1 B_{k-r} B'_{k-r}, \\ \beta_\perp \widehat{S}_{1,0} \widehat{N} \widehat{S}_{0,1} \beta_\perp &\xrightarrow{d} \int_0^1 B_{k-r} dB'_{k-r} \int_0^1 (dB_{k-r}) B'_{k-r}, \end{aligned}$$

the result (90) follows.

4.4 The VECM(p) case with intercepts due to initial values

The assumption $\pi_0 = 0$ is crucial for the results in Theorem 4. In the cases (30) and (37) the results of Theorem 4 change. I will only demonstrate this for the case (30). For the case (37), see Johansen (1995, Theorem 11.1). Recall from (35) that in VECM (30) the vector of intercept π_0 is due to initial values. Therefore, Δy_t is still a zero-mean stationary process. The only assumption that has to be dropped is the assumption that $u_t = 0$ for $t < 1$.

To demonstrate how the results in Theorem 4 change in the case (30) with $\pi_0 \neq 0$, write this model as

$$\begin{aligned} x_t &= \alpha\beta'y_{t-1} + (\pi_0, \Pi) \begin{pmatrix} 1 \\ X_{t-1} \end{pmatrix} + C_0 u_t \\ &= \alpha\beta'y_{t-1} + \Pi_* \tilde{X}_{t-1} + C_0 u_t \end{aligned}$$

where

$$\Pi_* = (\pi_0, \Pi), \quad \tilde{X}_{t-1} = (1, X'_{t-1})',$$

with X_{t-1} and Π defined by (39).

The main changes occur in the limiting distributions of the matrices $\hat{S}_{0,1}$ and $\hat{S}_{1,1}$.

4.4.1 The matrix $\hat{S}_{0,1}$

The matrix (44) now becomes

$$\begin{aligned} \hat{S}_{0,1} &= \alpha \left(\frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} y'_{t-1} \right) \\ &\quad - \alpha \left(\frac{1}{T} \sum_{t=1}^T \beta' y_{t-1} \tilde{X}'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} y'_{t-1} \right) \\ &\quad + C_0 \left(\frac{1}{T} \sum_{t=1}^T u_t y'_{t-1} \right) \\ &\quad - C_0 \left(\frac{1}{T} \sum_{t=1}^T u_t \tilde{X}'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} y'_{t-1} \right) \end{aligned} \tag{93}$$

The problem is that (64) does no longer hold if we replace X_{t-1} by \tilde{X}_{t-1} , so that the last term in (93) is no longer of order $o_p(1)$. Instead, we now have that

Lemma 7. *Under Assumption 1-2 and VECM (30),*

$$\begin{aligned} & \left(\frac{1}{T} \sum_{t=1}^T u_t \tilde{X}'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} y'_{t-1} \right) \\ &= \frac{1}{T} \sum_{t=1}^T u_t \frac{1}{T} \sum_{t=1}^T y'_{t-1} + o_p(1) \xrightarrow{d} B(1) \int_0^1 B(x)' dx C(1)' . \end{aligned}$$

Proof: It follows from the easy convergence result (see the Appendix)

$$\frac{1}{T\sqrt{T}} \sum_{t=1}^T y_{t-1} \xrightarrow{d} C(1) \int_0^1 B(x) dx \quad (94)$$

and (64) that

$$\begin{aligned} \begin{pmatrix} 1/\sqrt{T} & 0' \\ 0 & I_{k(p-1)} \end{pmatrix} \frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} y'_{t-1} &= \begin{pmatrix} \frac{1}{T\sqrt{T}} \sum_{t=1}^T y'_{t-1} \\ \frac{1}{T} \sum_{t=1}^T X_{t-1} y'_{t-1} \end{pmatrix} \\ &\xrightarrow{d} \begin{pmatrix} \int_0^1 B(x)' dx C(1)' \\ M_* \end{pmatrix} \end{aligned}$$

Moreover, it is easy to verify that under Assumption 1,

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T u_t \tilde{X}'_{t-1} \begin{pmatrix} \sqrt{T} & 0' \\ 0 & I_{k(p-1)} \end{pmatrix} &= \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T u_t, \frac{1}{T} \sum_{t=1}^T u_t X'_{t-1} \right) \\ &\xrightarrow{d} (B(1), O_{k,k(p-1)}) . \end{aligned}$$

Furthermore,

$$\begin{aligned} & \begin{pmatrix} 1/\sqrt{T} & 0' \\ 0 & I_{k(p-1)} \end{pmatrix} \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right)^{-1} \begin{pmatrix} \sqrt{T} & 0' \\ 0 & I_{k(p-1)} \end{pmatrix} \quad (95) \\ &= \left(\left(\begin{pmatrix} 1/\sqrt{T} & 0' \\ 0 & I_{k(p-1)} \end{pmatrix} \frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \begin{pmatrix} \sqrt{T} & 0' \\ 0 & I_{k(p-1)} \end{pmatrix} \right) \right)^{-1} \\ &= \left(\frac{1}{\frac{1}{T} \sum_{t=1}^T X_{t-1}} \frac{\frac{1}{T} \sum_{t=1}^T X'_{t-1}}{\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1}} \right)^{-1}, \end{aligned}$$

Since under Assumption 1, $p \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_{t-1} = 0$ and $p \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} = \Sigma_{XX}$, it follows from (95) that

$$\begin{aligned} p \lim_{T \rightarrow \infty} & \left(\begin{array}{cc} 1/\sqrt{T} & 0' \\ 0 & I_{k(p-1)} \end{array} \right) \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right)^{-1} \left(\begin{array}{cc} \sqrt{T} & 0' \\ 0 & I_{k(p-1)} \end{array} \right) \\ & = \left(\begin{array}{cc} 1 & 0' \\ 0 & \Sigma_{XX}^{-1} \end{array} \right). \end{aligned}$$

Lemma 7 follows straightforwardly from these convergence results. Q.E.D.

The previous result (66) now becomes

$$\alpha'_\perp \hat{S}_{0,1} = \alpha'_\perp C_0 \left(\frac{1}{T} \sum_{t=1}^T u_t \left(y_{t-1} - \frac{1}{T} \sum_{j=1}^T y_{j-1} \right)' \right) + o_p(1),$$

where

$$\frac{1}{T} \sum_{t=1}^T u_t \left(y_{t-1} - \frac{1}{T} \sum_{j=1}^T y_{j-1} \right)' \xrightarrow{d} \left(\int_0^1 (dB) B' - B(1) \int_0^1 B(x)' dx \right) C(1)'.$$

Note that, with $S_{t-1} = \sum_{j=1}^{t-1} u_j$ and $\bar{S}_{-1} = (1/T) \sum_{t=1}^T S_{t-1}$,

$$\frac{1}{T} \sum_{t=1}^T u_t (S_{t-1} - \bar{S}_{-1})' \xrightarrow{d} \int_0^1 (dB) B' - B(1) \int_0^1 B(x)' dx \quad (96)$$

Since

$$(S_{[xT]-1} - \bar{S}_{-1}) / \sqrt{T} \xrightarrow{d} B(x) - \int_0^1 B(y) dy = \bar{B}(x), \quad (97)$$

say, where $\bar{B}(x)$ is known as a demeaned k -variate standard Brownian motion, the right-hand side of (96) will be denoted by

$$\int_0^1 (dB) \bar{B}' \equiv \int_0^1 (dB) B' - B(1) \int_0^1 B(x)' dx. \quad (98)$$

With this change of notation, (73) reads

$$\begin{aligned} \Theta' & \left(\begin{array}{c} \beta'_\perp \\ \beta' \end{array} \right) \hat{S}_{1,0} \hat{S}_{0,0}^{-1} \hat{S}_{0,1} (\beta_\perp, \beta) \Theta \xrightarrow{d} \\ & \left(\begin{array}{cc} \beta_\perp C(1) \int_0^1 \bar{B} dB' C'_0 \alpha_\perp (\alpha'_\perp \Omega \alpha_\perp)^{-1} \alpha'_\perp C_0 \int_0^1 (dB) \bar{B}' C(1)' \beta_\perp & O \\ O & \Psi_* \end{array} \right) \end{aligned} \quad (99)$$

where Ψ_* is a nonrandom $r \times r$ matrix, similar (but not equal) to (75), and Θ defined similar to (70).

4.4.2 The matrix $\hat{S}_{1,1}$

The matrix (45) now becomes

$$\begin{aligned}\hat{S}_{1,1} &= \frac{1}{T} \sum_{t=1}^T y_{t-1} y'_{t-1} \\ &\quad - \left(\frac{1}{T} \sum_{t=1}^T y_{t-1} \tilde{X}'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} y'_{t-1} \right).\end{aligned}$$

Again, the result (77) does no longer hold, because similar to Lemma 7 we have that

Lemma 8. *Under Assumption 1 and VECM (30),*

$$\begin{aligned}&\frac{1}{T} \left(\frac{1}{T} \sum_{t=1}^T y_{t-1} \tilde{X}'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} y'_{t-1} \right) \\ &\xrightarrow{d} C(1) \int_0^1 B(x) dx \int_0^1 B(x)' dx C(1)'.\end{aligned}$$

Hence, (77) now becomes,

$$\begin{aligned}&\frac{1}{T} \hat{S}_{1,1} \xrightarrow{d} C(1) \left(\int_0^1 B(x) B(x)' dx - \int_0^1 B(x) dx \int_0^1 B(x)' dx \right) C(1)' \\ &= C(1) \int_0^1 \left(B(x) - \int_0^1 B(y) dy \right) \left(B(x) - \int_0^1 B(y) dy \right)' dx C(1)' \\ &= C(1) \int_0^1 \overline{B}(x) \overline{B}(x)' dx C(1)'.\end{aligned}$$

4.4.3 Limiting distributions of the general eigenvalues and the LR test statistics

It is now easy to verify that

Theorem 5. *If we change (91) to*

$$\det \left(\lambda \int_0^1 \overline{B}_{k-r} \overline{B}'_{k-r} - \int_0^1 \overline{B}_{k-r} dB'_{k-r} \int_0^1 (dB_{k-r}) \overline{B}'_{k-r} \right) = 0$$

and (92) to

$$\sum_{i=1}^{k-r} \tilde{\lambda}_i = \text{trace} \left(\int_0^1 (dB_{k-r}) \bar{B}'_{k-r} \left(\int_0^1 \bar{B}_{k-r} \bar{B}'_{k-r} \right)^{-1} \int_0^1 \bar{B}_{k-r} dB'_{k-r} \right),$$

where⁶

$$\bar{B}_{k-r}(x) = B_{k-r}(x) - \int_0^1 B_{k-r}(y) dy,$$

then the results of Theorem 4 carry over to VECM (30).

5 Asymptotic properties of the ML estimators of α , β and Ω

The partially concentrated log-likelihood (43) can be written as

$$\begin{aligned} \ln L_T(\alpha, \beta, \Omega) &= \max_{\Pi} \ln L_T(\alpha, \beta, \Pi, \Omega) \\ &= -\frac{1}{2} T \cdot \text{trace} \left(\Omega^{-1} \frac{1}{T} \sum_{t=1}^T (R_{0,t} - \alpha \beta' R_{1,t}) (R_{0,t} - \alpha \beta' R_{1,t})' \right) \\ &\quad - \frac{1}{2} T \cdot \ln (\det \Omega) - T \cdot k \ln (\sqrt{2\pi}) \\ &= -\frac{1}{2} T \cdot \text{trace} \left(\Omega^{-1} \left(\hat{S}_{0,0} - \alpha \beta' \hat{S}_{1,0} - \hat{S}_{0,1} \beta \alpha' + \alpha \beta' \hat{S}_{1,1} \beta \alpha' \right) \right) \quad (100) \\ &\quad - \frac{1}{2} T \cdot \ln (\det \Omega) - T \cdot k \ln (\sqrt{2\pi}) \end{aligned}$$

with corresponding ML estimators

$$(\hat{\alpha}, \hat{\beta}, \hat{\Omega}) = \arg \max_{\alpha, \beta, \Omega} \ln L_T(\alpha, \beta, \Omega) \quad (101)$$

Although $\hat{\alpha}$ and $\hat{\beta}$ themselves are not unique, $\hat{\alpha} \hat{\beta}'$ is unique, and therefore $\hat{\Omega}$ is unique. The same applies to α and β , of course. Nevertheless, after suitable normalization these ML estimators are consistent:

⁶Again, $\int_0^1 \bar{B}_{k-r} \bar{B}'_{k-r}$ is a short-hand notation for $\int_0^1 \bar{B}_{k-r}(x) \bar{B}_{k-r}(x)' dx$.

Theorem 6. Let $(\hat{\alpha}, \hat{\beta}, \hat{\Omega})$ be ML estimators of (α, β, Ω) . Without loss of generality we may assume that

$$\hat{\beta}'\hat{\beta} = O_p(1), \quad (\hat{\beta}'\hat{\beta})^{-1} = O_p(1), \quad \beta'\hat{\beta} = O_p(1), \quad (102)$$

and that the columns of $\hat{\beta}$ and β have been rescaled such that

$$\det(\hat{\beta}'\hat{\beta}) = \det(\beta'\beta) = 1. \quad (103)$$

Then under Assumptions 1-2 and VECM (30),

$$\left(\det(\beta'\hat{\beta})\right)^2 = 1 + O_p(T^{-1}). \quad (104)$$

Consequently, $\tilde{\beta} = \hat{\beta}(\beta'\hat{\beta})^{-1}(\beta'\beta)$ exists with probability converging to 1.

The matrix $\tilde{\beta}$ of normalized estimated cointegrating vectors is super consistent:

$$\tilde{\beta} - \beta = \beta_\perp U_T, \text{ with } U_T = O_p(T^{-1}), \quad (105)$$

Moreover, $\tilde{\alpha} = \hat{\alpha}(\beta'\beta)^{-1}(\beta'\hat{\beta})$ is a consistent estimator of α and $\hat{\Omega}$ is a consistent estimator of Ω .

Proof: Appendix

Note that $(\tilde{\alpha}, \tilde{\beta}, \hat{\Omega})$ also maximizes the log-likelihood (100). The first-order conditions involved are

$$O_{r,k} = \tilde{\alpha}'\hat{\Omega}\left(\hat{S}_{0,1} - \tilde{\alpha}\tilde{\beta}'\hat{S}_{1,1}\right) \quad (106)$$

$$O_{k,r} = \left(\hat{S}_{0,1} - \tilde{\alpha}\tilde{\beta}'\hat{S}_{1,1}\right)\tilde{\beta} \quad (107)$$

$$\hat{\Omega} = \hat{S}_{0,0} - \tilde{\alpha}\tilde{\beta}'\hat{S}_{1,0} - \hat{S}_{0,1}\tilde{\beta}\tilde{\alpha}' + \tilde{\alpha}\tilde{\beta}'\hat{S}_{1,1}\tilde{\beta}\tilde{\alpha}' \quad (108)$$

This is not too hard to verify for the case $r = 1$, but these conditions hold for $1 \leq r < k$ as well.

The limiting distribution of $T(\tilde{\beta} - \beta)$ is given in Theorem 7:

Theorem 7. Let $\tilde{\beta}$ and β_{\perp} be the defined as in Theorem 6 and Lemma 2, respectively, and let Assumptions 1-2 hold. Then in the case of VECM (38),

$$\begin{aligned} T \left(\tilde{\beta} - \beta \right) &\xrightarrow{d} \beta_{\perp} (\beta'_{\perp} C(1) C(1)' \beta_{\perp})^{-1/2} \\ &\times \left(\int_0^1 B_{k-r} B'_{k-r} \right)^{-1} \left(\int_0^1 B_{k-r} dB'_{\alpha} \right) (\alpha' \Omega^{-1} \alpha)^{-1/2} \end{aligned}$$

whereas in the case of VECM (30),

$$\begin{aligned} T \left(\tilde{\beta} - \beta \right) &\xrightarrow{d} \beta_{\perp} (\beta'_{\perp} C(1) C(1)' \beta_{\perp})^{-1/2} \\ &\times \left(\int_0^1 \bar{B}_{k-r} \bar{B}'_{k-r} \right)^{-1} \left(\int_0^1 \bar{B}_{k-r} dB'_{\alpha} \right) (\alpha' \Omega^{-1} \alpha)^{-1/2} \end{aligned}$$

where B_{α} is an r -variate standard Brownian motion which is independent of B_{k-r} and \bar{B}_{k-r} .

Proof: Appendix

6 Drift

The two cases considered so far only differ regarding the treatment of initial values; the assumption that Δy_t is a zero-mean stationary process satisfying the conditions in Assumption 1 has been maintained in both cases. Recall from (26) that in this case the cointegrating relationship takes the form $\beta'y_t = \beta'(y_0 - v_0) + \beta'v_t$, where $\beta'v_t$ is a zero-mean stationary process. In the bivariate case $k = 2$ the time series look like in Figure 1 below. In this case the time series run parallel and approximately horizontal, where the distance between the time series is due to the initial values $\beta'(y_0 - v_0)$.

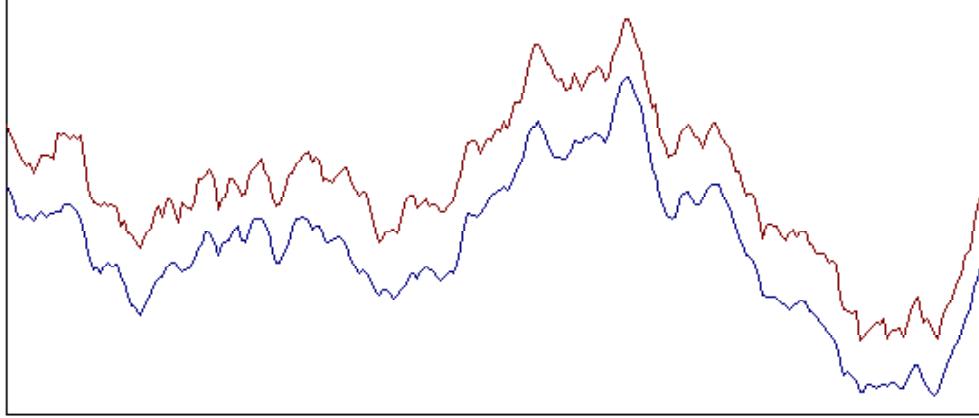


Figure 1. Cointegration when there is no drift

However, this pattern is rare for macro-economic time series. Most macro-economic time series have a trending pattern, due to drift: $E[\Delta y_t] = \mu \neq 0$, where μ is the vector of drift parameters. Thus, suppose that instead of (8) in Assumption 1,

$$y_t - y_{t-1} = \mu + C(L)u_t. \quad (109)$$

Then (12) becomes

$$\begin{aligned} y_t &= \sum_{j=1}^t (\mu + x_j) + y_0 = \mu t + \sum_{j=1}^t x_j + y_0 \\ &= (y_0 - v_0) + \mu t + C(1) \sum_{j=1}^t u_j + v_t \end{aligned} \quad (110)$$

Thus the expectation vector $\mu = E[\Delta y_t]$ now becomes a vector of trend parameters! Moreover,

$$\beta' y_t = \beta' (y_0 - v_0) + \beta' \mu t + \beta' v_t \quad (111)$$

so that $\beta' y_t$ is now trend stationary. If so, in the bivariate case $k = 2$ the time series look like in Figure 2.

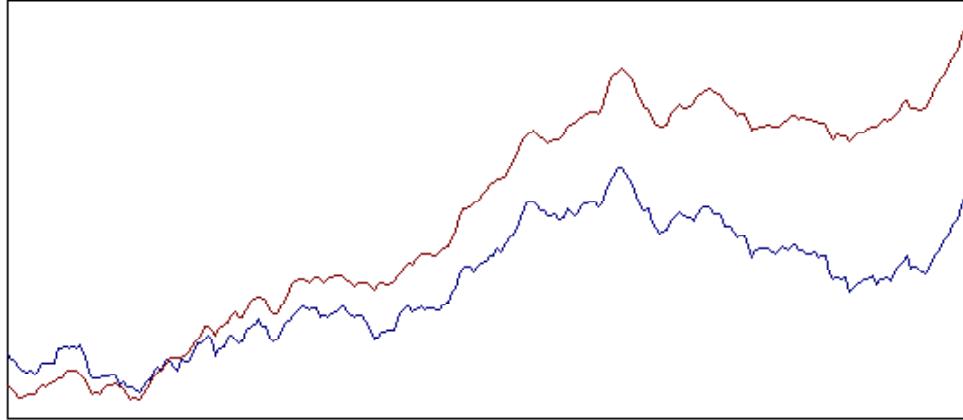


Figure 2: Drift, with trend in the cointegrating relationship

In this case the two time series drift apart, due to the time trend in the cointegrating relationship, and both are (upwards)⁷ sloping due to the drift parameters. However, a more common pattern is displayed in Figure 3:

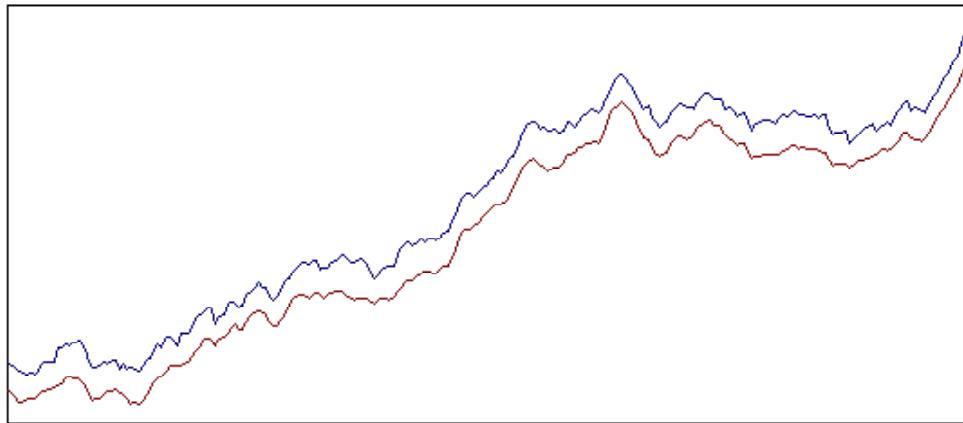


Figure 3: Drift, with constant in the cointegrating relationship

In this case the two time series run approximately parallel, but still (upwards) sloping due to the drift parameters. This pattern is only possible if in (111),

$$\beta' \mu = 0 \quad (112)$$

⁷Drift can also generate downwards sloping patterns, but that never happens for macroeconomic time series.

so that the cointegrating relationship becomes

$$\beta' y_t = \beta' (y_0 - v_0) + \beta' v_t. \quad (113)$$

It is easy to verify that the Granger representation theorem carries over if we replace y_t in (27) with $y_t - \mu.t$, which then gives rise to a VECM(p) model of the form

$$\Delta y_t - \mu = \pi_0 + \alpha \beta' (y_{t-1} - (t-1)\mu) + \sum_{j=1}^{p-1} \Pi_j (\Delta y_{t-j} - \mu) + C_0 u_t, \quad (114)$$

or equivalently,

$$\Delta y_t = \pi_{00} + \alpha \beta' (y_{t-1} - (t-1)\mu) + \sum_{j=1}^{p-1} \Pi_j \Delta y_{t-j} + C_0 u_t, \quad (115)$$

where

$$\pi_{00} = \pi_0 + \left(I_k - \sum_{j=1}^{p-1} \Pi_j \right) \mu.$$

This VECM(p) corresponds to the case (109) as displayed in Figure 2.

In the case (112) the model becomes

$$\Delta y_t = \pi_{00} + \alpha \beta' y_{t-1} + \sum_{j=1}^{p-1} \Pi_j \Delta y_{t-j} + C_0 u_t, \quad (116)$$

which at first sight looks the same as (27). However, the crucial difference is that now the vector π_{00} of intercepts depends on the drift parameters, which generates the sloping patterns as in Figure 3, whereas in the case (27) the vector of intercepts π_0 depends only on initial values. These initial values are not able to generate drift, as illustrated in Figure 1.

Due to the drift the results in Theorem 5 change, in different ways depending on whether condition (112) is imposed or not. See Johansen (1995, Theorem 11.1).

7 Appendix

7.1 Convergence of generalized eigenvalues

The following lemma is a corollary of Lemma 2 of Anderson, Brons and Jensen (1983):

Lemma A.1. Let A_T be a positive definite $m \times m$ matrix and let B_T be a symmetric $m \times m$ matrix, satisfying $(A_T, B_T) \xrightarrow{d} (A, B)$, where $\det(A) > 0$. Let $\lambda_{1,T} \geq \lambda_{2,T} \geq \dots \geq \lambda_{m,T}$ be the ordered solutions of the generalized eigenvalue problem $\det(\lambda A_T - B_T) = 0$, and let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$ be the ordered solutions of the generalized eigenvalue problem $\det(\lambda A - B) = 0$. Then $(\lambda_{1,T}, \lambda_{2,T}, \dots, \lambda_{m,T})' \xrightarrow{d} (\lambda_1, \lambda_2, \dots, \lambda_m)'$.

7.2 Derivation of (94)

Recall from (12) that $y_t = C(1) \sum_{j=1}^t u_t + v_t + y_0 - v_0$, so that

$$\begin{aligned} \frac{1}{T\sqrt{T}} \sum_{t=1}^T y_{t-1} &= C(1) \frac{1}{T} \sum_{t=1}^T \left(\frac{1}{\sqrt{T}} \sum_{j=1}^{t-1} u_t \right) + \frac{1}{T\sqrt{T}} \sum_{t=1}^T v_{t-1} \\ &\quad + \frac{1}{\sqrt{T}} (y_0 - v_0) \\ &= C(1) \frac{1}{T} \sum_{t=1}^T \left(\frac{1}{\sqrt{T}} \sum_{j=1}^{t-1} u_t \right) + O_p(T^{-1/2}) \\ &= C(1) \frac{1}{T} \sum_{t=1}^T \int_{t-1}^t \left(\frac{1}{\sqrt{T}} \sum_{j=1}^{[z]} u_t \right) dz + O_p(T^{-1/2}) \\ &= C(1) \int_0^T \left(\frac{1}{\sqrt{T}} \sum_{j=1}^{[z]} u_t \right) d(z/T) + O_p(T^{-1/2}) \\ &= C(1) \int_0^1 \left(\frac{1}{\sqrt{T}} \sum_{j=1}^{[xT]} u_t \right) dx + O_p(T^{-1/2}) \\ &\xrightarrow{d} C(1) \int_0^1 B(x) dx \end{aligned}$$

The latter follows from

$$\frac{1}{\sqrt{T}} \sum_{j=1}^{[xT]} u_t \Rightarrow B(x)$$

7.3 Proof of Theorem 6

7.3.1 Proof of (104)

Recall from (53) that the columns $\widehat{\beta}_1, \dots, \widehat{\beta}_r$ of $\widehat{\beta}$ are the eigenvectors corresponding to the r largest solutions of (49): $\widehat{\lambda}_i \widehat{S}_{1,1} \widehat{\beta}_i = \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \widehat{\beta}_i$, $i = 1, 2, \dots, r$. Hence, denoting $\widehat{\Lambda}_r = \text{diag}(\widehat{\lambda}_1, \dots, \widehat{\lambda}_r)$, it follows that

$$\widehat{S}_{1,1} \widehat{\beta} \widehat{\Lambda}_r = \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \widehat{\beta} \quad (117)$$

Next, multiply equation (117) from the left side by $(\beta, T^{-1/2} \beta_{\perp})$, and use the fact that

$$(\beta, T^{-1/2} \beta_{\perp})^{-1} \widehat{\beta} = \begin{pmatrix} \bar{\beta}' \widehat{\beta} \\ T^{1/2} \bar{\beta}'_{\perp} \widehat{\beta} \end{pmatrix} \quad (118)$$

Then (117) becomes

$$\begin{aligned} & \begin{pmatrix} \beta' \widehat{S}_{1,1} \beta & T^{-1/2} \beta' \widehat{S}_{1,1} \beta_{\perp} \\ T^{-1/2} \beta'_{\perp} \widehat{S}_{1,1} \beta & T^{-1} \beta'_{\perp} \widehat{S}_{1,1} \beta_{\perp} \end{pmatrix} \begin{pmatrix} \bar{\beta}' \widehat{\beta} \\ T^{1/2} \bar{\beta}'_{\perp} \widehat{\beta} \end{pmatrix} \widehat{\Lambda}_r \\ &= \begin{pmatrix} \beta' \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta & T^{-1/2} \beta' \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta_{\perp} \\ T^{-1/2} \beta'_{\perp} \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta & T^{-1} \beta'_{\perp} \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta_{\perp} \end{pmatrix} \begin{pmatrix} \bar{\beta}' \widehat{\beta} \\ T^{1/2} \bar{\beta}'_{\perp} \widehat{\beta} \end{pmatrix} \end{aligned}$$

which implies

$$\begin{aligned} & \beta' \widehat{S}_{1,1} \beta (\bar{\beta}' \widehat{\beta}) \widehat{\Lambda}_r + T^{-1/2} \beta' \widehat{S}_{1,1} \beta_{\perp} T^{1/2} \bar{\beta}'_{\perp} \widehat{\beta} \widehat{\Lambda}_r \\ &= T^{-1/2} \beta'_{\perp} \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta (\bar{\beta}' \widehat{\beta}) + T^{-1/2} \beta'_{\perp} \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta_{\perp} (\bar{\beta}'_{\perp} \widehat{\beta}) \end{aligned}$$

and

$$\begin{aligned} & T^{-1/2} \beta'_{\perp} \widehat{S}_{1,1} \beta (\bar{\beta}' \widehat{\beta}) \widehat{\Lambda}_r + T^{-1} \beta'_{\perp} \widehat{S}_{1,1} \beta_{\perp} T^{1/2} \bar{\beta}'_{\perp} \widehat{\beta} \widehat{\Lambda}_r \\ &= T^{-1/2} \beta'_{\perp} \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta (\bar{\beta}' \widehat{\beta}) + T^{-1/2} \beta'_{\perp} \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta_{\perp} \bar{\beta}'_{\perp} \widehat{\beta} \end{aligned}$$

The latter equality can be rewritten as

$$\begin{aligned} & T \bar{\beta}'_{\perp} \widehat{\beta} = \left(T^{-1} \beta'_{\perp} \widehat{S}_{1,1} \beta_{\perp} \right)^{-1} \quad (119) \\ & \times \left(\beta'_{\perp} \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta (\bar{\beta}' \widehat{\beta}) \widehat{\Lambda}_r^{-1} + \beta'_{\perp} \widehat{S}_{1,0} \widehat{S}_{0,0}^{-1} \widehat{S}_{0,1} \beta_{\perp} \bar{\beta}'_{\perp} \widehat{\beta} \widehat{\Lambda}_r^{-1} \right. \\ & \quad \left. - \beta'_{\perp} \widehat{S}_{1,1} \beta (\bar{\beta}' \widehat{\beta}) \right) \end{aligned}$$

It follows now from Lemmas 5 and 6 and Theorem 4 that the right-hand side of (119) is of order $O_p(1)$, hence $T\bar{\beta}'_{\perp}\hat{\beta} = O_p(1)$ and thus

$$\beta'_{\perp}\hat{\beta} = O_p(T^{-1}). \quad (120)$$

We can always write $\hat{\beta}$ as

$$\hat{\beta} = \beta(\beta'\beta)^{-1}(\beta'\hat{\beta}) + \beta_{\perp}(\beta'_{\perp}\beta_{\perp})^{-1}(\beta'_{\perp}\hat{\beta}), \quad (121)$$

Hence

$$I_r = (\hat{\beta}'\hat{\beta})^{-1/2}(\hat{\beta}'\beta)(\beta'\beta)^{-1}(\beta'\hat{\beta})(\hat{\beta}'\hat{\beta})^{-1/2} + O_p(T^{-1}), \quad (122)$$

where the $O_p(T^{-1})$ term follows from (102) and (120). Now (104) follows easily from (122) and the normalizations (103).

7.3.2 Proof of (105)

Recall that $(\beta'_{\perp}\hat{\beta})^{-1} = (\det(\beta'_{\perp}\hat{\beta}))^{-1}(\beta'_{\perp}\hat{\beta})_{Adj}$, where $(\beta'_{\perp}\hat{\beta})_{Adj}$ is the adjoint of $\beta'_{\perp}\hat{\beta}$, i.e., the matrix of cofactors of $(\beta'_{\perp}\hat{\beta})'$. Since $\beta'_{\perp}\hat{\beta} = O_p(1)$ implies $(\beta'_{\perp}\hat{\beta})_{Adj} = O_p(1)$, it follows from (104) that

$$(\beta'_{\perp}\hat{\beta})^{-1} = O_p(1).$$

It follows now from (120) and (121) that

$$\tilde{\beta} = \hat{\beta}(\beta'\hat{\beta})^{-1}(\beta'\beta) = \beta + \beta_{\perp}U_T,$$

where

$$U_T = (\beta'_{\perp}\beta_{\perp})^{-1}(\beta'_{\perp}\hat{\beta})(\beta'\hat{\beta})^{-1}(\beta'\beta) = O_p(T^{-1}).$$

7.3.3 Consistency of $\tilde{\alpha}$ and $\hat{\Omega}$

Recall from (79) and Lemma 8 that $\beta'\hat{S}_{1,1} = O_p(1)$, and from the proofs of Lemmas 5 and 7 that $\hat{S}_{0,1} = O_p(1)$. Hence it follows from Lemma 9 and (105)

that in the case (38),

$$\begin{aligned}\tilde{\beta}' \hat{S}_{1,1} \tilde{\beta} &= (\beta' + O_p(T^{-1})) \hat{S}_{1,1} (\beta + O_p(T^{-1})) \\ &= \beta' \hat{S}_{1,1} \beta + O_p(T^{-1}) = \Sigma_{\beta\beta}^* + o_p(1)\end{aligned}\quad (123)$$

$$\begin{aligned}\hat{S}_{0,1} \tilde{\beta} &= \hat{S}_{0,1} (\beta + O_p(T^{-1/2})) = \hat{S}_{0,1} \beta + O_p(T^{-1}) \\ &= \alpha \Sigma_{\beta\beta}^* + o_p(1)\end{aligned}\quad (124)$$

where the last equality in (123) follows from Lemma 6 and the last equality in (124) follows from Lemma 5. It follows therefore from (107) that

$$\tilde{\alpha} = \hat{S}_{0,1} \tilde{\beta} \left(\tilde{\beta}' \hat{S}_{1,1} \tilde{\beta} \right)^{-1} = \alpha + o_p(1) \quad (125)$$

This result carries over to the case of VECM (30).

It follows now from (108), (123), (124), (125) and Lemma 4 that

$$\hat{\Omega} = \hat{S}_{0,0} - \alpha \Sigma_{\beta\beta}^* \alpha' + o_p(1) = \Omega + o_p(1) \quad (126)$$

Again, this result carries over to the case of VECM (30).

7.4 Proof of Theorem 7

It follows from (106) that

$$\begin{aligned}O_{r,k-r} &= \tilde{\alpha}' \hat{\Omega}^{-1} \left(\hat{S}_{0,1} - \tilde{\alpha} \tilde{\beta}' \hat{S}_{1,1} \right) \beta_{\perp} \\ &= \tilde{\alpha}' \hat{\Omega}^{-1} \left(\hat{S}_{0,1} - \alpha \beta' \hat{S}_{1,1} - (\tilde{\alpha} - \alpha) \tilde{\beta}' \hat{S}_{1,1} - \alpha (\tilde{\beta} - \beta)' \hat{S}_{1,1} \right) \beta_{\perp} \\ &= \tilde{\alpha}' \hat{\Omega}^{-1} \left(\hat{S}_{0,1} - \alpha \beta' \hat{S}_{1,1} \beta_{\perp} \right) \beta_{\perp} - \tilde{\alpha}' \hat{\Omega}^{-1} (\tilde{\alpha} - \alpha) (\tilde{\beta} - \beta)' \hat{S}_{1,1} \beta_{\perp} \\ &\quad - \tilde{\alpha}' \hat{\Omega}^{-1} (\tilde{\alpha} - \alpha) \beta' \hat{S}_{1,1} \beta_{\perp} - \tilde{\alpha}' \hat{\Omega}^{-1} \alpha (\tilde{\beta} - \beta)' \hat{S}_{1,1} \beta_{\perp}\end{aligned}$$

Substituting (105) in this equation and multiplying from the right side by $(\tilde{\alpha}' \hat{\Omega}^{-1} \alpha)^{-1}$ yield

$$\begin{aligned}U'_T \left(\beta'_{\perp} \hat{S}_{1,1} \beta_{\perp} \right) &= \left(\tilde{\alpha}' \hat{\Omega}^{-1} \alpha \right)^{-1} \tilde{\alpha}' \hat{\Omega}^{-1} \left(\hat{S}_{0,1} - \alpha \beta' \hat{S}_{1,1} \right) \beta_{\perp} \\ &\quad - \left(\tilde{\alpha}' \hat{\Omega}^{-1} \alpha \right)^{-1} \tilde{\alpha}' \hat{\Omega}^{-1} (\tilde{\alpha} - \alpha) T U'_T \left(T^{-1} \beta'_{\perp} \hat{S}_{1,1} \beta_{\perp} \right) \\ &\quad - \left(\tilde{\alpha}' \hat{\Omega}^{-1} \alpha \right)^{-1} \tilde{\alpha}' \hat{\Omega}^{-1} (\tilde{\alpha} - \alpha) \beta' \hat{S}_{1,1} \beta_{\perp}\end{aligned}$$

Since by (79), $\beta' \hat{S}_{1,1} \beta_\perp = O_p(1)$, it follows from Lemma 9 that the last two terms are of order $o_p(1)$. Moreover, $\hat{S}_{0,1} \beta_\perp - \alpha \beta' \hat{S}_{1,1} \beta_\perp = O_p(1)$ and $(\tilde{\alpha}' \hat{\Omega}^{-1} \alpha)^{-1} \tilde{\alpha}' \hat{\Omega}^{-1} = (\alpha' \Omega^{-1} \alpha)^{-1} \alpha' \Omega^{-1} + o_p(1)$. Thus

$$\begin{aligned} T.U_T &= \left(T^{-1} \beta'_\perp \hat{S}_{1,1} \beta_\perp \right)^{-1} \beta'_\perp \left(\hat{S}_{1,0} - \hat{S}_{1,1} \beta \alpha' \right) \Omega^{-1} \alpha (\alpha' \Omega^{-1} \alpha)^{-1} \\ &\quad + o_p(1) \end{aligned}$$

If follows from Lemma 2, Lemma 6 and Lemma A.2 below that in the case of VECM (38),

$$T.U_T \xrightarrow{d} \left(\int_0^1 B_{k-r} B'_{k-r} \right)^{-1} \int_0^1 B_{k-r} dB'_\alpha (\alpha' \Omega^{-1} \alpha)^{-1/2}$$

where

$$B_\alpha = (\alpha' \Omega^{-1} \alpha)^{-1/2} \alpha' \Omega^{-1} C_0 B$$

is an r variate standard Brownian motion, which is independent of $B_{k-r} = (\alpha'_\perp \Omega \alpha_\perp)^{-1/2} \alpha'_\perp C_0 B$, whereas in the case of VECM (30),

$$\begin{aligned} T.U_T &\xrightarrow{d} \left(\int_0^1 \bar{B}_{k-r} \bar{B}'_{k-r} \right)^{-1} \left(\int_0^1 B_{k-r}(x) dB_\alpha(x)' - \int_0^1 B_{k-r}(x) dx B_\alpha(1)' \right) \\ &\quad \times (\alpha' \Omega^{-1} \alpha)^{-1/2} \\ &= \left(\int_0^1 \bar{B}_{k-r} \bar{B}'_{k-r} \right)^{-1} \left(\int_0^1 \bar{B}_{k-r} dB'_\alpha \right) (\alpha' \Omega^{-1} \alpha)^{-1/2} \end{aligned}$$

where \bar{B}_{k-r} is defined in Theorem 5 and the equality follows from

$$\begin{aligned} \int_0^1 \bar{B}_{k-r} dB'_\alpha &= \int_0^1 \left(B_{k-r}(x) - \int_0^1 B_{k-r}(y) dy \right) dB_\alpha(x)' \\ &= \int_0^1 B_{k-r}(x) dB_\alpha(x)' - \int_0^1 B_{k-r}(y) dy \int_0^1 dB_\alpha(x)' \\ &= \int_0^1 B_{k-r}(x) dB_\alpha(x)' - \int_0^1 B_{k-r}(y) dy B_\alpha(1)' \end{aligned}$$

Theorem 7 now follows from (105).

Lemma A.2. Under Assumptions 1-2,

$$\widehat{S}_{1,0} - \widehat{S}_{1,1}\beta\alpha' \xrightarrow{d} C(1) \left(\int_0^1 BdB' \right) C'_0 \quad (127)$$

in the case of VECM (38), and

$$\widehat{S}_{1,0} - \widehat{S}_{1,1}\beta\alpha' \xrightarrow{d} C(1) \left(\int_0^1 BdB' - \int_0^1 B(x)dx B(1)' \right) C'_0 \quad (128)$$

in the case of VECM (30).

Proof: First, consider the case of VECM (38): $x_t = \alpha\beta'y_{t-1} + \Pi X_{t-1} + C_0 u_t$.

Recall that $\widehat{S}_{0,1} = \frac{1}{T} \sum_{t=1}^T R_{0,t} R'_{1,t}$ and $\widehat{S}_{1,1} = \frac{1}{T} \sum_{t=1}^T R_{1,t} R'_{1,t}$, so that

$$\widehat{S}_{1,0} - \widehat{S}_{1,1}\beta\alpha' = \frac{1}{T} \sum_{t=1}^T R_{1,t} (R'_{0,t} - R'_{1,t}\beta\alpha') ,$$

where $R_{1,t}$ is the residual of the regression of y_{t-1} on X_{t-1} : $R_{1,t} = y_{t-1} - \widehat{\Delta}X_{t-1}$, with

$$\widehat{\Delta} = \left(\frac{1}{T} \sum_{t=1}^T y_{t-1} X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} = O_p(1)$$

and $R_{0,t}$ is the residual of the regression of x_t on X_{t-1} : $R_{0,t} = x_t - \widehat{\Upsilon}X_{t-1}$, with

$$\widehat{\Upsilon} = \left(\frac{1}{T} \sum_{t=1}^T x_t X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} = O_p(1)$$

Hence, $R_{0,t} - \alpha\beta'R_{1,t}$ is the residual of the regression of $x_t - \alpha\beta'y_{t-1} = \Pi X_{t-1} + C_0 u_t$ on X_{t-1} . But the latter regression has the same residual as the regression of $C_0 u_t$ on X_{t-1} :

$$R_{0,t} - \alpha\beta'R_{1,t} = C_0 u_t - \widehat{\Gamma}X_{t-1}$$

where

$$\widehat{\Gamma} = C_0 \left(\frac{1}{T} \sum_{t=1}^T u_t X'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right)^{-1} = O_p(T^{-1/2}) \quad (129)$$

Thus

$$\begin{aligned}
\widehat{S}_{1,0} - \widehat{S}_{1,1}\beta\alpha' &= \frac{1}{T} \sum_{t=1}^T (y_{t-1} - \widehat{\Delta}X_{t-1}) (C_0 u_t - \widehat{\Gamma}X_{t-1})' \\
&= \frac{1}{T} \sum_{t=1}^T y_{t-1} u'_t C'_0 - T^{-1/2} \widehat{\Delta} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T X_{t-1} u'_t \right) C'_0 \\
&\quad - \left(\frac{1}{T} \sum_{t=1}^T y_{t-1} X'_{t-1} \right) \widehat{\Gamma}' + \widehat{\Delta} \left(\frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1} \right) \widehat{\Gamma}' \\
&= \frac{1}{T} \sum_{t=1}^T y_{t-1} u'_t C'_0 + O_p(T^{-1/2})
\end{aligned}$$

where the $O_p(T^{-1/2})$ is due to (129), $\frac{1}{\sqrt{T}} \sum_{t=1}^T X_{t-1} u'_t = O_p(1)$, Theorem 3 and Lemma 3. It follows now from Theorem 2 that $\widehat{S}_{1,0} - \widehat{S}_{1,1}\beta\alpha' \xrightarrow{d} C(1) \left(\int_0^1 B dB' \right) C'_0$.

Next, consider the case of VECM (30) with $\pi_0 \neq 0$.

Recall that this model can be written as $x_t = \alpha\beta'y_{t-1} + \Pi_* \tilde{X}_{t-1} + C_0 u_t$ where $\Pi_* = (\pi_0, \Pi)$, $\tilde{X}_{t-1} = (1, X'_{t-1})'$. Then $R_{1,t} = y_{t-1} - \widehat{\Delta} \tilde{X}_{t-1}$, where

$$\begin{aligned}
\widehat{\Delta} &= \left(\frac{1}{T} \sum_{t=1}^T y_{t-1} \tilde{X}'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right)^{-1} \\
&= \left(\frac{1}{T} \sum_{t=1}^T y_{t-1}, \frac{1}{T} \sum_{t=1}^T y_{t-1} X'_{t-1} \right) \left(\frac{1}{\bar{X}_{-1}} \frac{\bar{X}'_{-1}}{\widehat{\Sigma}_{XX}} \right)^{-1}
\end{aligned}$$

where

$$\bar{X}_{-1} = \frac{1}{T} \sum_{t=1}^T X_{t-1}, \quad \widehat{\Sigma}_{XX} = \frac{1}{T} \sum_{t=1}^T X_{t-1} X'_{t-1},$$

It is a standard linear algebra exercise to verify that

$$\left(\frac{1}{\bar{X}_{-1}} \frac{\bar{X}'_{-1}}{\widehat{\Sigma}_{XX}} \right)^{-1} = \left(\begin{array}{cc} \widehat{\sigma} & -\widehat{\sigma} \bar{X}'_{-1} \widehat{\Sigma}_{XX}^{-1} \\ -\widehat{\sigma} \widehat{\Sigma}_{XX}^{-1} \bar{X}_{-1} & \widehat{\Sigma}_{XX}^{-1} + \widehat{\sigma} \widehat{\Sigma}_{XX}^{-1} \bar{X}_{-1} \bar{X}'_{-1} \widehat{\Sigma}_{XX}^{-1} \end{array} \right)$$

where

$$\widehat{\sigma} = \left(1 - \bar{X}'_{-1} \widehat{\Sigma}_{XX}^{-1} \bar{X}_{-1} \right)^{-1}$$

hence

$$\begin{aligned}
\tilde{\Delta} &= \left(\hat{\sigma} \frac{1}{T} \sum_{t=1}^T y_{t-1} - \hat{\sigma} \frac{1}{T} \sum_{t=1}^T y_{t-1} X'_{t-1} \hat{\Sigma}_{XX}^{-1} \bar{X}_{-1}, \right. \\
&\quad \left. - \hat{\sigma} \frac{1}{T} \sum_{t=1}^T y_{t-1} (\bar{X}'_{-1} \hat{\Sigma}_{XX}^{-1}) + \frac{1}{T} \sum_{t=1}^T y_{t-1} X'_{t-1} (\hat{\Sigma}_{XX}^{-1} + \hat{\sigma} \hat{\Sigma}_{XX}^{-1} \bar{X}_{-1} \bar{X}'_{-1} \hat{\Sigma}_{XX}^{-1}) \right) \\
&= \sqrt{T} \tilde{\Delta}_1 + \tilde{\Delta}_2
\end{aligned} \tag{130}$$

where

$$\begin{aligned}
\tilde{\Delta}_1 &= \left(\frac{1}{T\sqrt{T}} \sum_{t=1}^T y_{t-1} \right) \left(\hat{\sigma}, -\hat{\sigma} \bar{X}'_{-1} \hat{\Sigma}_{XX}^{-1} \right) \\
\tilde{\Delta}_2 &= \left(\frac{1}{T} \sum_{t=1}^T y_{t-1} X'_{t-1} \right) \left(-\hat{\sigma} \hat{\Sigma}_{XX}^{-1} \bar{X}_{-1}, \hat{\Sigma}_{XX}^{-1} + \hat{\sigma} \hat{\Sigma}_{XX}^{-1} \bar{X}_{-1} \bar{X}'_{-1} \hat{\Sigma}_{XX}^{-1} \right)
\end{aligned}$$

Moreover, $R_{0,t} - \alpha\beta'R_{1,t}$ is now the residual of the regression of $C_0 u_t$ on \tilde{X}_{t-1} : $R_{0,t} - \alpha\beta'R_{1,t} = C_0 u_t - \tilde{\Gamma} X_{t-1}$, where similar to (130),

$$\begin{aligned}
\tilde{\Gamma} &= C_0 \left(\frac{1}{T} \sum_{t=1}^T u_t \tilde{X}'_{t-1} \right) \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right)^{-1} \\
&= T^{-1/2} \tilde{\Gamma}_1 + T^{-1/2} \tilde{\Gamma}_2
\end{aligned}$$

with

$$\begin{aligned}
\tilde{\Gamma}_1 &= C_0 \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T u_t \right) \left(\hat{\sigma}, -\hat{\sigma} \bar{X}'_{-1} \hat{\Sigma}_{XX}^{-1} \right) \\
\tilde{\Gamma}_2 &= C_0 \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T u_t X'_{t-1} \right) \left(-\hat{\sigma} \hat{\Sigma}_{XX}^{-1} \bar{X}_{-1}, \hat{\Sigma}_{XX}^{-1} + \hat{\sigma} \hat{\Sigma}_{XX}^{-1} \bar{X}_{-1} \bar{X}'_{-1} \hat{\Sigma}_{XX}^{-1} \right)
\end{aligned}$$

Note that by Assumption 1 and Lemma 3,

$$\begin{aligned}
p \lim_{T \rightarrow \infty} \bar{X}_{-1} &= E[X_{t-1}] = 0 \\
p \lim_{T \rightarrow \infty} \hat{\Sigma}_{XX} &= E[X_{t-1} X'_{t-1}] = \Sigma_{XX} \\
p \lim_{T \rightarrow \infty} \hat{\sigma} &= 1
\end{aligned}$$

and recall that

$$\begin{aligned} \frac{1}{T\sqrt{T}} \sum_{t=1}^T y_{t-1} &\xrightarrow{d} C(1) \int_0^1 B(x) dx, \\ \frac{1}{T} \sum_{t=1}^T y_{t-1} X'_{t-1} &= O_p(1), \\ \frac{1}{\sqrt{T}} \sum_{t=1}^T u_t &\xrightarrow{d} B(1), \\ \frac{1}{\sqrt{T}} \sum_{t=1}^T u_t X'_{t-1} &= O_p(1), \end{aligned}$$

so that

$$\begin{aligned} \tilde{\Delta}_1 &= \left(\frac{1}{T\sqrt{T}} \sum_{t=1}^T y_{t-1} \right) (1, O_{1,(p-1)k}) + o_p(1) \xrightarrow{d} C(1) \int_0^1 B(x) dx \\ \tilde{\Delta}_2 &= O_p(1) \\ \tilde{\Gamma}_1 &= C_0 \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T u_t \right) (1, O_{1,(p-1)k}) + o_p(1) \xrightarrow{d} C_0 B(1) \\ \tilde{\Gamma}_2 &= C_0 \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T u_t X'_{t-1} \right) \begin{pmatrix} 1 & O_{1,k} \\ O_{k,1} & \Sigma_{XX}^{-1} \end{pmatrix} (O_{1,k}, I_k) + o_p(1) \end{aligned}$$

It follows from the previous part that

$$\frac{1}{T} \sum_{t=1}^T y_{t-1} u'_t \xrightarrow{d} C(1) \int_0^1 B dB'$$

Moreover, it is not too hard to verify that

$$\begin{aligned} \tilde{\Delta}_1 \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \tilde{X}_{t-1} u'_t \right) C'_0 &\xrightarrow{d} C(1) \int_0^1 B(x) dx B(1)' C'_0, \\ \left(\frac{1}{T\sqrt{T}} \sum_{t=1}^T y_{t-1} \tilde{X}'_{t-1} \right) \tilde{\Gamma}'_1 &\xrightarrow{d} C(1) \int_0^1 B(x) dx B(1)' C'_0 \\ \tilde{\Delta}_1 \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right) \tilde{\Gamma}'_1 &\xrightarrow{d} C(1) \int_0^1 B(x) dx B(1)' C'_0 \end{aligned}$$

$$\begin{aligned}\tilde{\Delta}_1 \left(\frac{1}{T} \sum_{t=1}^T \tilde{X}_{t-1} \tilde{X}'_{t-1} \right) \tilde{\Gamma}'_2 &= o_p(1) \\ \left(\frac{1}{T\sqrt{T}} \sum_{t=1}^T y_{t-1} \tilde{X}'_{t-1} \right) \tilde{\Gamma}'_2 &= o_p(1)\end{aligned}$$

It follows therefore straightforwardly that

$$\begin{aligned}\hat{S}_{1,0} - \hat{S}_{1,1}\beta\alpha' &= \frac{1}{T} \sum_{t=1}^T \left(y_{t-1} - T^{1/2} \tilde{\Delta}_1 \tilde{X}_{t-1} - \tilde{\Delta}_2 \tilde{X}_{t-1} \right) \\ &\quad \times \left(u'_t C'_0 - T^{-1/2} \tilde{X}'_{t-1} \tilde{\Gamma}'_1 - T^{-1/2} \tilde{X}'_{t-1} \tilde{\Gamma}'_2 \right) \\ &\xrightarrow{d} C(1) \left(\int_0^1 BdB'C'_0 - \int_0^1 B(x)dx B(1)' \right) C'_0\end{aligned}$$

8 References

- Anderson, S. A., H. K. Brons and S. T. Jensen (1983), "Distribution of Eigenvalues in Multivariate Statistical Analysis", *Annals of Statistics* 11, 392-415.
- Bierens, H. J. (1994), *Topics in Advanced Econometrics: Estimation, Testing and Specification of Cross-Section and Time Series Models*, Cambridge University Press.
- Bierens, H. J. (2004), *Introduction to the Mathematical and Statistical Foundations of Econometrics*, Cambridge University Press.
- Engle, R. F. and C. W. J. Granger (1987), "Cointegration and Error Correction: Representation, Estimation, and Testing", *Econometrica* 55, 251-276.
- Johansen, S. (1988), "Statistical Analysis of Cointegrated Vectors", *Journal of Economic Dynamics and Control* 12, 231-254.
- Johansen, S. (1991), "Estimation and Hypothesis Testing of Cointegrated Vectors in Gaussian Vector Autoregressive Models", *Econometrica* 59, 1551-1580.
- Johansen, S. (1995), *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press.
- Phillips, P. C. B. (1986), "Time Series Regression with a Unit Root", *Econometrica* 55, 277-301.
- Phillips, P. C. B. (1988), "Weak Convergence to the Matrix Stochastic Integral $\int_0^1 BdB'$ ", *Journal of Multivariate Time Series Analysis* 24, 252-264.

Introduction to Hilbert Spaces

Herman J. Bierens

Pennsylvania State University

(June 24, 2007)

1. Vector spaces

The notion of a vector space should be known from linear algebra:

Definition 1. Let V be a set endowed with two operations, the operation "addition", denoted by "+", which maps each pair (x,y) in $V \times V$ into V , and the operation "scalar multiplication", denoted by a dot (\cdot) , which maps each pair (c,x) in $\mathbb{R} \times V$ [or $\mathbb{C} \times V$] into V . Thus, a scalar is a real or complex number. The set V is called a real [complex] **vector space** if the addition and multiplication operations involved satisfy the following rules, for all x, y and z in V , and all scalars c, c_1 and c_2 in \mathbb{R} [\mathbb{C}]: :

- (a) $x + y = y + x;$
- (b) $x + (y + z) = (x + y) + z;$
- (c) There is a unique zero vector 0 in V such that $x + 0 = x;$
- (d) For each x there exists a unique vector $-x$ in V such that $x + (-x) = 0$;¹
- (e) $1 \cdot x = x;$
- (f) $(c_1 c_2) \cdot x = c_1 \cdot (c_2 \cdot x);$
- (g) $c \cdot (x + y) = c \cdot x + c \cdot y;$
- (h) $(c_1 + c_2) \cdot x = c_1 \cdot x + c_2 \cdot x.$

It is trivial to verify that the Euclidean space \mathbb{R}^n is a real vector space. However, the notion of a vector space is much more general. For example, let V be the space of all continuous functions on \mathbb{R} , with pointwise addition and scalar multiplication defined the same way as for real numbers. Then it is easy to verify that this space is a real vector space.

¹

In the sequel, $x + (-y)$ will be denoted by $x - y$.

Another (but weird) example of a vector space is the space V of positive real numbers endowed with the "addition" operation $x + y = x.y$ and the "scalar multiplication" $c.x = x^c$. In this case the null vector 0 is the number 1 , and $-x = 1/x$.

Definition 2. A subspace V_0 of a vector space V is a non-empty subset of V which satisfies the following two requirements:

- (a) For any pair x, y in V_0 , $x + y$ is in V_0 ;
- (b) For any x in V_0 and any scalar c , $c.x$ is in V_0 .

Thus, a subspace V_0 of a vector space is closed under linear combinations: any linear combination of elements in V_0 is an element of V_0 .

It is not hard to verify that a subspace of a vector space is a vector space itself, because the rules (a) through (h) in Definition 1 are inherited from the "host" vector space V . In particular, any subspace contains the null vector 0 , as follows from part (b) of Definition 2 with $c = 0$.

Definition 3. An inner product on a real vector space V is a real function $\langle x, y \rangle: V \times V \rightarrow \mathbb{R}$ such that for all x, y, z in V and all c in \mathbb{R} ,

- (1) $\langle x, y \rangle = \langle y, x \rangle$
- (2) $\langle cx, y \rangle = c\langle x, y \rangle$
- (3) $\langle x+y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- (4) $\langle x, x \rangle > 0$ when $x \neq 0$.

An inner product on a complex vector space is defined similarly. The inner product is then complex-valued, $\langle x, y \rangle: V \times V \rightarrow \mathbb{C}$. Condition (1) then becomes

$$(1^*) \quad \langle x, y \rangle = \overline{\langle y, x \rangle}^2,$$

and (2) now holds for all complex numbers c . Note that in both cases, $\langle x, x \rangle$ is real valued.

Finally, the norm of x in V is defined as $\|x\| = \sqrt{\langle x, x \rangle}$.

² The bar denotes the complex conjugate: for $z = a + i.b$, $\bar{z} = a - i.b$.

For example, in the space $C[0,1]$ of continuous real functions on $[0,1]$, the integral $\langle f,g \rangle = \int_0^1 f(t)g(t)dt$ is an inner product. Moreover, in the vector space of zero-mean random variables with finite second moments the expectation $\langle X,Y \rangle = E[X.Y]$ is an inner product.

As is well-known from linear algebra, for vectors $x, y \in \mathbb{R}^n$, $|x^T y| \leq \|x\| \cdot \|y\|$, which is known as the Cauchy-Schwarz inequality. This inequality carries over to general inner products:

Theorem 1. (Cauchy-Schwarz inequality) $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$.

Proof: Let the vector space involved be real. Then for any real λ ,

$$\begin{aligned} 0 &\leq \langle x + \lambda y, x + \lambda y \rangle = \langle x, x + \lambda y \rangle + \lambda \langle y, x + \lambda y \rangle = \langle x, x \rangle + 2\lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle \\ &= \|x\|^2 + 2\lambda \langle x, y \rangle + \lambda^2 \|y\|^2. \end{aligned}$$

Minimizing the latter to λ yields the result. The complex case is similar. Q.E.D.

Given the norm $\|x\| = \sqrt{\langle x, x \rangle}$, the following properties hold:

$$\|x\| > 0 \text{ if } x \neq 0; \tag{1}$$

$$\|cx\| = |c| \cdot \|x\|; \tag{2}$$

$$\|x+y\| \leq \|x\| + \|y\|. \quad [\text{Triangular inequality}] \tag{3}$$

The properties (1) and (2) follow trivially from Definition 3. In the case of a real vector space the triangular inequality (3) follows from

$$\begin{aligned} \|x+y\|^2 &= \langle x+y, x+y \rangle = \langle x, x+y \rangle + \langle y, x+y \rangle = \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \\ &= \|x\|^2 + 2\langle x, y \rangle + \|y\|^2 \leq \|x\|^2 + 2|\langle x, y \rangle| + \|y\|^2 \leq \|x\|^2 + 2\|x\| \cdot \|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2 \end{aligned}$$

where the last inequality is due to Theorem 1.

A norm can also be defined directly:

Definition 4. A norm on a vector space V is a mapping $\|\cdot\|: V \rightarrow [0, \infty)$ such that for all x and y in V and all scalars c , (1), (2) and (3) hold. A vector space endowed with a norm is called a normed

space.

A norm $\|.\|$ defines a metric $d(x,y) = \|x-y\|$ on V , i.e., a function that measures the “distance” between two elements x and y of V , for which (trivially) the following four properties hold. For all x, y and z in V ,

$$d(x,y) = d(y,x); \quad (4)$$

$$d(x,y) > 0 \text{ if } x \neq y; \quad (5)$$

$$d(x,x) = 0; \quad (6)$$

$$d(x,z) \leq d(x,y) + d(y,z) \text{ [Triangular inequality].} \quad (7)$$

More generally,

Definition 5. A metric on a space V is a mapping $d(.,.): V \times V \rightarrow [0,\infty)$ satisfying the properties (4) through (7) for all x, y and z in V . A space endowed with a metric is called a metric space.

In this definition the space V is not necessarily a vector space: Any space endowed with a metric is a metric space. Moreover, inner products and norms may not be defined on metric spaces.

Definition 6. A sequence of elements x_n of a metric space with metric $d(.,.)$ is called a Cauchy sequence if for every $\varepsilon > 0$ there exists an n_0 such that for all $k, m \geq n_0$, $d(x_k, x_m) < \varepsilon$.

The notion of a Cauchy sequence plays a critical role in defining Hilbert spaces. See the next section.

Theorem 2. Every Cauchy sequence in \mathbb{R}^ℓ or \mathbb{C}^ℓ , $\ell < \infty$, has a limit in the space involved.

Proof: Consider first the case \mathbb{R} . Let $\bar{x} = \limsup_{n \rightarrow \infty} x_n$, where x_n is a Cauchy sequence. I will show first that $\bar{x} < \infty$. There exists a subsequence n_k such that $\bar{x} = \lim_{k \rightarrow \infty} x_{n_k}$. Note that x_{n_k} is also a Cauchy sequence. For arbitrary $\varepsilon > 0$ there exists an index k_0 such that $|x_{n_k} - x_{n_m}| < \varepsilon$ if

$k, m \geq k_0$. Keeping k fixed and letting $m \rightarrow \infty$ it follows that $|x_{n_k} - \bar{x}| < \varepsilon$, hence $\bar{x} < \infty$. Similarly, $\underline{x} = \liminf_{n \rightarrow \infty} x_n > -\infty$. Now we can find an index k_0 and subsequences n_k and n_m such that for $k, m \geq k_0$, $|x_{n_k} - \bar{x}| < \varepsilon$, $|x_{n_m} - \underline{x}| < \varepsilon$, and $|x_{n_k} - x_{n_m}| < \varepsilon$, hence $|\underline{x} - \bar{x}| < 3\varepsilon$. Since ε is arbitrary, we must have $\underline{x} = \bar{x} = \lim_{n \rightarrow \infty} x_n$. Applying this argument to the real and imaginary parts of a complex Cauchy sequence the result for the case \mathbb{C} follows, and applying the argument to each component of a (complex) vector valued Cauchy sequence the result for the cases \mathbb{R}^ℓ and \mathbb{C}^ℓ follow. Q.E.D.

2. Hilbert spaces

A Euclidean space \mathbb{R}^n is a vector space endowed with the inner product $\langle x, y \rangle = x^T y$, norm $\|x\| = \sqrt{x^T x} = \sqrt{\langle x, x \rangle}$ and associated metric $\|x - y\|$, such that every Cauchy sequence takes a limit in \mathbb{R}^n . This makes \mathbb{R}^n a Hilbert space:

Definition 7. A Hilbert space H is a vector space endowed with an inner product and associated norm and metric, such that every Cauchy sequence in H has a limit in H .

A Hilbert space is also a Banach space:

Definition 8. A Banach space B is a normed space with associated metric $d(x, y) = \|x - y\|$ such that every Cauchy sequence in B has a limit in B .

The difference between a Banach space and a Hilbert space is the source of the norm. In the Hilbert space case the norm is defined via the inner product, $\|x\| = \sqrt{\langle x, x \rangle}$, whereas in the Banach space case the norm is defined directly, by Definition 4. Thus, a Hilbert space is a Banach space, but the other way around may not be true, because in some cases the norm cannot be associated with an inner product.

An example of a Hilbert space is the space $L^2(a, b)$:

Definition 9. The space $L^2(a,b)$ is the collection of Borel measurable real or complex valued square integrable functions f on (a,b) , i.e., $\int_a^b |f(t)|^2 dt < \infty$, endowed with inner product $\langle f, g \rangle = \int_a^b f(t) \overline{g(t)} dt$, and associated norm and metric

$$\|f\| = \sqrt{\int_a^b |f(t)|^2 dt}, \quad d(f, g) = \|f - g\| = \sqrt{\int_a^b |f(t) - g(t)|^2 dt},$$

respectively, where the integrals involved are Lebesgue integrals.

Note that in this case f and g are interpreted as being equal if they differ on (a,b) only on a set with Lebesgue measure zero. The proof that $L^2(a,b)$ is a Hilbert space will be given in Section 6 below.

Recall that vectors x and y in \mathbb{R}^n are orthogonal if $x^T y = 0$. More generally,

Definition 10. Elements x, y of a Hilbert space are orthogonal if $\langle x, y \rangle = 0$, also denoted by $x \perp y$, and orthonormal if in addition $\|x\| = 1$ and $\|y\| = 1$.

Note that in a Banach space orthogonality is a non-existing property, because an inner product is not defined on a Banach space. This is the main reason for working with Hilbert spaces.

Definition 11. An orthonormal sequence $e_n, n=1,2,3,\dots$ in a Hilbert space H is complete if the only member of H which is orthogonal to all e_n is the zero vector.

Theorem 3. Let $e_n, n=1,2,3,\dots$ be a complete orthonormal sequence in a Hilbert space H . Then for every x in H , $x = \sum_{n=1}^{\infty} \langle x, e_n \rangle e_n$ and $\|x\|^2 = \sum_{n=1}^{\infty} |\langle x, e_n \rangle|^2$.

Proof: Observe that

$$\begin{aligned} 0 &\leq \|x - \sum_{n=1}^m \langle x, e_n \rangle e_n\|^2 = \left\langle x - \sum_{n=1}^m \langle x, e_n \rangle e_n, x - \sum_{n=1}^m \langle x, e_n \rangle e_n \right\rangle \\ &= \left\langle x, x - \sum_{n=1}^m \langle x, e_n \rangle e_n \right\rangle - \sum_{n=1}^m \langle x, e_n \rangle \left\langle e_n, x - \sum_{n=1}^m \langle x, e_n \rangle e_n \right\rangle \\ &= \|x\|^2 - \sum_{n=1}^m |\langle x, e_n \rangle|^2 \end{aligned} \tag{8}$$

hence, letting $m \rightarrow \infty$, we have $\sum_{n=1}^{\infty} |\langle x, e_n \rangle|^2 \leq \|x\|^2 < \infty$. Therefore,

$$\lim_{m \rightarrow \infty} \sum_{n=m}^{\infty} |\langle x, e_n \rangle|^2 = 0 \quad (9)$$

Let $y_m = \sum_{n=1}^m \langle x, e_n \rangle e_n$ and note that

$$\|x - y_m\|^2 = \min_{\beta_1, \dots, \beta_m} \|x - \sum_{n=1}^m \beta_n e_n\|^2. \quad (10)$$

Thus, we can write

$$x = y_m + z_m = \sum_{n=1}^m \langle x, e_n \rangle e_n + z_m, \text{ where } \langle z_m, e_n \rangle = 0 \text{ for all } n \leq m, \quad (11)$$

hence $\langle y_m, z_m \rangle = 0$ and therefore

$$\|x\|^2 = \|y_m\|^2 + \|z_m\|^2. \quad (12)$$

Moreover,

$$\|y_m - y_k\|^2 = \sum_{n=\min(k,m)}^{\max(k,m)} |\langle x, e_n \rangle|^2 \leq \sum_{n=\min(k,m)}^{\infty} |\langle x, e_n \rangle|^2 \rightarrow 0 \quad (13)$$

for $\min(k,m) \rightarrow \infty$, hence it follows from (9) that y_m is a Cauchy sequence. Therefore, it follows from the properties of a Hilbert space that y_n converges to a limit $y \in H$:

$$\lim_{m \rightarrow \infty} \|y - y_m\| = \lim_{m \rightarrow \infty} \|\sum_{n=1}^m \langle x, e_n \rangle e_n\| = 0. \quad (14)$$

Thus, we can write y as

$$y = \sum_{n=1}^m \langle x, e_n \rangle e_n + u_m = y_m + u_m, \text{ where } \lim_{m \rightarrow \infty} \|u_m\| = 0. \quad (15)$$

This result gives rise to the notation

$$y = \sum_{n=1}^{\infty} \langle x, e_n \rangle e_n. \quad (16)$$

However, keep in mind that the actual definition of y is given by (15). On the other hand, without loss of generality we may treat y as the right-hand side of (16), as will be demonstrated below.

Let $z = x - y$. If $\langle z, e_k \rangle = 0$ for all k then it follows from the completeness of $\{e_k\}$ that $z = 0$, hence $x = y$. To prove this, note that by (15), $z = x - \sum_{n=1}^m \langle x, e_n \rangle e_n - u_m$ for all m . Then for all $m > k$,

$$\begin{aligned} \langle z, e_k \rangle &= \langle x, e_k \rangle - \sum_{n=1}^m \langle x, e_n \rangle \langle e_k, e_n \rangle - \langle e_k, u_m \rangle \\ &= \langle x, e_k \rangle - \langle x, e_k \rangle \langle e_k, e_k \rangle - \langle e_k, u_m \rangle \\ &= \langle x, e_k \rangle - \langle x, e_k \rangle - \langle e_k, u_m \rangle = -\langle e_k, u_m \rangle \end{aligned} \quad (17)$$

hence by Theorem 1 and (15),

$$|\langle z, e_k \rangle| = \lim_{m \rightarrow \infty} |\langle e_k, u_m \rangle| \leq \lim_{m \rightarrow \infty} \|u_m\| \cdot \|e_k\| = \lim_{m \rightarrow \infty} \|u_m\| = 0. \quad (18)$$

Thus, $x = y$. It follows now from (11) and (15) that $z_m = u_m$, hence

$$x = \sum_{n=1}^m \langle x, e_n \rangle e_n + u_m, \text{ where } \langle e_n, u_m \rangle = 0 \text{ for all } n \leq m, \text{ and } \lim_{m \rightarrow \infty} \|u_m\| = 0. \quad (19)$$

Again, this result is denoted by

$$x = \sum_{n=1}^{\infty} \langle x, e_n \rangle e_n. \quad (20)$$

This proves the first part of the theorem. The second part follows now easily from (8). Q.E.D.

Note that the result (18) could have been obtained more directly by treating y as the right-hand side of (16), because then $z = x - \sum_{n=1}^{\infty} \langle x, e_n \rangle e_n$, hence

$$\langle z, e_k \rangle = \langle x, e_k \rangle - \sum_{n=1}^{\infty} \langle x, e_n \rangle \langle e_k, e_n \rangle = \langle x, e_k \rangle - \langle x, e_k \rangle \langle e_k, e_k \rangle = \langle x, e_k \rangle - \langle x, e_k \rangle = 0.$$

Similarly, (19) implies that for any $v \in H$,

$$\begin{aligned} \langle x, v \rangle &= \lim_{m \rightarrow \infty} \sum_{n=1}^m \langle x, e_n \rangle \langle e_n, v \rangle + \lim_{m \rightarrow \infty} \langle u_m, v \rangle \\ &= \sum_{n=1}^{\infty} \langle x, e_n \rangle \langle e_n, v \rangle + \lim_{m \rightarrow \infty} \langle u_m, v \rangle = \sum_{n=1}^{\infty} \langle x, e_n \rangle \langle e_n, v \rangle \end{aligned} \quad (21)$$

where the last equality follows from the fact that by Theorem 1 and (19),

$$|\lim_{m \rightarrow \infty} \langle u_m, v \rangle| \leq \lim_{m \rightarrow \infty} \|u_m\| \cdot \|v\| = 0,$$

whereas the result (21) would have followed trivially if we has used (20).

Definition 12. *The coefficients $\langle x, e_n \rangle$ in the series representation $x = \sum_{n=1}^{\infty} \langle x, e_n \rangle e_n$ are called the Fourier coefficients of x .*

Definition 13. *Let A be a set of vectors in a Hilbert space H . Then $\text{Lin}(A)$ is the intersection of all sub-spaces of H which contain A , and $\text{Clin}(A)$ is the closure of $\text{Lin}(A)$.*

In other words, $\text{Lin}(A)$ is the set of all linear combinations of the elements of A . If A is a finite then $\text{Clin}(A) = \text{Lin}(A)$. The same applies if $A \subset \mathbb{R}^n$, because there are only a finite number of linear independent vectors in A , which can be made orthonormal..

Theorem 4. *Let e_n , $n = 1, 2, 3, \dots$ be an orthonormal sequence in a Hilbert space H . Then the following statements are equivalent.*

- (1) e_n is complete;
- (2) $H = \text{Clin}(\{e_n, n \geq 1\})$.

$$(3) \quad \text{For all } x \text{ in } H, \|x\|^2 = \sum_{n=1}^{\infty} |\langle x, e_n \rangle|^2.$$

Proof: (1) \Rightarrow (2) and (1) \Rightarrow (3) has already been proved in Theorem 3.

Proof of (3) \Rightarrow (1): Suppose that the sequence e_n is not complete. Then there exists a $z \neq 0$ in H such that $\langle z, e_n \rangle = 0$ for all n . But then by part (3), $\|z\|^2 = \sum_{n=1}^{\infty} |\langle z, e_n \rangle|^2 = 0$, which contradicts $z \neq 0$.

Proof of (2) \Rightarrow (1): Let $\langle z, e_n \rangle = 0$ for all n , and define $E = \{x \in H : \langle x, z \rangle = 0\}$. Then E contains all e_n , and for fixed z , E is a subspace, so that $\text{Lin}(\{e_n, n \geq 1\}) \subset E$. Moreover, it is not hard to verify from the continuity of the inner product that E is closed, hence $H = \text{Clin}(\{e_n, n \geq 1\}) \subset E$. But $z \in H$, and thus $z \in E$, and consequently, $\langle z, z \rangle = 0$. This proves (1). Q.E.D.

3. Fourier analysis

The following theorem implies that for every Borel measurable real or complex valued function f in $L^2(-\pi, \pi)$ has the series expansion

$$f(x) = (2\pi)^{-1} \sum_{n=-\infty}^{\infty} c_n \exp(i.n.x), \text{ where } c_n = \int_{-\pi}^{\pi} f(y) \exp(-i.n.y) dy. \quad (22)$$

Recall from the proof of Theorem 3 that (22) should be interpreted as:

$$\begin{aligned} \text{For all } m > 0, f(x) &= (2\pi)^{-1} \sum_{n=-m}^m c_n \exp(i.n.x) + \varepsilon_m(x), \text{ where} \\ &\lim_{m \rightarrow \infty} \int_{-\pi}^{\pi} |\varepsilon_m(x)|^2 dx = 0. \end{aligned} \quad (23)$$

Theorem 5. *The complex functions $e_n(x) = (2\pi)^{-1/2} \exp(i.n.x)$, $n = 0, \pm 1, \pm 2, \dots$ form a complete orthonormal sequence in $L^2(-\pi, \pi)$.*

The orthonormality of the sequence $\{e_n(x)\}$ follows from the fact that

$$\begin{aligned} \langle e_n, e_m \rangle &= (2\pi)^{-1} \int_{-\pi}^{\pi} \exp(i.(n-m).x) dx = (2\pi)^{-1} \int_{-\pi}^{\pi} \cos((n-m).x) dx + i.(2\pi)^{-1} \int_{-\pi}^{\pi} \sin((n-m).x) dx \\ &= (2\pi)^{-1} \int_{-\pi}^{\pi} \cos((n-m).x) dx = \begin{cases} \frac{\sin((n-m)\pi)}{(n-m)\pi} = 0 & \text{if } n \neq m, \\ 1 & \text{if } n = m. \end{cases} \end{aligned}$$

The rest of the proof of Theorem 5 is too long and is therefore given in the Appendix, Section A.1.

Since every function g in $L^2(a,b)$ can be converted to a function f in $L^2(-\pi,\pi)$, namely $f(x) = g(a+(b-a)(x+\pi)/(2\pi))$ and vice versa, $g(x) = f(\pi-2\pi(b-x)/(b-a))$, it follows from (22) that

$$g(x) = (2\pi)^{-1/2} \sum_{n=-\infty}^{\infty} c_n \exp[i.n.(\pi-2\pi(b-x)/(b-a))] \quad (24)$$

where

$$\begin{aligned} c_n &= (2\pi)^{-1/2} \int_{-\pi}^{\pi} f(y) \exp(-i.n.y) dy = (2\pi)^{-1/2} \int_{-\pi}^{\pi} g(a+(b-a)(y+\pi)/(2\pi)) \exp(-i.n.y) dy \\ &= \frac{\sqrt{2\pi}}{b-a} \int_a^b g(x) \exp[-i.n.(\pi-2\pi(b-x)/(b-a))] dx \end{aligned} \quad (25)$$

Therefore,

Theorem 6. Every function g in $L^2(a,b)$, $-\infty < a < b < \infty$, can be written as

$$g(x) = \sum_{n=-\infty}^{\infty} \gamma_n \exp[2\pi n.i.x/(b-a)]$$

$$\text{where } \gamma_n = (b-a)^{-1} \int_a^b g(x) \exp[-2\pi n.i.x/(b-a)] dx.$$

If g is real valued, then

$$\begin{aligned} \gamma_n &= (b-a)^{-1} \int_a^b g(x) \cos[2\pi n.i.x/(b-a)] dx - i(b-a)^{-1} \int_a^b g(x) \sin[2\pi n.i.x/(b-a)] dx \\ &= \alpha_n - i\beta_n, \end{aligned} \quad (26)$$

say, so that the result of Theorem 6 reads

$$\begin{aligned} g(x) &= \gamma_0 + \sum_{n=1}^{\infty} \gamma_n \exp[2\pi n.i.x/(b-a)] + \sum_{n=1}^{\infty} \overline{\gamma_n} \exp[-2\pi n.i.x/(b-a)] \\ &= \alpha_0 + 2 \sum_{n=1}^{\infty} \alpha_n \cos[2\pi n.x/(b-a)] + 2 \sum_{n=1}^{\infty} \beta_n \sin[2\pi n.x/(b-a)] \text{ a.e. on } (a,b). \end{aligned} \quad (27)$$

4. Functions of two or more variables

Let $g(x_1, x_2)$ be a function in $L^2((a_1, b_1) \times (a_2, b_2))$, $-\infty < a_j < b_j < \infty$, $j = 1, 2$. Since $\int_{a_1}^{b_1} \int_{a_2}^{b_2} |g(x_1, x_2)|^2 dx_1 dx_2 < \infty$, the set $N_2 = \{x_2 \in (a_2, b_2) : \int_{a_1}^{b_1} |g(x_1, x_2)|^2 dx_1 = \infty\}$ has Lebesgue measure zero. Consequently, for every fixed $x_2 \in (a_2, b_2) \setminus N_2$ the function $g(x_1, x_2)$ is a member of $L^2(a_1, b_1)$. Applying Theorem 6 then yields

$$g(x_1, x_2) = \sum_{n=-\infty}^{\infty} \gamma_n(x_2) \exp[2\pi n i x_1 / (b_1 - a_1)] \text{ a.e. on } (a_1, b_1), \quad (28)$$

where

$$\gamma_n(x_2) = (b_1 - a_1)^{-1} \int_{a_1}^{b_1} g(x_1, x_2) \exp[-2\pi n i x_1 / (b_1 - a_1)] dx_1. \quad (29)$$

But (29) is a member of $L^2(a_2, b_2)$, so that by Theorem 6,

$$\gamma_n(x_2) = \sum_{m=-\infty}^{\infty} \gamma_{n,m} \exp[2\pi m i x_2 / (b_2 - a_2)] \text{ a.e. on } (a_2, b_2), \quad (30)$$

where

$$\begin{aligned} \gamma_{n,m} &= (b_1 - a_1)^{-1} (b_2 - a_2)^{-1} \\ &\times \int_{a_1}^{b_1} \int_{a_2}^{b_2} g(x_1, x_2) \exp[-2\pi n i x_1 / (b_1 - a_1)] \exp[-2\pi m i x_2 / (b_2 - a_2)] dx_1 dx_2. \end{aligned} \quad (31)$$

Substituting (30) in (28) yields

$$g(x_1, x_2) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \gamma_{n,m} \exp[2\pi n i x_1 / (b_1 - a_1)] \exp[2\pi m i x_2 / (b_2 - a_2)]. \quad (32)$$

This result can be extended to functions of more than two variables:

Theorem 7. Let $C = \times_{j=1}^k (a_j, b_j)$, $-\infty < a_j < b_j < \infty$, let D be the diagonal matrix with diagonal elements $b_1 - a_1, \dots, b_k - a_k$, and denote by \mathbb{I}^k the set of k -dimensional vectors with integer components. Then for every function $g(x)$ in $L^2(C)$, $g(x) = \sum_{n \in \mathbb{I}^k} \gamma(n) \exp[2\pi i n^T D^{-1} x]$ where for $n \in \mathbb{I}^k$, $\gamma(n) = \det[D^{-1}] \int_C g(y) \exp[-2\pi i n^T D^{-1} y] dy$.

Again, we can write $\gamma(n) = \alpha(n) - i\beta(n)$, where

$$\begin{aligned} \alpha(n) &= \det[D^{-1}] \int_C g(y) \cos[2\pi n^T D^{-1} y] dy, \\ \beta(n) &= \det[D^{-1}] \int_C g(y) \sin[2\pi n^T D^{-1} y] dy, \end{aligned} \quad (33)$$

so that if g is real valued,

$$\begin{aligned}
g(x) &= \sum_{n \in \mathbb{I}^k} (\alpha(n) - i \cdot \beta(n)) (\cos[2\pi n^T D^{-1} x] + i \sin[2\pi n^T D^{-1} x]) \\
&= \alpha(0) + \sum_{n \in \mathbb{I}^k \setminus \{0\}} \alpha(n) \cos[2\pi n^T D^{-1} x] + \sum_{n \in \mathbb{I}^k \setminus \{0\}} \beta(n) \sin[2\pi n^T D^{-1} x].
\end{aligned} \tag{34}$$

5. Hilbert spaces of random variables

As mentioned before, for zero-mean random variables X and Y with finite second moments the expectation $E[X \cdot Y]$ can be interpreted as an inner product:

Theorem 8. *Let H be the space of zero-mean random variables with finite second moments defined on a common probability space $\{\Omega, \mathcal{F}, P\}$, endowed with the inner product $\langle X, Y \rangle = E[X \cdot Y]$, norm $\|X\| = \sqrt{E[X^2]}$ and metric $\|X - Y\|$. Then H is a Hilbert space.*

Proof: It is trivial to verify that the space of these random variables is a vector space. Therefore, we only need to show that every Cauchy sequence $X_n, n \geq 1$, has a limit in H , as follows. Since by Chebishev's inequality,

$$P[|X_n - X_m| > \varepsilon] \leq E[(X_n - X_m)^2]/\varepsilon^2 = \|X_n - X_m\|^2/\varepsilon^2 \rightarrow 0 \text{ as } n, m \rightarrow \infty$$

for every $\varepsilon > 0$, it follows that $|X_n - X_m| \rightarrow 0$ in probability as $n, m \rightarrow \infty$. Therefore, there exists a subsequence n_k such that $|X_{n_k} - X_m| \rightarrow 0$ a.s. as $n, m \rightarrow \infty$.³ The latter implies that there exists a null set N such that for every $\omega \in \Omega \setminus N$, $X_{n_k}(\omega)$ is a Cauchy sequence in \mathbb{R} , hence

$\lim_{k \rightarrow \infty} X_{n_k}(\omega) = X(\omega)$ exists for every $\omega \in \Omega \setminus N$. Now for every fixed m ,

$(X_{n_k} - X_m)^2 \rightarrow (X - X_m)^2$ a.s. as $k \rightarrow \infty$. By Fatou's lemma (see the Appendix, Section A.2) and the Cauchy property the latter implies that $\|X - X_m\|^2 = E[(X - X_m)^2] \leq \liminf_{k \rightarrow \infty} E[(X_{n_k} - X_m)^2] \rightarrow 0$ as $m \rightarrow \infty$. Moreover, it is easy to verify that $E[X] = 0$ and $E[X^2] < \infty$. Q.E.D.

Note that the result of Theorem 8 can be translated in the following way:

³ This follows from the fact that if $X_n \rightarrow 0$ in probability then every subsequence n_k contains a further subsequence $n_k(m)$ such that $X_{n_k(m)} \rightarrow 0$ a.s. as $m \rightarrow \infty$.

Corollary 1. Let $\{\Omega, \mathcal{F}, P\}$ be a probability space, and let $L_0^2(P)$ be the space of measurable functions⁴ $X(\omega): \Omega \rightarrow \mathbb{R}$ satisfying $\int X(\omega)^2 dP(\omega) < \infty$, $\int X(\omega) dP(\omega) = 0$, endowed with the inner product $\langle X, Y \rangle = \int X(\omega)Y(\omega) dP(\omega)$ and associated norm $\|X\| = \sqrt{\langle X, X \rangle}$ and metric $\|X - Y\|$. Then $L_0^2(P)$ is a Hilbert space.

Moreover, if we take $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}$ and $P = \mu$, where \mathcal{B} is the Euclidean Borel field and μ is a probability measure on \mathcal{B} , then Corollary 1 reads:

Corollary 2. The space $L_0^2(\mu)$ of Borel measurable real functions on \mathbb{R} satisfying $\int f(x)^2 d\mu(x) < \infty$ and $\int f(x) d\mu(x) = 0$, endowed with the inner product $\langle f, g \rangle = \int f(x)g(x) d\mu(x)$ and associated norm $\|f\| = \sqrt{\langle f, f \rangle}$ and metric $\|f - g\|$, is a Hilbert space.

6. The space $L^2(a,b)$

The results in Section 5 can be used to prove that the space $L^2(0,1)$ is a Hilbert space, which then implies that $L^2(a,b)$ is a Hilbert space. First, consider the subspace $L_0^2(0,1)$ of Borel measurable functions f in $L^2(0,1)$ satisfying $\int_0^1 f(x) dx = 0$. It follows straightforwardly from Corollary 2 that $L_0^2(0,1)$ is a Hilbert space. Now let f_n be a Cauchy sequence in $L^2(0,1)$. Then $g_n(x) = f_n(x) - \int_0^1 f_n(u) du$ is a Cauchy sequence in $L_0^2(0,1)$: there exists a g in $L_0^2(0,1)$ such that $\lim_{n \rightarrow \infty} \|g_n - g\| = 0$ because

$$\left| \int_0^1 f_k(u) du - \int_0^1 f_m(u) du \right| \leq \int_0^1 |f_k(u) - f_m(u)| du \leq \sqrt{\int_0^1 |f_k(u) - f_m(u)|^2 du} = \|f_k - f_m\|, \quad (35)$$

hence $\|g_k - g_m\| \leq 2\|f_k - f_m\|$. But inequality (35) also implies that $\int_0^1 f_n(u) du$ is a Cauchy sequence in the Hilbert space \mathbb{R} , and therefore

$$\mu = \lim_{n \rightarrow \infty} \int_0^1 f_n(u) du \in \mathbb{R}.$$

Next, let $f = g + \mu$. Then $f \in L^2(0,1)$ because $g \in L_0^2(0,1) \subset L^2(0,1)$ and the constant function μ is a member of $L^2(0,1)$. Moreover,

⁴ Recall that $X(\omega): \Omega \rightarrow \mathbb{R}$ is measurable if for all Borel sets B the sets $\{\omega \in \Omega: X(\omega) \in B\}$ are members of \mathcal{F} .

$$\|f_n - f\| \leq \|g_n - g\| + \left| \int_0^1 f_n(u) du - \mu \right| \rightarrow 0$$

as $n \rightarrow \infty$. Thus, $L^2(0,1)$ is a Hilbert space.

Along the same lines it is easy to show that:

Theorem 9. *The space $L^2(a,b)$ is a Hilbert space,*

and more generally that:

Theorem 10. *The space $L^2(\mu)$ of Borel measurable real functions on \mathbb{R} satisfying $\int f(x)^2 d\mu(x) < \infty$, endowed with the inner product $\langle f, g \rangle = \int f(x)g(x) d\mu(x)$ and associated norm $\|f\| = \sqrt{\langle f, f \rangle}$ and metric $\|f - g\|$, is a Hilbert space.*

Appendix

A.1 Proof of Theorem 5

The proof of Theorem 5 involves the following two steps:

- (1) Show that any continuous 2π -periodic real function f on $[-\pi, \pi]$, confined to $(-\pi, \pi)$, is contained in $\text{Clin}(\{e_n, n = 0, \pm 1, \pm 2, \dots\})$. Then the same applies to complex continuous 2π -periodic functions on $[-\pi, \pi]$, by applying the argument to the real and imaginary parts of f .
- (2) Show that the subset $C_P [-\pi, \pi]$ of continuous 2π -periodic functions on $[-\pi, \pi]$, confined to $(-\pi, \pi)$, is a dense subset of $L^2(-\pi, \pi)$, i.e., $L^2(-\pi, \pi)$ is the closure of $L^2(-\pi, \pi) \cap C_P [-\pi, \pi]$. Then the result follows from Theorem 4.

Step 1. Let f be a continuous *real* function on $[-\pi, \pi]$, and let $f_m = \sum_{n=-m}^m \langle f, e_n \rangle e_n$ and

$$F_m = (m+1)^{-1} \sum_{n=0}^m f_n.$$

Clearly, $F_m \in \text{Lin}(\{e_n, n = 0, \pm 1, \pm 2, \dots\})$. Therefore, $f \in \text{Clin}(\{e_n, n = 0, \pm 1, \pm 2, \dots\})$ if $F_m \rightarrow f$, so I will show that the latter is true.

Observe that $\langle f, e_n \rangle = \overline{\langle e_n, f \rangle} = (2\pi)^{-1/2} \int_{-\pi}^{\pi} f(x) \exp(-i.n.x) dx$, hence
 $f_m(y) = (2\pi)^{-1} \int_{-\pi}^{\pi} f(x) \sum_{n=-m}^m \exp(i.n.(y-x)) dx$

and thus

$$F_m(y) = (2\pi)^{-1} \int_{-\pi}^{\pi} f(x) \left((m+1)^{-1} \sum_{j=0}^m \sum_{n=-j}^j \exp(i.n.(y-x)) \right) dx = (2\pi)^{-1} \int_{-\pi}^{\pi} f(x) K_m(y-x) dx,$$

say, where

$$K_m(t) = (m+1)^{-1} \sum_{j=0}^m \sum_{n=-j}^j \exp(i.n.t), \quad t \in \mathbb{R}.$$

The latter function is known as the Fejer kernel. Note that for t such that $\exp(i.t) = 1$,

$$K_m(t) = (m+1)^{-1} \sum_{j=0}^m (1+2j) = 1 + m,$$

whereas for all t such that $\exp(i.t) \neq 1$,

$$\begin{aligned} K_m(t) &= (m+1)^{-1} \sum_{j=0}^m \left(1 + \sum_{n=1}^j \exp(i.n.t) + \sum_{n=1}^j \exp(-i.n.t) \right) \\ &= (m+1)^{-1} \sum_{j=0}^m \left(\sum_{n=0}^j (\exp(i.t))^n + \sum_{n=0}^j (\exp(-i.t))^n - 1 \right) \\ &= (m+1)^{-1} \sum_{j=0}^m \left(\frac{1 - \exp(i.t(j+1))}{1 - \exp(i.t)} + \frac{1 - \exp(-i.t(j+1))}{1 - \exp(-i.t)} - 1 \right) \\ &= \frac{1}{1 - \exp(i.t)} + \frac{1}{1 - \exp(-i.t)} - 1 \\ &\quad - (m+1)^{-1} \frac{\exp(i.t)}{1 - \exp(i.t)} \sum_{j=0}^m \exp(i.t.j) - (m+1)^{-1} \frac{\exp(-i.t)}{1 - \exp(-i.t)} \sum_{j=0}^m \exp(-i.t.j) \\ &= -(m+1)^{-1} \frac{\exp(i.t)}{1 - \exp(i.t)} \times \frac{1 - \exp(i.t.(m+1))}{1 - \exp(i.t.)} - (m+1)^{-1} \frac{\exp(-i.t)}{1 - \exp(-i.t)} \times \frac{1 - \exp(-i.t.(m+1))}{1 - \exp(-i.t.)} \\ &= (m+1)^{-1} \frac{2 - \exp(i.t.(m+1)) - \exp(-i.t.(m+1))}{(1 - \exp(i.t))(1 - \exp(-i.t.))} \\ &= (m+1)^{-1} \frac{2 - 2\cos(t.(m+1))}{2 - 2\cos(t)} = \frac{1}{m+1} \left(\frac{\sin((m+1)t/2)}{\sin(t/2)} \right)^2. \end{aligned}$$

Thus, $K_m(t) \geq 0$ and $K_m(t) = K_m(-t)$. Moreover, it is not hard to show that

$$\int_{-\pi}^{\pi} K_m(t) dt = 2\pi, \quad (36)$$

and that for any fixed $\delta \in (0, \pi)$,

$$\lim_{m \rightarrow \infty} \left(\int_{-\pi}^{-\delta} K_m(t) dt + \int_{\delta}^{\pi} K_m(t) dt \right) = 0. \quad (37)$$

Note that (36) implies, replacing t by $y - x$, that

$$\int_{y-\pi}^{y+\pi} K_m(y-x) dx = 2\pi,$$

hence for $y \in [-\pi, \pi]$,

$$f(y) = (2\pi)^{-1} \int_{y-\pi}^{y+\pi} f(y) K_m(y-x) dx. \quad (38)$$

Since f and K_m are both 2π -periodic, it follows for each y , $f(x)K_m(y-x)$ is 2π -periodic. Since a 2π -periodic function has the same integral over $[y-\pi, y+\pi]$ as over $[-\pi, \pi]$, we have

$$F_m(y) = (2\pi)^{-1} \int_{-\pi}^{\pi} f(x) K_m(y-x) dx = (2\pi)^{-1} \int_{y-\pi}^{y+\pi} f(x) K_m(x-y) dx \quad (39)$$

Hence, it follows from (38) and (39) that for any $\delta \in (0, \pi)$,

$$\begin{aligned} |F_m(y) - f(y)| &\leq (2\pi)^{-1} \int_{y-\pi}^{y+\pi} |f(x) - f(y)| K_m(x-y) dx \\ &= (2\pi)^{-1} \int_{y-\pi}^{y-\delta} |f(x) - f(y)| K_m(x-y) dx + (2\pi)^{-1} \int_{y-\delta}^{y+\delta} |f(x) - f(y)| K_m(x-y) dx \\ &\quad + (2\pi)^{-1} \int_{y+\delta}^{y+\pi} |f(x) - f(y)| K_m(x-y) dx \quad (40) \\ &= \pi^{-1} \sup_{-\pi \leq x \leq \pi} |f(x)| \int_{-\pi}^{-\delta} K_m(t) dt + \pi^{-1} \sup_{-\pi \leq x \leq \pi} |f(x)| \int_{\delta}^{\pi} K_m(t) dt \\ &\quad + (2\pi)^{-1} \int_{y-\delta}^{y+\delta} |f(x) - f(y)| K_m(x-y) dx \end{aligned}$$

Since f is bounded, it follows from (37) that the first and second integral at the right hand side of (40) converge to zero as $m \rightarrow \infty$, whereas the third term can be bounded by

$$\begin{aligned} (2\pi)^{-1} \int_{y-\delta}^{y+\delta} |f(x) - f(y)| K_m(x-y) dx &\leq \sup_{|x-y| \leq \delta} |f(x) - f(y)| (2\pi)^{-1} \int_{-\pi}^{\pi} K_m(t) dt \\ &= \sup_{|x-y| \leq \delta} |f(x) - f(y)|, \quad (41) \end{aligned}$$

where the equality follows from (36). Since continuous functions on a compact interval are uniformly continuous, it follows that f is uniformly continuous on $[-\pi, \pi]$, and since f is 2π -periodic it follows that f is uniformly continuous on \mathbb{R} . It follows therefore from (40) and (41) that

$$\lim_{m \rightarrow \infty} \sup_{-\pi \leq y \leq \pi} |F_m(y) - f(y)| = 0. \quad (42)$$

Note that (42) implies $\lim_{m \rightarrow \infty} \|F_m - f\| = 0$, which completes step 1 of the proof.

Step 2. First, note that every continuous real function on $[-\pi, \pi]$ can be written as a limit of 2π -periodic functions on $[-\pi, \pi]$. In particular, define for a continuous real function f on $[-\pi, \pi]$,

$$f_n(x) = \begin{cases} f(x) & \text{if } -\pi + n^{-1} \leq x \leq \pi - n^{-1}, \\ n \cdot f(-\pi + n^{-1})(x + \pi) & \text{if } 0 \leq x < -\pi + n^{-1}, \\ n \cdot f(\pi - n^{-1})(\pi - x) & \text{if } \pi - n^{-1} < x \leq \pi. \end{cases}$$

Then f_n is continuous on $[-\pi, \pi]$, and since $f_n(-\pi) = f_n(\pi) (= 0)$, it can be extended to the real line as a 2π -periodic function. Moreover,

$$\|f_n - f\|^2 = \int_{-\pi}^{-\pi+1/n} (f_n(x) - f(x))^2 dx + \int_{\pi-1/n}^{\pi} (f_n(x) - f(x))^2 dx \leq \frac{2}{n} \sup_{-\pi \leq x \leq \pi} f(x)^2 \rightarrow 0$$

as $n \rightarrow \infty$. A similar result holds for complex-valued continuous functions. Consequently, the closure of the space $C_p [-\pi, \pi]$ of continuous 2π -periodic functions on $[-\pi, \pi]$ contains all the continuous functions on $[-\pi, \pi]$.

Next, let B be an arbitrary Borel subset of $[-\pi, \pi]$, and let for $n = 1, 2, 3, \dots$,

$$f_n(x) = \exp\left(-n^{-1} \inf_{y \in \bar{B}} |x - y|\right) - \exp\left(-n^{-1} \inf_{y \in \bar{B} \setminus B} |x - y|\right), \quad x \in [-\pi, \pi],$$

where \bar{B} is the closure of B . Note that $f_n(x)$ is continuous on $[-\pi, \pi]$ because $\inf_{y \in \bar{B}} |x - y|$ is a continuous function on $[-\pi, \pi]$, and so is $\inf_{y \in \bar{B} \setminus B} |x - y|$. To see this, observe that for any pair $x_1, x_2 \in [-\pi, \pi]$, $\inf_{y \in \bar{B}} |x_1 - y| \leq |x_1 - x_2| + \inf_{y \in \bar{B}} |x_2 - y|$ and $\inf_{y \in \bar{B}} |x_2 - y| \leq |x_1 - x_2| + \inf_{y \in \bar{B}} |x_1 - y|$, hence

$$\left| \inf_{y \in \bar{B}} |x_1 - y| - \inf_{y \in \bar{B}} |x_2 - y| \right| \leq |x_1 - x_2|,$$

and similarly,

$$\left| \inf_{y \in \bar{B} \setminus B} |x_1 - y| - \inf_{y \in \bar{B} \setminus B} |x_2 - y| \right| \leq |x_1 - x_2|.$$

Since $\lim_{n \rightarrow \infty} f_n(x) = I(x \in B)$, where $I(\cdot)$ is the indicator function, and therefore by bounded convergence, $\lim_{n \rightarrow \infty} \|f_n(x) - I(x \in B)\| = 0$, the function $I(x \in B)$ is contained in the closure of the space $C_p [-\pi, \pi]$, and consequently, all real simple functions on $[-\pi, \pi]$ are included as well. Finally, since a Borel measurable function is a limit of a sequence of simple functions, it is easy to verify from the dominated convergence theorem that all Borel measurable functions f satisfying $\int_{-\pi}^{\pi} |f(t)|^2 dt < \infty$ are included in the closure of the space $C_p [-\pi, \pi]$. This completes step 2 of the proof. Q.E.D.

A.2 Fatou's Lemma

Fatou's lemma states:

For a sequence $X_n, n \geq 1$, of non-negative random variables, $E[\liminf_{n \rightarrow \infty} X_n] \leq \liminf_{n \rightarrow \infty} E[X_n]$.

Proof: Put $X = \liminf_{n \rightarrow \infty} X_n$ and let $\varphi(x)$ be a simple function satisfying $0 \leq \varphi(x) \leq x$.

Moreover, put $Y_n = \min(\varphi(X), X_n)$. Then $Y_n \xrightarrow{p} \varphi(X)$ because for arbitrary $\varepsilon > 0$,

$$P[|Y_n - \varphi(X)| > \varepsilon] = P[X_n < \varphi(X) - \varepsilon] \leq P[X_n < X - \varepsilon] \rightarrow 0.$$

Since $E[\varphi(X)] < \infty$ because φ is a simple function, and $Y_n \leq \varphi(X)$, it follows from $Y_n \xrightarrow{p} \varphi(X)$ and the dominated convergence theorem that

$$E[\varphi(X)] = \lim_{n \rightarrow \infty} E[Y_n] = \liminf_{n \rightarrow \infty} E[Y_n] \leq \liminf_{n \rightarrow \infty} E[X_n]. \quad (43)$$

Taking the supremum over all simple functions φ satisfying $0 \leq \varphi(x) \leq x$ it follows now from (43) and the definition of $E[X]$ that $E[X] \leq \liminf_{n \rightarrow \infty} E[X_n]$.

Finally, note that in the case $E[\liminf_{n \rightarrow \infty} X_n] = \infty$ Fatou's lemma states that then also $\liminf_{n \rightarrow \infty} E[X_n] = \infty$. Q.E.D.

Hilbert Space Theory and Its Applications to Semi-Nonparametric Modeling and Inference

Herman J. Bierens

Pennsylvania State University

Current version: January 25, 2012

Chapter 1

Introduction

As is well known, every vector in a Euclidean space can be represented as a linear combination of orthonormal vectors. Similarly, using Hilbert space theory, we can represent certain classes of Borel measurable functions¹ by countable infinite linear combinations of orthonormal functions, which allows us to approximate these functions arbitrarily close by finite linear combinations of these orthonormal functions. This is the basis for semi-nonparametric (SNP) modeling, where only a part of the model involved is parametrized, and the non-specified part is an unknown function which is approximated by a series expansion. See for example Chen (2007) for a recent survey, and Bickel et al (1998). There is also a substantial literature on estimation of semi-nonparametric models using nonparametric kernel density and/or regression estimators (see for example Horowitz 1998), but these approaches are beyond our scope.

Gallant (1981) was the first econometrician to proposed Fourier series expansions as a way to model unknown functions. Gallant's approach is actually nonparametric in that no Euclidean parameters are involved. See also Eastwood and Gallant (1991) and the references therein. However, the use of Fourier series expansions to model unknown functions has been proposed earlier in the statistics literature. See for example Kronmal and Tarter (1968).

Gallant and Nychka (1987) consider SNP estimation of Heckman's (1979) sample selection model, where the bivariate error distribution of the latent

¹See for example Bierens (2004, Ch. 2) for the definition of Borel measurability of functions.

variable equations is modeled semi-nonparametrically using an Hermite expansion of the error density.

Another example of a semi-nonparametric model is the mixed proportional hazard (MPH) model proposed by Lancaster (1979). In this model the hazard function is the product of three factors, the baseline hazard which depends only on the duration, the systematic hazard which is a function of the observable covariates, and an unobserved non-negative random variable representing neglected heterogeneity. Elbers and Ridder (1982) have shown that under some mild conditions and normalizations the MPH model is nonparametrically identified. Heckman and Singer (1984) propose to estimate the distribution function of the unobserved heterogeneity variable by a discrete distribution. Bierens (2008) and Bierens and Carvalho (2007) use orthonormal Legendre polynomials to model semi-nonparametrically the unobserved heterogeneity distribution of interval-censored mixed proportional hazard models and bivariate mixed proportional hazard models, respectively.

In chapter 2 I will explain what a Hilbert space is, and provide examples of non-Euclidean Hilbert spaces, in particular Hilbert spaces of Borel measurable functions and random variables. In chapter 3 I will discuss projections on sub-Hilbert spaces and their properties. One of the results involved is the famous Wold (1938) decomposition theorem, which will be derived first in general terms and then for covariance stationary time series. Also, the fundamental role of the Wold decomposition in time series analysis and empirical macro-econometrics will be pointed out.

The main focus of this book, however, is on Hilbert spaces of square integrable Borel measurable real functions and the various orthonormal sequences that span these Hilbert spaces, as the basis for semi-nonparametric modeling and inference. Therefore, following Hamming (1973), in chapter 4 I will review the various ways one can construct orthonormal polynomials that span a given Hilbert space of functions. In chapter 5 I will show that any square integrable Borel measurable real function on the unit interval can be written as a linear combination of the cosine series $\{\cos(k\pi u)\}_{k=0}^{\infty}$, $u \in [0, 1]$. This result is related to classical Fourier analysis, which will also be reviewed. The significance of this result is that it yields closed form series representations of arbitrary density and distribution functions, as will be shown in chapter 6, whereas in the approach of Gallant and Nychka (1987), which is based on Hermite polynomials, and the approach of Bierens (2008) and Bierens and Carvalho (2007), which is based on Legendre polynomials, the computation of their density and distribution functions has to be done

iteratively. In chapter 7 I will show how to construct compact metric spaces of density and distribution functions based on the cosine series expansion.

The applications to semi-nonparametric models, based on Bierens. (2011), will be added to this manuscript in due course.

Throughout this manuscript the set of positive integers will be denoted by \mathbb{N} , and the set of non-negative integers by \mathbb{N}_0 . Moreover, the well-known indicator function will be denoted by $I(\cdot)$, and $\mathbf{i} = \sqrt{-1}$.

Part I

Hilbert spaces

Chapter 2

Introduction to Hilbert spaces

In this chapter I will review the concepts of vector spaces, inner products and Cauchy sequences, and provide examples of Hilbert spaces.

2.1 Vector spaces

The notion of a vector space should be known from linear algebra:

Definition 2.1. Let \mathcal{V} be a set endowed with two operations, the operation "addition", denoted by "+", which maps each pair (x, y) in $\mathcal{V} \times \mathcal{V}$ into \mathcal{V} , and the operation "scalar multiplication", denoted by a dot (\cdot) , which maps each pair (c, x) in $\mathbb{R} \times \mathcal{V}$ [or $\mathbb{C} \times \mathcal{V}$] into \mathcal{V} . Thus, a scalar is a real or complex number. The set \mathcal{V} is called a real [complex] vector space if the addition and multiplication operations involved satisfy the following rules, for all x, y and z in \mathcal{V} , and all scalars c, c_1 and c_2 in \mathbb{R} [\mathbb{C}]:

- (a) $x + y = y + x$;
- (b) $x + (y + z) = (x + y) + z$;
- (c) There is a unique zero vector 0 in \mathcal{V} such that $x + 0 = x$;
- (d) For each x there exists a unique vector $-x$ in \mathcal{V} such that $x + (-x) = 0$;¹
- (e) $1.x = x$;
- (f) $(c_1 c_2).x = c_1.(c_2.x)$;
- (g) $c.(x + y) = c.x + c.y$;
- (h) $(c_1 + c_2).x = c_1.x + c_2.x$.

¹Also denoted by $x - x = 0$.

It is trivial to verify that the Euclidean space \mathbb{R}^n is a real vector space. However, the notion of a vector space is much more general. For example, let \mathcal{V} be the space of all continuous functions on \mathbb{R}^n , with pointwise addition and scalar multiplication defined the same way as for real numbers. Then it is easy to verify that this space is a real vector space.

Another (but weird) example of a vector space is the space \mathcal{V} of positive real numbers endowed with the "addition" operation $x + y = x \cdot y$ and the "scalar multiplication" $c \cdot x = x^c$. In this case the null vector 0 is the number 1, and $-x = 1/x$.

Definition 2.2. A subspace \mathcal{V}_0 of a vector space \mathcal{V} is a non-empty subset of \mathcal{V} which satisfies the following two requirements:

- (a) For any pair x, y in \mathcal{V}_0 , $x + y$ is in \mathcal{V}_0 ;
- (b) For any x in \mathcal{V}_0 and any scalar c , $c \cdot x$ is in \mathcal{V}_0 .

Thus, a subspace \mathcal{V}_0 of a vector space is closed under linear combinations: any linear combination of elements in \mathcal{V}_0 is an element of \mathcal{V}_0 .

It is not hard to verify that a subspace of a vector space is a vector space itself, because the rules (a) through (h) in Definition 2.1 are inherited from the "host" vector space \mathcal{V} . In particular, any subspace contains the null vector 0, as follows from part (b) of Definition 2.2 with $c = 0$.

2.2 Inner product and norm

As is well-known, in a Euclidean space \mathbb{R}^n the inner product of a pair of vectors x and y is defined as $x'y$, which is a mapping $\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ with the following properties:

- (a) $x'y = y'x$,
- (b) $(cx)'y = c(x'y)$ for arbitrary $c \in \mathbb{R}$,
- (c) $(x + y)'z = x'z + y'z$,
- (d) $x'x > 0$ if and only if $x \neq 0$.

Moreover, the norm of a vector $x \in \mathbb{R}^n$ is defined as $\|x\| = \sqrt{x'x}$. Of course, in \mathbb{R} the inner product is the ordinary product $x \cdot y$.

Mimicking these four properties, we can define more general inner products with associated norms as follows.

Definition 2.3. An inner product on a real vector space \mathcal{V} is a real function $\langle x, y \rangle: \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ such that for all x, y, z in \mathcal{V} and all c in \mathbb{R} ,

- (1) $\langle x, y \rangle = \langle y, x \rangle$
- (2) $\langle cx, y \rangle = c \langle x, y \rangle$
- (3) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- (4) $\langle x, x \rangle > 0$ if and only if $x \neq 0$.

An inner product on a complex vector space is defined similarly. The inner product is then complex-valued, $\langle x, y \rangle: \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{C}$. Condition (1) then becomes

$$(1^*) \langle x, y \rangle = \overline{\langle y, x \rangle},^2$$

and (2) now holds for all complex and real numbers c . Note that also in this case $\langle x, x \rangle$ is real valued.³ A vector space endowed with an inner product is called an inner product space. Finally, the norm of x in \mathcal{V} is defined as $\|x\| = \sqrt{\langle x, x \rangle}$

For example, in the vector space $C[0, 1]$ of continuous real functions on $[0, 1]$, the integral $\langle f, g \rangle = \int_0^1 f(u)g(u) du$ is an inner product, with norm $\|f\| = \sqrt{\int_0^1 f(u)^2 du}$. Moreover, in the vector space of zero-mean random variables with finite second moments the covariance $\langle X, Y \rangle = E[X.Y]$ is an inner product, with norm $\|X\| = \sqrt{E[X^2]}$.

As is well-known from linear algebra, for vectors $x, y \in \mathbb{R}^n$, $|x'y| \leq \|x\| \cdot \|y\|$, which is known as the Cauchy-Schwarz inequality. This inequality carries over to general inner products:

Theorem 2.1. (Cauchy-Schwarz inequality) $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$.

Given the norm $\|x\| = \sqrt{\langle x, x \rangle}$, the following properties hold:

$$\|x\| > 0 \text{ if } x \neq 0; \tag{2.1}$$

$$\|c.x\| = |c| \cdot \|x\|; \tag{2.2}$$

$$\|x + y\| \leq \|x\| + \|y\|. \tag{2.3}$$

The latter is known as the triangular inequality.

²The bar denotes the complex conjugate: for $z = a + i.b$, $\bar{z} = a - i.b$.

³Because $\langle x, x \rangle = \overline{\langle x, x \rangle}$ implies that $\langle x, x \rangle \in \mathbb{R}$.

The properties (2.1) and (2.2) follow trivially from Definition 2.3. In the case of a real vector space the triangular inequality (2.3) follows from

$$\begin{aligned}\|x + y\|^2 &= \langle x + y, x + y \rangle = \langle x, x \rangle + 2\langle x, y \rangle + \langle y, y \rangle \\ &= \|x\|^2 + 2\langle x, y \rangle + \|y\|^2 \leq \|x\|^2 + 2|\langle x, y \rangle| + \|y\|^2 \\ &\leq \|x\|^2 + 2\|x\|\cdot\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2\end{aligned}$$

where the last inequality is due to Theorem 2.1.

In a Euclidean space, a pair x, y of vectors is orthogonal if $x'y = 0$, and orthonormal if also $\|x\| = \|y\| = 1$. Similarly,

Definition 2.4. *Elements x and y in a inner product space with associated norm are orthogonal if $\langle x, y \rangle = 0$, which is also denoted by $x \perp y$, and are orthonormal if in addition $\|x\| = \|y\| = 1$.*

A norm can also be defined directly:

Definition 2.5. *A norm on a vector space \mathcal{V} is a mapping $\|\cdot\|: \mathcal{V} \rightarrow [0, \infty)$ such that for all x and y in \mathcal{V} and all scalars c the properties (2.1), (2.2) and (2.3) hold. A vector space endowed with a norm is called a normed space.*

2.3 Metric spaces

A norm $\|\cdot\|$ defines a metric $d(x, y) = \|x - y\|$ on \mathcal{V} , i.e., a function that measures the distance between two elements x and y of \mathcal{V} , for which (trivially) the following four properties hold. For all x, y and z in \mathcal{V} ,

$$d(x, y) = d(y, x) \tag{2.4}$$

$$d(x, y) > 0 \text{ if } x \neq y; \tag{2.5}$$

$$d(x, x) = 0; \tag{2.6}$$

$$d(x, z) \leq d(x, y) + d(y, z). \tag{2.7}$$

Again, the property (2.7) is known as the triangular inequality.

A metric can also be defined directly:

Definition 2.6. *A metric on a space \mathcal{M} is a mapping $d(\cdot, \cdot): \mathcal{M} \times \mathcal{M} \rightarrow [0, \infty)$ satisfying the properties (2.4) through (2.7) for all x, y and z in \mathcal{M} . A space endowed with a metric is called a metric space.*

In this definition the space \mathcal{M} is not necessarily a vector space: Any space endowed with a metric is a metric space. For example, let \mathcal{M} be the space of density functions on $[0, 1]$, endowed with the metric

$$d(f, g) = \int_0^1 \left(\sqrt{f(u)} - \sqrt{g(u)} \right)^2 du.$$

This space is not a vector space, and it is not possible to define an inner product on it.

2.4 Convergence of Cauchy sequences

A vector space \mathcal{V} endowed with an inner product $\langle x, y \rangle$ and associated norm $\|x\| = \sqrt{\langle x, y \rangle}$ and metric $\|x - y\|$ is called a *pre-Hilbert space*. The reason for the "pre" is that a fundamental property is still missing, namely that every Cauchy sequence has a limit in \mathcal{V} .

Definition 2.7. A sequence of elements x_n of a metric space with metric $d(\cdot, \cdot)$ is called a *Cauchy sequence* if for every $\varepsilon > 0$ there exists an $n_0(\varepsilon)$ such that for all $k, m \geq n_0(\varepsilon)$, $d(x_k, x_m) < \varepsilon$.

For example, in the Euclidean space \mathbb{R}^p with finite dimension p every Cauchy sequence converges to a limit in \mathbb{R}^p , and the same applies to the space \mathbb{C}^p of p -dimensional vectors with complex-valued components, endowed with the inner product

$$\begin{aligned} \langle x, y \rangle &= \bar{x}' y = (\operatorname{Re}(x) - \mathbf{i} \cdot \operatorname{Im}(x))' (\operatorname{Re}(y) + \mathbf{i} \cdot \operatorname{Im}(y)) \\ &= (\operatorname{Re}(x)' \operatorname{Re}(y) + \operatorname{Im}(x)' \operatorname{Im}(y)) \\ &\quad + \mathbf{i} \cdot (\operatorname{Re}(x)' \operatorname{Im}(y) - \operatorname{Im}(x)' \operatorname{Re}(y)) \end{aligned} \tag{2.8}$$

and associated norm and metric. It is an easy exercise to check that (2.8) satisfies the conditions in Definition 2.3. Thus,

Theorem 2.2. Every Cauchy sequence in \mathbb{R}^p or \mathbb{C}^p has a limit in that space.

To demonstrate the role of the Cauchy convergence property, consider the space $C[0, 1]$ of continuous real functions on $[0, 1]$, i.e., each $f \in C[0, 1]$ is

continuous on $(0, 1)$, and $f(0) = \lim_{u \downarrow 0} f(u)$ and $f(1) = \lim_{u \uparrow 1} f(u)$ are finite. Endow this space with the inner product $\langle f, g \rangle = \int_0^1 f(u)g(u)du$ and associated norm $\|f\| = \sqrt{\langle f, f \rangle}$ and metric $\|f - g\|$. Now consider the following sequence of functions in $C[0, 1]$:

$$f_n(u) = \begin{cases} 0 & \text{for } 0 \leq u < 0.5 \\ 2^n(u - 0.5) & \text{for } 0.5 \leq u < 0.5 + 2^{-n} \\ 1 & \text{for } 0.5 + 2^{-n} \leq u \leq 1, \end{cases}$$

$$n = 1, 2, 3, \dots.$$

It is an easy calculus exercise to verify that $\|f_k - f_m\|^2 = \int_0^1 (f_k(u) - f_m(u))^2 du < \frac{1}{3}(2^{-k} + 2^{-m})$, hence f_n is a Cauchy sequence in $C[0, 1]$. Moreover, it follows from the bounded convergence theorem that $\lim_{n \rightarrow \infty} \|f_n - f\| = 0$, where $f(u) = I(u > 0.5)$. However, this limit $f(u)$ is discontinuous in $u = 0.5$, and thus $f \notin C[0, 1]$. Therefore, the space $C[0, 1]$ is not closed under convergence.

2.5 Hilbert spaces and sub-Hilbert spaces

2.5.1 Hilbert spaces versus Banach spaces

It is usually quite easy to define an inner product on a vector space, and the same vector space can often be endowed with different inner products. For example, for the space of square integrable Borel measurable functions on $[0, 1]$ we can define an inner product by $\langle f, g \rangle = \int_0^1 f(u)g(u)du$ but also by $\langle f, g \rangle = \int_0^1 uf(u)g(u)du$, for example. However, to make such a space a Hilbert space the inner product must be chosen such that every Cauchy sequence converges to a limit in that space. The requirement that every Cauchy sequence in a Hilbert space has a limit in that space makes a Hilbert space closed under convergence, which generates all kinds of useful properties, similar to Euclidean spaces.

Definition 2.8. A Hilbert space \mathcal{H} is a vector space endowed with an inner product and associated norm and metric such that every Cauchy sequence in \mathcal{H} has a limit in \mathcal{H} . The way the inner product $\langle x, y \rangle$ is defined, together with the associated norm and metric, will be called the topology of \mathcal{H} .

Note that the limit of a Cauchy sequence in a Hilbert space is unique. To see this, suppose that a Cauchy sequence $x_n \in \mathcal{H}$ has two limits: $\lim_{n \rightarrow \infty} \|x_n - x\| = 0$ and $\lim_{n \rightarrow \infty} \|x_n - x_*\| = 0$. Then $\|x_* - x\| = \|x_* - x_n + x_n - x\| \leq \|x_n - x\| + \|x_n - x_*\| \rightarrow 0$ as $n \rightarrow \infty$. Conversely, any convergent sequence in a Hilbert space is a Cauchy sequence, because $\lim_{n \rightarrow \infty} \|x_n - x\| = 0$ implies that $\|x_k - x_m\| = \|(x_k - x) + (x - x_m)\| \leq \|x_k - x\| + \|x_m - x\| \rightarrow 0$ as $\min(k, m) \rightarrow \infty$.

In Definition 2.5 the norm $\|\cdot\|$ was defined directly, without reference to an inner product, giving rise to the definition of a normed space \mathcal{N} , for example. If we endow \mathcal{N} with the metric $\|x - y\|$ and require that every Cauchy sequence in \mathcal{N} has a limit in \mathcal{N} then the space \mathcal{N} becomes a Banach space. The difference between a Hilbert space and a Banach space is the source of the norm: In an Hilbert space the norm is defined on the basis of an inner product whereas in the case of a Banach space the norm is defined directly. Consequently, in a Banach space the notion of inner product is nonexisting, and so is the notion of orthogonality.

2.5.2 Linear manifolds and sub-Hilbert spaces

Because a Hilbert space is a vector space, we can define a subspace of a Hilbert space in the same way as for vector spaces (see Definition 2.2), and endow it with the same inner product, norm and metric as Hilbert space involved. Such a subspace is called a linear manifold:

Definition 2.9. *A linear manifold \mathcal{M} of a Hilbert space \mathcal{H} is a subspace of \mathcal{H} endowed with the topology of \mathcal{H} .*

However, a linear manifold \mathcal{M} is not necessarily a Hilbert space itself. In general there is no guarantee that every Cauchy sequence in \mathcal{M} takes a limit in \mathcal{M} . If so the linear manifold \mathcal{M} needs to be extended by augmenting it with the limits of all Cauchy sequence in \mathcal{M} . The resulting extended linear manifold coincides with the closure $\overline{\mathcal{M}}$ of \mathcal{M} . Recall that \mathcal{M} is a subset of the metric space \mathcal{H} , and that a point of closure of \mathcal{M} is an element \bar{x} such that for each $\varepsilon > 0$ there exists a $z \in \mathcal{M}$ and a $y \in \mathcal{H} \setminus \mathcal{M}$ such that $\|\bar{x} - z\| < \varepsilon$ and $\|\bar{x} - y\| < \varepsilon$. The set of all points of closure of \mathcal{M} is called the border of \mathcal{M} , denoted by $\partial\mathcal{M}$, and the closure of \mathcal{M} , denoted by $\overline{\mathcal{M}}$, is the union of \mathcal{M} and its border: $\overline{\mathcal{M}} = \mathcal{M} \cup \partial\mathcal{M}$.

Theorem 2.3. *The closure $\overline{\mathcal{M}}$ of a linear manifold \mathcal{M} is a Hilbert space.*

In other words, $\overline{\mathcal{M}}$ is a sub-Hilbert space.

2.5.3 Hilbert spaces spanned by a sequence

Let \mathcal{H} be a Hilbert space and let $\{x_k\}_{k=1}^{\infty}$ be a sequence of elements of \mathcal{H} . Let \mathcal{M}_m be the linear manifold spanned by x_1, \dots, x_m , i.e., \mathcal{M}_m consists of all linear combinations of x_1, \dots, x_m . Then it follows similar to the proof of Theorem 2.3 that

Lemma 2.1. *\mathcal{M}_m is a Hilbert space.*

Definition 2.10. *The space $\mathcal{M}_{\infty} = \overline{\cup_{n=1}^{\infty} \mathcal{M}_n}$ is called the space spanned by $\{x_j\}_{j=1}^{\infty}$, and is also denoted by $\text{span}(\{x_j\}_{j=1}^{\infty})$.*

It follows similar to the proof of Theorem 2.3 that

Lemma 2.2. *\mathcal{M}_{∞} is a Hilbert space.*

Remark. In the sequel a sub-Hilbert space will be referred to as a "sub-space".

Definition 2.11. *A sequence $\{x_k\}_{k=1}^{\infty}$ in a Hilbert space \mathcal{H} is called complete if $\mathcal{H} = \text{span}(\{x_j\}_{j=1}^{\infty})$.*

2.6 Examples of non-Euclidean Hilbert spaces

2.6.1 A Hilbert space of random variables

Consider the space \mathcal{R} of random variables defined on a common probability space $\{\Omega, \mathcal{F}, P\}$ with finite second moments, endowed with the inner product $\langle X, Y \rangle = E[X.Y]$ and associated norm $\|X\| = \sqrt{\langle X, X \rangle} = \sqrt{E[X^2]}$ and metric $\|X - Y\|$. Then

Theorem 2.4. *The space \mathcal{R} is a Hilbert space.*

2.6.2 Hilbert spaces of functions

Let $w(x)$ be a probability density on \mathbb{R} and let $L^2(w)$ be the space of Borel measurable real functions f on \mathbb{R} satisfying

$$\int_{-\infty}^{\infty} f(x)^2 w(x) dx < \infty$$

where the integral is the Lebesgue integral, endowed with the inner product

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x)g(x)w(x)dx$$

and associated norm $\|f\| = \sqrt{\langle f, f \rangle} = \sqrt{\int_{-\infty}^{\infty} f(x)^2 w(x) dx}$ and metric $\|f - g\|$. Then for $f, g \in L^2(w)$, $\langle f, g \rangle = E[f(X)g(X)]$, where X is a random drawing from the distribution with density $w(x)$, hence it follows from Theorem 2.4 that $L^2(w)$ is a Hilbert space.

2.7 Appendix: Proofs

2.7.1 Theorem 2.1

Let the vector space involved be complex. It follows from the properties (1)-(4) in Definition 2.3 that for any complex valued λ ,

$$\begin{aligned} 0 &\leq \langle x + \lambda y, x + \lambda y \rangle \\ &= \langle x, x \rangle + \langle \lambda y, x \rangle + \langle x, \lambda y \rangle + \langle \lambda y, \lambda y \rangle \\ &= \|x\|^2 + \lambda \overline{\langle y, x \rangle} + \overline{\langle \lambda y, x \rangle} + \lambda \langle y, \lambda y \rangle \\ &= \|x\|^2 + \lambda \overline{\langle x, y \rangle} + \lambda \overline{\langle y, x \rangle} + \lambda \overline{\langle \lambda y, y \rangle} \\ &= \|x\|^2 + \lambda \overline{\langle x, y \rangle} + \overline{\lambda} \cdot \overline{\langle y, x \rangle} + \lambda \cdot \overline{\lambda} \overline{\langle y, y \rangle} \\ &= \|x\|^2 + \lambda \overline{\langle x, y \rangle} + \overline{\lambda} \cdot \langle x, y \rangle + \lambda \cdot \overline{\lambda} \langle y, y \rangle \\ &= \|x\|^2 + \lambda \overline{\langle x, y \rangle} + \overline{\lambda} \cdot \langle x, y \rangle + \lambda \cdot \overline{\lambda} \|y\|^2 \end{aligned}$$

Next, note that

$$\begin{aligned} \lambda \overline{\langle x, y \rangle} + \overline{\lambda} \cdot \langle x, y \rangle &= (\operatorname{Re}(\lambda) + \mathbf{i} \cdot \operatorname{Im}(\lambda)) (\operatorname{Re}(\langle x, y \rangle) - \mathbf{i} \cdot \operatorname{Im}(\langle x, y \rangle)) \\ &\quad + (\operatorname{Re}(\lambda) - \mathbf{i} \cdot \operatorname{Im}(\lambda)) (\operatorname{Re}(\langle x, y \rangle) + \mathbf{i} \cdot \operatorname{Im}(\langle x, y \rangle)) \\ &= 2(\operatorname{Re}(\lambda) \operatorname{Re}(\langle x, y \rangle) + \operatorname{Im}(\lambda) \operatorname{Im}(\langle x, y \rangle)) \end{aligned}$$

and

$$\begin{aligned}\lambda \cdot \bar{\lambda} &= (\operatorname{Re}(\lambda) + \mathbf{i} \cdot \operatorname{Im}(\lambda)) (\operatorname{Re}(\lambda) - \mathbf{i} \cdot \operatorname{Im}(\lambda)) \\ &= (\operatorname{Re}(\lambda))^2 + (\operatorname{Im}(\lambda))^2\end{aligned}$$

Hence

$$\begin{aligned}0 &\leq \|x\|^2 + 2(\operatorname{Re}(\lambda) \operatorname{Re}(\langle x, y \rangle) + \operatorname{Im}(\lambda) \operatorname{Im}(\langle x, y \rangle)) \\ &\quad + ((\operatorname{Re}(\lambda))^2 + (\operatorname{Im}(\lambda))^2) \cdot \|y\|^2\end{aligned}\tag{2.9}$$

The latter is minimal for

$$\operatorname{Re}(\lambda) = -\frac{\operatorname{Re}(\langle x, y \rangle)}{\|y\|^2}, \quad \operatorname{Im}(\lambda) = -\frac{\operatorname{Im}(\langle x, y \rangle)}{\|y\|^2}.$$

Substituting these solutions in (2.9) yields

$$0 \leq \|x\|^2 - \frac{1}{\|y\|^2} ((\operatorname{Re}(\langle x, y \rangle))^2 + (\operatorname{Im}(\langle x, y \rangle))^2) = \|x\|^2 - \frac{|\langle x, y \rangle|^2}{\|y\|^2}$$

and thus $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$.

2.7.2 Theorem 2.2

Let x_n be a Cauchy sequence in \mathbb{R} , and denote $\bar{x} = \limsup_{n \rightarrow \infty} x_n$. Let us first show that $\bar{x} < \infty$, as follows. By the definition of "limsup" there exists a subsequence n_k such that $\bar{x} = \lim_{k \rightarrow \infty} x_{n_k}$. Note that this x_{n_k} is also a Cauchy sequence, hence for arbitrary $\varepsilon > 0$ there exists an index k_0 such that $|x_{n_k} - x_{n_m}| < \varepsilon$ for all $k, m \geq k_0$. Keeping $m \geq k_0$ fixed and letting $k \rightarrow \infty$ it follows that $|\bar{x} - x_{n_m}| < \varepsilon$, hence $\bar{x} < \infty$. By a similar argument it follows that $\underline{x} = \liminf_{n \rightarrow \infty} x_n > -\infty$. Thus, we can find an index k_0 and subsequences $n_{1,k}$ and $n_{2,m}$ such that for all $k, m \geq k_0$, $|\bar{x} - x_{n_{1,m}}| < \varepsilon$, $|\underline{x} - x_{n_{2,m}}| < \varepsilon$ and $|x_{n_{1,m}} - x_{n_{2,m}}| < \varepsilon$, hence $|\bar{x} - \underline{x}| < 3\varepsilon$. Since ε was arbitrary, it follows now that $\bar{x} = \underline{x} = x$, which implies that $\lim_{n \rightarrow \infty} x_n = x \in \mathbb{R}$. By applying this argument to the real and imaginary parts of a complex valued Cauchy sequence x_n it follows that $\lim_{n \rightarrow \infty} x_n = x \in \mathbb{C}$. Moreover, applying this argument to each component of a (complex) vector valued Cauchy sequence the results for the cases \mathbb{R}^p and \mathbb{C}^p follow.

2.7.3 Theorem 2.3

Let x_n be a Cauchy sequence in $\overline{\mathcal{M}} \subset \mathcal{H}$. Then x_n has a limit $\bar{x} \in \mathcal{H}$, i.e., $\lim_{n \rightarrow \infty} \|x_n - \bar{x}\| = 0$. Suppose that $\bar{x} \notin \overline{\mathcal{M}}$. Since $\overline{\mathcal{M}}$ is closed there exists an $\varepsilon > 0$ such that the set $\mathcal{N}(\bar{x}, \varepsilon) = \{x \in \mathcal{H} : \|x - \bar{x}\| < \varepsilon\}$ is completely outside $\overline{\mathcal{M}}$: $\mathcal{N}(\bar{x}, \varepsilon) \cap \overline{\mathcal{M}} = \emptyset$. But $\lim_{n \rightarrow \infty} \|x_n - \bar{x}\| = 0$ implies that there exists an $\underline{n}(\varepsilon)$ such that $x_n \in \mathcal{N}(\bar{x}, \varepsilon)$ for all $n > \underline{n}(\varepsilon)$, hence $x_n \notin \overline{\mathcal{M}}$ for all $n > \underline{n}(\varepsilon)$, which contradicts $x_n \in \overline{\mathcal{M}}$ for all n .

2.7.4 Lemma 2.1

Without loss of generality we may assume that the $m \times m$ matrix X_m with elements $\langle x_i, x_j \rangle$ is nonsingular, as otherwise we can re-arrange the x_j 's such that $\mathcal{M}_m = \mathcal{M}_r$ with $r = \text{rank}(X_m)$. Let $y_{n,m} = \sum_{j=1}^m c_{j,n} x_j$ be a Cauchy sequence in \mathcal{M}_m . Then

$$\begin{aligned} \|y_{n_1,m} - y_{n_2,m}\|^2 &= \left\| \sum_{j=1}^m (c_{j,n_1} - c_{j,n_2}) x_j \right\|^2 \\ &= \sum_{i=1}^m \sum_{j=1}^m (c_{i,n_1} - c_{i,n_2})(c_{j,n_1} - c_{j,n_2}) \langle x_i, x_j \rangle \rightarrow 0 \end{aligned}$$

as $\min(n_1, n_2) \rightarrow \infty$. This is only possible if for $j = 1, 2, \dots, m$, $\lim_{\min(n_1, n_2) \rightarrow \infty} |c_{j,n_1} - c_{j,n_2}| = 0$. Thus, the $c_{j,n}$'s are Cauchy sequences in \mathbb{R} , and therefore converge to limits c_j . Denoting $y_m = \sum_{j=1}^m c_j x_j$, which is an element of \mathcal{M}_m , it follows now easily that $\lim_{n \rightarrow \infty} \|y_{n,m} - y_m\| = 0$. Thus, every Cauchy sequence in \mathcal{M}_m converges to a limit in \mathcal{M}_m .

2.7.5 Theorem 2.4

Let X_n be a Cauchy sequence in \mathcal{R} . Then

$$\|X_n - X_m\|^2 = E [(X_n - X_m)^2] \rightarrow 0$$

as $\min(n, m) \rightarrow \infty$, so that by Chebyshev's inequality,

$$\plim_{\min(n,m) \rightarrow \infty} |X_n - X_m| = 0.$$

As is well-known, convergence in probability is equivalent to almost sure (a.s.) convergence along a further subsequence of an arbitrary subsequence⁴, hence there exists a subsequence n_k such that for $\min(k, m) \rightarrow \infty$,

$$|X_{n_k} - X_{n_m}| \xrightarrow{a.s.} 0.$$

In its turn this result is equivalent to the statement that there exists a set $N \in \mathcal{F}$ with $P(N) = 0$, called a null set, such that for all $\omega \in \Omega \setminus N$,

$$\lim_{\min(k, m) \rightarrow \infty} |X_{n_k}(\omega) - X_{n_m}(\omega)| = 0$$

Now $X_{n_k}(\omega)$ is a Cauchy sequence in \mathbb{R} and thus converges to a limit $X(\omega)$ in \mathbb{R} , which is measurable \mathcal{F} ,⁵ so that X is a random variable defined on $\{\Omega, \mathcal{F}, P\}$. Hence, for fixed m and $k \rightarrow \infty$

$$(X_{n_k} - X_m)^2 \xrightarrow{a.s.} (X - X_m)^2. \quad (2.10)$$

Finally, it follows from (2.10), Fatou's lemma⁶ and the Cauchy property that

$$\begin{aligned} \|X - X_m\|^2 &= E[(X - X_m)^2] = E\left[\lim_{k \rightarrow \infty} (X_{n_k} - X_m)^2\right] \\ &\leq \liminf_{k \rightarrow \infty} E[(X_{n_k} - X_m)^2] \rightarrow 0 \end{aligned}$$

for $m \rightarrow \infty$.

⁴See for example Bierens (2004, Theorem 6.B.3, p. 168).

⁵The latter follows from the well-known property that the limsup and liminf of a sequence of random variables are random variables themselves. See for example Bierens (2004, Theorem 2.13, p. 47).

⁶Fatou's lemma states: *For a sequence X_n of non-negative random variables, $E[\liminf_{n \rightarrow \infty} X_n] \leq \liminf_{n \rightarrow \infty} E[X_n]$.* See for example Bierens (2004, Lemma 7.A.1, p. 201).

Chapter 3

Projections

3.1 The projection theorem

As is well-known from linear algebra and econometrics, the projection of a vector $y \in \mathbb{R}^n$ on the subspace spanned by vectors x_1, \dots, x_k in \mathbb{R}^n is a linear combination $\hat{y} = \sum_{j=1}^k \beta_j x_j$ such that $\|y - \hat{y}\|$ is minimal. This is a linear regression problem: Minimize

$$\|y - \hat{y}\|^2 = y'y - 2y'X\beta + \beta'X'X\beta$$

to $\beta = (\beta_1, \dots, \beta_k)'$, where $X = (x_1, \dots, x_k)$. If $k \leq n$ and the vectors x_1, \dots, x_k are linear independent then the solution is $\beta = (X'X)^{-1}X'y$, hence $\hat{y} = X\beta = X(X'X)^{-1}X'y$.

If x_1, \dots, x_k are not linear independent then $\text{rank}(X) = m < k$. In that case we can rearrange x_1, \dots, x_k such that the matrix $X_1 = (x_1, \dots, x_m)$ has rank m and $X_2 = (x_{m+1}, \dots, x_k) = X_1C$ for some $(k-m) \times (k-m)$ matrix C . Partition β accordingly as $\beta = (b'_1, b'_2)'$. Then

$$\|y - \hat{y}\|^2 = y'y - 2y'X_1(b_1 - Cb_2) + (b_1 - Cb_2)'X'_1X_1(b_1 - Cb_2)$$

which is minimal for $(b_1 - Cb_2) = (X'_1X_1)^{-1}X'_1y$, hence

$$\hat{y} = X_1b_1 + X_2b_2 = X_1(b_1 - Cb_2) + X_1(X'_1X_1)^{-1}X'_1y,$$

which is unique. The latter follows from Theorem 3.1 below.

The notion of a projection for Hilbert spaces is similar:

Definition 3.1. *The projection \hat{y} of an element y of a Hilbert space \mathcal{H} on a subspace \mathcal{S} is an element $\hat{y} \in \mathcal{S}$ such that $\|y - \hat{y}\| = \inf_{z \in \mathcal{S}} \|y - z\|$.*

However, we still have to show that $\hat{y} \in \mathcal{S}$ is possible and unique. This follows from the fundamental projection theorem:

Theorem 3.1. (Projection theorem) *If \mathcal{S} is a subspace of a Hilbert space \mathcal{H} and y an element of \mathcal{H} then there exists a unique element $\hat{y} \in \mathcal{S}$ such that $\|y - \hat{y}\| = \inf_{z \in \mathcal{S}} \|y - z\|$. Moreover the residual $u = y - \hat{y}$ is orthogonal to any $z \in \mathcal{S}$: $\langle u, z \rangle = 0$.*

3.2 Projections in terms of angles

As is well known, the angle $\varphi(x, y)$ between two vectors x and y in a Euclidean space is defined by the cosine formula

$$\cos(\varphi(x, y)) = \frac{\|x\|^2 + \|y\|^2 - \|x - y\|^2}{2\|x\|\cdot\|y\|} = \frac{x'y}{\|x\|\cdot\|y\|},$$

due to the Law of Cosines.¹ Clearly, this formula carries over to elements x and y of a Hilbert space \mathcal{H} , simply by replacing the Euclidean inner product $x'y$ and norm $\|x\| = \sqrt{x'x}$ by $\langle x, y \rangle$ and $\|x\| = \sqrt{\langle x, x \rangle}$, respectively. Thus, the angle $\varphi(x, y)$ between two elements x and y of a Hilbert space is defined by the cosine formula

$$\cos(\varphi(x, y)) = \frac{\langle x, y \rangle}{\|x\|\cdot\|y\|}. \quad (3.1)$$

Let \mathcal{S} , y and \hat{y} be as before, and let x be any element of \mathcal{S} . Then it follows from the cosine formula (3.1) and the orthogonality condition $\langle x, y - \hat{y} \rangle = 0$ that

$$\cos(\varphi(x, y)) = \frac{\langle x, y \rangle}{\|x\|\cdot\|y\|} = \frac{\langle x, \hat{y} \rangle}{\|x\|\cdot\|y\|} = \frac{\|\hat{y}\|}{\|y\|} \cos(\varphi(x, \hat{y}))$$

¹Consider a triangle ABC , let φ be the angle between the legs $C \rightarrow A$ and $C \rightarrow B$, and denote the lengths of the legs opposite to the points A , B and C by α , β , and γ , respectively. Then $\gamma^2 = \alpha^2 + \beta^2 - 2\alpha\beta \cos(\varphi)$.

which is maximal if $\cos(\varphi(x, \hat{y})) = 1$. The latter is true if $x = c\hat{y}$ for some constant $c > 0$. Consequently,

$$\cos(\varphi(y, \hat{y})) = \max_{x \in \mathcal{S}} \cos(\varphi(x, y)) = \frac{\|\hat{y}\|}{\|y\|}. \quad (3.2)$$

3.3 Projections on subspaces spanned by a sequence

Let \mathcal{H} be a Hilbert space and let $\{x_k\}_{k=1}^\infty$ be a sequence of elements of \mathcal{H} . Let \mathcal{M}_n be the linear manifold spanned by x_1, \dots, x_n : $\mathcal{M}_n = \text{span}(\{x_k\}_{k=1}^n)$. As we have seen from Lemma 2.1, \mathcal{M}_n is a Hilbert space.

Consider the projection \hat{y}_n of an element $y \in \mathcal{H}$ on \mathcal{M}_n . Then \hat{y}_n takes the form $\hat{y}_n = \sum_{k=1}^n \theta_{n,k} x_k$, where the $\theta_{n,k}$'s are the solutions of the minimization problem

$$\begin{aligned} \min_{\theta_1, \theta_2, \dots, \theta_n} & \left\| y - \sum_{k=1}^n \theta_k x_k \right\|^2 \\ &= \min_{\theta_1, \theta_2, \dots, \theta_n} \left(\|y\|^2 - 2 \sum_{k=1}^n \theta_k \langle x_k, y \rangle + \sum_{k=1}^n \sum_{m=1}^n \theta_k \theta_m \langle x_k, x_m \rangle \right) \end{aligned}$$

Similar to linear regression, the first-order conditions involved are the normal equations

$$\sum_{m=1}^n \langle x_k, x_m \rangle \theta_{n,m} = \langle x_k, y \rangle, \quad k = 1, 2, \dots, n,$$

which can be written in matrix-vector form as $\Sigma_{n,xx} \theta_n = \Sigma_{n,xy}$, for example. To solve this system uniquely as $\theta_n = \Sigma_{n,xx}^{-1} \Sigma_{n,xy}$ we need to impose a similar condition as linear independence in Euclidean spaces, namely **regularity**:

Definition 3.2. Let $\{x_k\}_{k=1}^\infty$ be a sequence of elements of a Hilbert space \mathcal{H} . Denote the projection of x_k on $\text{span}(\{x_j\}_{j=k+1}^\infty)$ by \hat{x}_k , and let $u_k = x_k - \hat{x}_k$. The sequence $\{x_k\}_{k=1}^\infty$ is said to be regular if $\|u_k\| > 0$ for all $k \geq 1$.

Exercise: Given a regular sequence $\{x_k\}_{k=1}^\infty$, prove that for $n = 1, 2, 3, \dots$ the $n \times n$ matrices $\Sigma_{n,xx}$ with elements $\langle x_i, x_j \rangle$ are nonsingular.

Lemma 3.1. For $z \in \text{span}(\{x_k\}_{k=1}^\infty)$ let \hat{z}_n be the projection of z on $\text{span}(\{x_k\}_{k=1}^n)$. Then $\lim_{n \rightarrow \infty} \|z - \hat{z}_n\| = 0$.

More generally we have:

Theorem 3.2. For $z \in \mathcal{H}$, let \hat{z} be the projection of z on $\text{span}(\{x_k\}_{k=1}^\infty)$ and let \hat{z}_n be the projection of z on $\text{span}(\{x_k\}_{k=1}^n)$. Then $\lim_{n \rightarrow \infty} \|\hat{z} - \hat{z}_n\| = 0$.

Although each projection \hat{z}_n is a linear combination of x_1, \dots, x_n , in general the result of Theorem 3.2 does **not** imply that there exists a sequence $\{\theta_j\}_{j=1}^\infty$ such that $\hat{z} = \sum_{j=1}^\infty \theta_j x_j$. As an example of such a case, consider the Hilbert space \mathcal{R}_0 of zero-mean random variables with finite second moments, endowed with the inner product $\langle X, Y \rangle = E[X.Y]$ and associated norm and metric. Let

$$X_t = V_t - V_{t-1},$$

where V_t is distributed i.i.d. $N(0, 1)$. This is clearly a zero-mean covariance stationary process, with covariance function $\gamma(0) = 2$, $\gamma(1) = -1$, $\gamma(m) = 0$ for $m \geq 2$. Hence $X_t \in \mathcal{R}_0$ for all t .

For given t , let $\mathcal{M}_{-\infty}^{t-1} = \text{span}(\{X_{t-m}\}_{m=1}^\infty)$, $\mathcal{M}_{t-n}^{t-1} = \text{span}(X_{t-1}, \dots, X_{t-n})$. The projection $\hat{X}_{t,n}$ of X_t on \mathcal{M}_{t-n}^{t-1} takes the form

$$\hat{X}_{t,n} = \sum_{j=1}^n \theta_{n,j} X_{t-j}$$

where the coefficients $\theta_{n,j}$ are the solutions of the normal equations

$$\gamma(m) = \sum_{k=1}^n \gamma(|k - m|) \theta_{n,k}, \quad m = 1, \dots, n.$$

hence for $n \geq 3$,

$$\begin{aligned} -1 &= 2\theta_{n,1} - \theta_{n,2} \\ 0 &= -\theta_{n,1} + 2\theta_{n,2} - \theta_{n,3} \\ 0 &= -\theta_{n,2} + 2\theta_{n,3} - \theta_{n,4} \\ &\vdots \\ 0 &= -\theta_{n,n-2} + 2\theta_{n,n-1} - \theta_{n,n} \\ 0 &= -\theta_{n,n-1} + 2\theta_{n,n} \end{aligned}$$

The solutions of these normal equations are

$$\theta_{n,j} = \frac{j}{n+1} - 1, \quad j = 1, \dots, n,$$

hence

$$\widehat{X}_{t,n} = \sum_{j=1}^n \left(\frac{j}{n+1} - 1 \right) X_{t-j} \quad (3.3)$$

Next, let \widehat{X}_t be the projection of X_t on $\mathcal{M}_{-\infty}^{t-1}$, and suppose that there exists a sequence $\{\theta_j\}_{j=1}^\infty$ such that $\widehat{X}_t = \sum_{j=1}^\infty \theta_j X_{t-j}$. Note that the latter is merely a short-hand notation for

$$\lim_{n \rightarrow \infty} \left\| \widehat{X}_t - \sum_{j=1}^n \theta_j X_{t-j} \right\|^2 = \lim_{n \rightarrow \infty} E \left[\left(\widehat{X}_t - \sum_{j=1}^n \theta_j X_{t-j} \right)^2 \right] = 0 \quad (3.4)$$

If so, it follows from Theorem 3.2 and (3.3) that

$$\begin{aligned} 0 &= \lim_{n \rightarrow \infty} E \left[\left(\sum_{j=1}^n \theta_j X_{t-j} - \sum_{j=1}^n \left(\frac{j}{n+1} - 1 \right) X_{t-j} \right)^2 \right] \quad (3.5) \\ &= \lim_{n \rightarrow \infty} E \left[\left(\sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) X_{t-j} \right)^2 \right] \end{aligned}$$

But

$$\begin{aligned} \sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) X_{t-j} &= \sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) (V_{t-j} - V_{t-j-1}) \\ &= - \left(\frac{n}{n+1} + \theta_1 \right) V_{t-1} - \sum_{j=1}^{n-1} \left(\theta_{j+1} - \theta_j - \frac{1}{n+1} \right) V_{t-j-1} \\ &\quad + \left(\frac{1}{n+1} + \theta_n \right) V_{t-n-1} \end{aligned}$$

hence

$$\begin{aligned} E \left[\left(\sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) X_{t-j} \right)^2 \right] &= \left(\frac{n}{n+1} + \theta_1 \right)^2 \\ &\quad + \sum_{j=1}^{n-1} \left(\theta_{j+1} - \theta_j - \frac{1}{n+1} \right)^2 + \left(\frac{1}{n+1} + \theta_n \right)^2 \quad (3.6) \end{aligned}$$

This equality implies that for arbitrary integers $m \geq 1$,

$$\begin{aligned} & \liminf_{n \rightarrow \infty} E \left[\left(\sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) X_{t-j} \right)^2 \right] \\ & \geq \liminf_{n \rightarrow \infty} \left(\frac{n}{n+1} + \theta_1 \right)^2 + \liminf_{n \rightarrow \infty} \left(\theta_{m+1} - \theta_m - \frac{1}{n+1} \right)^2 \\ & = (\theta_1 + 1)^2 + (\theta_{m+1} - \theta_m)^2. \end{aligned}$$

Therefore, a necessary condition for (3.5) is that $\theta_m = -1$ for $m = 1, 2, 3, \dots$. But then it follows from (3.6) that

$$\lim_{n \rightarrow \infty} E \left[\left(\sum_{j=1}^n \left(\frac{j}{n+1} - 1 - \theta_j \right) X_{t-j} \right)^2 \right] = \lim_{n \rightarrow \infty} \left(\frac{1}{n+1} - 1 \right)^2 = 1$$

which contradicts (3.5). Thus, in this case there does **not** exist a sequence $\{\theta_j\}_{j=1}^\infty$ such that (3.4) holds.

The problem that for the projection \hat{z} on $\text{span}(\{x_j\}_{j=1}^\infty)$ there does not always exist a sequence $\{\theta_j\}_{j=1}^\infty$ such that $\hat{z} = \sum_{j=1}^\infty \theta_j x_j$ only occurs if the sequence $\{x_j\}_{j=1}^\infty$ is not orthogonal:

Theorem 3.3. *If a sequence $\{x_j\}_{j=1}^\infty$ in a Hilbert space \mathcal{H} is orthonormal, i.e.,*

$$\langle x_i, x_j \rangle = I(i=j), \quad (3.7)$$

then the projection \hat{z} of $z \in \mathcal{H}$ on $\text{span}(\{x_j\}_{j=1}^\infty)$ takes the form $\hat{z} = \sum_{j=1}^\infty \theta_j x_j$ (in the sense that $\lim_{n \rightarrow \infty} \|\hat{z} - \sum_{j=1}^n \theta_j x_j\|$), where $\theta_j = \langle z, x_j \rangle$ with $\sum_{j=1}^\infty \theta_j^2 < \infty$.

3.4 The Wold decomposition

Let $\mathcal{S}_1, \dots, \mathcal{S}_n$ be subspaces of a Hilbert space \mathcal{H} . Then similar to Definition 2.10,

Definition 3.3. *$\text{Span}(\mathcal{S}_1, \dots, \mathcal{S}_n)$ is the closure of the space of all linear combinations $\sum_{j=1}^n c_j x_j$, where $x_j \in \mathcal{S}_j$.*

We also need the definition of orthogonal complement:

Definition 3.4. *The orthogonal complement of a subspace \mathcal{S} of a Hilbert space \mathcal{H} , denoted by \mathcal{S}^\perp , is the subset of \mathcal{H} such that for each $x \in S$ and $y \in \mathcal{S}^\perp$, $\langle x, y \rangle = 0$.*

Lemma 3.2. *Orthogonal complements are subspaces.*

We can now formulate the following general version of the Wold decomposition:

Theorem 3.4. *Given a regular sequence $\{x_k\}_{k=1}^\infty$ in a Hilbert space, every $x \in \mathcal{S} = \text{span}(\{x_k\}_{k=1}^\infty)$ can be written as $x = \sum_{k=1}^\infty \alpha_k e_k + w$, in the sense that $\lim_{n \rightarrow \infty} \|x - w - \sum_{k=1}^n \alpha_k e_k\| = 0$, where $\{e_k\}_{k=1}^\infty$ is an orthonormal sequence in \mathcal{S} , $\alpha_k = \langle x, e_k \rangle$, $\sum_{k=1}^\infty \alpha_k^2 < \infty$, and*

$$w \in \mathcal{S}_\infty \cap \mathcal{U}_\infty^\perp, \quad (3.8)$$

with $\mathcal{S}_\infty = \bigcap_{n=1}^\infty \text{span}(\{x_k\}_{k=n}^\infty)$ and \mathcal{U}_∞^\perp the orthogonal complement of $\mathcal{U}_\infty = \text{span}(\{e_k\}_{k=1}^\infty)$. Note that (3.8) implies that w is orthogonal to all the e_k 's: $\langle e_k, w \rangle = 0$ for $k = 1, 2, 3, \dots$.

In the case of the Hilbert space \mathcal{R}_0 of zero-mean random variables with finite second moments, with inner product $\langle X, Y \rangle = E[X.Y]$ and associated norm and metric, the results of Theorem 3.4 translate as follows:

Theorem 3.5. (Wold decomposition theorem) *Let X_t be a regular univariate zero-mean covariance stationary time series process. Then X_t can be written as*

$$X_t = \sum_{j=0}^{\infty} \alpha_j U_{t-j} + W_t \text{ a.s.}, \quad (3.9)$$

where U_t is a zero-mean uncorrelated process with variance 1,

$$\alpha_j = E[X_t U_{t-j}], \quad \sum_{j=0}^{\infty} \alpha_j^2 < \infty, \quad (3.10)$$

and W_t is a zero-mean covariance stationary process satisfying

$$W_t \in \mathcal{U}_t^\perp \cap \mathcal{S}_{-\infty}, \quad (3.11)$$

where $\mathcal{S}_{-\infty} = \cap_n \text{span}(\{X_{n-k}\}_{k=1}^{\infty})$ and \mathcal{U}_t^\perp is the orthogonal complement of $\mathcal{U}_t = \text{span}(\{U_{t-k}\}_{k=0}^{\infty})$. The result (3.11) implies that

$$W_t \in \text{span}(\{W_{t-m}\}_{m=1}^{\infty}), \quad (3.12)$$

which in its turn implies that W_t is perfectly predictable from the past values $W_{t-1}, W_{t-2}, W_{t-3}, \dots$. Moreover, (3.11) implies that

$$E[W_t U_{t-m}] = 0 \quad (3.13)$$

for all leads and lags m .

The condition $\text{var}(U_t) = 1$ is not essential as long as X_t is regular. Without loss of generality we may then replace U_t with $\tilde{U}_t = \sigma U_t$, $\sigma > 0$, and α_k with $\tilde{\alpha}_k/\sigma$, where σ can be pinned down by normalizing $\tilde{\alpha}_0 = 1$.

The Wold decomposition carries over to k -variate covariance stationary processes X_t , as follows. Consider the Hilbert space \mathcal{R}_k of zero mean random vectors in \mathbb{R}^k with finite second moment matrices, endowed with the inner product $\langle X, Y \rangle = E[X'Y]$ and associated norm and metric. Let \hat{X}_t be the projection of X_t on $\text{span}(\{X_{t-j}\}_{j=1}^{\infty})$, with residual vector $V_t = X_t - \hat{X}_t$, and let $\Sigma = E[V_t V_t']$. In this case we need to extend the notion of regularity by requiring that Σ is positive definite rather than only $\|V_t\|^2 = E[V_t' V_t] > 0$, so that we can define $U_t = \Sigma^{-1/2} V_t$. Then the projection \tilde{X}_t of X_t on $\text{span}(\{U_{t-j}\}_{j=0}^n)$ takes the form $\tilde{X}_t = \sum_{j=1}^n A_j U_{t-j}$, where $A_j = E[X_t U_{t-j}']$. It follows now straightforwardly from the proofs of Theorems 3.4 and 3.5 that

$$X_t = \sum_{j=1}^{\infty} A_j U_{t-j} + W_t \text{ a.s.},$$

where the process U_t is uncorrelated with zero expectation vector and variance matrix I_k , and $W_t \in \mathcal{U}_t^\perp \cap \mathcal{S}_{-\infty}$, with \mathcal{U}_t^\perp and $\mathcal{S}_{-\infty}$ defined in Theorem 3.5.

It should be stressed that the deterministic process W_t is not necessarily nonrandom. For example let $W_t = a \cdot \cos(\lambda t) + b \cdot \sin(\lambda t)$, where a and b are independent random drawings from the standard normal distribution and $\lambda \in (-\pi/2, \pi/2)$ is a constant. Then $E[W_t] = 0$ and $E[W_t W_{t-m}] = \cos(\lambda m)$, hence W_t is a zero-mean covariance stationary process. If we observe W_{t-1} , W_{t-2} and W_{t-3} then we can solve a , b and λ , hence W_t is then determined for all t .

The question now arises under which conditions the deterministic process W_t is identical to zero. Since $W_t \in \cap_n \text{span}(\{X_{n-j}\}_{j=0}^\infty)$, it follows that W_t is measurable with respect to the remote σ -algebra of the process X_t :

Definition 3.5. Let $\mathcal{F}_t = \sigma(\{X_{t-j}\}_{j=0}^\infty)$ be the σ -algebra generated by $\{X_{t-j}\}_{j=0}^\infty$. The σ -algebra $\mathcal{F}_{-\infty} = \cap_t \mathcal{F}_t$ is called the remote σ -algebra of the process X_t .

If the process X_t is independent then it follows from Kolmogorov's zero-one law² that the sets in $\mathcal{F}_{-\infty}$ have either probability one or zero, so that the information in $\mathcal{F}_{-\infty}$ is non-informative. In other words, the memory of the remote past of X_t has vanished. However, this result carries over to certain dependent processes, for example α -mixing processes.³ This gives rise to the notion of vanishing memory:

Definition 3.6. A time series process is said to have a vanishing memory if the sets in its remote σ -algebra $\mathcal{F}_{-\infty}$ have either probability one or zero, i.e., $A \in \mathcal{F}_{-\infty}$ implies $P[A] = 1$ or $P[A] = 0$.

In that case $E[W_t | \mathcal{F}_{-\infty}] = E[W_t]$ a.s.⁴ However, since W_t is measurable $\mathcal{F}_{-\infty}$, we also have $E[W_t | \mathcal{F}_{-\infty}] = W_t$ a.s. Thus, $W_t = E[W_t] = 0$ a.s., where the second equality follows from the condition that $E[X_t] = 0$. Consequently,

Theorem 3.6. If the zero-mean covariance stationary process X_t has a vanishing memory then the deterministic term W_t in its Wold decomposition is zero with probability 1.

The Wold decomposition theorem in the form of Theorem 3.6 is the basis for time series analysis. In particular, for a univariate covariance stationary process X_t with a vanishing memory and expectation $E[X_t] = \mu$ the Wold decomposition can be written as

$$X_t = \mu + \alpha(L) U_t$$

where L is the lag operator and $\alpha(L) = 1 + \sum_{k=1}^{\infty} \alpha_k L^k$. The function $\alpha(L)$ can be approximated arbitrarily close by a ratio of two lag polynomials,

²See for example Bierens (2004, Theorem 7.5, p.185).

³See for example Bierens (2004, Theorem 7.6, p.186).

⁴See for example Bierens (2004, Exercise 3 in Section 7.6).

$\psi_q(L) = 1 + \sum_{k=1}^q \theta_k L^k$ and $\varphi_p(L) = 1 - \sum_{k=1}^p \gamma_k L^k$, of orders q and p , respectively, where at least $\varphi_p(L)$ is invertible with inverse $\varphi_p^{-1}(L)$.⁵ In particular, for arbitrary $\varepsilon > 0$ there exist lag polynomials $\psi_q(L)$ and $\varphi_p(L)$ such that

$$E \left[((\alpha(L) - \varphi_p^{-1}(L) \psi_q(L)) U_t)^2 \right] < \varepsilon.$$

This gives rise to the well-known ARMA(p, q) models, for which it is assumed that $\alpha(L)$ is exactly of the form $\alpha(L) = \varphi_p^{-1}(L) \psi_q(L)$, so that $\varphi_p(L) X_t = \gamma + \psi_q(L) U_t$ with $\gamma_0 = \varphi_p(1) \mu$. Thus,

$$X_t = \gamma_0 + \sum_{k=1}^p \gamma_k X_{t-k} + U_t + \sum_{m=1}^q \theta_m U_{t-m}.$$

Moreover, if also $\psi_q(L)$ is invertible then X_t has the representation

$$\psi_q^{-1}(L) \varphi_p(L) X_t = \beta_0 + U_t,$$

where $\beta_0 = \mu \cdot \varphi_p(1) / \psi_q(1)$. The lag function $\psi_q^{-1}(L) \varphi_p(L)$ can be written as $\psi_q^{-1}(L) \varphi_p(L) = 1 - \sum_{k=1}^{\infty} \beta_k L^k$, so that then X_t has the AR(∞) representation

$$X_t = \beta_0 + \sum_{k=1}^{\infty} \beta_k X_{t-k} + U_t.$$

This representation plays a key role in forecasting.

An important econometric application of the multivariate version of the Wold decomposition is Sims' (1980) innovation response analysis. Sims' (1980) landmark paper has changed the way empirical macroeconomics is conducted nowadays. His idea is the following. Let $X_t \in \mathbb{R}^k$ be a covariance stationary process of economic variables generated by a stationary VAR(p) process:

$$X_t = b_0 + \sum_{k=1}^p B_k X_{t-k} + U_t$$

Assume that the error vectors U_t are i.i.d. $N_k [0, \Sigma]$, where Σ is nonsingular. Stationarity of this process is equivalent to the requirement that the matrix-valued lag polynomial $B(L) = I_k - \sum_{k=1}^p B_k L^k$ is invertible.⁶ The latter

⁵I.e., $\varphi_p(z) = 0$ for some $z \in \mathbb{C}$ implies $|z| > 1$.

⁶Which in its turn is equivalent to the condition that the roots of the polynomial $\det(B(z))$ are all outside the complex unit circle: $\det(B(z)) = 0$ implies $|z| > 1$.

condition also guarantees that X_t has a vanishing memory. It follows then from the Wold decomposition that X_t can be decomposed as

$$X_t = \mu + \sum_{m=0}^{\infty} A_m U_{t-m},$$

where $\mu = E[X_t]$ and $A_0 = I_k$. The parameters Σ , μ , and A_m can be estimated by estimating the VAR(p) for X_t by ordinary least squares, and then inverting the VAR(p) lag polynomial.

The variance matrix Σ of the U_t 's can be written as $\Sigma = \Delta \cdot \Delta'$, where Δ is a $k \times k$ lower-triangular matrix, so that U_t can be written as $U_t = \Delta e_t$, where now $e_t \sim N_k [0, I_k]$. Sims proposes to interpret the components of e_t as the unpredictable parts of policy interventions in the corresponding components of X_t . To trace the effect of these policy innovations on the future path of X_t , project X_{t+m} for $m \geq 0$ on component $e_{i,t}$ of e_t . These projections take the form $A_m \delta_i e_{i,t}$, where δ_i is column i of Δ , and may be interpreted as the response of X_{t+m} to the innovation $e_{i,t}$. Since the scale of $e_{i,t}$ does not matter, the responses of X_{t+m} for $m = 0, 1, 2, \dots$ to a unit shock in $e_{i,t}$ are now $A_m \delta_i = E[X_{t+m}|e_{i,t} = 1] - E[X_{t+m}]$, which are usually presented in the form of graphs.

For more on the Wold decomposition and its time series applications, see for example Anderson (1994).

3.5 Projections on a random subspace

Because a Hilbert space \mathcal{H} is a metric space, we can define open sets in \mathcal{H} in the usual way. Therefore, similar to the Euclidean Borel field, we can define the Borel field $\mathcal{B}_{\mathcal{H}}$ of subsets of \mathcal{H} as the smallest σ -algebra containing the collection of all open sets in \mathcal{H} , and call its elements Borel sets. Moreover, given a probability space $\{\Omega, \mathcal{F}, P\}$, where \mathcal{F} is a σ -algebra of subsets of the sample space Ω and P is a probability measure on \mathcal{F} , a random element $X \in \mathcal{H}$ can now be defined as a mapping $X(\cdot) : \Omega \rightarrow \mathcal{H}$ such that for all Borel sets $B \in \mathcal{B}_{\mathcal{H}}$, $\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$. In particular, if X_n is a sequence of random elements of \mathcal{H} defined on a common probability space, the notion of convergence in probability of X_n to a non-random element x of \mathcal{H} , denoted by $\text{plim}_{n \rightarrow \infty} \|X_n - x\|$, can be defined in the same way as for

sequences of random vectors in a Euclidean space, i.e., for an arbitrary $\varepsilon > 0$,

$$\begin{aligned} \Pr [| | X_n - x | | < \varepsilon] &\stackrel{\text{def.}}{=} P(\{\omega \in \Omega : X(\omega) \in \{z \in \mathcal{H} : | | z - x | | < \varepsilon\}\}) \\ &\rightarrow 1 \text{ as } n \rightarrow \infty, \end{aligned}$$

because $\{z \in \mathcal{H} : | | z - x | | < \varepsilon\}$ is an open set and therefore a Borel set.

Now the question I will address is the following. Let Y_N and $X_{1,N}, X_{2,N}, \dots, X_{n,N}$ be random elements of \mathcal{H} depending on a sample of size N , where $n = n_N$ is a subsequence of N . Let $\hat{Y}_{n,N}$ be the projection of Y_N on $\text{span}(X_{1,N}, \dots, X_{n,N})$, and let $U_{n,N} = Y_N - \hat{Y}_{n,N}$ be the residual. Under what conditions do $\hat{Y}_{n,N}$ and $U_{n,N}$ converge in probability? The answer to this question is crucial for proving asymptotic normality of semi-nonparametric sieve estimators.

The answer is given in the following theorem.

Theorem 3.7. *Let Y_N and $X_{1,N}, X_{2,N}, \dots, X_{n,N}$ be random elements of a Hilbert space \mathcal{H} on the basis on a sample of size N , where n is a subsequence of N . Let $\hat{Y}_{n,N}$ be the projection of Y_N on $\text{span}(\{X_{m,N}\}_{m=1}^n)$, with residual $U_{n,N} = Y_N - \hat{Y}_{n,N}$. Suppose that the following conditions hold.*

(a) *There exists a non-random element y of \mathcal{H} such that*

$$\text{plim}_{N \rightarrow \infty} | | Y_N - y | | = 0. \quad (3.14)$$

(b) *There exist a sequence $\{x_m\}_{m=1}^\infty$ of non-random elements of \mathcal{H} and a sequence $\{\rho_m\}_{m=1}^\infty$ of positive numbers such that*

$$\text{plim}_{N \rightarrow \infty} \sum_{m=1}^n \rho_m | | X_{m,N} - x_m | | = 0 \quad (3.15)$$

and

$$\liminf_{n \rightarrow \infty} \left\| \sum_{m=1}^n \rho_m x_m \right\| > 0. \quad (3.16)$$

Then $\text{plim}_{N \rightarrow \infty} | | \hat{Y}_{n,N} - \hat{y} | | = 0$ and $\text{plim}_{N \rightarrow \infty} | | U_{n,N} - u | | = 0$, where \hat{y} is the projection of y on $\text{span}(\{x_m\}_{m=1}^\infty)$ and $u = y - \hat{y}$ is the residual involved.

3.6 Appendix: Proofs

3.6.1 Theorem 3.1

Recall that "subspace" means a sub-Hilbert space. Thus, \mathcal{S} is a Hilbert space.

Pick a sequence $z_n \in \mathcal{S}$ such that

$$\|y - z_n\| \leq \|y - \hat{y}\| + n^{-1}. \quad (3.17)$$

This is always possible because otherwise $\|y - z\| > \|y - \hat{y}\| + n^{-1}$ for all $z \in \mathcal{S}$ so that $\inf_{z \in \mathcal{S}} \|y - z\| \geq \|y - \hat{y}\| + n^{-1}$. Then

$$\lim_{n \rightarrow \infty} \|y - z_n\|^2 = \|y - \hat{y}\|^2 = \delta. \quad (3.18)$$

say. The first step is to show that z_n is a Cauchy sequence. Observe that

$$\begin{aligned} \|z_n - z_m\|^2 &= \|(z_n - y) - (z_m - y)\|^2 \\ &= \|z_n - y\|^2 - 2\langle z_n - y, z_m - y \rangle + \|z_m - y\|^2 \end{aligned}$$

and

$$\begin{aligned} 4.\|0.5(z_n + z_m) - y\|^2 &= \|(z_n - y) + (z_m - y)\|^2 \\ &= \|z_n - y\|^2 + 2\langle z_n - y, z_m - y \rangle + \|z_m - y\|^2 \end{aligned}$$

Adding these two equation up yields

$$\|z_n - z_m\|^2 = 2\|z_n - y\|^2 + 2\|z_m - y\|^2 - 4.\|0.5(z_n + z_m) - y\|^2 \quad (3.19)$$

Because $0.5(z_n + z_m) \in \mathcal{S}$, it follows that $\|0.5(z_n + z_m) - y\|^2 \geq \delta^2$, whereas by (3.17) and (3.18), $\|z_n - y\|^2 \leq (\delta + n^{-1})^2$ and $\|z_m - y\|^2 \leq (\delta + m^{-1})^2$. Therefore, it follows from (3.19) that

$$\begin{aligned} \|z_n - z_m\|^2 &\leq 2(\delta + n^{-1})^2 + 2(\delta + m^{-1})^2 - 4\delta^2 \\ &= 4\delta/n + 2n^{-2} + 4\delta/m + 2m^{-2}. \end{aligned}$$

Thus, z_n is a Cauchy sequence in \mathcal{S} and therefore takes a limit \hat{y} in \mathcal{S} .

The next step is to show that for all $z \in \mathcal{S}$, $\langle y - \hat{y}, z \rangle = 0$, as follows. Note that for any real scalar c , $\hat{y} + c.z \in \mathcal{S}$ and therefore

$$\|y - \hat{y}\|^2 \leq \|y - \hat{y} - c.z\|^2 = \|y - \hat{y}\|^2 - 2c.\langle y - \hat{y}, z \rangle + c^2\|z\|^2$$

The right-hand side is minimal for $c = \langle y - \hat{y}, z \rangle / \|z\|^2$, hence

$$0 \leq -\frac{(\langle y - \hat{y}, z \rangle)^2}{\|z\|^2}$$

and thus $\langle y - \hat{y}, z \rangle = 0$.

Note that this argument only applies if the Hilbert space \mathcal{H} is real. If \mathcal{H} is complex this orthogonality proof can be adapted similar to the proof of Theorem 2.1.

Finally, we need to show that \hat{y} is unique. Suppose that there exists another projection $\tilde{y} \in \mathcal{S}$. Then also $\langle y - \tilde{y}, z \rangle = 0$, and thus $\langle y - \tilde{y}, z \rangle - \langle y - \hat{y}, z \rangle = \langle \hat{y} - \tilde{y}, z \rangle = 0$. But $z = y - \tilde{y} \in \mathcal{S}$ so that $\|\hat{y} - \tilde{y}\|^2 = \langle \hat{y} - \tilde{y}, \hat{y} - \tilde{y} \rangle = 0$. Consequently, \hat{y} is unique.

3.6.2 Lemma 3.1

Let $\mathcal{M}_n = \text{span}(\{x_k\}_{k=1}^n)$ and $\mathcal{M}_\infty = \text{span}(\{x_k\}_{k=1}^\infty) = \overline{\cup_{n=1}^\infty \mathcal{M}_n}$. If $z \in \cup_{n=1}^\infty \mathcal{M}_n$ then there exists an n_0 such that $z \in \mathcal{M}_{n_0}$, hence for $n \geq n_0$, $\hat{z}_n = z$ and thus $\lim_{n \rightarrow \infty} \|z - \hat{z}_n\| = 0$. Now let $z \in \mathcal{M}_\infty \setminus (\cup_{n=1}^\infty \mathcal{M}_n)$. Since $\mathcal{M}_\infty = \overline{\cup_{n=1}^\infty \mathcal{M}_n}$ is closed and $\mathcal{M}_n \subset \mathcal{M}_{n+1}$, for each n there exists an $z_n \in \mathcal{M}_n$ such that $\lim_{n \rightarrow \infty} \|z - z_n\|^2 = 0$, hence for $n \rightarrow \infty$, $\|z - \hat{z}_n\|^2 \leq \|z - z_n\|^2 \rightarrow 0$.

3.6.3 Theorem 3.2

Adopting the notation in the proof of Lemma 3.1, we may without loss of generality assume that $\hat{z} \in \mathcal{M}_\infty \setminus (\cup_{n=1}^\infty \mathcal{M}_n)$, as otherwise the result of Theorem 3.2 holds trivially. Since \mathcal{M}_∞ is closed this assumption implies that for each n we can select a $z_n \in \mathcal{M}_n$ such that

$$\lim_{n \rightarrow \infty} \|\hat{z} - z_n\| = 0. \quad (3.20)$$

Let $\|z - \hat{z}\| = \delta$ and $\|z - \hat{z}_n\| = \delta_n$, and note that $\delta_n \geq \delta$. Since

$$\begin{aligned} \delta_n^2 &= \|z - \hat{z}_n\|^2 \leq \|z - z_n\|^2 = \|z - \hat{z} + \hat{z} - z_n\|^2 \\ &= \|z - \hat{z}\|^2 + \|\hat{z} - z_n\|^2 + 2 \langle z - \hat{z}, \hat{z} - z_n \rangle \\ &= \delta^2 + \|\hat{z} - z_n\|^2 \end{aligned}$$

it follows from (3.20) that

$$\lim_{n \rightarrow \infty} \delta_n = \delta. \quad (3.21)$$

Recall that $z = \hat{z} + u$, where $\langle u, x \rangle = 0$ for all $x \in \mathcal{M}_\infty$. Hence

$$\begin{aligned} \|\hat{z} - \hat{z}_n\|^2 &= \|z - \hat{z}_n - u\|^2 = \|z - \hat{z}_n\|^2 + \|u\|^2 - 2\langle z - \hat{z}_n, u \rangle \\ &= \|z - \hat{z}_n\|^2 + \|u\|^2 - 2\langle z, u \rangle = \delta_n^2 - \delta^2 \end{aligned} \quad (3.22)$$

where the last equality follows from $\langle z, u \rangle - \langle u, u \rangle = \langle \hat{z}, u \rangle = 0$ and $\langle u, u \rangle = \|u\|^2 = \delta^2$. The theorem now follows from (3.21) and (3.22).

3.6.4 Theorem 3.3

Due to the orthonormality condition (3.7), the projection \hat{z}_n of z on $\mathcal{M}_n = \text{span}(\{x_j\}_{j=1}^n)$ takes the form

$$\hat{z}_n = \sum_{j=1}^n \theta_j x_j, \text{ where } \theta_j = \langle z, x_j \rangle. \quad (3.23)$$

Moreover, denoting $u_n = z - \hat{z}_n$, it follows from (3.7) and (3.23) that

$$\begin{aligned} \|u_n\|^2 &= \left\| z - \sum_{j=1}^n \theta_j x_j \right\|^2 = \|z\|^2 - 2 \sum_{j=1}^n \theta_j \langle z, x_j \rangle + \sum_{j=1}^n \sum_{i=1}^n \theta_j \theta_i \langle x_j, x_i \rangle \\ &= \|z\|^2 - \sum_{j=1}^n \theta_j^2 \geq 0 \end{aligned} \quad (3.24)$$

hence $\sum_{j=1}^n \theta_j^2 \leq \|z\|^2$ for all n and thus $\sum_{j=1}^\infty \theta_j^2 < \infty$. Finally, it follows from Theorem 3.2 that

$$\lim_{n \rightarrow \infty} \left\| \hat{z} - \sum_{j=1}^n \theta_j x_j \right\|^2 = \lim_{n \rightarrow \infty} \|\hat{z} - \hat{z}_n\|^2 = 0$$

so that we can write $\hat{z} = \sum_{j=1}^\infty \theta_j x_j$.

3.6.5 Lemma 3.2

Let x be an arbitrary element of a subspace \mathcal{S} of an Hilbert space \mathcal{H} and let y_n be a Cauchy sequence in \mathcal{S}^\perp . Then there exists an $y \in \mathcal{H}$ such that $\lim_{n \rightarrow \infty} \|y - y_n\| = 0$. Since $\langle x, y_n \rangle = 0$ we have $\langle x, y \rangle = \langle x, y - y_n \rangle$. It follows now from the Cauchy-Schwarz inequality that $|\langle x, y \rangle| = |\langle x, y - y_n \rangle| \leq \|x\| \cdot \|y - y_n\| \rightarrow 0$. Hence $y \in \mathcal{S}^\perp$.

3.6.6 Theorem 3.4

Denote $\mathcal{S}_n = \text{span}(\{x_k\}_{k=n}^\infty)$. Project each x_k on \mathcal{S}_{k+1} , so that $x_k = \hat{x}_k + u_k$ with projection $\hat{x}_k \in \mathcal{S}_{k+1}$ and residual u_k . Recall that by the regularity condition, $\|u_k\| > 0$, hence $e_k = u_k/\|u_k\|$ is well defined. It is not hard to verify that the residuals u_k are orthogonal, so that the e_k 's are orthonormal. Next, denote

$$\mathcal{U}_n = \text{span}(e_1, \dots, e_n) = \text{span}(u_1, \dots, u_n),$$

and let \mathcal{U}_n^\perp be the orthogonal complement of \mathcal{U}_n . Note that

$$\mathcal{U}_{n+1}^\perp \subset \mathcal{U}_n^\perp. \quad (3.25)$$

To see this, let $z \in \mathcal{U}_{n+1}^\perp$. Then for all $x \in \mathcal{U}_{n+1}$, $\langle z, x \rangle = 0$, and because obviously $\mathcal{U}_n \subset \mathcal{U}_{n+1}$, it follows that also $\langle z, x \rangle = 0$ for all $x \in \mathcal{U}_n$. Hence, $z \in \mathcal{U}_n^\perp$.

As before, let $\mathcal{M}_n = \text{span}(\{x_k\}_{k=1}^n)$.

The theorem under review will be proved in six steps:

Step 1. First I will show that

$$\mathcal{M}_n \subset \text{span}(\mathcal{U}_n, \mathcal{U}_n^\perp \cap \mathcal{S}_2). \quad (3.26)$$

Proof. Let $z \in \mathcal{M}_n$ be arbitrary. Recall that z takes the form $z = \sum_{k=1}^n c_k x_k$. Substituting $x_k = \hat{x}_k + u_k = \hat{x}_k + \|u_k\|e_k$ we can write z as

$$\begin{aligned} z &= \sum_{k=1}^n c_k (\hat{x}_k + u_k) = \sum_{k=1}^n c_k u_k + \sum_{k=1}^n c_k \hat{x}_k \\ &= \sum_{k=1}^n c_k \|u_k\| e_k + \sum_{k=1}^n c_k \hat{x}_k \end{aligned}$$

Note that

$$\sum_{k=1}^n c_k \hat{x}_k \in \mathcal{S}_2 \quad (3.27)$$

because $\hat{x}_k \in \mathcal{S}_{k+1} \subset \mathcal{S}_2$.

Next, project $\sum_{k=1}^n c_k \hat{x}_k$ on \mathcal{U}_n . This projection takes the form $\hat{p}_n = \sum_{k=1}^n d_k e_k$ with residual $w_{n+1} \in \mathcal{S}_2$. The latter follows from (3.27). But since

w_{n+1} is a residual of a projection on \mathcal{U}_n we also have $\langle e_k, w_{n+1} \rangle = 0$ for $k = 1, \dots, n$, hence $w_{n+1} \in \mathcal{U}_n^\perp$. Thus,

$$w_{n+1} \in \mathcal{U}_n^\perp \cap \mathcal{S}_2.$$

Denoting $\alpha_k = c_k \|u_k\| + d_k$, we can now write

$$z = \sum_{k=1}^n \alpha_k e_k + w_{n+1}, \text{ where } w_{n+1} \in \mathcal{U}_n^\perp \cap \mathcal{S}_2.$$

Therefore, (3.26) holds.

Step 2. I will now show that

$$\text{span}(\mathcal{U}_n, \mathcal{U}_n^\perp \cap \mathcal{S}_2) = \text{span}(\mathcal{U}_n, \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1}). \quad (3.28)$$

Proof. Denote

$$\mathcal{S}_{k,m} = \text{span}(\{x_j\}_{j=k}^m)$$

for $m \geq k$ and let $z \in \mathcal{U}_n^\perp \cap \mathcal{S}_{2,m}$ for some $m \geq 2$. Consider first the case $m > n$. Since $z \in \mathcal{S}_{2,m}$ there exists constants c_k such that

$$\begin{aligned} z &= \sum_{k=2}^m c_k x_k = \sum_{k=2}^n c_k (\hat{x}_k + u_k) + \sum_{k=n+1}^m c_k x_k \\ &= \sum_{k=2}^n c_k \|u_k\| e_k + \sum_{k=2}^n c_k \hat{x}_k + \sum_{k=n+1}^m c_k x_k. \end{aligned}$$

Moreover, since $z \in \mathcal{U}_n^\perp$ it follows that $\langle z, e_k \rangle = 0$ for $k = 1, \dots, n$. In particular,

$$\begin{aligned} 0 &= \langle z, e_2 \rangle = c_2 \|u_2\| + \sum_{k=2}^n c_k \langle \hat{x}_k, e_2 \rangle + \sum_{k=n+1}^m c_k \langle x_k, e_2 \rangle \\ &= c_2 \|u_2\| \end{aligned}$$

because $\sum_{k=2}^n c_k \hat{x}_k \in S_3$, $\sum_{k=n+1}^m c_k x_k \in S_{n+1}$, and e_2 is orthogonal to S_3 and S_{n+1} . Hence $c_2 = 0$ and thus

$$z = \sum_{k=3}^n c_k \|u_k\| e_k + \sum_{k=3}^n c_k \hat{x}_k + \sum_{k=n+1}^m c_k x_k.$$

It follows now similarly that $c_k = 0$ for $k = 3, \dots, n$, hence

$$z = \sum_{k=n+1}^m c_k x_k \in \mathcal{S}_{n+1,m}.$$

Because $z \in \mathcal{U}_n^\perp$ as well, it follows now that

$$z \in \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1,m},$$

which implies

$$\mathcal{U}_n^\perp \cap \mathcal{S}_{2,m} \subset \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1,m}$$

because $z \in \mathcal{U}_n^\perp \cap \mathcal{S}_{2,m}$ was arbitrary. However, $\mathcal{S}_{n+1,m} \subset \mathcal{S}_{2,m}$ and therefore

$$\mathcal{U}_n^\perp \cap \mathcal{S}_{n+1,m} \subset \mathcal{U}_n^\perp \cap \mathcal{S}_{2,m},$$

so that

$$\mathcal{U}_n^\perp \cap \mathcal{S}_{2,m} = \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1,m} \text{ for } m > n.$$

This result implies that

$$\mathcal{U}_n^\perp \cap (\cup_{m=n+1}^\infty \mathcal{S}_{2,m}) = \mathcal{U}_n^\perp \cap (\cup_{m=n+1}^\infty \mathcal{S}_{n+1,m}) \quad (3.29)$$

In the case $m \leq n$, $z \in \mathcal{U}_n^\perp \cap \mathcal{S}_{2,m}$ implies that $z = 0$, as can be straightforwardly verified from the above argument, so that $\mathcal{U}_n^\perp \cap \mathcal{S}_{2,m} = \{0\}$ for $m = 2, 3, \dots, n$. Since Hilbert spaces are vector spaces and therefore always contain the null element it follows that

$$\begin{aligned} \cup_{m=2}^\infty \mathcal{U}_n^\perp \cap \mathcal{S}_{2,m} &= \{0\} \cup (\cup_{m=n+1}^\infty \mathcal{U}_n^\perp \cap \mathcal{S}_{2,m}) \\ &= \cup_{m=n+1}^\infty \mathcal{U}_n^\perp \cap \mathcal{S}_{2,m}, \end{aligned}$$

hence

$$\mathcal{U}_n^\perp \cap (\cup_{m=2}^\infty \mathcal{S}_{2,m}) = \mathcal{U}_n^\perp \cap (\cup_{m=n+1}^\infty \mathcal{S}_{2,m}). \quad (3.30)$$

Since by Definition 2.10,

$$\mathcal{S}_2 = \overline{\cup_{m=2}^\infty \mathcal{S}_{2,m}}, \quad \mathcal{S}_{n+1} = \overline{\cup_{m=n+1}^\infty \mathcal{S}_{n+1,m}}$$

it follows now from (3.30) that

$$\begin{aligned} \mathcal{U}_n^\perp \cap \mathcal{S}_2 &= \mathcal{U}_n^\perp \cap \overline{\cup_{m=2}^\infty \mathcal{S}_{2,m}} \\ &= \mathcal{U}_n^\perp \cap \overline{\cup_{m=n+1}^\infty \mathcal{S}_{n+1,m}} = \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1} \end{aligned}$$

which implies that (3.28) holds.

Step 3. Denote $\mathcal{R}_n = \text{span}(\mathcal{U}_n, \mathcal{U}_n^\perp \cap \mathcal{S}_{n+1})$. Then

$$\mathcal{S}_1 = \overline{\cup_{n=1}^{\infty} \mathcal{R}_n}. \quad (3.31)$$

Proof. Combining (3.26) and (3.28) yields $\mathcal{M}_n \subset \mathcal{R}_n$, hence

$$\mathcal{S}_1 = \overline{\cup_{n=1}^{\infty} \mathcal{M}_n} \subset \overline{\cup_{n=1}^{\infty} \mathcal{R}_n}, \quad (3.32)$$

where the equality follows from Definition 2.10. However, we also have $\mathcal{R}_n \subset \mathcal{S}_1$, as is not hard to verify, hence

$$\overline{\cup_{n=1}^{\infty} \mathcal{R}_n} \subset \mathcal{S}_1. \quad (3.33)$$

Thus, the result (3.31) follows from (3.32) and (3.33).

Step 4. For an $x \in \mathcal{S}_1$, let \hat{x}_n be the projection of x on \mathcal{R}_n . Then

$$\hat{x}_n = \sum_{j=1}^n \alpha_j e_j + w_{n+1} \quad (3.34)$$

where $\alpha_j = \langle x, e_j \rangle$ and w_{n+1} is the projection of x on $\mathcal{U}_n^\perp \cap S_{n+1}$. Moreover,

$$\sum_{j=1}^{\infty} \alpha_j^2 < \infty. \quad (3.35)$$

Furthermore,

$$\lim_{n \rightarrow \infty} \left\| x - \sum_{j=1}^n \alpha_j e_j - w_{n+1} \right\| = 0. \quad (3.36)$$

Proof. By the definition of \mathcal{R}_n and by Definition 3.3, $\hat{x}_n = \sum_{j=1}^n \theta_j e_j + w$ for some constants θ_j and a $w \in \mathcal{U}_n^\perp \cap S_{n+1}$. To determine the θ_j 's and w , note that

$$\left\| x - \sum_{j=1}^n \theta_j e_j - w \right\|^2 = \|x - w\|^2 - 2 \sum_{j=1}^n \theta_j \langle e_j, x \rangle + 2 \sum_{j=1}^n \theta_j \langle e_j, w \rangle$$

$$\begin{aligned}
& + \left\| \sum_{j=1}^n \theta_j e_j \right\|^2 \\
& = \|x - w\|^2 - 2 \sum_{j=1}^n \theta_j \langle e_j, x \rangle + \sum_{j=1}^n \theta_j^2
\end{aligned}$$

because $w \in \mathcal{U}_n^\perp \cap S_{n+1} \subset \mathcal{U}_n^\perp$ implies $\langle e_j, w \rangle = 0$ and

$$\left\| \sum_{j=1}^n \theta_j e_j \right\|^2 = \sum_{j=1}^n \sum_{i=1}^n \theta_j \theta_i \langle e_j, e_i \rangle = \sum_{j=1}^n \theta_j^2 \langle e_j, e_j \rangle = \sum_{j=1}^n \theta_j^2.$$

Thus

$$\begin{aligned}
\|x - \hat{x}_n\|^2 & = \inf_{\theta_1, \dots, \theta_n, w \in \mathcal{U}_n^\perp \cap S_{n+1}} \left\| x - \sum_{j=1}^n \theta_j e_j - w \right\|^2 \\
& = \inf_{\theta_1, \dots, \theta_n, w \in \mathcal{U}_n^\perp \cap S_{n+1}} \left(\|x - w\|^2 - 2 \sum_{j=1}^n \theta_j \langle e_j, x \rangle + \sum_{j=1}^n \theta_j^2 \right) \\
& = \inf_{w \in \mathcal{U}_n^\perp \cap S_{n+1}} \|x - w\|^2 - \sum_{j=1}^n \alpha_j^2 \\
& = \|x - w_{n+1}\|^2 - \sum_{j=1}^n \alpha_j^2
\end{aligned} \tag{3.37}$$

where $\alpha_j = \langle x, e_j \rangle$ and w_{n+1} is the projection of x on $\mathcal{U}_n^\perp \cap S_{n+1}$.

This result implies that for all n ,

$$\sum_{j=1}^n \alpha_j^2 \leq \|x - w_{n+1}\|^2 \leq \|x\|^2 \tag{3.38}$$

so that (3.35) holds.

Finally, to prove (3.36), let \hat{x} be the projection of x on $\overline{\cup_{n=1}^{\infty} \mathcal{R}_n}$. Then it follows from Theorem 3.2 that $\lim_{n \rightarrow \infty} \|\hat{x}_n - \hat{x}\| = 0$. But (3.31) implies $\hat{x} \in \mathcal{S}_1$, hence $x = \hat{x}$, so that $\lim_{n \rightarrow \infty} \|\hat{x}_n - x\| = 0$.

Step 5. Let $z_n = \sum_{j=1}^n \alpha_j e_j$. Then

$$\lim_{n \rightarrow \infty} \|z - z_n\| = 0, \text{ where } z \in \mathcal{U}_\infty. \tag{3.39}$$

Proof. This follows from the fact that z_n is a Cauchy sequence in $\mathcal{U}_\infty = \text{span}(\{e_k\}_{k=1}^\infty)$ because

$$\begin{aligned}\|z_n - z_m\|^2 &= \left\| \sum_{j=\min(m,n)+1}^{\max(m,n)} \alpha_j e_j \right\|^2 \\ &= \sum_{j=\min(m,n)+1}^{\max(m,n)} \alpha_j^2 \leq \sum_{j=\min(m,n)+1}^{\infty} \alpha_j^2 \\ &\rightarrow 0\end{aligned}$$

as $\min(m, n) \rightarrow \infty$, where the latter is due to $\sum_{j=1}^\infty \alpha_j^2 < \infty$.

Step 6. There exists a $w \in \mathcal{U}_\infty^\perp \cap S_\infty$ such that

$$\lim_{n \rightarrow \infty} \|w_{n+1} - w\| = 0. \quad (3.40)$$

Proof. Recall from Step 4 that

$$w_{n+1} \in \mathcal{U}_n^\perp \cap S_{n+1}.$$

Moreover, it follows from (3.25) and the definition of S_{n+1} that for an arbitrary $k \geq 1$,

$$\mathcal{U}_n^\perp \cap S_{n+1} \subset \mathcal{U}_k^\perp \cap S_{k+1} \text{ for } n \geq k$$

hence

$$w_{n+1} \in \mathcal{U}_k^\perp \cap S_{k+1} \text{ for } n \geq k.$$

Furthermore for $n \geq k$, w_{n+1} is a Cauchy sequence in $\mathcal{U}_k^\perp \cap S_{k+1}$ because

$$\begin{aligned}\|w_{n+1} - w_{m+1}\| &= \|\widehat{x}_n - z_n - \widehat{x}_m + z_m\| \\ &\leq \|\widehat{x}_n - \widehat{x}_m\| + \|z_n - z_m\| \\ &\leq \|\widehat{x}_n - x\| + \|\widehat{x}_m - x\| + \|z_n - z_m\| \\ &\rightarrow 0\end{aligned}$$

as $\min(m, n) \rightarrow \infty$. Thus, there exists a $w \in \mathcal{U}_k^\perp \cap S_{k+1}$ such that (3.40) holds. Since k was arbitrary we have $w \in \cap_{k=1}^\infty \mathcal{U}_k^\perp = \mathcal{U}_\infty^\perp$ and $w \in \cap_{k=1}^\infty S_{k+1} = S_\infty$, hence

$$w \in \mathcal{U}_\infty^\perp \cap S_\infty.$$

This completes the proof of Step 6.

The theorem now follows from (3.35), (3.39), (3.40) and the fact that $w \in \mathcal{U}_\infty^\perp \cap S_\infty \subset \mathcal{U}_\infty^\perp$, which implies that $\langle w, e_k \rangle = 0$ for $k \in \mathbb{N}$.

3.6.7 Theorem 3.5

Recall that $U_t = \tilde{U}_t / \sqrt{E[\tilde{U}_t^2]}$, where $\tilde{U}_t = X_t - \hat{X}_t$ with \hat{X}_t the projection of X_t on $\text{span}(\{X_{t-j}\}_{j=1}^\infty)$. The uncorrelatedness of the \tilde{U}_t 's follows from Theorem 3.4, but we still need to show that $E[\tilde{U}_t] = 0$ and $E[\tilde{U}_t^2] = \sigma^2$ for all t .

Proof of $E[\tilde{U}_t] = 0$

Let $\hat{X}_{t,n}$ be the projection of X_t on $\text{span}(\{X_{t-j}\}_{j=1}^n)$. Then $\hat{X}_{t,n}$ takes the form

$$\hat{X}_{t,n} = \sum_{j=1}^n \beta_{j,n} X_{t-j},$$

where the $\beta_{j,n}$'s do not depend on t . The latter follows from the fact that the $\beta_{j,n}$'s are the solutions of the normal equations

$$\sum_{j=1}^n \beta_{j,n} \gamma(i-j) = \gamma(i), \quad i = 1, 2, \dots, n,$$

where $\gamma(i) = E[X_t X_{t-i}]$ is the covariance function of X_t . Hence $E[\hat{X}_{t,n}] = 0$.

It follows from Theorem 3.2 that

$$\lim_{n \rightarrow \infty} \|\hat{X}_{t,n} - \hat{X}_t\|^2 = \lim_{n \rightarrow \infty} E[(\hat{X}_{t,n} - \hat{X}_t)^2] = 0 \quad (3.41)$$

so that by Liapounov's inequality and $E[\hat{X}_{t,n}] = 0$,

$$\begin{aligned} \lim_{n \rightarrow \infty} |E[\hat{X}_t]| &= \lim_{n \rightarrow \infty} |E[\hat{X}_t - \hat{X}_{t,n}]| \leq \lim_{n \rightarrow \infty} E[|\hat{X}_t - \hat{X}_{t,n}|] \\ &\leq \sqrt{\lim_{n \rightarrow \infty} E[(\hat{X}_{t,n} - \hat{X}_t)^2]} = 0. \end{aligned}$$

Thus $E[\hat{X}_t] = 0$ and therefore $E[\tilde{U}_t] = E[X_t - \hat{X}_t] = 0$.

Proof of $E[\tilde{U}_t^2] = \sigma^2$

Let $\tilde{U}_{t,n} = X_t - \hat{X}_{t,n}$. It follows from (3.41) that

$$\lim_{n \rightarrow \infty} E \left[(\tilde{U}_t - \tilde{U}_{t,n})^2 \right] = \lim_{n \rightarrow \infty} E \left[(\hat{X}_{t,n} - \hat{X}_t)^2 \right] = 0 \quad (3.42)$$

Moreover,

$$\begin{aligned} E \left[\tilde{U}_{t,n}^2 \right] &= \|X_t - \hat{X}_{t,n}\|^2 = E \left[\left(X_t - \sum_{j=1}^n \beta_{j,n} X_{t-j} \right)^2 \right] \\ &= \gamma(0) - 2 \sum_{j=1}^n \beta_{j,n} \gamma(j) + \sum_{j=1}^n \sum_{i=1}^n \beta_{j,n} \beta_{i,n} \gamma(i-j) \\ &= \sigma_n^2 \end{aligned}$$

say, which does not depend on t . Furthermore, note that σ_n^2 is non-increasing in n , so that

$$\lim_{n \rightarrow \infty} \sigma_n^2 = \sigma^2$$

exists, and that

$$\begin{aligned} E \left[(\tilde{U}_t - \tilde{U}_{t,n})^2 \right] &= \|\hat{X}_{t,n} - \hat{X}_t\|^2 = \|\hat{X}_{t,n} - X_t + \tilde{U}_t\|^2 \\ &= \|\hat{X}_{t,n} - X_t\|^2 + 2 \langle \hat{X}_{t,n} - X_t, \tilde{U}_t \rangle + \|\tilde{U}_t\|^2 \\ &= \|\tilde{U}_{t,n}\|^2 - 2 \langle X_t, \tilde{U}_t \rangle + \|\tilde{U}_t\|^2 \\ &= \|\tilde{U}_{t,n}\|^2 - 2 \langle \hat{X}_t + \tilde{U}_t, \tilde{U}_t \rangle + \|\tilde{U}_t\|^2 \\ &= \|\tilde{U}_{t,n}\|^2 - 2 \langle \tilde{U}_t, \tilde{U}_t \rangle + \|\tilde{U}_t\|^2 \\ &= \|\tilde{U}_{t,n}\|^2 - \|\tilde{U}_t\|^2 \\ &= E \left[\tilde{U}_{t,n}^2 \right] - E \left[\tilde{U}_t^2 \right]. \end{aligned}$$

Thus,

$$E \left[\tilde{U}_t^2 \right] = \sigma_n^2 - E \left[(\tilde{U}_t - \tilde{U}_{t,n})^2 \right] \rightarrow \sigma^2.$$

Proof of (3.10), (3.11) and (3.13)

The result of Theorem 3.4 can now be translated as

$$\lim_{n \rightarrow \infty} \left\| X_t - \sum_{j=0}^n \alpha_j U_{t-j} - W_t \right\| = 0, \quad (3.43)$$

where U_t is a zero-mean uncorrelated covariance stationary process with unit variance, and $\alpha_k = \langle X_t, U_{t-k} \rangle = E[X_t U_{t-k}]$ with $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.

We still need to prove that the α_k 's do not depend on t , as follows. Recall from the proof of $E[\tilde{U}_t^2] = \sigma^2$ that $\tilde{U}_{t,n} = X_t - \sum_{j=1}^n \beta_{j,n} X_{t-j}$, so that

$$E[X_{t+k} \tilde{U}_{t,n}] = \gamma(k) - \sum_{j=1}^n \beta_{j,n} \gamma(k+j),$$

which does not depend on t . Moreover, by the Cauchy-Schwarz inequality and (3.42),

$$\lim_{n \rightarrow \infty} \left| E[X_{t+k} (\tilde{U}_{t,n} - \tilde{U}_t)] \right|^2 \leq \gamma(0) \lim_{n \rightarrow \infty} E[(\tilde{U}_{t,n} - \tilde{U}_t)^2] = 0.$$

Thus $E[X_{t+k} \tilde{U}_t] = \lim_{n \rightarrow \infty} E[X_{t+k} \tilde{U}_{t,n}]$. Since the latter does not depend on t , neither does $\alpha_k = E[X_{t+k} U_t] = E[X_{t+k} \tilde{U}_t / ||\tilde{U}_t||]$.

The results (3.11) and (3.13) follow straightforwardly from Theorem 3.4.

Proof of (3.9)

The result (3.43) implies, by Chebyshev's inequality, that

$$X_t = \text{plim}_{n \rightarrow \infty} \sum_{j=0}^n \alpha_j U_{t-j} + W_t. \quad (3.44)$$

Recall that convergence in probability for $n \rightarrow \infty$ is equivalent to a.s. convergence along a further subsequence k_m of an arbitrary subsequence of n . See for example Bierens (2004, Theorem 6.B.3, p. 168). Thus for such a subsequence k_m ,

$$\sum_{j=0}^{k_m} \alpha_j U_{t-j} \xrightarrow{a.s.} X_t - W_t \quad (3.45)$$

as $m \rightarrow \infty$, and the same holds for any further subsequence of k_m .

Without loss of generality we may choose $k_0 = 0$. Then for each $n > 0$ we can find an m_n such that

$$k_{m_{n-1}} < n \leq k_{m_n}. \quad (3.46)$$

Moreover, (3.45) implies that

$$\sum_{j=0}^{k_{m_n}} \alpha_j U_{t-j} \xrightarrow{a.s.} X_t - W_t \text{ as } n \rightarrow \infty. \quad (3.47)$$

Due to (3.46),

$$\begin{aligned} \sum_{n=1}^{\infty} E \left[\left(\sum_{j=0}^{k_{m_n}} \alpha_j U_{t-j} - \sum_{j=0}^n \alpha_j U_{t-j} \right)^2 \right] &= \sum_{n=1}^{\infty} E \left[\left(\sum_{j=n+1}^{k_{m_n}} \alpha_j U_{t-j} \right)^2 \right] \\ &\leq \sum_{n=1}^{\infty} \sum_{j=k_{m_{n-1}}+1}^{k_{m_n}} \alpha_j^2 \leq \sum_{j=0}^{\infty} \alpha_j^2 < \infty, \end{aligned}$$

so that by Chebyshev's inequality, for arbitrary $\varepsilon > 0$,

$$\sum_{n=0}^{\infty} \Pr \left[\left| \sum_{j=0}^{k_{m_n}} \alpha_j U_{t-j} - \sum_{j=0}^n \alpha_j U_{t-j} \right| > \varepsilon \right] < \infty.$$

This result implies, by the Borel-Cantelli lemma,⁷ that

$$\sum_{j=0}^{k_{m_n}} \alpha_j U_{t-j} - \sum_{j=0}^n \alpha_j U_{t-j} \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty. \quad (3.48)$$

Combining (3.47) and (3.48) it follows now that

$$\sum_{j=0}^n \alpha_j U_{t-j} \xrightarrow{a.s.} X_t - W_t \text{ as } n \rightarrow \infty. \quad (3.49)$$

Since $\sum_{j=0}^{\infty} \alpha_j U_{t-j}$ is defined as $\lim_{n \rightarrow \infty} \sum_{j=0}^n \alpha_j U_{t-j}$, (3.9) is equivalent to (3.49).

⁷See for example Bierens (2004, Theorem 6.B.2, p. 168).

The zero-mean covariance stationarity of W_t

It follows now trivially from (3.9) that $E[W_t] = 0$. Moreover, it is left as an exercise to show that for $m \geq 0$,

$$E[W_t W_{t-m}] = \gamma(m) - \sum_{j=0}^{\infty} \alpha_{j+m} \alpha_j. \quad (3.50)$$

Proof of (3.12)

Finally, $W_t \in \cap_n \text{span}(\{X_{n-j}\}_{j=0}^{\infty})$ implies that $W_t \in \text{span}(\{X_{t-j}\}_{j=1}^{\infty})$, hence the projection of W_t on $\text{span}(\{X_{t-j}\}_{j=1}^{\infty})$ is W_t itself. Since by (3.9),

$$\text{span}(\{X_{t-j}\}_{j=1}^{\infty}) = \text{span}(\text{span}(\{U_{t-j}\}_{j=1}^{\infty}), \text{span}(\{W_{t-j}\}_{j=1}^{\infty}))$$

and the projection of W_t on $\text{span}(\{U_{t-j}\}_{j=1}^{\infty})$ is zero, it follows that the projection of W_t on $\text{span}(\{W_{t-j}\}_{j=1}^{\infty})$ is W_t itself, which proves (3.12). ■

3.6.8 Theorem 3.7

Note that $\|\hat{Y}_{n,N} - \hat{y}\| = \|(u - U_{n,N}) - (Y_N - y)\| \leq \|U_{n,N} - u\| + \|Y_N - y\|$, hence

$$\|\hat{Y}_{n,N} - \hat{y}\| \leq \|U_{n,N} - u\| + o_p(1),$$

where the $o_p(1)$ term follows from condition (a). Therefore it suffices to prove $\|U_{n,N} - u\| = o_p(1)$, as follows.

Let $\tilde{Y}_{n,N}$ be the projection of y on $\text{span}(\{X_{m,N}\}_{m=1}^n)$, with residual $\tilde{U}_{n,N} = y - \tilde{Y}_{n,N}$, and let \hat{y}_n be the projection of y on $\text{span}(\{x_m\}_{m=1}^n)$, with residual $u_n = y - \hat{y}_n$. Then by the triangular inequality,

$$\|U_{n,N} - u_n\| \leq \|U_{n,N} - \tilde{U}_{n,N}\| + \|u_n - \tilde{U}_{n,N}\|.$$

It will be shown that

$$\|U_{n,N} - \tilde{U}_{n,N}\| = o_p(1) \quad (3.51)$$

and

$$\|\hat{y}_n - \tilde{Y}_{n,N}\| = \|u_n - \tilde{U}_{n,N}\| = o_p(1). \quad (3.52)$$

Since $\lim_{n \rightarrow \infty} \|\hat{y}_n - \hat{y}\| = 0$ and thus $\lim_{n \rightarrow \infty} \|u_n - u\| = 0$, the result of the theorem under review then follows from (3.51) and (3.52).

Proof of (3.51)

Denote the angle between two elements x and y of \mathcal{H} by $\varphi(x, y)$. Recall that

$$\begin{aligned}\sin^2(\varphi(Y_N, \widehat{Y}_{n,N})) &= \|U_{n,N}\|^2/\|Y_N\|^2 \\ \sin^2(\varphi(y, \widetilde{Y}_{n,N})) &= \|\widetilde{U}_{n,N}\|^2/\|y\|^2 \\ \cos(\varphi(Y_N, \widehat{Y}_{n,N})) &= \|\widehat{Y}_{n,N}\|/\|Y_N\| \\ \cos(\varphi(y, \widetilde{Y}_{n,N})) &= \|\widetilde{Y}_{n,N}\|/\|y\|.\end{aligned}$$

Using these formulas we can write

$$\begin{aligned}\|U_{n,N} - \widetilde{U}_{n,N}\|^2 &= \|U_{n,N}\|^2 + \|\widetilde{U}_{n,N}\|^2 - 2\langle U_{n,N}, \widetilde{U}_{n,N} \rangle \\ &= \|Y_N\|^2 \sin^2(\varphi(Y_N, \widehat{Y}_{n,N})) + \|y\|^2 \sin^2(\varphi(y, \widetilde{Y}_{n,N})) \\ &\quad - 2\langle U_{n,N}, \widetilde{U}_{n,N} \rangle \\ &= \|Y_N\|^2 + \|y\|^2 \\ &\quad - \|Y_N\|^2 \cos^2(\varphi(Y_N, \widehat{Y}_{n,N})) - \|y\|^2 \cos^2(\varphi(y, \widetilde{Y}_{n,N})) \\ &\quad - 2\langle U_{n,N}, \widetilde{U}_{n,N} \rangle \\ &= \|Y_N - y\|^2 - \|Y_N\|^2 \cos^2(\varphi(Y_N, \widehat{Y}_{n,N})) \\ &\quad - \|y\|^2 \cos^2(\varphi(y, \widetilde{Y}_{n,N})) + 2\langle Y_N, y \rangle - 2\langle U_{n,N}, \widetilde{U}_{n,N} \rangle\end{aligned}$$

and

$$\begin{aligned}\langle U_{n,N}, \widetilde{U}_{n,N} \rangle &= \langle U_{n,N}, \widetilde{U}_{n,N} + \widetilde{Y}_{n,N} \rangle = \langle U_{n,N}, y \rangle \\ &= \langle U_{n,N} + \widehat{Y}_{n,N}, y \rangle - \langle \widehat{Y}_{n,N}, y \rangle \\ &= \langle Y_N, y \rangle - \langle \widehat{Y}_{n,N}, y \rangle \\ &= \langle Y_N, y \rangle - \cos(\varphi(y, \widehat{Y}_{n,N})) \|\widehat{Y}_{n,N}\| \cdot \|y\| \\ &= \langle Y_N, y \rangle - \cos(\varphi(y, \widehat{Y}_{n,N})) \cos(\varphi(Y_N, \widehat{Y}_{n,N})) \|y\| \cdot \|Y_N\| \\ &\geq \langle Y_N, y \rangle - \cos(\varphi(y, \widetilde{Y}_{n,N})) \cos(\varphi(Y_N, \widehat{Y}_{n,N})) \|y\| \cdot \|Y_N\|\end{aligned}$$

where the inequality follows from $\cos(\varphi(y, \hat{Y}_{n,N})) \leq \cos(\varphi(y, \tilde{Y}_{n,N}))$. Thus

$$\begin{aligned} \|U_{n,N} - \tilde{U}_{n,N}\|^2 &\leq \|Y_N - y\|^2 \\ &\quad - \left(\|Y_N\| \cos(\varphi(Y_N, \hat{Y}_{n,N})) - \|y\| \cos(\varphi(y, \tilde{Y}_{n,N})) \right)^2 \\ &\leq \|Y_N - y\|^2 = o_p(1) \end{aligned}$$

where the $o_p(1)$ term is due to condition (3.14). This proves (3.51).

Proof of (3.52)

Let $r_1 = x_1$ and for $m \geq 2$, let r_m be the residual of the projection of x_m on $\text{span}(x_1, \dots, x_{m-1})$. Denote $e_m = \|r_m\|^{-1}r_m$ if $\|r_m\| > 0$ and $e_m = 0$ if $\|r_m\| = 0$. Similarly, let $R_{1,N} = X_{1,N}$ and for $m = 2, \dots, n$, let $R_{m,N}$ be the residual of the projection of $X_{m,N}$ on $\text{span}(X_{1,N}, \dots, X_{m-1,N})$. Denote $\hat{e}_{m,N} = \|R_{m,N}\|^{-1}R_{m,N}$ if $\|R_{m,N}\| > 0$, and $\hat{e}_{m,N} = 0$ if $\|R_{m,N}\| = 0$. Then we can write

$$\hat{y}_n = \sum_{m=1}^n \alpha_m e_m, \text{ where } \alpha_m = \langle y, e_m \rangle \text{ and } \sum_{m=1}^{\infty} \alpha_m^2 < \infty \quad (3.53)$$

$$\tilde{Y}_{n,N} = \sum_{m=1}^n \hat{\alpha}_{m,N} \hat{e}_{m,N}, \text{ where } \hat{\alpha}_{m,N} = \langle y, \hat{e}_{m,N} \rangle. \quad (3.54)$$

It follows from the trivial equalities $\|\hat{y}_n - \tilde{Y}_{n,N}\|^2 = \|\tilde{Y}_{n,N}\|^2 + \|\hat{y}_n\|^2 - 2 \langle \hat{y}_n, \tilde{Y}_{n,N} \rangle$ and $\langle \hat{y}_n, \tilde{Y}_{n,N} \rangle = \langle \hat{y}_n, y - \tilde{U}_{n,N} \rangle = \|\hat{y}_n\|^2 - \langle \hat{y}_n, \tilde{U}_{n,N} \rangle$ that

$$\|\hat{y}_n - \tilde{Y}_{n,N}\|^2 = \|\tilde{Y}_{n,N}\|^2 - \|\hat{y}_n\|^2 + 2 \langle \hat{y}_n, \tilde{U}_{n,N} \rangle.$$

Moreover, using the Cauchy-Schwarz inequality and the fact that $\|\tilde{U}_{n,N}\| \leq \|y\|$, it follows that

$$\begin{aligned} \left| \langle \hat{y}_n, \tilde{U}_{n,N} \rangle \right| &= \left| \left\langle \sum_{m=1}^n \alpha_m e_m, \tilde{U}_{n,N} \right\rangle \right| = \left| \left\langle \sum_{m=1}^n \alpha_m (e_m - \hat{e}_{m,N}), \tilde{U}_{n,N} \right\rangle \right| \\ &\leq \|\tilde{U}_{n,N}\| \cdot \left\| \sum_{m=1}^n \alpha_m (e_m - \hat{e}_{m,N}) \right\| \\ &\leq \|y\| \cdot \left\| \sum_{m=1}^n \alpha_m (e_m - \hat{e}_{m,N}) \right\| \end{aligned}$$

$$\leq \|y\| \cdot \left\| \sum_{m=1}^k \alpha_m (e_m - \hat{e}_{m,N}) \right\| + 2\|y\| \sqrt{\sum_{m=k+1}^{\infty} \alpha_m^2}$$

Given an arbitrary $\varepsilon > 0$ we can choose k so large that $2\|y\| \sqrt{\sum_{m=k+1}^{\infty} \alpha_m^2} < \varepsilon$, and for this k , $\left\| \sum_{m=1}^k \alpha_m (e_m - \hat{e}_{m,N}) \right\| = o_p(1)$, as is easy to verify from condition (3.15). Consequently,

$$\langle \hat{y}_n, \tilde{U}_{n,N} \rangle = o_p(1)$$

and thus

$$\|\hat{y}_n - \tilde{Y}_{n,N}\|^2 = \|\tilde{Y}_{n,N}\|^2 - \|\hat{y}_n\|^2 + o_p(1). \quad (3.55)$$

The next step is to show that

$$\|\tilde{Y}_{n,N}\| \leq \|\hat{y}_n\| + o_p(1), \quad (3.56)$$

as follows. Note that

$$\begin{aligned} \|\tilde{U}_{n,N}\| &= \inf_{\beta_1, \dots, \beta_n} \left\| y - \sum_{m=1}^n \beta_m X_{m,N} \right\|^2 \\ &= \inf_{(\xi_1, \dots, \xi_n)' \in \mathbb{X}_{m=1}^n [-\rho_m, \rho_m]} \inf_{\lambda} \left\| y - \lambda \sum_{m=1}^n \xi_m X_{m,N} \right\|^2 \\ &= \inf_{(\xi_1, \dots, \xi_n)' \in \mathbb{X}_{m=1}^n [-\rho_m, \rho_m]} \left\{ \|y\|^2 - \left(\frac{\langle y, \sum_{m=1}^n \xi_m X_{m,N} \rangle}{\|\sum_{m=1}^n \xi_m X_{m,N}\|} \right)^2 \right\} \\ &= \|y\|^2 - \sup_{(\xi_1, \dots, \xi_n)' \in \mathbb{X}_{m=1}^n [-\rho_m, \rho_m]} \frac{\langle y, \sum_{m=1}^n \xi_m X_{m,N} \rangle^2}{\|\sum_{m=1}^n \xi_m X_{m,N}\|^2} \end{aligned}$$

hence

$$\|\tilde{Y}_{n,N}\| = \sup_{(\xi_1, \dots, \xi_n)' \in \mathbb{X}_{m=1}^n [-\rho_m, \rho_m]} \frac{\langle y, \sum_{m=1}^n \xi_m X_{m,N} \rangle}{\|\sum_{m=1}^n \xi_m X_{m,N}\|} \quad (3.57)$$

and similarly,

$$\|\hat{y}_n\| = \sup_{(\xi_1, \dots, \xi_n)' \in \mathbb{X}_{m=1}^n [-\rho_m, \rho_m]} \frac{\langle y, \sum_{m=1}^n \xi_m x_m \rangle}{\|\sum_{m=1}^n \xi_m x_m\|} \quad (3.58)$$

Note that by condition (3.16) at least one x_m is non-zero, so that (3.57) and (3.58) are well-defined for sufficiently large n .

Since the ratios in (3.57) and (3.58) are scale-invariant, we may without loss of generality impose the normalization

$$\left\| \sum_{m=1}^n \xi_m x_m \right\| = M_n = \frac{1}{2} \left\| \sum_{m=1}^n \rho_m x_m \right\|, \quad (3.59)$$

for example. Note that (3.59) is compatible with $(\xi_1, \dots, \xi_n)' \in \mathbb{X}_{m=1}^n [-\rho_m, \rho_m]$. Thus, denoting

$$\Xi_n = \left\{ (\xi_1, \dots, \xi_n)' \in \mathbb{X}_{m=1}^n [-\rho_m, \rho_m] : \left\| \sum_{m=1}^n \xi_m x_m \right\| = M_n \right\},$$

the expressions (3.57) and (3.58) are equivalent to

$$\|\tilde{Y}_{n,N}\| = \sup_{(\xi_1, \dots, \xi_n)' \in \Xi_n} \frac{\langle y, \sum_{m=1}^n \xi_m X_{m,N} \rangle}{\left\| \sum_{m=1}^n \xi_m X_{m,N} \right\|} \quad (3.60)$$

and

$$\|\hat{y}_n\| = \sup_{(\xi_1, \dots, \xi_n)' \in \Xi_n} \frac{\langle y, \sum_{m=1}^n \xi_m x_m \rangle}{\left\| \sum_{m=1}^n \xi_m x_m \right\|}, \quad (3.61)$$

respectively.

Finally, observe from (3.59), (3.60) and (3.61) that

$$\begin{aligned} \|\hat{y}_n\| &= \sup_{(\xi_1, \dots, \xi_n)' \in \Xi_n} \left\{ \frac{\langle y, \sum_{m=1}^n \xi_m X_{m,N} \rangle}{\left\| \sum_{m=1}^n \xi_m X_{m,N} \right\|} \times \frac{\left\| \sum_{m=1}^n \xi_m X_{m,N} \right\|}{\left\| \sum_{m=1}^n \xi_m x_m \right\|} \right. \\ &\quad \left. - \frac{\langle y, \sum_{m=1}^n \xi_m (X_{m,N} - x_m) \rangle}{\left\| \sum_{m=1}^n \xi_m x_m \right\|} \right\} \\ &= \sup_{(\xi_1, \dots, \xi_n)' \in \Xi_n} \left\{ \frac{\langle y, \sum_{m=1}^n \xi_m X_{m,N} \rangle}{\left\| \sum_{m=1}^n \xi_m X_{m,N} \right\|} \times \frac{\left\| \sum_{m=1}^n \xi_m X_{m,N} \right\|}{M_n} \right. \\ &\quad \left. - \frac{\langle y, \sum_{m=1}^n \xi_m (X_{m,N} - x_m) \rangle}{M_n} \right\} \\ &\geq \left(1 - \frac{\sum_{m=1}^n \rho_m \|X_{m,N} - x_m\|}{M_n} \right) \sup_{(\xi_1, \dots, \xi_n)' \in \Xi_n} \frac{\langle y, \sum_{m=1}^n \xi_m X_{m,N} \rangle}{\left\| \sum_{m=1}^n \xi_m X_{m,N} \right\|} \\ &\quad - \|y\| \frac{\sum_{m=1}^n \rho_m \|X_{m,N} - x_m\|}{M_n} \end{aligned}$$

$$\begin{aligned}
&= \left(1 - \frac{\sum_{m=1}^n \rho_m \|X_{m,N} - x_m\|}{M_n}\right) \|\tilde{Y}_{n,N}\| - \|y\| \frac{\sum_{m=1}^n \rho_m \|X_{m,N} - x_m\|}{M_n} \\
&\geq \|\tilde{Y}_{n,N}\| - 2\|y\| \frac{\sum_{m=1}^n \rho_m \|X_{m,N} - x_m\|}{M_n},
\end{aligned} \tag{3.62}$$

where the last inequality follows from $\|\tilde{Y}_{n,N}\| \leq \|y\|$. Since by condition (3.16), $\liminf_{n \rightarrow \infty} M_n > 0$, it follows from (3.15) and (3.62) that (3.56) holds. The latter together with (3.55) imply (3.52).

Chapter 4

Orthogonal polynomials

4.1 Introduction

Let $w(x)$ be a non-negative Borel measurable real-valued function on \mathbb{R} satisfying

$$\int_{-\infty}^{\infty} |x|^k w(x) dx \in (0, \infty) \text{ for } k \in \mathbb{N}_0$$

where the integral involved is the Lebesgue integral. Without loss of generality we may assume that w is a density function with finite absolute moments of any order. Let

$$p_k(x|w) = \sum_{j=0}^k \alpha_{k,j} x^j, \quad \alpha_{k,k} = 1, \quad k \in \mathbb{N}_0 \quad (4.1)$$

be a sequence of polynomials in $x \in \mathbb{R}$ such that

$$\int_{-\infty}^{\infty} p_k(x|w) p_m(x|w) w(x) dx = 0 \text{ if } k \neq m. \quad (4.2)$$

In words, the polynomials $p_k(x|w)$ are *orthogonal* with respect to the weight function $w(x)$.

Defining

$$\bar{p}_k(x|w) = \frac{p_k(x|w)}{\sqrt{\int_{-\infty}^{\infty} p_k(y|w)^2 w(y) dy}} \quad (4.3)$$

yields a sequence of *orthonormal* polynomials w.r.t. $w(x)$:

$$\int_{-\infty}^{\infty} \bar{p}_k(x|w) \bar{p}_m(x|w) w(x) dx = \begin{cases} 0 & \text{if } k \neq m, \\ 1 & \text{if } k = m. \end{cases} \quad (4.4)$$

This sequence is uniquely determined by $w(x)$, except for signs. In other words, $|\bar{p}_k(x|w)|$ is unique. To show this, suppose that there exists another sequence $\bar{p}_k^*(x|w)$ of orthonormal polynomials w.r.t. $w(x)$. Since $\bar{p}_k^*(x|w)$ is a polynomial of order k , we can write $\bar{p}_k^*(x|w) = \sum_{m=0}^k \beta_{m,k} \bar{p}_m(x|w)$. Similarly, we can write $\bar{p}_k(x|w) = \sum_{m=0}^k \alpha_{m,k} \bar{p}_m^*(x|w)$. Then for $j < k$,

$$\begin{aligned} \int_{-\infty}^{\infty} \bar{p}_k^*(x|w) \bar{p}_j(x|w) w(x) dx &= \sum_{m=0}^j \alpha_{m,j} \int_{-\infty}^{\infty} \bar{p}_k^*(x|w) \bar{p}_m(x|w) w(x) dx \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \int_{-\infty}^{\infty} \bar{p}_k^*(x|w) \bar{p}_j(x|w) w(x) dx &= \sum_{m=0}^k \beta_{m,k} \int_{-\infty}^{\infty} \bar{p}_m(x|w) \bar{p}_j(x|w) w(x) dx \\ &= \beta_{j,k} \int_{-\infty}^{\infty} \bar{p}_j(x|w)^2 w(x) dx = \beta_{j,k}, \end{aligned}$$

hence $\beta_{j,k} = 0$ for $j < k$ and thus

$$\bar{p}_k^*(x|w) = \beta_{k,k} \bar{p}_k(x|w).$$

Moreover, by normality,

$$1 = \int_{-\infty}^{\infty} \bar{p}_k^*(x|w)^2 w(x) dx = \beta_{k,k}^2 \int_{-\infty}^{\infty} \bar{p}_k(x|w)^2 w(x) dx = \beta_{k,k}^2,$$

so that $\bar{p}_k^*(x|w) = \pm \bar{p}_k(x|w)$. Consequently, $|\bar{p}_k(x|w)|$ is unique.

The reason for considering orthonormal polynomials is the following.

Theorem 4.1. *Let $w(x)$ be a density function with support (a, b) , $-\infty \leq a < b \leq \infty$, satisfying the moment conditions*

$$\int_{-\infty}^{\infty} |x|^k w(x) dx < \infty \quad (4.5)$$

for $k \in \mathbb{N}_0$. Denote by $L^2(w)$ be the Hilbert space of Borel measurable real functions f on (a, b) satisfying $\int_a^b f(x)^2 w(x) dx < \infty$, with inner product $\langle f, g \rangle = \int_a^b f(x)g(x)w(x)dx$ and associated norm $\|f\| = \sqrt{\langle f, f \rangle}$ and metric $\|f - g\|$. For an arbitrary function $f \in L^2(w)$, let

$$f_n(x) = \sum_{k=0}^n \gamma_k \bar{p}_k(x|w),$$

where

$$\gamma_k = \langle f, \bar{p}_k \rangle = \int_a^b f(x) \bar{p}_k(x|w) w(x) dx. \quad (4.6)$$

Then

$$\lim_{n \rightarrow \infty} \|f - f_n\| = 0. \quad (4.7)$$

This result implies that every function $f \in L^2(w)$ can be written as

$$f(x) = \sum_{k=0}^{\infty} \gamma_k \bar{p}_k(x|w) \text{ a.e. on } (a, b). \quad (4.8)$$

Note that condition (4.5) holds trivially if the support (a, b) of $w(x)$ is bounded. However, as is well-known, condition (4.5) also holds for the standard normal density, the exponential density and more generally the density of the Gamma distribution, for example.

Since for every density $w(x)$ with support (a, b) , $\int_a^b f(x)^2 dx < \infty$ implies that $f(x)/\sqrt{w(x)} \in L^2(w)$, the following corollary of Theorem 4.1 holds trivially.

Corollary 4.1. Let $L^2(a, b)$, $-\infty \leq a < b \leq \infty$, be the Hilbert space of square integrable Borel measurable real functions on (a, b) , with inner product $\langle f, g \rangle = \int_a^b f(x)g(x)dx$ and associated norm and metric. Every function $f \in L^2(a, b)$ can be written as

$$f(x) = \sqrt{w(x)} \left(\sum_{k=0}^{\infty} \gamma_k \bar{p}_k(x|w) \right) \text{ a.e. on } (a, b),$$

where w is a density with support (a, b) satisfying the moment conditions (4.5), the $\bar{p}_k(x|w)$'s are the orthonormal polynomials generated by $w(x)$ and

the γ_k 's are the Fourier coefficients of $f(x)/\sqrt{w(x)}$, i.e.,

$$\gamma_k = \int_a^b f(x) \bar{p}_k(x|w) \sqrt{w(x)} dx.$$

This result implies that the functions

$$\psi_k(x|w) = \bar{p}_k(x|w) \sqrt{w(x)}, \quad k \in \mathbb{N},$$

form a complete orthonormal sequence in $L^2(a, b)$:

$$L^2(a, b) = \text{span} \left(\left\{ \bar{p}_k(x|w) \sqrt{w(x)} \right\}_{k=0}^{\infty} \right).$$

Of course, the $\psi_k(x|w)$'s are no longer polynomials.

If $\max(|a|, |b|) < \infty$ then there is another way to construct a complete orthonormal sequence in $L^2(a, b)$, as follows. Let $W(x)$ be the distribution function of a density w with bounded support (a, b) . Then

$$G(x) = a + (b - a) W(x)$$

is a one-to-one mapping of (a, b) onto (a, b) , with inverse

$$G^{-1}(y) = W^{-1}((y - a) / (b - a))$$

where W^{-1} is the inverse of $W(x)$. For every $f \in L^2(a, b)$,

$$(b - a) \int_a^b f(G(x))^2 w(x) dx = \int_a^b f(G(x))^2 dG(x) = \int_a^b f(x)^2 dx < \infty.$$

Hence $f(G(x)) \in L^2(w)$ and thus by Theorem 4.1,

$$f(G(x)) = \sum_{k=0}^{\infty} \gamma_k \bar{p}_k(x|w) \text{ a.e. on } (a, b)$$

where

$$\begin{aligned} \gamma_k &= \int_a^b f(G(x)) \bar{p}_k(x|w) w(x) dx = \frac{1}{b - a} \int_a^b f(G(x)) \bar{p}_k(x|w) dG(x) \\ &= \frac{1}{b - a} \int_a^b f(x) \bar{p}_k(G^{-1}(x)|w) dx \end{aligned}$$

Consequently

$$f(x) = f(G(G^{-1}(x))) = \sum_{k=0}^{\infty} \gamma_k \bar{p}_k(G^{-1}(x)|w) \text{ a.e. on } (a, b)$$

Note that $dG^{-1}(x)/dx = dG^{-1}(x)/dG(G^{-1}(x)) = 1/G'(G^{-1}(x))$, so that

$$\begin{aligned} & \int_a^b \bar{p}_k(G^{-1}(x)|w) \bar{p}_m(G^{-1}(x)|w) dx \\ &= \int_a^b \bar{p}_k(G^{-1}(x)|w) \bar{p}_m(G^{-1}(x)|w) G'(G^{-1}(x)) dG^{-1}(x) \\ &= \int_a^b \bar{p}_k(x|w) \bar{p}_m(x|w) G'(x) dx \\ &= (b-a) \int_a^b \bar{p}_k(x|w) \bar{p}_m(x|w) w(x) dx = (b-a) I(k=m) \end{aligned}$$

Thus,

Corollary 4.2. *Let w be a density with bounded support (a, b) , satisfying the moment conditions (4.5). Let W be the c.d.f. of w , with inverse W^{-1} . Then the functions*

$$\psi_k(x|w) = \bar{p}_k(W^{-1}((x-a)/(b-a))|w) / \sqrt{(b-a)}, \quad k \in \mathbb{N}_0,$$

form a complete orthonormal sequence in $L^2(a, b)$, i.e., every $f \in L^2(a, b)$ can be written as $f(x) = \sum_{k=0}^{\infty} \alpha_k \psi_k(x|w)$ a.e. on (a, b) , where $\alpha_k = \int_a^b f(x) \psi_k(x|w) dx$.

4.2 The three-term recurrence relation

It follows from (4.1) that $p_0(x|w) \equiv 1$, and it follows from (4.2) that $p_1(x|w) = \alpha_{1,0} + x$ can be constructed by solving $\int_{-\infty}^{\infty} (\alpha_{1,0} + x) w(x) dx = 0$. Hence, given that $w(x)$ is a density, $\alpha_{1,0} = -\int_{-\infty}^{\infty} x w(x) dx$. The question now arises how to construct these orthogonal polynomials further for $k \geq 2$.

The answer is the following.

Theorem 4.2. Every sequence of polynomials $p_k(x|w) = \sum_{j=0}^k \alpha_{k,j} x^j$, with $\alpha_{k,k} = 1$, satisfying the orthogonality condition (4.2), with $w(x)$ satisfying the moment conditions (4.5), can be generated recursively by the three-term recurrence relation (hereafter referred to as TTRR)

$$p_{k+1}(x|w) + (b_k - x)p_k(x|w) + c_k p_{k-1}(x|w) = 0, \quad k \in \mathbb{N}, \quad (4.9)$$

where

$$b_k = \frac{\int_{-\infty}^{\infty} x.p_k(x|w)^2 w(x) dx}{\int_{-\infty}^{\infty} p_k(x|w)^2 w(x) dx} \quad (4.10)$$

and

$$c_k = \frac{\int_{-\infty}^{\infty} p_k(x|w)^2 w(x) dx}{\int_{-\infty}^{\infty} p_{k-1}(x|w)^2 w(x) dx} \quad (4.11)$$

Next, let $d_k = \sqrt{\int_{-\infty}^{\infty} p_k(x|w)^2 w(x) dx}$, so that $\bar{p}_k(x|w) = p_k(x|w)/d_k$ is a sequence of orthonormal polynomials. Substituting $p_k(x|w) = d_k \cdot \bar{p}_k(x|w)$ in (4.9), (4.10) and (4.11) yields

$$\frac{d_{k+1}}{d_k} \bar{p}_{k+1}(x|w) + (b_k - x) \bar{p}_k(x|w) + c_k \frac{d_{k-1}}{d_k} \bar{p}_{k-1}(x|w) = 0, \quad k \geq 1,$$

where $b_k = \int_{-\infty}^{\infty} x \cdot \bar{p}_k(x|w)^2 w(x) dx$ and $c_k = d_k^2/d_{k-1}^2$, hence

$$\frac{d_{k+1}}{d_k} \bar{p}_{k+1}(x|w) + (b_k - x) \bar{p}_k(x|w) + \frac{d_k}{d_{k-1}} \bar{p}_{k-1}(x|w) = 0, \quad k \geq 1.$$

Moreover, note that

$$\lim_{|x| \rightarrow \infty} \frac{x \bar{p}_{k-1}(x|w)}{\bar{p}_k(x|w)} = \frac{d_k}{d_{k-1}} \lim_{|x| \rightarrow \infty} \frac{x.p_{k-1}(x|w)}{p_k(x|w)} = \frac{d_k}{d_{k-1}},$$

where the latter equality is due to the normalization $\alpha_{k,k} = 1$ in Theorem 4.2. Thus:

Theorem 4.3. Every sequence $\bar{p}_k(x|w)$ of orthonormal polynomials with respect to a density function $w(x)$ satisfying the moment conditions (4.5) can be generated recursively by the TTRR

$$a_{k+1} \cdot \bar{p}_{k+1}(x|w) + (b_k - x) \bar{p}_k(x|w) + a_k \cdot \bar{p}_{k-1}(x|w) = 0, \quad k \in \mathbb{N}. \quad (4.12)$$

where

$$a_k = \left| \lim_{|x| \rightarrow \infty} \frac{x \cdot \bar{p}_{k-1}(x|w)}{\bar{p}_k(x|w)} \right| \quad (4.13)$$

and

$$b_k = \int_{-\infty}^{\infty} x \cdot \bar{p}_k(x|w)^2 w(x) dx. \quad (4.14)$$

4.3 Examples of orthonormal polynomials

4.3.1 Hermite polynomials

If $w(x)$ is the density of the standard normal distribution,

$$w_{\mathcal{N}[0,1]}(x) = \exp(-x^2/2) / \sqrt{2\pi},$$

the orthonormal polynomials involved satisfy the TTRR

$$\sqrt{k+1} \bar{p}_{k+1}(x|w_{\mathcal{N}[0,1]}) - x \cdot \bar{p}_k(x|w_{\mathcal{N}[0,1]}) + \sqrt{k} \bar{p}_{k-1}(x|w_{\mathcal{N}[0,1]}) = 0, \quad k \in \mathbb{N},$$

starting from $\bar{p}_0(x|w_{\mathcal{N}[0,1]}) = 1$, $\bar{p}_1(x|w_{\mathcal{N}[0,1]}) = x$. Thus in this case $a_k = \sqrt{k}$ and $b_k = 0$ in (4.12). These polynomials are known as Hermite¹ polynomials.

It follows from Theorem 4.1 that the Hermite polynomials span the Hilbert space $L^2(w_{\mathcal{N}[0,1]})$, and it follows from Corollary 4.1 that

$$L^2(\mathbb{R}) = \text{span} \left(\left\{ \sqrt{w_{\mathcal{N}[0,1]}(x)} \bar{p}_k(x|w_{\mathcal{N}[0,1]}) \right\}_{k=0}^{\infty} \right).$$

Consequently, any density $f(x)$ on \mathbb{R} can be represented by

$$f(x) = w_{\mathcal{N}[0,1]}(x) \left(\sum_{k=0}^{\infty} \gamma_k \bar{p}_k(x|w_{\mathcal{N}[0,1]}) \right)^2$$

where $\sum_{k=0}^{\infty} \gamma_k^2 = 1$.

¹Charles Hermite (1822-1901).

4.3.2 Laguerre polynomials

The standard exponential density function

$$w_{\text{Exp}}(x) = I(x \geq 0) \exp(-x)$$

gives rise to the orthonormal Laguerre² polynomials, with TTRR

$$(k+1)\bar{p}_{k+1}(x|w_{\text{Exp}}) + (2k+1-x)\bar{p}_k(x|w_{\text{Exp}}) + k\bar{p}_{k-1}(x|w_{\text{Exp}}) = 0,$$

for $k \in \mathbb{N}$, starting from $\bar{p}_0(x|w_{\text{Exp}}) = 1$, $\bar{p}_1(x|w_{\text{Exp}}) = x - 1$. Thus in this case $a_k = k$ and $b_k = 2k + 1$.

Since the moment conditions (4.5) hold for $w_{\text{Exp}}(x)$, it follows from Theorem 4.1 that any Borel measurable real function $f(x)$ satisfying $\int_0^\infty \exp(-x) f(x)^2 dx < \infty$ can be written as $f(x) = \sum_{k=0}^{\infty} \gamma_k \bar{p}_k(x|w_{\text{Exp}})$ a.e. on $[0, \infty)$, where $\gamma_k = \int_0^\infty \exp(-x) \bar{p}_{k+1}(x|w) f(x) dx$.

Again, it follows from Corollary 4.1 that

$$L^2(0, \infty) = \text{span}(\{\exp(-x/2) \bar{p}_k(x|w_{\text{Exp}})\}_{k=0}^{\infty}),$$

hence any density $f(x)$ on $[0, \infty)$ can be written as

$$f(x) = \exp(-x) \left(\sum_{k=0}^{\infty} \gamma_k \bar{p}_k(x|w_{\text{Exp}}) \right)^2, \quad \text{with } \sum_{k=0}^{\infty} \gamma_k^2 = 1.$$

4.3.3 Legendre polynomials

The uniform density on $[-1, 1]$,

$$w_{\mathcal{U}[-1,1]}(x) = \frac{1}{2}I(|x| \leq 1),$$

generates the orthonormal Legendre³ polynomials on $[-1, 1]$, with TTRR

$$\begin{aligned} & \frac{k+1}{\sqrt{2k+3}\sqrt{2k+1}}\bar{p}_{k+1}(x|w_{\mathcal{U}[-1,1]}) - x\bar{p}_k(x|w_{\mathcal{U}[-1,1]}) \\ & + \frac{k}{\sqrt{2k+1}\sqrt{2k-1}}\bar{p}_{k-1}(x|w_{\mathcal{U}[-1,1]}) = 0, \end{aligned} \quad (4.15)$$

²Edmund Nicolas Laguerre (1834-1886)

³Adrien-Marie Legendre (1752-1833)

for $k \in \mathbb{N}$, starting from $\bar{p}_0(x|w_{\mathcal{U}[-1,1]}) = 1$, $\bar{p}_1(x|w_{\mathcal{U}[-1,1]}) = \sqrt{3}x$.

Note that the orthonormal Legendre polynomials $\bar{p}_k(x|w_{\mathcal{U}[-1,1]})$ satisfy

$$\begin{aligned} & \int_0^1 \bar{p}_k(2u - 1|w_{\mathcal{U}[-1,1]}) \bar{p}_m(2u - 1|w_{\mathcal{U}[-1,1]}) du \\ &= \frac{1}{2} \int_0^1 \bar{p}_k(2u - 1|w_{\mathcal{U}[-1,1]}) \bar{p}_m(2u - 1|w_{\mathcal{U}[-1,1]}) d(2u - 1) \\ &= \frac{1}{2} \int_{-1}^1 \bar{p}_k(x|w_{\mathcal{U}[-1,1]}) \bar{p}_m(x|w_{\mathcal{U}[-1,1]}) dx = I(k = m) \end{aligned}$$

Hence,

$$\bar{p}_k(u|w_{\mathcal{U}[0,1]}) = \bar{p}_k(2u - 1|w_{\mathcal{U}[-1,1]}), \quad k \in \mathbb{N}_0,$$

is a sequence of orthonormal polynomials w.r.t. the uniform density on $[0, 1]$,

$$w_{\mathcal{U}[0,1]}(u) = I(0 \leq u \leq 1)$$

The $\bar{p}_k(u|w_{\mathcal{U}[0,1]})$'s are known as the shifted Legendre polynomials, also called the orthonormal Legendre polynomials on the unit interval $[0, 1]$. Substituting $x = 2u - 1$ and $\bar{p}_k(x|w_{\mathcal{U}[-1,1]}) = \bar{p}_k(u|w_{\mathcal{U}[0,1]})$ in (4.15) yields the TTRR

$$\begin{aligned} & \frac{(k+1)/2}{\sqrt{2k+3}\sqrt{2k+1}} \rho_{k+1}(u|w_{\mathcal{U}[0,1]}) + (0.5 - u) \cdot \rho_k(u|w_{\mathcal{U}[0,1]}) \\ &+ \frac{k/2}{\sqrt{2k+1}\sqrt{2k-1}} \rho_{k-1}(u|w_{\mathcal{U}[0,1]}) = 0, \quad k \in \mathbb{N}, \end{aligned}$$

starting from $\rho_0(u) = 1$, $\rho_1(u) = \sqrt{3}(2u - 1)$.

Again, it follows from Theorem 4.1 that any Borel measurable real function $f(x)$ on $[0, 1]$ can be written as $f(x) = \sum_{k=0}^{\infty} \gamma_k \rho_k(u|w_{\mathcal{U}[0,1]})$, where $\gamma_k = \int_0^1 f(x) \rho_k(u|w_{\mathcal{U}[0,1]}) dx$, hence $L^2(0, 1) = \text{span}(\{\rho_k(u|w_{\mathcal{U}[0,1]})\}_{k=0}^{\infty})$.

These shifted Legendre polynomials have been used by Bierens (2008) and Bierens and Carvalho (2007) to model semi-nonparametrically the unobserved heterogeneity of interval-censored mixed proportional hazard models and bivariate mixed proportional hazard models, respectively.

4.3.4 Chebyshev polynomials

Chebyshev polynomials on $[-1, 1]$

Chebyshev polynomials on $[-1, 1]$ are generated by the weight function

$$w_{C[-1,1]}(x) = \frac{1}{\pi\sqrt{1-x^2}} I(|x| < 1). \quad (4.16)$$

This is a density function on $(-1, 1)$. To see this, let $\theta = \arccos(x)$, so that $x = \cos(\theta)$, and observe that

$$\frac{dx}{d\theta} = -\sin(\theta) = -\sqrt{1 - \cos^2(\theta)} = -\sqrt{1 - x^2},$$

hence

$$\frac{d \arccos(x)}{dx} = \frac{-1}{\sqrt{1-x^2}} \quad (4.17)$$

Then

$$\begin{aligned} \int_{-1}^1 \frac{1}{\pi\sqrt{1-x^2}} dx &= -\frac{1}{\pi} \int_{-1}^1 d \arccos(x) \\ &= \frac{\arccos(-1) - \arccos(1)}{\pi} = 1 \end{aligned}$$

because $\arccos(-1) = \pi$ and $\arccos(1) = 0$. Clearly, the corresponding distribution function is

$$W_{C[-1,1]}(x) = \frac{\arccos(-1) - \arccos(x)}{\pi}, \quad x \in [-1, 1].$$

The orthogonal (but not orthonormal) Chebyshev polynomials $p_k(x|w_{C[-1,1]})$ satisfy the TTRR

$$p_{k+1}(x|w_{C[-1,1]}) - 2xp_k(x|w_{C[-1,1]}) + p_{k-1}(x|w_{C[-1,1]}) = 0, \quad k \in \mathbb{N}, \quad (4.18)$$

starting from $p_0(x|w_{C[-1,1]}) = 1$, $p_1(x|w_{C[-1,1]}) = x$, with orthogonality properties

$$\int_{-1}^1 \frac{p_k(x|w_{C[-1,1]})p_m(x|w_{C[-1,1]})}{\pi\sqrt{1-x^2}} dx = \begin{cases} 0 & \text{if } k \neq m, \\ 1/2 & \text{if } k = m > 0, \\ 1 & \text{if } k = m = 0. \end{cases}$$

An important practical difference with the other polynomials discussed so far is that Chebyshev polynomials have the closed form⁴:

$$p_k(x|w_{C[-1,1]}) = \cos(k \cdot \arccos(x)). \quad (4.19)$$

To see this, observe from (4.17) and the well-known sine-cosine formulas that

$$\begin{aligned} & \int_{-1}^1 \frac{\cos(k \cdot \arccos(x)) \cos(m \cdot \arccos(x))}{\pi \sqrt{1-x^2}} dx \\ &= -\frac{1}{\pi} \int_{-1}^1 \cos(k \cdot \arccos(x)) \cos(m \cdot \arccos(x)) d \arccos(x) \\ &= \frac{1}{\pi} \int_0^\pi \cos(k \cdot \theta) \cos(m \cdot \theta) d\theta \\ &= \frac{1}{2\pi} \int_0^\pi \cos((k+m)\theta) d\theta + \frac{1}{2\pi} \int_0^\pi \cos((k-m)\theta) d\theta \\ &= \frac{1}{2} \left(\frac{\sin((k+m)\pi)}{(k+m)\pi} + \frac{\sin((k-m)\pi)}{(k-m)\pi} \right) \\ &= \begin{cases} 0 & \text{if } k \neq m, \\ 1/2 & \text{if } k = m > 0, \\ 1 & \text{if } k = m = 0. \end{cases} \end{aligned} \quad (4.20)$$

Moreover, the TTRR (4.18) follows from

$$\begin{aligned} & \cos((k+1)\theta) - 2\cos(\theta)\cos(k\theta) + \cos((k-1)\theta) \\ &= \cos(k\theta)\cos(\theta) - \sin(k\theta)\sin(\theta) - 2\cos(\theta)\cos(k\theta) \\ &\quad + \cos(k\theta)\cos(\theta) + \sin(k\theta)\sin(\theta) = 0. \end{aligned}$$

Hence, the functions (4.19) satisfy the TTRR (4.18) and are therefore genuine polynomials.

In view of (4.20) we can now define the orthonormal Chebyshev polynomials as

$$\bar{p}_k(x|w_{C[-1,1]}) = \begin{cases} 1 & \text{for } k = 0, \\ \sqrt{2} \cos(k \cdot \arccos(x)) & \text{for } k \in \mathbb{N}. \end{cases}$$

⁴Note that $\arccos(x) = \text{atan}(-x/\sqrt{1-x^2}) + \frac{1}{2}\pi$, where $\text{atan}(x)$ is the inverse of the tangents function $\tan(\theta) = \sin(\theta)/\cos(\theta)$, $\theta \in (-\pi/2, \pi/2)$. In most programming languages the function $\text{atan}(x)$ is an intrinsic function. For example, in Visual Basic this function is the `ATN(x)` function.

It is trivial to verify that the density (4.16) satisfies the moment condition (4.5), so that the Chebyshev polynomials form a complete orthonormal sequence in the Hilbert space $L^2(w_{C[-1,1]})$ involved.

Further properties of Chebyshev polynomials

Because $p_n(x|w_{C[-1,1]})$ is a polynomial of order n in $x \in [-1, 1]$, it has at most n real roots in $[-1, 1]$. Obviously, these roots are

$$x_{n,k} = \cos(\pi(k - 0.5)/n), \quad k = 1, 2, \dots, n$$

Moreover,

Lemma 4.1. *For $j_1, j_2 = 0, 1, 2, \dots, n - 1$,*

$$\begin{aligned} & \sum_{k=1}^n \cos(\pi j_1(k - 0.5)/n) \cos(\pi j_2(k - 0.5)/n) \\ &= \sum_{k=1}^n p_{j_1}(x_{n,k}|w_{C[-1,1]}) p_{j_2}(x_{n,k}|w_{C[-1,1]}) = \begin{cases} 0 & \text{if } j_1 \neq j_2, \\ n/2 & \text{if } j_1 = j_2 > 0, \\ n & \text{if } j_1 = j_2 = 0. \end{cases} \end{aligned}$$

Now interpret k in Lemma 4.1 as a time index: $k = t = 1, \dots, n$, and denote

$$\begin{aligned} P_{0,n}(t) &\equiv 1, \quad P_{j,n}(t) = \sqrt{2} \cos(j\pi(t - 0.5)/n), \\ j &= 1, 2, \dots, n - 1, \quad t = 1, 2, \dots, n. \end{aligned}$$

The $P_{j,n}(t)$'s are known as Chebyshev time polynomials, which by Lemma 4.1 satisfy

$$\frac{1}{n} \sum_{t=1}^n P_{i,n}(t) P_{j,n}(t) = I(i = j), \quad i, j = 0, 1, 2, \dots, n - 1.$$

Consequently, any function $g(t)$ of time $t = 1, 2, \dots, n$ can be represented by

$$g(t) = \sum_{j=1}^{n-1} c_{j,n} P_{j,n}(t), \quad \text{where } c_{j,n} = \frac{1}{n} \sum_{k=1}^n g(k) P_{j,n}(k).$$

In particular, if $g(t)$ is smooth then

$$g(t) \approx \sum_{j=1}^m c_{j,n} P_{j,n}(t)$$

for modest values of m . This approximation has been used in Bierens (1997) to test the unit root hypothesis against nonlinear trend stationarity, and in Bierens and Martins (2010) to test for time varying cointegration.

Shifted Chebyshev polynomials

Substituting $x = 2u - 1$ for $u \in [0, 1]$ in (4.16) yields

$$w_{\mathcal{C}[0,1]}(u) = \frac{2}{\pi \sqrt{1 - (2u - 1)^2}} = \frac{1}{\pi \sqrt{u(1-u)}}. \quad (4.21)$$

with corresponding distribution function

$$W_{\mathcal{C}[0,1]}(u) = 1 - \pi^{-1} \arccos(2u - 1), \quad (4.22)$$

and shifted Chebyshev polynomials

$$\bar{p}_k(u|w_{\mathcal{C}[0,1]}) = \begin{cases} 1 & \text{for } k = 0, \\ \sqrt{2} \cos(k \cdot \arccos(2u - 1)) & \text{for } k \in \mathbb{N}. \end{cases} \quad (4.23)$$

Again, it follows from Corollary 4.1 that the orthonormal sequence

$$\psi_k(u) = \begin{cases} \sqrt{w_{\mathcal{C}[0,1]}(u)} & \text{for } k = 0, \\ \sqrt{2} \sqrt{w_{\mathcal{C}[0,1]}(u)} \cos(k \cdot \arccos(2u - 1)) & \text{for } k \in \mathbb{N}, \end{cases}$$

is complete in $L^2(0, 1)$. Thus, every function $f \in L^2(0, 1)$ can be written as

$$f(u) = \sum_{k=0}^{\infty} \gamma_k \psi_k(u) \quad (4.24)$$

a.e. on $(0, 1)$, where $\gamma_k = \int_0^1 f(u) \psi_k(u) du$.

4.4 Bivariate functions

Let $w_1(x)$ and $w_2(y)$ be densities with supports (a_1, b_1) and (a_2, b_2) , respectively, where $\infty \leq a_i < b_i \leq \infty$, $i = 1, 2$, satisfying the conditions of Theorem 4.1. Consider the space $L^2(w_1 \times w_2)$ of bivariate Borel measurable functions $f(x, y)$ satisfying

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} w_1(x)w_2(y)f(x, y)^2 dxdy < \infty, \quad (4.25)$$

endowed with the inner product

$$\langle f, g \rangle = \int_{a_1}^{b_1} \int_{a_2}^{b_2} w_1(x)w_2(y)f(x, y)g(x, y)dxdy$$

and associated norm $\|f\| = \sqrt{\langle f, f \rangle}$ and metric $\|f - g\|$. Then for any fixed $y \in (a_2, b_2)$ for which

$$\int_{a_1}^{b_1} w_1(x)f(x, y)^2 dx < \infty, \quad (4.26)$$

we have $f(x, y) \in L^2(w_1)$, hence

$$f(x, y) = \sum_{k=0}^{\infty} \gamma_k(y) \bar{p}_k(x|w_1) \text{ a.e. on } (a_1, b_1). \quad (4.27)$$

where $\gamma_k(y) = \int_{a_1}^{b_1} w_1(x)f(x, y)\bar{p}_k(x|w_1)dx$ and $\sum_{k=0}^{\infty} \gamma_k(y)^2 < \infty$.

Note that by the Cauchy-Schwarz inequality and (4.25),

$$\begin{aligned} \int_{a_2}^{b_2} w_2(y)\gamma_k(y)^2 dy &= \int_{a_2}^{b_2} w_2(y) \left(\int_{a_1}^{b_1} w_1(x)\bar{p}_k(x|w_1)f(x, y)dx \right)^2 dy \\ &\leq \int_{a_1}^{b_1} w_1(x)\bar{p}_k(x|w_1)^2 dx \int_{a_1}^{b_1} \int_{a_2}^{b_2} w_1(x)w_2(y)f(x, y)^2 dxdy \\ &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} w_1(x)w_2(y)f(x, y)^2 dxdy < \infty \end{aligned}$$

where the second equality follows from the fact that $\int_{a_1}^{b_1} w_1(x)\bar{p}_k(x|w_1)^2 dx = 1$, so that $\gamma_k(y) \in L^2(w_2)$. Consequently, for each $k \in \mathbb{N}_0$ we have

$$\gamma_k(y) = \sum_{m=0}^{\infty} \gamma_{k,m} \bar{p}_m(y|w_2) \text{ a.e. on } (a_2, b_2), \quad (4.28)$$

where

$$\begin{aligned}\gamma_{k,m} &= \int_{a_2}^{b_2} w_2(y) \gamma_k(y) \bar{p}_m(y|w_2) dy \\ &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} w_1(x) w_2(y) f(x, y) \bar{p}_k(x|w_1) \bar{p}_m(y|w_2) dx dy\end{aligned}\quad (4.29)$$

and $\sum_{m=0}^{\infty} \gamma_{k,m}^2 < \infty$.

Moreover, note that due to (4.25) the restriction (4.26) holds a.e. on (a_2, b_2) , so that

$$f(x, y) = \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \gamma_{k,m} \bar{p}_k(x|w_1) \bar{p}_m(y|w_2) \text{ a.e. on } (a_1, b_1) \times (a_2, b_2),$$

where the double-array $\gamma_{k,m}$ of Fourier coefficients are given by (4.29) and satisfies $\sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \gamma_{k,m}^2 < \infty$. Consequently, the space $L^2(w_1 \times w_2)$ is a Hilbert space.

Recall that in the case

$$w_1(x) = w_2(x) = \exp(-x^2/2)/\sqrt{2\pi} = w_{\mathcal{N}[0,1]}(x)$$

the polynomials $\bar{p}_k(x|w_{\mathcal{N}[0,1]})$ are the Hermite polynomials. Then every density $f(x, y)$ on \mathbb{R}^2 can be written as

$$\begin{aligned}f(x, y) &= \frac{\exp\left(-\frac{1}{2}(x^2 + y^2)\right)}{2\pi} \\ &\times \left(\sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \gamma_{k,m} \bar{p}_k(x|w_{\mathcal{N}[0,1]}) \bar{p}_m(y|w_{\mathcal{N}[0,1]}) \right)^2 \\ &\text{a.e. on } \mathbb{R}^2, \text{ where } \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \gamma_{k,m}^2 = 1.\end{aligned}\quad (4.30)$$

This is the approach taken by Gallant and Nychka (1987). They consider SNP estimation of Heckman's (1979) sample selection model, where the bivariate error distribution of the latent variable equations involved is modeled semi-nonparametrically via the Hermite expansion (4.30) of the error density.

4.5 Appendix: Proofs

4.5.1 Theorem 4.1

Let $\bar{f}_n(x) = \sum_{k=0}^n \gamma_k \bar{p}_k(x|w)$, where $\gamma_k = \int_a^b \bar{p}_k(x|w) f(x) w(x) dx$, and observe that due to condition (4.5), $\bar{f}_n \in L^2(w)$. Next, observe that

$$\begin{aligned} \|f - \bar{f}_n\|^2 &= \int_a^b \left(f(x) - \sum_{k=0}^n \gamma_k \bar{p}_k(x|w) \right)^2 w(x) dx \\ &= \int_a^b f(x)^2 w(x) dx - 2 \sum_{k=0}^n \gamma_k \int_a^b \bar{p}_k(x|w) f(x) w(x) dx \\ &\quad + \sum_{k_1=0}^n \sum_{k_2=0}^n \gamma_{k_1} \gamma_{k_2} \int_a^b \bar{p}_{k_1}(x|w) \bar{p}_{k_2}(x|w) w(x) dx \\ &= \int_a^b f(x)^2 w(x) dx - \sum_{k=0}^n \gamma_k^2 \geq 0. \end{aligned} \tag{4.31}$$

Hence $\sum_{k=0}^n \gamma_k^2 \leq \int_a^b f(x)^2 w(x) dx < \infty$ for all $n \geq 0$, and thus

$$\sum_{k=0}^{\infty} \gamma_k^2 < \infty. \tag{4.32}$$

The latter implies that \bar{f}_n is a Cauchy sequence in $L^2(w)$ because

$$\lim_{\min(n,m) \rightarrow \infty} \|\bar{f}_n - \bar{f}_m\|^2 = \lim_{\min(n,m) \rightarrow \infty} \sum_{k=\min(n,m)+1}^{\max(n,m)} \gamma_k^2 \leq \lim_{m \rightarrow \infty} \sum_{k=m}^{\infty} \gamma_k^2 = 0.$$

Therefore, there exists a function $\bar{f} \in L^2(w)$ such that

$$\lim_{n \rightarrow \infty} \|\bar{f}_n - \bar{f}\| = 0. \tag{4.33}$$

This limit function \bar{f} can be written as

$$\bar{f}(x) = \sum_{k=0}^n \gamma_k \bar{p}_k(x|w) + \varepsilon_n(x) \tag{4.34}$$

for all $n \in \mathbb{N}$, where

$$\lim_{n \rightarrow \infty} \int_a^b \varepsilon_n(x)^2 w(x) dx = 0. \quad (4.35)$$

Proof of (4.7)

To prove (4.7), it suffices to show that

$$\int_a^b \exp(\mathbf{i}.t.x) (f(x) - \bar{f}(x)) w(x) dx = 0 \quad (4.36)$$

for all $t \in R$, because (4.36) implies that $f(x) = \bar{f}(x)$ a.e. on (a, b) , due to the uniqueness of the Fourier transform.⁵.

It follows from the definition of γ_m and \bar{f} that for $m \leq n$,

$$\begin{aligned} \left| \int_a^b (f(x) - \bar{f}(x)) \bar{p}_m(x|w) w(x) dx \right| &= \left| \int_a^b \varepsilon_n(x) \bar{p}_m(x|w) w(x) dx \right| \\ &\leq \sqrt{\int_a^b \varepsilon_n(x)^2 w(x) dx}, \end{aligned}$$

hence by (4.35),

$$\int_a^b (f(x) - \bar{f}(x)) \bar{p}_m(x|w) w(x) dx = 0 \quad (4.37)$$

for all $m \in \mathbb{N}$. This result implies, by induction, that

$$\int_a^b (f(x) - \bar{f}(x)) x^m w(x) dx = 0 \text{ for all } m \in \mathbb{N}. \quad (4.38)$$

In its turn (4.38) implies, together with the well-known equality $\exp(\mathbf{i}.t.x) = \sum_{m=0}^{\infty} (\mathbf{i}.t.x)^m / m!$, that for $t \in \mathbb{R}$ and all $n \in \mathbb{N}$,

$$\begin{aligned} &\int_a^b \exp(\mathbf{i}.t.x) (f(x) - \bar{f}(x)) w(x) dx \\ &= \int_a^b \sum_{m=0}^n \frac{(\mathbf{i}.t.x)^m}{m!} (f(x) - \bar{f}(x)) w(x) dx \end{aligned}$$

⁵See for example Bierens (1994, Theorem 3.1.1, p.50).

$$\begin{aligned}
& + \int_a^b \left(\sum_{m=n+1}^{\infty} \frac{(\mathbf{i}.t.x)^m}{m!} \right) (f(x) - \bar{f}(x)) w(x) dx \\
& = \int_a^b \left(\sum_{m=n+1}^{\infty} \frac{(\mathbf{i}.t.x)^m}{m!} \right) (f(x) - \bar{f}(x)) w(x) dx
\end{aligned}$$

If $-\infty < a < b < \infty$ then by dominated convergence,

$$\begin{aligned}
& \int_a^b \exp(\mathbf{i}.t.x) (f(x) - \bar{f}(x)) w(x) dx \\
& = \int_a^b \left(\lim_{n \rightarrow \infty} \sum_{m=n+1}^{\infty} \frac{(\mathbf{i}.t.x)^m}{m!} \right) (f(x) - \bar{f}(x)) w(x) dx = 0
\end{aligned}$$

If $a = -\infty$ and/or $b = \infty$ we can find for arbitrary $\varepsilon > 0$ a finite lower bound $a(\varepsilon) > a$ and a finite upper bound $b(\varepsilon) < b$ such that

$$\begin{aligned}
\left| \int_a^{a(\varepsilon)} \exp(\mathbf{i}.t.x) (f(x) - \bar{f}(x)) w(x) dx \right| & < \varepsilon/2 \\
\left| \int_{b(\varepsilon)}^b \exp(\mathbf{i}.t.x) (f(x) - \bar{f}(x)) w(x) dx \right| & < \varepsilon/2
\end{aligned}$$

whereas by dominated convergence

$$\begin{aligned}
& \int_{a(\varepsilon)}^{b(\varepsilon)} \exp(\mathbf{i}.t.x) (f(x) - \bar{f}(x)) w(x) dx \\
& = \int_{a(\varepsilon)}^{b(\varepsilon)} \left(\lim_{n \rightarrow \infty} \sum_{m=n+1}^{\infty} \frac{(\mathbf{i}.t.x)^m}{m!} \right) (f(x) - \bar{f}(x)) w(x) dx = 0
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we therefore have in either case that (4.36) holds. It therefore follows from (4.34) and (4.35) that

$$\lim_{n \rightarrow \infty} \int_a^b \left(f(x) - \sum_{k=0}^n \gamma_k \bar{p}_k(x|w) \right)^2 w(x) dx = 0. \quad (4.39)$$

This completes the proof of (4.7).

Proof of (4.8)

To prove that (4.7) implies (4.8), let X be a random drawing from $w(x)$. Then by Chebyshev's inequality, (4.39) implies

$$f(X) = \operatorname{plim}_{n \rightarrow \infty} \sum_{k=0}^n \gamma_k \bar{p}_k(X|w) \quad (4.40)$$

As is well-known⁶, convergence in probability is equivalent to almost sure (a.s.) convergence along a further subsequence of an arbitrary subsequence of n . Thus it follows from (4.40) that for any subsequence n_j in \mathbb{N} there exists a further subsequence n_{j_m} such that for $m \rightarrow \infty$,

$$\sum_{k=0}^{n_{j_m}} \gamma_k \bar{p}_k(X|w) \xrightarrow{\text{a.s.}} f(X). \quad (4.41)$$

For each n there exists an m such that $n_{j_{m-1}} \leq n < n_{j_m}$. Hence, there exists a further subsequence j_n of n_{j_m} such that for $j_{n-1} \leq n < j_n$ and $n \rightarrow \infty$,

$$\sum_{k=0}^{j_n} \gamma_k \bar{p}_k(X|w) \xrightarrow{\text{a.s.}} f(X). \quad (4.42)$$

The latter implies that

$$\begin{aligned} E \left[\left(\sum_{k=0}^{j_n} \gamma_k \bar{p}_k(X|w) - \sum_{k=0}^n \gamma_k \bar{p}_k(X|w) \right)^2 \right] &= E \left(\sum_{k=n+1}^{j_n} \gamma_k \bar{p}_k(X|w) \right)^2 \\ &\leq \sum_{k=j_{n-1}+1}^{j_n} \gamma_k^2 \end{aligned}$$

so that

$$\begin{aligned} \sum_{n=1}^{\infty} E \left[\left(\sum_{k=0}^{j_n} \gamma_k \bar{p}_k(X|w) - \sum_{k=0}^n \gamma_k \bar{p}_k(X|w) \right)^2 \right] &\leq \sum_{n=1}^{\infty} \sum_{k=j_{n-1}+1}^{j_n} \gamma_k^2 \\ &\leq \sum_{k=0}^{\infty} \gamma_k^2 < \infty \end{aligned}$$

⁶See for example Bierens (2004, Theorem 6.B.3, p.168).

Then by Chebyshev's inequality,

$$\sum_{n=1}^{\infty} \Pr \left[\left| \sum_{k=0}^{j_n} \gamma_k \bar{p}_k(X|w) - \sum_{k=0}^n \gamma_k \bar{p}_k(X|w) \right| > \varepsilon \right] < \infty$$

for all $\varepsilon > 0$, which by the Borel-Cantelli lemma⁷ implies that for $n \rightarrow \infty$

$$\sum_{k=0}^{j_n} \gamma_k \bar{p}_k(X|w) - \sum_{k=0}^n \gamma_k \bar{p}_k(X|w) \xrightarrow{a.s.} 0. \quad (4.43)$$

Combining (4.42) and (4.43), it follows that $\sum_{k=0}^n \gamma_k \bar{p}_k(X|w) \xrightarrow{a.s.} f(X)$ as $n \rightarrow \infty$, which is equivalent to (4.8) because the support of $w(x)$ was assumed to be (a, b) .

4.5.2 Theorem 4.2

Due to the normalization $\alpha_{k,k} = 1$ it follows that $p_{k+1}(x|w) - x.p_k(x|w)$ is a polynomial of order k , which can be written as a linear combination of $p_0(x|w), p_1(x|w), \dots, p_k(x|w)$:

$$p_{k+1}(x|w) - x.p_k(x|w) = \sum_{j=0}^k \delta_{j,k} p_j(x|w) \quad (4.44)$$

for example. Then for $m \leq k$,

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} p_{k+1}(x|w) p_m(x|w) w(x) dx - \int_{-\infty}^{\infty} x.p_k(x|w) p_m(x|w) w(x) dx \\ &\quad - \sum_{j=0}^k \delta_{j,k} \int_{-\infty}^{\infty} p_j(x|w) p_m(x|w) w(x) dx \\ &= - \int_{-\infty}^{\infty} x.p_k(x|w) p_m(x|w) w(x) dx - \delta_{m,k} \int_{-\infty}^{\infty} p_m(x|w)^2 w(x) dx \end{aligned}$$

so that

$$\delta_{m,k} = -\frac{\int_{-\infty}^{\infty} (x.p_m(x|w)) p_k(x|w) w(x) dx}{\int_{-\infty}^{\infty} p_m(x|w)^2 w(x) dx}, \quad m = 0, 1, 2, \dots, k.$$

⁷See for example Bierens (2004, Theorem 2.B.2, p. 168).

Because $x.p_m(x|w)$ is a polynomial of order $m + 1$, it follows that for $m \leq k - 2$, $x.p_m(x|w)$ is orthogonal to $p_k(x|w)$, hence $\delta_{m,k} = 0$ for $m = 0, 1, \dots, k - 2$. Thus it follows from (4.44) that

$$\begin{aligned} p_{k+1}(x|w) - x.p_k(x|w) &= \delta_{k,k}p_k(x|w) + \delta_{k-1,k}p_{k-1}(x|w) \\ &= -b_k p_k(x|w) - c_k p_{k-1}(x|w) \end{aligned}$$

where

$$b_k = -\delta_{k,k} = \frac{\int_{-\infty}^{\infty} x.p_k(x|w)^2 w(x) dx}{\int_{-\infty}^{\infty} p_k(x|w)^2 w(x) dx}$$

and

$$\begin{aligned} c_k = -\delta_{k-1,k} &= \frac{\int_{-\infty}^{\infty} x.p_{k-1}(x|w).p_k(x|w) w(x) dx}{\int_{-\infty}^{\infty} p_{k-1}(x|w)^2 w(x) dx} \\ &= \frac{\int_{-\infty}^{\infty} p_k(x|w)^2 w(x) dx}{\int_{-\infty}^{\infty} p_{k-1}(x|w)^2 w(x) dx} \end{aligned}$$

The last equality follows from the fact that $x.p_{k-1}(x|w)$ can be written as $x.p_{k-1}(x|w) = \sum_{m=0}^k \beta_{m,k} p_m(x|w)$, where $\beta_{k,k} = 1$, so that

$$\begin{aligned} \int_{-\infty}^{\infty} x.p_{k-1}(x|w).p_k(x|w) w(x) dx &= \sum_{m=0}^k \beta_{m,k} \int_{-\infty}^{\infty} p_m(x|w) p_k(x|w) w(x) dx \\ &= \beta_{k,k} \int_{-\infty}^{\infty} p_k(x|w)^2 w(x) dx \\ &= \int_{-\infty}^{\infty} p_k(x|w)^2 w(x) dx. \end{aligned}$$

4.5.3 Lemma 4.1

Using the well-known cosine formulas $2 \cos(a) \cos(b) = \cos(a+b) + \cos(a-b)$ and $\cos(a-b) = \cos(a) \cos(b) + \sin(a) \sin(b)$ we can write

$$\begin{aligned} &\sum_{k=1}^n p_{j_1}(x_{n,k}|w_{\mathcal{C}[-1,1]}) p_{j_2}(x_{n,k}|w_{\mathcal{C}[-1,1]}) \\ &= \sum_{k=1}^n \cos(\pi j_1(k-0.5)/n) \cos(\pi j_2(k-0.5)/n) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \sum_{k=1}^n \cos(\pi(j_1 + j_2)(k - 0.5)/n) + \frac{1}{2} \sum_{k=1}^n \cos(\pi(j_1 - j_2)(k - 0.5)/n) \\
&= \frac{1}{2} \cos(0.5\pi(j_1 + j_2)/n) \sum_{k=1}^n \cos(\pi(j_1 + j_2)k/n) \\
&\quad + \frac{1}{2} \sin(0.5\pi(j_1 + j_2)/n) \sum_{k=1}^n \sin(\pi(j_1 + j_2)k/n) \\
&\quad + \frac{1}{2} \cos(0.5\pi(j_1 - j_2)/n) \sum_{k=1}^n \cos(\pi(j_1 - j_2)k/n) \\
&\quad + \frac{1}{2} \sin(0.5\pi(j_1 - j_2)/n) \sum_{k=1}^n \sin(\pi(j_1 - j_2)k/n)
\end{aligned}$$

Moreover, using the well-known De Moivre formula $\exp(\mathbf{i}.a) = \cos(a) + \mathbf{i}.\sin(a)$ it follows that

$$\begin{aligned}
&\frac{1}{2} \sum_{k=1}^n \cos(\pi.m.k/n) \\
&= \sum_{k=1}^n \exp(\mathbf{i}.\pi m.k/n) + \sum_{k=1}^n \exp(-\mathbf{i}.\pi m.k/n) \\
&= \sum_{k=1}^n (\exp(\mathbf{i}.\pi m/n))^k + \sum_{k=1}^n (\exp(-\mathbf{i}.\pi m/n))^k \\
&= \exp(\mathbf{i}.\pi m/n) \frac{\exp(\mathbf{i}.\pi m) - 1}{\exp(\mathbf{i}.\pi m/n) - 1} + \exp(-\mathbf{i}.\pi m/n) \frac{\exp(-\mathbf{i}.\pi m) - 1}{\exp(-\mathbf{i}.\pi m/n) - 1} \\
&= \frac{\exp(\mathbf{i}.\pi m/(2n))}{\exp(\mathbf{i}.\pi m/(2n)) - \exp(-\mathbf{i}.\pi m/(2n))} (\cos(\pi m) - 1) \\
&\quad - \frac{\exp(-\mathbf{i}.\pi m/(2n))}{\exp(\mathbf{i}.\pi m/(2n)) - \exp(-\mathbf{i}.\pi m/(2n))} (\cos(\pi m) - 1) \\
&= \cos(\pi m) - 1
\end{aligned}$$

and similarly for $m \neq 0$,

$$\frac{1}{2} \sum_{k=1}^n \sin(\pi.m.k/n)$$

$$\begin{aligned}
&= \frac{1}{i} \sum_{k=1}^n \exp(i.\pi m.k/n) - \frac{1}{i} \sum_{k=1}^n \exp(-i.\pi m.k/n) \\
&= \frac{1}{i} \sum_{k=1}^n (\exp(i.\pi m/n))^k - \frac{1}{i} \sum_{k=1}^n (\exp(-i.\pi m/n))^k \\
&= \frac{1}{i} \exp(i.\pi m/n) \frac{\exp(i.\pi m) - 1}{\exp(i.\pi m/n) - 1} - \frac{1}{i} \exp(-i.\pi m/n) \frac{\exp(-i.\pi m) - 1}{\exp(-i.\pi m/n) - 1} \\
&= \frac{1}{i} \frac{\exp(i.\pi m/(2n)) + \exp(-i.\pi m/(2n))}{\exp(i.\pi m/(2n)) - \exp(-i.\pi m/(2n))} (\cos(\pi m) - 1) \\
&= -\frac{\cos(\pi m/(2n))}{\sin(\pi m/(2n))} (\cos(\pi m) - 1)
\end{aligned}$$

Thus, for $j_1 \neq j_2$,

$$\begin{aligned}
&\sum_{k=1}^n p_{j_1}(x_{n,k}|w_{\mathcal{C}[-1,1]}) p_{j_2}(x_{n,k}|w_{\mathcal{C}[-1,1]}) \\
&= \frac{1}{2} \cos(0.5\pi(j_1 + j_2)/n) (\cos(\pi(j_1 + j_2)) - 1) \\
&\quad - \frac{1}{2} \cos(\pi(j_1 + j_2)/(2n)) (\cos(\pi(j_1 + j_2)) - 1) \\
&\quad + \frac{1}{2} \cos(0.5\pi(j_1 - j_2)/n) (\cos(\pi(j_1 - j_2)) - 1) \\
&\quad - \frac{1}{2} \cos(\pi(j_1 - j_2)/(2n)) (\cos(\pi(j_1 - j_2)) - 1) \\
&= 0
\end{aligned}$$

whereas for $j_1 = j_2 = j > 0$,

$$\begin{aligned}
&\sum_{k=1}^n p_j(x_{n,k}|w_{\mathcal{C}[-1,1]}) p_j(x_{n,k}|w_{\mathcal{C}[-1,1]}) \\
&= \frac{1}{2} \cos(\pi.j/n) \sum_{k=1}^n \cos(2\pi.j.k/n) + \frac{1}{2} \sin(\pi.j/n) \sum_{k=1}^n \sin(2\pi.j.k/n) \\
&\quad + \frac{1}{2}n = \frac{1}{2}n
\end{aligned}$$

The case $j_1 = j_2 = 0$ is trivial.

Chapter 5

Trigonometric series

5.1 Cosine series representation

Note that the distribution function $W_{\mathcal{C}[0,1]}(u)$ defined by (4.22) has inverse

$$W_{\mathcal{C}[0,1]}^{-1}(u) = (1 + \cos(\pi(1 - u))) / 2. \quad (5.1)$$

It is now easy to verify from Corollary 4.2, (5.1) and (4.24) that every function $f \in L^2(0, 1)$ can be written as

$$\begin{aligned} f(u) &= \gamma_0 + \sum_{k=1}^{\infty} \gamma_k \sqrt{2} \cos(k \cdot \arccos(2W_c^{-1}(u) - 1)) \\ &= \gamma_0 + \sum_{k=1}^{\infty} \gamma_k \sqrt{2} \cos(k\pi(1 - u)) \\ &= \gamma_0 + \sum_{k=1}^{\infty} \gamma_k (-1)^k \sqrt{2} \cos(k\pi u) \\ &= \alpha_0 + \sum_{k=1}^{\infty} \alpha_k \sqrt{2} \cos(k\pi u) \end{aligned}$$

where

$$\begin{aligned} \alpha_k &= \gamma_k (-1)^k = (-1)^k \int_0^1 f(u) p_k \left(\frac{1}{2} (1 + \cos(\pi(1 - u))) \Big| w_{\mathcal{C}[0,1]} \right) du \\ &= \begin{cases} \int_0^1 f(u) du & \text{if } k = 0, \\ \int_0^1 f(u) \sqrt{2} \cos(k\pi u) du & \text{if } k \in \mathbb{N}. \end{cases} \end{aligned}$$

Consequently,

Theorem 5.1. *The functions*

$$\kappa_k(u) = \begin{cases} 1 & \text{if } k = 0, \\ \sqrt{2} \cos(k\pi u) & \text{if } k \in \mathbb{N}, \end{cases}$$

form a complete orthonormal sequence in $L^2(0, 1)$. Thus, given a function $f \in L^2(0, 1)$, let

$$f_n(u) = \alpha_0 + \sum_{k=1}^n \alpha_k \sqrt{2} \cos(k\pi u)$$

where $\alpha_k = \int_0^1 f(u) \kappa_k(u) du$. Then $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ and

$$\lim_{n \rightarrow \infty} \int_0^1 (f(u) - f_n(u))^2 du = \lim_{n \rightarrow \infty} \sum_{k=n+1}^{\infty} \alpha_k^2 = 0.$$

Consequently, similar to Theorem 4.1, f can be written as

$$f(u) = \alpha_0 + \sum_{k=1}^{\infty} \alpha_k \sqrt{2} \cos(k\pi u) \text{ a.e. on } (0, 1). \quad (5.2)$$

5.2 Fourier analysis

Consider the following sequence of functions on $[-1, 1]$:

$$\begin{aligned} \varphi_0(x) &= 1 \\ \varphi_{2k-1}(x) &= \sqrt{2} \sin(k\pi x), \quad \varphi_{2k}(x) = \sqrt{2} \cos(k\pi x), \quad k \in \mathbb{N}. \end{aligned} \quad (5.3)$$

These functions are known as the Fourier series on $[-1, 1]$. It is easy to verify that these functions are orthonormal with respect to the weight function $w(x) = \frac{1}{2}I(|x| \leq 1)$, i.e.,

$$\frac{1}{2} \int_{-1}^1 \varphi_m(x) \varphi_k(x) dx = I(m = k)$$

It is a classical Fourier analysis result that

Theorem 5.2. *The Fourier series $\{\varphi_n\}_{n=0}^{\infty}$ is complete in $L^2(-1, 1)$.*

The "official" proof of this result is long and tedious. See for example Young (1988). However, using Theorem 5.1 this result can be proved somewhat easier, as follows.

We need to show that for an arbitrary function $g \in L^2(-1, 1)$,

$$\lim_{n \rightarrow \infty} \frac{1}{2} \int_{-1}^1 (g(x) - g_n(x))^2 dx, \quad (5.4)$$

where

$$g_n(x) = \alpha_0 + \sum_{k=1}^n \alpha_k \sqrt{2} \cos(k\pi x) + \sum_{k=1}^n \beta_k \sqrt{2} \sin(k\pi x) \quad (5.5)$$

with Fourier coefficients

$$\begin{aligned} \alpha_0 &= \frac{1}{2} \int_{-1}^1 g(x) dx \\ \alpha_k &= \frac{1}{2} \int_{-1}^1 \sqrt{2} \cos(k\pi x) g(x) dx \\ \beta_k &= \frac{1}{2} \int_{-1}^1 \sqrt{2} \sin(k\pi x) g(x) dx. \end{aligned}$$

Let $x = 2u - 1$ for $u \in [0, 1]$, and denote

$$f(u) = g(2u - 1), \quad f_n(u) = g_n(2u - 1)$$

Then it follows from the well-known sine-cosine equalities that

$$\begin{aligned} \alpha_0 &= \int_0^1 g(2u - 1) du = \int_0^1 f(u) du \\ \alpha_k &= \int_0^1 \sqrt{2} \cos(k\pi(2u - 1)) g(2u - 1) du \\ &= (-)^k \int_0^1 \sqrt{2} \cos(2k\pi u) f(u) du \\ \beta_k &= \int_0^1 \sqrt{2} \sin(k\pi(2u - 1)) g(2u - 1) du \\ &= (-)^k \int_0^1 \sqrt{2} \sin(2k\pi u) f(u) du \end{aligned}$$

and

$$\begin{aligned}
 f_n(u) &= g_n(2u - 1) \\
 &= \alpha_0 + \sum_{k=1}^n \alpha_k \sqrt{2} \cos(k\pi(2u - 1)) + \sum_{k=1}^n \beta_k \sqrt{2} \sin(k\pi(2u - 1)) \\
 &= \alpha_0 + \sum_{k=1}^n \alpha_k (-)^k \sqrt{2} \cos(2k\pi u) + \sum_{k=1}^n \beta_k (-)^k \sqrt{2} \sin(2k\pi u)
 \end{aligned}$$

Thus, (5.4) is true if and only

$$\lim_{n \rightarrow \infty} \int_0^1 (f(u) - f_n(u))^2 du = 0.$$

Theorem 5.2 follows now from the following result, which will be proved in the appendix to this chapter.

Theorem 5.3. *The functions $\bar{\varphi}_0(u) = 1$, $\bar{\varphi}_k(u) = \sqrt{2} \sin(2k\pi u)$ if $k \geq 1$ is odd, $\bar{\varphi}_k(u) = \sqrt{2} \cos(2k\pi u)$ if $k \geq 2$ is even, form a complete orthonormal sequence in $L^2(0, 1)$.*

Although Theorem 5.1 was used to prove Theorem 5.2, Theorem 5.2 can also be proved independently. See for example Young (1988). Then Theorem 5.1 becomes a corollary of Theorem 5.2, as follows.

Let $f(u) \in L^2(0, 1)$ be arbitrary, and let $g(x) = f(|x|)$. Then $g(x) \in L^2(-1, 1)$, with Fourier coefficients

$$\begin{aligned}
 \alpha_0 &= \frac{1}{2} \int_{-1}^1 f(|x|) dx = \int_0^1 f(u) du \\
 \alpha_k &= \frac{1}{2} \int_{-1}^1 \sqrt{2} \cos(k\pi x) f(|x|) dx = \int_0^1 \sqrt{2} \cos(k\pi u) f(u) du \\
 \beta_k &= \frac{1}{2} \int_{-1}^1 \sqrt{2} \sin(k\pi x) f(|x|) dx = 0
 \end{aligned}$$

Hence it follows from Theorem 5.2 that

$$\lim_{n \rightarrow \infty} \int_0^1 \left(f(u) - \alpha_0 - \sum_{k=1}^n \alpha_k \sqrt{2} \cos(k\pi u) \right)^2 du \quad (5.6)$$

$$\begin{aligned}
&= \frac{1}{2} \lim_{n \rightarrow \infty} \int_{-1}^1 \left(f(|x|) - \alpha_0 - \sum_{k=1}^n \alpha_k \sqrt{2} \cos(k\pi x) \right)^2 dx \\
&= 0
\end{aligned}$$

Similar to the proof of Theorem 4.1 is follows now from (5.6) that

$$f(u) = \alpha_0 + \sum_{k=1}^{\infty} \alpha_k \sqrt{2} \cos(k\pi u) \text{ a.e. on } (0, 1),$$

where $\alpha_0 = \int_0^1 f(u) du$ and $\alpha_k = \int_0^1 \sqrt{2} \cos(k\pi u) f(u) du$ for $k \geq 1$, which is just the result in Theorem 5.1.

5.3 Sine series representation

Let $f(x)$ be a square integrable function on $[-1, 1]$ such that $f(x) = -f(-x)$, with a possible discontinuity at $x = 0$. Then

$$\begin{aligned}
\beta_k &= \frac{1}{2} \int_{-1}^1 f(x) \sqrt{2} \sin(k\pi x) du = \int_0^1 f(u) \sqrt{2} \sin(k\pi u) du \\
0 &= \frac{1}{2} \int_{-1}^1 f(x) \sqrt{2} \cos(k\pi x) dx \\
0 &= \frac{1}{2} \int_{-1}^1 f(x) dx = 0
\end{aligned}$$

Hence by Theorem 5.2, $\lim_{n \rightarrow \infty} \frac{1}{2} \int_{-1}^1 \left(f(x) - \sum_{k=1}^n \beta_k \sqrt{2} \sin(k\pi x) \right)^2 dx = 0$, which by the condition $f(x) = -f(-x)$ implies

$$\lim_{n \rightarrow \infty} \int_0^1 (f(u) - f_n(u))^2 du = 0,$$

where

$$f_n(u) = \sum_{k=1}^n \beta_k \sqrt{2} \sin(k\pi u)$$

Moreover, it is easy to verify that

$$\int_0^1 \sqrt{2} \sin(k\pi u) \sqrt{2} \sin(m\pi u) du = I(k = m).$$

Thus, we have the following corollary of Theorem 5.2.

Theorem 5.4. *The sine series $\{\sqrt{2} \sin(k\pi u)\}_{k=1}^{\infty}$ is a complete orthonormal sequence in $L^2(0, 1)$. Consequently, any function $f \in L^2(0, 1)$ can be written as*

$$f(u) = \sum_{k=1}^{\infty} \beta_k \sqrt{2} \sin(k\pi u) \text{ a.e. on } (0, 1),$$

where $\beta_k = \int_0^1 f(u) \sqrt{2} \sin(k\pi u) du$.

Note however that $f_n(u)$ will be a poor approximation of $f(u)$ for u close to zero or one because $f_n(0) = f_n(1) = 0$ whereas $f(0)$ and $f(1)$ may be nonzero. The reason is that in general $\lim_{u \rightarrow u_0} \lim_{n \rightarrow \infty} f_n(u) \neq \lim_{n \rightarrow \infty} \lim_{u \rightarrow u_0} f_n(u)$.

5.4 How well does the cosine series fit?

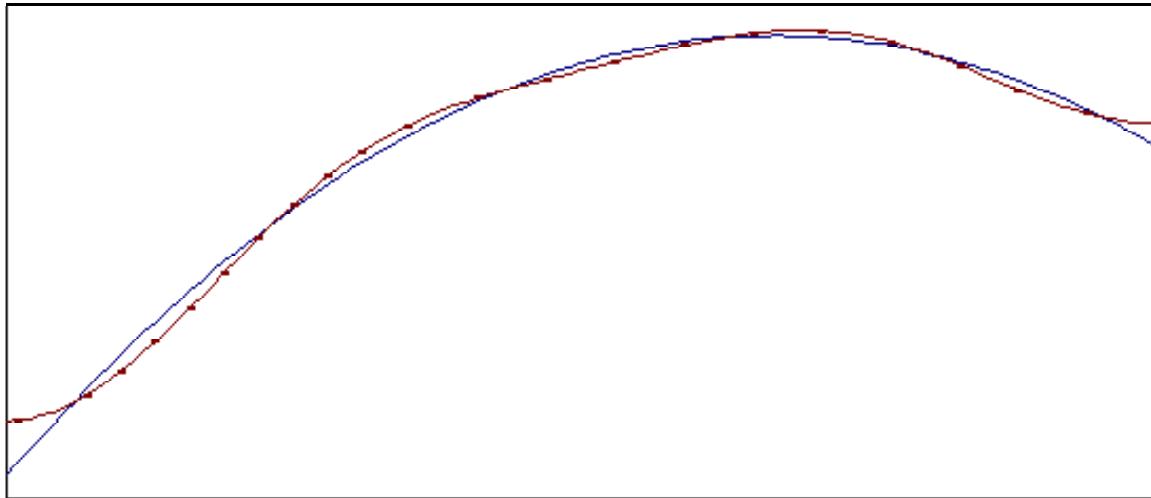
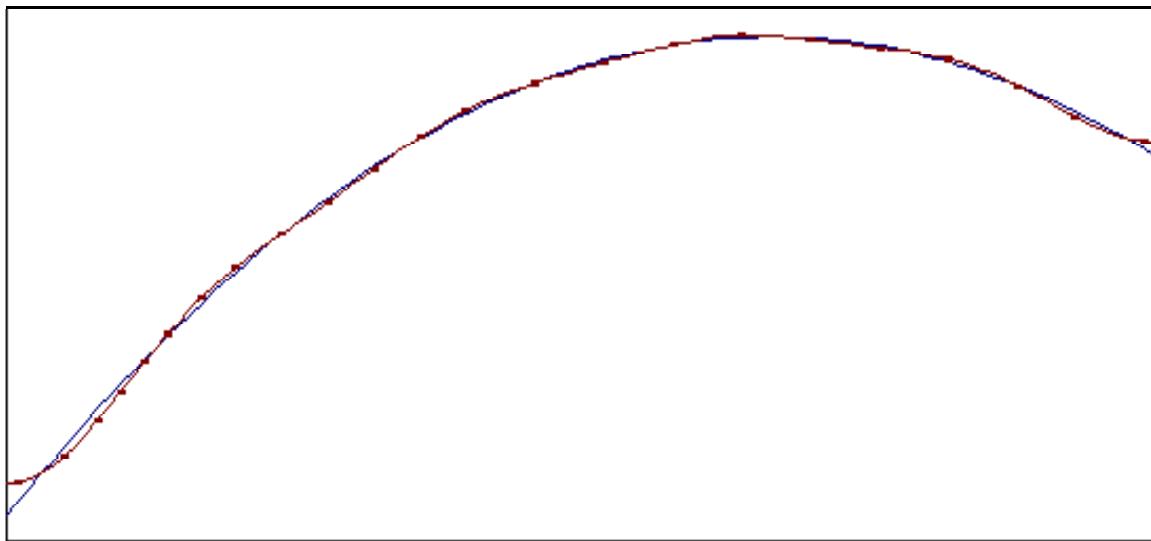
5.4.1 Exact Fourier coefficients

To check how well the cosine series fit, consider the function $f(u) = u(4 - 3u)$ on $[0, 1]$. Note that this is a density function. For this function we can derive the Fourier coefficients involved analytically, as

$$\begin{aligned} \alpha_0 &= \int_0^1 f(u) du = 1, \\ \alpha_k &= \int_0^1 f(u) \sqrt{2} \cos(k\pi u) du = -2\sqrt{2}(k\pi)^{-2} ((-1)^k + 1) \end{aligned}$$

This way of approximating densities directly by a series expansion has been advocated by Kronmal and Tarter (1968). However, a potential problem with this approach is that in general there is no guarantee that $f_n(u) \geq 0$.

In the following figures the function $f(u) = u(4 - 3u)$ is compared with its SNP approximation $f_n(u) = 1 + \sum_{k=1}^n \alpha_k \sqrt{2} \cos(k\pi u)$ (dotted curve) for $n = 4, 8, 12$.

Figure 5.1: $f(u) = u(4 - 3u)$ compared with $f_n(u)$ for $n = 4$ Figure 5.2: $f(u) = u(4 - 3u)$ compared with $f_n(u)$ for $n = 8$

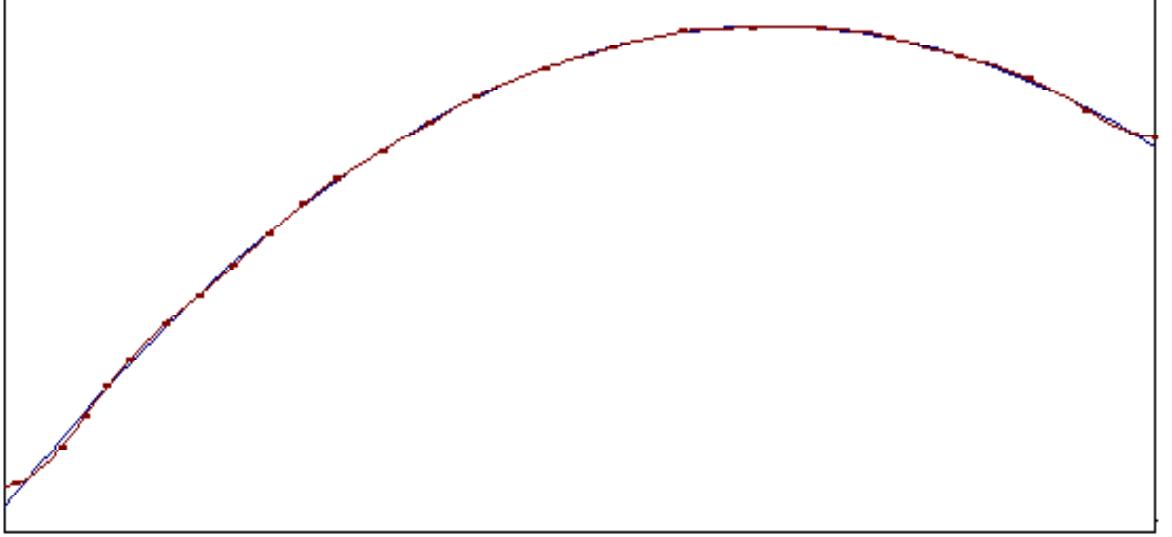


Figure 5.3: $f(u) = u(4 - 3u)$ compared with $f_n(u)$ for $n = 12$

We see that $f_n(u)$ approximates $f(u)$ quite well, even for $n = 4$, except for the tails of $f_n(u)$ in the latter case. The reason is that $f'_n(u) = -\sum_{k=1}^n \alpha_k k \pi \sqrt{2} \sin(k \pi u)$, so that $f'_n(0) = f'_n(1) = 0$. As expected, the tail fit becomes better for larger truncation orders n .

5.4.2 Bivariate SNP regression

Let $(Y, X) \in \mathbb{R}^2$ be a pair of absolutely continuous random variables satisfying

$$E[Y^2] < \infty, \quad E[X^2] < \infty. \quad (5.7)$$

We can always write

$$E[Y|X] = f(X) = \alpha + \beta X + X^2 r(X), \quad (5.8)$$

where $\alpha + \beta X$ is the linear projection of Y on 1 and X , with residual $X^2 r(X)$. Moreover, given an absolutely continuous distribution function $G(x)$ with density $g(x) > 0$ on \mathbb{R} and inverse $G^{-1}(u)$, $u \in [0, 1]$, we can write

$$r(x) = \varphi(G(x)) \quad (5.9)$$

where

$$\varphi(u) = r(G^{-1}(u)) \quad (5.10)$$

Now let us assume that

$$\int_0^1 \varphi(u)^2 du = \int_{-\infty}^{\infty} r(x)^2 g(x) dx < \infty \quad (5.11)$$

so that $\varphi \in L^2(0, 1)$. Then by Theorem 5.1, φ has the series expansion

$$\varphi(u) = \gamma + \sum_{k=1}^{\infty} \delta_k \sqrt{2} \cos(k\pi u) \text{ a.e. on } [0, 1],$$

where

$$\gamma = \int_0^1 \varphi(u) du, \quad \delta_k = \int_0^1 \sqrt{2} \cos(k\pi u) \varphi(u) du.$$

Consequently,

$$f(X) = E[Y|X] = \alpha + \beta X + \gamma X^2 + X^2 \sum_{k=1}^{\infty} \delta_k \sqrt{2} \cos(k\pi G(X)) \text{ a.s.}$$

Next, let

$$f_n(X) = \alpha + \beta X + \gamma X^2 + X^2 \sum_{k=1}^n \delta_k \sqrt{2} \cos(k\pi G(X)),$$

and denote $r_n(x) = \sum_{k=1}^n \delta_k \sqrt{2} \cos(k\pi G(X))$. Since by Theorem 5.1,

$$\lim_{n \rightarrow \infty} r_n(x) = r(x) \text{ a.e.,}$$

it follows that

$$\lim_{n \rightarrow \infty} f_n(X) = f(X) \text{ a.s.}$$

In principle we could specify $f(X)$ directly as $f(X) = \varphi(G(x))$, but if $f(x)$ is linear then we need the full series expansion of φ to fit $f(x) = \alpha + \beta X$, whereas in the case (5.8) the linear regression model corresponds to $r(x) \equiv 0$.

A convenient choice for G is the logistic distribution function

$$G(x) = (1 + \exp(-x))^{-1},$$

which has density $g(x) = G(x)(1 - G(x))$ and inverse $G^{-1}(u) = \ln(u/(1-u))$. Since all the moments of the Logistic distribution are finite, the condition (5.11) allows $r(x)$ to be a polynomial of any order.

To check how well $f_n(X)$ fits, let $Y = f(X) + U$, where X and U independent standard normally distributed, and

$$f(x) = (|x| - 1/4)^3 (I(x > 1/4) - I(x < -1/4)).$$

The reason for this choice of $f(x)$ is to check whether the cosine series expansion is able to capture the horizontal part of $f(x)$ for $|x| \leq 1/4$.

The following three figures compare $f(x)$ with the SNP-OLS estimates $\hat{f}_n(x)$ of $f_n(x)$ for $n = 4, 8, 12$ and $x \in [-2, 2]$, with G the Logistic distribution function, on the basis of a random sample of size 500 from (Y, X) .

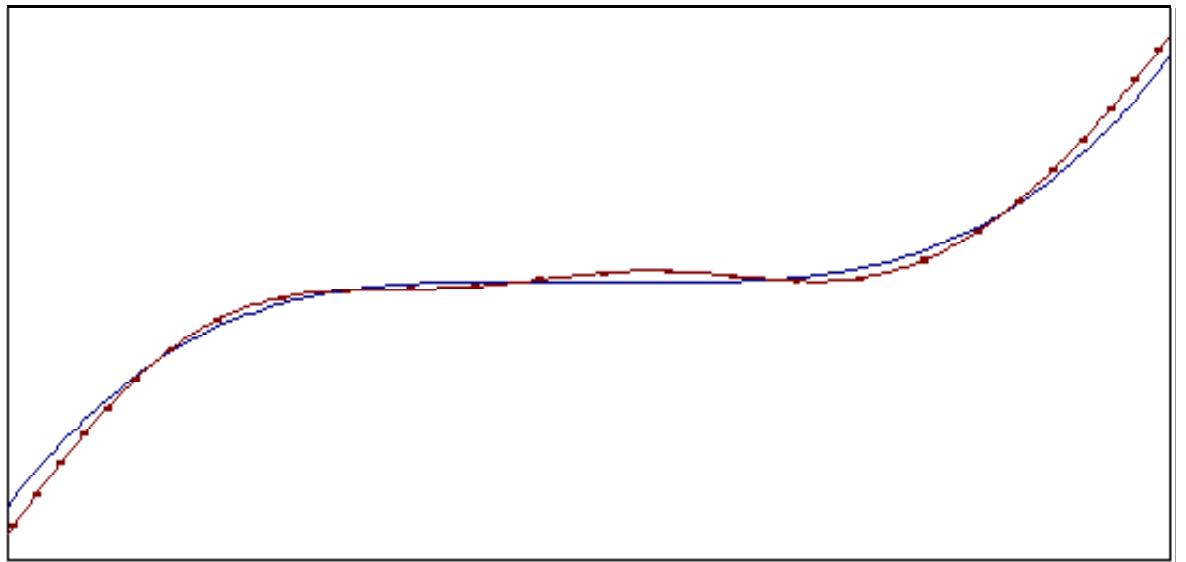


Figure 5.4: $f(x)$ compared with its SNP-OLS estimate $\hat{f}_4(x)$ on $[-2, 2]$

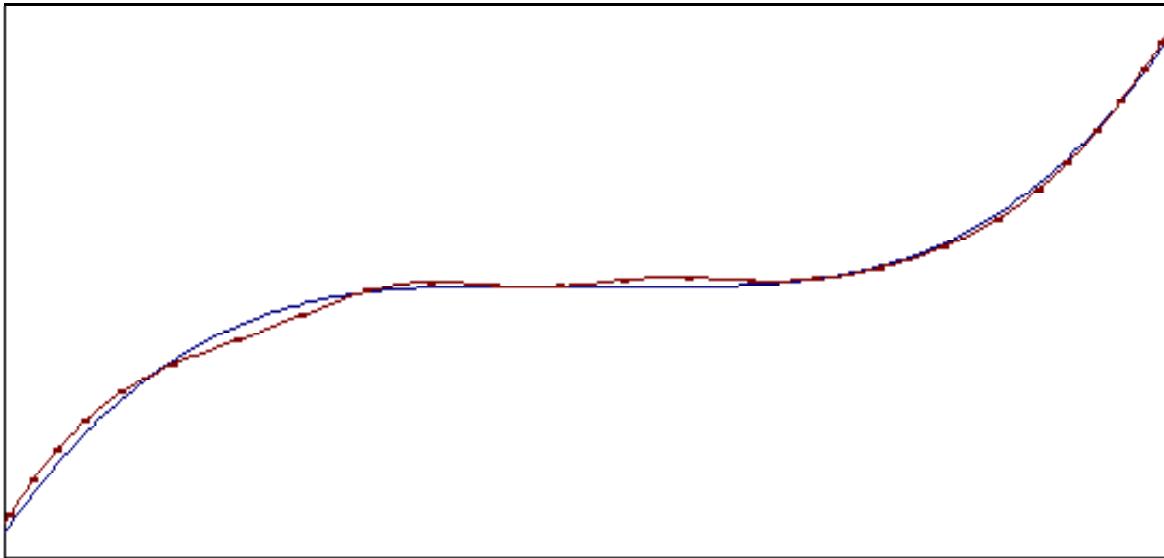


Figure 5.5: $f(x)$ compared with its SNP-OLS estimate $\hat{f}_8(x)$ on $[-2, 2]$

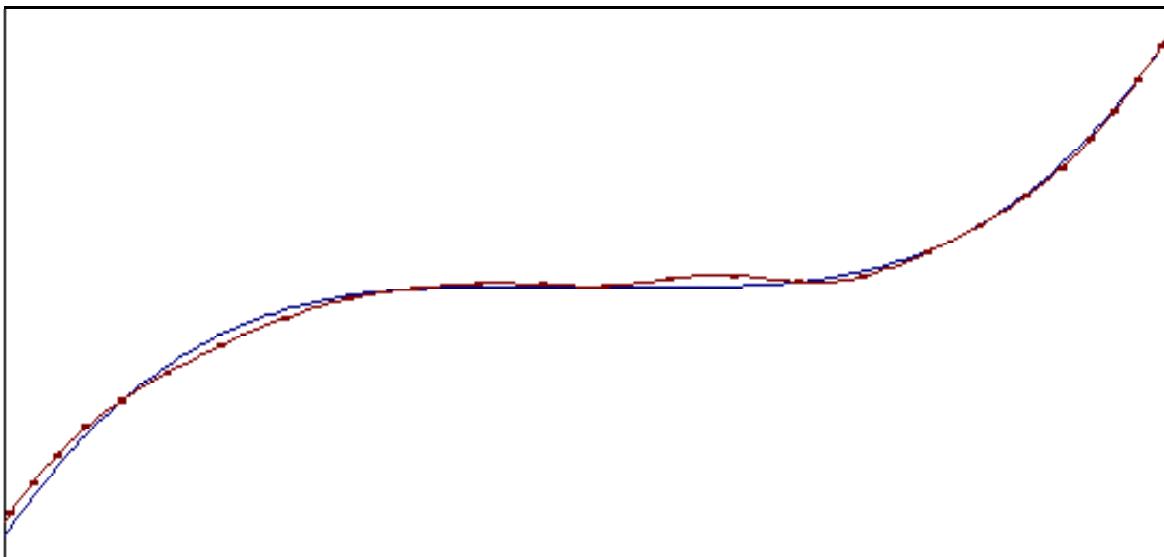


Figure 5.6: $f(x)$ compared with its SNP-OLS estimate $\hat{f}_{12}(x)$ on $[-2, 2]$

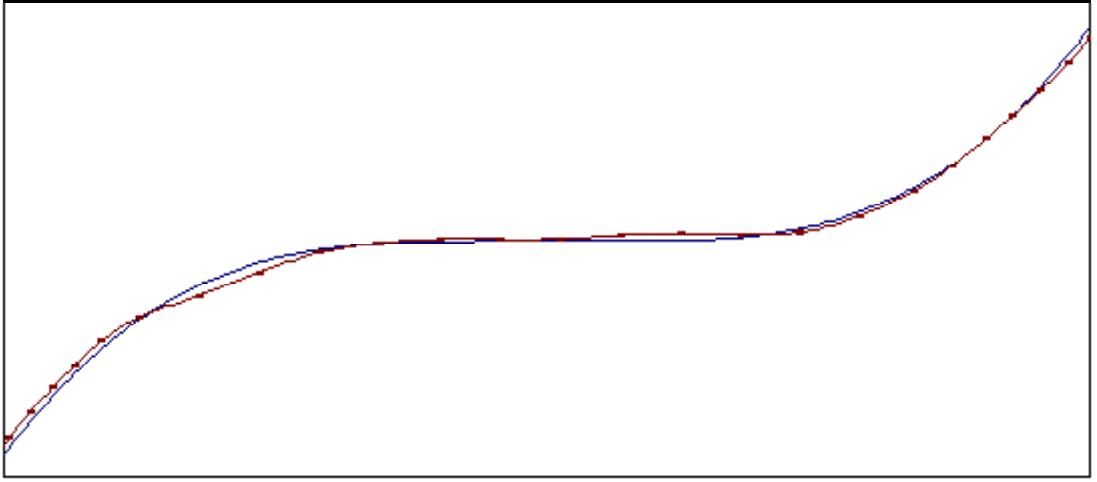


Figure 5.7: Comparison of $f(x)$ with its nonparametric kernel regression estimate $\tilde{f}(x)$ on $[-2, 2]$

As a comparison I have also estimated $f(x)$ by nonparametric kernel regression, similar to Bierens and Pott-Buter (1990), with standard normal kernel and bandwidth constant determined by in-sample leaving-one-out cross-validation over the interval $[0.1, 2]$. The result for $x \in [-2, 2]$ is displayed in Figure 5.7.

As to the SNP results, note the slight wiggle of $\hat{f}_n(x)$ in the flat area $|x| < 1/4$, whereas the nonparametric kernel regression estimator $\tilde{f}(x)$ is smoother in this area. However, in view of the fact that this flat part of $f(x)$ has been approximated via a linear combination of cosine functions the SNP approach works better than I expected.

5.5 Appendix: Proof of Theorem 5.3

The orthonormality of the sequence $\{\overline{\varphi}_n\}_{n=0}^{\infty}$ is easy to verify. The completeness proof employs the following steps.

Step 1. Let $C_0[0, 1]$ be the space of continuous functions $f(u)$ on $[0, 1]$ satisfying $\int_0^1 f(u)du = 0$, endowed with the $L^2(0, 1)$ topology, and let $C_{0,1}[0, 1]$ be the space of continuously differentiable functions $F(u)$ on $[0, 1]$ satisfying $F(0) = F(1) = 0$, also endowed with the $L^2(0, 1)$ topology. Note that the

functions in $C_{0,1}[0, 1]$ take the form $F(u) = \int_0^u f(x) dx$ with $f(u) = F'(u)$. It will be shown that $C_{0,1}[0, 1] \subset \text{span}(\{\bar{\varphi}_n\}_{n=0}^\infty)$.

Step 2. It will be shown that $C_0[0, 1]$ is the closure of $C_{0,1}[0, 1]$, hence $C_0[0, 1] \subset \text{span}(\{\bar{\varphi}_n\}_{n=0}^\infty)$. It follows then trivially that the space $C[0, 1]$ of continuous functions on $[0, 1]$ is contained in $\text{span}(\{\bar{\varphi}_n\}_{n=0}^\infty)$.

Step 3. Finally, it will be shown that every function in $L^2(0, 1)$ can be written as a limit of a sequence of continuous functions, hence $L^2(0, 1)$ is the closure of $C[0, 1]$, so that $L^2(0, 1) = \text{span}(\{\bar{\varphi}_n\}_{n=0}^\infty)$.

Proof of Step 1

Let $f_n(u)$ and $f(u)$ be the same as in Theorem 5.1, except that due to the condition $\int_0^1 f(u) du = 0$, $\alpha_0 = 0$, and let $F_n(u) = \int_0^u f_n(x) dx$. Then

$$\begin{aligned} F_n(u) &= \sum_{k=1}^n \frac{\alpha_k}{k\pi} \sqrt{2} \sin(k\pi u) \\ &= \sum_{k=1}^{[(n+1)/2]} \frac{\alpha_{2k-1}}{(2k-1)\pi} \sqrt{2} \sin((2k-1)\pi u) + \sum_{k=1}^{[n/2]} \frac{\alpha_{2k}}{2k\pi} \sqrt{2} \sin(2k\pi u) \end{aligned}$$

and

$$\begin{aligned} \sup_{0 \leq u \leq 1} |F(u) - F_n(u)| &\leq \int_0^1 |f(x) - f_n(x)| dx \\ &\leq \sqrt{\int_0^1 (f(x) - f_n(x))^2 dx} = o(1) \quad (5.12) \end{aligned}$$

Next, observe that

$$\begin{aligned} \int_0^1 \sqrt{2} \sin((2k-1)\pi u) du &= \frac{-2\sqrt{2}}{(2k-1)\pi} = \gamma_{0,k} \\ \int_0^1 \sqrt{2} \sin((2k-1)\pi u) \sqrt{2} \cos((2m-1)\pi u) du &= 0 \\ \int_0^1 \sqrt{2} \sin((2k-1)\pi u) \sqrt{2} \cos(2m\pi u) du \\ &= \frac{-2}{(2(k+m)-1)\pi} + \frac{-2}{(2(k-m)-1)\pi} \\ &= -\frac{2}{\pi} \frac{4k-2}{(2(k+m)-1)(2(k-m)-1)} \end{aligned}$$

$$= -\frac{2}{\pi} \frac{k-1/2}{(k-1/2)^2 - m^2} = \gamma_{m,k}$$

Hence

$$\sqrt{2} \sin((2k-1)\pi u) = \gamma_{0,k} + \sum_{m=1}^{\infty} \gamma_{m,k} \sqrt{2} \cos(2m\pi u)$$

a.e. on $[0, 1]$. Now let

$$\begin{aligned} \tilde{F}_n(u) &= \sum_{k=1}^{[(n+1)/2]} \frac{\alpha_{2k-1}}{(2k-1)\pi} \gamma_{0,k} + \sum_{k=1}^{[n/2]} \frac{\alpha_{2k}}{2k\pi} \sqrt{2} \sin(2k\pi u) \\ &\quad + \sum_{m=1}^N \left(\sum_{k=1}^{[(n+1)/2]} \frac{\alpha_{2k-1}}{(2k-1)\pi} \gamma_{m,k} \right) \sqrt{2} \cos(2m\pi u) \\ &= -2\sqrt{2} \sum_{k=1}^{[(n+1)/2]} \frac{\alpha_{2k-1}}{(2k-1)^2 \pi^2} + \sum_{k=1}^{[n/2]} \frac{\alpha_{2k}}{2k\pi} \sqrt{2} \sin(2k\pi u) \\ &\quad - \frac{1}{\pi^2} \sum_{m=1}^N \left(\sum_{k=1}^{[(n+1)/2]} \frac{\alpha_{2k-1}}{(k-1/2)^2 - m^2} \right) \sqrt{2} \cos(2m\pi u) \end{aligned}$$

where $N \geq [(n+1)/2]$. Then

$$\begin{aligned} \int_0^1 (\tilde{F}_n(u) - F_n(u))^2 du &= \frac{1}{\pi^4} \sum_{m=N+1}^{\infty} \left(\sum_{k=1}^{[(n+1)/2]} \frac{\alpha_{2k-1}}{m^2 - (k-1/2)^2} \right)^2 \\ &\leq \frac{1}{\pi^4} \sum_{m=N+1}^{\infty} \left(\sum_{k=1}^{[(n+1)/2]} \frac{|\alpha_{2k-1}|}{m^2 - (k-1/2)^2} \right)^2 \\ &\leq \frac{1}{\pi^4} \sum_{m=N+1}^{\infty} \left(\frac{\sum_{k=1}^{[(n+1)/2]} |\alpha_{2k-1}|}{m^2 - ([n/2])^2} \right)^2 \\ &= \frac{1}{\pi^4} \sum_{m=N+1}^{\infty} \left(\frac{\sum_{k=1}^{[(n+1)/2]} |\alpha_{2k-1}|}{(m - [n/2])(m + [n/2])} \right)^2 \\ &\leq \frac{1}{4\pi^4} \sum_{m=N+1}^{\infty} \left(\frac{\frac{1}{[n/2]} \sum_{k=1}^{[(n+1)/2]} |\alpha_{2k-1}|}{m - [n/2]} \right)^2 \end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{4\pi^4} \sum_{m=[n/2]+1}^{\infty} \left(\frac{\frac{1}{[n/2]} \sum_{k=1}^{[(n+1)/2]} |\alpha_{2k-1}|}{m - [n/2]} \right)^2 \\
&= \frac{1}{4\pi^4} \left(\sum_{m=1}^{\infty} \frac{1}{m^2} \right) \left(\frac{1}{[n/2]} \sum_{k=1}^{[(n+1)/2]} |\alpha_{2k-1}| \right)^2 \\
&\leq \frac{1}{4\pi^4} \left(\sum_{m=1}^{\infty} \frac{1}{m^2} \right) \frac{1}{[n/2]} \sum_{k=1}^{\infty} \alpha_{2k-1}^2 \\
&= O(1/n)
\end{aligned} \tag{5.13}$$

Hence by (5.12) and (5.13),

$$\lim_{n \rightarrow \infty} \int_0^1 (\tilde{F}_n(u) - F(u))^2 du = 0$$

Since $\tilde{F}_n \in \text{span}(\{\bar{\varphi}_n\}_{n=0}^{\infty})$ it follows that $F \in \text{span}(\{\bar{\varphi}_n\}_{n=0}^{\infty})$, hence $C_{0,1}[0, 1] \subset \text{span}(\{\bar{\varphi}_n\}_{n=0}^{\infty})$.

Proof of Step 2

Choose an arbitrary function $f \in C_0[0, 1]$, and extend $f(x)$ for $x > 1$ as $f(x) = f(1)$. Let $F(u) = \int_0^u f(x) dx$ and

$$f_n(u) = \frac{(F(u + n^{-1}) - F(u))}{n^{-1}} = \frac{1}{n} \int_u^{u+1/n} f(x) dx$$

Then by continuity

$$\lim_{n \rightarrow \infty} |f_n(u) - f(u)| \leq \lim_{n \rightarrow \infty} \sup_{u \leq x \leq u+1/n} |f(x) - f(u)| = 0$$

pointwise in $u \in [0, 1]$. Moreover,

$$\sup_{0 \leq u \leq 1} |f_n(u) - f(u)| \leq 2 \sup_{0 \leq u \leq 1} |f(u)| < \infty$$

Therefore it follows by bounded convergence that

$$\lim_{n \rightarrow \infty} \int_0^1 (f_n(u) - f(u))^2 du = 0$$

Since $f_n \in C_{0,1}[0, 1] \subset \text{span}(\{\bar{\varphi}_n\}_{n=0}^\infty)$ it follows now that $C_0[0, 1] \subset \text{span}(\{\bar{\varphi}_n\}_{n=0}^\infty)$.

Because the functions in $C[0, 1]$ differ from the functions in $C_0[0, 1]$ by constants only, it follows that $C[0, 1] \subset \text{span}(\{\bar{\varphi}_n\}_{n=0}^\infty)$.

Proof of Step 3

Let B be an arbitrary Borel subset of $[0, 1]$ and let

$$f_n(u) = \exp\left(-n^{-1} \inf_{x \in \overline{B}} |x - u|\right) - \exp\left(-n^{-1} \inf_{x \in \overline{B} \setminus B} |x - u|\right),$$

where \overline{B} is the closure of B . This function is continuous on $[0, 1]$. To see this, note that for $u_1, u_2 \in [0, 1]$,

$$\begin{aligned} \inf_{x \in B} |x - u_1| &\leq |u_2 - u_1| + \inf_{x \in B} |x - u_2| \\ \inf_{x \in \overline{B} \setminus B} |x - u_2| &\leq |u_2 - u_1| + \inf_{x \in \overline{B} \setminus B} |x - u_1| \end{aligned}$$

hence

$$\left| \inf_{x \in \overline{B}} |x - u_2| - \inf_{x \in \overline{B}} |x - u_1| \right| \leq |u_2 - u_1|$$

and similarly,

$$\left| \inf_{x \in \overline{B} \setminus B} |x - u_2| - \inf_{x \in \overline{B} \setminus B} |x - u_1| \right| \leq |u_2 - u_1|$$

For $u \in B$, $\inf_{x \in \overline{B}} |x - u| = 0$ and $\inf_{x \in \overline{B} \setminus B} |x - u| > 0$, hence $\lim_{n \rightarrow \infty} f_n(u) = 1$. For $u \in B \setminus \overline{B}$, $\inf_{x \in \overline{B}} |x - u| = 0$ and $\inf_{x \in \overline{B} \setminus B} |x - u| = 0$, hence $f_n(u) = 0$, and for $u \in [0, 1] \setminus \overline{B}$, $\inf_{x \in \overline{B}} |x - u| > 0$ and $\inf_{x \in \overline{B} \setminus B} |x - u| > 0$, hence $\lim_{n \rightarrow \infty} f_n(u) = 0$. Thus

$$\lim_{n \rightarrow \infty} f_n(u) = I(x \in B).$$

Since $f_n(u) \in C[0, 1] \subset \text{span}(\{\bar{\varphi}_n\}_{n=0}^\infty)$ it follows now that for arbitrary Borel sets B , $I(x \in B) \in \text{span}(\{\bar{\varphi}_n\}_{n=0}^\infty)$ and so are all simple functions on $[0, 1]$. Because functions are Borel measurable if and only if they are limits of sequences of simple functions, it follows that $L^2(0, 1) = \text{span}(\{\bar{\varphi}_n\}_{n=0}^\infty)$.

■

Chapter 6

Density and distribution functions

6.1 Density functions on the unit interval

It follows from Theorem 5.1 that for any density function $h(u)$ on $[0, 1]$ there exists a sequence $\{\alpha_k\}_{k=0}^{\infty}$ satisfying $\sum_{k=0}^{\infty} \alpha_k^2 = 1$ such that

$$h(u) = \left(\alpha_0 + \sum_{k=1}^{\infty} \alpha_k \sqrt{2} \cos(k\pi u) \right)^2 \text{ a.e. on } (0, 1). \quad (6.1)$$

The square guarantees that $h(u) \geq 0$. Gallant and Nychka (1987) proposed a similar series expansion on the basis of Hermite polynomials.

Note that the α_k 's in (6.1) are no longer unique. For example, we can always write $h(u) = f_B(u)^2$, where for an arbitrary Borel set B in $[0, 1]$,

$$f_B(u) = (I(u \in B) - I(u \notin B)) \sqrt{h(u)}. \quad (6.2)$$

Then the α_k 's in (6.1) take the form

$$\alpha_k = \int_B \sqrt{h(u)} \kappa_k(u) du - \int_{[0,1] \setminus B} \sqrt{h(u)} \kappa_k(u) du$$

In particular, we may choose for α_0 any

$$\alpha_0 \in \left[- \int_0^1 \sqrt{h(u)} du, \int_0^1 \sqrt{h(u)} du \right]. \quad (6.3)$$

If we choose $\alpha_0 \in \left(0, \int_0^1 \sqrt{h(u)} du\right]$ then we can reparametrize the Fourier coefficients α_k as

$$\begin{aligned}\alpha_0 &= \frac{1}{\sqrt{1 + \sum_{m=1}^{\infty} \delta_m^2}} \\ \alpha_k &= \frac{\delta_k}{\sqrt{1 + \sum_{m=1}^{\infty} \delta_m^2}}, \quad k \in \mathbb{N},\end{aligned}$$

where $\sum_{m=0}^{\infty} \delta_m^2 < \infty$. Hence,

Theorem 6.1. *For any density function $h(u)$ on $[0, 1]$ there exist possibly uncountable many sequences $\{\delta_m\}_{m=1}^{\infty}$ satisfying $\sum_{m=0}^{\infty} \delta_m^2 < \infty$ such that*

$$h(u) = \frac{(1 + \sum_{k=1}^{\infty} \delta_k \sqrt{2} \cos(k\pi u))^2}{1 + \sum_{m=1}^{\infty} \delta_m^2} \text{ a.e. on } (0, 1). \quad (6.4)$$

In particular, (6.4) holds for all sequences δ_k of the form

$$\delta_k = \frac{\int_0^1 (I(u \in B) - I(u \notin B)) \sqrt{h(u)} \sqrt{2} \cos(k\pi u) du}{\int_0^1 (I(u \in B) - I(u \notin B)) \sqrt{h(u)} du}, \quad (6.5)$$

where B is any Borel set in $[0, 1]$ satisfying

$$\int_0^1 (I(u \in B) - I(u \notin B)) \sqrt{h(u)} du > 0.$$

Moreover, the corresponding SNP densities

$$h_n(u) = \frac{(1 + \sqrt{2} \sum_{k=1}^n \delta_k \cos(k\pi u))^2}{1 + \sum_{m=1}^n \delta_m^2} \quad (6.6)$$

satisfy

$$\int_0^1 |h(u) - h_n(u)| du \leq \sqrt{5 \sum_{k=n+1}^{\infty} \delta_k^2} \rightarrow 0 \quad (6.7)$$

Furthermore, the corresponding SNP distribution functions have the closed form expressions

$$H_n(u) = u$$

$$\begin{aligned}
& + \frac{1}{1 + \sum_{m=1}^n \delta_m^2} \left[2\sqrt{2} \sum_{k=1}^n \delta_k \frac{\sin(k\pi u)}{k\pi} + \sum_{m=1}^n \delta_m^2 \frac{\sin(2m\pi u)}{2m\pi} \right. \\
& \left. + 2 \sum_{k=2}^n \sum_{m=1}^{k-1} \delta_k \delta_m \frac{\sin((k+m)\pi u)}{(k+m)\pi} + 2 \sum_{k=2}^n \sum_{m=1}^{k-1} \delta_k \delta_m \frac{\sin((k-m)\pi u)}{(k-m)\pi} \right], \tag{6.8}
\end{aligned}$$

and satisfy

$$\sup_{0 \leq u \leq 1} |H(u) - H_n(u)| \leq \sqrt{5 \sum_{k=n+1}^{\infty} \delta_k^2} \rightarrow 0. \tag{6.9}$$

6.2 Uniqueness of the series representation

The density $h(u)$ in Theorem 6.1 can be written as $h(u) = \eta(u)^2 / \int_0^1 \eta(v)^2 dv$, where

$$\eta(u) = 1 + \sum_{m=1}^{\infty} \delta_m \sqrt{2} \cos(m\pi u) \text{ a.e. on } (0, 1). \tag{6.10}$$

Moreover, recall that in general,

$$\begin{aligned}
\delta_m &= \frac{\int_0^1 (I(u \in B) - I(u \in [0, 1] \setminus B)) \sqrt{2} \cos(m\pi u) \sqrt{h(u)} du}{\int_0^1 (I(u \in B) - I(u \in [0, 1] \setminus B)) \sqrt{h(u)} du}, \\
\frac{1}{\sqrt{1 + \sum_{m=1}^{\infty} \delta_m^2}} &= \int_0^1 (I(u \in B) - I(u \in [0, 1] \setminus B)) \sqrt{h(u)} du.
\end{aligned}$$

for some Borel set B satisfying $\int_0^1 (I(u \in B) - I(u \in [0, 1] \setminus B)) \sqrt{h(u)} du > 0$, hence

$$\eta(u) = (I(u \in B) - I(u \in [0, 1] \setminus B)) \sqrt{h(u)} \sqrt{1 + \sum_{m=1}^{\infty} \delta_m^2} \tag{6.11}$$

Similarly, given this Borel set B and the corresponding δ_m 's, the SNP density (6.6) can be written as $h_n(u) = \eta_n(u)^2 / \int_0^1 \eta_n(v)^2 dv$, where

$$\eta_n(u) = 1 + \sum_{m=1}^n \delta_m \sqrt{2} \cos(m\pi u)$$

$$= (I(u \in B) - I(u \in [0, 1] \setminus B)) \sqrt{h_n(u)} \sqrt{1 + \sum_{m=1}^n \delta_m^2} \quad (6.12)$$

Now suppose that $h(u)$ is continuous and positive on $(0, 1)$. Moreover, let $S \subset [0, 1]$ be the set with Lebesgue measure zero on which $h(u) = \lim_{n \rightarrow \infty} h_n(u)$ fails to hold. Then for any $u_0 \in (0, 1) \setminus S$, $\lim_{n \rightarrow \infty} h_n(u_0) = h(u_0) > 0$, hence for sufficient large n , $h_n(u_0) > 0$. Because obviously $h_n(u)$ and $\eta_n(u)$ are continuous on $(0, 1)$, for such an n there exists a small $\varepsilon_n(u_0) > 0$ such that $h_n(u) > 0$ for all $u \in (u_0 - \varepsilon_n(u_0), u_0 + \varepsilon_n(u_0)) \cap (0, 1)$, and therefore

$$I(u \in B) - I(u \in [0, 1] \setminus B) = \frac{\eta_n(u)}{\sqrt{h_n(u)} \sqrt{1 + \sum_{m=1}^n \delta_m^2}} \quad (6.13)$$

is continuous on $(u_0 - \varepsilon_n(u_0), u_0 + \varepsilon_n(u_0)) \cap (0, 1)$. Substituting (6.13) in (6.11) it follows now that $\eta(u)$ is continuous on $(u_0 - \varepsilon_n(u_0), u_0 + \varepsilon_n(u_0)) \cap (0, 1)$, hence by the arbitrariness of $u_0 \in (0, 1) \setminus S$, $\eta(u)$ is continuous on $(0, 1)$.

Next, suppose that $\eta(u)$ takes positive and negative values on $(0, 1)$. Then by the continuity of $\eta(u)$ on $(0, 1)$ there exists a $u_0 \in (0, 1)$ for which $\eta(u_0) = 0$ and thus $h(u_0) = 0$, which however is excluded by the condition that $h(u) > 0$ on $(0, 1)$. Therefore, either $\eta(u) > 0$ for all $u \in (0, 1)$ or $\eta(u) < 0$ for all $u \in (0, 1)$. However, the latter is excluded because by (6.10), $\int_0^1 \eta(u) du = 1$. Thus, $\eta(u) > 0$ on $(0, 1)$, so that by (6.11), $I(u \in B) - I(u \in [0, 1] \setminus B) = 1$ on $(0, 1)$.

Consequently,

Theorem 6.2. *For every continuous and positive valued density $h(u)$ on $(0, 1)$ the sequence $\{\delta_m\}_{m=1}^\infty$ in Theorem 6.1 is unique, with*

$$\delta_m = \frac{\int_0^1 \sqrt{2} \cos(m\pi u) \sqrt{h(u)} du}{\int_0^1 \sqrt{h(u)} du}.$$

6.3 General representation

Given a continuous distribution function $G(x)$ with support $\Xi \subset \mathbb{R}$, any distribution function $F(x)$ with support contained in Ξ can be written as

$F(x) = H(G(x))$, where $H(u) = F(G^{-1}(u))$ is a distribution function on $[0, 1]$. Moreover, if F and G are absolutely continuous with densities f and g , respectively, then H is absolutely continuous with density $h(u)$, and $f(x) = h(G(x))g(x)$. Therefore, $f(x)$ can be estimated semiparametrically by estimating $h(u)$ semiparametrically.

In general, the role of the a priori chosen distribution function G is three-fold:

1. G specifies the support of the unknown distribution functions F in the semi-nonparametric model;
2. G maps the parameter space \mathcal{F} of candidate distributions for F one-to-one onto a space $\mathcal{H}(0, 1)$ of distribution functions on the unit interval, which enables us to develop a unified inference approach for a wide range of semi-nonparametric models;
3. G serves as an initial guess for $F(x) = H(G(x))$. If the guess is right then $H(u) = u$. A related interpretation of G is that it serves as a (non-Bayesian) "prior" for F , with the estimate \hat{H} of H playing the role of correction mechanism which converts the prior G into a "posterior" \hat{F} for F on the basis of data evidence. Another related interpretation is that $F = G$ represents a standard parametric model for which the semi-nonparametric model is a generalization.

It follows now from Theorem 6.1 and (6.22) that

Theorem 6.3. *Given an absolutely continuous distribution function $G(x)$ on \mathbb{R} with density $g(x)$, any density function $f(x)$ with support contained in the support of g (i.e., $\{x : f(x) > 0\} \subset \{x : g(x) > 0\}$) can be written as*

$$f(x) = g(x) \frac{\left(1 + \sqrt{2} \sum_{k=1}^{\infty} \delta_k \cos(k\pi G(x))\right)^2}{1 + \sum_{m=1}^{\infty} \delta_m^2} \quad (6.14)$$

a.e. on $\{x : f(x) > 0\}$. Moreover, the corresponding SNP densities

$$f_n(x) = g(x) \frac{\left(1 + \sqrt{2} \sum_{k=1}^n \delta_k \cos(k\pi G(x))\right)^2}{1 + \sum_{m=1}^n \delta_m^2} \quad (6.15)$$

satisfy

$$\int_{-\infty}^{\infty} |f(x) - f_n(x)| dx \leq \sqrt{5 \sum_{k=n+1}^{\infty} \delta_k^2} \rightarrow 0.$$

Furthermore, the SNP distribution function $F_n(x) = H_n(G(x))$ satisfies

$$\sup_x |F(x) - F_n(x)| \leq \sqrt{5 \sum_{k=n+1}^{\infty} \delta_k^2} \rightarrow 0,$$

where $F(x) = \int_{-\infty}^x f(z) dz$ and $H_n(u)$ is defined by (6.8).

6.4 Smoothness

The non-Euclidean parameter of a semi-nonparametric econometric model often takes the form of a density function $f(x)$. Usually it is assumed that $f(x)$ has certain smoothness and regularity properties, like boundedness, continuity and differentiability. Also, usually the semiparametric model involved requires that the support of $f(x)$ is connected, i.e.,

$$\{x \in \mathbb{R} : f(x) > 0\} = (a, b),$$

where possibly $a = -\infty$ and/or $b = \infty$. To impose these conditions, we need to impose corresponding smoothness and regularity conditions on the density $g(x)$ of the a priori chosen distribution function $G(x)$ and on the density $h(u)$ in the transformation $f(x) = h(G(x))g(x)$.

Denoting $u = G(x)$, we can write

$$h(u) = \frac{f(G^{-1}(u))}{g(G^{-1}(u))}$$

Given that G is chosen such that f and g have the same support (a, b) , it follows that $h(u)$ must have support $(0, 1)$, i.e.,

$$h(u) > 0 \text{ on } (0, 1), \tag{6.16}$$

and if f and g are continuous on (a, b) then $h(u)$ is continuous on $(0, 1)$. Moreover, if it is known that $f(x) < \infty$ for each $x \in (a, b)$, then $f(x)/g(x) < \infty$ for each $x \in (a, b)$, hence $h(u) < \infty$ for each $u \in (0, 1)$. Furthermore,

since $g(x)$ is an initial guess of $f(x)$, it is reasonable to assume that $g(x)$ is sufficiently close to $f(x)$ to guarantee that

$$\lim_{x \downarrow a} f(x)/g(x) < \infty, \quad \lim_{x \uparrow b} f(x)/g(x) < \infty$$

These tail conditions, together with the condition that $f(x) < \infty$ for each $x \in (a, b)$, are equivalent to $\sup_{0 \leq u \leq 1} h(u) < \infty$. A sufficient condition for the latter is that the δ_k 's in (6.4) satisfy

$$\sum_{k=1}^{\infty} |\delta_k| < \infty. \quad (6.17)$$

This condition is stronger than necessary for $\sup_{0 \leq u \leq 1} h(u) < \infty$ only, because:

Theorem 6.4. *Condition (6.17) implies that $\sum_{k=1}^{\infty} \delta_k \sqrt{2} \cos(k\pi u)$ is uniformly continuous on $[0, 1]$, hence the corresponding density function $h(u)$ in (6.4) is then uniformly continuous on $[0, 1]$.*¹

Note that Theorems 6.2 and 6.4 imply the following corollary.

Theorem 6.5. *Suppose that $h(u)$ has support $(0, 1)$. If the δ_k 's in (6.4) are confined to those for which $\sum_{k=1}^{\infty} |\delta_k| < \infty$ then they are unique.*

Next, suppose that f and g are continuously differentiable on (a, b) . Then $h(u)$ is continuously differentiable on $(0, 1)$. A sufficient condition for the latter is that

$$\sum_{k=1}^{\infty} k |\delta_k| < \infty. \quad (6.18)$$

To see this, pick any $u \in [0, 1]$ and let $\varepsilon \neq 0$ be so small that $u + \varepsilon \in [0, 1]$. Then by the mean value theorem there exists a sequence $\lambda_k(u, \varepsilon) \in [0, 1]$ such that

$$\limsup_{\varepsilon \rightarrow 0} \left| \frac{1}{\varepsilon} \sum_{k=1}^{\infty} \delta_k (\cos(k\pi(u + \varepsilon)) - \cos(k\pi u)) + \pi \sum_{k=1}^{\infty} k \delta_k \sin(k\pi u) \right|$$

¹Which implies that $\sup_{0 \leq u \leq 1} h(u) < \infty$.

$$\begin{aligned}
&\leq \pi \limsup_{\varepsilon \rightarrow 0} \sum_{k=1}^{\infty} k |\delta_k| \cdot |\sin(k\pi u) - \sin(k\pi(u + \lambda_k(u, \varepsilon)\varepsilon))| \\
&\leq \pi \limsup_{\varepsilon \rightarrow 0} \sum_{k=1}^n k |\delta_k| \cdot |\sin(k\pi u) - \sin(k\pi(u + \lambda_k(u, \varepsilon)\varepsilon))| + 2\pi \sum_{k=n+1}^{\infty} k |\delta_k| \\
&= 2\pi \sum_{k=n+1}^{\infty} k |\delta_k| \rightarrow 0 \text{ as } n \rightarrow \infty,
\end{aligned}$$

hence

$$\frac{d}{du} \left(\sum_{k=1}^{\infty} \delta_k \cos(k\pi u) \right) = \sum_{k=1}^{\infty} \delta_k \frac{d \cos(k\pi u)}{du} = -\pi \sum_{k=1}^{\infty} k \delta_k \sin(k\pi u)$$

and thus

$$h'(u) = -\frac{2\pi \left(1 + \sqrt{2} \sum_{k=1}^{\infty} \delta_k \sqrt{2} \cos(k\pi u) \right) \left(\sum_{k=1}^{\infty} k \delta_k \sqrt{2} \sin(k\pi u) \right)}{1 + \sum_{i=1}^{\infty} \delta_i^2}.$$

Note that $h'(0) = h'(1) = 0$. Moreover, it follows similar to Theorem 6.4 that $h'(u)$ is uniformly continuous on $[0, 1]$.

Along the same lines it can be shown that

Theorem 6.6. *If for some natural number $\ell \geq 1$, $\sum_{k=1}^{\infty} k^{\ell} |\delta_k| < \infty$, then the density function $h(u)$ in (6.4) is ℓ -times continuously differentiable on $[0, 1]$.*

6.5 Bivariate densities

Similar to (4.30) and (6.1), any bivariate density $h(u, v)$ on $[0, 1] \times [0, 1]$ can be written as

$$\begin{aligned}
h(u, v) &= \left(\alpha_{0,0} + \sum_{k=1}^{\infty} \alpha_{k,0} \sqrt{2} \cos(k\pi u) + \sum_{m=1}^{\infty} \alpha_{0,m} \sqrt{2} \cos(m\pi v) \right. \\
&\quad \left. + \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \alpha_{k,m} \sqrt{2} \cos(k\pi u) \sqrt{2} \cos(m\pi v) \right)^2
\end{aligned}$$

a.e. on $[0, 1] \times [0, 1]$, where $\sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \alpha_{k,m} = 1$, and similar to (6.4) we can reparametrize the $\alpha_{k,m}$'s such that $h(u, v)$ becomes

$$h(u, v) = \frac{1}{1 + \sum_{k=1}^{\infty} \delta_{k,0}^2 + \sum_{m=1}^{\infty} \delta_{0,m}^2 + \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \delta_{k,m}^2}$$

$$\begin{aligned} & \times \left(1 + \sum_{k=1}^{\infty} \delta_{k,0} \sqrt{2} \cos(k\pi u) + \sum_{m=1}^{\infty} \delta_{0,m} \sqrt{2} \cos(m\pi v) \right. \\ & \quad \left. + \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \delta_{k,m} \sqrt{2} \cos(k\pi u) \sqrt{2} \cos(m\pi v) \right)^2 \end{aligned}$$

a.e. on $[0, 1] \times [0, 1]$, where

$$\sum_{k=1}^{\infty} \delta_{k,0}^2 + \sum_{m=1}^{\infty} \delta_{0,m}^2 + \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \delta_{k,m}^2 < \infty. \quad (6.19)$$

Note that the marginal densities of $h(u, v)$ take the form of a weighted sum of univariate densities. In particular, denoting

$$\begin{aligned} h_1(u) &= \int_0^1 h(u, v) dv \\ h_{1,0}(u) &= \frac{1 + \sqrt{2} \sum_{k=1}^{\infty} \delta_{k,0} \cos(k\pi u)}{1 + \sum_{k=1}^{\infty} \delta_{k,0}^2} \\ h_{1,m}(u) &= \frac{(\delta_{0,m} + \sum_{k=1}^{\infty} \delta_{k,m} \sqrt{2} \cos(k\pi u))^2}{\sum_{k=0}^{\infty} \delta_{k,m}^2} \end{aligned}$$

it can be shown that

$$h_1(u) = \frac{(1 + \sum_{k=1}^{\infty} \delta_{k,0}^2) h_{1,0}(u) + \sum_{m=1}^{\infty} (\sum_{k=0}^{\infty} \delta_{k,m}^2) h_{1,m}(u)}{1 + \sum_{k=1}^{\infty} \delta_{k,0}^2 + \sum_{m=1}^{\infty} \delta_{0,m}^2 + \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \delta_{k,m}^2}.$$

Of course, the $\delta_{k,m}$'s can be reparametrized such that $h_1(u)$ takes the form (6.4).

Similar to (6.6), let

$$\begin{aligned} h_n(u, v) &= \frac{1}{1 + \sum_{k=1}^n \delta_{k,0}^2 + \sum_{m=1}^n \delta_{0,m}^2 + \sum_{k=1}^n \sum_{m=1}^n \delta_{k,m}^2} \\ &\quad \times \left(1 + \sum_{k=1}^n \delta_{k,0} \sqrt{2} \cos(k\pi u) + \sum_{m=1}^n \delta_{0,m} \sqrt{2} \cos(m\pi v) \right. \\ & \quad \left. + \sum_{k=1}^n \sum_{m=1}^n \delta_{k,m} \sqrt{2} \cos(k\pi u) \sqrt{2} \cos(m\pi v) \right)^2 \end{aligned}$$

be the truncated version of $h(u, v)$. It is not hard to verify that

$$\lim_{n \rightarrow \infty} \int_0^1 \int_0^1 |h(u, v) - h_n(u, v)| dudv = 0.$$

Finally, note that any density $f(x, y)$ with support $\Xi_x \times \Xi_y \subset \mathbb{R}^2$ can be represented by

$$f(x, y) = g_x(x)g_y(y)h(G_x(x), G_y(y)) \text{ a.e. on } \Xi_x \times \Xi_y,$$

with truncated version

$$f_n(x, y) = g_x(x)g_y(y)h_n(G_x(x), G_y(y)),$$

where G_x is an a priori chosen absolutely continuous distribution function with density g_x and support Ξ_x , and G_y is an a priori chosen absolutely continuous distribution function with density g_y and support Ξ_y .

6.6 Appendix: Proofs

6.6.1 Theorem 6.1

The result (6.8) follows from the well-known sine-cosine formulas. To prove (6.7), denote

$$\begin{aligned} f(u) &= \frac{1 + \sum_{k=1}^{\infty} \delta_k \sqrt{2} \cos(k\pi u)}{\sqrt{1 + \sum_{m=1}^{\infty} \delta_m^2}}, \\ f_n(u) &= \frac{1 + \sum_{k=1}^n \delta_k \sqrt{2} \cos(k\pi u)}{\sqrt{1 + \sum_{m=1}^n \delta_m^2}} \end{aligned}$$

It follows from the Cauchy-Schwarz inequality that

$$\begin{aligned} \int_0^1 |f(u)^2 - f_n(u)^2| du &= \int_0^1 |f(u) - f_n(u)| \cdot |f(u) + f_n(u)| du \\ &\leq 2 \sqrt{\int_0^1 (f(u) - f_n(u))^2 du} \end{aligned} \tag{6.20}$$

Moreover,

$$\begin{aligned}
& \int_0^1 (f(u) - f_n(u))^2 du \\
&= \left(1 + \sum_{k=1}^n \delta_k^2 \right) \left(\frac{1}{\sqrt{1 + \sum_{m=1}^{\infty} \delta_m^2}} - \frac{1}{\sqrt{1 + \sum_{m=1}^n \delta_m^2}} \right)^2 \\
&+ \frac{\sum_{k=n+1}^{\infty} \delta_k^2}{1 + \sum_{m=1}^{\infty} \delta_m^2} \leq \frac{5}{4} \sum_{k=n+1}^{\infty} \delta_k^2
\end{aligned} \tag{6.21}$$

as is not hard to verify. The result (6.7) now follows from (6.20) and (6.21). Finally, (6.9) follows from

$$\sup_{0 \leq u \leq 1} |H(u) - H_n(u)| \leq \int_0^1 |h(x) - h_n(x)| dx \leq \sqrt{5 \sum_{k=n+1}^{\infty} \delta_k^2} \rightarrow 0. \tag{6.22}$$

6.6.2 Theorem 6.4

Pick any $u \in [0, 1]$ and let $\varepsilon \neq 0$ be so small that $u + \varepsilon \in [0, 1]$. Then

$$\begin{aligned}
& \limsup_{\varepsilon \rightarrow 0} \left| \sum_{k=1}^{\infty} \delta_k (\cos(k\pi(u + \varepsilon)) - \cos(k\pi u)) \right| \\
&= \limsup_{\varepsilon \rightarrow 0} \left| \sum_{k=1}^{\infty} \delta_k ((\cos(k\pi\varepsilon) - 1) \cos(k\pi u) - \sin(k\pi\varepsilon) \sin(k\pi u)) \right| \\
&\leq \limsup_{\varepsilon \rightarrow 0} \sum_{k=1}^n |\delta_k| (|1 - \cos(k\pi\varepsilon)| + |\sin(k\pi\varepsilon)|) + 3 \sum_{k=n+1}^{\infty} |\delta_k| \\
&= 3 \sum_{k=n+1}^{\infty} |\delta_k| \rightarrow 0 \text{ as } n \rightarrow \infty.
\end{aligned}$$

By the compactness of $[0, 1]$ this result implies that $\sum_{k=1}^{\infty} \delta_k \cos(k\pi u)$ is uniformly continuous on $[0, 1]$, and so is $h(u)$.

Chapter 7

Compactness

As said before, the non-Euclidean parameter of a semi-nonparametric econometric model often takes the form of a density and/or distribution function. See the next section for an example. Similar to parametric nonlinear estimation, these non-Euclidean parameters need to be confined to a compact metric space. In this chapter it will be shown how to construct such compact metric spaces.

7.1 General density and distribution functions

Recall that the results in Theorem 6.1 read more generally as follows. Given a complete orthonormal sequence $\{\rho_k\}_{k=0}^{\infty}$ in $L^2(0, 1)$ with $\rho_0(u) \equiv 1$, for every density function $h(u)$ on $[0, 1]$ there exist uncountable many sequences $\{\delta_m\}_{m=1}^{\infty}$ satisfying

$$\sum_{m=1}^{\infty} \delta_m^2 < \infty \quad (7.1)$$

such that

$$h(u) = \frac{(1 + \sum_{m=1}^{\infty} \delta_m \rho_m(u))^2}{1 + \sum_{m=1}^{\infty} \delta_m^2} \text{ a.e.} \quad (7.2)$$

Moreover, recall that this representation does not require any smoothness conditions. Thus (7.2) holds if $h(u)$ is merely Borel measurable. Furthermore, denoting

$$h_n(u) = \frac{(1 + \sum_{m=1}^n \delta_m \rho_m(u))^2}{1 + \sum_{m=1}^n \delta_m^2} \quad (7.3)$$

for $n \geq 1$, it follows that

$$\int_0^1 |h(u) - h_n(u)| du \leq \sqrt{5 \sum_{k=n+1}^{\infty} \delta_k^2} \rightarrow 0 \quad (7.4)$$

as $n \rightarrow \infty$.

The condition (7.1) can be imposed by imposing the restrictions $|\delta_k| \leq \bar{\delta}_k$, where $\bar{\delta}_k$ is an a priori chosen positive sequence such that $\sum_{k=1}^{\infty} \bar{\delta}_k^2 < \infty$. For example, let

$$\bar{\delta}_k = \frac{c}{1 + \sqrt{k} \ln(k)}, \quad (7.5)$$

for some constant $c > 0$. It is easy to verify that then $\sum_{k=1}^{\infty} \bar{\delta}_k^2 < c^2 + c^2 / \ln(2)$.

These restrictions on the δ_k 's also play a key-role in proving compactness:

Theorem 7.1. *Let $\mathcal{D}(0, 1)$ be the space of densities of the type (7.2) subject to the restrictions $|\delta_k| \leq \bar{\delta}_k$ for some a priori chosen positive sequence $\bar{\delta}_k$ satisfying $\sum_{k=1}^{\infty} \bar{\delta}_k^2 < \infty$, endowed with the L^1 metric*

$$\|h_1 - h_2\|_1 = \int_0^1 |h_1(u) - h_2(u)| du.$$

Then $\mathcal{D}(0, 1)$ is compact. Consequently, the space

$$\mathcal{H}(0, 1) = \left\{ H(u) = \int_0^u h(v) dv, \ h \in \mathcal{D}(0, 1) \right\}$$

endowed with the "sup" metric

$$\|H_1 - H_2\|_{\sup} = \sup_{0 \leq u \leq 1} |H_1(u) - H_2(u)|$$

is compact as well. Moreover, let $\mathcal{D}_n(0, 1)$ be the space of SNP densities of the type (7.3), with $\mathcal{D}_0(0, 1)$ the singleton $\{h(u) \equiv 1\}$, subject to the same restrictions on the δ_k 's, and endowed with the same metric as $\mathcal{D}(0, 1)$. Then the sequence $\mathcal{D}_n(0, 1)$ is dense in $\mathcal{D}(0, 1)$:

$$\mathcal{D}(0, 1) = \overline{\cup_{n=0}^{\infty} \mathcal{D}_n(0, 1)}.$$

Consequently, the spaces

$$\mathcal{H}_n(0, 1) = \left\{ H_n(u) = \int_0^u h_n(v) dv, \ h_n \in \mathcal{D}_n(0, 1) \right\}$$

endowed with the sup metric are dense in $\mathcal{H}(0, 1)$:

$$\mathcal{H}(0, 1) = \overline{\cup_{n=0}^{\infty} \mathcal{H}_n(0, 1)}.$$

Bierens (2008, Theorem 8) proved this result for the case where the $\rho_m(u)$ are Legendre polynomials and $\bar{\delta}_k$ is given by (7.5). However, as will be shown below these results hold for any complete orthonormal sequence $\rho_n(u)$ and any positive sequence $\bar{\delta}_k$ satisfying $\sum_{k=1}^{\infty} \bar{\delta}_k^2 < \infty$.

Similarly, it is easy to construct compact metric spaces of general density and distribution functions on \mathbb{R} . In particular, recall that any density $f(x)$ with support $\mathbb{X} \subset \mathbb{R}$ can be written as $f(x) = h(G(x))g(x)$, where $G(x)$ is a given absolutely continuous distribution function with density $g(x)$ and support containing \mathbb{X} : $\mathbb{X} \subset \{x \in \mathbb{R} : g(x) > 0\}$. Thus, denoting

$$\begin{aligned}\mathcal{D}(G) &= \{f(x) = h(G(x))g(x) : h \in \mathcal{D}(0, 1)\} \\ \mathcal{F}(G) &= \left\{ F(x) = \int_{-\infty}^x f(z)dz : f \in \mathcal{D}(G) \right\}\end{aligned}$$

it follows trivially from Theorem 7.1 that $\mathcal{D}(G)$ is a compact metric space of densities with metric $\int_{-\infty}^{\infty} |f_1(x) - f_2(x)| dx$, and $\mathcal{F}(G)$ is a compact metric space of absolutely continuous distribution functions with metric $\sup_{x \in \mathbb{R}} |F_1(x) - F_2(x)|$. Moreover, denoting

$$\begin{aligned}\mathcal{D}_n(G) &= \{f(x) = h(G(x))g(x) : h \in \mathcal{D}_n(0, 1)\} \\ \mathcal{F}_n(G) &= \left\{ F(x) = \int_{-\infty}^x f(z)dz : f \in \mathcal{D}_n(G) \right\}\end{aligned}$$

it follows from Theorem 7.1 that $\mathcal{D}(G) = \overline{\cup_{n=0}^{\infty} \mathcal{D}_n(G)}$ and $\mathcal{F}(G) = \overline{\cup_{n=0}^{\infty} \mathcal{F}_n(G)}$.

The compactness part of Theorem 7.1 follows from the following two lemmas and the fact that similar to (6.7), for each pair $h_1, h_2 \in \mathcal{D}(0, 1)$ there exist sequences $\delta_1 = \{\delta_{1,k}\}_{k=1}^{\infty}$ and $\delta_2 = \{\delta_{2,k}\}_{k=1}^{\infty}$ such that

$$\int_0^1 |h_1(u) - h_2(u)| du \leq \sqrt{5} \sqrt{\sum_{k=1}^{\infty} (\delta_{1,k} - \delta_{2,k})^2}.$$

Lemma 7.1. Let $\{\bar{\delta}_k\}_{k=1}^{\infty}$ be an a priori chosen positive sequence satisfying $\sum_{k=1}^{\infty} \bar{\delta}_k^2 < \infty$, and let $\Delta = X_{k=1}^{\infty}[-\bar{\delta}_k, \bar{\delta}_k]$. Endow the space Δ with the metric

$$d(\delta_1, \delta_2) = \sqrt{\sum_{k=1}^{\infty} (\delta_{1,k} - \delta_{2,k})^2},$$

where $\delta_1 = \{\delta_{1,k}\}_{k=1}^{\infty} \in \Delta$, $\delta_2 = \{\delta_{2,k}\}_{k=1}^{\infty} \in \Delta$. Then Δ is compact.

Lemma 7.2. Let $s(\delta_1, \delta_2)$ be another metric on Δ such that for some constant $c > 0$, $s(\delta_1, \delta_2) \leq c \cdot d(\delta_1, \delta_2)$. Then under the conditions of Lemma 7.2, the space Δ endowed with the metric s is compact as well.

7.2 Smooth densities on the unit interval

Note that if we replace the condition $\sum_{k=1}^{\infty} \bar{\delta}_k^2 < \infty$ in Lemma 7.1 by $\sum_{k=1}^{\infty} k^{\ell} \bar{\delta}_k < \infty$ for some integer $\ell \geq 0$ and the metric $d(\delta_1, \delta_2)$ by

$$d(\delta_1, \delta_2) = \sum_{k=1}^{\infty} k^{\ell} |\delta_{1,k} - \delta_{2,k}|$$

then the result of Lemma 7.1 carries over. Consequently, the following results hold.

Theorem 7.2. Let $\mathcal{D}_{\ell}(0, 1)$ be the space of densities of the type (6.4) subject to the restrictions $|\delta_k| \leq \bar{\delta}_k$ for some a priori chosen positive sequence $\bar{\delta}_k$ satisfying $\sum_{k=1}^{\infty} k^{\ell} \bar{\delta}_k < \infty$ for some integer $\ell \geq 0$. Endow $\mathcal{D}_{\ell}(0, 1)$ with the Sobolev¹ metric

$$\|h_1 - h_2\|_{\ell} = \max_{0 \leq m \leq \ell} \sup_{0 \leq u \leq 1} |h_1^{(m)}(u) - h_2^{(m)}(u)|, \quad (7.6)$$

where $h^{(m)}(u) = d^m h(u)/(du)^m$ for $m \geq 1$, $h^{(0)}(u) = h(u)$. Then $\mathcal{D}_{\ell}(0, 1)$ is compact. Moreover, let $\mathcal{D}_{\ell,n}(0, 1)$ be the space of SNP densities of the type (6.6), subject to the same restrictions on the δ_k 's, and endowed with the same metric as $\mathcal{D}_{\ell}(0, 1)$. Again, $\mathcal{D}_{\ell,0}(0, 1)$ is the singleton $\{h(u) \equiv 1\}$. Then the sequence $\mathcal{D}_{\ell,n}(0, 1)$ is dense in $\mathcal{D}_{\ell}(0, 1)$: $\mathcal{D}_{\ell}(0, 1) = \overline{\cup_{n=0}^{\infty} \mathcal{D}_{\ell,n}(0, 1)}$.

¹See for example Adams and Fournier (2003).

This result follows from the fact that for each pair $h_1, h_2 \in \mathcal{D}_\ell(0, 1)$ with corresponding sequences $\{\delta_{1,k}\}_{k=1}^\infty$ and $\{\delta_{2,k}\}_{k=1}^\infty$ we have

$$\|h_1 - h_2\|_\ell = O \left(\sum_{k=1}^{\infty} k^\ell |\delta_{1,k} - \delta_{2,k}| \right), \quad (7.7)$$

as is not hard to verify.

7.3 Appendix: Proofs

7.3.1 Lemma 7.1

To prove the compactness of Δ it suffices to prove that Δ is complete and totally bounded. See Royden (1968, Proposition 15, p.164).

Completeness means that every Cauchy sequence in Δ takes a limit in Δ . To show this, let $\delta_n = \{\delta_{n,k}\}_{k=1}^\infty$ be an arbitrary Cauchy sequence in Δ , i.e.,

$$\lim_{\min(n,m) \rightarrow \infty} d(\delta_n, \delta_m) = \lim_{\min(n,m) \rightarrow \infty} \sqrt{\sum_{k=1}^{\infty} (\delta_{n,k} - \delta_{m,k})^2} = 0.$$

Then for each $k \geq 1$, $\lim_{\min(n,m) \rightarrow \infty} |\delta_{n,k} - \delta_{m,k}| = 0$, hence $\delta_{n,k}$ is a Cauchy sequence in $[-\bar{\delta}_k, \bar{\delta}_k]$ and therefore takes a limit $\delta_k \in [-\bar{\delta}_k, \bar{\delta}_k]$. Consequently, $\delta = \{\delta_k\}_{k=1}^\infty \in \Delta$ and $\lim_{n \rightarrow \infty} d(\delta_n, \delta) = 0$, where the latter follows from

$$\begin{aligned} \limsup_{n \rightarrow \infty} (d(\delta_n, \delta))^2 &= \limsup_{n \rightarrow \infty} \sum_{k=1}^{\infty} (\delta_{n,k} - \delta_k)^2 \\ &\leq \limsup_{n \rightarrow \infty} \sum_{k=1}^m (\delta_{n,k} - \delta_k)^2 + 4 \sum_{k=m+1}^{\infty} \bar{\delta}_k^2 \\ &= 4 \sum_{k=m+1}^{\infty} \bar{\delta}_k^2 \rightarrow 0 \text{ as } m \rightarrow \infty. \end{aligned}$$

Thus, Δ is complete.

To prove total boundedness, let $\varepsilon > 0$ be arbitrary, and choose an n so large that $\sqrt{\sum_{k=n+1}^{\infty} \bar{\delta}_k^2} < \varepsilon/4$. Denote

$$\Delta_n = (\mathbb{X}_{k=1}^n [-\bar{\delta}_k, \bar{\delta}_k]) \times (\mathbb{X}_{k=n+1}^\infty \{0\}). \quad (7.8)$$

Since $\bigcup_{k=1}^n [-\bar{\delta}_k, \bar{\delta}_k]$ is a closed and bounded subset of \mathbb{R}^n it is compact, hence Δ_n is compact. Therefore, there exist elements $\delta_1, \dots, \delta_M$ of Δ_n such that $\Delta_n \subset \bigcup_{i=1}^M \{\delta_* \in \Delta_n : d(\delta, \delta_i) < \varepsilon/2\}$. Since for each $\delta \in \Delta$ there exists a $\delta_* \in \Delta_n$ such that $d(\delta, \delta_*) \leq 2\sqrt{\sum_{k=n+1}^{\infty} \bar{\delta}_k^2} < \varepsilon/2$, it follows now that each $\delta \in \Delta$ belongs to one of the open sets $\{\delta \in \Delta : d(\delta, \delta_i) < \varepsilon\}$, hence $\Delta \subset \bigcup_{i=1}^M \{\delta \in \Delta : d(\delta, \delta_i) < \varepsilon\}$. Thus, Δ is totally bounded.

7.3.2 Lemma 7.2

Let Δ_O be a set which is open under the metric $s(., .)$ but not under the metric $d(., .)$. Let $\underline{\delta} \in \Delta_O$ be a point of closure under the d -metric. Note that by assumption, $\underline{\delta}$ is an interior point of Δ_O under the s -metric. Then for every $\varepsilon > 0$ there exists a $\delta \notin \Delta_O$ such that $d(\underline{\delta}, \delta) < \varepsilon$. But then $s(\underline{\delta}, \delta) < \varepsilon/c$, which would imply that $\underline{\delta}$ is a point of closure under the s -metric as well. This contradiction implies that open sets under the s -metric are also open under the d -metric. Consequently, any open covering of Δ under the s -metric is an open covering under the d -metric. Since in the latter case Δ is compact, there exists a finite sub-covering of Δ , which is also a finite sub-covering under the s -metric. Hence Δ is compact under the s -metric.

Part II

Semi-Nonparametric models (To be done)

References

- Adams, R.A. & J.J.F. Fournier (2003), *Sobolev Spaces*. Academic Press.
- Andrews, D.W.K. (1994), "Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity", *Econometrica* 62, 43-72.
- Bickel, P.J., C.A.J. Klaassen, Y. Ritov & J.A. Wellner (1998), *Efficient and Adaptive Estimation for Semiparametric Models*, Springer.
- Bierens, H. J. (1997), "Testing the Unit Root with Drift Hypothesis Against Nonlinear Trend Stationarity, with an Application to the U.S. Price Level and Interest Rate", *Journal of Econometrics* 81, 29-64.
- Bierens, H.J. (2004), *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Cambridge University Press.
- Bierens, H.J. (2008), "Semi-Nonparametric Interval-Censored Mixed Proportional Hazard Models: Identification and Consistency Results", *Econometric Theory* 24, 749-794.
- Bierens, H.J. (2011), "Consistency and Asymptotic Normality of Sieve Estimators Under Weak and Verifiable Conditions". <http://econ.la.psu.edu/~hbierens/SNPMODELS.PDF>
- Bierens, H.J. & J. R. Carvalho (2007), "Semi-Nonparametric Competing Risks Analysis of Recidivism", *Journal of Applied Econometrics* 22, 971-993.
- Bierens, H. J. & L. F. Martins (2010), "Time Varying Cointegration", *Econometric Theory* 26, 1453-1490.
- Bierens, H.J. & H. Pott-Buter (1990), "Specification of Engel Curves by Nonparametric Regression (with discussion)", *Econometric Reviews* 9, 123-184.
- Billingsley, P. (1968), *Convergence of Probability Measures*. John Wiley.
- Chen, X. (2007), "Large sample sieve estimation of semi-nonparametric models". In J.J. Heckman & E. Leamer (eds.), *Handbook of Econometrics*, Vol. 6, Ch. 76. Elsevier.
- Chen, X., O. Linton & I. Van Keilegom (2003), "Estimation of Semiparametric Models when the Criterion Function is Not Smooth", *Econometrica*, 71, 1591-1608.
- Eastwood, B.J. & A.R. Gallant (1991), "Adaptive Rules for Semi-Nonparametric Estimators that Achieve Asymptotic Normality", *Econometric Theory* 7, 307-340.
- Elbers, C. & G. Ridder (1982), "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model", *Review of Economic Studies* 49, 403-409.

- Gallant, A. R. (1981), "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form", *Journal of Econometrics* 15, 211-245.
- Gabler, S, F. Laisney & M. Lechner (1993), "Seminonparametric Estimation of Binary-Choice Models with an Application to Labor-Force Participation", *Journal of Business & Economic Statistics* 11, 61-80.
- Gallant, A.R. & D.W. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation", *Econometrica* 55, 363-390.
- Gill, R.D. (1989), "Non- and Semi-Parametric Maximum Likelihood Estimators and the Von Mises Method (Part 1)", *Scandinavian Journal of Statistics* 16, 97-128.
- Grenander, U. (1981), *Abstract Inference*. Wiley.
- Hamming, R.W. (1973), *Numerical Methods for Scientists and Engineers*. Dover Publications.
- Hannan, E.J., & B.G. Quinn (1979), "The Determination of the Order of an Autoregression", *Journal of the Royal Statistical Society, Series B*, 41, 190-195.
- Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error", *Econometrica* 47, 153-161.
- Horowitz, J.L. (1998), *Semiparametric Methods in Econometrics*, Springer.
- Jennrich, R.I. (1969), "Asymptotic Properties of Nonlinear Least Squares Estimators", *Annals of Mathematical Statistics* 40, 633-643.
- Kronmal, R. & M. Tarter (1968), "The Estimation of Densities and Cumulatives by Fourier Series Methods", *Journal of the American Statistical Association* 63, 925-952.
- Lancaster, T. (1979), "Econometric Methods for the Duration of Unemployment", *Econometrica* 47, 939-956.
- Manski, C.F. (1988), "Identification of Binary Response Models", *Journal of the American Statistical Association* 83, 729-738.
- Newey, W.K. (1997), "Convergence Rates and Asymptotic Normality for Series Estimators", *Journal of Econometrics* 79, 147-168.
- Royden, H.L. (1968), *Real Analysis*. Macmillan.
- Schwarz, G. (1978), "Estimating the Dimension of a Model", *Annals of Statistics* 6, 461-464.
- Young, N. (1988), *An Introduction to Hilbert Space*. Cambridge University Press.
- Shen, X. (1997), "On the Method of Sieves and Penalization", *Annals of Statistics* 25, 2555-2591.

Sims, C.A. (1980), "Macroeconomics and Reality", *Econometrica* 48, 1-48.

White, H. & J. Wooldridge (1991), "Some Results on Sieve Estimation with Dependent Observations". In W.A. Barnett, J. Powell & G. Tauchen (eds.), *Non-parametric and Semi-parametric Methods in Econometrics and Statistics*, Ch. 18, Cambridge University Press.

Wold, H. (1938), *A Study in the Analysis of Stationary Time Series*. Almqvist and Wiksell, Sweden.

Semi-Nonparametric Identification of the Right Censored Mixed Proportional Hazard Model

Herman J. Bierens
Department of Economics
Pennsylvania State University

November 3, 2008

Abstract

Elbers and Ridder (1982) and Heckman and Singer (1984) have shown that under mild conditions the mixed proportional hazard (MPH) model is nonparametrically identified. However, Elbers and Ridder did not consider the case of right censored MPH models, whereas Heckman and Singer only considered right censoring in the Weibull baseline hazard case. In this lecture note I will explain the Elbers-Ridder approach and extend it to the case of right censoring.

1 The mixed proportional hazard model

Let \tilde{T} be a duration, and let X be a vector of covariates. The conditional hazard function is defined as

$$\lim_{\delta \downarrow 0} \frac{\Pr[\tilde{T} \in (t, t + \delta) | \tilde{T} > t, X]}{\delta} = \frac{f(t|X)}{1 - F(t|X)} = \lambda(t, X),$$

where $F(t|X) = \Pr[\tilde{T} \leq t|X]$, and $f(t|X)$ is the corresponding conditional density function. Since

$$\frac{\partial \ln(1 - F(t|X))}{\partial t} = \frac{-f(t|X)}{1 - F(t|X)} = -\lambda(t, X),$$

it follows that

$$1 - F(t|X) = \exp\left(-\int_0^t \lambda(\tau, X)d\tau\right)$$

The proportional hazard model assumes that

$$\lambda(t, X) = \varphi(X) \lambda_0(t),$$

where $\varphi(X) > 0$ is called the systematic hazard, and $\lambda_0(t)$ is called the baseline hazard. Usually, $\varphi(X)$ is parametrized as

$$\varphi(X) = \exp(\beta'_0 X), \quad (1)$$

so that the conditional survival function takes the form

$$\begin{aligned} S(t|X) &= \Pr[\tilde{T} > t|X] = \exp\left(-\exp(\beta'_0 X) \int_0^t \lambda_0(\tau)d\tau\right) \\ &= \exp(-\exp(\beta'_0 X) \Lambda_0(t)), \end{aligned}$$

where

$$\Lambda_0(t) = \int_0^t \lambda_0(\tau)d\tau$$

is the integrated baseline hazard. Note that for $\lim_{t \rightarrow \infty} \Pr[\tilde{T} > t|X] = 0$ we need to require that $\lim_{t \rightarrow \infty} \Lambda_0(t) = \int_0^\infty \lambda_0(\tau)d\tau = \infty$, and for $\Pr[\tilde{T} > t|X]$ to be monotonic decreasing on $(0, \infty)$ we need to require that $\lambda_0(t) > 0$ for $t \in (0, \infty)$.

Adopting the specification (1) for the systematic hazard, the mixed proportional hazard (MPH) model, proposed by Lancaster (1979), assumes that the conditional survival function takes the form

$$S(t|X) = E[\exp(-\exp(\beta'_0 X + Y) \Lambda_0(t))|X], \quad (2)$$

where Y represents unobserved heterogeneity, which is assumed to be independent of X . Denoting the distribution function of $V = \exp(Y)$ by $G_0(v)$, we have

$$\begin{aligned} S(t|X) &= \int_0^\infty \exp(-v \cdot \exp(\beta'_0 X) \Lambda_0(t)) dG_0(v) \\ &= \int_0^\infty (\exp(-\exp(\beta'_0 X) \Lambda_0(t)))^v dG_0(v) \\ &= H_0(\exp(-\exp(\beta'_0 X) \Lambda_0(t))), \end{aligned} \quad (3)$$

where

$$H_0(u) = \int_0^\infty u^v dG_0(v), \quad u \in [0, 1], \quad (4)$$

is a distribution function on $[0, 1]$. Since for $u \in (0, 1)$ and $n = 1, 2, 3, \dots$,

$$\sup_{v \geq 0} \left| \frac{\partial^n (u^v)}{(\partial u)^n} \right| < \infty$$

it follows from the dominated convergence theorem that

$$d^n H_0(u) / (du)^n = u^{-n} \int_0^\infty \prod_{j=0}^n (v - j) u^v dG_0(v)$$

Consequently, $H_0(u)$ is absolutely continuous, with density

$$h_0(u) = \int_0^\infty v u^{v-1} dG_0(v), \quad (5)$$

which is continuously differentiable of any order on $(0, 1)$. Note that if $\lim_{v \uparrow 1} G_0(v) > 0$ then $\lim_{u \downarrow 0} h_0(u) = \infty$ but $\lim_{u \downarrow 0} u \cdot h_0(u) = 0$, whereas $\lim_{u \uparrow 1} h_0(u) = \int_0^\infty v dG_0(v)$. Moreover, note that absence of unobserved heterogeneity, i.e., $\Pr[V = 1] = 1$, is equivalent to the case $h_0(u) \equiv 1$.

2 Right-censoring

Usually the duration \tilde{T} is only observed up to an upper bound \bar{T} , which may vary per individual. This is called right-censoring, which is indicated by the observable dummy variable $C = I(\tilde{T} > \bar{T})$, where $I(\cdot)$ is the indicator function.¹ As usual, it will be assumed that

Assumption 1. *Conditional on the covariates X , the actual duration \tilde{T} and the censoring time \bar{T} are independent. The observed duration T is equal to \tilde{T} if $\tilde{T} \leq \bar{T}$ and is equal to \bar{T} if $\tilde{T} > \bar{T}$. These events are observable in the form of a dummy variable $C = I(\tilde{T} > \bar{T})$.*

Also, for later reference it will be assumed that

¹ $I(true) = 1$, $I(false) = 0$.

Assumption 2. *The support of the distribution of \bar{T} is $(0, \bar{t})$,*

where possibly but not necessarily $\bar{t} = \infty$. Then

$$\Pr[C = 1|X, \bar{T}] = S(\bar{T}|X) = H_0\left(\exp\left(-\exp(\beta'_0 X)\Lambda_0(\bar{T})\right)\right) \quad (6)$$

and

$$\begin{aligned} \Pr[T \leq t|X, \bar{T}, C = 0] &= \frac{\Pr[T \leq t, T \leq \bar{T}|X]}{\Pr[C = 0|X, \bar{T}]} = \frac{\Pr[T \leq \min(t, \bar{T})|X]}{\Pr[C = 0|X, \bar{T}]} \\ &= \frac{1 - H_0\left(\exp\left(-\exp(\beta'_0 X)\Lambda_0\left(\min(t, \bar{T})\right)\right)\right)}{1 - H_0\left(\exp\left(-\exp(\beta'_0 X)\Lambda_0(\bar{T})\right)\right)}. \end{aligned}$$

Thus the survival function of T given X, \bar{T} and $C = 0$ takes the form:

$$\begin{aligned} S(t|X, \bar{T}, C = 0) &\quad (7) \\ &= \frac{H_0\left(\exp\left(-\exp(\beta'_0 X)\Lambda_0(t)\right)\right) - H_0\left(\exp\left(-\exp(\beta'_0 X)\Lambda_0(\bar{T})\right)\right)}{1 - H_0\left(\exp\left(-\exp(\beta'_0 X)\Lambda_0(\bar{T})\right)\right)} \\ &\quad \times I(\bar{T} > t) \end{aligned}$$

3 Nonparametric identification

Elbers and Ridder (1982) have shown that under some conditions the MPH model is nonparametrically identified. Heckman and Singer (1984) provide an alternative identification proof based on the more general conditions in Kiefer and Wolfowitz (1956), and propose to parametrize G_0 as a discrete distribution: $G_0(v) = \sum_{i=1}^q I(v \leq \theta_i)p_i$, with $I(\cdot)$ the indicator function, where $\theta_i > 0$, $p_i > 0$, and $\sum_{i=1}^q p_i = 1$. Thus, Heckman and Singer (1984) implicitly specify $h_0(u) = \sum_{i=1}^q \theta_i u^{\theta_i-1} p_i$. However, they assume that the support S of the systematic hazard $\exp(\beta'_0 X)$ contains an open interval, which excludes the case that all the components of X are discrete. Elbers and Ridder (1982) derive the nonparametric identification of the MPH model under this condition as well as for the case that X is a dummy variable, with $\beta_0 \neq 0$.

In this section I will explain the Elbers and Ridder (1982) approach, in a slightly different way than in that paper.

3.1 Identification of the systematic hazard

Suppose that there exist a parameter vector β , an integrated baseline hazard $\Lambda(t)$, and a distribution function $H(u)$ on $[0, 1]$ with density $h(u)$ such that

$$\begin{aligned} \sup_{t \in [0, \bar{T}]} |H(\exp(-\exp(\beta'X).\Lambda(t))) - H_0(\exp(-\exp(\beta'_0X).\Lambda_0(t)))| \\ = 0 \text{ a.s.} \end{aligned} \quad (8)$$

Then

Lemma 1. *Under Assumption 2, (8) implies that for all $t \in (0, \bar{t})$*

$$H(\exp(-\exp(\beta'X).\Lambda(t))) = H_0(\exp(-\exp(\beta'_0X).\Lambda_0(t))) \text{ a.s.} \quad (9)$$

Proof: Appendix

Taking derivatives of both sides of (9) to $t \in (0, \bar{t})$ yield

$$\begin{aligned} & h(\exp(-\exp(\beta'X).\Lambda(t))) \exp(-\exp(\beta'X).\Lambda(t)) \\ & \times \exp(\beta'X)\lambda(t) \\ & = h_0(\exp(-\exp(\beta'_0X).\Lambda_0(t))) \exp(-\exp(\beta'_0X)\Lambda_0(t)) \\ & \times \exp(\beta'_0X)\lambda_0(t) \end{aligned}$$

hence

$$\begin{aligned} & \frac{h(\exp(-\exp(\beta'X).\Lambda(t)))}{h_0(\exp(-\exp(\beta'_0X).\Lambda_0(t)))} \\ & = \frac{\exp(-\exp(\beta'_0X).\Lambda_0(t))}{\exp(-\exp(\beta'X).\Lambda(t))} \exp((\beta_0 - \beta)'X) \frac{\lambda_0(t)}{\lambda(t)}. \end{aligned} \quad (10)$$

Following Elbers and Ridder (1982) and Bierens (2008) I will now assume that

Assumption 3: *The distribution function $G_0(v)$ of the unobserved heterogeneity variable V is confined to the class \mathcal{G} of distribution functions G on $(0, \infty)$ satisfying $\int_0^\infty v dG(v) = 1$.*

The actual assumption involved is that for all $G \in \mathcal{G}$, $\int_0^\infty v dG(v) = \mu$ for some common $\mu \in (0, \infty)$. Then without loss of generality we may normalize $\mu = 1$.

With $h_0(u) = \int_0^\infty vu^{v-1}dG_0(v)$ and $h(u) = \int_0^\infty vu^{v-1}dG(v)$, where $G_0, G \in \mathcal{G}$, it follows from Assumption 3 that

$$h_0(1) = h(1) = 1. \quad (11)$$

Regardless whether or not $h(u)$ is of the form $\int_0^\infty vu^{v-1}dG(v)$ with $G \in \mathcal{G}$, it will be assumed that condition (11) holds. How to implement (11) in practice has been shown in Bierens (2008).

Taking the limit of (10) for $t \downarrow 0$, it follows now from (11) that

$$1 = \frac{h(1)}{h_0(1)} = \exp\left((\beta_0 - \beta)' X\right) \lim_{t \downarrow 0} \frac{\lambda_0(t)}{\lambda(t)}. \quad (12)$$

hence

$$(\beta - \beta_0)' X = \ln\left(\lim_{t \downarrow 0} \frac{\lambda_0(t)}{\lambda(t)}\right) \text{ a.s.} \quad (13)$$

Next, following Bierens (2008), suppose that

Assumption 4: $E[X'X] < \infty$ and $\Sigma = E[(X - E[X])(X - E[X])']$ is non-singular.²

The condition $E[X'X] < \infty$ implies that $E[X]$ exists and is finite, hence $\ln(\lim_{t \downarrow 0} \lambda_0(t)/\lambda(t)) = (\beta - \beta_0)' E[X]$ exists and is finite and thus by (13)

$$(\beta - \beta_0)' (X - E[X]) = 0 \text{ a.s.} \quad (14)$$

It follows now from (14) and Assumption 4 that $(\beta - \beta_0)' \Sigma (\beta - \beta_0) = 0$, hence $\beta = \beta_0$ and

$$\lim_{t \downarrow 0} \frac{\lambda_0(t)}{\lambda(t)} = 1. \quad (15)$$

3.2 Identification of the baseline hazard in the Weibull case

Note that if $\lambda_0(t)$ and $\lambda(t)$ are of the Weibull type,

$$\lambda_0(t) = \exp(\alpha_0) \omega_0 t^{\omega_0-1}, \quad \lambda(t) = \exp(\alpha) \omega t^{\omega-1},$$

²Note that the nonsingularity of Σ excludes the presence of a constant covariate.

where α_0 and α are scale parameters, and $\omega_0 > 0$, $\omega > 0$, then (15) implies $\alpha = \alpha_0$, $\omega = \omega_0$, hence $\lambda(t) = \lambda_0(t)$ for all $t > 0$. Thus in this case (9) reads

$$\begin{aligned} H\left(\exp\left(-\exp(\alpha_0 + \beta'_0 X) \cdot \bar{\Lambda}_0(t)\right)\right) \\ = H_0\left(\exp\left(-\exp(\alpha_0 + \beta'_0 X) \cdot \bar{\Lambda}_0(t)\right)\right) \text{ a.s.} \end{aligned} \quad (16)$$

for all $t \in [0, \bar{t})$, where

$$\bar{\Lambda}_0(t) = t^{\omega_0}.$$

However, this result by itself does not pin down α_0 , because (16) also holds for $H^*(u) = H(u^{1/c})$ and $H_0^*(u) = H_0(u^{1/c})$, where $c > 0$ is arbitrary:

$$\begin{aligned} H^*\left(\exp\left(-\exp(\alpha_0 + \ln(c) + \beta'_0 X) \cdot \bar{\Lambda}_0(t)\right)\right) \\ = H_0^*\left(\exp\left(-\exp(\alpha_0 + \ln(c) + \beta'_0 X) \cdot \bar{\Lambda}_0(t)\right)\right) \text{ a.s.} \end{aligned}$$

On the other hand, if $H_0^*(u) = \int_0^\infty u^v dG_0^*(v)$ where $G_0^* \in \mathcal{G}$, then by Assumption 3 and (11),

$$1 = \lim_{u \uparrow 1} \frac{dH_0^*(u)}{du} = \lim_{u \uparrow 1} h_0(u^{1/c}) \frac{1}{c} u^{1/c-1} = \frac{1}{c}.$$

and similarly for H^* . Thus, Assumption 3 pins down the scale of the integrated hazard in the Weibull case.

3.3 Identification of the integrated hazard

3.3.1 Continuous covariates

Under Assumptions 2-3, (9) now reads

$$H(\exp(-Z \cdot \Lambda(t))) = H_0(\exp(-Z \cdot \Lambda_0(t))) \quad (17)$$

a.s. for $t \in (0, \bar{t})$, where $Z = \exp(\beta'_0 X)$. Elbers and Ridder (1982) assume in first instance that the support of Z contains an open interval, for example the interval $(z_0 - \varepsilon, z_0 + \varepsilon)$. Then

$$H(\exp(-z \cdot \Lambda(t))) = H_0(\exp(-z \cdot \Lambda_0(t)))$$

on $(z_0 - \varepsilon, z_0 + \varepsilon) \times (0, \bar{t})$. Hence, taking the derivative to $z \in (z_0 - \varepsilon, z_0 + \varepsilon)$ and letting $z \rightarrow z_0$, it follows that

$$\begin{aligned} h(\exp(-z_0 \cdot \Lambda(t))) \exp(-z_0 \cdot \Lambda(t)) \Lambda(t) \\ = h_0(\exp(-z_0 \cdot \Lambda_0(t))) \exp(-z_0 \cdot \Lambda_0(t)) \Lambda_0(t), \end{aligned} \quad (18)$$

whereas if we take the derivative to $t \in (0, \bar{t})$ and set $z = z_0$ then

$$\begin{aligned} h(\exp(-z_0 \Lambda(t))) \exp(-z_0 \Lambda(t)) \lambda(t) \\ = h_0(\exp(-z_0 \Lambda_0(t))) \exp(-z_0 \Lambda_0(t)) \lambda_0(t). \end{aligned} \quad (19)$$

Dividing (19) by (18) yields

$$\frac{\lambda(t)}{\Lambda(t)} = \frac{\lambda_0(t)}{\Lambda_0(t)} \text{ for } t \in (0, \bar{t}),$$

which is equivalent to

$$\frac{d \ln(\Lambda(t))}{dt} = \frac{d \Lambda_0(t)}{dt} \text{ for } t \in (0, \bar{t}).$$

Integrating both derivatives and using the boundary condition $\Lambda(0) = \Lambda_0(0) = 0$ it follows that

$$\Lambda(t) = \Lambda_0(t) \text{ for } t \in [0, \bar{t}]. \quad (20)$$

Thus, the integrated baseline hazard is now identified on $[0, \bar{t}]$ up to a multiplicative constant. However, this constant is identified. So see this, denote

$$\begin{aligned} \alpha_0 &= \ln(\Lambda_0(1)) = \ln(\Lambda(1)), \\ \bar{\Lambda}_0(t) &= \Lambda_0(t)/\Lambda_0(1), \quad \bar{\Lambda}_0(t) = \Lambda_0(t)/\Lambda_0(1) \end{aligned}$$

so that $\bar{\Lambda}(1) = \bar{\Lambda}_0(1) = 1$. Then it follows similar to the Weibull case that α_0 is identified.

3.3.2 Discrete covariates

In the Appendix of their paper, Elbers and Ridder (1982) also consider the case where the systematic hazard takes two values. Thus, let again $Z = \exp(\beta'_0 X)$ and suppose that there exists a pair z_1, z_2 such that

$$\Pr[Z = z_1] > 0, \quad \Pr[Z = z_2] > 0, \quad 0 < z_1 < z_2 < \infty.$$

Then for $0 \leq t \leq \bar{t}$ and $j = 1, 2$,

$$H(\exp(-z_j \Lambda(t))) = H_0(\exp(-z_j \Lambda_0(t))),$$

hence

$$\Lambda(t) = \frac{\ln(H^{-1}(H_0(\exp(-z_j \cdot \Lambda_0(t)))))}{-z_j}, \quad j = 1, 2. \quad (21)$$

Next, assume that

Assumption 5. $\lambda_0(t) > 0$ on $(0, \infty)$.

Then $\Lambda_0(t)$ is strictly monotonic increasing on $(0, \infty)$, and since $H_0(u)$ and $H(u)$ are strictly monotonic increasing on $(0, 1)$ it follows from (21) that $\Lambda(t)$ is strictly monotonic increasing on $(0, \infty)$. Therefore, both $\Lambda_0(t)$ and $\Lambda(t)$ are invertible on $(0, \infty)$, with inverses denoted by $\Lambda_0^{-1}(\cdot)$ and $\Lambda^{-1}(\cdot)$, respectively.

Given an arbitrary $t \in (0, \bar{t})$, let

$$t_2 = \Lambda_0^{-1}\left(\frac{z_1}{z_2} \Lambda_0(t)\right) = \Lambda_0^{-1}(\rho \Lambda_0(t)), \quad (22)$$

where

$$\rho = z_1/z_2 < 1,$$

and note that by the monotonicity of $\Lambda_0(t)$, $t_2 < t$. Then it follows from (21) that

$$\begin{aligned} \Lambda(t_2) &= \Lambda\left(\Lambda_0^{-1}(\rho \Lambda_0(t))\right) \\ &= \frac{\ln(H^{-1}(H_0(\exp(-z_2 \cdot \Lambda_0(t_2)))))}{-z_2} \\ &= \frac{\ln(H^{-1}(H_0(\exp(-z_2 \cdot \Lambda_0(\Lambda_0^{-1}(\frac{z_1}{z_2} \Lambda_0(t)))))))}{-z_2} \\ &= \frac{\ln(H^{-1}(H_0(\exp(-z_1 \cdot \Lambda_0(t)))))}{-z_2} = \frac{z_1}{z_2} \Lambda(t) \\ &= \rho \Lambda(t), \end{aligned} \quad (23)$$

hence

$$\rho \Lambda(t) = \Lambda\left(\Lambda_0^{-1}(\rho \Lambda_0(t))\right) \quad (24)$$

and thus also

$$\rho \Lambda_0(t) = \Lambda_0\left(\Lambda^{-1}(\rho \Lambda(t))\right). \quad (25)$$

More generally, it follows by induction that

Lemma 2. For $n = 1, 2, 3, \dots$ and $t \in (0, \bar{t})$, $\rho^n \Lambda_0(t) = \Lambda_0(\Lambda^{-1}(\rho^n \Lambda(t)))$.

Proof: Appendix.

Lemma 2 implies that

$$\frac{\Lambda_0(t)}{\Lambda(t)} = \frac{\Lambda_0(\Lambda^{-1}(\rho^n \Lambda(t)))}{\rho^n \Lambda(t)}.$$

Taking the limit for $n \rightarrow \infty$ yields

$$\begin{aligned} \frac{\Lambda_0(t)}{\Lambda(t)} &= \lim_{n \rightarrow \infty} \frac{\Lambda_0(\Lambda^{-1}(\rho^n \Lambda(t)))}{\rho^n \Lambda(t)} \\ &= \lim_{n \rightarrow \infty} \frac{\Lambda_0(\Lambda^{-1}(\rho^n \Lambda(t))) - \Lambda_0(\Lambda^{-1}(0))}{\rho^n \Lambda(t)} \\ &= \lim_{\tau \downarrow 0} \frac{d\Lambda_0(\Lambda^{-1}(\tau))}{d\tau} \Big| \\ &= \lim_{\tau \downarrow 0} \frac{d\Lambda_0(\Lambda^{-1}(\tau))}{d\Lambda^{-1}(\tau)} \times \frac{d\Lambda^{-1}(\tau)}{d\Lambda(\Lambda^{-1}(\tau))} \Big| \\ &= \lim_{\tau \downarrow 0} \frac{\lambda_0(\Lambda^{-1}(\tau))}{\lambda(\Lambda^{-1}(\tau))} \Big| = \lim_{\tau \downarrow 0} \frac{\lambda_0(\tau)}{\lambda(\tau)} \Big| \\ &= 1 \end{aligned}$$

where the latter follows from (15). Since $t \in (0, \bar{t})$ was chosen arbitrarily, it follows that $\Lambda(t) = \Lambda_0(t)$ on $[0, \bar{t}]$. Consequently, the result (27) carries over.

3.3.3 Mixed continuous-discrete covariates

As is well-known [see for example Chung (1974, Sect. 1.3)], the distribution function $\Psi(z)$ of $Z = \exp(\beta'_0 X)$ can always be written as a unique convex combination

$$\Psi(z) = \Pr[Z \leq z] = \delta_1 \Psi_c(z) + \delta_2 \Psi_d(z) + \delta_3 \Psi_s(z)$$

where $\delta_1 \geq 0$, $\delta_2 \geq 0$, $\delta_3 \geq 0$, $\delta_1 + \delta_2 + \delta_3 = 1$, $\Psi_c(z)$ is an absolutely continuous distribution function, $\Psi_d(z)$ is a discrete distribution function, and $\Psi_s(z)$ is a singular continuous distribution function.

Elbers and Ridder (1982) assume that either the support of Z contains an open interval, which corresponds to $\delta_1 > 0$ and $\Psi'_c(z) > 0$ on an open interval, or that $\delta_1 = 0$ and $\delta_2 > 0$ where $\Psi_d(z)$ is non-degenerated (i.e., $\Psi_d(z)$ is not the distribution function of a constant, so that $\Psi_d(z)$ has at least two jumps). These two assumptions can be combined as follows:

Assumption 6. *The distribution function $\Psi(z)$ of $Z = \exp(\beta'_0 X)$ satisfies $\Psi(z) = \delta_1 \Psi_c(z) + \delta_2 \Psi_d(z)$, where $\delta_1 \geq 0$, $\delta_2 \geq 0$, $\delta_1 + \delta_2 = 1$, $\Psi_c(z)$ is an absolutely continuous distribution function on $(0, \infty)$ with support containing an open interval, and $\Psi_d(z)$ is a non-degenerated discrete distribution function.*

Summarizing, it has been shown:

Theorem 1. *Let Assumptions 1-6 hold. Suppose that the conditional survival function has two equivalent representations,*

$$\begin{aligned} S(t|X) &= H_0(\exp(-\exp(\alpha_0 + \beta'_0 X) \Lambda_0(t))) \\ &= H(\exp(-\exp(\alpha + \beta' X) \Lambda(t))) \end{aligned} \quad (26)$$

for all $t \in (0, \bar{t})$, where \bar{t} is the upper bound of the support of the censoring time \bar{T} , the integrated hazards $\Lambda_0(t)$ and $\Lambda(t)$ are normalized such that $\Lambda(1) = \Lambda_0(1) = 1$, and the density h of the distribution function H is normalized by $h(1) = 1$. Then $\alpha = \alpha$, $\beta = \beta_0$ and $\Lambda(t) = \Lambda_0(t)$ for all $t \in (0, \bar{t})$.

3.4 Identification of the unobserved heterogeneity distribution

It remains to show that under the conditions of Theorem 1,

$$\begin{aligned} \Pr \left[\sup_{t \in [0, \bar{t}]} |H(\exp(-\exp(\alpha_0 + \beta'_0 X) \Lambda_0(t))) - H_0(\exp(-\exp(\alpha_0 + \beta'_0 X) \Lambda_0(t)))| = 0 \mid X \right] \\ = 1 \end{aligned} \quad (27)$$

implies $H(u) = H_0(u)$ on $[0, 1]$. If $\bar{t} = \infty$ this is trivial. If $\bar{t} < \infty$ but $H(u)$ is of the type $H(u) = \int_0^\infty u^v dG(v)$ then $H = H_0$ follows from the following lemma.

Lemma 3. *Let $H(u) = \int_0^\infty u^v dG(v)$ and $H_0(u) = \int_0^\infty u^v dG_0(v)$ for all $u \in [0, 1]$, where $G(v)$ and $G_0(v)$ are distribution functions with non-negative support. If $H(u) = H_0(u)$ on an arbitrary interval $(\underline{u}, \bar{u}) \subset [0, 1]$ then $G(v) = G_0(v)$ on $[0, \infty)$, hence $H(u) = H_0(u)$ on $[0, 1]$.*

Proof: Appendix

Remark. Note that the condition that H takes the form $H(u) = \int_0^\infty u^v dG(v)$ for all $u \in [0, 1]$ is necessary. If $H_0(u) = \int_0^\infty u^v dG_0(v)$ a.e. on $[0, 1]$ but only $H(u) = \int_0^\infty u^v dG(v) = H_0(u)$ a.e. on (\underline{u}, \bar{u}) then Lemma 3 implies that $H(u) = \int_0^\infty u^v dG_0(v)$ a.e. on (\underline{u}, \bar{u}) but not a.e. on $[0, 1]$ if $\underline{u} > 0$ or $\bar{u} < 1$. The latter follows easily from the fact that we can always extend $H(u)$ beyond (\underline{u}, \bar{u}) such that $H(u) \neq \int_0^\infty u^v dG_0(v)$ with positive Lebesgue measure.

Theorem 2. *In addition to the conditions of Theorem 1, suppose that either the support of the censoring time \bar{T} is $(0, \infty)$, or $H(u) = \int_0^\infty u^v dG(v)$, where G has nonnegative support. Then $H(u) = H_0(u)$ for all $u \in [0, 1]$.*

4 Proofs

4.1 Lemma 1

It is trivial that (8) implies that for any $\tau \in (0, \bar{t})$,

$$\begin{aligned} & \sup_{t \in [0, \tau]} |H(\exp(-\exp(\beta' X) \cdot \Lambda(t))) - H_0(\exp(-\exp(\beta'_0 X) \cdot \Lambda_0(t)))| \\ & \quad \times I(\bar{T} \leq \tau) = 0 \text{ a.s.} \end{aligned}$$

Letting $\tau \rightarrow \bar{t}$, the lemma now follows from the fact that by Assumption 2, $I(\bar{T} \leq \bar{t}) = 1$ a.s.

4.2 Lemma 2

Suppose that for an $n \geq 1$, $\rho^n \Lambda_0(t) = \Lambda_0(\Lambda^{-1}(\rho^n \Lambda(t)))$ on $(0, \bar{t})$. Let $t_2 = \Lambda^{-1}(\rho^n \Lambda(t))$ and note that $t_2 \leq t$ because $\rho \in (0, 1)$. Then

$$\begin{aligned}\rho^{n+1} \Lambda_0(t) &= \rho^n \cdot \Lambda_0(\Lambda_0^{-1}(\rho \Lambda_0(t))) \\ &= \rho^n \cdot \Lambda_0(t_2) = \Lambda_0(\Lambda^{-1}(\rho^n \Lambda(t_2))) \\ &= \Lambda_0(\Lambda^{-1}(\rho^{n+1} \Lambda(t))).\end{aligned}$$

4.3 Lemma 3

The proof of Lemma 3 is an adaptation of the results of Abbring and van den Berg (2003). It is given in the separate appendix (Bierens and Carvalho 2007b, Lemma A.1) to Bierens and Carvalho (2007a), but will be reproduced here again.

First, observe that for $u \in (0, 1)$ and non-negative integers m ,

$$\sup_{v \geq 0} v^m u^{v-1} < \infty. \quad (28)$$

Take the derivative of $H(u)$ and $H_0(u)$ to $u \in (\underline{u}, \bar{u})$. Then by (28) and dominated convergence we may take the derivatives inside the integrals involved:

$$\int_0^\infty v u^{v-1} dG(v) = \int_0^\infty v u^{v-1} dG_0(v). \quad (29)$$

Multiply (29) by u , and then take the derivatives to $u \in (\underline{u}, \bar{u})$ again, which by (28) implies that

$$\int_0^\infty v^2 u^{v-1} dG(v) = \int_0^\infty v^2 u^{v-1} dG_0(v).$$

Repeating this procedure it follows by induction that

$$\int_0^\infty v^m u^v dG(v) = \int_0^\infty v^m u^v dG_0(v) \text{ for } m = 0, 1, 2, \dots \quad (30)$$

hence

$$\int_0^\infty \sum_{m=0}^k \frac{(t \cdot v)^m}{m!} u^v dG(v) = \int_0^\infty \sum_{m=0}^k \frac{(t \cdot v)^m}{m!} u^v dG_0(v) \text{ for } k = 0, 1, 2, \dots \quad (31)$$

Since

$$\begin{aligned} \sup_{k \geq 1} \left| \sum_{m=0}^k \frac{(t.v)^m}{m!} u^v \right| &\leq \sum_{m=0}^{\infty} \frac{(|t|.v)^m}{m!} \exp(-v \cdot \ln(1/u)) \\ &= \exp((|t| - \ln(1/u)) \cdot v) \\ &\leq 1 \text{ if } |t| < \ln(1/u) \end{aligned}$$

it follows from (31) and bounded convergence that

$$\int_0^\infty \exp(t.v) u^v dG(v) = \int_0^\infty \exp(t.v) u^v dG_0(v) \text{ for } |t| < \ln(1/u). \quad (32)$$

Now denote

$$F(x|u) = \frac{\int_0^x u^v dG(v)}{\int_0^\infty u^v dG(v)}, \quad F_0(x|u) = \frac{\int_0^x u^v dG_0(v)}{\int_0^\infty u^v dG_0(v)} \quad (33)$$

for $u \in (\underline{u}, \bar{u})$ and $x > 0$. Then it follows from (32) and (33) that

$$\int_0^\infty \exp(t.v) dF(v|u) = \int_0^\infty \exp(t.v) dF_0(v|u) \text{ for } |t| < \ln(1/u).$$

Hence it follows from the uniqueness of moment-generating functions that $F(x|u) = F_0(x|u)$ for $u \in (\underline{u}, \bar{u})$ and $x > 0$, and thus

$$\int_0^x u^v dG(v) = \int_0^x u^v dG_0(v). \quad (34)$$

Moreover, similar to (30) it follows from (34) that for $x > 0$, $m, k = 0, 1, 2, \dots$ and $u \in (\underline{u}, \bar{u})$,

$$\int_0^x v^{m+k} u^v dG(v) = \int_0^x v^{m+k} u^v dG_0(v),$$

hence

$$\begin{aligned} \int_0^x v^m dG(v) &= \int_0^x v^m \sum_{k=0}^{\infty} \frac{(v \cdot \ln(1/u))^k}{k!} u^v dG(v) \\ &= \sum_{k=0}^{\infty} \frac{(\ln(1/u))^k}{k!} \int_0^x v^{m+k} u^v dG(v) \\ &= \sum_{k=0}^{\infty} \frac{(\ln(1/u))^k}{k!} \int_0^x v^{m+k} u^v dG_0(v) \end{aligned}$$

$$\begin{aligned}
&= \int_0^x \sum_{k=0}^{\infty} \frac{(v \cdot \ln(1/u))^k}{k!} v^m u^v dG_0(v) \\
&= \int_0^x \exp(-v \cdot \ln(u)) u^v v^m dG_0(v) \\
&= \int_0^x v^m dG_0(v).
\end{aligned}$$

Thus, for $x > 0$ and $m = 0, 1, 2, \dots$,

$$\int_0^\infty (v \cdot I(v \leq x))^m dG(v) = \int_0^\infty (v \cdot I(v \leq x))^m dG_0(v), \quad (35)$$

where $I(\cdot)$ is the indicator function.

Now use the well-known fact that distributions of bounded random variables are equal if and only if all their moments are equal. Then, with V a random drawing from $G(v)$ and V_0 a random drawing from $G_0(v)$, it follows from (35) that for $x > v > 0$,

$$\Pr[V \cdot I(V \leq x) \leq v] = \Pr[V_0 \cdot I(V_0 \leq x) \leq v].$$

This implies that $G(x) - G(v) = G_0(x) - G_0(v)$. Hence, letting $x \rightarrow \infty$, it follows that

$$G(v) = G_0(v) \text{ for } v \geq 0.$$

References

- Abbring, J. H., and G. J. van den Berg (2003), "The Identifiability of the Mixed Proportional Hazards Competing Risks Model", *Journal of the Royal Statistical Society B* 65, 701-710.
- Bierens, H. J. (2008), "Semi-Nonparametric Interval Censored Mixed Proportional Hazard Models: Identification and Consistency Results", *Econometric Theory* 24, 749-794.
- Bierens, H.J. and J. R. Carvalho (2007a), "Semi-Nonparametric Competing Risks Analysis of Recidivism", *Journal of Applied Econometrics* 22, 971-993.
- Bierens, H.J. and J. R. Carvalho (2007b), "Separate appendix to: Semi-Nonparametric Competing Risks Analysis of Recidivism" (http://econ.la.psu.edu/~hbierens/RECIDIVISM_APP.PDF).
- Chung, K. L. (1974), *A Course in Probability Theory*, New York, Academic Press.

Elbers, C., and G. Ridder (1982), "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model", *Review of Economic Studies*, 49, 403-409.

Heckman, J. J., and B. Singer (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data", *Econometrica*, 52, 271-320.

Kiefer, J., and J. Wolfowitz (1956), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters", *The Annals of Mathematical Statistic*, 27, 887-906.

Lancaster, T. (1979), "Econometric Methods for the Duration of Unemployment", *Econometrica*, 47, 939-956.

The Integrated Conditional Moment Test

Herman J. Bierens

Abstract

In this lecture I will review the integrated conditional moment (ICM) test for functional form of a conditional expectation model. This is a consistent test: the ICM test has asymptotic power 1 against all deviations from the null hypothesis. Moreover, this test has non-trivial \sqrt{n} local power.

I will focus on the mathematical foundations of the ICM approach, in particular the consistency proof and the derivation of the asymptotic null distribution.

1 The Fourier transform of a Borel measurable function

Let $g(x)$ be a Borel measurable real function on \mathbb{R}^k . The Fourier transform of $g(x)$ relative to a probability measure $\mu_X(\cdot)$ on the Borel sets in \mathbb{R}^k is defined by

$$\varphi(\xi) = \int \exp(i.\xi'x) g(x) d\mu_X(x), \quad i = \sqrt{-1},$$

provided that $\int |g(x)| d\mu_X(x) < \infty$.

LEMMA 1: Let $g_1(x)$ and $g_2(x)$ be Borel measurable functions on \mathbb{R}^k satisfying $\int |g_1(x)| d\mu_X(x) < \infty$, $\int |g_2(x)| d\mu_X(x) < \infty$, with Fourier transforms $\varphi_1(\xi)$, $\varphi_2(\xi)$, respectively, relative to a probability measure $\mu(\cdot)$ on the Borel sets in \mathbb{R}^k . Then $g_1(x) = g_2(x)$ a.s. $\mu_X(\cdot)$, i.e.,

$$B_0 = \{x \in \mathbb{R}^k : g_1(x) - g_2(x) = 0\} \Rightarrow \mu_X(B_0) = 1, \text{ if and only if } \varphi_1(\xi) \equiv \varphi_2(\xi).$$

Proof: Suppose $\varphi_1(\xi) \equiv \varphi_2(\xi)$ and $\mu_X(B_0) < 1$. Let

$$\begin{aligned} r_1(x) &= \max(0, g_1(x) - g_2(x)), \\ r_2(x) &= \max(0, -g_1(x) + g_2(x)) \end{aligned}$$

Then $g_1(x) - g_2(x) = r_1(x) - r_2(x)$ and

$$\begin{aligned} &\int \exp(i.\xi' x) r_1(x) d\mu_X(x) \\ &- \int \exp(i.\xi' x) r_2(x) d\mu_X(x) \\ &= \varphi_1(\xi) - \varphi_2(\xi) \equiv 0 \end{aligned}$$

Substituting $\xi = 0$ yields

$$\int r_1(x) d\mu_X(x) = \int r_2(x) d\mu_X(x) = c \geq 0.$$

If $c = 0$ then $r_1(x) = r_2(x) = 0$ a.s. $\mu(\cdot)$, hence $g_1(x) = g_2(x)$ a.s. $\mu(\cdot)$.

Therefore, assume that $c > 0$. Then we can define the probability measures

$$v_m(B) = \frac{1}{c} \int_B r_m(x) d\mu_X(x), \quad m = 1, 2,$$

with corresponding characteristic functions

$$\begin{aligned}\psi_m(\xi) &= \int \exp(i.\xi'y) dv_m(y) \\ &= \frac{1}{c} \int \exp(i.\xi'x) r_m(x) d\mu_X(x)\end{aligned}$$

for $m = 1, 2$. But I have just established that

$\int \exp(i.\xi'x) r_1(x) d\mu_X(x) \equiv \int \exp(i.\xi'x) r_2(x) d\mu_X(x)$, hence $\psi_1(\xi) \equiv \psi_2(\xi)$, which by the uniqueness of characteristic functions implies that $v_1(B) = v_2(B)$ for all Borel sets $B \subset \mathbb{R}^k$. It is now an easy (ECON 501) exercise to verify that the latter implies $r_1(x) = r_2(x)$ a.s. $\mu_X(\cdot)$, hence $g_1(x) = g_2(x)$ a.s. $\mu_X(\cdot)$.

Corollary:

LEMMA 2: *Let U be a random variable satisfying $E[|U|] < \infty$, and let $X \in \mathbb{R}^k$ be a random vector. If $P[E(U|X) = 0] < 1$ then there exists a $\xi \in \mathbb{R}^k$ such that $E[U \exp(i.\xi'X)] \neq 0$.*

Question: Where to look for such a ξ ?

LEMMA 3: If X is bounded then under the conditions of Lemma 2, for each $\varepsilon > 0$ there exists a ξ satisfying $\|\xi\| < \varepsilon$ such that $E[U \exp(i.\xi'X)] \neq 0$.

Proof: Let $X \in \mathbb{R}$. Then

$$\begin{aligned} E[U \exp(i.\xi X)] &= E\left[U \sum_{m=0}^{\infty} \frac{i^m \xi^m X^m}{m!}\right] \\ &= \sum_{m=0}^{\infty} \frac{i^m \xi^m}{m!} E[U.X^m] \end{aligned}$$

Since $E[U \exp(i.\xi X)] \neq 0$ for some ξ we must have that $E[U.X^m] \neq 0$ for some integer $m \geq 0$. Let m_0 be the smallest m for which $E[U.X^m] \neq 0$. Then

$$\left. \frac{d^{m_0} E[U \exp(i.\xi X)]}{(d\xi)^{m_0}} \right|_{\xi=0} = i^{m_0} E[U.X^{m_0}] \neq 0$$

which implies that $E[U \exp(i.\xi X)] \neq 0$ for $\xi \neq 0$ arbitrarily close to zero.

LEMMA 4: *Under the conditions of Lemma 3, the set $S_0 = \{\xi \in \mathbb{R}^k : E[U \exp(i \cdot \xi' X)] = 0\}$ has Lebesgue measure zero and is nowhere dense.*

Proof: Let $k = 1$ and $\xi_0 \in S_0$. Define $U_0 = U \exp(i \cdot \xi_0 X)$. Then $P(E[U_0 | X] = 0) < 1$, hence for an arbitrarily small $\varepsilon > 0$ there exists a $\xi \in (-\varepsilon, 0) \cup (0, \varepsilon)$ such that

$$E[U \exp(i \cdot \xi_0 X) \exp(i \cdot \xi X)] \neq 0.$$

By continuity it follows now that for each $\xi_0 \in S_0$ there exists an $\varepsilon > 0$ such that

$\xi \notin S_0$ for all $\xi \in (\xi_0 - \varepsilon, \xi_0) \cup (\xi_0, \xi_0 + \varepsilon)$. Consequently, in the case $k = 1$, the set S_0 is countable and is nowhere dense. In the general case $k \geq 1$, S_0 has Lebesgue measure zero and is nowhere dense.

More generally, we have:

LEMMA 5: *Let $w(u)$ be a real or complex valued function of the type*

$$w(u) = \sum_{s=0}^{\infty} (\gamma_s / s!) u^s$$

where $|\gamma_s| < \infty$ and at most a finite number of γ_s 's are zero. Then under the conditions of Lemma 3, the set

$$S_0 = \{ \xi \in \mathbb{R}^k : E [U.w(\xi'X)] = 0 \}$$

has Lebesgue measure zero and is nowhere dense.

For example, let $w(u) = \cos(u) + \sin(u)$, or $w(u) = \exp(u)$.

The condition that the random vector X is bounded can be get rid of by replacing X with $\Phi(X)$, where Φ is a Borel measurable bounded one-to-one mapping, because the σ -algebra generated by X is then the same as the σ -algebra generated by $\Phi(X)$, hence conditioning on $\Phi(X)$ is equivalent to conditioning on X .

THEOREM 1: Let U be a random variable satisfying $E[|U|] < \infty$ and let $X \in \mathbb{R}^k$ be a random vector. Denote

$S = \{\xi \in \mathbb{R}^k : E[U.w(\xi' \Phi(X))] = 0\},$
 where $w(\cdot)$ is defined in Lemma 5, and $\Phi(\cdot)$ is a Borel measurable bounded one-to-one mapping. If $P[E(U|X) = 0] < 1$ then S has Lebesgue measure zero and is nowhere dense, whereas if $P[E(U|X) = 0] = 1$ then $S = \mathbb{R}^k$.

2 The ICM test

Given a random sample (Y_j, X_j) , $j = 1, \dots, n$, $X_j \in \mathbb{R}^k$, and a conditional expectation model

$$E(Y_j|X_j) = g(X_j, \theta_0), \quad \theta_0 \in \Theta,$$

where $\Theta \subset \mathbb{R}^m$ is the parameter space, Theorem 1 suggests to test the correctness of the functional specification of this model on the basis of following ICM statistic:

$$\int |\widehat{z}(\xi)|^2 d\mu(\xi)$$

In this expression, $\mu(\xi)$ is an absolutely continuous (w.r.t. Lebesgue measure) probability measure with compact support $\Xi \subset \mathbb{R}^k$, and

$$\widehat{z}(\xi) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \widehat{U}_j w(\xi' \Phi(X_j)).$$

where $\widehat{U}_j = Y_j - g(X_j, \widehat{\theta})$, with $\widehat{\theta}$ the NLLS estimator of θ_0 , Φ is a bounded one-to-one mapping, and $w(\cdot)$ is a weight function satisfying the conditions of Theorem 1.

More formally, the null hypothesis to be tested is that

H_0 : There exists a $\theta_0 \in \Theta$ such that

$$P [E(y_j|x_j) = g(x_j, \theta_0)] = 1,$$

and the alternative hypothesis is that H_0 is false:

H_1 : For all $\theta \in \Theta$,

$$P [E(y_j|x_j) = g(x_j, \theta)] < 1,$$

Under the null hypothesis and standard regularity conditions,

$$\begin{aligned}\sqrt{n} \left(\widehat{\theta} - \theta_0 \right) &= A^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^n \left. \frac{\partial g(X_j, \theta)}{\partial \theta'} \right|_{\theta=\theta_0} U_j \\ &\quad + o_p(1)\end{aligned}$$

where

$$A = p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \left. \left(\frac{\partial g(X_j, \theta)}{\partial \theta'} \right) \left(\frac{\partial g(X_j, \theta)}{\partial \theta'} \right)' \right|_{\theta=\theta_0}$$

Hence, by the uniform law of large numbers,

$$\begin{aligned}\widehat{z}(\xi) &= \frac{1}{\sqrt{n}} \sum_{j=1}^n \widehat{U}_j w(\xi' \Phi(X_j)) \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^n U_j w(\xi' \Phi(X_j)) \\ &\quad - \frac{1}{\sqrt{n}} \sum_{j=1}^n \left(g(X_j, \widehat{\theta}) - g(X_j, \theta_0) \right) w(\xi' \Phi(X_j)) \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^n U_j \phi_j(\xi) + o_p(1),\end{aligned}$$

say, where

$$\begin{aligned}\phi_j(\xi) = & w(\xi' \Phi(X_j)) \\ & - b(\xi)' A^{-1} \left. \frac{\partial g(X_j, \theta)}{\partial \theta'} \right|_{\theta=\theta_0}\end{aligned}$$

with

$$b(\xi) = p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \left. \frac{\partial g(X_j, \theta)}{\partial \theta'} \right|_{\theta=\theta_0} w(\xi' \Phi(X_j)),$$

and $o_p(1)$ is uniform in $\xi \in \Xi$.

THEOREM 2: *Under the null hypothesis and some regularity conditions (one of these conditions is that Ξ is compact),*

$$\widehat{z}(\xi) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \widehat{U}_j w(\xi' \Phi(X_j)) \Rightarrow z(\xi) \text{ on } \Xi,$$

where $z(\xi)$ is a zero-mean Gaussian process on Ξ , with covariance function

$$\Gamma(\xi_1, \xi_2) = E[z(\xi_1)z(\xi_2)].$$

Hence by the continuous mapping theorem,

$$\int |\widehat{z}(\xi)|^2 d\mu(\xi) \rightarrow_d \int |z(\xi)|^2 d\mu(\xi).$$

Under the alternative that the null is false,

$\widehat{z}(\xi)/\sqrt{n} \rightarrow_p \eta(\xi)$ uniformly on Ξ , where $\eta(\xi) \neq 0$ except on a set with zero Lebesgue measure,

so that

$$(1/n) \int |\widehat{z}(\xi)|^2 d\mu(\xi) \rightarrow_p \int |\eta(\xi)|^2 d\mu(\xi) > 0,$$

provided that $\mu(\xi)$ is absolutely continuous w.r.t. Lebesgue measure and its support Ξ has positive Lebesgue measure.

3 The null distribution of the ICM test

If we choose the weight function w real-valued, for example $w(u) = \cos(u) + \sin(u)$, then $z(\xi)$ is a real-valued zero-mean Gaussian process on Ξ , with real-valued covariance function

$$\begin{aligned}\Gamma(\xi_1, \xi_2) &= E [z(\xi_1)z(\xi_2)] \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n E [U_j^2 \phi_j(\xi_1) \phi_j(\xi_2)].\end{aligned}$$

This covariance function is symmetric and positive semidefinite, in the following sense:

$$\int \int \psi(\xi_1) \Gamma(\xi_1, \xi_2) \psi(\xi_2) d\mu(\xi_1) d\mu(\xi_2) \geq 0$$

for all Lebesgue integrable functions $\psi(\xi)$ on Ξ . Such functions have non-negative eigenvalues and corresponding orthonormal eigenfunctions:

THEOREM 3: *The functional eigenvalue problem* $\int \Gamma(\xi_1, \xi_2) \psi(\xi_2) d\mu(\xi_2) = \lambda \cdot \psi(\xi_1)$ *a.e. on* Ξ *has a countable number of solutions*

$$\int \Gamma(\xi_1, \xi_2) \psi_j(\xi_2) d\mu(\xi_2) = \lambda_j \cdot \psi_j(\xi_1) \text{ a.e. on } \Xi,$$

$$j = 1, 2, \dots$$

where

$$\lambda_j \geq 0, \quad \sum_{j=1}^{\infty} \lambda_j < \infty,$$

$$\int_{\Xi} \psi_{j_1}(\xi) \psi_{j_2}(\xi) d\mu(\xi) \quad \begin{cases} = 0 & \text{if } j_1 \neq j_2 \\ = 1 & \text{if } j_1 = j_2 \end{cases}$$

Moreover,

THEOREM 4 (Mercer's Theorem): *The covariance function* $\Gamma(\xi_1, \xi_2)$ *can be written as* $\Gamma(\xi_1, \xi_2) = \sum_{j=1}^{\infty} \lambda_j \psi_j(\xi_1) \psi_j(\xi_2)$.

The sequence $\{\psi_t(\xi)\}_{t=1}^{\infty}$ is an orthonormal basis for a Hilbert space $\mathcal{H}(\mu)$ of Lebesgue integrable functions on Ξ , with inner product

$$\langle f, g \rangle = \int f(\xi) g(\xi) d\mu(\xi).$$

so that every function f in $\mathcal{H}(\mu)$ can be written as

$$f(\xi) = \sum_{t=1}^{\infty} \gamma_t \psi_t(\xi), \quad \sum_{t=1}^{\infty} \gamma_t^2 < \infty, \text{ where}$$

$$\gamma_t = \langle f, \psi_t \rangle, \quad t = 1, 2, 3, \dots$$

It can be shown that $z(\xi)$ is a random element of $\mathcal{H}(\mu)$, hence

$$z(\xi) = \sum_{t=1}^{\infty} z_t \psi_t(\xi), \text{ where}$$

$$z_t = \int_{\Xi} z(\xi) \psi_t(\xi) d\mu(\xi), \quad t = 1, 2, 3, \dots$$

Consequently,

$$\begin{aligned}
\int |z(\xi)|^2 d\mu(\xi) &= \int \left(\sum_{t=1}^{\infty} z_t \psi_t(\xi) \right)^2 d\mu(\xi) \\
&= \sum_{t_1=1}^{\infty} \sum_{t_2=1}^{\infty} z_{t_1} z_{t_2} \int \psi_{t_1}(\xi) \psi_{t_2}(\xi) d\mu(\xi) \\
&= \sum_{t_1=1}^{\infty} \sum_{t_2=1}^{\infty} z_{t_1} z_{t_2} I(t_1 = t_2) = \sum_{t=1}^{\infty} z_t^2
\end{aligned}$$

The sequence z_t is a zero-mean Gaussian process, with variance function

$$\begin{aligned}
E[z_t^2] &= E \left[\left(\int z(\xi) \psi_t(\xi) d\mu(\xi) \right)^2 \right] \\
&= \int \int E[z(\xi_1) z(\xi_2)] \psi_t(\xi_1) \psi_t(\xi_2) d\mu(\xi_1) d\mu(\xi_2) \\
&= \int \int \left(\sum_{j=1}^{\infty} \lambda_j \psi_j(\xi_1) \psi_j(\xi_2) \right) \psi_t(\xi_1) \psi_t(\xi_2) \\
&\quad \times d\mu(\xi_1) d\mu(\xi_2)
\end{aligned}$$

where the latter equality follows from Mercer's theorem.

Thus by the orthonormality of the eigenfunctions,

$$\begin{aligned} E[z_t^2] &= \sum_{j=1}^{\infty} \lambda_j \left(\int \psi_j(\xi) \psi_t(\xi) d\mu(\xi) \right)^2 \\ &= \sum_{j=1}^{\infty} \lambda_j I(j=t) = \lambda_t. \end{aligned}$$

Moreover, by a similar argument it follows that

$$\begin{aligned} E[z_{t_1} z_{t_2}] &= \sum_{j=1}^{\infty} \lambda_j I(j=t_1) I(j=t_2) \\ &= 0 \text{ if } t_1 \neq t_2. \end{aligned}$$

Hence, denoting $\varepsilon_t = z_t / \sqrt{\lambda_t}$ if $\lambda_t > 0$, we have:

THEOREM 5: $\int |z(\xi)|^2 d\mu(\xi) = \sum_{t=1}^{\infty} \lambda_t \varepsilon_t^2$, where the ε_t 's are i.i.d. $N(0, 1)$ and the λ_t 's are the eigenvalues of the covariance function Γ .

4 Critical values

The problem is that the eigenvalues λ_t are case-dependent: They depend on the distribution of the regressors, the functional form of the NLLS model, and the conditional variance of the errors. Therefore, the distribution of $\int |z(\xi)|^2 d\mu(\xi)$ cannot be tabulated. A possible way to get around this problem is to bootstrap this distribution. However, a convenient way to get around this problem is to derive upper bounds of the critical values, as follows.

Without loss of generality we may assume that the λ_t 's are positive and arranged in decreasing order. Moreover, it follows from Mercer's theorem that

$$\begin{aligned} \int \Gamma(\xi, \xi) d\mu(\xi) &= \sum_{j=1}^{\infty} \lambda_j \int \psi_j(\xi)^2 d\mu(\xi) \\ &= \sum_{j=1}^{\infty} \lambda_j \end{aligned}$$

THEOREM 6: Denoting $p_t = \lambda_t / \sum_{j=1}^{\infty} \lambda_j$,

we have

$$\begin{aligned} \frac{\int |z(\xi)|^2 d\mu(\xi)}{\int \Gamma(\xi, \xi) d\mu(\xi)} &= \sum_{t=1}^{\infty} p_t \varepsilon_t^2 \\ &\leq \sup_{p_1 \geq p_2 \geq \dots, \sum_{t=1}^{\infty} p_t = 1} \sum_{t=1}^{\infty} p_t \varepsilon_t^2 \\ &= \sup_{m \geq 1} \frac{1}{m} \sum_{i=1}^m \varepsilon_i^2 = \bar{T}, \end{aligned}$$

say,

so that asymptotic critical values can be derived from the latter distribution. The actual test statistic of the ICM test is therefore

$$\widehat{T}_{ICM} = \frac{\int |\widehat{z}(\xi)|^2 d\mu(\xi)}{\int \widehat{\Gamma}(\xi, \xi) d\mu(\xi)},$$

where $\widehat{\Gamma}(\xi_1, \xi_2)$ is a consistent estimator of $\Gamma(\xi_2, \xi_2)$, uniformly on $\Xi \times \Xi$.

5 Local power of the ICM test

Consider the local alternative hypothesis

$$H_1^L : E[Y_j|X_j] = g(X_j, \theta_0) + \frac{h(X_j)}{\sqrt{n}} \text{ a.s.,}$$

where $h(X_j)$ is not constant:

$$P[h(X_j) = E(h(X_j))] < 1.$$

Then under H_1^L ,

$$\hat{z}(\xi) \Rightarrow z(\xi) + \omega(\xi) \text{ on } \Xi,$$

where $z(\xi)$ is the same zero-mean Gaussian process on Ξ as before, and $\omega(\xi)$ is a deterministic mean function satisfying $0 < \int \omega(\xi)^2 d\mu(\xi) < \infty$.

Similar to the case under the null hypothesis, we can write

$$z(\xi) + \omega(\xi) = \sum_{t=1}^{\infty} z_t \psi_t(\xi)$$

where now

$$z_t = \varepsilon_t \sqrt{\lambda_t} + \omega_t$$

with ε_t i.i.d. $N(0, 1)$ and $\omega_t = \int \omega(\xi) \psi_t(\xi) d\mu(\xi)$.

Hence,

$$\begin{aligned} \int |\widehat{z}(\xi)|^2 d\mu(\xi) &\rightarrow_d \int |z(\xi) + \omega(\xi)|^2 d\mu(\xi) \\ &= \sum_{t=1}^{\infty} (\varepsilon_t \sqrt{\lambda_t} + \omega_t)^2 \end{aligned}$$

THEOREM 7: *The ICM test has nontrivial \sqrt{n} -local power in the sense that for every $K > 0$,*

$$\begin{aligned} P \left[\sum_{t=1}^{\infty} (\varepsilon_t \sqrt{\lambda_t} + \omega_t)^2 \leq K \right] \\ < P \left[\sum_{t=1}^{\infty} \lambda_t \varepsilon_t^2 \leq K \right] \end{aligned}$$

Proof:

Let

$$C_n = \sum_{t=1}^{\infty} (\varepsilon_t \sqrt{\lambda_t} + \omega_t)^2 - (\varepsilon_n \sqrt{\lambda_n} + \omega_n)^2$$

and suppose that $\omega_n \neq 0$. Then

$$\begin{aligned} & P \left[\sum_{t=1}^{\infty} (\varepsilon_t \sqrt{\lambda_t} + \omega_t)^2 \leq K \right] \\ &= P \left[(\varepsilon_n \sqrt{\lambda_n} + \omega_n)^2 \leq K - C_n \text{ and } C_n \leq K \right] \\ &= P \left[-\sqrt{K - C_n} \leq \varepsilon_n \sqrt{\lambda_n} + \omega_n \leq \sqrt{K - C_n} \right. \\ &\quad \left. \text{and } C_n \leq K \right] \\ &< P \left[-\sqrt{K - C_n} \leq \varepsilon_n \sqrt{\lambda_n} \leq \sqrt{K - C_n} \right. \\ &\quad \left. \text{and } C_n \leq K \right] \\ &= P \left[\varepsilon_n^2 \lambda_n + C_n \leq K \text{ and } C_n \leq K \right] \\ &= P \left[\varepsilon_n^2 \lambda_n + C_n \leq K \right] \end{aligned}$$

where the inequality is due to the symmetry and unimodality of the $N(0, \lambda_n)$ distribution. The result of Theorem 7 follows now by induction.

6 Bibliography

- Bierens, H. J. (1982): "Consistent Model Specification Tests", *Journal of Econometrics* 20, 105-134.
- Bierens, H. J. (1984): "Model Specification Testing of Time Series Regressions", *Journal of Econometrics* 26, 323-353.
- Bierens, H. J. (1990): "A Consistent Conditional Moment Test of Functional Form", *Econometrica* 58, 1443-1458.
- Bierens, H. J. and W. Ploberger (1997): "Asymptotic Theory of Integrated Conditional Moment Tests", *Econometrica* 65, 1129-1151.
- De Jong, R. M. (1996): "On the Bierens Test Under Data Dependence", *Journal of Econometrics* 72, 1-32.
- Stinchcombe, M. B., and H. White (1998): "Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative", *Econometric Theory* 14, 295-325.

Review of the Integrated Conditional Moment Test and Its Implementation in EasyReg International

Herman J. Bierens
Pennsylvania State University

April 21, 2006

1 The ICM test

The ICM test is based on the following theorem:

THEOREM 1: Let u be a random variable satisfying $E|u| < \infty$, and $P[E(u|x) = 0] < 1$, where $x \in \mathbb{R}^k$ is a bounded random vector.

(a) Let $w(u)$ be a complex or real valued function that is infinitely many times differentiable in $u = 0$ and satisfies the condition that $(d/du)^s w(u)|_{u=0} \neq 0$ for all but a finite number of natural numbers s . Then for every $\varepsilon > 0$ there exists a $\xi \in \mathbb{R}^k$ such that $E[u.w(\xi'x)] \neq 0$ and $\|\xi\| < \varepsilon$.

(b) If in addition $w(u)$ is a power series in an open neighborhood of $u = 0$, i.e., for some $\delta > 0$, $w(u) = \sum_{s=0}^{\infty} (\gamma_s / s!) u^s$ for $|u| < \delta$, where $\gamma_s = (d/du)^s w(u)|_{u=0}$, then the set $\{\xi \in \mathbb{R}^k : E[u.w(\xi'x)] = 0\}$ has Lebesgue measure zero and is nowhere dense.

Proof: See Bierens (1982) for part (a) with $w(u) = \exp(i.u)$, Bierens (1990) for the case $w(u) = \exp(u)$, and Bierens and Ploberger (1997) for the general case. Examples of suitable functions $w(u)$ in the general case are $w(u) = \cos(u) + \sin(u)$, and $w(u) = 1/[1 + \exp(c - u)]$ for $c \neq 0$. See also Stinchcombe and White (1998) for further elaborations on this theorem, and Bierens (1994, Ch. 3) for the details of the proof of Theorem 1 for the cases $w(u) = \exp(i.u)$ and $w(u) = \exp(u)$.

The condition that the random vector x is bounded can be get rid of by replacing x with $\Phi(x)$, where Φ is a Borel measurable bounded one-to-one mapping, because the σ -algebra generated by x is then the same as the σ -algebra generated by $\Phi(x)$, hence conditioning on $\Phi(x)$ is equivalent to conditioning on x . See Bierens (1982, 1990, 1994, Ch. 3).

Theorem 1 suggests that, given a random sample (y_j, x_j) , $j = 1, \dots, n$, $x_j \in \mathbb{R}^k$, and a conditional expectation model

$$E(y_j|x_j) = g(x_j, \theta_0), \theta_0 \in \Theta,$$

where $\Theta \subset \mathbb{R}^k$ is the parameter space, the null hypothesis

$$H_0: \text{There exists a } \theta_0 \in \Theta \text{ such that } P[E(y_j|x_j) = g(x_j, \theta_0)] = 1,$$

can be consistently tested against the general alternative hypothesis that the null hypothesis is false, i.e.,

$$H_1: \text{For all } \theta \in \Theta, \quad P[E(y_j|x_j) = g(x_j, \theta)] < 1,$$

on the basis of the Integrated Conditional Moment (ICM) statistic

$$\int |\hat{z}(\xi)|^2 d\mu(\xi).$$

In this integral,

$$\hat{z}(\xi) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \hat{u}_j w(\xi' \Phi(x_j)). \quad (1)$$

where \hat{u}_j is the nonlinear least squares residual: $\hat{u}_j = y_j - g(x_j, \hat{\theta})$, with $\hat{\theta}$ the nonlinear least squares estimator of θ_0 , Φ is a bounded one-to-one mapping, $w(\cdot)$ is a weight function satisfying the conditions of Theorem 1, and μ is a probability measure on a compact set $\Xi \subset \mathbb{R}^k$ with positive Lebesgue measure, which is absolute continuous with respect to Lebesgue measure.

This ICM statistic was proposed first by Bierens (1982), for the case $w(u) = \exp(i.u)$, Ξ a hypercube in \mathbb{R}^k , μ the Lebesgue measure on Ξ , and i.i.d. observations (y_j, x_j) .

2 The asymptotic null distribution of the ICM statistic

Under the null hypothesis and standard regularity conditions,

$$\sqrt{n} (\hat{\theta} - \theta_0) = A^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^n \left. \frac{\partial g(x_j, \theta)}{\partial \theta'} \right|_{\theta=\theta_0} u_j + o_p(1)$$

where

$$u_j = y_j - g(x_j, \theta_0) = y_j - E[y_j | x_j]$$

and

$$A = p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \left(\frac{\partial g(x_j, \theta)}{\partial \theta'} \right) \left(\frac{\partial g(x_j, \theta)}{\partial \theta'} \right)' \Big|_{\theta=\theta_0}$$

Hence,

$$\hat{z}(\xi) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \hat{u}_j w(\xi' \Phi(x_j)) = \frac{1}{\sqrt{n}} \sum_{j=1}^n u_j \phi_j(\xi) + o_p(1),$$

where

$$\phi_j(\xi) = w(\xi' \Phi(x_j)) - b(\xi)' A^{-1} \left. \frac{\partial g(x_j, \theta)}{\partial \theta'} \right|_{\theta=\theta_0}$$

with

$$b(\xi) = p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \left. \frac{\partial g(x_j, \theta)}{\partial \theta'} \right|_{\theta=\theta_0} w(\xi' \Phi(x_j)),$$

and $o_p(1)$ is uniform in $\xi \in \Xi$.

It has been shown by Bierens (1990) and Bierens and Ploberger (1997) that under some mild regularity conditions (among which the assumption that the function $w(\cdot)$ is real-valued), and the null hypothesis involved,

$$\tilde{z}(\xi) = \frac{1}{\sqrt{n}} \sum_{j=1}^n u_j \phi_j(\xi)$$

converges weakly to a zero-mean Gaussian process $z(\xi)$ on Ξ , with covariance function

$$\Gamma(\xi_1, \xi_2) = E[z(\xi_1) z(\xi_2)] = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n E[u_j^2 \phi_j(\xi_1) \phi_j(\xi_2)] \quad (2)$$

Consequently, under the null hypothesis

$$\int |\widehat{z}(\xi)|^2 d\mu(\xi) \rightarrow \int |z(\xi)|^2 d\mu(\xi)$$

in distribution, whereas under the general alternative that the null is false,

$$\widehat{z}(\xi)/\sqrt{n} \rightarrow \eta(\xi) = p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \widehat{u}_j w(\xi' \Phi(x_j))$$

in probability, uniformly on Ξ , where $\eta(\xi) \neq 0$ except on a set with zero Lebesgue measure. Thus, under the alternative hypothesis,

$$(1/n) \int |\widehat{z}(\xi)|^2 d\mu(\xi) \rightarrow \int |\eta(\xi)|^2 d\mu(\xi) > 0,$$

a.s.

Moreover, Bierens and Ploberger (1997) have shown that the asymptotic null distribution of the ICM statistic is of the type

$$\int |z(\xi)|^2 d\mu(\xi) = \sum_{i=1}^{\infty} \lambda_i \varepsilon_i^2,$$

where the ε_i 's are i.i.d. $N(0, 1)$ and the λ_i 's are the eigenvalues of the covariance function Γ , and that

$$\int \sigma^2(\xi) d\mu(\xi) = \sum_{i=1}^{\infty} \lambda_i,$$

where

$$\sigma^2(\xi) = \Gamma(\xi, \xi),$$

is the variance function of $z(\xi)$. Furthermore, they have shown that

$$\frac{\int |z(\xi)|^2 d\mu(\xi)}{\int \sigma^2(\xi) d\mu(\xi)} = \frac{\sum_{i=1}^{\infty} \lambda_i \varepsilon_i^2}{\sum_{i=1}^{\infty} \lambda_i} \leq \sup_{m \geq 1} \frac{1}{m} \sum_{i=1}^m \varepsilon_i^2 = \bar{T},$$

say, so that asymptotic critical values can be derived from the latter distribution.

As to the choice of the probability measure μ , Boning and Sowell (1999) have shown that the uniform probability measure on Ξ is optimal. The actual test statistic of the ICM test is therefore

$$\widehat{T}_{ICM} = \frac{\int_{\Xi} |\widehat{z}(\xi)|^2 d\xi}{\int_{\Xi} \widehat{\sigma}^2(\xi) d\xi}, \quad (3)$$

where $\widehat{\sigma}^2(\xi)$ is a consistent estimator of the variance function $\sigma^2(\xi)$, uniformly on Ξ .

The asymptotic null distribution of \widehat{T}_{ICM} is case-dependent, because the eigenvalues λ_i depend on the distribution of (y_t, x_t) and the conditional expectation model $g(x_t, \theta_0)$, but is dominated by the distribution of \bar{T} . Thus, denoting the $1 - \alpha$ quantile of \bar{T} by T_α , i.e., $P(\bar{T} \geq T_\alpha) = \alpha$, the null hypothesis is rejected at the $\alpha \times 100\%$ significance level if $\widehat{T}_{ICM} \geq T_\alpha$. The values of T_α for $\alpha = 0.10$ and $\alpha = 0.05$ are:

$$\begin{aligned} T_{0.10} &= 3.23 \\ T_{0.05} &= 4.26 \end{aligned}$$

3 The ICM test of the martingale difference hypothesis

The Bierens-Ploberger version of the ICM test allows for consistently testing of linear and nonlinear ARX models, but not for ARMAX models, because, given a $k + 1$ -variate vector time series process

$$z_t = (y_t, x'_{t+1})' \in \mathbb{R} \times \mathbb{R}^{k-1}, \quad (4)$$

an ARMAX model represents the conditional expectation of the dependent variable y_t relative to *all* lagged z_t 's. Bierens (1984) and De Jong (1996) have, in different ways, extended the ICM test to the case where z is infinite dimensional, i.e., $z = (z'_{t-1}, z'_{t-2}, \dots)'$, in order to accommodate conditioning on the infinite past of a k -variate time series process z_t . In this paper I shall review the approach of De Jong (1996).

The space $(\Xi, \|\cdot\|)$ defined in De Jong (1996) is given as follows. For two infinite sequences of points in $\mathbb{R}^k \times \mathbb{R}^\infty$, ξ and ζ , given by $\xi = (\xi'_1, \xi'_2, \dots)'$ and $\zeta = (\zeta'_1, \zeta'_2, \dots)'$, where $\xi_j, \zeta_j \in \mathbb{R}^k$, define the norm

$$\|\xi - \zeta\| = \sqrt{\sum_{j=1}^{\infty} j^2 |\xi_j - \zeta_j|^2}, \quad (5)$$

where $|\xi_j - \zeta_j|$ is the Euclidean norm on \mathbb{R}^k . Next, define the space Ξ as

$$\Xi = \{\xi \in \mathbb{R}^{k+1} \times \mathbb{R}^\infty : a_j \leq \xi_j \leq b_j, \forall j \geq 1\}, \quad (6)$$

where $a_j < b_j$ and $|a_j|, |b_j| \leq cj^{-2}$ for some constant $c > 0$. Note that Ξ has finite Lebesgue measure. With this definition $(\Xi, \|\cdot\|)$ is a compact metric space, and therefore it is totally bounded.

Following Bierens (1990), De Jong now proposes to use the weight function

$$w_t(\xi) = \exp \left(\sum_{j=1}^{t-1} \xi'_j \Phi(z_{t-j}) \right),$$

and the Lebesgue measure on Ξ as the measure μ . However, in view of Theorem 1, De Jong's results carry over to the more general case

$$w_t(\xi) = w \left(\sum_{j=1}^{t-1} \xi'_j \Phi(z_{t-j}) \right), \quad (7)$$

where $w(\cdot)$ is a real-valued function satisfying the conditions of Theorem 1(b), and Φ is a bounded one-to-one mapping. If $w(\cdot)$ is real valued but only satisfies the conditions of Theorem 1(a), we have to choose $a_j < 0 < b_j, \forall j$. In this case the results of Theorem 1(b) read:

THEOREM 2: *Let u_t be a random variable satisfying $E|u_t| < \infty$, and let z_t be a k -variate time series process, such that (u_t, z_t) is stationary. Let $(\Xi, \|\cdot\|)$ be defined by (5) and (6), and let*

$$\bar{w}_t(\xi) = w \left(\sum_{j=1}^{\infty} \xi'_j \Phi(z_{t-j}) \right), \quad (8)$$

where $w(\cdot)$ satisfies the conditions of Theorem 1. Then

$$P [E(u_t | z_{t-1}, z_{t-2}, \dots) = 0] < 1$$

if and only if the set $\{\xi \in \Xi : E(u_t \bar{w}_t(\xi)) = 0\}$ has Lebesgue measure zero and is nowhere dense in Ξ .

Proof: De Jong (1997).

The actual test statistic is now similar to (3), and the upper bounds of the critical values still apply.

4 The ICM test in EasyReg International

4.1 Cross-section data

If your model is a (nonlinear) cross-section regression model

$$y_j = g(x_j, \theta_0) + u_j, \quad j = 1, \dots, n, \quad (9)$$

the default instrumental variables are the components of x_j . You may remove some of these components from the list of instrumental variables, or add other variables to the list. Once you have confirmed the list of instrumental variables, they will be standardized by taking them in deviation of their sample means, and then dividing them by their sample standard errors. Thus, each component $x_{i,j}$ of x_j is standardized as

$$\tilde{x}_{i,j} = (x_{i,j} - \bar{x}_i) / s_i,$$

where

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{i,j}, \quad s_i^2 = \frac{1}{n-1} \sum_{j=1}^n (x_{i,j} - \bar{x}_i)^2.$$

Next, replace x_t in (1) with $\tilde{x}_t = (\tilde{x}_{1,t}, \dots, \tilde{x}_{k,t})'$:

$$\hat{z}(\xi) = \frac{1}{\sqrt{n}} \sum_{t=1}^n \hat{u}_t w(\xi' \Phi(\tilde{x}_t)).$$

The reason will be given below, after discussing the choice of Φ .

The components of the bounded one-to-one mapping Φ in EasyReg are $\arctan(\cdot)$. Thus

$$\Phi(\tilde{x}_t) = \begin{pmatrix} \arctan(\tilde{x}_{1,t}) \\ \vdots \\ \arctan(\tilde{x}_{k,t}) \end{pmatrix}.$$

Now without standardization, $\arctan(x_{i,t}) \approx \pi/2$ if all the $x_{i,t}$ take large positive values, or $\arctan(x_{i,t}) \approx -\pi/2$ if all the $x_{i,t}$ take large negative values, which would destroy the consistency of the ICM test.

As to the function $w(\cdot)$, EasyReg provides two options,

$$w(\cdot) = \cos(\cdot) + \sin(\cdot), \quad (10)$$

which is the default option, and

$$w(\cdot) = \exp(\cdot), \quad (11)$$

which was used in Bierens (1990).

Finally, the only option for the set Ξ in (3) is

$$\Xi(c) = \times_{\ell=1}^k [-c, c], \quad (12)$$

where $c > 0$ has to be chosen.

4.2 Computation of the ICM test statistic

Note that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n E [u_j^2 \phi_j(\xi)^2] \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n E \left[u_j^2 \left(w(\xi' \Phi(\tilde{x}_j)) - b(\xi)' A^{-1} \frac{\partial g(x_j, \theta)}{\partial \theta'} \Big|_{\theta=\theta_0} \right)^2 \right] \\ &= p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n u_j^2 w(\xi' \Phi(\tilde{x}_j))^2 \\ & \quad - 2b(\xi)' A^{-1} \left(p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n u_j^2 \frac{\partial g(x_j, \theta)}{\partial \theta'} \Big|_{\theta=\theta_0} w(\xi' \Phi(\tilde{x}_j)) \right) \\ & \quad + b(\xi)' A^{-1} \left(p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n u_j^2 \frac{\partial g(x_j, \theta)}{\partial \theta'} \Big|_{\theta=\theta_0} \frac{\partial g(x_j, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right) A^{-1} b(\xi) \\ &= p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n u_j^2 w(\xi' \Phi(\tilde{x}_j))^2 + b(\xi)' A^{-1} B A^{-1} b(\xi) - 2b(\xi)' A^{-1} c(\xi) \end{aligned}$$

where

$$\begin{aligned} c(\xi) &= p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n u_j^2 \frac{\partial g(x_j, \theta)}{\partial \theta'} \Big|_{\theta=\theta_0} w(\xi' \Phi(x_j)) \\ B &= p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n u_j^2 \frac{\partial g(x_j, \theta)}{\partial \theta'} \Big|_{\theta=\theta_0} \frac{\partial g(x_j, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \end{aligned}$$

Thus,

$$\begin{aligned}
\int_{\Xi} \sigma^2(\xi) d\xi &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n E \left[u_j^2 \int_{\Xi} \phi_j(\xi)^2 d\xi \right] \\
&= p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n u_j^2 \int_{\Xi} w(\xi' \Phi(\tilde{x}_j))^2 d\xi \\
&\quad + \int_{\Xi} b(\xi)' A^{-1} B A^{-1} b(\xi) d\xi - 2 \int_{\Xi} b(\xi)' A^{-1} c(\xi) d\xi
\end{aligned}$$

Denoting,

$$\hat{X}_j = \frac{\partial g(x_j, \theta)}{\partial \theta'} \Big|_{\theta=\hat{\theta}},$$

the matrices A and B can be consistently estimated by

$$\hat{A} = \frac{1}{n} \sum_{j=1}^n \hat{X}_j \hat{X}_j', \quad \hat{B} = \frac{1}{n} \sum_{j=1}^n \hat{u}_j^2 \hat{X}_j \hat{X}_j',$$

respectively, $b(\xi)$ can be consistently estimated by

$$\hat{b}(\xi) = \frac{1}{n} \sum_{j=1}^n \hat{X}_j w(\xi' \Phi(x_j)),$$

and $c(\xi)$ can be consistently estimated by

$$\hat{c}(\xi) = \frac{1}{n} \sum_{j=1}^n \hat{u}_j^2 \hat{X}_j w(\xi' \Phi(x_j)).$$

Hence,

$$\begin{aligned}
\int_{\Xi(c)} \hat{\sigma}^2(\xi) d\xi &= \frac{1}{n} \sum_{j=1}^n \hat{u}_j^2 \int_{\Xi(c)} w(\xi' \Phi(x_j))^2 d\xi \\
&\quad + \int_{\Xi(c)} \hat{b}(\xi)' \hat{A}^{-1} \hat{B} \hat{A}^{-1} b(\xi) d\xi - 2 \int_{\Xi(c)} \hat{b}(\xi)' \hat{A}^{-1} \hat{c}(\xi) d\xi.
\end{aligned}$$

Next, denote

$$\omega_{j_1, j_2}(c) = (2c)^{-k} \int_{\Xi(c)} w(\xi' \Phi(\tilde{x}_{j_1})) w(\xi' \Phi(\tilde{x}_{j_2})) d\xi$$

Then the ICM test statistic takes the form

$$\widehat{T}_{ICM}(c) = \frac{\widehat{T}_1(c)}{\widehat{T}_2(c)} \quad (13)$$

where

$$\widehat{T}_1(c) = \frac{\int_{\Xi(c)} |\widehat{z}(\xi)|^2 d\xi}{\int_{\Xi(c)} d\xi} = \frac{1}{n} \sum_{j_1=1}^n \sum_{j_2=1}^n \widehat{u}_{j_1} \widehat{u}_{j_2} \omega_{j_1, j_2}(c) \quad (14)$$

and

$$\begin{aligned} \widehat{T}_2(c) &= \frac{\int_{\Xi(c)} \widehat{\sigma}^2(\xi) d\xi}{\int_{\Xi(c)} d\xi} = \frac{1}{n} \sum_{j=1}^n \widehat{u}_j^2 \omega_{j,j}(c) \\ &\quad + \frac{1}{n^2} \sum_{j_1=1}^n \sum_{j_2=1}^n \widehat{X}'_{j_1} \widehat{A}^{-1} \widehat{B} \widehat{A}^{-1} \widehat{X}_{j_2} \omega_{j_1, j_2}(c) \\ &\quad - 2 \frac{1}{n^2} \sum_{j_1=1}^n \sum_{j_2=1}^n \widehat{X}'_{j_1} \widehat{A}^{-1} \widehat{X}_{j_2} \widehat{u}_{j_2}^2 \omega_{j_1, j_2}(c). \end{aligned} \quad (15)$$

To derive $\omega_{j_1, j_2}(c)$ let

$$\xi' \Phi(\widetilde{x}_j) = \sum_{\ell=1}^k \xi_\ell \phi_{\ell, j}.$$

Then it is easy to verify that

$$\begin{aligned} \omega_{j_1, j_2}(c) &= \prod_{\ell=1}^k \frac{\sin(c(\phi_{\ell, j_1} - \phi_{\ell, j_2}))}{c(\phi_{\ell, j_1} - \phi_{\ell, j_2})} \text{ if } w(.) = \cos(.) + \sin(.), \\ \omega_{j_1, j_2}(c) &= \prod_{\ell=1}^k \frac{\exp(c(\phi_{\ell, j_1} + \phi_{\ell, j_2})) - \exp(-c(\phi_{\ell, j_1} + \phi_{\ell, j_2}))}{2.c(\phi_{\ell, j_1} + \phi_{\ell, j_2})} \\ &\quad \text{if } w(.) = \exp(.). \end{aligned}$$

Note that in these cases $\omega_{j_1, j_2}(0) = 1$, so that

$$\widehat{T}_1(0) = \frac{1}{n} \sum_{j_1=1}^n \sum_{j_2=1}^n \widehat{u}_{j_1} \widehat{u}_{j_2} = \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n \widehat{u}_j \right)^2,$$

$$\begin{aligned}\widehat{T}_2(0) &= \frac{1}{n} \sum_{j=1}^n \widehat{u}_j^2 + \left(\frac{1}{n} \sum_{j=1}^n \widehat{X}_j \right)' \widehat{A}^{-1} \widehat{B} \widehat{A}^{-1} \left(\frac{1}{n} \sum_{j=1}^n \widehat{X}_j \right) \\ &\quad - 2 \left(\frac{1}{n} \sum_{j=1}^n \widehat{X}_j \right)' \widehat{A}^{-1} \left(\frac{1}{n} \sum_{j=1}^n \widehat{X}_j \widehat{u}_j^2 \right),\end{aligned}$$

If the model contains a constant term then by the first-order conditions for NLLS, $\sum_{j=1}^n \widehat{u}_j = 0$, hence $\widehat{T}_1(0) = 0$ and thus $T_{ICM}(0) = 0$. Therefore, c should not be chosen too close to 0.

4.3 Time series data

If your data consist of time series, EasyReg automatically assumes that you want to use De Jong's (1996) version of the ICM test. The default instrumental variables are then all the lagged dependent variables and the current and lagged X variables, if any. If there are X variables in your model, for example if your model is a (nonlinear) ARX(p) model,

$$y_t = g(y_{t-1}, \dots, y_{t-p}, x_t, \theta_0) + u_t, \quad (16)$$

or a (nonlinear) ARMAX(p, q) model,

$$y_t = g(y_{t-1}, \dots, y_{t-p}, x_t, \theta_0) + u_t - \gamma_1 u_{t-1} - \dots - \gamma_q u_{t-q}, \quad (17)$$

where u_t is a martingale difference process, and $x_t \in \mathbb{R}^k$ is a vector of exogenous (X) variables, then the default instrumental variables are y_{t-j} , $j \geq 1$, and all (lagged) x_t . However, only the latter are shown in the list of default instrumental variables. You may remove some or all of the components of x_t from the list of instrumental variables, or add other variables to the list of instrumental variables, but you cannot remove the lagged dependent variables. Once you have confirmed the choice of instruments other than the lagged dependent variables, you have to specify their minimum lags. In the cases (16) and (17) I recommend that you adopt the default instruments x_t , and specify zero as the minimum lag. Then the ICM test tests the martingale difference hypothesis

$$P(E[u_t | z_{t-1}, z_{t-2}, z_{t-3}, \dots] = 0) = 1,$$

where

$$z_{t-1} = (y_{t-1}, x_t')'.$$

Similarly to the cross-section case, the z_{t-j} 's in (7) are standardized as \tilde{z}_{t-j} by taking them in deviation of the sample mean, and then dividing them by the sample standard error. Moreover, the \tilde{z}_t 's that are not observable are set equal to zero vectors, say for $t < 1$. Then the actual weight functions are

$$w_t(\xi) = w \left(\sum_{j=1}^{t-1} \xi_j' \Phi(\tilde{z}_{t-j}) \right),$$

with Φ the same as in the cross-section case. Moreover, the set Σ is now infinite dimensional,

$$\Xi(c) = \times_{j=1}^{\infty} \left\{ \times_{\ell=1}^k [-cj^{-2}, cj^{-2}] \right\}$$

where $c > 0$ has to be chosen.

Finally, note that the same limiting null distribution and critical values as in the cross-section case apply.

4.4 Computational options

It is clear from (14) and (15) that if the sample size n is large, the computation of the integrals involved will take a long time on a PC. Therefore EasyReg offers the option (Option 1) to compute (14) and (15) by Monte Carlo integration:

$$\widehat{T}_1(c) \approx \frac{1}{M} \sum_{j=1}^M |\widehat{z}(\xi_j)|^2, \quad \widehat{T}_2(c) \approx \frac{1}{M} \sum_{j=1}^M \widehat{\sigma}^2(\xi_j),$$

where the ξ_j are random drawings from the uniform distribution $\Xi(c)$. However, if you choose the option of computing the ICM test exactly (Option 2), you can specify a time period after which you will get a message asking whether you want to continue, or to switch to Monte Carlo integration.

REFERENCES

- Bierens, H. J. (1982): "Consistent Model Specification Tests", *Journal of Econometrics* 20, 105-134.
- Bierens, H. J. (1984): "Model Specification Testing of Time Series Regressions", *Journal of Econometrics* 26, 323-353.
- Bierens, H. J. (1990): "A Consistent Conditional Moment Test of Functional Form", *Econometrica* 58, 1443-1458.

Bierens, H. J. (1994): *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*, Cambridge: Cambridge University Press

Bierens, H. J. and W. Ploberger (1997): "Asymptotic Theory of Integrated Conditional Moment Tests", *Econometrica* 65, 1129-1151.

Boning, W. B. and F. Sowell (1999): "Optimality for the Integrated Conditional Moment Test", *Econometric Theory* 15, 710-718.

De Jong, R. M. (1996): "On the Bierens Test Under Data Dependence", *Journal of Econometrics* 72, 1-32.

Stinchcombe, M. B., and H. White (1998): "Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative", *Econometric Theory* 14, 295-325.

The Inverse of a Partitioned Matrix

Herman J. Bierens

September 6, 2014

Consider a pair A, B of $n \times n$ matrices, partitioned as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where A_{11} and B_{11} are $k \times k$ matrices. Suppose that A is nonsingular and $B = A^{-1}$. In this note it will be shown how to derive the B_{ij} 's in terms of the A_{ij} 's, given that

$$\det(A_{11}) \neq 0 \text{ and } \det(A_{22}) \neq 0. \quad (1)$$

If $B = A^{-1}$ then

$$\begin{aligned} AB &= \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} & (2) \\ &= \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix} \\ &= \begin{pmatrix} I_k & O_{k,n-k} \\ O_{n-k,k} & I_{n-k} \end{pmatrix}, \end{aligned}$$

where as usual I denotes the unit matrix and O a zero matrix, with sizes indicated by the subscripts involved.

To solve (2), we need to solve four matrix equations:

$$A_{11}B_{11} + A_{12}B_{21} = I_k \quad (3)$$

$$A_{11}B_{12} + A_{12}B_{22} = O_{k,n-k} \quad (4)$$

$$A_{21}B_{11} + A_{22}B_{21} = O_{n-k,k} \quad (5)$$

$$A_{21}B_{12} + A_{22}B_{22} = I_{n-k} \quad (6)$$

It follows from (4) and (5) that

$$B_{12} = -A_{11}^{-1} A_{12} B_{22}, \quad (7)$$

$$B_{21} = -A_{22}^{-1} A_{21} B_{11}, \quad (8)$$

so that (3) and (6) become

$$\begin{aligned} (A_{11} - A_{12} A_{22}^{-1} A_{21}) B_{11} &= I_k \\ (A_{22} - A_{21} A_{11}^{-1} A_{12}) B_{22} &= I_{n-k} \end{aligned}$$

Hence

$$\begin{aligned} B_{11} &= (A_{11} - A_{12} A_{22}^{-1} A_{21})^{-1} \\ B_{22} &= (A_{22} - A_{21} A_{11}^{-1} A_{12})^{-1} \end{aligned}$$

Substituting these solutions in (7) and (8) it follows that

$$\begin{aligned} B_{12} &= -A_{11}^{-1} A_{12} (A_{22} - A_{21} A_{11}^{-1} A_{12})^{-1} \\ B_{21} &= -A_{22}^{-1} A_{21} (A_{11} - A_{12} A_{22}^{-1} A_{21})^{-1} \end{aligned}$$

Thus,

$$A^{-1} = \begin{pmatrix} (A_{11} - A_{12} A_{22}^{-1} A_{21})^{-1} & -A_{11}^{-1} A_{12} (A_{22} - A_{21} A_{11}^{-1} A_{12})^{-1} \\ -A_{22}^{-1} A_{21} (A_{11} - A_{12} A_{22}^{-1} A_{21})^{-1} & (A_{22} - A_{21} A_{11}^{-1} A_{12})^{-1} \end{pmatrix}$$

Moreover, since $A \cdot A^{-1} = I_n$ implies $A^{-1}A = I_n$, we also have

$$A^{-1} = \begin{pmatrix} (A_{11} - A_{12} A_{22}^{-1} A_{21})^{-1} & - (A_{11} - A_{12} A_{22}^{-1} A_{21})^{-1} A_{12} A_{22}^{-1} \\ - (A_{22} - A_{21} A_{11}^{-1} A_{12})^{-1} A_{21} A_{11}^{-1} & (A_{22} - A_{21} A_{11}^{-1} A_{12})^{-1} \end{pmatrix}$$

The matrix norm $\|A\| = \sqrt{\text{trace}(AA')}$

Herman Bierens

November 16, 2009

Let A be an $k \times m$ matrix. Define the matrix norm $\|\cdot\|$ by

$$\|A\| = \sqrt{\text{trace}(AA')} \quad (1)$$

Recall that a norm needs to satisfy three conditions

$$\|A\| = 0 \text{ if and only if } A = O, \quad (2)$$

$$\|c.A\| = |c|.\|A\| \text{ for any scalar } c, \quad (3)$$

$$\|A + B\| \leq \|A\| + \|B\| \text{ (triangular inequality),} \quad (4)$$

where of course in the latter case the matrix B has the same size as A . Conditions (2) and (3) follow trivially from (1), but condition (4) is not obvious and will be shown below. We also have

$$\|A\| = \|A'\| \quad (5)$$

because $\text{trace}(AA') = \text{trace}(A'A)$. Moreover, it will be shown that for conformable matrices A and B ,

$$\|AB\| \leq \|A\|.\|B\| \quad (6)$$

and thus for $b \in \mathbb{R}^m$,

$$\|Ab\| \leq \|A\|.\|b\|.$$

Proof of (4): To check the triangular inequality, let B be a $k \times m$ matrix. Then

$$\|A + B\|^2 = \text{trace}((A + B)(A' + B'))$$

$$\begin{aligned}
&= \text{trace}(AA') + \text{trace}(AB') + \text{trace}(BA') + \text{trace}(BB') \\
&= \|A\|^2 + 2.\text{trace}(B'A) + \|B\|^2 \\
&\leq \|A\|^2 + 2.\|A\|\cdot\|B\| + \|B\|^2 \\
&= (\|A\| + \|B\|)^2
\end{aligned}$$

which implies (4). The inequality follows from Schwarz inequality, applied twice, as follows. Let $a_{i,j}$ and $b_{i,j}$ be the typical elements of A and B , respectively. Then

$$\begin{aligned}
|\text{trace}(B'A)| &= \left| \sum_{j=1}^m \sum_{i=1}^k b_{j,i} a_{i,j} \right| \leq \sum_{j=1}^m k \left| \frac{1}{k} \sum_{i=1}^k b_{j,i} a_{i,j} \right| \\
&\leq \sum_{j=1}^m k \sqrt{\frac{1}{k} \sum_{i=1}^k b_{j,i}^2} \sqrt{\frac{1}{k} \sum_{i=1}^k a_{i,j}^2} = \sum_{j=1}^m \sqrt{\sum_{i=1}^k b_{j,i}^2} \sqrt{\sum_{i=1}^k a_{i,j}^2} \\
&\leq \sqrt{\sum_{j=1}^m \sum_{i=1}^k b_{j,i}^2} \sqrt{\sum_{j=1}^m \sum_{i=1}^k a_{i,j}^2} = \sqrt{\text{trace}(BB')} \sqrt{\text{trace}(AA')} \\
&= \|B\| \cdot \|A\|
\end{aligned}$$

Proof of (6): Let B be an $m \times \ell$ matrix. Then

$$\|AB\|^2 = \text{trace}(ABB'A') = \text{trace}((BB')(A'A))$$

We can write $BB' = \sum_{i=1}^m \lambda_i q_i q_i'$, where the λ_i 's are the (nonnegative) eigenvalues of BB' and the q_i 's are the corresponding orthonormal eigenvectors. Note that $\sum_{i=1}^m q_i q_i' = I_m$ and $\text{trace}(BB') = \sum_{i=1}^m \lambda_i \text{trace}(q_i q_i') = \sum_{i=1}^m \lambda_i (q_i' q_i) = \sum_{i=1}^m \lambda_i$. Then

$$\begin{aligned}
\|AB\|^2 &= \text{trace}((BB')(A'A)) \\
&= \sum_{i=1}^m \lambda_i \text{trace}(q_i q_i' A' A) = \sum_{i=1}^m \lambda_i q_i' A' A q_i \\
&\leq \sum_{i=1}^m \lambda_i \sum_{i=1}^m q_i' A' A q_i \\
&= \text{trace}(BB') \sum_{i=1}^m \text{trace}(A' A q_i q_i')
\end{aligned}$$

$$\begin{aligned}
&= \text{trace}(BB').\text{trace}\left(A'A \sum_{i=1}^m q_i q'_i\right) \\
&= \text{trace}(BB').\text{trace}(A'A) \\
&= \text{trace}(BB').\text{trace}(AA') = \|B\|^2 \cdot \|A\|^2
\end{aligned}$$