

# Cointegration

Rob Hayward

June 20, 2013

## 1 Introduction

This paper will examine time series with a focus on stationarity, integration and cointegration. The first part discusses tests for stationarity, the second looks at cointegration and the methods used to identify cointegrated series.

## 2 Stationary data

A standard series to be investigated can take the form of

$$y_t = TD_t + z_t \quad (1)$$

Where  $y_t$  is the series of attention,  $TD_t$  is the deterministic component that takes the form of  $TD_t = \beta_0 + \beta_1 t$  and  $z_t$  is the stochastic part that is assumed to be an autoregressive-moving average process of the form  $\Phi L(z_t) = \theta L(z_t)\varepsilon_t$ , with  $\varepsilon_t \sim iid$ .

It is possible to differentiate between *trend stationary*

$$y_t = y_{t-1} + \mu = y_0 + \mu t \quad (2)$$

and *difference stationary* processes.

$$y_t = y_{t-1} + \varepsilon = y_0 + \sum_{i=0}^t \varepsilon_i \quad (3)$$

If all the roots of the autoregressive polynomial  $\phi_p(z)$  lie outside the unit circle, the process is stationary (possibly trend stationary); if at least one of the roots lies on the unit circle and there is a unit root and then the process is difference stationary.

$$\phi_p(z) = 1 - \phi_1(z) - \phi_2(z)^2 - \phi_3(z)^3 \dots \phi_p(z)^p \quad (4)$$

It is possible to create and plot these different types of time series.

```
set.seed(123456)
e <- rnorm(500)
rw.nd <- cumsum(e)
```

After setting the seed and generating 500 normal random variables ( $e$ ).

```
trd <- 1:500
```

The *random walk* ( $rw.nd$ ) is the cumulation ( $cumsum(e)$ ) of the normal random variable.

```
rw.wd <- 0.5 * trd + cumsum(e)
```

By creating a trend ( $trd$ ) a *random walk with drift* can be established with a combination of the cumulative shock and a constant drift ( $rw.wd$ ).

```
dt <- e + 0.5 * trd
```

A *deterministic trend with noise* ( $dt$ ) combines the trend ( $trd$ ) with noise ( $e$ ). Now plot the three series.

```
par(mar = rep(5, 4))
plot.ts(dt, lty = 1, ylab = "", xlab = "")
lines(rw.wd, lty = 2)
par(new = T)
plot.ts(rw.nd, lty = 3, axes = FALSE)
axis(4, pretty(range(rw.nd)))
lines(rw.nd, lty = 3)
legend(10, 18.7, legend = c("det. trend + noise (ls)", "rw drift (ls)", "rw (rs)"),
      lty = c(1, 2, 3))
```

There are also a series of tests that can be used to determine the nature of the time series. There are three types of stationary series to be identified: *trend stationary*, *difference stationary* and *difference stationary with drift*.

## 2.1 Dickey-Fuller Tests

Equation 5 can be used to estimate all three types of series.

$$y_t = \beta_1 + \beta_2 t + \rho y_{t-1} + \sum_{j=1}^k \gamma_j \Delta y_i + u_{1t} \quad (5)$$

However, rather than testing the unit root as  $\rho$  being equal to unity, it is more usual to take  $y_{t-1}$  is taken from each side to produce the following adaption of Equation 5.

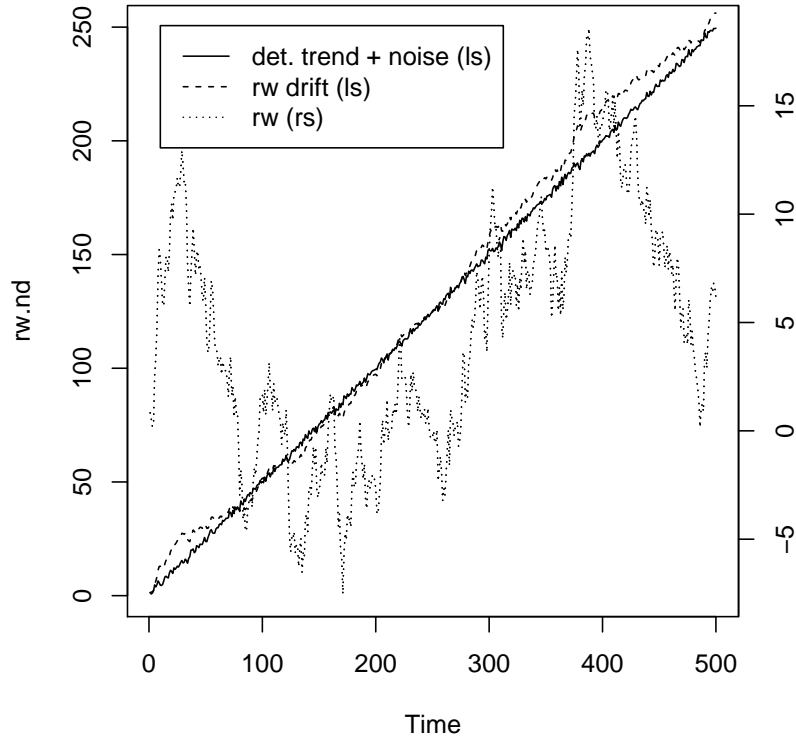


Figure 1: Three Series

$$\Delta y_t = \beta_1 + \beta_2 t + \pi y_{t-1} + \sum_{j=1}^k \gamma_j \Delta y_i + u_{1t} \quad (6)$$

where  $\pi = 1 - \rho$  and therefore if  $\pi$  is significantly different from zero,  $\rho$  cannot be one and there is no unit root.

Lags of the dependent variable are used to remove any serial correlation in the residuals. *Information Criteria* and t-statistics can be used to assess the appropriate number of lags

Using the usca package and the ur.df function on UK real consumer spending data (lc). Set up the data as a timeseries.

```
library(urca)

## Warning: package 'urca' was built under R version 2.15.3

library(xtable)
data(Raotbl3)
lc <- ts(Raotbl3$lc, start = c(1966, 4), end = c(1991, 2), frequency = 4)
```

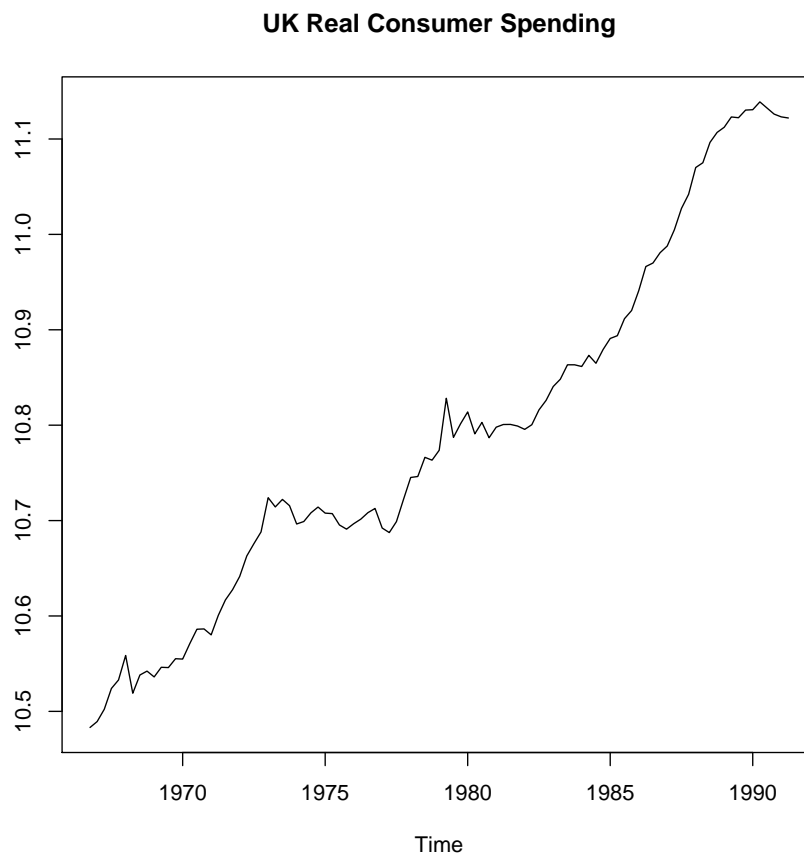


Figure 2: Log UK Consumer Spending

Conduct the Augmented Dickey-Fuller test on (lc.ct) trend and (lc.co) drift using three lags.

```
lc.ct <- ur.df(lc, lags = 3, type = "trend")
lc.co <- ur.df(lc, lags = 3, type = "drift")
```

The three different equations are tested by 'trend', 'drift' or 'none' and there are two tests that take place.

The first ( $\tau_3$ ) tests whether  $\pi$  is equal to zero. The test is the usual t-value on the lagged dependent variable. This can be seen in the `summary()` function or the SlotName "teststat". The critical values for the test statistics are in the slotName "cval".

$$\Delta y_t = \beta_0 + \beta_1 t + \pi y_{t-1} + \sum_{j=1}^k \gamma_j \delta y_{t-j} \quad (7)$$

The following code extracts the relevant values and puts them into a table.

```
a <- cbind(t(lc.ct@teststat), lc.ct@cval)
print(xtable(a, digits = 2, caption = "DW and F-tests"))
```

	statistic	1pct	5pct	10pct
tau3	-2.24	-4.04	-3.45	-3.15
phi2	3.74	6.50	4.88	4.16
phi3	2.60	8.73	6.49	5.47

Table 1: DW and F-tests

The  $\tau_3$  test statistic is the test of the null hypothesis that the coefficient on the difference of the lagged dependent variable is equal to zero and that there is a *unit root* as  $\rho$  is equal to one.

The critical value for a sample size of 100 comes from (Fuller 1976).

An F-test of the null hypothesis that the coefficients on the lagged change in the dependent variable and the coefficient on the time trend are jointly equal to zero is also supplied ( $\phi_3$ ). The critical values come from Table VI (Dickey & Fuller 1981) testing the null  $(\alpha, \beta, \rho) = (\alpha, 0, 1)$ . It seems that unit root and lack of time trend cannot be rejected. A joint test of the null that the coefficients on the drift, time trend and lagged difference of the dependent variable is supposed in ( $\phi_2$ ). The critical values come from Table V (Dickey & Fuller 1981) testing the null  $(\alpha, \beta, \rho) = (0, 0, 1)$ .

As the consumption series does not appear to be trend stationary, a test without the trend can be carried out. This is equivalent to setting  $\beta_2$  in Equation 5 to zero. `lc.co` is the test of the series with drift.

The F-Test is of the hypothesis  $(\beta_1, \beta_2, \pi) = (\beta_1, 0, 0)$ . It is conducted by imposing restrictions on the equation and assesses whether these restrictions have a significant effect on the residuals. The null hypothesis is that the restrictions hold.

```
a <- cbind(t(lc.co@teststat), lc.co@cval)
print(xtable(a, digits = 2, caption = "DW and F-tests 2"))
```

	statistic	1pct	5pct	10pct
tau2	-0.09	-3.51	-2.89	-2.58
phi1	2.88	6.70	4.71	3.86

Table 2: DW and F-tests 2

The critical value of 2.88 ( $\phi_1$ ) is a test of the null that the coefficients on the drift and lagged difference of the dependend variable are jointly equal to zero. This cannot be rejected. Therefore, it seems that the log of UK consumer spending is a random walk.

To complete the picture, the change in consumer spending is tested to make sure that the series are I(1) rather than I(2). First create the difference series lc2.ct.

```
lc2 <- diff(lc)
lc2.ct <- ur.df(lc2, type = "trend", lags = 3)
```

```
a <- cbind(t(lc2.ct@teststat), lc2.ct@cval)
print(xtable(a, digits = 2, caption = "DW and F-tests 3"))
```

	statistic	1pct	5pct	10pct
tau3	-4.39	-4.04	-3.45	-3.15
phi2	6.45	6.50	4.88	4.16
phi3	9.62	8.73	6.49	5.47

Table 3: DW and F-tests 3

This shows that the null of a unit root can be rejected and indicates that the UK consumer spending data are difference stationary.

## 2.2 KPSS

There are a number of other tests of a unit root in the Bernhard Pfaff text (pages 94 to 102). These include the *Phillips-Peron*, *Elliot-Rootenbergs-Stock* and *Schmidt-Phillips* tests which are implemented by ur.pp, ur.ers and ur.sp respectively in the urca package. However, these all test the null of a unit root. The *Kwiatkowski-Phillips-Schmidt-Shin Test* (Kwiatkowski et al. 1992) tests the null stationarity. This is a much more powerful test and can be used in conjunction with the more conventional tests. If the other tests suggest a unit root but the KPSS rejects a unit root, it is probably best to consider the data as stationary.

The KPSS test is of the form

$$y_t = \zeta t + r_t + \varepsilon_t \quad (8)$$

$$r_t = r_{t-1} + u_t \quad (9)$$

The test statistic is calculated by running the regression of  $y$  on a constant and trend as in Equation 8 or on just a constant as in equation 8 with  $\zeta$  equal to zero.

$$LM = \frac{\sum_{i=1}^T S_t^2}{\hat{\sigma}_t^2} \quad (10)$$

where

$$S_t = \sum_{i=1}^t \hat{\varepsilon}_i, t = 1, 2, \dots, T \quad (11)$$

and the estimate of the error variance

$$\hat{\sigma}_\varepsilon^2 = s^2(l) = T^{-1} \sum_{t=1}^T \varepsilon_t^2 + 2T - 1 \sum_{s=1}^l 1 - \frac{s}{l+1} \sum_{t=s+1}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-1} \quad (12)$$

Using the `urca` package and the data for US interest rates and nominal wages, the KPSS test is either on level stationary ( $\text{type} = \mu$ ) or trend stationary ( $\text{type} = \tau$ ) and the lags for the error term are either specified (as below) or set to "short"  $\sqrt[4]{4 \times (n/100)}$  or "long"  $\sqrt[4]{12 \times (n/100)}$ .

```
data(nporg)
ir <- na.omit(nporg[, "bnd"])
wg <- log(na.omit(nporg[, "wg.n"]))
```

Plot the data

```
par(mfrow = c(2, 1))
plot.ts(ir, main = "US interest rates")
plot.ts(wg, main = "US nominal wages")
```

```
ir.kpss <- ur.kpss(ir, type = "mu", use.lag = 8)
wg.kpss <- ur.kpss(wg, type = "tau", use.lag = 8)
```

And the appropriate data can be extracted and placed into a table using the following.

```
a <- cbind(ir.kpss@teststat, ir.kpss@cval)
b <- cbind(wg.kpss@teststat, wg.kpss@cval)
ab <- rbind(a, b)
colnames(ab) <- c("CV", "10pct", "5pct", "2.5pct", "1.0pct")
rownames(ab) <- c("ir", "wg")
print(xtable(ab, digits = 2, caption = "KPSS and critical values"))
```

This shows that the null hypothesis of level stationarity for the interest rate series and trend stationarity for the wage series cannot be rejected.

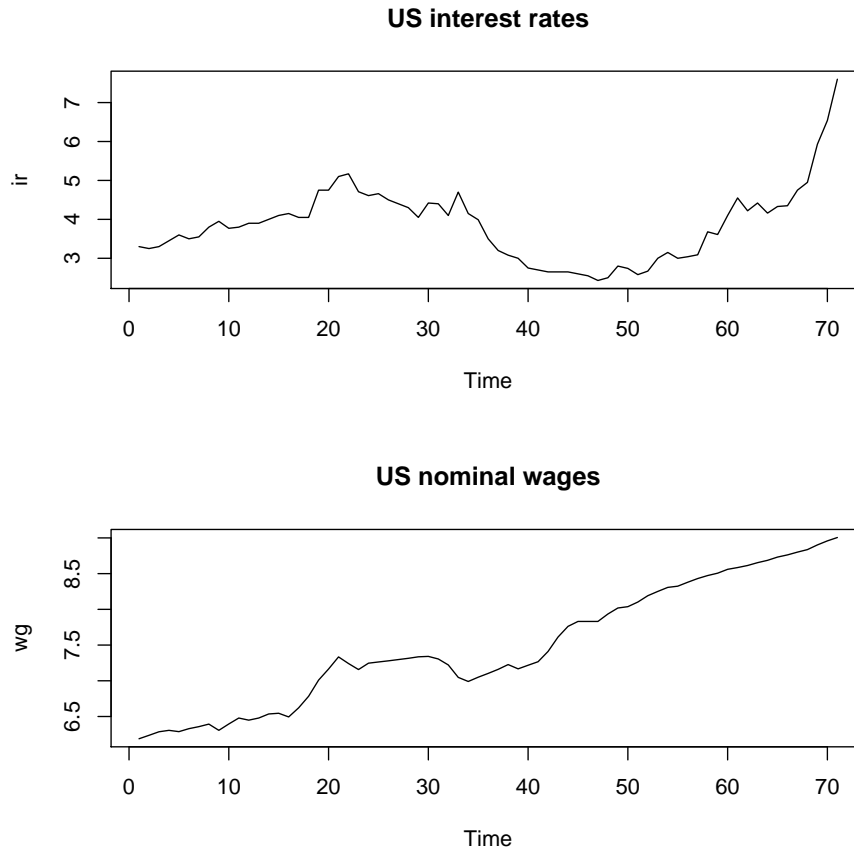


Figure 3: US interest rate and wage data

## 2.3 Dealing with lack of stationarity

If the data are trend-stationary, one way to deal with the lack of stationarity would be to remove the trend. One method is described in the footnote on page 53 of Pfaff. This takes the residuals from a regression of a series that is the same length as the log of consumption.

```
detrended <- residuals(lm(lc ~ seq(along = lc)))
```

Which takes the following form.

```
plot(detrended, type = "l", main = "De-trended series")
```



	CV	10pct	5pct	2.5pct	1.0pct
ir	0.13	0.35	0.46	0.57	0.74
wg	0.10	0.12	0.15	0.18	0.22

Table 4: KPSS and critical values

### 3 Cointegration

This is the overview of cointegration and the methods use to analyse cointegrated relationships. Non-stationary data may exhibit *spurious regression*. If two normal random variables are created (e1 and e2) and two series (y1 and y2) have a trend plus a random shock.

```
library(lmtest)
library(xtable)
set.seed(123456)
e1 <- rnorm(500)
e2 <- rnorm(500)
trd <- 1:500
y1 <- 0.8 * trd + cumsum(e1)
y2 <- 0.6 * trd + cumsum(e2)
```

Now plot the two series.

```
plot(y1, type = "l", main = "Plot of y1 and y2", lty = 1, ylab = "y1, y2")
lines(y2, lty = 2)
legend("topleft", legend = c("y1", "y2"), lty = c(1, 2))
```

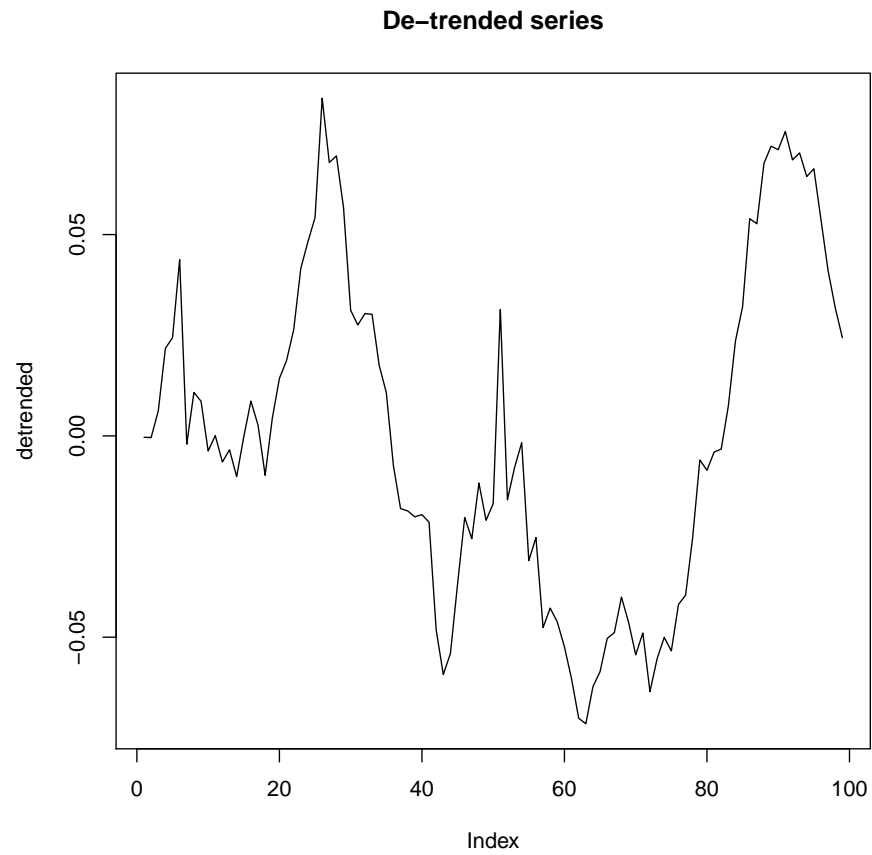


Figure 4: Detrended Plot



Run a regression of  $y_1$  on  $y_2$  and it appears that there is a strong relationship.

```
sr.reg <- lm(y1 ~ y2)
print(xtable(sr.reg, caption = "Regresson results"))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-29.3270	1.3672	-21.45	0.0000
y2	1.4408	0.0075	191.62	0.0000

Table 5: Regresson results

However, the Durbin-Watson statistics shows there is a large amount of auto-correlation in the residuals.

```
sr.dw <- dwtest(sr.reg)$statistic
sr.dw

##      DW
## 0.01715
```

The statistic will be around 2 if there is no autocorrelation. As a general rule, there are groups for suspicion if the  $R^2$  is larger than the Durbin-Watson statistic.

The main idea of cointegration is that a combination of one or more non-stationary variables will show a stationary relationship. Pfaff provides the following definition.

“The components of a vector  $\mathbf{x}_t$ , are said to be cointegrated of order  $b, d$ ; denoted  $x \sim CI(b, d)$  if (a) all components of  $x_t$  are  $I(d)$  and (b) a vector  $\alpha (\neq 0)$  exists so that  $z_t = \alpha' x_t \sim I(d - b), b > 0$ . The vector  $\alpha$  is called the cointegrating vector.”

(Pfaff 2008, p. 75)

If two or more non-stationary series are cointegrated, a linear combination of the two may be cointegrated and this combination can be included in the regression. The aim is to have a system of the form

$$\Delta y_t = \psi_0 + \gamma_1 z_{t-1} + \sum_{i=1}^k \psi_i \Delta x_{t-i} + \sum_{i=1}^k \psi_i \Delta y_{t-i} + \varepsilon_{1,t} \quad (13a)$$

$$\Delta x_t = \psi_0 + \gamma_1 z_{t-1} + \sum_{i=1}^k \psi_i \Delta x_{t-i} + \sum_{i=1}^k \psi_i \Delta y_{t-i} + \varepsilon_{2,t} \quad (13b)$$

Where  $z$  is the cointegrated relationship.  $y$  and  $x$  are difference stationary. One way to estimate this model is to use the two-step *Engle-Granger* method (Engle & Granger 1987).

For an example of this, create two non-stationary series ( $y_1$ ) and ( $y_2$ ) with a long-run relationship where  $y_2$  is equal to 0.6  $y_1$ .

```
set.seed(123456)
e1 <- rnorm(100)
e2 <- rnorm(100)
y1 <- cumsum(e1)
y2 <- 0.6 * y1 + e2
```

Plot these series.

```
plot(y1, type = "l", lty = 1, main = "Plot y1 and y2")
lines(y2, lty = 2)
legend("topleft", legend = c("y1", "y2"), lty = c(1, 2))
```

Now run the regression on the long-run relationship and save the residuals from that regression. The residuals are the deviations from the long run relationship.

```
lr.reg <- lm(y2 ~ y1)
error <- residuals(lr.reg)
```

The residual show the divergence from the long run relationship between  $y_1$  and  $y_2$ .

```
plot(error, type = "l", main = "Divergence from long-run y1-y2 relationship")
```

Now create the lagged error term and differences in  $y_1$  and  $y_2$  to allow each variable to respond to the deviation from the long-run relationship. The `embed()` function will created the lagged dataframe.

```
error.lagged <- error[-c(1, 100)]
dy1 <- diff(y1)
dy2 <- diff(y2)
diff.dat <- data.frame(embed(cbind(dy1, dy2), 2))
colnames(diff.dat) <- c("dy1", "dy2", "dy1.1", "dy2.1")
```

```
ecm.reg <- lm(dy2 ~ error.lagged + dy1.1 + dy2.1, data = diff.dat)
print(xtable(summary(ecm.reg), caption = "Engle-Granger Regression Result"))
```

The results show that most of the disturbance from equilibrium is corrected swiftly with the coefficient on the lagged error at 0.97.

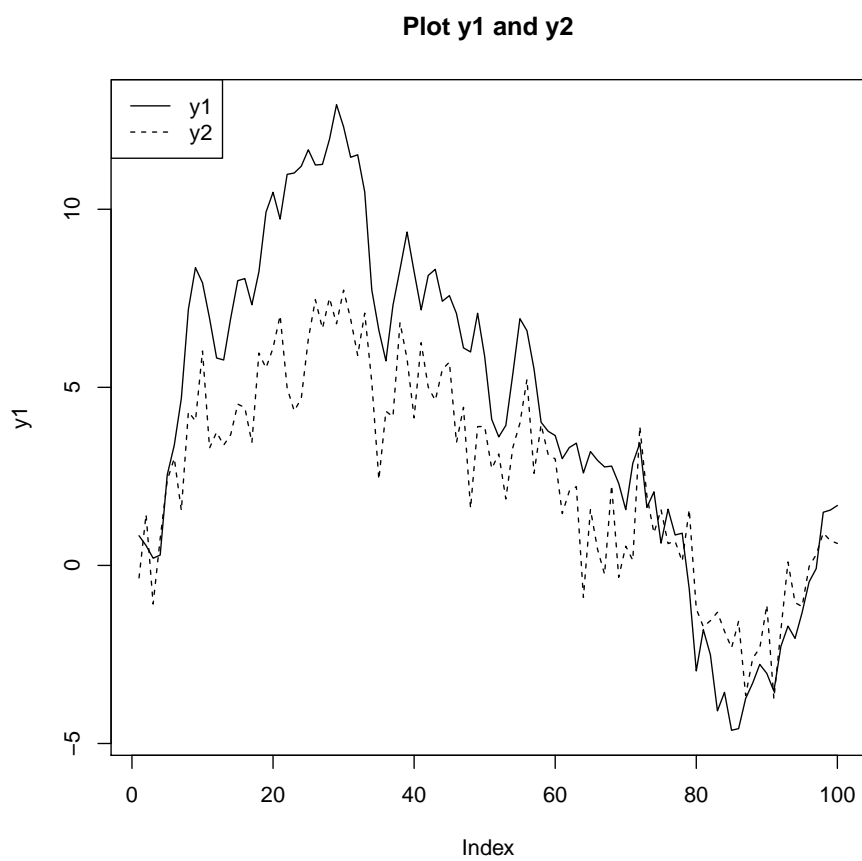


Figure 5: Plot y1 and y2

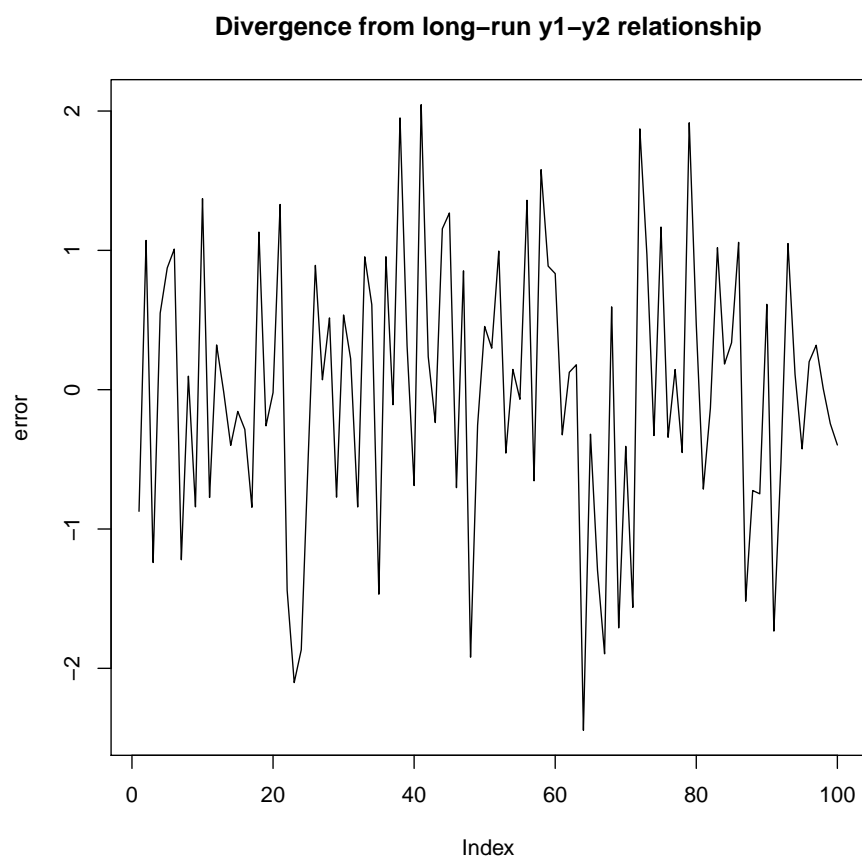


Figure 6: Plot Error

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.0034	0.1036	0.03	0.9739
error.lagged	-0.9688	0.1586	-6.11	0.0000
dy1.1	0.2456	0.1270	1.93	0.0561
dy2.1	-0.0901	0.1059	-0.85	0.3971

Table 6: Engle-Granger Regression Result

### 3.1 Use of cointegration

These cointegration techniques can be used to test whether there is a stable relationship between variables. For example, one test could be whether cash and futures prices are cointegrated. A Dickey-Fuller test cannot be made of the residuals of a regression of cash on futures because the process of creating the residuals through the regression will bias the result by selecting parameters that reduce the variability of the estimated errors as much as possible. Therefore, the Engle-Granger residuals should be used. These are interpolations from the response surface in (MacKinnon 1996). These are in Table C (Enders 2010, p. 490).

If the tests indicate that the residuals from the regression are stationary, the coefficient from this regression gives the relationship between cash and futures. Similarly, the efficiency of pairs trades can be assessed with the error-correction model. This also gives an indication of how swiftly the relationship will return to normal after a disturbance.

This technique could also be used to make a pairs test of two securities. The cointegration is the relationship between the two and the error correction can show the speed at which the relationship will return to the long run equilibrium. Alternatively, an analysis of the bond market could use the yield curve (long term and short term interest rates) as a cointegrating factor.

### 3.2 Example

This example uses the artificial data from the Enders textbook (Enders 2010, pp. 377-379). The simulated series are created in the following way.

	$(y_t)$	$(z_t)$	$(w_t)$
Trend	$\mu_{yt} = \mu_{yt} + \varepsilon_{zt}$	$\mu_{zt} = \mu_{zt} + \varepsilon_{zt}$	$\mu_{wt} = \mu_{yt} + \mu_{zt}$
Irregular	$\delta_{yt} = 0.5\delta_{yt-1} + \eta_{yt}$	$\delta_{zt} = 0.5\delta_{zt-1} + \eta_{zt}$	$\delta_{wt} = 0.5\delta_{wt-1} + \eta_{wt}$
Series	$y_t = \mu_{yt} + \delta_{yt}$	$z_t = \mu_{zt} + \delta_{zt}$	$w_t = \mu_{wt} + \delta_{wt} + 0.5\delta_{yt} + 0.5\delta_{zt}$

The data are available on the Enders companion web site. Load the data and plot.

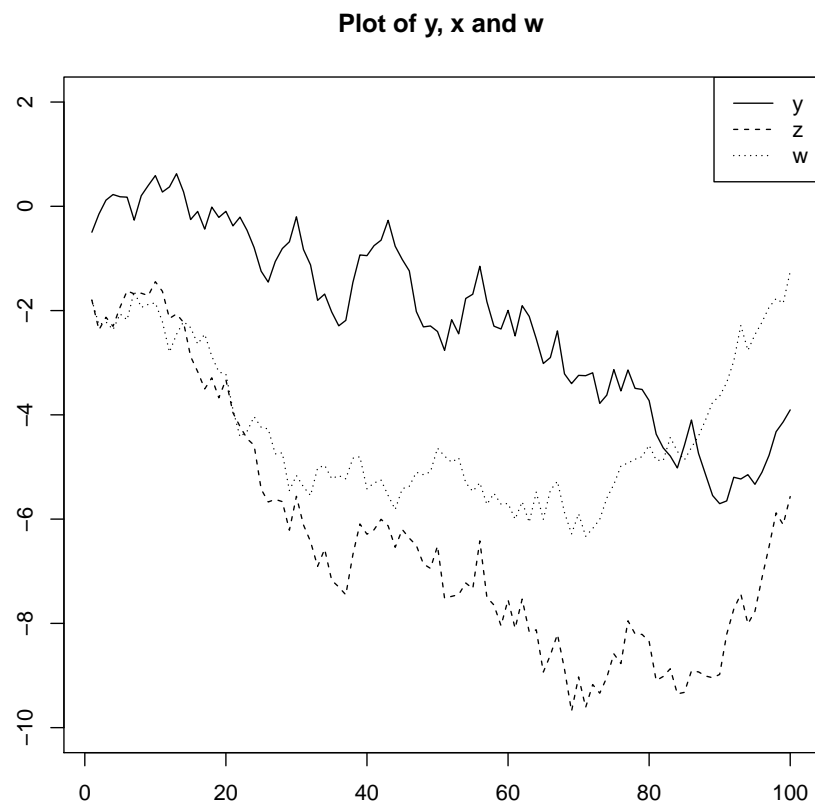
```
da <- read.csv("coint6.csv", header = TRUE)
head(da)

##           y           z           w
```

```
## 1 -1.802 -0.4970 -1.800
## 2 -2.270 -0.1491 -2.370
## 3 -2.217  0.1182 -2.128
## 4 -2.366  0.2260 -2.302
## 5 -2.064  0.1829 -1.934
## 6 -2.178  0.1755 -1.622
```

Plot data

```
plot(da$z, lty = 1, ylab = "", xlab = "", type = "l", main = "Plot of y, x and w",
      ylim = c(-10, 2))
lines(da$w, lty = 2)
lines(da$y, lty = 3)
legend("topright", legend = c("y", "z", "w"), lty = c(1, 2, 3))
```



Now test whether the series are stationary using the augmented Dickey-Fuller test.



```
eq1 <- lm(da$y ~ da$z + da$w)
print(xtable(summary(eq1), caption = "Least Squares y on z and w"))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.0484	0.0842	-0.58	0.5666
da\$z	-0.9273	0.0243	-38.10	0.0000
da\$w	0.9769	0.0183	53.46	0.0000

Table 7: Least Squares y on z and w

This can be repeated for each of the three possible relationships and the errors that are collected can be saved and tested for stationarity.

```
resy <- eq1$residuals
eq2 <- lm(da$z ~ da$y + da$w)
resz <- eq2$residuals
eq3 <- lm(da$w ~ da$z + da$y)
resw <- eq3$residuals
```

The test is of the form.

$$\Delta y_t = \alpha_0 + \alpha_1 y_{t-1} + \sum_{i=1}^p \alpha_{i+1} \Delta y_{t-1} + \varepsilon_t \quad (14)$$

	No Lags	4 Lags
e_yt	-0.4430	-0.5953
SE(y)	-5.1749	-4.0742
e_zt	-0.4520	-0.5934
SE(z)	-5.3788	-4.2268
e_wt	-0.4553	-0.6071
SE(w)	-5.3896	-4.2248

Table 8: Estimated  $a_1$

The figures displayed here are the same as those in Table 6.2 of Enders (Enders 2010, p. 379). The critical value comes from MacKinnon. The programme does not appear to give as much flexibility as the talbes in Enders. Sample code below.

```
xtable(unitrootTable(trend = "nc"), digits = 4, caption = "MacKinnon unit root table")
```

	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
25	-2.6610	-2.2720	-1.9550	-1.6090	0.9250	1.3370	1.7010	2.1390
50	-2.6120	-2.2490	-1.9480	-1.6130	0.9060	1.3090	1.6600	2.0730
100	-2.5880	-2.2380	-1.9440	-1.6150	0.8970	1.2960	1.6410	2.0430
250	-2.5740	-2.2320	-1.9420	-1.6160	0.8910	1.2880	1.6300	2.0260
500	-2.5700	-2.2290	-1.9410	-1.6160	0.8900	1.2860	1.6270	2.0210
Inf	-2.5650	-2.2270	-1.9410	-1.6170	0.8880	1.2840	1.6240	2.0150

Table 9: MacKinnon unit root table

### 3.3 Johansen Procedure

When there are more than two series there may still be a relationship that creates a stationary, linear combination that can be used in a regression.

For example, a simulated series can be created in the following fashion. First create the three series and put them into a dataframe.

```
set.seed(12345)
e1 <- rnorm(250, 0, 0.5)
e2 <- rnorm(250, 0, 0.5)
e3 <- rnorm(250, 0, 0.5)
u1.ar1 <- arima.sim(model = list(ar = 0.75), innov = e1, n = 250)
u2.ar1 <- arima.sim(model = list(ar = 0.3), innov = e2, n = 250)
y3 <- cumsum(e3)
y1 <- 0.8 * y3 + u1.ar1
y2 <- -0.3 * y3 + u2.ar1
y.mat <- data.frame(y1, y2, y3)
```

Take a look at the series that have been created.

```
plot(y3, main = "Three series", lty = 3, type = "l")
lines(y2, lty = 1, type = "l")
lines(y1, lty = 2, type = "l")
legend("topleft", legend = c("y1", "y2", "y3"), lty = c(1, 2, 3))
```

```
require(xtable)
vecm <- ca.jo(y.mat)
jo.results <- summary(vecm)
```

```
a <- cbind(jo.results@teststat, jo.results@cval)
colnames(a) <- c("CV", "10pct", "5pct", "1pct")
print(xtable(a, digits = 2, caption = "Johansen Test and Critical Values"))
```

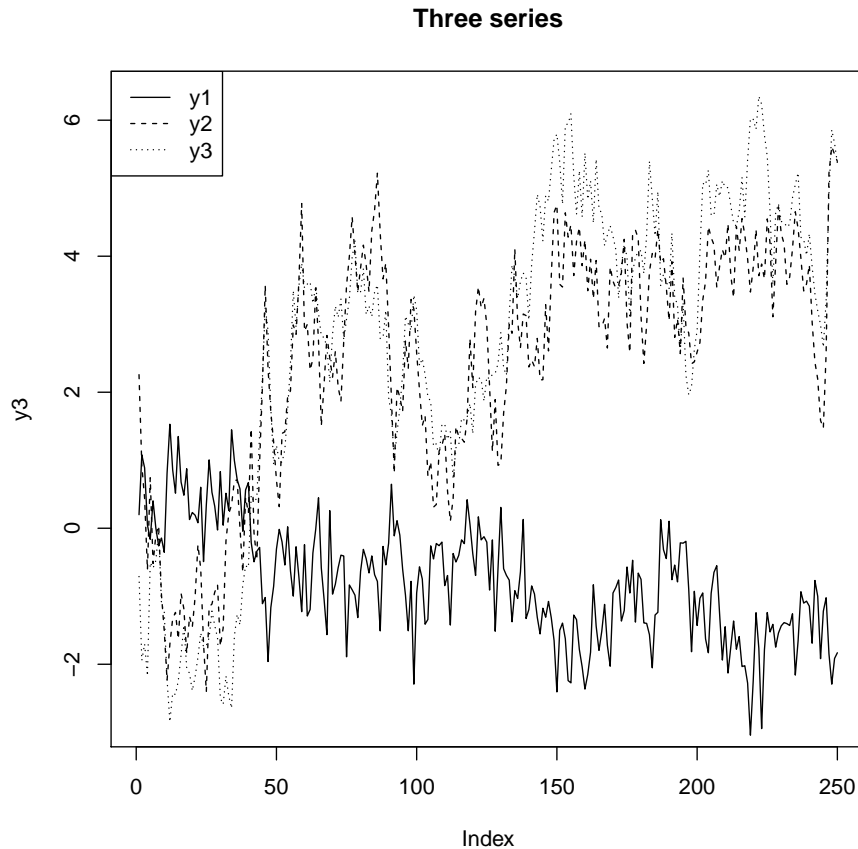


Figure 7: Three Simulated Series

## References

- Dickey, D. & Fuller, W. (1981), 'Likelihood ratio statistics for autoregressive time series with a unit root', *Econometrica* **49**, 1057 – 1072.
- Enders, W. (2010), *Applied Econometric Time Series*, 3 edn, Wiley.
- Engle, R. F. & Granger, C. W. (1987), 'Co-integration and error correction representaiton, estimation and testing', *Econometrica* **55**(2), 251 – 276.
- Fuller, W. (1976), *Introducton to Statistical Time Series*, John Wiley and Sons, New York.
- Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. & Shin, Y. (1992), 'Testing the null hypothesis of stationarity against the alternative of a unit root: How sure

	CV	10pct	5pct	1pct
r $\leq$ 2	4.72	6.50	8.18	11.65
r $\leq$ 1	41.69	12.91	14.90	19.19
r = 0	78.17	18.90	21.07	25.75

Table 10: Johansen Test and Critical Values

are we that economic time series have a unit root?', *Journal of Econometrics* **54**, 159 – 178.

MacKinnon, J. G. (1996), 'Numerical distribution functions for unit root and cointegration tests', *Journal of Applied Econometrics* **11**(6), 601–618.

Pfaff, B. (2008), *Analysis of Integrated and Cointegrated Time Series with R*, second edn, Springer, New York.