Cointegration and Error Correction Model

Hüseyin Taştan[1]

[1]Yıldız Technical University
Department of Economics

Econometrics II

Lecture Plan

- Spurious regression problem when variables are I(1)
- Cointegration
- Engle-Granger cointegration test
- Error Correction Model (ECM)

Spurious Regression

- We have seen spurious regression in a variety of contexts so far.
- In cross-sectional data, for example, results may be spurious if we ignore an important variable. For example, we may find that x has a significant impact on y, but when we add a third variable, z, x becomes insignificant. Thus, the relationship between x and y is spurious.
- As an example consider the recent news in the media that says "going to opera helps people live longer" suggesting a positive correlation between the two variables. But this may not be interpreted as a causal statement as there may be another factor, income level for example, that may be related to both. People going to opera may be predominantly of high-income strata and can get high-quality health care that helps them live longer.

Spurious Regression

- A similar situation may also arise when we run a regression of I(0) variables.
- In time series framework, we discussed the possibility of spurious regression when we have trending variables.
- Ignoring trend may lead to significant relationship when in fact there is none. This problem is similar to the omitted variable bias where the omitted variable is just the trend (or more specifically, the correct specification of trend).
- A similar phenomenon occurs when we have nonstationary variables in a regression model.
- Even though variables may not have a clear trend, results may be spurious if they are highly persistent (think of random walks).

Spurious Regression

- To be more specific, consider the following independent random walks:

$$y_t = y_{t-1} + \epsilon_{1t},$$
$$x_t = x_{t-1} + \epsilon_{2t}$$

where ϵ_{1t} and ϵ_{2t} are two independent white noise processes with means 0, and variances σ_1^2, and σ_2^2, respectively.

- Note that these two variables do not have a trending mean. Consider the regression of y_t on x_t

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

- In this regression, because y_t on x_t are independent by construction, we expect that in the sample regression function

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$$

$plim(\hat{\beta}_1) = 0.$

Spurious Regression

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$$

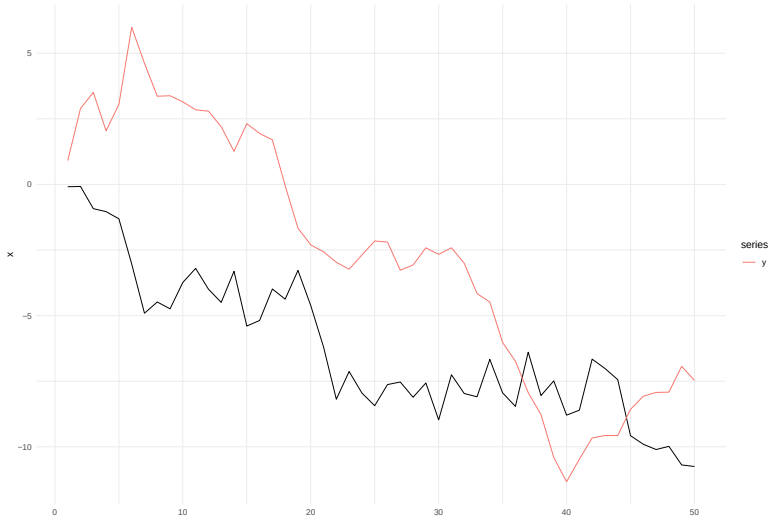
- More specifically, consider the usual significance test where the null and alternative is given by

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$$

using $\alpha = 0.05$ significance level.

- We expect that the t statistic on $\hat{\beta}_1$ will be significant (i.e. reject the null hypothesis) 5% of the time and insignificant 95% of the time.
- Granger and Newbold (1974) showed that t statistic is statistically significant a large percentage of the time, much larger than the nominal significance level (that is α).
- They called this the **spurious regression problem**

A simulation of two independent random walks



A simulation of two independent random walks

Running the regression of y on x we get the following R output

```
> summary( lm(y~x) )

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.5538     1.0926   5.083 6.07e-06 ***
x             1.3232     0.1602   8.261 8.91e-11 ***
--------
Residual standard error: 3.138 on 48 degrees of freedom
Multiple R-squared:  0.5871,Adjusted R-squared:  0.5785
F-statistic: 68.24 on 1 and 48 DF,  p-value: 8.913e-11
```

The coefficient on x is statistically significant although x and y are sampled independently. The output above is obtained from a single simulation. We can repeat the same procedure a large number of times and inspect the behavior of the t-test. This is what we will do next.

Simulating the Spurious Regression Problem

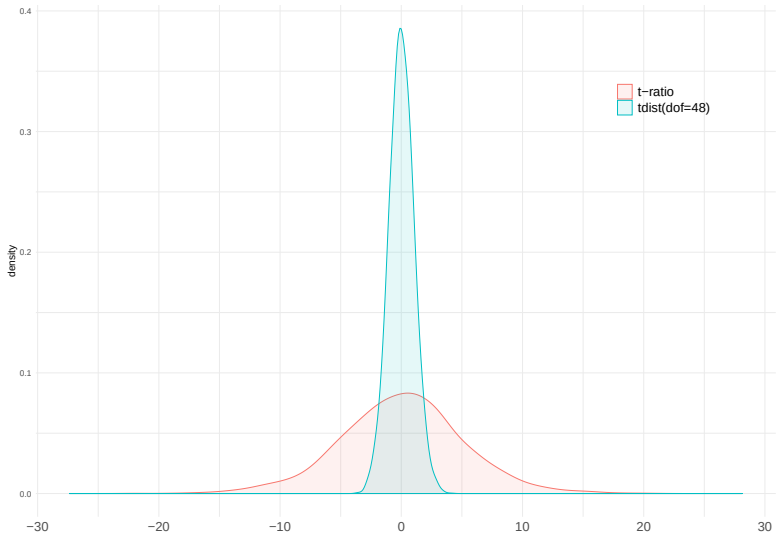
$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$$

1. Generate two independent random walks of size $n = 50$ and run the regression of y_t on x_t.
2. Compute the t-statistic on $\hat{\beta}_1$ and R^2, and save them.
3. Repeat this 10000 times and save t ratios and R^2s
4. Compute the fraction of samples in which t ratio leads to the rejection of the null $H_0 : \beta_1 = 0$

Summary of Results:

- Running the experiment we see that the null is rejected approximately 66% of the time, instead of 5%
- The t statistic does not follow the usual t distribution.
- R^2 tends to be arbitrarily large.

Actual Distribution of the t Statistic



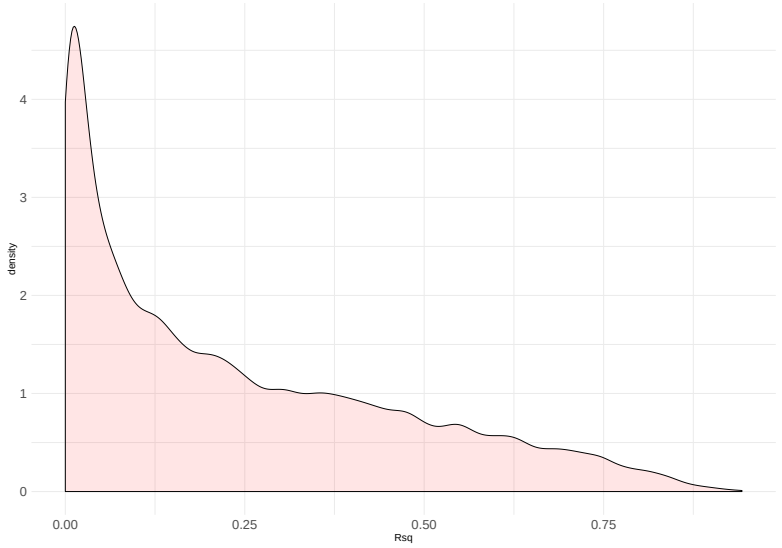
Spurious Regression Problem

- Why do we find significant t statistics more often than implied by the 5% significance level?
- The reason is that under the null hypothesis $H_0 : \beta_1 = 0$ we have

$$y_t = \beta_0 + u_t$$

- Because y_t follows a random walk process, u_t also follows a random walk process. This implies the Gauss-Markov assumptions do not hold.
- The t ratio does not follow the t distribution even asymptotically. This means that the usual decision rule is invalid. As $n \to \infty$, t statistic increases to ∞
- Also, R^2 does not converge to the population R-squared, $1 - \sigma_u^2/\sigma_y^2$. So in practice it can be arbitrarily large.

Distribution of R^2 Under Spurious Regression



Spurious Regression Problem

- The spurious regression problem implies that we should be careful when we have I(1) variables in our regression model.
- A regression model involving I(1) variables may be informative, i.e., not spurious, under certain conditions.
- In particular, under what conditions the models including levels of variables (which are nonstationary) provide us economically meaningful interpretations?
- We learned that when we have I(1) variables we can use first differences in our regression because they will be stationary.
- Although one can follow this strategy, differencing leads to throwing out valuable information regarding the relationship between the levels of the variables. Thus, always differencing may limit the type of questions we can answer.

Cointegration

- The concept of cointegration was introduced by Engle and Granger (1987) (They shared the Nobel prize in economics in 2003 for their contributions to time series econometrics)
- When variables in a regression model are all I(1), i.e., their first differences are stationary, then there may be a meaningful relationship among I(1) variables if they are cointegrated (share a common trend).
- To fix ideas, let $\{y_t : \ t = 1, 2, \ldots\}$ and $\{x_t : \ t = 1, 2, \ldots\}$ be two I(1) variables.
- If there is a nonzero β parameter such that $y_t - \beta x_t$ is stationary then we say the they are cointegrated. In other words,

$$y_t - \beta x_t \ \sim \ I(0), \quad \beta \neq 0$$

- β is called the cointegration parameter.

Cointegration

$$y_t - \beta x_t \ \sim \ I(0), \quad \beta \neq 0$$

- It is possible to write $x_t - (1/\beta) y_t$ which is also I(0).
- This implies that the linear combination of y_t and x_t is not unique.
- To prevent this, we fix the parameter on y_t to unity so that cointegration relationship is unique. Note that the parameter vector is $(1, \ -\beta)^\top$
- If I(1) variables are related in such a way that the regression reflects long-run relationship, in other words, if they are cointegrated, we can be sure that we do not have spurious regression.

Cointegration

$$y_t - \beta x_t \ \sim \ I(0), \quad \beta \neq 0$$

- The cointegration relationship can be interpreted as reflecting a long run equilibrium relationship.
- Cointegrated variables tend to move together
- In the short run, there will be deviations from the economic equilibrium relationship but they will be temporary and short-lived and equilibrium relationship will be attained at a certain speed.
- Examples: The Law of One Price (LOP), Purchasing Power Parity (PPP)

Engle-Granger Cointegration Test

- How do we know if two series are cointegrated?
- Engle-Granger suggested a simple regression-based test for this
- Both y_t and x_t must be I(1) variables. This can be checked by ADF test before the Engle-Granger cointegration test.
- In the first step we apply OLS to estimate the following model

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$$

- If they are cointegrated then the residual $\hat{u}_t = y_t - \hat{\alpha} - \hat{\beta} x_t$ must be I(0).
- Thus, in the second step, we apply ADF test on the residual to see if it is stationary.

Engle-Granger Cointegration Test

- The null hypothesis states that

$$H_0 : u_t \text{ is nonstationary} \quad \text{(NO COINTEGRATION)}$$

  against the alternative

$$H_1 : u_t \text{ is stationary} \quad \text{(COINTEGRATION)}$$

- Note that under the null hypothesis we have a spurious regression.
- If we reject H_0, we say that y_t and x_t are cointegrated. Otherwise, we have spurious regression. In that case, we should take the first differences of variables.

Distribution of the Engle-Granger Cointegration Test

- Can we use the usual ADF critical values in our decision?
- The answer is NO. The fact that we first estimated the parameter vector and then apply the unit root test complicates the asymptotic distribution.
- It can be approximated using simulation. Critical values depend on whether the model has a trend or not as shown in the following tables.

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$$

TABLE 18.4 Asymptotic Critical Values for Cointegration Test: No Time Trend

Significance level	1%	2.5%	5%	10%
Critical value	−3.90	−3.59	−3.34	−3.04

Note that these critical values are larger in absolute value than their ADF counterparts. The decision rule is the same as ADF's.

Distribution of the Engle-Granger Cointegration Test

- When the cointegration relationship includes a time trend we have the following relationship in the first step:

$$\hat{y}_t = \hat{\alpha} + \hat{\eta}\, t + \hat{\beta} x_t$$

- In this case, the appropriate table of critical values are given in the table below.

TABLE 18.5 Asymptotic Critical Values for Cointegration Test: Linear Time Trend

Significance level	1%	2.5%	5%	10%
Critical value	−4.32	−4.03	−3.78	−3.50

Cointegration Example: Fertility Equation

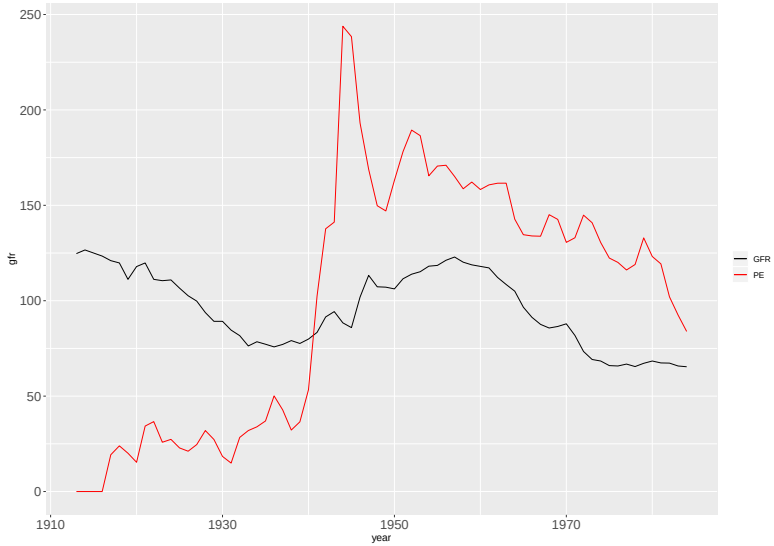
- Consider the following static regression

$$gfr_t = \alpha + \gamma t + \beta pe_t + u_t$$

gfr: gross fertility rate, pe: personal tax exemption

- We have seen this model in our previous classes. Now the question is are they cointegrated? Or, to put it differently, is the relationship spurious or genuine?
- Running the ADF tests we see that both gfr_t and pe_t are nonstationary, i.e., I(1) variables (see the R Lab notes)
- Because the regression above involves I(1) variables there may be a potentially spurious relationship.
- We can apply the Engle-Granger cointegration test to sort this out.

Gross Fertility Rate (GFR) and Personal Exemptions (PE)



Cointegration Example: Fertility Equation

- Running the regression using OLS we obtain the following results

$$gfr_t = 109.9 - 0.91t + 0.19pe_t + \hat{u}_t$$

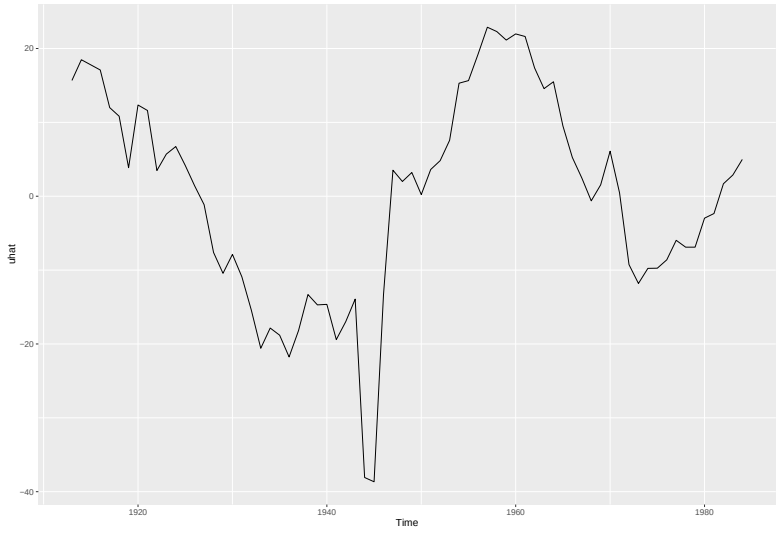
- The Engle-Granger (EG) cointegration test statistic is simply the ADF unit root test statistic for the residuals. The ADF test regression is

$$\widehat{\Delta \hat{u}_t} = -0.18 - 0.12\hat{u}_{t-1} + 0.24\Delta\hat{u}_{t-1}$$
$$(0.671) \quad (.049) \quad (.117)$$

$$EG = \frac{-0.12}{0.049} = -2.43$$

- From Table 18.5 we see that the 10% critical value is -3.50. Because EG is larger than the critical value we **fail to reject** the null hypothesis. There is no cointegration.

Plot of residuals from the regression of gfr on pe



Cointegration Example: Fertility Equation

- The EG test suggests that the gross fertility rate and personal tax exemptions are **not cointegrated**.
- This implies that the static OLS regression in levels suffers from the spurious regression problem. Thus, the results cannot be trusted.
- In such cases, we can always estimate a model after taking the first differences of variables. Because each variable is I(1), their first differences will be I(0).
- However, the new model must be interpreted accordingly (in terms of changes or growth rates).
- For example, we can estimate an FDL(2) model in first differences (see ch.11 for details):

$$\widehat{\Delta \text{gfr}_t} = \underset{(0.468)}{-0.964} - \underset{(0.027)}{0.036}\,\Delta pe_t - \underset{(0.028)}{0.014}\,\Delta pe_{t-1} + \underset{(0.027)}{0.110}\,\Delta pe_{t-2}$$

$$n = 69 \quad R^2 = 0.233$$

Example: Are 3-month and 6-month interest rates cointegrated?

- $r6_t$ annualized interest rate for six-month T-bills (at the end of quarter t)
- $r3_t$ annualized interest rate for three-month T-bills. (These are also known as "bond equivalent yields")
- Both $r6_t$ and $r3_t$ are I(1) variables (according to ADF tests)
- Let $spr_t = r6_t - r3_t$ be the spread between the two rates.
- Because of the simple arbitrage relationship, spr_t will not wander away from its mean value. In other words it will be an I(0) variable.
- If spr_t continues to grow then investors would shift away from three-month and toward six-month T-bills. The price of six-month T-bills will go up. But because interest rates are inversely related to price, this would lower $r6$ and increase $r3$, until the spread is reduced.

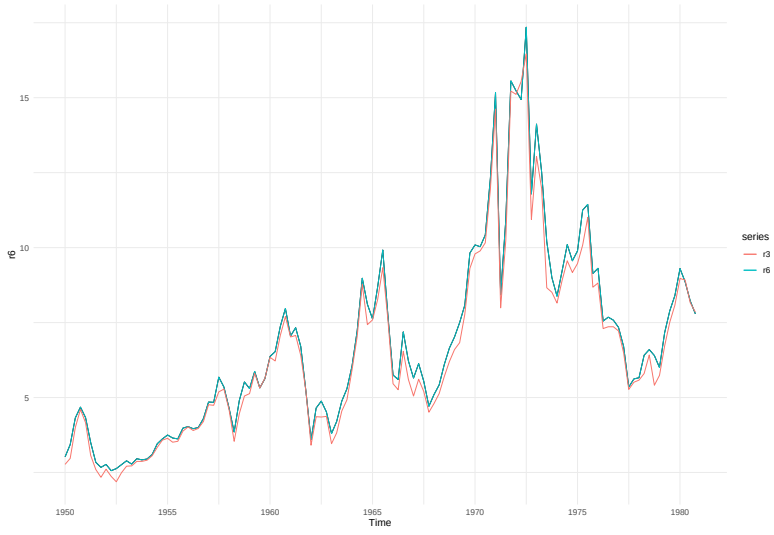
Example: Are 3-month and 6-month interest rates cointegrated?

- The arbitrage argument implies that $r6_t$ and $r3_t$ will be cointegrated in the long run.
- The relationship can be written as

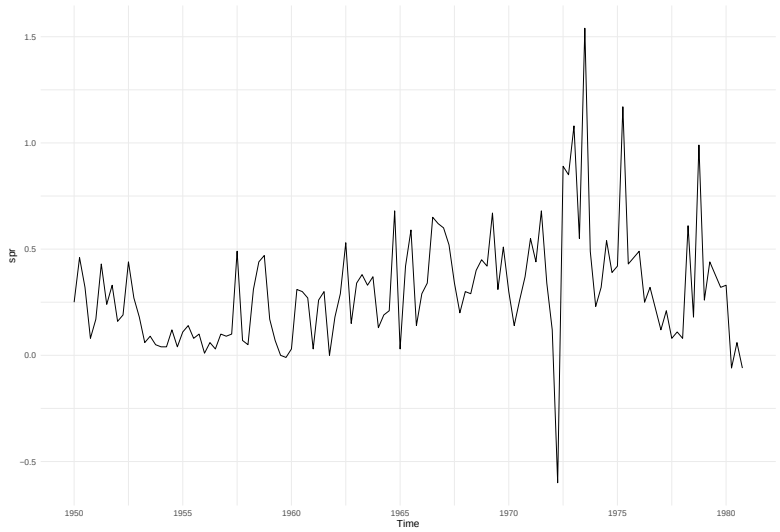
$$r6_t = \alpha + \beta\, r3_t + u_t$$

- Economic theory suggests that they are cointegrated (u_t is I(0)) with $\beta = 1$

Plots of $r6_t$ and $r3_t$



Plot of spread $spr_t = r6_t - r3_t$



Are they cointegrated?

- The regression of $r6_t$ on $r3_t$ produces the following R output

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.135374   0.054867   2.467    0.015
r3          1.025899   0.007709 133.081   <2e-16
```

in equation form

$$\widehat{r6}_t = 0.135 + 1.03 r3_t$$

- The coefficient on $r3_t$ is very close to unity, suggesting a one-to-one relationship as expected.
- Applying the EG test we have

$$\widehat{\Delta \hat{u}_t} = -0.004 - 0.632 \hat{u}_{t-1} - 0.146 \Delta \hat{u}_{t-1}$$
$$(0.023) \quad (.112) \quad (.091)$$

$$EG = \frac{-0.632}{0.112} = -5.64$$

This is less than the 1% critical value from Table 18.4, thus we reject the null hypothesis. The two series are cointegrated.

Error Correction Model (ECM)

- Let's assume that y_t and x_t are I(1) but they are not cointegrated.
- In that case, we may estimate a dynamic model in first differences. For example,

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + u_t$$

where u_t is mean zero given all right hand side variables.

- This is an example of autoregressive distributed lag model of order 1 in first differences.
- Because all variables are I(0), OLS estimation poses no problems.
- On the other hand, if y_t and x_t are cointegrated, we can estimate richer dynamic models. In particular, we can augment the FDL by deviations (errors) from the long run equilibrium relationship.

Error Correction Model (ECM)

- Now assume that y_t and x_t are I(1) and they are cointegrated.
- Furthermore, assume that the cointegration relationship is given by

$$s_t = y_t - \beta x_t \sim \ I(0)$$

where s_t is stationary by definition.

- Now we can add the first lag of s_t into our dynamic model:

$$\begin{aligned}\Delta y_t &= \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta s_{t-1} + u_t \\ &= \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta \left(y_{t-1} - \beta x_{t-1}\right) + u_t\end{aligned}$$

- This model is called ECM. The term $\delta \left(y_{t-1} - \beta x_{t-1}\right)$ is called the *error correction term*.
- The EC parameter, $\delta < 0$, is also known as the *speed of adjustment* parameter.
- In some EC model, the contemporaneous variables (Δx_t) may be excluded (in forecasting models, for example).

Error Correction Model (ECM)

- For simplicity, assume that there are no lagged terms:

$$\Delta y_t = \alpha_0 + \gamma_0 \Delta x_t + \delta \left(y_{t-1} - \beta x_{t-1} \right) + u_t$$

  where $\delta < 0$
- The novelty of ECM is that the EC parameter δ governs how y_t responds to deviations from the long run equilibrium relationship. Some people prefer using **Equilibrium** Correction instead of **Error** Correction.
- If y_{t-1} is larger than βx_{t-1}, that is $s_{t-1} > 0$, then because $\delta < 0$, y_t will be forced to return back to the equilibrium. Note that $\Delta y_t < 0$ in that case.
- In the opposite case where $y_{t-1} < \beta x_{t-1}$, or $s_{t-1} < 0$, error correction occurs in the opposite direction. This will induce a positive change in y_t, again, pushing it back to the equilibrium.

Engle-Granger Two-Step Procedure

- In practice, the cointegration parameter β is rarely known.
- In that case, we can first estimate it in the first step, then obtain the cointegration vector and use it in the second step to estimate the ECM.
- This is known as the Engle-Granger Two-Step Procedure
- The general model now can be written as

$$\Delta y_t = \alpha_0 + \delta \left(y_{t-1} - \hat{\beta} x_{t-1} \right) + \sum_{j=1}^{p} \alpha_j \Delta y_{t-j} + \sum_{j=0}^{q} \gamma_j \Delta x_{t-j} + u_t$$

- The lag lengths p and q can be chosen using data-dependent information criteria (such as AIC or BIC)

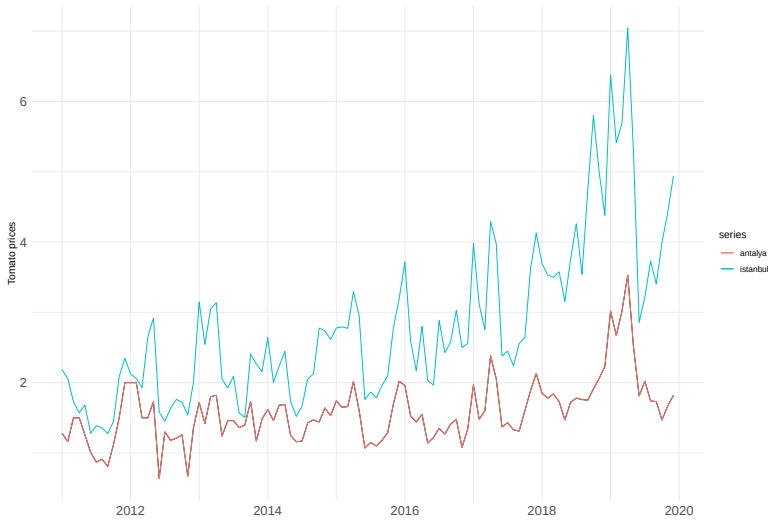
Example: Tomato prices in Antalya and Istanbul

- Law of One Price (LOP) states that due to spatial arbitrage, prices of a homogenous product in two different locations will be the same after accounting for the transaction costs.
- LOP is only valid if there are no restrictions on free trade, markets are perfectly competitive (no market power so that prices are freely determined)
- As stated, LOP implies that prices at two locations will not diverge from each other.
- Although there may be deviations from the long run equilibrium, market participants will recognize the arbitrage opportunity and will drive prices to equilibrium.
- In this application, we will examine a (more or less) homogenous agricultural product (tomato) and test if the prices in two different locations are cointegrated.

Example: Tomato prices in Antalya and Istanbul

- As a representative of the producer region we will use Antalya prices.
- Antalya region is one of the biggest producers of tomato in Turkey. The region supplies tomatoes and many other fresh agricultural products to several locations.
- Producer prices for tomato in Antalya and consumer prices for Istanbul were obtained from Turkish Statistical Institute. The data set is monthly and covers 2011.01-2019.12
- As can be inspected in the next plot, Istanbul prices are always above Antalya prices. The difference reflects various transaction costs including transport cost, insurance, profits and commissions of intermediaries, taxes and fees, etc.

Tomato prices in Antalya and Istanbul



Cointegration Equation

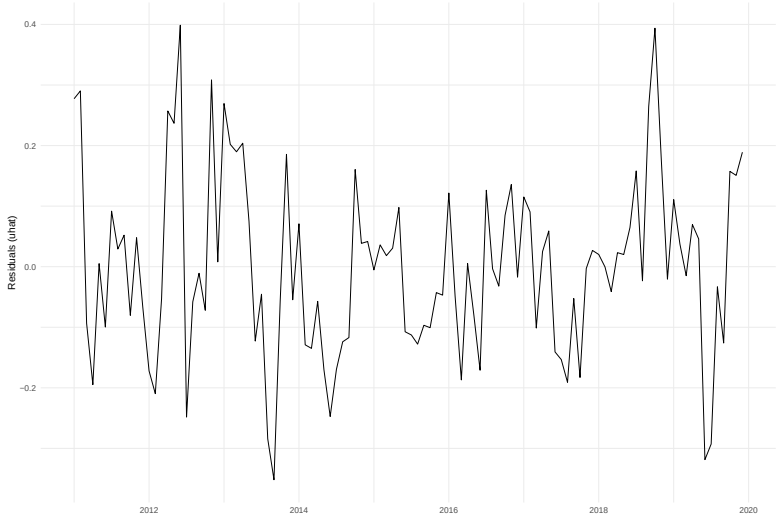
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.303658	0.031171	9.742	2.33e-16	***
trend	0.070355	0.006603	10.656	< 2e-16	***
lantalya	0.787020	0.062464	12.600	< 2e-16	***

- The regression of log(istanbul) on log(antalya) produced the R output above (together with time trend).

$$\widehat{\log(istanbul)}_t = 0.304 + 0.787 \log(antalya)_t + 0.070t$$

- According to these results the elasticity of Istanbul tomato prices with respect to Antalya producer prices is about 0.79%. If Antalya prices increases 10% then Istanbul prices are predicted to increase by 7.87%
- This interpretation is only valid if the regression above is not spurious. In other words, if the two price series are cointegrated.

Plot of the residuals



Tomato Prices: EG Test

- The Engle-Granger cointegration test statistic is $EG = -5.75$ which is smaller than the critical value at 1% level.
- Thus, Istanbul and Antalya prices are cointegrated and the regression in levels reflect a long run equilibrium relationship.
- The ECM estimates are

$$\begin{aligned}\widehat{\Delta \log(istanbul)}_t =& 0.006 - 0.52\hat{s}_{t-1} + 0.08\Delta \log(antalya)_{t-1} \\ & (0.020)\ \ (.138)\qquad (.087)\end{aligned}$$

  where $\hat{s}_{t-1} = \log(istanbul)_{t-1} - 0.304 - 0.787 \log(antalya)_{t-1} - 0.070(t-1)$

- The error correction parameter (speed of adjustment) is -0.52 and statistically significant.
- If istanbul price is above the equilibrium relationship by 1 percentage point then istanbul price falls by 0.52 percentage point in the next month. About half of the deviation from the equilibrium relationship will be corrected within a month.

Generalization of Cointegration to Multiple Variables

Let $\boldsymbol{Y}_t = \begin{bmatrix} y_{1t} & y_{2t} & \cdots & y_{kt} \end{bmatrix}^\top$ be $k \times 1$ vector of $I(1)$ variables. If there exists $k \times 1$ nonzero vector $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \ldots & \beta_k \end{bmatrix}^\top$ such that

$$\boldsymbol{\beta}^\top \boldsymbol{Y}_t = \beta_1 y_{1t} + \beta_2 y_{2t} + \ldots + \beta_k y_{kt} \sim I(0)$$

then these variables are cointegrated. Normalized with respect to y_{1t} it can be written as:

$$y_{1t} = \beta_2 y_{2t} + \beta_3 y_{3t} + \ldots + \beta_k y_{kt} + u_t$$

Defining the cointegrating vector as $\boldsymbol{\beta} = \begin{bmatrix} 1 & -\beta_2 & \ldots & -\beta_k \end{bmatrix}^\top$ the long run relationship can be written as:

$$\boldsymbol{\beta}^\top \boldsymbol{Y}_t = y_{1t} - \beta_2 y_{2t} - \beta_3 y_{3t} - \ldots - \beta_k y_{kt} = u_t \sim I(0)$$

There may be $0 < r < k$ linearly independent cointegration vectors.

Cointegration with Multiple Variables

- If the variables are cointegrated then their short run behavior can be modeled using a **Vector Error Correction** model (VEC)
- A VEC is a special case of **Vector Autoregression** (VAR) models.
- In multiple time series contexts, there are other ways testing and estimating cointegration relationships.
- One of the popular methods is Johansen's approach in which the variables are modeled using a VAR in levels
- The Johansen's cointegration tests are developed within the Maximum Likelihood framework.
- For technical details of these tests see more advanced texts such as Hamilton (1994) and Lutkepohl (2005).