Χρονικές σειρές

$10^{\text{ο}}$ μάθημα: Μη στάσιμα μοντέλα ARIMA
Μεθοδολογία Box-Jenkins

Εαρινό εξάμηνο 2018-2019
Τμήμα Μαθηματικών ΑΠΘ

Διδάσκουσα: **Αγγελική Παπάνα**

Μεταδιδακτορική Ερευνήτρια

Πολυτεχνική σχολή, Α.Π.Θ. & Οικονομικό Τμήμα, Πανεπιστήμιο Μακεδονίας

http://users.auth.gr/~agpapana/

Εισαγωγή

Τα γραμμικά στοχαστικά μοντέλα που είδαμε μέχρι τώρα αναφέρονται σε **στάσιμες** στοχαστικές διαδικασίες. Αυτό σημαίνει ότι ο **μέσος**, η **διακύμανση** και οι **αυτοσυνδιακυμάνσεις δεν εξαρτώνται από τον χρόνο t**.

Δηλαδή ο **μέσος** και η **διακύμανση** δεν μεταβάλλονται (είναι σταθερά), ενώ οι **αυτοσυνδιακυμάνσεις εξαρτώνται μόνο από την υστέρηση s**.

Θα εξετάσουμε τώρα διαδικασίες που δεν είναι στάσιμες.

Τυχαία διαδρομή

Έστω η αυτοπαλίνδρομη διαδικασίας πρώτης τάξης:

$$\boxed{\Upsilon_t = \alpha_1 \Upsilon_{t-1} + \varepsilon_t}$$

όπου ε_t είναι λευκός θόρυβος, δηλαδή $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{Var}(\varepsilon_t) = \sigma^2$.

- Για $\boxed{\alpha_1 = 1}$, το παραπάνω μοντέλο γίνεται: $\boxed{\Upsilon_t = \Upsilon_{t-1} + \varepsilon_t}$

το οποίο είναι γνωστό ως **τυχαία διαδρομή** ή **τυχαίος περίπατος**.

- Αν υπάρχει **σταθερός όρος**, δηλαδή $\boxed{\Upsilon_t = \alpha + \Upsilon_{t-1} + \varepsilon_t}$

το μοντέλο είναι γνωστό ως **τυχαία διαδρομή με περιπλάνηση.**

Μια στοχαστική διαδικασία που ακολουθεί την τυχαία διαδρομή δεν είναι στάσιμη.

Απόδειξη

Στην σχέση $\boxed{\Upsilon_t = \Upsilon_{t-1} + \varepsilon_t}$

κάνουμε διαδοχικές αντικαταστάσεις για t $= 1, .., t$:

$$\Upsilon_1 = \Upsilon_0 + \varepsilon_1$$

$$\Upsilon_2 = \Upsilon_1 + \varepsilon_2$$

$$\Upsilon_3 = \Upsilon_2 + \varepsilon_3$$

....

$$\Upsilon_t = \Upsilon_{t-1} + \varepsilon_t$$

------------------------------------ (+)

$$\Upsilon_t = \Upsilon_0 + \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t$$

Οπότε $\boxed{\mathrm{E}(\Upsilon_t) = \Upsilon_0}$ και $\boxed{\mathrm{Var}(\Upsilon_t) = t\sigma^2}$

Βλέπουμε ότι ο μέσος είναι σταθερός $\boxed{\mathrm{E}(\Upsilon_t) = \Upsilon_0}$

αλλά η διακύμανση εξαρτάται από τον χρόνο $\boxed{\mathrm{Var}(\Upsilon_t) = t\sigma^2}$.

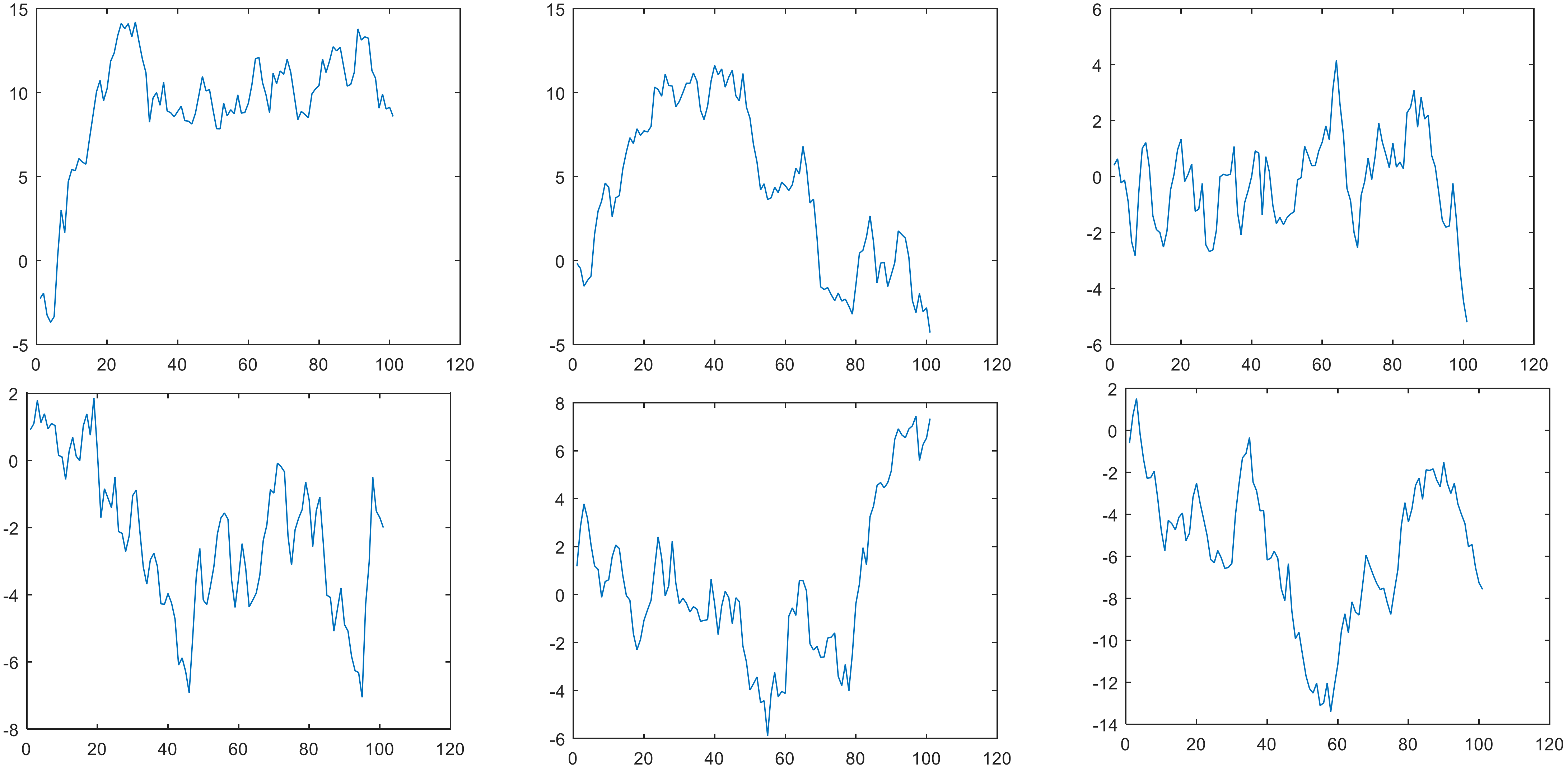
Επομένως η σειρά είναι **μη στάσιμη**.

Επειδή η μη στασιμότητα οφείλεται στην διακύμανση, λέμε ότι η σειρά είναι **μη στάσιμη ως προς την διακύμανση**.

Αν υποθέσουμε ότι $\varepsilon_t \sim \mathrm{WN}(\mu, \sigma^2)$, τότε και ο μέσος δεν είναι σταθερός, και η σειρά είναι **μη στάσιμη** και **ως προς τον μέσο**, αφού:

$$\boxed{\mathrm{E}(\Upsilon_t) = \Upsilon_0 + t\mu}$$

Πραγματοποιήσεις της διαδικασίας τυχαίου περιπάτου



Αν πάρουμε τις πρώτες διαφορές της διαδικασίας τυχαίου περιπάτου:

$$\boxed{\Delta Y_t = Y_t - Y_{t-1} = \varepsilon_t}$$

τότε η σειρά που προκύπτει είναι **στάσιμη**, εφόσον υποτίθεται ότι η ε_t είναι στάσιμη.

Ολοκληρωμένες διαδικασίες

Όταν μια σειρά μετατρέπεται σε **στάσιμη** παίρνοντας **πρώτες διαφορές**, λέμε ότι είναι **ολοκληρωμένη πρώτης τάξεως** (integrated first order) και συμβολίζεται ως $\mathrm{I}(1)$.

Αν η σειρά μετατρέπεται σε **στάσιμη** παίρνοντας τις **δεύτερες διαφορές**, λέμε ότι είναι **ολοκληρωμένη δεύτερης τάξεως** (integrated second order) και συμβολίζεται ως $\mathrm{I}(2)$, κ.ο.κ.

Γενικά αν $\boldsymbol{d}$ είναι ο αριθμός των διαφορών που μετατρέπει μια σειρά σε **στάσιμη**, η σειρά καλείται **ολοκληρωμένη $\boldsymbol{d}$ τάξεως** και συμβολίζεται ως $\mathrm{I}(d)$.

Χρησιμοποιώντας τον τελεστή υστερήσεως, οι **πρώτες διαφορές** ορίζονται ως

$$\boxed{\Upsilon_t - \Upsilon_{t-1} = (1 - \mathrm{L})\Upsilon_t = \Delta\Upsilon_t}$$

- Το $\Delta = (1 - \mathrm{L})$ είναι ο τελεστής πρώτων διαφορών.
- Το $\Delta^2 = (1 - \mathrm{L})^2$ είναι ο τελεστής δεύτερων διαφορών, όπου

$$\Delta^2\Upsilon_t = (1 - \mathrm{L})^2\Upsilon_t = (1 - 2\mathrm{L} + \mathrm{L}^2)\Upsilon_t =$$
$$= \Upsilon_t - 2\mathrm{L}\Upsilon_t + \mathrm{L}^2\Upsilon_t \quad \text{ή}$$

$$\Delta^2\Upsilon_t = \Delta(\Delta\Upsilon_t) = \Delta(\Upsilon_t - \Upsilon_{t-1}) = \Delta\Upsilon_t - \Delta\Upsilon_{t-1} =$$
$$= (\Upsilon_t - \Upsilon_{t-1}) - (\Upsilon_{t-1} - \Upsilon_{t-2}) =$$
$$= \Upsilon_t - 2\Upsilon_{t-1} + \Upsilon_{t-2}$$

- Γενικά, το $\Delta^d = (1 - \mathrm{L})^d$ είναι ο τελεστής d διαφορών.

Παρατήρηση

Η διαφορά ανάμεσα στην Υ_t και στην Υ_{t-d} μπορεί να παρασταθεί ως $\Delta_d \Upsilon_t = \Upsilon_t - \Upsilon_{t-d}$, αλλά δεν πρέπει να συγχέεται με τον τελεστή διαφορών $\Delta^d \Upsilon_t$.

- Το μοντέλο τυχαίας διαδρομής είναι ολοκληρωμένο πρώτης τάξεως, αφού όπως είδαμε μετατρέπεται σε στάσιμο με τις πρώτες διαφορές.
- Μια στάσιμη σειρά, όπως ο λευκός θόρυβος, θεωρείται ολοκληρωμένη μηδενικής τάξεως, δηλαδή $\mathrm{I}(0)$.
- Ο όρος ολοκληρωμένη προέρχεται από τον τρόπο με τον οποίο μια μη στάσιμη διαδικασία προκύπτει από μια στάσιμη. Δηλαδή *αθροίζοντας ή ολοκληρώνοντας* d φορές.

Παράδειγμα Έστω η σειρά $\{\Upsilon_t\}$ και οι πρώτες διαφορές: $\mathrm{z}_t = \Upsilon_t - \Upsilon_{t-1}$.

Είναι: $\mathrm{z}_t + \mathrm{z}_{t-1} + \mathrm{z}_{t-2} + \cdots =$

$$= (\Upsilon_t - \Upsilon_{t-1}) + (\Upsilon_{t-1} - \Upsilon_{t-2}) + (\Upsilon_{t-2} - \Upsilon_{t-3}) + \cdots = \Upsilon_t$$

Δηλαδή το Υ_t είναι το άθροισμα (ολοκλήρωμα) όλων των διαφορών.

Παράδειγμα Έστω οι δεύτερες διαφορές: $\mathrm{w}_t = \mathrm{z}_t - \mathrm{z}_{t-1}$ της σειράς $\{\Upsilon_t\}$.

Είναι: $\mathrm{w}_t + \mathrm{w}_{t-1} + \mathrm{w}_{t-2} + \cdots = (\mathrm{z}_t - \mathrm{z}_{t-1}) + (\mathrm{z}_{t-1} - \mathrm{z}_{t-2}) + \cdots = \mathrm{z}_t$

Αθροίζουμε τώρα τα z_t, οπότε: $\mathrm{z}_t + \mathrm{z}_{t-1} + \mathrm{z}_{t-2} + \cdots =$

$$= (\Upsilon_t - \Upsilon_{t-1}) + (\Upsilon_{t-1} - \Upsilon_{t-2}) + (\Upsilon_{t-2} - \Upsilon_{t-3}) + \cdots = \Upsilon_t$$

Δηλαδή για να βρούμε το Υ_t αθροίζουμε δυο φορές.

- Γενικά, αθροίζουμε d φορές μια στάσιμη σειρά για να προκύψει η αρχική μη στάσιμη σειρά.

Μη στάσιμα μοντέλα ARIMA

Ένα μοντέλο ARMA(p, q) που εφαρμόζεται σε μια ολοκληρωμένη σειρά d τάξεως, ονομάζεται αυτοπαλίνδρομο – κινητού μέσου ολοκληρωμένο μοντέλο τάξεως (p, d, q) (Autoregressive Integrated Moving average) και συμβολίζεται **ARIMA** $(\boldsymbol{p}, \boldsymbol{d}, \boldsymbol{q})$.

Π.χ. ARIMA $(2,1,2)$ σημαίνει ότι η σειρά καθίσταται στάσιμη με τις πρώτες διαφορές και στην προκύπτουσα σειρά των πρώτων διαφορών εφαρμόζεται το ARMA(2,2).

Στην γενική της μορφή μια διαδικασία **ARMA**$(\boldsymbol{p}, \boldsymbol{q})$ γράφεται ως

$$\boxed{\mathrm{z}_t = \alpha_1 \mathrm{z}_{t-1} + \cdots + \alpha_p \mathrm{z}_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}}$$

Μια διαδικασία **ARIMA**$(\boldsymbol{p}, \boldsymbol{d}, \boldsymbol{q})$ μπορεί να διατυπωθεί με τρεις διαφορετικές μορφές, δηλ. η μη στάσιμη αρχική χρονοσειρά Υ_t γράφεται:

1. Ως συνάρτηση των παρελθουσών τιμών της και των τιμών του διαταρακτικού όρου, τρέχουσας και παρελθουσών. Η μορφή αυτή είναι γνωστή ως **εξίσωση διαφοράς**.

   π.χ. αν $\mathrm{z}_t = \Upsilon_t - \Upsilon_{t-1}$, η εξίσωση διαφοράς για την **ARIMA**$(\boldsymbol{p}, \mathbf{1}, \boldsymbol{q})$, τότε η εξίσωση διαφοράς της διαδικασίας Υ_t είναι η ακόλουθη:

$$\boxed{\begin{aligned}\Upsilon_t = (1+\alpha_1)\Upsilon_{t-1} + (\alpha_2 - \alpha_1)\Upsilon_{t-2} + \cdots + \left(\alpha_p - \alpha_{p-1}\right)\Upsilon_{t-p} - \alpha_p \Upsilon_{t-p-1} + \\ \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}\end{aligned}}$$

   Η σχέση αυτή προκύπτει από την ARMA(p, q) μορφή της z_t (στάσιμη)

$$\boxed{\mathrm{z}_t = \alpha_1 \mathrm{z}_{t-1} + \cdots + \alpha_p \mathrm{z}_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}}$$

   αντικαθιστώντας το z_t με το ίσο του, δηλ. $\mathrm{z}_t = \Upsilon_t - \Upsilon_{t-1}$.

Απόδειξη

Από τις σχέσεις $z_t = \Upsilon_t - \Upsilon_{t-1}$ και

$$z_t = \alpha_1 z_{t-1} + \cdots + \alpha_p z_{t-p} + \underbrace{\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}}_{A}$$

προκύπτει:

$$\Upsilon_t - \Upsilon_{t-1} = \alpha_1(\Upsilon_{t-1} - \Upsilon_{t-2}) + \alpha_2(\Upsilon_{t-2} - \Upsilon_{t-3}) + \cdots + \alpha_p(\Upsilon_{t-p} - \Upsilon_{t-p-1}) + A \Leftrightarrow$$

$$\Upsilon_t = \Upsilon_{t-1} + \alpha_1\Upsilon_{t-1} - \alpha_1\Upsilon_{t-2} + \alpha_2\Upsilon_{t-2} - \alpha_2\Upsilon_{t-3} + \cdots + \alpha_p\Upsilon_{t-p} - \alpha_p\Upsilon_{t-p-1} + A \Leftrightarrow$$

$$\Upsilon_t = (1 + \alpha_1)\Upsilon_{t-1} + (\alpha_2 - \alpha_1)\Upsilon_{t-2} + \cdots + (\alpha_p - \alpha_{p-1})\Upsilon_{t-p} - \alpha_p\Upsilon_{t-p-1} + A$$

2. Ως συνάρτηση των παρελθουσών τιμών της και της τρέχουσας τιμής του διαταρακτικού όρου.

Η μορφή αυτή της Υ_t είναι γνωστή ως **αντίστροφη μορφή**.

π.χ. η αντίστροφη μορφή μιας **ARIMA$(\mathbf{0, 1, 1})$** διαδικασίας είναι:

Η σχέση αυ$\Upsilon_t = (1 + \theta_1)\Upsilon_{t-1} - \theta_1(1 + \theta_1)\Upsilon_{t-2} + \theta_1^2(1 + \theta_1)\Upsilon_{t-3} - \theta_1^3(1 + \theta_1)\Upsilon_{t-4} + + \cdots + \varepsilon_t$

τή προκύπτει από την εξίσωση διαφοράς με διαδοχικές αντικαταστάσεις για $\varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}, \ldots$

Απόδειξη

Για μια **ARIMA**$(\mathbf{0},\mathbf{1},\mathbf{1})$ διαδικασία η εξίσωση διαφοράς είναι:

$$\Upsilon_t = \Upsilon_{t-1} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \ (1) \Leftrightarrow$$

$$\varepsilon_t = \Upsilon_t - \Upsilon_{t-1} - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}$$

Οπότε $\quad \varepsilon_{t-1} = \Upsilon_{t-1} - \Upsilon_{t-2} - \theta_1\varepsilon_{t-2} - \theta_2\varepsilon_{t-3} - \cdots - \theta_q\varepsilon_{t-q-1}$ (2)

$\varepsilon_{t-2} = \Upsilon_{t-2} - \Upsilon_{t-3} - \theta_1\varepsilon_{t-3} - \theta_2\varepsilon_{t-4} - \cdots - \theta_q\varepsilon_{t-q-2}$ (3) κ.ο.κ

Στην (1) αντικαθιστούμε διαδοχικά τα ε_{t-1}, ε_{t-2},… από τις (2), (3) κτλ

$$\Upsilon_t = \Upsilon_{t-1} + \varepsilon_t + \theta_1\left(\Upsilon_{t-1} - \Upsilon_{t-2} - \theta_1\varepsilon_{t-2} - \theta_2\varepsilon_{t-3} - \cdots - \theta_q\varepsilon_{t-q-1}\right) +$$

$$\theta_2\left(\Upsilon_{t-2} - \Upsilon_{t-3} - \theta_1\varepsilon_{t-3} - \theta_2\varepsilon_{t-4} - \cdots - \theta_q\varepsilon_{t-q-2}\right) + \cdots \Leftrightarrow \ldots$$

Συνεχίζοντας τις αντικαταστάσεις και τις πράξεις, προκύπτει η ζητούμενη σχέση.

3. Ως συνάρτηση μόνο των τιμών του διαταρακτικού όρου, τρέχουσας και παρελθουσών.

Η μορφή αυτή της Υ_t είναι γνωστή ως **τυχαία διαταραχή**.

π.χ. η μορφή μιας **ARIMA$(\mathbf{0}, \mathbf{1}, \mathbf{1})$** διαδικασίας είναι:

$\Upsilon_t = \varepsilon_t + (1 + \theta_1)\varepsilon_{t-1} + (1 + \theta_1)\varepsilon_{t-2} + \cdots$

Η σχέση αυτή προκύπτει από την εξίσωση διαφοράς με διαδοχικές αντικαταστάσεις για $\Upsilon_{t-1}, \Upsilon_{t-2}, \Upsilon_{t-3}, \ldots$

Απόδειξη

Από την εξίσωση διαφοράς

$$\Upsilon_t = \Upsilon_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (1)$$

έχουμε:

$$\Upsilon_{t-1} = \Upsilon_{t-2} + \varepsilon_{t-1} + \theta_1 \varepsilon_{t-2} + \theta_2 \varepsilon_{t-3} + \cdots + \theta_q \varepsilon_{t-q-1}$$

$$\Upsilon_{t-2} = \Upsilon_{t-3} + \varepsilon_{t-2} + \theta_1 \varepsilon_{t-3} + \theta_2 \varepsilon_{t-4} + \cdots + \theta_q \varepsilon_{t-q-2} \quad \text{κ.ο.κ.}$$

Αντικαθιστούμε στην (1) διαδοχικά τα $\Upsilon_{t-1}, \Upsilon_{t-2}, \ldots$, οπότε προκύπτει η ζητούμενη σχέση.

Μεθοδολογία Box-Jenkins

Η προσέγγιση **Box-Jenkins** στην ανάλυση χρονολογικών σειρών είναι μια μέθοδος εύρεσης ενός στατιστικού μοντέλου (ARIMA) που να παριστάνει ικανοποιητικά τη στοχαστική διαδικασία που παρήγαγε τα δεδομένα (δείγμα).

Η μέθοδος περιλαμβάνει τα εξής τρία στάδια:

1) **Την ταυτοποίηση**
2) **Την εκτίμηση**
3) **Το διαγνωστικό έλεγχο**
   α) Έλεγχος Καταλοίπων
   β) Έλεγχος Τάξεως Υποδείγματος
   γ) Κριτήρια Επιλογής Υποδείγματος

1) Ταυτοποίηση

Το πρώτο στάδιο είναι η εξειδίκευση ενός ARIMA μοντέλου με βάση τις πληροφορίες του διαθέσιμου δείγματος. Αυτό σημαίνει καθορισμός των κατάλληλων τιμών $\boldsymbol{d}, \boldsymbol{p}$ και $\boldsymbol{q}$.

Δηλαδή, καθορισμός του αριθμού $(\boldsymbol{d})$ των διαφορών που απαιτούνται προκειμένου η σειρά να μετατραπεί σε στάσιμη, αν δεν είναι.

Στη συνέχεια πρέπει να καθοριστεί η τάξη $\boldsymbol{p}$ της αυτοπαλίνδρομης διαδικασίας και η τάξη $\boldsymbol{q}$ της διαδικασίας κινητού μέσου.

Άρα πρώτα πρέπει να διαπιστωθεί αν η σειρά είναι στάσιμη.

Η διαπίστωση αυτή βασίζεται στη συμπεριφορά της **δειγματικής συνάρτησης αυτοσυσχέτισης**. Αν οι αυτοσυσχετίσεις συγκλίνουν ταχύτατα προς το μηδέν, αυτό είναι σοβαρή ένδειξη ότι η σειρά είναι στάσιμη.

Αντίθετα αν οι αυτοσυσχετίσεις φθίνουν με αργό ρυθμό, είναι ένδειξη ότι η σειρά είναι μη στάσιμη.

Για τον έλεγχο στασιμότητας, χρησιμοποιούνται οι γνωστοί ως **έλεγχοι μοναδιαίας ρίζας**, τους οποίους θα δούμε στην επόμενη ενότητα.

Αν η σειρά δεν είναι στάσιμη, μετατρέπεται σε στάσιμη παίρνοντας τις πρώτες διαφορές ή τις δεύτερες διαφορές κτλ. Με τα δεδομένα της στάσιμης χρονοσειράς, στη συνέχεια, προσδιορίζεται η **τάξη** του μοντέλου **ARIMA**. Δηλαδή προσδιορίζονται οι τιμές $\boldsymbol{p}$ και $\boldsymbol{q}$. Ο προσδιορισμός βασίζεται στις δειγματικές, απλές και μερικές αυτοσυσχετίσεις, όπως είδαμε στην προηγούμενη ενότητα.

2) Εκτίμηση

Μετά την εξειδίκευση του μοντέλου ακολουθεί η εκτίμηση των $\boldsymbol{p}$ παραμέτρων $\alpha_1, \alpha_2, .., \alpha_p$ της AR διαδικασίας και των $\boldsymbol{q}$ παραμέτρων $\theta_1, \theta_2, \ldots, \theta_q$ της MA διαδικασίας.

Αν η σειρά είναι μόνο αυτοπαλίνδρομη (AR), οι παράμετροι $\alpha_1, \alpha_2, .., \alpha_p$ μπορούν να εκτιμηθούν με την μέθοδο των ελαχίστων τετραγώνων ή με τις εξισώσεις Yule-Walker.

Αν περιέχονται και όροι κινητού μέσου (MA) στο μοντέλο, τότε για την εκτίμηση των παραμέτρων συνήθως χρησιμοποιούνται μη γραμμικές μέθοδοι ή με τις εξισώσεις Yule-Walker.

3) Διαγνωστικός έλεγχος

Στο στάδιο αυτό γίνεται έλεγχος καλής προσαρμογής του μοντέλου που εκτιμήθηκε.

Δηλαδή ελέγχουμε πόσο καλά «ταιριάζει» το εκτιμώμενο μοντέλο με τα δεδομένα, αφού είναι πιθανό κάποιο άλλο μοντέλο ARIMA να προσαρμόζεται καλύτερα.

Ο διαγνωστικός έλεγχος περιλαμβάνει στατιστικούς ελέγχους για τη συμπεριφορά των καταλοίπων, για την τάξη του υποδείγματος και την σημαντικότητα των συντελεστών του μοντέλου.

Έλεγχος καταλοίπων

Αν το εκτιμώμενο ARIMA πράγματι εκφράζει ικανοποιητικά τη διαδικασία από την οποία προέρχονται τα δεδομένα, τότε τα κατάλοιπα πρέπει να συμπεριφέρονται ως μια διαδικασία λευκού θορύβου. Δηλαδή τα κατάλοιπα δεν πρέπει να αυτοσυσχετίζονται.

Ο έλεγχος των καταλοίπων γίνεται με την στατιστική $\boldsymbol{Q}$ των Box-Pierce με το οποίο ελέγχεται η σημαντικότητα από κοινού ενός αριθμού συντελεστών αυτοσυσχετίσεως, έστω $\boldsymbol{m}$. Δηλαδή ελέγχεται η μηδενική υπόθεση $\mathbf{H_0}$: $\boldsymbol{\rho_1 = \rho_2 = \cdots = \rho_m = 0}$.

Η στατιστική $\boldsymbol{Q}$ ορίζεται ως:

$$\boxed{\boldsymbol{Q = \mathrm{T} \sum_{s=1}^{m} \widehat{\rho}_s^2}}$$

όπου $\boldsymbol{\widehat{\rho}_s}$ είναι οι δειγματικές αυτοσυσχετίσεις των καταλοίπων και $\mathbf{T}$ ο αριθμός των παρατηρήσεων (καταλοίπων).

Συνήθως, ο αριθμός των αυτοσυσχετίσεων των καταλοίπων που χρησιμοποιούνται για τον υπολογισμό της στατιστικής $\boldsymbol{Q}$ ισούται με την τετραγωνική ρίζα του αριθμού των παρατηρήσεων: $\boldsymbol{m} = \sqrt{\mathbf{T}}$.

Η στατιστική $\boldsymbol{Q}$ ακολουθεί προσεγγιστικά την κατανομή $\boldsymbol{X}^2$ με $\boldsymbol{m} - \boldsymbol{p} - \boldsymbol{q}$ βαθμούς ελευθερίας. Για δεδομένο επίπεδο σημαντικότητας $\boldsymbol{\alpha}$, η μηδέν υπόθεση ότι όλοι οι συντελεστές αυτοσυσχέτισης είναι μηδέν απορρίπτεται αν η τιμή της $\boldsymbol{Q}$ υπερβαίνει την κρίσιμη τιμή της $\boldsymbol{X}^2$, δηλαδή αν $\boldsymbol{Q} > \boldsymbol{X}^2$.

Μια τροποποιημένη μορφή της παραπάνω στατιστικής για μικρά δείγματα είναι αυτή των Ljung και Box, η οποία ορίζεται ως

$$\boxed{\boldsymbol{Q} = \mathbf{T}(\mathbf{T}+2)\sum_{s=1}^{m}\frac{\widehat{\rho}_s^2}{\mathbf{T}-\mathbf{s}}}$$

η οποία ακολουθεί πάλι προσεγγιστικά την κατανομή $\boldsymbol{X}^2$ με $\boldsymbol{m} - \boldsymbol{p} - \boldsymbol{q}$ βαθμούς ελευθερίας.

Έλεγχος της τάξης του υποδείγματος

Η καταλληλότητα του εκτιμώμενου μοντέλου ελέγχεται συγκρίνοντας το με ένα άλλο υπόδειγμα μεγαλύτερης τάξης.

Δηλαδή το εκτιμώμενο μοντέλο ARMA$(\boldsymbol{p}, \boldsymbol{q})$ συγκρίνεται με τα υποδείγματα ARMA$(\boldsymbol{p} + \mathbf{1}, \boldsymbol{q})$ και ARMA$(\boldsymbol{p}, \boldsymbol{q} + \mathbf{1})$.

Αν το υπόδειγμα που εκτιμήθηκε περιγράφει τη διαδικασία που παρήγαγε τα δεδομένα, οι επιπλέον συντελεστές στα μεγαλύτερα υποδείγματα δεν θα πρέπει να είναι στατιστικά διαφορετικοί από το μηδέν. Η παραπάνω διαδικασία ελέγχου ονομάζεται **υπερπροσαρμογή** (overfitting).

Κριτήριο επιλογής μοντέλου

Είναι φανερό ότι αυξάνοντας την τάξη του εκτιμώμενου μοντέλου, δηλαδή προσθέτοντας υστερήσεις για το αυτοπαλίνδρομο τμήμα ή και για το τμήμα κινητού μέσου, θα μειώνεται το άθροισμα των τετραγώνων των καταλοίπων, αλλά θα μειώνονται και οι βαθμοί ελευθερίας, αφού εκτιμώνται περισσότεροι παράμετροι.

Για την σύγκριση της ερμηνευτικής ικανότητας εναλλακτικών υποδειγμάτων, που διαφέρουν ως προς τον αριθμό των παραμέτρων ή και το μέγεθος του δείγματος, χρησιμοποιούνται συνηθέστερα τα παρακάτω κριτήρια (μέγιστης πιθανοφάνειας):

1. Κριτήριο πληροφοριών Akaike (AIC)

$$\text{AIC} = ln\frac{\sum\hat{\varepsilon}^2}{\mathrm{T}} + \frac{2K}{\mathrm{T}}$$

2. Μπαϊεσιανό κριτήριο Schwartz (SBC)

$$\text{SBC} = ln\frac{\sum\hat{\varepsilon}^2}{\mathrm{T}} + \frac{K}{\mathrm{T}}ln\mathrm{T}$$

όπου

$\sum\hat{\varepsilon}^2$: άθροισμα τετραγώνων καταλοίπων

Κ: αριθμός παραμέτρων που εκτιμούνται ($\boldsymbol{p}+\boldsymbol{q}+\mathbf{1}$)

Τ: αριθμός παρατηρήσεων

- Με βάση τα παραπάνω κριτήρια, επιλέγεται το μοντέλο με τη μικρότερη τιμή.
- Τα κριτήρια αυτά μπορούν να πάρουν και αρνητικές τιμές.

Άσκηση 1

Να δείξετε ότι αν Υ_t είναι λευκός θόρυβος ($\Upsilon_t \sim W\mathrm{N}(\mu, \sigma^2)$), τότε οι πρώτες διαφορές $\Delta\Upsilon_t$ είναι $\mathrm{I}(0)$.

Υπενθύμιση

$\mathrm{Cov}(\mathrm{X}, \Upsilon + \mathrm{Z}) = \mathrm{Cov}(\mathrm{X}, \mathrm{Z}) + \mathrm{Cov}(\Upsilon, \mathrm{Z})$

Λύση

Είναι: $\Delta\Upsilon_t = \Upsilon_t - \Upsilon_{t-1}$

- $\mathrm{E}(\Delta\Upsilon_t) = \mathrm{E}(\Upsilon_t - \Upsilon_{t-1}) = \mathrm{E}(\Upsilon_t) - \mathrm{E}(\Upsilon_{t-1}) = 0$ σταθερό
- $\mathrm{Var}(\Delta\Upsilon_t) = \mathrm{Var}(\Upsilon_t - \Upsilon_{t-1}) = \mathrm{Var}(\Upsilon_t) + \mathrm{Var}(\Upsilon_{t-1}) - 2\mathrm{Cov}(\Upsilon_t, \Upsilon_{t-1})$
  $= \sigma^2 + \sigma^2 = 2\sigma^2$ σταθερό
- $\mathrm{Cov}\big(\Delta\Upsilon_t, \mathrm{Var}(\Delta\Upsilon_{t-s})\big) = \mathrm{Cov}(\Upsilon_t - \Upsilon_{t-1}, \Upsilon_{t-s} - \Upsilon_{t-s-1}) =$

$= \mathrm{Cov}(Y_t, Y_{t-s}) - \mathrm{Cov}(Y_t, Y_{t-s-1}) - \mathrm{Cov}(Y_{t-1}, Y_{t-s}) + \mathrm{Cov}(\mathrm{Cov}(Y_{t-1}, Y_{t-s-1})$
$= 0$

Άρα η σειρά πρώτων διαφορών είναι στάσιμη δηλαδή $\mathrm{I}(0)$ και μάλιστα είναι λευκός θόρυβος αφού οι συνδιακυμάνσεις είναι μηδενικές.

Άσκηση 2

Έστω ότι για μια δεδομένη χρονοσειρά προτείνονται τα ακόλουθα δύο μοντέλα:

$$\Upsilon_t = \delta + 0.86\Upsilon_{t-1} + 0.10\Upsilon_{t-2} + \varepsilon_t$$

$$\Upsilon_t = \Upsilon_{t-1} + \varepsilon_t - 0.10\varepsilon_{t-1}$$

α. Να δοθεί η αντίστροφη μορφή του δεύτερου μοντέλου.

β. Είναι πράγματι πολύ διαφορετικά αυτά τα δύο μοντέλα;

Λύση

α. Το δεύτερο μοντέλο είναι η εξίσωση διαφοράς μιας ARIMA$(0,1,1)$ διαδικασίας, αφού: $\Upsilon_t = \Upsilon_{t-1} + \varepsilon_t - 0.10\varepsilon_{t-1}$ ($|\alpha_1| = 1$ μη στάσιμη)

$\Upsilon_t - \Upsilon_{t-1} = \varepsilon_t - 0.10\varepsilon_{t-1}$ (πρώτες διαφορές)

$z_t = \varepsilon_t - 0.10\varepsilon_{t-1}$ (ARMA(0,1) με $\theta_1 = -0.1$)

Άρα η αρχική χρονοσειρά Υ_t (μη στάσιμη) προέρχεται από μια ARIMA$(0,1,1)$, αφού οι πρώτες διαφορές της Υ_t, δηλαδή η z_t (στάσιμη) είναι μια ARMA$(0,1)$ διαδικασία.

Η αντίστροφη μορφή μιας ARIMA$(0,1,1)$ διαδικασίας είναι γνωστή:

$$\Upsilon_t = (1+\theta_1)\Upsilon_{t-1} - \theta_1(1+\theta_1)\Upsilon_{t-2} + \theta_1^2(1+\theta_1)\Upsilon_{t-3} - \theta_1^3(1+\theta_1)\Upsilon_{t-4} + \varepsilon_t$$

Οπότε για $\theta_1 = -0.1$:

$$\Upsilon_t = (1-0.1)\,\Upsilon_{t-1} + 0.1(1-0.1)\Upsilon_{t-2} + 0.1^2(1-0.1)\Upsilon_{t-3} + \cdots + \varepsilon_t$$

$$\boldsymbol{\Upsilon_t = 0.9\,\Upsilon_{t-1} + 0.09\Upsilon_{t-2} + 0.009\Upsilon_{t-3} + \cdots + \varepsilon_t}$$

Βλέπουμε ότι η σπουδαιότητα των συντελεστών μειώνεται ταχύτατα.

β. Τα μοντέλα

$$\Upsilon_t = \delta + 0.86\Upsilon_{t-1} + 0.10\Upsilon_{t-2} + \varepsilon_t$$

$$\Upsilon_t = \Upsilon_{t-1} + \varepsilon_t - 0.10\varepsilon_{t-2}$$

είναι διαφορετικά, αλλά η ARIMA(0,1,1) διαδικασία της οποίας το δεύτερο μοντέλο αποτελεί την εξίσωση διαφοράς της, μπορεί να θεωρηθεί ότι προσεγγίζεται ικανοποιητικά με την ARMA(2,0) διαδικασία που περιγράφει το πρώτο μοντέλο, καθώς γράφεται

$$\Upsilon_t = 0.9\,\Upsilon_{t-1} + 0.09\Upsilon_{t-2} + 0.009\Upsilon_{t-3} + \cdots + \varepsilon_t$$

όπου βλέπουμε ότι οι συντελεστές α_1, α_2 στα δυο μοντέλα διαφέρουν πολύ λίγο και οι τιμές των υπόλοιπων συντελεστών της αντίστροφης μορφής του ARIMA(0,1,1) είναι πολύ μικροί και φθίνουν εκθετικά.

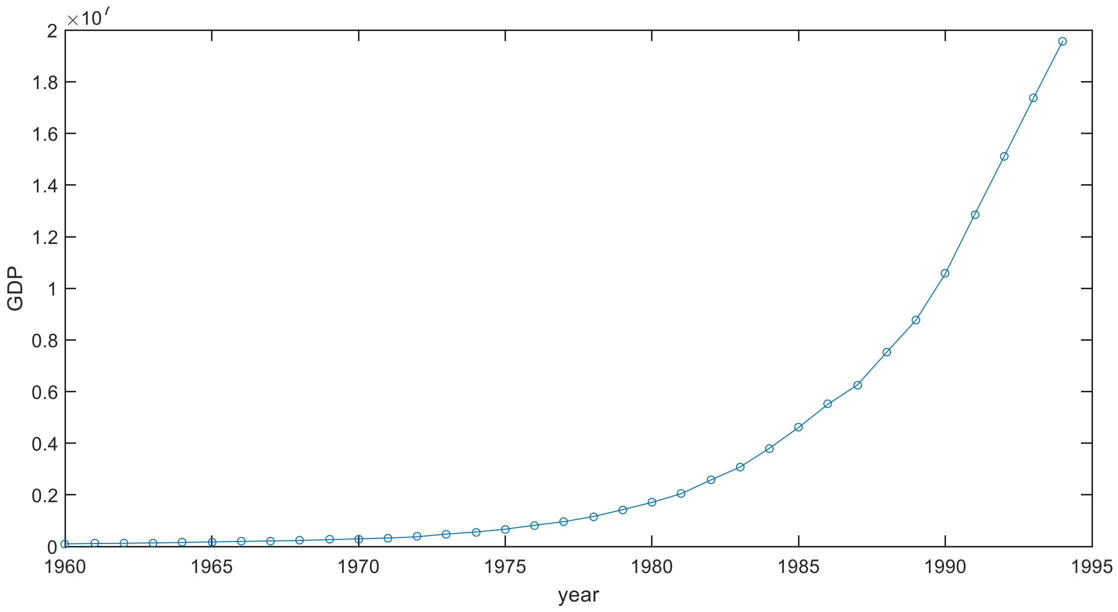
Παράδειγμα ARIMA

Θα εφαρμόσουμε ην μεθοδολογία Box-Jenkins εκτιμώντας ένα ARIMA μοντέλο για το ακαθάριστο προϊόν (GDP) της Ελλάδας για την περίοδο 1960-1994. Τα δεδομένα δίνονται στο αρχείο GDP.dat.

α) Ταυτοποίηση του μοντέλου

Στο παρακάτω σχήμα δίνεται η γραφική παράσταση της χρονοσειράς.

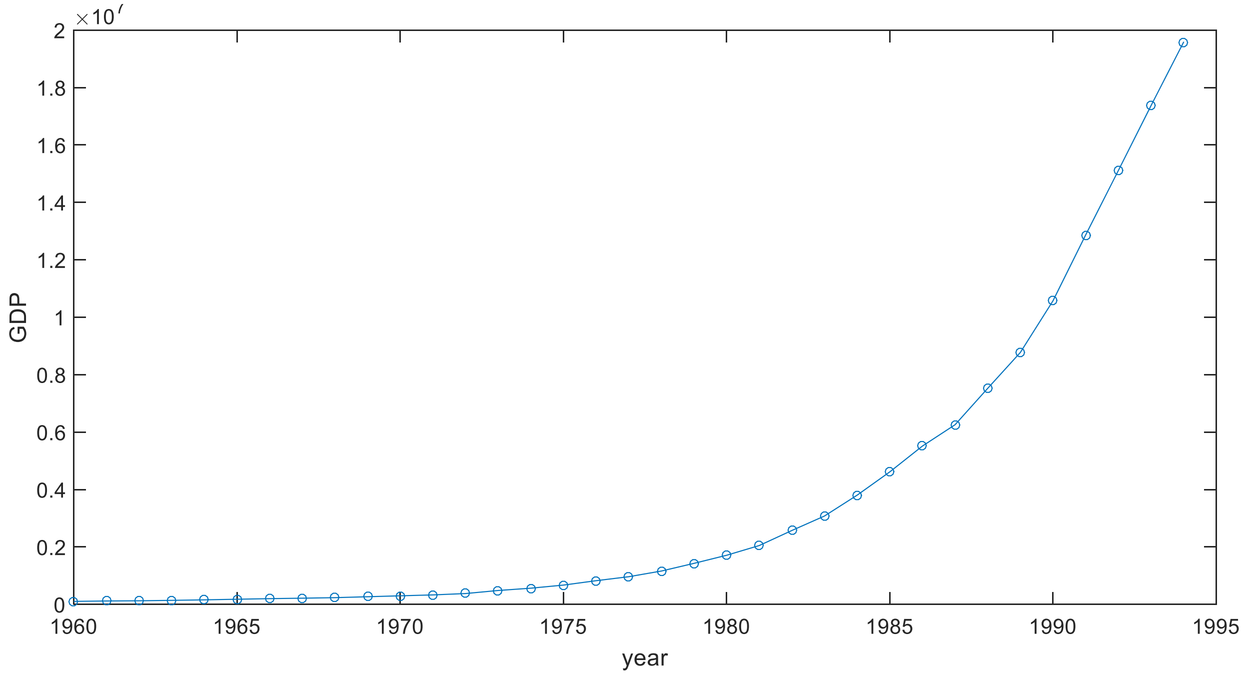
```
>> plot((1960:1994),GDP,'o-')
```



Με βάση τη γραφική παράσταση της χρονοσειράς, η σειρά δεν μπορεί να είναι στάσιμη, αφού οι τιμές της αυξάνονται συνεχώς.

Η σειρά χαρακτηρίζεται από έντονη τάση.

Η μη στασιμότητα της σειράς επιβεβαιώνεται αν εκτιμήσουμε τις αυτοσυσχετίσεις της.



Στα παρακάτω σχήματα δίνονται οι γραφικές παραστάσεις της συνάρτησης αυτοσυσχέτισης

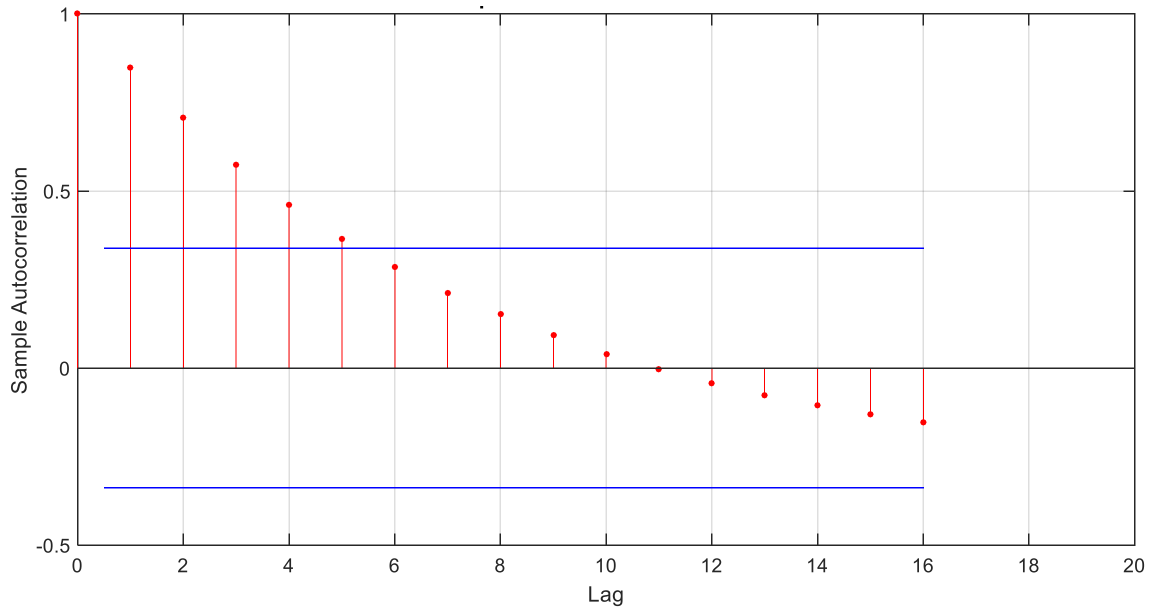
```
>> acM = autocorr(GDP,16);
```

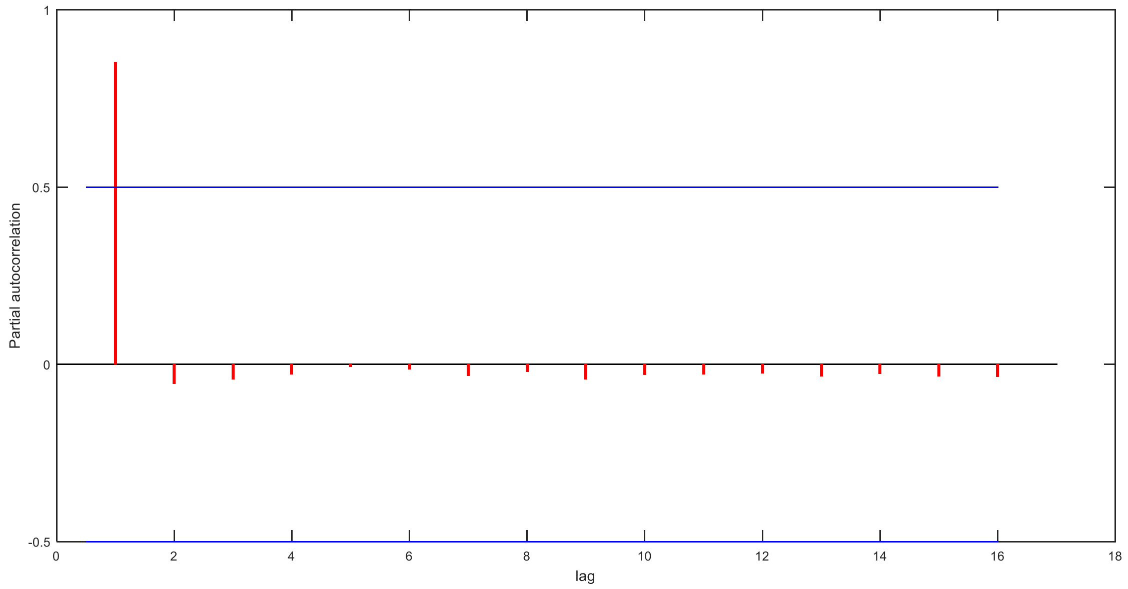
ή

```
>> acM = autocorrelation(GDP, 16);
```

και μερικής αυτοσυσχέτισης της χρονοσειράς

```
>> pacfV = acf2pacf(acM(2:end),1)
```



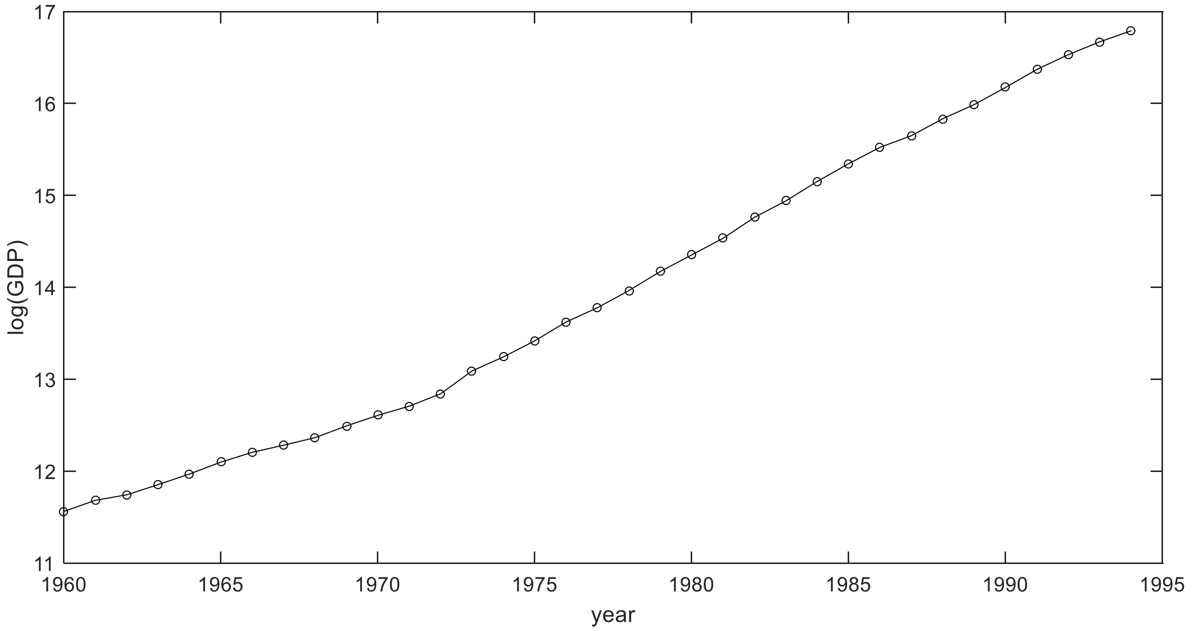


Οι αυτοσυσχετίσεις της σειράς είναι σημαντικές μέχρι και την $6^{η}$ υστέρηση και η συνάρτηση αυτοσυσχέτισης φθίνει με αργό ρυθμό.

Σε αυτές τις περιπτώσεις ενδείκνυται η χρησιμοποίηση των λογαρίθμων αντί των αρχικών της χρονοσειράς.

Στο παρακάτω σχήμα δίνεται η γραφική παράσταση των λογαρίθμων της χρονοσειράς.

```
>> plot((1960:1994),log(GDP),'o-')
```

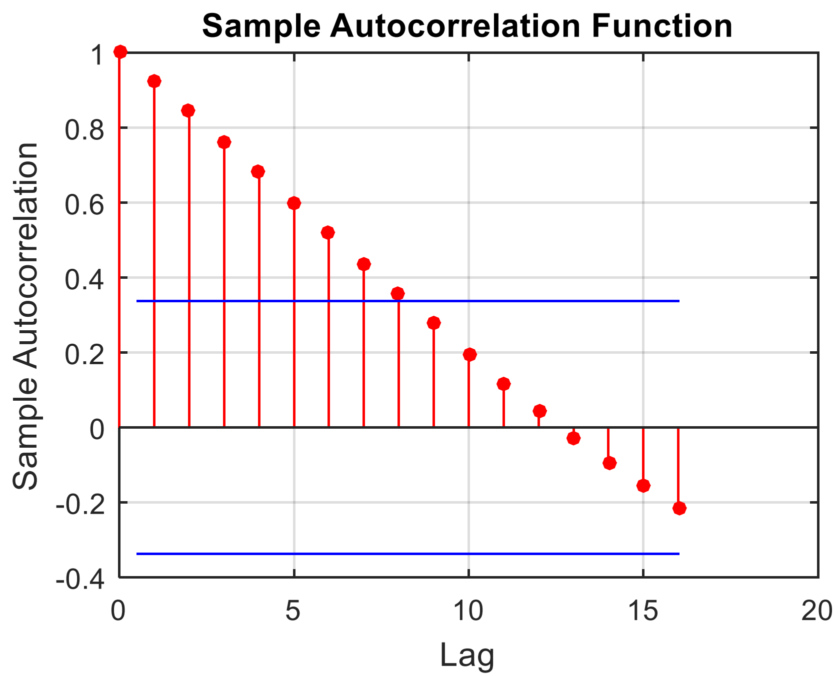


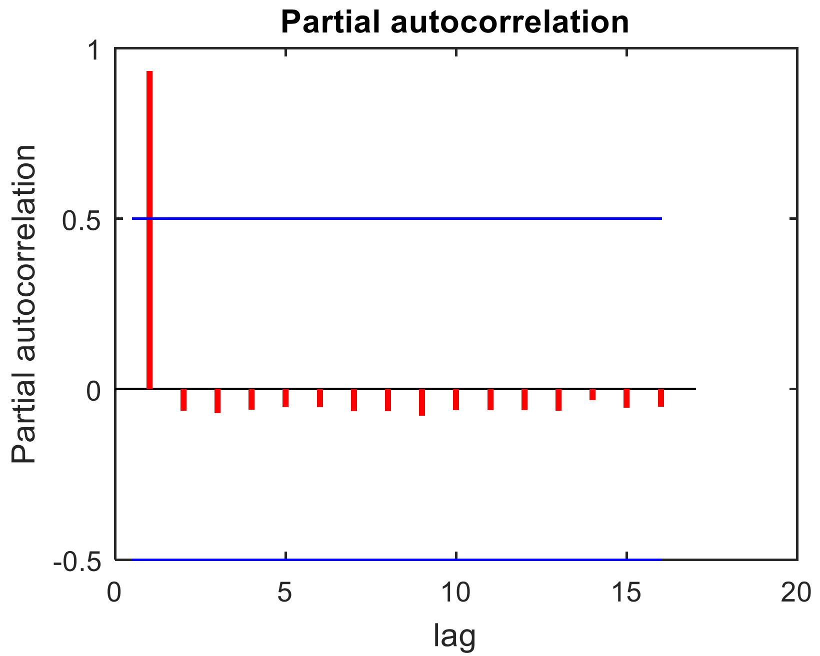
Στο παρακάτω σχήμα δίνονται οι γραφικές παραστάσεις της συνάρτησης αυτοσυσχέτισης

>> autocorr(log(GDP),16)

και μερικής αυτοσυσχέτισης της λογαριθμικής χρονοσειράς

>> pacfV = acf2pacf(acM(2:end),1)



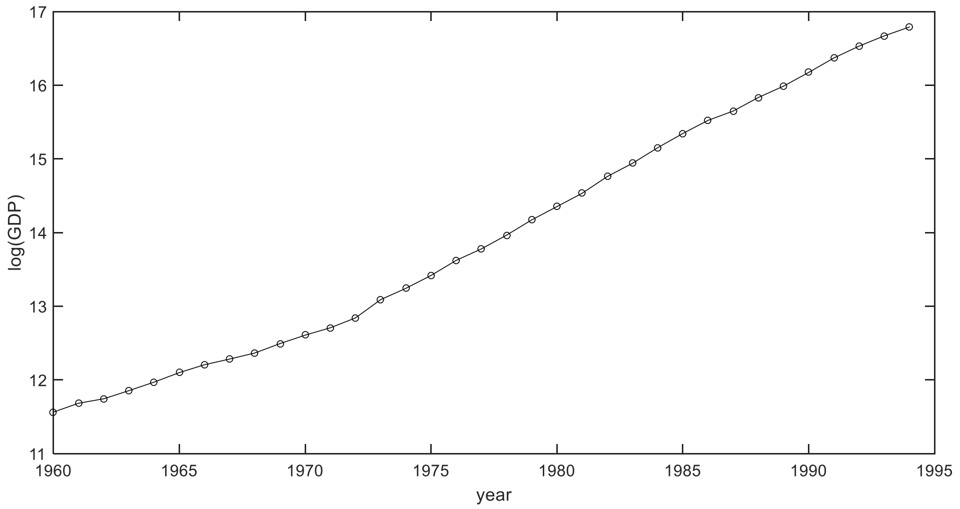


Προφανώς ούτε η λογαριθμική χρονοσειρά μπορεί να θεωρηθεί ως στάσιμη.

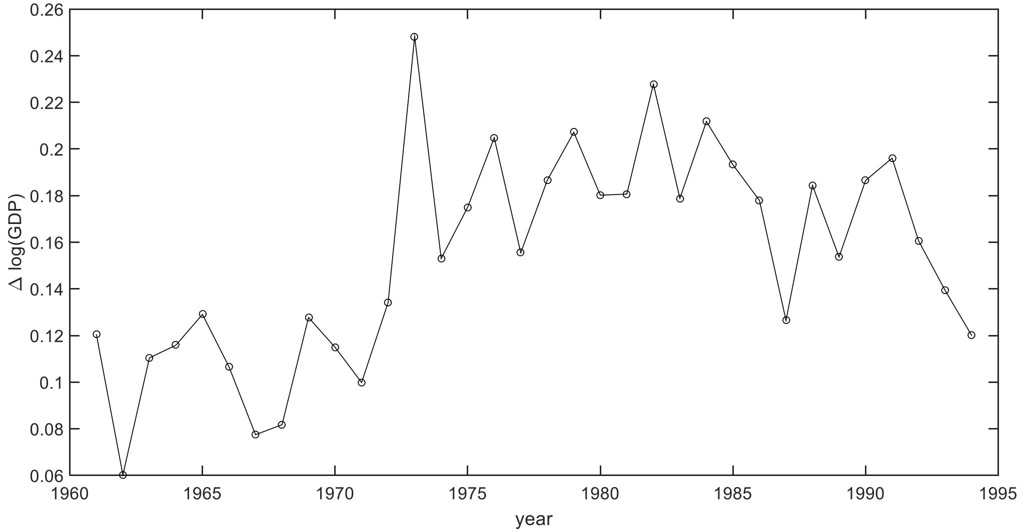
Για αυτό παίρνουμε τις πρώτες διαφορές.

Δηλαδή εξετάζουμε την σειρά:

$x_t = \Delta\log(GDP)_t = \log(GDP)_t - \log(GDP)_{t-1}$




```
>> xV = diff(log(GDP));
```

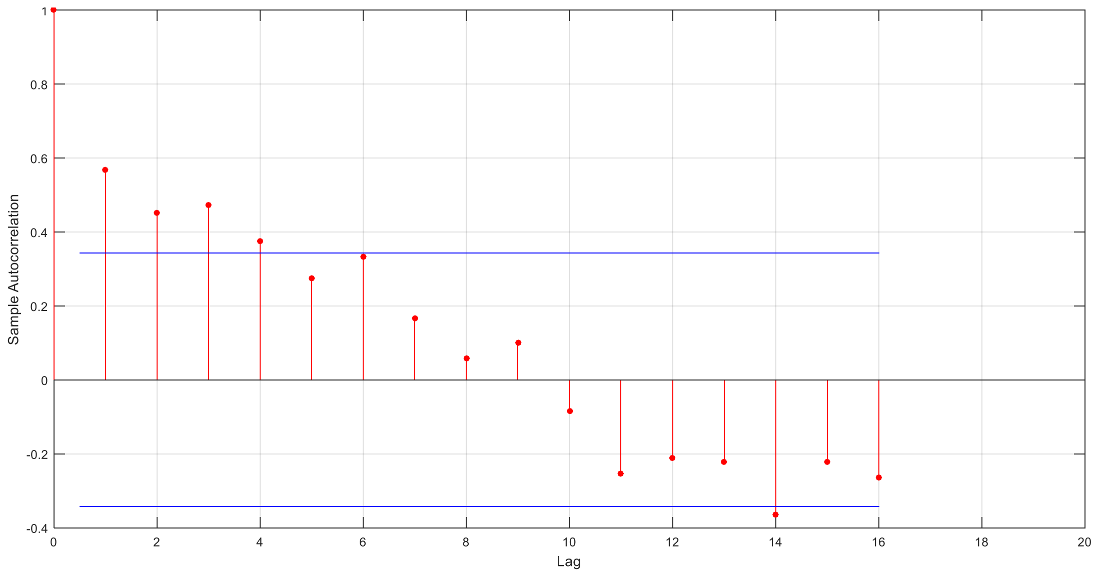


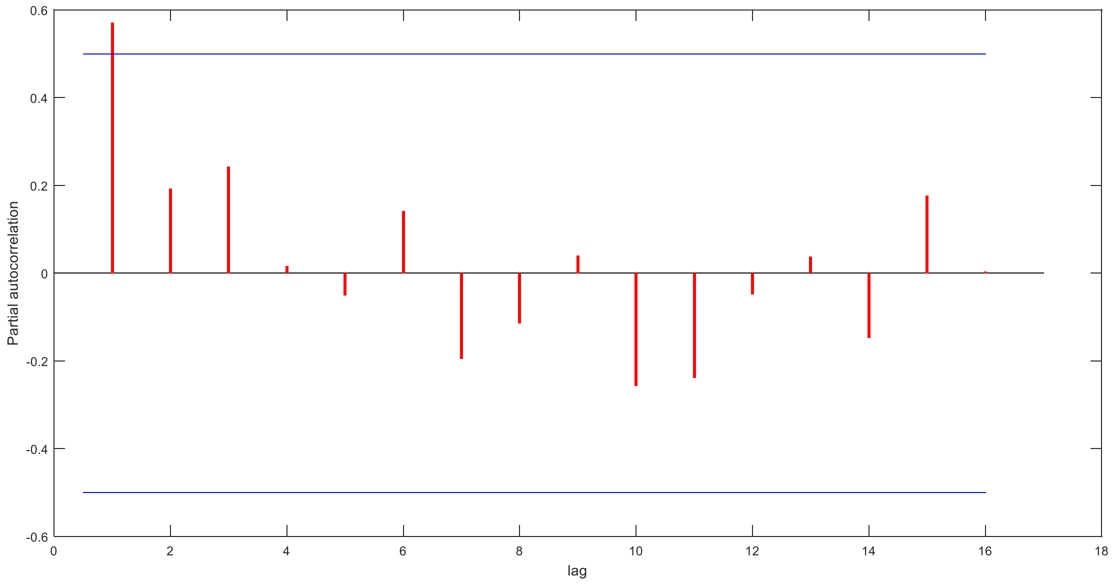
Στο παρακάτω σχήμα δίνονται οι γραφικές παραστάσεις της συνάρτησης αυτοσυσχέτισης

```
>> acM = autocorr(xV,16);
```

και μερικής αυτοσυσχέτισης της λογαριθμικής χρονοσειράς πρώτων διαφορών

```
>> pacfV = acf2pacf(acM(2:end),1)
```





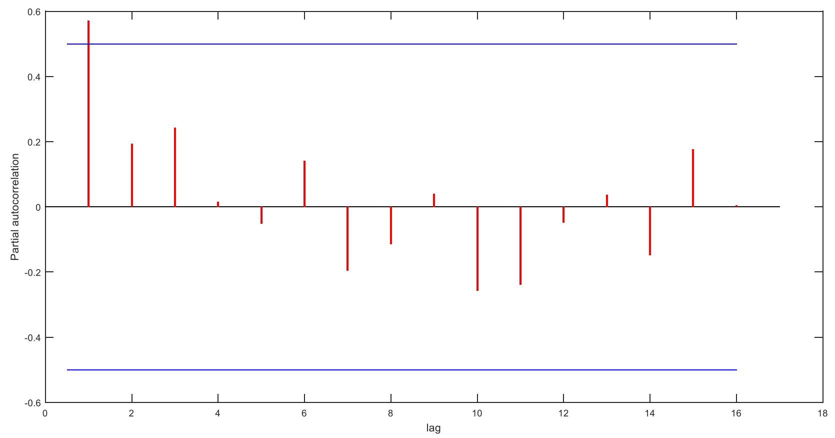
Επειδή η συνάρτηση αυτοσυσχέτισης της λογαριθμικής χρονοσειράς πρώτων διαφορών φθίνει πολύ γρήγορα σε σχέση με την αρχική μη στάσιμη χρονοσειρά, δεν χρειάζεται να πάρουμε δεύτερες διαφορές.

Μπορούμε δηλαδή να θεωρήσουμε ότι η σειρά μας είναι στάσιμη (το οποίο θα επιβεβαιώσουμε με κάποιον έλεγχο μοναδιαίας ρίζας).

Παρατήρηση: Η δειγματική συνάρτηση αυτοσυσχέτισης δεν μπορεί να μιμείται επακριβώς τη συμπεριφορά της θεωρητικής συνάρτησης.

Από την συνάρτηση μερικής αυτοσυσχέτισης βλέπουμε ότι μόνο για υστέρηση 1 είναι σημαντική.

Η συμπεριφορά αυτή είναι χαρακτηριστική ενός μοντέλου AR(1). Επομένως ως πρώτη προσέγγιση μπορούμε να θεωρήσουμε ότι η σειρά ακολουθεί ένα μοντέλο ARIMA(1,1,0).



Υπενθύμιση

Εκτίμηση δειγματικών αυτοσυσχετίσεων μιας σειράς X_t:

$r_s = \hat{\rho}_s = \frac{\sum_{t=1}^{T}(X_t-\overline{X})(X_{t+s}-\overline{X})}{\sum_{t=1}^{T}(X_t-\overline{X})^2}$ ή

$r_s = \hat{\rho}_s = \frac{T-1}{T-s}\frac{\sum_{t=1}^{T}(X_t-\overline{X})(X_{t+s}-\overline{X})}{\sum_{t=1}^{T}(X_t-\overline{X})^2}$ για μικρά δείγματα

Εκτίμηση δειγματικών μερικών αυτοσυσχετίσεων μιας σειράς X_t:

$r_{11} = \hat{\rho}_{11} = \hat{\rho}_1$

$r_{22} = \hat{\rho}_{22} = \frac{\begin{vmatrix} 1 & \hat{\rho}_1 \\ \hat{\rho}_1 & \hat{\rho}_2 \end{vmatrix}}{\begin{vmatrix} 1 & \hat{\rho}_1 \\ \hat{\rho}_1 & 1 \end{vmatrix}} = \frac{\hat{\rho}_2-\hat{\rho}_1^2}{1-\hat{\rho}_1^2}$

…

$$r_{ss} = \frac{\begin{vmatrix} 1 & \hat{\rho}_1 & .. & \hat{\rho}_1 \\ \hat{\rho}_1 & 1 & .. & \hat{\rho}_2 \\ .. & .. & .. & .. \\ \hat{\rho}_{s-1} & \hat{\rho}_{s-2} & .. & \hat{\rho}_s \end{vmatrix}}{\begin{vmatrix} 1 & \hat{\rho}_1 & .. & \hat{\rho}_{s-1} \\ \hat{\rho}_1 & 1 & .. & \hat{\rho}_{s-2} \\ .. & .. & .. & .. \\ \hat{\rho}_{s-1} & \hat{\rho}_{s-2} & .. & 1 \end{vmatrix}}$$

Υπενθύμιση

Στατιστικά σημαντικές δειγματικές αυτοσυσχετίσεις μιας σειράς X_t:

$$r_s > \frac{-2}{\sqrt{\mathrm{T}}}$$

Στατιστικά σημαντικές μερικές δειγματικές αυτοσυσχετίσεις μιας σειράς X_t:

$$r_{ss} > \frac{-2}{\sqrt{\mathrm{T}}}$$

β) Εκτίμηση μοντέλου

Τα αποτελέσματα της εκτίμησης ενός μοντέλου ARMA(1,0) στις πρώτες διαφορές προκύπτουν ως εξής:

```
>> [nrmseV,phiV,thetaV,SDz,aicS,fpeS,armamodel]=fitARMA(xV,1,0,0)
```

Οι συντελεστές του AR(1) είναι:

phiV = 0.0648   0.5785

AIC = -6.4997 (Akaike information criterion)

FPE = 0.0015 (Final prediction error)

MSE: 0.00133 (Mean square error)

Εκτιμώμενο AR(1) μοντέλο για την στάσιμη χρονοσειρά xV (x_t):

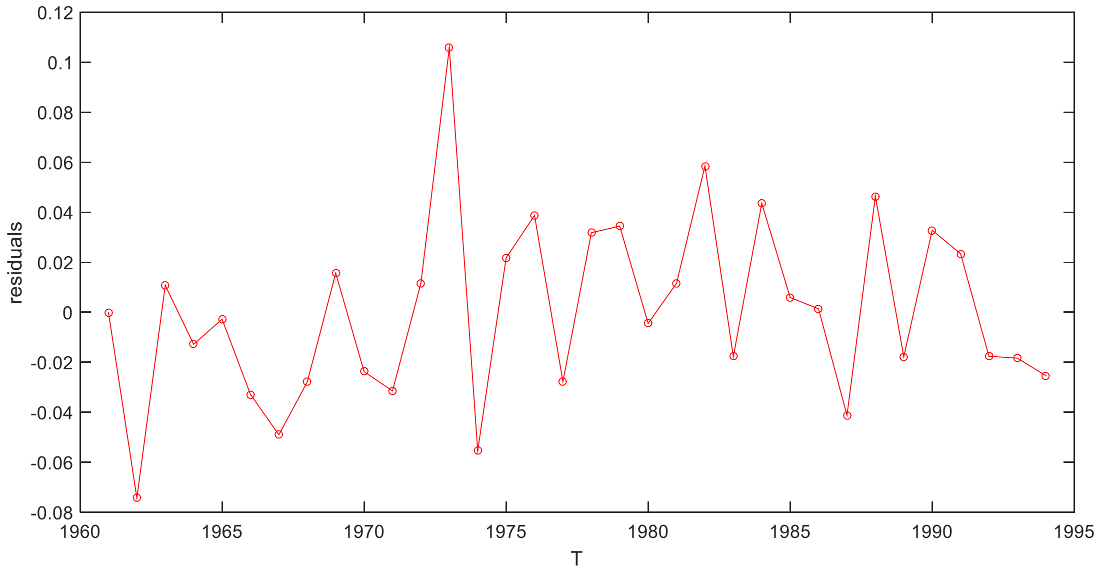
$$\hat{x}_t = 0.0648 + 0.5785 x_{t-1}$$

```
>> T = length(xV);
>> for t=2:T
       yV(t) = 0.0648 + 0.5785*xV(t-1);
   end
```

Οπότε τα κατάλοιπα υπολογίζονται ως:

$$\hat{\varepsilon}_t = x_t - \hat{x}_t$$

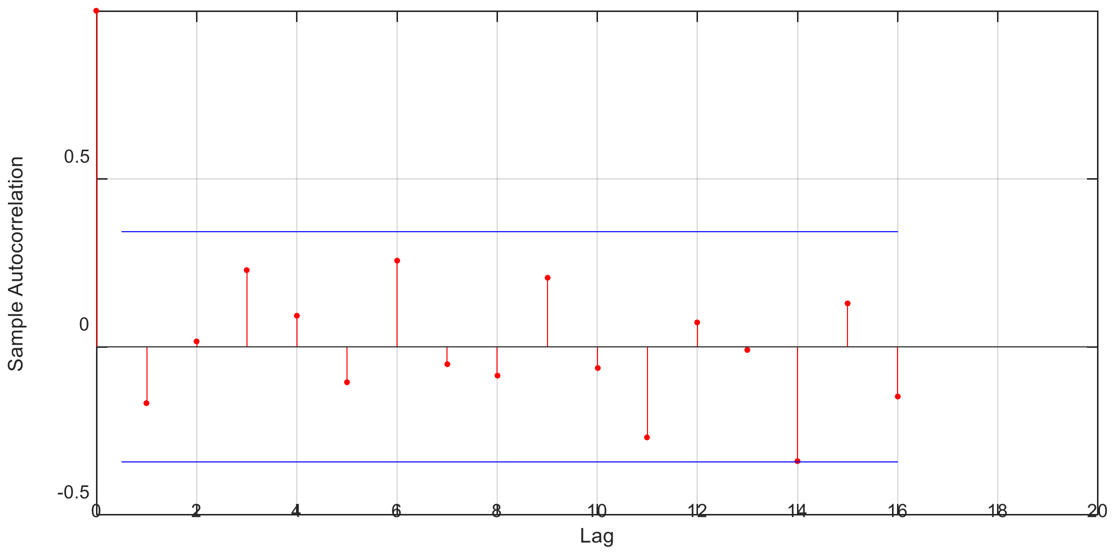
```
>> eV = xV - yV;
```

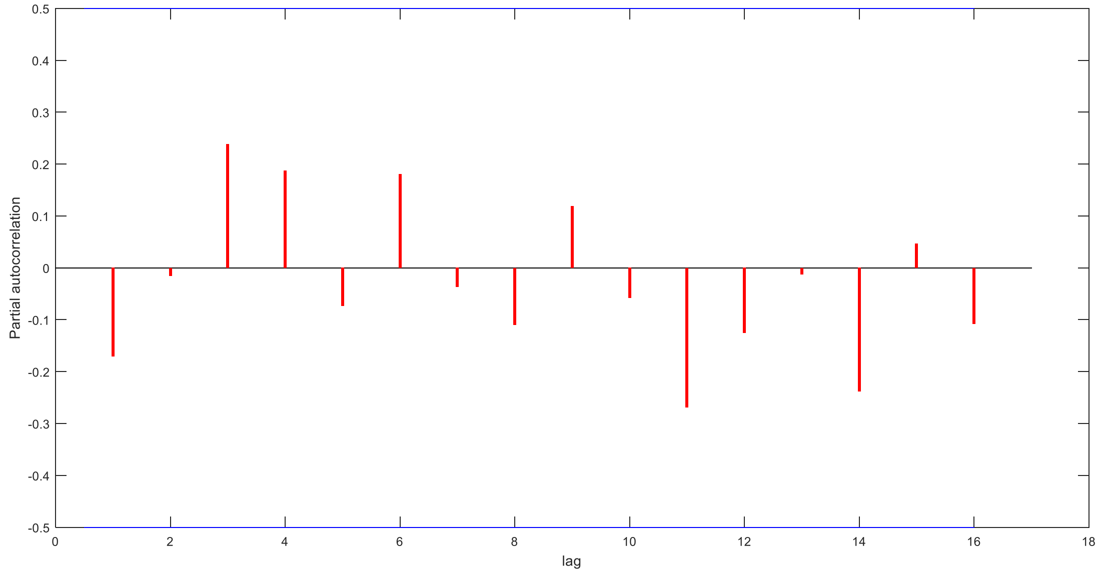


Για να θεωρήσουμε ότι το μοντέλο που εκτιμήσαμε είναι κατάλληλο, πρέπει να μην υπάρχουν αυτοσυσχετίσεις στα κατάλοιπα.

Πράγματι, δεν υπάρχουν αυτοσυσχετίσεις στα κατάλοιπα:

```
>> [acM] = autocorrelation(eV, 16);
>> autocorr(eV, 16)
>> pacfV = acf2pacf(acM(2:end,2),1);
```





Βιβλιογραφία

1. Ε. Μπόρα – Σέντα, Χ. Μωυσιάδης. Εφαρμοσμένη στατιστική, Β’ έκδοση, Εκδόσεις Ζήτη, 1995.
2. Γ. Κ. Χρήστου. Εισαγωγή στην Οικονομετρία, Β τόμος (Γ’ έκδοση), Εκδόσεις Gutenberg, 2007.
3. Δ. Κουγιουμτζής. Σημειώσεις μαθήματος Χρονοσειρών. Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, ΑΠΘ.