Σημειώσεις Στοχαστικές Διεργασίες - 11ο Μάθημα

Τρίτη 9 ιανουαρίου 2024

15. Χρονοσειρές

Τι είναι μία χρονοσειρά;

Στη βιβλιογραφία συναντώνται και οι δύο παρακάτω εκδοχές:

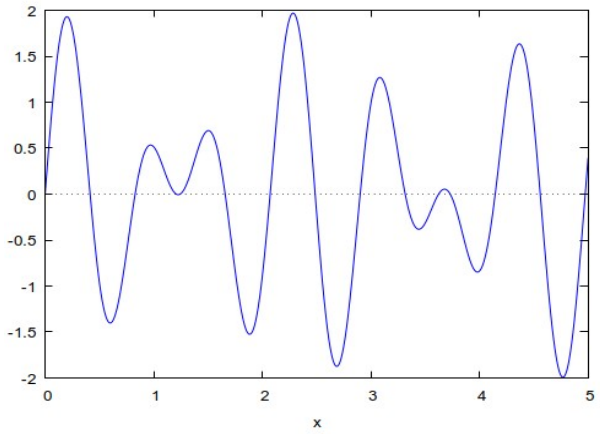
- Χρονοσειρά είναι μία ακολουθία αριθμών που αντιπροσωπεύουν τις τιμές μίας μεταβλητής σε διαφορετικές χρονικές στιγμές.

- Χρονοσειρά είναι μία Στοχαστική Διεργασία Διακριτού Χρόνου {X(t), t ϵ N} με τιμές πραγματικούς αριθμούς.

Στη συνέχεια, με τον όρο χρονοσειρά, θα εννοείται είτε η ίδια η διεργασία {X(t), t ϵ N} είτε μία αριθμητική ακολουθία τιμών που αναπαριστά μία από τις πιθανές υλοποιήσεις της {X(t), t ϵ N}.

Ντετερμινιστική χρονοσειρά

Μία χρονοσειρά μπορεί να είναι ντετερμινιστική αν καθορίζεται από κάποιο γνωστό κανόνα, δηλαδή, αν οι τιμές της προκύπτουν ως αποτέλεσμα μίας γνωστής συνάρτησης όπως για παράδειγμα η



$$f(t) = \sin(2\pi t) + \sin(2\sqrt{2}\pi t)$$

Τυχαία χρονοσειρά

Τυχαία ή μη ντετερμινιστική χρονοσειρά είναι αυτή που δεν μπορεί να περιγραφεί με κάποια συγκεκριμένη συναρτησιακή έκφραση. Μια χρονοσειρά μπορεί να είναι μη ντετερμινιστική επειδή:

- Δεν είναι γνωστές *όλες οι πηγές* της μεταβλητότητας των τιμών της, όπως π.χ. το πλήθος των ατόμων που θα επισκεφτούν ένα αρχαιολογικό χώρο μία μέρα του Ιούνη.

- Η φύση της διαδικασίας που περιγράφεται με αυτήν είναι εγγενώς τυχαία, όπως π.χ. το πλήθος των σωματιδίων που καταφτάνουν σε έναν ανιχνευτή στραμμένο προς το διάστημα.

Καθώς οι μη ντετερμινιστικές χρονοσειρές εξελίσσονται με τυχαίο τρόπο, η περιγραφή της εξέλιξής τους αρμόζει να γίνει με όρους της θεωρίας πιθανοτήτων, δηλαδή με χρήση κατανομών πιθανοτήτων.

Η αναγνώριση μίας χρονοσειράς ως μία από τις πιθανές υλοποιήσεις μίας στοχαστικής διεργασίας, επιτρέπει την εφαρμογή μεθόδων της Στατιστικής για τη μελέτη των τυχαίων χρονοσειρών. Στο παραπάνω πλαίσιο:

Το σύνολο των πιθανών εξελίξεων μιας στοχαστικής διεργασίας είναι όλος ο στατιστικός πληθυσμός.

Μία παρατηρούμενη χρονοσειρά είναι μια δειγματική μονάδα.

Ένα σύνολο παρατηρούμενων χρονοσειρών είναι ένα δείγμα του πληθυσμού.

Στόχοι της ανάλυσης μίας χρονοσειράς

Ο βασικός στόχος της ανάλυσης είναι ο προσδιορισμός ενός μοντέλου που περιγράφει το μοτίβο που ακολουθεί η χρονοσειρά. Ένα τέτοιο μοντέλο, θα μπορούσε να χρησιμοποιηθεί για:

- Να περιγράψει τα σημαντικά χαρακτηριστικά της χρονοσειράς.
- Να εξηγήσει πώς το παρελθόν επηρεάζει το μέλλον.
- Να προβλέψει τις μελλοντικές τιμές της σειράς.

Τύποι αναλυτικών μοντέλων

Υπάρχουν πολλοί τύποι αναλυτικών μοντέλων.

Παρακάτω θα περιγραφούν οι εξής:

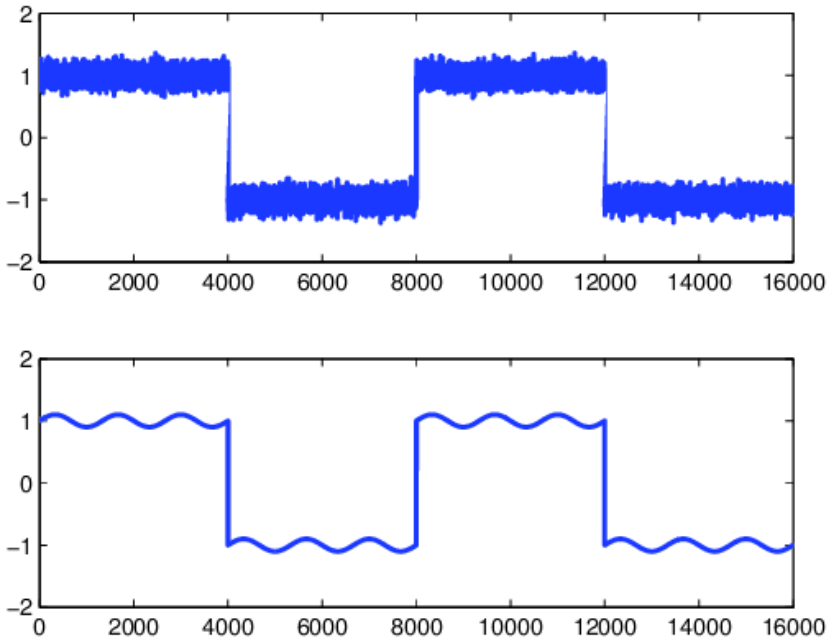
- Μοντέλα ARIMA (Autoregressive Integrated Moving Average) με τα οποία συσχετίζεται κάθε τιμή μιας χρονοσειράς με προηγούμενες τιμές και προηγούμενα σφάλματα πρόβλεψης.
- Μοντέλα γραμμικής παλινδρόμησης στα οποία ο χρόνος τοποθετείται στη θέση της ανεξάρτητης μεταβλητής x ή/και αλγεβρικούς όρους με επιθυμητές ιδιότητες.

Σημαντικά ερωτήματα

Μερικά σημαντικά ερωτήματα που προκύπτουν κατά την παρατήρηση μιας χρονοσειράς είναι:

- Υπάρχει τάση στη σειρά, δηλαδή, οι μετρήσεις τείνουν κατά μέσο όρο να αυξάνονται (ή να μειώνονται) με την πάροδο του χρόνου;
- Υπάρχει εποχικότητα, δηλαδή υπάρχει ένα επαναλαμβανόμενο μοτίβο με υψηλές και χαμηλές τιμές που να σχετίζονται με την χρονική στιγμή;
- Υπάρχουν ιδιάζουσες ή ακραίες τιμές, δηλαδή τιμές που απέχουν ασυνήθιστα πολύ από τις υπόλοιπες παρατηρήσεις;

- Υπάρχει μακροχρόνιος κύκλος ή κάποια άλλη περιοδικότητα που δεν σχετίζεται με εποχιακούς παράγοντες;
- Υπάρχει μεταβολή στη διακύμανση των τιμών με την πάροδο του χρόνου;
- Υπάρχουν απότομες μεταβολές στις τιμές της σειράς είτε στη διακύμανση αυτής;
- Οι τιμές της χρονοσειράς σε ένα χρονικό διάστημα συσχετίζονται με τις τιμές σε κάποιο προηγούμενο διάστημα; (υψηλή αυτοσυσχέτιση)



Εικόνα: Παραδείγματα σημάτων με υψηλή αυτοσυσχέτιση

15.1 Συνδιακύμανση και Συσχέτιση

Η **συνδιακύμανση Cov(X, Y)** είναι μία στατιστική ποσότητα που μας επιτρέπει να ποσοτικοποιήσουμε τον τρόπο με τον οποίο συμμεταβάλλονται δύο τυχαίες μεταβλητές X, Y. Η κανονικοποιημένη εκδοχή της συνδιακύμανσης είναι ο **συντελεστής συσχέτισης Pearson $\rho_{X,Y}$**, ο οποίος παίρνει τιμές από -1 έως +1.

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\operatorname{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

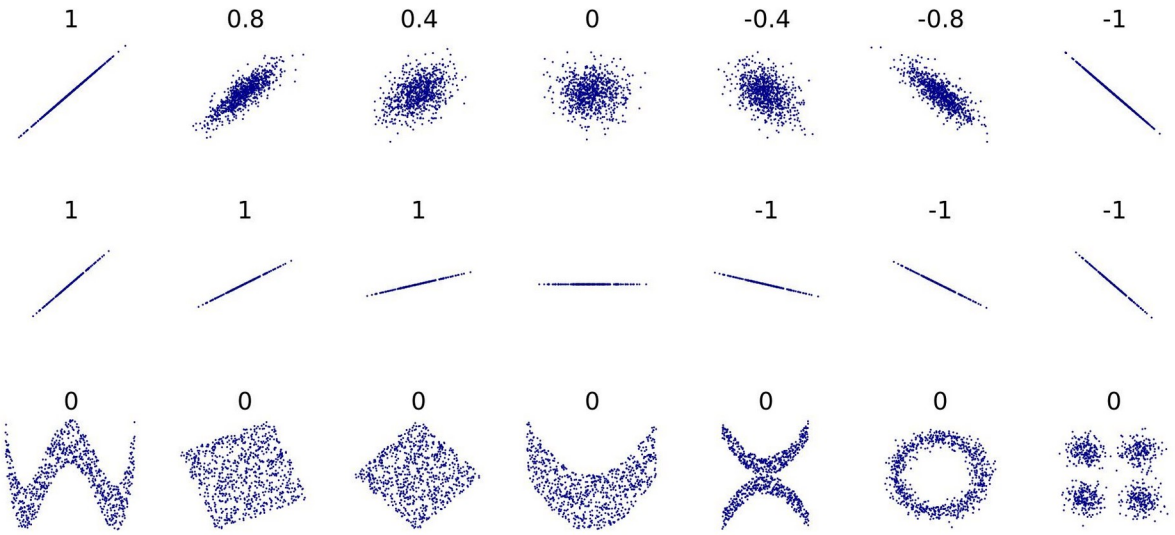
$$\rho_{X,Y} = \frac{\operatorname{E}(XY) - \operatorname{E}(X)\operatorname{E}(Y)}{\sqrt{\operatorname{E}(X^2) - \operatorname{E}(X)^2} \cdot \sqrt{\operatorname{E}(Y^2) - \operatorname{E}(Y)^2}}$$

Δειγματική Συσχέτιση

Στην περίπτωση όπου οι δύο τυχαίες μεταβλητές εκπροσωπούνται από ένα δείγμα τιμών, ο δειγματικός συντελεστής συσχέτισης Pearson συμβολίζεται με $r_{X,\,Y}$, και υπολογίζεται ως εξής:

$$r_{xy} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

Ενδεικτικά διαγράμματα διασποράς για δείγματα τιμών δύο τυχαίων μεταβλητών X, Y μαζί με τον αντίστοιχο δειγματικό συντελεστή συσχέτισης Pearson $r_{X,\,Y}$.



Αν $r_{X,\,Y} \approx 1$, τότε η εξίσωση $Y = \alpha X + \beta$, προσεγγίζει με μεγάλη αξιοπιστία την πραγματική σχέση των μεταβλητών X, Y.

Μερικός Συντελεστής Συσχέτισης

Ο μερικός συντελεστής συσχέτισης (partial correlation coefficient) υπολογίζει τη συσχέτιση δύο μεταβλητών X, Y ύστερα από την απομάκρυνση της μεταβλητότητας των X, Y που εξηγείται από μία τρίτη μεταβλητή Z.

$$\rho_{X,\,Y\,|\,Z} = \frac{\mathrm{Cov}(X,\,Y\,|\,Z)}{\sigma_{X\,|\,Z}\cdot\sigma_{Y\,|\,Z}}$$

Ο μερικός συντελεστής συσχέτισης δίνει περισσότερο ακριβή εικόνα για τη σχέση δύο μεταβλητών στην περίπτωση που είναι γνωστή μία τρίτη μεταβλητή που συσχετίζεται και με τις δύο.

Παράδειγμα

Σε ένα δείγμα μαθητών δημοτικού, αν

X: Ικανότητα Απομνημόνευσης, Y: Ρυθμός ομιλίας,

τότε ο απλός συντελεστής συσχέτισης Pearson $\rho_{X,\,Y}$, δίνει μία εκτίμηση για τη σχέση των δύο μεταβλητών. Καθώς όμως, τόσο η ικανότητα απομνημόνευσης όσο και ο ρυθμός ομιλίας εξαρτώνται από την ηλικία (Z) του παιδιού, ο μερικός συντελεστής συσχέτισης $\rho_{X,\,Y\,|\,Z}$ αποτελεί περισσότερο ακριβής εκτίμηση για την πραγματική σχέση των δύο μεταβλητών (το παράδειγμα αναπτύσσεται στην [1]).

[1] https://personal.utdallas.edu/~herve/Abdi-PartialRegressionCoefficient2007-pretty.pdf

Ο υπολογισμός του μερικού συντελεστή συσχέτισης μπορεί να γενικευθεί για περισσότερες της μίας μεταβλητής, $Z_1, Z_2, \ldots, Z_n$.

$$\rho_{X,\,Y\,|\,Z_1,\,Z_2,\,\ldots,\,Z_n} = \frac{\mathrm{Cov}\left(X,\,Y\,|\,Z_1,\,Z_2,\,\ldots,\,Z_n\right)}{\sigma_{X\,|\,Z_1,\,Z_2,\,\ldots,\,Z_n} \cdot \sigma_{Y\,|\,Z_1,\,Z_2,\,\ldots,\,Z_n}}$$

Ημιμερικός Συντελεστής Συσχέτισης

Στην περίπτωση όπου αφαιρείται η κοινή μεταβλητότητα της Z μόνο με τη X ή μόνο με τη Y, τότε ο συντελεστής καλείται **ημιμερικός συντελεστής συσχέτισης** (semipartial ή part correlation coefficient).

Υπολογισμός ημιμερικού συντελεστή συσχέτισης

Για τον υπολογισμό του δειγματικού μερικού συντελεστή συσχέτιση αρκεί να υπολογίσουμε τις γραμμικές εξισώσεις πρόβλεψης των X, Y από τις $Z_1, Z_2, \ldots, Z_n$.

$$X = a_1 + b_1 \cdot Z_1 + b_2 \cdot Z_2 + \ldots + b_n \cdot Z_n + e_1$$

$$Y = a_1 + b_1 \cdot Z_1 + b_2 \cdot Z_2 + \ldots + b_n \cdot Z_n + e_1$$

Στη συνέχεια, ο **δειγματικός μερικός συντελεστής συσχέτισης** υπολογίζεται ως ο συντελεστής συσχέτισης Pearson των υπολοίπων e_1, e_2 των δύο γραμμικών μοντέλων.

15.2 Αυτοδιακύμανση και Αυτοσυσχέτιση

Η αυτοδιακύμανση (autocovariance) και η αυτοσυσχέτιση (autocorrelation) είναι δύο στατιστικές ποσότητες που μας επιτρέπουν να ανιχνεύσουμε τυχόν περιοδικότητα μίας χρονοσειράς $\{X_n, n \in N^*\} = \{X_1, X_2, X_3, \ldots\}$ με ένα μετρήσιμο τρόπο.

Η **αυτοδιακύμανση** είναι η συνδιακύμανση της μεταβλητής X_t των τιμών της χρονοσειράς με ένα αντίγραφό της X_{t-h} χρονικά μετατοπισμένο κατά πλήθος χρονικών στιγμών h = 1, 2, …

$h = 1$: $Cov(X_t, X_{t-1}) = Cov(\{X_1, X_2, X_3, \ldots\}, \{X_2, X_3, X_4, \ldots\})$

$h = 2$: $Cov(X_t, X_{t-2}) = Cov(\{X_1, X_2, X_3, \ldots\}, \{X_3, X_4, X_5, \ldots\})$

….

$h = n$: $Cov(X_t, X_{t-n}) = Cov(\{X_1, X_2, X_3, \ldots\}, \{X_{n+1}, X_{n+2}, X_{n+3}, \ldots\})$

Σημείωση: Φανερά, κάθε μία αυτοδιακύμανση υπολογίζεται σε πλήθος παρατηρήσεων μειωμένο κατά ένα από την προηγούμενη.

Η αυτοσυσχέτιση είναι η κανονικοποιημένη ως προς το πιθανό εύρος τιμών εκδοχή της αυτοδιακύμανσης.

$$\rho_1 = Corr(X_t, X_{t-1}) = Cov(X_t, X_{t-1}) / [SD(X_t)SD(X_{t-1})]$$

$$\rho_2 = Corr(X_t, X_{t-2}) = Cov(X_t, X_{t-2}) / [SD(X_t)SD(X_{t-2})]$$

…

$$\rho_n = Corr(X_t, X_{t-n}) = Cov(X_t, X_{t-n}) /[SD(X_t)SD(X_{t-n})]$$

Στην περίπτωση όπου $\{x_1, x_2, x_3, \ldots\}$ είναι συγκεκριμένες αριθμητικές τιμές τότε με αυτές εκφράζεται μία εκδοχή της στοχαστικής διεργασίας $\{X_n, n \in N^*\} = \{X_1, X_2, X_3, \ldots\}$ που τις παράγει.

Στην περίπτωση αυτή η $Cov(x_t, x_{t-n})$ ονομάζεται **δειγματική αυτοδιακύμανση** και η $Corr(x_t, x_{t-n})$ **δειγματική αυτοσυσχέτιση**.

Δειγματική Συνάρτηση Αυτοσυσχέτισης (ACF)

Ο υπολογισμός της αυτοσυσχέτισης για όλες τις υστερήσεις Lag = 1, 2, 3, …, οδηγεί στον ορισμό της δειγματικής συνάρτησης αυτοσυσχέτισης (Autocorrelation Function – ACF). Πιο συγκεκριμένα, αν x_1, x_2, x_3, …, x_n είναι τα χρονικά δεδομένα, τότε η ACF είναι η συνάρτηση $r(h) = r_h: N \rightarrow R$, με

$r_0 = Corr(\{x_1, x_2, x_3, \ldots, x_n\}, \{x_1, x_2, x_3, \ldots, x_n\})$,

$r_1 = Corr(\{x_1, x_2, x_3, \ldots, x_{n-1}\}, \{x_2, x_3, x_4, \ldots, x_n\})$,

$r_2 = Corr(\{x_1, x_2, x_3, \ldots, x_{n-2}\}, \{x_3, x_4, x_5, \ldots, x_n\})$,

….

$r_h = Corr(\{x_1, x_2, x_3, \ldots, x_{n-h}\}, \{x_{h+1}, x_{h+2}, x_{h+3}, \ldots, x_n\})$,

Η αυτοσυσχέτιση ελέγχει κατά πόσο ταυτίζονται οι μεταβολές της χρονοσειράς σε όλο της το μήκος. Αυτοσυσχέτιση με υστέρηση 0, είναι η συσχέτιση της χρονοσειράς με ένα ακριβές αντίγραφό της, δηλαδή ίση με 1.

Υψηλή αυτοσυσχέτιση μίας χρονοσειράς με αντίγραφό της που υστερεί κατά h βήματα υποδεικνύει ότι ύστερα από h βήματα η μεταβολή των τιμών συμβαίνει κατά τον ίδιο τρόπο, δηλαδή είτε μαζί αυξάνεται η χρονοσειρά και το αντίγραφό της είτε μειώνονται μαζί, **υποδεικνύοντας με τον τρόπο αυτό κάποιου είδους περιοδικότητα στις τιμές της χρονοσειράς.**

Τρόπος υπολογισμού r_h για τη χρονοσειρά $\{x_1, x_2, x_3, \ldots, x_n\}$

1ο βήμα: Υπολογισμός γενικού αριθμητικού μέσου

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2ο βήμα: Υπολογισμός δειγματικής αυτοσυσχέτισης

$$r_h = \frac{Cov(x_n, x_{n-h})}{Var(x_n)} = \frac{\sum_{i=1+h}^{n}(x_i - \bar{x})(x_{i-h} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, h = 0, 1, , \ldots, n-1.$$

Στην πράξη, υπολογίζουμε την αυτοσυσχέτιση της χρονοσειράς με τον εαυτό της σε όλες τις πιθανές μετατοπίσεις και παρατηρούμε τις τιμές που προκύπτουν. Συνήθως, αυτές απεικονίζονται σε ένα διάγραμμα, **το Correlogram (διάγραμμα αυτοσυσχετίσεων ή διάγραμμα ACF)**

Δειγματική Συνάρτησης Μερικής Αυτοσυσχέτισης

Αν $x_1, x_2, x_3, \ldots, x_n$ είναι τα χρονικά δεδομένα, τότε η συνάρτηση μερικής αυτοσυσχέτισης φ_h με υστέρηση (lag) h, (Partial Autocorrelation Function – PACF) είναι η αυτοσυσχέτιση μεταξύ των $x_1, x_2, x_3, \ldots, x_{n-h}$ και $x_{h+1}, x_{h+2}, x_{h+3}, \ldots, x_n$ από την οποία αφαιρείται η γραμμική εξάρτηση όλων των ενδιάμεσων υστερήσεων 1, 2, …, h – 1.

Δηλαδή

$$\varphi_1 = \mathrm{Corr}(x_{t+1}, x_t) = r_1$$

$$\varphi_h = \mathrm{Corr}\left(x_{t+h} - \hat{x}_{t+h},\ x_t - \hat{x}_t\right),\ h \geq 2.$$

όπου $\hat{x}_{t+h}$ και $\hat{x}_t$ είναι γραμμικοί συνδυασμοί των $x_{t+1}, x_{t+2}, x_{t+3}, \ldots, x_{t+h-1}$ που ελαχιστοποιούν το μέσο τετράγωνο σφάλμα των x_{t+h} και x_t αντίστοιχα.

Στην πράξη, υπολογίζουμε την μερική αυτοσυσχέτιση της χρονοσειράς με τον εαυτό της σε όλες τις πιθανές μετατοπίσεις και παρατηρούμε τις τιμές που προκύπτουν. Συνήθως, αυτές απεικονίζονται σε ένα διάγραμμα, **το Partial Correlogram (διάγραμμα μερικών αυτοσυσχετίσεων ή διάγραμμα PACF).**

Από το διάγραμμα αυτοσυσχετίσεων (ACF) μπορούμε να δούμε αν η χρονοσειρά μας προσαρμόζεται ικανοποιητικά σε ένα αυτοπαλινδρομικό μοντέλο, ενώ από το διάγραμμα μερικών αυτοσυσχετίσεων (PACF), μπορούμε να δούμε πόσους όρους πρέπει να έχει αυτό το μοντέλο.

Ενδεικτικές περιπτώσεις αναγνώρισης της δομής μίας χρονοσειράς

Αν $|r_h| > 0{,}7$, για κάποιο h ϵ N, τότε υποδεικνύεται πως η χρονοσειρά προσαρμόζεται καλά σε ένα μοντέλο της μορφής $x_{t+h} = \alpha x_t + b$.

Αν $|r_h| > 0{,}7$ και $\varphi_h \approx 0$, για κάποιο h ϵ N, τότε υποδεικνύεται πως η χρονοσειρά προσαρμόζεται καλά σε ένα μοντέλο της μορφής $x_{t+h} = \alpha x_t + b$, αλλά είναι οι ενδιάμεσοι όροι που επεξηγούν αυτή τη σχέση. Συνεπώς, στην περίπτωση αυτή, αναζητούμε κάποιο μοντέλο που να περιέχει τους όρους με υστέρηση lag < h.

Αν $|r_h| > 0{,}7$ και $|\varphi_h| > 0{,}7$, για κάποιο h ϵ N, τότε υποδεικνύεται πως η χρονοσειρά προσαρμόζεται καλά σε ένα μοντέλο της μορφής $x_{t+h} = \alpha x_t + b$, και επιπλέον, πως η σχέση αυτή δεν επεξηγείται από τους ενδιάμεσους όρους. Συνεπώς, στην περίπτωση αυτή, αναζητούμε κάποιο μοντέλο που να περιέχει και κάποιους ή όλους τους όρους με υστέρηση lag < h.

Γενικές συμβουλές για την ερμηνεία του διαγράμματος συσχετίσεων

Τυχαία χρονοσειρά (Λευκός Θόρυβος)

Εάν μια χρονοσειρά είναι εντελώς τυχαία, τότε για ικανά μεγάλο δείγμα μεγέθους K, θα είναι $r_h \approx 0$, $h > 0$. Αυτό συμβαίνει γιατί σε μια τυχαία χρονοσειρά (αποδεικνύεται ότι) είναι περίπου $r_h \sim N(0, 1/K)$. Έτσι, αναμένουμε το 95% των τιμών (19 από τις 20 τιμές) να βρίσκονται μεταξύ $-2/K^{½}$ και $+2/K^{½}$.

Βραχυπρόθεσμη συσχέτιση

Οι στάσιμες χρονοσειρές συχνά εμφανίζουν βραχυπρόθεσμη συσχέτιση που χαρακτηρίζεται από μια αρκετά μεγάλη τιμή του r_1, η οποία ακολουθείται από από 2 ή 3 συντελεστές σημαντικά μεγαλύτερους από το μηδέν και στατιστικά μηδενικούς συντελεστές για μεγαλύτερη καθυστέρηση. Σε μια χρονοσειρά με τέτοιο συσχετισμό τιμών αυτοσυσχέτισης r_h, κάθε μια παρατήρηση πάνω από τον μέσο όρο τείνει να ακολουθείται από μία ή περισσότερες παρατηρήσεις πάνω από το μέσο όρο, και στη συνέχεια με παρατηρήσεις κάτω από το μέσο όρο. Ένα αυτοπαλινδρομικό μοντέλο ενδέχεται να είναι κατάλληλο για σειρές αυτού του τύπου.

Εναλλασσόμενη σειρά

Εάν μια χρονοσειρά έχει την τάση να εναλλάσσεται με διαδοχικές παρατηρήσεις σε διαφορετικές πλευρές του συνολικού μέσου όρου, τότε η συνάρτηση αυτοσυσχέτισης τείνει επίσης να εναλλάσσεται ως προς το πρόσημό της. Έτσι καθώς οι τιμές (x_1, x_2), (x_2, x_3), … θα βρίσκονται από διαφορετικές μεριές του μέσου όρου αναμένουμε να είναι $r_1 < 0$, ενώ αντίστοιχα, καθώς οι τιμές (x_1, x_3), (x_2, x_4), … θα βρίσκονται από την ίδια μεριά του μέσου όρου αναμένουμε να είναι $r_2 > 0$.

Μη στάσιμες σειρές

Εάν μια χρονοσειρά περιέχει μια τάση, τότε η τιμή του r_h δεν θα μειωθεί στο μηδέν παρά μόνο όταν η υστέρηση h γίνει πολύ μεγάλη. Αυτό οφείλεται στο γεγονός ότι ένας μεγάλος αριθμός συνεχόμενων παρατηρήσεων θα βρίσκεται στην ίδια πλευρά του μέσου όρου λόγω της τάσης.

Εποχιακές διακυμάνσεις

Εάν μια χρονοσειρά περιέχει μια εποχιακή διακύμανση, τότε η τιμής της αυτοσυσχέτισης r_h, θα παρουσιάζει επίσης μια ταλάντωση στην ίδια περίοδο της διακύμανσης. Ειδικότερα, αν η χρονοσειρά $\{x_n\}$ ακολουθεί ένα ημιτονοειδές μοτίβο, τότε το ίδιο θα συμβαίνει για την r_h. Στην περίπτωση αυτή, καλό είναι να αφαιρεθεί η εποχιακή διακύμανση και να μελετηθεί η σειρά των υπόλοιπων τιμών.

Σειρά με ιδιάζουσες ή ακραίες τιμές

Εάν μια χρονοσειρά περιέχει μία ή περισσότερες ιδιάζουσες (outliers) ή ακραίες (extreme) τιμές, το διάγραμμα αυτοσυσχετίσεων μπορεί να επηρεαστεί σοβαρά. Εάν υπάρχει μία ακραία τιμή στη χρονοσειρά, τότε η γραφική παράσταση του r_h θα περιέχει δύο ακραία

σημεία, τα οποία θα τείνουν να μειώνουν τους συντελεστές συσχέτισης του δείγματος προς το μηδέν. Εάν υπάρχουν δύο ακραίες τιμές, αυτή η επίδραση είναι πιο αισθητή.

Στα παρακάτω παραδείγματα, αξιοποιείται η συνάρτηση της R:

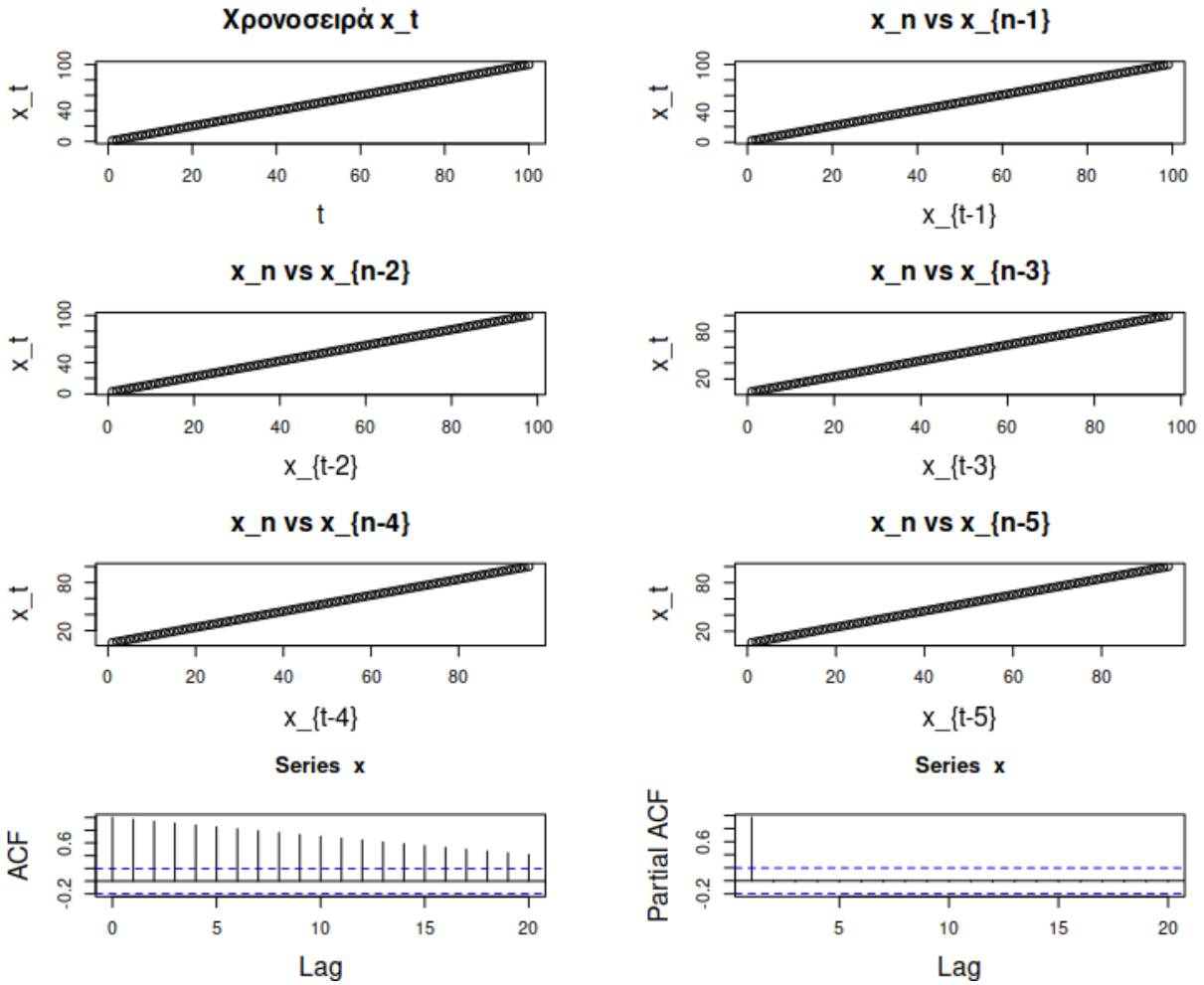
```
my.plot.ts = function(x){
  par(mfrow=c(4,2))
  par(mar = c(4, 4.5, 4, 4.5))
  plot(x, main = "Χρονοσειρά x_t", xlab = "t", ylab = "x_t", cex.lab = 1.5, cex.main = 1.5)
  for(l in 1:5){
    plot(na.omit(Lag(x, l)), x[(l+1):100], main = paste0('x_n vs x_{n-', l,"}"), ylab = 'x_t', xlab
= paste0('x_{t-', l,"}"), cex.main = 1.5,cex.lab = 1.5)
  }
  acf(x, pl=T, cex.lab = 1.5, cex.main = 1.5)
  pacf(x, pl=T, cex.lab = 1.5, cex.main = 1.5)
  par(mfrow=c(1,1))
}
```

Παράδειγμα 1 (Κώδικας R + output)

Χρονοσειρά x_t = t, t = 1, .., 100

x = c(rep(1:100))

my.plot.ts(x)



Σχόλια

Η θετική γραμμική σχέση μεταξύ των μετρήσεων με οποιαδήποτε υστέρηση (lag) είναι εμφανής και αντανακλάται στις μεγάλες τιμές της αυτοσυσχέτισης (διάγραμμα ACF). Η επιπλέον θετική συσχέτιση μεταξύ των ενδιάμεσων αντιγράφων της χρονοσειράς, αντανακλάται στις τιμές του PACF και υποδεικνύει πως η εξίσωση της μορφής

$$x_t = \alpha x_{t-1} + \beta,$$

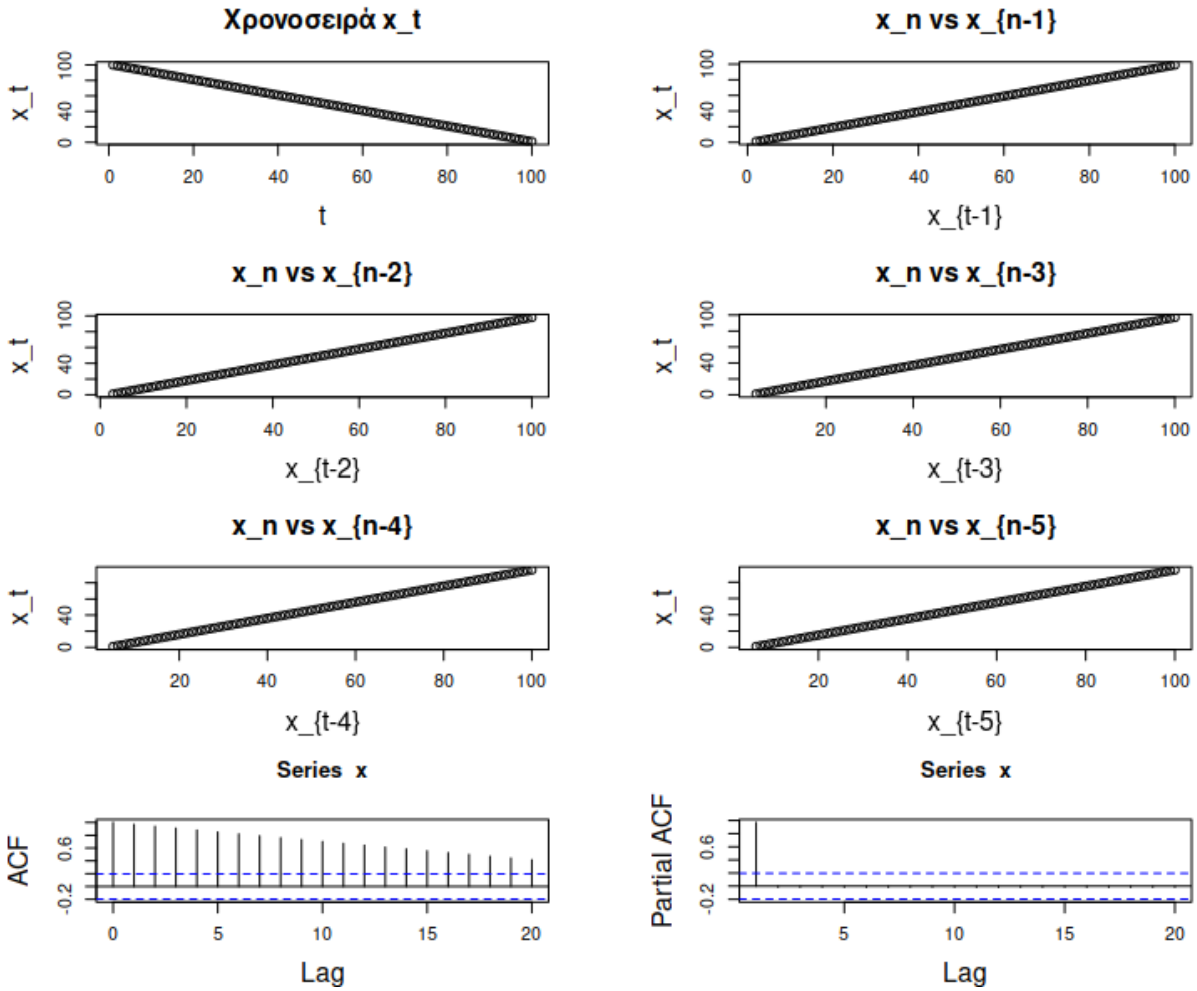
προσαρμόζεται καλά στα χρονικά δεδομένα.

Παράδειγμα 2 (Κώδικας R + output)

Χρονοσειρά x_t = 100 – t, t = 1, .., 100

x = c(rep(100:1))

my.plot.ts(x)



Σχόλια

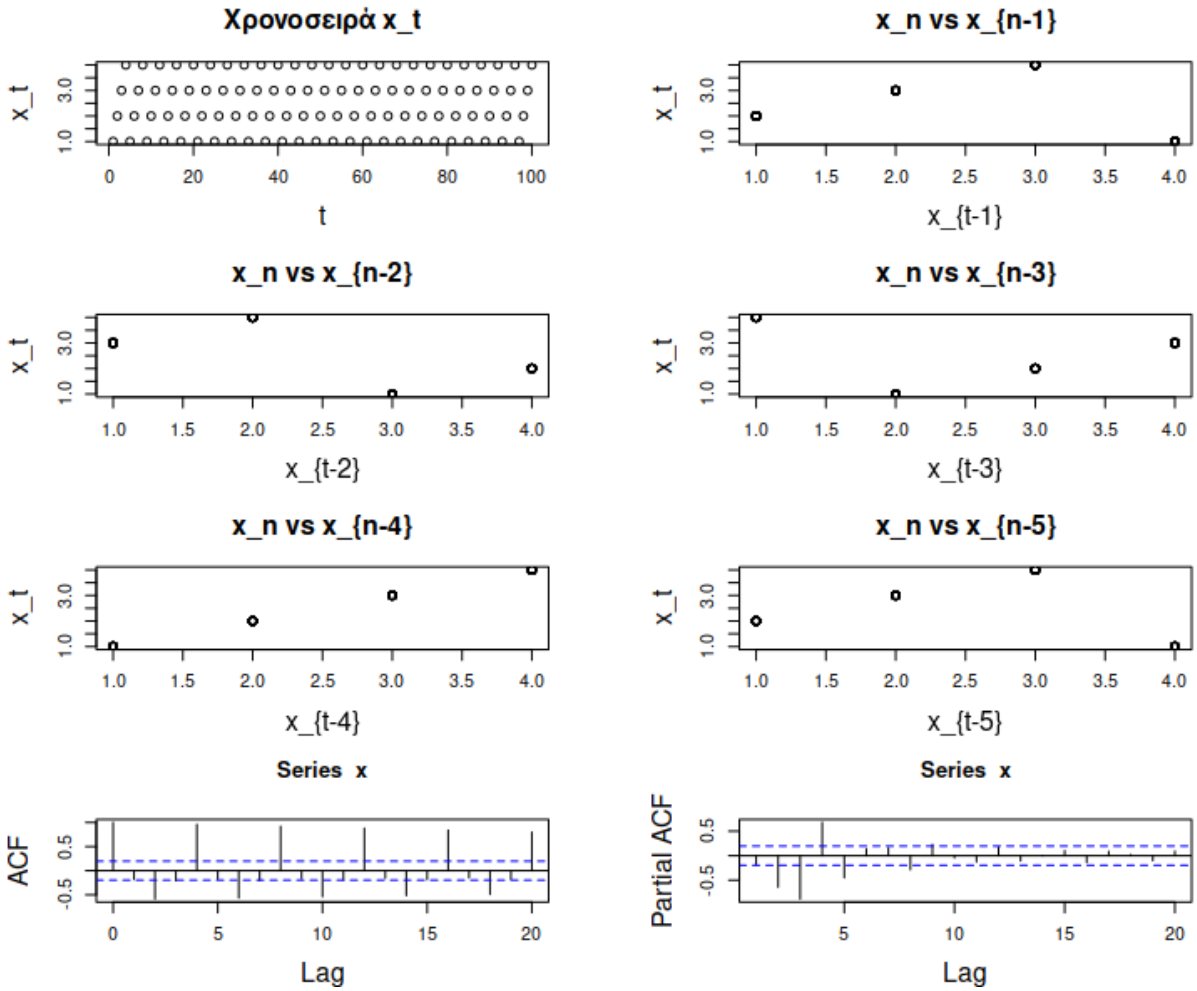
Ανάλογα με το παράδειγμα 1.

Παράδειγμα 3 (Κώδικας R + output)

Περιοδική Χρονοσειρά με τιμές 1, 2, 3, 4, περίοδο 4 και 100 τιμές

x = c(rep(1:4, 25))

my.plot.ts(x)



Σχόλια

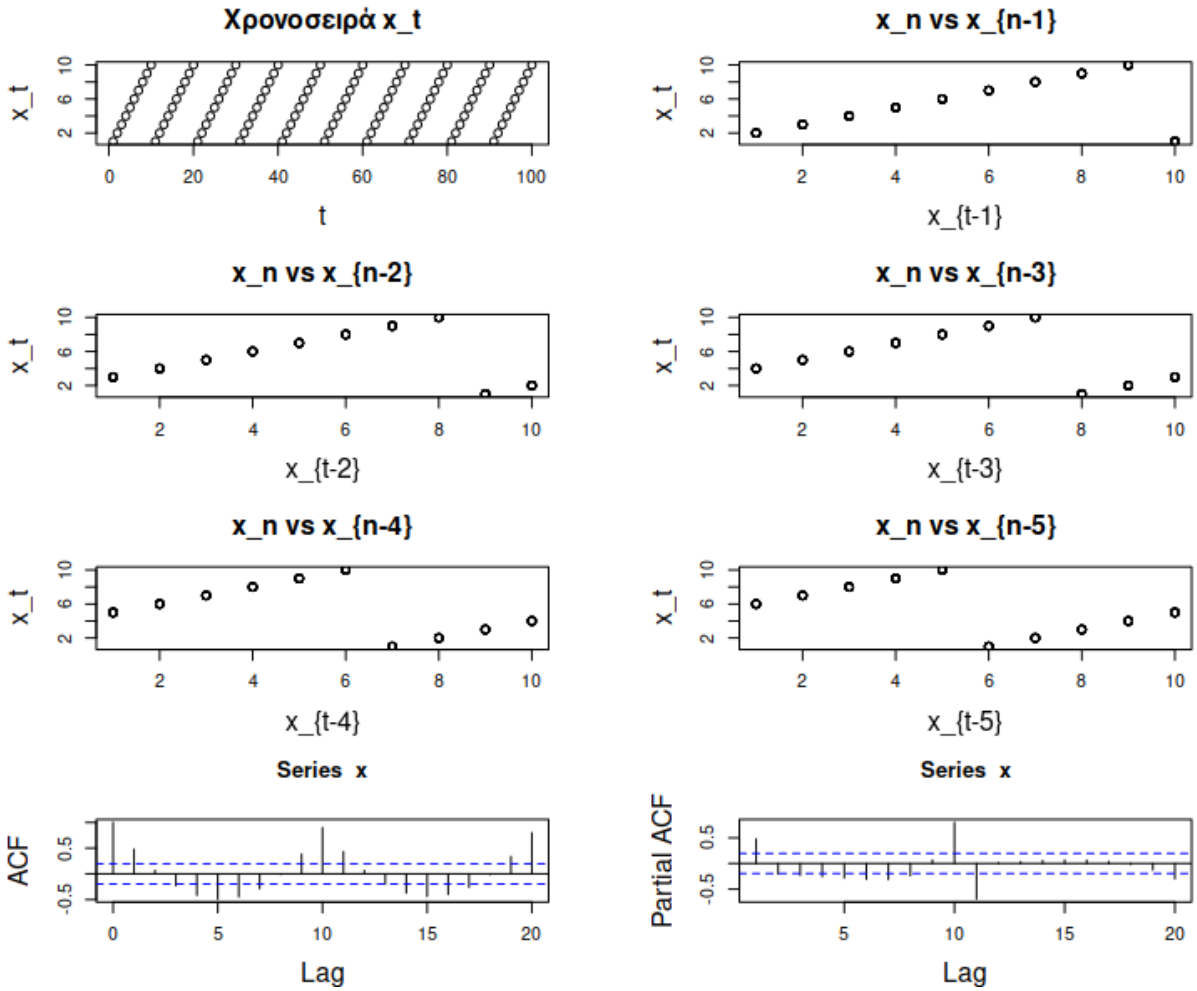
Η θετική γραμμική σχέση μεταξύ των μετρήσεων που απέχουν 4 θέσεις (όσες και η περίοδος) υποδεικνύει πως η χρονοσειρά είναι περιοδική με περίοδο 4.

Παράδειγμα 4 (Κώδικας R + output)

Περιοδική Χρονοσειρά με τιμές 1, 2, …, 10, περίοδο 10 και 100 τιμές

x = c(rep(1:10, 10))

my.plot.ts(x)



Σχόλια

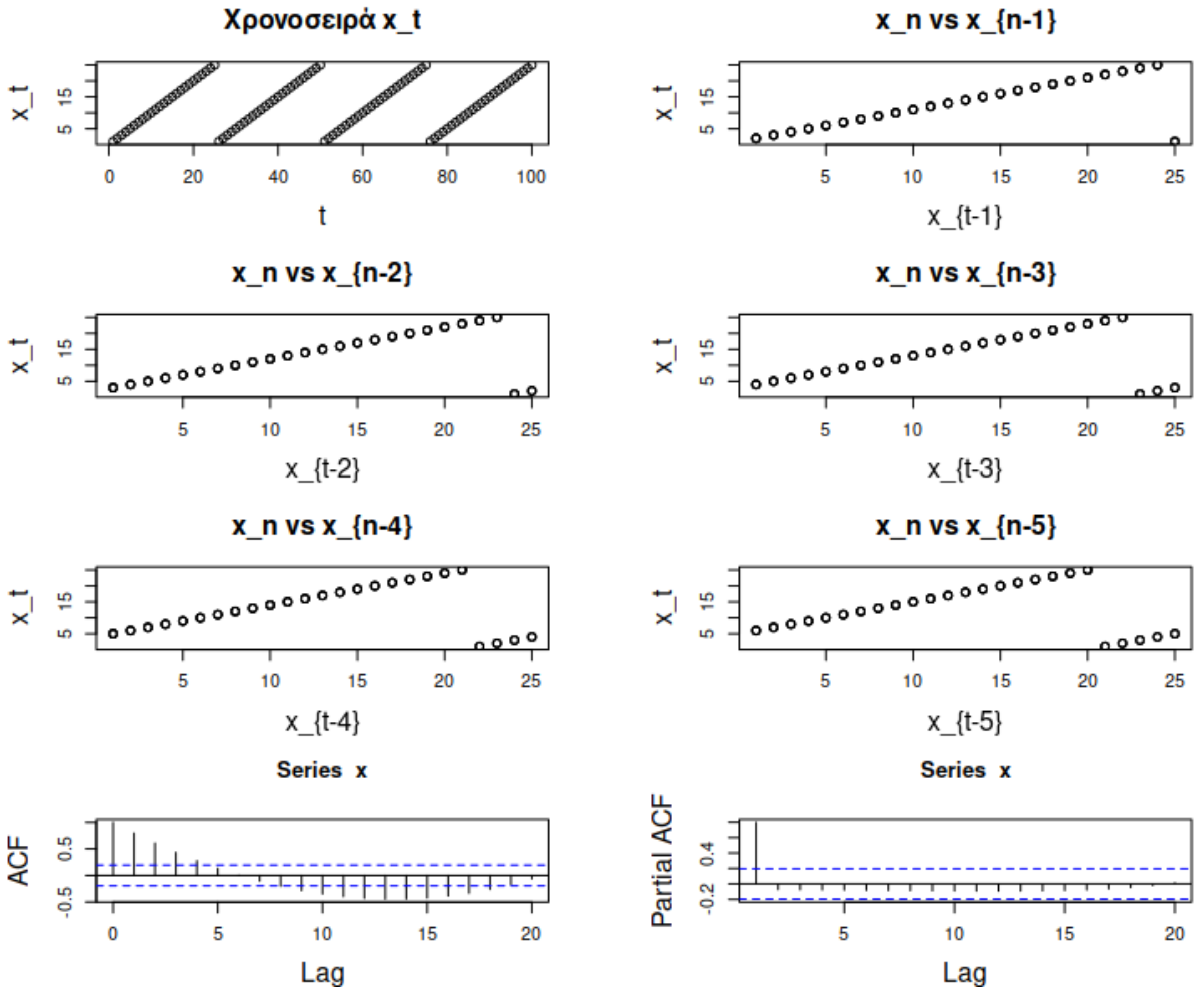
Η θετική γραμμική σχέση μεταξύ των μετρήσεων που απέχουν 10 θέσεις (όσες και η περίοδος) υποδεικνύει πως η χρονοσειρά είναι περιοδική με περίοδο 10.

Παράδειγμα 5 (Κώδικας R + output)

Περιοδική Χρονοσειρά με τιμές 1, 2, …, 25, περίοδο 25 και 100 τιμές

x = c(rep(1:25, 4))

my.plot.ts(x)



Σχόλια

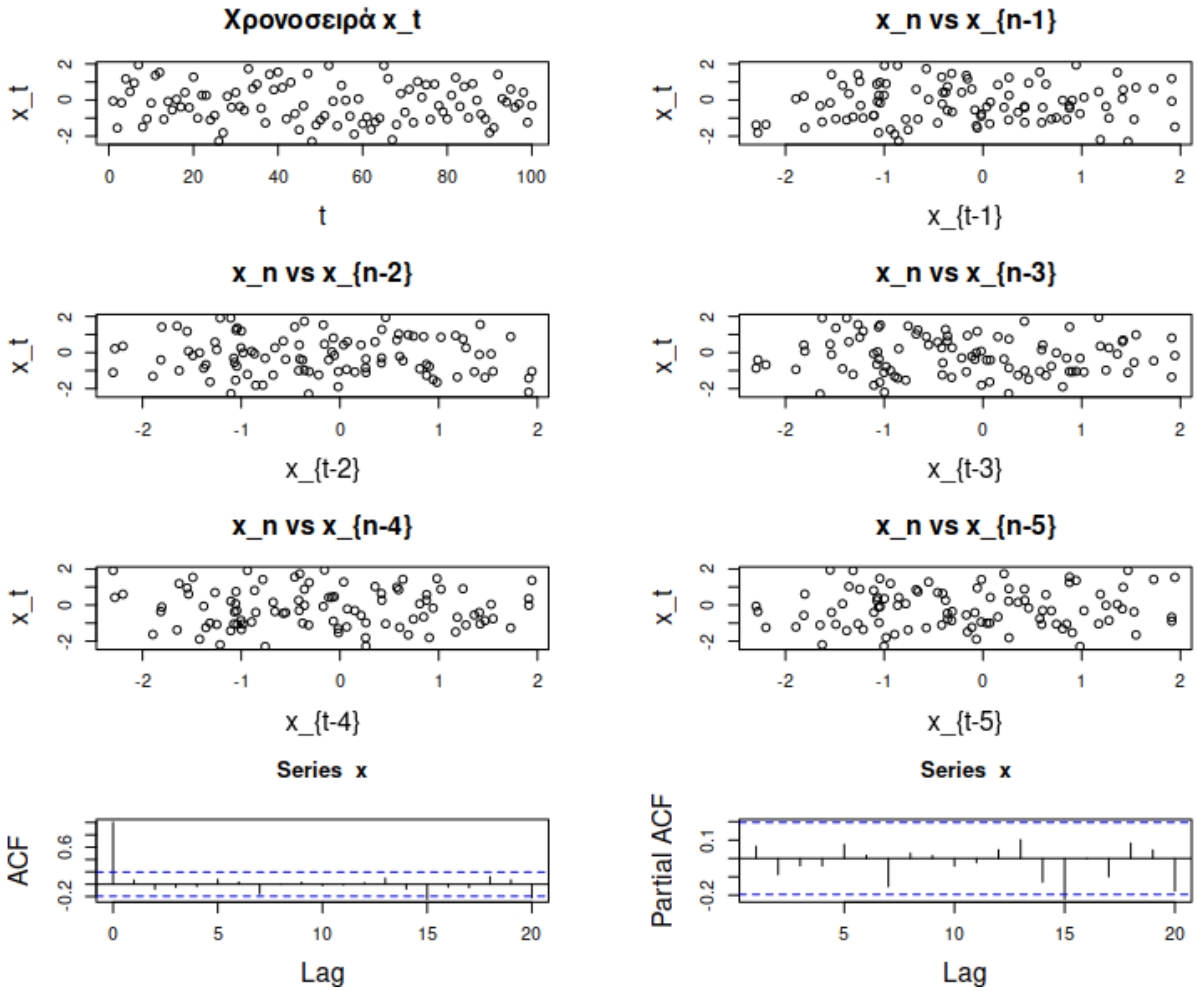
Η θετική γραμμική σχέση μεταξύ των μετρήσεων που απέχουν 25 θέσεις (όσες και η περίοδος) υποδεικνύει πως η χρονοσειρά είναι περιοδική με περίοδο 25.

Παράδειγμα 6 (Κώδικας R + output)

Χρονοσειρά 100 τυχαίων τιμών από την N(0, 1)

x = rnorm(100, 0, 1)

my.plot.ts(x)



Σχόλια

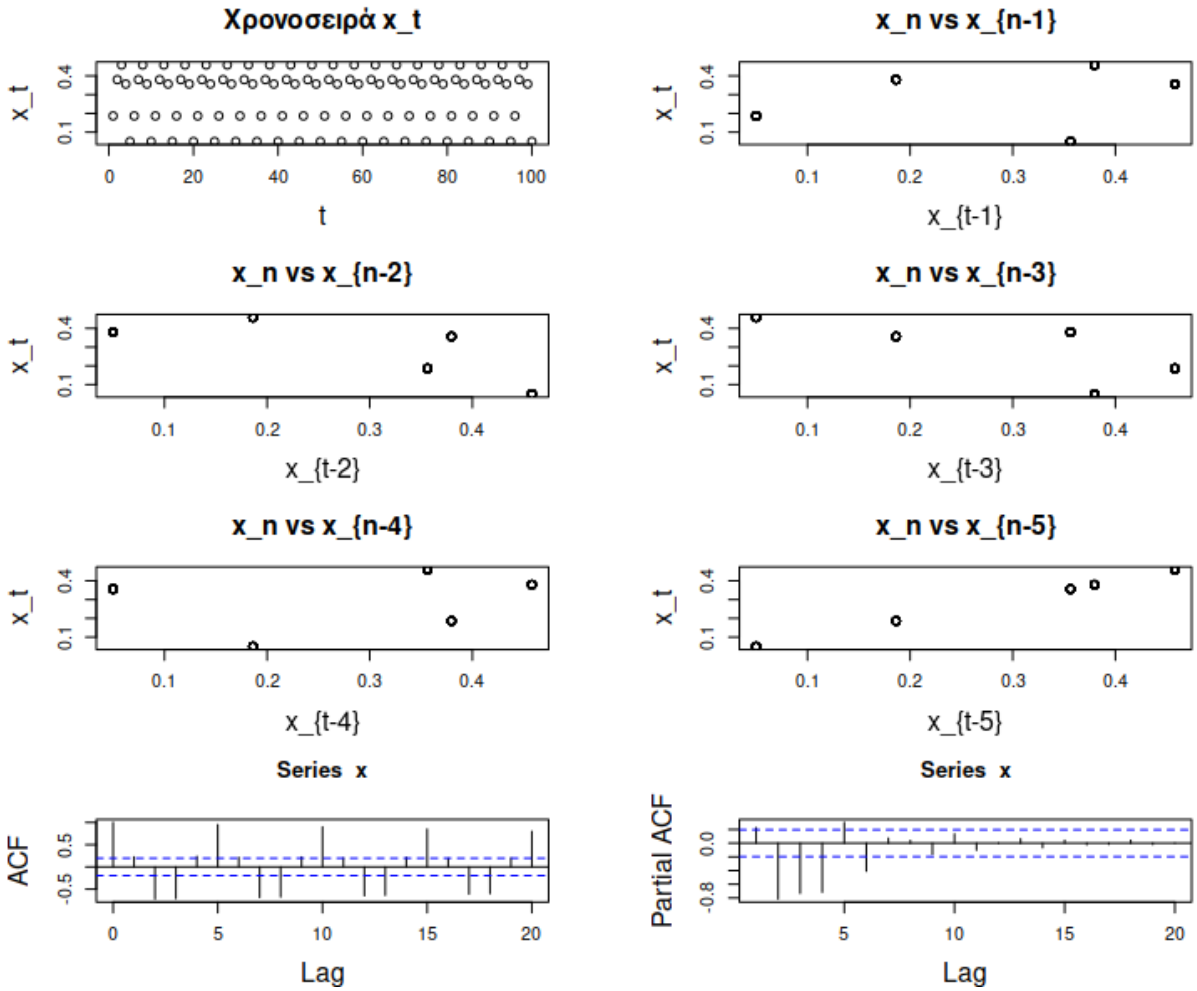
Οι τιμές εξελίσσονται τυχαία και ως εκ τούτου δεν υπάρχει σαφής θετική συσχέτιση μεταξύ της χρονοσειράς και των αντιγράφων της με οποιαδήποτε υστέρηση.

Παράδειγμα 7 (Κώδικας R + output)

Περιοδική Χρονοσειρά 5 τυχαίων τιμών από την U(0, 1), περίοδο 5 και 100 τιμές

x = rep(runif(5, 0, 1), 20)

my.plot.ts(x)

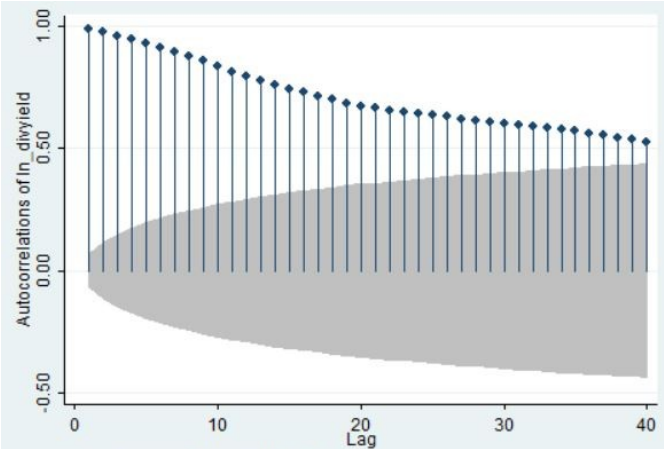


Σχόλια

Η θετική γραμμική σχέση μεταξύ των μετρήσεων που απέχουν 5 θέσεις (όσες και η περίοδος) υποδεικνύει πως η χρονοσειρά είναι περιοδική με περίοδο 5.

Δραστηριότητα

Σχολιάστε το παρακάτω διάγραμμα ACF μίας χρονοσειράς μηνιαίων τιμών.

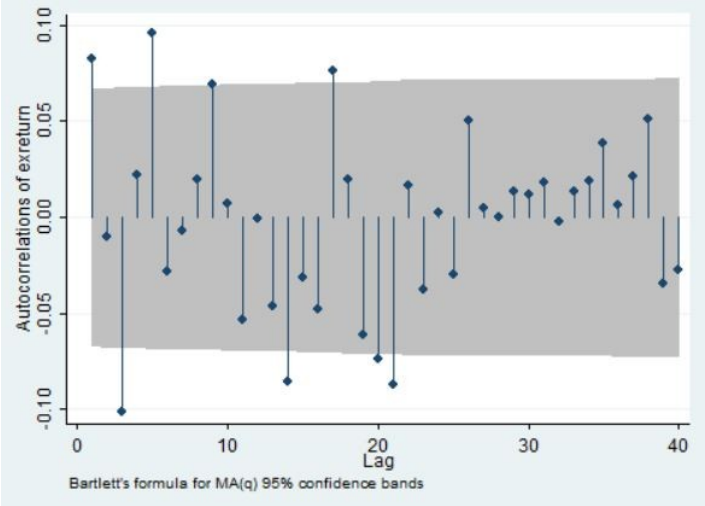


Σημείωση: Η γκρι περιοχή υποδηλώνει το 95% διάστημα εμπιστοσύνης για κάθε μία τιμή.

Απάντηση: Στο γράφημα, υπάρχουν υψηλές θετικές συσχετίσεις που μειώνονται με αργό ρυθμό, γεγονός που δηλώνει αυξημένη αυτοσυσχέτιση και άρα υπάρχει κάποιου είδους εξάρτηση (γραμμική ή μη γραμμική) μεταξύ των τιμών της χρονοσειράς. Ειδικότερα, οπωσδήποτε δεν πρόκειται για μία στάσιμη χρονοσειρά.

Δραστηριότητα

Σχολιάστε το παρακάτω διάγραμμα ACF μίας χρονοσειράς μηνιαίων τιμών.



Απάντηση: Οι συσχετίσεις είναι πολύ χαμηλές (μεταξύ -0,1 και 0,1) και δεν φαίνεται να έχουν κάποιο μοτίβο, ενώ οι περισσότερες είναι στατιστικά μη διαφορετικές από το 0. Συμπεραίνουμε, ότι κατά πάσα πιθανότητα, αυτή η χρονοσειρά προέρχεται από μία στάσιμη διαδικασία.

Πηγή: https://stats.stackexchange.com/questions/182627/interpretation-of-correlogram

15.3 Λευκός Θόρυβος (White Noise)

Ορισμός

Μία διεργασία {U(n), n ϵ N} για την οποία ισχύει:

(α) E(U(n)) = 0.

(β) Var(U(n)) = $\sigma^2 < +\infty$.

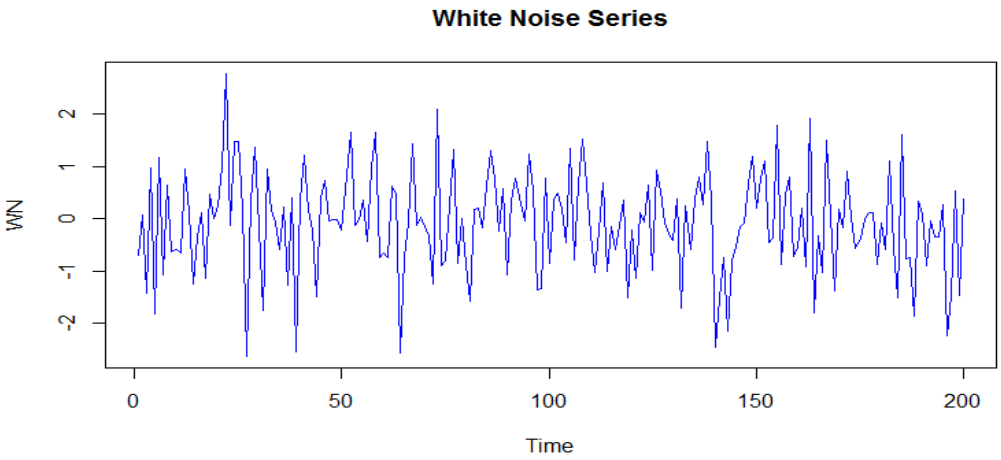
(γ) Cov(U(n), U(n – k)) = 0, για κάθε n, n – k ≥0.

ονομάζεται **Λευκός Θόρυβος (White Noise)** με μέση τιμή 0 και διακύμανση σ^2.

Γράφουμε

$$\mathbf{U(n) \sim WN(0, \sigma^2).}$$

Φανερά, ο Λευκός Θόρυβος είναι παράδειγμα ασθενώς στάσιμης διεργασίας.



Αναπαράσταση 200 σημεία διεργασίας Λευκού Θορύβου X(n) ~ WN(0, 1).

Παρατηρούμε ότι στην U(n) ~ WN(0, σ^2):

- Δεν υπάρχει σαφής τάση.
- Δεν υπάρχει περιοδικότητα.
- Δεν υπάρχουν εμφανείς ακραίες τιμές.
- Δεν υπάρχει σαφής μεταβολή στη διακύμανση των τιμών καθώς προχωράμε στο χρόνο.

Παρατηρήσεις

1. Καθώς Cov(X, Y) = E(XY) – E(X)E(Y) και E(U(t)) = 0, η συνθήκη

$$(\gamma)\ Cov(U(n), U(n-k)) = 0,\ n,\ n-k \geq 0$$

μπορεί ισοδύναμα να εκφραστεί ως

(γ)’ $E(U(t)\cdot U(h)) = 0$, για κάθε $t \neq h$.

2. Αν επιπλέον, οι U(t) είναι ανεξάρτητες και ισόνομες με μέση τιμή 0 και διακύμανση σ^2, γράφουμε ισοδύναμα **$U(n) \sim IWN(0, \sigma^2)$**.

3. Κάθε διεργασία που αποτελείται από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές αποτελεί Λευκό Θόρυβο μόνο στην περίπτωση όπου η τάση της είναι σταθερή και ίση με 0.

4. Αν {U(n), n ϵ N} είναι διεργασία Λευκού Θορύβου, τότε οι U(n) δεν είναι απαραίτητα ανεξάρτητες μεταξύ τους.

Παράδειγμα

Έστω {W(n)}, $n \geq 0$, τυχαίες μεταβλητές, ανεξάρτητες μεταξύ τους που ακολουθούν την κατανομή Rademacher, δηλαδή λαμβάνουν μόνο τις τιμές -1 και 1 με πιθανότητα 50% την κάθε μία.

Έστω επίσης {U(n), n ϵ N} στοχαστική διεργασία τέτοια ώστε $U(0) \sim N(0, 1)$, U(0) ανεξάρτητη με τις {W(n)}, $n \geq 0$, και $U(n) = U(n-1)W(n)$, δηλαδή

$$U(n) = U(0)W(1)W(2) \cdots W(n).$$

Τότε, οι {U(n), n ϵ N} αποτελείται εξ΄ ορισμού από εξαρτημένες τυχαίες μεταβλητές. Ωστόσο είναι μία διεργασία Λευκού Θορύβου καθώς:

(α) $P(U(n) \leq x) = E(P(U(0) \leq x \mid W(1)W(2) \cdots W(n))$

$$= P(U(0) \leq x)\, P(W(1)W(2) \cdots W(n) = 1)$$

$$+ P(-U(0) \leq x)\, P(W(1)W(2) \cdots W(n) = -1)$$

$$= \tfrac{1}{2}\, \Phi(x) + \tfrac{1}{2}\, \Phi(x)$$

$$= \Phi(x), \qquad (W(1)W(2) \cdots W(n) \sim \text{Rademacher, για κάθε } n \geq 0)$$

δηλαδή σε κάθε περίπτωση $U(t) \sim N(0, 1)$, ή ειδικότερα $E(U(t)) = 0$ και $Var(U(t)) = 1$.

(β) Έστω $t < h$, $m = \min(t, h)$, $M = \max(t, h)$

$$Cov(U(t), U(h)) = E(U(t) \cdot U(h))$$

$$= E(U^2(0)W(1)W(2) \cdots W(t)W(1)W(2) \cdots W(h))$$

$$= E(U^2(0))\, E(W(1)W(2) \cdots W(t)W(1)W(2) \cdots W(h))$$

(U(0) ανεξάρτητη με τις W(t), άρα $U^2(0)$ ανεξάρτητη με τις W(t))

$$= E(U^2(0))\, E(W^2(1))E(W^2(2)) \cdots E(W^2(m))E(W(m+1)) \cdots E(W(M))$$

$$= E(U^2(0)) \cdot 1 \cdot 1 \cdot \ldots \cdot 0 \cdot \ldots \cdot 0 = 0$$

Άρα, η {U(n), n ϵ N} είναι διεργασία Λευκού Θορύβου που αποτελείται από εξαρτημένες τυχαίες μεταβλητές.

5. Αν {U(n), n ϵ N} είναι διεργασία Λευκού Θορύβου, τότε η διεργασία δεν είναι απαραίτητα ισχυρά στάσιμη.

Παράδειγμα

Έστω W ~ U(0, 2π) και η διεργασία {U(n), n ϵ N*} με U(n) = cos(n·W), n = 1, 2, …

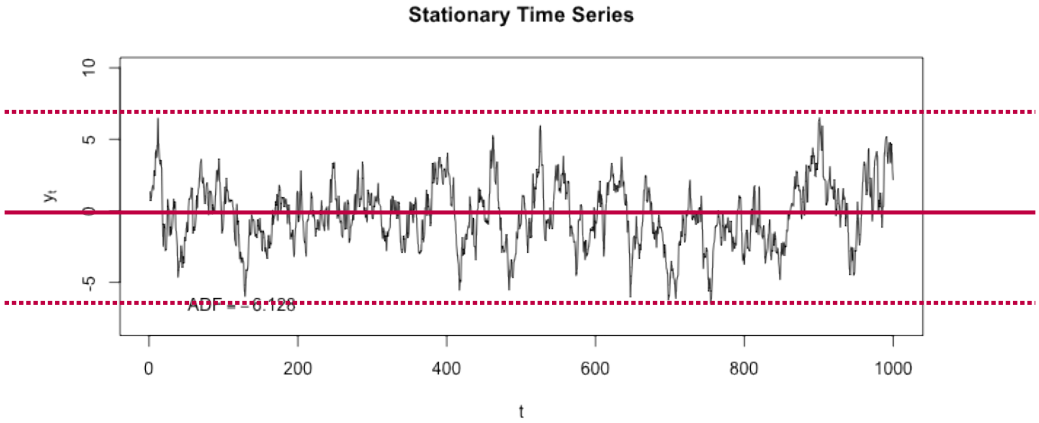
Τότε, είναι E(U(n)) = 0, Var(U(n)) = ½ και Cov(U(n), U(n – k)) = 0 για κάθε n = 1, 2, .. και n – k = 1, 2, …,

Δηλαδή U(n) ~ WN(0, ½).

Ωστόσο, μπορεί να αποδειχθεί ότι η διεργασία {U(n), n ϵ N*} δεν είναι ισχυρά στάσιμη.

Παράδειγμα χρονοσειράς που προέρχεται από στάσιμη στοχαστική διεργασία

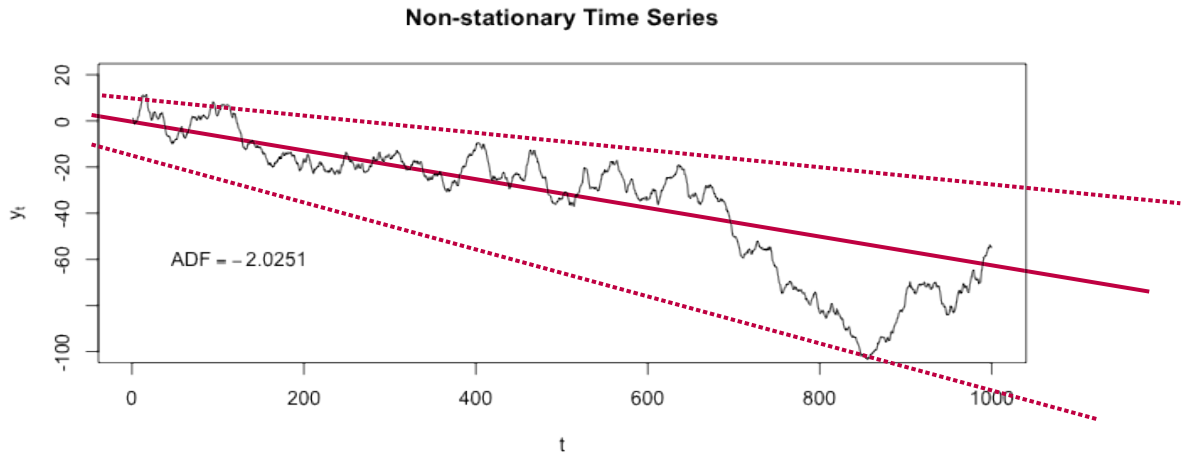
Η χρονοσειρά που απεικονίζεται είναι μία *εκδοχή των τιμών* που μπορεί να πάρει μία στάσιμη στοχαστική διεργασία



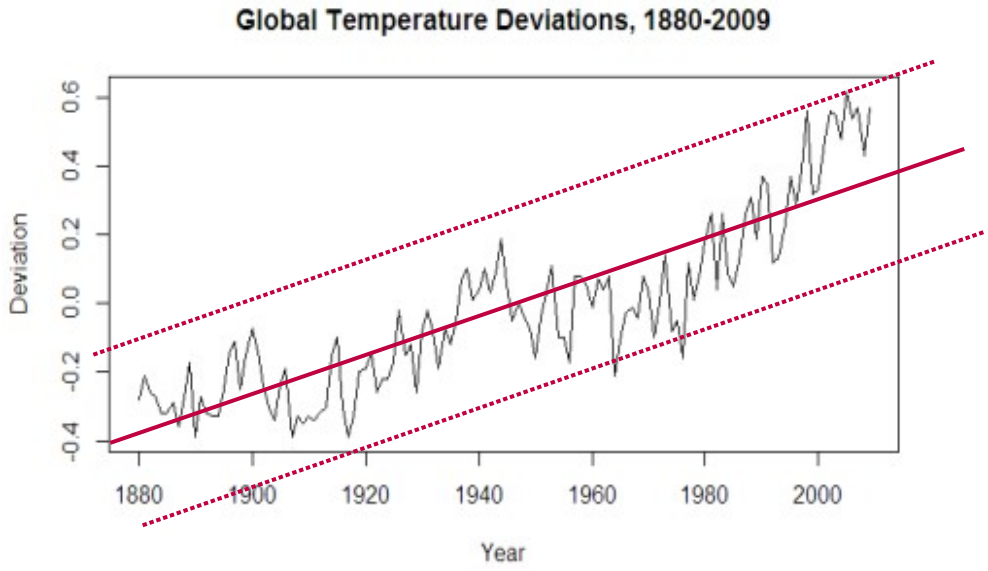
Σταθερή διακύμανση (στην πράξη) σημαίνει πως όλες οι εκδοχές της στάσιμης στοχαστικής διεργασίας αναμένονται να είναι μέσα σε ένα διάστημα που δεν μεταβάλλεται καθώς το t αυξάνει.

Παράδειγμα χρονοσειράς που προέρχεται από μη στάσιμη στοχαστική διεργασία

Η χρονοσειρά που απεικονίζεται είναι μία *εκδοχή των τιμών* που μπορεί να πάρει μία μη στάσιμη στοχαστική διεργασία



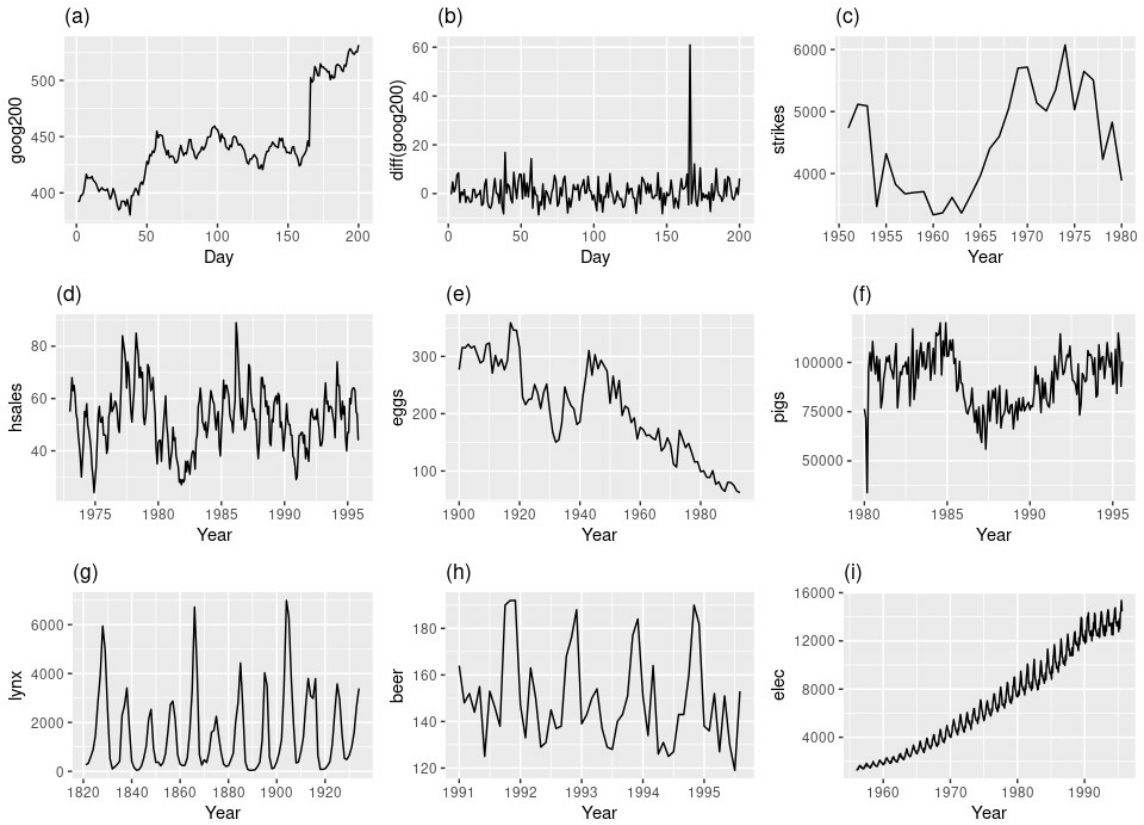
Η μέση τιμή και το εύρος των τιμών διαφοροποιείται με το χρόνο, γεγονός που υποδεικνύει μη στάσιμη διεργασία.



Η μέση τιμή διαφοροποιείται με το χρόνο γεγονός που δεν μπορεί να αντιστοιχεί σε στάσιμη διεργασία.

Δραστηριότητα

Ποιες από τις χρονοσειρές του διαγράμματος είναι πιθανό να προέρχονται από στάσιμες διεργασίες (ασθενώς ή ισχυρά);



(a) Τιμή μετοχής της Google για 200 συνεχόμενες ημέρες.

(b) Ημερήσια μεταβολή στην τιμή της μετοχής της Google για 200 συνεχόμενες ημέρες.

(c) Ετήσιος αριθμός απεργιών στις ΗΠΑ.

(d) Μηνιαίες πωλήσεις νέων μονοκατοικιών που πωλούνται στις ΗΠΑ.

(e) Ετήσια τιμή δωδεκάδας αυγών στις ΗΠΑ (δολάρια).

(f) Μηνιαίο πλήθος χοίρων που καταναλώνονται στη Βικτώρια της Αυστραλίας.

(g) Πληθυσμός λύγκα στην περιοχή του ποταμού McKenzie του βορειοδυτικού Καναδά.

(h) Μηνιαία παραγωγή μπύρας στην Αυστραλία.

(i) Μηνιαία παραγωγή ηλεκτρικής ενέργειας στην Αυστραλία.

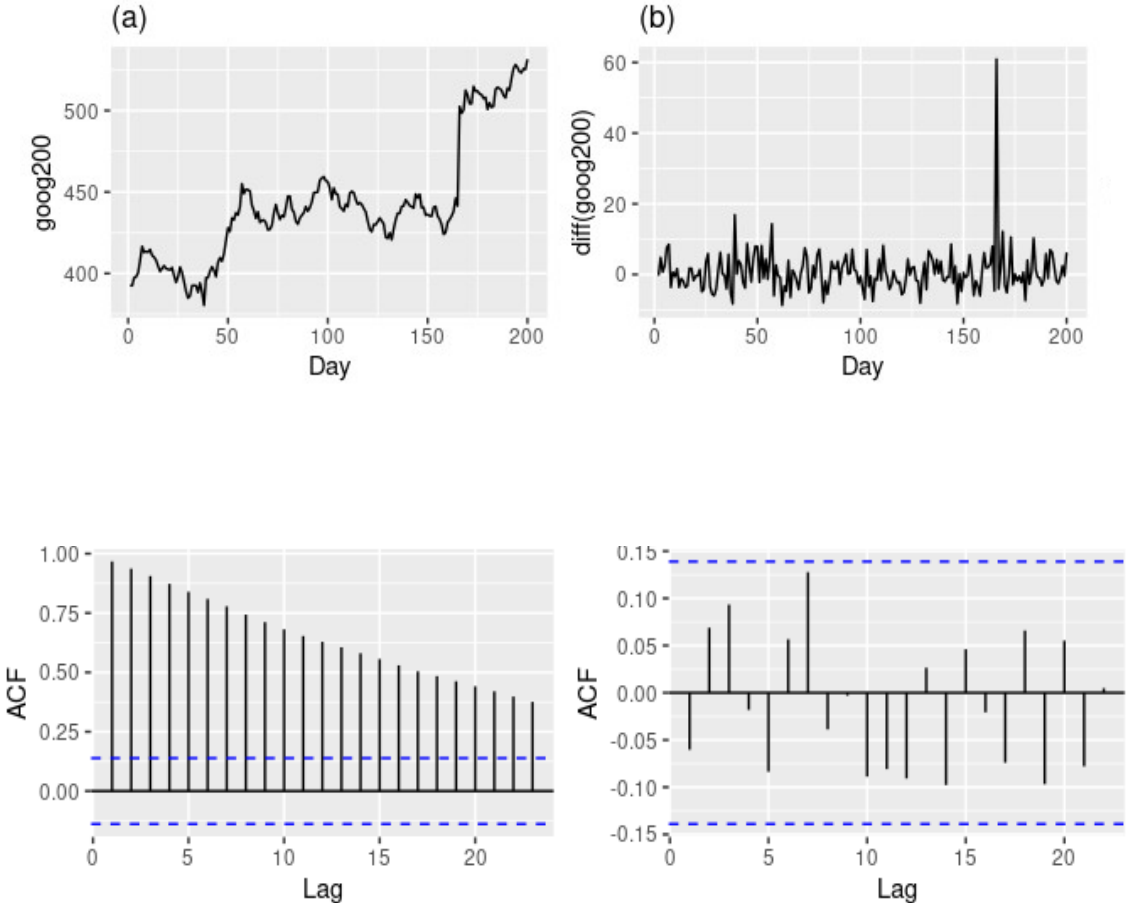
Απάντηση

Η φανερή εποχικότητα αποκλείει τις σειρές (d), (h) και (i). Οι τάσεις και τα μεταβαλλόμενα επίπεδα αποκλείουν τις σειρές (a), (c), (e), (f) και (i). Η αυξανόμενη διακύμανση αποκλείει επίσης το (i). Αυτό αφήνει μόνο τα (b) και (g) ως σταθερές σειρές.

Με την πρώτη ματιά, οι ισχυροί κύκλοι στη σειρά (g) μπορεί να φαίνεται ότι το καθιστούν μη στάσιμο. Αλλά αυτοί οι κύκλοι είναι απεριοδικοί — προκαλούνται όταν ο πληθυσμός του λύγκα γίνεται πολύ μεγάλος για τη διαθέσιμη τροφή, έτσι ώστε να σταματήσει η αναπαραγωγή και ο πληθυσμός πέφτει σε χαμηλούς αριθμούς, τότε η αναγέννηση των πηγών τροφής του επιτρέπει στον πληθυσμό να αυξηθεί ξανά. Μακροπρόθεσμα, ο χρονισμός αυτών των κύκλων δεν είναι προβλέψιμος. Ως εκ τούτου, η σειρά είναι στάσιμη.

15.4 Το διάγραμμα των τιμών αυτοσυσχέτισης (correlogram) αντανακλά τη στασιμότητα ή μη μίας χρονοσειράς.

Για μια στάσιμη χρονοσειρά, η τιμή της αυτοσυσχέτισης (ACF) θα μειωθεί σχετικά γρήγορα, ενώ η ACF μίας μη στάσιμης χρονοσειράς μειώνεται με αργό ρυθμό. Επίσης, για μία μη στάσιμη χρονοσειρά, η τιμή του r_1 είναι συχνά μεγάλη και θετική.



Διάγραμμα: Οι χρονοσειρές της μετοχής της Google και οι ημερήσιες διαφορές, μαζί με τα αντίστοιχα συσχετογράμματα.

15.5 Αυτοπαλινδρομικό μοντέλο 1ης τάξης (Μοντέλο AR(1))

Ένα από τα απλούστερα μοντέλα είναι το αυτοπαλινδρομικό μοντέλο τάξης 1, AR(1) (Autoregressive model of order 1) με το οποίο συσχετίζεται κάθε μία τιμή της σειράς $\{X_t, t \in N^*\}$ με την αμέσως προηγούμενη της. Η αναλυτική του έκφραση είναι

$$\mathbf{X_t = \delta + \varphi_1 \cdot X_{t-1} + w_t,}$$

όπου w_t είναι τα ανεξάρτητα και ισόνομα σφάλματα των προβλέψεων με κανονική κατανομή $N(0, \sigma^2_w)$, (δηλαδή white noise) τα οποία επιπλέον, δεν εξαρτώνται από την αντίστοιχη τιμή της σειράς x_t.

Μοντέλο AR(1) για μία ασθενώς στάσιμη χρονοσειρά

Υποθέτουμε ότι η χρονοσειρά $\{X_t, t \in N^*\}$ αποτελεί μία ασθενώς στάσιμη στοχαστική διεργασία, δηλαδή ότι, για κάθε $t \in N^*$:

- $E(X_t) = \mu$ = σταθερό,
- $Var(X_t) = \sigma^2$ = σταθερό,
- Η συνδιακύμανση $Cov(X_t, X_{t-h})$ εξαρτάται μόνο από το h = 1, 2, …και όχι από το t.

Στην περίπτωση αυτή, η συνάρτηση αυτοσυσχέτισης για τη χρονοσειρά γράφεται

$$\rho_h = Corr(X_t, X_{t-h}) = Cov(X_t, X_{t-h}) / [SD(X_t)\, SD(X_{t-h})]$$

$$= Cov(X_t, X_{t-h}) / Var(X_t), \qquad (SD(X_t) = SD(X_{t-h}) = \sigma)$$

δηλαδή, μία ασθενώς στάσιμη στοχαστική διεργασία αντιστοιχεί σε χρονοσειρά που για δεδομένη υστέρηση h, έχει σταθερή αυτοσυσχέτιση σε όλο το εύρος του χρονικού διαστήματος που περιγράφει.

Η υπόθεση της ασθενούς στασιμότητας δεν επαληθεύεται σε πολλά φυσικά συστήματα. Ωστόσο, όταν αυτή μπορεί να γίνει αποδεκτή, οδηγεί σε μία απλή μορφή του μοντέλου AR(1).

Αναμενόμενη τιμή ασθενούς στάσιμης χρονοσειράς AR(1)

Πράγματι, από την εξίσωση $X_t = \delta + \varphi_1 \cdot X_{t-1} + w_t$, παίρνουμε

$$E(X_t) = E(\delta + \varphi_1 \cdot X_{t-1} + w_t)$$

$$= \delta + E(\varphi_1 \cdot X_{t-1}) + E(w_t)$$

$$= \delta + \varphi_1 \cdot E(X_{t-1}) \qquad (E(w_t) = 0)$$

Όμως, $E(X_t) = E(X_{t-1}) = \mu$, άρα

$$\mu = \delta + \varphi_1 \cdot \mu$$

ή

$$\mathbf{E(X_t) = \mu = \delta / (1 - \varphi_1)}$$

Διακύμανση ασθενούς στάσιμης χρονοσειράς AR(1)

Περαιτέρω, από την εξίσωση $X_t = \delta + \varphi_1 \cdot X_{t-1} + w_t$, παίρνουμε

$Var(X_t) = Var(\delta + \varphi_1 \cdot X_{t-1} + w_t)$

$= Var(\varphi_1 \cdot X_{t-1}) + Var(w_t)$ (τα σφάλματα w_t είναι ανεξάρτητα από τις τιμές X_{t-1})

$= \varphi^2_1 \cdot Var(X_{t-1}) + \sigma^2_w$

Όμως, $Var(X_t) = Var(X_{t-1})$, άρα

$$Var(X_t) = \varphi^2_1 \cdot Var(X_t) + \sigma^2_w$$

ή

$$\mathbf{Var(X_t) = \sigma^2_w / (1 - \varphi^2_1)}$$

Αυτοσυσχέτιση (υστέρησης h) ασθενούς στάσιμης χρονοσειράς AR(1)

Υποθέτουμε ότι $E(X_t) = 0$, κάτι που συμβαίνει όταν $\delta = 0$ (αν $\delta \neq 0$, τότε με κατάλληλη αντικατάσταση $Y_t = \alpha + X_t$, η εξίσωση μπορεί να πάρει την μορφή $Y_t = \varphi_1 Y_{t-1} + w_t$)

Τότε,

$$Cov(X_t, X_{t-h}) = E(X_t, X_{t-h}) \text{ (γιατί;)}$$

ενώ η εξίσωση γράφεται $X_t = \varphi_1 \cdot X_{t-1} + w_t$. Από την τελευταία παίρνουμε:

$$X_{t-h} \cdot X_t = \varphi_1 \cdot X_{t-h} \cdot X_{t-1} + X_{t-h} \cdot w_t \text{ ή}$$

$E(X_{t-h} \cdot X_t) = E(\varphi_1 \cdot X_{t-h} \cdot X_{t-1}) + E(X_{t-h} \cdot w_t)$

$= \varphi_1 \cdot E(X_{t-h} \cdot X_{t-1})$

($E(X_{t-h} \cdot w_t) = 0$, γιατί τα σφάλματα w_t είναι ανεξάρτητα από τις X_{t-h} και $E(X_{t-h}) = 0$)

Αν $\gamma_h = E(X_{t-h} \cdot X_t)$ η παραπάνω σχέση γράφεται

$$\gamma_h = \varphi_1 \cdot \gamma_{h-1}$$

Παρατηρούμε ότι $\gamma_1 = E(X_{t-1} \cdot X_t)$

$$= E(X_{t-1} \cdot (\varphi_1 \cdot X_{t-1} + w_t))$$

$$= E(\varphi_1 \cdot X^2_{t-1} + X_{t-1} \cdot w_t)$$

$$= \varphi_1 \cdot E(X^2_{t-1}) = \varphi_1 \cdot Var(X_{t-1})$$

$$= \varphi_1 \cdot Var(X_t).$$

Από τις σχέσεις $\gamma_h = \varphi_1 \cdot \gamma_{h-1}$ και $\gamma_1 = \varphi_1 \cdot Var(X_t)$ με αναδρομικό συλλογισμό βρίσκουμε ότι

$$\mathbf{E(X_{t-h} \cdot X_t) = \gamma_h = \varphi^h_1 \cdot Var(X_t)}$$

Τώρα,

$\rho_h = Cov(X_t, X_{t-h}) / Var(X_t)$

$$= E(X_t, X_{t-h}) / Var(X_t)$$

$$= \varphi^h_1 \cdot Var(X_t) / Var(X_t)$$

$$= \varphi^h_1$$

Βρήκαμε ότι,

$$\boldsymbol{\rho_h = \varphi^h_1}$$

Το ίδιο αποδεικνύεται και στην περίπτωση όπου $E(X_t) \neq 0$.

Σύνοψη

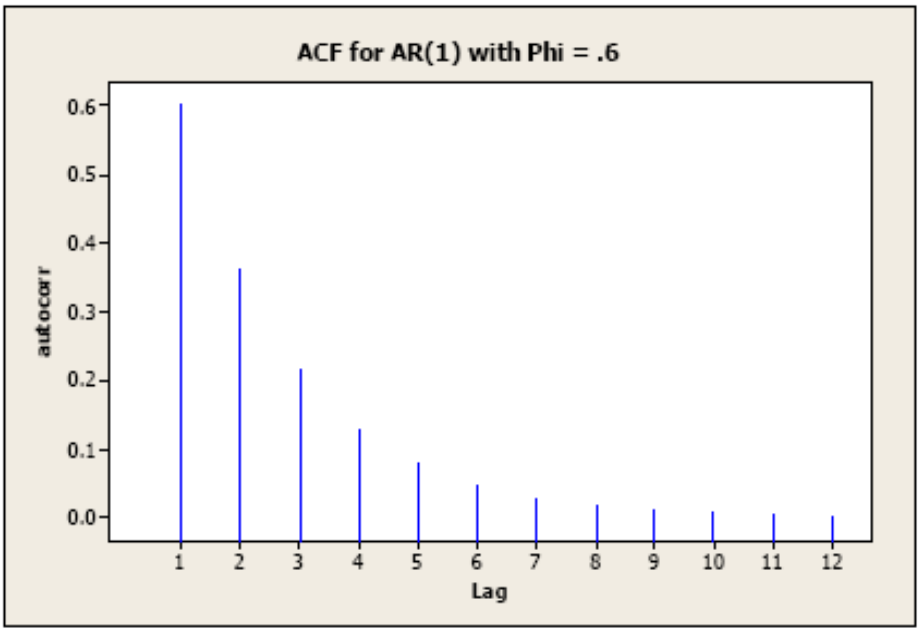
Σε κάθε ασθενώς στάσιμο AR(1) μοντέλο χρονοσειράς $X_t = \delta + \varphi_1 \cdot X_{t-1} + w_t$, ισχύει ότι:

- $E(X_t) = \mu = \delta / (1 - \varphi_1)$.
- $Var(X_t) = \sigma^2_w / (1 - \varphi^2_1)$.
- $Cov(X_t, X_{t-h}) = \rho_h = \varphi^h_1$.

Από την ισότητα $\rho_h = \varphi^h_1$ προκύπτει ότι $\rho_2 = \rho^2_1$, $\rho_3 = \rho^3_1$, ..., σχέσεις που προσφέρουν μία απλή μέθοδο αξιολόγησης μίας χρονοσειράς ως προς τη δυνατότητα προσαρμογής ενός AR(1) μοντέλου σε αυτήν.

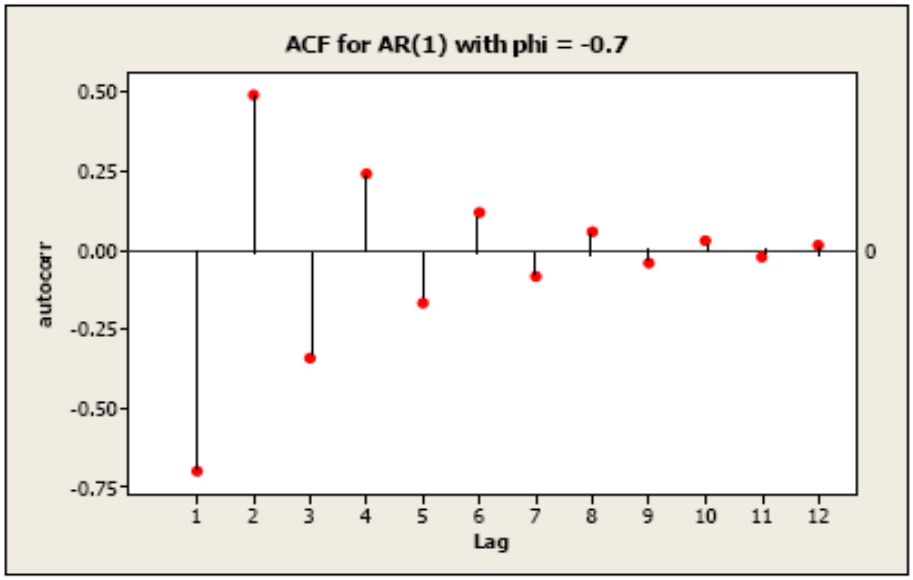
$$\rho_h = \varphi^h_1, \text{ για } \varphi_1 = 0{,}6$$

Ενδεικτικά, ένα διάγραμμα της συνάρτησης αυτοσυσχέτισης για $\varphi_1 = 0{,}6$, θα μοιάζει ως εξής:



$$\rho_h = \varphi^h_1, \text{ για } \varphi_1 = -0{,}7$$

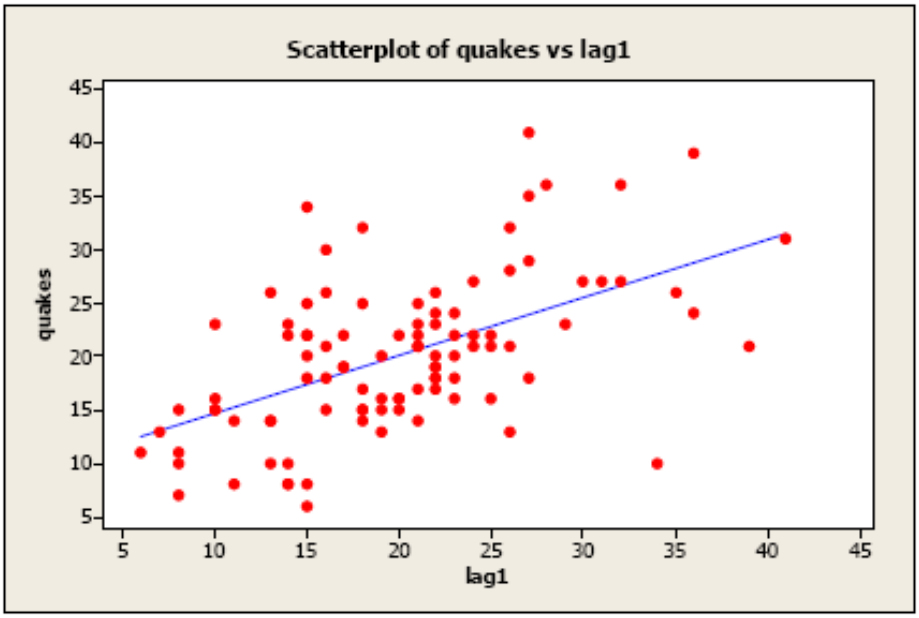
Ενδεικτικά, ένα διάγραμμα της συνάρτησης αυτοσυσχέτισης για $\varphi_1 = -0{,}7$, θα μοιάζει ως εξής:

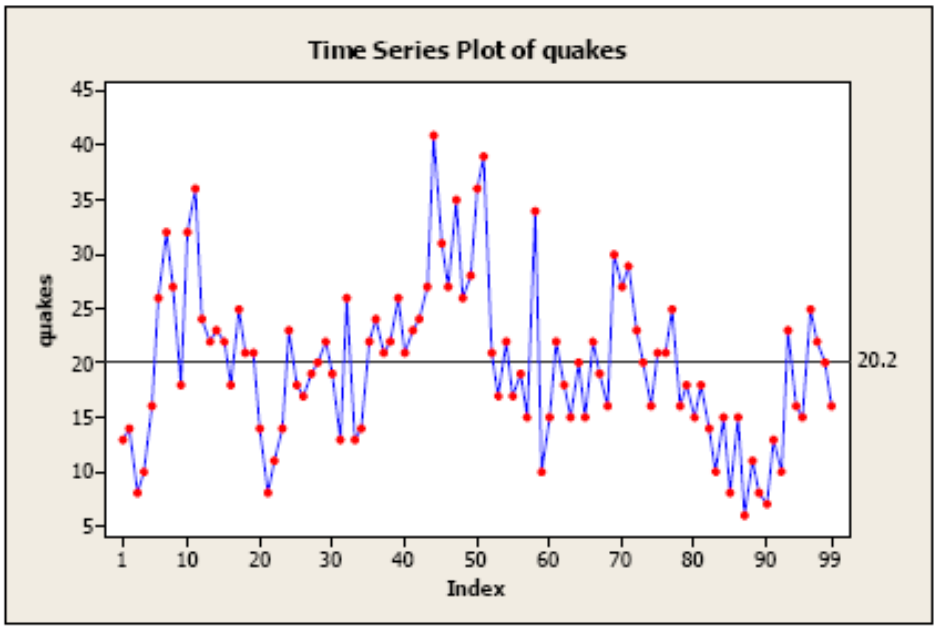


Παράδειγμα 1

Στην περίπτωση των σεισμών, το μοντέλο AR(1) είναι

$$\textbf{Πλήθος σεισμών } x_t = 9{,}19 + 0{,}543 \cdot x_{t-1},$$





Ο συντελεστής προσδιορισμού είναι

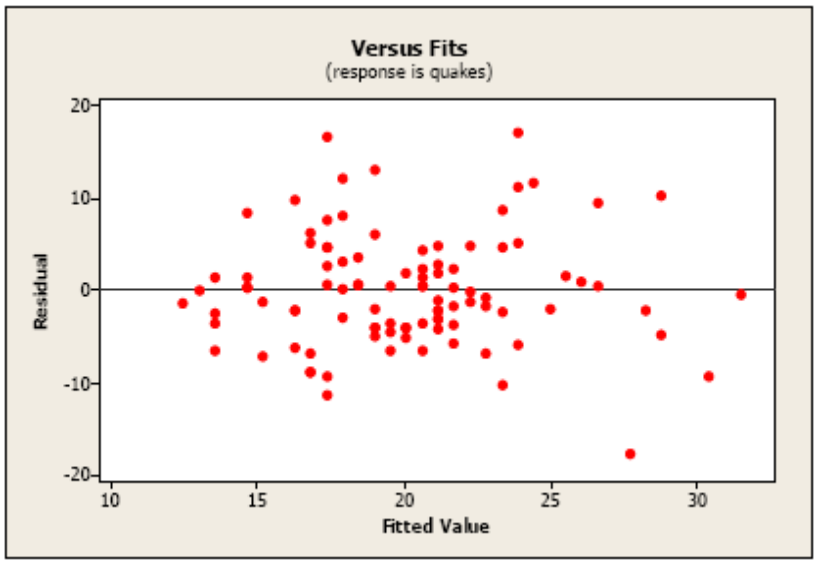
$$R^2 = 0{,}297$$

γεγονός που σημαίνει πως κάθε παρατήρηση προσδιορίζει τη μεταβλητότητα της επόμενης σε ποσοστό 29,7%, όχι ιδιαίτερα μεγάλο για αξιόπιστη πρόβλεψη.

Από το διάγραμμα των σφαλμάτων εκτίμησης (residuals) παρατηρούμε ότι τα σημεία:

- Κατανέμονται τυχαία πάνω και κάτω από γραμμή y = 0
- Βρίσκονται μέσα στα πλαίσια μίας ζώνης, χωρίς ακραίες παρατηρήσεις με πιθανή εξαίρεση μία ιδιάζουσα τιμή στο 28

Συνάγεται ότι η γραμμική εξίσωση δεν αποκλίνει σημαντικά από την αναμενόμενη μορφή.



Υστέρηση (lag)	r_h
1	0,542
2	0,419
3	0,398
4	0,324
5	0,237
6	0,172
7	0,190
8	0,061
9	-0,049
10	-0,107
11	-0,043
12	-0,072

Σε ένα μοντέλο AR(1), ισχύει

$$\rho_h = \varphi^h{}_1,$$

άρα θεωρητικά θα έπρεπε να είναι $\rho_2 = \rho^2{}_1$.

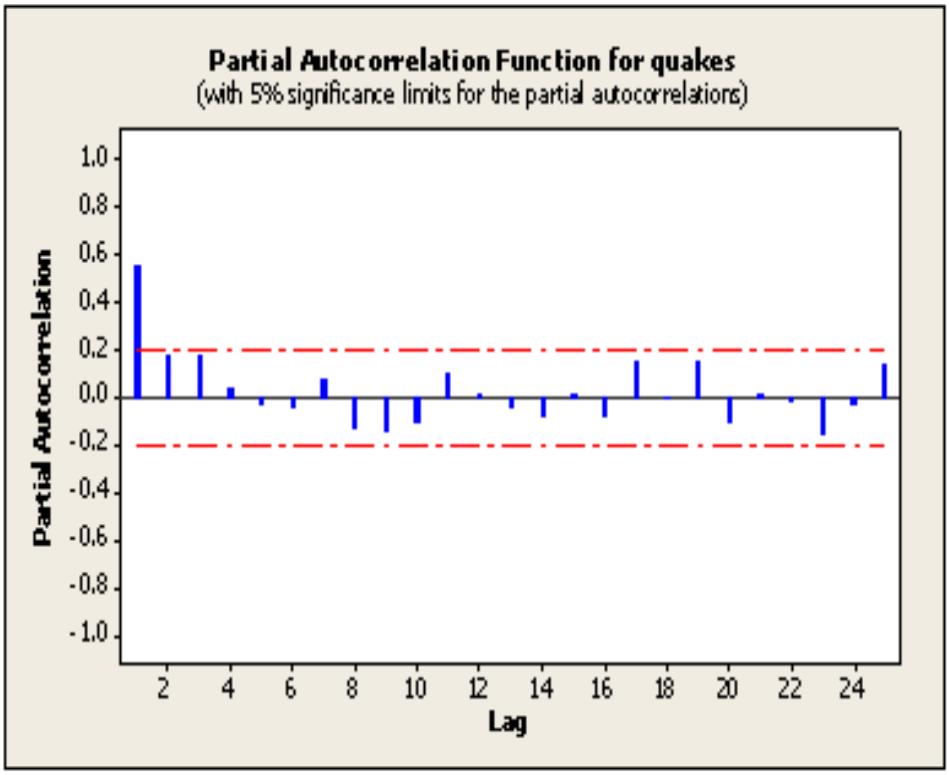
Οι αυτοσυσχετίσεις του δείγματος μειώνονται, αν και όχι τόσο γρήγορα όσο θα έπρεπε για ένα AR(1).

Στην περίπτωση του δείγματος των σεισμών, η παρατηρούμενη αυτοσυσχέτιση υστέρησης 2 (r_2) είναι ίση με 0,419, αρκετά μεγαλύτερη από το 0,293 = $0{,}542^2 = r^2{}_1$, γεγονός που υποδεικνύει πως, ένα ασθενώς στάσιμο μοντέλο AR(1) δεν προσαρμόζεται με τον καλύτερο δυνατό τρόπο.

Σημείωση

Οι κόκκινες γραμμές στο διάγραμμα χαρακτηρίζουν το 95% δ.ε. για τους υπολογισμούς της δειγματικής αυτοσυσχέτισης. Παρατηρήσεις που βρίσκονται εντός της ζώνης δεν διαφέρουν σημαντικά από το 0.

Ένα ακόμα κριτήριο προσαρμογής είναι η συνάρτηση μερικής αυτοσυσχέτισης (PACF - partial autocorrelation function). Παρατηρούμε ότι οι συντελεστές μερικής αυτοσυσχέτισης δεν διαφοροποιούνται σημαντικά από το 0 με την εξαίρεση του πρώτου συντελεστή που ταυτίζεται με τον 1ο συντελεστή αυτοσυσχέτισης, υπόδειξη πως το μοντέλο AR(1) είναι αποδεκτό για την περιγραφή των δεδομένων.



Παράδειγμα 2

Στο παράδειγμα της χρονοσειράς που αναπαριστάται στο διάγραμμα, η στοχαστική διεργασία που της αντιστοιχεί φανερά δεν είναι στάσιμη, άρα η εφαρμογή ενός AR(1) ασθενώς στάσιμου μοντέλου σε αυτήν δεν είναι κατάλληλη.

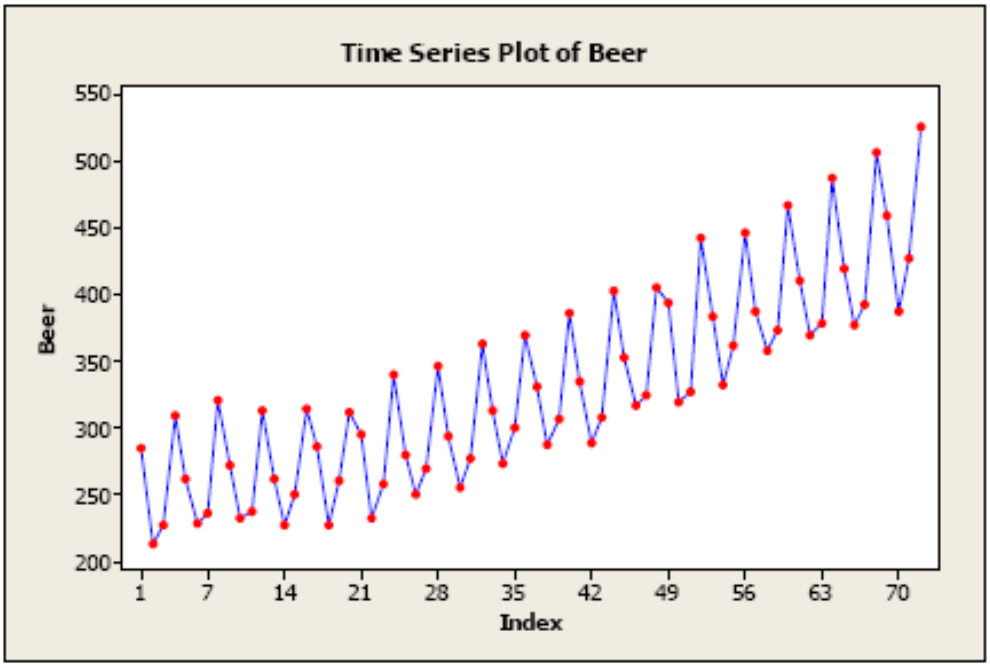
Μία από τις επιλογές, θα μπορούσε να είναι η προσαρμογή ενός γραμμικού μοντέλου της μορφής

$$X_t = \alpha \cdot t + \beta_1 \cdot S_1 + \beta_2 \cdot S_2 + \beta_3 \cdot S_3 + \beta_4 \cdot S_4 + e_t,$$

όπου S_1, S_2, S_3, S_4 μεταβλητές που παίρνουν την τιμή 1 ή 0 ανάλογα για το αν η παρατήρηση αντιστοιχεί στο 1°, 2° , 3° ή 4° τρίμηνο του έτους.

Επιπλέον, θα μπορούσε να προσαρμοστεί ένα μοντέλο της μορφής

$$X_t = \alpha \cdot t + \beta_1 \cdot \sin(\omega t)$$



Παράδειγμα 3

Στο διάγραμμα παρουσιάζονται οι θάνατοι από καρδιακό επεισόδιο στο LA μεταξύ 1970 και 1979.

Καθώς, η χρονοσειρά έχει αρνητική κλίση, δεν είναι στάσιμη, συνεπώς κατ' αρχήν δεν αρμόζει η προσαρμογή ενός ασθενώς στάσιμου AR(1) μοντέλου. Ωστόσο, η κατάσταση μπορεί να βελτιωθεί αν αντί των ίδιων των τιμών

$$\{x_1, x_2, x_3, \dots\}$$

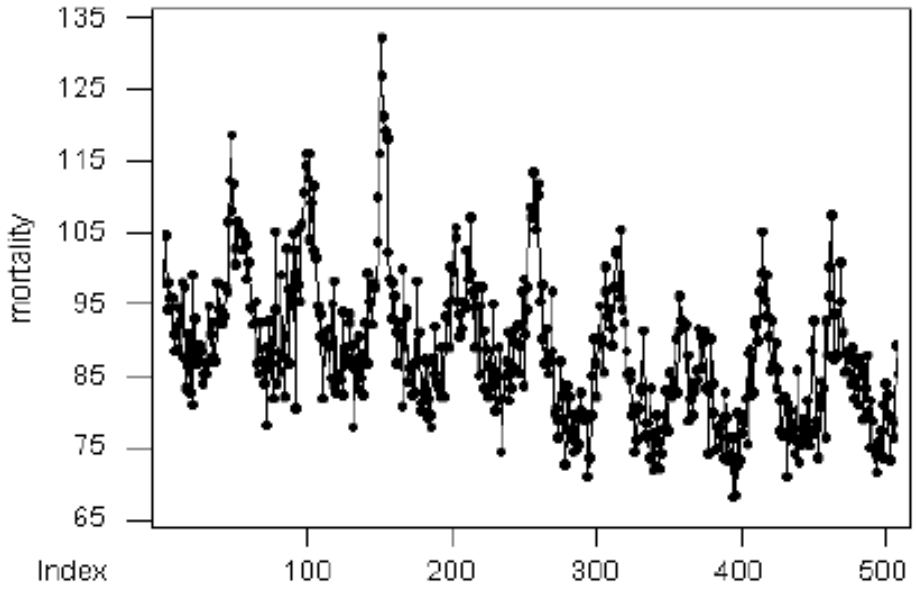
πάρουμε τις διαφορές πρώτης τάξης

$$\{z_1, z_2, z_3, \dots\} = \{x_2 - x_1, x_3 - x_2, x_4 - x_3, \dots\}.$$

Παρατηρούμε ότι η χρονοσειρά

$$\{z_1, z_2, z_3, \dots\}$$

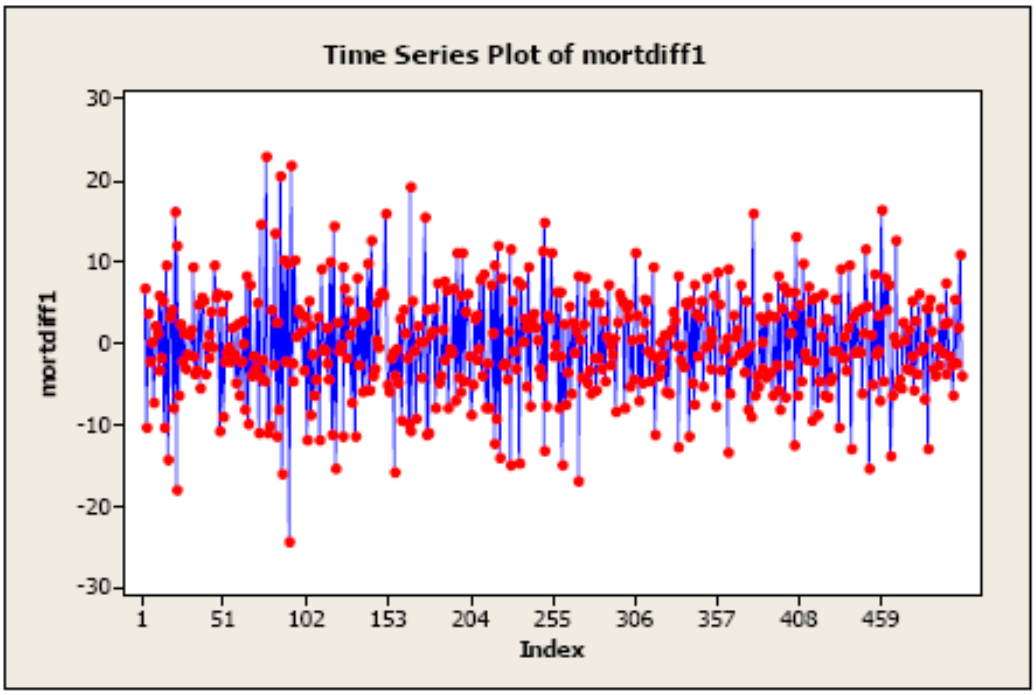
έχει σταθερή τάση.



Το γεγονός αυτό δεν είναι τυχαίο: Αν οι παρατηρήσεις x_t βρίσκονται κατά μήκος μίας ευθείας $y = \lambda x + \beta$ τότε οι διαφορές των διαδοχικών όρων πρέπει να είναι σταθερές και ίσες με το συντελεστή διεύθυνσης λ.

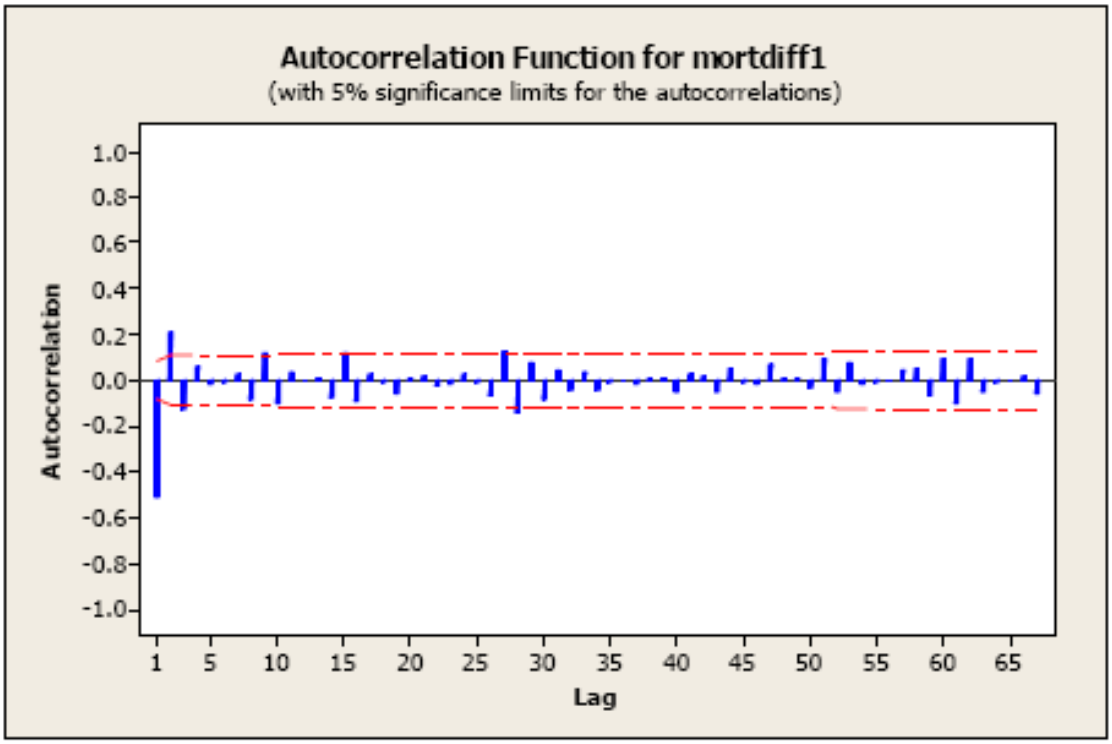
Η προσαρμογή ενός AR(1) μοντέλου σε αυτήν οδηγεί στην εξίσωση πρόβλεψης (των διαφορών πρώτης τάξης)

$$z_t = -0{,}046 - 0{,}506\, z_{t-1}.$$



Η καταλληλότητα του AR(1) μοντέλου για την ακολουθία των διαφορών αναδεικνύεται και από τις τιμές της αυτοσυσχέτισης οι οποίες στις πρώτες τους τιμές ικανοποιούν τις σχέσεις

$$\rho_2 = \rho^2_1,\ \rho_3 = \rho^3_1,\ \ \rho_4 = \rho^4_1.$$



15.6 Αυτοπαλινδρομικό μοντέλο 2ης τάξης (Μοντέλο AR(2))

Με το αυτοπαλινδρομικό μοντέλο τάξης 2, AR(2) (Autoregressive model of order 2) συσχετίζεται κάθε μία τιμή της σειράς $\{X_t, t \in N^*\}$ με την αμέσως προηγούμενη και την παρατήρηση που βρίσκεται δύο θέσεις πριν.

Η αναλυτική του έκφραση είναι

$$\mathbf{X_t = \delta + \varphi_1 \cdot X_{t-1} + \varphi_2 \cdot X_{t-2} + w_t,}$$

όπου w_t είναι τα ανεξάρτητα και ισόνομα σφάλματα των προβλέψεων με κανονική κατανομή $N(0, \sigma^2_w)$, τα οποία επιπλέον, δεν εξαρτώνται από την αντίστοιχη τιμή της σειράς x_t. Η αναλυτική του έκφραση είναι

$$X_t = \delta + \varphi_1 \cdot X_{t-1} + \varphi_2 \cdot X_{t-2} + w_t, \quad w_t \sim N(0, \sigma^2_w).$$

Αποδεικνύεται ότι για κάθε ασθενώς στάσιμο μοντέλο AR(2), οι αυτοσυσχετίσεις με υστέρηση 1, 2 και 3 αντίστοιχα δίνονται από τους τύπους

$$\rho_1 = \varphi_1 / (1 - \varphi_2).$$

$$\rho_2 = \varphi^2_1 / (1 - \varphi_2) + \varphi_2.$$

$$\rho_3 = (\varphi^3_1 + \varphi_1 \cdot \varphi_2) / (1 - \varphi_2) + \varphi_1 \cdot \varphi_2.$$

Αποδεικνύεται ότι για κάθε ασθενώς στάσιμο μοντέλο AR(2) ισχύει:

$$\varphi_1 = \rho_1(1 - \rho_2) / (1 - \rho^2_1),$$

$$\varphi_2 = (\rho_2 - \rho^2_1) / (1 - \rho^2_1),$$

όπου ρ_1, ρ_2, οι αυτοσυσχετίσεις με υστέρηση 1 και 2 αντίστοιχα.

Επιπλέον, $Var(X_t) = \sigma^2_w / (1 - \rho_1 \cdot \varphi_1 - \rho_2 \cdot \varphi_2)$, ή

$$\mathbf{Var(X_t) = (1 - \varphi_2)\sigma^2_w / [(1 + \varphi_2)(1 - \varphi_1 - \varphi_2)(1 + \varphi_1 - \varphi_2)]}$$

Από τη σχέση

$$Var(X_t) = (1 - \varphi_2)\sigma^2_w / [(1 + \varphi_2)(1 - \varphi_1 - \varphi_2)(1 + \varphi_1 - \varphi_2)]$$

καθώς, όλοι οι όροι της παράστασης οφείλουν να είναι θετικοί, έχουμε ότι στην περίπτωση της ασθενούς στάσιμης AR(2) χρονοσειράς, πρέπει:

- $-1 < \varphi_2 < 1$
- $\varphi_2 - \varphi_1 < 1$
- $\varphi_2 + \varphi_1 < 1$

16.7 Μοντέλο κινούμενου μέσου όρου (Moving Average)

Έστω $\{X_t, t \geq 1\}$ μία χρονοσειρά και $\{w_t\}$, $t \geq 1$, μία οικογένεια ισόνομων και ανεξάρτητων τυχαίων μεταβλητών που ακολουθούν την $N(0, \sigma^2_w)$ κατανομή (δηλαδή WN).

Το μοντέλο κινητού μέσου 1ης τάξης, που συμβολίζεται με MA(1) είναι:

$$X_t = \mu + w_t + \theta_1 \cdot w_{t-1}.$$

Το μοντέλο κινητού μέσου 2ης τάξης, που συμβολίζεται με MA(2) είναι:

$$X_t = \mu + w_t + \theta_1 \cdot w_{t-1} + \theta_2 \cdot w_{t-2}.$$

Το μοντέλο κινητού μέσου όρου q^{th} τάξης, που συμβολίζεται με MA(q) είναι:

$$X_t = \mu + w_t + \theta_1 \cdot w_{t-1} + \theta_2 \cdot w_{t-2} + \ldots + + \theta_q \cdot w_{t-q}.$$

Μοντέλο MA(1)

Το μοντέλο κινητού μέσου 1ης τάξης, που συμβολίζεται με MA(1) είναι:

$$X_t = \mu + w_t + \theta_1 \cdot w_{t-1},\ w_t \sim N(0, \sigma^2_w)$$

Εύκολα αποδεικνύεται ότι:

- $E(X_t) = \mu$.
- $Var(X_t) = (1 + \theta^2_1) \cdot \sigma^2_w$.

Επιπλέον, μπορεί να δειχθεί ότι

$$\rho_1 = \theta_1 / (1 + \theta^2_1) \text{ και } \rho_h = 0,\ h \geq 2.$$

Πράγματι: $Cov(X_t, X_{t-h}) = E[(X_t - \mu)\cdot(X_{t-h} - \mu)]$

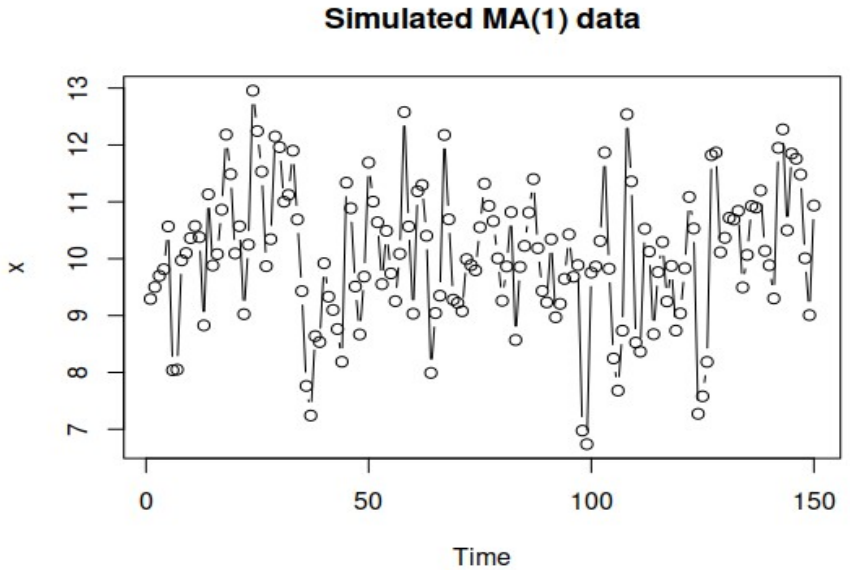
$$= E[(w_t + \theta_1 \cdot w_{t-1})\cdot(w_{t-h} + \theta_1 \cdot w_{t—h-1})]$$

$$= E(w_t w_{t-h} + \theta_1 w_{t-1} w_{t-h} + \theta_1 w_t w_{t—h-1} + \theta_1 w_{t-1} w_{t—h–1})$$

Προσέχουμε ότι για h = 1, είναι $Cov(X_t, X_{t-1}) = \theta_1 \cdot \sigma^2_w$, ενώ $Cov(X_t, X_{t-h}) = 0$, για h > 1. Με αυτές τις παρατηρήσεις, η απόδειξη του τύπου της συνδιακύμανσης προκύπτει εύκολα.

Παράδειγμα MA(1)

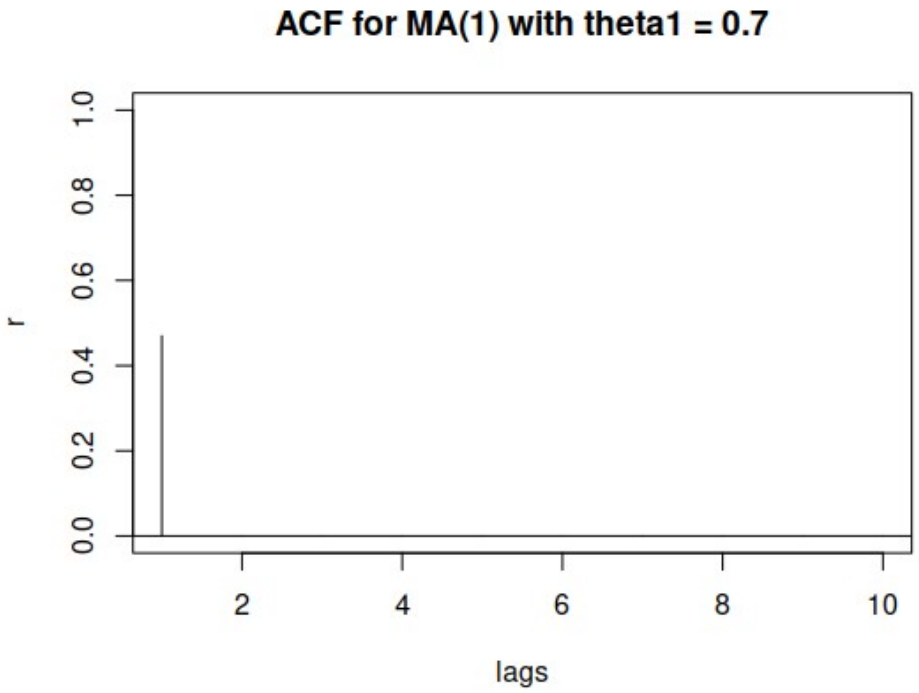
Θεωρητικό μοντέλο $X_t = 10 + w_t + 0.7 \cdot w_{t-1}$, όπου $w_t \sim N(0, 1)$.

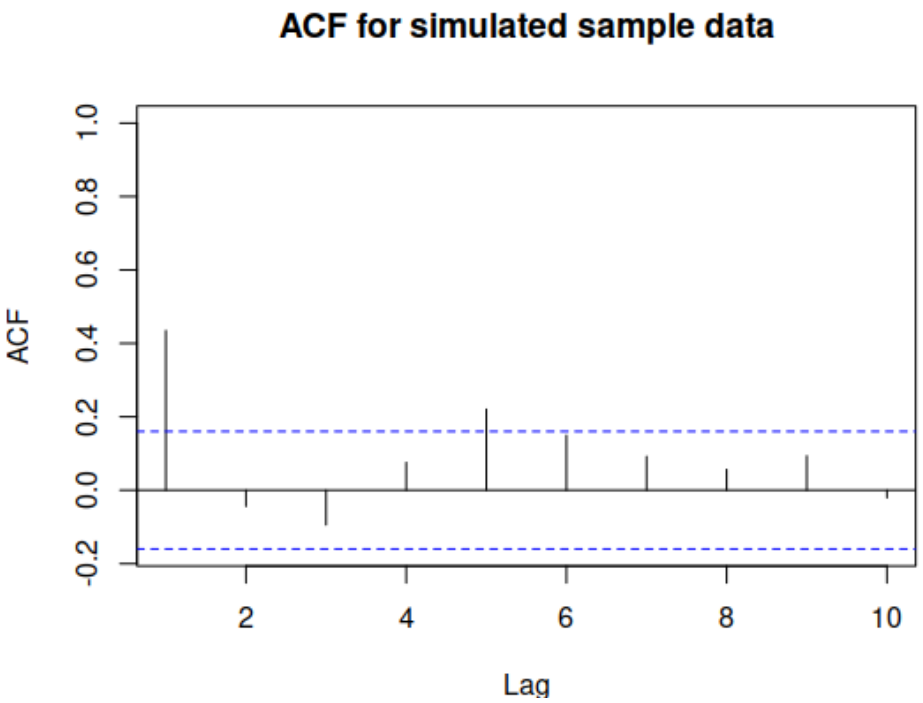


Υπολογίζουμε:

$\rho_1 = 0{,}7 / (1 + 0{,}7^2) = 0{,}470$ και

$\rho_h = 0$, h > 1.





Μοντέλο MA(2)

Το μοντέλο κινητού μέσου 2ης τάξης, που συμβολίζεται με MA(2) είναι:

$$X_t = \mu + w_t + \theta_1 \cdot w_{t-1} + \theta_2 \cdot w_{t-2}, \qquad w_t \sim N(0, \sigma^2_w)$$

Εύκολα αποδεικνύεται (πως;) ότι:

- $E(X_t) = \mu$.
- $Var(X_t) = (1 + \theta^2_1 + \theta^2_2) \cdot \sigma^2_w$.

Επιπλέον, μπορεί να δειχθεί ότι

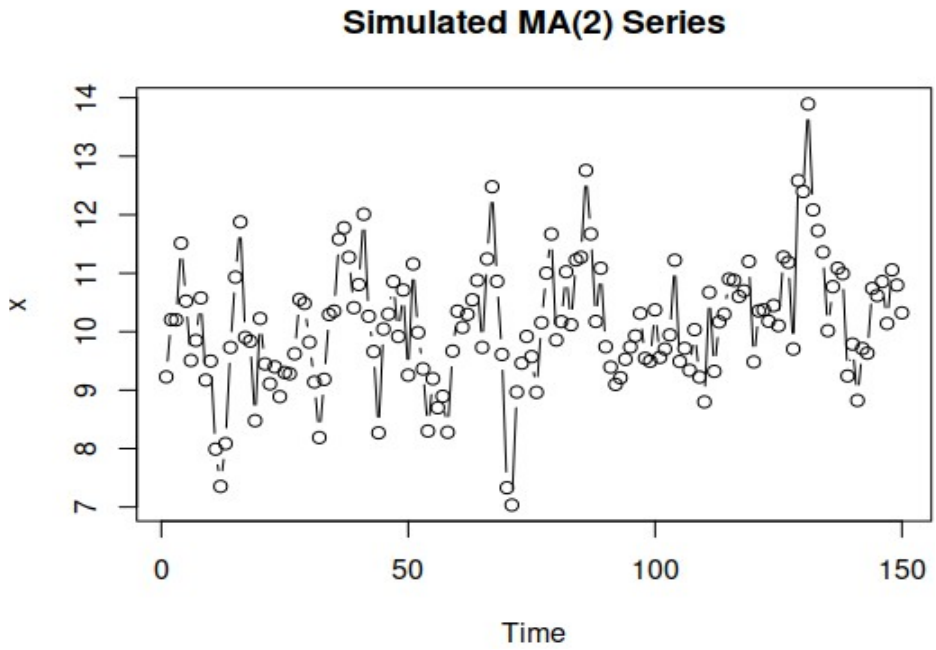
$$\rho_1 = (\theta_1 + \theta_2)/ (1 + \theta^2_1 + \theta^2_2),$$

$$\rho_2 = \theta_2 / (1 + \theta^2_1 + \theta^2_2),$$

$$\rho_h = 0, h \geq 3.$$

Παράδειγμα μοντέλου MA(2)

$X_t = 10 + w_t + 0.5w_{t-1} + 0.3w_{t-2}$, όπου $w_t \sim N(0, 1)$.

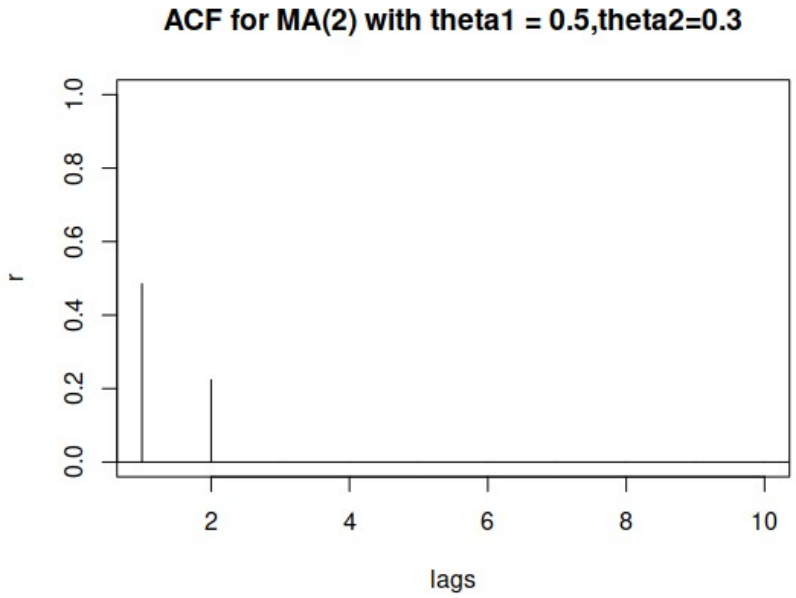


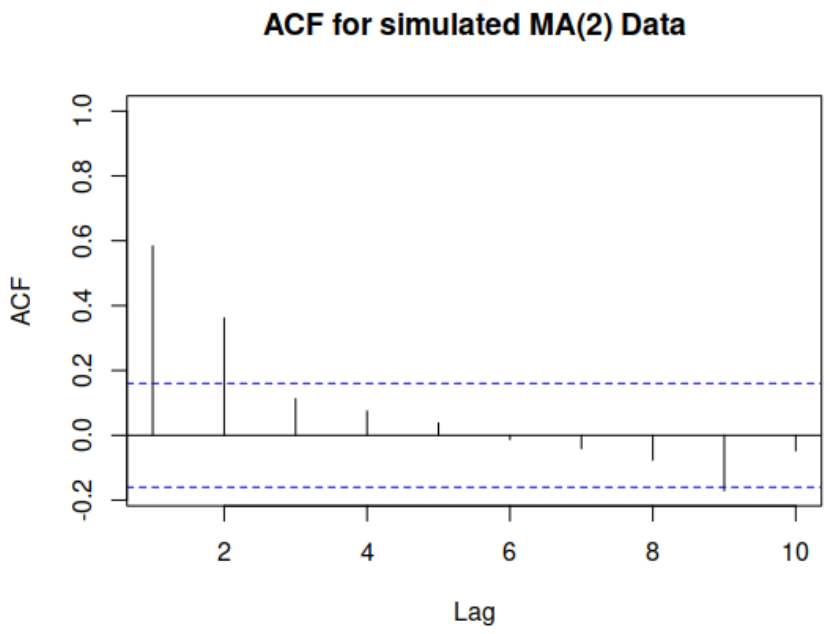
Υπολογίζουμε:

$\rho_1 = 0{,}485$,

$\rho_2 = 0{,}224$ και

$\rho_h = 0$, $h > 2$.





Σύνοψη

- Εάν υπάρχει μια εμφανής ανοδική ή καθοδική γραμμική τάση, ενδεχομένως απαιτείται ο υπολογισμός της πρώτης διαφοράς μεταξύ των τιμών.
- Αν οι αυτοσυσχετίσεις φαίνονται να ακολουθούν μία σχέση της μορφής

$$\rho_h = \varphi^h{}_1,$$

τότε ένα μοντέλο AR(1) πιθανώς να προσαρμόζεται καλώς στη χρονοσειρά.

- Αν

$$\rho_1 \neq 0 \text{ και } \rho_h = 0,\ h \geq 2,$$

τότε ένα μοντέλο MA(1) είναι αποδεκτή επιλογή.

- Αν

$$\rho_1 \neq 0,\ \ \rho_2 \neq 0 \text{ και } \rho_h = 0,\ h \geq 3,$$

τότε ένα μοντέλο MA(2) πιθανώς να προσαρμοστεί καλά στη χρονοσειρά.

15.7 Ολοκλήρωση (Integration) μίας σειράς (Διαφορές 1ης ή 2ης τάξης)

Εισαγωγική Δραστηριότητα

Επαληθεύστε ότι αν η χρονοσειρά $\{y_n\}$ ακολουθεί γραμμικό κανόνα $y_n = \alpha \cdot n + \beta$, τότε η διαφορά δύο διαδοχικών τιμών είναι σταθερή.

Αν η σειρά των διαφορών πρώτης τάξης δεν παρουσιάζει στάσιμη συμπεριφορά, τότε είναι δυνατός ο υπολογισμός των διαφορών δεύτερης τάξης.

Το ερώτημα που εύλογα προκύπτει είναι:

Πως μπορεί ο ερευνητής να αξιολογήσει τη στασιμότητα της στοχαστικής διεργασίας της οποίας η χρονοσειρά αποτελεί δείγμα τιμών;

Τα κύρια στοιχεία που θα συνηγορήσουν στη στασιμότητα ή μη της χρονοσειράς είναι η φύση της διαδικασίας.

Επιπλέον, υπάρχουν στατιστικοί έλεγχοι με τις οποίες ελέγχεται η υπόθεση πως η παρατηρούμενη χρονοσειρά αποτελεί εκδοχή μίας στάσιμης διεργασίας, όπως για παράδειγμα η δοκιμασία Kwiatkowski-Phillips-Schmidt-Shin (KPSS) ([1], [2]).

[1] 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 are we that economic time series have a unit root? Journal of Econometrics, 54(1-3), 159–178.

[2] https://en.wikipedia.org/wiki/KPSS_test

Μία διαφορά 1ης τάξης στα δεδομένα, σημαίνει πως η σειρά που μελετάται είναι αυτή που προκύπτει λαμβάνοντας τις διαφορές διαδοχικών όρων.

$$\{z_1, z_2, z_3, \dots\} = \{x_2 - x_1, x_3 - x_2, x_4 - x_3, \dots\}.$$

Η διαφορά 1ης τάξης επιλέγεται για την απομάκρυνση γραμμικής τάσης στη χρονοσειρά. Στην περίπτωση αυτή, η σειρά λέγεται ότι είναι I(1). (Ολοκληρωμένη – Integrated 1ης τάξης)

Μία διαφορά 2ης τάξης στα δεδομένα, σημαίνει πως η σειρά που μελετάται είναι αυτή που προκύπτει λαμβάνοντας τις διαφορές των διαφορών διαδοχικών όρων.

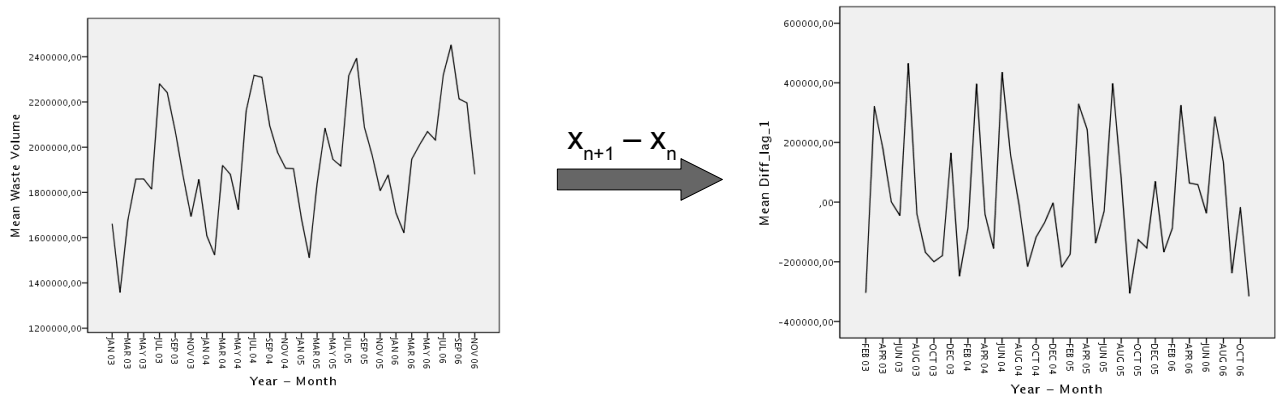
$$\{w_1, w_2, w_3, \dots\} = \{z_2 - z_1, z_3 - z_2, z_4 - z_3, \dots\}$$

$$= \{(x_3 - x_2) - (x_2 - x_1), (x_4 - x_3) - (x_3 - x_2), (x_5 - x_4) - (x_4 - x_3), \dots\}.$$

Η διαφορά 2ης τάξης επιλέγεται για την ερμηνεία χρονοσειράς με τετραγωνική τάση.

Στην περίπτωση αυτή, η σειρά λέγεται ότι είναι I(2). (Ολοκληρωμένη – Integrated 2ης τάξης)

Παράδειγμα



Ανεξαρτησία (Box-Ljung test):	$x^2(1) = 19.618$, $p < 0{,}001$	$x^2(1) = 0.001742$, $p = 0{,}967$
Στασιμότητα (Dickey-Fuller test):	DF.stat = -4.1194, p = 0.013	DF.stat = -2.7841, p = 0.262

Box-Ljung

H_0: Οι τιμές της χρονοσειράς είναι ανεξάρτητες μεταξύ τους ($X_t = \varepsilon_t$, όπου $\varepsilon_t \sim IWN(0, \sigma^2)$. Υπάρχει αυτοσυσχέτιση σημαντικά διαφορετική από το 0.

Dickey-Fuller

H_0: Οι διαφορές των τιμών της χρονοσειράς είναι ανεξάρτητες μεταξύ τους ($X_t - X_{t-1} = \varepsilon_t$, όπου $\varepsilon_t \sim IWN(0, \sigma^2)$.

15.8 Μοντέλα ARIMA(p, d, q)

Τα μοντέλα χρονοσειρών που είναι γνωστά ως μοντέλα ARIMA μπορεί να περιλαμβάνουν αυτοπαλινδρομικούς όρους, διαφορές πρώτης ή μεγαλύτερης τάξης μεταξύ των παρατηρήσεων και όρους κινητού μέσου όρου.

Τα στοιχεία στο μοντέλο καθορίζονται με τη σειρά (AR, διαφορές, MA).

Ένα μοντέλο με δύο όρους AR μόνο, θα προσδιορίζεται ως ARIMA τάξης (2,0,0).

Ένα μοντέλο MA(2) θα προσδιορίζεται ως ARIMA τάξης (0,0,2).

Ένα μοντέλο με έναν όρο AR, μια διαφορά πρώτης τάξης μεταξύ των τιμών και έναν όρο MA θα προσδιορίζεται ως (1,1,1).

Το μοντέλο ARIMA(p, d, q) είναι ένα μοντέλο ARMA(p, q) το οποίο προσαρμόζεται στις διαφορές d – τάξης της χρονοσειράς.

Παράδειγμα ARIMA(1, 1, 1): $Z_t = \mu + \alpha_1 Z_{t-1} + \theta_1 \varepsilon_{t-1}$, $Z_t = X_t - X_{t-1}$ ή

$$\mathbf{X_t - X_{t-1} = \mu + \alpha_1(X_{t-1} - X_{t-2}) + \theta_1 \varepsilon_{t-1},}$$

15.9 Γενική εξίσωση AR(p)

Έστω B ο τελεστής μετατόπισης της ακολουθίας μία θέση προς τα πίσω (ή τελεστής υστέρησης ή τελεστής ολίσθησης, backward shift) που στέλνει κάθε στοιχείο της ακολουθίας στο αμέσως προηγούμενό του:

$$\mathbf{B(x_n) = x_{n-1}}$$

Φανερά $\mathbf{B^{\kappa}(x_n) = x_{n-\kappa}}$. Με χρήση του τελεστή υστέρησης B, η γενική εξίσωση ενός AR(p) μοντέλου

$$\mathbf{X_t = \delta + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t,\ \ \varepsilon_t \sim IWN(0, \sigma^2),}$$

εκφράζεται ως εξής:

$$(1 - \alpha_1 B - \alpha_2 B^2 - \ldots - \alpha_p B^p) X_t = \delta + \varepsilon_t,$$

ή ακόμα και ως

$$\mathbf{\Phi(B) X_t = \delta + \varepsilon_t,}$$

όπου $\mathbf{\Phi(\omega) = 1 - \alpha_1}\omega - \alpha_2\omega^2 - \ldots - \alpha_p\omega^p$, το **χαρακτηριστικό πολυώνυμο** της AR(p).

Οι ρίζες του Φ(ω) προσδιορίζουν αν θα είναι στάσιμη ή όχι η AR(p) διεργασία.

Αν όλες οι ρίζες του Φ βρίσκονται έξω από τον μοναδιαίο δίσκο, τότε η AR(p) είναι στάσιμη.

16.10 Γενική εξίσωση MA(q)

Η γενική εξίσωση ενός MA(q) μοντέλου είναι ($\varepsilon_t \sim IWN(0, \sigma^2)$):

$$X_t = \mu + \varepsilon_t + \theta_1 \cdot \varepsilon_{t-1} + \theta_2 \cdot \varepsilon_{t-2} + \ldots + \theta_q \cdot \varepsilon_{t-q},$$

Χωρίς βλάβη της γενικότητας μπορεί να γραφεί

$$\mathbf{X_t = \mu + \varepsilon_t - \theta_1 \cdot \varepsilon_{t-1} - \theta_2 \cdot \varepsilon_{t-2} - \ldots - \theta_q \cdot \varepsilon_{t-q}.}$$

Με χρήση του τελεστή B η τελευταία σχέση εκφράζεται ως εξής:

$$X_t = \mu + (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)\varepsilon_t,$$

ή ακόμα και ως

$$\mathbf{X_t = \mu + \Theta(B)\varepsilon_t,}$$

όπου $\mathbf{\Theta(\omega) = 1 - \theta_1}\omega - \theta_{\mathbf{2}}\omega^{\mathbf{2}} - \ldots - \theta_{\mathbf{q}}\omega^{\mathbf{q}}$, το **χαρακτηριστικό πολυώνυμο** της MA(q).

15.10 Γενική εξίσωση ARMA(p, q)

Η γενική εξίσωση ενός ARMA(p, q) μοντέλου είναι ($\varepsilon_t \sim IWN(0, \sigma^2)$):

$$X_t = \delta + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q},$$

Χωρίς βλάβη της γενικότητας μπορεί να γραφεί

$$X_t = \delta + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q},$$

Με χρήση του τελεστή B η τελευταία σχέση εκφράζεται ως εξής:

$$\mathbf{\Phi(B)X_t = \delta + \Theta(B)\varepsilon_t}$$

15.11 Γενική εξίσωση I(d)

Παρατηρούμε ότι η διαφορά δύο διαδοχικών όρων μίας χρονοσειράς μπορεί να γραφεί ως:

$$\Delta X_t = X_t - X_{t-1} = X_t - B(X_t) = (I - B)X_t.$$

Άρα, η διαφορά 2ης τάξης των στοιχείων της χρονοσειράς γράφεται ως:

$$\Delta^2 X_t = Z_t - Z_{t-1} = Z_t - B(Z_t) = (I - B)Z_t = (I - B)^2 X_t.$$

Συμπεραίνουμε ανάλογα ότι ο υπολογισμός των διαφορών d τάξης μίας χρονοσειράς αποδίδει την ακολουθία

$$\mathbf{\Delta^d X_t = (I - B)^d X_t.}$$

15.12 Γενική εξίσωση ARIMA(p, d, q)

Καθώς, ένα ARIMA(p, d, q) μοντέλο προσαρμόζει ένα μοντέλο ARMA(p, q) στις διαφορές d τάξης της χρονοσειράς, συνάγουμε ότι η γενική εξίσωση του ARIMA(p, d, q) μοντέλου είναι:

$$\Phi(B)(I - B)^d X_t = \delta + \Theta(B)\varepsilon_t,$$

όπου $\Phi(\omega) = 1 - \alpha_1\omega - \alpha_2\omega^2 - \ldots - \alpha_p\omega^p$, το χαρακτηριστικό πολυώνυμο της AR(p) και $\Theta(\omega) = 1 - \theta_1\omega - \theta_2\omega^2 - \ldots - \theta_q\omega^q$, το χαρακτηριστικό πολυώνυμο της MA(q).

15.13 Επιλέγοντας ένα ARIMA μοντέλο

Για την επιλογή των κατάλληλων συνιστωσών που αποτελούν ένα ARIMA μοντέλο ο ερευνητής οφείλει να μελετήσει

(α) Το διάγραμμα των τιμών (x ο χρόνος και y η τιμή της χρονοσειράς)

(β) Το διάγραμμα αυτοσυσχέτισης (συνάρτηση ACF)

(γ) Το διάγραμμα μερικής αυτοσυσχέτισης (συνάρτηση PACF).

Εάν υπάρχει μια εμφανής ανοδική ή καθοδική γραμμική τάση, ενδεχομένως απαιτείται ο υπολογισμός της πρώτης διαφοράς μεταξύ των τιμών. Μια τετραγωνική τάση ενδεχομένως απαιτεί διαφορά $2^{ης}$ τάξης.

Τα διαγράμματα ACF και PACF είναι καλό να ερμηνευθούν μαζί. Η συνάρτηση PACF ενός αυτοπαλινδρομικού μοντέλου αναμένεται να έχει μη μηδενικές τιμές στους αντίστοιχους αυτοπαλινδρομικούς όρους που περιέχει το μοντέλο και μηδενικές τιμές στους υπόλοιπους όρους.

Εάν υπάρχει μια εμφανής ανοδική ή καθοδική γραμμική τάση, ενδεχομένως απαιτείται ο υπολογισμός της πρώτης διαφοράς μεταξύ των τιμών.

Αν οι αυτοσυσχετίσεις φαίνονται να ακολουθούν μία σχέση της μορφής

$$\rho_h = \varphi^h{}_1,$$

τότε ένα μοντέλο AR(1) πιθανώς να προσαρμόζεται καλώς στη χρονοσειρά.

Αν

$$\rho_1 \neq 0 \text{ και } \rho_h = 0,\ h \geq 2,$$

τότε ένα μοντέλο MA(1) είναι αποδεκτή επιλογή.

Αν

$$\rho_1 \neq 0, \ \ \rho_2 \neq 0 \text{ και } \rho_h = 0, \ h \geq 3,$$

τότε ένα μοντέλο MA(2) πιθανώς να προσαρμοστεί καλά στη χρονοσειρά.

15.14 Εκτίμηση διαγραμμάτων ACF και PACF

AR(p): (χρονοσειρά που συμφωνεί με την υπόθεση πως κάθε παρατήρηση προσδιορίζεται από τις p προηγούμενες)

Οι αυτοσυσχετίσεις (διάγραμμα ACF) μειώνονται σημαντικά (γεωμετρικά ή και ημιτονοειδώς) μετά τη θέση p.

Οι μερικές αυτοσυσχετίσεις (διάγραμμα PACF) είναι 0 μετά τη θέση p.

MA(q): (χρονοσειρά που συμφωνεί με την υπόθεση πως το σφάλμα κάθε μέτρησης προκύπτει ως συσσώρευση των q τελευταίων σφαλμάτων)

Οι αυτοσυσχετίσεις είναι 0 μετά τη θέση q.

Οι μερικές αυτοσυσχετίσεις μειώνονται γεωμετρικά μετά τη θέση q.

ARMA(p, q): (Οι τιμές της χρονοσειράς προβλέπονται από μία AR(p) διεργασία και τα σφάλματα με μία MA(q) διεργασία).

Οι αυτοσυσχετίσεις μειώνονται σημαντικά μετά τη θέση q.

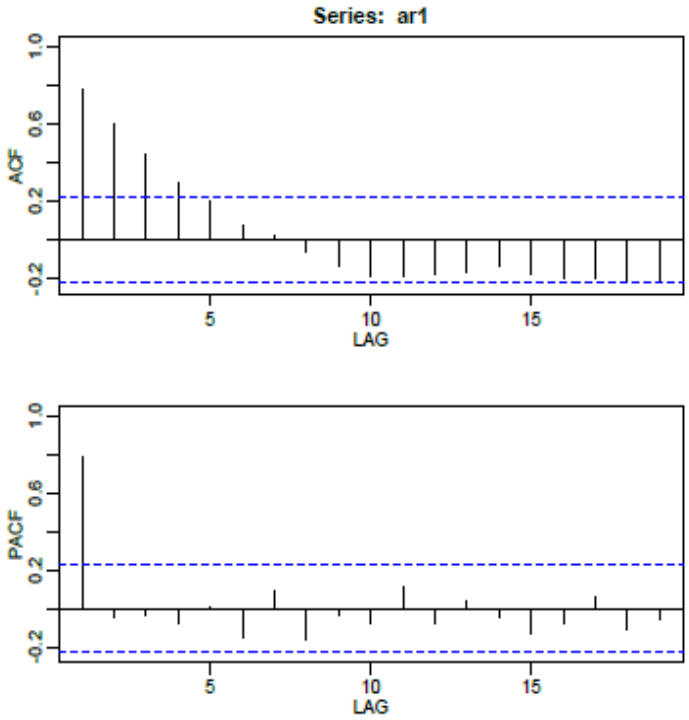
Οι μερικές αυτοσυσχετίσεις φθίνουν γεωμετρικά μετά τη θέση p.

Παράδειγμα 1

Το διάγραμμα της δειγματικής ACF είναι σύμφωνο με την υπόθεση πως η χρονοσειρά ακολουθεί ένα AR(1) μοντέλο, καθώς οι τιμές της ακολουθούν μία σχέση της μορφής

$$r_h = r^h_1$$

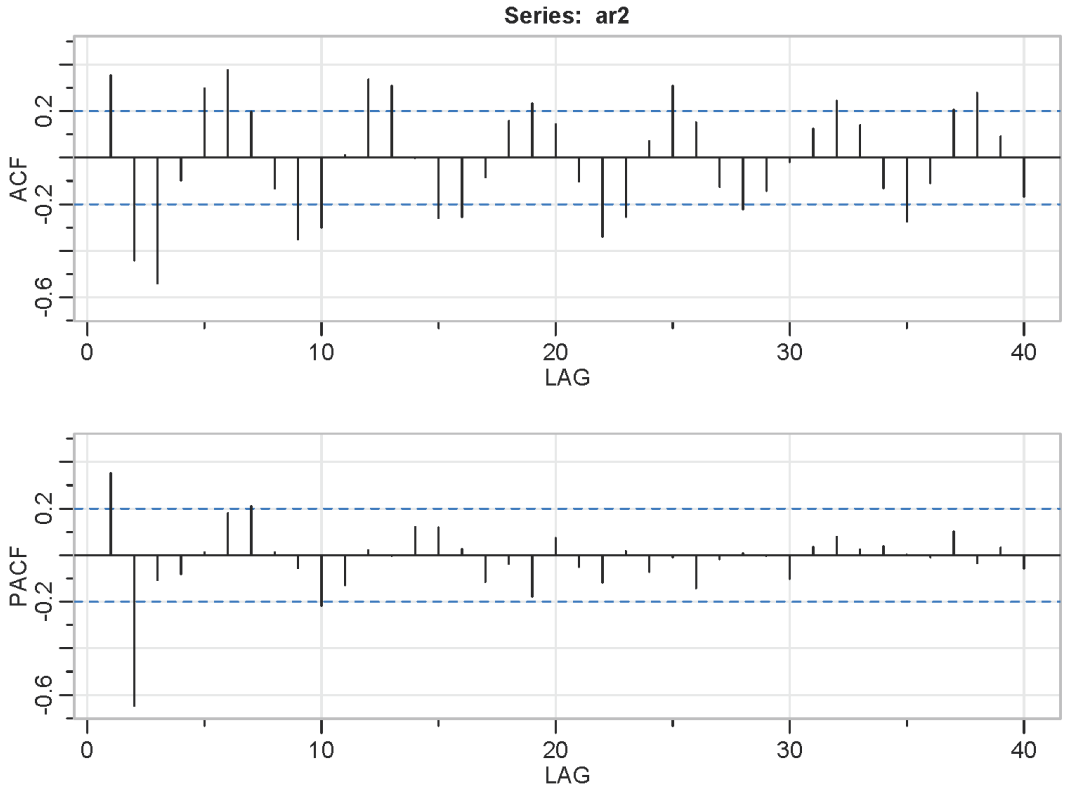
Το ίδιο υποστηρίζεται από το διάγραμμα της PACF, όπου μόνο η 1η μερική αυτοσυσχέτιση είναι σημαντικά διαφορετική από το 0, παρατήρηση συμβατή με το AR(1) μοντέλο.



Παράδειγμα 2

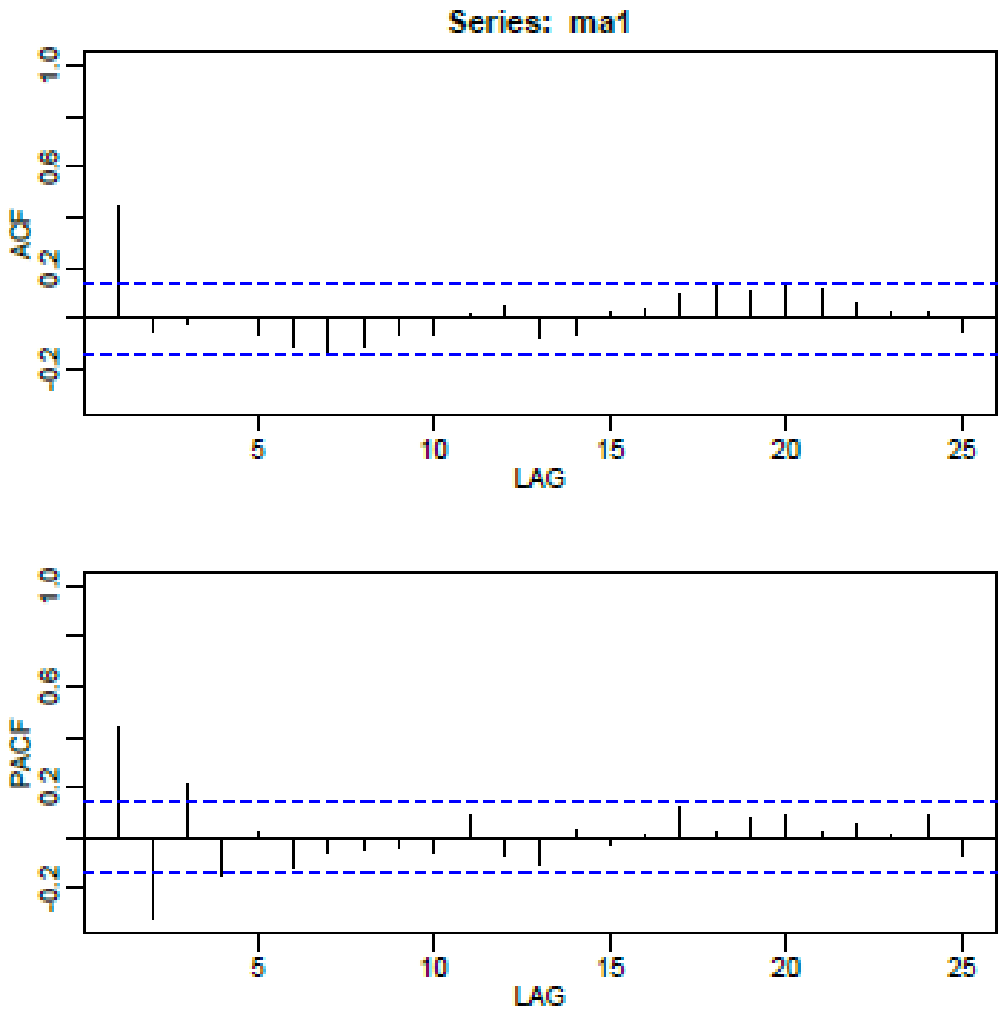
Το διάγραμμα της δειγματικής ACF προσαρμόζεται στην υπόθεση πως η χρονοσειρά ακολουθεί ένα AR(2) μοντέλο, καθώς οι τιμές της εναλλάσσονται μεταξύ θετικών και αρνητικών τιμών.

Το ίδιο υποστηρίζεται από το διάγραμμα της PACF, όπου οι πρώτες δύο μερικές αυτοσυσχετίσεις είναι σημαντικά διαφορετικές από το 0, παρατήρηση συμβατή με το AR(2) μοντέλο.



Παράδειγμα 3

Το συσχετόγραμμα (correlogram) υποδεικνύει πως η χρονοσειρά μπορεί να προσεγγιστεί από ένα MA(1) μοντέλο, καθώς μόνο η πρώτη αυτοσυσχέτιση είναι σημαντικά μεγαλύτερη του μηδενός.



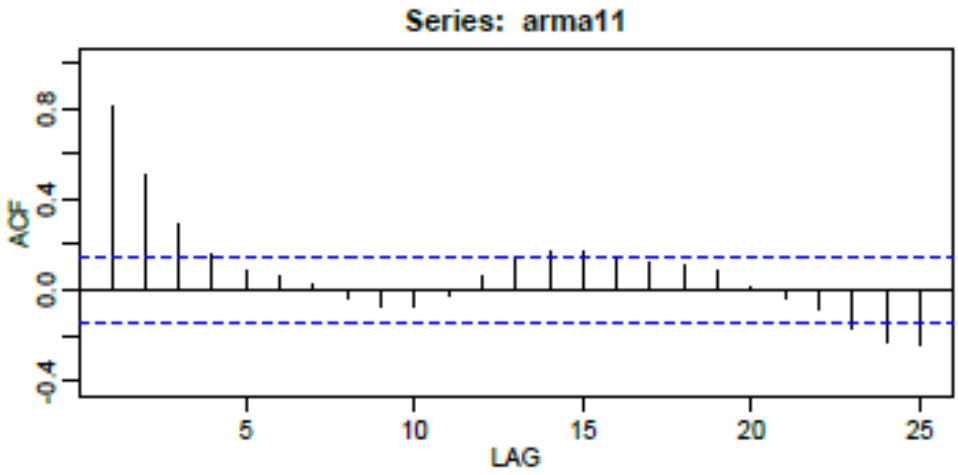
Παράδειγμα 4

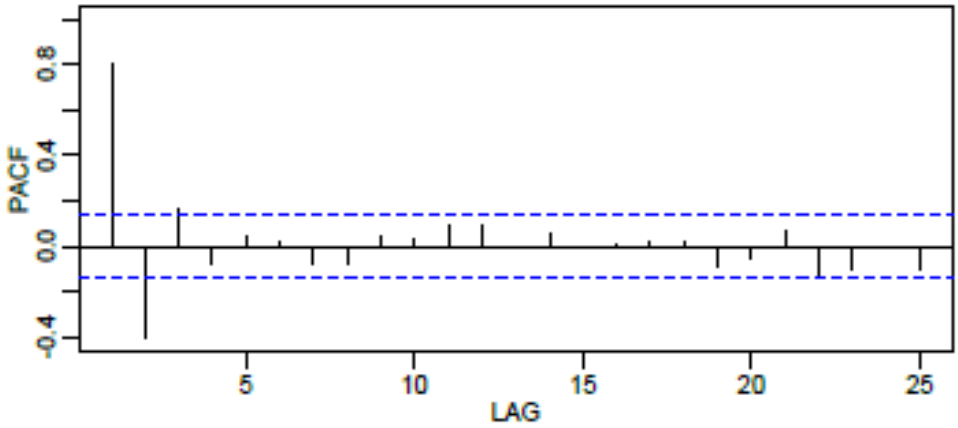
Οι αυτοσυσχετίσεις φθίνουν γεωμετρικά μετά την 1η ή τη 2η ενώ οι μερικές αυτοσυσχετίσεις φθίνουν μετά την 1η ή τη 2η.

Αυτό το δείγμα τιμών προήλθε από μία ARMA(1, 1).

Σε αυτό, θα μπορούσε να προσαρμοστεί και ένα ARMA(2, 2).

Μεταξύ των δύο, τελικά θα επιλέγαμε το πιο απλό, δηλαδή το ARMA(1, 1).





Παρατηρήσεις

- Εάν η αυτοσυσχέτιση (διάγραμμα ACF) και η μερική αυτοσυσχέτιση (διάγραμμα PACF) δεν φθίνουν προς το μηδέν, αλλά παραμένουν κοντά στο 1 για πολλές τιμές υστέρησης (lag) τότε, η σειρά είναι μη στάσιμη και πιθανώς απαιτείται ο υπολογισμός των διαφορών $1^{ης}$ τάξης και ενδεχομένως και μεγαλύτερης.
- Εάν το σύνολο των αυτοσυσχετίσεων δεν διαφοροποιείται σημαντικά από το 0, τότε η σειρά είναι τυχαία (λευκός θόρυβος), η σειρά των παρατηρήσεων έχει σημασία, αλλά τα δεδομένα εξελίσσονται με ανεξάρτητο τρόπο και είναι πανομοιότυπα κατανεμημένα.
- Εάν η σειρά $\{x_1, x_2, x_3, \dots\}$ έχει αντικατασταθεί με τις διαφορές $1^{ης}$ τάξης $\{z_1, z_2, z_3, \dots\} = \{x_2 - x_1, x_3 - x_2, x_4 - x_3, \dots\}$, και όλες οι αυτοσυσχετίσεις των διαφορών είναι μη σημαντικές, τότε η αρχική σειρά αποτελεί έναν τυχαίο περίπατο. Τα δεδομένα εξαρτώνται μεταξύ τους και δεν κατανέμονται πανομοιότυπα. Ο μέσος όρος όσο και η διακύμανση αυξάνονται με την πάροδο του χρόνου.

15.15 ARIMA με περιοδικότητα (ή SARIMA)

Στην περίπτωση όπου υπάρχει περιοδικότητα στη χρονοσειρά τότε αυτή εξ' ορισμού δεν είναι στάσιμη καθώς η μέση τιμή δεν είναι διαχρονικά σταθερή (λαμβάνει διαφορετική τιμή σε διαφορετικά χρονικά διαστήματα).

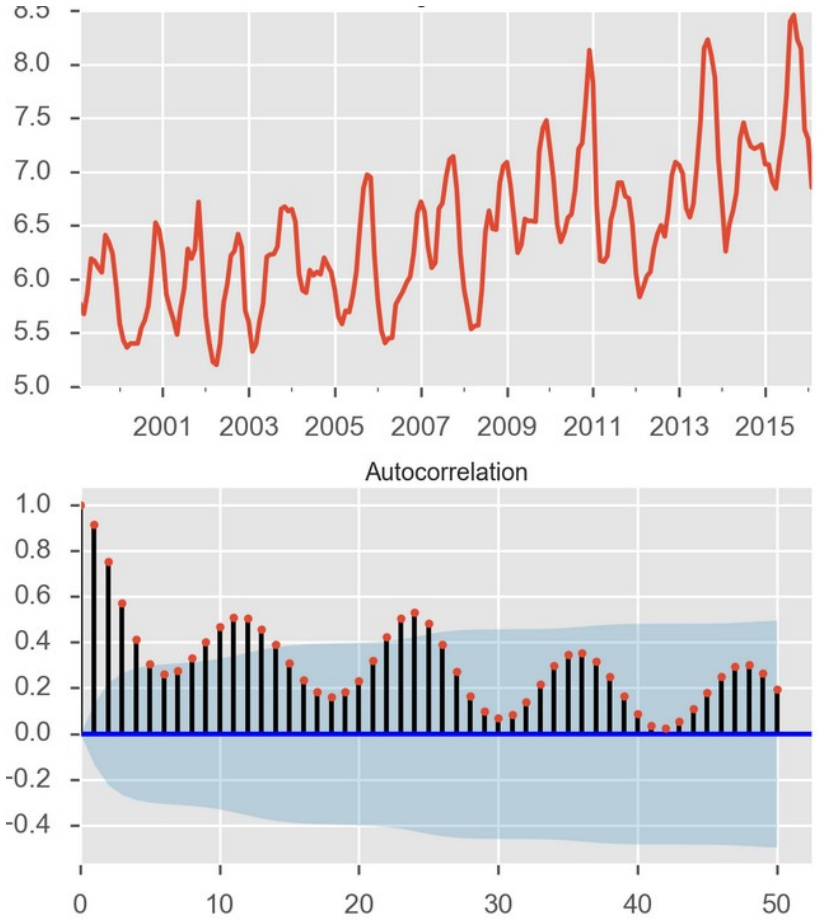
Η αναγνώριση της περιοδικότητας μπορεί να γίνει:

(α) Από το χρονοδιάγραμμα των τιμών με απλή παρατήρηση.

(β) Από το διάγραμμα αυτοσυσχετίσεων ACF: Θετική αυτοσυσχέτιση σε σταθερό πλήθος διαφορών δείχνει περιοδικότητα αντίστοιχης περιόδου.

Στην περίπτωση όπου υπάρχει περιοδικότητα στη χρονοσειρά τότε αυτή εξ' ορισμού δεν είναι στάσιμη καθώς η μέση τιμή δεν είναι διαχρονικά σταθερή.

Η πρώτη ενέργεια στην περίπτωση αυτή είναι ο εντοπισμός της εξίσωσης που θα εξαλείψει από τα δεδομένα την περιοδικότητα, ώστε στη συνέχεια να γίνει εφαρμογή ενός ARIMA μοντέλου στην χρονοσειρά που θα απομείνει ως υπόλοιπο μετά την εφαρμογή της εποχικής εξίσωσης.



Για την περιγραφή της περιοδικότητας μπορεί να χρησιμοποιηθούν ψευδομεταβλητές. Έτσι, αν η παρατηρούμενη περιοδικότητα είναι μήκους 4, η προσαρμογή του μοντέλου (με τη μέθοδο των ελαχίστων τετραγώνων)

$X_t = \gamma_1 D_{t,1} + \gamma_2 D_{t,2} + \gamma_3 D_{t,3} + \gamma_4 D_{t,4}$, όπου $D_{t,i} = 1$ αν $t = i + 4k$ και 0 αλλού,

μπορεί να οδηγήσει στην αφαίρεση της επιρροής της συγκεκριμένης περιοδικότητας.

(Στο παραπάνω μοντέλο οι συντελεστές γ_i, προσδιορίζουν τη διαφοροποίηση κάθε μίας από τις στιγμές της εποχικότητας)

Ένας άλλος τρόπος είναι η προσαρμογή ενός νέου ARIMA μοντέλου για την εξήγηση της εποχικής συνιστώσας της χρονοσειράς. Για την περιγραφή της περιοδικότητας μπορεί να χρησιμοποιηθούν αυτοπαλινδρομικοί όροι (πλήθος P) , διαφορές (πλήθος D) και κινούμενοι μέσοι όροι (πλήθος Q). Ένα τέτοιο μοντέλο το συμβολίζουμε

$$\textbf{ARIMA}(p, d, q)(P, D, Q)_m,$$

Όπου m το πλήθος των βημάτων που ορίζει την περίοδο της χρονοσειράς (π.χ. για χρονοσειρά που έχει μηνιαίες τιμές και εμφανίζει ετήσια περιοδικότητα, θα είναι m = 12, αν έχει μηνιαίες τιμές και εμφανίζει τριμηνιαία εποχικότητα, θα είναι m = 3 κλπ).

Στην παραπάνω υλοποίηση, η διαφορά D αντιστοιχεί σε όρο $(1 - B^m)^D$, η AR(P) αντιστοιχεί σε όρο της μορφής $\Phi_1(B^m)$ ο οποίος εφαρμόζεται, πριν το κύριο ARIMA(p, d, q), στις τιμές της χρονοσειράς ενώ ο MA(Q) όρος εφαρμόζεται στα υπόλοιπα της χρονοσειράς, πριν την εφαρμογή τυχόν MA(q) όρων του βασικού μοντέλου.

Για να γράψουμε την εξίσωση ενός ARIMA(p, d, q)(P, D, Q)$_m$ μοντέλου, συμφωνούμε να γράφουμε με κεφαλαία γράμματα τις εποχικές συνιστώσες του μοντέλου και με μικρά τις μη εποχικές.

Κάποιες φορές η αναλυτική καταγραφή μίας εξίσωσης SARIMA, είναι μία απλή διαδικασία.

Παράδειγμα

Ένα ARIMA(0, 2, 1)(0, 0, 1)$_{12}$ μοντέλο, χωρίς σταθερά, έχει εξίσωση της μορφής

$$(1 - B)^2 X_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\varepsilon_t, \text{ ή (γιατί;)}$$

$$\mathbf{X_t - 2X_{t-1} + X_{t-2} = \varepsilon_t + \Theta_1\varepsilon_{t-12} + \theta_1\varepsilon_{t-1} + \theta_1\Theta_1\varepsilon_{t-13}.}$$

Για μεγαλύτερες τιμές των p, d, q, P, D, Q η πολυπλοκότητα της εξίσωσης μεγαλώνει ανάλογα.

Παράδειγμα ARIMA(2, 1, 0)(0, 2, 2)$_{12}$

(πηγή: https://stats.stackexchange.com/questions/69407/how-do-i-write-a-mathematical-equation-for-arima-2-1-0-x-0-2-2-period-12)

$$(1 - \phi_1 B - \phi_2 B^2)(1 - B)(1 - B^{12})^2 Y_t = (1 + \Theta_1 B^{12} + \Theta_2 B^{24})e_t$$

$$(1 - \phi_1 B - \phi_2 B^2)(1 - B)(1 - 2B^{12} + B^{24})Y_t = (1 + \Theta_1 B^{12} + \Theta_2 B^{24})e_t$$

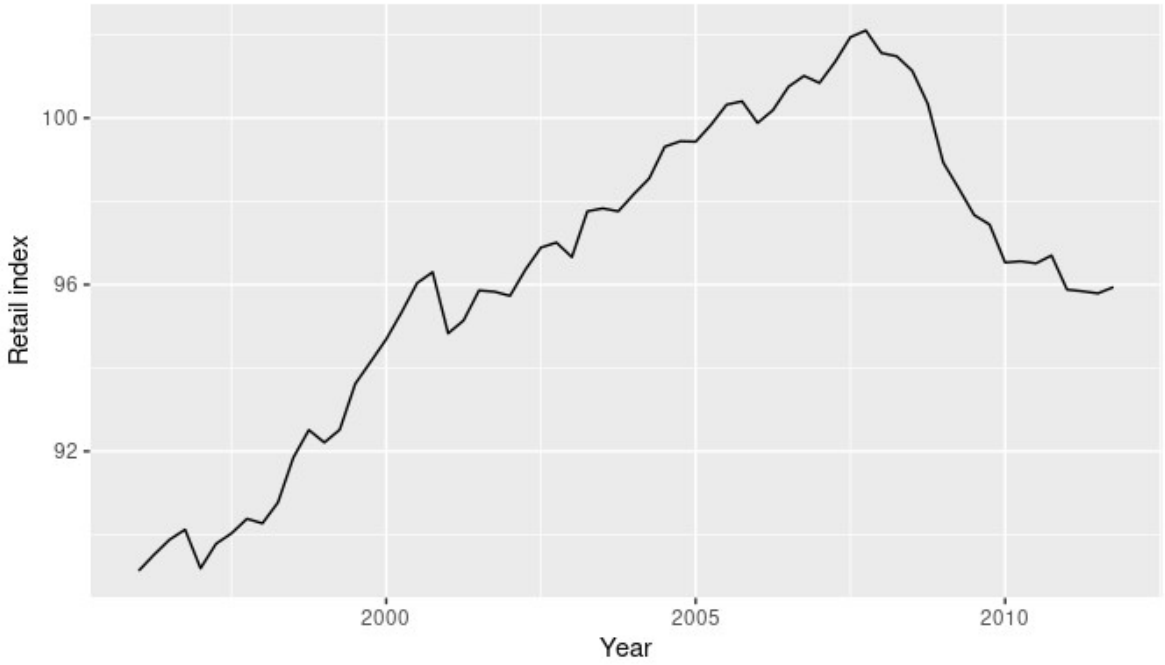
$$(1 - \phi_1 B - \phi_2 B^2 - B + \phi_1 B^2 + \phi_2 B^3)(1 - 2B^{12} + B^{24})Y_t = (1 + \Theta_1 B^{12} + \Theta_2 B^{24})e_t$$

$$(1 - \phi_1 B - \phi_2 B^2 - B + \phi_1 B^2 + \phi_2 B^3 - 2B^{12} + 2\phi_1 B^{13} + 2\phi_2 B^{14} + 2B^{13} - 2\phi_1 B^{14} - 2\phi_2 B^{15} + B^{24} - \phi_1 B^{25} - \phi_2 B^{26} - B^{25} + \phi_1 B^{26} + \phi_2 B^{27})Y_t = (1 + \Theta_1 B^{12} + \Theta_2 B^{24})e_t$$

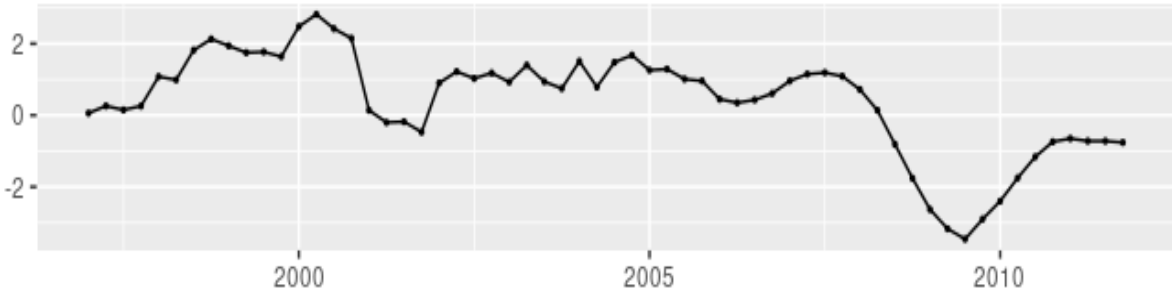
$$\boxed{Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - Y_{t-1} + \phi_1 Y_{t-2} + \phi_2 Y_{1-3} - 2Y_{t-12} + 2\phi_1 Y_{t-13} + 2\phi_2 Y_{t-14} + 2Y_{t-13} - 2\phi_1 Y_{t-14} - 2\phi_2 Y_{t-15} + Y_{t-24} - \phi_1 Y_{t-25} - \phi_2 Y_{t-26} - Y_{t-25} + \phi_1 Y_{t-26} + \phi_2 Y_{t-27} = e_t + \Theta_1 e_{t-12} + \Theta_2 e_{t-24}}$$

Παράδειγμα ARIMA με περιοδικότητα

Θα περιγράψουμε την εποχική διαδικασία μοντελοποίησης ARIMA χρησιμοποιώντας τριμηνιαία στοιχεία του ευρωπαϊκού λιανικού εμπορίου από το 1996 έως το 2011.



Τα δεδομένα είναι μη στάσιμα, ενώ παρουσιάζουν και εποχικότητα, οπότε θα πάρουμε πρώτα μια εποχιακή διαφορά τεσσάρων περιόδων (m = 4 τρίμηνα = 1 έτος)



Τα εποχιακά διαφοροποιημένα δεδομένα φαίνονται επίσης να είναι μη στάσιμα, επομένως παίρνουμε και μια πρώτη διαφορά, στις τιμές αυτής της σειράς.

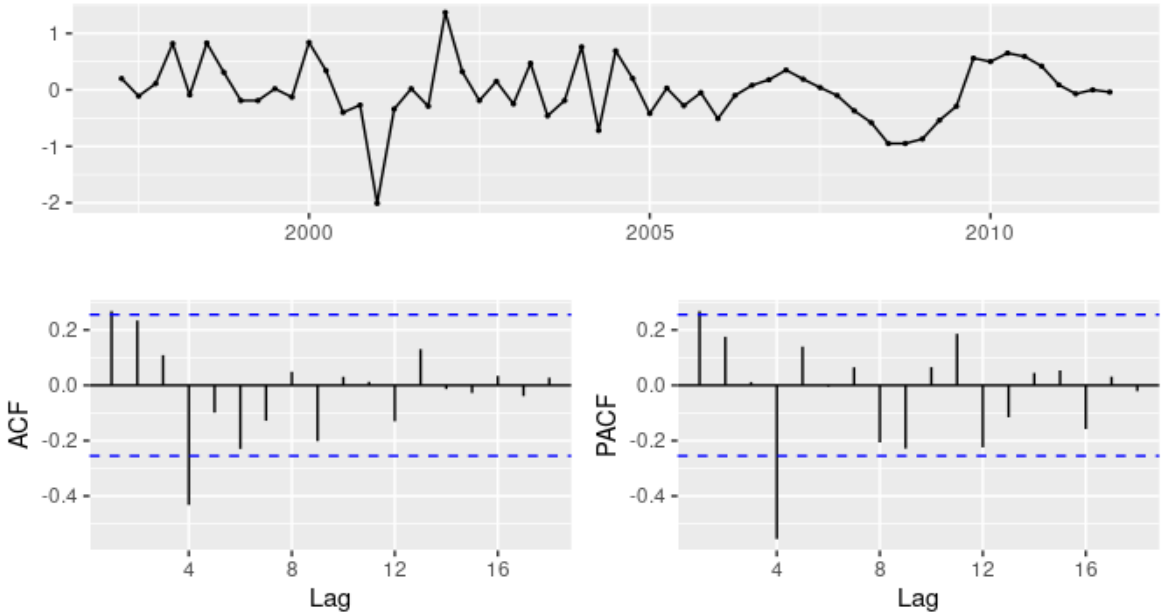
Η σειρά που προκύπτει έχει εύλογη στασιμότητα.

Το σημαντικό μέγεθος της αυτοσυσχέτισης με υστέρηση 1 υποδηλώνει ένα παράγοντα MA(1) στα μετασχηματισμένα δεδομένα ενώ η σημαντική αυτοσυσχέτιση με υστέρηση 4 υποδηλώνει ένα εποχιακό MA(1) παράγοντα.

Κατά συνέπεια, ξεκινάμε με ένα μοντέλο

$$\text{ARIMA}(0,1,1)(0,1,1)_4,$$

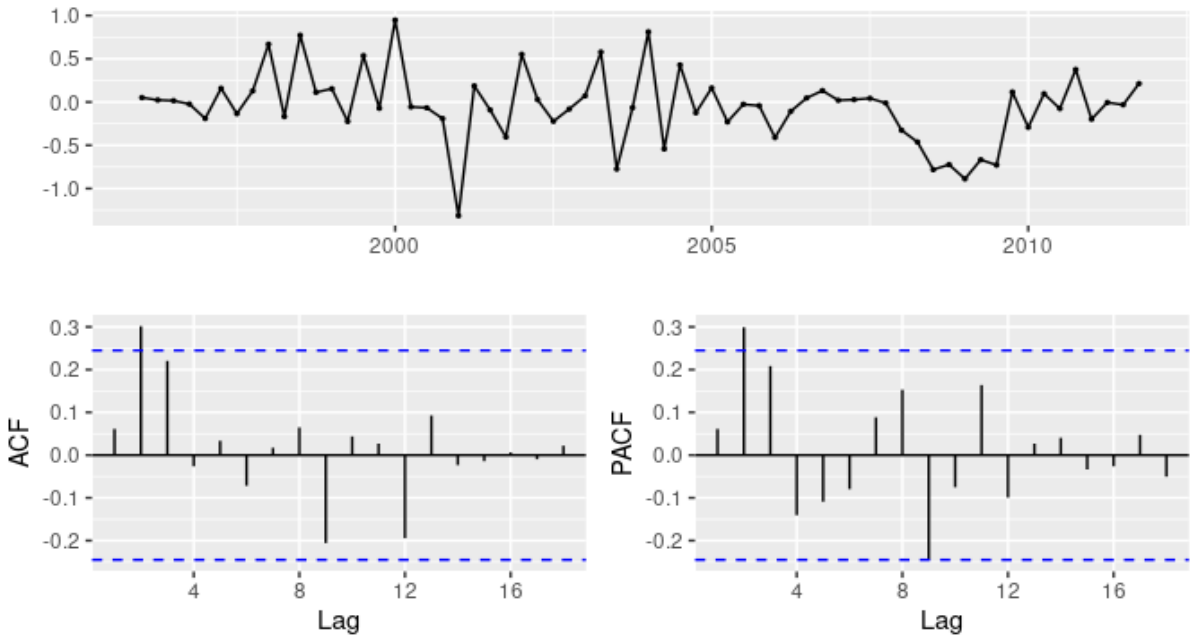
που υποδεικνύει μια πρώτη και εποχιακή διαφορά, και μη εποχικά και εποχιακά στοιχεία MA(1).



Τα υπόλοιπα του μοντέλου

$$\text{ARIMA}(0,1,1)(0,1,1)_4,$$

υποδεικνύουν ότι ορισμένοι πρόσθετοι μη εποχικοί όροι πρέπει να συμπεριληφθούν σε αυτό.

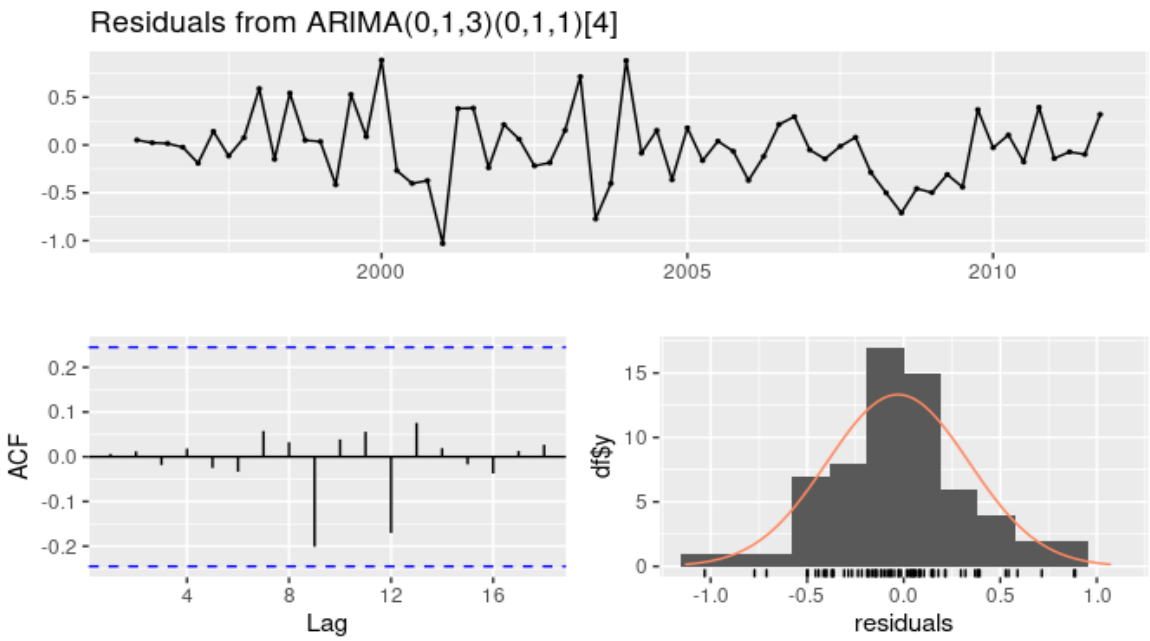


Ύστερα από δοκιμές επιλέγεται το μοντέλο

$$\text{ARIMA}(0,1,3)(0,1,1)_4,$$

Τα υπόλοιπά του απεικονίζονται στο Σχήμα 8.21.

Όλες οι τιμές της αυτοσυσχέτισης (ACF) είναι στατιστικά μη σημαντικές, δηλαδή τα υπόλοιπα φαίνεται να είναι λευκός θόρυβος.



15.16 Κριτήριο AIC (Akaike information criterion)

Το κριτήριο πληροφοριών Akaike (AIC) (Hirotugu Akaike(1927 – 2009) είναι ένας εκτιμητής του σφάλματος πρόβλεψης ενός μοντέλου. Συνεπώς, εκφράζεται με αυτό η σχετική ποιότητα πρόβλεψης ενός στατιστικού μοντέλου για ένα σύνολο δεδομένων.

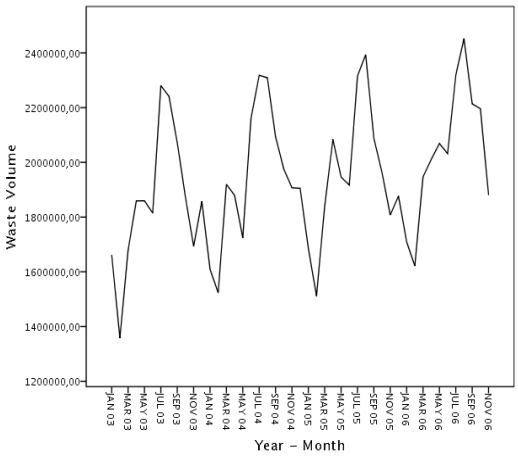
Hirotugu Akaike(1927 – 2009

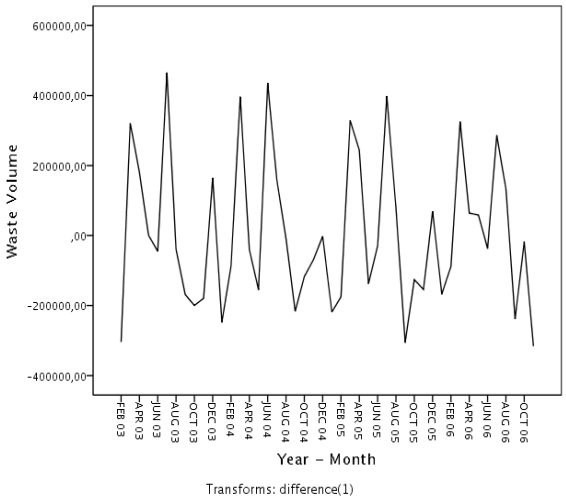
Υπολογίζοντας το AIC για όλα τα πιθανά μοντέλα, επιλέγεται αυτό που έχει τη μικρότερη τιμή σε συνδυασμό με την απλότητά του ορισμού του.

2^{ο} παράδειγμα ARIMA με περιοδικότητα

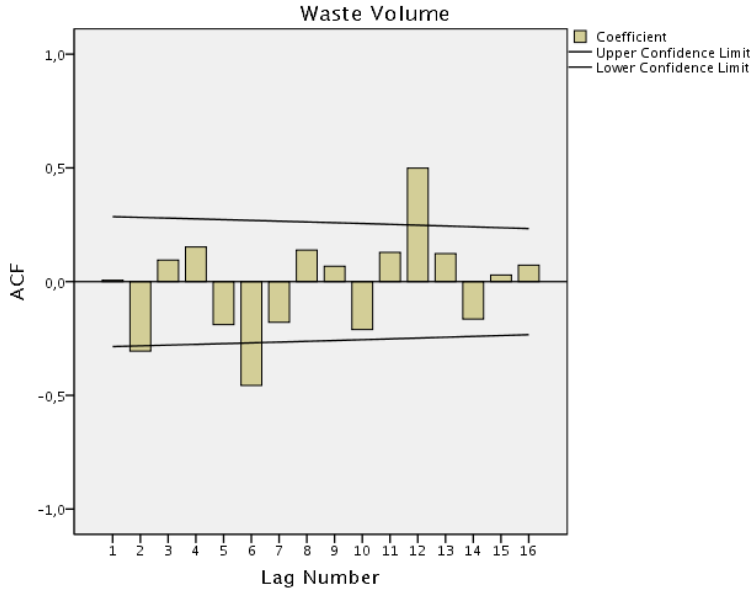
Στο διάγραμμα αναπαριστάται η χρονοσειρά με τιμές τον όγκο απορριμμάτων του Δήμου Κατερίνης τα έτη 2003 έως 2006.

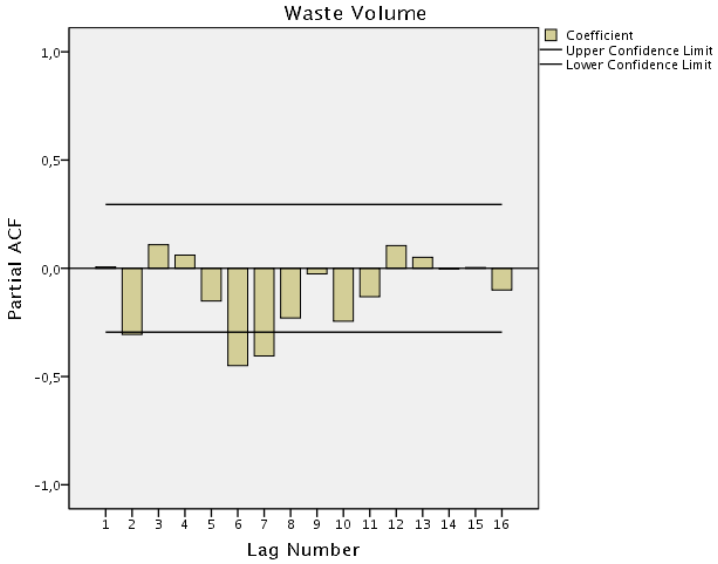
Η σαφής ύπαρξη γραμμικής τάσης υποδεικνύει την αντικατάσταση της χρονοσειράς με τις διαφορές πρώτης τάξης.





Στη χρονοσειρά των διαφορών πρώτης τάξης, παρατηρούμε ότι υπάρχει σημαντική περιοδικότητα στους 12 μήνες.



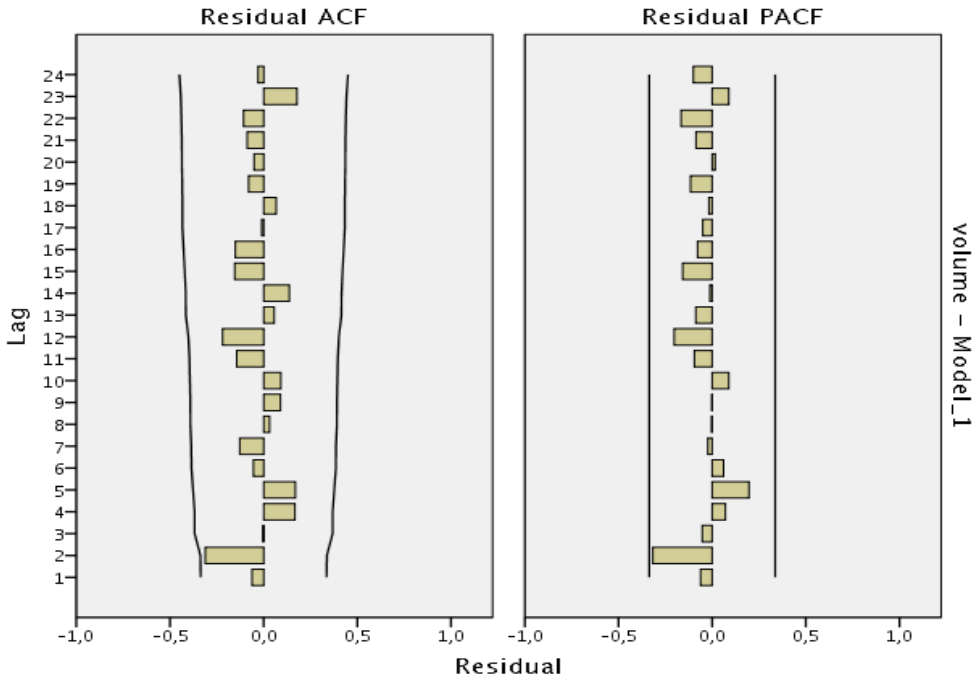


Από τις παραπάνω παρατηρήσεις συνάγεται ότι ένα μοντέλο της μορφής

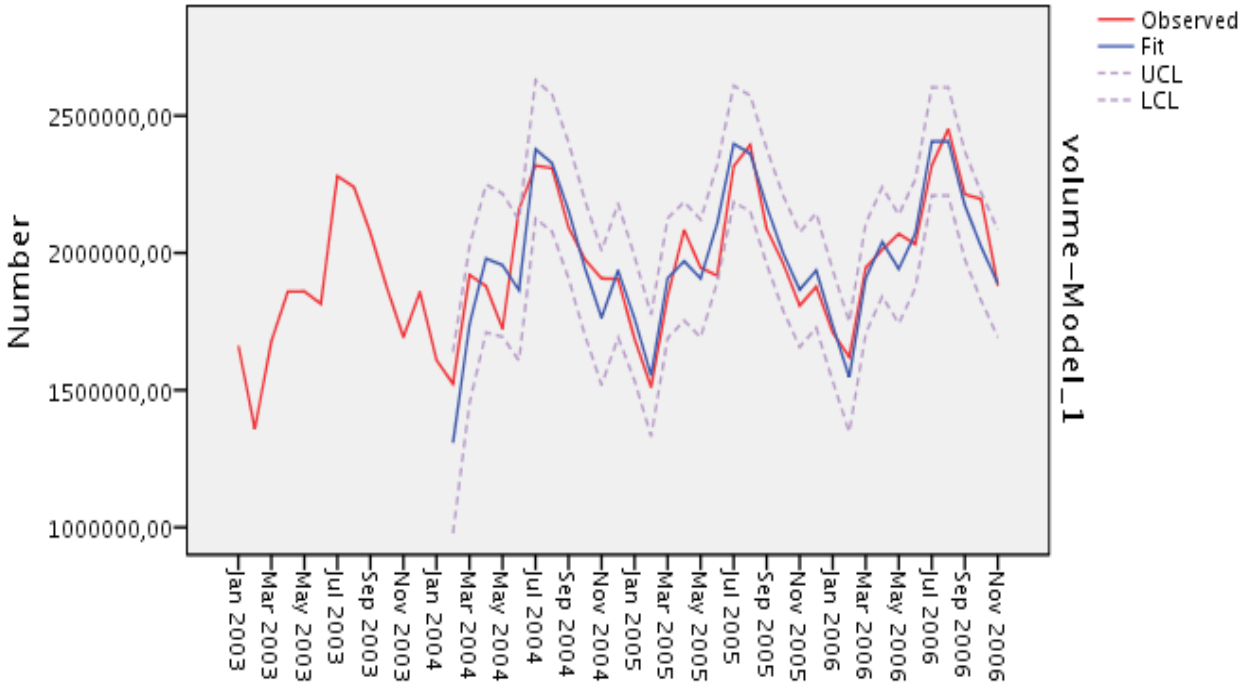
$$\text{ARIMA}(0,1,1)(0,1,1)_{12},$$

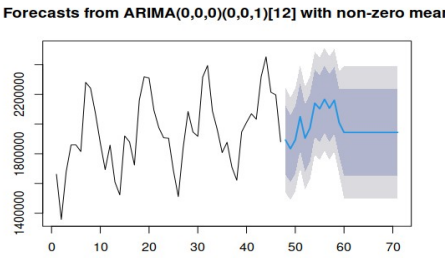
ενδέχεται να προσαρμοστεί ικανοποιητικά στα δεδομένα.

Πράγματι αυτό συμβαίνει καθώς οι αυτοσυσχετίσεις των υπολοίπων του μοντέλου είναι στο σύνολό τους στατιστικά μη διαφορετικές από το 0.

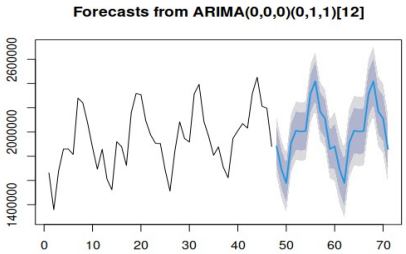


Η γραφική αναπαράσταση του μοντέλου ARIMA(0,1,1)(0,1,1)$_{12}$, όπως και τα αντίστοιχα διαστήματα εμπιστοσύνης αναπαριστώνται στο παρακάτω διάγραμμα.

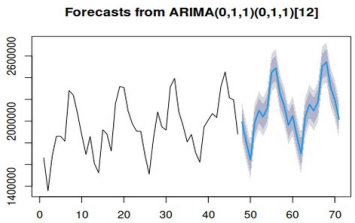




AIC = 1285.3



AIC = 924.9



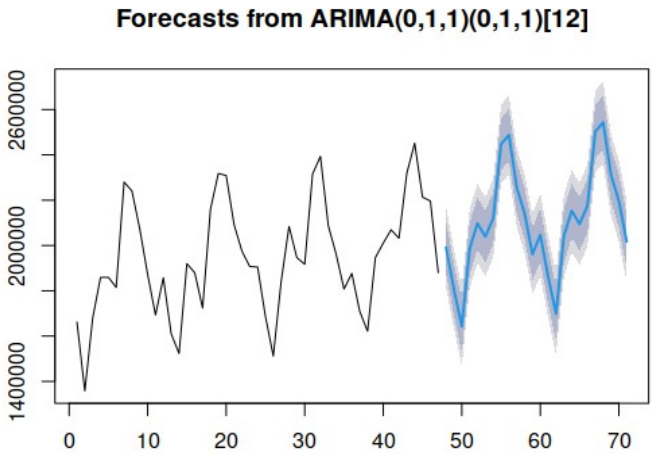
AIC = 887.2

Κώδικας R

```
mymodel1 = arima(volume, order = c(0,0,0), seasonal = list(order = c(0,0,1), period = 12))
plot(forecast(mymodel1))

mymodel2 = arima(volume, order = c(0,0,0), seasonal = list(order = c(0,1,1), period = 12))
plot(forecast(mymodel2))

mymodel3 = arima(volume, order = c(0,1,1), seasonal = list(order = c(0,1,1), period = 12))
plot(forecast(mymodel3))
```

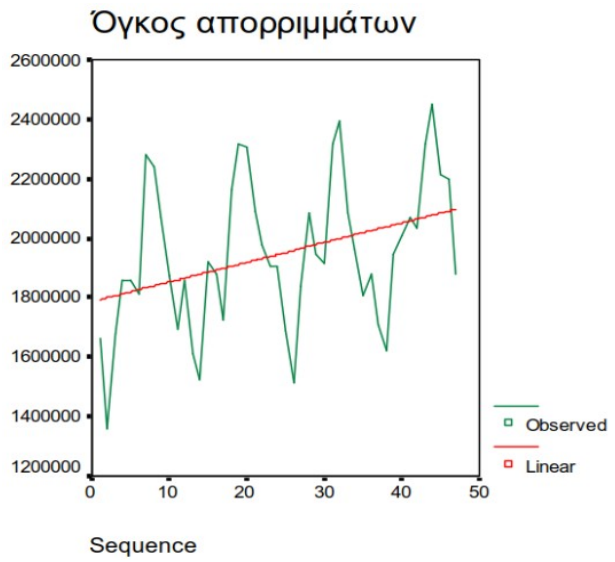


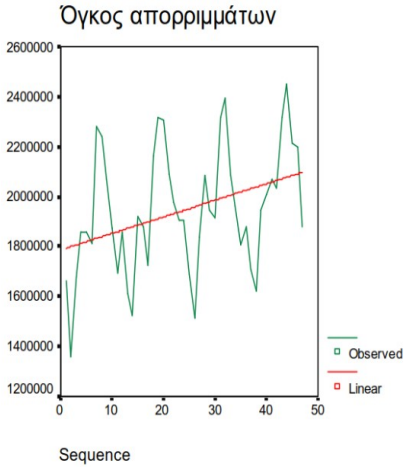
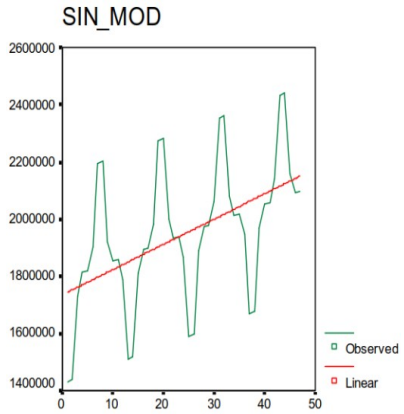
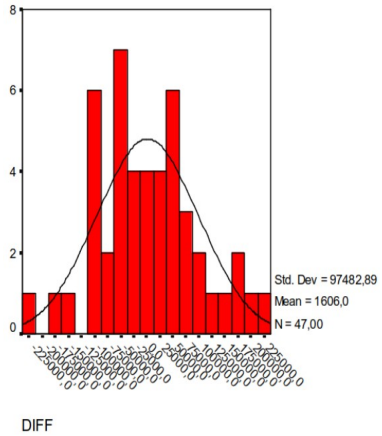
Point	Forecast	80% C.I Lower	80% C.I Upper	95% C.I Lower	95% C.I Upper
48	1991312	1875672	2106951	1814457	2168167
49	1806137	1692318	1919955	1632067	1980207
50	1642804	1528986	1756623	1468734	1816875
51	1985759	1871941	2099578	1811689	2159829
52	2097494	1983676	2211313	1923424	2271564
53	2039019	1925201	2152838	1864949	2213089
54	2120217	2006398	2234035	1946147	2294287

Είναι αξιοσημείωτο το γεγονός πως η συνεπής περιοδικότητα της χρονοσειράς επιτρέπει την προσαρμογή ενός απλού γραμμικού μοντέλου, στο οποίο έχει προστεθεί ένας κατάλληλος ημιτονοειδής παράγοντας.

Στην παρακάτω εξίσωση, ααμ είναι ο αύξων αριθμός του μήνα ξεκινώντας από τον Νοέμβριο του 2003 (όπου ορίζουμε να είναι ααμ = 0), ο αριθμός 430.000 είναι το μέσο ημιπλάτος του κύκλου των πραγματικών δεδομένων και το 3π/4 η χρονική υστέρηση που πρέπει να υπάρχει καθώς ξεκινάει η πρόβλεψη από μήνα Νοέμβριο.

$$M = 1.787.097 + 6.610 \cdot \text{ααμ} + 430.000 \cdot \sin^5\left(\text{ααμ} \cdot \pi/6 - 3\frac{\pi}{4}\right)$$



Το μέσο σφάλμα της εκτίμησης είναι μόλις 1.606 κιλά με τυπική απόκλιση 97.482 κιλά.

Επιπλέον παρατηρούμε ότι το 68% των σφαλμάτων είναι κατά απόλυτη τιμή μικρότερα από 100.000 κιλά, (μέγιστο σφάλμα 5% επί της πραγματικής τιμής) το 17% μεταξύ 100.000 και 150.000 κιλά (αντ. σφάλμα 7,5%)και το υπόλοιπο 15% από 150.000 μέχρι 225.000 (αντ. Σφάλμα 12,5%).