Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ)

Προπτυχιακό Πρόγραμμα Σπουδών

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Πτυχιακή Εργασία

Μοντέλα Χρονοσειρών και Εφαρμογές

Με Χρήση του Στατιστικού Πακέτου R

Επιβλέπον Καθηγητής: Φουσκάκης Δ.

Τριμελής Επιτροπή:

Φουσκάκης Δημήτρης, Αναπληρωτής Καθηγητής ΣΕΜΦΕ, ΕΜΠ

Λουλάκης Μιχαήλ, Αναπληρωτής Καθηγητής ΣΕΜΦΕ, ΕΜΠ

Παπαπαντολέων Αντώνης, Επίκουρος Καθηγητής ΣΕΜΦΕ, ΕΜΠ

Στέφανος Παπαχριστοδούλου
ΑΜ: 09111607
Φεβρουάριος 2018, Αθήνα

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον Δρ. Φουσκάκη Δημήτρη, που με την εμπιστοσύνη που μου έδειξε, μου έδωσε τη δυνατότητα να ασχοληθώ με ένα τόσο επίκαιρο και ενδιαφέρον θέμα, καθώς επίσης και για την ολική καθοδήγηση και βοήθεια που μου προσέφερε.

Θα ήθελα επίσης να ευχαριστήσω τους φίλους και συμφοιτητές μου, που στέκονταν δίπλα μου και με βοηθούσαν σε καθημερινό επίπεδο σε κάθε μου δυσκολία.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και την σύντροφό μου Βαρβάρα, για την αδιάκοπη υποστήριξη τους και για όλες τις θυσίες που έκαναν για εμένα καθόλα τα χρόνια της φοίτησής μου.

Περίληψη

Στην παρούσα διπλωματική, παρουσιάζονται κάποιες βασικές εισαγωγικές έννοιες της Ανάλυσης Χρονοσειρών, καθώς και τα συνηθέστερα μονοδιάστατα μοντέλα που προκύπτουν από μελέτες. Σχολιάζονται και αναλύονται οι διαφορές και οι ομοιότητες των μοντέλων αυτών με απώτερο σκοπό την κατανόηση των δυνατοτήτων τους, για να μπορέσουμε να τα συνδυάσουμε κατάλληλα ώστε να περιγράψουμε ικανοποιητικά τα δεδομένα που μελετάμε. Συγκεκριμένα, θα εστιάσουμε στις μεθόδους εντοπισμού μοντέλων, στις μεθόδους εκτιμήσεων των συντελεστών των εν λόγω μοντέλων και στους διαγνωστικούς ελέγχους για καλή προσαρμογή στα δεδομένα.

Επίσης θα ασχοληθούμε και με το συνηθέστερο δισδιάστατο μοντέλο, το μοντέλο αυτοπαλινδρόμησης κατανεμημένων υστερήσεων (ADL), που με την χρήση μιας επιπλέον επεξηγηματικής μεταβλητής μας δίνει την δυνατότητα να πετύχουμε καλύτερη ακρίβεια αλλά και να δώσουμε καλύτερη και ακριβέστερη ερμηνεία στο πως συνδέονται τα προς μελέτη μεγέθη.

Αφού παρουσιαστεί η απαραίτητη θεωρία, θα προχωρήσουμε σε εφαρμογές των μοντέλων αυτών σε πραγματικά δεδομένα με την χρήση του Στατιστικού Πακέτου R, κατά τις οποίες ο σκοπός μας θα είναι να δημιουργήσουμε κατάλληλα μοντέλα που περιγράφουν επαρκώς τα δεδομένα μας ώστε να τα χρησιμοποιήσουμε για μελλοντικές προβλέψεις.

Στις εφαρμογές αυτές θα γίνει προσπάθεια αναλυτικής και σταδιακής παρουσίασης της μεθοδολογίας που θα ακολουθηθεί, ώστε να επιτευχθεί πλήρης κατανόηση της όλης διαδικασίας κατασκευής των μοντέλων. Στο τέλος κάθε εφαρμογής, θα δίνονται κάποιες προβλέψεις οι οποίες θα αντιπαρατίθενται με τις πραγματικές τιμές που παρατηρήθηκαν, και θα παρουσιάζονται κάποιες ερμηνείες και αποτελέσματα βάση των μοντέλων που επιλέχθηκαν.

Abstract

In the current dissertation, we introduce some basic ideas of Time Series Analysis as well as the most common univariate models. We then proceed to discuss about the distinctions and similarities of those models, to explore their capabilities and by combining them accurately, to be able to produce models that represent our given data as best as possible. Specifically, we will focus in model identification techniques, methods of estimating the coefficients of the studied models and the diagnostic checks to confirm the model's good fit to the data.

Furthermore, the most common bivariate model will be presented and studied, the Autoregressive Distributive Lag Model (ADL), which, with the use of an additional exogenous explanatory variable to the model, enables us to achieve better accuracy but also to give a better and more concrete interpretation of the relationship between the main and the exogenous variable.

Once the necessary theory has been put forward, we will proceed to applications of the mentioned models into real data, using the Statistical Package R. Our main objective is to construct appropriate and well-behaved models that fit the data in order to be able to use them for future predictions.

In those applications, an attempt will be made to analytically present every step of the followed process, to achieve better understanding. At the end of each application we will compare forecasts that occurred from the model versus the actual values that were observed in the future, and in addition, we will include some interpretations based on the resulting model.

Περιεχόμενα

1. Εισαγωγή στην Ανάλυση Χρονοσειρών – Βασική Ιδέα

Πολλές φορές στην καθημερινότητα μας, προκύπτει η ανάγκη να προβλέψουμε την συμπεριφορά μιας μεταβλητής στο μέλλον. Στις περιπτώσεις που η μεταβλητή αυτή έχει ντετερμινιστική συμπεριφορά ή διέπεται από συγκεκριμένους **αυστηρούς νόμους**, που καθιστούν τις προβλέψεις μας αχρείαστες, τότε δεν χρειαζόμαστε την στατιστική. Για παράδειγμα δεν χρειάζεται η χρήση της στατιστικής για να υπολογίσουμε την ταχύτητα πρόσκρουσης ενός βαριδίου που αφήνετε ελεύθερο από συγκεκριμένο ύψος, καθώς μια τέτοια διαδικασία διέπεται από τους ισχυρούς νόμους της Φυσικής και η τυχαιότητα της είναι πρακτικά μηδενική.

Υπάρχουν όμως πολλά προβλήματα τα οποία δεν είναι ντετερμινιστικά και ούτε υπάρχουν αυστηροί νόμοι που τα διέπουν, και οι παράμετροι που τα επηρεάζουν είναι τόσο πολλοί που μπορούν να θεωρηθούν πρακτικά τυχαία γεγονότα. Παρόλα αυτά χρειάζεται με κάποιο τρόπο να διερευνήσουμε και να διερμηνεύσουμε την τυχαιότητα τους, και να δώσουμε κάποιες πιθανές εκτιμήσεις, που περιορίζουν την αβεβαιότητα των προβλημάτων όσο περισσότερο μπορούν.

Τα παραδείγματα τέτοιων προβλημάτων που μπορούμε να φανταστούμε είναι άπειρα, όπως η τιμή μιας μετοχής σε μια βδομάδα, η μέση θερμοκρασία του επόμενου έτους, ο εκτιμώμενος πληθυσμός της γης το 2030, οι πωλήσεις μιας εταιρίας για τον ερχόμενο μήνα και πολλά άλλα. Υπάρχουν δυο τρόποι που μπορούν να δώσουν απαντήσεις σε τέτοιου είδους προβλήματα.

Στην περίπτωση προβλημάτων που δεν υπάρχουν ιστορικά δεδομένα, τότε η στατιστική δεν μπορεί να μας δώσει αποτελέσματα καθώς δεν υπάρχουν πληροφορίες. Σε τέτοιες περιπτώσεις, και δεδομένου το ότι η φύση του προβλήματος είναι τέτοια που το επιτρέπει, χρησιμοποιούμε ποιοτικές προβλέψεις (**Qualitive Forecasting Methods**). Οι ποιοτικές προβλέψεις βασίζονται κυρίως σε κάποιο εμπειρογνώμονα ή ειδικό, που έχει κάποιου είδους εμπειρία σε παρόμοιους τομείς και εκτιμά χωρίς την ανάλυση δεδομένων μια ποσότητα.

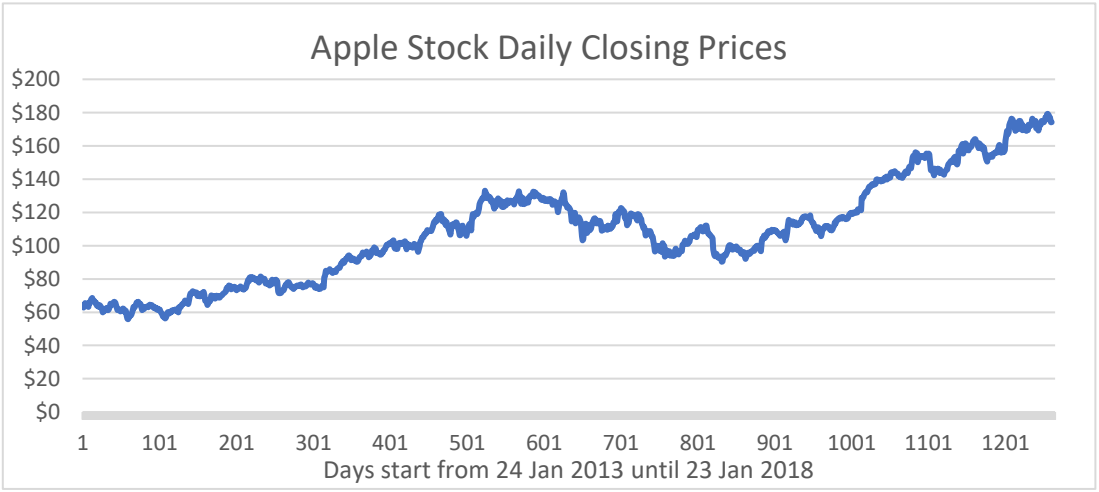
Σε περιπτώσεις όμως που υπάρχουν ιστορικά δεδομένα τότε χρησιμοποιούμε ποσοτικές μεθόδους πρόβλεψης (**Quantitative Forecasting Methods**) που λαμβάνουν υπόψη τα ιστορικά αυτά δεδομένα και με την χρήση της στατιστικής, μας επιτρέπουν να κάνουμε προβλέψεις που αν και μη ακριβής, δρουν με συστηματικότητα και μεθοδολογία που μας δίνει τα καλύτερα δυνατά αποτελέσματα.

Υπάρχουν δυο μέθοδοι ποσοτικών προβλέψεων. Η πρώτη είναι η ανάλυση χρονοσειρών, που προσπαθεί λαμβάνοντας υπόψη μόνο τις προηγούμενες παρατηρήσεις της μεταβλητής που μελετάμε, να αναγνωρίσει μοτίβα και συσχετίσεις και να εξάγει συμπεράσματα για την μελλοντική της συμπεριφορά. Η δεύτερη μέθοδος είναι τα αιτιολογικά μοντέλα (Causal Models) που προσπαθούν να εντοπίσουν και να εκτιμήσουν άλλες παραμέτρους που επηρεάζουν την προς μελέτη μεταβλητή, και βάση αυτής της σχέσης να κάνουν μελλοντικές προβλέψεις.

Αρκετές φορές τα αιτιολογικά μοντέλα, εφόσον προσδιοριστούν σωστά, μας δίνουν επιπρόσθετες πληροφορίες και καλύτερες εκτιμήσεις για την μεταβλητή που μελετάμε. Δυστυχώς όμως, ο προσδιορισμός ενός τέτοιου μοντέλου είναι πολύ δύσκολος και αρκετές φορές ανέφικτος καθώς δεν έχουμε γνώση των τιμών των μεταβλητών που επηρεάζουν την προς μελέτη μεταβλητή.

Τα μονοδιάστατα μοντέλα της ανάλυσης χρονοσειρών είναι πολύ πιο εύκολο να χρησιμοποιηθούν λόγω του ότι δεν χρειαζόμαστε καμία άλλη πληροφορία πέρα από το ιστορικό της μεταβλητής που μελετάμε. Αρκετές φορές επίσης, η γνώση των ιστορικών τιμών είναι αρκετά ικανοποιητική για να μας δώσει πολύ καλές εκτιμήσεις, καθώς υποκρύπτουν σχεδόν όλες τις απαραίτητες πληροφορίες που χρειαζόμαστε.

Μια χρονοσειρά ορίζεται ως μια συλλογή από διακεκριμένες και διαδοχικές παρατηρήσεις, σε τακτά και ίσα διαστήματα, μιας μεταβλητής. Για παράδειγμα, μια σειρά που απαρτίζεται από τις ημερήσιες τιμές στις οποίες έκλεισε η μετοχή της Apple είναι μια χρονοσειρά. Συμβολίζουμε συνήθως τις τιμές μια τέτοιας χρονοσειράς με τον όρο Z_i όπου το i παίρνει τιμές από το 1 μέχρι το T. Δηλαδή η τιμή του Z_3 είναι η τιμή της μεταβλητής Z που μελετάμε, την χρονική περίοδο 3. Στο Διάγραμμα 1.1 παριστάνεται γραφικά μια χρονοσειρά των τιμών Z_i ως προς την χρονική περίοδο παρατήρησης i.



Διάγραμμα 1.1

Μια γενική πρώτη ιδέα για να μελετήσουμε προβλήματα χρονοσειρών, είναι η διάσπαση της χρονοσειράς (Time Series Decomposition) σε όρους που είναι πιο εύκολο να ερμηνευθούν. Για παράδειγμα, μια πρώτη απλή διάσπαση μιας χρονοσειράς με την χρήση του κλασσικού προσθετικού μοντέλου, γίνεται με την ακόλουθη εξίσωση

$$Z_i = T_i + S_i + C_i + I_i\ ,\ \ i = 1,2. \ . \ . \ T\ ,$$

όπου με $\mathbf{T_i}$ συμβολίζουμε την τάση (**Trend**) των τιμών σε διάφορα σημεία (δηλαδή αν η σειρά τείνει να αυξάνεται, όπου το T_i θα είναι θετικό, ή να μειώνεται, με το T_i αρνητικό, την χρονική στιγμή i), $\mathbf{S_i}$ συμβολίζουμε την εποχικότητα (**Seasonality**) της τιμής κατά την περίοδο i (για παράδειγμα αν εκτιμούσαμε τον τουρισμό στην Ελλάδα θα αναμέναμε πέρα της γενικής ανοδικής ή πτωτικής τάσης να υπάρχει μεγάλη αύξηση κατά τους καλοκαιρινούς μήνες και πτώση κατά τους χειμερινούς), με τον όρο $\mathbf{C_i}$ συμβολίζουμε ένα κυκλικό παράγοντα (**Cycle**) και $\mathbf{I_i}$ μια τυχαία διακύμανση (**Irregular Component**) που δεν μπορεί να προβλεφθεί. Εφόσον οι πρώτες τρεις παράμετροι μπορούν να εκτιμηθούν και να υπολογισθούν, η αβεβαιότητα της μεταβλητής Z οφείλεται μόνο στον τελευταίο τυχαίο παράγοντα.

Σκοπός μιας τέτοιας διάσπασης είναι να εκτιμηθούν οι πιο πάνω όροι ξεχωριστά (με εξαίρεση του τυχαίου) και έπειτα να προστεθούν εκ νέου για να σχηματίσουν την σειρά που μελετάμε, και βάση των εκτιμήσεων των όρων αυτών, να εκτιμήσουμε πιθανές μελλοντικές τιμές της χρονοσειράς Y_t για χρόνους μετά το T (μέχρι το οποίο θεωρούμε τις τιμές γνωστές).

Η ιδέα αυτής της διάσπασης έγκειται στο γεγονός ότι αν αφαιρέσουμε από την σειρά τους πιο πάνω όρους τους οποίους μπορώ να εκτιμήσω, και καταλήξω σε μια τυχαία διαδικασία, τότε έχω αποσπάσει όλες τις πληροφορίες που μου παρείχαν οι ιστορικές τιμές και το μόνο που έχω αφήσει πίσω είναι ο λευκός θόρυβος (που εξορισμού δεν μπορεί να εκτιμηθεί).

Μπορούμε με πολλές μεθόδους να κάνουμε εκτιμήσεις για τους πιο πάνω παράγοντες του διασπασμένου μοντέλου, και σε περίπτωση που το αποτέλεσμα δεν μας ικανοποιεί, μπορούμε να χρησιμοποιήσουμε διάφορες άλλες παρόμοιες μορφές, όπως το απλό πολλαπλασιαστικό μοντέλο και άλλα.

Μια τακτική που χρησιμοποιούμε στην ανάλυση χρονοσειρών, ειδικά όταν έχουμε μεγάλο πλήθος ιστορικών δεδομένων, είναι να χωρίζουμε τα δεδομένα μας σε δυο μέρη. Το πρώτο μέρος να αποτελείτε από τις T-N πρώτες παρατηρήσεις και το δεύτερο μέρος να είναι οι τελευταίες N παρατηρήσεις.

Ακολούθως υπολογίζουμε τους όρους που μας ενδιαφέρουν σύμφωνα με την μέθοδο ελαχίστων τετραγώνων στο πρώτο μέρος, για να καταλήξουμε σε κάποιο μοντέλο που περιγράφει ικανοποιητικά τα δεδομένα μας. Έπειτα κάνουμε τις “προβλέψεις” για τις τελευταίες Ν παρατηρήσεις βάση του μοντέλου που βρήκαμε, και αφού οι πραγματικές τιμές της μεταβλητής είναι γνωστές, συγκρίνουμε τις εκτιμήσεις με τα πραγματικά δεδομένα για να δούμε αν οι προβλέψεις μας ήταν έγκυρες. Δεδομένου ότι οι προβλέψεις μας ήταν ικανοποιητικές, επαναπροσαρμόζουμε το μοντέλο σε όλα τα δεδομένα και χρησιμοποιούμε το νέο μοντέλο για να κάνουμε τις μελλοντικές εκτιμήσεις μας.

Φυσικά τα πιο πάνω αποτελούν μια απλούστευση μιας μεγάλης κατηγορίας προβλημάτων που μπορούν πλέον να επιλυθούν με μεγάλη ταχύτητα και ακρίβεια με την χρήση αρκετά πιο πολύπλοκων μοντέλων, μέσω σύγχρονων στατιστικών πακέτων όπως την R, Stata, SPSS, eViews και πολλά άλλα.

Στην παρούσα διπλωματική θα ασχοληθούμε με 5 διαφορετικά προβλήματα πραγματικών δεδομένων, από διαφορετικούς τομείς, με σκοπό την ανάδειξη της πληθώρας των εφαρμογών που βρίσκει η ανάλυση χρονοσειρών σήμερα, καθώς επίσης και την επίδειξη της διαδικασίας που ακολουθείτε, από την αρχή μέχρι το τέλος, δηλαδή από τα ωμά δεδομένα στις μελλοντικές προβλέψεις μεταβλητών.

2. Βασικά Μονοδιάστατα Μοντέλα

2.1 Βασικοί Τελεστές

Στα παρακάτω μοντέλα θα γίνει χρήση κάποιων βασικών τελεστών. Κυρίως θα ασχοληθούμε με τον τελεστή οπίσθιας μετάθεσης **(backward shift operator)**, που συμβολίζουμε με **B,** τον τελεστή εμπρόσθιας μετάθεσης **(forward shift operator)**, που συμβολίζουμε με **F**, και τέλος τον τελεστή οπίσθιας διαφοράς **(backward difference operator)**, με σύμβολο **Δ**.

Ορίζουμε τον τελεστή **B** ώστε $\mathbf{B}Z_t = Z_{t-1}$ και συνεπώς $\mathbf{B}^pZ_t = Z_{t-p}$.

Ο τελεστής **F** είναι ο ανάστροφος του **B**, δηλαδή $\mathbf{F}Z_t = Z_{t+1}$ και $\mathbf{F}^pZ_t = Z_{t+p}$.

Τέλος ο τελεστής **Δ** μας δίνει την διαφορά μεταξύ των τιμών δύο διαδοχικών παρατηρήσεων. Δηλαδή $\mathbf{\Delta}Z_t = Z_t - Z_{t-1} = (1 - \mathbf{B})Z_t$

2.2 Μοντέλο γραμμικού Φίλτρου

Μια χρονοσειρά Z_t της οποίας οι διαδοχικές της τιμές φανερώνουν μεγάλη εξάρτηση μεταξύ τους, μπορεί να θεωρηθεί ότι απλά παράγεται από το άθροισμα κάποιων τυχαίων διακυμάνσεων α_t (random shocks). Οι διακυμάνσεις αυτές προέρχονται από μια κανονική κατανομή με μέση τιμή 0 και διασπορά σ_α^2. Μια τέτοια ακολουθία από ανεξάρτητες τυχαίες μεταβλητές $\alpha_t, \alpha_{t-1}, \alpha_{t-2}, \ldots$ καλείτε διαδικασία λευκού θορύβου.

Ένα γραμμικό φίλτρο μετατρέπει μια διαδικασία λευκού θορύβου σε μια χρονοσειρά Z_t. Ουσιαστικά εκφράζουμε την τιμή του Z_t σαν ένα σταθμικό άθροισμα των προηγούμενων διακυμάνσεων. $Z_t = \mu + \alpha_t + \psi_1\alpha_{t-1} + \psi_2\alpha_{t-2} + \ldots = \mu + \psi(\boldsymbol{B})\alpha_t$

$$\text{όπου } \psi(\boldsymbol{B}) = 1 + \psi_1\boldsymbol{B} + \psi_2\boldsymbol{B}^2 + \ldots$$

Το μ είναι το “επίπεδο” στο οποίο κυμαίνεται η χρονοσειρά (αν η χρονοσειρά είναι στάσιμη τότε το μ ορίζει τον μέσο, γύρω από το οποίο οι τιμές πάλλονται) ενώ το $\psi(\boldsymbol{B})\alpha_t$ είναι ο γραμμικός τελεστής που μετατρέπει την διαδικασία α_t στη χρονοσειρά Z_t. Ο τελεστής αυτός καλείται συνάρτηση μεταφοράς (transfer function) του φίλτρου.

Η ακολουθία $\psi_1, \psi_2, \ldots$ που ορίζεται από τα “βάρη” του αθροίσματος, μπορεί να είναι πεπερασμένη ή άπειρη. Λέμε ότι το φίλτρο είναι σταθερό (**stable**) και η χρονοσειρά στάσιμη (**stationary**) αν η ακολουθία είναι πεπερασμένη, είτε άπειρη αλλά το άθροισμα των στοιχείων της συγκλίνει κατά απόλυτη τιμή, δηλαδή $\sum_{j=0}^{\infty}|\psi_j| < +\infty$.

2.3 Μοντέλο Αυτοπαλινδρόμησης

Ένα από τα βασικότερα μοντέλα που θα μελετηθεί είναι το μοντέλο Αυτοπαλινδρόμησης (**Auto-Regressive model - AR)**. Το μοντέλο αυτό συσχετίζει την τρέχουσα τιμή της μεταβλητής που μελετάται, με τις τιμές σε προηγούμενους χρόνους της ίδιας μεταβλητής. Συγκεκριμένα ένα AR μοντέλο τάξης p, συνδέει την τρέχουσα τιμή, με τις p προηγούμενες παρατηρήσεις. Ένας απλός τύπος που περιγράφει την σχέση αυτή είναι ο ακόλουθος:

$$Y_t = \lambda_0 + \lambda_1 Y_{t-1} + \lambda_2 Y_{t-2} + \ldots + \lambda_p Y_{t-p} + a_t \quad (2.1)\,,$$

όπου το λ_i είναι ο συντελεστής που εκφράζει το πόσο επηρεάζει η τιμή της παραμέτρου σε χρόνο i στο παρελθόν την τρέχουσα παρατήρηση, ενώ το α_t είναι μια τυχαία διακύμανση (random shock) που δεν μπορεί να εκτιμηθεί (σφάλμα). Για να είναι το μοντέλο μας κατάλληλο να περιγράψει επαρκώς την χρονοσειρά, κάνουμε την υπόθεση ότι τα σφάλματα είναι **IID (Independent and Identically Distributed)** δηλαδή ανεξάρτητα και ισόνομα κατανεμημένα, συγκεκριμένα ζητάμε να έχουν μέση τιμή 0 και σταθερή διασπορά ανεξάρτητη του t. Το λ_0 είναι η τιμή του Y_t αν δεν υπήρχε επιρροή από τις προηγούμενες παρατηρήσεις επομένως μπορεί να θεωρηθεί σαν ο μέσος όρος μ της χρονοσειράς.

Εύκολα μπορούμε να δούμε ότι η εκτίμηση για την τιμή του Y την χρονική στιγμή t δίνεται από την σχέση:

$$\widehat{Y_t} = E[Y_t] = E[\lambda_0 + \lambda_1 Y_{t-1} + \lambda_2 Y_{t-2} + \ldots + \lambda_p Y_{t-p} + a_t]$$

$$\widehat{Y_t} = E[\lambda_0] + \lambda_1 E[Y_{t-1}] + \lambda_2 E[Y_{t-2}] + \ldots + \lambda_p E[Y_{t-p}] + E[a_t]$$

$$\widehat{Y_t} = \lambda_0 + \lambda_1 \widehat{Y}_{t-1} + \lambda_2 \widehat{Y}_{t-2} + \ldots + \lambda_p \widehat{Y}_{t-p}.$$

Συνεπώς, αν έχουμε σαν δεδομένα τις αρχικές p τιμές της μεταβλητής Y, μπορούμε να εκτιμήσουμε την αμέσως επόμενη, και χρησιμοποιώντας αυτή να εκτιμήσουμε και την επόμενη και ούτω καθεξής. Μπορούμε δηλαδή, θεωρητικά, να εκτιμήσουμε την τιμή της Y για κάθε μελλοντική χρονική στιγμή. Φυσικά οι τυπικές αποκλίσεις των εκτιμήσεων θα αυξάνονται εκθετικά πράγμα που καθιστά μακρινές προβλέψεις αναξιόπιστες.

Αφού το λ_0 μπορεί να θεωρηθεί ο μέσος όρος της σειράς, μπορούμε να θεωρήσουμε μια νέα χρονοσειρά Z_t που ισούται με την Y_t – μ, δηλαδή εκφράζει τις αποκλίσεις των παρατηρήσεων από το μέσο. Μπορούμε να φέρουμε την μορφή της εξίσωσης 2.1 στην ισοδύναμη της:

$$Z_t = \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \ldots + \varphi_p Z_{t-p} + a_t \quad (2.2).$$

Χρησιμοποιώντας τον τελεστή **B** μπορούμε να φέρουμε τη μορφή της εξίσωσης 2.2 στην πιο κάτω:

$$Z_t = \varphi_1 \boldsymbol{B} Z_t + \varphi_2 \boldsymbol{B}^2 Z_t + \ldots + \varphi_p \boldsymbol{B}^p Z_t + a_t = \sum_{i=1}^{p} \varphi_i \boldsymbol{B}^i Z_t + a_t \ (2.3).$$

Φέρνοντας τους όρους που περιλαμβάνουν το Z_t αριστερά και βγάζοντας κοινό παράγοντα, μπορούμε να ορίσουμε το **AR πολυώνυμο**:

$$\varphi(\boldsymbol{B}) = 1 - \varphi_1 \boldsymbol{B} - \varphi_2 \boldsymbol{B}^2 - \ldots - \varphi_p B^p \ (2.4) .$$

Έτσι η εξίσωση μας μπορεί πλέον να γραφεί οικονομικά ως:

$$\varphi(\boldsymbol{B}) Z_t = a_t \ (2.5) .$$

Αυτό το μοντέλο περιλαμβάνει p+2 άγνωστες παραμέτρους, το μ, τα φ_i και το σ_α^2 (η διασπορά της διαδικασίας λευκού θορύβου α_t). Βασικός μας σκοπός είναι να εκτιμήσουμε τις τιμές των παραμέτρων, βάση των δεδομένων μας, ώστε να καθορίσουμε το μοντέλο.

Εύκολα μπορούμε να δούμε ότι το **AR** μοντέλο είναι μια ειδική περίπτωση του μοντέλου γραμμικού φίλτρου. Αν αντικαταστήσουμε το

$$Z_{t-1} = \varphi_1 Z_{t-2} + \varphi_2 Z_{t-3} + \ldots + \varphi_p Z_{t-p-1} + a_{t-1}$$

στην εξίσωση 2.2 και συνεχίζοντας διαδοχικά και κάνοντας m τέτοιες αντικαταστάσεις καταλήγουμε σε μια άπειρη σειρά που περιλαμβάνει της τυχαίες διακυμάνσεις α.

Για παράδειγμα ας εστιάσουμε στο απλό **AR(1)** μοντέλο $Z_t = \varphi Z_{t-1} + a_t$.

Κάνοντας m αντικαταστάσεις όπως πιο πάνω, καταλήγουμε:

$$Z_t = \varphi^{m+1} Z_{t-m-1} + a_t + \varphi a_{t-1} + \varphi^2 a_{t-2} + \ldots + \varphi^m a_{t-m} .$$

Παίρνοντας το όριο για m να τείνει στο άπειρο, καταλήγουμε στο ότι $Z_t = \sum_{j=0}^{\infty} \varphi^j a_{t-j}$ όπου είναι το μοντέλο γραμμικού φίλτρου με βάρη $\psi_j = \varphi^j$. Η σειρά αυτή συγκλίνει αν και μόνο αν $|\varphi| < 1$.

Στο γενικό **AR** μοντέλο μπορούμε συμβολικά, με ίδιες διαδικασίες να καταλήξουμε στην πιο κάτω εξίσωση αντιστρέφοντας τον γραμμικό τελεστή στην σχέση 2.5 έχουμε:

$$Z_t = \varphi^{-1}(\boldsymbol{B}) a_t = \psi(\boldsymbol{B}) a_t \ \ \text{όπου} \ \psi(\boldsymbol{B}) = \sum_{j=0}^{\infty} \psi_j \boldsymbol{B}^j .$$

Μια Auto Regressive διαδικασία μπορεί να είναι είτε στάσιμη είτε μη στάσιμη. Για να είναι στάσιμη η διαδικασία πρέπει οι συντελεστές φ_i να είναι τέτοιοι ώστε τα βάρη ψ_i να σχηματίζουν μια συγκλίνουσα σειρά.

Αποδεικνύετε ότι αυτό ισχύει όταν το AR πολυώνυμο (βαθμού p) έχει τις ρίζες του μεγαλύτερες από την μονάδα κατά απόλυτη τιμή (οι ρίζες μπορεί να είναι μιγαδικές), δηλαδή να βρίσκονται εκτός του μοναδιαίου κύκλου. Αυτό μας επιβεβαιώνει το αποτέλεσμα που βρήκαμε στο AR (1) μοντέλο πιο πάνω, δηλαδή η διαδικασία είναι στάσιμη για $|\varphi| < 1$.

2.4 Μοντέλα Κινούμενου Μέσου

Όπως είδαμε στα προηγούμενα, το AR μοντέλο εκφράζει την απόκλιση από τον μέσο της μεταβλητής Z_t της διαδικασίας, σαν ένα πεπερασμένο σταθμισμένο άθροισμα των p προηγούμενων αποκλίσεων, συν μια τυχαία διακύμανση α_t. Ισοδύναμα όμως μπορούμε να το μετατρέψουμε ώστε το Z_t να εκφράζεται από ένα άπειρο σταθμισμένο άθροισμα των διακυμάνσεων α_t, α_{t-1}, . . .

Στα Μοντέλα Κινούμενου Μέσου **(Moving Average - MA)** τάξης q, λαμβάνουμε υπόψη όχι τις προηγούμενες τιμές των παρατηρήσεων, αλλά τις q προηγούμενες διακυμάνσεις (σφάλματα). Δηλαδή εκτιμάμε την τιμή βάση των προηγούμενων q σφαλμάτων του μοντέλου. Συγκεκριμένα ισχύει:

$$Y_t = \mu + \ \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \quad (2.6).$$

Σημειώνουμε ότι θα μπορούσαμε να συμβολίσουμε το πιο πάνω μοντέλο σαν διαφορά και όχι σαν άθροισμα, με μόνη διαφορά το ότι οι συντελεστές θα ήταν αντίθετοι.

Η αναμενόμενη τιμή της τιμής Y σε χρόνο t είναι:

$$\widehat{Y_t} = E[Y_t] = E[\mu + \ \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}]$$

$$\widehat{Y_t} = E[\mu] - \ E[\varepsilon_t] - \theta_1 E[\varepsilon_{t-1}] - \theta_2 E[\varepsilon_{t-2}] - \ldots - \theta_q E\left[\varepsilon_{t-q}\right] = \mu \,,$$

μιας και η αναμενόμενη τιμή του ε_t είναι 0 για κάθε t (λόγω του ότι οι διακυμάνσεις είναι IID με μέσο 0 και διασπορά σ_ε^2. Σε αυτό το μοντέλο θεωρούμε ότι οι τιμές των ε_0, ε_{-1}, ε_{-2}, . . . ε_{-q+1} είναι ίσες με 0. Συνεπώς βλέπουμε ότι σε αντίθεση με τα AR μοντέλα, τα MA μοντέλα τάξης q, όπου q πεπερασμένος αριθμός, είναι εκ κατασκευής στάσιμα.

Η διασπορά του Y_t είναι:

$$Var[Y_t] = E[(Y_t - \mu)^2] = E[(\ \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q})^2]$$

$$Var[Y_t] = E\left[{\varepsilon_t}^2 + \sum_{i=1}^{q} \theta_i^2 \varepsilon_{t-i}^2 + \sum_{i=1}^{q} \sum_{j=1}^{q} \theta_i \, \theta_j \varepsilon_{t-i} \varepsilon_{t-q}\right].$$

Όμως λόγω του ότι τα ε_t είναι ανεξάρτητα έχουμε ότι η συνδιασπορά:

$$Cov[\varepsilon_i, \varepsilon_j] = \delta_{ij}\sigma_\varepsilon^2 \quad \text{δηλαδή } \mathrm{E}[\varepsilon_i . \varepsilon_j] = 0 \text{ για } i \neq j,$$

$$\text{όπου } \delta_{ij} \text{ το } \textbf{Kronecker } \boldsymbol{delta}, \text{δηλαδή } \delta_{ij} = \begin{cases} 1\,, & \text{όταν } i = j \\ 0\,, & \text{όταν } i \neq j \end{cases}.$$

Γνωρίζοντας ότι οι διασπορά του ε_i είναι εξ ορισμού σ_ε^2 καταλήγουμε:

$$Var[Y_t] = \left(1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2\right)\sigma_\varepsilon^2.$$

Βλέπουμε ότι η διασπορά του Y_t είναι σταθερή και ανεξάρτητη της χρονικής περιόδου του t.

Στην ειδική περίπτωση του MA μοντέλου άπειρης τάξης, είναι στάσιμο αν και μόνο αν το άπειρο άθροισμα των τετραγώνων τον συντελεστών συγκλίνει, δηλαδή:

$$1 + \theta_1^2 + \theta_2^2 + \ldots = \sum_{i=1}^{\infty} \theta_i^2 < +\infty\,.$$

Έχοντας στα χέρια μας τις αρχικές T παρατηρήσεις (T > q) και γνωρίζοντας τις τιμές των συντελεστών και του μέσου, μπορούμε να προβλέψουμε βάση του μοντέλου την τιμή της τυχαίας μεταβλητής Y_{T+1}. Μπορούμε να κάνουμε την διαδικασία αυτή μέχρι και q περιόδους στο μέλλον καθώς για μελλοντικές παρατηρήσεις δεν μπορούμε να εκτιμήσουμε το ε_{T+1} ούτε και τα υπόλοιπα γιατί δεν έχουμε την ακριβή τιμή της Y_{T+1}, αλλά ούτε και των επόμενων περιόδων. Δηλαδή όλες οι προβλέψεις για μεταγενέστερους χρόνους θα είναι ίσες με το μέσο μ.

Όπως και στην περίπτωση των AR μοντέλων, μπορούμε να μετατρέψουμε την μεταβλητή Y_t στην μεταβλητή Z_t αφαιρώντας από τις μετρήσεις τον μέσο της διαδικασίας. Δηλαδή η τιμές της Z_t εκφράζουν τις αποκλίσεις από τον μέσο όρο. Μπορούμε δηλαδή να γράψουμε την εξίσωση που περιγράφει το μοντέλο ως εξής:

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \quad (2.7).$$

Χρησιμοποιώντας και πάλι τον τελεστή οπίσθιας μετατόπισης μπορούμε να γράψουμε το μοντέλο μας οικονομικά σαν:

$$Z_t = \theta(\boldsymbol{B}) a_t\,,$$

Όπου με θ(**B**) συμβολίζουμε το πολυώνυμο τάξης q που έχει την μορφή:

$$\theta(\mathbf{B}) = 1 - \theta_1 \boldsymbol{B} - \theta_2 \boldsymbol{B}^2 - \ldots - \theta_q \boldsymbol{B}^q\,.$$

2.5 Συνδυασμένα Μοντέλα

Μπορούμε να συνδυάσουμε τα AR και MA μοντέλα για να πετύχουμε καλύτερη προσαρμογή στα δεδομένα μας. Τα μοντέλα **ARMA** περιέχουν όρους της μορφής $\varphi_i Z_{t-i}$ αλλά και $\theta_i a_{t-i}$. Ένα μοντέλο ARMA με p Auto Regressive όρους και q Moving Average όρους συμβολίζεται σαν ARMA(p,q) και έχει την μορφή:

$$Z_t = \varphi_1 Z_{t-1} + \ldots + \varphi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \quad (2.8).$$

Σύμφωνα με προηγούμενους συμβολισμούς μπορεί να γραφεί στην πολύ σύντομη μορφή:

$$\varphi(\boldsymbol{B}) Z_t = \theta(\boldsymbol{B}) a_t \quad (2.9).$$

Μπορούμε επίσης να αντιστρέψουμε το φ(**B**) και να το φέρουμε στην μορφή του γραμμικού φίλτρου:

$$Z_t = \varphi^{-1}(\boldsymbol{B})\theta(\boldsymbol{B}) a_t = \psi(\boldsymbol{B}) a_t \quad \mu\varepsilon\ \psi(\boldsymbol{B}) = \varphi^{-1}(\boldsymbol{B})\theta(\boldsymbol{B}).$$

Το μοντέλο ARMA (τάξης p,q) περιλαμβάνει p+q+2 άγνωστες παραμέτρους που πρέπει να εκτιμηθούν (p+q συντελεστές, το μέσο μ και την διασπορά των διακυμάνσεων σ_α^2). Συνήθως όμως, τα μοντέλα ARMA όπου τα p και q είναι μικρότερα ή ίσα του 3, είναι αρκετά για να περιγράψουν ικανοποιητικά μια πληθώρα από προβλήματα στα οποία οι μεταβλητή Z_t παρουσιάζει στασιμότητα.

Σε περιπτώσεις όμως που τα δεδομένα μας δεν φαίνονται να παραμένουν στάσιμα, δηλαδή υπάρχουν μεγάλες διακυμάνσεις των τιμών, τα μοντέλα ARMA δεν είναι κατάλληλα. Σε αυτές τις περιπτώσεις μια πολύ συνηθισμένη μέθοδος είναι να μετατρέψουμε κατάλληλα την χρονοσειρά ώστε να γίνει στάσιμη και να χρησιμοποιήσουμε μοντέλα ARMA.

2.6 Integrated Μοντέλα

Είναι αρκετά συνηθισμένο σε εφαρμογές της ανάλυσης χρονοσειρών, μια διαδικασία Y_t να μην είναι στάσιμη, δηλαδή να μην περιστρέφεται γύρω από κάποιο σταθερό μέσο μ, αλλά παρόλα αυτά να φανερώνει μια ομοιογένεια σχετικά με τον χρόνο t. Σε τέτοιες περιπτώσεις μπορούμε να ορίσουμε μια χρονοσειρά με τις διαφορές διαδοχικών παρατηρήσεων (είτε ποσοστιαίες διαφορές), περιορίζοντας έτσι την μεταβλητότητα της νέας διαδικασίας Z_t. Αν και η νέα διαδικασία Z_t είναι μη-στάσιμη τότε μπορούμε να ορίσουμε μια άλλη διαδικασία που να παίρνει διαδοχικές διαφορές των Z_t (δηλαδή οι διαφορές των διαφορών της αρχικής διαδικασίας Y_t).

Μπορούμε να συνεχίσουμε επαγωγικά με την ίδια μέθοδο μέχρι να καταλήξουμε σε στάσιμο μοντέλο. Συγκεκριμένα ένα Integrated μοντέλο τάξης d κάνει την πιο πάνω διαδικασία d-φορές. Για d=1 έχουμε,

$$Z_t = Y_t - Y_{t-1} = Y_t - \boldsymbol{B}Y_t = (1 - \boldsymbol{B})Y_t = \boldsymbol{\Delta}Y_t \,,$$

ενώ για d=2,

$$Z_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = (1 - 2\boldsymbol{B} + \boldsymbol{B^2})Y_t = \ (1 - \boldsymbol{B})^2 Y_t = \boldsymbol{\Delta}^2 Y_t \,.$$

Είναι εύκολο να δούμε επαγωγικά ότι για κάθε d ισχύει:

$$Z_t = (1 - \boldsymbol{B})^d Y_t = \boldsymbol{\Delta}^d Y_t \,.$$

Το απλό Integrated μοντέλο δηλαδή, περιγράφεται από την εξίσωση:

$$Z_t = (1 - \boldsymbol{B})^d Y_t = \mu + \varepsilon_t \,,$$

όπου το μ είναι το επίπεδο της σταθερής πλέον διαδικασίας Z_t και το ε_t είναι το τυχαίο σφάλμα.

Τα Integrated μοντέλα, αν και αρκετά απλοϊκά, είναι ένα πολύ ισχυρό εργαλείο που σε συνδυασμό με τα ARMA μοντέλα που είδαμε πιο πάνω, μας επιτρέπει να περιγράψουμε με αρκετή ακρίβεια πάρα πολλές χρονοσειρές (στάσιμες και μη) που προκύπτουν από πραγματικά δεδομένα.

2.7 Μοντέλα Auto-Regressive Integrated Moving Average (ARIMA)

Η λογική επέκταση των πιο πάνω μας οδηγεί στην δημιουργία ενός συνδυασμένου μοντέλου που περιλαμβάνει τις ιδέες των 3 μοντέλων που σχολιάστηκαν. Συγκεκριμένα το μοντέλο ARIMA τάξης p,d,q είναι το μοντέλο που παίρνει τις d-διαφορές, χρησιμοποιεί p AR-όρους και q MA-όρους. Αναλυτικά το μοντέλο έχει την μορφή:

$$Z_t = \varphi_1 Z_{t-1} + \ldots + \varphi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \ \ (\mu\varepsilon\ Z_t = \boldsymbol{\Delta}^d Y_t).$$

Σύμφωνα με προηγούμενους συμβολισμούς μπορούμε να συνοψίσουμε την πιο πάνω εξίσωση στην πολύ πιο απλή μορφή:

$$\varphi(\boldsymbol{B})Z_t = \theta(\boldsymbol{B})a_t \ \ ,\text{όπου } Z_t = (1 - \boldsymbol{B})^d Y_t \,.$$

Μπορούμε σε αυτό το σημείο να δούμε ότι γράφοντας την πιο πάνω σχέση χρησιμοποιώντας την αρχική χρονοσειρά Y_t μπορούμε να γράψουμε το μοντέλο σαν:

$$\Phi(\boldsymbol{B})Y_t = \varphi(\boldsymbol{B})(1 - \boldsymbol{B})^d Y_t = \theta(\boldsymbol{B})a_t \,.$$

Μπορούμε να δούμε δηλαδή ότι υπάρχει μια ισοδυναμία μεταξύ τον μοντέλων ARIMA με τα μοντέλα ARMA όπου το πολυώνυμο Φ(**B**) έχει ακριβώς d ρίζες ίσες με την μονάδα (που εμπίπτουν στην περιφέρεια του μοναδιαίου κύκλου). Όπως είχαμε δει αν όλες οι ρίζες ενός AR πολυωνύμου βρίσκονται εξωτερικά του μοναδιαίου κύκλου, τότε έχουμε μια στάσιμη διαδικασία, ενώ αν έστω και μια βρίσκεται στο εσωτερικό τότε έχουμε το μοντέλο να εκρήγνυται προς το άπειρο. Στην πολύ ειδική περίπτωση που έχουμε ρίζες ακριβώς ίσες με την μονάδα τότε τα μοντέλα ARMA συμπίπτουν με τα ARIMA.

2.8 Σύνοψη

Στο Πίνακα 2.1, συνοψίζονται κάποια γενικά συμπεράσματα για τα μοντέλα που σχολιάστηκαν στην ενότητα αυτή. Συγκεκριμένα σχολιάζεται η στασιμότητα των μοντέλων αλλά και οι ισοδυναμίες κάποιων ειδικών περιπτώσεών τους.

Μοντέλο	Συμπεράσματα
AR(p)	Στάσιμο αν οι ρίζες του AR πολυωνύμου βρίσκονται εκτός του μοναδιαίου κύκλου. Το AR(1) μοντέλο ονομάζεται μοντέλο Markov και όταν ο συντελεστής του είναι ίσος με την μονάδα τότε ονομάζεται τυχαίος περίπατος και συμπίπτει με το μοντέλο Integrated(1).
MA(q)	Εκ κατασκευής στάσιμο.
ARMA(p,q)	Στάσιμο αν οι ρίζες του AR πολυωνύμου βρίσκονται εκτός του μοναδιαίου κύκλου.
Integrated(d)	Εκ κατασκευής μη στάσιμο για d≠0. Είναι ειδικές περιπτώσεις των μη στάσιμων AR μοντέλων με όλες τις ρίζες τους να ανήκουν στην περιφέρεια του μοναδιαίου κύκλου.
ARIMA(p,d,q)	Εκ κατασκευής μη στάσιμο για d≠0. Συμπίπτει με το μοντέλο ARMA(p+d,q) όπου το πολυώνυμο AR έχει ακριβώς d ρίζες στο σύνορο του μοναδιαίου κύκλου.

Πίνακας 2.1

3. Μοντέλα Αυτοπαλινδρόμησης Κατανεμημένων Υστερήσεων

Όλα τα μοντέλα που συζητήθηκαν πιο πάνω είναι μονοδιάστατα (univariate), δηλαδή οι προβλέψεις βασίζονται σε μια μόνο τυχαία μεταβλητή, την Z_t και τις προηγούμενες τιμές της. Στην ανάλυση χρονοσειρών αρκετές φορές όμως, θέλουμε να συσχετίσουμε μια μεταβλητή, όχι μόνο με προηγούμενες τιμές της, άλλα και με τιμές άλλων επεξηγηματικών μεταβλητών (που μπορούν και αυτές να αποτελούν δικές τους χρονοσειρές).

Υπάρχει μια πληθώρα από τέτοια μοντέλα που όμως παρουσιάζουν δυσκολία στην εφαρμογή τους λόγω του ότι οι προβλέψεις μελλοντικών τιμών της μεταβλητής που μας ενδιαφέρει, απαιτούν και προβλέψεις για τις υπόλοιπες επεξηγηματικές μεταβλητές. Για τον λόγο αυτό αρκούμαστε συχνά στο να χρησιμοποιούμε περιορισμένες επεξηγηματικές μεταβλητές.

Ένα μοντέλο που βρίσκει αρκετές εφαρμογές σε προβλήματα, κυρίως στην οικονομετρία και σε άλλους χρηματοοικονομικούς κλάδους είναι το μοντέλο **Autoregressive Distributed Lag** (ARDL ή **ADL**). Ένα τέτοιο μοντέλο συνδέει την μεταβλητή που μελετάμε με προηγούμενες τιμές της, λαμβάνοντας όμως υπόψη και τις "ιστορικές" παρατηρήσεις μιας άλλης επεξηγηματικής μεταβλητής X_t. Ένα μοντέλο ADL λέγεται τάξης p,q όταν συνδέει την τρέχον τιμή της προς μελέτη μεταβλητής, με p σε πλήθος προηγούμενες παρατηρήσεις της ίδιας μεταβλητής και με την τρέχον τιμή της επεξηγηματικής μεταβλητής καθώς και τις q προηγούμενές της. Συγκεκριμένα το γενικό μοντέλο ADL τάξης p,q έχει την μορφή:

$$Z_t = \delta + \theta_1 Z_{t-1} + \ldots \theta_p Z_{t-p} + \varphi_0 X_t + \varphi_1 X_{t-1} \ldots + \varphi_q X_{t-q} + \varepsilon_t .$$

Όπως και στα προηγούμενα μοντέλα θεωρούμε ότι τα ε_t είναι ανεξάρτητα και ισόνομα με μέση τιμή 0 και διασπορά σ^2.

Αποδεικνύεται ότι για να είναι η διαδικασία Z_t στάσιμη, πρέπει και η X_t να είναι στάσιμη αλλά και επίσης οι ρίζες του AR πολυωνύμου να βρίσκονται εκτός του μοναδιαίου κύκλου, όπως ακριβώς ζητούσαμε και στα μοντέλα ARMA.

Αρκετές φορές έχουμε την δυνατότητα να επηρεάζουμε το επίπεδο των τιμών της επεξηγηματικής μεταβλητής X_t. Ένα απλό παράδειγμα είναι οι πωλήσεις μιας εταιρίας. Σαφώς και οι πωλήσεις στις προηγούμενες χρονικές περιόδους επηρεάζουν τις μελλοντικές πωλήσεις (λόγω σταθερότητας των καταναλωτών), όμως και οι διαφημίσεις των τελευταίων χρόνων επηρεάζουν τις πωλήσεις.

Θα μπορούσαμε δηλαδή να θεωρήσουμε ένα μοντέλο με Z_t τις πωλήσεις σε ευρώ μιας εταιρίας για την περίοδο t και X_t τα έξοδα τις εταιρίας για διαφημίσεις την χρονική περίοδο t (θεωρώντας πως τα έξοδα διαφήμισης δηλώνουν άμεσα την συνεισφορά των εν λόγω διαφημίσεων στις πωλήσεις).

Η μεταβλητή X_t σε αυτή την περίπτωση είναι απολύτως ελεγχόμενη και η τυχαιότητα της είναι περιορισμένη, όμως εύκολα μπορεί κανείς να σκεφτεί περιπτώσεις που η επεξηγηματική μεταβλητή είναι τυχαία αλλά μπορούμε να επηρεάσουμε με κάποιο τρόπο την κατανομή της, χωρίς να την ελέγχουμε απόλυτα.

Είναι επίσης φανερό ότι η συνεισφορά στις πωλήσεις από τις διαφημίσεις 2 ή και 3 χρόνων στον παρελθόν δεν είναι σημαντική, και λόγω αυτού πολλές φορές αρκούμαστε στην μελέτη του ειδικού αλλά πολύ χρήσιμου ADL (1,1) μοντέλου του οποίου η μορφή θα σχολιαστεί αργότερα.

Βάσει των πιο πάνω μας δημιουργούνται εύλογες απορίες που οι απαντήσεις τους είναι πολύ χρήσιμες για πρακτικές εφαρμογές.

Για παράδειγμα ποιες θα είναι οι μελλοντικές επιρροές στις πωλήσεις (μεταβλητή απόκρισης) αν αυξήσουμε τα έξοδα διαφημίσεων ενός χρόνου (επεξηγηματική μεταβλητή);

Ποιες θα είναι οι αντίστοιχες επιρροές στις πωλήσεις αν αυξήσουμε τα έξοδα διαφημίσεων για όλες τις υπόλοιπες χρονιές (πχ αύξηση κατά 10%);

Αν η διαδικασία Z_t είναι στάσιμη τότε πως θα επηρεαστεί το επίπεδο γύρω από το οποίο περιφέρονται οι τιμές της μακροχρόνια, δηλαδή ποιο θα είναι το νέο επίπεδο στασιμότητας (steady state);

Το ADL (1,1) μοντέλο έχει την μορφή:

$$Z_t = \delta + \theta Z_{t-1} + \varphi_0 X_t + \varphi_1 X_{t-1} + \varepsilon_t \quad (3.1).$$

Εφόσον ο τύπος αυτός είναι αναδρομικός, εύκολα μπορούμε να δούμε:

$$Z_{t+1} = \delta + \theta Z_t + \varphi_0 X_{t+1} + \varphi_1 X_t + \varepsilon_t$$

$$Z_{t+2} = \delta + \theta Z_{t+1} + \varphi_0 X_{t+2} + \varphi_1 X_{t+1} + \varepsilon_t \ . \ . \ .$$

Θέλοντας να δούμε πως επηρεάζει ένας παλμός (shock), δηλαδή μια μεμονωμένη αλλαγή στο X_t τις αμέσως επόμενες τιμές, ζητούμε ουσιαστικά τις τιμές των μερικών παραγώγων των τιμών Z ως προς την συγκεκριμένη X_t, που ονομάζουμε δυναμικούς πολλαπλασιαστές.

Συγκεκριμένα οι δυναμικοί πολλαπλασιαστές του μοντέλου ADL (1,1) παρουσιάζονται πιο κάτω:

$$\frac{\partial Z_t}{\partial X_t} = \varphi_0 .$$

$$\frac{\partial Z_{t+1}}{\partial X_t} = \varphi_1 + \theta \frac{\partial Z_t}{\partial X_t} = \varphi_1 + \theta\varphi_0 .$$

$$\frac{\partial Z_{t+2}}{\partial X_t} = \theta \frac{\partial Z_{t+1}}{\partial X_t} = \theta(\varphi_1 + \theta\varphi_0) .$$

Επαγωγικά καταλήγουμε στην σχέση που μας δίνει την επιρροή για k-χρονικές περιόδους στο μέλλον:

$$\frac{\partial Z_{t+k}}{\partial X_t} = \theta \frac{\partial Z_{t+k-1}}{\partial X_t} = \ldots = \theta^{k-1}(\varphi_1 + \theta\varphi_0) .$$

Γνωρίζουμε όμως ότι, προκειμένου η Z_t διαδικασία να είναι στάσιμη, η απόλυτη τιμή του θ πρέπει να είναι μικρότερη της μονάδας (όπως είδαμε στο AR μοντέλο τάξης 1). Το γεγονός αυτό μας λέει ότι σε βάθος χρόνου, η επιρροή ενός παλμού εξαφανίζεται καθώς:

$$\lim_{k\to\infty} \frac{\partial Z_{t+k}}{\partial X_t} = \lim_{k\to\infty} \theta^{k-1}(\varphi_1 + \theta\varphi_0) = 0 \; \gamma\iota\alpha |\theta| < 1.$$

Δηλαδή οι μεμονωμένες αλλαγές του X_t επηρεάζουν μόνο παροδικά τις μελλοντικές τιμές του Z_t.

Για μόνιμες όμως αλλαγές σε όλες τις τιμές των X_t πρέπει να υπολογίσουμε τον πολλαπλασιαστή βάθους χρόνου (Long-run Multiplier) που ορίζεται ως:

$$\beta = \frac{\partial E[Z_t]}{\partial E[X_t]}.$$

Χρησιμοποιώντας την σχέση 3.1 και παίρνοντας την αναμενόμενη τιμή βλέπουμε ότι:

$$E[Z_t] = \delta + \theta E[Z_{t-1}] + \varphi_0 E[X_t] + \varphi_1 E[X_{t-1}] + E[\varepsilon_t] .$$

Γνωρίζουμε όμως, εξ υποθέσεως, ότι και οι δυο διαδικασίες είναι στάσιμες, δηλαδή η αναμενόμενη τιμή στην τυχαία χρονική τιμή t-1 είναι ίδια με την αναμενόμενη τιμή στην αμέσως επόμενη χρονική στιγμή. Επίσης η αναμενόμενη τιμή του σφάλματος είναι μηδενική. Συνεπώς:

$$E[Z_t] = \delta + \theta E[Z_t] + \varphi_0 E[X_t] + \varphi_1 E[X_t]$$

$$E[Z_t] = \frac{\delta}{1-\theta} + \frac{(\varphi_0 + \varphi_1)}{1-\theta} E[X_t].$$

Επομένως προκύπτει ότι ο πολλαπλασιαστής βάθους χρόνου είναι:

$$\beta = \frac{\partial E[Z_t]}{\partial E[X_t]} = \frac{(\varphi_0 + \varphi_1)}{1-\theta}.$$

Δηλαδή αυξάνοντας τις τιμές του X_t για κάθε μελλοντική χρονική περίοδο (τα έξοδα διαφημίσεων στο παράδειγμα) οδηγούμαστε σε μια μακροχρόνια αύξηση του επιπέδου των τιμών του Z_t (πωλήσεις) κατά β-μονάδες.

Χρησιμοποιώντας αυτό σαν δεδομένο μπορούμε να βρούμε τη νέα στάσιμη κατάσταση του πιο πάνω μοντέλου, δηλαδή όταν ισχύει $Z_t = Z_{t-1}$ και $X_t = X_{t-1}$, όπου δίνεται από:

$$Z_t = \frac{\delta}{1-\theta} + \frac{(\varphi_0 + \varphi_1)}{1-\theta} X_t = \alpha + \beta X_t \ , \quad \text{όπου } \alpha = \frac{\delta}{1-\theta} \quad (3.2).$$

Μια διαφορετική ερμηνεία των πιο πάνω αποτελεσμάτων, μπορεί να μας δώσει το επόμενο μοντέλο που θα προκύψει άμεσα από τροποποίηση του στάσιμου μοντέλου ADL(1,1).

3.1 Μοντέλο Διόρθωσης Αποκλίσεων

Σε ένα στάσιμο ADL μοντέλο, όπως είδαμε πιο πάνω, υπάρχει ένα μακροχρόνιο σημείο ισορροπίας (Steady-State / Long Run Solution).

Μπορούμε να δημιουργήσουμε μια νέα χρονοσειρά, που να εκφράζει τις αποκλίσεις των τιμών, στις τυχαίες χρονικές στιγμές, από το μακροχρόνιο σημείο ισορροπίας. Μια τέτοια χρονοσειρά περιγράφεται κατάλληλα από το μοντέλο διόρθωσης αποκλίσεων **(Error Correction Model – ECM)**. Θεωρώντας και πάλι την εξίσωση 3.1 του μοντέλου ADL(1,1) έχουμε:

$$Z_t = \delta + \ \theta Z_{t-1} + \ \varphi_0 X_t + \ \varphi_1 X_{t-1} \ + \varepsilon_t.$$

Αφαιρώντας και από τις δύο πλευρές τον όρο Z_{t-1} και προσθαφαιρώντας στο δεξί μέλος τον όρο $\varphi_0 X_{t-1}$ έχουμε ισοδύναμα:

$$Z_t - Z_{t-1} = \delta + \ \theta Z_{t-1} - Z_{t-1} + \ \varphi_0 X_t - \varphi_0 X_{t-1} + \ \varphi_1 X_{t-1} + \varphi_0 X_{t-1} + \varepsilon_t \ ,$$

$$\boldsymbol{\Delta} Z_t = \delta + (\theta - 1) Z_{t-1} + \ \varphi_0 \boldsymbol{\Delta} X_t + (\varphi_0 + \ \varphi_1) X_{t-1} + \varepsilon_t \ .$$

Σε αυτό το σημείο βγάζουμε κοινό παράγοντα το θ-1 ή αλλιώς το –(1-θ) με εξαίρεση του όρους $\varphi_0 \boldsymbol{\Delta} X_t$ και ε_t και παίρνουμε:

$$\boldsymbol{\Delta} Z_t = -(1-\theta)\left\{ Z_{t-1} - \left[\frac{\delta}{1-\theta} + \frac{(\varphi_0 + \varphi_1)}{1-\theta} X_{t-1} \right] \right\} + \varphi_0 \boldsymbol{\Delta} X_t + \varepsilon_t$$

$$\boldsymbol{\Delta} Z_t = \gamma_1 [Z_{t-1} - (\alpha + \beta X_{t-1})] + \varphi_0 \boldsymbol{\Delta} X_t + \varepsilon_t \ , \ \ \text{όπου } \gamma_1 = (\theta - 1)$$

Βλέπουμε ότι εντός των αγκύλων στην τελευταία εξίσωση, έχουμε την απόκλιση της τιμής Z_{t-1} από το μακροχρόνιο σημείο ισορροπίας της όπως ορίστηκε στην εξίσωση 3.2 στο μοντέλο ADL(1,1).

Γνωρίζουμε όμως ότι στο στάσιμο μοντέλο ADL(1,1) η απόλυτη τιμή του θ είναι μικρότερη του 1 επομένως ο συντελεστής γ_1 ανήκει στο διάστημα (-2,0) και σαφώς είναι αρνητικός.

Το γεγονός ότι ο συντελεστής γ_1 είναι αρνητικός είναι διαισθητικά λογικό καθώς μας λέει ότι αν την χρονική στιγμή t-1 βρισκόμαστε πάνω από το σημείο ισορροπίας, η διαφορά εντός της παρένθεσης θα είναι θετική και έτσι πολλαπλασιάζοντας με αρνητικό συντελεστή θα οδηγηθούμε σε μια χαμηλότερη τιμή (πιο κοντά στο σημείο ισορροπίας). Στην αντίθετη περίπτωση που βρισκόμαστε κάτω από το μακροχρόνιο σημείο ισορροπίας, ο όρος εντός παρένθεσης θα είναι αρνητικός, και αφού πολλαπλασιάσουμε με τον αρνητικό συντελεστή του, θα γίνει θετικό πράγμα που μας δείχνει ότι η επόμενη τιμή του Z θα είναι ψηλότερη (προς την κατεύθυνση του σημείου ισορροπίας). Δηλαδή το σημείο ισορροπίας δρα σαν ελκυστής (**attractor**) των τιμών τις διαδικασίας.

Η πιο πάνω παράγραφος γίνεται ίσως πιο εύκολα αντιληπτή από την απεικόνιση του Διαγράμματος 3.1.

Διάγραμμα 3.1

4. Συνάρτηση Αυτοσυσχέτισης και Μερικής Αυτοσυσχέτισης

Ένα πολύ χρήσιμο εργαλείο, που μας βοηθά κυρίως στο να αναγνωρίσουμε το σωστό μοντέλο μιας χρονοσειράς, είναι η συνάρτηση Αυτοσυσχέτισης **(Autocorrelation Function)** που συμβολίζεται ως **ACF** και η συνάρτηση Μερικής Αυτοσυσχέτισης (**Partial Autocorrelation Function)** που συμβολίζεται με **PACF**.

Όπως είδαμε και πιο πάνω, μια διαδικασία Z_t καλείται απολύτως στάσιμη (strictly stationary) αν οι ιδιότητες της παραμένουν αναλλοίωτες σε μια αλλαγή του αρχικού της σημείου. Μαθηματικά αυτό μας λέει ότι η από κοινού κατανομή m+1 διαδοχικών παρατηρήσεων για τις χρονικές περιόδους t, t+1, t+2. . . t+m είναι ίδια με την από κοινού κατανομή m+1 διαδοχικών παρατηρήσεων για τις χρονικές περιόδους t+k, t+k+1, t+k+2…, t+k+m, για κάθε k.

Το πιο πάνω για m=0 μας λέει ότι η κατανομή για κάθε παρατήρηση είναι ίδια επομένως $p(Z_t) = p(Z)$ για κάθε t. Ο σταθερός μέσος γύρω από τον οποίο περιφέρεται η διαδικασία και η σταθερή διασπορά της διαδικασίας φαίνονται πιο κάτω:

$$\mu = E[Z_t] = \int_{-\infty}^{\infty} Z_t p(Z)\, dz\,.$$

$$\sigma^2 = E[(Z_t - \mu)^2] = \int_{-\infty}^{\infty} (Z_t - \mu)^2 p(z)\, dz\,.$$

Στην πράξη όμως έχουμε διακριτές, πεπερασμένες παρατηρήσεις της μεταβλητής Z_t και επομένως εκτιμούμε τις πιο πάνω ποσότητες με τον δειγματικό μέσο και την δειγματική διασπορά.

$$\hat{\mu} = \bar{Z} = \frac{1}{N}\sum_{t=1}^{N} Z_t \quad \kappa\alpha\iota \quad \hat{\sigma}_z^2 = \frac{1}{N}\sum_{t=1}^{N} (Z_t - \bar{Z})^2\,.$$

Από τα πιο πάνω έπεται άμεσα ότι η συνδιασπορά μεταξύ δυο παρατηρήσεων που απέχουν k χρονικές στιγμές μεταξύ τους είναι ίδια για κάθε t και την συμβολίζουμε με γ_k. Συγκεκριμένα έχουμε:

$$\gamma_k = Cov[Z_t, Z_{t+k}] = E[(Z_t - \mu)(Z_{t+k} - \mu)]\,.$$

Όμοια ο συντελεστής συσχέτισης (autocorrelation) ορίζεται ως:

$$\rho_k = \frac{E[(Z_t - \mu)(Z_{t+k} - \mu)]}{\sqrt{E[(Z_t - \mu)^2]\, E[(Z_{t+k} - \mu)^2]}} = \frac{\gamma_k}{\gamma_0}\,.$$

Στον παρονομαστή έχουμε απλά την κοινή διασπορά όλων των παρατηρήσεων που συμπίπτει με την τιμή του γ_0 όπως έχει οριστεί πιο πάνω. Εύκολα βλέπουμε ότι η τιμή του ρ_0 είναι ίση με την μονάδα όπως αναμέναμε.

Ένα γράφημα που μας δείχνει τις τιμές των ρ_k για κάθε τιμή του k (0,1,2,…) ονομάζεται **γράφημα ACF**. Όπως θα δούμε πιο κάτω, το γράφημα αυτό είναι βασικό εργαλείο για την ανίχνευση μοντέλων.

Φυσικά όμως οι πραγματικές (θεωρητικές) τιμές των ρ_k δεν μπορούν να υπολογιστούν από ένα δείγμα με πεπερασμένα δεδομένα. Δημιουργείται λοιπόν η ανάγκη να εντοπίσουμε κατάλληλες εκτιμήτριες των ποσοτήτων που μας ενδιαφέρουν αλλά και τα τυπικά σφάλματα των εκτιμητριών.

Ο δειγματικός μέσος είναι μια αμερόληπτη εκτιμήτρια του μέσου. Η διασπορά της εκτίμησης είναι:

$$Var[\bar{Z}] = \frac{1}{N^2}\sum_{t=1}^{N}\sum_{s=1}^{N}\gamma_{t-s} = \frac{\gamma_0}{N}\left[1 + 2\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)\rho_k\right].$$

Ίσως η καλύτερη και η συνηθέστερη εκτιμήτρια για τον συντελεστή συσχέτισης μεταξύ k (k=0,1,…) παρατηρήσεων είναι η συνάρτηση δειγματικών συσχετίσεων r_k:

$$r_k = \hat{\rho}_k = \frac{c_k}{c_0}\,, \qquad \text{όπου } \; c_k = \hat{\gamma}_k = \frac{1}{N}\sum_{t=1}^{N-k}(Z_t - \bar{Z})\,(Z_{t+k} - \bar{Z})\,.$$

Αποδεικνύεται ότι διασπορά της εκτίμησης είναι περίπου ίση με:

$$Var[r_k] \approx \frac{1}{N}\sum_{u=-\infty}^{\infty}(\rho_u^2 + \rho_{u+k}\rho_{u-k} - 4\rho_k\rho_u\rho_{u-k} + 2\rho_u^2\rho_k^2)$$

Στις διαδικασίες που οι συντελεστές συσχέτισης ρ_u είναι 0 (ή μπορούν να θεωρηθούν μηδενικοί) για κάθε u>q προκύπτει πώς πολλοί όροι της πιο πάνω έκφρασης μηδενίζονται για k>q. Καταλήγουμε λοιπόν:

$$Var[r_k] \approx \frac{1}{N}\left(1 + 2\sum_{u=1}^{q}\rho_u^2\right)\,, \quad k > q\,.$$

Αφού εκτιμήσουμε τα r_u για u=1,2,…q, το τυπικό σφάλμα της εκτίμησης r_k είναι:

$$SE[r_k] = \sqrt{\widehat{Var}[r_k]} \approx \sqrt{\frac{1}{N}\left(1 + 2\sum_{u=1}^{q}r_u^2\right)}\,, \quad k > q\,.$$

4.1 Συναρτήσεις ACF και PACF σε μοντέλα AR

Θα δούμε συγκεκριμένα τώρα σε μοντέλα AR την μορφή των ACF και PACF. Υπενθυμίζεται από την εξίσωση 2.2 ότι για το μοντέλο AR τάξης p έχουμε:

$$Z_t = \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \ldots + \varphi_p Z_{t-p} + a_t \, .$$

Πολλαπλασιάζοντας και τα δύο μέλη με τον όρο Z_{t-k} για k θετικό και παίρνοντας αναμενόμενες τιμές καταλήγουμε στην εξίσωση διαφορών:

$$\gamma_k = \varphi_1 \gamma_{k-1} + \varphi_2 \gamma_{k-2} + \ldots + \varphi_p \gamma_{k-p} \quad (4.1).$$

Μπορούμε να διαιρέσουμε όλους τους όρους με το γ_0 και να μετατρέψουμε την εξίσωση 4.1 στην ισοδύναμή της:

$$\rho_k = \varphi_1 \rho_{k-1} + \varphi_2 \rho_{k-2} + \ldots + \varphi_p \rho_{k-p} \quad (4.2).$$

Χρησιμοποιώντας προηγούμενους συμβολισμούς, και συγκεκριμένα το πολυώνυμο AR, μπορούμε να καταλήξουμε στο ότι:

$$\varphi(\boldsymbol{B})\rho_k = 0 \, , \qquad \text{όπου το } \boldsymbol{B} \text{ δεν δρά στο } t \text{ αλλά στο } k.$$

Το φ(**B**) που είναι πολυώνυμο βαθμού p, μπορεί γραφεί σαν γινόμενο των ριζών του, πιο συγκεκριμένα

$$\varphi(\boldsymbol{B}) = \prod_{i=1}^{p} (1 - G_i B)$$

Συνεπώς η γενική λύση της εξίσωσης διαφορών 4.2 είναι:

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \ldots + A_p G_p^k$$

Επειδή η διαδικασία είναι στάσιμη, έπεται ότι οι ρίζες του πολυωνύμου AR βρίσκονται εκτός του μοναδιαίου κύκλου, που ισοδύναμα μας λέει ότι η απόλυτη τιμή των G_i είναι μικρότερη του 1 για κάθε i. Επομένως βλέπουμε πως οι συντελεστές συσχέτισης ρ_k τείνουν στο 0 καθώς το k τείνει στο άπειρο.

Σε περιπτώσεις μιγαδικών ριζών του πολυωνύμου, η συνεισφορά τους στον συντελεστή συσχέτισης παίρνει την μορφή συγκλίνουσας ημιτονοειδούς σειράς που τείνει στο 0.

Δηλαδή η συνάρτηση συσχέτισης ACF ενός στάσιμου AR μοντέλου τάξης p έχει άπειρους όρους που συγκλίνουν στο 0. Παρόλα αυτά, εκ κατασκευής, μπορούμε να περιγράψουμε τη συσχέτιση αυτή με p στο πλήθος μη μηδενικές συναρτήσεις.

Το πιο πάνω γεγονός μας οδηγεί στην χρήση της μερικής συνάρτησης αυτοσυσχέτισης PACF. Συμβολίζουμε με φ_{kj} τον j συντελεστή μιας διαδικασίας AR τάξης k με j=1,…k. Από την εξίσωση διαφορών 4.2 έχουμε τις εξισώσεις:

$$\rho_j = \varphi_{k1}\rho_{j-1} + \varphi_{k2}\rho_{j-2} + \ldots + \varphi_{k(k-1)}\rho_{j-k+1} + \varphi_{kk}\rho_{j-k} \quad (j = 1, \ldots k) .$$

Ουσιαστικά πιο πάνω έχουμε ένα σύστημα από k εξισώσεις (μια για κάθε τιμή του j). Οι εξισώσεις τους συστήματος είναι γνωστές με την ονομασία εξισώσεις **Yule-Walker**. Το σύστημα μπορεί να γραφεί με την βοήθεια πινάκων ως εξής:

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \varphi_{k1} \\ \varphi_{k2} \\ \varphi_{k3} \\ \vdots \\ \varphi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_k \end{bmatrix} \rightarrow \boldsymbol{P}_k \boldsymbol{\varphi}_k = \boldsymbol{\rho}_k (4.3) .$$

Μπορούμε να λύσουμε διαδοχικά για διάφορες τιμές του k το πιο πάνω σύστημα και έχουμε ότι:

$$\varphi_{11} = \rho_1 \quad (\gamma\iota\alpha\ k = 1) .$$

$$\varphi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \quad (\gamma\iota\alpha\ k = 2) .$$

$$\varphi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}} \quad (\gamma\iota\alpha\ k = 3) .$$

Γενικά βλέπουμε ότι ο γενικός όρος φ_{kk} έχει σαν παρονομαστή την ορίζουσα του πίνακα $\mathbf{P}_k$ ενώ στο αριθμητή έχει την ίδια ορίζουσα με μόνη διαφορά ότι στην τελευταία στήλη έχει τα στοιχεία του διανύσματος $\boldsymbol{\rho}_k$. Η συνάρτηση με τιμές τα φ_{kk} για τους διάφορους χρόνους καθυστέρησης k καλείται συνάρτηση μερικής αυτοσυσχέτισης PACF.

Σε ένα μοντέλο AR τάξης p, ισχύει ότι οι τιμές των φ_{kk} είναι μη μηδενικές για $k \leq p$ ενώ θα ισούται με 0 για κάθε τιμή του $k > p$. Με άλλα λόγια περιμένουμε ότι σε ένα AR μοντέλο, οι εκτιμήσεις των τιμών της συνάρτησης PACF θα είναι διάφορες του μηδενός μέχρι και την τιμή του p. Συνεπώς μπορούμε να ανιχνεύσουμε την τιμή της τάξης του μοντέλου χρησιμοποιώντας τον στατιστικό έλεγχο αν η εκτιμώμενη τιμή του φ_{kk} είναι σημαντικά διάφορη του μηδενός.

Χρησιμοποιώντας την γνώστη μέθοδο ελαχίστων τετραγώνων, αποδεικνύετε ότι για μια στάσιμη διαδικασία με μέσο όρο 0, η καλύτερη γραμμική εκτίμηση για την τιμή Z_t βάση των παρατηρήσεων Z_{t-1}, Z_{t-2},…, Z_{t-k}, δηλαδή η εκτίμηση με το μικρότερο τυπικό σφάλμα, είναι η:

$$\hat{Z}_t = \varphi_{(k-1),1} Z_{t-1} + \varphi_{(k-1),2} Z_{t-2} + \ldots + \varphi_{(k-1),(k-1)} Z_{t-k+1}$$

Οι πραγματικές τιμές της συνάρτησης PACF είναι επίσης άγνωστες, επομένως χρειάζεται να βρούμε κατάλληλες εκτιμήσεις και τυπικά σφάλματα για τις εκτιμήσεις μας.

Μια μέθοδος που εφαρμόζεται πολύ συχνά, κυρίως στην σύγχρονη εποχή που αναπτύχθηκαν υπολογιστικά στατιστικά πακέτα, είναι να προσαρμόσουμε διαδοχικά πολλά μοντέλα AR τάξης 1,2, . . . με μεθόδους ελαχίστων τετραγώνων των σφαλμάτων, και να δούμε τον τελευταίο συντελεστή που προκύπτει σε κάθε μοντέλο. Ο τελευταίος συντελεστής του AR μοντέλου τάξης 1, θα είναι ο φ_{11}, ο τελευταίος συντελεστής του AR μοντέλου τάξης 2, θα είναι ο φ_{22} και ούτω καθεξής.

Εναλλακτικά μπορούμε να χρησιμοποιήσουμε τις εξισώσεις **Yule Walker**, να αντικαταστήσουμε τις τιμές των συντελεστών συσχέτισης με τις αντίστοιχες εκτιμήσεις τους, δηλαδή $\rho_j \rightarrow r_j$, και να λύσουμε το σύστημα για διάφορες τιμές του k=1,2,…

Όσο αφορά τα τυπικά σφάλματα των εκτιμήσεων μας, κάνοντας την υπόθεση ότι η διαδικασία που μελετάτε είναι μια διαδικασία AR τάξης p, οι όροι φ_{kk} για k ≥ p+1 είναι ανεξάρτητοι και ακολουθούν κανονική κατανομή με μέση τιμή 0 και διασπορά περίπου ίση με:

$$Var[\hat{\varphi}_{kk}] \approx \frac{1}{n} \text{ , όπου n ο αριθμός των παρατηρήσεων στο δείγμα.}$$

Επομένως τα τυπικά σφάλμα είναι:

$$SE[\hat{\varphi}_{kk}] \equiv \sqrt{\widehat{Var}[\hat{\varphi}_{kk}]} \approx \frac{1}{\sqrt{n}} \text{ , για } k \geq p+1 \text{ .}$$

4.2 Συναρτήσεις ACF και PACF σε μοντέλα MA

Αντίστοιχα, με ίδιες μεθόδους, πρέπει να βρούμε τις τιμές των συναρτήσεων ACF και PACF για το μοντέλο MA τάξης q. Υπενθυμίζεται ότι ένα τέτοιο μοντέλο έχει την μορφή της εξίσωσης 2.7.

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \text{ .}$$

Επομένως για την συνάρτηση αυτοσυσχέτισης έχουμε:

$$\gamma_\kappa = E[Z_t Z_{t-k}] = E\left[\left(a_t - \ldots - \theta_q a_{t-q}\right)\left(a_{t-k} - \ldots - \theta_q a_{t-k-q}\right)\right]$$

$$\gamma_\kappa = -\theta_k E[a_{t-k}^2] + \theta_1\theta_{k+1} E[a_{t-k-1}^2] + \ldots + \theta_{q-k}\theta_q E\left[a_{t-q}^2\right].$$

Αυτό ισχύει γιατί τα σφάλματα είναι ασυσχέτιστα και $\gamma_k = 0$ για $k > q$ λόγω της περιορισμένης μνήμης του μοντέλου MA.

Θέτοντας k=0 στο πιο πάνω τύπο παίρνουμε την διασπορά της διαδικασίας που ισούται με:

$$\gamma_0 = \left(1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2\right)\sigma_\alpha^2 .$$

Για k>0 οι τιμές των γ_k είναι 0 ενώ για τιμές k=1,2,…,q ισχύει

$$\gamma_k = \left(-\theta_k + \theta_1\theta_{k+1} + \ldots + \theta_{q-k}\theta_q\right)\sigma_\alpha^2 .$$

Αντίστοιχα τα ρ_k όπως και πριν ισούνται με τον λόγο γ_k ως προς γ_0.

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \ldots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2} & k = 1,2,\ldots,q \\ 0 & k > q \end{cases}.$$

Δηλαδή βλέπουμε ότι οι τιμές της ACF συνάρτησης, αποκόπτονται ακριβώς μετά το q. Επομένως θα μπορούσαμε εκτιμώντας τους όρους της ACF συνάρτησης για μια διαδικασία, να υπολογίσουμε την παράμετρο μετά από την οποία οι τιμές δεν διαφέρουν σημαντικά από το 0, και έτσι να εκτιμήσουμε την τάξη του MA μοντέλου που περιγράφει την διαδικασία.

Η συνάρτηση PACF που προκύπτει από τις λύσεις των Yule Walker εξισώσεων έχει μια αρκετά περίπλοκη αναλυτική μορφή. Για τους σκοπούς της μελέτης μας, μας αρκεί ότι η PACF συνάρτηση είτε φθίνει εκθετικά στο 0, είτε ακολουθά φθίνουσα ημιτονοειδές συνάρτηση που τείνει στο 0. Δηλαδή η PACF συνάρτηση μιας MA διαδικασίας μοιάζει σε συμπεριφορά με την ACF συνάρτηση μιας AR διαδικασίας.

4.3 Συναρτήσεις ACF και PACF σε μοντέλα ARMA

Τώρα μελετάμε τις ACF και PACF συναρτήσεις στην περίπτωση των μεικτών μοντέλων ARMA. Υπενθυμίζουμε ότι βάσει της εξίσωσης 2.8, η γενική εξίσωση του μοντέλου ARMA τάξης (p,q) είναι:

$$Z_t = \varphi_1 Z_{t-1} + \ldots + \varphi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} .$$

Πολλαπλασιάζοντας με την τιμή Z_{t-k} και παίρνοντας τις αναμενόμενες τιμές δεξιά και αριστερά της ισότητας, με πολύ όμοιες διαδικασίες όπως στα μοντέλα AR, καταλήγουμε στην εξίσωση:

$$\gamma_k = \varphi_1\gamma_{k-1} + \ldots + \varphi_p\gamma_{k-p} + \gamma_{za}(k) - \theta_1\gamma_{za}(k-1) - \ldots - \theta_q\gamma_{za}(k-q)\,,$$

όπου το $\gamma_{za}(k)$ είναι η συνάρτηση συνδιασποράς μεταξύ των τιμών Z_t και α_t, που ορίζεται ως η αναμενόμενη τιμή του γινομένου Z_{t-k} επί α_t.

Γνωρίζοντας ότι η τιμή του Z_{t-k} εξαρτάται από τις προηγούμενες διακυμάνσεις και χρησιμοποιώντας την μορφή του μοντέλου MA ως άθροισμα άπειρων διακυμάνσεων που φαίνεται στη Εξίσωση 4.4 καταλήγουμε στον υπολογισμό των $\gamma_{za}(k)$:

$$Z_{t-k} = \psi(\boldsymbol{B})a_{t-k} = \sum_{j=0}^{\infty}\psi_j a_{t-k-j} \quad (4.4)\,.$$

$$\gamma_{za}(k) = \begin{cases} 0 & k > 0 \\ \psi_{-k}\sigma_\alpha^2 & k \le 0 \end{cases}.$$

Τελικά, σύμφωνα με τα πιο πάνω καταλήγουμε:

$$\gamma_k = \varphi_1\gamma_{k-1} + \ldots + \varphi_p\gamma_{k-p} - \sigma_\alpha^2\left(\theta_k\psi_0 + \theta_{k+1}\psi_1 + \ldots + \theta_q\psi_{q-k}\right),\ \theta_0 = -1\,.$$

Βλέπουμε ότι για k > q έπεται ότι οι όροι της παρένθεσης δεν υπάρχουν.

$$\gamma_k = \varphi_1\gamma_{k-1} + \varphi_2\gamma_{k-2} + \ldots + \varphi_p\gamma_{k-p}\,.$$

$$\rho_k = \varphi_1\rho_{k-1} + \varphi_2\rho_{k-2} + \ldots + \rho_p\rho_{k-p}\,.$$

Σύμφωνα με τα πιο πάνω, όταν η τάξη των όρων AR είναι μεγαλύτερη της τάξης των όρων MA, δηλαδή p > q, τότε οι τιμές της εξίσωσης ACF αποτελούνται από άθροισμα φθινουσών εκθετικών και ημιτονοειδών συναρτήσεων. Δηλαδή αναμένουμε ένα γράφημα που συγκλίνει γεωμετρικά στο 0. Στην περίπτωση που το p ≤ q, περιμένουμε την ίδια συμπεριφορά με εξαίρεση όμως τους αρχικούς q-p+1 συντελεστές αυτοσυσχέτισης.

Για την συνάρτηση PACF υπενθυμίζουμε πως το μοντέλο ARMA γράφεται συνοπτικά στην μορφή:

$$\varphi(\boldsymbol{B})Z_t = \theta(\boldsymbol{B})a_t \rightarrow a_t = \theta^{-1}(\boldsymbol{B})\varphi(\boldsymbol{B})Z_t\,,$$

όπου το $\theta^{-1}(\boldsymbol{B})$ είναι μια άπειρη σειρά των **B**. Συνεπώς η συνάρτηση PACF έχει άπειρο μήκος (δεν διακόπτεται απότομα). Σε βάθος χρόνου η συνάρτηση PACF των μοντέλων ARMA έχει όμοια συμπεριφορά της συνάρτησης PACF στα μοντέλα MA. Δηλαδή περιμένουμε τις τιμές των PACF να συγκλίνουν γεωμετρικά στο 0.

4.4 Συμπεράσματα Συναρτήσεων ACF - PACF

Τέλος, κλείνουμε την ενότητα των συναρτήσεων ACF και PACF συνοψίζοντας στον Πίνακα 4.1 τις θεωρητικές συμπεριφορές των εν λόγω συναρτήσεων στα μοντέλα AR, MA και ARMA.

Συμπεριφορά συναρτήσεων ACF/PACF βάση των μοντέλων			
Συνάρτηση	**AR(p)**	**MA(q)**	**ARMA(p,q)**
ACF	**Άπειροι όροι** με σταδιακή απόσβεση με ή χωρίς ημιτονοειδή συμπεριφορά (**συγκλίνουν στο 0**).	**Πεπερασμένοι** μη μηδενικοί όροι (Αποκόπτονται απότομα μετά από **q περιόδους**).	**Άπειροι όροι** με απόσβεση με ή χωρίς ημιτονοειδή συμπεριφορά (**συγκλίνουν στο 0**) **με εξαίρεση τις πρώτες q-p περιόδους.**
PACF	**Πεπερασμένοι** μη μηδενικοί όροι (Αποκόπτονται απότομα μετά από **p περιόδους**).	**Άπειροι όροι** με σταδιακή απόσβεση με ή χωρίς ημιτονοειδή συμπεριφορά (**συγκλίνουν στο 0**).	**Άπειροι όροι** με απόσβεση με ή χωρίς ημιτονοειδή συμπεριφορά (**συγκλίνουν στο 0**) **με εξαίρεση τις πρώτες p-q περιόδους.**

Πίνακας 4.1

5. Κατασκευή Μοντέλου σε προβλήματα Χρονοσειρών

Σε αυτή την ενότητα θα ασχοληθούμε με την κατασκευή μοντέλων (**Model Building**) που περιγράφουν επαρκώς τα δεδομένα μας, ώστε να μπορέσουμε να τα χρησιμοποιήσουμε για μελλοντικές προβλέψεις προεκτείνοντας τα.

Συγκεκριμένα, θα ασχοληθούμε αρχικά με την ανίχνευση της γενικής μορφής του μοντέλου αλλά και την εκτίμηση του πλήθους των παραμέτρων του εν λόγω μοντέλου.

Έπειτα, με την χρήση εκτιμητριών μέγιστης πιθανοφάνειας, θα εκτιμήσουμε τους συντελεστές του υποψήφιου μοντέλου και ακολούθως θα χρησιμοποιήσουμε διαγνωστικούς ελέγχους υπολοίπων, για να επιβεβαιώσουμε ότι μοντέλο που χρησιμοποιήσαμε είναι κατάλληλο να περιγράψει ικανοποιητικά τα δεδομένα μας και συνεπώς οι μελλοντικές προβλέψεις που θα προκύψουν θα είναι βάσιμες.

Σε περίπτωση που οι διαγνωστικοί έλεγχοι μας δώσουν ενδείξεις ότι το μοντέλο που εκτιμήθηκε δεν είναι κατάλληλο, τότε θα προβούμε σε υπολογισμούς εναλλακτικών, παραπλήσιων μοντέλων έως ότου να καταλήξουμε σε ένα οικονομικό και ικανοποιητικό μοντέλο.

5.1 Εύρεση Γενικής Μορφής Μοντέλου

Για τα προβλήματα που θα ακολουθήσουν, θα χρησιμοποιήσουμε την μέθοδο της μετάβασης από ένα γενικό μοντέλο σε ειδικό μοντέλο.

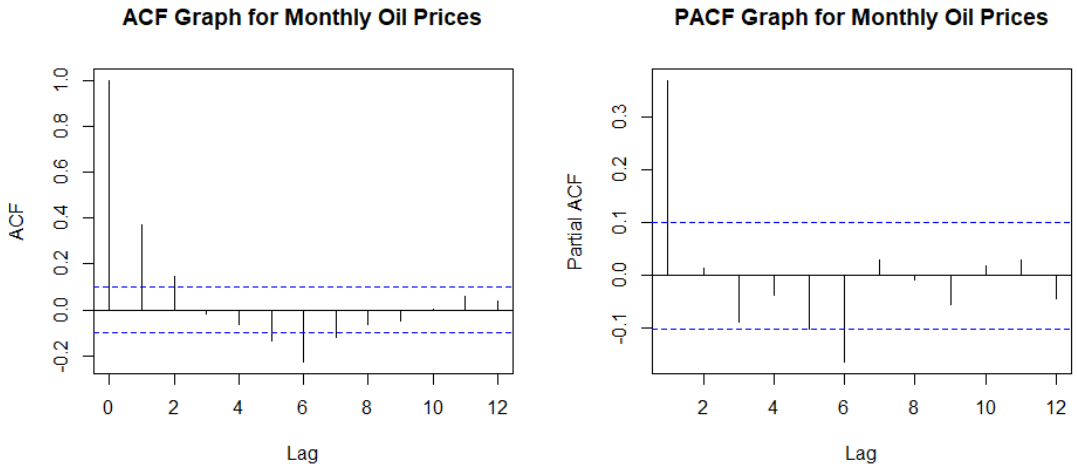
Αρχικά, μπορούμε με ένα διάγραμμα που απεικονίζει τις τιμές της προς μελέτη μεταβλητής σε σχέση με την χρονική περίοδο κατά την οποία παρατηρήθηκαν (ανάλογο του Διαγράμματος 1.1), να διαπιστώσουμε την ύπαρξη τάσεων, δηλαδή αν τα δεδομένα μας τείνουν να αυξάνονται ή να μειώνεται σε βάθος χρόνου. Σε περίπτωση που σε ένα τέτοιο διάγραμμα δεν υπάρχουν τάσεις, έχουμε μια ένδειξη ότι η εν λόγω χρονοσειρά παρουσιάζει στασιμότητα.

Πέρα από τον γραφικό αυτό έλεγχο, μπορούμε να γίνουμε πιο ακριβής χρησιμοποιώντας το **Augmented Dickey-Fuller test** που ελέγχει την υπόθεση ότι υπάρχει μια τουλάχιστον ρίζα του πολυωνύμου AR που βρίσκεται στο σύνορο ή εντός του μοναδιαίου κύκλου (που θα μας υποδείκνυε μη-στασιμότητα) έναντι της εναλλακτικής υπόθεσης ότι όλες οι ρίζες βρίσκονται εκτός του μοναδιαίου κύκλου.

Σε περίπτωση που το προς μελέτη μοντέλο είναι μη στάσιμο τότε παίρνουμε διαφορές πρώτης τάξης και επαναλαμβάνουμε το γραφικό έλεγχο και το Dickey Fuller test. Συνεχίζουμε με την ίδια διαδικασία μέχρι να καταλήξουμε σε στάσιμη σειρά. Ο αριθμός των επαναλήψεων της πιο διαδικασίας μας δίνει την τιμή του d (παράμετρος Integration) στο γενικό μοντέλο ARIMA (p,d,q).

Χρησιμοποιώντας πλέον την χρονοσειρά που προκύπτει από τις d-διαφορές, επιλέγουμε παραμέτρους p και q σχετικά μεγάλες ώστε να έχουμε ένα αρκετά περιγραφικό μοντέλο. Έπειτα, θέτουμε περιορισμούς στις τιμές των παραμέτρων ώστε σταδιακά να ελέγξουμε όλα τα απλούστερα, εμφωλευμένα μοντέλα (**nested models**) του αρχικού μεταξύ τους.

Μια καλή μέθοδος για να βρούμε τις αρχικές τιμές των παραμέτρων p και q για την πιο πάνω διαδικασία είναι η μέθοδος αποσύνθεσης Box-Jenkins **(Box-Jenkins Decomposition)**. Για το γενικό, στάσιμο πλέον μοντέλο ARMA, κάνουμε χρήση των ACF και PACF γραφημάτων και βλέπουμε σε ποιο σημείο αποκόπτονται οι τιμές τους (δηλαδή το σημείο πέρα του οποίου δεν είναι στατιστικά διάφορες του μηδενός) και εκτιμούμε τις τιμές των q και p αντίστοιχα όπως συζητήθηκε στην προηγούμενη ενότητα.



Διάγραμμα 5.1

Στο Διάγραμμα 5.1, όλες οι τιμές των ACF και PACF που εμπίπτουν εντός των δυο διακεκομμένων γραμμών δεν μπορούν να θεωρηθούν στατιστικά διάφορες του 0.

Επομένως στο γράφημα ACF βλέπουμε πως μετά την **τρίτη** παρατήρηση (τιμή της ACF συνάρτησης για παρατηρήσεις που απέχουν 2 περιόδους μεταξύ τους) οι τιμές φαίνονται να είναι στατιστικά ασήμαντες (με ίσως μοναδική εξαίρεση την τιμή για lag=6). Αυτή η συμπεριφορά, σύμφωνα με τον Πίνακα 5.1, είναι ανάλογη ενός μοντέλου MA τάξης 3 (q=3).

Επίσης από το γράφημα PACF βλέπουμε ότι μόνο η πρώτη παρατήρηση είναι στατιστικά σημαντική ενώ οι υπόλοιπες βρίσκονται εντός των διακεκομμένων γραμμών (και πάλι με εξαίρεση για lag=6). Αυτό, σύμφωνα με τον Πίνακα 5.1, είναι κάτι που θα αναμενόταν σε ένα μοντέλο AR τάξης 1 (p=1).

Συνεπώς σε αυτή την περίπτωση θα υποψιαζόμασταν ότι ένα πιθανό μοντέλο που θα μπορούσε να περιγράφει ικανοποιητικά τα δεδομένα μας είναι το ARMA (1,3).

Για σκοπούς ακόμη ορθότερης προσέγγισης μπορούμε να χρησιμοποιήσουμε σαν αρχικές παραμέτρους p+1 και q+1 όρους ώστε να διασφαλίσουμε καλύτερη ακρίβεια, εφόσον όλα τα εμφωλευμένα μοντέλα θα ελεγχθούν.

Αφού βρούμε τις τιμές των αρχικών παραμέτρων, προσαρμόζουμε το μοντέλο στα δεδομένα μας και ξεχωριστά για την εκτίμηση του κάθε συντελεστή ελέγχουμε με t-test την υπόθεση ότι είναι μηδενικός. Σε περίπτωση που ένας συντελεστής είναι στατιστικά μηδενικός τότε απλοποιούμε το μοντέλο αφαιρώντας αυτή την παράμετρο. Για παράδειγμα αν έχουμε ένα μοντέλο ARMA (p,q) και βρούμε ότι ο AR συντελεστής για την χρονική περίοδο t-p είναι στατιστικά ίσος με το μηδέν, προσαρμόζουμε εκ νέου το μοντέλο ARMA (p-1,q). Μπορούμε έτσι να καταλήξουμε στο απλούστερο μοντέλο όπου όλοι οι όροι του είναι σημαντικοί.

Φυσικά κατά την διαδικασία ελάττωσης της τάξης πρέπει να συγκρίνουμε όλα τα μοντέλα και με το αρχικό αλλά και μεταξύ τους ώστε να καταλήξουμε στο ορθότερο. Στην περίπτωση που έχουμε μοντέλα που δεν είναι εμφωλευμένα το ένα στο άλλο, όπως για τα μοντέλα ARMA (p-1,q) και ARMA (p,q-1), η σύγκριση αυτή γίνεται με πληροφοριακά κριτήρια (**Information Criteria**) όπως το Akaike Information Criterion **(AIC)** και Bayesian Information Criterion **(BIC)**. Τέτοια κριτήρια χρησιμοποιούν ποινικοποιημένες συναρτήσεις για να ισορροπήσουν την ακρίβεια (**fit**) και την απλότητα (**parsimony**) του μοντέλου. Επιλέγουμε το κατάλληλο μοντέλο με την μικρότερη τιμή AIC/BIC. Συγκεκριμένα οι τιμές AIC και BIC προκύπτουν από τις εξισώσεις 5.1 και 5.2 αντίστοιχα:

$$AIC = 2k - 2\ln(\hat{L}) \quad (5.1)\,,$$

$$BIC = k[\ln(n)] - 2\ln(\hat{L}) \quad (5.2)\,,$$

όπου με k συμβολίζουμε το πλήθος των προς εκτίμηση παραμέτρων, $\hat{L}$ είναι η μέγιστη τιμή της συνάρτησης πιθανοφάνειας και n το πλήθος του δείγματος (δηλαδή το πλήθος των παρατηρήσεων μας).

5.2 Εκτιμήσεις των συντελεστών στο γενικό μοντέλο ARIMA (p,d,q)

Είναι προφανές ότι τα μοντέλα ARMA, AR και MA είναι ειδικές περιπτώσεις του μοντέλου ARIMA, για παράδειγμα ένα μοντέλο ARIMA (p,0,0) είναι ταυτόσημο με το μοντέλο AR(p). Επόμενος αναλύοντας τις εκτιμήσεις των παραμέτρων για το γενικό μοντέλο ARIMA, καλύπτουμε όλες τις δυνατές περιπτώσεις που θα αντιμετωπίσουμε.

Αφού καταλήξουμε στην καταλληλότερη μορφή του μοντέλου, για να εκτιμήσουμε τους συντελεστές χρησιμοποιούμε συνήθως την μέθοδο Μέγιστης Πιθανοφάνειας (**Maximum Likelihood**) που είναι ισοδύναμη με την μέθοδο Ελαχίστων Τετραγώνων (**Least Squares**) στην δεσμευμένη περίπτωση.

5.2.1 Δεσμευμένες Εκτιμήτριες Μέγιστης Πιθανοφάνειας σε μοντέλα ARIMA(p,d,q)

Θεωρούμε ως γνωστές N (N = n + d) στο πλήθος διαδοχικές παρατηρήσεις της αρχικής σειράς προς μελέτη Y, και τις συμβολίζουμε ως Y_{-d+1}, Y_{-d+2}. . . Y_0, Y_1, Y_2 . . . Y_n.

Χρησιμοποιώντας τις παρατηρήσεις αυτές, μπορούμε να δημιουργήσουμε μια νέα σειρά πλήθους n (N-d) παίρνοντας τις διαφορές τάξης d, δηλαδή η νέα σειρά μας (που είναι στάσιμη) έχει τους όρους W_0, W_1, . . . W_n όπου $W_t = \mathbf{\Delta}^d(Y_t)$.

Οπότε το πρόβλημα εκτίμησης των συντελεστών στο αρχικό μη-στάσιμο μοντέλο ARIMA (p,d,q) είναι ισοδύναμο με τους υπολογισμούς των παραμέτρων στο στάσιμο και αντιστρέψιμο μοντέλο ARMA (p,q). Όπως είδαμε στην εξίσωση 2.8 ισχύει ότι:

$$Z_t = \varphi_1 Z_{t-1} + \ldots + \varphi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}$$

και λύνοντας ως προς α_t έχουμε

$$a_t = Z_t - \varphi_1 Z_{t-1} - \varphi_2 Z_{t-2} - \ldots - \varphi_p Z_{t-p} + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q} ,$$

όπου θέσαμε $Z_t = W_t - \mu = \mathbf{\Delta}^d(Y_t) - \mu$ όπου $\mu = E[W_t]$ που σύμφωνα με το δείγμα μας (αν είναι ικανοποιητικά μεγάλο) εκτιμάτε με το δειγματικό μέσο των τιμών W_t. Δηλαδή η νέα σειρά Z που δημιουργήσαμε, είναι στάσιμη με μέση τιμή 0, και ακολουθεί μοντέλο ARMA τάξης (p,q).

Οι τιμές των α_t στο μοντέλο Z ικανοποιούν την εξίσωση διαφορών που φαίνεται πιο πάνω. Οι τιμές όμως των α_t δεν μπορούν να υπολογιστούν άμεσα καθώς εξαρτώνται από τις προηγούμενες μεταβλητές.

Αν θεωρήσουμε ότι έχουμε την γνώση p-όρων της μεταβλητής Z_t (διανυσματικά **z***), και των q-όρων της μεταβλητής α_t (διανυσματικά **α***) **πριν την έναρξη της διαδικασίας** τότε θα μπορούσαμε αναδρομικά να υπολογίσουμε τις τιμές των $\alpha_1, \alpha_2, \ldots \alpha_n$ συναρτήσει των φ, των θ και των Z. Αφού οι τιμές των όρων α_t είναι ανεξάρτητες και ακολουθούν κανονική κατανομή, χρησιμοποιώντας την συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής ισχύει:

$$p(\alpha_1, \alpha_2, \ldots, \alpha_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} e^{-\frac{(a_i-0)^2}{2\sigma_\alpha^2}} = \frac{1}{\left(\sqrt{2\pi\sigma_\alpha^2}\right)^n} e^{-\left[\sum_{i=1}^{n}\left(\frac{a_i^2}{2\sigma_\alpha^2}\right)\right]}.$$

Συνεπώς, αν έχουμε συγκεκριμένα δεδομένα **Z** η λογαριθμημένη συνάρτηση πιθανοφάνειας, με παραμέτρους τις τιμές των **φ, θ** και σ^2_α που δεσμεύονται από την επιλογή των διανυσμάτων **z*** και **α*** είναι:

$$l^*(\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma_\alpha^2 \mid \boldsymbol{z}^*, \boldsymbol{a}^*) = -\frac{n}{2} ln(2\pi\sigma_\alpha^2) - \frac{S^*(\boldsymbol{\varphi}, \boldsymbol{\theta})}{2\sigma_\alpha^2} \quad , \quad \text{όπου}$$

$$S^*(\boldsymbol{\varphi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} a_t^2(\boldsymbol{\varphi}, \boldsymbol{\theta}, \boldsymbol{Z} \mid \boldsymbol{z}^*, \boldsymbol{a}^*) \quad \textit{το δεσμευμένο Άθροισμα Τετραγώνων.}$$

Εφόσον θέλουμε να μεγιστοποιήσουμε την συνάρτηση πιθανοφάνειας, μπορούμε να αγνοήσουμε τον όρο 2π εντός του λογαρίθμου καθώς μπορούμε να διασπάσουμε το γινόμενο ως άθροισμα λογαρίθμων και επομένως θα προκύψει ένας σταθερός όρος (που περιέχει μόνο το n και π). Επίσης σε περίπτωση που η διασπορά είναι σταθερή, τότε οι εκτιμήσεις βάσει της δεσμευμένης εκτιμήτριας μέγιστης πιθανοφάνειας και της εκτιμήτριας ελαχίστων δεσμευμένων αθροισμάτων τετραγώνων είναι ακριβώς οι ίδιες.

Σε περιπτώσεις που η χρονοσειρά παρουσιάζει στασιμότητα και συγκεκριμένα έχει μέση τιμή 0, τότε μπορούμε να θεωρήσουμε πως τα διανύσματα **z*** και **α***, που περιέχουν τις αρχικές συνθήκες, είναι μηδενικά.

Σε περιπτώσεις που το AR πολυώνυμο έχει ρίζες κοντά στο σύνορο του μοναδιαίου κύκλου, και συνεπώς η στασιμότητα του μοντέλου δεν είναι δεδομένη, τότε οι προβλέψεις βάσει αυτών των αρχικών συνθηκών τείνουν να σφάλλουν. Σε τέτοιες περιπτώσεις, υπάρχει μέθοδος οι οποία μας δίνει την εκτιμήτρια αδέσμευτης μέγιστης πιθανοφάνειας, που δεν χρειάζεται αρχικές υποθέσεις, και αντίστοιχα παίρνουμε την εκτιμήτρια ελαχίστων ανεξάρτητων αθροισμάτων τετραγώνων, που σε αυτή την περίπτωση διαφέρουν μεταξύ τους, όμως για μεγάλο δείγμα τείνουν να δίνουν παρόμοιες τιμές.

5.3 Διαγνωστικοί Έλεγχοι Υπολοίπων

Εφόσον εκτιμήσουμε το κατάλληλο μοντέλο με σωστό πλήθος παραμέτρων βάσει των μεθόδων που είδαμε σε προηγούμενη ενότητα και αφού υπολογίσουμε τους συντελεστές μέγιστης πιθανοφάνειας για το εν λόγω μοντέλο, πρέπει να ελέγξουμε την εγκυρότητα του μοντέλου και να εξακριβώσουμε αν οι υποθέσεις που κάναμε πληρούνται.

Οι διαφορές που προκύπτουν από τις πραγματικές τιμές των παρατηρήσεων και τις εκτιμήσεις βάσει του μοντέλου που εκτιμήσαμε, ονομάζονται υπόλοιπα. Συγκεκριμένα τα υπόλοιπα έχουν την μορφή:

$$\hat{a}_t = Z_t - \hat{\varphi}_1 Z_{t-1} - \hat{\varphi}_2 Z_{t-2} - \ldots - \hat{\varphi}_p Z_{t-p} + \hat{\theta}_1 \hat{a}_{t-1} + \hat{\theta}_2 \hat{a}_{t-2} + \ldots + \hat{\theta}_q \hat{a}_{t-q} ,$$

όπου τα $\hat{\varphi}_1, \hat{\varphi}_2, \ldots, \hat{\varphi}_p, \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q$ είναι οι εκτιμήσεις μέγιστης πιθανοφάνειας.

Οι τιμές των υπολοίπων υπολογίζονται αναδρομικά από την πιο πάνω σχέση, θεωρώντας είτε τα αρχικά q υπόλοιπα ως μηδενικά (δεσμευμένη μέθοδος) είτε χρησιμοποιώντας την ακριβή μέθοδο οπίσθιας πρόβλεψης (αδέσμευτη μέθοδος).

Σε περίπτωση που το μοντέλο είναι κατάλληλο μπορεί να δειχθεί ότι:

$$\hat{a}_t = a_t + \boldsymbol{O}\left(\frac{1}{\sqrt{n}}\right)$$

Δηλαδή καθώς το μέγεθος του δείγματος n μεγαλώνει τότε οι τιμές των υπολοίπων συγκλίνουν στις τυχαίες διακυμάνσεις, που όπως είχαμε πει και στο παρελθόν είναι εξορισμού ανεξάρτητες τυχαίες μεταβλητές με μέση τιμή μηδέν και διασπορά σ_α^2.

Επομένως θα θέλαμε τα υπόλοιπα να πληρούν τις υποθέσεις και να συμπεριφέρονται ως ανεξάρτητες και τυχαία εκλεγμένες τιμές από την κανονική κατανομή με μέση τιμή 0 και διασπορά σ_α^2. Διαιρώντας τα υπόλοιπα με την θεωρητική εκτίμηση της ρίζας της διασποράς τους παίρνουμε τα προσαρμοσμένα υπόλοιπα R_t που ακολουθούν την N(0,1) κατανομή.

$$R_t = \frac{\hat{a}_t}{\sqrt{\frac{\sum_{i=1}^{n} \hat{a}_t^2}{n}}}$$

Ένας καλός πρώτος έλεγχος για να ελέγξουμε την υπόθεση κανονικότητας των σφαλμάτων είναι να ελέγξουμε γραφικά τις τιμές των προσαρμοσμένων υπολοίπων αν όντως ακολουθούν την κατανομή N(0,1). Πέρα του γραφικού ελέγχου μπορεί να χρησιμοποιηθούν έλεγχοι καλής προσαρμογής για την κανονική κατανομή όπως ο έλεγχος Kolmogorov-Smirnov ή ο έλεγχος Shapiro-Wilk.

Στην περίπτωση που η μορφή του μοντέλου που επιλέξαμε είναι η πραγματική (κάτι που είναι πρακτικά αδύνατο) και οι πραγματικές τιμές των συντελεστών ήταν σε διανυσματική μορφή τα **φ** και **θ**, μπορεί να αποδειχθεί ότι οι αναμενόμενες αυτοσυσχετίσεις $r_k(\alpha)$ θα είναι επίσης ασυσχέτιστες με μέση τιμή μηδέν και διασπορά ίση με 1/n (δηλαδή με τυπικό σφάλμα $1/\sqrt{n}$).

Στην πραγματικότητα όμως οι συντελεστές που επιλέξαμε για τα φ και θ, δεν είναι οι ίδιοι με τους πραγματικούς. Επομένως μπορούμε να υπολογίσουμε τις τιμές αυτοσυσχέτισης μεταξύ των υπολοίπων $r_k(\hat{a})$, και βάση του τυπικού σφάλματος να ελέγξουμε την υπόθεση ότι οι τιμές είναι μηδενικές (δηλαδή δεν υπάρχει αυτοσυσχέτιση για καμιά τιμή του k). Γραφικά μπορούμε να χρησιμοποιήσουμε τη συνάρτηση αυτοσυσχέτισης ACF για τις τιμές των υπολοίπων.

5.3.1 Έλεγχος Box-Pierce και Ljung-Box

Κατόπιν πολλών εφαρμογών και ελέγχων, διάφοροι ερευνητές κατέληξαν στο συμπέρασμα ότι το να ελέγξουμε κάθε συντελεστή αυτοσυσχέτισης βάσει του θεωρητικού τυπικού σφάλματος είναι μια κακή πρακτική. Συγκεκριμένα ο **James Durbin (1970)** ήταν ο πρώτος που έκρινε την πρακτική αυτή ως επικίνδυνη, καθώς κατάφερε να δείξει ότι στο μοντέλο AR(1) με παράμετρο φ, η διασπορά του $r_k(\hat{a})$ είναι φ^2/n, που μπορεί αρκετές φορές να είναι πολύ πιο μικρή από το 1/n. Για τον λόγο αυτό αναπτύχθηκαν οι σύνθετοι **(portmanteau)** έλεγχοι **Box-Pierce (1970)** και **Ljung-Box (1978)** που ελέγχουν την υπόθεση ανεξαρτησίας των σφαλμάτων χωρίς να λαμβάνουν υπόψη την κάθε τιμή αυτοσυσχέτισης ξεχωριστά, αλλά κοιτάζοντας την γενική εικόνα των πρώτων K στο πλήθος αυτοσυσχετίσεων.

Συγκεκριμένα η πρώτη μέθοδος που αποδείχτηκε και εφαρμόστηκε στα μοντέλα ARIMA τάξης (p,d,q) με N παρατηρήσεις, ήταν ο έλεγχος Box-Pierce ο οποίος χρησιμοποιεί σαν στατιστικό την τιμή:

$$Q = n\sum_{k=1}^{K} r_k^2(\hat{a}) \quad \text{όπου } n = N - d$$

Η τυχαία μεταβλητή Q ακολουθεί προσεγγιστικά την κατανομή χ^2 (Chi-Squared) με K-p-q βαθμούς ελευθερίας και επομένως υπολογίζοντας την τιμή του στατιστικού στο μοντέλο μας και ελέγχοντας την θεωρητική τιμή της κατανομής μπορούμε να ελέγξουμε την υπόθεση ανεξαρτησίας σφαλμάτων.

Μετά την εύρεση του πιο πάνω στατιστικού, παρατηρήθηκε από τους Ljung&Box ότι η υπόθεση ανεξαρτησίας σφαλμάτων δεν συμφωνούσε απόλυτα με το στατιστικό Q και πρότειναν μια νέα στατιστική συνάρτηση Q* η οποία τροποποιεί ελαφρώς το αρχικό στατιστικό, που ακολουθεί προσεγγιστικά την ίδια κατανομή, και βάση μελλοντικών δοκιμών κατέληξαν στο ότι δίνει παρόμοια δε, αλλά ακριβέστερα αποτελέσματα.

Η νέα στατιστική συνάρτηση Q* για το μοντέλο ARIMA (p,d,q) με N παρατηρήσεις είναι η ακόλουθη.

$$Q^* = n(n+2)\sum_{k=1}^{K}\frac{r_k^2(\hat{a})}{n-k} \quad \text{όπου } n = N - d$$

Αργότερα προτάθηκαν και άλλοι έλεγχοι, που δεν χρησιμοποιούν τις αυτοσυσχετίσεις αλλά τις μερικές αυτοσυσχετίσεις, με στατιστικό εντελώς ίδιο με την περίπτωση των αυτοσυσχετίσεων με την διαφοροποίηση ότι αντικαταστούμε τις τιμές r_k με φ_{kk} και το στατιστικό ακολουθεί ίδια κατανομή. Οι βασικοί έλεγχοι όμως κρίθηκαν ως αρκετά ικανοποιητικοί και αποτελούν το βασικότερο εργαλείο ελέγχου καλής προσαρμογής του μοντέλου (Goodness of Fit).

5.3.2 Εναλλακτικοί Έλεγχοι Τυχαιότητας Υπολοίπων

Άλλοι τρεις έλεγχοι, που ίσως φαίνονται πιο απλοί και κατανοητοί, θα χρησιμοποιηθούν κατά την μελέτη προβλημάτων στις επόμενες ενότητες. Οι έλεγχοι αυτή μας βοηθάνε να επιβεβαιώσουμε την τυχαιότητα και την ανεξαρτησία των υπολοίπων.

Α) Έλεγχος σημείων καμπής (Turning Points Test)

Σε μια ακολουθία n παρατηρήσεων $y_1,y_{2,},\ldots y_n$ λέμε ότι υπάρχει σημείο καμπής στο χρόνο i (1< i < n) αν ισχύει ότι $y_{i-1} < y_i$ και $y_i > y_{i+1}$ ή ισχύει ότι $y_{i-1} > y_i$ και $y_i < y_{i+1}$. Αν οι τιμές $y_1,y_{2,},\ldots y_n$ είναι ανεξάρτητες και τυχαία κατανεμημένες με μέση τιμή 0 και διασπορά σ^2, η πιθανότητα να έχουμε σημείο καμπής στο τυχαίο χρόνο i είναι 2/3. Ο λόγος είναι πως για κάθε τριάδα πραγματικών τυχαίων αριθμών, που συμβολίζουμε με y_{i-1},y_i και y_{i+1}, ένας εκ των τριών αριθμών θα είναι ο ελάχιστος και ένας θα είναι ο μέγιστος. Συνεπώς η πιθανότητα να υπάρξει σημείο καμπής στο χρόνο i είναι ίση με την πιθανότητα το y_i να είναι ο μέγιστος των τριών διαδοχικών όρων (δηλαδή 1/3) συν την πιθανότητα να είναι ο ελάχιστος των τριών διαδοχικών όρων (1/3).

Αν αριθμήσουμε και συμβολίσουμε με T το πλήθος των σημείων καμπής μιας ανεξάρτητης κανονικής διαδικασίας πλήθους n με μέσο 0 και διασπορά σ^2 τότε έχουμε n-2 πιθανά σημεία καμπής (αποκλείουμε τον πρώτο και τελευταίο όρο της διαδικασίας) και επομένως

$$\mu_T = E[T] = \frac{2}{3}(n-2)$$

Επίσης για την ίδια διαδικασία, αποδεικνύεται ότι η διασπορά του T είναι:

$$\sigma_T^2 = Var[T] = E[T^2] - (E[T])^2 = \frac{(16n-29)}{90}$$

Όταν η τιμή της μεταβλητής T βρίσκεται αρκετά πάνω από την αναμενόμενη της τιμή μ_T, τότε το μοντέλο μας πάλλεται πολύ περισσότερο από ότι θα περιμέναμε από μια τυχαία διαδικασία. Στην αντίθετη όμως περίπτωση, όπου η τιμή του T είναι αρκετά κάτω από την αναμενόμενη της τιμή, μας υποδεικνύεται ότι υπάρχει θετική συσχέτιση μεταξύ διαδοχικών παρατηρήσεων.

Όταν το δείγμα μας n είναι αρκετά μεγάλο, τότε από το κεντρικό οριακό θεώρημα καταλήγουμε στο ότι η τυχαία μεταβλητή T ακολουθεί κανονική κατανομή με μέση τιμή μ_T και διασπορά σ_T^2. Επομένως μπορούμε με συνηθισμένο έλεγχο t να κρίνουμε αν το πλήθος των σημείων καμπής είναι στατιστικά συμβατό με μια τυχαία διαδικασία.

Β) Έλεγχος πρόσημου διαφορών (Difference Sign Test)

Στην ίδια διαδικασία που περιγράψαμε πιο πάνω, συμβολίζουμε με S το πλήθος των παρατηρήσεων που ικανοποιούν την σχέση $y_i > y_{i-1}$ με i=1,2,…n. Ισοδύναμα δηλαδή μετράμε το πλήθος των παρατηρήσεων για τις οποίες ισχύει $y_i - y_{i-1} > 0$ η ισοδύναμα ότι το πρόσημο της διαφοράς τους από την προηγούμενη παρατήρηση είναι θετικό (από όπου και παίρνει το όνομα του ο έλεγχος). Για μια τυχαία διαδικασία ξέρουμε πως η αναμενόμενη τιμή και η διασπορά του S είναι:

$$\mu_S = E[S] = \frac{1}{2}(n-1) \quad \kappa\alpha\iota \quad \sigma_S^2 = Var[S] = \frac{n+1}{12}$$

Σε περίπτωση που η τιμή του S αποκλίνει πολύ από την αναμενόμενη του τιμή μ_S, καταλαβαίνουμε ότι υπάρχει θετική η αρνητική τάση στα δεδομένα μας (πράγμα που δεν θέλουμε να ισχύει σε πραγματικά τυχαίες διαδικασίες). Επομένως όπως και πριν, σύμφωνα με το κεντρικό οριακό θεώρημα όταν οι παρατηρήσεις μας είναι σχετικά αρκετές, η τυχαία μεταβλητή S ακολουθεί κανονική κατανομή με μέσο και διασπορά όπως πιο πάνω. Όπως και πριν μπορούμε να χρησιμοποιήσουμε το one-sample t-test για να αποφανθούμε για την σημαντικότητα της απόκλισης του S από την αναμενόμενη της τιμή.

Γ) Έλεγχος τάξης (Rank Test)

Όπως και πριν ορίζουμε ως P, το πλήθος των ζευγών (i,j) τέτοια ώστε y_j να είναι μεγαλύτερο του y_i και j μεγαλύτερο του i όπου i=1,…,n-1. Υπάρχουν n ανά 2 ζεύγη όπου το j > i και για τυχαία διαδικασία η πιθανότητα $p(y_j > y_i) = ½$. Επομένως η μέση τιμή του P και η διασπορά του φαίνονται πιο κάτω:

$$\mu_P = E[P] = \frac{1}{4}n(n-1) \quad \kappa\alpha\iota \quad \sigma_P^2 = Var[P] = \frac{n(n-1)(2n+5)}{72}$$

Μεγάλες θετικές η αρνητικές αποκλίσεις της τιμής P από την αναμενόμενη της τιμή μ_P μας υποδεικνύουν θετικές η αρνητικές τάσης στα δεδομένα. Θέλοντας να αποφύγουμε τέτοιες περιπτώσεις, ελέγχουμε πάλι με one-sample t-test την υπόθεση ότι η απόκλιση του S είναι στατιστικά σημαντική, πράγμα που θα έδειχνε ότι τα δεδομένα μας δεν είναι τυχαία.

5.4. Αλγόριθμος ARIMA Μοντέλων

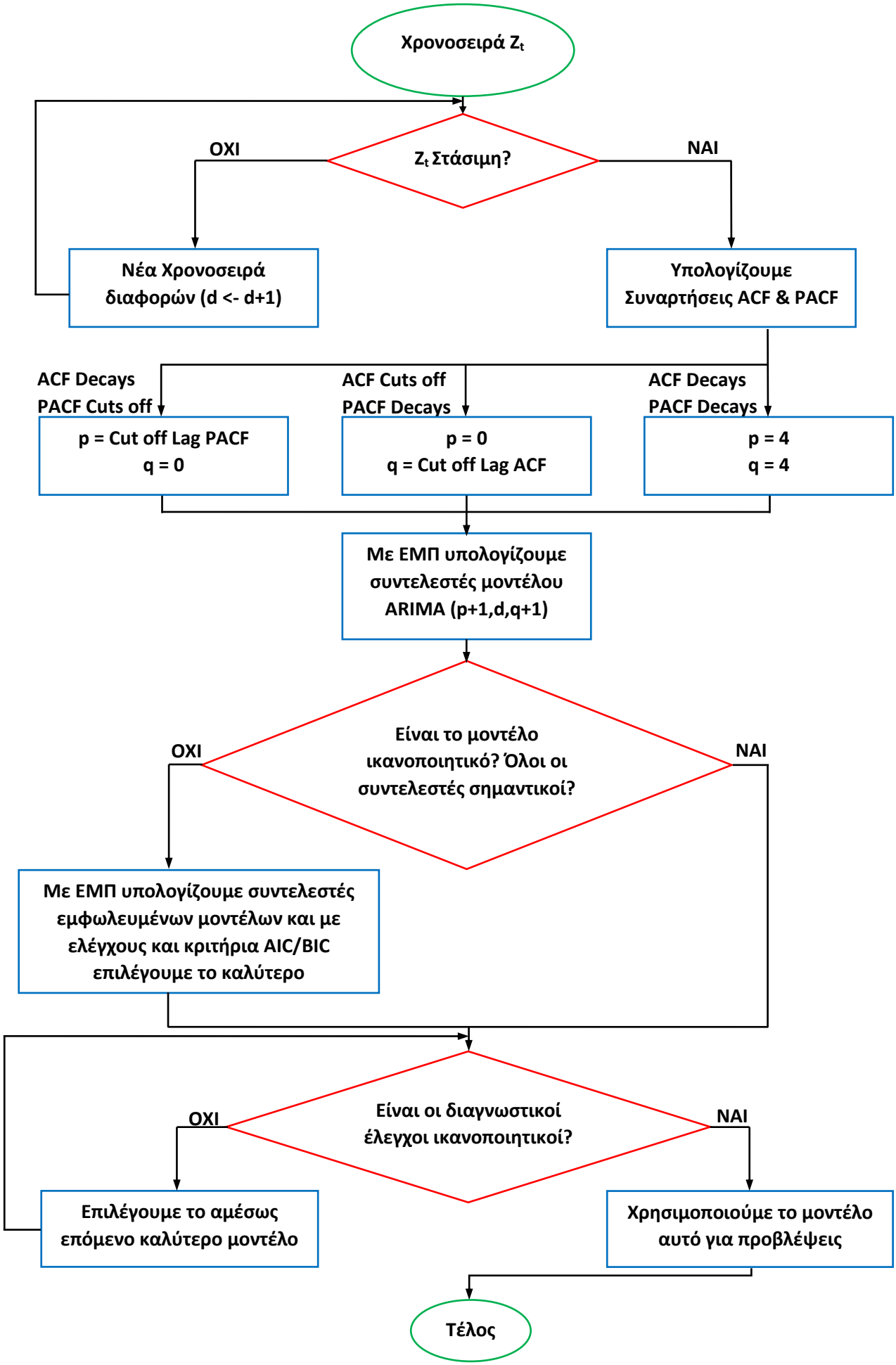
Το Διάγραμμα 5.2 της επόμενης σελίδας συνοψίζει όλα αυτά που έχουν συζητηθεί σε αυτή την ενότητα.

Αρχίζουμε με τον έλεγχο αν η διαδικασία είναι στάσιμη και σε περίπτωση που δεν είναι, παίρνουμε διαφορές μέχρι να καταλήξουμε σε στάσιμη σειρά. Το πόσες φορές θα επαναλάβουμε τη διαδικασία αυτή μέχρι να πετύχουμε στασιμότητα, μας καθορίζει το πλήθος της παραμέτρου d. Έπειτα με χρήση των γραφημάτων ACF και PACF λαμβάνουμε μια πρώτη εκτίμηση για τις παραμέτρους p και q.

Στην περίπτωση που τα γραφήματα ACF και PACF δεν μας παρουσιάζουν μια προφανή εικόνα, θέτουμε αυθαίρετα τις τιμές των μεταβλητών p και q να είναι ίσες με το 4, που είναι αρκετές ώστε να μπορούν να περιγράψουν κάθε είδους προβλήματα (συνήθως προτιμάμε πιο οικονομικά μοντέλα με πιο μικρό πλήθος παραμέτρων).

Στο επόμενο βήμα προσαρμόζουμε μοντέλα ARIMA (p+1,d,q+1) για να διευρύνουμε το πλήθος των μοντέλων που θα μελετηθούν καθώς και το υποψήφιο μοντέλο ARIMA (p,d,q) θα ελεγχθεί ως εμφωλευμένο του πρώτου.

Μέσα από αυτή την διαδικασία, επιλέγουμε το καλύτερο μοντέλο βάσει κριτηρίων, και δεδομένου ότι όλοι του οι συντελεστές είναι σημαντικοί, προχωρούμε σε διαγνωστικούς ελέγχους υπολοίπων. Σε περίπτωση που οι διαγνωστικοί έλεγχοι μας φανερώσουν καλή προσαρμογή, συνεχίζουμε στις προβλέψεις, εναλλακτικά επιλέγουμε ένα άλλο υποψήφιο μοντέλο και συνεχίζουμε με τις ίδιες διαδικασίες.



Διάγραμμα 5.2

6. Εφαρμοσμένα Προβλήματα

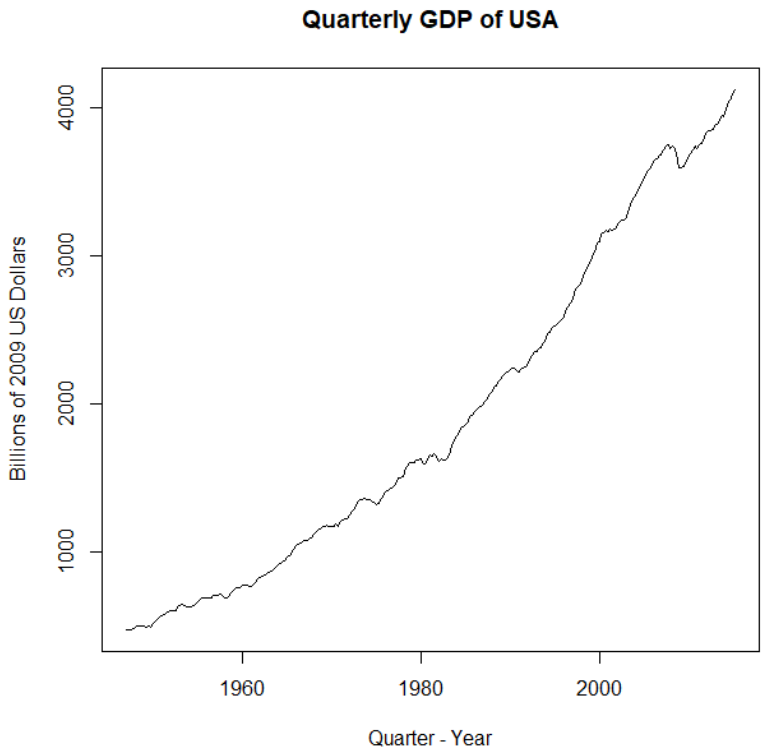
Στην ενότητα αυτή θα μελετήσουμε προβλήματα χρονοσειρών πραγματικών δεδομένων, χρησιμοποιώντας τις μεθόδους που σχολιάστηκαν στις προηγούμενες ενότητες.

Θα μελετηθούν τρία προβλήματα διαφορετικής φύσεως για τα οποία θα χρησιμοποιήσουμε μονοδιάστατα μοντέλα (ARIMA) και δύο προβλήματα για τα οποία θα χρησιμοποιήσουμε μια εξωγενή επεξηγηματική μεταβλητή και επομένως θα χρησιμοποιήσουμε μοντέλα ADL.

6.1 Μελέτη του ΑΕΠ για τις ΗΠΑ

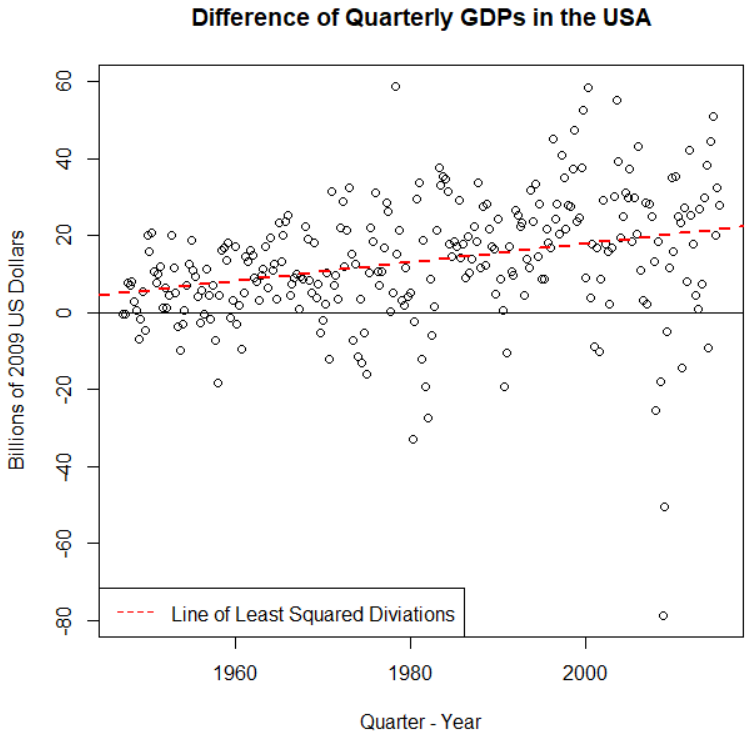
Τα δεδομένα μας σε αυτή την μελέτη περιλαμβάνουν το τριμηνιαίο Ακαθάριστο Εγχώριο Προϊόν (ΑΕΠ), καταγεγραμμένο σε τρίμηνα από το 1947 μέχρι και το 2017. Τα δεδομένα έχουν προσαρμοστεί ως προς την εποχικότητα τους, και έχουν αποπληθωριστεί, με έτος αναφοράς αξίας του δολαρίου, για το 2009. Η πηγή από όπου αντλήθηκαν είναι η ιστοσελίδα της Federal Reserve Bank of St. Louis.

Θέλοντας να ελέγξουμε αν οι προβλέψεις μας ανταποκρίνονται στα πραγματικά δεδομένα, αφαιρέσαμε τις τελευταίες 10 τιμές (των τελευταίων 2.5 ετών) και θα τις θεωρούμε άγνωστες μέχρι και το τέλος της ανάλυσης μας, όπου θα τις συγκρίνουμε με τις προβλέψεις.



Διάγραμμα 6.1

Το Διάγραμμα 6.1 παρουσιάζει το ποσό του ΑΕΠ ανά χρονική περίοδο (τρίμηνο). Από την γραφική παράσταση αυτή, είναι προφανές ότι τα δεδομένα μας δεν παρουσιάζουν στασιμότητα καθώς υπάρχει προφανή ανοδική τάση, επομένως στάσιμα μοντέλα δεν μπορούν να τα περιγράψουν. Όπως έχουμε συζητήσει σε προηγούμενες ενότητες, θα χρησιμοποιήσουμε μοντέλα ARIMA για να περιγράψουμε τέτοιες διαδικασίες. Δημιουργούμε συνεπώς μια νέα διαδικασία ή οποία παίρνει τις διαφορές πρώτης τάξης αυτής της χρονοσειράς και δημιουργούμε την αντίστοιχη γραφική παράσταση που φαίνεται στο Διάγραμμα 6.2.



Διάγραμμα 6.2

Σε αυτό το σημείο καλό θα ήταν να σημειώσουμε ότι κάθε φορά που δημιουργούμε νέα χρονοσειρά διαφορών, το πλήθος των παρατηρήσεων μας μειώνεται κατά ένα (η πρώτη παρατήρηση χάνεται).

Στο Διάγραμμα 6.2, λόγω του ότι δεν ήταν προφανές αν τα δεδομένα μας αποτελούν μια στάσιμη διαδικασία, προσθέσαμε την ευθεία ελαχίστων τετραγώνων των παρατηρήσεων (με διακεκομμένη γραμμή) για να αντιληφθούμε καλύτερα αν τα δεδομένα είναι στάσιμα και ποιο είναι περίπου το επίπεδο της διακύμανσής τους.

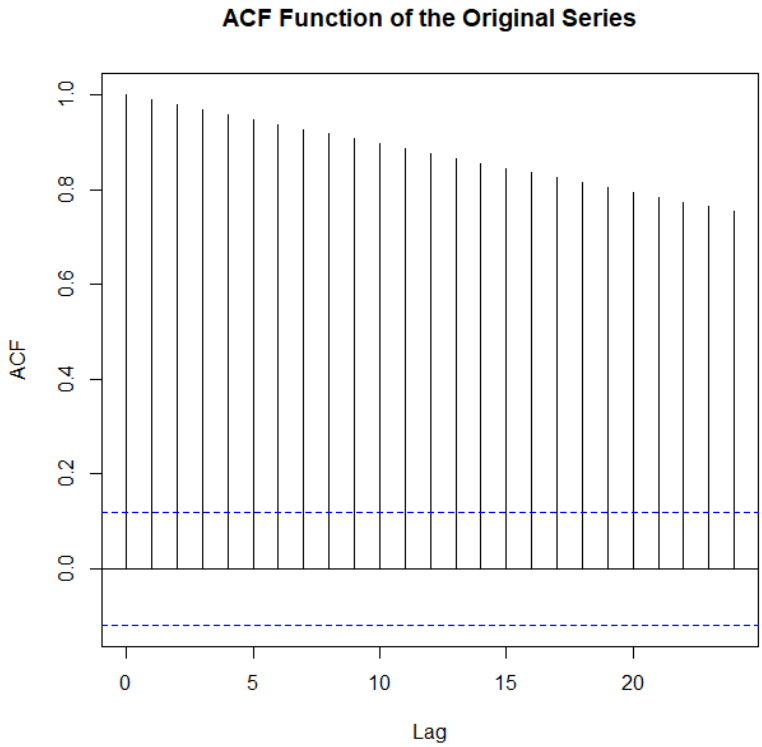
Λαμβάνοντας υπόψη τα παραπάνω, τα δεδομένα μας φαίνεται να έχουν μια ανοδική τάση και συνεπώς η στασιμότητα δεν είναι δεδομένη. Επίσης το επίπεδο της διαδικασίας φαίνεται υψηλό (γύρω στις 20 μονάδες). Για τον λόγω αυτό, δημιουργούμε το αντίστοιχο Διάγραμμα 6.3 με τις διαφορές δεύτερης τάξης.

Διάγραμμα 6.3

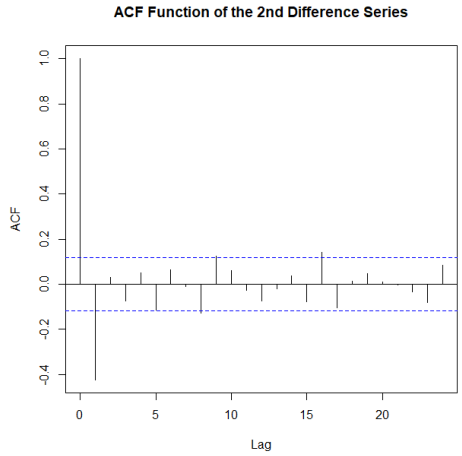
Είναι προφανές ότι τα δεδομένα παρουσιάζουν στασιμότητα (γύρω από το 0). Αυτό αποδεικνύεται πέρα από το Διάγραμμα 6.3 και από το **Augmented Dickey Fuller Test** που θεωρεί ως μηδενική υπόθεση ότι τα δεδομένα μας δεν παρουσιάζουν στασιμότητα και έχει p-value μικρότερη του 0.01, και συνεπώς με αρκετή βεβαιότητα δεχόμαστε την εναλλακτική υπόθεση στασιμότητας. Συγκεκριμένα ο έλεγχος αυτός γίνεται με την εντολή adf.test της βιβλιοθήκης tseries της R. Επομένως ο βαθμός Integration στο μοντέλο ARIMA(p,d,q), δηλαδή η τιμή της παραμέτρου d, θα είναι 2.

Το γεγονός ότι τα αρχικά δεδομένα μας δεν παρουσιάζουν στασιμότητα φαίνεται και από το γράφημα ACF των αρχικών παρατηρήσεων, που φαίνεται στο Διάγραμμα 6.4, που φθίνει μεν, αλλά με πάρα πολύ αργό ρυθμό (θα έπρεπε να έφθινε εκθετικά αν η διαδικασία ήταν στάσιμη).

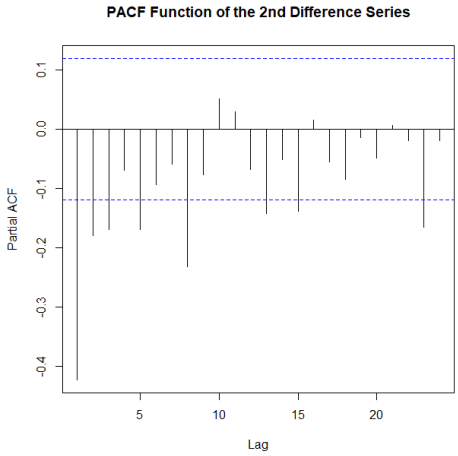
Θέλοντας να εκτιμήσουμε τώρα το πλήθος των συντελεστών p και q που θα χρησιμοποιήσουμε, βλέπουμε στα Διάγραμμα 6.5 και 6.6 τα γραφήματα ACF και PACF αντίστοιχα της στάσιμης διαδικασίας των δεύτερων διαφορών. Σύμφωνα με αυτά που σχολιάστηκαν σε προηγούμενες ενότητες, μπορούμε αναλύοντας την συμπεριφορά των γραφημάτων αυτών, να κάνουμε αρχικές εκτιμήσεις για το πλήθος των παραμέτρων του μοντέλου.



Διάγραμμα 6.4



Διάγραμμα 6.5



Διάγραμμα 6.6

Από το Διάγραμμα 6.5 βλέπουμε ότι έχουμε τους πρώτους 2 όρους της ACF συνάρτησης να είναι σημαντικά διάφοροι του μηδενός, πράγμα που μας υποδεικνύει ότι η αρχική πιθανή τιμή του q είναι 2. Αντίστοιχα από το Διάγραμμα 6.6 βλέπουμε ότι ο πρώτος όρος είναι προφανώς διάφορος του μηδενός, ενώ οι 2 μεταγενέστεροι όροι είναι πιθανό να είναι διάφοροι του μηδενός, επομένως περιμένουμε ότι στο μοντέλο μας το p θα είναι 1, 2 ή 3.

Η εικόνα που παίρνουμε από τα γραφήματα αυτά είναι λίγο θολή. Επομένως θεωρούμε ορθότερο να ξεκινήσουμε από το γενικό μοντέλο ARIMA (3,2,3) και να ελέγξουμε τα εμφωλευμένα μοντέλα του.

Επομένως προσαρμόζοντας διαδοχικά τα πιο κάτω μοντέλα παίρνουμε τα εξής αποτελέσματα:

ARIMA (3,2,3):

```
> arima(price, order=c(3,2,3))

Call:
arima(x = price, order = c(3, 2, 3))

Coefficients:
          ar1     ar2     ar3      ma1      ma2     ma3
      -0.2408  0.4403  0.0347  -0.4204  -0.6498  0.1019
s.e.   0.8584  0.2407  0.2346   0.8561   0.6596  0.3268

sigma^2 estimated as 213.2:  log likelihood = -1116.29,  aic = 2246.59
```

Λόγω μεγάλων τυπικών σφαλμάτων των εκτιμήσεων, δεν μπορούμε για κανένα συντελεστή να πούμε με βεβαιότητα ότι είναι διάφορος του μηδενός. Επομένως δοκιμάζουμε να μειώσουμε τους συντελεστές p kai q.

ARIMA (2,2,3):

```
> arima(price, order=c(2,2,3))

Call:
arima(x = price, order = c(2, 2, 3))

Coefficients:
          ar1     ar2      ma1      ma2     ma3
      -0.1196  0.4416  -0.5406  -0.5730  0.1416
s.e.   0.3085  0.2305   0.3192   0.4194  0.1745

sigma^2 estimated as 213.2:  log likelihood = -1116.3,  aic = 2244.59
```

Αρκετά μεγάλα τυπικά σφάλματα. Ο μοναδικός συντελεστής που είναι διάφορος του μηδενός (με οριακή σημαντικότητα) είναι ο δεύτερος συντελεστής AR. Επομένως αντίστοιχα θα δοκιμάσουμε να ελαττώσουμε το πλήθος των παραμέτρων.

ARIMA (1,2,3):

```
> arima(price, order=c(1,2,3))

Call:
arima(x = price, order = c(1, 2, 3))

Coefficients:
         ar1      ma1     ma2      ma3
      0.5377  -1.2025  0.3077  -0.0860
s.e.  0.1743   0.1800  0.1421   0.0803

sigma^2 estimated as 213.6:  log likelihood = -1116.54,  aic = 2243.08
```

Αρκετά καλή προσαρμογή. Έχουμε σημαντικότητα στον όρο AR και στους 2 πρώτους MA όρους. Επίσης έχουμε ελάττωση των κριτηρίων AIC και BIC. Θα δοκιμάσουμε να ελαττώσουμε το μοντέλο κατά ένα MA όρο.

ARIMA (3,2,2):

```
> arima(price, order=c(3,2,2))

Call:
arima(x = price, order = c(3, 2, 2))

Coefficients:
          ar1     ar2     ar3      ma1      ma2
      -0.3936  0.3888  0.0863  -0.2656  -0.6962
s.e.   0.6168  0.1962  0.1338   0.6111   0.5985

sigma^2 estimated as 213.3:  log likelihood = -1116.35,  aic = 2244.7
```

Όμοια σχόλια με το ARIMA (2,2,3).

ARIMA (3,2,1):

```
> arima(price, order=c(3,2,1))

Call:
arima(x = price, order = c(3, 2, 1))

Coefficients:
         ar1     ar2      ar3      ma1
      0.3145  0.1602  -0.0065  -0.9780
s.e.  0.0616  0.0633   0.0613   0.0128

sigma^2 estimated as 213.6:  log likelihood = -1116.53,  aic = 2243.06
```

Όμοια σχόλια με το ARIMA (1,2,3) αλλά με μη σημαντικό τον τρίτο όρο AR. Και πάλι θα δοκιμάσουμε να ελαττώσουμε το p κατά 1.

ARIMA (2,2,2):

```
> arima(price, order=c(2,2,2))

Call:
arima(x = price, order = c(2, 2, 2))

Coefficients:
         ar1     ar2      ma1      ma2
      0.2514  0.1807  -0.9152  -0.0612
s.e.  0.4662  0.1749   0.4758   0.4627

sigma^2 estimated as 213.6:  log likelihood = -1116.53,  aic = 2243.06
```

Όμοια σχόλια με το ARIMA (2,2,3).

ARIMA (2,2,1):

```
> arima(price, order=c(2,2,1))

Call:
arima(x = price, order = c(2, 2, 1))

Coefficients:
         ar1     ar2      ma1
      0.3136  0.1584  -0.9782
s.e.  0.0611  0.0610   0.0125

sigma^2 estimated as 213.6:  log likelihood = -1116.54,  aic = 2241.07
```

Αυτό είναι το πρώτο μοντέλο που έχει όλους τους συντελεστές του σημαντικά διάφορους του 0. Σε σχέση με τα μοντέλα ARIMA (3,2,1) και (1,2,3), το κριτήριο AIC έχει μειωθεί κατά 2 μονάδες ενώ το κριτήριο BIC έχει μειωθεί κατά 6 περίπου μονάδες. Είναι συνεπώς υποψήφιο μοντέλο προς επιλογή.

ARIMA (1,2,2):

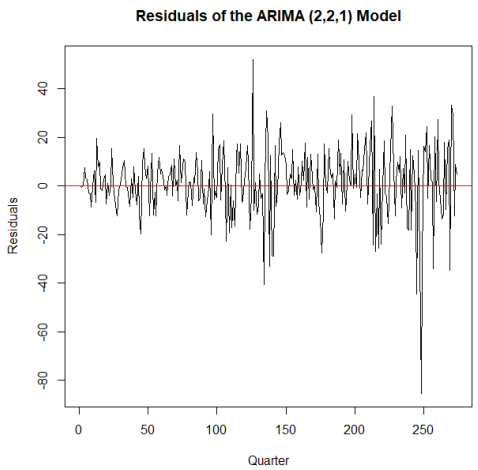
```
> arima(price, order=c(1,2,2))

Call:
arima(x = price, order = c(1, 2, 2))

Coefficients:
         ar1      ma1     ma2
      0.6536  -1.3040  0.3182
s.e.  0.1044   0.1243  0.1197

sigma^2 estimated as 214.5:  log likelihood = -1117.08,  aic = 2242.15
```

Ένα άλλο πολύ καλό μοντέλο με όλους τους συντελεστές σημαντικούς και ελάχιστα ψηλότερα επίπεδα κριτηρίου επιλογής AIC και BIC με το μοντέλο ARIMA (2,2,1).

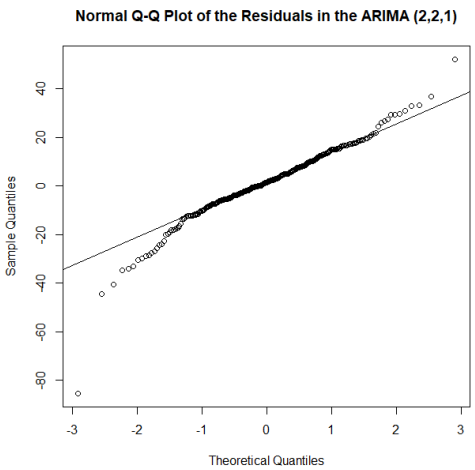
Επομένως θα μπορούσαμε να πούμε ότι τα δυο τελευταία μοντέλα που σχολιάστηκαν είναι ικανοποιητικά. Θα αρχίσουμε από το μοντέλο ARIMA (2,2,1), που αφού γνωρίζουμε τους συντελεστές του, μπορούμε να το γράψουμε ως:

$$Z_t = 0.3136Z_{t-1} + 0.1584Z_{t-2} + a_t - 0.97a_{t-1} \ \ (\mu\varepsilon\ Z_t = \boldsymbol{\Delta}^2 Y_t).$$

Στο Διάγραμμα 6.7 βλέπουμε ότι τα υπόλοιπα φαίνεται να κατανέμονται σχετικά κανονικά γύρω από το 0, με μοναδική εξαίρεση ότι ίσως η διασπορά τους να αυξάνεται σε βάθος χρόνου. Με την χρήση του κλασικού QQ Plot για τα υπόλοιπα (Διάγραμμα 6.8), βλέπουμε ότι τα υπόλοιπα φαίνονται κανονικά με κάποιες αποκλίσεις στις ουρές. Χρησιμοποιώντας το Shapiro-Wilkinson Normality Test για τα σφάλματα έχουμε p-value της τάξης του $1x10^{-8}$. Το αντίστοιχο One-Sample Kolmogorov-Smirnov Test για την κανονική κατανομή μας δίνει p-value της τάξης του $1x10^{-16}$. Επομένως δεχόμαστε ότι τα υπόλοιπα βάση του μοντέλου αυτού, είναι κανονικά κατανεμημένα.

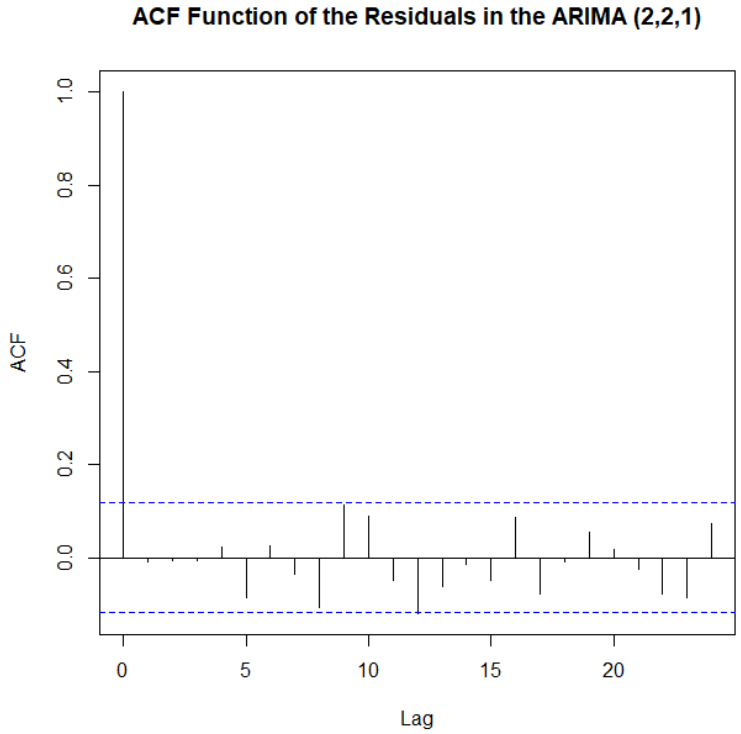


Διάγραμμα 6.7



Διάγραμμα 6.8

Μπορούμε να υπολογίζουμε τα προσαρμοσμένα υπόλοιπα (διαιρώντας με την τυπική τους απόκλιση) και να κοιτάξουμε το γράφημα ACF για τα προσαρμοσμένα υπόλοιπα για να ελέγξουμε την υπόθεση ανεξαρτησίας των υπολοίπων.



Διάγραμμα 6.9

Στο Διάγραμμα 6.9 βλέπουμε ότι καμία τιμή των αυτοσυσχετίσεων των υπολοίπων δεν είναι σημαντική (πέρα της πρώτης που πάντα είναι ίση με την μονάδα). Όλες οι τιμές βρίσκονται εντός του διαστήματος εμπιστοσύνης που παριστάνεται με τις μπλε διακεκομμένες γραμμές. Έτσι έχουμε μια αρχική ένδειξη ότι τα υπόλοιπα του μοντέλου μας παρουσιάζουν ανεξαρτησία.

Ορθότερο όμως είναι, όπως σχολιάστηκε και στην ενότητα 5.3.1, να μην κοιτάζουμε τον κάθε όρο της ACF συνάρτησης ξεχωριστά, αλλά να τους δούμε συνολικά, και με την χρήση των ελέγχων Ljung-Box και Box-Pierce να αποφανθούμε για την ανεξαρτησία των υπολοίπων. Ο έλεγχος Ljung-Box μας δίνει p-value 0.289 με μηδενική υπόθεση την ανεξαρτησία των υπολοίπων και εναλλακτική το ότι τα υπόλοιπα μας δεν είναι ανεξάρτητα. Επομένως δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση της ανεξαρτησίας. Αντίστοιχα ο έλεγχος Box-Pierce μας δίνει p-value ίση με 0.3193 και πάλι δεν απορρίπτουμε την μηδενική υπόθεση ανεξαρτησίας υπολοίπων.

Τέλος, κάνοντας τους διαγνωστικούς ελέγχους της ενότητας 5.3.2, κανένας έλεγχος δεν μπορεί να απορρίψει την υπόθεση κανονικότητας και ούτε να αποδείξει την ύπαρξη τάσεων στα υπόλοιπα.

Επομένως καταλήγουμε στο ότι το μοντέλο ARIMA (2,2,1) είναι ένα αρκετά καλό μοντέλο που περιγράφει ικανοποιητικά τα δεδομένα μας και ακολούθως θα είναι χρήσιμο και συνεπές για προβλέψεις μελλοντικών παρατηρήσεων.

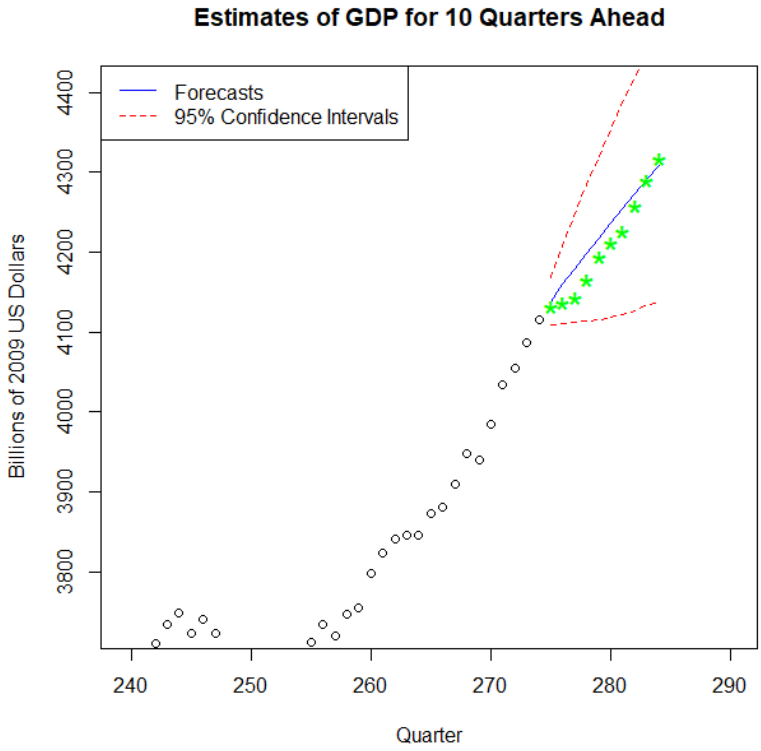
Μπορούμε να προβλέψουμε βάση του μοντέλου αυτού τις τιμές της αρχικής σειράς (ΑΕΠ) για τις μελλοντικές 10 παρατηρήσεις (τα επόμενα 2.5 έτη). Συγκεκριμένα, μπορούμε να υπολογίσουμε τα ποσά του Πίνακα 6.1 με την χρήση της εντολής forecast από το πακέτο Forecast της R. Βλέπουμε στις πρώτες στήλες τις σημειακές εκτιμήσεις καθώς και τα όρια των 95% και 80% διαστημάτων εμπιστοσύνης, και στην τελευταία στήλη τις πραγματικές τιμές που παρατηρήθηκαν στα τελευταία 10 τρίμηνα.

Date	Forecast	80% Low	80% High	95% Low	95% High	Actual
3rd 2015	4138.63	4119.90	4157.37	4109.99	4167.28	4131.90
4th 2015	4159.95	4129.37	4191.20	4112.16	4207.74	4136.91
1st 2016	4179.91	4136.35	4223.48	4113.28	4246.55	4142.89
2nd 2016	4199.13	4144.21	4254.05	4115.13	4283.13	4165.88
3rd 2016	4217.90	4152.41	4283.38	4117.74	4318.05	4194.54
4th 2016	4236.40	4161.06	4311.74	4121.17	4351.63	4212.86
1st 2017	4254.75	4170.13	4339.37	4125.34	4384.17	4225.81
2nd 2017	4273.02	4179.60	4366.43	4130.15	4415.88	4257.77
3rd 2017	4291.23	4189.40	4393.05	4135.50	4446.96	4290.97
4th 2017	4309.41	4199.50	4419.32	4141.31	4477.51	4318.12

Πίνακας 6.1

Οι προβλέψεις μας επομένως φαίνονται να συμβαδίζουν αρκετά με τις πραγματικές τιμές. Τέλος δίνεται μια γραφική παράσταση στο Διάγραμμα 6.10, που μας δείχνει όλες μας τις παρατηρήσεις, με μπλε γραμμή δείχνει τις σημειακές μας προβλέψεις, οι κόκκινες γραμμές δείχνουν τα όρια των 95% διαστημάτων εμπιστοσύνης, και με πράσινους αστερίσκους συμβολίζουμε τις πραγματικές τιμές που παρατηρήθηκαν.

Οι ίδιοι έλεγχοι και προβλέψεις έγιναν και για το μοντέλο ARIMA (1,2,2) που είχε ανάλογη προσαρμογή με το μοντέλο που χρησιμοποιήθηκε, ενώ είχε ελάχιστα (μικρότερες της μονάδας) υψηλότερες τιμές κριτηρίων AIC και BIC και καταλήξαμε σε πάρα πολύ παρόμοια αποτελέσματα και προβλέψεις, επομένως όποιο μοντέλο εκ των δυο χρησιμοποιηθεί, θα έχει αρκετά αξιόπιστα αποτελέσματα. Λόγω του ότι τα δυο μοντέλα συμπεριφέρονταν με σχεδόν ίδιο τρόπο, αποφασίσαμε να μην συμπεριλάβουμε τα αποτελέσματα και από το δεύτερο μοντέλο.



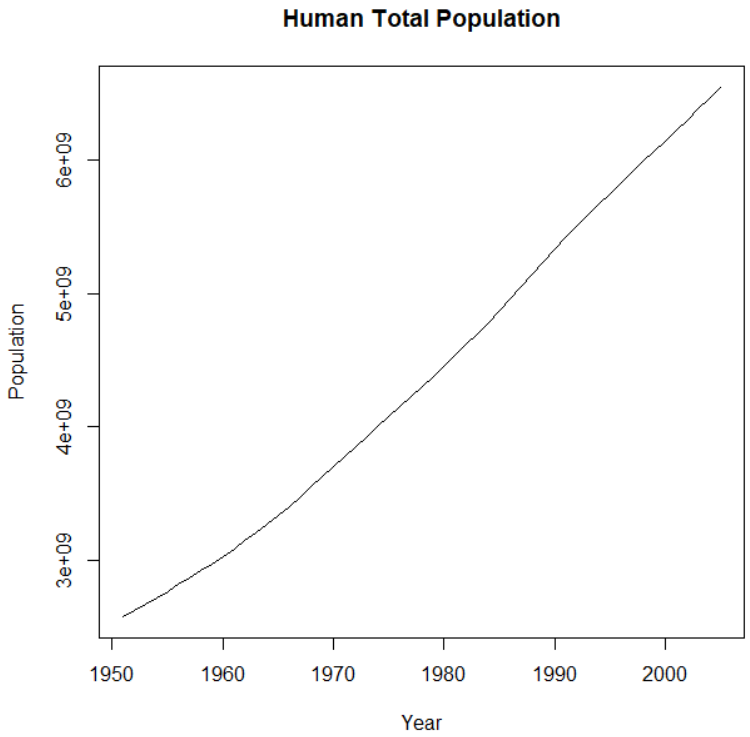
Διάγραμμα 6.10

6.2 Μελέτη για τον Συνολικό Πληθυσμό της Γης

Σε αυτό το πρόβλημα, θα ασχοληθούμε με τις προβλέψεις του μελλοντικού συνολικού πληθυσμού της Γης. Τα δεδομένα που θα χρησιμοποιηθούν είναι ετήσια με αρχή το 1951 έως το 2017. Όπως και στις προηγούμενες ενότητες, θα αγνοήσουμε αρχικά τις τελευταίες 12 παρατηρήσεις και θα χρησιμοποιήσουμε τις παρατηρήσεις μέχρι και το 2005. Έπειτα θα εκτιμήσουμε βάσει του μοντέλου το επίπεδο του πληθυσμού για τα έτη 2006-2017 και θα τα συγκρίνουμε με τα δεδομένα που παρατηρήθηκαν τις αντίστοιχες χρονιές. Όλα τα δεδομένα έχουν αντληθεί από την ιστοσελίδα www.worldometers.info.

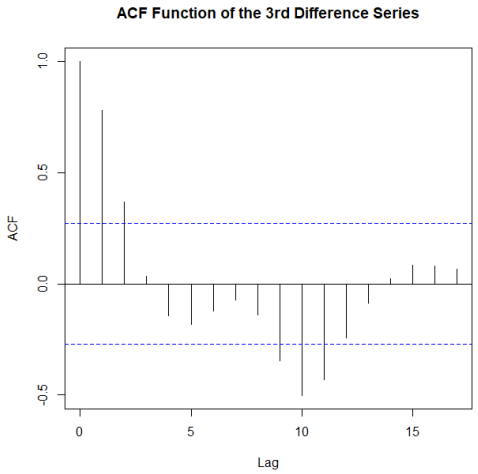
Όπως και πριν θα ελέγξουμε κατά πόσο τα δεδομένα μας παρουσιάζουν στασιμότητα. Βλέποντας το Διάγραμμα 6.11 είναι προφανές ότι υπάρχει μεγάλη αυξητική τάση στα δεδομένα και συνεπώς δεν παρουσιάζουν στασιμότητα. Μπορούμε να επιβεβαιώσουμε το γεγονός αυτό κάνοντας τον έλεγχο Augmented Dickey Fuller.

Επομένως αφού τα δεδομένα μας δεν είναι στάσιμα, παίρνουμε διαφορές και επαναλαμβάνουμε τους ελέγχους. Διαπιστώνουμε ότι για να πετύχουμε στασιμότητα σε αυτή την περίπτωση, απαιτείτε να πάρουμε διαφορές και τρίτης τάξης. Επομένως η παράμετρος Integration, δηλαδή το d στο μοντέλο ARIMA (p,d,q), θα είναι 3.

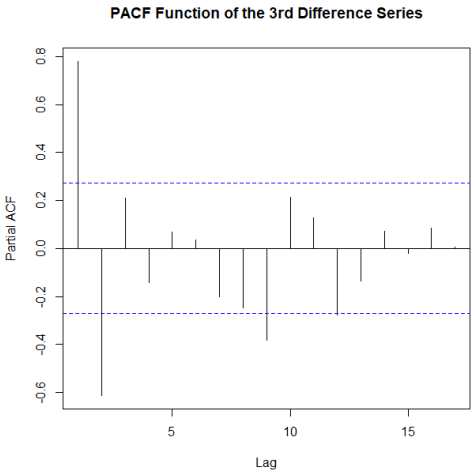


Διάγραμμα 6.11

Έπειτα, για να εκτιμήσουμε το πλήθος των παραμέτρων p και q, κοιτάμε την συμπεριφορά των ACF και PACF συναρτήσεων μέσω των ανάλογων γραφημάτων που παρουσιάζονται στα Διαγράμματα 6.12 και 6.13 αντίστοιχα.



Διάγραμμα 6.12



Διάγραμμα 6.13

Επομένως η αρχική μας εκτίμηση για την τιμή του q είναι 3 λόγω του ότι στο Διάγραμμα 6.12 παρουσιάζονται σημαντικές οι πρώτες 3 τιμές ενώ οι υπόλοιπες θα μπορούσαν να θεωρηθούν μηδενικές (με εξαίρεση τις τιμές για lag=9-11 που βγαίνουν εκτός των θεωρητικών ορίων).

Αντίστοιχα από το Διάγραμμα 6.13 μπορούμε να εξαγάγουμε το συμπέρασμα ότι το πλήθος των AR όρων, δηλαδή η τιμή του p στο μοντέλο ARIMA (p,d,q), θα είναι 2.

Σύμφωνα με τα πιο πάνω το πρώτο υποψήφιο μοντέλο είναι το ARIMA (2,3,3). Προσαρμόζοντας μέσο της R το μοντέλο αυτό παίρνουμε τον εξής πίνακα αναφοράς:

```
> arima(population, order=c(2,3,3))

Call:
arima(x = population, order = c(2, 3, 3))

Coefficients:
          ar1     ar2     ma1     ma2     ma3
      -0.3599  0.3213  2.2670  1.8845  0.5015
s.e.   0.4015  0.1420  0.4506  0.8020  0.4523

sigma^2 estimated as 2.812e+10:  log likelihood = -703.61,  aic = 1419.23
```

Βλέπουμε πως για τον όρο AR1 και για τον όρο MA3 δεν μπορούμε σε επίπεδο σημαντικότητας 95% να αποφανθούμε ότι είναι διάφοροι του μηδενός. Συνεπώς είναι πολύ πιθανό να υπάρχει ένα απλούστερο μοντέλο (εμφωλευμένο σε αυτό) που να εξηγεί το ίδιο καλά τα δεδομένα. Σημειώνουμε επίσης ότι το κριτήριο επιλογής BIC για το μοντέλο αυτό έχει την τιμή 1430.93. Επομένως μελετάμε τα πιο κάτω μοντέλα:

ARIMA (2,3,2):

```
> arima(population, order=c(2,3,2))

Call:
arima(x = population, order = c(2, 3, 2))

Coefficients:
         ar1     ar2     ma1     ma2
      0.1325  0.2188  1.7384  0.9974
s.e.  0.1512  0.1589  0.0866  0.0833

sigma^2 estimated as 2.874e+10:  log likelihood = -703.93,  aic = 1417.86
```

Στο μοντέλο αυτό οι δυο όροι AR φαίνονται μη-σημαντικοί σε αντίθεση με τους 2 MA όρους που είναι σημαντικά διάφοροι του μηδενός. Τα κριτήρια AIC και BIC μειώθηκαν και συνεπώς συνεχίζουμε την μελέτη μας.

ARIMA (1,3,3):

```
> arima(population, order=c(1,3,3))

Call:
arima(x = population, order = c(1, 3, 3))

Coefficients:
         ar1     ma1     ma2      ma3
      0.5248  1.3915  0.4507  -0.3178
s.e.  0.3005  0.3307  0.5574   0.3217

sigma^2 estimated as 2.925e+10:  log likelihood = -704.53,  aic = 1419.06
```

Ο όρος AR1 οριακά δεν είναι σημαντικός, το ίδιο και οι 2 τελευταίοι όροι MA. Αν και οι τιμές BIC και AIC έχουν μειωθεί η μελέτη μας συνεχίζεται.

ARIMA (1,3,2):

```
> arima(population, order=c(1,3,2))

Call:
arima(x = population, order = c(1, 3, 2))

Coefficients:
         ar1     ma1     ma2
      0.2098  1.7046  0.9955
s.e.  0.1385  0.1009  0.1066

sigma^2 estimated as 2.964e+10:  log likelihood = -704.9,  aic = 1417.8
```

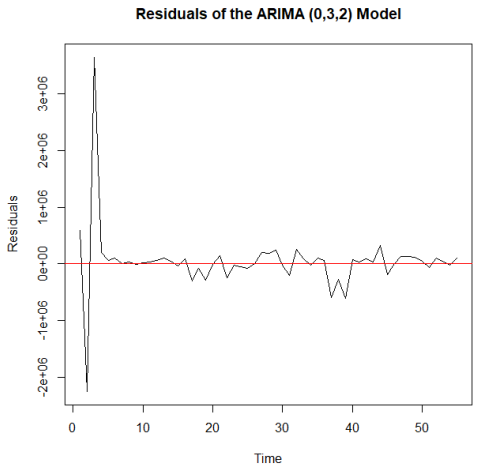
Το μοντέλο αυτό είναι καλύτερο από το ARIMA (1,3,3), έχει μικρότερες τιμές AIC και BIC, και έχει σημαντικούς και τους 2 MA συντελεστές του. Παρόλα αυτά για τον AR συντελεστή του δεν μπορούμε σε επίπεδο σημαντικότητας 95% να δεχθούμε ότι είναι διάφορος του μηδενός. Επομένως πρέπει να ελέγξουμε και μοντέλα χωρίς AR όρους.

ARIMA (0,3,2):

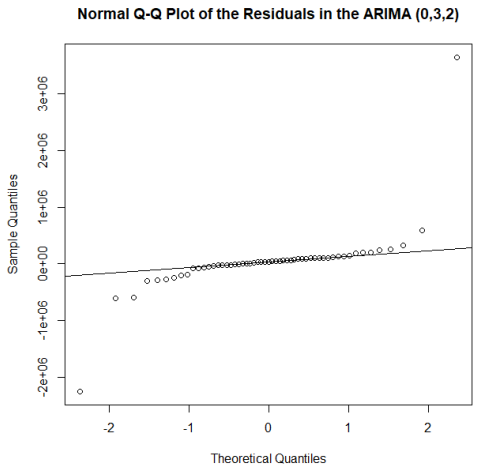
```
> arima(population, order=c(0,3,2))

Call:
arima(x = population, order = c(0, 3, 2))

Coefficients:
         ma1     ma2
      1.7238  0.9977
s.e.  0.0829  0.0827

sigma^2 estimated as 3.123e+10:  log likelihood = -705.98,  aic = 1417.97
```

Το μοντέλο αυτό είναι το μοντέλο με το χαμηλότερο BIC και έχει όλους τους συντελεστές τους σημαντικούς. Φυσικά η διαφορά της BIC τιμής του έναντι το προηγούμενου μοντέλου είναι μόλις 2 μονάδες. Επομένως και τα δυο μοντέλα πιθανώς να περιγράφουν ικανοποιητικά τα δεδομένα μας.

Θα συνεχίσουμε με την χρήση του μοντέλου ARIMA (0,3,2). Ελέγχουμε επομένως αν τα υπόλοιπα του μοντέλου αυτού είναι ανεξάρτητα και ακολουθούν κανονική κατανομή με μέση τιμή 0.



Διάγραμμα 6.13

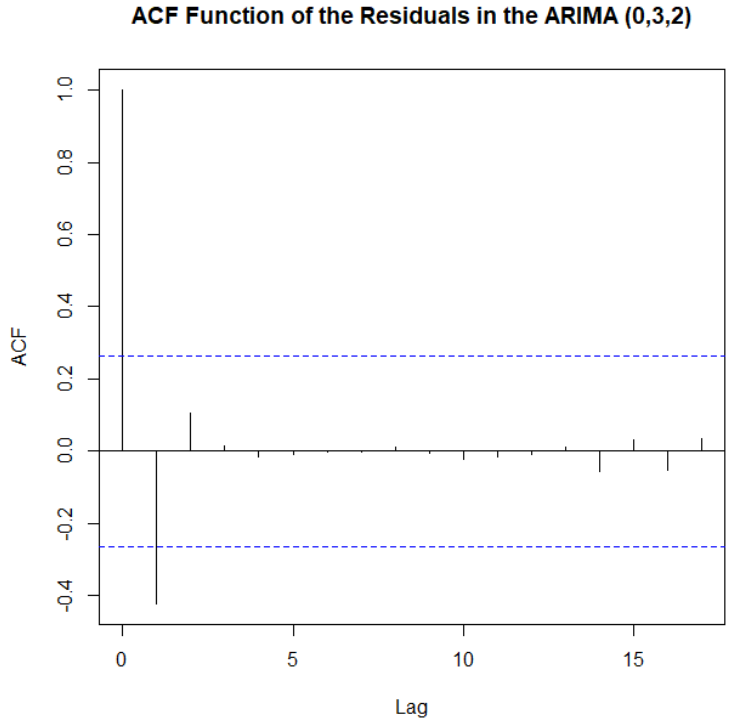


Διάγραμμα 6.14

Στο Διάγραμμα 6.13 βλέπουμε ότι για τις αρχικές παρατηρήσεις υπάρχουν μεγάλες αποκλίσεις βάσει του μοντέλου. Αυτό μπορεί να οφείλεται στο γεγονός ότι ο ρυθμός αύξησης των τελευταίων χρόνων επιτεύχθηκε μετά το 1960. Επομένως αν εξαιρέσουμε τις αρχικές τιμές, τα δεδομένα μας φαίνονται κανονικά κατανεμημένα με μέση τιμή 0. Επίσης στο Διάγραμμα 6.14 έχουμε το κλασσικό qq-plot από το οποίο και πάλι μπορούμε να δεχθούμε την κανονικότητα των υπολοίπων με εξαίρεση τις δυο ουρές (που οφείλονται στις ίδιες παρατηρήσεις που σχολιάστηκαν και πριν).

Κάνοντας τους ελέγχους κανονικότητας Shapiro-Wilk και Kolmogorov-Smirnov παίρνουμε πολύ μικρές p-τιμές και επομένως όντως δεχόμαστε ότι τα υπόλοιπα έχουν μέση τιμή μηδέν και είναι κανονικά κατανεμημένα.

Η αμέσως επόμενη υπόθεση που πρέπει να ελεγχθεί είναι η ανεξαρτησία των υπολοίπων. Μπορούμε αρχικά με την χρήση του Διαγράμματος 6.15 να δούμε ότι οι τιμές της ACF συνάρτησης είναι στατιστικά μη-διάφορες του μηδενός με εξαίρεση την τιμή για q=1.



Διάγραμμα 6.15

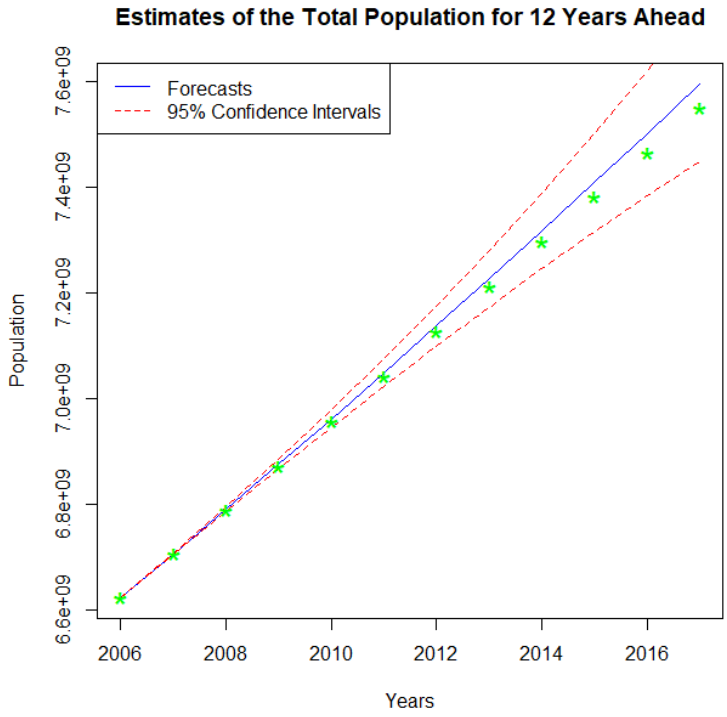
Όπως έχει ήδη σχολιαστεί, για να εξακριβώσουμε με καλύτερη ακρίβεια αν τα υπόλοιπα είναι ανεξάρτητα, θα πρέπει να γίνουν οι έλεγχοι Box-Pierce και Ljung-Box. Και οι δυο αυτοί έλεγχοι δίνουν πολύ μεγάλες p-τιμές (μεγαλύτερες του 90%) και επομένως μπορούμε να δεχθούμε την μηδενική υπόθεση ανεξαρτησίας υπολοίπων.

Συνεχίζουμε με τους ελέγχους της ενότητας 5.3.2 για τα υπόλοιπα. Οι συναρτήσεις που χρησιμοποιήθηκαν στην R ανήκουν στην βιβλιοθήκη spgs. Συγκεκριμένα, τα ονόματα των συναρτήσεων και τα αποτελέσματα τις R φαίνονται πιο κάτω:

```
> turningpoint.test(normresid)

        Turning point test of independence

data:  normresid
T = 0.54201, p-value = 0.5878

> diffsign.test(normresid)

        Difference-sign test of independence

data:  normresid
D = -0.92582, p-value = 0.3545

> rank.test(normresid)

        Rank test of independence

data:  normresid
R = 0.021779, p-value = 0.9826
```

Κανένας από τους πιο πάνω ελέγχους δεν φαίνεται να αποκλείει την υπόθεση ανεξαρτησίας και της κανονικότητας των υπολοίπων και επομένως το μοντέλο μας φαίνεται ικανό να περιγράψει τα δεδομένα μας και ως επέκταση οι μελλοντικές προβλέψεις βάσει αυτού έχουν καλές προοπτικές.

Στο Διάγραμμα 6.16 παρουσιάζονται οι προβλέψεις και τα διαστήματα εμπιστοσύνης καθώς επίσης και οι πραγματικές τιμές του πληθυσμού για τα έτη 2006-2017 που συμβολίζονται με αστερίσκους.



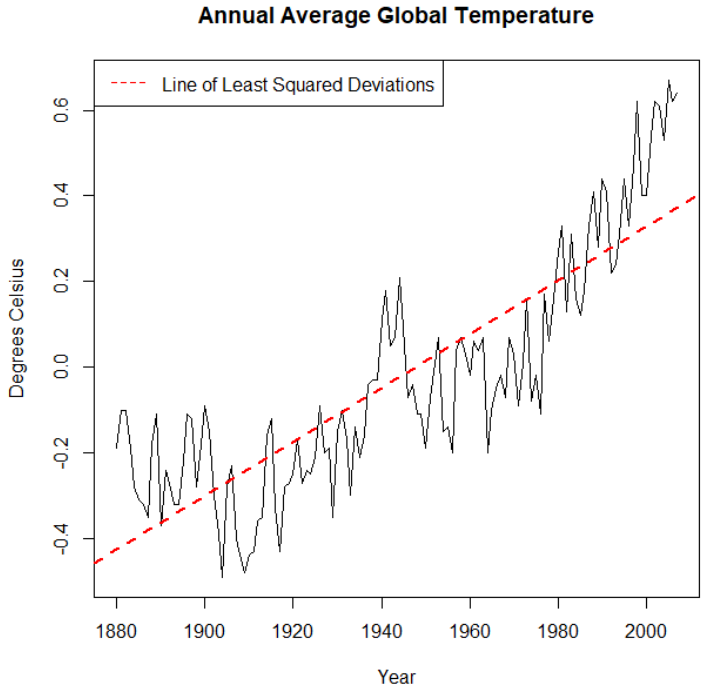
Διάγραμμα 6.16

Επομένως βλέπουμε ότι οι προβλεπόμενες τιμές είναι αρκετά κοντά στις πραγματικές τιμές. Το μοντέλο είναι αρκετά αξιόπιστο και επομένως θα μπορούσε να χρησιμοποιηθεί για μελλοντικές προβλέψεις με αρκετά καλή ακρίβεια. Αξιοσημείωτο είναι επίσης το γεγονός ότι τα διαστήματα εμπιστοσύνης είναι αρκετά μικρά σε σχέση με την προηγούμενη μελέτη. Ο λόγος για τον οποίο ισχύει αυτό, είναι η μικρή αβεβαιότητα που παρουσιάζουν τα δεδομένα μας. Δεν υπάρχουν μεγάλες αποκλίσεις από τις προσαρμοσμένες τιμές και επομένως μπορούμε να περιορίσουμε την αβεβαιότητα τον προβλέψεων μας αρκετά. Σημειώνουμε επίσης ότι οι προβλέψεις έγιναν για 12 περιόδους στο μέλλον, έχοντας ως δεδομένα μόλις 55 παρατηρήσεις.

6.3 Μελέτη Μέσης Θερμοκρασίας της Γης

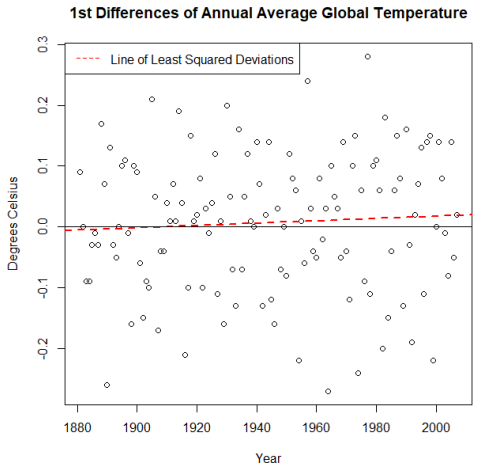
Σε αυτό το πρόβλημα, θα μας απασχολήσει η ετήσια μέση θερμοκρασία της Γης, μετρημένη σε βαθμούς Κελσίου. Τα δεδομένα που θα χρησιμοποιήσουμε είναι αντλημένα από την ιστοσελίδα της NASA (www.climate.nasa.gov) και αφορούν παρατηρήσεις για τα έτη 1880-2017. Σύμφωνα με όσα έχουν σχολιαστεί και σε προηγούμενα προβλήματα, θα θεωρήσουμε τις παρατηρήσεις των τελευταίων δέκα ετών ως άγνωστες, για να μπορέσουμε να συγκρίνουμε τις προβλέψεις μας με τις πραγματικές παρατηρήσεις.

Το Διάγραμμα 6.17 παρουσιάζει την μέση θερμοκρασία της Γης ανά έτος. Είναι εύκολο να διαπιστώσουμε ότι υπάρχει μια αυξητική τάση στα δεδομένα. Πράγματι και ο έλεγχος ADF επιβεβαιώνει ότι τα δεδομένα μας δεν παρουσιάζουν στασιμότητα (p-value = 0.67).

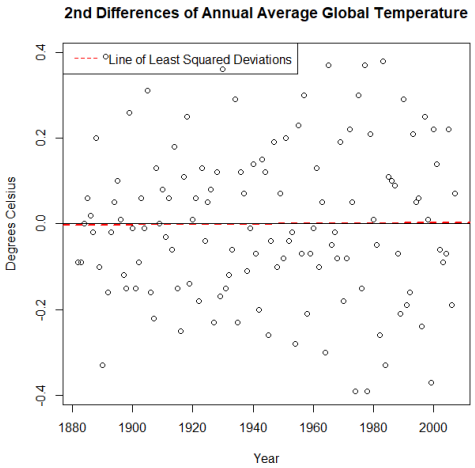


Διάγραμμα 6.17

Επομένως, για να προκύψει μια στάσιμη χρονοσειρά χρειάζεται να δημιουργήσουμε την χρονοσειρά πρώτων διαφορών, που φαίνεται στο Διάγραμμα 6.18. Ο αντίστοιχος ADF έλεγχος μας ενδεικνύει στασιμότητα, αν και φαίνεται να υπάρχει μια πολύ μικρή θετική τάση στα δεδομένα. Στο Διάγραμμα 6.19 παρουσιάζεται η ίδια γραφική παράσταση παίρνοντας όμως τις διαφορές δεύτερης τάξης των αρχικών μας παρατηρήσεων. Επομένως στην περίπτωση αυτή θα μελετηθούν αρχικά μοντέλα με παράμετρο Integration 1, αλλά σε περίπτωση που αντιμετωπίσουμε δυσκολίες, καλό θα ήταν να ελεγχθούν και μοντέλα με παράμετρο Integration 2 ώστε να επιλεχθεί το καλύτερο (d=1 ή d=2).

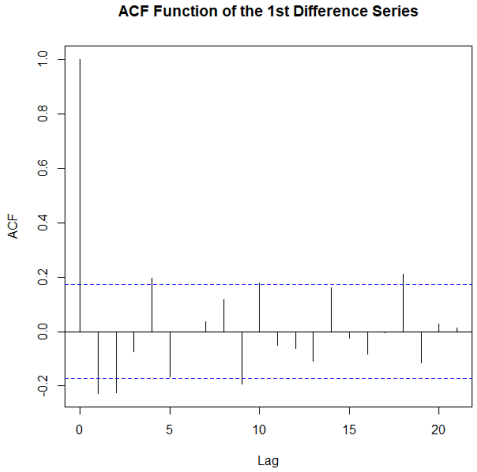


Διάγραμμα 6.18

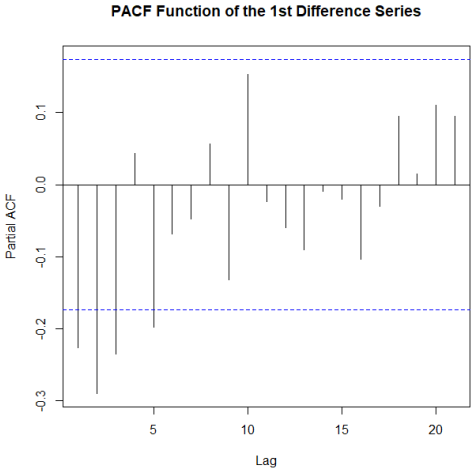


Διάγραμμα 6.19

Αρχικά θα ασχοληθούμε για την περίπτωση των διαφορών πρώτης τάξης (d=1). Με την χρήση των γραφημάτων ACF και PACF για αυτή την χρονοσειρά, που φαίνονται στα Διαγράμματα 6.20 και 6.21, καταλήγουμε ότι οι υποψήφιες τιμές των παραμέτρων (p,q) είναι (3,3) με όμως οριακή σημαντικότητα.



Διάγραμμα 6.18



Διάγραμμα 6.19

Συνεχίζουμε την μελέτη μας προσαρμόζοντας το μοντέλο ARIMA (3,1,3) για την αρχική χρονοσειρά και ελέγχοντας την σημαντικότητα των συντελεστών και τα κριτήρια επιλογής AIC και BIC για το μοντέλο αυτό, αλλά και τα εμφωλευμένα του, επιλέγουμε το καταλληλότερο μοντέλο.

ARIMA (3,1,3):

```
> arima(temp, order=c(3,1,3))

Call:
arima(x = temp, order = c(3, 1, 3))

Coefficients:
          ar1      ar2      ar3     ma1      ma2      ma3
      -0.9610  -0.2193  -0.1933  0.6473  -0.3721  -0.1805
s.e.   0.2863   0.2557   0.2073  0.2881   0.2039   0.2399

sigma^2 estimated as 0.01017:  log likelihood = 110.82,  aic = -207.65
```

Είναι προφανές ότι τα τυπικά σφάλματα των συντελεστών είναι αρκετά μεγάλα και επομένως οι μόνοι συντελεστές που είναι σημαντικά διάφοροι του μηδενός είναι ο AR1 και MA1. Συνεπώς απλούστερα μοντέλα θα ήταν καλό να ελεγχθούν έναντι του αρχικού.

ARIMA (2,1,2):

```
> arima(temp, order=c(2,1,2))

Call:
arima(x = temp, order = c(2, 1, 2))

Coefficients:
         ar1      ar2      ma1     ma2
      0.2915  -0.2084  -0.6493  0.0597
s.e.  0.4145   0.1881   0.4181  0.3017

sigma^2 estimated as 0.01072:  log likelihood = 107.63,  aic = -205.25
```

Το μοντέλο αυτό δεν έχει κανένα συντελεστή σημαντικά διάφορο του μηδενός αλλά ενώ το κριτήριο AIC μειώθηκε κατά 2 μονάδες, το κριτήριο BIC αυξήθηκε κατά 4 μονάδες. Επομένως θα λέγαμε ότι το μοντέλο αυτό είναι χειρότερο από το αρχικό.

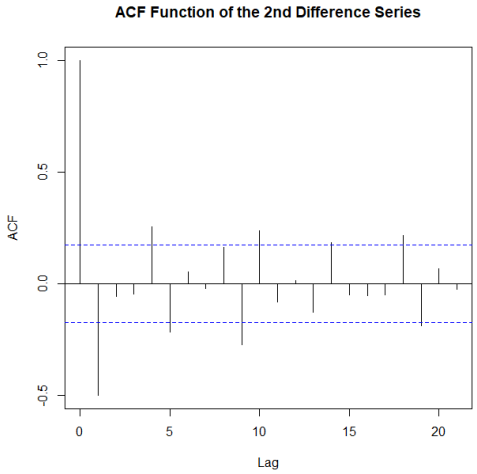
ARIMA (1,1,1):

```
> arima(temp, order=c(1,1,1))

Call:
arima(x = temp, order = c(1, 1, 1))

Coefficients:
         ar1      ma1
      0.3353  -0.7279
s.e.  0.1376   0.0940

sigma^2 estimated as 0.01093:  log likelihood = 106.44,  aic = -206.88
```

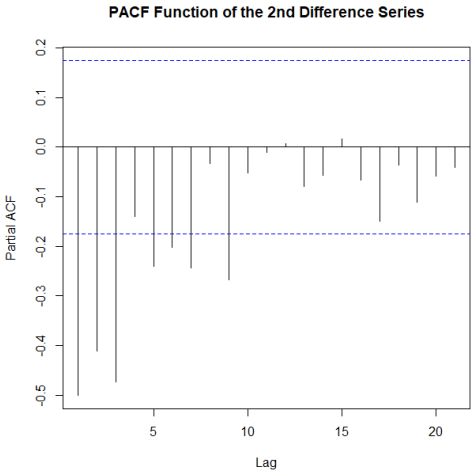
Εδώ έχουμε σημαντικότητα και στους δύο συντελεστές, όμως το κριτήριο AIC είναι σε ίδια επίπεδα με το αρχικό μοντέλο αλλά το κριτήριο BIC αυξήθηκε κατά 8.5 μονάδες. Επομένως ούτε αυτό το μοντέλο φαίνεται κατάλληλο.

Σημειώνουμε στο σημείο αυτό ότι ελέγχθηκαν και όλα τα υπόλοιπα εμφωλευμένα μοντέλα στην περίπτωση που το d=1 και κανένα από αυτά τα μοντέλα δεν φαίνεται να προσαρμόζεται ικανοποιητικά στα δεδομένα μας.

Συνεπώς θα ελέγξουμε μοντέλα για τα οποία το d=2. Όπως και πριν δημιουργούμε τα Διαγράμματα 6.20 και 6.21 που παρουσιάζουν τα γραφήματα ACF και PACF για την χρονοσειρά των διαφορών 2ης τάξης.



Διάγραμμα 6.20



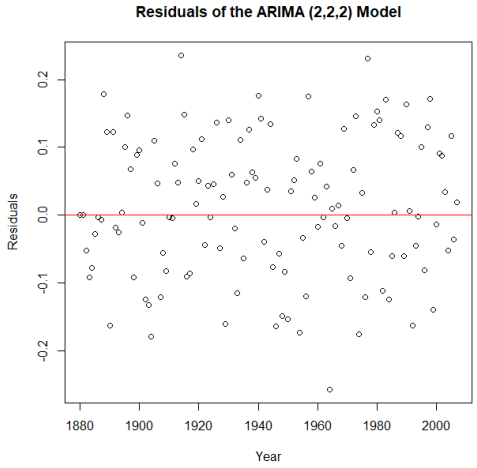
Διάγραμμα 6.21

Βάσει της εικόνας που παρουσιάζουν τα γραφήματα ACF και PACF, μπορούμε να αποφανθούμε ότι ένα πρώτο πιθανό μοντέλο στην περίπτωση που ο βαθμός Integration είναι 2, είναι το ARIMA (3,2,2). Ξεκινούμε όμως την μελέτη μας από το γενικότερο μοντέλο ARIMA (4,2,3) για να συμπεριλάβουμε περισσότερα μοντέλα.

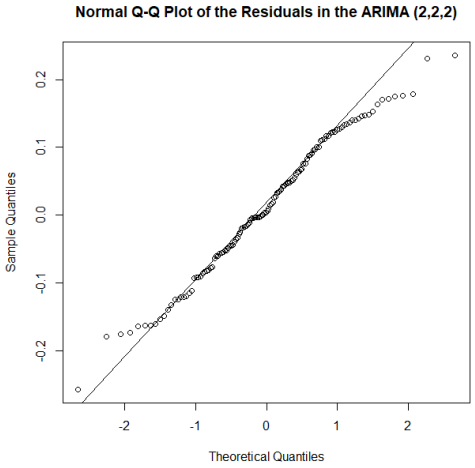
Έχοντας μελετήσει όλα τα εμφωλευμένα μοντέλα, καταλήγουμε στο ότι τα καλύτερα βάσει κριτηρίων AIC και BIC αλλά και της σημαντικότητας των συντελεστών, είναι τα ARIMA (2,2,0) και (1,2,0). Παρόλα αυτά και τα δυο μοντέλα αποτυγχάνουν στους διαγνωστικούς ελέγχους αφού παρουσιάζουν εξαρτημένα υπόλοιπα. Στο ίδιο συμπέρασμα καταλήγουμε και με τα μοντέλα ARIMA (2,2,1) και (2,2,1). Επομένως το απλούστερο μοντέλο που περιγράφει ικανοποιητικά τα δεδομένα μας είναι το ARIMA (2,2,2) και κάποια στοιχεία του φαίνονται πιο κάτω:

```
> fit <-arima(temp, order=c(2,2,2))
> arima(temp, order=c(2,2,2))

Call:
arima(x = temp, order = c(2, 2, 2))

Coefficients:
         ar1      ar2      ma1     ma2
      0.3389  -0.1364  -1.7083  0.7188
s.e.  0.1779   0.1198   0.1674  0.1660

sigma^2 estimated as 0.01064:  log likelihood = 104.86,  aic = -199.73
> BIC(fit)
[1] -185.5477
```

Δηλαδή ο μόνος συντελεστής που δεν είναι σημαντικά διάφορος του μηδενός είναι ο AR2. Επίσης σε σχέση με το μοντέλο ARIMA (3,1,3) έχουμε σημαντική μείωση του κριτηρίου AIC (8 περίπου μονάδες), αλλά έχουμε και μια μικρή μείωση του κριτηρίου BIC κατά 2.5 μονάδες. Επομένως φαίνεται το μοντέλο ARIMA (2,2,2) να είναι το καλύτερο από τα μοντέλα που προσαρμόστηκαν επομένως θα προχωρήσουμε με αυτό στους διαγνωστικούς ελέγχους υπολοίπων.



Διάγραμμα 6.22



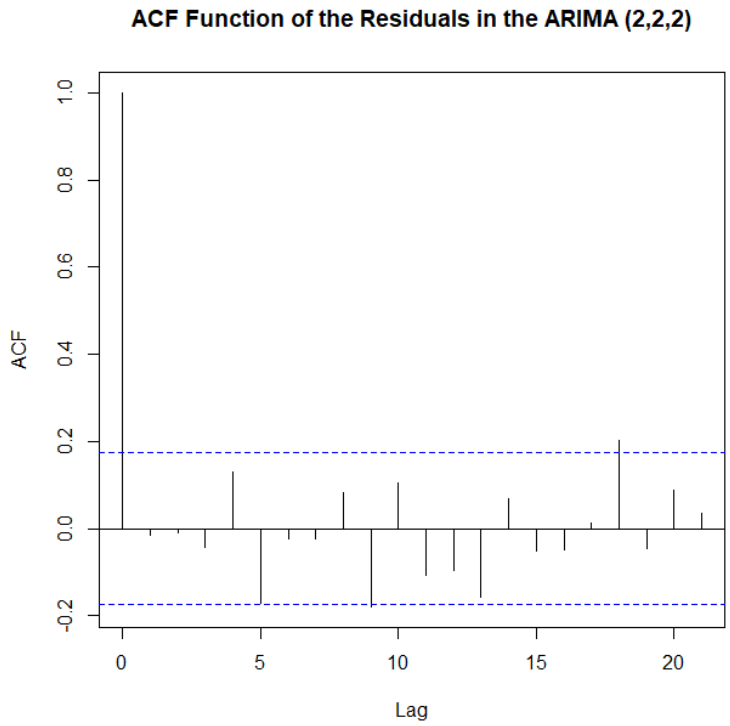
Διάγραμμα 6.23

Σύμφωνα με τα Διαγράμματα 6.22 και 6.23, δεν μπορώ να απορρίψουμε την υπόθεση κανονικότητας των υπολοίπων. Ο έλεγχος κανονικότητας Kolmogorov-Smirnov ενδεικνύει κανονικότητα στα υπόλοιπα όμως ο αντίστοιχος έλεγχος Shapiro-Wilk εκφέρει κάποιες αμφιβολίες με σχετικά ψηλότερη p-τιμή.

Για την υπόθεση ανεξαρτησίας των υπολοίπων, μπορούμε αρχικά να μελετήσουμε το γράφημα ACF των υπολοίπων, όπως φαίνεται στο Διάγραμμα 6.24. Φαίνεται πως για τις περισσότερες τιμές της παραμέτρου lag, οι τιμές της συνάρτησης βρίσκονται εντός των ορίων ανεξαρτησίας. Υπάρχουν όμως κάποιες περιπτώσεις για τις οποίες οι τιμές είναι αρκετά κοντά στα όρια και μάλιστα για κάποιες τιμές, ξεπερνούν τα όρια. Επομένως οι έλεγχοι Box-Pierce και Ljung-Box είναι απαραίτητοι.

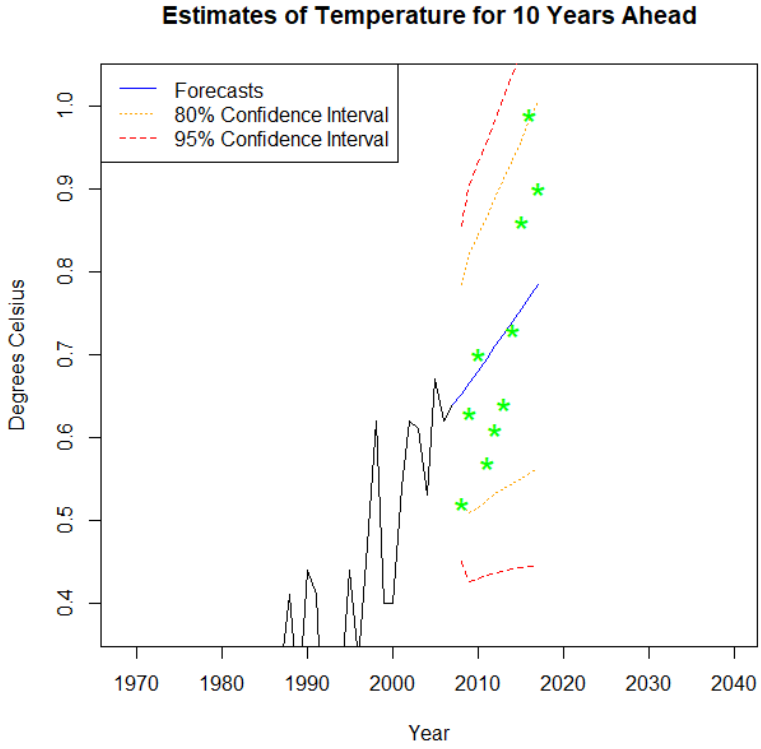
Τα p-value και των δύο ελέγχων είναι μεγαλύτερα του 0.10 και επομένως δεν μπορώ να απορρίψω την μηδενική υπόθεση ανεξαρτησίας των υπολοίπων σε 90% επίπεδο σημαντ.

```
> Box.test(fit_resid,lag=15,type="Box-Pierce")

        Box-Pierce test

data:  fit_resid
X-squared = 19.479, df = 15, p-value = 0.1928

> Box.test(fit_resid,lag=15,type="Ljung-Box")

        Box-Ljung test

data:  fit_resid
X-squared = 21.265, df = 15, p-value = 0.1286
```



Διάγραμμα 6.24

Στα ίδια αποτελέσματα καταλήγουμε και με τους εναλλακτικούς ελέγχους της ενότητας 5.3.2 και επομένως δεχόμαστε ότι τα υπόλοιπα του μοντέλου ARIMA (2,2,2) είναι κανονικά και ανεξάρτητα κατανεμημένα, με μηδενική μέση τιμή. Συνεπώς μπορούμε βάσει του μοντέλου αυτού να προβούμε σε προβλέψεις για τα έτη 2008 έως 2017.



Διάγραμμα 6.25

Βλέποντας το Διάγραμμα 6.25, μπορούμε να διακρίνουμε μια σχετικά μεγάλη απόκλιση τον πραγματικών παρατηρήσεων από τις σημειακές προβλέψεις. Παρόλα αυτά, προβλέφθηκε σωστά η αυξητική τάση των τιμών καθώς επίσης και όλες οι τιμές βρίσκονται εντός των 95% διαστημάτων εμπιστοσύνης και σχεδόν όλες (με εξαίρεση της θερμοκρασίας που παρατηρήθηκε το 2016) να βρίσκονται και εντός των 80% διαστημάτων εμπιστοσύνης. Το μοντέλο αυτό δίνει μια καλή προσέγγιση για τα δεδομένα μας αλλά θα συνιστούσαμε να κρατάμε κάποια επιφύλαξη για τις προβλέψεις βάσει αυτού καθώς παρουσιάζονται μεγάλες διακυμάνσεις των τιμών.

6.4 Καμπύλη Phillips

Ο οικονομολόγος Alban William Housego Phillips, ήταν ο πρώτος που ισχυρίστηκε ότι υπάρχει μια αντιστρόφως ανάλογη σχέση μεταξύ του ποσοστού ανεργίας και των αντίστοιχων επιπέδων των μισθών. Την σχέση αυτή επιβεβαίωσε μελετώντας τα αντίστοιχα δεδομένα του Ηνωμένου Βασιλείου για τα έτη 1861 έως 1957. Η σχέση όμως αυτή δεν περιγράφεται από μια αυστηρή μαθηματική εξίσωση, και πηγάζει κυρίως από εμπειρικές παρατηρήσεις.

Αργότερα, το 1967, ο οικονομολόγος Milton Friedman επέκτεινε λογικά την σχέση που ισχυρίστηκε ο Phillips, και σύνδεσε το ποσοστό ανεργίας με το επίπεδο πληθωρισμού. Αυτή η σχέση είναι γνωστή ως η Καμπύλη Phillips.

Πολλοί μετέπειτα οικονομολόγοι ισχυρίστηκαν ότι η σχέση της καμπύλης Phillips ισχύει μόνο βραχυπρόθεσμα και όχι μακροχρόνια, ενώ άλλοι υποστήριζαν ότι η σχέση αυτή ίσχυε ίσως στις οικονομίες του 19^{ου} και στις αρχές του 20^{ου} αιώνα, αλλά πλέον στις σύγχρονες οικονομίες δεν υπάρχει σημαντική συσχέτιση μεταξύ των δυο οικονομικών μεγεθών.

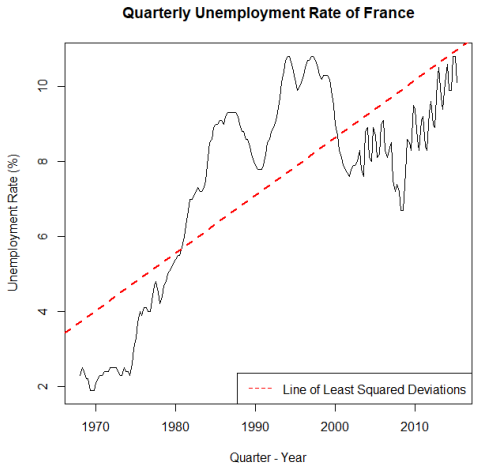
Θέλοντας να ερευνήσουμε κατά πόσο μια τέτοια σχέση συνδέει τα 2 αυτά μεγέθη, μπορούμε να προσαρμόσουμε μοντέλα Autoregressive Distributed Lag (ADL) και μελετώντας τα, να αποφανθούμε για την σημαντικότητα της συσχέτισης σε μια σύγχρονη οικονομία και ακολούθως, σε περίπτωση που μια τέτοια σχέση υπάρχει, να καταλήξουμε σε κάποια συμπεράσματα που πιθανόν να ήταν χρήσιμο να ληφθούν υπόψη στον καθορισμό οικονομικής πολιτικής της χώρας.

Τα δεδομένα που θα χρησιμοποιήσουμε σε αυτό το πρόβλημα, είναι το ποσοστό ανεργίας (άνεργος πληθυσμός έναντι του εργατικού δυναμικού της χώρας) και ο πληθωρισμός (διαφοροποίηση στον δείκτη τιμών καταναλωτή CPI) στην Γαλλία για τα έτη 1968 έως 2017, ανά τρίμηνο.

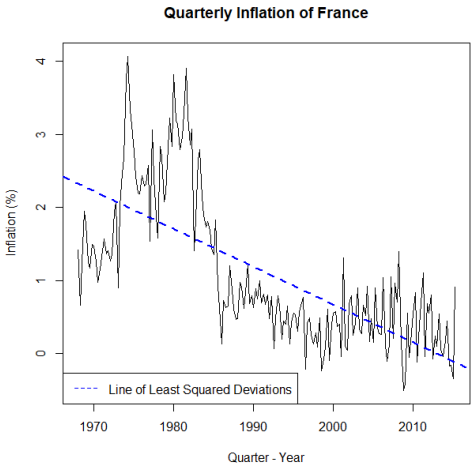
Ένας πρώτος, απλός έλεγχος, που μπορούμε να κάνουμε είναι ο έλεγχος Pearson's Product-Moment Correlation. Ο έλεγχος αυτός φυσικά ελέγχει αν υπάρχει στατιστικά σημαντική συσχέτιση (διάφορη του μηδενός) μεταξύ των δεδομένων μας. Η σημειακή εκτίμηση το συντελεστή συσχέτισης είναι -0.647151 με p-value μικρότερη του εκατομμυριοστού.

Σε περίπτωση όμως που υπήρχε συσχέτιση μεταξύ ενός μεγέθους και κάποιας υστέρησης του άλλου, τότε ο απλός αυτός έλεγχος, δεν θα εντόπιζε την συσχέτιση αυτή. Συνεπώς περαιτέρω μελέτη είναι αναγκαία για τον εντοπισμό τέτοιων συσχετίσεων.

Κοιτάζοντας τις γραφικές παραστάσεις που παριστάνουν τα δεδομένα μας, στα Διαγράμματα 6.26 και 6.27, βλέπουμε ότι τα δεδομένα μας δεν παρουσιάζονται να είναι στάσιμα. Το ίδιο αποτέλεσμα συμπεραίνουμε και με την χρήση του επαυξημένου ελέγχου Dickey Fuller που μας δίνει p-τιμές 0.616 για την ανεργία και 0.409 για τον πληθωρισμό. Επομένως και στις δύο περιπτώσεις, δεχόμαστε την μηδενική υπόθεση της μη-στασιμότητας.

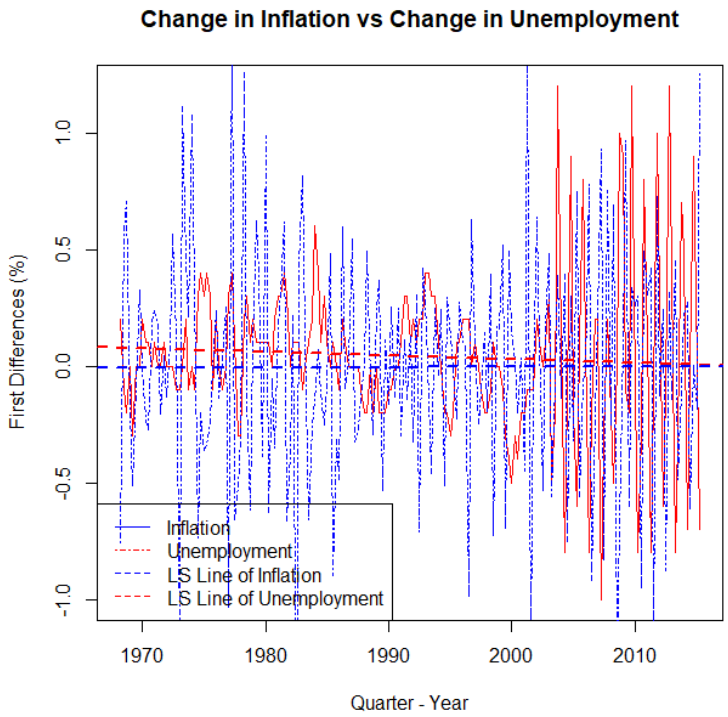


Διάγραμμα 6.26



Διάγραμμα 6.27

Παίρνοντας διαφορές πρώτης τάξης και για τα δύο μεγέθη που μελετάμε, και κάνοντας τους ελέγχους ADF, καταλήγουμε στο ότι τα δεδομένα μας φαίνονται ικανοποιητικά στάσιμα αφού και για τις δυο χρονοσειρές ο έλεγχος ADF με εναλλακτική υπόθεση την στασιμότητα των παρατηρήσεων, επιστρέφει p-τιμή μικρότερη της τάξης του 1%. Επίσης, συνοψίζοντας τις δυο γραφικές παραστάσεις στο ίδιο γράφημα (όπως φαίνεται στο Διάγραμμα 6.28 της επόμενης σελίδας), μπορούμε και γραφικά να επιβεβαιώσουμε την στασιμότητα αυτή.



Διάγραμμα 6.28

Πλέον τα μεγέθη που μελετάμε, πέρα από στάσιμα, έχουν και την πραγματική ερμηνεία που μελετάμε. Η αρνητική συσχέτιση που παρατήρησε ο Phillips αφορούσε αυξομειώσεις πληθωρισμού και ανεργίας. Όπως και πριν, με χρήση του ελέγχου συντελεστή συσχέτισης Pearson για τα δύο μεγέθη, παίρνουμε μια αρνητική μεν, αλλά πολύ μικρή τιμή δε, της τάξεως του -0.04014, με διάστημα εμπιστοσύνης που περιλαμβάνει το 0. Επομένως, εκ πρώτης όψεως, δεν μπορούμε να αποφανθούμε ότι η συσχέτιση αυτή είναι στατιστικά σημαντική.

Ξεκινούμε την εκτενή μελέτη μας προσαρμόζοντας το απλούστερο ADL μοντέλο, τάξης (1,0) που συνδέει την διαφορά στον πληθωρισμό με την διαφορά στον πληθωρισμό στην αμέσως προηγούμενη χρονική περίοδο και με την αυξομείωση του επιπέδου ανεργίας στην τρέχον περίοδο. Βάσει των εκτιμητριών ελαχιστοποίησης των τετραγώνων των σφαλμάτων, και συμβολίζοντας με Z τις διαφορές των τιμών του πληθωρισμού και με X τις διαφορές των τιμών στο επίπεδο ανεργίας, το μοντέλο αυτό παίρνει την μορφή της εξίσωσης 6.1:

$$Z_t = -0.00067 - 0.39395\, Z_{t-1} - 0.196\, X_t + \varepsilon_t \quad (6.1).$$

Βάσει των τυπικών σφαλμάτων των συντελεστών του μοντέλου αυτού, μπορούμε να διαπιστώσουμε ότι ο συντελεστής της υστέρησης του πληθωρισμού είναι διάφορος του μηδενός σε επίπεδο σημαντικότητας μικρότερο του 0.1%, ενώ ο συντελεστής που μας ενδιαφέρει, δηλαδή που συσχετίζει τις διαφορές πληθωρισμού με διαφορές ανεργίας, είναι αρνητικός (όπως θα προέβλεπε η καμπύλη Phillips) και είναι στατιστικά σημαντικός σε επίπεδο σημαντικότητας 10%.

Παρόλα αυτά, κάνοντας χρήση του ελέγχου Breusch-Godfrey για συσχέτιση υπολοίπων που διαφέρουν p-περιόδους (που βρίσκεται στην βιβλιοθήκη lmtest της R), βλέπουμε ότι όντως τα διαδοχικά υπόλοιπα του πιο πάνω μοντέλου είναι εξαρτημένα. Επομένως το μοντέλο αυτό δεν περιγράφει ικανοποιητικά τα δεδομένα που έχουμε. Τα αποτελέσματα αυτά παρουσιάζονται στην R με τον πιο κάτω τρόπο:

```
> mod1<-lm(L0.Inflation~L1.Inflation+L0.Unemp)
> summary(mod1)

Call:
lm(formula = L0.Inflation ~ L1.Inflation + L0.Unemp)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56068 -0.30066  0.00157  0.27855  1.19602

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0006707  0.0357409  -0.019    0.9850
L1.Inflation -0.3939502  0.0718228  -5.485 1.36e-07 ***
L0.Unemp     -0.1959927  0.1042380  -1.880    0.0617 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4839 on 183 degrees of freedom
Multiple R-squared:  0.142,     Adjusted R-squared:  0.1326
F-statistic: 15.14 on 2 and 183 DF,  p-value: 8.246e-07

> res.mod1<-ts(mod1$residuals,frequency=4)
> bgtest(mod1,fill=NA,order=1)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  mod1
LM test = 6.9841, df = 1, p-value = 0.008224
```

Ακολούθως, προσαρμόζουμε το αμέσως επόμενο μοντέλο, ADL (1,1) και όπως και πριν βλέπουμε τον συντελεστή που συνδέει τις διαφορές των τιμών του πληθωρισμού με τις διαφορές στο επίπεδο ανεργίας της ίδιας χρονικής περιόδου. Και πάλι έχουμε αρνητικό συντελεστή, και μάλιστα είναι στατιστικά διάφορος του μηδενός σε επίπεδο σημαντικότητας μικρότερο του 2%. Όπως και πριν όμως, παραβιάζεται η υπόθεση ανεξαρτησίας υπολοίπων και συνεπώς μελετάμε το αμέσως επόμενο μοντέλο, το ADL (2,0).

Συνεχίζουμε την διαδικασία αυτή και απορρίπτοντας όλα τα μοντέλα με εξαρτημένα υπόλοιπα, καταλήγουμε στο απλούστερο μοντέλο που περιγράφει ικανοποιητικά τα δεδομένα και έχει ανεξάρτητα υπόλοιπα. Το μοντέλο αυτό είναι το ADL τάξης (3,2) και συγκεκριμένα έχει την μορφή της εξίσωσης 6.2:

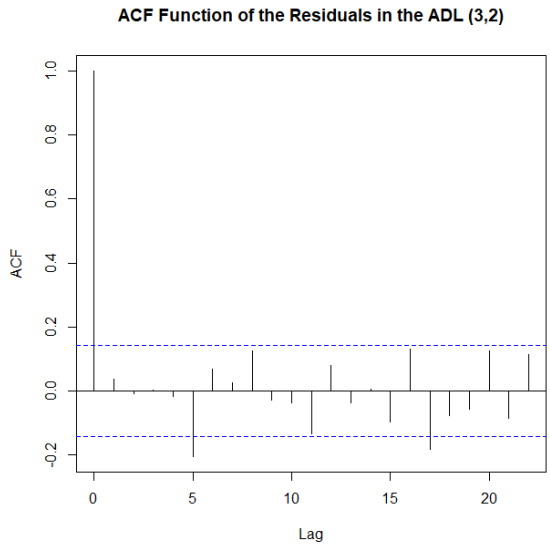
$$Z_t = -0.005 - 0.547\, Z_{t-1} - 0.359\, Z_{t-2} - 0.297\, Z_{t-3} - 0.258\, X_t + 0.120\, X_{t-1} - 0.076\, X_{t-2} + \varepsilon_t \quad (6.2).$$

Οι στατιστικά σημαντικοί συντελεστές στο πιο πάνω μοντέλο είναι οι συντελεστές των υστερήσεων του πληθωρισμού (σε επίπεδο σημαντικότητας μικρότερο του 0,1%) και ο προς μελέτη συντελεστής του X_t (σε επίπεδο σημαντικότητας μικρότερο του 3%). Συγκεκριμένα, τα 95% διαστήματα εμπιστοσύνης των συντελεστών του πιο πάνω μοντέλου, παρουσιάζονται στο Πίνακα 6.2:

	2.5% Percentile	97.5% Percentile
(Intercept)	-0.07242154	0.06143830
L1.Inflation	-0.69175747	-0.40141613
L2.Inflation	-0.51619500	-0.20153314
L3.Inflation	-0.44252811	-0.15235096
L0.Unemp	-0.48859166	-0.02782039
L1.Unemp	-0.08942979	0.33036818
L2.Unemp	-0.30643949	0.15347835

Πίνακας 6.2

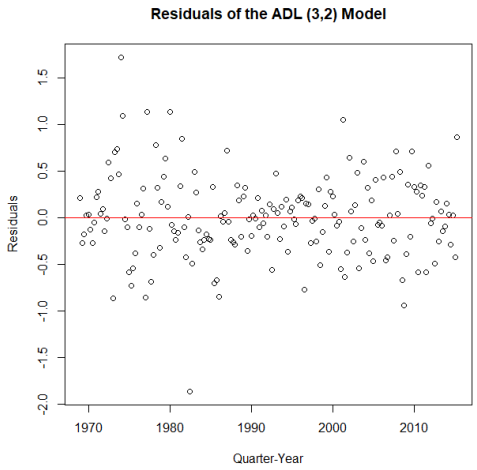
Μελετώντας τις συσχετίσεις διαδοχικών υπολοίπων μέχρι και 5ης τάξης, δεν μπορούμε να απορρίψουμε την υπόθεση ανεξαρτησίας υπολοίπων. Με την χρήση του γνωστού ελέγχου Box-Ljung για την συνολική εικόνα των υπολοίπων, δεν έχουμε και πάλι σοβαρές ενδείξεις κατά της ανεξαρτησίας τους. Γραφικά, μπορούμε να δούμε στο Διάγραμμα 6.29 με χρήση του γραφήματος της συνάρτησης ACF των υπολοίπων, ότι κάποια υπόλοιπα βρίσκονται ελαφρώς εκτός των ορίων σημαντικότητας, αλλά αυτό όπως έχει σχολιαστεί σε προηγούμενες ενότητες, είναι ανεπαρκές κριτήριο για να αποφανθούμε για την ανεξαρτησία των υπολοίπων.



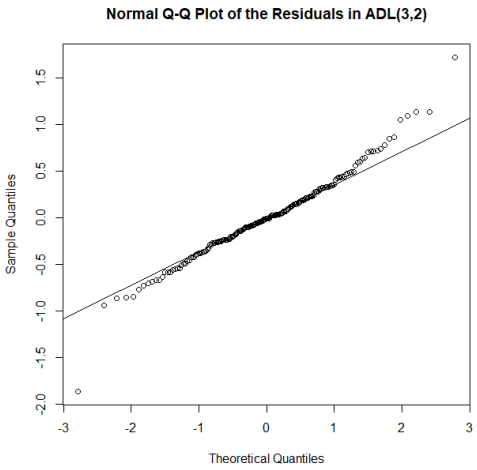
Διάγραμμα 6.29

Εξετάζουμε έπειτα την υπόθεση κανονικότητας των υπολοίπων στο ίδιο μοντέλο με τους ίδιου τρόπους που χρησιμοποιήσαμε στα προηγούμενα προβλήματα.

Συγκεκριμένα μπορούμε να δούμε γραφικά στο Διάγραμμα 6.30 ότι τα υπόλοιπα μας φαίνονται τυχαία κατανεμημένα γύρω από τον μηδενικό άξονα και δεν παρουσιάζουν μοτίβα. Αντίστοιχα στο qq-plot του Διαγράμματος 6.31, αν και τα δεδομένα παρουσιάζουν μικρές αποκλίσεις στις ουρές, φαίνονται να εμπίπτουν κατά πλειοψηφία στη γραμμή κανονικότητας.



Διάγραμμα 6.30



Διάγραμμα 6.31

Πέρα των γραφικών ελέγχων, τα υπόλοιπα φαίνονται να ακολουθούν κανονική κατανομή με μέση τιμή μηδέν και σύμφωνα με τους στατιστικούς ελέγχους κανονικότητας Shapiro-Wilk και Kolmogorov Smirnov, σε επίπεδο σημαντικότητας μικρότερο της τάξης του 0.2%. Στα ίδια αποτελέσματα οδηγούμαστε και με την χρήση των τριών ελέγχων τις ενότητας 5.3.2.

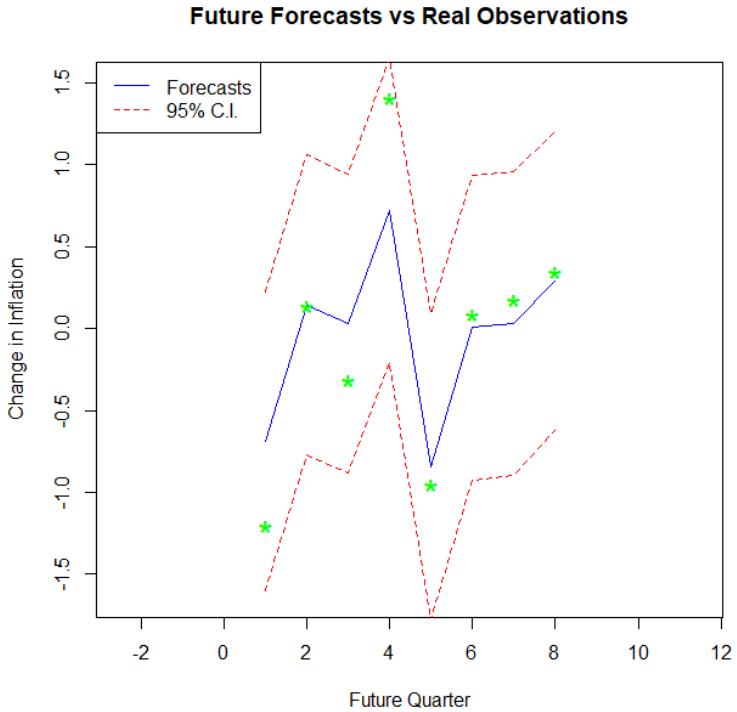
```
> turningpoint.test(normresid)

        Turning point test of independence

data:  normresid
T = 0.058252, p-value = 0.9535

> diffsign.test(normresid)

        Difference-sign test of independence

data:  normresid
D = 0.12666, p-value = 0.8992

> rank.test(normresid)

        Rank test of independence

data:  normresid
R = 0.015313, p-value = 0.9878
```

```
> shapiro.test(fit_resid)

        Shapiro-Wilk normality test

data:  fit_resid
W = 0.97365, p-value = 0.00138

> ks.test(fit_resid,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  fit_resid
D = 0.21483, p-value = 7e-08
alternative hypothesis: two-sided
```

Βάση των πιο πάνω, το μοντέλο ADL (3,2) φαίνεται να περιγράφει ικανοποιητικά τα δεδομένα μας και ταυτόχρονα πληροί τις προϋποθέσεις κανονικότητας και ανεξαρτησίας υπολοίπων. Επίσης, φαίνεται να επαληθεύουν την **αρνητική** συσχέτιση που είναι γνωστή ως η καμπύλη Phillips, καθώς μια **αύξηση** του ποσοστού ανεργίας κατά 1% θα οδηγήσει (βάσει της εξίσωσης 6.2) σε **μείωση** του ποσοστού πληθωρισμού κατά 0.258%.

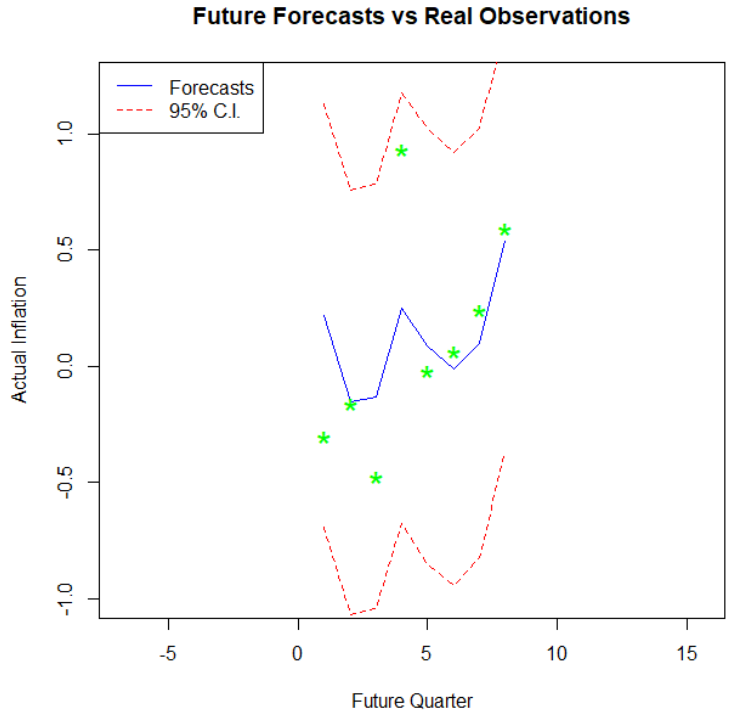
Στις πιο πάνω αναλύσεις, δεν θεωρήθηκαν γνωστές οι παρατηρήσεις τις ανεργίας και του πληθωρισμού για τα 8 τελευταία τρίμηνα, δηλαδή για τα έτη 2016 και 2017. Με χρήση δηλαδή των προηγούμενων παρατηρήσεων, και επεκτείνοντας το μοντέλο στο μέλλον, μπορούμε να κάνουμε προβλέψεις για τα 8 τελευταία τρίμηνα και ακολούθως να τα συγκρίνουμε με τις πραγματικές παρατηρήσεις.

Για να αποφύγουμε την διόγκωση των τυπικών σφαλμάτων των εκτιμήσεων, δεν θα εκτιμήσουμε και τις μελλοντικές τιμές τις ανεργίας, αλλά θα τις θεωρούμε δοσμένες για την τρέχουσα χρονική περίοδο και ταυτόχρονα, θα θεωρούμε ως δεδομένα τα πραγματικά ποσοστά πληθωρισμού για τα προηγούμενα εξάμηνα. Δηλαδή οι προβλέψεις μας θα αφορούν μια περίοδο στο μέλλον και χρησιμοποιούμε όλα τα διαθέσιμα πραγματικά δεδομένα που θα γίνονταν γνωστά σε εμάς την αμέσως προηγούμενη χρονική περίοδο. Τέτοιες προβλέψεις ονομάζονται **one-step ahead updated forecasts**. Στο Διάγραμμα 6.32, βλέπουμε με μπλε γραμμή τις σημειακές μας εκτιμήσεις, οι κόκκινες γραμμές σχηματίζουν τα όρια του 95% διαστήματος εμπιστοσύνης ενώ με πράσινους αστερίσκους συμβολίζουμε τις πραγματικές παρατηρήσεις της μεταβολής του επιπέδου πληθωρισμού.



Διάγραμμα 6.32

Συνήθως όμως, δεν μας ενδιαφέρει η διαφοροποίηση του πληθωρισμού, αλλά μας ενδιαφέρει το ποσοστό πληθωρισμού. Επομένως βάσει των πιο πάνω παρατηρήσεων, μπορούμε να σχηματίσουμε τις αντίστοιχες εκτιμήσεις για το ποσοστό πληθωρισμού σε κάθε χρονική περίοδο και να τις συγκρίνουμε με τις πραγματικές τιμές του πληθωρισμού που παρατηρήθηκαν. Στο Διάγραμμα 6.33 που φαίνεται πιο κάτω, κρατήσαμε ίδιους χρωματικούς συμβολισμούς με πριν, το μόνο που διαφέρει είναι το ότι οι προβλέψεις αφορούν το ποσοστό πληθωρισμού και όχι τις διαφορές στο ποσοστό πληθωρισμού.



Διάγραμμα 6.33

Και πάλι φαίνεται το μοντέλο να ανταποκρίνεται ικανοποιητικά στα πραγματικά δεδομένα και μάλιστα με πραγματικές παρατηρήσεις που βρίσκονται αρκετά κοντά στις σημειακές εκτιμήσεις (με εξαίρεση την πρόβλεψη για το 4ο μελλοντικό εξάμηνο).

Επομένως, σύμφωνα με τα πιο πάνω, το μοντέλο που υπολογίστηκε, προσαρμόζεται αξιόπιστα στα δεδομένα, πληροί τις αναγκαίες θεωρητικές προϋποθέσεις, και επιβεβαιώνει την θεωρητική αντίστροφη συσχέτιση πληθωρισμού και ανεργίας.

Παρόλα αυτά, μας δημιουργείτε η απορία, πως θα επηρεαστούν οι μελλοντικές τιμές του πληθωρισμού σε περίπτωση που η ανεργία ενός έτους, μειωθεί ή αυξηθεί μεμονωμένα (κάτι που ίσως μπορεί σκόπιμα να επιδιώξει μια κυβέρνηση). Αναμένουμε, σύμφωνα με τα όσα έχουμε πει στην ενότητα 3, ότι μια τέτοια μεμονωμένη διακύμανση θα έχει μια παροδική επιρροή στις τιμές του πληθωρισμού που σταδιακά θα εξαφανιστεί. Ποιο είναι όμως το μέγεθος αυτής της παροδικής επιρροής.

Μια επέκταση του πιο πάνω ερωτήματος, είναι ποια θα είναι η μόνιμη αλλαγή στο επίπεδο του πληθωρισμού σε περίπτωση που υπάρξει μια μόνιμη, διαχρονική αλλαγή στο επίπεδο ανεργίας μιας χώρας (όπως και πριν είναι κάτι που μπορεί να επιτευχθεί από την κυβέρνηση της χώρας).

Στην ενότητα 3, συζητήθηκαν οι δυναμικοί πολλαπλασιαστές αλλά και ο πολλαπλασιαστής βάθους χρόνου στο μοντέλο ADL (1,1). Στην δική μας περίπτωσή μπορούμε να υπολογίσουμε τους αντίστοιχους πολ/στές στο μοντέλο ADL (3,2).

Για την μελέτη αυτή θεωρούμε το γενικό μοντέλο ADL (3,2) που έχει την μορφή:

$$Z_t = \delta + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \alpha_3 Z_{t-3} + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t \, .$$

Επομένως οι δυναμική πολλαπλασιαστές προκύπτουν από τις σχέσεις:

$$\frac{\partial Z_t}{\partial X_t} = \beta_0 \, ,$$

$$\frac{\partial Z_{t+1}}{\partial X_t} = \alpha_1 \frac{\partial Z_t}{\partial X_t} + \beta_1 \, ,$$

$$\frac{\partial Z_{t+2}}{\partial X_t} = \alpha_1 \frac{\partial Z_{t+1}}{\partial X_t} + \alpha_2 \frac{\partial Z_t}{\partial X_t} + \beta_2 \, ,$$

$$\frac{\partial Z_{t+3}}{\partial X_t} = \alpha_1 \frac{\partial Z_{t+2}}{\partial X_t} + \alpha_2 \frac{\partial Z_{t+1}}{\partial X_t} + \alpha_3 \frac{\partial Z_t}{\partial X_t} \, . \, . \, .$$

$$\frac{\partial Z_{t+k}}{\partial X_t} = \alpha_1 \frac{\partial Z_{t+k-1}}{\partial X_t} + \alpha_2 \frac{\partial Z_{t+k-2}}{\partial X_t} + \alpha_3 \frac{\partial Z_{t+k-3}}{\partial X_t} \, .$$

Ενώ ο πολλαπλασιαστής βάθος χρόνου, αποδεικνύεται ότι στο γενικό μοντέλο ADL (p,q) προκύπτει από την πιο κάτω σχέση:

$$\beta = \frac{\partial E[Z_t]}{\partial E[X_t]} = \frac{\sum_{i=0}^{q} b_i}{1 - \sum_{j=1}^{p} a_j} \, .$$

Γνωρίζοντας όμως την ακριβή μορφή του μοντέλου που περιγράφει ικανοποιητικά τα δεδομένα μας, μπορούμε να απαντήσουμε στα πιο πάνω ερωτήματα, υπολογίζοντας ακριβοί νούμερα που παρουσιάζονται συγκεντρωτικά στον Πίνακα 6.3 που βρίσκεται στην επόμενη σελίδα. Στον πίνακα αυτό παρουσιάζουμε την επιρροή μιας διακύμανσης μιας τιμή της επεξηγηματικής μεταβλητής στις μελλοντικές τιμές της μεταβλητής απόκρισης.

Μια μεμονωμένη διακύμανση στην τιμή X_t (Ανεργία)	
Περίοδος	**Επιρροή στο Z_t (Πληθωρισμό)**
0	-0.25821
1	0.261601
2	-0.12681
3	0.052233
4	-0.06085
5	0.052235
6	-0.02225
7	0.011516
8	-0.01385
9	0.010054
10+	**<0.005%** (κατά απόλυτη τιμή)

Πίνακας 6.3

Για παράδειγμα, σε περίπτωση που το ποσοστό ανεργίας για ένα και μόνο έτος αυξηθεί κατά 2%, ενώ όλοι οι υπόλοιποι παράγοντες παραμείνουν σταθεροί, αναμένουμε ότι ο πληθωρισμός για το ίδιο έτος θα μειωθεί κατά 0.516% , ενώ στο επόμενο έτος θα αναμένουμε αύξηση του πληθωρισμού κατά 0.523%. Σε 2 χρόνια μπροστά περιμένουμε μια μείωση και πάλι της τάξεως του 0.254% και αυτή η “ταλαντωτική” συμπεριφορά συνεχίζει επ’άπειρον. Φυσικά μετά την δέκατη μελλοντική περίοδο, η επιρροή θα είναι αμελητέα, κάτι το οποίο περιμέναμε.

Φυσικά σε περίπτωση που υπάρξει μια μόνιμη αλλαγή στο επίπεδο ανεργίας, για παράδειγμα αν προσληφθεί επιπλέον 1% του εργατικού δυναμικού σε μόνιμες θέσεις εργασίας επ’αόριστον, τότε θα αναμέναμε ότι το γενικό ποσοστό πληθωρισμού θα αυξηθεί κατά τον πολλαπλασιαστεί βάθος χρόνου, δηλαδή κατά 0.097%.

6.5 Συσχέτιση Πωλήσεων και Διαφημιστικών Δαπανών.

Ένα πολύ συχνό πρόβλημα που αντιμετωπίζουν σε καθημερινό επίπεδο πολλές εταιρίες, είναι η εκτίμηση των μελλοντικών τους πωλήσεων. Υπάρχει μια μεγάλη ποικιλία σε μεθόδους που χρησιμοποιούνται για να δώσουν την απάντηση σε αυτό το ερώτημα (ανάμεσα τους και η μονοδιάστατη ανάλυση χρονοσειρών).

Μια όμως καλύτερη προσέγγιση μπορεί να επιτευχθεί αν χρησιμοποιήσουμε περισσότερες μεταβλητές (για δεδομένα στα οποία εύκολα έχουμε πρόσβαση), τα οποία συνδέονται άμεσα με τις πωλήσεις. Μια τέτοια επεξηγηματική μεταβλητή είναι τα έξοδα διαφημίσεων.

Πέρα της πιθανής καλύτερης ακρίβειας που μπορούμε να πετύχουμε με την μελέτη ενός μοντέλου που περιλαμβάνει τα έξοδα διαφημίσεων, μπορούμε να εξαγάγουμε χρήσιμες πληροφορίες για την σημαντικότητα των διαφημίσεων στις πωλήσεις. Θα μπορέσουμε επίσης να υπολογίσουμε μελλοντικές επιρροές στις πωλήσεις για διάφορα σενάρια σχετικά με τα έξοδα διαφήμισης (όπως έχει συζητηθεί και σε προηγούμενες ενότητες/προβλήματα). Η γνώση τέτοιων πληροφοριών θα είναι αρκετά χρήσιμη και στον καθορισμό της μελλοντικής διαφημιστικής πολιτικής της εταιρίας.

Για την μελέτη αυτή, θα χρησιμοποιηθούν τα ετήσια δεδομένα (έξοδα διαφημίσεων και αντίστοιχες πωλήσεις) της εταιρίας Procter and Gamble για τα έτη 1987 έως 2015. Όπως και σε προηγούμενες ενότητες, θα θεωρήσουμε τα δεδομένα για τα τελευταία 3 έτη (2013-2015) άγνωστα, ώστε να μπορέσουμε να επαληθεύσουμε την εγκυρότητα των προβλέψεων μας βάση του μοντέλου που θα υπολογιστεί.

Κοιτάζοντας αρχικά τα δεδομένα μας, μπορούμε να κάνουμε ένα απλό υπολογισμό του συντελεστή συσχέτισης μεταξύ των εξόδων διαφήμισης και των πωλήσεων, και να δούμε με ένα Pearson’s product-moment correlation test αν αυτή η συσχέτιση είναι στατιστικά διάφορη του μηδενός. Κάνοντας τον έλεγχο αυτό παίρνουμε ότι η συσχέτιση είναι σημαντική (σε επίπεδο σημαντικότητας μέχρι και 99.9%) και συγκεκριμένα η σημειακή μας εκτίμηση είναι 0.9965772 (δηλαδή απίστευτα κοντά στην μονάδα).

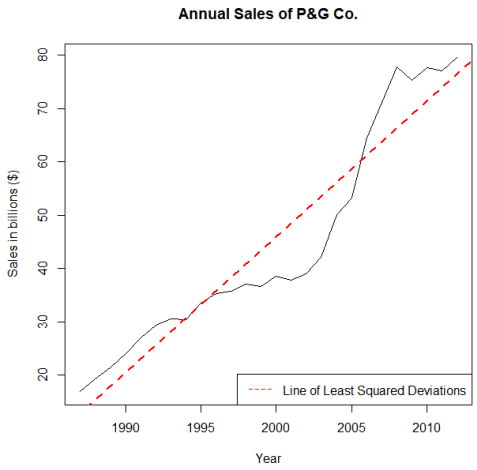
Σε αυτό το σημείο ίσως να διερωτώμασταν, εφόσον ο συντελεστής συσχέτισης μας είναι τόσο υψηλός, γιατί να μην χρησιμοποιήσουμε ένα απλό γραμμικό μοντέλο παλινδρόμησης που να συνδέει τις τρέχον πωλήσεις με τα τρέχων έξοδα διαφήμισης και να αποφύγουμε πλήρως την ανάλυση χρονοσειρών.

Κάτι τέτοιο θα ήταν αρκετά επιπόλαιο, καθώς θα θεωρούσαμε ότι τα έξοδα διαφήμισης είναι ο μοναδικός παράγοντας που ελέγχει απόλυτα τις πωλήσεις (πράγμα που σαφώς δεν ισχύει) καθώς και ότι οι διαφημίσεις των προηγούμενων ετών δεν επηρεάζουν το τρέχον έτος. Για να ισχύουν οι πιο πάνω υποθέσεις θα έπρεπε οι καταναλωτές να είχαν μηδενική μνήμη (και ως προς τις διαφημίσεις αλλά και ως προς τις προτιμήσεις τους) κάτι που κάνει πλέον φανερό ότι το απλό γραμμικό μοντέλο αν και φαινομενικά ίσως να έδινε σχετικά καλά αποτελέσματα, δεν θα είχε καμιά λογική υπόσταση και τα συμπεράσματα αλλά και οι προβλέψεις βάσει αυτού, πολύ πιθανό να έσφαλλαν.

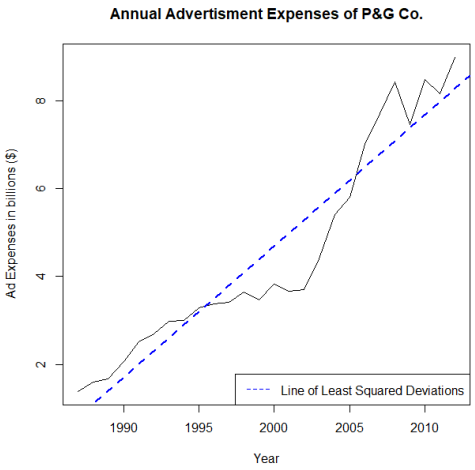
Έχοντας αναφέρει αυτό όμως, δεν μπορούμε να αρνηθούμε ότι μια τόσο ισχυρή συσχέτιση έχει σαφή ρόλο στους όγκους πωλήσεων. Επομένως θα χρησιμοποιήσουμε μοντέλα ADL που περιλαμβάνουν πωλήσεις σε προηγούμενους περιόδους αλλά και τα έξοδα διαφήμισης στην τρέχον και σε προηγούμενες περιόδους.

Όπως και στην προηγούμενη ενότητα ξεκινούμε με τα απλούστερα μοντέλα και σιγά σιγά προσθέτουμε περαιτέρω παραμέτρους μέχρι να καταλήξουμε σε ένα μοντέλο που προσαρμόζεται καλά στα δεδομένα μας.

Αρχικά βλέπουμε ότι ούτε οι πωλήσεις αλλά ούτε και τα έξοδα διαφήμισης είναι στάσιμα. Αυτό φαίνεται και στα Διαγράμματα 6.30 και 6.31 αλλά και από τους ελέγχους Augmented Dickey Fuller που απορρίπτουν την στασιμότητα.



Διάγραμμα 6.30

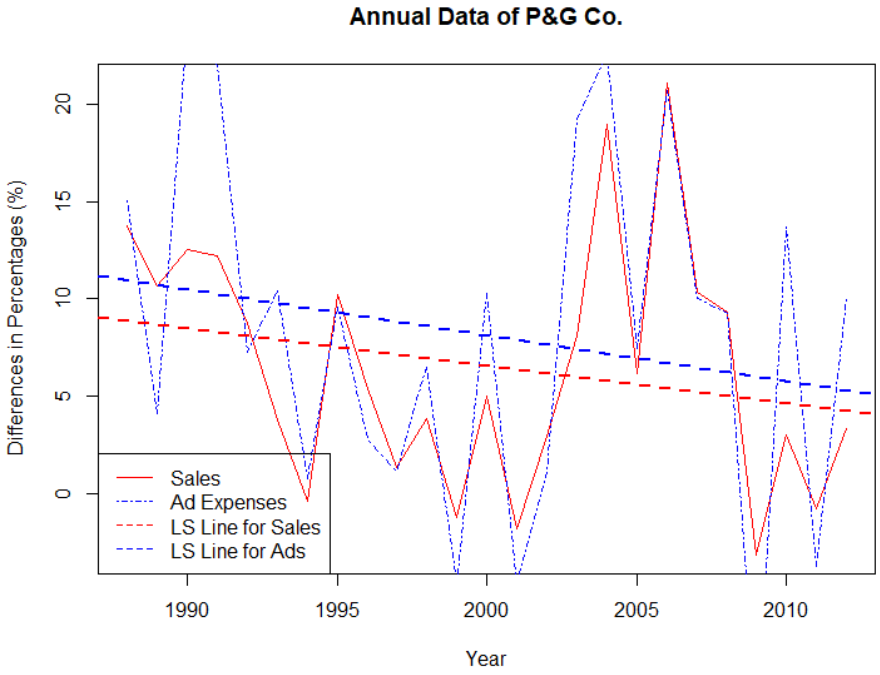


Διάγραμμα 6.31

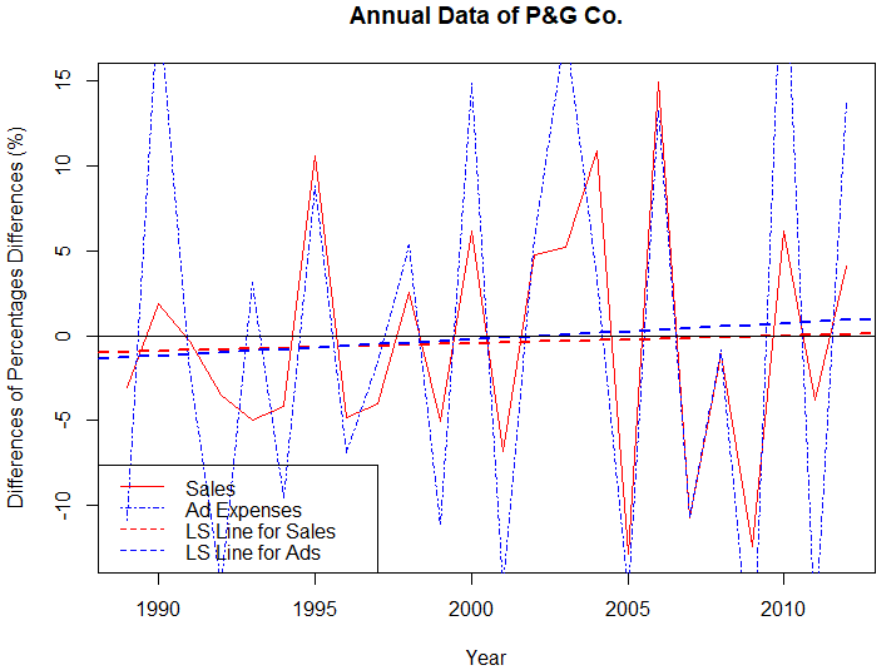
Η μεγάλη θετική συσχέτιση των μεγεθών γίνεται εύκολα αντιληπτή και από την εμφανή ομοιότητα των πιο πάνω γραφημάτων. Εφόσον όμως δεν παρουσιάζουν στασιμότητα τα δεδομένα μας, παίρνουμε τις ποσοστιαίες διαφορές μεταξύ διαδοχικών περιόδων και μελετάμε τις προσκόπτουσες νέες χρονοσειρές ως προς την στασιμότητα τους.

Το Διάγραμμα 6.32 συνοψίζει και τις δυο νέες χρονοσειρές και παρατηρούμε ότι και πάλι δεν έχουμε στασιμότητα. Το ίδιο επαληθεύει και ο έλεγχος Augmented Dickey Fuller.

Παίρνοντας τις διαφορές πρώτης τάξης των ποσοστιαίων διαφορών (αντίστοιχες των διαφορών δεύτερης τάξης), παίρνουμε πλέον στάσιμες σειρές, που φαίνονται στο Διάγραμμα 6.33. Σημαντικό είναι το γεγονός ότι η αυτοσυσχέτιση των μεγεθών παραμένει σημαντική και στις διαφορές αυτές.



Διάγραμμα 6.32



Διάγραμμα 6.33

Ο έλεγχος Augmented Dickey Fuller για την χρονοσειρά των εξόδων διαφήμισης μας υποδεικνύει στασιμότητα σε επίπεδο σημαντικότητας μεγαλύτερο του 95% ενώ ο αντίστοιχος έλεγχος για τις πωλήσεις μας παρουσιάζει ψηλότερη p-τιμή και θα μπορούσε να θεωρηθεί ως μη στάσιμη. Αυτό ίσως να οφείλεται στον μικρό όγκο των δεδομένων ή σε ανακρίβειες του ελέγχου καθώς οι πωλήσεις φαίνονται να είναι πιο “στάσιμες” από τα διαφημιστικά έξοδα στην περίπτωση αυτή.

Επομένως θα εργαστούμε από εδώ και πέρα με αυτές τις διαφορές μιας και είναι οι διαφορές μικρότερης τάξης που παρουσιάζουν ικανοποιητική στασιμότητα. Προσαρμόζοντας αρχικά το απλό μοντέλο ADL (1,0) που συνδέει τις πωλήσεις με τις πωλήσεις στην αμέσως προηγούμενη περίοδο και τα έξοδα διαφήμισης για την τρέχον περίοδο.

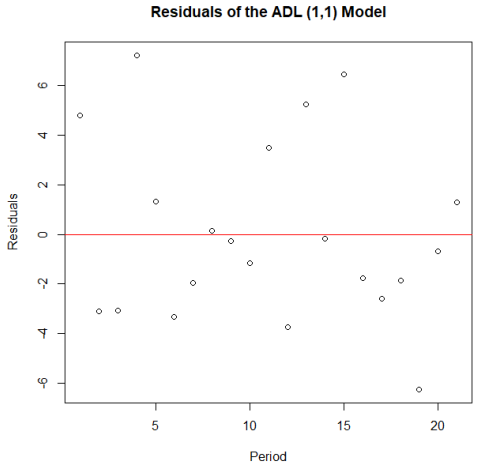
Το μοντέλο αυτό παρουσιάζει σημαντικά προβλήματα καθώς ο συντελεστής που αφορά την υστέρηση των πωλήσεων είναι μη σημαντικός και τα υπόλοιπα παρουσιάζουν εξάρτηση μεταξύ τους. Συνεπώς προχωρούμε στο επόμενο μοντέλο που θα δοκιμαστεί, το ADL (1,1). Το μοντέλο αυτό συνδέει τις πωλήσεις με τις πωλήσεις στην αμέσως προηγούμενη χρονική περίοδο καθώς και με τα έξοδα των διαφημίσεων για τις δυο τελευταίες χρονικές περιόδους.

Το μοντέλο αυτό φαίνεται να περιγράφει ικανοποιητικά τα δεδομένα μας και να έχει όλους του τους συντελεστές σημαντικά διάφορους του μηδέν (με εξαίρεση του σταθερού όρου). Επίσης οι έλεγχοι Breusch-Godfrey δεν μας απορρίπτουν την υπόθεση ανεξαρτησίας για διαδοχικά υπόλοιπα τάξης 1 έως 5. Ο γενικός έλεγχος Ljung-Box για την συνολική εικόνα των υπολοίπων μας δίνει μια πολύ μεγάλη p-τιμή και σαφώς δεν μπορούμε να απορρίψουμε την ανεξαρτησία των υπολοίπων.

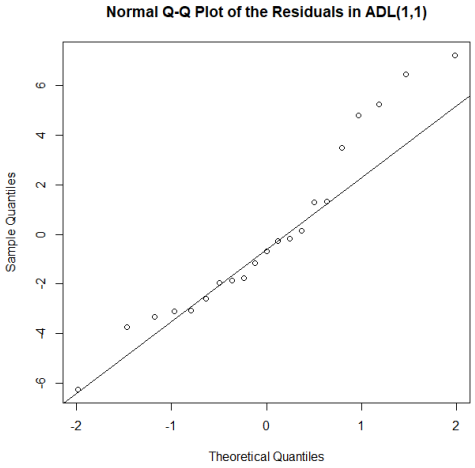
```
> mod2<-lm(L0.sales ~L1.sales +L0.adds+L1.adds)
> summary(mod2)

Call:
lm(formula = L0.sales ~ L1.sales + L0.adds + L1.adds)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2446 -2.5823 -0.6837  1.3380  7.2141

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01285    0.86985   0.015  0.98838
L1.sales    -0.55452    0.19559  -2.835  0.01143 *
L0.adds      0.52259    0.08662   6.033 1.34e-05 ***
L1.adds      0.37191    0.12553   2.963  0.00872 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.924 on 17 degrees of freedom
Multiple R-squared:  0.7837,    Adjusted R-squared:  0.7455
F-statistic: 20.53 on 3 and 17 DF,  p-value: 6.871e-06

> confint(mod2)
                2.5 %     97.5 %
(Intercept) -1.8223623  1.8480681
L1.sales    -0.9671804 -0.1418501
L0.adds      0.3398286  0.7053524
L1.adds      0.1070720  0.6367507

> Box.test(res.mod2,lag=20,type="Ljung-Box")

        Box-Ljung test

data:  res.mod2
X-squared = 11.864, df = 20, p-value = 0.9207
```

Επομένως το μοντέλο ADL (1,1) αξίζει να μελετηθεί περισσότερο ώστε να εξακριβώσουμε αν πληρούνται όλες οι προϋποθέσεις που απαιτούμε για την καταλληλότητα του μοντέλου. Συνεχίζουμε όμως την μελέτη μας ερευνώντας κατά πόσο μοντέλα ανώτερης τάξης μας παρουσιάζουν καλύτερη συμπεριφορά. Κοιτάζοντας τα μοντέλα ADL (2,0), (2,1) και (2,2) διαπιστώνουμε ότι τι μοναδικό που φαίνεται να πληροί την υπόθεση ανεξαρτησίας υπολοίπων είναι το (2,2) που όμως έχει μη-σημαντικούς συντελεστές και ως εκ τούτου θα προτιμηθεί το μοντέλο ADL (1,1).

Στο Διάγραμμα 6.34, βλέπουμε ότι τα υπόλοιπα φαίνονται να είναι τυχαία και χωρίς μοτίβα σκορπισμένα γύρω από τον μηδενικό άξονα, που είναι μια καλή ένδειξη κανονικότητας υπολοίπων. Στο Διάγραμμα 6.35 βλέπουμε τον γραφικό έλεγχο qq-plot για τα υπόλοιπα. Τα δεδομένα μας είναι λίγα και φαίνεται να έχουν κάποιες αποκλίσεις από την γραμμή κανονικότητας.



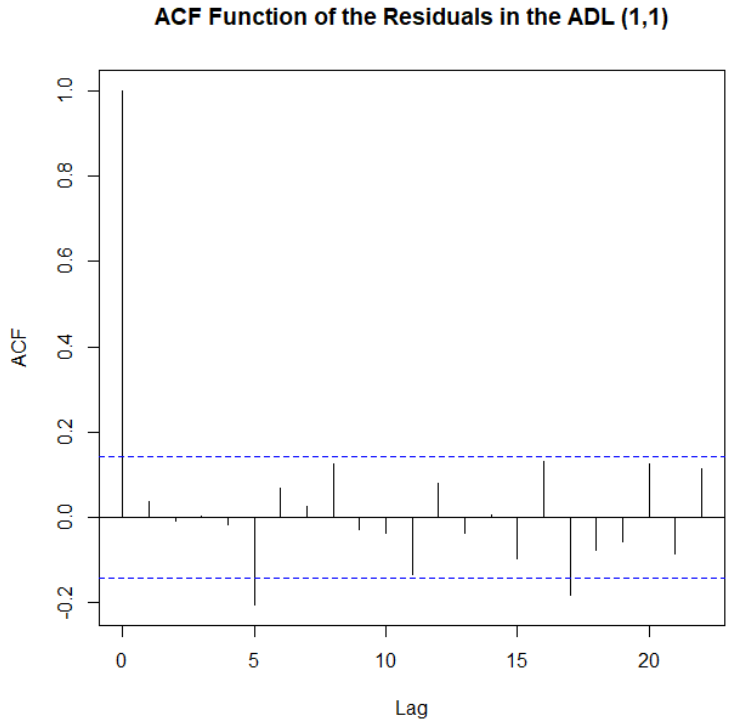
Διάγραμμα 6.34



Διάγραμμα 6.35

Για την καλύτερη εικόνα της κανονικότητας των υπολοίπων, χρησιμοποιούμε τους ελέγχους κανονικότητας Shapiro-Wilk και Kolmogorov-Smirnov. Ο πρώτος έλεγχος μας δίνει p-τιμή της τάξεως του 20%, κάτι που υποδεικνύει μη-κανονική συμπεριφορά των υπολοίπων. Ο δεύτερος όμως έλεγχος μας δίνει p-τιμή της τάξεως το 0.2%, που μας υποδεικνύει την κανονικότητα των υπολοίπων. Η μεγάλη απόκλιση των δυο ελέγχων είναι αρκετά ασυνήθιστή και πιθανόν και πάλι να οφείλεται στο μικρό πλήθος των παρατηρήσεων μας.

Το γράφημα που περιλαμβάνει τις τιμές της συνάρτησης αυτοσυσχέτισης (ACF) για τα υπόλοιπα, που φαίνεται στο Διάγραμμα 6.36, επιβεβαιώνει την ανεξαρτησία που είχαμε ελέγξει και πριν με τους ελέγχους Ljung-Box και Breusch-Godfrey.



Διάγραμμα 6.36

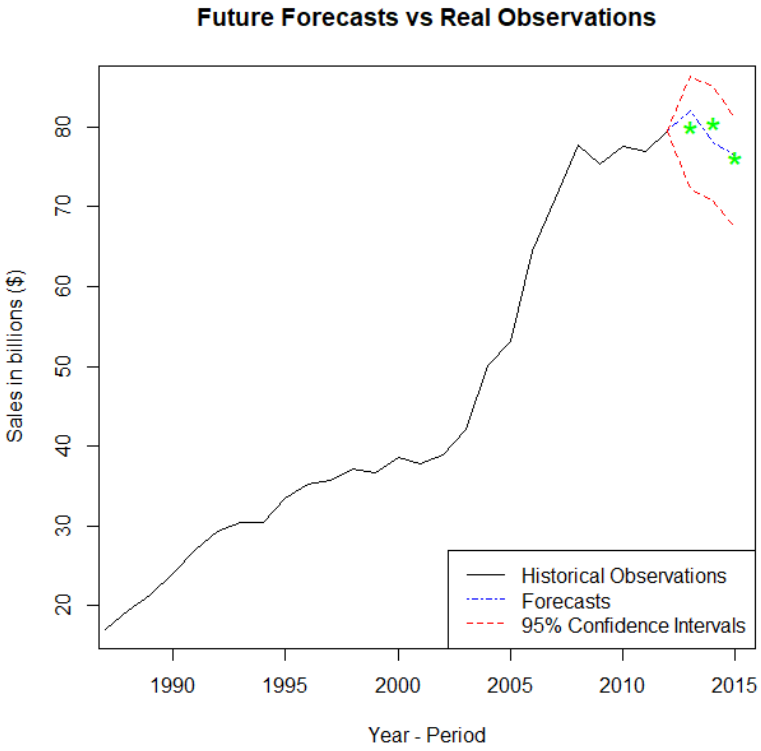
Τέλος, ελέγχοντας και για την κανονικότητα και την ανεξαρτησία των υπολοίπων, χρησιμοποιούμε τους ελέγχους που συζητήθηκαν στην ενότητα 5.3.2, που και οι 3 μας δίνουν αρκετά μεγάλες p-τιμές και συνεπώς δεν καταρρίπτουν την υπόθεση κανονικότητας και ανεξαρτησίας σφαλμάτων. Επομένως το μοντέλο ADL (1,1) φαίνεται ικανοποιητικό για τα δεδομένα μας και μπορεί να χρησιμοποιηθεί για προβλέψεις.

Συμβολίζοντας με Z_t τις τιμές των διαφορών των πωλήσεων στην χρονική στιγμή t, και με X_t τις αντίστοιχες τιμές των διαφορών των εξόδων διαφήμισης, η ακριβής μορφή του εν λόγω μοντέλου φαίνεται στην πιο κάτω εξίσωση:

$$Z_t = 0.01285 - 0.55452Z_{t-1} + 0.52259X_t + 0.37191X_{t-1} + \varepsilon_t .$$

Όπως έχει ήδη σχολιαστεί, όλοι οι συντελεστές είναι σημαντικοί με εξαίρεση του σταθερού όρου. Χρησιμοποιώντας το μοντέλο αυτό μπορούμε να κάνουμε προβλέψεις σχετικά με τις τιμές των πωλήσεων για τα έτη 2013 έως 2015.

Στο Διάγραμμα 6.37 που φαίνεται στην επόμενη σελίδα, παρουσιάζονται με μαύρη γραμμή όλα τα δεδομένα που έχουμε για τις πωλήσεις (μέχρι και το 2012), με μπλε γραμμή συμβολίζουμε τις σημειακές εκτιμήσεις βάσει του πιο πάνω μοντέλου, με κόκκινες γραμμές συμβολίζουμε τα άκρα του 95% διαστήματος εμπιστοσύνης για τις προβλέψεις μας, και με πράσινους αστερίσκους σημειώνουμε τις πραγματικές πωλήσεις που επιτεύχθηκαν κατά τα έτη 2013-2015.



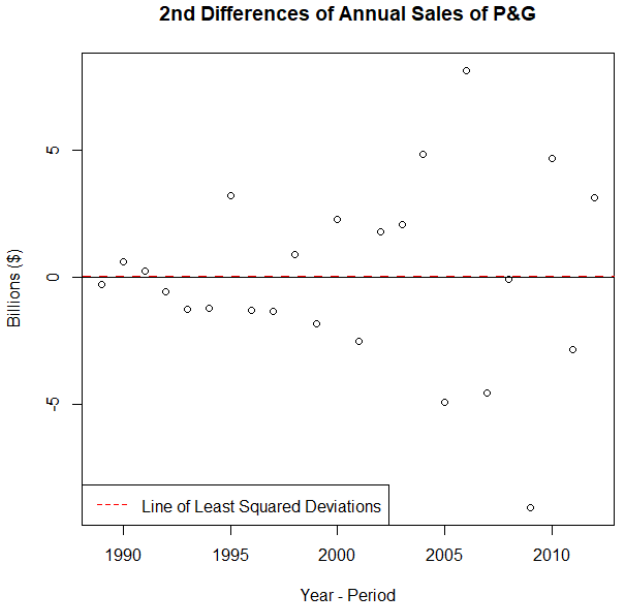
Διάγραμμα 6.37

Βλέπουμε λοιπόν ότι οι προβλέψεις μας θα ήταν αρκετά ικανοποιητικές και συνεπώς το μοντέλο αυτό περιγράφει ικανοποιητικά τα δεδομένα μας και θα είναι χρήσιμο και για μελλοντικές προβλέψεις.

Θα μπορούσαμε να μελετήσουμε μόνο την μεταβλητή πωλήσεων ως μια χρονοσειρά και να αγνοήσουμε τα έξοδα διαφημίσεων πλήρως, και να κάνουμε μια παρόμοια μελέτη με προηγούμενα προβλήματα, δηλαδή με χρήση μονοδιάστατων μοντέλων ARIMA.

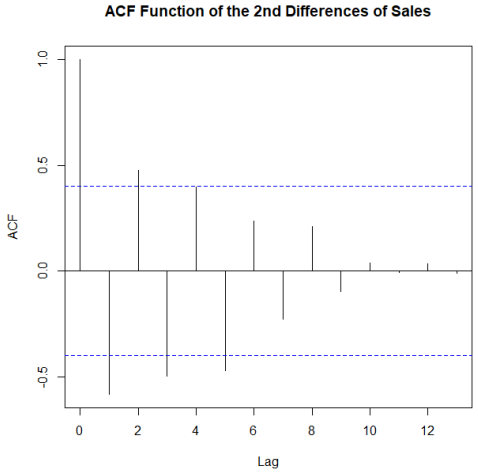
Όπως και πριν οι πωλήσεις δεν παρουσιάζουν στασιμότητα και για τον λόγο αυτό χρησιμοποιούμε διαφορές πρώτης τάξης. Παρατηρούμε και γραφικά αλλά και με τον έλεγχο Augmented Dickey Fuller ότι τα δεδομένα μας δεν παρουσιάζουν στασιμότητα.

Παίρνοντας διαφορές δεύτερης τάξης καταλήγουμε σε χρονοσειρά με στάσιμη συμπεριφορά, πράγμα που μας υποδεικνύει ότι ο βαθμός integration θα είναι 2. Συγκεκριμένα στο Διάγραμμα 6.38 μπορούμε να δούμε διαφορές δεύτερης τάξης των ετήσιων πωλήσεων. Η κόκκινη διακεκομμένη γραμμή είναι η ευθεία ελαχίστων τετραγώνων και συμπίπτει σχεδόν πλήρως με την μηδενική ευθεία, πράγμα που μας υποδεικνύει ότι τα δεδομένα μας παρουσιάζουν στασιμότητα (εφόσον δεν υπάρχουν θετικές ή αρνητικές τάσεις).

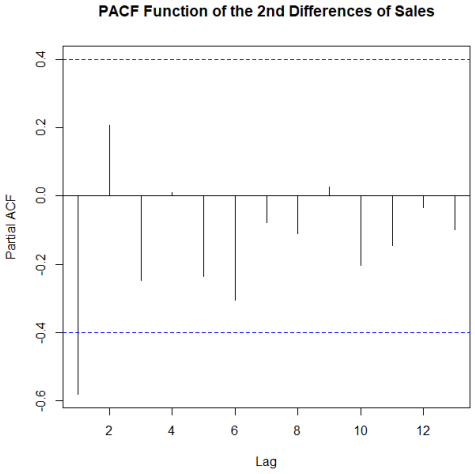


Διάγραμμα 6.38

Για να εκτιμήσουμε τις αρχικές τιμές του πλήθους των παραμέτρων p και q στο γενικό μοντέλο ARIMA (p,d,q), βλέπουμε τις πιο κάτω γραφικές παραστάσεις που απεικονίζουν τις τιμές των συναρτήσεων ACF και PACF για τις τιμές των δευτέρων διαφορών των πωλήσεων.



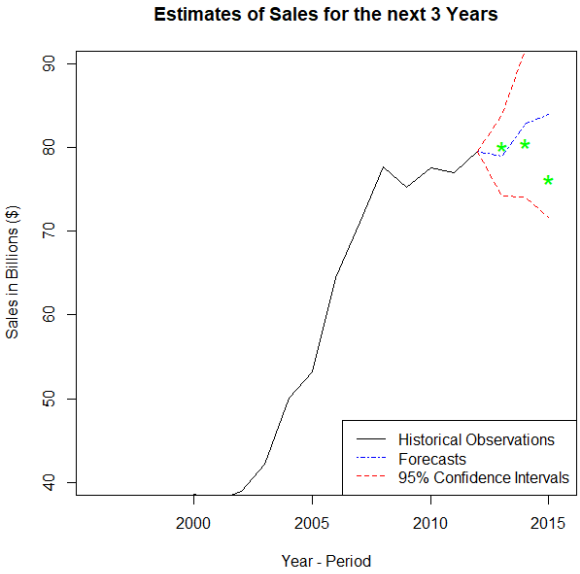
Διάγραμμα 6.39



Διάγραμμα 6.40

Παρατηρούμε από το Διάγραμμα 6.39 ότι θα ήταν καλό να ελεγχθούν αρχικά μοντέλα με 4 όρους MA, και από το Διάγραμμα 6.40 ότι το πιο πιθανό ένας όρος AR θα ήταν ικανοποιητικός. Πράγματι, προσαρμόζοντας αρκετά μοντέλα ARIMA (p,d,q) καταλήγουμε ότι καλύτερη προσαρμογή έχει το μοντέλο ARIMA (1,2,3).

Μελετώντας περαιτέρω το μοντέλο αυτό καταλήγουμε στο ότι τα υπόλοιπα είναι ικανοποιητικώς ανεξάρτητα και κανονικά κατανεμημένα, και συνεπώς το μοντέλο φαίνεται κατάλληλο για προβλέψεις. Οι προβλέψεις για τα έτη 2013-2015 φαίνονται στο Διάγραμμα 6.41 κρατώντας ίδιους χρωματικούς συμβολισμούς όπως στο Διάγραμμα 6.37.



Διάγραμμα 6.41

Βλέπουμε ότι οι προβλέψεις μας δεν είναι το ίδιο ικανοποιητικές και ότι τα διαστήματα εμπιστοσύνης έχουν μεγαλώσει. Ο λόγος είναι ότι στο μονοδιάστατο μοντέλο ARIMA έχουμε μεγαλύτερα τυπικά σφάλματα καθώς έχουμε λιγότερες σχετικές πληροφορίες.

Επομένως το μοντέλο ADL (1,1) είναι καλύτερο στο να περιγράψει τα δεδομένα και να χρησιμοποιηθεί για μελλοντικές προβλέψεις. Πέρα αυτού όμως, οι τιμές των συντελεστών του μοντέλου αυτού μπορούν να μας δώσουν χρήσιμες πληροφορίες για την αντιμετώπιση της αγοράς στις διαφημίσεις της εταιρίας. Όπως συζητήθηκε και στην ενότητα 3, θα μπορούσαμε να υπολογίσουμε τους δυναμικούς πολλαπλασιαστές αλλά και τον πολλαπλασιαστή βάθος χρόνου.

Η γνώση των ποσοτήτων αυτών είναι πολύ χρήσιμη καθώς θα δώσουν απαντήσεις στα ερωτήματα “ποια θα είναι η επιρροή στις πωλήσεις των επόμενων χρόνων από μια μεμονωμένη αύξηση στα έξοδα διαφήμισης για ένα έτος;” και “πόσο θα επηρεαστεί το επίπεδο των πωλήσεων μας, αν αυξήσουμε τα διαφημιστικά έξοδα επ’αόριστον κατά κάποιο ποσό;”. Γνωρίζοντας τα πιο πάνω θα βρισκόμαστε σε θέση να κρίνουμε κατά πόσο θα είναι κερδοφόρο για την εταιρία να προβεί σε τέτοιες αλλαγές, αλλά θα μπορούμε επίσης να υπολογίσουμε το ποσό το οποίο αν επενδύσουμε στα έξοδα διαφήμισης, θα μεγιστοποιήσει το κέρδος της.

Ο πολλαπλασιαστής βάθος χρόνου υπολογίζεται όπως έχει συζητηθεί στην ενότητα 3, και είναι ίσος με 0.57542, ενώ οι υπόλοιποι πολλαπλασιαστές παρουσιάζονται στον συγκεντρωτικό Πίνακα 6.4 που βρίσκεται στην επόμενη σελίδα.

Μια μεμονωμένη διακύμανση στην τιμή X_t (Έξοδα Διαφημίσεων)	
Περίοδος	**Επιρροή στο Y_t (Πωλήσεις)**
0	0.52259
1	0.08212
2	-0.04554
3	0.025252
4	-0.014
5+	**<0.01%** (κατά απόλυτη τιμή)

Πίνακας 6.4

Συνεπώς μπορούμε να εκτιμήσουμε βάση του μοντέλου μας, ότι σε περίπτωση που επιλέξουμε να αυξήσουμε τα διαφημιστικά έξοδα για το τρέχον έτος κατά 10% θα αναμέναμε 5.226% αύξηση στις πωλήσεις του ίδιου έτους, ενώ θα περιμέναμε 0.82% αύξηση στις πωλήσεις και του επόμενου έτους που θα ακολουθούνταν από μείωση στις πωλήσεις στην μεθεπόμενη περίοδο κατά 0.455% ενώ για τις μετέπειτα περιόδους οι διαφοροποιήσεις στις πωλήσεις λόγω αυτής της μεμονωμένης επένδυσης θα είναι αμελητέες.

Το γεγονός ότι η αύξηση στα έξοδα διαφημίσεων θα επιφέρει μείωση στις πωλήσεις μετά από 2 περιόδους είναι όντως κάτι περίεργο (που το πιο πιθανό δεν ισχύει) που προκύπτει λόγο τις προσεγγιστικής φύσης του μοντέλου. Γνωρίζοντας τις πληροφορίες αυτές θα μπορούσαμε να αποφανθούμε για το κατά πόσο μας συμφέρει ή όχι να αυξήσουμε τα έξοδα των διαφημίσεων για κάποια χρονική περίοδο.

Σε περίπτωση που η εταιρεία θα ενδιαφερόταν στο να αυξήσει μόνιμα τα έξοδα διαφημίσεων κατά 10% τότε θα αναμέναμε ότι το επίπεδο των πωλήσεων μας για κάθε μελλοντικό έτος, θα ήταν αυξημένο (λόγω τις επένδυσης αυτής) κατά 5.7542%.

7. Συμπεράσματα

Στην παρούσα διπλωματική εργασία, είδαμε και μελετήσαμε κάποια βασικά μοντέλα Ανάλυσης Χρονοσειρών και τις εφαρμογές τους. Πράγματι τα μοντέλα αυτά είναι αρκετά χρήσιμα και μπορούν να δώσουν λύσεις σε πολλά προβλήματα που προκύπτουν στην καθημερινότητα.

Παρόλα αυτά όμως, υπάρχει μια τεράστια ποικιλία μοντέλων που είναι ικανά να περιγράψουν πιο περίπλοκα και ιδιόρρυθμα δεδομένα.

Ένα απλό παράδειγμα είναι τα μοντέλα **ARCH** (Auto-Regressive Conditionally Heteroskedastic Models), που είναι εκ κατασκευής ετεροσκεδαστικά. Τέτοια μοντέλα μπορούν να χρησιμοποιηθούν σε χρονοσειρές που η διασπορά των τιμών τους μεταβάλλεται στον χρόνο.

Επίσης τα δεδομένα που έχουν επιλεγεί στην ενότητα των εφαρμογών, έχουν επιλεγεί σε κατάλληλα χρονικά διαστήματα ώστε να μην υπάρχει εποχικότητα. Σε περίπτωση που μας ενδιαφέραν προβλέψεις που τα δεδομένα μας παρουσίαζαν εποχικότητα, θα έπρεπε να χρησιμοποιήσουμε μια άλλη μεγάλη κλάση μοντέλων, τα **SARIMA** (Seasonal Auto Regressive Integrated Moving Average Models) που είναι ικανά, είτε με την χρήση διορθωτικών παραγόντων, είτε με την χρήση τριγωνομετρικών συναρτήσεων, να περιγράψουν κατάλληλα δεδομένα αυτού του τύπου.

Σε περιπτώσεις που μας ενδιαφέρει ένα σύνολο παρατηρήσεων σε διάφορες χρονικές στιγμές, θα χρησιμοποιούσαμε αντίστοιχα μοντέλα για διανύσματα, όπως είναι τα VAR (Vector Auto-Regression Models).

Πέρα αυτών, η ανάπτυξη αυτοδιορθωτικών μοντέλων βάσει πολύπλοκων αλγορίθμων και συγκεκριμένα η ανάλυση χρονοσειρών με την χρήση εκμάθησης μηχανής (Machine Learning Time Series Analysis) είναι ένα πολύ επίκαιρο θέμα που παρουσιάζει εξέλιξη και απασχολεί εκατοντάδες ερευνητές.

Η Ανάλυση Χρονοσειρών λοιπόν, είναι ένας πολύ μεγάλος, ενδιαφέρον και σύγχρονος κλάδος της Στατιστικής που παρουσιάζει ραγδαία εξέλιξη τις τελευταίες δεκαετίες με την ανάπτυξη ισχυρών στατιστικών πακέτων και μεθόδων.

Τα θέματα που συζητήθηκαν στην διπλωματική αυτή, δεν είναι τίποτα παραπάνω από μια εισαγωγή στην Ανάλυση Χρονοσειρών.

Βιβλιογραφία

Α) Διεθνής Βιβλιογραφία

Blanchard O. & Johnson D.R. (2011). *Macroeconomics.* Pearson Education Inc. Boston.

Box G.E.P., Jenkins G.M. & Reinsel G.C. (2008). *Time Series Analysis – Forecasting and Control.* John Wiley & Sons. New Jersey.

Box G.E.P. & Pierce D.A. (1970). Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association,* **65***,* 1509-1526*.*

Brockwell P.J. & Richard A.D. (2002). *Introduction to Time and Forecasting*. Springer. New York.

Dickey D.A. & Fuller W.A. (1979). Distribution of the estimates for autoregressive time series with a unit root. *Journal of the American Statistical Association,* **74***,* 427-431*.*

Durbin J. (1970). Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. *Econometrica,* **38**, 410-421.

Frees E.W. (1996). *Data Analysis Using Regression Models – The Business Perspective.* Prentice Hall. New Jersey.

Hamilton J.D. (1994). *Time Series Analysis.* Princeton University Press. New Jersey.

Hanke J.E. & Wichern D.W. (2009). *Business Forecasting.* Prentice Hall. New Jersey.

Hewins. R.D. (2014). *Management Mathematics*. University of London. London.

Levine D.M., Berenson M.L. & Stephan D. (1999). *Statistics for Managers Using Microsoft Excel.* Prentice Hall. New Jersey.

Ljung G.M. & Box G.E.P. (1978). On a Measure of a Lack of Fit in Time Series Models. *Biometrica,* **65**, 297-303.

Nkoro E. & Uko A.K. (2016). Autoregressive Distributed Lag (ARDL) Cointegration Technique: Application and Interpretation. *Journal of Statistical and Econometrics Methods*, **5**, 61-91.

Β) Ελληνική Βιβλιογραφία

Βόντα Ι. & Καραγρηγορίου Α. (2012). *Εφαρμοσμένη Στατιστική Ανάλυση & Στοιχεία Πιθανοτήτων.* Mike Printings. Λάρνακα.

Κοκολάκης Γ. & Φουσκάκης Δ. (2009). *Στατιστική Θεωρία & Εφαρμογές.* Εκδόσεις Συμεών. Αθήνα.

Οικονόμου Π. & Καρώνη Χ. (2010). *Στατιστικά Μοντέλα Παλινδρόμησης.* Εκδόσεις Συμεών. Αθήνα.

Φουσκάκης Δ. (2013). *Ανάλυση Δεδομένων με Χρήση της R.* Εκδόσεις Τσότρας. Αθήνα.

Χρυσαφίνου Ο. (2012). *Εισαγωγή Στις Στοχαστικές Ανελίξεις.* Εκδόσεις σοφία. Θεσσαλονίκη.