ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΑΛΙΝΔΡΟΜΙΚΑ ΜΟΝΤΕΛΑ ΓΙΑ ΔΥΑΔΙΚΕΣ ΚΑΙ ΑΠΑΡΙΘΜΗΤΕΣ ΧΡΟΝΟΣΕΙΡΕΣ

Βαρυπάτη Σοφία

ΑΘΗΝΑ, ΑΠΡΙΛΙΟΣ 2012

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΑΛΙΝΔΡΟΜΙΚΑ ΜΟΝΤΕΛΑ ΓΙΑ ΔΥΑΔΙΚΕΣ ΚΑΙ ΑΠΑΡΙΘΜΗΤΕΣ ΧΡΟΝΟΣΕΙΡΕΣ

Ονοματεπώνυμο: Βαρυπάτη Σοφία

Επιβλέπουσα καθηγήτρια: Καρώνη Χ. – Αναπλ. Καθηγήτρια

Τριμελής επιτροπή: Καρώνη Χ. – Αναπλ. Καθηγήτρια
Παπανικολάου Β. – Καθηγητής
Σπηλιώτης Ι. – Αναπλ. Καθηγητής

ΑΘΗΝΑ, ΑΠΡΙΛΙΟΣ 2012

Στους γονείς μου,

Μενέλαο και Ιωάννα

ΠΡΟΛΟΓΟΣ

Παραδείγματα χρονοσειρών μπορούμε να συναντήσουμε στα περισσότερα επιστημονικά πεδία και είναι γνωστό πως αποτελούν σημαντικό στατιστικό εργαλείο για τη μελέτη πολλών φαινομένων και προβλημάτων. Οι δυαδικές *(binary)* και οι απαριθμητές *(count)* χρονοσειρές, αποτελούν δύο ειδικές περιπτώσεις χρονοσειρών που πολύ συχνά επιλέγονται για διάφορες εφαρμογές. Είναι όμως γεγονός, ότι κατά τη μοντελοποίηση μιας χρονοσειράς γενικότερα, αλλά και άλλων διαδικασιών όπως η εκτίμηση των παραμέτρων ή οι έλεγχοι υποθέσεων και καταλληλότητας του μοντέλου, αντιμετωπίζονται αρκετά προβλήματα τα οποία ισχύουν εξίσου και για τα δύο αυτά, συγκεκριμένα είδη χρονοσειρών.

Σκοπός λοιπόν της παρούσας εργασίας, είναι να παρουσιάσουμε έναν τρόπο μοντελοποίησης των χρονοσειρών αυτών, με τον οποίο να μπορούν να ξεπεραστούν τα προαναφερθέντα προβλήματα. Αυτό θα γίνει με τη χρήση συγκεκριμένων παλινδρομικών μοντέλων, των οποίων η χρηστικότητα βασίζεται αφενός στη δομή των ***γενικευμένων γραμμικών μοντέλων*** και αφετέρου στη συμπερασματολογία της ***μεθόδου της μερικής πιθανοφάνειας***, η οποία επιτρέπει την ύπαρξη των χρονοεξαρτώμενων δεδομένων που περιέχει μια χρονοσειρά. Πιο συγκεκριμένα, θα δούμε τη διαδικασία μοντελοποίησης των δυαδικών χρονοσειρών, μέσω του λογιστικού παλινδρομικού μοντέλου και των απαριθμητών χρονοσειρών, μέσω του παλινδρομικού μοντέλου Poisson.

Μπορούμε να σκεφτούμε την εργασία, χωρισμένη σε δύο βασικά μέρη, στα δύο πρώτα κεφάλαια και στα δύο επόμενα. Στο πρώτο κεφάλαιο, ξεκινάμε με την ανάλυση των γενικευμένων γραμμικών μοντέλων και την περιγραφή της "κλασσικής" λογιστικής και Poisson παλινδρόμησης, ενώ στο δεύτερο κεφάλαιο παρουσιάζονται βασικές έννοιες της ανάλυσης χρονοσειρών και τα βασικά στάδια μοντελοποίησης τους. Στο δεύτερο μέρος, το οποίο αποτελείται από το τρίτο και το τέταρτο κεφάλαιο, ουσιαστικά "συνδυάζονται" οι έννοιες των δύο κεφαλαίων που προηγήθηκαν. Δηλαδή, στο τρίτο κεφάλαιο αναλύεται ο τρόπος με τον οποίο μπορούν να προσαρμοστούν οι χρονοσειρές στα γενικευμένα γραμμικά μοντέλα και παρουσιάζεται η σημαντική μέθοδος της μερικής πιθανοφάνειας, μέσω της οποίας επιτυγχάνεται η συμπερασματολογία για τα μοντέλα αυτά. Ενώ το τέταρτο κεφάλαιο, συνιστά την πρακτική εφαρμογή μοντελοποίησης δυαδικής και απαριθμητής χρονοσειράς μέσω των αντίστοιχων παλινδρομικών μοντέλων, του λογιστικού και του Poisson.

Για τις εφαρμογές αυτές, αλλά και για τα μικρότερα παραδείγματα που παρουσιάζονται στο πρώτο και στο δεύτερο κεφάλαιο, έχει χρησιμοποιηθεί το στατιστικό πρόγραμμα R, το οποίο μπορεί να μην είναι τόσο φιλικό στο χρήστη συγκριτικά με άλλα προγράμματα, αλλά έχει το πλεονέκτημα του προγράμματος ανοιχτού κώδικα *(open source),* της πολύ μεγάλης "γκάμας" δυνατοτήτων μέσω των βιβλιοθηκών *(libraries)* και των εντολών που διαθέτει, και ότι φυσικά μπορεί ο οποιοσδήποτε ενδιαφερόμενος να το εγκαταστήσει δωρεάν στον υπολογιστή του, μέσω Διαδικτύου.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω την κυρία Χρυσηίδα Καρώνη, Αναπληρώτρια Καθηγήτρια στο Ε.Μ.Π και επιβλέπουσα καθηγήτρια της διπλωματικής μου εργασίας, για τη συνεχή καθοδήγηση που μου παρείχε μέχρι και την ολοκλήρωση της παρούσης εργασίας. Θέλω να ευχαριστήσω επίσης την οικογένεια μου, για την αμέριστη συμπαράσταση τους καθ'όλη τη διάρκεια των σπουδών μου, καθώς και τους φίλους μου, που υπήρξαν ο καθένας με τον τρόπο του, σημαντικό στήριγμα σε όλη αυτή τη διαδρομή.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ, ΛΟΓΙΣΤΙΚΗ ΚΑΙ POISSON ΠΑΛΙΝΔΡΟΜΗΣΗ

1.1 ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

1.1.1 Γενικά στοιχεία

Η ανάλυση παλινδρόμησης αποτελεί ένα σημαντικότατο εργαλείο της στατιστικής, και χρησιμοποιείται, όταν ενδιαφέρει η εύρεση της όποιας σχέσης μεταξύ κάποιων μεταβλητών. Πιο συγκεκριμένα, μέσω της ανάλυσης παλινδρόμησης, διαθέτουμε τις απαραίτητες τεχνικές ώστε να κατασκευάσουμε το κατάλληλο μοντέλο που περιγράφει τη σχέση μεταξύ μιας εξαρτημένης μεταβλητής και κάποιων ανεξάρτητων μεταβλητών (που μπορεί να είναι μία ή και περισσότερες). Έτσι, θα είμαστε σε θέση να προβλέψουμε, πώς και αν αλλάζει η χαρακτηριστική τιμή της εξαρτημένης μεταβλητής όταν αλλάζει κάθε μια από τις ανεξάρτητες.

Να διευκρινίσουμε ότι ανεξάρτητες μεταβλητές *(X)*, ονομάζονται αυτές που καθορίζουμε εμείς τις τιμές τους, όταν το πείραμα είναι ελεγχόμενο, ενώ η εξαρτημένη μεταβλητή *(Y)*, είναι αυτή στην οποία αντανακλάται το αποτέλεσμα των μεταβολών των ανεξάρτητων μεταβλητών και της οποίας μας ενδιαφέρει η πρόβλεψη.

Οι μεταβλητές γενικώς, μπορούν να είναι είτε διακριτές, όπως είναι πχ ο αριθμός κάποιων ατυχημάτων, είτε κατηγορικές, όπως είναι πχ το φύλο, είτε συνεχείς, να παίρνουν δηλαδή οποιαδήποτε τιμή σε κάποιο διάστημα, και αναλόγως ακολουθείται και η κατάλληλη παλινδρόμηση. Για παράδειγμα, στην περίπτωση που δεν έχουμε μία συνεχή μεταβλητή απόκρισης, αλλά μία διακριτή, τότε δεν μπορούμε να χρησιμοποιήσουμε γραμμική παλινδρόμηση, αλλά θα πρέπει να επιλέξουμε ίσως την Poisson την οποία θα δούμε αναλυτικά παρακάτω.

Με βάση τα παραπάνω καταλαβαίνουμε ότι ένα μοντέλο παλινδρόμησης ουσιαστικά αποτελείται από μια συνάρτηση που συνδέει την εξαρτημένη μεταβλητή με τις ανεξάρτητες:

$$\Upsilon \approx f(\mathrm{X}, \beta)$$

όπου,

Y, η εξαρτημένη μεταβλητή (ή αλλιώς μεταβλητή απόκρισης)

X, η ανεξάρτητη μεταβλητή, που μπορεί να είναι και διάνυσμα (ή αλλιώς επεξηγηματική μεταβλητή)

β, οι άγνωστοι συντελεστές ή αλλιώς παράμετροι του μοντέλου, τις οποίες και εκτιμούμε μέσω της παλινδρόμησης και ελέγχουν τη συμπεριφορά του μοντέλου

1.1.2 Γραμμικό μοντέλο παλινδρόμησης

Συνήθως, η μοντελοποίηση της σχέσης μεταξύ κάποιων μεταβλητών, έχει στόχο να περιγράψει, πώς η μέση τιμή της εξαρτημένης μεταβλητής E(Y) αλλάζει, όταν μεταβάλλονται οι συνθήκες, υποθέτοντας ότι η διακύμανση παραμένει σταθερή και συνεπώς και το "σχήμα" της συνάρτησης παραμένει αμετάβλητο.

Για παράδειγμα, το πιο απλό μοντέλο παλινδρόμησης, περιγράφεται από τη συναρτησιακή σχέση:

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i \qquad (1.1)$$

Το απλό γραμμικό μοντέλο αποτελείται από μια μόνο ανεξάρτητη μεταβλητή. Όπως βλέπουμε από την παραπάνω σχέση, η μέση τιμή της μεταβλητής Y μεταβάλλεται με σταθερό ρυθμό β_1 όταν αυξάνεται ή μειώνεται η X κατά μία μονάδα. Είναι επίσης προφανές ότι αναπαριστάται από μία ευθεία γραμμή της οποίας η κλίση δίνεται από τη μεταβλητή β_1 και ότι υπάρχει μια γραμμική εξάρτηση μεταξύ της μέσης τιμής και των συντελεστών β_0 και β_1. Γραφική αναπαράσταση παραδείγματος μιας απλής γραμμικής παλινδρόμησης, βλέπουμε παρακάτω, στο Γράφημα 1.1.

Οι παρατηρήσεις της μεταβλητής Y θεωρούμε πως είναι τυχαίες, από διαφορετικούς πληθυσμούς τυχαίων μεταβλητών, και έχουν μέση τιμή $E(y_i)$, ίση με τη μέση τιμή του πληθυσμού από τον οποίο προέρχονται. Έτσι, για να μπορέσουμε να "δείξουμε" την απόκλιση μιας παρατήρησης y από τη μέση τιμή του πληθυσμού της, θα πρέπει να προσθέσουμε ένα τυχαίο σφάλμα, και έτσι θα έχουμε πλέον το στατιστικό μοντέλο: (Rawlings, et al., 1998)

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i} \qquad (1.2)$$

Η υπόθεση που έχουμε για το παραπάνω μοντέλο, και γενικότερα στα γραμμικά μοντέλα παλινδρόμησης, σε σχέση με τα τυχαία σφάλματα είναι ότι τα ε_i είναι ανεξάρτητα και ακολουθούν όλα την ίδια κατανομή $N(0,\sigma^2)$. Χρησιμοποιήσαμε το παράδειγμα του απλού γραμμικού μοντέλου γιατί είναι η πιο απλουστευμένη μορφή παλινδρόμησης, αλλά αποτελεί επίσης και τη βάση για το πως διαμορφώνονται και όλα τα υπόλοιπα, πιο σύνθετα μοντέλα.

Με το ίδιο σκεπτικό, σύμφωνα με τα παραπάνω, το γραμμικό μοντέλο με περισσότερες από μία επεξηγηματική μεταβλητή δίνεται από τη σχέση:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip} + \varepsilon_i$$

και με τη χρήση πινάκων, έχουμε την εξής αναπαράσταση:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}}_{\text{πίνακας X}} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

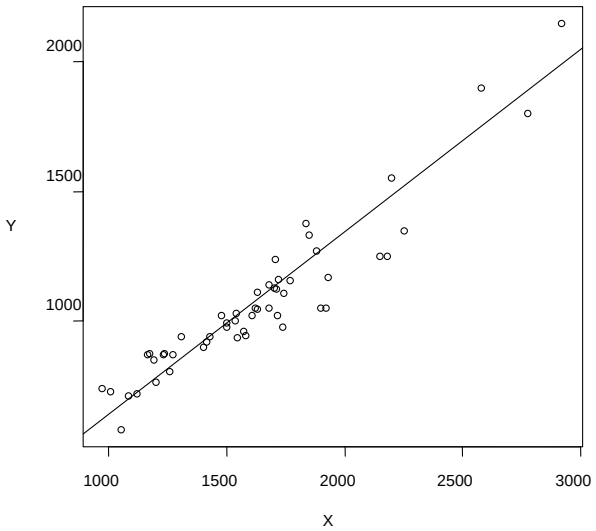
όπου:

$y = (y_1, \cdots, y_n)^T$, η στήλη των παρατηρήσεων της εξαρτημένης μεταβλητής *Y*

X, ο πίνακας του οποίου κάθε στήλη περιέχει τις τιμές των p επεξηγηματικών μεταβλητών και του οποίου η διάσταση είναι $(n \times p')$ με $p' = p + 1$, γιατί εκτός από τις *p*-στήλες των μεταβλητών, περιέχει και την πρώτη στήλη των μονάδων.

$b = (b_0, b_1, ..., b_p)^T$, η στήλη των αγνώστων παραμέτρων, διάστασης $(p' \times 1)$

$\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$, η στήλη των τυχαίων σφαλμάτων τα οποία όπως αναφέρθηκε είναι ανεξάρτητα και ακολουθούν την $N(0, \sigma^2)$.



Γράφημα 1.1: Παράδειγμα γραφήματος απλού γραμμικού μοντέλου παλινδρόμησης

1.1.3 Μοντέλα Παλινδρόμησης

Τα μοντέλα που μπορούμε να συναντήσουμε στην ανάλυση παλινδρόμησης χωρίζονται σε δύο μεγάλες κατηγορίες, τα γραμμικά μοντέλα και τα μη γραμμικά μοντέλα. Τα μη γραμμικά μοντέλα είναι συνήθως και τα πιο ρεαλιστικά ως προς τα δεδομένα, είναι δηλαδή καταλληλότερα για την πλειοψηφία των στατιστικών αναλύσεων, καθώς τις περισσότερες φορές η σχέση εξάρτησης μεταξύ των μεταβλητών δεν είναι γραμμική. Επίσης, χωρίζονται και αυτά με τη σειρά τους σε δύο κατηγορίες, σε μοντέλα που μπορούν να αναχθούν σε γραμμικά μέσω κάποιων μετασχηματισμών, και σε μοντέλα τα οποία δεν μπορούν να γραμμικοποιηθούν. (Montgomery, et al., 2006)

Εμείς θα ασχοληθούμε παρακάτω, μετά από αυτή τη μικρή εισαγωγή για το τι σημαίνει ανάλυση παλινδρόμησης, με μια συγκεκριμένη κατηγορία μοντέλων, τα οποία καλούνται Γενικευμένα Γραμμικά Μοντέλα.

Πριν κλείσουμε αυτή την ενότητα, να τονίσουμε πως η συντριπτική πλειοψηφία των στατιστικών προβλημάτων μπορούν να διατυπωθούν ως παλινδρομικά

μοντέλα. Η γενική διαδικασία που ακολουθείται αφού επιλέξουμε το πιθανό μοντέλο παλινδρόμησης για τα δεδομένα μας, είναι η ίδια είτε στη γραμμική παλινδρόμηση, είτε στα γενικευμένα γραμμικά μοντέλα με τα οποία θα ασχοληθούμε εμείς. Δηλαδή,

- καθορίζουμε τις εξισώσεις που συνδέουν την εξαρτημένη μεταβλητή με την/τις ανεξάρτητες μεταβλητές, καθώς και την κατανομή πιθανότητας της εξαρτημένης μεταβλητής,
- εκτιμούμε τις παραμέτρους που χρησιμοποιούνται στο μοντέλο ακολουθώντας μέθοδο κατάλληλη για το μοντέλο (πχ για ένα γραμμικό μοντέλο χρησιμοποιείται η μέθοδος των ελαχίστων τετραγώνων),
- ελέγχουμε πόσο καλά προσαρμόζεται το μοντέλο με τα δεδομένα μας με διάφορους ελέγχους (αναλόγως με το τι μοντέλο έχουμε κάθε φορά)
- και τέλος καταλήγουμε σε κάποια συμπεράσματα και προβλέψεις μέσω κατασκευής διαστημάτων εμπιστοσύνης ή ελέγχους υποθέσεων.

1.2 ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

1.2.1 Γενικά στοιχεία

Τα **Γενικευμένα Γραμμικά Μοντέλα** (*Generalized Linear Models- GLM*) αποτελούν μια μεγάλη και σημαντική ομάδα στατιστικών μοντέλων. Ονομάζονται έτσι, γιατί γενικεύουν τα κλασικά γραμμικά μοντέλα, που βασίζονται στην κανονική κατανομή, των οποίων τη γενική αναπαράσταση δείξαμε παραπάνω. Η γενίκευση αυτή, του γραμμικού μοντέλου, συντελείται στην κατάργηση δύο βασικών προϋποθέσεων της γραμμικής παλινδρόμησης: (Montgomery, et al., 2006)

- στο είδος της κατανομής, που ήταν απαραίτητο για τη μεταβλητή απόκρισης Y να ακολουθεί την κανονική,
- και στη σταθερότητα της διακύμανσης *(Var)* της μεταβλητής Y

Σε αντίθεση λοιπόν με τη γνωστή γραμμική παλινδρόμηση των κλασικών μοντέλων, μοναδική προϋπόθεση για τη μεταβλητή *Y*, είναι να ανήκει σε μία συγκεκριμένη οικογένεια κατανομών, την εκθετική. Στην εκθετική οικογένεια ανήκουν αρκετές κατανομές όπως είναι η κανονική, η Poisson, η διωνυμική, η γάμμα, η εκθετική, η γεωμετρική.

Όπως καταλαβαίνουμε, η πρόοδος της στατιστικής θεωρίας και των στατιστικών πακέτων με τη χρήση του υπολογιστή, βοήθησαν στο να είμαστε σε θέση να εφαρμόζουμε τις τεχνικές που ήδη είχαν αναπτυχθεί για τα γραμμικά μοντέλα και σε ένα άλλο, πολύ μεγαλύτερο φάσμα στατιστικών προβλημάτων όπου οι μεταβλητή απόκρισης δεν χρειάζεται να ακολουθεί την κανονική κατανομή, ενώ η σχέση μεταξύ των μεταβλητών, απόκρισης και επεξηγηματικών, παύει να είναι απαραιτήτως γραμμική.

Συγκεκριμένα, το “βήμα” της στατιστικής θεωρίας που βοήθησε προς αυτή την κατεύθυνση, ήταν το ότι αποδείχθηκε, πως πολλές από τις “καλές” για μια

στατιστική ανάλυση, ιδιότητες της κανονικής κατανομής, παρατηρούνται εξίσου και στην εκθετική οικογένεια κατανομών. Παράλληλα, η εξέλιξη των στατιστικών προγραμμάτων βοήθησε στον υπολογισμό και στη χρήση αριθμητικών μεθόδων για την εκτίμηση των άγνωστων παραμέτρων β_i στην περίπτωση που η σχέση που συνδέει το μέσο $\mathrm{E}(Y_i)$, με το γραμμικό μέρος των ανεξάρτητων μεταβλητών $x_i^T\beta$, δεν είναι μια γραμμική συνάρτηση όπως στην περίπτωση του γραμμικού μοντέλου (1.1), αλλά μια μη γραμμική συνάρτηση, η οποία ονομάζεται συνάρτηση σύνδεσης *(link function)* και θα την αναλύσουμε παρακάτω. (McCullagh & Nelder, 1989)

Μέσω της τεχνικής των επαναληπτικών σταθμισμένων τετραγώνων, μπορούμε να υπολογίσουμε τις εκτιμήσεις της μέγιστης πιθανοφάνειας των παραμέτρων, με παρατηρήσεις που ακολουθούν κατανομή από την εκθετική οικογένεια, όπως σε αντιστοιχία, κάνουμε στη γραμμική παλινδρόμηση μέσω των ελαχίστων τετραγώνων. Στα γενικευμένα γραμμικά μοντέλα, συναντάμε ουσιαστικά μια γενίκευση της ανάλυσης διασποράς χρησιμοποιώντας λογαριθμικές πιθανοφάνειες *(log-likelihoods)*.

1.2.2 Εκθετική οικογένεια κατανομών

Πριν συνεχίσουμε με την ανάλυση της δομής των γενικευμένων γραμμικών μοντέλων, είναι σημαντικό να αναφέρουμε κάποια βασικά πράγματα για την εκθετική οικογένεια κατανομών.

Ας υποθέσουμε πως έχουμε μία μεταβλητή Y, της οποίας η κατανομή πιθανότητας εξαρτάται μόνο από μια παράμετρο θ. Η κατανομή αυτή, ανήκει στην εκθετική οικογένεια κατανομών αν μπορεί να γραφεί στη μορφή: (Dobson, 2002)

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \tag{1.3}$$

όπου a, b, s, t, θεωρούνται γνωστές συναρτήσεις.

Η παραπάνω εξίσωση, μπορεί να γραφεί και με την εξής μορφή:

$$\boxed{f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]} \tag{1.4}$$

όπου $\mathrm{s(y)} = \exp[\mathrm{d(y)}]$ και $\mathrm{t}(\theta) = \exp[\mathrm{c}(\theta)]$ από την (1.3).

Στην περίπτωση που $a(y) = y$ τότε η κατανομή λέμε ότι βρίσκεται στην **«κανονική»** της μορφή, δηλαδή:

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \tag{1.5}$$

Η $b(\theta)$ καλείται συχνά *"φυσική"* παράμετρος της κατανομής. Αν υπάρχουν παραπάνω παράμετροι, εκτός της παραμέτρου θ που μας ενδιαφέρει, αυτές ονομάζονται παράμετροι *"κλίμακας"*, εμφανίζονται ως μέρη των συναρτήσεων a, b, c, d και σε αυτήν την περίπτωση θα αντιμετωπίζονται σα να είναι γνωστές.

Παράδειγμα κατανομής που έχει και άλλη παράμετρο, είναι η κανονική που εκτός του μ_i έχει και τη διασπορά σ^2 οπότε η συνάρτηση της έχει τη μορφή:

$$f(y_i;\mu_i;\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(y_i-\mu_i)^2}{2\sigma^2}}$$

όπου μ είναι η παράμετρος που μας ενδιαφέρει, και σ^2 είναι παράμετρος κλίμακας.

Όπως αναφέραμε και παραπάνω, στην εκθετική οικογένεια κατανομών ανήκουν πολλές γνωστές κατανομές, (κανονική, Poisson, διωνυμική, κλπ), και όλες αυτές μπορούν να γραφούν σύμφωνα με τον τύπο της κανονικής μορφής (1.5). Θα παραθέσουμε μερικά παραδείγματα τέτοιων κατανομών, για το πως γράφονται στην κανονική μορφή, τη μορφή έχουν δηλαδή οι συναρτήσεις $b(\theta)$, $c(\theta)$ και $d(\theta)$ για την κάθε περίπτωση:

ΚΑΤΑΝΟΜΗ	$b(\theta)$	$c(\theta)$	$d(\theta)$
Poisson	$\log(\theta)$	$-\theta$	$-\log y!$
Κανονική	$\frac{\mu}{\sigma^2}$	$\frac{-\mu}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)$	$-\frac{y^2}{2\sigma^2}$
Διωνυμική	$\log\left(\frac{\pi}{1-\pi}\right)$	$n\log(1-\pi)$	$\log\binom{n}{y}$

Πίνακας 2.1: Περιγραφή τριών κατανομών ως μέλη της εκθετικής οικογένειας

Πριν κλείσουμε αυτή τη μικρή ενότητα περί εκθετικής οικογένειας, να αναφέρουμε και μερικές σημαντικές ιδιότητες της συγκεκριμένης οικογένειας κατανομών: (Dobson, 2002)

- Για τη συνάρτηση $a(Y)$, χρειαζόμαστε τις αναμενόμενες τιμές και τη διακύμανση, χρειαζόμαστε δηλαδή, τις εκφράσεις των $E[a(y)]$ και $Var[a(y)]$.

Εξ'ορισμού για τη συνάρτηση πυκνότητας πιθανότητας έχουμε ότι:

$$\int f(y;\theta)dy=1 \Rightarrow$$

$$\frac{d}{d\theta}\int f(y;\theta)dy=\frac{d}{d\theta}1=0 \Rightarrow \int\frac{df(y;\theta)}{d\theta}dy=0 \Rightarrow \int\frac{d^2f(y;\theta)}{d\theta^2}dy=0 \qquad (1.6)$$

Με βάση τις σχέσεις αυτές, για τη συνάρτηση εκθετικής κατανομής, θα έχουμε αντίστοιχα:

$$\frac{df(y;\theta)}{d\theta}=[\alpha(y)b'(\theta)+c'(\theta)]f(y;\theta) \Rightarrow \int[a(y)b'(\theta)+c'(\theta)]f(y;\theta)dy=0 \qquad (1.7)$$

και αν λάβουμε υπόψη ότι:

$$\int a(y)f(y;\theta)dy=E[a(y)] \quad \text{και} \quad \int c'(\theta)f(y;\theta)dy=c'(\theta)$$

τότε η τελευταία σχέση (1.6) μπορεί να γραφεί απλοποιημένα ως:

$$b'(\theta)\mathrm{E}[a(y)]+c'(\theta)=0 \Rightarrow \boxed{E[a(\mathrm{Y})]=-c'(\theta)/b'(\theta)} \quad (1.8)$$

η σχέση (1.8) αποτελεί την έκφραση της μέσης τιμής της συνάρτησης $a(y)$.

Αντίστοιχα για την $Var[a(y)]$:

Με βάση και πάλι τις πρώτες σχέσεις, και συγκεκριμένα τη (1.6), και ότι $\int\{a(y)-E[a(Y)]\}^2 f(y;\theta)dy=\mathrm{var}[a(Y)]$, τότε θα έχουμε ότι:

$$\int\frac{d^2 f(y;\theta)}{d\theta^2}dy = b''(\theta)\mathrm{E}[\alpha(\mathrm{Y})]+c''(\theta)+[b'(\theta)]^2\,\mathrm{var}[a(Y)]=0$$

Τότε έχοντας και το αποτέλεσμα για τη μέση τιμή από τη σχέση (1.8), η τελευταία γίνεται:

$$\boxed{\mathrm{var}[a(Y)]=\frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{[b'(\theta)]^3}} \quad (1.9)$$

η σχέση (1.9) αποτελεί την έκφραση της διασποράς της συνάρτησης $a(y)$.

- Μας ενδιαφέρουν επίσης οι εκφράσεις της μέσης τιμής και της διασποράς της παραγώγου της λογαριθμικής συνάρτησης πιθανοφάνειας, δηλαδή οι εκφράσεις των $\mathrm{E}(U)$ και $Var[U]$. Οι σχέσεις αυτές είναι σημαντικές, γιατί η συνάρτηση U, χρησιμοποιείται στην εξαγωγή συμπερασμάτων σχετικά με τις τιμές των παραμέτρων των γενικευμένων γραμμικών μοντέλων, είτε μέσω κατασκευής διαστημάτων εμπιστοσύνης, είτε μέσω ελέγχων υποθέσεων.

Η λογαριθμική συνάρτηση πιθανοφάνειας για την εκθετική κατανομή είναι:

$$l(\theta;y)=a(y)b(\theta)+c(\theta)+d(y)$$

Ενώ η παράγωγος αυτής ως προς θ θα είναι:

$$U(\theta;y)=\frac{dl(\theta;y)}{d\theta}=a(y)b'(\theta)+c'(\theta)\Rightarrow \quad (1.10)$$

$$E(U)=b'(\theta)E[a(Y)]+c'(\theta)$$

Όμως από τη σχέση (1.8) θα έχουμε ότι:

$$E(U)=b'(\theta)[-\frac{c'(\theta)}{b'(\theta)}]+c'(\theta) \Rightarrow \boxed{E(U)=0} \quad (1.11)$$

Η σχέση (1.11) αποτελεί την έκφραση της μέσης τιμής της λογαριθμικής συνάρτησης πιθανοφάνειας για μία εκθετική κατανομή.

Για τη διασπορά, λαμβάνοντας υπόψη τη σχέση (1.10) θα έχουμε:

$$\mathrm{var}(U)=[b'(\theta)^2]\,\mathrm{var}[a(Y)]$$

Όμως από τη σχέση (1.9) θα έχουμε:

$$\boxed{\text{var}(U) = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta)} \qquad (1.12)$$

Η σχέση (1.12) αποτελεί αντίστοιχα την έκφραση της διασποράς για τη λογαριθμική συνάρτηση πιθανοφάνειας.

1.2.3 Δομή Γενικευμένων Γραμμικών Μοντέλων

Τα γενικευμένα γραμμικά μοντέλα, μπορούν να αναλυθούν σε **τρία βασικά δομικά μέρη** από τα οποία αποτελούνται: (Lindsey, 1997)

Α] Τη μεταβλητή απόκρισης Y, όπου οι παρατηρήσεις y_i είναι τυχαίες και ανεξάρτητες. Οι y_i $(i = 1,...,n)$, με μέσους μ_i, ακολουθούν κατανομή από την εκθετική οικογένεια κατανομών με σταθερή παράμετρο κλίμακας.

Με βάση λοιπόν τα όσα αναφέρθηκαν για την εκθετική οικογένεια κατανομών, η μεταβλητή απόκρισης Y, ενός γενικευμένου γραμμικού μοντέλου θα ακολουθεί κατανομή που έχει τη μορφή:

$$f(y;\theta) = \exp[yb(\theta) + c(\theta) + d(y)]$$

Πολύ συχνά η παραπάνω μορφή, πιο γενικευμένα, γράφεται και ως:

$$\boxed{f(y;\theta;\varphi) = \exp\left[\frac{y\theta - b(\theta)}{a(\varphi)} + c(y;\varphi)\right]} \qquad (1.13)$$

όπου αντίστοιχα η θ είναι η φυσική παράμετρος, η φ η παράμετρος κλίμακας και οι συναρτήσεις a, b, c είναι γνωστές. Όπως παρατηρούμε στη συγκεκριμένη μορφή (1.13), υπάρχει και η επιπλέον παράμετρος φ, εκτός της φυσικής, που όπως αναφέραμε παραπάνω, αν θεωρηθεί γνωστή, έχουμε τη γνωστή γραμμική εκθετική οικογένεια σε κανονική μορφή, στην οποία μπορούν να γραφούν όλες οι κατανομές που ανήκουν στην εκθετική οικογένεια. (Οικονόμου & Καρώνη , 2010)

Με τα αντίστοιχα επιχειρήματα όπως στην προηγούμενη ενότητα, μπορούμε να δείξουμε πως οι σχέσεις της μέσης τιμής και της διασποράς με βάση τη μορφή (1.13) έχουν ως εξής:

$$\boxed{E[y] = b'(\theta)} \quad \text{και} \quad \boxed{Var[y] = a(\varphi)b''(\theta)}$$

και αντίστοιχα για την παράγωγο της πιθανοφάνειας:

$$\boxed{E(U) = 0} \quad \text{και} \quad \boxed{Var[U] = \frac{b''(\theta)}{a(\varphi)}}$$

Η μορφή (1.13), και οι σχέσεις που προκύπτουν από αυτήν, είναι ισοδύναμες με τις σχέσεις που αποδείξαμε αναλυτικά στην ενότητα 2.1 περί εκθετικής οικογένειας, και χρησιμοποιούνται αμφότερες.

Β] Το σύνολο των **επεξηγηματικών** μεταβλητών $X_{n\times p}=[x_1^T,...,x_n^T]^T$ και των **άγνωστων παραμέτρων** $\beta^T=[\beta_1,...,\beta_p]$ με $p<n$, οι οποίες αποτελούν το γραμμικό μέρος του μοντέλου, σχηματίζοντας τη λεγόμενη γραμμική προβλέπουσα (linear predictor) η:

$$\eta = \mathrm{X}\beta \tag{1.14}$$

Γ] Τη συνάρτηση σύνδεσης (link function). Όπως παρατηρούμε από τη σχέση (1.14) και όπως είχαμε αναφέρει και παραπάνω, η σχέση που συνδέει το μέσο $\mathrm{E}(Y_i)$ με το γραμμικό μέρος των ανεξάρτητων μεταβλητών $x_i^T\beta$, δεν είναι απαραίτητα μια γραμμική συνάρτηση όπως έχουμε στη γραμμική παλινδρόμηση:

$$E[y_i]=\mu_i=x_i^T\beta$$

Από τη στιγμή που η μεταβλητή απόκρισης μπορεί να ακολουθεί κάποια κατανομή που να μην εξασφαλίζει την παραπάνω σχέση, θα πρέπει να εισάγουμε μια συνάρτηση που θα μπορεί και να εξασφαλίζει την ισότητα, και αυτομάτως να επιτυγχάνεται και η γραμμικότητα. Για παράδειγμα, αν η μεταβλητή Y ακολουθεί την κατανομή Poisson τότε θα πρέπει $E[y_i]=\mu_i>0$, και άρα δε θα μπορούμε να έχουμε ότι $\mu_i=x_i^T\beta$, αφού στο δεξί μέλος της ισότητας δεν υπάρχει ο ίδιος περιορισμός.

Το παραπάνω "εμπόδιο", ξεπερνάμε με τη χρήση της συνάρτησης σύνδεσης g_i:

$$\boxed{\eta_i=g_i(\mu_i)=x_i^T\beta} \tag{1.15}$$

Παρατηρώντας τις σχέσεις (1.14) και (1.15), καταλαβαίνουμε ότι η συνάρτηση g_i συνδέει γραμμικά πλέον, τη μέση τιμή της μεταβλητής απόκρισης με την $x_i^T\beta$, κάτι που δε θα μπορούσε να επιτευχθεί διαφορετικά.

Η συνάρτηση σύνδεσης g_i πρέπει να είναι μία μονότονη και διαφορίσιμη συνάρτηση. Η πιο συνηθισμένη επιλογή συνάρτησης σύνδεσης είναι η ίδια συνάρτηση που χρησιμοποιείται για τη "μετατροπή" του μέσου κάθε κατανομής που ανήκει στην εκθετική οικογένεια, ώστε να τη φέρουμε στην κανονική της μορφή, δηλαδή είναι η ίδια με τη φυσική παράμετρο θ της σχέσης (1.13). Αν ισχύει λοιπόν ότι $\eta=g(\mu)=\theta$, τότε η συνάρτηση σύνδεσης λέγεται ***"κανονική"***.

Κατανομή	**Κανονική συνάρτηση σύνδεσης**	
Poisson	log	$\eta_i = g(\mu_i) = \log(\mu_i)$
Διωνυμική	logit	$\eta_i = g(\mu_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$
Κανονική	identity (ταυτοτική)	$\eta_i = \mu_i$

Πίνακας 2.2: Παραδείγματα συνάρτησης σύνδεσης για τις κατανομές Poisson, Διωνυμική και Κανονική

Αν λοιπόν η μεταβλητή απόκρισης Y ακολουθεί κατανομή από την εκθετική οικογένεια, και η μεταβλητή κλίμακας φ θεωρείται γνωστή, τότε με την επιλογή μιας κανονικής συνάρτησης σύνδεσης g_i είμαστε σε θέση να "γραμμικοποιήσουμε" το μοντέλο μας και όλες οι άγνωστες παράμετροι του γραμμικού μέρους να έχουν επαρκή στατιστικά στοιχεία ώστε να μπορούν να υπολογιστούν μέσω της παλινδρόμησης.

Όπως καταλαβαίνουμε, η συνάρτηση σύνδεσης "κατασκευάζεται" με σκοπό να απλοποιήσει τις αριθμητικές μεθόδους για την εκτίμηση των άγνωστων παραμέτρων μοντέλων που περιέχουν γραμμικά μέρη, άρα για αυστηρώς μη γραμμικά παλινδρομικά μοντέλα χάνει τη χρησιμότητα της.

1.3 ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

1.3.1 Γενικά στοιχεία

Η λογιστική παλινδρόμηση αποτελεί μία από τις πιο σημαντικές τεχνικές που βασίζεται στα γενικευμένα γραμμικά μοντέλα και θα μας απασχολήσει παρακάτω, γι'αυτό και είναι χρήσιμο να αναλυθούν κάποια βασικά στοιχεία της.

Πολύ συχνά, υπάρχουν περιπτώσεις όπου η μεταβλητή απόκρισης είναι κατηγορική και έχει μόνο δύο πιθανά αποτελέσματα, τα οποία συχνά καλούνται αποτυχία και επιτυχία και συμβολίζονται με 0 και 1 αντίστοιχα. Παραδείγματα τέτοιων περιπτώσεων αποτελούν η μελέτη για την επιτυχημένη ή όχι θεραπεία των ασθενών, η επιτυχία ή όχι των φοιτητών και μαθητών σε εξετάσεις, ή η μελέτη της καλής λειτουργίας ή όχι των μηχανημάτων, κλπ.

Όπως καταλαβαίνουμε, δεν μπορεί να χρησιμοποιηθεί η γραμμική παλινδρόμηση και το μοντέλο της σχέσης (1.2), γιατί η επιλογή μιας γραμμικής συνάρτησης για τη μεταβλητή απόκρισης Y, θα είχε σαν αποτέλεσμα να έχουμε προβλέψεις, μικρότερες του 0 ή μεγαλύτερες του 1, πράγμα άτοπο στην περίπτωση μας. Θα πρέπει λοιπόν να επιλέξουμε κατάλληλη κατανομή και κατάλληλο μοντέλο ώστε να έχει νόημα για το πρόβλημα μας.

Η μεταβλητή απόκρισης y ακολουθεί τη διωνυμική κατανομή, μέλος της εκθετικής οικογένειας κατανομών, που μπορεί να περιγράψει μεταβλητές τέτοιου είδους: (Οικονόμου & Καρώνη , 2010)

$$y \sim b(n, p) \text{ τότε } \quad f(y) = \binom{n}{p} p^y (1-p)^{n-y} \ , \ y = 0,1,\ldots,n$$

$$\text{με} \qquad E(y) = np \quad \text{και} \quad V(y) = np(1-p)$$

όπου p η παράμετρος της κατανομής που συμβολίζει την πιθανότητα επιτυχίας. (στην ειδική περίπτωση των **δυαδικών δεδομένων** τότε έχουμε n=1, που είναι και η περίπτωση που μας ενδιαφέρει εμάς).

Η κατασκευή του μοντέλου παλινδρόμησης που συνδέει τη μεταβλητή απόκρισης με τις επεξηγηματικές μεταβλητές βασίζεται στη δομή του γενικευμένου γραμμικού μοντέλου που αναλύσαμε στην προηγούμενη ενότητα και άρα θα έχει τη μορφή:

$$\eta_i = g(E(y_i)) = g(\mu_i) = x_i^T \beta$$

όπου όπως είδαμε παραπάνω, για τη διωνυμική κατανομή ως συνάρτηση σύνδεσης επιλέγουμε την *logit*:

$$\eta_i = g(\mu_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$$

Και έτσι, προκύπτει ολοκληρωμένο το μοντέλο της λογιστικής παλινδρόμησης που δίνεται από τη σχέση:

$$\boxed{\eta_i = \ln\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta}$$

Αν αντιστρέψουμε τη συνάρτηση σύνδεσης, τότε θα έχουμε ότι:

$$p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

με p_i όπως αναφέραμε να είναι η πιθανότητα επιτυχίας. Είναι προφανές ότι έχουμε πετύχει το ζητούμενο, δηλαδή την απαραίτητη προϋπόθεση, η πιθανότητα να περιορίζεται μεταξύ του 0 και του 1.

Έτσι, η τελευταία σχέση γίνεται:

$$p_i = p_{x_i} = \frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)}$$

Η τελευταία σχέση, στην περίπτωση που έχουμε κ διαφορετικές συμμεταβλητές τότε προφανώς μπορεί να γραφεί στην εξής αθροιστική μορφή:

$$p_i = p_{x_i} = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_\kappa x_{i\kappa})}{1+\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_\kappa x_{i\kappa})}$$

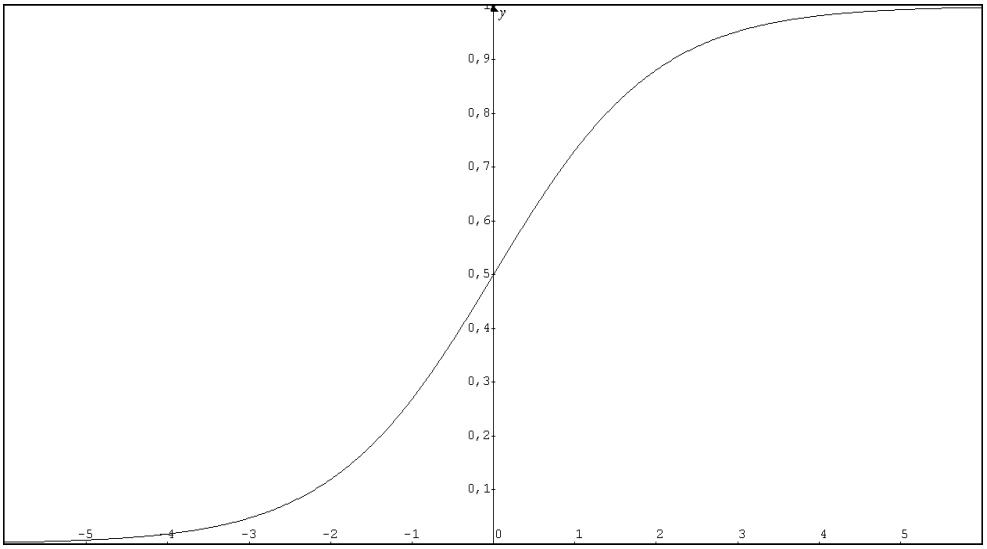
Επίσης, στην περίπτωση δυαδικών δεδομένων (0 ή 1) έχουμε ότι:

$$E(y_i) = 1(p_i) + 0(1 - p_i) = p_i$$

Άρα για τις παρατηρήσεις y_i που αποτελούν τυχαίες και ανεξάρτητες μεταβλητές της διωνυμικής κατανομής, θα έχουμε τις εξής αναμενόμενες τιμές που δίνονται από τη συνάρτηση της μέσης τιμής:

$$E(y_i) = p_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} = \frac{1}{1 + \exp(-x_i^T \beta)}$$

Τα παραπάνω, αποτελούν την περιγραφή του μοντέλου της λογιστικής παλινδρόμησης και θα χρησιμοποιηθούν αργότερα, στην ανάλυση του συγκεκριμένου μοντέλου, στην περίπτωση δυαδικών χρονοσειρών.



Γράφημα: 3.1 Λογιστική συνάρτηση

1.3.2 Παράδειγμα μοντέλου λογιστικής παλινδρόμησης

Θα παρουσιάσουμε ένα απλό παράδειγμα χρήσης της λογιστικής παλινδρόμησης, ώστε να έχουμε μια εικόνα για το πώς χρησιμοποιείται το συγκεκριμένο είδος παλινδρόμησης.

Θα χρησιμοποιήσουμε ένα μικρό αριθμό δεδομένων, τα οποία βλέπουμε στον παρακάτω πίνακα, και αφορούν δείγμα 20 τραπεζών και την χρηματοοικονομική τους κατάσταση, έτσι όπως έχουν κριθεί από ειδικούς. Οι δύο τελευταίες στήλες αποτελούν δύο πολύ συνηθισμένους δείκτες/ποσοστά που χρησιμοποιούνται για τη χρηματοοικονομική ανάλυση των τραπεζών, το δείκτη των συνολικών δανείων και χρηματοδοτικών μισθώσεων προς το σύνολο ενεργητικού και το δείκτη των συνολικών εξόδων προς το σύνολο του ενεργητικού αντίστοιχα, κάθε τράπεζας.

Παρατήρηση	Χρηματ/νομική κατάσταση	Σύνολο δανείων / Σύνολο ενεργητικού	Σύνολο εξόδων/ Σύνολο ενεργητικού
1	1	0.64	0.13
2	1	1.04	0.10
3	1	0.66	0.11
4	1	0.80	0.09
5	1	0.69	0.11
6	1	0.74	0.09
7	1	0.63	0.11
8	1	0.75	0.14
9	1	0.56	0.12
10	1	0.65	0.12
11	0	0.55	0.16
12	0	0.46	0.12
13	0	0.72	0.10
14	0	0.43	0.08
15	0	0.52	0.08
16	0	0.54	0.08
17	0	0.30	0.07
18	0	0.67	0.08
19	0	0.51	0.09
20	0	0.79	0.13

Πίνακας 3.1: Οικονομική κατάσταση 20 τραπεζών.
Πηγή δεδομένων: Πανεπιστήμιο MIT

Έχουμε λοιπόν:

- Μεταβλητή απόκρισης y, την χρηματοοικονομική κατάσταση της τράπεζας που παίρνει τις τιμές:

  0, αν η χρημ/κή κατάσταση είναι κακή
  1, αν η χρημ/κή κατάσταση είναι καλή

- Επεξηγηματική μεταβλητή x_1, το δείκτη των συνολικών δανείων και μισθώσεων
- Επεξηγηματική μεταβλητή x_2, το δείκτη των συνολικών εξόδων

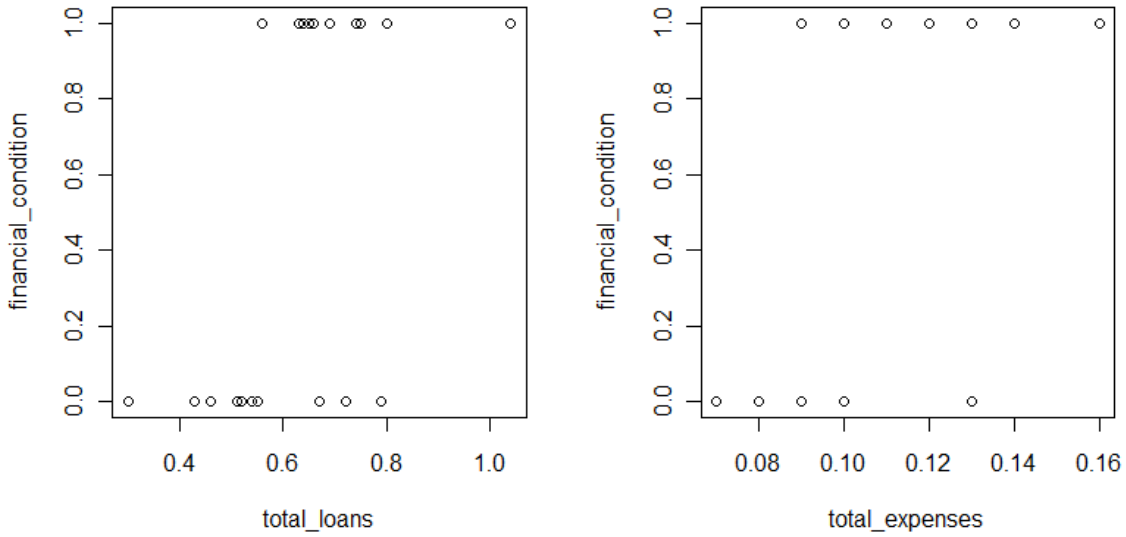
Μας ενδιαφέρει να ελέγξουμε αν και κατά πόσο, η κάθε επεξηγηματική μεταβλητή επηρεάζει τη χρηματοοικονομική κατάσταση μιας τράπεζας και να επιλέξουμε το κατάλληλο μοντέλο λογιστικής παλινδρόμησης που θα περιγράφει τελικά τη μεταξύ τους σχέση. Για την επεξεργασία των δεδομένων μας, θα χρησιμοποιήσουμε το πρόγραμμα της R.

Εισάγουμε τα δεδομένα μας στην R, και φτιάχνουμε με αυτά ένα πλαίσιο, στο οποίο έχουμε τις 3 στήλες που μας ενδιαφέρουν, τη μεταβλητή απόκρισης "financial_condition", την πρώτη επεξηγηματική μεταβλητή "total_loans" και τη δεύτερη "total_expenses".

Εισάγοντας τις εντολές:

> plot(total_loans,financial_condition)
> plot(total_expenses,financial_condition)

έχουμε μια πρώτη γραφική αναπαράσταση των μεταβλητών μας:



Αρχικά θα κατασκευάσουμε το μοντέλο λογιστικής παλινδρόμησης που περιέχει και τις δύο επεξηγηματικές μεταβλητές, ώστε να ελέγξουμε τη σημαντικότητα των δύο μεταβλητών και συνεπώς την καταλληλότητα του μοντέλου.

Εισάγοντας τις εντολές:

```
> mylogit<- glm(financial_condition~total_loans, family=binomial(link="logit"),
na.action=na.pass)
> summary(mylogit)
```

έχουμε τα αποτελέσματα για την προσαρμογή του μοντέλου λογιστικής παλινδρόμησης με συνάρτηση σύνδεσης logit:

```
Call:
glm(formula = financial_condition ~ total_loans + total_expenses,
family=binomial(link = "logit"), na.action = na.pass)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.64035  -0.35514   0.02079   0.53234   1.03373

Coefficients:
                Estimate   Std. Error   z value      Pr(>|z|)
(Intercept)     -14.188      6.122       -2.317      0.0205 *
total_loans       9.173       6.864       1.336      0.1814
total_expenses   79.964   39.263          2.037      0.0417 *
```

```
(Dispersion parameter for binomial family taken to be 1)
Null deviance:      27.726  on 19  degrees of freedom
Residual deviance: 12.831   on 17  degrees of freedom
AIC: 18.831
Number of Fisher Scoring iterations: 6
```

Από τα παραπάνω αποτελέσματα, βλέπουμε ότι η επεξηγηματική μεταβλητή "total_loans" , δηλαδή η μεταβλητή του δείκτη των συνολικών δανείων της τράπεζας, δεν είναι στατιστικά σημαντική. Αντίθετα η μεταβλητή "total_expenses" φαίνεται να συμβάλει στο μοντέλο και να καθορίζει τη χρηματοοικονομική κατάσταση της τράπεζας. Τα συμπεράσματα αυτά προέκυψαν από τις p-τιμές των μεταβλητών, αφού μόνο για τη μεταβλητή του δείκτη των συνολικών εξόδων είναι 0.0417, πράγμα που σημαίνει ότι επηρεάζει τη μεταβλητή απόκρισης (έλεγχος Wald).

Στο τέλος του παραπάνω πίνακα των αποτελεσμάτων της παλινδρόμησης βλέπουμε την τιμή του κριτηρίου AIC. Το κριτήριο AIC αποτελεί κριτήριο επιλογής του βέλτιστου μοντέλου παλινδρόμησης στη γενική περίπτωση των γενικευμένων γραμμικών, και χρησιμοποιείται και στην περίπτωση της λογιστικής παλινδρόμησης. Όσο μικρότερη είναι η τιμή του δείκτη AIC, τόσο καταλληλότερο είναι και το μοντέλο μας. Στο συγκεκριμένο μοντέλο, η τιμή του AIC είναι 18.83.

Με την παρακάτω εντολή, μπορούμε να δούμε τις προβλεπόμενες τιμές για κάθε μία παρατήρηση χωριστά, με βάση το μοντέλο παλινδρόμησης που εφαρμόσαμε:

```
>fitted(mylogit)
        1         2         3         4         5         6         7
0.88880751 0.96607160 0.65988984 0.58607781 0.71869544 0.97802390 0.76624821
        8         9        10        11        12        13        14
0.90788100 0.97688030 0.79749494 0.24124065 0.02736463 0.23402492 0.02091905
       15        16        17        18        19        20
0.02145720 0.05536240 0.01421931 0.07987680 0.09009646 0.96936803
```

Όπως αναφέρθηκε και παραπάνω, για την εξαγωγή συμπερασμάτων από μια στατιστική ανάλυση, χρησιμοποιούμε τα αποτελέσματα των διαστημάτων εμπιστοσύνης (συνήθως 95%). Έτσι, για να κατασκευάσουμε διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου μας, αλλά και τα αντίστοιχα διαστήματα των $\exp(\beta_j)$, χρησιμοποιούμε τις παρακάτω εντολές που μας δίνουν και τα αντίστοιχα αποτελέσματα:

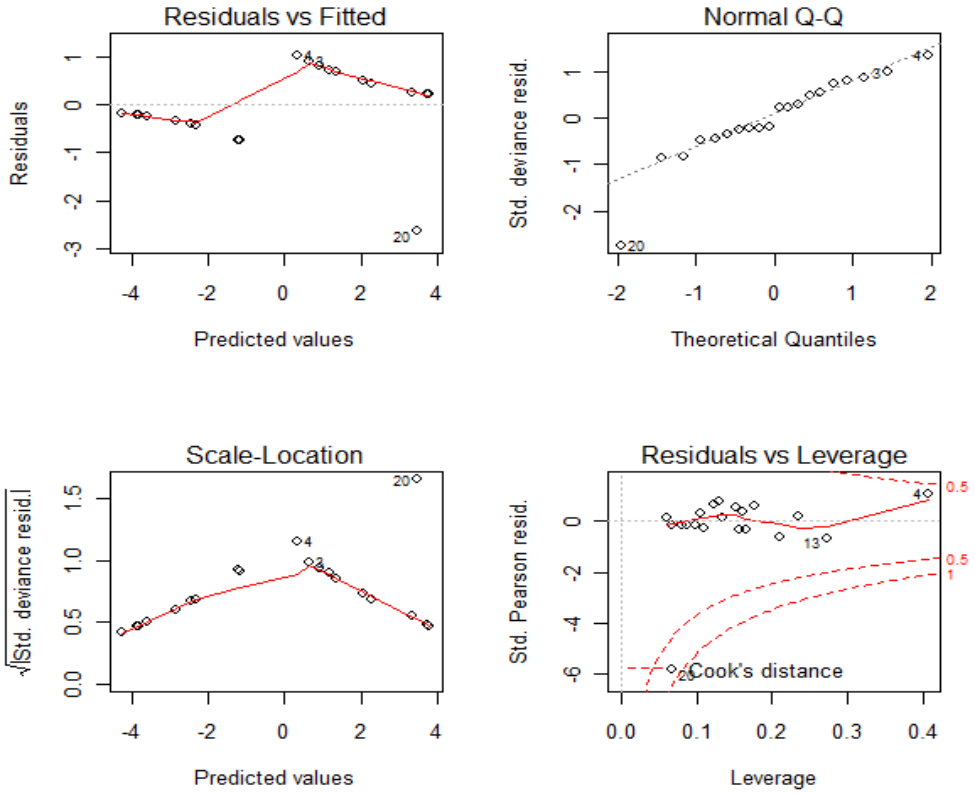
```
> confint(mylogit)

                    2.5 %       97.5 %
(Intercept)     -32.2230470  -5.38442
total_loans      -0.6689217   28.49168
total_expenses   19.4962744  181.70156
```

> exp(confint(mylogit))

	2.5 %	97.5 %
(Intercept)	1.013231e-14	4.587500e-03
total_loans	5.122607e-01	2.364716e+12
total_expenses	2.931733e+08	8.165562e+78

Τέλος, για τη γραφική αναπαράσταση των αποτελεσμάτων του μοντέλου μας, εισάγουμε την εντολή:

> plot(mylogit), και έτσι έχουμε:



Όπως παρατηρήσαμε στα αποτελέσματα που προέκυψαν για το μοντέλο λογιστικής παλινδρόμησης που εφαρμόσαμε μέσω της εντολής >*summary (mylogit)*, η μεταβλητή "total_loans" δε συμβάλει στο μοντέλο. Με βάση αυτό το αποτέλεσμα θα κατασκευάσουμε και ένα δεύτερο μοντέλο παλινδρόμησης, αυτή τη φορά μόνο με την επεξηγηματική μεταβλητή "total_expenses", για να δούμε τις διαφορές καθώς και το πιο από τα δύο είναι καταλληλότερο.

Με τις αντίστοιχες εντολές που χρησιμοποιήσαμε και πριν, έχουμε το δεύτερο μοντέλο με τα παρακάτω αποτελέσματα:

```
Call:
glm(formula = financial_condition ~ total_expenses, family = binomial(link= "logit"),
na.action = na.pass)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.3428  -0.4946  -0.1122   0.5701   1.6638

Coefficients:
                  Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)        -9.587     3.944     -2.431     0.0151 *
total_expenses     94.345     38.890    2.426      0.0153 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

Null deviance:         27.726  on 19  degrees of freedom
Residual deviance: 16.036  on 18  degrees of freedom
AIC: 20.036
Number of Fisher Scoring iterations: 5
```

Παρατηρούμε ότι η μεταβλητή "total_expenses", είναι στατιστικά σημαντική με το μοντέλο μας, με p-value μικρότερη από ότι στο προηγούμενο (0.0153<0.0417), ενώ ο δείκτης AIC που αναφέραμε παραπάνω, είναι κατά λίγο μεγαλύτερος σε σχέση με του άλλου μοντέλου (20.036>18.831).

Για τις προβλεπόμενες τιμές έχουμε κι εδώ αντίστοιχα τις παρακάτω:

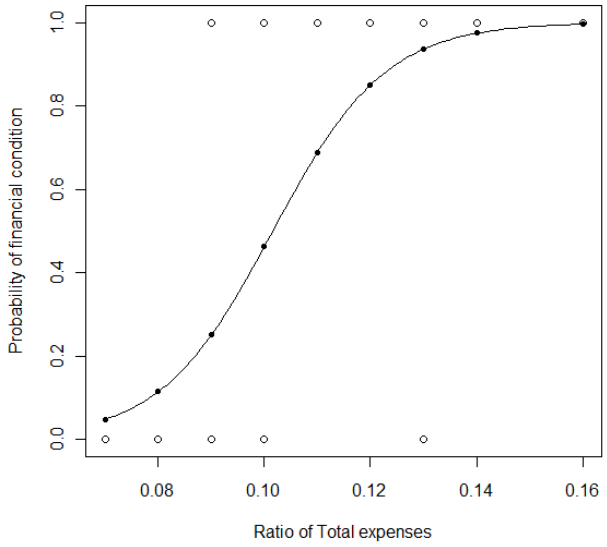
```
> fitted(mylogit)
     1          2          3          4          5          6          7          8
0.9357162  0.4619842  0.6880668  0.2505256  0.6880668  0.9739529  0.8499936  0.8499936
     9         10         11         12         13         14         15         16
0.9959636  0.8499936  0.4619842  0.1151416  0.1151416  0.1151416  0.0482127  0.1151416
    17         18         19         20
0.2505256  0.0482127  0.2505256  0.9357162
```

Σε αυτό το μοντέλο, θα κατασκευάσουμε το γράφημα των δεδομένων μας, με την εκτιμώμενη λογιστική καμπύλη που προκύπτει από την παλινδρόμηση. Έχει ενδιαφέρον μέσω του συγκεκριμένου γραφήματος, να δούμε και το πως κατανέμονται οι παρατηρήσεις μας πάνω σε αυτήν, προσαρμοσμένες με βάση το γενικευμένο μοντέλο που χρησιμοποιήσαμε (έχοντας πάντα υπόψη ότι με 0 συμβολίζουμε την «κακή» χρηματοοικονομική κατάσταση και με 1 την «καλή»).

Εισάγοντας τις εντολές:

```
> plot(total_loans,financial_condition)
> curve(predict(mylogit,data.frame(total_loans=x),type="resp"),add=TRUE)
> points(total_loans,fitted(mylogit),pch=20)
```

έχουμε την παρακάτω λογιστική καμπύλη:



Στη λογιστική παλινδρόμηση υπάρχουν πολλοί τρόποι για να ελέγξουμε την προσαρμογή ενός μοντέλου ή την καταλληλόλητα του σε σχέση με κάποιο άλλο, όπως είναι η ελεγχοσυνάρτηση deviance ή ο έλεγχος του Pearson. Δείκτη καταλληλότητας αποτελεί και αυτό του AIC που αναφέραμε παραπάνω. Στη δική μας όμως περίπτωση, που μας αφορά συγκεκριμένα η δυαδική λογιστική παλινδρόμηση, έχουμε δηλαδή μεταβλητή απόκρισης που παίρνει τιμές 0 ή 1, η ελεγχοσυνάρτηση deviance και ο έλεγχος Pearson, δεν αποτελούν κατάλληλους ελέγχους για τα μοντέλα μας και δεν μπορούν να χρησιμοποιηθούν. Ένας κατάλληλος έλεγχος για να δούμε την καταλληλότητα των μοντέλων μας, είναι αυτός του Hosmer-Lemeshow.

Πριν τον εφαρμόσουμε, να αναφέρουμε απλώς ότι για τον έλεγχο Hosmer-Lemeshow, οι παρατηρήσεις μας ταξινομούνται σε αύξουσα σειρά με βάση την εκτιμώμενη πιθανότητα τους και χωρίζονται σε ομάδες με ίσο περίπου αριθμό παρατηρήσεων στην κάθε μία. Η ελεγχοσυνάρτηση Hosmer-Lemeshow, όπως και αυτή του Pearson, είναι ουσιαστικά ένας έλεγχος χ^2 που δίνεται από τον παρακάτω τύπο:

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{(O_i - N_i \bar{\pi}_i)^2}{N_i \bar{\pi}_i (1-\bar{\pi}_i)}$$

Όπου: g , ο αριθμός των ομάδων

N_i, ο αριθμός των παρατηρήσεων της ομάδας i

O_i, ο αριθμός των επιτυχιών της i ομάδας, δηλ το άθροισμα των y_i παρατηρήσεων της ομάδας i.

$\bar{\pi}_i$, η μέση πιθανότητα επιτυχίας της i ομάδας.

Κατασκευάζουμε λοιπόν στην R την παρακάτω συνάρτηση, μέσω της οποίας θα γίνει ο έλεγχος Hosmer-Lemeshow:

```
> hosmerlem = function(y, yhat, g=5) {
  cutyhat = cut(yhat, breaks = quantile(yhat, probs=seq(0,1, 1/g)),
   include.lowest=TRUE)
  obs = xtabs(cbind(1 - y, y) ~ cutyhat)
  expect = xtabs(cbind(1 - yhat, yhat) ~ cutyhat)
  chisq = sum((obs - expect)^2/expect)
  P = 1 - pchisq(chisq, g - 2)
  return(list(chisq=chisq,p.value=P)) }
```

Καλώντας την παραπάνω συνάρτηση έχουμε τα αποτελέσματα του ελέγχου για το πρώτο μοντέλο μας:

```
> hosmerlem(y=financial_condition, yhat=fitted(mylogit))
$chisq
[1] 9.43619
$p.value
[1] 0.02401998
```

Και για το δεύτερο μοντέλο μας αντίστοιχα:

```
> hosmerlem(y=financial_condition, yhat=fitted(mylogit2))
$chisq
[1] 6.400199
$p.value
[1] 0.09368262
```

Με βάση τα αποτελέσματα των ελέγχων, αν και το δείγμα δεδομένων μας είναι πολύ μικρό, παρατηρούμε ότι το δεύτερο μοντέλο, με βάση την p-value (0.09368), έχει καλύτερη προσαρμογή από ότι το πρώτο μοντέλο. Αυτό σημαίνει ότι η πραγματική οικονομική κατάσταση κάθε τράπεζας δε διαφέρει σημαντικά από τις προβλεπόμενες τιμές που μας δίνει το μοντέλο μας. Σε αντίθεση με το πρώτο μοντέλο του οποίου η προσαρμογή δεν είναι πολύ ικανοποιητική. Παρ'όλα αυτά, από τη στιγμή που τον αριθμό των ομάδων στις οποίες χωρίζονται οι παρατηρήσεις μας, τον καθορίζουμε εμείς, ο έλεγχος Hosmer-Lemeshow δεν αποτελεί επίσημο μέτρο εκτίμησης του αν ένα μοντέλο προσαρμόζεται καλά, αλλά μία ένδειξη.

Ο σκοπός των παραπάνω, ήταν απλώς να παρουσιάσουμε κάποια βασικά βήματα που μπορούμε να ακολουθήσουμε για την επεξεργασία των δεδομένων μας στην περίπτωση της λογιστικής παλινδρόμησης και την κατασκευή μοντέλων μέσω της R, και σίγουρα δεν πρόκειται για ολοκληρωμένη στατιστική ανάλυση. Σε επόμενο κεφάλαιο, θα δούμε εφαρμογή συγκεκριμένης λογιστικής παλινδρόμησης σε δυαδικές χρονοσειρές, όπου τα όσα αναλύθηκαν μέχρι στιγμής αποτελούν απαραίτητο υπόβαθρο για την κατανόηση της όλης διαδικασίας.

1.4 ΠΑΛΙΝΔΡΟΜΗΣΗ POISSON

1.4.1 Γενικά στοιχεία

Η παλινδρόμηση Poisson βασίζεται και αυτή στα γενικευμένα γραμμικά μοντέλα και είναι μια εξίσου σημαντική και χρήσιμη τεχνική, που χρησιμοποιούμε όταν έχουμε να κάνουμε με δεδομένα συχνοτήτων *(counts)*, δηλαδή με δεδομένα που αφορούν τον αριθμό εμφάνισης κάποιου γεγονότος. Είναι το δεύτερο είδος παλινδρόμησης, μαζί με τη λογιστική, που θα ασχοληθούμε παρακάτω, γι'αυτό και θα κάνουμε μια περιγραφή των βασικών χαρακτηριστικών της.

Όταν η μεταβλητή απόκρισης y, περιγράφει αριθμό εμφανίσεων κάποιων γεγονότων σε συγκεκριμένο χρόνο ή χώρο, τότε η κατάλληλη κατανομή για την ανάλυση των συγκεκριμένων δεδομένων είναι αυτή της Poisson. Παραδείγματα τέτοιων δεδομένων μπορεί να είναι ο αριθμός τροχαίων ατυχημάτων ανά μήνα, ο αριθμός θανάτων από συγκεκριμένη ασθένεια ανά έτος, ή ο αριθμός των πελατών σε μια τράπεζα ανά ώρα, κλπ.

Από τη στιγμή που η μεταβλητή απόκρισης είναι μεν ποσοτική αλλά δεν είναι σε καμία περίπτωση συνεχής, είναι αντίθετα διακριτή, παίρνοντας μόνο μη αρνητικές ακέραιες τιμές $(0,1,2,\ldots)$, είναι σίγουρο πως δεν μπορούμε να χρησιμοποιήσουμε γραμμική παλινδρόμηση. Επίσης, εάν υποθέσουμε ότι χρησιμοποιούσαμε τη γραμμική παλινδρόμηση και την κανονική κατανομή, τότε θα υπήρχε η απαίτηση της σταθερής διασποράς ($Var(y)=\sigma^2$) για όλες τις παρατηρήσεις μας, πράγμα που προφανώς δεν ισχύει σε αυτή την περίπτωση. Άρα, πρέπει όπως και στην περίπτωση της λογιστικής παλινδρόμησης, να βρούμε μια κατανομή που θα μοντελοποιεί κατάλληλα το συγκεκριμένο είδος δεδομένων, και γι'αυτό το λόγο επιλέγουμε την κατανομή Poisson.

Έτσι, η μεταβλητή απόκρισης y, θα ακολουθεί κατανομή Poisson, η οποία δίνεται από τον παρακάτω τύπο: (Οικονόμου & Καρώνη , 2010)

$$y \sim Poisson(\mu) \quad \text{τότε} \quad f(y)=\frac{e^{-y}\mu^{y}}{y!}\ ,\ \mu>0\ ,\ y=0,1,2,\ldots$$

$$\text{με} \qquad E(y)=\mu \qquad \text{και} \qquad Var(y)=\mu$$

όπου μ, είναι η παράμετρος της κατανομής και η μέση τιμή της, και πολλές φορές δίνεται σαν το «ρυθμό» ενός γεγονότος, δηλαδή, για παράδειγμα, ο αριθμός των θανάτων ανά 1000 ασθενείς που πάσχουν από καρκίνο του πνεύμονα. (Εάν έχουμε για παράδειγμα έναν αναμενόμενο αριθμό τροχαίων ατυχημάτων λ, ανά μονάδα χρόνου, τότε η τυχαία μεταβλητή που "μετράει" τον αριθμό των ατυχημάτων σε ένα χρονικό διάστημα *t*, ακολουθεί κατανομή Poisson με $\mu=\lambda t$.)

Αφού η παλινδρόμηση Poisson, αποτελεί μια μορφή των γενικευμένων γραμμικών μοντέλων, η κατασκευή του μοντέλου της, που συνδέει τη μεταβλητή απόκρισης y με τις επεξηγηματικές μεταβλητές, βασίζεται και αυτό στην εξής δομή:

$$\eta_i = g(E(y_i)) = g(\mu_i) = x_i^T \beta$$

Για την κατανομή Poisson, η συνάρτηση σύνδεσης είναι η λογαριθμική (log), δηλαδή:

$$\eta_i = g(\mu_i) = \log(\mu_i)$$

και έτσι η μέση τιμή του μοντέλου έχει τη μορφή:

$$\boxed{E(y_i) = \mu_i = \mathrm{e}^{x^T \beta}}$$

Άρα οι αναμενόμενες τιμές των παρατηρήσεων y_i, που αποτελούν τυχαίες και ανεξάρτητες μεταβλητές της κατανομής Poisson, δίνονται από τη συνάρτηση της μέσης τιμής:

$$\boxed{\log_e(E(y_i)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n}$$

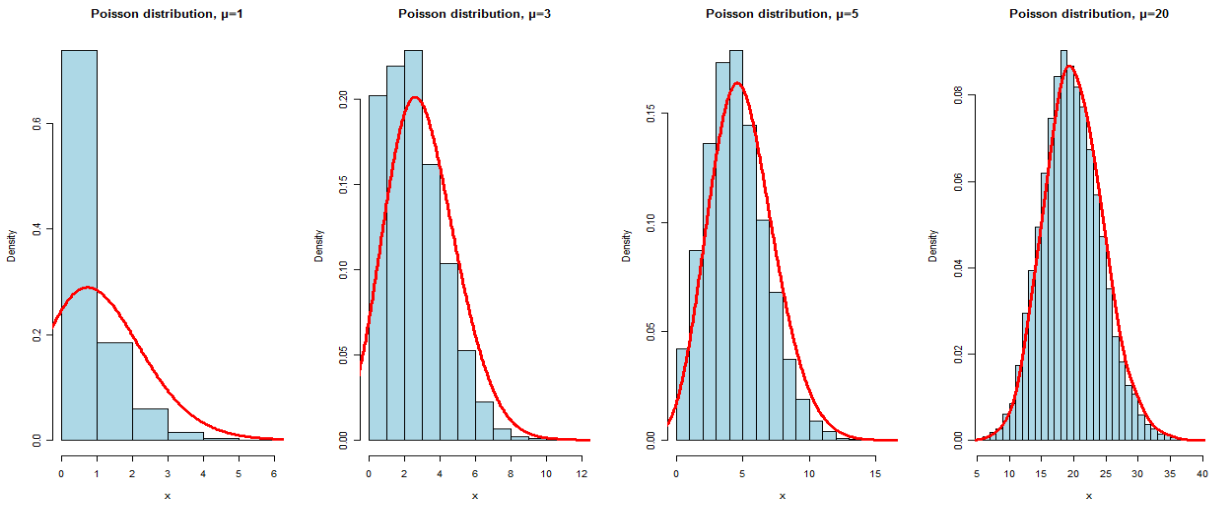
ή ισοδύναμα:

$$E(y_i) = (e^{\beta_0})(e^{\beta_1 x_1})(e^{\beta_2 x_2})\ldots(e^{\beta_n x_n})$$

Έχοντας κατασκευάσει το μοντέλο, μπορούμε να δούμε συνολικά κάποιες παραδοχές που ισχύουν για το μοντέλο παλινδρόμησης Poisson:

- Ο λογάριθμος της μέσης τιμής της μεταβλητής απόκρισης μεταβάλλεται γραμμικά σε σχέση με τις επεξηγηματικές μεταβλητές.
- Οι μεταβολές της μέσης τιμής της μεταβλητής απόκρισης, από τις συνολικές επιπτώσεις των επεξηγηματικών μεταβλητών σε αυτήν, είναι πολλαπλασιαστικές.
- Οι παρατηρήσεις είναι ανεξάρτητες.
- Η παλινδρόμηση Poisson είναι χρήσιμη για "σπάνια" γεγονότα, δηλαδή για μεταβλητές που είναι σχετικά μικροί ακέραιοι αριθμοί, συμπεριλαμβανομένου και του μηδενός.

Στο Γράφημα 4.1 παρατηρούμε την κατανομή Poisson, όπου η παράμετρος μ παίρνει τις εξής τιμές μ=1, μ=3, μ=5 και μ=20.



Γράφημα 4.1 Κατανομής Poisson, μέσω της R, για 4 διαφορετικές τιμές της παραμέτρου μ (μ=1,μ=3,μ=5,μ=20)

1.4.2 Παράδειγμα μοντέλου παλινδρόμησης Poisson

Όπως και στην περίπτωση της λογιστικής παλινδρόμησης, θα είναι χρήσιμο για αργότερα, να παρουσιάσουμε μια γενική εικόνα του πως χρησιμοποιούμε την παλινδρόμηση Poisson, μέσω ενός απλού παραδείγματος.

Θα χρησιμοποιήσουμε τα παρακάτω δεδομένα, τα οποία βλέπουμε στον Πίνακα 4.1 ο οποίος αποτελείται από τέσσερις στήλες. Τα δεδομένα μας, έχουν να κάνουν με ομάδες καπνιστών και μη καπνιστών που παρατηρήθηκαν για κάποια χρόνια, καθώς και με τις περιπτώσεις καρκίνου του πνεύμονα που διαγνώστηκαν στις συγκεκριμένες ομάδες ατόμων. Ο σκοπός της παλινδρόμησης, είναι να εξετάσουμε αν και πως επηρεάζει το κάπνισμα την εμφάνιση καρκίνου του πνεύμονα.

Αριθμός τσιγάρων ανά ημέρα	Χρόνια καπνίσματος	Αριθμός ατόμων που παρατηρήθηκαν	Αριθμός περιπτώσεων καρκίνου του πνεύμονα
0	15	10366	1
0	25	5969	0
0	35	3512	0
0	45	1421	0
0	55	826	2
5	15	3121	0
5	25	2288	0
5	35	1648	1
5	45	927	0
5	55	606	0
11	15	3577	0
11	25	2546	1
11	35	1826	0
11	45	988	2
11	55	449	3
16	15	4317	0
16	25	3185	0
16	35	1893	0
16	45	849	2
16	55	280	5
20	15	5683	2
20	25	5483	3
20	35	3646	5
20	45	1567	9
20	55	416	7
27	15	3042	2
27	25	4290	5
27	35	3529	9
27	45	1409	10
27	55	284	3
40	15	670	1
40	25	1482	3
40	35	1336	6
40	45	556	7
40	55	104	1

Πίνακας 4.1: Δεδομένα καπνιστών και καρκίνου του πνεύμονα
Πηγή δεδομένων: Πανεπιστήμιο Princeton

Στον πίνακα έχουμε ουσιαστικά 35 ομάδες ατόμων όπου στην πρώτη στήλη, έχουμε για κάθε ομάδα, τον αριθμό των τσιγάρων που κατανάλωναν καθημερινά, στη δεύτερη τον αριθμό των χρόνων που κάπνιζαν, στην τρίτη τον αριθμό των ατόμων που ανήκουν στην ομάδα και τέλος, στην τέταρτη, τον αριθμό εμφανίσεων καρκίνου του πνεύμονα στην κάθε ομάδα.

Έχουμε λοιπόν:

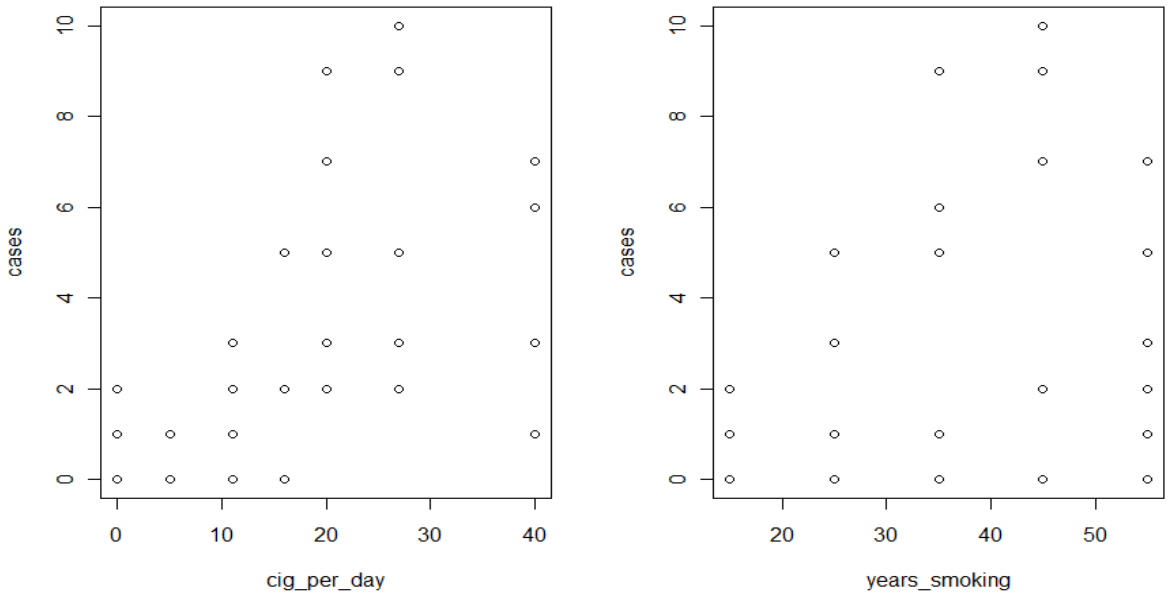
- Μεταβλητή απόκρισης y, τον αριθμό εμφανίσεων καρκίνου του πνεύμονα σε κάθε ομάδα
- Επεξηγηματική μεταβλητή x_1, τον ημερήσιο αριθμό τσιγάρων κάθε ομάδας
- Επεξηγηματική μεταβλητή x_2, τα χρόνια καπνίσματος των ατόμων που ανήκουν στην κάθε ομάδα
- Επεξηγηματική μεταβλητή x_3, ο αριθμός των ατόμων που παρατηρήθηκαν σε κάθε ομάδα.

Για την επεξεργασία και την ανάλυση των δεδομένων μας, θα χρησιμοποιήσουμε και πάλι το πρόγραμμα της R.
Όπως και στο προηγούμενο παράδειγμα 3.2, εισάγουμε τα δεδομένα στην R, και φτιάχνουμε με αυτά ένα πλαίσιο, στο οποίο έχουμε 4 στήλες, τη μεταβλητή απόκρισης “cases”, την πρώτη επεξηγηματική μεταβλητή “cig_per_day”, τη δεύτερη “years_smoking ”, και την τρίτη “persons”.

Εισάγοντας τις εντολές:

```
> plot(cig_per_day,cases)
> plot(years_smoking,cases)
```

έχουμε μια πρώτη περιγραφική εικόνα του πως σχετίζονται οι δύο επεξηγηματικές μας μεταβλητές με τη μεταβλητή απόκρισης μέσω των δύο παρακάτω γραφημάτων:



Από τα γραφήματα, παρατηρούμε ότι και το καθημερινό κάπνισμα, που υποδηλώνει η μεταβλητή "cig_per_day", αλλά και το διαχρονικό κάπνισμα, που υποδηλώνει η μεταβλητή "years_smoking ", φαίνεται αρχικά, να σχετίζονται με την εμφάνιση καρκίνου στον πνεύμονα, καθώς υπάρχει μια αυξητική τάση των περιπτώσεων του καρκίνου ενώ αυξάνονται η ημερήσια κατανάλωση τσιγάρων και τα συνολικά χρόνια καπνίσματος αντίστοιχα.

Θα κατασκευάσουμε λοιπόν το μοντέλο παλινδρόμησης Poisson που θα περιέχει τις τρείς επεξηγηματικές μας μεταβλητές, για να ελέγξουμε αν όντως επηρεάζουν με κάποιο τρόπο την μεταβλητή απόκρισης.

Κατασκευάζουμε το μοντέλο εισάγοντας την εντολή:
>poisson<-glm(cases~persons+cig_per_day+years_smoking, family="poisson", dataframe)

ενώ με την εντολή:
> summary(poisson)
έχουμε τα παρακάτω αποτελέσματα προσαρμογής του μοντέλου παλινδρόμησης:

```
Call:
glm(formula = cases ~ persons + cig_per_day + years_smoking,
family = "poisson", data = dataframe)

Deviance Residuals:
Min      1Q  Median      3Q     Max
-4.0300  -1.3749  -0.1274   0.4626   2.7066

Coefficients:
                  Estimate    Std. Error    z value     Pr(>|z|)
(Intercept)   -3.252e+00  8.042e-01    -4.043    5.27e-05 ***
persons          3.220e-04    8.872e-05   3.629    0.000284 ***
cig_per_day   5.914e-02  9.631e-03     6.140    8.23e-10 ***
years_smoking  6.002e-02  1.324e-02   4.533   5.82e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)

Null deviance:          119.648  on 34  degrees of freedom
Residual deviance:  64.286  on 31  degrees of freedom
AIC: 141.3
Number of Fisher Scoring iterations: 5
```

Παρατηρούμε, με βάση τις p-τιμές και των ελέγχων Wald, πως και οι τρεις επεξηγηματικές μεταβλητές, είναι στατιστικά σημαντικές και συμβάλλουν στο μοντέλο, δηλαδή επηρεάζουν όλες με κάποιο τρόπο, την εμφάνιση καρκίνου στον πνεύμονα. Επιβεβαιώνεται λοιπόν η αρχική υποψία που είχαμε μέσω των γραφημάτων, ότι οι επεξηγηματικές μεταβλητές σχετίζονται όντως με την y.

Τη σημαντικότητα και την καλή προσαρμογή του μοντέλου, στην περίπτωση της παλινδρόμησης Poisson, μπορούμε να την ελέγξουμε μέσω της ελεγχοσυνάρτησης Deviance. Πριν εκτελέσουμε τις κατάλληλες εντολές στην R για να δούμε τα αποτελέσματα αυτού του ελέγχου, να πούμε ότι η ελεγχοσυνάρτηση deviance ορίζεται ως η συνάρτηση με τύπο:

$$D(\hat{\beta}) = -2\{l(\hat{\beta}) - \tilde{l}\}$$

όπου, $l(\hat{\beta})$ η μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας για το μοντέλο με τις συμμεταβλητές ($\mu_i = e^{x_i^T \beta}$)

και $\tilde{l}$ η μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας για το μοντέλο ($\mu_i = y_i$), το λεγόμενο "κορεσμένο μοντέλο".

Η συνάρτηση της λογαριθμικής πιθανοφάνειας για το μοντέλο Poisson δίνεται από τη σχέση:

$$l = \sum_{i=1}^{n} [-\mu_i + y_i \ln \mu_i - \ln(y_i!)]$$

κι έτσι έχουμε:

$$\tilde{l} = -\sum_{i=1}^{n} -y_i + \sum_{i=1}^{n} y_i \ln y_i - \sum_{i=1}^{n} \ln(y_i!) \quad \text{και} \quad l(\hat{\beta}) = \sum_{i=1}^{n} [-e^{x_i^T \beta} + y_i x_i^T \beta - \ln(y_i!)]$$

οπότε μετά από πράξεις η συνάρτηση deviance για τη συνάρτηση Poisson απλοποιείται ως εξής:

$$\boxed{D(\hat{\beta}) = 2\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) \sim \chi^2_{n-p}}, \text{ όπου } \hat{\mu}_i = e^{x_i^T \beta}$$

Η συνάρτηση deviance κατανέμεται ασυμπτωτικά με την κατανομή χ^2 με *n-p* βαθμούς ελευθερίας, όπου *p* το πλήθος των παραμέτρων του μοντέλου. Οπότε, ουσιαστικά πρόκειται για έναν έλεγχο χ^2, που μέσω της p-τιμής που προκύπτει μπορούμε να διαπιστώσουμε αν το μοντέλο μας είναι ικανοποιητικό ή όχι. Αν η p-τιμή που θα προκύψει είναι στατιστικά σημαντική, σημαίνει ότι η προσαρμογή του μοντέλου που επιλέξαμε διαφέρει αρκετά από το κορεσμένο και δεν μπορούμε να το δεχτούμε ως ικανοποιητικό. Η σύγκριση που γίνεται με το κορεσμένο μοντέλο, για το οποίο ισχύει ότι έχει ίσο αριθμό παρατηρήσεων και παραμέτρων, οφείλεται στο ότι διαθέτει την καλύτερη δυνατή προσαρμογή. Έτσι, μέσω της deviance, όταν εμείς επιβάλουμε τη συνάρτηση σύνδεσης και τις συμμεταβλητές, μπορούμε να μετρήσουμε την απώλεια της προσαρμογής σε σχέση με αυτή του κορεσμένου, κι αν αυτή δεν είναι σημαντική, τότε το μοντέλο που επιλέξαμε μας ικανοποιεί.

Εισάγοντας λοιπόν την παρακάτω εντολή, έχουμε και τα αποτελέσματα της ελεγχοσυνάρτησης deviance:

```
>1-pchisq(poisson$deviance,poisson$df.residual)
 [1] 0.0004092999
```

Από τα αποτελέσματα του ελέγχου λοιπόν, συμπεραίνουμε ότι η προσαρμογή του μοντέλου δεν είναι ικανοποιητική αφού έχουμε στατιστικά σημαντική απώλεια προσαρμογής από αυτή του κορεσμένου.

Για τον έλεγχο της καταλληλότητας του μοντέλου, χρησιμοποιούμε επίσης διάφορα γραφήματα υπολοίπων. Μέσω των γραφημάτων αυτών, είμαστε σε θέση να εντοπίσουμε τους λόγους για τους οποίους δεν προσαρμόζεται καλά το μοντέλο μας, όπως προέκυψε από τα αποτελέσματα της ελεγχοσυνάρτησης deviance. Αυτό, οφείλεται στο ότι τα υπόλοιπα μας δείχνουν το κατά πόσον οι παρατηρήσεις y_i του δείγματος μας, συμπίπτουν με τις προβλεπόμενες από το μοντέλο μας τιμές $\hat{y}_i$. Τα υπόλοιπα που χρησιμοποιούμε συνήθως είναι τα Pearson και deviance.

Να αναφέρουμε απλώς ότι για τον υπολογισμό αυτών των υπολοίπων χρησιμοποιούνται οι παρακάτω τύποι:

- $res_i^P = \dfrac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$, $(\hat{y}_i = \hat{\mu}_i)$
- $res_i^D = \text{sgn}(y_i - \hat{y}_i)\{d_i(\hat{\beta})\}^{1/2}$, $(\hat{y}_i = \hat{\mu}_i)$

Ο τελευταίος τύπος, μας δίνει τα υπόλοιπα deviance που είναι ίσα με την ποσότητα $\{d_i(\hat{\beta})\}^{1/2}$ και με πρόσημο ανάλογα με το πρόσημο της ποσότητας $(y_i - \hat{y}_i)$. Αν $y_i - \hat{y}_i > 0$ τότε έχουμε και το αντίστοιχο θετικό πρόσημο ενώ αν $y_i - \hat{y}_i < 0$ τότε έχουμε το αντίστοιχο αρνητικό.

Μέσω των παρακάτω εντολών, "κατασκευάζουμε" τα υπόλοιπα Pearson και τα υπόλοιπα της deviance:

```
> res.dev<-residuals(poisson)
> res.pears<-residuals(poisson,type="pearson")
```

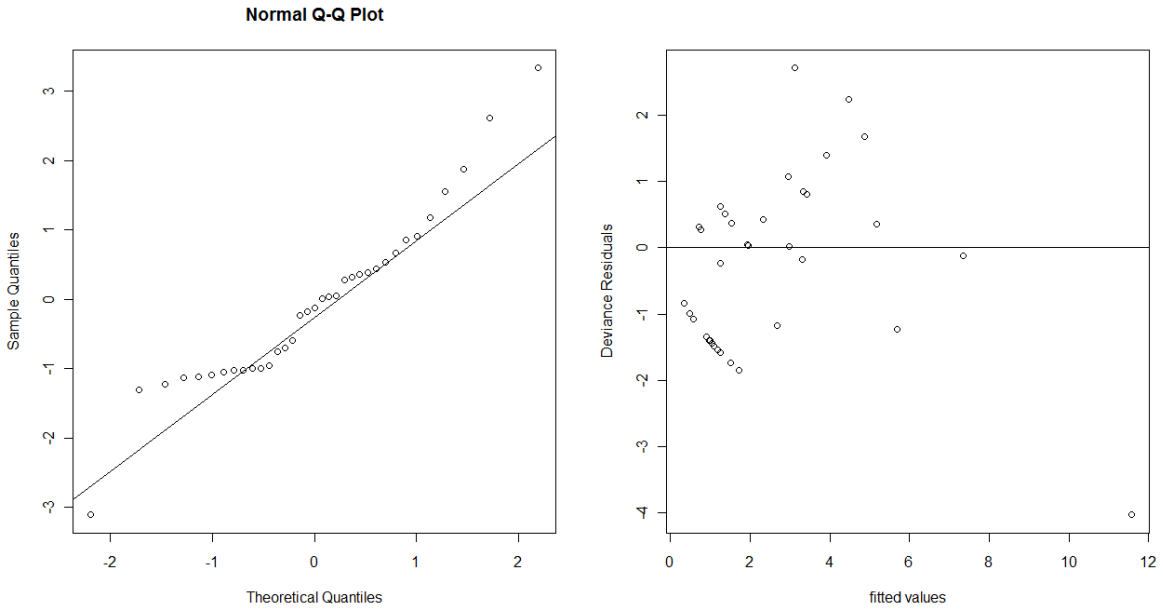
και χρησιμοποιώντας τις εντολές:

```
> qqnorm(r.pears)
> qqplot(r.pears)
```

και τις:

```
> plot(fitted.values(poisson),r.dev,xlab="fitted values", ylab="Deviance   Residuals")
> abline(h=0)
```

έχουμε τα αντίστοιχα γραφήματα για τα υπόλοιπα Pearson και τα υπόλοιπα deviance:



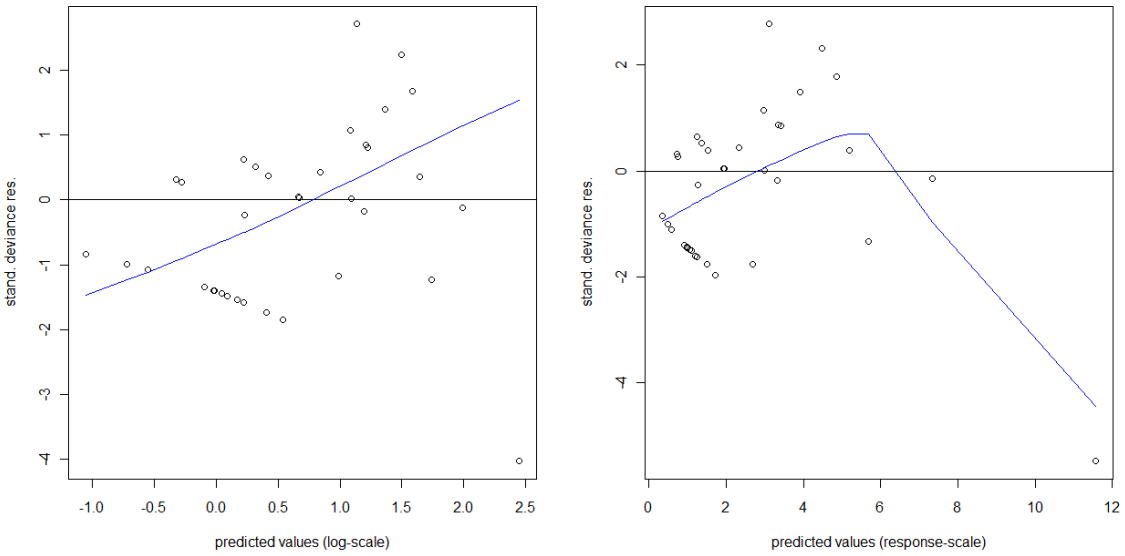
Από τη γραφική παράσταση των υπολοίπων Pearson παρατηρούμε πως η πλειοψηφία των σημείων σχηματίζει μια ευθεία γραμμή, υπάρχουν όμως αρκετά σημεία τα οποία απέχουν από τη "νοητή" ευθεία που σχηματίζεται. Η ύπαρξη αυτών των απόμακρων σημείων, θα μπορούσε να μας οδηγήσει ίσως σε έναν από τους λόγους για τους οποίους το μοντέλο μας δεν προσαρμόζεται καλά, με βάση τα προηγούμενα αποτελέσματα της ελεγχοσυνάρτησης deviance. Από τη δεύτερη γραφική παράσταση των υπολοίπων deviance, παρατηρούμε πως κατανέμονται τυχαία συναρτήσει των προβλεπόμενων τιμών, πράγμα που σημαίνει ότι η υπόθεση της ανεξαρτησίας των παρατηρήσεων μας, φαίνεται να ισχύει.

Οι λόγοι για τους οποίους το μοντέλο μας δεν περιγράφει ικανοποιητικά τα δεδομένα μας και κυρίως στις πολύ χαμηλές και υψηλές τιμές y, θα μπορούσαν να είναι διάφοροι, όπως για παράδειγμα το ότι η μεταβλητή απόκρισης "cases" εξαρτάται και από άλλες επεξηγηματικές μεταβλητές και όχι μόνο από τις ήδη υπάρχουσες στο μοντέλο.

Τέλος, εισάγοντας τις παρακάτω εντολές στην R:

```
> predict.link <- predict(poisson, type = "link")
> plot(res.dev ~ predict.link, ylab = "stand. deviance res.", xlab = "predicted values
  (log-scale)")
> abline(h = 0)
> lines(lowess(res.dev ~ predict.link, f = 1), col = "blue")
> predict.fit <- predict(poisson, type = "response")
> plot(res ~ predict.fit, ylab = "stand. deviance res.", xlab = "predicted values
(response-scale)")
> abline(h = 0)
> lines(lowess(res.dev ~ predict.fit, f = 1), col = "blue")
```

προκύπτουν τα παρακάτω διαγράμματα, στα οποία φαίνονται και πάλι τα υπόλοιπα deviance συναρτήσει των προβλεπόμενων τιμών, αλλά στη μία περίπτωση οι τιμές αυτές βρίσκονται σε λογαριθμική κλίμακα, με βάση τη συνάρτηση σύνδεσης που χρησιμοποιούμε στο μοντέλο παλινδρόμησης Poisson, ενώ στη δεύτερη περίπτωση έχουμε το γράφημα που είδαμε και παραπάνω, όπου οι προβλεπόμενες τιμές βρίσκονται στην ίδια κλίμακα με τη μεταβλητή απόκρισης:



Όπως και την περίπτωση του παραδείγματος της λογιστικής παλινδρόμησης, είδαμε μια περιγραφή των βασικών χαρακτηριστικών ενός μοντέλου παλινδρόμησης Poisson και το πως μπορεί να χρησιμοποιηθεί αυτού του είδους η παλινδρόμηση μέσω της R, για να επεξεργαστούμε συγκεκριμένο είδος δεδομένων. Και για την περίπτωση της παλινδρόμησης Poisson, θα δούμε σε επόμενο κεφάλαιο, αντίστοιχη εφαρμογή για απαριθμητή χρονοσειρά, δηλαδή χρονοσειρά που αφορά τον αριθμό εμφάνισης συγκεκριμένου γεγονότος.

ΚΕΦΑΛΑΙΟ 2

ΑΝΑΛΥΣΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΧΡΟΝΟΣΕΙΡΩΝ

2.1 ΕΙΣΑΓΩΓΗ

Η ανάλυση δεδομένων που έχουν να κάνουν με το χρόνο, και η πρόβλεψη μελλοντικών τιμών των αντίστοιχων αυτών μεταβλητών, αποτελεί σημαντικό πρόβλημα για πολλές εφαρμογές σε διάφορους τομείς όπως είναι τα οικονομικά, η μηχανική, η βιολογία, η ιατρική, η κοινωνιολογία. Η επίλυση αυτών των προβλημάτων και η επεξεργασία τέτοιων δεδομένων, γίνεται μέσω των χρονοσειρών, για των οποίων τη μοντελοποίηση και την ανάλυση, χρησιμοποιούμε διάφορες στατιστικές μεθόδους, γραφικές και αριθμητικές.

Να εξηγήσουμε αρχικά, ότι μια *χρονοσειρά*, είναι μια ακολουθία παρατηρήσεων x_t με $t = 1, \ldots, \mathrm{T}$, όπου κάθε μία από αυτές, καταγράφηκε σε μια δεδομένη χρονική στιγμή (ή διάστημα) t, και αποτελούν ένα δείγμα με ισαπέχοντα και διαδοχικά χρονικά σημεία ή χρονικά διαστήματα. Μια χρονοσειρά, εκφράζει την εξέλιξη ενός *στοχαστικού* συστήματος. Ένα στοχαστικό σύστημα είναι ένα σύστημα του οποίου η συμπεριφορά είναι τυχαία, σε αντίθεση με ένα *ντετερμινιστικό* σύστημα του οποίου η συμπεριφορά περιγράφεται και προσδιορίζεται, συνήθως μέσω διαφορικών εξισώσεων. (Κοκολάκης, 2007)

Η εξέλιξη ενός στοχαστικού συστήματος περιγράφεται πιθανοθεωρητικά μέσω μιας στοχαστικής ανέλιξης, δηλαδή μιας οικογένειας τυχαίων μεταβλητών $\{X_t, t \in T\}$ που ορίζεται πάνω σε ένα χώρο πιθανότητας (Ω, F, P) (όπου Ω ο δειγματοληπτικός χώρος, F η σ-άλγεβρα και P η συνάρτηση πιθανότητας). Όταν η παράμετρος t εκφράζει το χρόνο, τότε η στοχαστική αυτή διαδικασία καλείται χρονοσειρά. Ο όρος λοιπόν της χρονοσειράς, αναφέρεται ουσιαστικά σε μια στοχαστική ανέλιξη $\{X_t, t \in T\}$ ή σε μια *τροχιά* αυτής $\{x_t = X_t(\omega) : t \in T\}$.

Καταλαβαίνουμε, ότι ένα σύνολο παρατηρήσεων x_t μιας χρονοσειράς, διαφέρει πολύ από ένα οποιοδήποτε άλλο σύνολο δεδομένων από τη στιγμή που υπάρχει χρονική εξάρτηση, η διασπορά μεταβάλλεται στο χρόνο και συνήθως επηρεάζονται από τάσεις και περιοδικότητες. Επομένως, οι στατιστικές μέθοδοι και διεργασίες που χρησιμοποιούνται για ανεξάρτητα ή ομοιόμορφα κατανεμημένα δεδομένα, δεν είναι κατάλληλες στην περίπτωση των χρονοσειρών. Τις διαδικασίες που ακολουθούνται, θα δούμε παρακάτω.

2.2 ΔΙΑΔΙΚΑΣΙΑ ΑΝΑΛΥΣΗΣ ΚΑΙ ΠΡΟΒΛΕΨΗΣ ΧΡΟΝΟΣΕΙΡΩΝ

Εάν υποθέσουμε ότι μας ενδιαφέρει μια μεταβλητή x, και έχουμε ένα σύνολο x_t παρατηρήσεων της μεταβλητής αυτής από $t = 1, \ldots, \mathrm{T}$ διαφορετικές χρονικές στιγμές

(μέρες, μήνες, έτη, κλπ), τότε το σύνολο των σημείων αυτών των παρατηρήσεων, δημιουργεί ένα χρονοδιάγραμμα, η μελέτη του οποίου μπορεί να μας δώσει μια γενική εικόνα του πώς εξελίσσεται διαχρονικά η μεταβλητή x, ή αντίστοιχα το φαινόμενο που βρίσκεται υπό έρευνα. Παράδειγμα τέτοιου χρονοδιαγράμματος αποτελεί το Γράφημα 2.1. Η αντιμετώπιση τέτοιου είδους έρευνας ή προβλήματος, μπορεί να χωριστεί σε δύο βασικά στάδια, αυτό της ανάλυσης και αυτό της πρόβλεψης. Το κομμάτι της ανάλυσης, στοχεύει στο να περιγράψει, να επεξηγήσει και να ελέγξει τον τρόπο με τον οποίο συμπεριφέρεται η μεταβλητή στο χρόνο κι έτσι να μελετηθούν οι μηχανισμοί δημιουργίας της χρονοσειράς. Στο κομμάτι της πρόβλεψης, μπορούμε μέσω της μοντελοποίησης της χρονοσειράς, να υπολογίσουμε μελλοντικές τιμές της και να κατασκευάσουμε διαστήματα πρόβλεψης για χρόνο μεταγενέστερο των παρατηρήσεων που διαθέτουμε μέχρι στιγμής.

Εμείς, θα ασχοληθούμε με συγκεκριμένο κομμάτι της μοντελοποίησης μιας χρονοσειράς, αλλά είναι σημαντικό να δούμε σύντομα τα βασικά βήματα της όλης διαδικασίας, ώστε να έχουμε μια ολοκληρωμένη εικόνα για το πως φτάνουμε στη μοντελοποίηση και στην πρόβλεψη μιας χρονοσειράς, (Montgomery et al., 2008; Brockwell & Davis, 2002):

- Πρώτο βήμα αποτελεί η γραφική αναπαράσταση της χρονοσειράς μας, ώστε να διαπιστώσουμε τα βασικά χαρακτηριστικά της, εάν δηλαδή υπάρχει κάποιου είδους τάση, εποχικότητα ή περιοδικότητα, ή ακόμα, αν υπάρχουν κάποιες μεγάλες αλλαγές στη συμπροφορά της χρονοσειράς σε κάποιο χρονικό διάστημα.

- Δεύτερο και σημαντικότατο βήμα αποτελεί η εξάλειψη κάθε είδους τάσης και εποχικότητας που παρουσιάστηκε στο πρώτο βήμα. Αυτό μπορεί να επιτευχθεί με πολλούς τρόπους, όπως μέσω της μεθόδου των διαφορών (αντικαθιστώντας δηλ. την αρχική χρονοσειρά $\{X_t\}$ με την $\{Z_t := X_t - X_{t-d}\}$ για κάποιο θετικό ακέραιο d), είτε μέσω κατάλληλου μετασχηματισμού της μεταβλητής, είτε μέσω της προσαρμογής κάποιου παλινδρομικού μοντέλου στα δεδομένα μας. Ο σκοπός όλων αυτών, είναι τελικά να καταλήξουμε σε μια στάσιμη χρονοσειρά, της οποίας οι τιμές αναφέρονται ως **υπόλοιπα**. Ο όρος της στασιμότητας θα αναλυθεί παρακάτω.

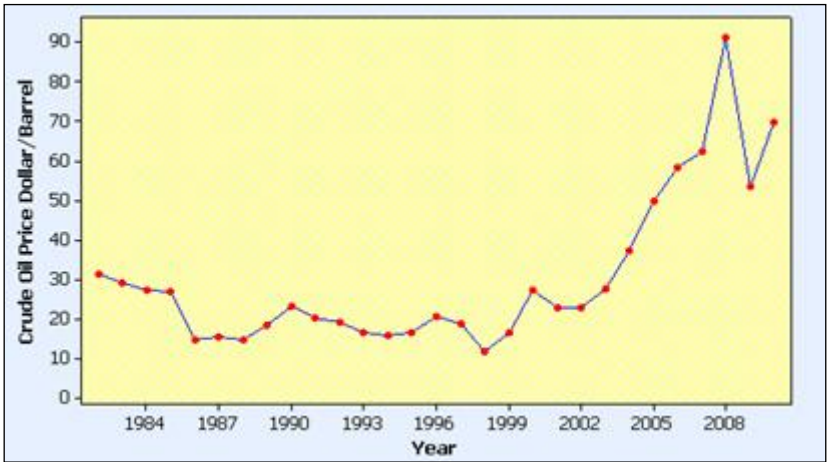
- Τρίτο βήμα είναι η επιλογή κατάλληλου μοντέλου για την προσαρμογή των υπολοίπων της χρονοσειράς που προέκυψε από το προηγούμενο βήμα. Κατά τη διαδικασία μοντελοποίησης της χρονοσειράς γίνεται χρήση γραφημάτων της αυτοσυνδιακύμανσης ή της αυτοσυσχέτισης, βασικών εννοιών που θα αναλυθούν παρακάτω. Συνηθισμένα μοντέλα χρονοσειρών αποτελούν τα αυτοπαλινδρομούμενα (Autoregressive-AR), αυτά του κινητού μέσου (Moving Average-MA) και τα μικτά (Autoregressive Moving Average-ARIMA). Ο έλεγχος της καταλληλότητας του μοντέλου που εφαρμόσαμε, γίνεται μέσω των υπολοίπων που προκύπτουν από το μοντέλο, τα οποία θα πρέπει να έχουν τη μορφή μιας χρονοσειράς λευκού θορύβου, δηλαδή μορφή τυχαίας σειράς.

- Τελευταίο βήμα αποτελεί η πρόβλεψη, μέσω της οποίας είμαστε σε θέση να υπολογίσουμε μελλοντικές τιμές της χρονοσειράς μας. Δηλαδή, αν υποθέσουμε πως έχουμε μια χρονοσειρά $\{x_t\}$ με $t=1,\ldots,T$, τότε ξεκινώντας από μια χρονική στιγμή h, η οποία αποτελεί την "αρχή της πρόβλεψης" *(forecast origin)*, μπορούμε να προβλέψουμε την τιμή x_{h+l}, όπου ο θετικός ακέραιος l αποτελεί το χρονικό ορίζοντα της πρόβλεψης *(forecast horizon)*. Οι παρακάτω δύο σχέσεις, δηλώνουν την πρόβλεψη των l-βημάτων μπροστά από την χρονοσειρά $\{x_t\}$ με αρχή πρόβλεψης την στιγμή h και το αντίστοιχο σφάλμα πρόβλεψης:

$$\hat{x}_h(l) = f(x_h,\ldots,x_{h-(T-1)}) \quad \text{με} \quad e_h(l) = x_{h+l} - \hat{x}_h(l)$$

Να αναφέρουμε απλώς, ότι για τη διενέργεια προβλέψεων υπάρχει ένας μεγάλος αριθμός μεθόδων που μπορούν να χρησιμοποιηθούν ανάλογα με την περίπτωση και το μοντέλο που έχουμε επιλέξει για τη χρονοσειρά μας. Οι μέθοδοι εξομαλύνσεις *(smoothing methods)*, η διάσπαση χρονοσειρών *(time series decomposition)*, η κατά Box-Jenkins ανάλυση, αποτελούν παραδείγματα τεχνικών που ακολουθούνται για την πρόβλεψη μελλοντικών τιμών μιας χρονοσειράς. Το κομμάτι όμως της πρόβλεψης δε θα μας απασχολήσει στην παρούσα εργασία, γι'αυτό και δε θα αναλυθεί περεταίρω.

Στο επόμενο κεφάλαιο, θα ασχοληθούμε με δύο συγκεκριμένα είδη χρονοσειρών. Χρονοσειρές που έχουν να κάνουν με αριθμό εμφάνισης κάποιου γεγονότος ***(count time series)*** και χρονοσειρές των οποίων οι παρατηρήσεις μπορούν να πάρουν δύο μόνο τιμές, 0 και 1, συμβολίζοντας την "αποτυχία" ή την "επιτυχία" κάποιου γεγονότος αντίστοιχα ***(binary time series)***. Σε αυτές, θα προσαρμόσουμε δύο μοντέλα παλινδρόμησης, τα οποία αναλύσαμε στο προηγούμενο κεφάλαιο, αυτό της Poisson και αυτό της λογιστικής, για την κάθε περίπτωση αντίστοιχα. Πριν φτάσουμε όμως στην ανάλυση των δύο αυτών μοντέλων, είναι απαραίτητο στο παρόν κεφάλαιο, να δούμε κάποιες βασικές έννοιες και μεθόδους της ανάλυσης χρονοσειρών που χωρίς αυτές δεν θα γίνει κατανοητή η περεταίρω μελέτη των παλινδρομικών μας μοντέλων.



Γράφημα 2.1: Παράδειγμα χρονοδιαγράμματος όπου φαίνεται διαχρονικά η τιμή του πετρελαίου ανά βαρέλι, απο το 1982 εως το 2010

2.2 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΟΡΙΣΜΟΙ

ΣΤΑΣΙΜΟΤΗΤΑ

Αναφερθήκαμε προηγουμένως στην έννοια της στασιμότητας, λέγοντας ότι προϋπόθεση για τη μοντελοποίηση μιας χρονοσειράς, αποτελεί η εξάλειψη κάθε είδους πιθανής τάσης, με σκοπό να μετατρέψουμε τη χρονοσειρά μας σε στάσιμη. Παρατηρώντας τη χρονοσειρά $\{X_t, t \in T\}$ από μία χρονική στιγμή $t = 0$, έστω μέχρι και την παρούσα στιγμή $t = s$, τότε γνωρίζουμε τη συγκεκριμένη αυτή τροχιά της $\{x_t : 0 \le t \le s\}$. Εάν θέλουμε να προβλέψουμε κάποιες μελλοντικές της τιμές, X_{s+h} για κάποιο ακέραιο h, τότε θα πρέπει να βασιστούμε στις ήδη γνωστές μας τιμές και στην όποια εξάρτηση υπάρχει μεταξύ της τροχιάς και των μελλοντικών τιμών. Αυτό όμως, προϋποθέτει ότι κάποια βασικά πιθανοθεωρητικά χαρακτηριστικά της χρονοσειράς παραμένουν αμετάβλητα στις χρονικές μετατοπίσεις.

Συγκεκριμένα, λαμβάνοντας υπόψη τα βασικότερα χαρακτηριστικά μιας χρονοσειράς $\{X_t, t \in T\}$ με $E(X_t^2) < \infty$, (Κοκολάκης,2007):

τη **συνάρτηση μέσου** που είναι η $\mu_X(t) = E(X_t)$,
τη **συνάρτηση διασποράς** που είναι η $\sigma^2(t) = V[X_t] = E[(X_t - \mu_t)^2]$ και
τη **συνάρτηση αυτοσυνδιακύμανσης** (ACVF) που είναι η

$$\boxed{\gamma(t,h) = Cov(X_t, X_{t+h}) = E[(X_t - \mu_t)(X_{t+h} - \mu_{t+h})],} \quad t, h \in T$$

Τότε μπορούμε να διατυπώσουμε τον ορισμό της (ασθενούς) **στασιμότητας** μιας χρονοσειράς με βάση τον οποίο η $\{X_t, t \in T\}$ είναι (ασθενώς) στάσιμη όταν έχει πεπερασμένη διασπορά και ισχύουν τα παρακάτω:

- $\mu = E(X_t)$, δηλαδή η μέση τιμή είναι ανεξάρτητη του χρόνου t
- $\gamma(h) = Cov(X_t, X_{t+h})$, δηλαδή η συνάρτηση αυτοσυνδιακύμανσης είναι ανεξάρτητη του χρόνου t για κάθε h.

Στον παραπάνω ορισμό, σημειώθηκε ότι η συγκεκριμένη διατύπωση αντιστοιχεί στην **"ασθενή"** στασιμότητα καθώς υπάρχει και η έννοια της **"αυστηρής"** στασιμότητας. Με βάση τον ορισμό της αυστηρής στασιμότητας η χρονοσειρά $\{X_t, t \in T\}$ είναι αυστηρώς στάσιμη αν όλα τα πιθανοθεωρητικά της χαρακτηριστικά παραμένουν αμετάβλητα στο χρόνο, δηλαδή η $(X_{t_1}, ..., X_{t_n})$ κατανέμεται όπως η $(X_{t_{1+h}}, ..., X_{t_{n+h}})$ $\forall n \in N, t_i \in T, (i = 1, ..., n), h \in T$. Παρ'όλα αυτά, όταν αναφερόμαστε σε στάσιμη χρονοσειρά εννοούμε την ασθενή στασιμότητα.

ΣΥΝΑΡΤΗΣΕΙΣ ΑΥΤΟΣΥΝΔΙΑΚΥΜΑΝΣΗΣ ΚΑΙ ΑΥΤΟΣΥΣΧΕΤΙΣΗΣ

Έστω ότι έχουμε μια στάσιμη χρονοσειρά $\{X_t\}$. Ορίσαμε λίγο παραπάνω τη συνάρτηση αυτοσυνδιακύμανσης $\gamma(t,h)$, η οποία στην περίπτωση της στάσιμης χρονοσειράς είναι $\gamma(h) = Cov(X_t, X_{t+h})$ με *υστέρηση (lag)* h, $h \in T$ και αποτελεί βασικό εργαλείο για την περιγραφή της. Παρατηρούμε ότι για υστέρηση $h = 0$,

τότε η αυτοσυνδιακύμανση είναι ίση με τη διασπορά, δηλαδή $\gamma(0)=\sigma^2$, ενώ ισχύει πως για κάθε $h\in T$, $\gamma(h)=\gamma(-h)$.
Με βάση τη συνάρτηση αυτοσυνδιακύμανσης ορίζεται η συνάρτηση αυτοσυσχέτισης (ACF), η οποία, μέσω της διασποράς, αποτελεί την κανονικοποιημένη μορφή της αυτοσυνδιακύμανσης:

$$\boxed{\rho(h)=\frac{\gamma(h)}{\sigma^2}=\frac{\gamma(h)}{\gamma(0)}}\ ,\quad h\in T$$

Παρατηρούμε ότι $\rho(0)=1$ και $\rho(h)=\rho(-h)$ για κάθε $h\in T$.

Οι συναρτήσεις αυτοσυνδιακύμανσης $\gamma(h)$ και αυτοσυσχέτισης $\rho(h)$, αποτελούν μέτρα της γραμμικής εξάρτησης μεταξύ των παρατηρήσεων που έχουν καταγραφεί σε χρονική υστέρηση h. Γι'αυτό το λόγο, για τη διερεύνηση των δεδομένων, αλλά και για την υπόδειξη κατάλληλων μοντέλων, χρησιμοποιούμε τις γραφικές παραστάσεις των δύο αυτών συναρτήσεων οι οποίες καταλαβαίνουμε ότι αποτελούν δύο πολύ σημαντικές έννοιες στην ανάλυση χρονοσειρών.

ΤΑΣΗ ΚΑΙ ΠΕΡΙΟΔΙΚΟΤΗΤΑ

Η στασιμότητα, προφανώς και δεν αποτελεί χαρακτηριστικό για όλες τις χρονοσειρές. Πολύ συχνά, στην αναπαράσταση της χρονοσειράς παρατηρείται κάποιο «μοτίβο», είτε μέσω κάποιας συγκεκριμένης τάσης, αυξητικής ή φθίνουσας, είτε μέσω κάποιων επαναλαμβανόμενων εναλλαγών που δημιουργούν ένα είδος περιοδικότητας /κυκλικότητας με αποτέλεσμα, να έχουμε έτσι μια διακύμανση των τιμών με περίοδο κάποιο χρονικό διάστημα (εβδομάδες, μήνες, έτος κλπ). Όπως αναφέραμε και στην προηγούμενη παράγραφο, πριν προχωρήσουμε στη μοντελοποίηση μιας χρονοσειράς, θα πρέπει να ελέγξουμε αν εμφανίζει κάποια τέτοια συμπεριφορά, αν δηλαδή παραβιάζεται η προϋπόθεση της στασιμότητας.

Το γενικό μοντέλο που αναπαριστούμε μια χρονοσειρά η οποία έχει και τάση και εποχικότητα είναι το παρακάτω (Brockwell & Davis, 2002):

$$\boxed{X_t=m_t+s_t+Y_t}\qquad t\in\Re$$

Όπου:
m_t είναι η συνιστώσα μέσω της οποίας εισάγεται η τάση και είναι μια χαμηλών μεταβολών συνάρτηση του *t*. Η συνάρτηση αυτή μπορεί να είναι γραμμική, εκθετική ή ένα πολυώνυμο χαμηλού βαθμού (πχ $m_t=a+bt$ ή $m_t=a\exp\{bt\}$ ή $m_t=a_0+a_1t+a_2t^2+\ldots+a_kt^k$).

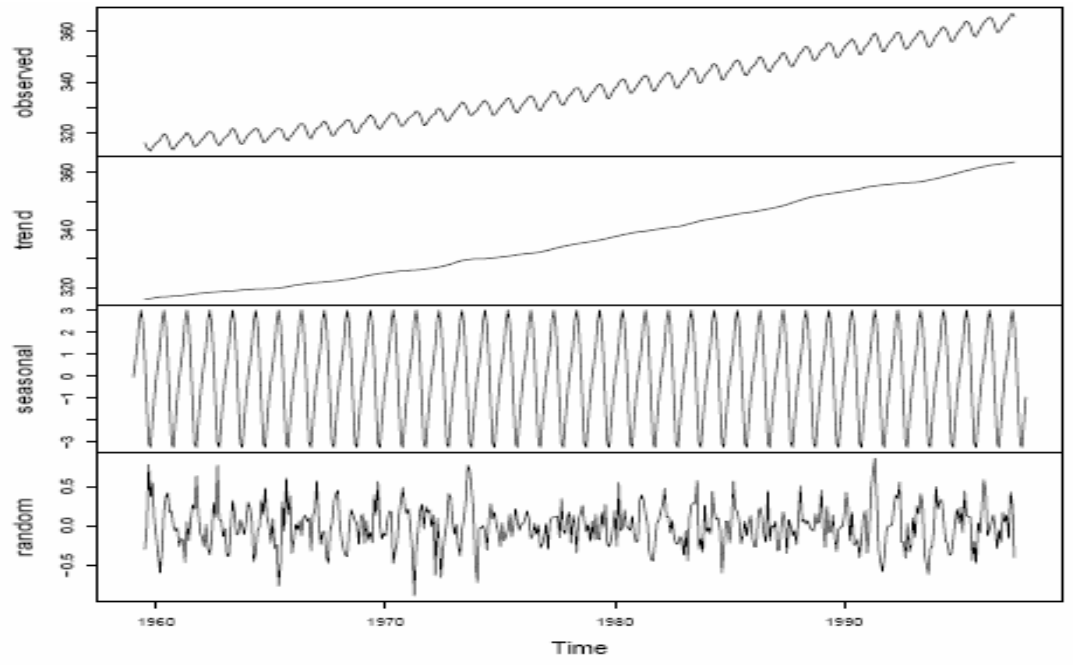
s_t, είναι η συνιστώσα μέσω της οποίας εισάγεται η εποχικότητα και είναι μια συνάρτηση με γνωστή περίοδο d. Η συνάρτηση αυτή μπορεί να έχει τη γενικότερη μορφή $s_t=\sum_{k=1}^{n}a_k\sin(2\pi\omega_kt+\theta_k)$.

Y_t, αποτελεί το *θόρυβο*, και είναι το στάσιμο κομμάτι της χρονοσειράς.

Το παραπάνω αθροιστικό μοντέλο αναπαράστασης μιας χρονοσειράς μπορεί να το συναντήσουμε κάποιες φορές και με έναν επιπλέον όρο εκτός των άλλων, αυτόν της κυκλικότητας, c_t. Η συνιστώσα της εποχικότητας s_t σε αυτή την περίπτωση, εκφράζει την κυκλική κύμανση της χρονοσειράς με περίοδο ανά έτος, ενώ η συνιστώσα της κυκλικότητας c_t , εκφράζει την κυκλική κύμανση με περίοδο μεγαλύτερη του ενός έτους.

Λευκός θόρυβος:
Μια χρονοσειρά $\{X_t\}$ λέγεται *λευκός θόρυβος* όταν $X_t \sim WN(0,\sigma^2)$, δηλαδή όταν: $E(X_t)=0$, $Var(X_t)=\sigma^2<\infty$ και $Cov(X_t,X_s)=0$, $\forall t,s\in Z$ και $t\neq s$. Παρατηρούμε μέσω του ορισμού ότι αποτελεί όντως μια στάσιμη χρονοσειρά. Να υπενθυμίσουμε ότι αποτελεί μια σημαντική έννοια, αυτή του θορύβου, καθώς όπως αναφέραμε στην προηγούμενη παράγραφο, αποτελεί κριτήριο για την επιτυχημένη μοντελοποίηση μιας χρονοσειράς, αφού ελέγχουμε γραφικά το αν τελικά η μορφή της χρονοσειράς θυμίζει αυτή του λευκού θορύβου. Η μορφή του λευκού θορύβου αντιστοιχεί στο γράφημα "random" του Γραφήματος 2.2, στο οποίο μπορούμε να δούμε τις γραφικές αναπαράστασεις όλων των συνιστωσών του γενικού μοντέλου μιας χρονοσειράς.

Για την εκτίμηση ή την απαλοιφή της όποιας τάσης και περιοδικότητας που παρουσιάζει μια χρονοσειρά, εφαρμόζονται διάφορες μέθοδοι, όπως είναι η μέθοδος των διαφορών ή η μέθοδος κινητού μέσου ή μέσω κάποιου κατάλληλου μετασχηματισμού κλπ. Παρ'όλα αυτά δε θα αναλύσουμε κάποιον από τους τρόπους με τους οποίους επιτυγχάνεται τελικά η στασιμότητα μιας χρονοσειράς με τάση και εποχικότητα, καθώς δεν αποτελούν ζητούμενο της παρούσας εργασίας.



Γράφημα 2.2: Γραφικές αναπαραστάσεις των συνιστωσών του γενικού μοντέλου μιας παρατηρούμενης χρονοσειράς, τάση-εποχικότητα-θόρυβος

ΜΟΝΤΕΛΑ AR, MA, ARMA

Πριν κλείσουμε αυτή την ενότητα, είναι σημαντικό να αναφέρουμε κάποια βασικά μοντέλα χρονοσειρών τα οποία παρέχουν το γενικό πλαίσιο για τη μελέτη στάσιμων διαδικασιών. Τα μοντέλα αυτά, που μπορούν να περιγράψουν στάσιμες χρονοσειρές χωρίζονται σε τρείς μεγάλες κατηγορίες, τα αυτοπαλινδρομούμενα μοντέλα (AR), τα μοντέλα κινητού μέσου (MA) και τα μεικτά μοντέλα (ARMA). Από τα τρία μοντέλα, εμάς μας αφορά κυρίως η έννοια του αυτοπαλινδρομούμενου μοντέλου, μιας και θα αναφερθεί και παρακάτω και είναι χρήσιμο να έχουμε κατανοήσει τον ορισμό του.

Μια γραμμική στοχαστική διαδικασία, δηλαδή μια γραμμική χρονοσειρά, ορίζεται για κάθε χρονική στιγμή t ως ένα άθροισμα τυχαίων και ασυσχέτιστων μεταβλητών (Montgomery et al., 2008):

$$X_t = \mu + \sum_{i=-\infty}^{\infty} \psi_i Z_{t-i} \ , \tag{2.1}$$

όπου $Z_t \sim WN(0, \sigma^2)$ και $\{\psi_j\}$ μια ακολουθία συντελεστών με $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

Αυτοπαλινδρομούμενα μοντέλα- $AR(p)$

Η αυτοπαλινδρομούμενη διαδικασία τάξης p ορίζεται από τον περιορισμό του αθροίσματος της σχέσης (2.1), στους p πρώτους όρους:

$$X_t = \delta + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots + \varphi_p X_{t-p} + Z_t \ , \quad Z_t \sim WN(0, \sigma^2) \tag{2.2}$$

Θα δείξουμε για το απλό, πρώτης τάξης μοντέλο $AR(1)$ το πως προκύπτει η γενικότερη σχέση (2.2).

Την έκφραση της σχέσης (2.1) μπορούμε να τη γράψουμε ως

$$X_t = \mu + \sum_{i=-\infty}^{\infty} \psi_i Z_{t-i} = \mu + \sum_{i=-\infty}^{\infty} \psi_i B^i Z_t = \mu + \Psi(B) Z_t$$

(όπου B ο οπισθοδρομικός τελεστής με τη ιδιότητα: $Bx_t = x_{t-1}$)
Και έτσι, να γραφεί ως

$$X_t = \mu + Z_t + \varphi Z_{t-1} + \varphi^2 Z_{t-2} + \ldots = \mu + \sum_{i=0}^{\infty} \varphi^i Z_{t-i} \tag{2.3}$$

Από αυτήν την τελευταία σχέση προκύπτει επίσης ότι

$$X_{t-1} = \mu + Z_{t-1} + \varphi Z_{t-2} + \varphi^2 Z_{t-3} + \ldots \tag{2.4}$$

Έτσι, συνδυάζοντας τις σχέσεις (2.3) και (2.4) έχουμε την

$$X_t = \mu + Z_t + \underbrace{\varphi Z_{t-1} + \varphi^2 Z_{t-2} + \ldots}_{=\varphi X_{t-1} + \varphi\mu} = \mu - \varphi\mu + \varphi X_{t-1} + Z_t = \delta + \varphi X_{t-1} + Z_t \tag{2.5}$$

όπου $\delta = \varphi X_{t-1} + Z_t$.

Η σχέση (2.5) αποτελεί την έκφραση του μοντέλου AR τάξης 1, δηλαδή για $p=1$. Μπορούμε να καταλάβουμε τώρα, πως προκύπτει και η γενική σχέση (2.2) για οποιαδήποτε τάξη p.

Από την έκφραση (2.2), καταλαβαίνουμε ότι μέσω του $AR(p)$, η X_t ορίζεται ως ένας γραμμικός συνδυασμός των προηγούμενων p τιμών της χρονοσειράς, $X_{t-1},...,X_{t-p}$, διαταραγμένο από το λευκό θόρυβο Z_t. Γίνεται επίσης κατανοητό πλέον για ποιό λόγο καλείται "αυτοπαλινδρομούμενο" το μοντέλο, καθώς παρατηρώντας για παράδειγμα τη σχέση (2.5), μπορούμε να τη δούμε σαν μια παλινδρόμηση της X_t με την X_{t-1}. Με την ίδια λογική, το γενικό μοντέλο $AR(p)$ αποτελεί μια παλινδρόμηση της X_t με τις $X_{t-1}, X_{t-2,...,}X_{t-p}$, δηλαδή της τρέχουσας τιμής της χρονοσειράς με τις p προηγούμενες.

Το μοντέλο $AR(p)$ μπορεί να εκφραστεί και μέσω του οπισθοδρομικού τελεστή που αναφέραμε λίγο παραπάνω ως εξής:

$$\Phi(B)X_t = \delta + Z_t \quad \text{όπου} \quad \Phi(B) = 1-\varphi_1 B-\varphi_2 B^2-...-\varphi_p B^p$$

Με βάση την παραπάνω έκφραση, η διαδικασία $AR(p)$ είναι στάσιμη όταν οι ρίζες του χαρακτηριστικού πολυωνύμου

$$\Phi(B) = 1-\varphi_1 B-\varphi_2 B^2-...-\varphi_p B^p$$

βρίσκονται εκτός του μοναδιαίου κύκλου, ή αντίστοιχα με βάση την έκφραση (2.2), οι ρίζες της εξίσωσης

$$m^p-\varphi_1 m^{p-1}-\varphi_2 m^{p-2}-...-\varphi_p = 0$$

είναι μικρότερες της μονάδας σε απόλυτη τιμή.

Μοντέλα κινητού μέσου- $MA(q)$

Η δεύτερη κατηγορία είναι αυτή των μοντέλων κινητού μέσου τα οποία προκύπτουν περιορίζοντας τους όρους του λευκού θορύβου στους q πιο πρόσφατους (*Ιστοσελίδα Δ. Κουγιουμτζή, Αριστοτέλειο Πανεπιστήμιο*):

$$X_t = \mu + Z_t - \theta_1 Z_{t-1} - \theta_2 Z_{t-2} - ... - \theta_q Z_{t-q} \quad , \quad Z_t \sim WN(0,\sigma^2)$$

Ενώ η αντίστοιχη έκφραση με τη χρήση του οπισθοδρομικού τελεστή είναι:

$$X_t = \mu + \theta(B)Z_t$$

Σε αντίθεση με τα μοντέλα $AR(p)$, τα μοντέλα $MA(q)$ είναι πάντα στάσιμα ως πεπερασμένο άθροισμα ορών λευκού θορύβου. Επίσης σε αυτήν την περίπτωση, η X_t ορίζεται ως ένας γραμμικός συνδυασμός των πιο πρόσφατων τιμών q του λευκού θορύβου $Z_{t-1},...,Z_{t-q}$ χωρίς να δίνεται κάποια άλλη πληροφορία για την X_t πέραν από τις τυχαίες διαταράξεις των $q+1$ τελευταίων χρονικών στιγμών, γι'αυτό και η προσαρμογή των παραμέτρων στο μοντέλο MA είναι πιο περίπλοκη απ'ότι στο μοντέλο AR.

Στην περίπτωση των μοντέλων $MA(q)$ δε μας απασχολεί η έννοια της στασιμότητας από τη στιγμή που είναι εξασφαλισμένη, μας αφορά όμως η έννοια της αντιστρεψιμότητας, αν δηλαδή μπορούμε να εκφράσουμε τον θόρυβο Z_t γνωρίζοντας την X_t και όλες τις προηγούμενες τιμές της. Αυτό μπορεί να συμβεί αν το πολυώνυμο $\theta(B)$ είναι αντιστρέψιμο, έτσι ώστε $Z_t = \theta^{-1}(B)X_t - \mu$, οπότε σε αντιστοιχία με τα μοντέλα $AR(p)$, οι ρίζες του $\theta(B)$ θα πρέπει να βρίσκονται εκτός του μοναδιαίου κύκλου.

Μικτά Μοντέλα- $ARMA(p,q)$

Τα μοντέλα $ARMA$ (Αυτοπαλινδρομούμενα μοντέλα κινούμενου μέσου), αποτελούν τη σύνθεση των μοντέλων AR και MA:

$$X_t = \delta + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots + \varphi_p X_{t-p} + Z_t - \theta_1 Z_{t-1} - \theta_2 Z_{t-2} - \ldots - \theta_q Z_{t-q}$$

Και με βάση τους οπισθοδρομικούς τελεστές έχουμε την ισοδύναμη έκφραση:

$$\Phi(B)X_t = \delta + \Theta(B)Z_t \text{ , όπου } Z_t \sim WN(0,\sigma^2)$$

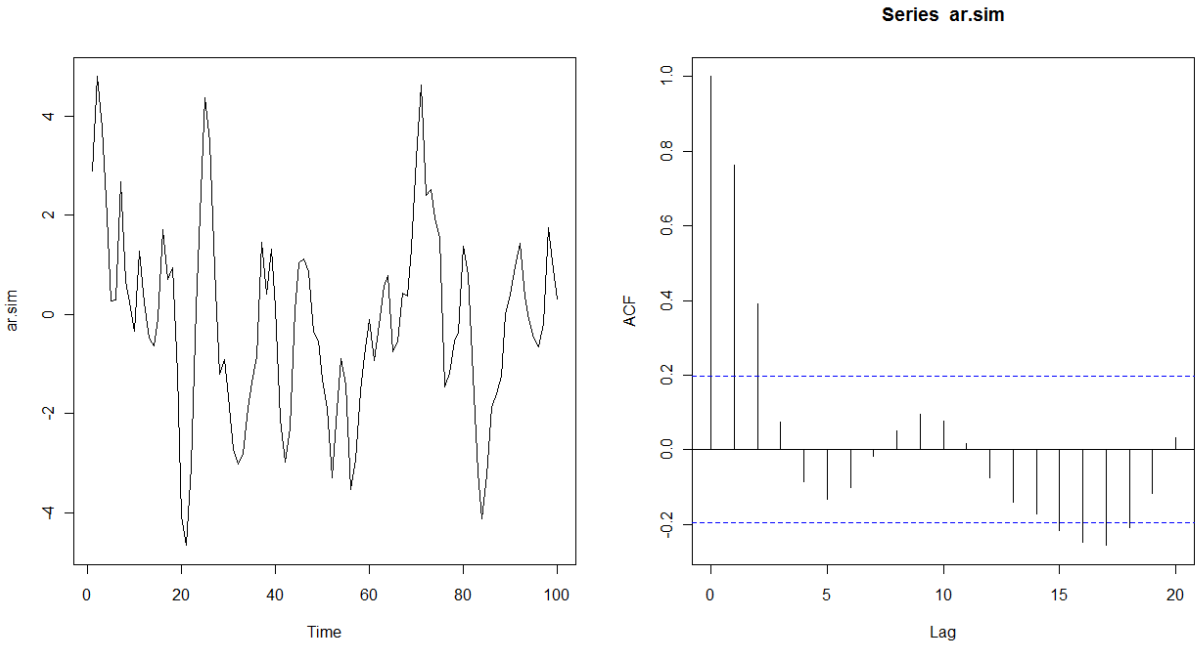
Οι διαδικασίες $ARMA(p,q)$ περιέχουν και το μέρος AR τάξης p και το μέρος MA τάξης q, με αποτέλεσμα η στασιμότητα της χρονοσειράς να καθορίζεται από το AR μέρος και τις ρίζες του πολυωνύμου $\varphi(B)$, ενώ η αντιστρεψιμότητα από το MA μέρος και τις ρίζες του πολυωνύμου $\theta(B)$.

Εκτός από τις τρεις αυτές βασικές κατηγορίες μοντέλων υπάρχουν και άλλα, τα οποία όμως βασίζονται στα παραπάνω, όπως τα *ARIMA (Autoregressive integrated moving average* - που είναι κατάλληλα για αρχικά μη στάσιμες χρονοσειρές), τα *GARCH (Generalized autoregressive conditional heteroscedastic)* ή τα *GARMA (Generalized autoregressive moving average).*

Κλείνοντας αυτή την παράγραφο να αναφέρουμε επίσης ότι η επιλογή του κατάλληλου μοντέλου (AR, MA, $ARMA$) καθορίζεται κατά ένα μεγάλο βαθμό από τη μορφή της συνάρτησης αυτοσυσχέτισης (ACF), που διαφέρει σε κάθε περίπτωση. Παρακάτω, στα Γραφήματα 1.3, 1.4, 1.5 βλέπουμε τις γραφικές αναπαραστάσεις των μοντέλων $AR(2), MA(2), ARMA(2,2)$ καθώς και το γράφημα της συνάρτησης αυτοσυσχέτισης για την κάθε περίπτωση. Μέσω της R και των παρακάτω εντολών, προσομοιώσαμε 100 τιμές από κάθε διαδικασία $AR, MA, ARMA$ τάξης 2, δηλαδή $p = 2, q = 2$ αντίστοιχα.

Για το $AR(2)$:

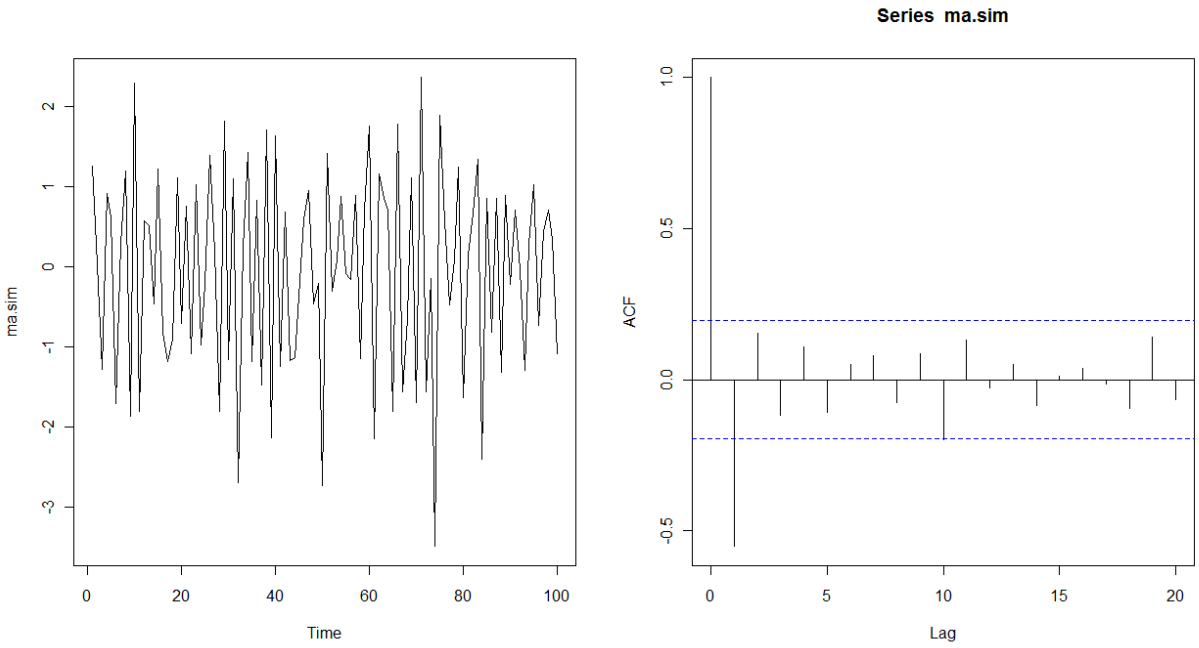
```
>ar.sim<-arima.sim(model=list(ar=c(.9,-.2)),n=100)
> ts.plot(ar.sim)
> ar.acf<-acf(ar.sim,type="correlation",plot=T)
```



Γράφημα 2.3: Γράφημα διαδικασίας AR(2) και γράφημα συνάρτησης αυτοσυσχέτισης

Για το *MA*(2):

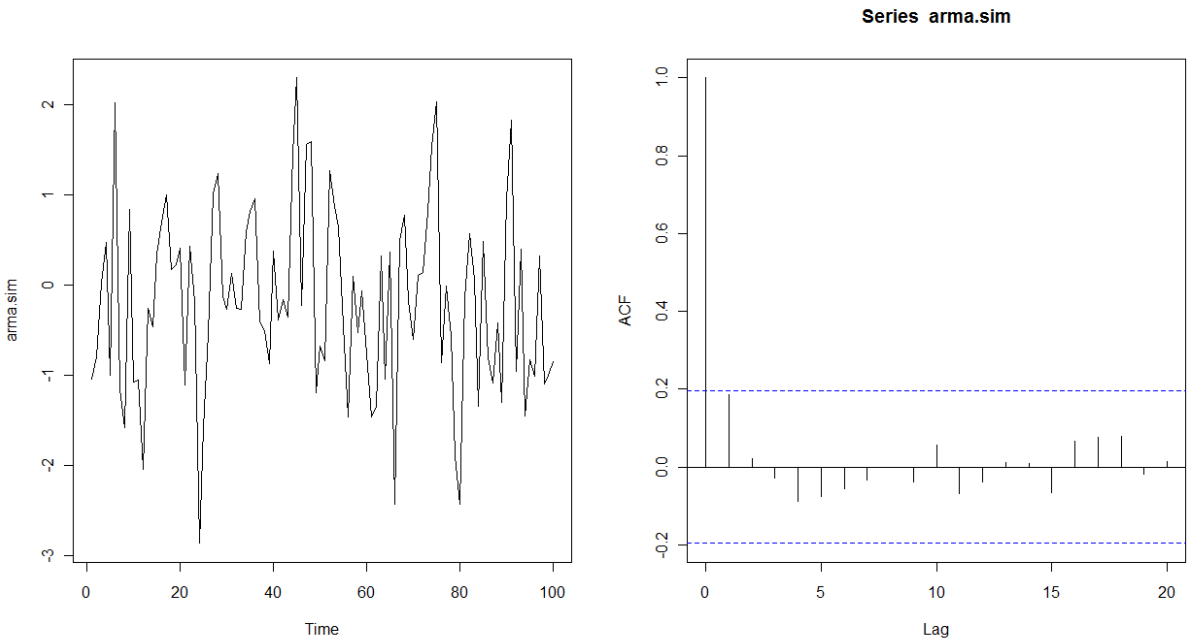
```
>ma.sim<-arima.sim(model=list(ma=c(-.7,.1)),n=100)
> ts.plot(ma.sim)
> ma.acf<-acf(ma.sim,type="correlation",plot=T)
```



Γράφημα 2.4: Γράφημα διαδικασίας MA(2) και γράφημα συνάρτησης αυτοσυσχέτισης

Για το $ARMA(2,2)$:

```
> arma.sim<-arima.sim(model=list(ar=c(.9,-.2),ma=c(-.7,.1)),n=100)
> ts.plot(arma.sim)
> arma.acf<-acf(arma.sim,type="correlation",plot=T)
```



Γράφημα 2.5: Γράφημα διαδικασίας ARMA(2,2) και γράφημα συνάρτησης αυτοσυσχέτισης

Από τις γραφικές αναπαραστάσεις 2.3, 2.4, και 2.5 που προέκυψαν για τα τρία διαφορετικά μοντέλα, παρατηρούμε ότι υπάρχουν εμφανείς διαφορές μεταξύ των γραφημάτων αυτοσυσχέτισης και ειδικότερα μεταξύ του μοντέλου AR και MA.

2.3 ΠΑΡΑΔΕΙΓΜΑ ΜΟΝΤΕΛΟΠΟΙΗΣΗΣ ΧΡΟΝΟΣΕΙΡΑΣ

Για να κατανοήσουμε το πως χρησιμοποιούνται στην πράξη, οι βασικές έννοιες της ανάλυσης χρονοσειρών που αναφέραμε σε αυτό το κεφάλαιο, και να δούμε σε γενικές γραμμές τη διαδικασία που ακολουθείται για τη μοντελοποίηση μιας χρονοσειράς, θα παραθέσουμε ένα μικρό παράδειγμα, βασισμένο σε αυτό των «Κ.Φωκιανός-Χ.Χαραλάμπους» *(Ιστοσελίδα Κ.Φωκιανού, Πανεπιστήμιο Κύπρου).*

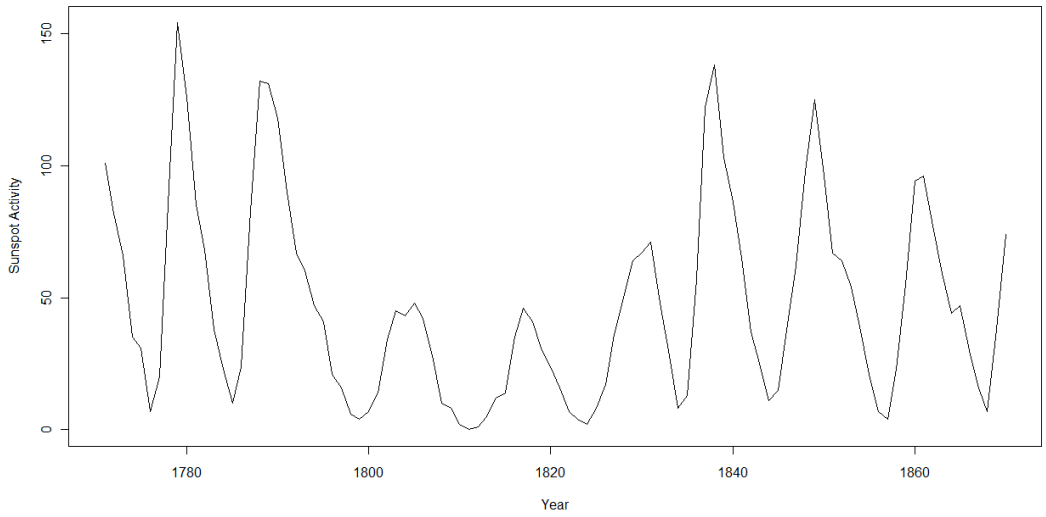
Τα δεδομένα μας έχουν να κάνουν με τον αριθμό ηλιακών κηλίδων από το 1771 έως το 1870. Εισάγουμε τα δεδομένα στην R, και μέσω της παρακάτω εντολής "κατασκευάζουμε" ένα αντικείμενο χρονοσειράς (time series-ts) με τιμές τα δεδομένα μας, με αρχή το έτος 1771, τέλος το 1870 και συχνότητα ίση με 1:

```
> sunspot<-ts(data,start=1771,end=1870,frequency=1)
```

```
> sunspot
Time Series:
Start = 1771
End = 1870
Frequency = 1
  [1] 101  82  66  35  31   7  20  92 154 126  85  68  38  23  10  24  83 132
 [19] 131 118  90  67  60  47  41  21  16   6   4   7  14  34  45  43  48  42
 [37]  28  10   8   2   0   1   5  12  14  35  46  41  30  24  16   7   4   2
 [55]   8  17  36  50  64  67  71  48  28   8  13  57 122 138 103  86  65  37
 [73]  24  11  15  40  62  98 125  96  67  64  54  39  21   7   4  23  55  94
 [91]  96  77  59  44  47  30  16   7  38  74
```

Όπως αναφέραμε και παραπάνω, το πρώτο και βασικό βήμα για την ανάλυση και μοντελοποίηση μιας χρονοσειράς είναι η γραφική αναπαράσταση της, ώστε να αποκτήσουμε μια πρώτη εικόνα περί στασιμότητας, τάσης ή περιοδικότητας. Με την παρακάτω εντολή έχουμε το γράφημα της χρονοσειράς *sunspot*:

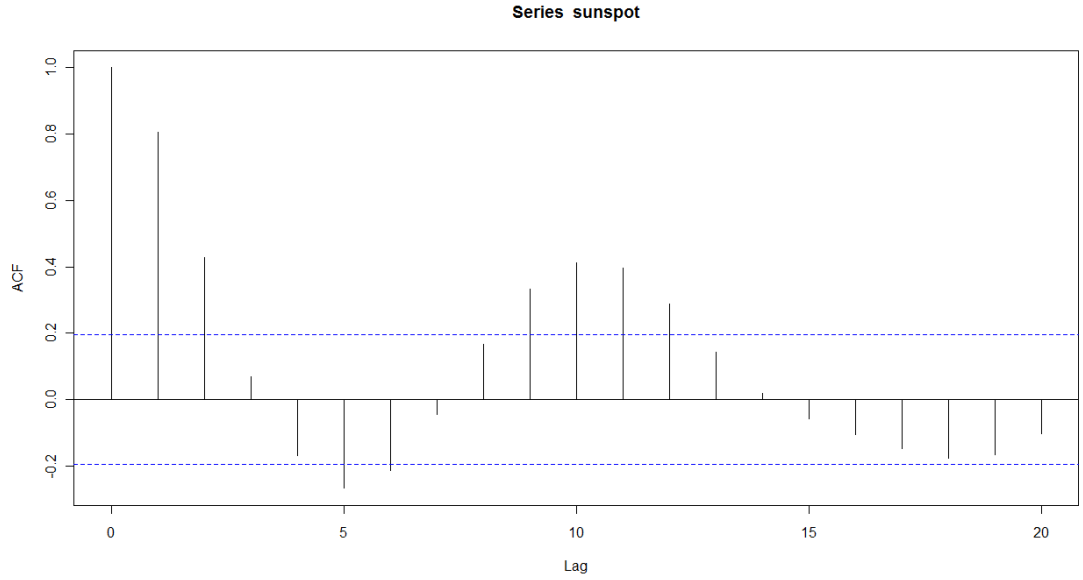
```
> ts.plot(sunspot,xlab="Year",ylab="Sunspot Activity")
```



Παρατηρώντας το γράφημα, φαίνεται να υπάρχει μια περιοδικότητα στα δεδομένα μας αλλά σίγουρα δεν παρατηρείται κάποια συγκεκριμένη τάση. Από τη στιγμή που παρατηρείται κάποια περιοδικότητα, καταλαβαίνουμε από το γράφημα και μόνο, ότι δεν πρόκειται για μια στάσιμη χρονοσειρά, και πριν την όποια προσπάθεια μοντελοποίησης της θα πρέπει να εξαλείψουμε με κάποιο τρόπο την περιοδικότητα.

Δεύτερο βήμα της διαδικασίας, αποτελεί το γράφημα της συνάρτησης αυτοσυσχέτισης (ACF), η οποία όπως εξηγήσαμε, αποτελεί βασικό στοιχείο της περιγραφής μιας χρονοσειράς, καθώς επίσης συντελεί και στην επιλογή του κατάλληλου μοντέλου προσαρμογής της. Εισάγοντας την παρακάτω εντολή προκύπτει το γράφημα της συνάρτησης αυτοσυσχέτισης για την χρονοσειρά *sunspot* έχοντας ορίσει ως μέγιστο αριθμό υστερήσεων *(lags)*, για τις οποίες θα γίνει η γραφική παράσταση, τις 20 (ο αριθμός 20, επιλέχθηκε γιατί μετά από δοκιμές, μας έδινε μια αντιπροσωπευτική εικόνα της χρονοσειράς):

>acf(sunspot,lag.max=20,type="correlation")



Το γράφημα της αυτοσυχέτισης παρουσιάζει μια ημιτονοειδής μορφή, πράγμα που επιβεβαιώνει την περιοδικότητα που είχε παρατηρηθεί από πριν, ενώ η περίοδος φαίνεται να είναι τα 10 με 11 χρόνια. Οι δύο οριζόντιες γραμμές που παρατηρούμε στο γράφημα, μας δείχνουν το 95% διάστημα εμπιστοσύνης για τον έλεγχο της υπόθεσης $H_0 : \rho = 0$ (η αυτοσυσχέτιση να είναι μηδενική).

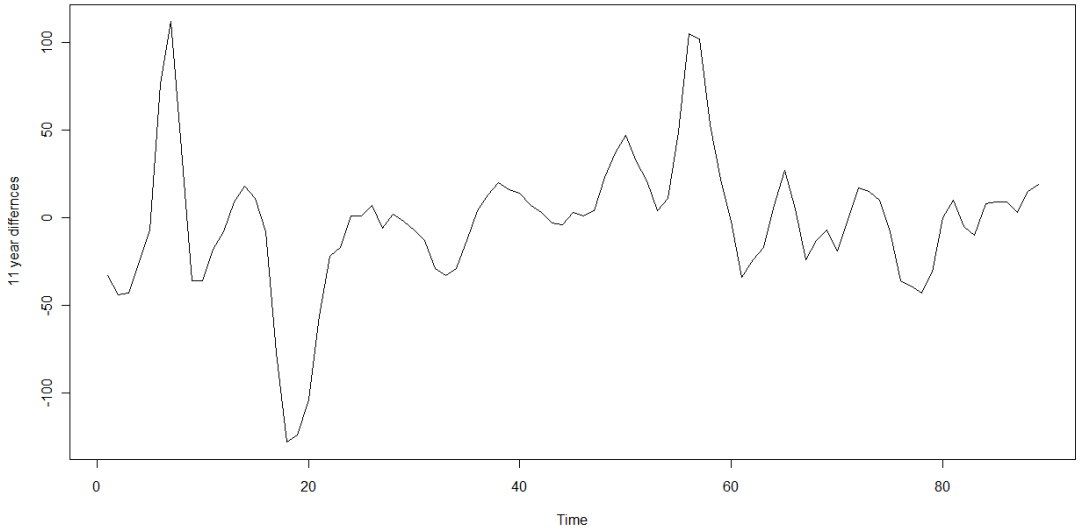
Στο επόμενο βήμα, πριν την προσαρμογή κάποιου μοντέλου στη χρονοσειρά μας, θα πρέπει να γίνει απαλοιφή της περιοδικότητας που παρατηρήθηκε. Όπως αναφέραμε και παραπάνω, υπάρχουν πολλοί τρόποι για να γίνει αυτό, αλλά εμείς θα χρησιμοποιήσουμε τη μέθοδο των διαφορών. Κατά τη μέθοδο αυτή, επειδή στις περιοδικές χρονοσειρές ισχύει η σχέση $s_t = s_{t+d} \quad t = 1, 2, \ldots$, όπου d η περίοδος, λαμβάνουμε τις διαφορές $X_t - X_{t-d}$ για $t = d, d+1, \ldots, n$ και έτσι κατασκευάζουμε ουσιαστικά μια νέα χρονοσειρά, όπου η συνιστώσα της περιοδικότητας s_t εξαλείφεται.

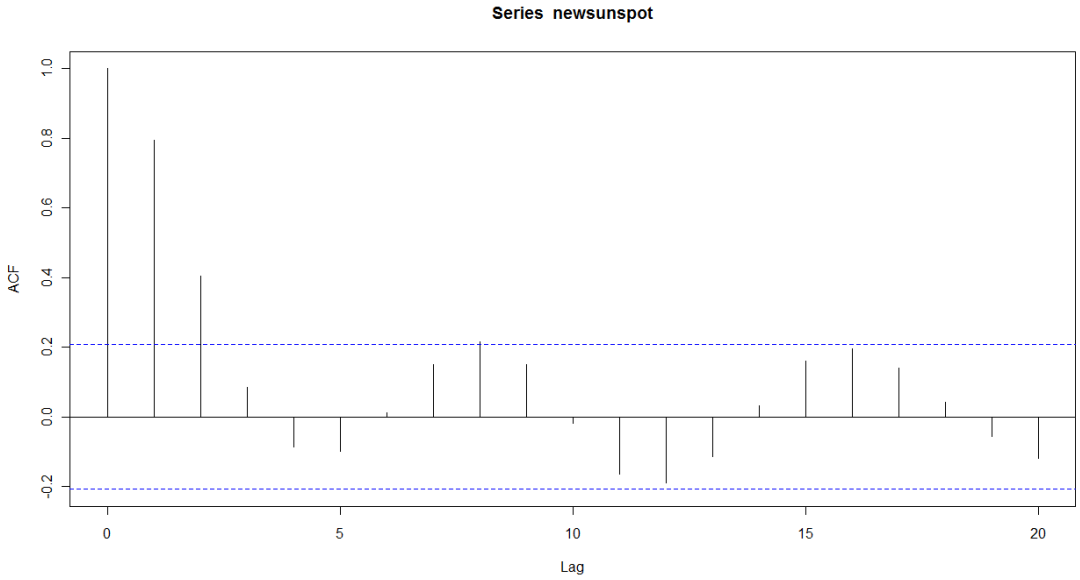
Υποθέτοντας λοιπόν, με βάση το γράφημα της αυτοσυσχέτισης, ότι έχουμε μια περιοδικότητα 11 χρόνων, ορίζουμε τη νέα χρονοσειρά:

$$y_t = x_t - x_{t-11}$$

Για την κατασκευή της νέας αυτής χρονοσειράς *newsunspot* μέσω της R, εισάγουμε τις παρακάτω εντολές, μέσω των οποίων προκύπτει η γραφική αναπαράσταση της χρονοσειράς και το γράφημα της συνάρτησης αυτοσυσχέτισης της:

```
> newsunspot<-sunspot[12:100]-sunspot[1:89]
> ts.plot(newsunspot,xlab="Time",ylab="11 year differnces")
> acf(newsunspot,lag.max=20)
```





Και από τα δύο γραφήματα της νέας χρονοσειράς που προέκυψαν, συμπεραίνουμε πως η περιοδικότητα έχει μειωθεί αισθητά σε σχέση με την αρχική. Στο γράφημα της συνάρτησης αυτοσυσχέτισης, παρατηρείται μεν μια ημιτονοειδής μορφή αλλά όχι σε τόσο μεγάλο βαθμό όπως πριν.

Τελευταίο βήμα, αποτελεί η μοντελοποίηση της καινούργιας πλέον χρονοσειράς *newsunspot*. Με βάση το γράφημα της συνάρτησης αυτοσυσχέτισης, ένα κατάλληλο μοντέλο που θα μπορούσαμε να εφαρμόσουμε στη χρονοσειρά, φαίνεται να είναι ένα μοντέλο αυτοπαλινδρόμησης AR. Χρησιμοποιώντας την παρακάτω εντολή προκύπτει το αυτοπαλινδρομούμενο μοντέλο για την χρονοσειρά μας:

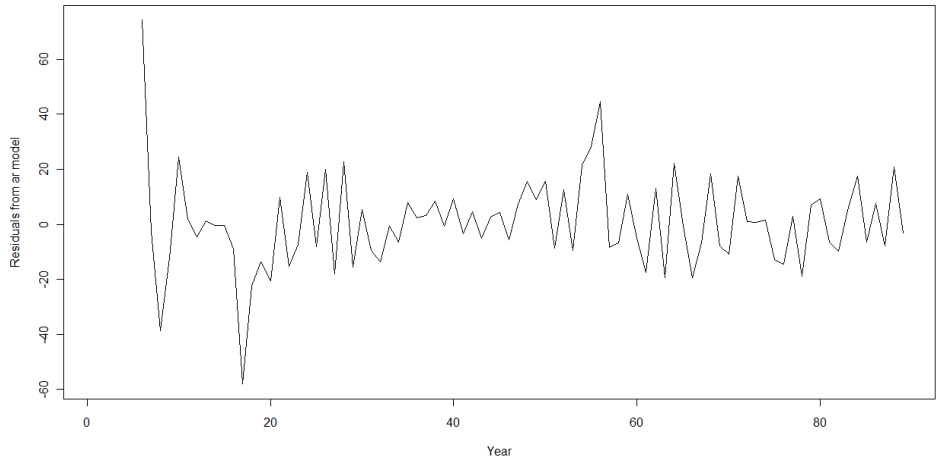
```
> sunspot.ar<-ar(newsunspot)

> sunspot.ar
Call:
ar(x = newsunspot)
```

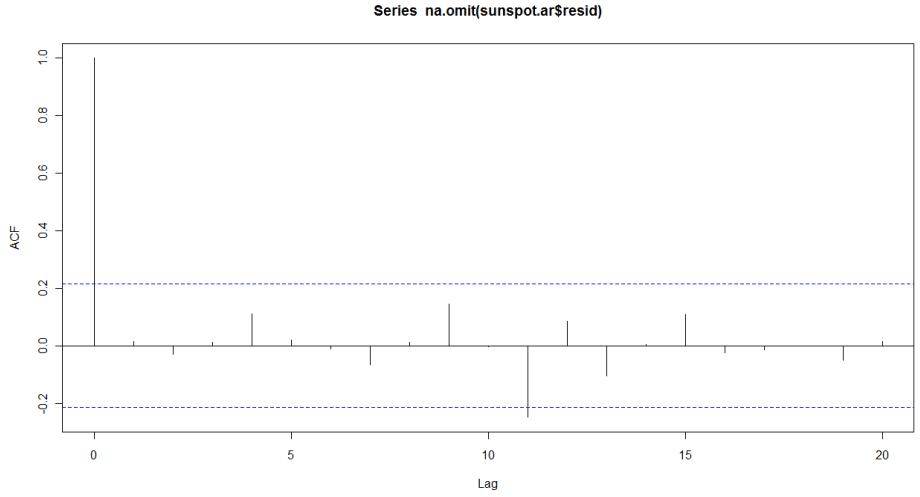
```
Coefficients:
      1        2        3        4        5
 1.4985  -1.1434   0.6369  -0.4126   0.1961
Order selected 5  sigma^2 estimated as  313.4
```

Με βάση τα αποτελέσματα του μοντέλου που προέκυψε, συμπεραίνουμε ότι η τάξη p του μοντέλου AR που προσαρμόζεται καλύτερα στα δεδομένα μας, εκτιμήθηκε να είναι $p=5$. Από τα αποτελέσματα μπορούμε επίσης να δούμε τους εκτιμώμενους συντελεστές και τη διασπορά του σφάλματος του μοντέλου.

Για τον έλεγχο της καταλληλότητας του μοντέλου θα χρησιμοποιήσουμε το γράφημα των υπολοίπων καθώς και το γράφημα της συνάρτησης αυτοσυσχέτισης των υπολοίπων. Όπως αναφέραμε και στην αρχή του κεφαλαίου, εάν το μοντέλο εφαρμόζει ικανοποιητικά στα δεδομένα μας, τότε τα υπόλοιπα θα έχουν τη μορφή χρονοσειράς λευκού θορύβου. Εισάγοντας λοιπόν τις παρακάτω εντολές στην R, έχουμε το γράφημα των υπολοίπων και το γράφημα της συνάρτησης αυτοσυσχέτισης τους, αντίστοιχα:

```
> ts.plot(sunspot.ar$resid,xlab="Year",ylab="Residuals from ar model")
> acf(na.omit(sun.ar$resid),lag.max=20)
```

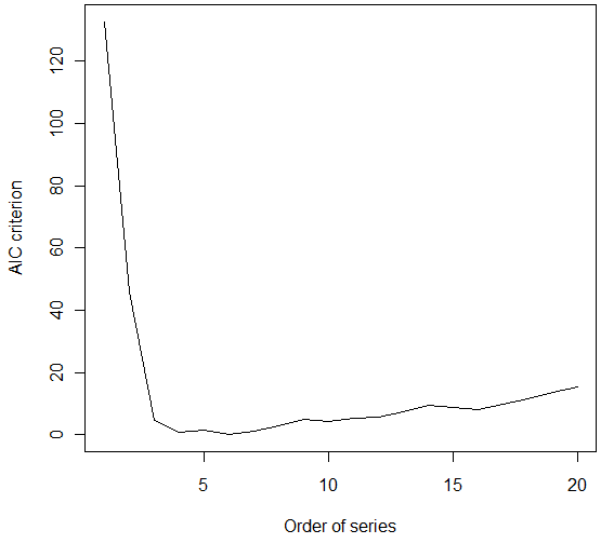




Από τις γραφικές παραστάσεις που προκύπτουν και βλέπουμε παραπάνω, το μοντέλο τελικά φαίνεται να είναι κατάλληλο, καθώς τα υπόλοιπα δε παρουσιάζουν κάποια προφανή δομή αλλά έχουν τη μορφή μιας τυχαίας σειράς. Επίσης, το ίδιο συμπεραίνουμε και από το γράφημα της συνάρτησης αυτοσυσχέτισης, το οποίο δεν παρουσιάζει καμία ημιτονοειδή μορφή, ενώ σχεδόν για όλες τις υστερήσεις οι τιμές της αυτοσυσχέτισης βρίσκονται εντός των ορίων του 95% διαστήματος εμπιστοσύνης, που σχηματίζουν οι δύο οριζόντιες γραμμές του γραφήματος.

Κλείνοντας, ένας άλλος έλεγχος, αποτελεί και το γνωστό από το προηγούμενο κεφάλαιο, κριτήριο *AIC*, μέσω του οποίου μπορούμε να ελέγξουμε την τάξη του μοντέλου μας. Εάν η τάξη που έχει ήδη εκτιμηθεί από το AR μοντέλο που προσαρμόσαμε μέσω της R, ως $p=5$, είναι σωστή, τότε θα πρέπει να ελαχιστοποιεί και το κριτήριο *AIC*. Εισάγοντας την παρακάτω εντολή, έχουμε το αντίστοιχο γράφημα του κριτηρίου:

```
> ts.plot(sunspot.ar$aic,xlab="Order of series",ylab="AIC criterion")
```



Σύμφωνα με το γράφημα, το AIC ελαχιστοποιείται στην τιμή 6, αλλά επειδή η γραφική παράσταση του κριτηρίου ξεκινάει από το σημείο 1, που αντιστοιχεί στο μοντέλο τάξης 0, συμπεραίνουμε πως η εκτιμώμενη τάξη 5, του αυτοπαλινδρομούμενου μας μοντέλου, όντως επιβεβαιώνεται ως η καταλληλότερη.

ΚΕΦΑΛΑΙΟ 3

ΧΡΟΝΟΣΕΙΡΕΣ ΚΑΙ ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

3.1 ΕΙΣΑΓΩΓΗ

Στα προηγούμενα δύο κεφάλαια, αναλύσαμε αρχικά, τα γενικευμένα γραμμικά μοντέλα, με βάση τα οποία προκύπτουν δύο σημαντικά είδη παλινδρόμησης, αυτό της λογιστικής και αυτό της Poisson. Στη συνέχεια, περιγράφηκαν οι πολύ σημαντικές έννοιες για την κατανόηση της ανάλυσης χρονοσειρών καθώς και οι βασικές διαδικασίες για τη μοντελοποίηση μιας χρονοσειράς. Αφού λοιπόν έχουμε καθορίσει το απαραίτητο υπόβαθρο, είμαστε σε θέση να προχωρήσουμε στο ερώτημα που προκύπτει, για το αν μπορούμε να χρησιμοποιήσουμε τα γενικευμένα γραμμικά μοντέλα, για τη μοντελοποίηση χρονοσειρών.

Η αναγκαιότητα του συγκεκριμένου είδους μοντελοποίησης έγκειται στους ίδιους λόγους οι οποίοι ισχύουν και για τα απλά δεδομένα, δηλαδή στο ότι υπάρχουν περιπτώσεις, όπου οι παρατηρήσεις των χρονοσειρών είναι μη κανονικές, όπως είναι για παράδειγμα τα δυαδικά ή τα απαριθμητά δεδομένα. Το εμπόδιο που προκύπτει στην εφαρμογή της μεθοδολογίας των γενικευμένων γραμμικών μοντέλων στις χρονοσειρές, είναι ότι στην περίπτωση των χρονοσειρών, τα δεδομένα των μεταβλητών είναι χρονικά εξαρτημένα.

Η δυσκολία αυτή, ξεπερνάται με τη χρήση της *μερικής πιθανοφάνειας (partial likelihood),* με τη βοήθεια της οποίας μπορούμε να επεκτείνουμε τα χαρακτηριστικά των γενικευμένων γραμμικών μοντέλων, τα οποία είναι κατάλληλα για ανεξάρτητα δεδομένα, στις εξαρτημένες παρατηρήσεις των χρονοσειρών, χωρίς να απαιτείται η στασιμότητα. Μέσω της μερικής πιθανοφάνειας, επιτυγχάνεται μια ευέλικτη και συνεχή συμπερασματολογία, βασισμένη στα όσα είναι γνωστά κατά την περίοδο της παρατήρησης μιας χρονοσειράς, επιτρέποντας την παρουσία αυτοπαλινδρομούμενων μερών και αλληλεπιδράσεων μεταξύ των συμμεταβλητών, κάτι που δε θα ήταν εφικτό διαφορετικά. (Kedem & Fokianos, 2002)

Εμάς, θα μας απασχολήσουν τα δύο συγκεκριμένα μοντέλα που ήδη έχουμε περιγράψει, το λογιστικό και το Poisson. Θα δούμε δηλαδή τα συγκεκριμένα είδη παλινδρομήσεων, ως αυτοπαλινδρομούμενα μοντέλα για τα αντίστοιχα είδη χρονοσειρών, δηλαδή για δυαδικές και απαριθμητές χρονοσειρές αντίστοιχα. Πριν όμως προχωρήσουμε στην ανάλυση των δύο αυτών μοντέλων, θα πρέπει να κάνουμε μια επισκόπηση στο γενικότερο υπόβαθρο, της θεωρίας και της μεθοδολογίας, στο οποίο στηρίζονται οι χρονοσειρές που ακολουθούν γενικευμένα γραμμικά μοντέλα. Στις επόμενες παραγράφους λοιπόν, θα παρουσιάσουμε την έννοια της μερικής πιθανοφάνειας, τη δομή και τη μορφή του γενικευμένου γραμμικού μοντέλου στην περίπτωση της χρονοσειράς, καθώς και κάποιους βασικούς διαγνωστικούς ελέγχους που χρησιμοποιούνται για την προσαρμογή των συγκεκριμένων μοντέλων.

3.2 Η ΕΝΝΟΙΑ ΤΗΣ ΜΕΡΙΚΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

3.2.1 Γενικά στοιχεία

Η θεωρία της μερικής πιθανοφάνειας παίζει καθοριστικό ρόλο στη μοντελοποίηση χρονοσειρών διακριτών δεδομένων, είτε αυτά είναι δυαδικά (0 και 1), είτε είναι απαριθμητά. Αρχικά, η προσέγγιση αυτού του είδους χρονοσειρών, έγινε μέσω Μαρκοβιανών ανελίξεων και συγκεκριμένα μέσω ομογενών Μαρκοβιανών αλυσίδων. Αυτός όμως ο τρόπος ανάλυσης αντιμετώπιζε διάφορα προβλήματα, όπως το ότι χωρίς επιπλέον περιορισμούς ο αριθμός των παραμέτρων αυξανόταν εκθετικά καθώς μεγάλωνε η τάξη της αλυσίδας (Fahrmeir & Tutz, 2001).

Τα προβλήματα αυτά που παρουσιάστηκαν στα Μαρκοβιανά μοντέλα, αντιμετωπίστηκαν σε μεγάλο βαθμό από τα μοντέλα παλινδρόμησης που αναπτύχθηκαν βασισμένα στη θεωρία των γενικευμένων γραμμικών μοντέλων. Η ανάπτυξη αυτών των παλινδρομικών μοντέλων στηρίχθηκε όπως ήδη αναφέραμε, στη θεωρία της μερικής πιθανοφάνειας, η οποία αποτελεί μια σημαντική μέθοδο συμπερασματολογίας για εξαρτημένα δεδομένα αλλά και μια προσέγγιση για τη μοντελοποίηση στοχαστικών διαδικασιών χωρίς να απαιτείται η στασιμότητα ή η μαρκοβιανή ιδιότητα. Να αναφέρουμε, ότι μια στοχαστική διαδικασία έχει τη *μαρκοβιανή ιδιότητα* όταν η δεσμευμένη κατανομή πιθανότητας των μελλοντικών τιμών της διαδικασίας, εξαρτάται μόνο από την παρούσα κατάσταση και όχι από την ακολουθία των γεγονότων που προηγήθηκαν. (Kedem & Fokianos, 2002)

Η μερική πιθανοφάνεια αποτελεί μια γενίκευση της συνάρτησης πιθανοφάνειας και της δεσμευμένης πιθανοφάνειας. Συγκεκριμένα, ο τύπος της μερικής πιθανοφάνειας απλοποιείται στον τύπο της "απλής" πιθανοφάνειας, όταν δεν υπάρχουν εξωγενείς μεταβλητές και τα δεδομένα είναι ανεξάρτητα. Ενώ παίρνει τη μορφή της δεσμευμένης πιθανοφάνειας, όταν οι συμμεταβλητές *(covariates)* είναι ντετερμινιστικές, είναι δηλαδή γνωστές καθ'όλη τη διάρκεια της παρατήρησης. Αντίθετα με τη δεσμευμένη, η μερική πιθανοφανεία στηρίζεται ουσιαστικά μόνο σε αυτά που είναι γνωστά στον παρατηρητή μέχρι και την χρονική στιγμή $t-1$ της παρατήρησης (δηλαδή στα γεγονότα που έχουν προηγηθεί) , χωρίς να χρειάζεται πλήρης γνώση της από κοινού κατανομής της μεταβλητής απόκρισης ή των συμμεταβλητών.

Στις δύο επόμενες ενότητες αυτής της παραγράφου, θα δούμε τον ορισμό της μερικής πιθανοφάνειας, στον οποίο θα καταλήξουμε με τη βοήθεια της δεσμευμένης πιθανοφάνειας *(conditional likelihood)* και της πλήρους πιθανοφάνειας *(full liklehood)* , καθώς και τη λειτουργικότητα της εν τέλει, στη μοντελοποίηση των χρονοσειρών.

3.2.2 Δεσμευμένη και Μερική Πιθανοφάνεια

Αρχικά, να υπενθυμίσουμε τον ορισμό της ***συνάρτησης πιθανοφάνειας***, η οποία αποτελεί θεμελιώδη έννοια της στατιστικής και σε αυτή όπως είναι φυσικό, βασίζονται και τα όσα θα αναλύσουμε παρακάτω.

Έστω $X=(X_1,....,X_n)$ δείγμα από έναν πληθυσμό με συνάρτηση πυκνότητας πιθανότητας $f(x;\theta)$. Τότε η συνάρτηση $L(\theta)=\prod_{i=1}^{n} f(x_i;\theta)$ της αγνώστου παραμέτρου θ, αποτελεί τη συνάρτηση πιθανοφάνειας και εκφράζει το πόσο πιθανοφανείς ή αλλιώς πόσο σύμφωνες με το συγκεκριμένο δείγμα , είναι οι διάφορες τιμές της παραμέτρου θ (Κοκολάκης & Φουσκάκης, 2009).

Όπως ήδη έχουμε εξηγήσει, μια χρονοσειρά αποτελεί την τροχιά μιας στοχαστικής ανέλιξης, η οποία ουσιαστικά αποτελεί ένα δείγμα ενός διαχρονικού στοχαστικού φαινομένου. Αν λοιπόν υποθέσουμε ότι έχουμε μια χρονοσειρά $\{Y_t\}$, $t=1,...,N$ τότε για να έχουμε το μέγιστο βαθμό στοχαστικής πληροφόρησης θα πρέπει να γνωρίζουμε την από κοινού κατανομή της:

$$f(y_1,y_2,...,y_N;\theta) \text{ ή αλλιώς } f_\theta(y_1,y_2,...,y_N)$$

Σε αυτή όμως την περίπτωση, η αναγωγή της πιθανοφάνειας στη γνωστή και εύχρηστη σχέση

$$f(y_1,y_2,...,y_N;\theta)=\prod_{t=1}^{N} f(y_t;\theta)$$

δεν είναι εφικτή λόγω της εξάρτησης των δεδομένων. Αρχικά, για την αντιμετώπιση αυτού του προβλήματος, χρησιμοποιήθηκε η μέθοδος της δεσμευμένης πιθανοφάνειας, ξεκινώντας με την ακόλουθη διάσπαση της συνάρτησης πιθανοφάνειας

$$f_\theta(y_1,...,y_N)=f_\theta(y_1)\prod_{t=2}^{N} f_\theta(y_t \mid y_1,y_2,...,y_{t-1}) \qquad (3.1)$$

Η σχέση όμως (3.1) εμφανίζει μια βασική δυσκολία, γιατί χωρίς να γίνει κάποια άλλη υπόθεση, η παραπάνω δεσμευμένη κατανομή, για κάθε χρονική στιγμή εξαρτάται από ένα διαφορετικό σύνολο πληροφοριών το οποίο όσο εξελίσσεται το φαινόμενο, αυτό διευρύνεται. Αυτό συμβαίνει γιατί, όσο αυξάνεται το μέγεθος N της χρονοσειράς, τόσο αυξάνεται και το πλήθος των παραμέτρων θ. Έτσι, αντί να λαμβάνουμε περισσότερη πληροφορία για ένα σύνολο σταθερών παραμέτρων, δεχόμαστε πληροφορία για ένα αυξανόμενο αριθμό παραμέτρων, πράγμα που δυσκολεύει τη μοντελοποίηση των προβλημάτων.

Για την αντιμετώπιση αυτού του προβλήματος, θα βοηθήσει η παρακάτω υπόθεση. Αν υποθέσουμε πως η χρονοσειρά Y_t είναι μια τάξης 1, Μακοβιανή και στάσιμη στοχαστική διαδικασία, τότε η σχέση (3.1) μπορεί να απλοποιηθεί και έτσι η από κοινού κατανομή να γραφεί ως εξής:

$$\boxed{f_\theta(y_1,...,y_N)=f_\theta(y_1)\prod_{t=2}^{N} f_\theta(y_t \mid y_{t-1})} \qquad (3.2)$$

Αγνοώντας τον όρο $f_\theta(y_1)$ αφού είναι ανεξάρτητη του N, η μοντελοποίηση της Y_t και η συμπερασματολογία όσον αφορά τη μεταβλητή θ, μπορεί να στηριχτεί στο γινόμενο της σχέσης (3.2), δηλαδή στη δεσμευμένη κατανομή $f_\theta(y_t \mid y_{t-1})$. Αυτό,

αποτελεί παράδειγμα ***δεσμευμένης πιθανοφάνειας*** η οποία προκύπτει από εξαρτημένες παρατηρήσεις, εκφρασμένες ως γινόμενο δεσμευμένων κατανομών.

Έχοντας εξηγήσει την έννοια της δεσμευμένης πιθανοφάνειας, μπορούμε να καταλάβουμε ότι παρότι λύνει πολλά από τα προβλήματα, (όπως το ότι πλέον η μεταβλητή θ είναι σταθερή και ανεξάρτητη του μεγέθους N), δεσμεύει το πρόβλημα, από τη στιγμή που προϋποθέτει τη μαρκοβιανή ιδιότητα και τη στασιμότητα, που στην πραγματικότητα, τις περισσότερες φορές δεν μπορούν να εξασφαλιστούν. Η μέθοδος της μερικής πιθανοφάνειας, βασίζεται στην ιδέα που εισήγαγε ο Cox (Cox, 1975), να χρησιμοποιηθεί ένα μόνο μέρος της ολοκληρωμένης σχέσης (3.1) της πιθανοφάνειας, ως μια συνηθισμένη πιθανοφάνεια. Η αξία της ιδέας αυτής, έγκειται ακριβώς στο γεγονός ότι δεν απαιτεί την ισχύ των προαναφερθέντων υποθέσεων.

Έτσι, εάν υποθέσουμε πως έχουμε δύο χρονοσειρές, Y_t και X_t με $t=1,...,N$, όπου η $\{Y_t\}$ αποτελεί τη μεταβλητή απόκρισης και η $\{X_t\}$ μια εξαρτημένη μεταβλητή, τότε με βάση τη σχέση (3.1) η από κοινού κατανομή όλων των παρατηρήσεων μπορεί να εκφραστεί ως

$$f_\theta(x_1, y_1, ..., x_N, y_N) = f_\theta(x_1)\left[\prod_{t=2}^{N} f_\theta(x_t \mid d_t)\right]\left[\prod_{t=1}^{N} f_\theta(y_t \mid c_t)\right], \qquad (3.3)$$

όπου $d_t = (y_1, x_1, ..., y_{t-1}, x_{t-1})$ και $c_t = (y_1, x_1, ..., y_{t-1}, x_{t-1}, x_t)$.

Η σχέση (3.3) μπορεί να θεωρηθεί ως η ***πλήρης πιθανοφάνεια*** *(full likelihood)* του δείγματος. Το δεύτερο γινόμενο της σχέσης (3.3) συνιστά τη ***μερική πιθανοφάνεια*** του δείγματος, η οποία με κατάλληλες τροποποιήσεις που θα δούμε αμέσως μετά, καθίσταται κατάλληλη για συμπερασματολογία, ακόμα και με την απώλεια πληροφορίας για τη μεταβλητή θ που υπάρχει, παραλείποντας το πρώτο από τα δύο γινόμενα της σχέσης (Kedem & Fokianos, 2002).

3.2.3 Η μερική πιθανοφάνεια στο στατιστικό μοντέλο

Έχοντας εξηγήσει σύμφωνα με τα παραπάνω, το γενικό πλαίσιο στο οποίο βασίζεται η όλη "ιδέα" και η αναγκαιότητα της μερικής πιθανοφάνειας, μπορούμε να προχωρήσουμε στον ολοκληρωμένο ορισμό της έννοιας. Θα καταλήξουμε στη διατύπωση του ορισμού της μερικής πιθανοφάνειας, εξηγώντας τη λειτουργία της κατά τη μοντελοποίηση μιας χρονοσειράς μιας και η κατανόηση του ορισμού στηρίζεται σε έννοιες που εμπεριέχονται στη διαδικασία διαμόρφωσης ενός τέτοιου στατιστικού μοντέλου.

Ας συνεχίσουμε να βασιζόμαστε στην υπόθεση που ξεκινήσαμε από την προηγούμενη παράγραφο, ότι έχουμε δηλαδή δύο χρονοσειρές Y_t και X_t με $t=1,...,N$, όπου η $\{Y_t\}$ αποτελεί τη μεταβλητή απόκρισης και η $\{X_t\}$ μια εξαρτημένη μεταβλητή. Τότε το στατιστικό μοντέλο για την Y_t θα χτιστεί με βάση τις δεσμευμένες ροπές

$$E(Y_t \mid c_t) \text{ και } Var(Y_t \mid c_t)$$

όπου $c_t = (y_1, x_1, ..., y_{t-1}, x_{t-1}, x_t)$, όπως συμβολίσαμε και στη σχέση (3.3), και αποτελεί ουσιαστικά το σύνολο της πληροφορίας του συγκεκριμένου φαινομένου (Δρυμώνης, 2005).

Έστω λοιπόν ότι ισχύει:

$$E(Y_t \mid c_t) = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \beta_3 X_{t-1}$$
$$Var(Y_t \mid c_t) = \sigma^2$$

Τότε καταλήγουμε σε ένα μοντέλο παλινδρόμησης:

$$\begin{cases} Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \beta_3 X_{t-1} \\ Var(Y_t \mid c_t) = \sigma^2 \end{cases}$$

Οι παράμετροι του παλινδρομικού αυτού μοντέλου, θα εκτιμηθούν μέσω της συνάρτησης μερικής πιθανοφάνειας, η οποία με βάση το δεύτερο γινόμενο της σχέσης (3.3) και τα όσα εξηγήσαμε, μπορεί να γραφεί με την εξής μορφή:

$$PL(\theta; y_1, y_2, ..., y_N) = \prod_{t=1}^{N} f_\theta(y_t \mid c_t)$$

Η σχέση όμως αυτή, χωρίς κάποιες τροποποιήσεις, θα αντιμετώπιζε το ίδιο πρόβλημα με αυτό που είχαμε αναφέρει και παραπάνω στην περίπτωση της δεσμευμένης πιθανοφάνειας, δηλαδή όσο θα αυξανόταν το μέγεθος N της χρονοσειράς, θα υπήρχε και η αντίστοιχη αύξηση του πλήθους των παραμέτρων θ. Για την επίλυση αυτού του εμποδίου, τροποποιούμε τη δεσμευμένη κατανομή $f(y_t \mid c_t)$, $c_t = (y_1, x_1, ..., y_{t-1}, x_{t-1}, x_t)$ και την αντικαθιστούμε με την $f(y_t \mid F_{t-1})$.

Η F_{t-1} είναι η σ-άλγεβρα, η οποία αποτελείται από τις παρελθοντικές τιμές της μεταβλητής απόκρισης Y_t, και από τις παρελθοντικές ή ακόμα και από παροντικές τιμές (αν αυτές είναι γνωστές) των όποιων επεξηγηματικών μεταβλητών $X_t, W_t, ...$ κλπ, που υπάρχουν στο πρόβλημα. Με λίγα λόγια, η F_{t-1} περιέχει μέσα της οποιαδήποτε πληροφορία σχετικά με το φαινόμενο που είναι γνωστή στον παρατηρητή μέχρι και την χρονική στιγμή $t-1$.

Στην τυποποίηση των παραπάνω και στην έκφραση της F_{t-1} θα βοηθήσει το $p-$ διάστατο διάνυσμα

$$Z_{t-1} = (Z_{(t-1)_1}, ..., Z_{(t-1)_p})' \quad , \quad t = 1, ..., N$$

Το διάνυσμα Z_{t-1} περιέχει παρελθοντικές επεξηγηματικές μεταβλητές και θεωρούμε πως είναι αυτό, που κάθε χρονική στιγμή t, διαμορφώνει την τιμή της μεταβλητής Y_t. Ουσιαστικά στο Z_{t-1}, εμπεριέχεται η εξέλιξη του συστήματος όπως έχει διαμορφωθεί μέχρι το χρόνο $t-1$. Αποτελεί δηλαδή, μια διανυσματική στοχαστική ανέλιξη την οποία καλούμε *συμμεταβλητή διαδικασία (covariate process)*.

Μπορούμε λοιπόν να εκφράσουμε την σ-άλγεβρα F_{t-1} με την εξής γενική μορφή:

$$\boxed{F_{t-1} = \sigma\{Y_{t-1}, Y_{t-2}, ..., X_t, W_t, ..., Z_{t-1}, Z_{t-2}, ...\}} \quad (3.4)$$

Όπου οι Y_{t-1}, Y_{t-2} είναι η πρώτη και η δεύτερη υστέρηση της χρονοσειράς $\{Y_t\}$, ενώ οι X_t, W_t είναι επεξηγηματικές μεταβλητές που είναι γνωστές τη χρονική στιγμή $t-1$. Σε πολλές περιπτώσεις, είναι δυνατό και χρήσιμο στη μοντελοποίηση, να συμπεριλάβουμε τις επεξηγηματικές μεταβλητές καθώς επίσης και παρελθοντικές τιμές της μεταβλητής απόκρισης, μέσα στην $\{Z_{t-1}\}$. Για παράδειγμα, το διάνυσμα Z_{t-1} μπορεί να έχει τη μορφή $Z_{t-1} = (Y_{t-1}, Y_{t-2}, X_t, W_t)'$, όπου σε αυτήν την περίπτωση οι επεξηγηματικές μεταβλητές X_t, W_t είναι ντετερμινιστικές, οι τιμές τους δηλαδή για τη χρονική περίοδο t είναι ήδη γνωστές από την χρονική περίοδο $t-1$.

Έχοντας ορίσει πλέον την έννοια της F_{t-1}, μπορούμε να καταλάβουμε ότι κατά την παρατήρηση ενός φαινομένου, θα υπάρχει μια αύξουσα ακολουθία σ-αλγεβρών $F_0 \subset F_1 \subset F_2 ...$ για τους αντίστοιχους χρόνους $t = 0, 1, 2, ...$, αφού η πληροφορία κάθε χρονικής στιγμής περιέχει ουσιαστικά και όλη την προηγούμενη πληροφορία για την εξελικτική πορεία του φαινομένου, κατά τη διάρκεια του χρόνου. Οπότε, αν για παράδειγμα έχουμε την τιμή Y_{t-1}, που αντιστοιχεί προφανώς στο χρόνο $t-1$, εάν θέλουμε να προβλέψουμε την τιμή Y_t, τότε θα πρέπει να στηριχτούμε στη σ-άλγεβρα F_{t-1} που θα περιέχει όλη την ιστορία του φαινομένου μέχρι εκείνη τη στιγμή. Άρα λοιπόν, το στατιστικό μοντέλο που θα κατασκευάσουμε για τη χρονοσειρά $\{Y_t\}$ θα βασιστεί στις δεσμευμένες ροπές

$$\mu_t = E[Y_t \mid F_{t-1}] \text{ και } \sigma_t^2 = Var[Y_t \mid F_{t-1}],$$

οι οποίες κάθε χρονική στιγμή θα εξαρτώνται από τις όποιες μεταβλητές περιέχει το διάνυσμα Z_{t-1}. Όπως είναι επόμενο, αναλόγως με τις μεταβλητές αυτές που περιέχονται στο Z_{t-1}, καθορίζονται και οι αντίστοιχες παράμετροι του μοντέλου.

Έτσι λοιπόν, έχοντας εξηγήσει τον τρόπο με τον οποίο θα καταλήξουμε στο στατιστικό μοντέλο μιας χρονοσειράς, είμαστε σε θέση να δώσουμε τον ορισμό της μερικής πιθανοφάνειας, με τη χρήση της οποίας, θα εκτιμηθεί το σταθερό διάνυσμα β, στο οποίο περιλαμβάνονται οι παράμετροι του μοντέλου (στον παρακάτω γενικό ορισμό, το παραμετρικό διάνυσμα συμβολίζεται με θ αντί του β που έχουμε χρησιμοποιήσει εμείς).

Ορισμός μερικής πιθανοφάνειας:
Έστω F_{t-1}, $t = 1, 2, ...$, μια αύξουσα ακολουθία σ-αλγεβρών, $F_0 \subset F_1 \subset F_2 ...$ και έστω $Y_1, Y_2, ...$ μια ακολουθία τυχαίων μεταβλητών ορισμένων σε ένα κοινό χώρο πιθανότητας έτσι ώστε η Y_t να είναι F_t μετρήσιμη. Συμβολίζοντας τη συνάρτηση πυκνότητας πιθανότητας του Y_t δοθέντος F_{t-1}, με $f_t(y_t; \theta)$, όπου $\theta \in R^p$ είναι ένα

σταθερό διάνυσμα. Τότε η συνάρτηση PL που σχετίζεται με το θ, F_t και τα δεδομένα $Y_1, Y_2, ..., Y_N$ δίνεται από τον παρακάτω τύπο: (Kedem & Fokianos, 2002)

$$PL(\theta; y_1, y_2, ..., y_N) = \prod_{t=1}^{N} f_t(y_t; \theta)$$

Παρατηρώντας τον ορισμό και μόνο, καταλαβαίνουμε τα όσα ειπώθηκαν αρχικά για τη μερική πιθανοφάνεια και το ότι η συμπερασματολογία με βάση αυτήν, στηρίζεται αποκλειστικά σε ότι είναι γνωστό από το παρελθόν, εννοώντας παρελθοντικές τιμές της μεταβλητής απόκρισης και των συμμεταβλητών. Το διάνυσμα που μεγιστοποιεί την παραπάνω σχέση, καλείται εκτιμητής μέγιστης πιθανοφάνειας (MPLE) και είναι σημαντικό να αναφέρουμε ότι ικανοποιεί τις ίδιες χρήσιμες ιδιότητες με έναν συνήθη εκτιμητή μέγιστης πιθανοφάνειας, όπως είναι αυτή της ***συνέπειας*** *(consistency)* και της ***ασυμπτωτικής κανονικότητας*** *(asymptotic normality)*.

Πριν περάσουμε στην επόμενη παράγραφο, να υπενθυμίσουμε ότι στην περίπτωση των ντετερμινιστικών συμμεταβλητών, η μερική πιθανοφάνεια παίρνει τη μορφή της δεσμευμένης πιθανοφάνειας της σχέσης (3.2). Η υπενθύμιση αυτή γίνεται με αφορμή τις εφαρμογές που θα ακολουθήσουν στο Κεφάλαιο 4, όπου οι συμμεταβλητές που συμμετέχουν στα μοντέλα μας αφορούν είτε χρονικές υστερήσεις, είτε την περιοδικότητα, πράγμα που σημαίνει ότι στα συγκεκριμένα αυτά μοντέλα, χρησιμοποιείται ουσιαστικά μια ειδική περίπτωση της πλήρους πιθανοφάνειας της σχέσης (3.3).

3.3 ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΧΡΟΝΟΣΕΙΡΕΣ

Στο πρώτο κεφάλαιο, περιγράψαμε αναλυτικά τη χρησιμότητα, τη λειτουργία και τη δομή των γενικευμένων γραμμικών μοντέλων. Έτσι, βασισμένοι στο θεωρητικό υπόβαθρο που έχει ήδη αναφερθεί, θα δούμε πως επεκτείνονται όλα αυτά στην περίπτωση που οι μεταβλητές μας είναι χρονοσειρές.

Υπενθυμίζουμε τη γενική μορφή του γενικευμένου γραμμικού μοντέλου καθώς και τη σχέση της συνάρτησης σύνδεσης, που δίνονται στις σχέσεις (1.13) και (1.15) αντίστοιχα, μαζί με τις ερμηνείες των μεταβλητών που χρησιμοποιούνται σε αυτές:

$$f(y;\theta;\varphi) = \exp\left[\frac{y\theta - b(\theta)}{a(\varphi)} + c(y;\varphi)\right] \quad \text{και} \quad \eta_i = g_i(\mu_i) = x_i^T \beta$$

Έστω λοιπόν ότι διαθέτουμε μια χρονοσειρά Y_t, $t = 1, ..., N$. Αρχικά, για να ακολουθεί η χρονοσειρά αυτή κάποιο γενικευμένο γραμμικό μοντέλο, θα πρέπει η $Y_t \mid F_{t-1}$ να ανήκει στην εκθετική οικογένεια κατανομών (Ε.Ο.Κ), δηλαδή θα πρέπει για τη δεσμευμένη κατανομή της μεταβλητής Y_t δοθέντος του παρελθόντος F_{t-1}, να ισχύει ότι: (Kedem & Fokianos, 2002)

$$\boxed{f(y_t;\theta_t;\varphi \mid F_{t-1}) = \exp\left[\frac{y_t\theta_t - b(\theta_t)}{a_t(\varphi)} + c(y_t;\varphi)\right]} \tag{3.5}$$

Αφού η $Y_t \mid F_{t-1} \in$ Ε.Ο.Κ , με βάση τις σχέσεις που προκύπτουν από την (1.13) για τη μέση τιμή και τη διασπορά, θα έχουμε για τις δεσμευμένες ροπές του στατιστικού μοντέλου ότι:

$$\mu_t = E[Y_t \mid F_{t-1}] = b'(\theta_t) \tag{3.6}$$

και

$$\sigma^2(t) = Var[Y_t \mid F_{t-1}] = a_t(\varphi)b''(\theta_t) \tag{3.7}$$

Από τη σχέση (3.6), αφού $b''(\theta_t) > 0$ και άρα $b'(\theta_t)$ αντιστρέψιμη, προκύπτει κατά τα γνωστά, ότι η φυσική παράμετρος θ_t αποτελεί μονότονη συνάρτηση της μ_t οπότε μπορεί να χρησιμοποιηθεί για τον καθορισμό της συνάρτησης σύνδεσης, εφόσον ισχύει ότι:

$$\theta_t = (b')^{-1}(\mu_t)$$

Επομένως, βασισμένοι και στη γνωστή μας σχέση (1.15), για τη συνάρτηση σύνδεσης, η οποία υποδεικνύει και το μοντέλο μας, θα ισχύει ότι:

$$\boxed{g(\mu_t) = \theta_t(\mu_t) = \eta_t = \sum_{j=1}^{p} \beta_j Z_{(t-1)j} = Z'_{t-1}\beta} \tag{3.8}$$

Όπου $t = 1,\ldots,N$, β το διάνυσμα των παραμέτρων, και Z_{t-1} το διάνυσμα των μεταβλητών. Η παραπάνω συνήθης έκφραση, αποτελεί την κανονική συνάρτηση σύνδεσης για την οποία ισχύει ότι $g = \mu^{-1} \equiv (b')^{-1}$. Καταλαβαίνουμε τώρα από τη σχέση (3.8), πόσο σημαντική είναι η επιλογή των μεταβλητών που καθορίζουν αυτομάτως τα διανύσματα Z_{t-1} και β, στην προσαρμογή αλλά και στα αποτελέσματα του αντίστοιχου μοντέλου.

Επιλογές για το $Z'_{t-1}\beta$ που συναντώνται συχνά, όταν διαθέτουμε μια μεταβλητή απόκρισης $\{Y_t\}$ και κάποια επεξηγηματική μεταβλητή $\{X_t\}$, αποτελούν τα παρακάτω παραδείγματα:

$$Z'_{t-1}\beta = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 X_t \cos(\omega_0 t) \tag{3.9}$$

ή

$$Z'_{t-1}\beta = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-1} X_t + \beta_4 Y_{t-2} X_{t-1}$$

Πριν κλείσουμε τη συγκεκριμένη ενότητα, να αναφέρουμε απλώς, και μια άλλη γνωστή μορφή της συνάρτησης σύνδεσης, στην οποία στηρίζεται μια κατηγορία μοντέλων που ονομάζονται *Γενικευμένα Αυτοπαλινδρομούμενα Μοντέλα Κινητού Μέσου (GARMA(p,q))*:

$$g(\mu_t) = \eta_t = X'_t\gamma + \sum_{i=1}^{p} \varphi_i (g(Y_{t-i}) - X'_{t-i}\gamma) + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{3.10}$$

όπου $\varepsilon_{t-i} = g(Y_{t-i}) - \eta_{t-i}$.

Το μοντέλο της σχέσης (3.10) βασίζεται στις παρακάτω εκφράσεις των Z_{t-1} και β αντίστοιχα:

$$Z_{t-1} = (X_t, H_1(Y_{t-1}), ..., H_p(Y_{t-p)}, D_1(\mu_{t-1}), ..., D_q(\mu_{t-q}))'$$

και

$$\beta = (\gamma', \varphi_1, ..., \varphi_p, \theta_1, ..., \theta_q)'$$

όπου $H_i(\cdot)$ και $D(\cdot)$ είναι γνωστές συναρτήσεις για κάθε i.

Εμείς, για το $Z'_{t-1}\beta$, θα χρησιμοποιήσουμε εκφράσεις αντίστοιχες με τα παραδείγματα στην (3.9), αφού στο λογιστικό και στο Poisson μοντέλο με τα οποία θα ασχοληθούμε και θα δούμε παρακάτω, θα χρησιμοποιήσουμε την κανονική συνάρτηση σύνδεσης για την κάθε περίπτωση αντίστοιχα.

3.4 ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΜΕ ΤΗ ΧΡΗΣΗ ΤΗΣ ΜΕΡΙΚΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

Έχοντας αναλύσει στην προηγούμενη ενότητα την τροποποιημένη δομή των γενικευμένων γραμμικών μοντέλων στην περίπτωση των χρονικά εξαρτώμενων δεδομένων, από τα οποία αποτελείται μια χρονοσειρά, μπορούμε πλέον να παρουσιάσουμε τη διαδικασία εκτίμησης των παραμέτρων του μοντέλου (3.8), η οποία θα οδηγήσει και στη συμπερασματολογία. Όπως ήδη έχουμε εξηγήσει, η διαδικασία αυτή, στηρίζεται στη μέθοδο της μέγιστης πιθανοφάνειας.

Έστω λοιπόν, ότι διαθέτουμε μια χρονοσειρά $\{Y_t\}$, $t = 1, ..., N$ για την οποία ισχύει η σχέση (3.5), δηλαδή η δεσμευμένη κατανομή της δοθέντος της F_{t-1} ανήκει στην εκθετική οικογένεια κατανομών. Για τη συνάρτηση μερικής πιθανοφάνειας για την Y_t με βάση τον ορισμό που δόθηκε παραπάνω, θα ισχύει:

$$PL(\beta) = \prod_{t=1}^{N} f(y_t; \theta_t, \varphi \mid F_{t-1})$$

Άρα με βάση τη σχέση (3.5), για το λογάριθμο της μερικής πιθανοφάνειας θα έχουμε ότι: (Kedem & Fokianos, 2002)

$$l(\beta) = \sum_{t=1}^{N} \log f(y_t; \theta_t, \varphi \mid F_{t-1}) = \sum_{t=1}^{N} \left\{ \frac{y_t\theta_t - b(\theta_t)}{a_t(\varphi)} + c(y_t, \varphi) \right\}$$

$$= \sum_{t=1}^{N} \left\{ \frac{y_t u(z'_{t-1}\beta) - b(u(z'_{t-1}\beta))}{\alpha_t(\varphi)} + c(y_t, \varphi) \right\} \equiv \sum_{t=1}^{N} l_t$$

όπου

$$u(\cdot) \equiv (g \circ \mu(\cdot))^{-1} = \mu^{-1}(g^{-1}(\cdot)) \qquad (3.11)$$

Η μεταβλητή $u(\cdot)$, της οποίας η έκφραση δίνεται μέσω της σχέσης (3.11), εισάγεται για να εκφράσουμε την εξάρτηση της μερικής πιθανοφάνειας από το παραμετρικό διάνυσμα β. Συμβολίζει ουσιαστικά τη φυσική παράμετρο, η οποία μπορεί να γραφεί ως σύνθεση των g^{-1} και μ^{-1} με ανεξάρτητη μεταβλητή την $\eta_t = z'_{t-1}\beta$. Γενικά για κάθε συνάρτηση σύνδεσης, και όχι μόνο για την κανονική ισχύει ότι:

$$\theta_t = (b')^{-1}(g^{-1}(\eta_t)) = \mu^{-1}(g^{-1}(\eta_t))$$

Άρα με βάση την τελευταία σχέση και την (3.11), έχουμε ότι:

$$\theta_t = u(z'_{t-1}\beta).$$

Για να καταλήξουμε στις εκτιμήσεις των παραμέτρων του διανύσματος β, θα πρέπει να υπολογίσουμε τις συνιστώσες του p-διάστατου διανύσματος:

$$\nabla l(\beta) = \left(\frac{\partial l(\beta)}{\partial \beta_1}, \frac{\partial l(\beta)}{\partial \beta_2} ..., \frac{\partial l(\beta)}{\partial \beta_p} \right)' \tag{3.12}$$

Το διάνυσμα της σχέσης (3.12) καλείται *μερικό σκορ (partial score)* και για τον υπολογισμό των $\frac{\partial l_t(\beta)}{\partial \beta_j}$, $j = 1, 2, ..., p$ θα χρησιμοποιήσουμε τον παρακάτω κανόνα αλυσίδας:

$$\frac{\partial l_t(\beta)}{\partial \beta_j} = \frac{\partial l_t}{\partial \theta_t} \frac{\partial \theta_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial \beta_j} \tag{3.13}$$

Με βάση τα όσα έχουν αναφερθεί και συγκεκριμένα τις σχέσεις (3.6), (3.7), (3.8), έχουμε τα εξής:

$$\frac{\partial l_t}{\partial \theta_t} = \frac{(y - b'(\theta_t))}{a_t(\varphi)} = \frac{(y_t - \mu_t)}{a_t(\varphi)}, \qquad \frac{\partial \theta_t}{\partial \mu_t} = \frac{1}{b''(\theta_t)} = \frac{a_t(\varphi)}{Var[Y_t \mid F_{t-1}]}, \qquad \frac{\partial \eta_t}{\partial \beta_j} = z_{(t-1)j}$$

Συνεπώς, η (3.13) μπορεί να γραφεί:

$$\frac{\partial l_t(\beta)}{\partial \beta_j} = \frac{(y_t - \mu_t)}{Var[Y_t \mid F_{t-1}]} \frac{\partial \mu_t}{\partial \eta_t} z_{(t-1)j}, \quad j = 1, 2, ..., p$$

Έτσι, συμβολίζοντας το διάνυσμα των μερικών σκορ $\nabla l(\beta)$ με $S_N(\beta)$, (το οποίο αντιστοιχεί στο διάνυσμα των σκορ $u(\beta)$ που συναντάμε στην περίπτωση των κλασικών γενικευμένων γραμμικών μοντέλων), έχουμε ότι:

$$S_N(\beta) = \nabla l(\beta) = 0 \tag{3.14}$$

όπου $S_N(\beta) \equiv \nabla l(\beta) = \sum_{t=1}^{N} Z_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{(Y_t - \mu_t(\beta))}{\sigma_t^2(\beta)}$ με $\sigma_t^2(\beta) = Var[Y_t \mid F_{t-1}]$

Αντίστοιχα, η διανυσματική στοχαστική ανέλιξη $\{S_t(\beta)\}$, $t = 1, ..., N$ θα ορίζεται απο τα μερικά αθροίσματα:

$$S_t(\beta) = \sum_{s=1}^{N} Z_{s-1} \frac{\partial \mu_s}{\partial \eta_s} \frac{(Y_s - \mu_s(\beta))}{\sigma_s^2(\beta)}$$

Για την $S_N(\beta)$ ισχύει ότι:

$$\mathrm{E}[S_N(\beta)] = 0$$

το οποίο οφείλεται στο ότι $E\left[Z_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{(Y_t - \mu_t(\beta))}{\sigma_t^2(\beta)} \mid F_{t-1} \right] = 0$

Ομοίως, αποδεικνύεται επίσης ότι:

$$E\left[Z_{s-1} \frac{\partial \mu_s}{\partial \eta_s} \frac{(Y_s - \mu_s(\beta))}{\sigma_s^2(\beta)} Z'_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{(Y_t - \mu_t(\beta))}{\sigma_t^2(\beta)} \right] = 0 \ , \quad s < t$$

Η λύση λοιπόν της εξίσωσης

$$S_N(\beta) = \nabla \log PL(\beta) = \nabla l(\beta) = 0$$

θα μας δώσει το διάνυσμα $\hat{\beta}$, το οποίο αποτελεί τον εκτιμητή μέγιστης μερικής πιθανοφάνειας του β και θα περιέχει τις εκτιμήσεις των παραμέτρων του διανύσματος β. Το σύστημα των εξισώσεων που προκύπτει από την (3.14), είναι μη γραμμικό, γι'αυτό και επιλύεται συνήθως με τη χρήση της μεθόδου "scoring" του Fisher (Fisher scoring), η οποία αποτελεί έναν επαναληπτικό αλγόριθμο παρόμοιο με αυτόν της διαδικασίας Newton-Raphson. Η μέθοδος αυτή, καλείται *μέθοδος επαναληπτικών σταθμισμένων ελαχίστων τετραγώνων (iteratively reweighted least squares)*, γιατί η όλη διαδικασία εκτίμησης των παραμέτρων θυμίζει αυτήν των σταθμισμένων ελαχίστων τετραγώνων.

Στην παρούσα εργασία δε θα αναλυθεί η λειτουργία του συγκεκριμένου αλγόριθμου, παρ'όλα αυτά θα παραθέσουμε κάποιους σημαντικούς πίνακες οι οποίοι συναντώνται κατά τη χρήση του αλγορίθμου, αλλά διαδραματίζουν επίσης και σημαντικό ρόλο στη συμπερασματολογία με τη χρήση της μερικής πιθανοφάνειας.

Αρχικά, ορίζουμε τον *πίνακα δεσμευμένης πληροφορίας* $G_N(\beta)$ *(cumulative conditional information matrix)* , ως το παρακάτω άθροισμα των δεσμευμένων συνδιακυμάνσεων: (Kedem & Fokianos, 2002)

$$G_N(\beta) = \sum_{t=1}^{N} Cov\left[Z_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{(Y_t - \mu_t(\beta))}{\sigma_t^2(\beta)} \mid F_{t-1} \right]$$

$$= \sum_{t=1}^{N} Z_{t-1} \left(\frac{\partial \mu_t}{\partial \eta_t} \right)^2 \frac{1}{\sigma_t^2(\beta)} Z'_{t-1} = Z'W(\beta)Z \qquad (3.15)$$

Όπου $Z = \begin{bmatrix} Z'_0 \\ Z'_1 \\ \vdots \\ Z'_{N-1} \end{bmatrix}$ ένας $N \times p$ πίνακας, και

$W(\beta) = diag(w_1, w_2, ..., w_n)$ με $w_t = \left(\frac{\partial \mu_t}{\partial \eta_t}\right)^2 \frac{1}{\sigma_t^2(\beta)}, t = 1, ..., N$, ένας $N \times N$ πίνακας

Ο *αδέσμευτος πίνακας πληροφορίας* $F_N(\beta)$ *(unconditional information matrix)* ορίζεται ως εξής:

$$Cov(S_N(\beta)) = F_N(\beta) = E[G_N(\beta)]$$

Η τιμή της παραπάνω σχέσης, δεν είναι εύκολο να υπολογιστεί, γι'αυτό και κάτω από κάποιες κατάλληλες προϋποθέσεις, εκτιμάται μέσω του πίνακα της δεσμευμένης πληροφορίας $G_N(\beta)$.

Τέλος, έχουμε τον *πίνακα παρατηρούμενης πληροφορίας* $H_N(\beta)$ *(observed information matrix)*, ο οποίος ορίζεται ως εξής:

$$H_N(\beta) \equiv -\nabla\nabla' l(\beta)$$

Για τον πίνακα $H_N(\beta)$ ισχύει επίσης η ακόλουθη σχέση:

$$H_N(\beta) = G_N(\beta) - R_N(\beta) \tag{3.16}$$

όπου $R_N(\beta) = \frac{1}{a_t(\varphi)} \sum_{t=1}^{N} Z_{t-1} d_t(\beta) Z'_{t-1}(Y_t - \mu_t(\beta))$ και $d_t(\beta) = \left[\partial^2 u(\eta_t) / \partial \eta_t^2\right]$

Κλείνοντας την παρούσα ενότητα, είναι σημαντικό να εξηγήσουμε ότι στην περίπτωση της χρήσης της κανονικής σύνδεσης η_t στο γενικευμένο γραμμικό μοντέλο, όλες οι παραπάνω γενικές σχέσεις απλοποιούνται. Υπενθυμίζουμε πως για την κανονική σύνδεση ισχύει ότι:

$$\eta_t = \theta_t \quad \text{οπότε} \quad \frac{\partial \mu_t}{\partial \eta_t} = \frac{\partial \mu_t}{\partial \theta_t} = b''(\theta_t) \tag{3.17}$$

Και επίσης $u(\eta_t) = \eta_t$ οπότε $d_t(\beta) = 0$ (3.18)

Οπότε με βάση τις (3.17) και (3.18), οι σχέσεις (3.14), (3.15) και (3.16) απλοποιούνται αντίστοιχα, ως εξής:

$$S_N(\beta) = \frac{1}{a_t(\varphi)} \sum_{t=1}^{N} Z_{t-1}(Y_t - \mu_t(\beta)) \tag{3.19}$$

$$G_N(\beta) = \frac{1}{a_t^2(\varphi)} \sum_{t=1}^{N} Z_{t-1} \sigma_t^2(\beta) Z'_{t-1} \tag{3.20}$$

$$H_N(\beta) = G_N(\beta) \tag{3.21}$$

Άρα συμπεραίνουμε ότι για το λογιστικό και το Poisson μοντέλο, όπου χρησιμοποιείται η κανονική συνάρτηση σύνδεσης και θα δούμε στο επόμενο

κεφάλαιο για δυαδικά και απαριθμητά δεδομένα χρονοσειρών αντίστοιχα, θα ισχύουν οι παραπάνω απλοποιημένες σχέσεις.

Πριν κλείσουμε την παράγραφο αυτή, έχοντας πλέον αποκτήσει μια καλή αντίληψη της προσαρμογής μιας χρονοσειράς σε ένα γενικευμένο μοντέλο και τη διαδικασία συμπερασματολογίας μέσω αυτού, είναι σημαντικό να αναφέρουμε πως πέρα από τη θεωρητική υπόσταση όλων αυτών, υπάρχει και η πρακτική εφαρμογή τους. Η εκτίμηση των παραμέτρων ενός γενικευμένου γραμμικού μοντέλου με τη χρήση της μερικής πιθανοφάνειας, υποστηρίζεται από πολλά στατιστικά προγράμματα, όπως είναι η R ή το SAS. Μπορούμε δηλαδή μέσω αυτών των προγραμμάτων και μέσω των κατάλληλων εντολών να υπολογίσουμε τις εκτιμήσεις της μέγιστης μερικής πιθανοφάνειας για τις παραμέτρους του μοντέλου κι έτσι να οδηγηθούμε σε κατάλληλα συμπεράσματα, πράγμα που οφείλεται στο ότι η εξίσωση των μερικών σκορ (3.14), είναι η ίδια με την αντίστοιχη εξίσωση που έχουμε για ανεξάρτητα δεδομένα.

3.5 ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ ΚΑΙ ΔΙΑΓΝΩΣΤΙΚΟΙ ΕΛΕΓΧΟΙ

Πριν κλείσουμε αυτό το κεφάλαιο, θα αναφερθούμε επιγραμματικά σε κάποιους συνηθισμένους στατιστικούς ελέγχους που χρησιμοποιούνται για τον έλεγχο των συντελεστών β, και σε κάποιες διαγνωστικές τεχνικές μέσω των οποίων μπορούμε να ελέγξουμε την καλή προσαρμογή του μοντέλου μας. Κάποιους από τους παρακάτω ελέγχους θα τους δούμε αναλυτικότερα και στο επόμενο κεφάλαιο, εδώ ο σκοπός είναι η απλή αναφορά τους ώστε να γνωρίζουμε ποιοί επιλέγονται στην περίπτωση των γενικευμένων γραμμικών μοντέλων. (Οικονόμου & Καρώνη , 2010; Kedem & Fokianos, 2002)

Α] ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ

Πολύ συχνά, θέλουμε να ελέγξουμε τις τιμές των συντελεστών β_j, ώστε μα βάση τα αποτελέσματα να βγάλουμε συμπεράσματα για τις μεταβλητές οι οποίες συμμετέχουν στο μοντέλο μας. Για παράδειγμα, μέσω του ελέγχου της πιο συνηθισμένης υπόθεσης:

$$H_0 : \beta_j = 0 \quad \text{με εναλλακτική την} \quad H_1 : \beta_j \neq 0 \tag{3.22}$$

συμπεραίνουμε τη σημαντικότητα της μεταβλητής με δείκτη j στο μοντέλο και έτσι μπορούμε να αποφασίσουμε για τον αν θα συμπεριληφθεί ή όχι μέσα σε αυτό. Γενικότερα, σε πολλές εφαρμογές, μπορεί να προκύψει η παρακάτω υπόθεση:

$$H_0 : C\beta = \beta_0 \quad \text{με εναλλακτική την} \quad H_1 : C\beta \neq \beta_0 \tag{3.23}$$

όπου C ένας κατάλληλος, γνωστός πίνακας.

Για τον έλεγχο λοιπόν της υπόθεσης της σχέσης (3.22), αν υποθέσουμε πως $\tilde{\beta}$ είναι η εκτιμήτρια μέγιστης μερικής πιθανοφάνειας του μοντέλου χωρίς τις όποιες μεταβλητές j ενώ $\hat{\beta}$ η αντίστοιχη εκτιμήτρια του αρχικού, ολοκληρωμένου μοντέλου, χρησιμοποιούνται κυρίως οι παρακάτω γνωστοί στατιστικοί έλεγχοι:

- ✓ Έλεγχος του λόγου μερικής πιθανοφάνειας (Log-partial likelihood ratio statistic):

$$\lambda_{\mathrm{N}} = 2\{l(\hat{\beta}) - l(\tilde{\beta})\}$$

- ✓ Έλεγχος του σκορ (score statistic):

$$c_N = \frac{1}{N} S_N'(\tilde{\beta}) G^{-1}(\tilde{\beta}) S_N(\tilde{\beta})$$

- ✓ Έλεγχος του Wald (του οποίου παραθέτουμε τη μορφή, στην περίπτωση της υπόθεσης της σχέσης (3.23)):

$$w_N = \{C\hat{\beta} - \beta_0\}' \{CG^{-1}(\hat{\beta})C'\}^{-1} \{C\hat{\beta} - \beta_0\}$$

Και οι τρείς στατιστικοί έλεγχοι ακολουθούν ασυμπτωτικά την κατανομή χ^2 (με βαθμούς ελευθερίας που καθορίζονται από τις παραμέτρους του μοντέλου), και στατιστικά σημαντική $p-$value οδηγεί στην απόρριψη της H_0.

Β] ΔΙΑΓΝΩΣΤΙΚΟΙ ΕΛΕΓΧΟΙ

Μέσω των παρακάτω διαγνωστικών μεθόδων ελέγχουμε την καταλληλότητα του μοντέλου και το πόσο καλά προσαρμόζονται τα δεδομένα μας σε αυτό. Κάποιους από αυτούς τους ελέγχους, τους έχουμε ήδη συναντήσει και από τα προηγούμενα κεφάλαια.

- ✓ Ελεγχοσυνάρτηση Deviance

$$D \equiv 2\{l(y;y) - l(\hat{\mu};y)\}$$

Όπου, όπως έχουμε εξηγήσει (βλέπε πρώτο κεφάλαιο, παράγραφος 1.4.2):
$l(y;y)$ η μέγιστη τιμή του λογαρίθμου της μερικής πιθανοφάνειας για το *κορεσμένο* μοντέλο, το οποίο περιέχει ίσο αριθμό παρατηρήσεων και παραμέτρων (κάθε μ_t εκτιμάται άμεσα απο τις παρατηρήσεις $Y_1,\ldots,Y_N$)
$l(\hat{\mu};y)$ η αντίστοιχη τιμή για το "μειωμένο" μοντέλο που περιέχει τη συνάρτηση σύνδεσης και τις συμμεταβλητές.
Η $D \sim \chi^2_{N-p}$ ασυμπτωτικά, όπου p ο αριθμός των παραμέτρων, και μετράει ουσιαστικά την απώλεια της προσαρμογής του μοντέλου μας σε σχέση με αυτή του κορεσμένου. Χρησιμοποιείται επίσης και για τη σύγκριση δύο (μη κορεσμένων) μοντέλων ($D_0 - D_1 \sim \chi_{p-q}$).

- ✓ Κριτήριο AIC

$$AIC(p) = -2\log PL(\hat{\beta}) + 2p \qquad (3.24)$$

Όπου $\hat{\beta}$ ως γνωστόν, ο εκτιμητής μέγιστης μερικής πιθανοφάνειας του β και p η τάξη του μοντέλου, $p = \dim(\beta)$. Μέσω του AIC επιλέγουμε το καταλληλότερο μοντέλο, με βάση αυτό το p που ελαχιστοποιεί την (3.24).

✓ Κριτήριο BIC

$$BIC(p) = -2\log PL(\hat{\beta}) + p\log N$$

Το κριτήριο BIC αποτελεί μια τροποποιημένη μορφή του AIC, κατάλληλη για μεγάλα δείγματα και μοντέλα πολύ μεγάλης τάξης p, ώστε να δίνει καλύτερες εκτιμήσεις του p.

✓ Υπόλοιπα

Για να εξετάσουμε την καλή προσαρμογή του μοντέλου μας, γνωρίζουμε ότι εκτός των άλλων χρησιμοποιούμε και γραφικές παραστάσεις διαφόρων τύπων υπολοίπων, όπως είναι τα υπόλοιπα Pearson, τα υπόλοιπα deviance ή τα υπόλοιπα της μεταβλητής απόκρισης ($\hat{e}_t = Y_t - \hat{\mu}_t$ όπου $\hat{\mu}_t = \mu_t(\hat{\beta})$).

✓ Έλεγχος λευκού θορύβου

Όπως έχουμε ήδη αναφέρει, στην ιδανική περίπτωση που η γραφική αναπαράσταση των υπολοίπων του μοντέλου μας έχει τη μορφή χρονοσειράς λευκού θορύβου, τότε η επιλογή του μοντέλου είναι κατάλληλη. Για τον έλεγχο αυτόν, πέραν του γραφήματος των υπολοίπων, χρησιμοποιούμε επίσης το γράφημα της συνάρτησης αυτοσυσχέτισης (βλέπε παράγραφο 2.3).

ΚΕΦΑΛΑΙΟ 4

ΠΑΛΙΝΔΡΟΜΙΚΑ ΜΟΝΤΕΛΑ ΓΙΑ ΔΥΑΔΙΚΕΣ ΚΑΙ ΑΠΑΡΙΘΜΗΤΕΣ ΧΡΟΝΟΣΕΙΡΕΣ

4.1 ΕΙΣΑΓΩΓΗ

Στο παρόν κεφάλαιο, όπως έχει ήδη αναφερθεί, θα ασχοληθούμε με τη στατιστική ανάλυση δύο πολύ χρήσιμων ειδών χρονοσειρών, των δυαδικών και των απαριθμητών χρονοσειρών, μέσω της προσαρμογής τους στο λογιστικό και στο Poisson παλινδρομικό μοντέλο αντίστοιχα. Στο προηγούμενο κεφάλαιο, εξηγήσαμε τη χρησιμότητα της μοντελοποίησης των χρονοσειρών μέσω των γενικευμένων γραμμικών μοντέλων, ενώ αναλύσαμε εκτενώς το θεωρητικό υπόβαθρο στο οποίο στηρίζεται η δυνατότητα αυτή, αναφερόμενοι στη συμπερασματολογία μέσω της δεσμευμένης και μερικής πιθανοφάνειας. Είναι προφανές, ότι τα όσα αναφέρθηκαν για τη δομή που αποκτούν τα γενικευμένα γραμμικά μοντέλα στην περίπτωση των χρονοεξαρτώμενων δεδομένων, από τα οποία αποτελείται μια χρονοσειρά, επεκτείνονται και στις δύο αυτές συγκεκριμένες περιπτώσεις, του λογιστικού και του Poisson μοντέλου.

Στις επόμενες παραγράφους που ακολουθούν, θα δούμε δύο εκτενείς εφαρμογές των προαναφερθέντων μοντέλων, για δυαδικές χρονοσειρές που αφορούν το φαινόμενο της βροχόπτωσης στη μία περίπτωση, και για απαριθμητές χρονοσειρές που αφορούν τον αριθμό αυτοκτονιών στην άλλη περίπτωση. Πριν από την κάθε εφαρμογή, θα παρουσιάσουμε μια σύντομη περιγραφή της δομής των δύο μοντέλων στην περίπτωση αυτή, όπου οι μεταβλητές αποτελούν χρονοσειρές, καθώς η αναλυτική περιγραφή για τη γενική μορφή της λογιστικής και της Poisson παλινδρόμησης έχει προηγηθεί στο πρώτο κεφάλαιο.

4.2 ΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ ΓΙΑ ΔΥΑΔΙΚΕΣ ΧΡΟΝΟΣΕΙΡΕΣ

4.2.1 Δομή λογιστικού μοντέλου

Έστω ότι μας ενδιαφέρει ένα διαχρονικό φαινόμενο στο οποίο σε κάθε επανάληψη είναι δυνατόν να συμβεί είτε “επιτυχία”, είτε “αποτυχία”, καταγράφοντας έτσι τις παρατηρήσεις μας ως 1 και 0 αντίστοιχα, από τις οποίες θα αποτελείται και η χρονοσειρά μας Y_t, που έχει το ρόλο της μεταβλητής απόκρισης στο πρόβλημα μας. Γνωρίζουμε ότι η δεσμευμένη κατανομή $Y_t \mid F_{t-1}$ ακολουθεί την κατανομή *Bernoulli* η οποία μπορεί να γραφεί στην εξής μορφή:

$$f(y_t;\pi_t \mid F_{t-1}) = \exp\left\{ y_t \log\left(\frac{\pi_t(\beta)}{1-\pi_t(\beta)} \right) + \log(1-\pi_t(\beta)) \right\}$$

όπου $\pi_t(\beta)$ αποτελεί τη δεσμευμένη πιθανότητα επιτυχίας:

$$\pi_t(\beta) = P_\beta = (Y_t = 1 | F_{t-1}) \equiv \mu_t(\beta)$$

με β να είναι το διάνυσμα των παραμέτρων και F_{t-1} η γνωστή μας πλέον σ-άλγεβρα που περιέχει τις πληροφορίες που είναι γνωστές για το φαινόμενο μέχρι και τη στιγμή $t-1$.

Ο σκοπός είναι να μοντελοποιήσουμε την παραπάνω δεσμευμένη πιθανότητα μέσω του λογιστικού μοντέλου παλινδρόμησης το οποίο θα εξαρτάται από το παραμετρικό διάνυσμα β. Έτσι δεδομένης της χρονοσειράς $\{Y_t\}$, $t = 1,\ldots,N$ και μιας συγκεκριμένης κάθε φορά διανυσματικής στοχαστικής ανέλιξης $\{Z_{t-1}\}$, $t = 1,\ldots,N$, θα προκύπτουν και οι εκτιμήσεις του β. Επιλέγοντας την κανονική συνάρτηση σύνδεσης, δηλαδή το σύνδεσμο *logit*:

$$\theta_t = g(\pi_t) = \log\left(\frac{\pi_t}{1-\pi_t}\right)$$

προκύπτει για τη χρονοσειρά μας, η μορφή του λογιστικού μοντέλου η οποία στηρίζεται στη σχέση (3.8) από το προηγούμενο κεφάλαιο:

$$\boxed{\text{logit}(\pi_t(\beta)) = \log\left(\frac{\pi_t}{1-\pi_t}\right) = \mathrm{Z}'_{t-1}\beta} \qquad (4.1)$$

Η σχέση (4.1), μπορεί φυσικά να γραφεί και στην εξής ισοδύναμη μορφή:

$$\boxed{\pi_t(\beta) = \frac{1}{1+\exp[-\beta'\,\mathrm{Z}_{t-1}]}} \qquad (4.2)$$

Πέραν του λογιστικού, υπάρχουν και εναλλακτικά μοντέλα στα οποία θα μπορούσε να στηριχθεί η μοντελοποίηση μιας δίτιμης χρονοσειράς, στα οποία θα οδηγούμασταν εάν είχαμε επιλέξει κάποια άλλη συνάρτηση σύνδεσης (πχ την *probit*) και όχι την κανονική. Εμείς όμως, στην εφαρμογή που θα ακολουθήσει θα χρησιμοποιήσουμε το λογιστικό μοντέλο, που αποτελεί και τη συνηθέστερη επιλογή, του οποίου η μορφή δίνεται μέσω της σχέσης (4.1).

Προφανώς, η συμπερασματολογία για το λογιστικό μοντέλο παλινδρόμησης μιας χρονοσειράς Y_t, θα στηρίζεται στη θεωρία της δεσμευμένης και της μερικής πιθανοφάνειας, και από τη στιγμή που για το μοντέλο χρησιμοποιείται η κανονική συνάρτηση σύνδεσης, υπενθυμίζουμε ότι για τη συμπερασματολογία, θα ισχύουν οι απλοποιημένες σχέσεις (3.19), (3.20), (3.21):

$$S_N(\beta) = \frac{1}{a_t(\varphi)}\sum_{t=1}^{N} Z_{t-1}(Y_t - \mu_t(\beta)) \quad , \qquad G_N(\beta) = \frac{1}{a_t^2(\varphi)}\sum_{t=1}^{N} Z_{t-1}\sigma_t^2(\beta)Z'_{t-1}$$

$$H_N(\beta) = G_N(\beta)$$

Όπου όπως έχουμε εξηγήσει στο προηγούμενο κεφάλαιο, μέσω της εξίσωσης $S_N(\beta) = 0$, υπολογίζεται το διάνυσμα $\hat{\beta}$, που περιέχει τις εκτιμήσεις μέγιστης

μερικής πιθανοφάνειας του β, ενώ οι $G_N(\beta), H_N(\beta)$ αποτελούν τους πίνακες δεσμευμένης και παρατηρούμενης πληροφορίας αντίστοιχα.

Για τους ελέγχους υποθέσεων και καταλληλότητας του μοντέλου, ισχύουν τα όσα έχουμε αναφέρει στην παράγραφο 3.5, με εξαίρεση την ελεγχοσυνάρτηση Deviance, που όπως έχουμε εξηγήσει στο πρώτο κεφάλαιο, δεν εφαρμόζεται σε δυαδικές χρονοσειρές.

4.2.2 Εφαρμογή λογιστικού παλινδρομικού μοντέλου για δυαδική χρονοσειρά

Σε αυτήν την εφαρμογή θα ασχοληθούμε με την προσπάθεια μοντελοποίησης χρονοσειράς δυαδικών δεδομένων. Με βάση τα όσα αναφέραμε παραπάνω για τη δομή του λογιστικού παλινδρομικού μοντέλου στην περίπτωση των χρονοσειρών και λαμβάνοντας πάντα υπόψη, τα όσα έχουμε εξηγήσει στο Κεφάλαιο 2, θα μπορέσουμε να καταλάβουμε τη λογική στην οποία στηρίζεται η προσαρμογή μιας δίτιμης χρονοσειράς στο συγκεκριμένο μοντέλο, στην πράξη και όχι μόνο θεωρητικά.

Συγκεκριμένα, θα προσπαθήσουμε να επιλέξουμε κάποιο μοντέλο, το οποίο θα μπορεί να περιγράψει σε κάποιο βαθμό τη συμπεριφορά του φαινομένου της βροχόπτωσης και για την ακρίβεια της υψηλής βροχόπτωσης. Στη διάθεσή μας έχουμε τα δεδομένα δύο χρονοσειρών. Και οι δύο χρονοσειρές αφορούν την ύπαρξη υψηλής βροχόπτωσης (0 όχι/1 ναι) ανά μήνα κατά τη διάρκεια 21 ετών, από το 1972 έως και το 1992. Η πρώτη χρονοσειρά αποτελείται από τις παρατηρήσεις υψηλής βροχόπτωσης σε σύγκριση με το 75ο ποσοστιαίο σημείο, ενώ η δεύτερη, από τις παρατηρήσεις σε σύγκριση με το 90ο ποσοστιαίο σημείο. Ουσιαστικά, στην πρώτη περίπτωση έχουμε τις παρατηρήσεις για την εμφάνιση ή όχι υψηλής βροχόπτωσης σε κάθε μήνα, ενώ στη δεύτερη έχουμε τις αντίστοιχες παρατηρήσεις, αλλά για ακόμα πιο υψηλή βροχόπτωση από ότι στην πρώτη.

Συγκεκριμένα, η τιμή των ποσοστιαίων σημείων είναι 42.0 για το 75ο και 58.8 για το 90ο, δηλαδή για την πρώτη περίπτωση το 75% των παρατηρήσεων είναι κάτω από την τιμή 42.0 και το 25% πάνω από το 42.0, ενώ για τη δεύτερη περίπτωση το 90% των παρατηρήσεων βρίσκεται κάτω από την τιμή 58.8 και το 10% πάνω από το 58.8. Αυτό σημαίνει, πως στην πρώτη χρονοσειρά έχουμε 252 παρατηρήσεις για κάθε μήνα, οι οποίες παίρνουν την τιμή 0, αν στο 75% του δείγματος μας η τιμή για τη βροχόπτωση ήταν κάτω από 42.0, ενώ παίρνουν την τιμή 1, αν η τιμή ήταν πάνω από 42.0, πράγμα που θα χαρακτηρίζει το υπόλοιπο 25% για το αν έχουμε ή όχι υψηλή βροχόπτωση στον εκάστοτε μήνα. Αντίστοιχα στη δεύτερη χρονοσειρά, η κάθε παρατήρηση παίρνει την τιμή 0 αν στο 90% του δείγματος η τιμή για τη βροχόπτωση ήταν κάτω από 58.8 και 1, αν ήταν πάνω από 58.8 (δηλ υψηλότερη βροχόπτωση σε σχέση με την προηγούμενη χρονοσειρά), διαμορφώνοντας έτσι την εικόνα της πολύ υψηλής βροχόπτωσης για το υπόλοιπο 10% του δείγματος, για κάθε μήνα.

Στην απαραίτητη περιγραφική ανάλυση, που θα ακολουθήσει σε πρώτο βήμα, θα επεξεργαζόμαστε τις δύο αυτές χρονοσειρές παράλληλα, ώστε να μπορούμε να διακρίνουμε και τις όποιες μεταξύ τους διαφορές σχετικά με τη συμπεριφορά των παρατηρήσεων τους. Η επεξεργασία των δεδομένων μας θα γίνει μέσω της R, όπου

εισάγουμε τα δεδομένα δημιουργώντας ένα πλαίσιο, το οποίο περιέχει δύο στήλες, μία για κάθε χρονοσειρά. Κατασκευάζουμε τις δύο χρονοσειρές μας για το 75ο και 90ο ποσοστιαίο σημείο αντίστοιχα, μεγέθους 252 παρατηρήσεων:

>rainfall_top25<-ts(y1,start=1972,frequency=12)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1972	1	0	0	1	0	0	0	0	0	1	0	0
1973	0	0	0	0	0	0	0	0	0	0	1	1
1974	0	0	0	1	0	0	0	0	0	0	1	0

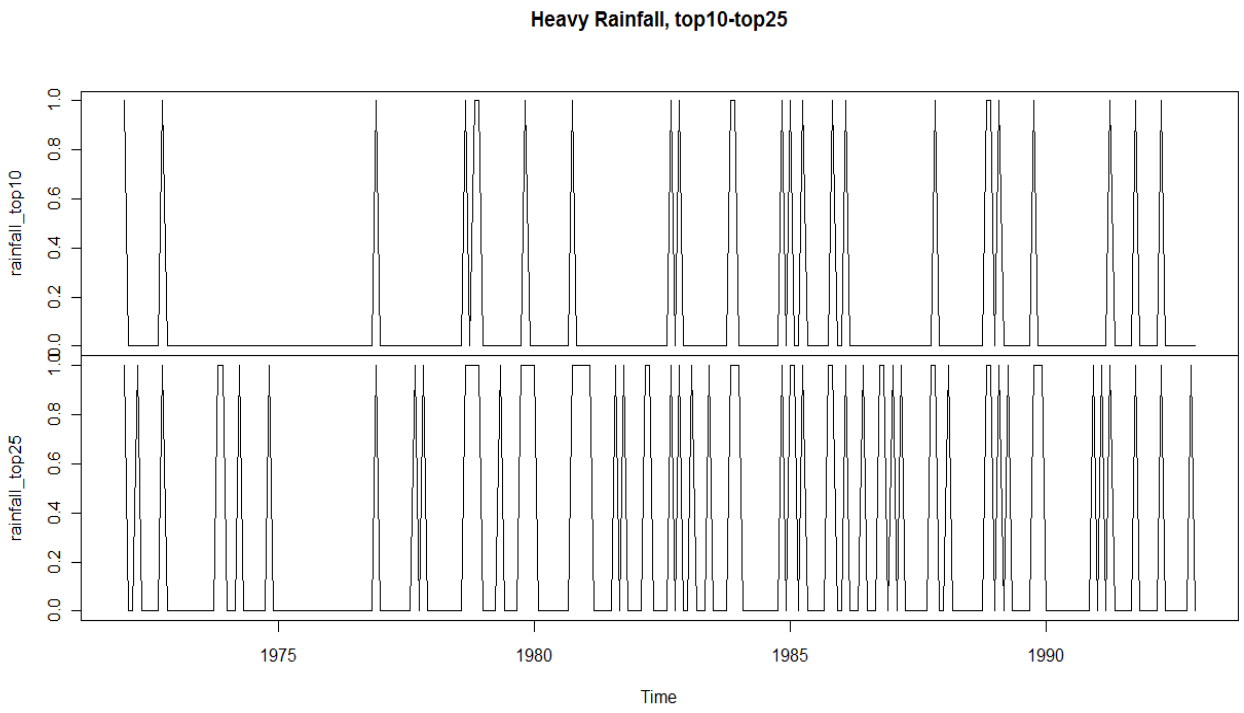
> rainfall_top10<-ts(y2,start=1972,frequency=12)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1972	1	0	0	0	0	0	0	0	0	1	0	0
1973	0	0	0	0	0	0	0	0	0	0	0	0
1974	0	0	0	0	0	0	0	0	0	0	0	0

(Παραθέτουμε μόνο τα τρία πρώτα έτη, απλώς για να δούμε τη μορφή της χονοσειράς. Πηγή δεδομένων: Από μελέτη ΕΜΠ)

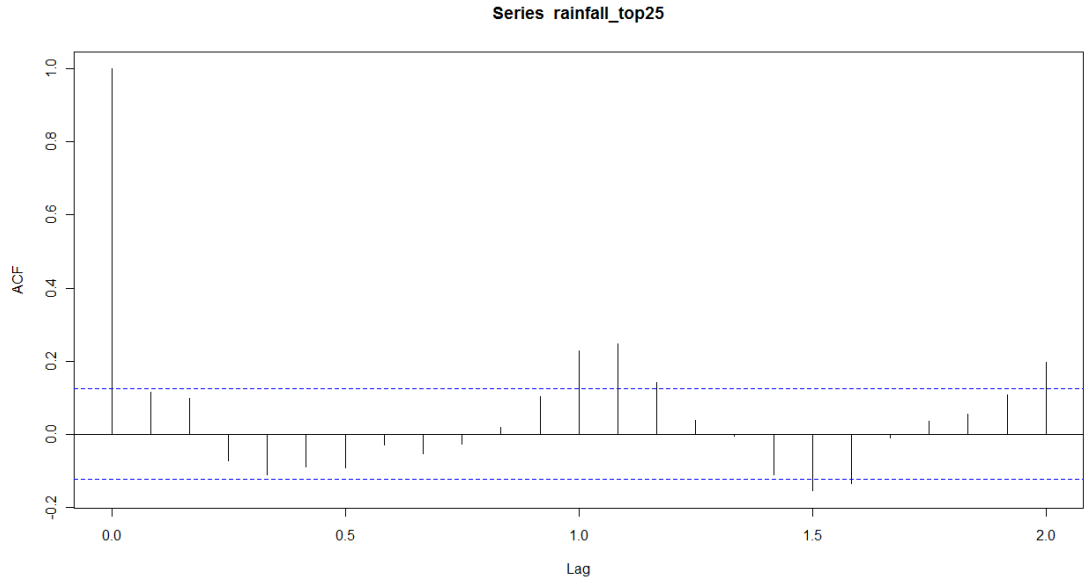
Με τη χρήση των παρακάτω εντολών έχουμε τη γραφική αναπαράσταση και των δύο χρονοσειρών, καθώς και τα γραφήματα της αυτοσυσχέτισης τους, μέσω των οποίων αποκτούμε μια αντιπροσωπευτική εικόνα του πως συμπεριφέρονται οι παρατηρήσεις μας στο χρόνο:

> union<-ts.union(rainfall_top10,rainfall_top25)
> plot(union,main="Heavy Rainfall, top10-top25")



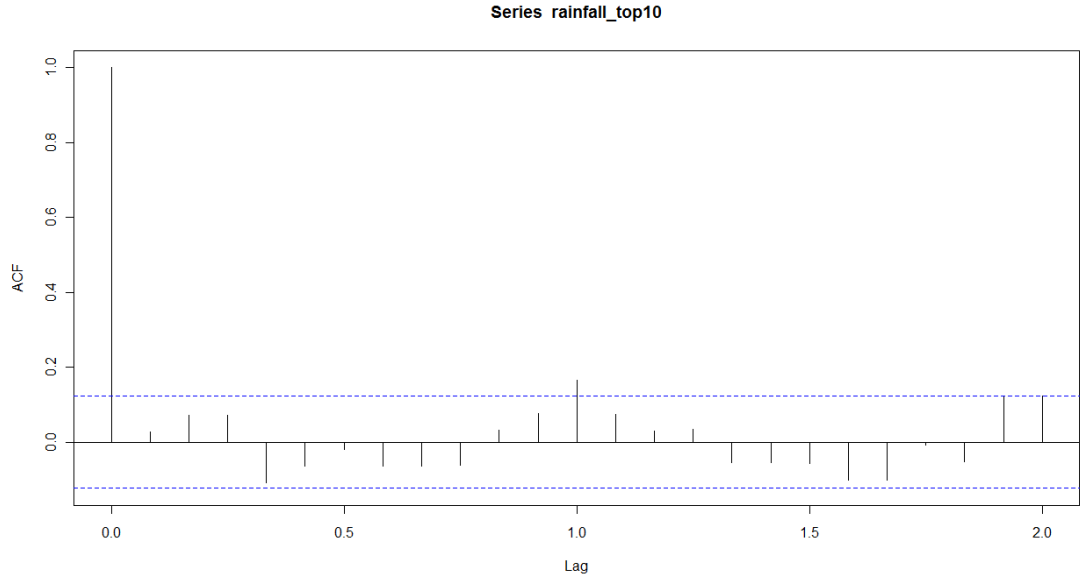
Η εικόνα που παρουσιάζουν οι γραφικές αναπαραστάσεις των χρονοσειρών μας, ήταν μάλλον αναμενόμενες, καθώς η χρονοσειρά η οποία αφορά την πολύ υψηλή βροχόπτωση, έχει εμφανώς περισσότερες μηδενικές παρατηρήσεις. Είναι λογικό, το φαινόμενο της υψηλής βροχόπτωσης να είναι αρκετά συχνότερο από αυτό της ακόμα πιο έντονης βροχόπτωσης. Χαρακτηριστική περίοδος, είναι αυτή από το 1973 έως το 1977, η οποία αποτελεί τη μακρύτερη χρονική περίοδο που δεν παρατηρείται ιδιαίτερα υψηλή βροχόπτωση *(γράφημα rainfall_top10)*, ενώ από το 1975 έως το 1977 είναι η αντίστοιχη μακρύτερη χρονική περίοδος, που δεν παρατηρείται καθόλου υψηλή βροχόπτωση *(γράφημα rainfall_top25)*. Πέραν όμως των συγκεκριμένων ετών, οι χρονοσειρές δεν παρουσιάζουν κάποια τάση κατά τη διάρκεια των 21 ετών, αφού υπάρχει συνεχής εναλλαγή των τιμών μεταξύ 0 και 1.

```
> acf(rainfall_top25,lag.max=20,type="correlation")
```



Γράφημα 4.1

```
> acf(rainfall_top10,lag.max=20,type="correlation")
```



Γράφημα 4.2

Το γράφημα της αυτοσυσχέτισης 4.1, που αντιστοιχεί στη χρονοσειρά υψηλής βροχόπτωσης, υποδεικνύει την ύπαρξη μιας εμφανούς περιοδικότητας μεταξύ των παρατηρήσεων, όπως μπορούμε να πούμε ότι συμβαίνει και στο γράφημα 4.2, της περισσότερο υψηλής βροχόπτωσης, αλλά σίγουρα σε μικρότερο βαθμό από ότι στο πρώτο. Επίσης, μέσω του δεύτερου γραφήματος και των δύο οριζόντιων γραμμών του, οι οποίες δείχνουν το 95% διάστημα εμπιστοσύνης για τον έλεγχο της υπόθεσης $H_0 : \rho = 0$, παρατηρούμε πως για τη χρονοσειρά της ιδιαίτερα υψηλής βροχόπτωσης, μάλλον δεν υπάρχει ένδειξη αυτοσυσχέτισης.

Για την περιοδικότητα που παρατηρήθηκε μέσω των γραφημάτων αυτοσυσχέτισης, θα είχε ενδιαφέρον να δούμε αναλυτικά τις εκτιμήσεις που προκύπτουν γι'αυτήν σε κάθε μήνα χωριστά, μέσω των οποίων θα μπορούσαμε να ελέγξουμε τον τρόπο, με τον οποίο κυμαίνονται διαχρονικά οι παρατηρήσεις μας. Όπως μπορούμε να δούμε παρακάτω, οι εκτιμήσεις αυτές, μπορούν να υπολογιστούν μέσω της εντολής *decompose*, με τη χρήση της οποίας αποθηκεύουμε τις εκτιμήσεις όλων των συνιστωσών της χρονοσειράς στη μεταβλητή *components* και με την εντολή *components$seasonal,* μπορούμε να εμφανίσουμε τα αποτελέσματα μόνο για τη συνιστώσα που μας ενδιαφέρει, που σε αυτήν την περίπτωση είναι η περιοδικότητα. Οι τιμές που θα προκύψουν αφορούν τον κάθε μήνα ξεχωριστά, από τον Ιανουάριο μέχρι το Δεκέμβριο, και είναι η τιμή του κάθε μήνα είναι ίδια για όλα τα χρόνια.

Έτσι, για τη χρονοσειρά της υψηλής βροχόπτωσης έχουμε τα εξής αποτελέσματα:

```
> components <- decompose(rainfall_top25)
> components$seasonal
```

	Jan	Feb	Mar	Apr	May	Jun
1972	-0.0002	0.0998	-0.1501	0.0519	-0.1980	-0.1501
	Jul	Aug	Sep	Oct	Nov	Dec
1972	-0.2522	-0.2001	-0.1001	0.2498	0.4498	0.1998

Πίνακας 4.1

Και αντίστοιχα για τη χρονοσειρά της πολύ υψηλής βροχόπτωσης:

```
> components <- decompose(rainfall_top10)
> components$seasonal
```

	Jan	Feb	Mar	Apr	May	Jun
1972	-0.0519	-0.0019	-0.1019	0.0501	-0.0977	-0.0977
	Jul	Aug	Sep	Oct	Nov	Dec
1972	-0.0998	-0.0977	0.0022	0.1001	0.2980	0.0980

Πίνακας 4.2

Σύμφωνα με τον πίνακα 4.1, για τη χρονοσειρά της υψηλής βροχόπτωσης, η μεγαλύτερη τιμή σημειώνεται το Νοέμβριο, ενώ οι μικρότερες τον Ιούλιο και τον Αύγουστο, το οποίο σημαίνει ότι διαχρονικά, η υψηλή βροχόπτωση παρατηρείται με μεγάλη συχνότητα τον Νοέμβριο και σπανιότερα/ή καθόλου τον Ιούλιο και τον

Αύγουστο. Αντίστοιχα, από τον πίνακα 4.2 για τη χρονοσειρά που αφορά την ακόμα πιο έντονη βροχόπτωση, ο μήνας, στον οποίο παρατηρείται συχνότερα το φαινόμενο αυτό, είναι και πάλι ο Νοέμβριος, ενώ η μικρότερη τιμή σε αυτήν την περίπτωση παρατηρείται το Μάρτιο, με πολύ μικρή όμως διαφορά από όλους τους θερινούς μήνες από το Μάιο μέχρι και τον Αύγουστο, που είναι και λογικό. *(Παραθέτουμε τις τιμές μόνο για τη μία χρονιά σε κάθε μήνα καθώς είναι η ίδια για όλες τις επόμενες)*

Έχοντας αποκτήσει πλέον μια γενική εικόνα των δύο χρονοσειρών, μπορούμε να προχωρήσουμε στην προσπάθεια μοντελοποίησης τους. Μια συχνή προσέγγιση για τη μοντελοποίηση του φαινομένου της βροχόπτωσης γενικότερα, είναι εκείνη που στηρίζεται στις χρονικές υστερήσεις της μεταβλητής Y_t. Θα προσαρμόσουμε δηλαδή ένα αυτοπαλινδρομούμενο λογιστικό μοντέλο, υποθέτοντας ότι η βροχόπτωση εξαρτάται από κάποιους συγκεκριμένους, προηγούμενους μήνες. Για να ελέγξουμε το αν όντως ισχύει αυτό, θα πρέπει να κατασκευάσουμε το μοντέλο που θα περιέχει έναν αριθμό χρονικών υστερήσεων (lags). Θα ξεκινήσουμε με την χρονοσειρά που αφορά την υψηλή βροχόπτωση (75ο ποσοστιαίο σημείο) και θα ακολουθήσει η χρονοσειρά της πολύ υψηλής βροχόπτωσης (90ο ποσοστιαίο σημείο).

Για να σχηματίσουμε μια άποψη για τον αριθμό των lags, που πρέπει να επιλεχθεί για την χρονοσειρά μας, χρησιμοποιούμε την εντολή *ar(suicides)* η οποία θα μας υποδείξει ποιά θα ήταν η ιδανική τάξη (*p*) για ένα αυτοπαλινδρομούμενο μοντέλο, δηλαδή τον καταλληλότερο αριθμό χρονικών υστερήσεων. Σύμφωνα με τα παρακάτω αποτελέσματα, ο αριθμός των lags που επιλέγεται για τη συγκεκριμένη χρονοσειρά, είναι 14:

```
> ar(rainfall_top25)
Call:
ar(x = rainfall_top25)
Coefficients:
     1       2       3       4       5       6       7       8       9      10
0.0151  0.0630 -0.0761 -0.0839 -0.0424 -0.0453  0.0180 -0.0132 -0.0118  0.0082
    11      12      13      14
0.0574  0.1731  0.1966  0.0916
Order selected 14  sigma^2 estimated as  0.1718
```

Για τη δημιουργία των χρονικών αυτών υστερήσεων κατασκευάζουμε με τον παρακάτω κώδικα τις συναρτήσεις *lagmatrix* και *lag,* μέσω των οποίων θα μπορούμε να δημιουργήσουμε έναν πίνακα, που κάθε στήλη του θα αντιστοιχεί και σε ένα επιπλέον lag της χρονοσειράς. Ο συνολικός αριθμός των στηλών του πίνακα θα ισούται με $n+1$ (όπου n ο αριθμός των lags που θέλουμε κάθε φορά), γιατί στην πρώτη στήλη θα περιέχεται η αρχική χρονοσειρά και στις υπόλοιπες οι χρονικές υστερήσεις σε αύξουσα σειρά:

```
> lagmatrix <- function(x,max.lag){embed(c(rep(NA,max.lag),x),max.lag)}
>  lag <- function(x,lag) {
+  out<-lagmatrix(x,lag+1)[,lag]
+  return(out[1:length(out)-1]) }
```

Οι τιμές της συνάρτησης *lagmatrix,* για τα 14 lags που θέλουμε, αποθηκεύονται σε έναν πίνακα *Y,* από τον οποίο αφαιρούμε τις γραμμές με τις κενές τιμές (*NA*) που δημιουργούνται λόγω των lags. Αυτά επιτυγχάνονται, με την εισαγωγή των παρακάτω εντολών:

```
> Y<-lagmatrix(rainfall_top25,15)
> Y<-Y[16:dim(Y)[1],]
```

Δεν πρέπει όμως να ξεχνάμε, ότι στη δική μας περίπτωση έχουμε παρατηρήσει εξ'αρχής, από τα γραφήματα της αυτοσυσχέτισης και των δύο χρονοσειρών, την ύπαρξη περιοδικότητας. Άρα λοιπόν, δεν θα είχε νόημα να φτιάξουμε ένα μοντέλο από το οποίο θα έλειπε ο όρος που θα υποδείκνυε την υπάρχουσα περιοδικότητα, ώστε να είμαστε και σε θέση να ελέγξουμε, αν όντως καθορίζει την εμφάνιση υψηλής βροχόπτωσης. Για την εισαγωγή της περιοδικότητας στο μοντέλο μας, χρησιμοποιούμε τον παρακάτω όρο του συνημίτονου:

$$\cos\left(\frac{2\pi t}{12}\right)$$

Το διάνυσμα *t* αποτελείται από τις 252 χρονικές στιγμές της χρονοσειράς μας, μία για κάθε παρατήρηση. Επειδή όμως το τελικό διάνυσμα που θα συμμετέχει στο μοντέλο μας θα πρέπει να "συμβαδίζει" με τα lags, που θα χρησιμοποιήσουμε, αφαιρούμε τις πρώτες 14 χρονικές στιγμές ώστε να έχει την ίδια διάσταση με τις στήλες του πίνακα *Y*:

```
> t<-as.vector(seq(1:252))
> t<-(2*pi*t)/12
>cos<-cos(t)
> cos<-tail(cos,-14)
```

Μπορούμε τώρα να κατασκευάσουμε, το μοντέλο το οποίο θα περιέχει σαν μεταβλητή απόκρισης τη χρονοσειρά μας, η οποία περιέχεται στην πρώτη στήλη του πίνακα *Y* και σαν επεξηγηματικές μεταβλητές τις 14 χρονικές υστερήσεις, που μας υπέδειξε η εντολή *ar(rainfall_top25),* καθώς επίσης και τον όρο της περιοδικότητας. Θα έχουμε δηλαδή το παρακάτω λογιστικό παλινδρομικό μοντέλο:

$$\text{logit}(\pi_t(\beta)) = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_{14} Y_{t-14} + \beta_{15} \cos\left(\frac{2\pi t}{12}\right)$$

$$\text{Με } \beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{14}, \beta_{15})' \text{ και } Z_{t-1} = \left(1, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-14}, \cos\left(\frac{2\pi t}{12}\right)\right)'$$

Με τη χρήση των παρακάτω εντολών προκύπτουν τα αποτελέσματα για το παραπάνω μοντέλο:

```
>logit<-dyn$glm(Y[,1]~Y[,2]+Y[,3]+Y[,4]+Y[,5]+Y[,6]+Y[,7]+Y[,8]+Y[,9]+
 Y[,10]+Y[,11]+Y[,12]+Y[,13]+Y[,14]+Y[,15]+cos, family=binomial(link="logit"))

> summary(logit)
```

(Να εξηγήσουμε ότι η εντολή dyn$glm χρησιμοποιείται, όταν θέλουμε να κάνουμε χρήση κάποιου γενικευμένου γραμμικού μοντέλου και τα δεδομένα μας είναι χρονοεξαρτώμενα. Σε αυτήν την περίπτωση λόγω του τρόπου που ορίσαμε τον πίνακα με τα lags, θα είχαμε τα ίδια αποτελέσματα και με την απλή, γνωστή μας εντολή glm. Για τη χρήση της dyn$glm χρειάζονται οι libraries zoo και dyn.)

```
Call:
glm(formula = dyn(Y[, 1] ~ Y[, 2] + Y[, 3] + Y[, 4] + Y[, 5] + Y[, 6] + Y[, 7] + Y[, 8] +
Y[, 9] + Y[, 10] + Y[, 11] + Y[, 12] + Y[, 13] + Y[, 14] + Y[, 15] + cos),
family = binomial(link = "logit"))

Deviance Residuals:
   Min      1Q  Median    3Q     Max
-4.2818 -1.2668 -1.1003  0.6235 16.1760

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -1.869557   0.446466  -4.187   2.82e-05 ***
Y[, 2]      -0.340083   0.423710  -0.803   0.42219
Y[, 3]       0.240428   0.406815   0.591   0.55452
Y[, 4]      -0.624705   0.434278  -1.438   0.15029
Y[, 5]      -0.142350   0.450747  -0.316   0.75215
Y[, 6]       0.254197   0.463522   0.548   0.58341
Y[, 7]       0.073607   0.474496   0.155   0.87672
Y[, 8]       0.717011   0.453053   1.583   0.11351
Y[, 9]       0.306751   0.433603   0.707   0.47929
Y[, 10]     -0.122914   0.415602  -0.296   0.76742
Y[, 11]     -0.055389   0.402498  -0.138   0.89055
Y[, 12]     -0.009415   0.386817  -0.024   0.98058
Y[, 13]      0.514401   0.376230   1.367   0.17155
Y[, 14]      0.711136   0.385304   1.846   0.06494 .
Y[, 15]      0.358994   0.404067   0.888    0.37430
cos          1.440159   0.402850   3.575   0.00035 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance:      268.76  on 237  degrees of freedom
Residual deviance: 216.89  on 222  degrees of freedom
AIC: 248.89
Number of Fisher Scoring iterations: 5
```

Αποτελέσματα 4.1

Έτσι, η εκτιμήτρια μέγιστης μερικής πιθανοφάνειας, που προκύπτει για το β, είναι:

$$\hat{\beta} = (-1.869, -0.340, 0.240, \ldots, 0.711, 0.358, 1.440)'$$

Η στήλη των τυπικών σφαλμάτων των $\hat{\beta}$ (Std.errors), αποτελούν ουσιαστικά την τετραγωνική ρίζα της διαγωνίου του δεσμευμένου πίνακα πληροφορίας $G_N(\beta)$, για τον οποίο μιλήσαμε στο προηγούμενο κεφάλαιο και τον οποίο μπορούμε να δούμε, εισάγοντας την εντολή *>vcov(poisson).*

Από τα αποτελέσματα 4.1, συμπεραίνουμε πως η περιοδικότητα είναι αυτή που κατά κύριο λόγο επηρεάζει τη μεταβλητή απόκρισης, αφού είναι η μοναδική στατιστικά σημαντική μεταβλητή που προέκυψε κατά την παλινδρόμηση, ίσως μαζί με τη 13η χρονική υστέρηση η οποία φαίνεται να συμβάλει κι αυτή στο μοντέλο, σε μικρότερο όμως βαθμό. Η δοκιμή που κάναμε αφαιρώντας τον όρο του συνημίτονου και διατηρώντας μόνο τις χρονικές υστερήσεις, είχε ως αποτέλεσμα ένα μοντέλο με *AIC: 261.21,* μία τιμή η οποία είναι μεγαλύτερη από την αντίστοιχη του μοντέλου *(AIC: 248.89),* επιβεβαιώνοντας έτσι, πως η ύπαρξη της περιοδικότητας στο μοντέλο κρίνεται απαραίτητη.

Το παραπάνω όμως μοντέλο αποδεικνύεται μάλλον μη αποδοτικό, αφού διαθέτει αρκετές επεξηγηματικές μεταβλητές και επιπλέον οι περισσότερες από αυτές, δεν φαίνεται να συμβάλουν στη μοντελοποίηση της υψηλής βροχόπτωσης. Γι'αυτό το λόγο, θα χρησιμοποιήσουμε τη μέθοδο stepwise, η οποία μπορεί να μας υποδείξει το καταλληλότερο μοντέλο με βάση το κριτήριο AIC, ξεκινώντας από το αρχικό μας μοντέλο και αφαιρώντας σε κάθε βήμα *(step)* κάποιες επεξηγηματικές μεταβλητές. Έτσι καταλήγουμε σε ένα τελικό μοντέλο με τον "ιδανικότερο" συνδυασμό από τις αρχικές επεξηγηματικές μεταβλητές, αφαιρώντας τις περιττές και διατηρώντας αυτές που επηρεάζουν τη μεταβλητή μας. Για να γίνει αυτό, εισάγουμε τις παρακάτω εντολές:

(Για τη χρήση της μεθόδου stepAIC, χρειάζεται η βιβλιοθήκη MASS Δεν παραθέτονται τα αποτελέσματα κάθε step της διαδικασίας, που προκύπτουν από την πρώτη εντολή, αφού αυτό που μας αφορά, είναι το τελικό μοντέλο.)

```
> step <- stepAIC(logit, direction="backward")
> step$anova
Stepwise Model Path
Analysis of Deviance Table
Initial Model:
Y[, 1] ~ Y[, 2] + Y[, 3] + Y[, 4] + Y[, 5] + Y[, 6] + Y[, 7] +Y[, 8] + Y[, 9] + Y[, 10] +
Y[, 11] + Y[, 12] + Y[, 13] + Y[, 14] + Y[, 15] + cos

Final Model:
Y[, 1] ~ Y[, 4] + Y[, 8] + Y[, 14] + cos
```

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1			222	216.8890	248.8890
2 - Y[, 12]	1	0.0005926414	223	216.8896	246.8896
3 - Y[, 11]	1	0.0193583016	224	216.9089	244.9089
4 - Y[, 7]	1	0.0241478437	225	216.9331	242.9331
5 - Y[, 5]	1	0.0976987735	226	217.0308	241.0308
6 - Y[, 10]	1	0.1073241741	227	217.1381	239.1381
7 - Y[, 6]	1	0.2955086085	228	217.4336	237.4336
8 - Y[, 3]	1	0.2716251170	229	217.7052	235.7052
9 - Y[, 9]	1	0.3944354860	230	218.0997	234.0997
10- Y[, 2]	1	0.5861721404	231	218.6858	232.6858
11- Y[, 15]	1	0.8472538729	232	219.5331	231.5331
12- Y[, 13]	1	1.8692644373	233	221.4023	**231.4023**

Όπως μάλλον αναμενόταν, τα αποτελέσματα που προέκυψαν, υποδεικνύουν ένα καλύτερο μοντέλο με *AIC:231.40,* μικρότερο από αυτό του αρχικού μοντέλου *(AIC: 248.89)* και με πολύ λιγότερες επεξηγηματικές μεταβλητές. Οι μεταβλητές οι οποίες τελικά διατηρούνται στο μοντέλο είναι ο όρος της περιοδικότητας, η 3η, η 7η και η 13η χρονική υστέρηση. Άρα, καταλαβαίνουμε πως στην περίπτωση που μοντελοποιούσαμε τη χρονοσειρά μας μόνο με βάση την περιοδικότητα, θα είχαμε σίγουρα ένα λιγότερο αποδοτικό μοντέλο, αφού στο τελικό αποτέλεσμα της stepwise μεθόδου, στο οποίο αντιστοιχεί και η μικρότερη τιμή *AIC,* συνεχίζουν να περιέχονται και κάποιες χρονικές υστερήσεις.

Έχοντας καταλήξει σε ένα πιο εύχρηστο και ευέλικτο μοντέλο απ'ότι το αρχικό, με τις επεξηγηματικές μεταβλητές οι οποίες όντως επηρεάζουν τη μεταβλητή απόκρισης, κατασκευάζουμε ξανά το μοντέλο με τις αντίστοιχες μεταβλητές, ώστε να μπορέσουμε να ελέγξουμε την καταλληλότητα του για την περιγραφή της χρονοσειράς μας. Έχουμε έτσι, το εξής μοντέλο με τα αποτελέσματα 4.2:

```
>logit2<-dyn$glm(Y[,1]~Y[,4]+Y[,8]+Y[,14]+cos, family=binomial(link="logit"))
> summary(logit2)
```

```
Call:
glm(formula = dyn(Y[, 1] ~ Y[, 4] + Y[, 8] + Y[, 14] + cos),
   family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8084  -1.2353  -1.0993   0.6941  19.2643

Coefficients:
             Estimate   Std. Error z value   Pr(>|z|)
(Intercept)  -1.5809     0.2686  -5.886    3.95e-09 ***
Y[, 4]       -0.5950     0.4093  -1.454    0.1460
Y[, 8]        0.7256     0.4339   1.672    0.0945 .
Y[, 14]       0.7153     0.3651   1.959    0.0501 .
cos           1.4581     0.2924   4.986    6.15e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance:         268.76  on 237  degrees of freedom
Residual deviance: 221.40  on 233  degrees of freedom
AIC: 231.40
Number of Fisher Scoring iterations: 5
```

Αποτελέσματα 4.2

Όπως έχουμε εξηγήσει και στο πρώτο κεφάλαιο, στην παράγραφο της λογιστικής παλινδρόμησης, όταν η μεταβλητή απόκρισης είναι δίτιμη, όπως στη συγκεκριμένη περίπτωση, ο έλεγχος μέσω της συνάρτησης Deviance δε μπορεί να χρησιμοποιηθεί, καθώς δε μπορεί να αποτελέσει δείκτη καλής προσαρμογής του μοντέλου συγκριτικά με το κορεσμένο. (Οικονόμου & Καρώνη, 2010; Collett, 2003).

Παρ'όλα αυτά, μπορούμε να χρησιμοποιήσουμε την ελεγχοσυνάρτηση Deviance για τη σύγκριση δύο μοντέλων, μη κορεσμένων, ακόμα και στην περίπτωση της δίτιμης εξαρτημένης μεταβλητής. Θα ήταν λοιπόν χρήσιμο να ελέγξουμε, πέραν της stepwise μεθόδου και της τιμής του *AIC*, αν όντως το δεύτερο μοντέλο *logit2*, είναι καταλληλότερο σε σχέση με το αρχικό και αν πράγματι μπορεί να επιλεχθεί, μέσω των μεταβολών της Deviance που θα προκύψουν για τα δύο αυτά μοντέλα. Εισάγοντας την παρακάτω εντολή έχουμε και τα αντίστοιχα αποτελέσματα:

```
> anova.glm(logit2, logit, test="Chisq")

Analysis of Deviance Table
Model 1: Y[, 1] ~ Y[, 4] + Y[, 8] + Y[, 14] + cos

Model 2: Y[, 1] ~ Y[, 2] + Y[, 3] + Y[, 4] + Y[, 5] + Y[, 6] + Y[, 7] + Y[, 8] + Y[, 9] +
Y[,10]+ Y[, 11] + Y[, 12] + Y[, 13] + Y[, 14] + Y[, 15] + cos

  Resid. Df Resid.   Dev     Df  Deviance P(>|Chi|)
1       233   221.402
2       222   216.889  11  4.513       0.9524
```

Από τα παραπάνω αποτελέσματα, συμπεραίνουμε ότι η προσθήκη όλων των υπόλοιπων χρονικών υστερήσεων, οδηγεί μεν σε μια μικρότερη τιμή της Deviance αλλά χωρίς να δημιουργείται κάποια στατιστικά σημαντική διαφοροποίηση. Άρα το μοντέλο στο οποίο οδηγηθήκαμε τελικά, μέσω της stepwise μεθόδου, δείχνει να είναι ικανοποιητικό, με την έννοια του ότι η προσαρμογή της χρονοσειράς μας σε αυτό, δε φαίνεται να διαφέρει από αυτήν του αρχικού μοντέλου, παρά το ότι αποτελούνταν από αρκετά περισσότερες επεξηγηματικές μεταβλητές.

Καταλήγουμε λοιπόν στο συμπέρασμα, ότι η περιγραφή του φαινομένου της υψηλής βροχόπτωσης μπορεί να στηριχτεί κατά βάση στην περιοδικότητα εμφάνισης του μεταξύ των ετών, αλλά και σε κάποιες συγκεκριμένες τρείς χρονικές υστερήσεις, με την προσθήκη των οποίων οδηγούμαστε σε ένα αποδοτικότερο μοντέλο. Ερμηνευτικά βέβαια, ίσως θα ήταν καλύτερο, το μοντέλο που εκτός της περιοδικότητας θα περιείχε μόνο την 13η χρονική υστέρηση, η οποία αρχικά ήταν και η μοναδική στατιστικά σημαντική μεταβλητή του μοντέλου και όχι και τις άλλες δύο (3η και 7η). Γιατί τότε, η υψηλή βροχόπτωση σε κάθε μήνα θα καθοριζόταν από την περιοδικότητα γενικότερα, και από το τις παρατηρούμενες τιμές του ακριβώς προηγούμενου χρόνου ειδικότερα (μπορούμε προσεγγιστικά να πούμε ότι η η προηγούμενη 13η χρονική περίοδος, δηλαδή ο προηγούμενος 13ος μήνας, συμπίπτει με την αντίστοιχη περίοδο του προηγούμενου έτους κάθε φορά), το οποίο θα οδηγεί σε μια πιο "κατανοητή" ερμηνεία του φαινομένου. Στην περίπτωση όμως που επιλέξουμε το συγκεκριμένο μοντέλο, τότε θα έχουμε *AIC:234.43,* η τιμή του οποίου είναι κατά λίγο μεγαλύτερη από αυτή του μοντέλου που περιέχει και τις άλλες δύο χρονικές υστερήσεις *(AIC: 231.40).* Τα αποτελέσματα του μοντέλου αυτού, παραθέτονται παρακάτω:

```
> logit3<-dyn$glm(Y[,1]~Y[,13]+cos, family=binomial(link="logit"))
> summary(logit3)
```

```
Call:
glm(formula = dyn(Y[, 1] ~ Y[, 13] + cos), family = binomial(link = "logit"))

Deviance Residuals:
    Min      1Q      Median      3Q      Max
-2.4682  -1.2295  -1.1180   0.9957  17.4753
Coefficients:
             Estimate Std. Error z value    Pr(>|z|)
(Intercept)  -1.4717    0.2102   -7.001   2.54e-12 ***
Y[, 13]       0.5256    0.3572    1.472   0.141
cos           1.3301    0.2757    4.825   1.40e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance:       268.76  on 237  degrees of freedom
Residual deviance: 228.43  on 235  degrees of freedom
AIC: 234.43
Number of Fisher Scoring iterations: 5
```

Αποτελέσματα 4.3

Η διαφορά που υπάρχει στην τιμή του κριτηρίου *AIC* μεταξύ των δύο μοντέλων δεν είναι τόσο μεγάλη, ώστε να κριθεί "απαγορευτική" η χρήση αυτού του μοντέλου συγκριτικά με το προηγούμενο (Αποτελέσματα 4.2), που περιείχε και τις άλλες δύο χρονικές υστερήσεις. Η επιλογή του μοντέλου καθορίζεται από τις ανάγκες της εκάστοτε μελέτης και σύμφωνα με τα αποτελέσματα που έχουν προκύψει, θα μπορούσαν να χρησιμοποιηθούν πιθανότατα και τα δύο προαναφερθέντα μοντέλα. Ωστόσο, παραθέτουμε απλώς την αντίστοιχη σύγκριση των τιμών της ελεγχοσυνάρτησης Deviance μεταξύ των προαναφερθέντων μοντέλων, όπως κάναμε και παραπάνω, η οποία όμως αυτή τη φορά επιβεβαιώνει ότι η προσθήκη και των άλλων χρονικών υστερήσεων οδηγεί σε ένα πιο "ακριβές" μοντέλο για τη χρονοσειρά μας:

```
> anova.glm(logit3,logit2,test="Chisq")
Analysis of Deviance Table
Model 1: Y[, 1] ~ Y[, 13] + cos

Model 2: Y[, 1] ~ Y[, 4] + Y[, 8] + Y[, 14] + cos

  Resid. Df Resid.   Dev    Df  Deviance  P(>|Chi|)
1         235     228.43
2         233     221.40    2    7.03       0.02974 *
```

Πριν προχωρήσουμε στην αντίστοιχη διαδικασία για τη δεύτερη χρονοσειρά, που αφορά την πολύ υψηλή βροχόπτωση, έχει ενδιαφέρον να δούμε τι θα συνέβαινε, εάν δεν είχαμε επιλέξει αρχικά ένα αυτοπαλινδρομούμενο μοντέλο με έναν όρο περιοδικότητας για τη μοντελοποίηση της χρονοσειράς μας, αλλά ένα λογιστικό παλινδρομικό μοντέλο που θα περιείχε ως επεξηγηματικές μεταβλητές κάθε μήνα χωριστά *(dummy variables).* Ένα τέτοιο μοντέλο μπορούμε να κατασκευάσουμε με τη χρήση των παρακάτω εντολών, και συγκεκριμένα μέσω της

seasonaldummy(rainfall_top25), η οποία "απομονώνει" τις εποχές, στις οποίες είναι χωρισμένη η χρονοσειρά μας, δηλαδή τους μήνες του έτους και τις αποθηκεύει στη μεταβλητή *season*:

```
> season<-seasonaldummy(rainfall_top25)
```
(Για τη χρήση της εντολής seasonaldummy χρειάζεται η βιβλιοθήκη forecast)

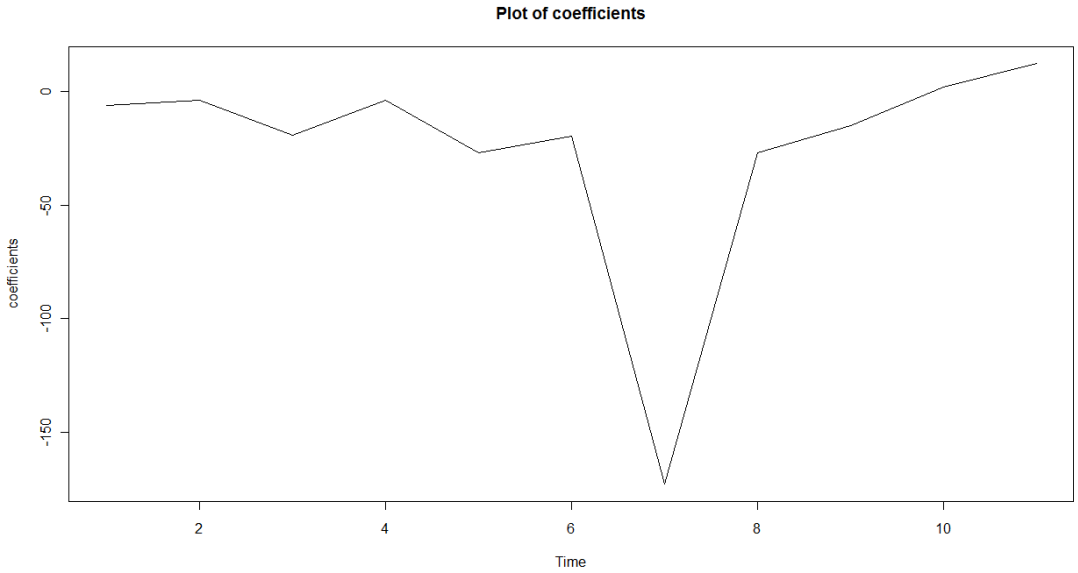
Κι έτσι προκύπτει το εξής μοντέλο:
```
> logit<-glm(rainfall_top25~season, family=binomial(link="logit"))
> summary(logit)
```

```
Call:
glm(formula = rainfall_top25 ~ season, family = binomial(link = "logit"))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.5829  -0.8203  -0.4474   0.2049   2.4676

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2877     0.4410  -0.652   0.5141
seasonJan    -0.6286     0.6540  -0.961   0.3365
seasonFeb    -0.4055     0.6393  -0.634   0.5259
seasonMar    -1.9636     0.8643  -2.272   0.0231 *
seasonApr    -0.4055     0.6393  -0.634   0.5259
seasonMay    -2.7081     1.1155  -2.428   0.0152 *
seasonJun    -1.9636     0.8643  -2.272   0.0231 *
seasonJul   -17.2784   863.309 -0.020   0.9840
seasonAug    -2.7081     1.1155  -2.428  0.0152 *
seasonSep    -1.5041     0.7638  -1.969   0.0489 *
seasonOct     0.1924     0.6208   0.310   0.7566
seasonNov     1.2040     0.6540   1.841   0.0657 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance:        283.42  on 251  degrees of freedom
Residual deviance: 221.19  on 240  degrees of freedom
AIC: 245.19
Number of Fisher Scoring iterations: 16
```

Αποτελέσματα 4.4

Σύμφωνα με τα αποτελέσματα 4.4, σίγουρα δεν πρόκειται για καλύτερο μοντέλο συγκριτικά με τα προηγούμενα δύο μοντέλα, στα οποία καταλήξαμε με τη χρήση των μεταβλητών της περιοδικότητας και των χρονικών υστερήσεων, αφού η τιμή του κριτηρίου AIC είναι εμφανώς μεγαλύτερη. Παρ'όλα αυτά, μπορούμε να χρησιμοποιήσουμε τους εκτιμώμενους συντελεστές, που προκύπτουν για τον κάθε μήνα και να δημιουργήσουμε το αντίστοιχο γράφημα, ώστε να παρατηρήσουμε το πώς επηρεάζει κάθε μήνας με βάση το συντελεστή του, την ύπαρξη της υψηλής βροχόπτωσης και να ελέγξουμε αν η ερμηνεία που θα προκύψει με βάση το

συγκεκριμένο μοντέλο, συμπίπτει με αυτή που είχε βασιστεί στις εκτιμήσεις της περιοδικότητας για τον κάθε μήνα:



Γράφημα 4.3

Σύμφωνα με το γράφημα 4.3, όπως ήταν και το αναμενόμενο, οι μήνες στους οποίους παρατηρείται συντελεστής με θετικό πρόσημο, άρα συμβάλουν θετικά στην ύπαρξη του φαινομένου της υψηλής βροχόπτωσης, είναι ο Οκτώβριος, ο Νοέμβριος και ο Δεκέμβριος. Μάλιστα από το Σεπτέμβριο και μετά, μέχρι και το Νοέμβριο υπάρχει μια συνεχόμενα ανοδική πορεία τις τιμής των συντελεστών που αντιστοιχούν στον κάθε μήνα. Αντίθετα, από το Φεβρουάριο και μετά, έχουμε μια καθοδική πορεία στην ήδη αρνητική τιμή των συντελεστών, με μια μικρή εξαίρεση κατά τον Απρίλιο. Η ακραία τιμή της καθοδικής αυτής πορείας, παρατηρείται κατά τον Ιούλιο όπου εκεί το φαινόμενο είναι ανύπαρκτο και στα 21 έτη που μελετάμε.

Η μεγάλη αυτή απόκλιση του γραφήματος που παρατηρούμε στον Ιούλιο, μπορεί να εξηγηθεί αν σκεφτούμε ότι κατά τη περίοδο της παρατήρησης, δεν υπήρξε καθόλου υψηλή βροχόπτωση κατά τον Ιούλιο, πράγμα που, σύμφωνα με τα όσα έχουμε αναφέρει στην προηγούμενη παράγραφο 4.1, σημαίνει ότι για το συγκεκριμένο αυτόν μήνα:

$$\pi = P(1) = 0$$

Άρα με βάση τη σχέση (4.2)*:*

$$\text{logit}(\pi_t(\beta)) = \log\left(\frac{\pi_t}{1-\pi_t}\right) = \mathrm{Z}'_{t-1}\beta$$

καταλαβαίνουμε ότι ο συντελεστής της ψευδομεταβλητής του Ιουλίου στο μοντέλο μας, θα πρέπει να απειρίζεται, προκαλώντας έτσι και την αντίστοιχη εικόνα που έχουμε στη γραφική παράσταση. Την "προβληματική" αυτή συμπεριφορά του συγκεκριμένου μήνα, υποδεικνύει και το ιδιαίτερα μεγάλο τυπικό σφάλμα που βλέπουμε στα αποτελέσματα του μοντέλου *(863.309).*

Καταλήγουμε προφανώς στο συμπέρασμα, ότι η μοντελοποίηση της χρονοσειράς μας μέσω ενός τέτοιου μοντέλου δε θα μπορούσε να γίνει, καθώς αντιμετωπίζει προβλήματα προσαρμογής σε αυτό.

Την ίδια ακριβώς διαδικασία ακολουθούμε και για την χρονοσειρά της ακόμα πιο υψηλής βροχόπτωσης. Ξεκινώντας με τη γνωστή εντολή *ar,* θα δούμε αν μπορούμε στη συγκεκριμένη περίπτωση να χρησιμοποιήσουμε κάποιο αυτοπαλινδρομούμενο μοντέλο, και αν ναι, πόσες χρονικές υστερήσεις θα περιέχει το αρχικό μοντέλο.

```
> ar(rainfall_top10)
Call:
ar(x = rainfall_top10)
Order selected 0  sigma^2 estimated as  0.08972
```

Σύμφωνα με τα αποτελέσματα που προέκυψαν στη συγκεκριμένη χρονοσειρά, μάλλον δεν έχει νόημα να προσαρμόσουμε μοντέλο, το οποίο θα στηρίζεται σε χρονικές υστερήσεις. Για να είμαστε όμως σίγουροι, δοκιμάσαμε ένα αυτοπαλινδρομούμενο λογιστικό μοντέλο με τον ίδιο τρόπο όπως και παραπάνω, αλλά και έτσι, καμία χρονική υστέρηση δεν αποτέλεσε στατιστικά σημαντική μεταβλητή ακόμα και χωρίς την ύπαρξη άλλης επεξηγηματικής μεταβλητής όπως ο όρος της περιοδικότητας. Αυτό ίσως και να ήταν αναμενόμενο αν θυμηθούμε το γράφημα αυτοσυσχέτισης της χρονοσειράς.

Άρα, επόμενο βήμα είναι να κατασκευάσουμε ένα μοντέλο, το οποίο θα στηρίζεται αποκλειστικά στην περιοδικότητα, η οποία υπενθυμίζουμε ότι έχει παρατηρηθεί και για τη χρονοσειρά της ιδιαίτερα υψηλής βροχόπτωσης, αν και σε μικρότερο βαθμό από ότι στην περίπτωση της υψηλής βροχόπτωσης. Για να ελέγξουμε αν όντως μπορεί να μοντελοποιηθεί η συγκεκριμένη χρονοσειρά με βάση τον όρο της περιοδικότητας, δημιουργούμε το αντίστοιχο μοντέλο:

```
> logit<-dyn$glm(rainfall_top10~cos, family=binomial(link="logit"))
```

```
Call:
glm(formula = dyn(rainfall_top10 ~ cos), family = binomial(link = "logit"))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.7585  -0.5350  -0.2539  -0.1920   2.6270

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  -2.6451     0.3065   -8.631   < 2e-16 ***
cos           1.5464     0.4097    3.774  0.000160 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance:        162.96  on 251  degrees of freedom
Residual deviance: 143.35  on 250  degrees of freedom
AIC: 147.35
Number of Fisher Scoring iterations: 6
```

Αποτελέσματα 4.5

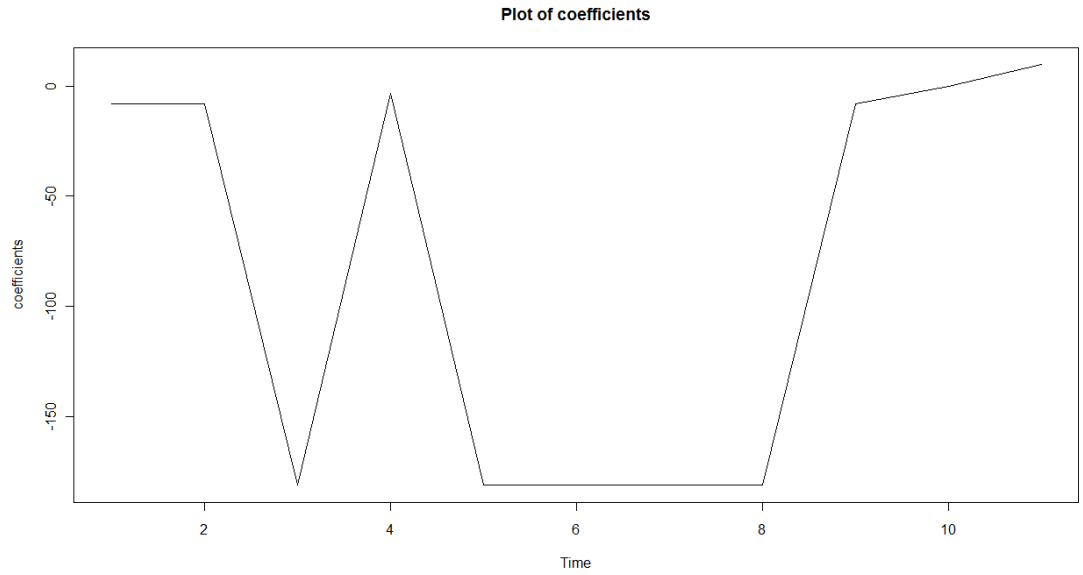
Από τα αποτελέσματα 4.5, συμπεραίνουμε ότι η περιοδικότητα πράγματι επηρεάζει το φαινόμενο της πολύ υψηλής βροχόπτωσης και ότι η επιλογή ενός λογιστικού παλινδρομικού μοντέλου με έναν όρο περιοδικότητας, φαίνεται να είναι ικανοποιητική για τη χρονοσειρά μας.

Γνωρίζουμε πλέον, από την προηγούμενη περίπτωση, της υψηλής βροχόπτωσης, ότι ένα μοντέλο που θα περιέχει ως επεξηγηματικές μεταβλητές τη ψευδομεταβλητή του κάθε μήνα *(dummy variables),* θα παρουσιάσει προβλήματα, υποθέτοντας έτσι ότι ένα αντίστοιχο μοντέλο δε θα μπορούσε να επιλεχθεί ούτε γι' αυτήν τη χρονοσειρά, της υψηλότερης βροχόπτωσης. Παρ'όλα αυτά παραθέτουμε τα αποτελέσματα που θα είχε το συγκεκριμένο μοντέλο σε αυτή την περίπτωση, καθώς και το διάγραμμα των συντελεστών που προκύπτουν για τον κάθε μήνα, απλά και μόνο για να επιβεβαιώσουμε την υπόθεση μας:

```
> season<-seasonaldummy(rainfall_top10)
> logit<-glm(rainfall_top10~season, family=binomial(link="logit"))
```

```
Call:
glm(formula = rainfall_top10 ~ season, family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-9.794e-01  -5.552e-01  -7.976e-05  -7.976e-05   2.169e+00

Coefficients:
             Estimate   Std. Error    z value    Pr(>|z|)
(Intercept) -1.447e+00  5.557e-01   -2.604     0.00922 **
seasonJan   -8.044e-01   9.281e-01   -0.867     0.38614
seasonFeb   -8.044e-01   9.281e-01   -0.867     0.38614
seasonMar   -1.812e+01  2.347e+03   -0.008     0.99384
seasonApr   -3.448e-01   8.353e-01   -0.413     0.67973
seasonMay   -1.812e+01  2.347e+03    -0.008     0.99384
seasonJun   -1.812e+01  2.347e+03    -0.008     0.99384
seasonJul    -1.812e+01  2.347e+03    -0.008     0.99384
seasonAug   -1.812e+01  2.347e+03    -0.008     0.99384
seasonSep    -8.044e-01   9.281e-01    -0.867     0.38614
seasonOct    1.827e-15   7.859e-01   2.32e-15 1.00000
seasonNov   9.614e-01   7.147e-01     1.345     0.17854
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance:         162.96  on 251  degrees of freedom
Residual deviance: 125.66  on 240  degrees of freedom
AIC: 149.66
Number of Fisher Scoring iterations: 18
```

Από το γράφημα 4.4, επιβεβαιώνουμε πως την ίδια "ιδιαίτερη" συμπεριφορά που είχε ο Ιούλιος στην προηγούμενη περίπτωση, εδώ έχουν και άλλοι μήνες κατά τους οποίους το φαινόμενο έχει μηδενικές παρατηρήσεις:



Γράφημα 4.4

Το φαινόμενο της βροχόπτωσης γενικότερα είναι ένα πολύ σύνθετο φαινόμενο και εξαρτάται από πολλές παραμέτρους, καθιστώντας πολύ δύσκολή τη μοντελοποίηση του. Από όλη την ανάλυση που προηγήθηκε, επιβεβαιώνεται πως δεν είναι εύκολη η μοντελοποίηση των βροχοπτώσεων μόνο μέσω αυτοπαλινδρομούμενων λογιστικών μοντέλων, βασιζόμενοι δηλαδή μόνο στο χρόνο και στην όποια περιοδικότητα παρουσιάζει η εκάστοτε χρονοσειρά. Καταλαβαίνουμε πως η προσαρμογή μοντέλων, που θα περιείχαν και άλλες συμμεταβλητές, όπως θερμοκρασία, ατμοσφαιρική πίεση κλπ, θα ήταν πιθανόν πολύ καλύτερη.

Παρ'όλα αυτά η χρήση του λογιστικού παλινδρομικού μοντέλου με τη βοήθεια των χρονικών υστερήσεων και του όρου της περιοδικότητας μπορεί να μην είναι αρκετό για μια ακριβή πρόβλεψη, αλλά όπως είδαμε, μπορεί σίγουρα να περιγράψει σε κάποιο βαθμό τη διαχρονική συμπεριφορά των φαινομένων της υψηλής και της πολύ υψηλής βροχόπτωσης. Συγκεκριμένα για τις δύο περιπτώσεις, μπορούμε να πούμε ότι η συμπεριφορά της χρονοσειράς της υψηλής βροχόπτωσης μπορεί να περιγραφεί ικανοποιητικά μέσω συγκεκριμένων χρονικών υστερήσεων και του συνημιτονοειδούς όρου της περιοδικότητας, ενώ αντίστοιχα για τη χρονοσειρά της πολύ υψηλής βροχόπτωσης, στηριχτήκαμε αποκλειστικά στην περιοδικότητα του φαινομένου. Η πολύ υψηλή βροχόπτωση φάνηκε να είναι, ένα μάλλον πιο "απρόβλεπτο" φαινόμενο συγκριτικά με την υψηλή βροχόπτωση, αφού οι μεταπτώσεις μεταξύ των μηνών είναι αρκετά εντονότερες, το οποίο ίσως και να ήταν αναμενόμενο αν σκεφτούμε πως η ιδιαίτερα υψηλή βροχόπτωση δεν είναι ένα πολύ συνηθισμένο φαινόμενο για τα ελληνικά δεδομένα, με την έννοια του ότι δεν αποτελεί βασικό και βέβαιο χαρακτηριστικό κάποια εποχής το οποίο μπορούμε να πούμε ότι συμβαίνει με την περίπτωση της υψηλής βροχόπτωσης. Τέλος, να αναφέρουμε ότι στην περίπτωση που είχαμε στη διάθεση μας ημερήσια δεδομένα των μηνών, θα είχε ενδιαφέρον ο διαχωρισμός των χειμερινών και θερινών μηνών για την εύρεση, πιθανώς κάποιων καλύτερων μοντέλων για κάθε περίπτωση, πράγμα που δεν μπορεί να συμβεί στην παρούσα εφαρμογή καθώς τα δεδομένα μας είναι μηνιαία.

4.3 POISSON ΜΟΝΤΕΛΟ ΓΙΑ ΑΠΑΡΙΘΜΗΤΕΣ ΧΡΟΝΟΣΕΙΡΕΣ

4.3.1 Δομή μοντέλου Poisson

Αν υποθέσουμε ότι μας ενδιαφέρει ο αριθμός εμφάνισης κάποιου γεγονότος διαχρονικά, τότε η μεταβλητή απόκρισης του προβλήματος μας, θα αντιστοιχεί σε μια χρονοσειρά Y_t, η οποία λόγω της φύσεως των παρατηρήσεων *(counts)*, θα αποτελείται από ακέραιες και μη αρνητικές τιμές. Έχοντας υπόψη μας το πρώτο κεφάλαιο, ξέρουμε ότι σε αυτή την περίπτωση η δεσμευμένη κατανομή $Y_t \mid F_{t-1}$ ακολουθεί την κατανομή Poisson η οποία μπορεί να γραφεί στην εξής μορφή:

$$f(y_t; \mu_t \mid F_{t-1}) = \exp\{(y_t \log \mu_t - \mu_t) + \log y_t!\},\ t = 1, \ldots, N$$

και με την ιδιότητα:

$$\mathrm{E}[Y_t \mid F_{t-1}] = Var[Y_t \mid F_{t-1}] = \mu_t$$

όπου F_{t-1} η σ-άλγεβρα που περιέχει οτιδήποτε είναι γνωστό στον παρατηρητή σχετικά με το γεγονός, μέχρι και τη στιγμή $t-1$.

Για την κατασκευή του μοντέλου, επιλέγουμε και πάλι την κανονική συνάρτηση σύνδεσης, η οποία στην περίπτωση της κατανομής poisson είναι ο σύνδεσμος *log*:

$$\theta_t = g(\mu_t) = \log(\mu_t)$$

και στηριζόμενοι στη σχέση *(3.8)* του τρίτου κεφαλαίου, θα προκύψει η μορφή του παλινδρομικού μοντέλου Poisson για τη χρονοσειρά $\{Y_t\}$, $t = 1, \ldots, N$ και μια συγκεκριμένη διανυσματική ανέλιξη $\{Z_{t-1}\}$, $t = 1, \ldots, N$:

$$\boxed{\mu_t(\beta) = \exp(Z'_{t-1}\beta)} \tag{4.3}$$

όπου β αποτελεί προφανώς το διάνυσμα των παραμέτρων από το οποίο εξαρτάται το μοντέλο μας.

Όπως ίσχυε και για τις δυαδικές χρονοσειρές, υπάρχουν και σε αυτή την περίπτωση, των απαριθμητών χρονοσειρών, διαφορετικοί τρόποι μοντελοποίησης τους πέραν του μοντέλου Poisson που προέκυψε με τη χρήση της κανονικού συνδέσμου. Σε εναλλακτικά μοντέλα, κατάλληλα για μια απαριθμητή χρονοσειρά, μπορούμε να καταλήξουμε είτε μέσω ενός άλλου συνδέσμου (πχ $h(\eta_t) = \sqrt{\eta_t}$), είτε μέσω κάποιας τροποποιημένης μορφής του μοντέλου Poisson (πχ μοντέλο Zeger-Qaqish). Στην εφαρμογή που ακολουθεί, εμείς θα χρησιμοποιήσουμε το μοντέλο poisson, του οποίου η μορφή δίνεται μέσω της σχέσης (4.3).

Όσον αφορά τη συμπερασματολογία, τους ελέγχους υποθέσεων και τους διαγνωστικούς ελέγχους, ισχύουν τα όσα έχουμε εξηγήσει αναλυτικά στο τρίτο κεφάλαιο, στην παράγραφο 3.5.

4.3.2 Εφαρμογή παλινδρομικού μοντέλου Poisson για απαριθμητή χρονοσειρά

Στη παρούσα εφαρμογή θα προσπαθήσουμε με αντίστοιχο τρόπο, όπως και στην περίπτωση της δυαδικής χρονοσειράς που προηγήθηκε, να επιλέξουμε κάποιο μοντέλο για χρονοσειρά, της οποίας οι παρατηρήσεις έχουν να κάνουν με τον αριθμό εμφάνισης κάποιου γεγονότος. Έχοντας αναλύσει το θεωρητικό υπόβαθρο του παλινδρομικού μοντέλου Poisson, στην περίπτωση που η μεταβλητή μας είναι μια χρονοσειρά $\{Y_t\}$ και λαμβάνοντας πάντα υπόψη τα όσα έχουμε αναφέρει περί χρονοσειρών στο Κεφάλαιο 2, είμαστε σε θέση να κατανοήσουμε στην πράξη την προσαρμογή μιας απαριθμητής χρονοσειράς, σε ένα τέτοιο μοντέλο.

Τα δεδομένα μας έχουν να κάνουν με τον αριθμό αυτοκτονιών ανά μήνα στην Ελλάδα, κατά τη διάρκεια 11 ετών, από το 1997 έως και το 2007. Εκτός από το συνολικό αριθμό των αυτοκτονιών έχουμε στη διάθεση μας χωριστά τον αριθμό των αυτοκτονιών των ανδρών και των γυναικών στον κάθε μήνα αντίστοιχα. Μπορούμε λοιπόν να κατασκευάσουμε τρείς χρονοσειρές, μια για κάθε περίπτωση, ώστε να δούμε και τι ακριβώς συμβαίνει στο σύνολο των αυτοκτονιών διαχρονικά, αλλά και να ελέγξουμε, αν υπάρχουν διαφορές μεταξύ των αυτοκτονιών των ανδρών και των γυναικών.

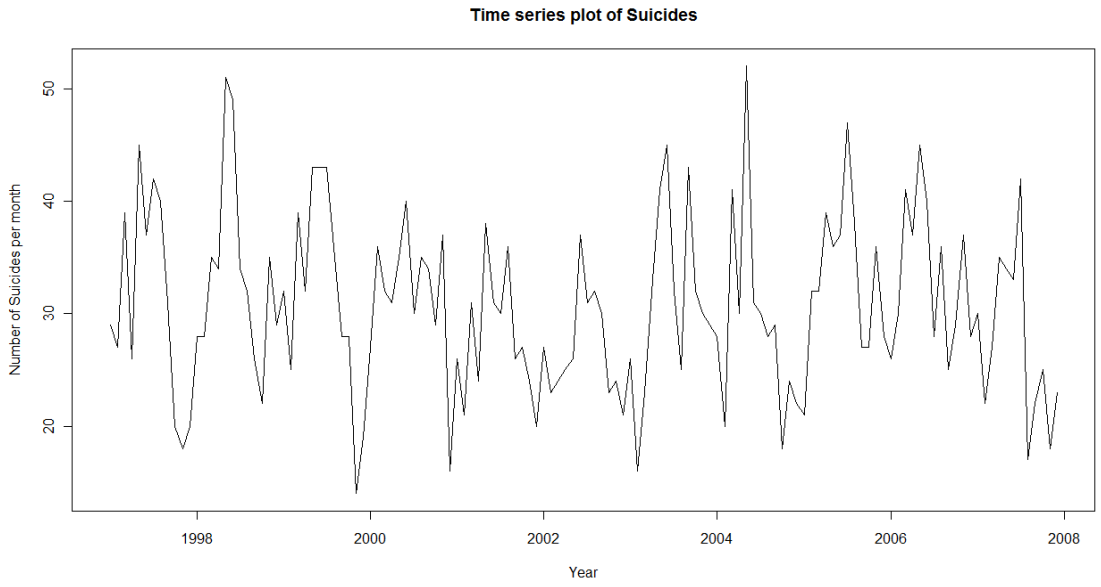
Αρχικά, πριν φτάσουμε στη διαμόρφωση κάποιου μοντέλου, πρέπει να δούμε κάποια απαραίτητα περιγραφικά γραφήματα ώστε να αντιληφθούμε τη συμπεριφορά των χρονοσειρών μας. Η επεξεργασία των δεδομένων μας θα γίνει και πάλι μέσω της R, εισάγοντας τα δεδομένα μας και δημιουργώντας ένα πλαίσιο, του οποίου κάθε στήλη αντιστοιχεί και σε μία χρονοσειρά. Ξεκινάμε, κατασκευάζοντας τη χρονοσειρά των συνολικών αυτοκτονιών ανά μήνα, μεγέθους 132 παρατηρήσεων:

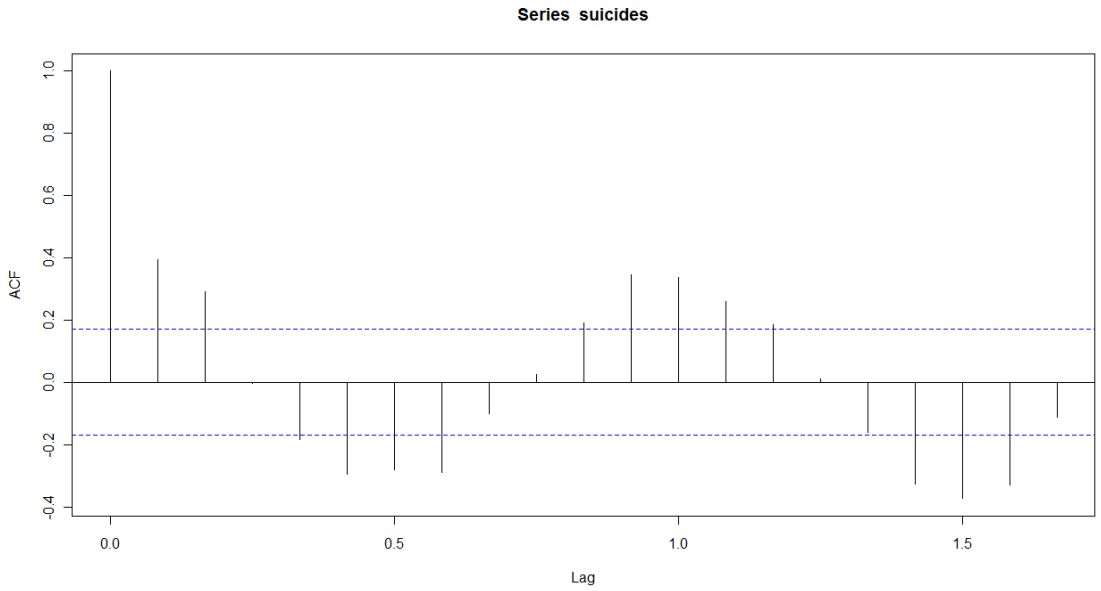
```
> y<-dataframe[,1]
> suicides<-ts(y,start=1997,frequency=12)
> suicides
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
1997  29  27  39  26  45  37  42  40  31  20  18  20
1998  28  28  35  34  51  49  34  32  26  22  35  29
1999  32  25  39  32  43  43  43  35  28  28  14  19
```

(Παραθέτουμε μόνο τα τρία πρώτα έτη, απλώς για να δούμε τη μορφή της χρονοσειράς. Πηγή δεδομένων: ΕΛ.ΣΤΑΤ.)

Πρώτο βήμα, όπως πάντα, είναι η γραφική αναπαράσταση της χρονοσειράς και το γράφημα της αυτοσυσχέτισης της, που προκύπτουν αντίστοιχα μέσω των παρακάτω εντολών :

```
> ts.plot(suicides,xlab="Year",ylab="Number of Suicides per month")
> acf(suicides,lag.max=20,type="correlation")
```

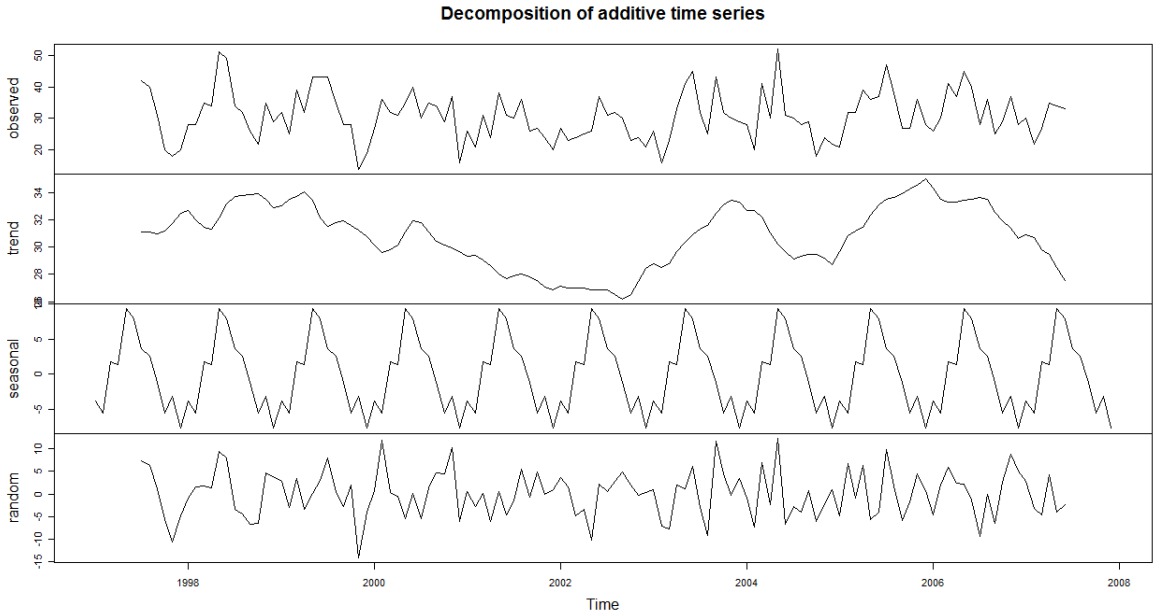




Το γράφημα της χρονοσειράς μάς υποδεικνύει μια προφανή περιοδικότητα των δεδομένων, πράγμα που επιβεβαιώνεται μέσω της ημιτονοειδούς μορφής της γραφικής παράστασης της αυτοσυσχέτισης. Δεν υπάρχει κάποια συγκεκριμένη τάση στην χρονοσειρά, όπως είναι και το λογικό, καθώς μεταξύ των μηνών και των ετών υπάρχει συνεχής αυξομείωση στον αριθμό των αυτοκτονιών. Στο γράφημα της αυτοσυσχέτισης, όπως έχουμε εξηγήσει ξανά, οι δύο οριζόντιες γραμμές που παρατηρούμε, μας δείχνουν το 95% διάστημα εμπιστοσύνης για τον έλεγχο της υπόθεσης $H_0 : \rho = 0$ (να έχουμε δηλαδή μηδενική αυτοσυσχέτιση), πράγμα που εδώ εύκολα φαίνεται πως απορρίπτεται.

Η εμφανής περιοδικότητα της χρονοσειράς μπορεί να παρατηρηθεί και μέσω της ακόλουθης εντολής, από την οποία προκύπτει το γράφημα 4.5 που περιέχει τις γραφικές αναπαραστάσεις της συνιστώσας της τάσης, της περιοδικότητας και του θορύβου.

```
> plot(decompose(suicides))
```



Γράφημα 4.5

Έχει επίσης ενδιαφέρον, να δούμε αναλυτικά τις εκτιμήσεις της περιοδικότητας, από τις οποίες προκύπτει και το παραπάνω γράφημα, ώστε να βγάλουμε κάποια συμπεράσματα για το πώς κυμαίνονται διαχρονικά οι παρατηρήσεις μας. Με την πρώτη εντολή *decompose(suicides)*, αποθηκεύουμε τις εκτιμήσεις όλων των συνιστωσών της χρονοσειράς στη μεταβλητή *components*, ενώ με τη χρήση της *components$seasonal,* εμφανίζουμε μόνο τα αποτελέσματα για τη συνιστώσα της περιοδικότητας, που μας ενδιαφέρει να παρατηρήσουμε:

```
> components <- decompose(suicides)
> components$seasonal
```

	Jan	Feb	Mar	Apr	May	Jun
1997	-3.805	-5.509	1.823	1.340	9.419	7.906
	Jul	Aug	Sep	Oct	Nov	Dec
1997	3.606	2.623	-1.105	-5.493	-3.084	-7.722

Πίνακας 4.3

Οι εκτιμώμενες τιμές της περιοδικότητας, που προκύπτουν, δίνονται για κάθε μήνα από τον Ιανουάριο μέχρι το Δεκέμβριο και σε κάθε μήνα η τιμή που προκύπτει είναι ίδια για όλα τα χρόνια. Παρατηρούμε από τον πίνακα 4.3, ότι η μεγαλύτερη τιμή αφορά το μήνα Μάιο, ενώ η μικρότερη το μήνα Δεκέμβριο, πράγμα που υποδεικνύει διαχρονικά, μια αύξηση των αυτοκτονιών στο Μάιο και μια μείωση στο Δεκέμβριο. *(Παραθέτουμε τις τιμές μόνο για τη μία χρονιά σε κάθε μήνα καθώς είναι η ίδια για όλες τις επόμενες.)*

Για τη μοντελοποίηση της χρονοσειράς θα χρησιμοποιήσουμε αρχικά, ένα αυτοπαλινδρομούμενο μοντέλο Poisson, θέλοντας να ελέγξουμε αν ο αριθμός των αυτοκτονιών κάθε μήνα, επηρεάζεται από κάποιους προηγούμενους μήνες. Το

αρχικό μας αυτό μοντέλο, θα χρησιμοποιηθεί πιο πολύ για να επιβεβαιώσουμε την υπόθεση που μπορούμε να κάνουμε, ότι οι χρονικές υστερήσεις δεν θα καθορίζουν τη μεταβλητή απόκρισης και άρα δε θα συμβάλλουν τελικά στο μοντέλο. Η υπόθεση αυτή, στηρίζεται στην εικόνα που έχουμε από το γράφημα της αυτοσυσχέτισης της χρονοσειράς, το οποίο μας έχει προϊδεάσει για το ότι η μοντελοποίηση της θα στηριχτεί μάλλον στην έντονη περιοδικότητα του φαινομένου, παρά στις όποιες χρονικές υστερήσεις. Μέσω της εντολής *ar(suicides),* θα δούμε τον ιδανικό αριθμό των lags που θα επιλεγόταν για την χρονοσειρά μας, στην περίπτωση ενός αυτοπαλινδρομούμενου μοντέλου:

```
> ar(suicides)
Call:
ar(x = suicides)
Coefficients:
    1          2          3          4          5
 0.2862   0.2573  -0.0782  -0.1657  -0.1537
Order selected 5  sigma^2 estimated as  46.26
```

Άρα θα ξεκινήσουμε τη μοντελοποίηση με αριθμό lags ίσο με 5. Στη δημιουργία των χρονικών υστερήσεων θα χρησιμοποιήσουμε και πάλι τις συναρτήσεις *lagmatrix* και *lag,* των οποίων το τρόπο λειτουργίας έχουμε ήδη εξηγήσει στην προηγούμενη παράγραφο, στην εφαρμογή για το λογιστικό μοντέλο παλινδρόμησης:

```
> lagmatrix <- function(x,max.lag){embed(c(rep(NA,max.lag),x),max.lag)}
>   lag <- function(x,lag) {
+   out<-lagmatrix(x,lag+1)[,lag]
+   return(out[1:length(out)-1]) }
```

Αποθηκεύουμε τις τιμές των 5 lags σε έναν πίνακα *Y,* και με τη δεύτερη εντολή αφαιρούμε τις γραμμές με τις κενές τιμές (*NA*) που δημιουργούνται λόγω των lags:

```
> Y<-lagmatrix(suicides,6)
> Y<-Y[7:dim(Y)[1],]
```

Φυσικά, πέραν των χρονικών υστερήσεων, θα πρέπει να εισάγουμε στο μοντέλο μας και τον πολύ σημαντικό όρο που θα υποδεικνύει την περιοδικότητα, η οποία έχει παρατηρηθεί στη χρονοσειρά. Αυτό θα γίνει, με τη χρήση του όρου:

$$\cos\left(\frac{2\pi t}{12}\right)$$

Το διάνυσμα *t* θα περιέχει τις 132 χρονικές στιγμές οι οποίες αντιστοιχούν σε κάθε παρατήρηση της χρονοσειράς μας. Για το τελικό διάνυσμα αφαιρούμε τις πρώτες 5 χρονικές στιγμές:

```
> t<-as.vector(seq(1:132))
> t<-(2*pi*t)/12
>cos<-cos(t)
>cos<-tail(cos,-5)
```

Κατασκευάζουμε το αρχικό μας μοντέλο το οποίο θα έχει σαν μεταβλητή απόκρισης τη χρονοσειρά μας, η οποία, όπως έχουμε εξηγήσει, περιέχεται στην πρώτη στήλη του πίνακα *Y,* και επεξηγηματικές μεταβλητές τις 5 χρονικές υστερήσεις και τον όρο της περιοδικότητας. Θα έχουμε δηλαδή το εξής παλινδρομικό μοντέλο Poisson:

$$\log \mu_t(\beta) = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + \beta_5 Y_{t-5} + \cos\left(\frac{2\pi t}{12}\right)$$

Με $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$ και $Z_{t-1} = \left(1, Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, Y_{t-5}, \cos\left(\frac{2\pi t}{12}\right)\right)'$

Εισάγοντας τις παρακάτω εντολές προκύπτει και το αντίστοιχο μοντέλο:

```
> poisson<-dyn$glm(Y[,1]~Y[,2]+Y[,3]+Y[,4]+Y[,5]+Y[,6]+cos, family=poisson)
> summary(poisson)
```

(Υπενθυμίζουμε ότι για τη χρήση της dyn$glm χρειάζονται οι libraries zoo και dyn.)

```
Call:
glm(formula = dyn(Y[, 1] ~ Y[, 2] + Y[, 3] + Y[, 4] + Y[, 5] + Y[, 6] + cos),
family = poisson)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.499270  -0.146944   0.003153   0.123052   0.538981

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  3.322769   0.113426  29.294  < 2e-16 ***
Y[, 2]       0.003316   0.002666   1.244    0.214
Y[, 3]       0.003498   0.002552   1.370    0.171
Y[, 4]      -0.003247   0.002530  -1.283    0.199
Y[, 5]      -0.002047   0.002540  -0.806    0.420
Y[, 6]       0.001289   0.002721   0.474    0.636
cos         -0.189376   0.034461  -5.495    3.9e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 248.35  on 126  degrees of freedom
Residual deviance: 148.62  on 120  degrees of freedom
AIC: 827.15
Number of Fisher Scoring iterations: 4
```

Αποτελέσματα 4.6

Έτσι, η εκτιμήτρια μέγιστης μερικής πιθανοφάνειας του β είναι:

$$\hat{\beta} = (3.323, 0.003, 0.0035, -0.0032, -0.002, 0.001, -0.189)'$$

Εισάγοντας την εντολή *>vcov(poisson),* μπορούμε να επιβεβαιώσουμε, όπως έχουμε αναφέρει ξανά ότι η στήλη των τυπικών σφαλμάτων των $\hat{\beta}$, ταυτίζεται με την τετραγωνική ρίζα της διαγωνίου του δεσμευμένου πίνακα πληροφορίας $G_N(\beta)$.

Σύμφωνα με τα αποτελέσματα που προέκυψαν, όπως και περιμέναμε, η μόνη στατιστικά σημαντική μεταβλητή, που συμβάλει στο μοντέλο και επηρεάζει τη μεταβλητή απόκρισης, είναι ο όρος της περιοδικότητας και καμία χρονική υστέρηση.

Με τη χρήση των παρακάτω εντολών προκύπτει ο ακόλουθος πίνακας, που περιέχει τα διάφορα αποτελέσματα (βαθμοί ελευθερίας, p-values), όσον αφορά τον έλεγχο του μοντέλου μέσω της ελεγχοσυνάρτησης Deviance:

```
> DD<-poisson$null.deviance-poisson$deviance
> df<-poisson$df.null-poisson$df.residual
> 1-pchisq(DD,df)
> poisson$deviance
> poisson$df.residual
> 1-pchisq(poisson$deviance,poisson$df.residual)
```

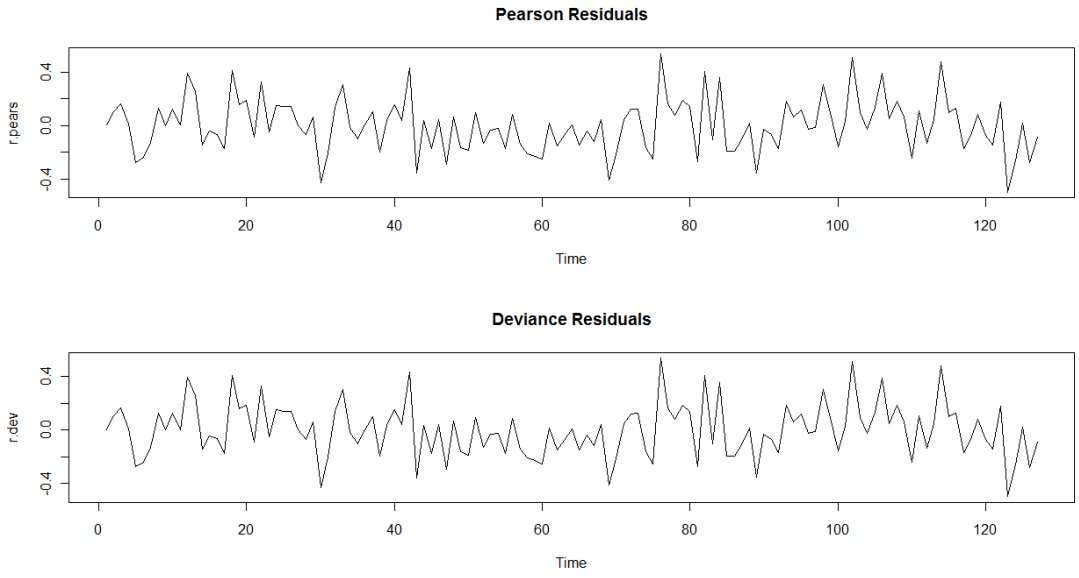
Ανάλυση της Deviance			
	Deviance	df	p-value
Μοντέλο	99,733	6	0,0000
Residual	148,62	120	0,0392
Σύνολο	248.35	126	

Πίνακας 4.4

Με βάση τον πίνακα 4.4 και την p-τιμή της deviance του μοντέλου (*p-value=0*) οδηγούμαστε στο συμπέρασμα ότι υπάρχει συσχέτιση μεταξύ των μεταβλητών, οπότε πρέπει να απορρίψουμε την υπόθεση του κενού μοντέλου χωρίς συμμεταβλητές. Παρ'όλα αυτά, η δεύτερη p-τιμή *(p-value=0.0392)* υποδεικνύει ότι το μοντέλο δεν μπορεί να θεωρηθεί ικανοποιητικό, αφού η προσαρμογή του είναι στατιστικά διαφορετική από αυτή του κορεσμένου μοντέλου.

Μπορούμε επίσης μέσω των παρακάτω εντολών να δούμε τα γραφήματα 4.6 και 4.7 των χρονοσειρών των υπολοίπων Pearson και deviance του μοντέλου, αντίστοιχα:

```
> r.pears<-residuals(poisson,type="pearson")
> r.dev<-residuals(poisson)
```



Γραφήματα 4.6 και 4.7

Σύμφωνα και με τα παρακάτω αποτελέσματα της stepwise μεθόδου και του ελέγχου AIC, επιβεβαιώνουμε ότι το καταλληλότερο μοντέλο για τη χρονοσειρά σε σύγκριση με το αρχικό, είναι τελικά αυτό που περιέχει μόνο τον όρο της περιοδικότητας, με *AIC=821.91*, κατά λίγο μικρότερο από αυτόν του πρώτου μοντέλου. *(Υπενθυμίζουμε ότι για τη χρήση της μεθόδου stepAIC χρειάζεται η βιβλιοθήκη MASS. Παρουσιάζουμε το αποτέλεσμα του ελέγχου και όχι κάθε step που παράγει η μέθοδος μέχρι το τελικό μοντέλο.)*

```
> step <- stepAIC(poisson, direction="backward")
> step$anova
```

Stepwise Model Path
Analysis of Deviance Table
Initial Model:
Y[, 1] ~ Y[, 2] + Y[, 3] + Y[, 4] + Y[, 5] + Y[, 6] + cos
Final Model:
Y[, 1] ~ cos

Συμπεραίνουμε λοιπόν, ότι το αυτοπαλινδρομούμενο μοντέλο που επιλέξαμε για τη χρονοσειρά μας δεν ήταν κατάλληλο, καθώς οι διάφορες χρονικές υστερήσεις δεν συμβάλλουν τελικά στο μοντέλο. Δηλαδή, οι αυτοκτονίες των προηγούμενων μηνών $(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-5})$, δεν επηρεάζουν τον αριθμό των αυτοκτονιών στο παρόντα μήνα Y_t. Αυτό που καθορίζει τη μεταβλητή απόκρισης μας είναι φυσικά η περιοδικότητα. Στο πέρασμα των ετών υπάρχει μια επαναληψιμότητα στο πως κυμαίνεται ο αριθμός των αυτοκτονιών ανά μήνα. Γίνεται λοιπόν σαφές ότι η εποχή/οι μήνες καθορίζουν το ποσοστό των ανθρώπων που αυτοκτονούν. Παρά το ότι ίσως θα περιμέναμε το αντίθετο, ο μεγαλύτερος αριθμός αυτοκτονιών εμφανίζεται στους καλοκαιρινούς μήνες (Μάιος-Ιούνιος) και όχι στους χειμερινούς, ενώ οι λιγότερες αυτοκτονίες παρατηρούνται το Δεκέμβριο το οποίο θα μπορούσε ίσως να εξηγηθεί λόγω του πιο ευχάριστου κλίματος που επιβάλλουν συνήθως οι

γιορτές (σε αντίθεση με το "μύθο" περί μελαγχολίας κατά την περίοδο των γιορτών ο οποίος έχει καταρριφθεί και από πρόσφατες έρευνες). Βεβαίως και υπάρχουν και πολλοί άλλοι παράγοντες που πιθανώς να επηρεάζουν ένα κοινωνικό φαινόμενο, όπως είναι η αυτοκτονία.

Έχοντας πλέον καταλήξει στο ότι η περιοδικότητα είναι αυτή που καθορίζει τον αριθμό των αυτοκτονιών, μπορούμε να ελέγξουμε αν και κατά πόσο υπάρχει κάποια διαφοροποίηση μεταξύ των μηνών, ώστε να καταλήξουμε σε ένα τελικό μοντέλο, στο οποίο πιθανώς να προσαρμόζονται καλύτερα τα δεδομένα μας. Δηλαδή, να ελέγξουμε, αν υπάρχουν κάποιοι μήνες που συμβάλλουν στη διαμόρφωση της μεταβλητής απόκρισης με διαφορετικό τρόπο από αυτόν που θα προβλεπόταν σύμφωνα με την περιοδικότητα και τον όρο του συνημιτόνου.

Για να γίνει αυτό, θα πρέπει αρχικά να κατασκευάσουμε ένα μοντέλο το οποίο θα έχει ως επεξηγηματικές μεταβλητές κάθε μήνα ξεχωριστά *(dummy variables)*, ώστε να παρατηρήσουμε ποιοί μήνες και με ποιο τρόπο επηρεάζουν τον αριθμό των αυτοκτονιών, χωρίς τον όρο της περιοδικότητας. Εισάγοντας την εντολή *seasonaldummy(suicides),* αποθηκεύουμε αυτομάτως στην μεταβλητή *seasons* τις "εποχές", στις οποίες είναι χωρισμένη η χρονοσειρά μας.

```
> seasons<-seasonaldummy(suicides)
```
(για την εντολή *seasonaldummy γίνεται χρήση της βιβλιοθήκης forecast)*

Έχουμε έτσι, τα αποτελέσματα 4.7:

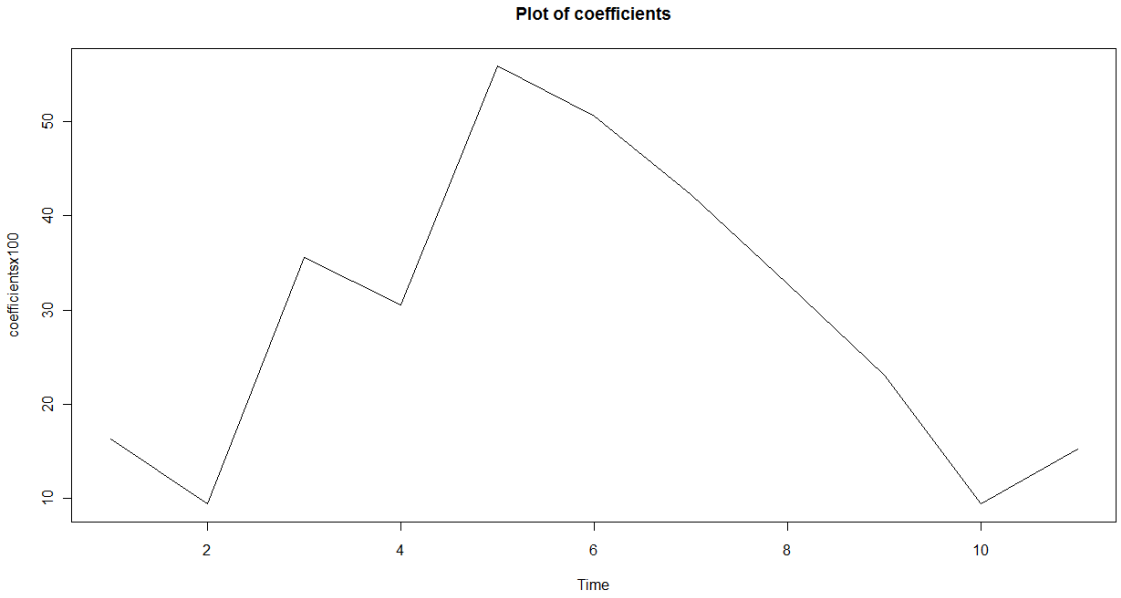
```
> poissonseasonal<-glm(suicides~seasons,family="poisson")
> summary(poissonseasonal)
```

```
Call:
glm(formula = suicides ~ seasons, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.94368  -0.73332  -0.05231   0.68732   2.38787

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.14337    0.06262  50.196  < 2e-16 ***
seasonJan    0.16252    0.08518   1.908  0.056386 .
seasonFeb    0.09353    0.08656   1.080  0.279942
seasonMar   0.35589    0.08166   4.358  1.31e-05 ***
seasonApr    0.30518    0.08253   3.698  0.000218 ***
seasonMay   0.55906    0.07851   7.121  1.07e-12 ***
seasonJun    0.50611    0.07928   6.384  1.73e-10 ***
seasonJul    0.42232    0.08057   5.241  1.59e-07 ***
seasonAug    0.32803    0.08214   3.994  6.50e-05 ***
seasonSep    0.23018    0.08389   2.744 0.006071 **
seasonOct    0.09353    0.08656   1.080  0.279942
seasonNov    0.15247    0.08537   1.786  0.074114 .
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 257.58  on 131  degrees of freedom
Residual deviance: 142.31  on 120  degrees of freedom
AIC: 857.44
Number of Fisher Scoring iterations: 4
```

Αποτελέσματα 4.7

Ήταν μάλλον αναμενόμενο, πως από το συγκεκριμένη παλινδρόμηση οι μήνες που θα έβγαιναν στατιστικά σημαντικοί, θα ήταν κυρίως οι θερινοί, στους οποίους έχει παρατηρηθεί η μεγάλη αύξηση των αυτοκτονιών. Αυτό που μας ενδιαφέρει από τα συγκεκριμένα αποτελέσματα, είναι οι εκτιμήσεις των συντελεστών $\hat{\beta}$ για τον κάθε μήνα, ώστε να προκύψει το γράφημα 4.8:



Γράφημα 4.8

Από τη γραφική αναπαράσταση των συντελεστών, παρατηρούμε πως από τη συνημιτονοειδή μορφή, “ξεφεύγει” ο Μάρτιος ο οποίος βρίσκεται σχετικά ψηλότερα από όσο θα έπρεπε, ενώ το ίδιο θα μπορούσαμε να ισχυριστούμε σε μικρότερο βαθμό για τον Ιανουάριο και το Νοέμβριο. Η ψηλότερη τιμή που αποτελεί την κορυφή του γραφήματος 4.8, δε μας ξαφνιάζει καθώς συμπίπτει με το Μάιο στον οποίο όπως είπαμε, παρατηρείται διαχρονικά ο μεγαλύτερος αριθμός αυτοκτονιών. Το αν οι προαναφερθέντες τρείς μήνες, συμβάλλουν όντως στη μοντελοποίηση της χρονοσειράς μας, πέραν της περιοδικότητας, θα το ελέγξουμε κατασκευάζοντας το αντίστοιχο μοντέλο.

Πριν την προσθήκη των συγκεκριμένων μηνών ως επεξηγηματικές μεταβλητές, θα δούμε τα αποτελέσματα του μοντέλου που περιέχει ως μοναδική επεξηγηματική μεταβλητή την περιοδικότητα (για την οποία γνωρίζουμε πως σίγουρα καθορίζει τη μεταβλητή μας), ώστε να μπορούμε να συγκρίνουμε τα δύο αυτά μοντέλα.

Οπότε, σύμφωνα με τις γνωστές μας πλέον εντολές, έχουμε τα παρακάτω αποτελέσματα:

```
> poisson1<-glm(suicides~cos,family="poisson")
Call:
glm(formula = suicides ~ cos, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.21226  -0.82779  -0.03525   0.73314   2.36938

Coefficients:
             Estimate Std. Error  z value     Pr(>|z|)
(Intercept)  3.41280    0.01589 214.716   <2e-16 ***
cos         -0.22037    0.02241  -9.833    <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance:        257.58  on 131  degrees of freedom
Residual deviance: 160.00  on 130  degrees of freedom
AIC: 855.14
Number of Fisher Scoring iterations: 4
```

Αποτελέσματα 4.8

Και αντίστοιχα για το δεύτερο μοντέλο μας, στο οποίο προσθέτουμε τους τρείς μήνες του Ιανουαρίου, του Μαρτίου και του Νοεμβρίου:

```
>poisson2<-glm(suicides~cos+january+march+november,family="poisson")
Call:
glm(formula = suicides ~ cos + january + march + november, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.13464  -0.72229  -0.09511   0.77156   2.38130

Coefficients:
             Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)  3.37468    0.01994  169.264  <2e-16 ***
cos         -0.26700    0.02775   -9.623   <2e-16 ***
january      0.16243    0.06872    2.364    0.0181 *
march        0.12458    0.05608    2.222    0.0263 *
november     0.15238    0.06896    2.210    0.0271 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
Null deviance:        257.58  on 131  degrees of freedom
Residual deviance: 149.07  on 127  degrees of freedom
AIC: 850.2
Number of Fisher Scoring iterations: 4
```

Αποτελέσματα 4.9

Αρχικά, βλέπουμε πως οι τιμές του AIC και για τα δύο μοντέλα, είναι μικρότερες από αυτήν του παλινδρομικού μοντέλου που προηγήθηκε με επεξηγηματικές μεταβλητές κάθε μήνα χωριστά *(AIC=857.44)*, πράγμα που σημαίνει ότι ο όρος της περιοδικότητας και μόνο, σίγουρα προσαρμόζει καλύτερα τα δεδομένα μας. Από τα αποτελέσματα 4.9 που προέκυψαν για το δεύτερο μοντέλο, παρατηρούμε ότι οι τρείς συγκεκριμένοι μήνες που προσθέσαμε, αποτελούν όντως στατιστικά σημαντικές μεταβλητές και μαζί με την περιοδικότητα επηρεάζουν διαχρονικά τον αριθμό των αυτοκτονιών. Επίσης, το κριτήριο AIC για το δεύτερο μοντέλο είναι μικρότερο από αυτό που περιέχει μόνο τον όρο της περιοδικότητας (Αποτελέσματα 4.9), πράγμα που υποδεικνύει ότι ίσως αυτό, να είναι πράγματι ένα καλύτερο μοντέλο για τη χρονοσειρά μας.

Αυτό όμως για να επιβεβαιωθεί, θα πρέπει να δούμε τα αποτελέσματα της ελεγχοσυνάρτησης Deviance για τα δύο μοντέλα, με τον ίδιο τρόπο που κάναμε και παραπάνω:

Ανάλυση της Deviance για το 1ο και 2ο μοντέλο							
1ο	Dev	df	p-value	**2ο**	Dev	df	p-value
Μοντέλο	97,574	1	0,0000	Μοντέλο	108,51	4	0,0000
Residual	160,00	130	0,0379	Residual	149,06	127	0,0880
Σύνολο	257,58	131		Σύνολο	257,58	131	

Πίνακας 4.5

Σύμφωνα με τον πίνακα 4.5 και την p-τιμή *(p-value=0.088)* που προκύπτει από τον έλεγχο Deviance για το μοντέλο που περιέχει ως επεξηγηματικές μεταβλητές την περιοδικότητα και τους τρείς μήνες, καταλήγουμε στο ότι δεν μπορούμε να το απορρίψουμε ως μη ικανοποιητικό, αφού η προσαρμογή του δεν φαίνεται να είναι στατιστικά διαφορετική από αυτή του κορεσμένου μοντέλου, σε αντίθεση με το μοντέλο που περιέχει μόνο τον όρο της περιοδικότητας *(p-value=0.037).*

Εφαρμόζοντας και τη μέθοδο stepwise για το δεύτερο μοντέλο, επιβεβαιώνεται πως αποτελεί ένα ικανοποιητικό μοντέλο για τη χρονοσειρά, καθώς δεν αφαιρείται καμία επεξηγηματική μεταβλητή όπως βλέπουμε από τα παρακάτω αποτελέσματα:

```
> step <- stepAIC(poisson2, direction="backward")
> step$anova
Stepwise Model Path
Analysis of Deviance Table
Initial Model:
suicides ~ cos + january + march + november
Final Model:
suicides ~ cos + january + march + november
```

Καταλήγουμε λοιπόν, στο ότι ένα καλύτερο μοντέλο για τη χρονοσειρά μας αποτελεί αυτό, που εκτός από τον όρο της περιοδικότητας, περιέχει και τους τρείς μήνες των οποίων οι συντελεστές, σύμφωνα με το αντίστοιχο διάγραμμα, έχουν τιμές, οι οποίες βρίσκονται ψηλότερα από τη συνημιτονοειδή μορφή που περιμένουμε. Ωστόσο, πριν αρκεστούμε στο μοντέλο που περιέχει αυτούς τους

συγκεκριμένους μήνες, για τους οποίους είχαμε κάνει και την αρχική υπόθεση παρατηρώντας τη γραφική παράσταση των συντελεστών, έγιναν πολλές δοκιμές και για άλλους μήνες πριν μπορέσουμε να καταλήξουμε στο ότι όντως η περιοδικότητα με τους τρείς αυτούς μήνες μπορούν να περιγράψουν ικανοποιητικά τη χρονοσειρά μας.

Πριν κλείσουμε τη συγκεκριμένη εφαρμογή, έχει ενδιαφέρον να δούμε τις διαφοροποιήσεις που ίσως υπάρχουν στον αριθμό των αυτοκτονιών αναλόγως με το φύλο. Σε αυτήν την περίπτωση, μας ενδιαφέρει πιο πολύ να παρατηρήσουμε τη συμπεριφορά των δύο αυτών χρονοσειρών και λιγότερο τη διαδικασία μοντελοποίησης τους, αφού περιμένουμε να στηρίζονται και αυτές στο ίδιο σκεπτικό του μοντέλου, το οποίο τελικά προέκυψε για τη χρονοσειρά των συνολικών αυτοκτονιών (οι συνολικές αυτοκτονίες αποτελούν το άθροισμα των αυτοκτονιών των δύο φύλων). Έτσι, δημιουργούμε τις ακόλουθες δύο χρονοσειρές για τους άντρες και τις γυναίκες αντίστοιχα:

```
>ym<-dataframe[,2]
>suicides_m<-ts(ym,start=1997,frequency=12)

>yf<-dataframe[,3]
>suicides_f<-ts(yf,start=1997,frequency=12)
```

Με βάση τις δύο χρονοσειρές, προκύπτουν τα γραφήματα 4.9 και 4.10 (πάνω για τους άντρες, κάτω για τις γυναίκες):

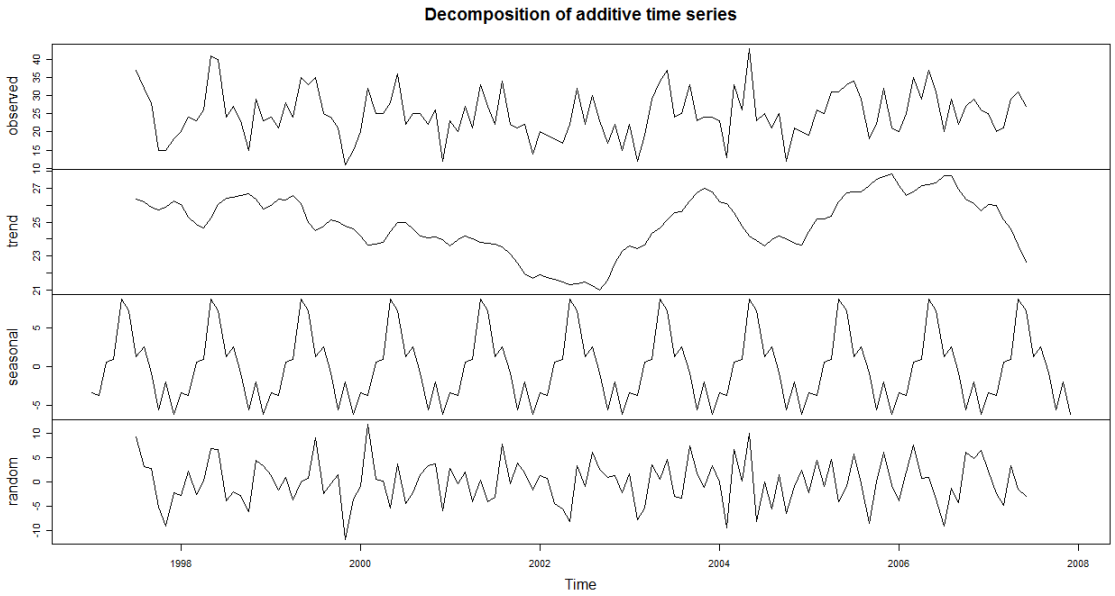
Γραφήματα 4.9 και 4.10

Από τις παραπάνω γραφικές παραστάσεις των χρονοσειρών για τα δύο φύλα, παρατηρούμε πως ο αριθμός αυτοκτονιών στους άντρες είναι διαχρονικά πολύ

μεγαλύτερος από αυτόν στις γυναίκες ενώ και στις δύο περιπτώσεις παρατηρείται μια περιοδικότητα στα δεδομένα.

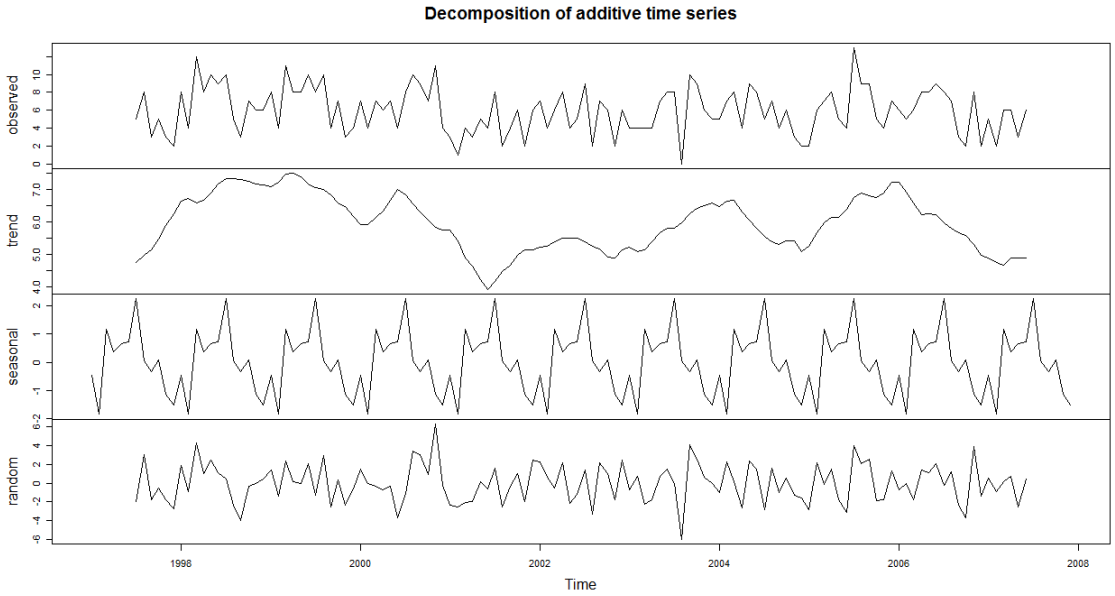
Όπως και για την αρχική μας χρονοσειρά, μπορούμε μέσω των γραφημάτων 4.11 και 4.12 να επιβεβαιώσουμε την περιοδικότητα και να δούμε καλύτερα τις συνιστώσες των δύο χρονοσειρών αντίστοιχα.

Για τους άντρες:



Γράφημα 4.11

Για τις γυναίκες:



Γράφημα 4.12

Και στις δύο περιπτώσεις είναι εμφανής η περιοδικότητα ανά έτος και δεν υπάρχει κάποια συγκεκριμένη, συνολική τάση στα δεδομένα. Παρ'όλα αυτά, στην περίπτωση των αντρών παρατηρούμε μια συνεχόμενη μείωση στον αριθμό των

αυτοκτονιών από το 1999 μέχρι το 2002 περίπου, ενώ από το 2003 και μετά υπάρχει μια αύξηση, με μια εξαίρεση κατά τη διάρκεια του 2004-2005. Στην περίπτωση των γυναικών, στο γράφημα 4.10, δεν παρατηρείται κάτι αντίστοιχο, παρά μόνο μια απότομη μείωση κατά τη διάρκεια του 2001.

Επίσης, ο πίνακας που υποδεικνύει την περιοδικότητα στον κάθε μήνα, για την περίπτωση των αντρών, έχει ως εξής:

	Jan	Feb	Mar	Apr	May	Jun
1997	-3.355	-3.664	0.656	0.981	8.769	7.177
	Jul	Aug	Sep	Oct	Nov	Dec
1997	1.352	2.569	-0.776	-5.564	-1.951	-6.193

Πίνακας 4.6

Ενώ για τις γυναίκες:

	Jan	Feb	Mar	Apr	May	Jun
1997	-0.449	-1.845	1.167	0.358	0.650	0.729
	Jul	Aug	Sep	Oct	Nov	Dec
1997	2.254	0.054	-0.328	0.071	-1.132	-1.528

Πίνακας 4.7

Από τον πίνακα 4.6, παρατηρούμε ότι στην περίπτωση των αντρών, ο μήνας που οι αυτοκτονίες αυξάνονται είναι ο Μάιος, ενώ ο μήνας που μειώνονται είναι ο Δεκέμβριος, όπως συνέβαινε και στη χρονοσειρά των συνολικών αυτοκτονιών. Στην περίπτωση των γυναικών όμως, με βάση τον πίνακα 4.7, ο μήνας με τις περισσότερες αυτοκτονίες είναι ο Ιούλιος, ενώ αυτός με τις λιγότερες είναι ο Φεβρουάριος.

Μπορούμε να πούμε πως υπάρχουν αρκετές διαφοροποιήσεις μεταξύ των δύο φύλων, με σημαντικότερη όλων το ότι ο αριθμός των αντρών που αυτοκτονούν, είναι σταθερά μεγαλύτερος στο πέρασμα των ετών. Βέβαια, υπάρχει και μια βασική ομοιότητα στις δύο χρονοσειρές, η περιοδικότητα που παρατηρείται μεταξύ των παρατηρήσεων. Καταλαβαίνουμε, ότι και στις δύο περιπτώσεις, όπως και στην αρχική χρονοσειρά μας, η μεταβλητή απόκρισης που είναι ο αριθμός των αυτοκτονιών, θα επηρεάζεται μάλλον από τον όρο της περιοδικότητας, $\cos\left(\frac{2\pi t}{12}\right)$.

Με το ίδιο σκεπτικό όπως και παραπάνω, θα κατασκευάσουμε για καθένα από τα δύο φύλα, το μοντέλο που περιέχει την περιοδικότητα, ώστε να μπορέσουμε να δούμε αν ο συγκεκριμένος όρος επηρεάζει στον ίδο βαθμό τον αριθμό των αυτοκτονιών στους άντρες και στις γυναίκες και να συγκρίνουμε τα αποτελέσματα, που θα προκύψουν τώρα, με τα αποτελέσματα των αμέσως επόμενων μοντέλων.

Έτσι, για τους άντρες έχουμε τα εξής αποτελέσματα:

```
> poisson_m<-glm(suicides_m~cos,family="poisson")
```

```
Call:
glm(formula = suicides_m ~ cos, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.84088  -0.69159   0.06809   0.68238   2.42059

Coefficients:
            Estimate Std. Error  z value  Pr(>|z|)
(Intercept)  3.20041    0.01768 181.016   <2e-16 ***
cos         -0.22563    0.02493  -9.052   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
Null deviance: 236.53  on 131  degrees of freedom
Residual deviance: 153.81  on 130  degrees of freedom
AIC: 820.52
Number of Fisher Scoring iterations: 4
```

Αποτελέσματα 4.10

Και τα αντίστοιχα για τις γυναίκες:

```
> poisson_f<-glm(suicides_f~cos,family="poisson")
```

```
Call:
glm(formula = suicides_f ~ cos, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.58083  -0.79469   0.06448   0.60393   2.36917

Coefficients:
             Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)  1.75898    0.03630  48.461  < 2e-16 ***
cos         -0.19813    0.05121  -3.869   0.000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
Null deviance: 152.58  on 131  degrees of freedom
Residual deviance: 137.50  on 130  degrees of freedom
AIC: 606.15
Number of Fisher Scoring iterations: 4
```

Αποτελέσματα 4.11

Από τα αποτελέσματα 4.10 και 4.11, παρατηρούμε πως και στα δύο φύλα, η περιοδικότητα όντως καθορίζει τον αριθμό των αυτοκτονιών, συγκριτικά όμως, η p-τιμή που προκύπτει για το συγκεκριμένο όρο στην περίπτωση των αντρών, είναι

αρκετά μικρότερη από αυτήν στην περίπτωση των γυναικών, χωρίς βέβαια αυτό να σημαίνει ότι δεν αποτελεί μια στατιστικά σημαντική μεταβλητή και για τη χρονοσειρά των γυναικών.

Κατασκευάζουμε, επίσης, τα παλινδρομικά μοντέλα Poisson, που περιέχουν ως επεξηγηματικές μεταβλητές κάθε μήνα, χωριστά από τους συντελεστές των οποίων προκύπτουν και τα αντίστοιχα διαγράμματα που βλέπουμε παρακάτω. Βασισμένοι στα διαγράμματα αυτά, θα ελέγξουμε αν υπάρχουν κάποιοι μήνες που θα πρέπει να προστεθούν ως επεξηγηματικές μεταβλητές στο μοντέλο μας για την καλύτερη περιγραφή των χρονοσειρών, όπως συνέβη και στην αρχική χρονοσειρά των συνολικών αυτοκτονιών.

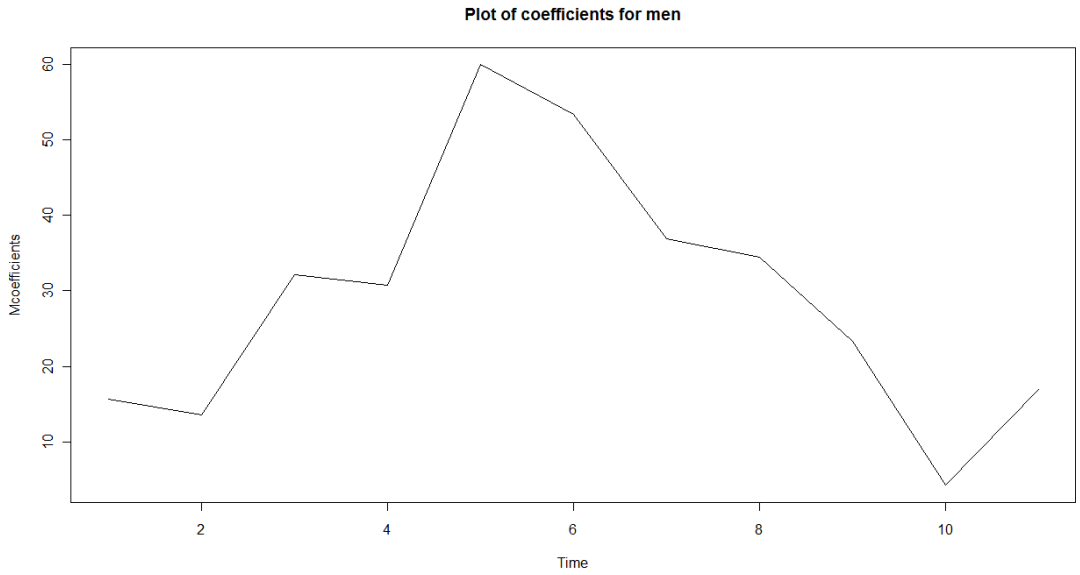
Για την χρονοσειρά των αντρών προκύπτουν τα αποτελέσματα 4.12:

(Παραθέτονται κατευθείαν τα αποτελέσματα και δεν επαναλαμβάνονται οι εντολές για την κατασκευή των μοντέλων, καθώς θεωρούνται πλέον γνωστές)

```
Call:
glm(formula = suicides_m ~ season_m, family = "poisson")

Deviance Residuals:
     Min            1Q        Median            3Q           Max
-2.6628391  -0.5931706   0.0002545   0.6778683   2.1204364

Coefficients:
             Estimate   Std. Error z value    Pr(>|z|)
(Intercept)   2.92998    0.06967  42.053     < 2e-16 ***
season_mJan  0.15692    0.09489   1.654    0.098179 .
season_mFeb  0.13596    0.09535   1.426    0.153910
season_mMar  0.32110    0.09152   3.509    0.000450 ***
season_mApr   0.30691    0.09179   3.344   0.000827 ***
season_mMay  0.59905    0.08672   6.908    4.93e-12 ***
season_mJun   0.53291    0.08777   6.072    1.27e-09 ***
season_mJul   0.36922    0.09061   4.075    4.60e-05 ***
season_mAug   0.34545    0.09105   3.794    0.000148 ***
season_mSep   0.23281    0.09328   2.496    0.012565 *
season_mOct   0.04276    0.09750   0.43     0.660952
season_mNov   0.16929    0.09462   1.78     0.073583 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance:       236.53  on 131  degrees of freedom
Residual deviance: 133.75  on 120  degrees of freedom
AIC: 820.47
Number of Fisher Scoring iterations: 4
```

Αποτελέσματα 4.12

Κι έτσι έχουμε και τη γραφική αναπαράσταση των συντελεστών:



Γράφημα 4.13

Από το γράφημα 4.13, παρατηρούμε ότι όπως και προηγουμένως στη γενική χρονοσειρά, ο Ιανουάριος και ο Νοέμβριος βρίσκονται ψηλότερα από όσο θα έπρεπε, δείχνοντας να μην ακολουθούν τη συνημιτονοειδή μορφή της χρονοσειράς. Ο Μάρτιος σε αντίθεση με την περίπτωση των συνολικών αυτοκτονιών δε φαίνεται να δημιουργεί κάποιο πρόβλημα, ωστόσο στο μοντέλο που θα κατασκευάσουμε θα προσθέσουμε και αυτόν, ώστε να επιβεβαιώσουμε την υπόθεση μας:

```
Call:
glm(formula = suicides_m ~ cos + january + march + november,
family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.78363  -0.72653   0.04387   0.69625   2.31285

Coefficients:
             Estimate Std. Error  z value   Pr(>|z|)
(Intercept)  3.16407    0.02217  142.696   <2e-16 ***
cos         -0.27412    0.03084   -8.887   <2e-16 ***
january      0.16022    0.07662    2.091    0.0365 *
march        0.08701    0.06335    1.373    0.1696
november     0.17259    0.07629    2.262    0.0237 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance:         236.53  on 131  degrees of freedom
Residual deviance: 145.74  on 127  degrees of freedom
AIC: 818.46
Number of Fisher Scoring iterations: 4
```

Αποτελέσματα 4.13

Σύμφωνα με τα αποτελέσματα 4.13 για το συγκεκριμένο μοντέλο, επιβεβαιώθηκε η υπόθεση μας, ότι ο Ιανουάριος και ο Νοέμβριος συμβάλλουν στη διαμόρφωση της μεταβλητής απόκρισης μαζί με την περιοδικότητα, ενώ ο Μάρτιος όντως δεν αποτελεί στατιστικά σημαντική μεταβλητή.

Το ίδιο προκύπτει και από τα αποτελέσματα της stepwise μεθόδου για το μοντέλο:

```
> step$anova
Stepwise Model Path
Analysis of Deviance Table
Initial Model:
suicides_m ~ cos + january + march + november

Final Model:
suicides_m ~ cos + january + november

     Step Df  Deviance Resid. Df Resid. Dev      AIC
1                             127   145.7423 818.4598
2 - march  1 1.845747       128   147.5880 818.3056
```

Άρα, το τελικό μοντέλο για τη χρονοσειρά, που αφορά τον αριθμό αυτοκτονιών των αντρών, θα περιέχει ως επεξηγηματικές μεταβλητές την περιοδικότητα και τους δύο μήνες του Ιανουαρίου και του Νοεμβρίου για τους οποίους ελέγξαμε ότι όντως επηρεάζουν τη συμπεριφορά της χρονοσειράς μας, δημιουργώντας ένα αποδοτικότερο μοντέλο σε σχέση με τα υπόλοιπα, καθώς το κριτήριο AIC *(AIC: 818.3)* είναι μικρότερο και από αυτό του μοντέλου που περιέχει μόνο τον όρο της περιοδικότητας *(AIC: 820.5)* , αλλά και από αυτό που περιέχει ως επεξηγηματική μεταβλητή κάθε μήνα χωριστά *(AIC: 820.5).*

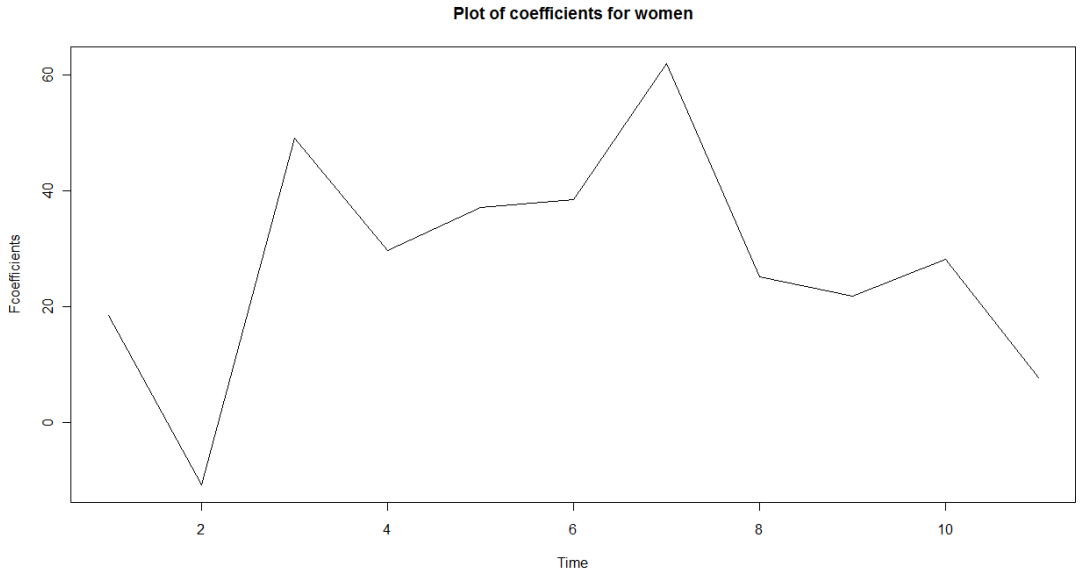
Τα παραπάνω βήματα ακολουθούμε αντίστοιχα και για τη χρονοσειρά των αυτοκτονιών των γυναικών, για την οποία έχουμε τα αποτελέσματα 4.14:

```
Call:
glm(formula = suicides_f ~ season_f, family = "poisson")

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.3845  -0.7356  0.0000  0.6317  2.4078

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.49393    0.14286  10.457  < 2e-16 ***
season_fJan   0.18572    0.19328   0.961 0.336618
season_fFeb  -0.10763    0.20769  -0.518 0.604301
season_fMar   0.49021    0.18141   2.702 0.006887 **
season_fApr   0.29783    0.18857   1.579 0.114241
season_fMay   0.37086    0.18572   1.997 0.045841 *
season_fJun   0.38485    0.18519   2.078 0.037704 *
season_fJul   0.61904    0.17719   3.494 0.000477 ***
```

```
season_fAug   0.25131    0.19048   1.319    0.187035
season_fSep    0.21905    0.19184   1.142    0.253508
season_fOct    0.28257    0.18919   1.494    0.135290
season_fNov   0.07847    0.19818   0.396    0.692136
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance:        152.58  on 131  degrees of freedom
Residual deviance: 123.66  on 120  degrees of freedom
AIC: 612.31
Number of Fisher Scoring iterations: 5
```

Αποτελέσματα 4.14

Όπως επίσης και τη γραφική αναπαράσταση των εκτιμώμενων συντελεστών:



Γράφημα 4.14

Στην περίπτωση των γυναικών, παρατηρούμε μια τελείως διαφορετική εικόνα από αυτή που είχαμε στο αντίστοιχο γράφημα 4.8 για τη χρονοσειρά των συνολικών αυτοκτονιών, και από αυτή στο 4.13 για τη χρονοσειρά των αυτοκτονιών των αντρών. Στις δύο προηγούμενες περιπτώσεις, είχαμε μια πολύ ακραία τιμή στο διάγραμμα, η οποία ταυτιζόταν με το μήνα του υψηλότερου αριθμού αυτοκτονιών, και την οποία μπορούσαμε να “εντάξουμε” στη συνημιτονοειδή μορφή, στην οποία βασίζεται η μοντελοποίηση και των τριών χρονοσειρών μας, ως την κορυφή της. Εδώ, οι πολύ υψηλές τιμές που παρατηρούνται το Μάρτιο και τον Ιούλιο, που αποτελούν τις δύο ακραίες τιμές του γραφήματος, θα πρέπει μάλλον να προστεθούν στο μοντέλο μας. Επίσης, σε αυτήν την περίπτωση, μεγαλύτερο πρόβλημα φαίνεται να προκαλεί η ιδιαίτερα χαμηλή τιμή του Φεβρουαρίου και όχι η τιμή του Ιανουαρίου, όπως συνέβαινε στις προηγούμενες περιπτώσεις. Άρα λοιπόν, θα κατασκευάσουμε το παρακάτω μοντέλο ώστε να ελέγξουμε αν αυτές μας οι υποθέσεις είναι βάσιμες:

```
Call:
glm(formula = suicides_f ~ cos + feb + march + july, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.477917 -0.771256 -0.008243  0.647983  2.302590

Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  1.73142   0.04234    40.891   <2e-16 ***
cos         -0.13660   0.05700    -2.396    0.0166 *
feb         -0.27683   0.15952    -1.735    0.0827 .
march        0.25271   0.11955     2.114    0.0345 *
july         0.26325   0.12257     2.148    0.0317 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 152.58  on 131  degrees of freedom
Residual deviance: 125.57  on 127  degrees of freedom
AIC: 600.22
Number of Fisher Scoring iterations: 4
```

Αποτελέσματα 4.15

Από τα αποτελέσματα 4.15 που προέκυψαν, συμπεραίνουμε ότι από τους μήνες που επιλέχθηκαν, κυρίως ο Μάρτιος και ο Ιούλιος συμβάλλουν στο μοντέλο μας, αλλά και ο Φεβρουάριος φαίνεται να επηρεάζει σε κάποιο μικρό βαθμό τη μεταβλητή απόκρισης. Το αξιοσημείωτο από τα συγκεκριμένα αποτελέσματα είναι η p-τιμή που προκύπτει για τον όρο του συνημιτόνου, η οποία μπορεί να παραμένει στατιστικά σημαντική αλλά είναι πολύ μεγαλύτερη από την αντίστοιχη p-τιμή στο μοντέλο για τη χρονοσειρά των αντρών. Αυτό σημαίνει ότι στην περίπτωση των αυτοκτονιών των γυναικών, μάλλον ο όρος του συνημιτόνου και της περιοδικότητας καθορίζουν μεν τη συμπεριφορά της χρονοσειράς, αλλά όχι στον ίδιο βαθμό που συμβαίνει στην περίπτωση των αντρών. Παρ'όλα αυτά, όπως βλέπουμε, το κριτήριο AIC *(AIC: 600.22)* είναι μικρότερο από αυτό του μοντέλου που δεν περιέχει καθόλου την περιοδικότητα και καθιστά κάθε μήνα ως επεξηγηματική μεταβλητή *(AIC: 612.31)*, όπως επίσης και από αυτό του μοντέλου που περιέχει ως μοναδικό όρο, τον όρο του συνημιτόνου, στο οποίο αντιστοιχούν τα αποτελέσματα 4.11 που βρίσκονται παραπάνω *(AIC: 606.15).*

Στο ίδιο μοντέλο καταλήγουμε και αν χρησιμοποιήσουμε τη μέθοδο stepwise, αφού, όπως μπορούμε να δούμε παρακάτω, η μεταβλητή του Φεβρουαρίου δεν αφαιρείται, παρ' ότι στα παραπάνω αποτελέσματα δεν ήταν στατιστικά σημαντική, πράγμα που σημαίνει ότι πρέπει να συμπεριληφθεί τελικά στο μοντέλο για την καλύτερη περιγραφή της χρονοσειράς:

```
> step$anova
Stepwise Model Path
Analysis of Deviance Table
Initial Model:
suicides_f ~ cos  + feb + march + july

Final Model:
suicides_f ~ cos + feb + march + july
Step  Df  Deviance Resid.  Df   Resid. Dev     AIC
 1                          27   125.5652   600.2181
```

Για το συγκεκριμένο μοντέλο, όπως συνέβη και στα αντίστοιχα προηγούμενα των άλλων δύο χρονοσειρών, έγιναν δοκιμές και για άλλους μήνες (όπως πχ ο Ιανουάριος), αλλά επιβεβαιωνόταν ότι αυτοί που συμβάλλουν τελικά στο μοντέλο μας, είναι αυτοί, που είχαμε υποθέσει αρχικά βασιζόμενοι στο διάγραμμα των συντελεστών.

Κλείνοντας, μετά από τα διάφορα αποτελέσματα που προέκυψαν για τα δύο φύλα, συμπεραίνουμε ότι η μοντελοποίηση για τις δύο αυτές χρονοσειρές διαφέρει, κι αυτό ίσως να οφείλεται στο ότι μάλλον η συμπεριφορά της χρονοσειράς για τις αυτοκτονίες των γυναικών, είναι πιο "απρόβλεπτη" από αυτή των αντρών. Στην περίπτωση των αντρών, η περιοδικότητα και άρα η διαχρονική επαναληψιμότητα, δείχνει να παίζει σημαντικότερο ρόλο στη μοντελοποίηση της χρονοσειράς, από ότι στην περίπτωση των γυναικών. Επίσης, το πιθανό μοντέλο στο οποίο καταλήξαμε για τους άντρες, είναι πιο κοντά σε αυτό για το οποίο είχαμε καταλήξει και για τη χρονοσειρά των συνολικών αυτοκτονιών, το οποίο όμως ήταν μάλλον αναμενόμενο, καθώς η μεγάλη πλειοψηφία των αυτοκτονιών ανήκει στους άντρες και όχι στις γυναίκες. Παρ'όλα αυτά, και στην περίπτωση των γυναικών έχουμε μια στατιστικά σημαντική περιοδικότητα, μόνο που υπάρχουν μήνες οι οποίοι παρουσιάζουν απότομα, ακραίες τιμές, με αποτέλεσμα να μην υπάρχουν περίοδοι κατά τη διάρκεια του χρόνου, που η χρονοσειρά να παρουσιάζει συνεχόμενα ανοδική ή καθοδική πορεία, πράγμα που μπορεί να ειπωθεί για την περίπτωση των αντρών.

Συνολικά, έχοντας δει την εφαρμογή των παλινδρομικών μοντέλων Poisson για κάθε μία από τις τρείς απαριθμητές χρονοσειρές μας, μπορούμε να καταλάβουμε πλέον τον τρόπο με τον οποίο μπορούν να χρησιμοποιηθούν στην πράξη. Εμείς, όπως και στην περίπτωση των βροχοπτώσεων και του λογιστικού μοντέλου, ασχοληθήκαμε μόνο με τη συμπεριφορά του φαινομένου στο χρόνο και το πώς επηρεάζεται από αυτόν. Αν όμως είχαμε στη διάθεση μας και κάποιες άλλες μεταβλητές, που πιθανόν επηρεάζουν τις αυτοκτονίες είτε συνολικά, είτε για κάθε φύλο χωριστά (πχ ψυχολογική, οικονομική κατάσταση), θα μπορούσαμε ίσως να φτιάξουμε ακόμα πιο ακριβή μοντέλα μέσω των οποίων θα κατανοούσαμε σε μεγαλύτερο βάθος το πώς και με βάση ποιούς παράγοντες, διαμορφώνεται διαχρονικά, ο αριθμός των αυτοκτονιών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΞΕΝΟΓΛΩΣΣΗ ΚΑΙ ΕΛΛΗΝΙΚΗ:

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to Time Series and Forecasting.* NewYork: Springer.

Collett, D. (2003). *Modelling Binary Data.* 2nd ed.,Boca Raton: Chapman and Hall.

Cox, D. (1975). Partial likelihood. *Biometrica,**62***, σσ. 69-76.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models.* Boca Raton: Chapman&Hall/CRC.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* 2nd ed.,New York: Springer.

Kedem, B., & Fokianos, K. (2002). *Regressions Models for Time Series Analysis.* New Jersey: Wiley.

Lindsey, J. K. (1997). *Applying Generalized Linear Models.* New York: Springer.

McCullagh, P., & Nelder, J. A. (2nd ed.,1989). *Generalized Linear Models.* Cambridge: Chapman and Hall.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2008). *Introduction to Time Series Analysis and Forecasting.* New Jersey: Wiley.

Montgomery, D. C., Peck, E. A., & Vinning, G. G. (2006). *Introduction to Linear Regression Analysis.* 4th ed.,New Jersey: Wiley.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). *Applied Regression Analysis:A Research Tool.* 2nd ed.,New York: Springer.

Δρυμώνης, Μ. (2005). *Ανάλυση Κατηγορικών Χρονοσειρών.* Πειραιάς.

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

Κοκολάκης, Γ. Ε. (2007). *Σημειώσεις Ανάλυσης Χρονοσειρών.* Εκδόσεις ΕΜΠ.

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

ΔΙΑΔΙΚΤΥΑΚΟΙ ΤΟΠΟΙ:

http://users.auth.gr/dkugiu/Teach/TimeSeries/

http://cran.r-project.org/doc/manuals/R-intro.html

http://www2.ucy.ac.cy/~fokianos/GreekRbook/indexRbook.htm

http://www.stat.pitt.edu/stoffer/tsa2/R_time_series_quick_fix.htm

http://www.imamu.edu.sa/Scientific_selections/abstracts/Math/Analysis of rainfall variability using generalized linear models.pdf

http://www.ats.ucla.edu/stat/r/dae/default.htm

http://www.metu.edu.tr/~ceylan/TimeSeriesR2004.pdf