ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

# ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΧΡΟΝΟΣΕΙΡΩΝ ΜΕ ΑΠΑΡΙΘΜΗΤΑ ΔΕΔΟΜΕΝΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΘΕΟΔΟΣΟΠΟΥΛΟΥ ΚΩΝΣΤΑΝΤΙΝΟΥ

**ΕΠΙΒΛΕΠΩΝ:** ΚΑΡΩΝΗ ΧΡΥΣΗΣ, ΚΑΘΗΓΗΤΡΙΑ Ε.Μ.Π

**ΕΠΙΤΡΟΠΗ:** ΙΩΑΝΝΗΣ ΠΟΛΥΡΑΚΗΣ, ΦΙΛΙΑ ΒΟΝΤΑ

ΑΘΗΝΑ, ΜΑΡΤΙΟΣ 2017

# Ευχαριστίες

Πρώτα απ' όλα, θέλω να ευχαριστήσω την επιβλέπουσα της διπλωματικής εργασίας μου, Καθηγήτρια κ. Χρυσηίς Καρώνη, για την πολύτιμη βοήθεια και καθοδήγησή της κατά τη διάρκεια της δουλειάς μου. Επίσης, είμαι ευγνώμων στα υπόλοιπα μέλη της εξεταστικής επιτροπής της διπλωματικής εργασίας μου, Καθηγητές κ. Ιωάννη Πολυράκη και κ. Φιλία Βόντα για την προσεκτική ανάγνωση της εργασίας μου. Ευχαριστώ τη συνάδελφο μου Σπυριδούλα Αλμπάνη, την αδερφή μου Ζωή Θεοδοσοπούλου και το φίλο μου Γιάννη Κοντέα για την ηθική τους υποστήριξη αλλά και για την συνδρομή τους σε κάποια σημεία της εργασίας. Τέλος, είμαι ευγνώμων και στους γονείς μου, Νικόλαο Θεοδοσόπουλο και Παρασκευή Σταυροπούλου για την ολόψυχη αγάπη και υποστήριξή τους όλα αυτά τα χρόνια. Αφιερώνω αυτή την εργασία σ'όλους τους παραπάνω.

Κων/νος Θεοδοσόπουλος

**Περίληψη**

Η παρούσα διπλωματική εργασία πραγματεύεται την ανάπτυξη μοντέλων παλινδρόμησης χρονοσειρών όπου τα δεδομένα είναι απαριθμητά. Για το σκοπό αυτό θα εστιάσουμε στην παλινδρόμηση Poisson η οποία ως ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων είναι η καταλληλη για την μοντελοποίηση απαριθμητών χρονοσειρών. Τα μοντέλα παλινδρόμησης χρονοσειρών είναι ένα χρήσιμο και απαραίτητο εργαλείο το οποίο πολλοί επιστήμονες μπορούν να χρήσιμοποιήσουν στις μελέτες τους. Η θεωρία αναλύεται λεπτομερώς και η εφαρμογή τους σε απαριθμητά δεδομένα μέσω πολλών παραδειγμάτων.

Αρχικά επισημαίνουμε την μεγάλη σημασία που έχει η πρόβλεψη σε διάφορους τομείς. Στην συνέχεα εισάγουμε την έννοια της χρονοσειράς καθώς και τα είδη χρονοσειρών τα οποία γίνονται περισσότερο κατανοητά με τη χρήση κατάλληλων παραδειγμάτων και γραφικών παραστάσεων. Είναι απαραίτητο να ορίσουμε τα μοντέλα χρονοσειρών όπως MA, AR(1), AR(p) και ARMA και εισάγουμε τις ιδέες των γενικευμένων γραμμικών μοντέλων για τη μοντελοποίηση χρονοσειρών με τη χρήση της μερικής πιθανοφάνειας , καθώς διάφρων στατιστικών ιδιοτήτων και διάφορων τεχνικών ανάλυσης αυτών των μοντέλων. Ιδιαίτερη προσοχή δίνουμε στους διαγνωστικούς ελέγχους, ελέγχους υποθέσεων, στην ανάπτυξη βέλτιστων μοντέλων και στην Intervation ανάλυση. Στα παραδείγματα αυτά παρουσιάζεται η χρήση των των στατιστικών προγραμμάτων MINITAB και R αλλά και η ερμηνεία των αποτελεσμάτων.

**Abstract**

This thesis deals with the development of time series regression models for counts data. For this purpose we focus on Poisson regression which is the special case of generalized linear models that is most suitable for modeling time series of counts. Regression modelling of time series is a useful and necessary tool that many scientists can apply in their studies.

We note the great importance of the forecasting in the various sectors. We introduce the definition of countinuous time series and the types of time series using appropriate examples and graphical presentation. We define time series models such as MA, AR (1), AR (p) and ARMA and mention the ideas of generalized linear models for modeling time series using the partial likelihood, as well as statistical properties and various analysis techniques for these models. Special attention is given to goodness of fit, hypothesis testing, development of optimal models and Intervation analysis. The examples illustrate the use of the statistical programs such as MINITAB and R, and the interpretation of results.

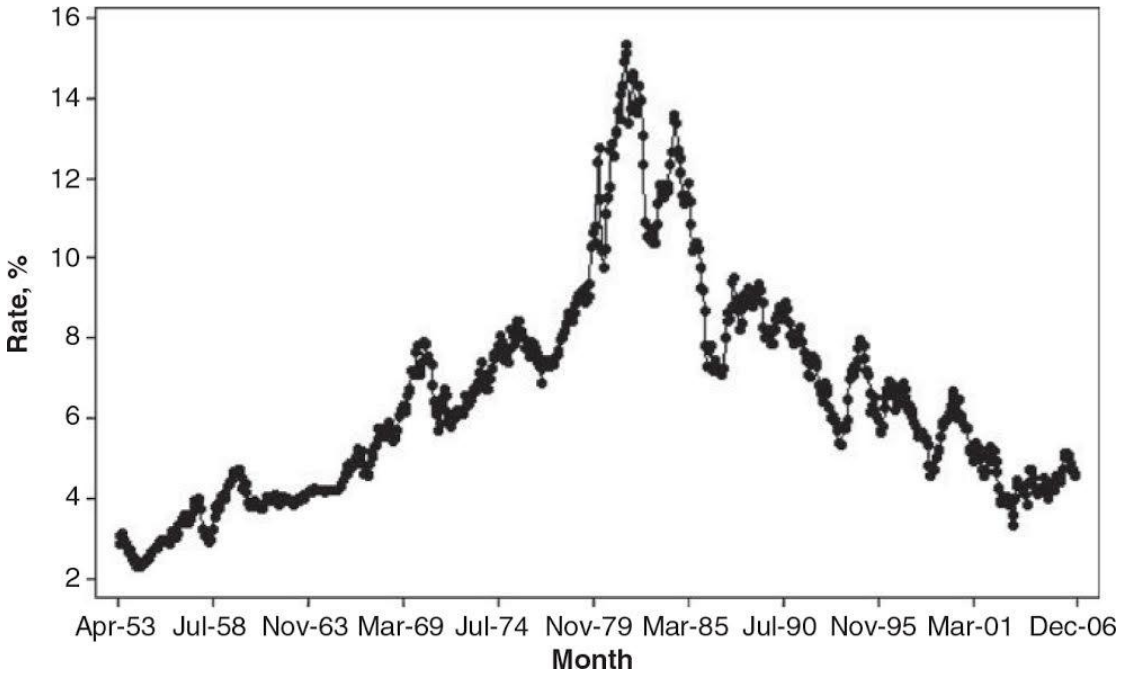
**ΠΕΡΙΕΧΟΜΕΝΑ**

# ΚΕΦΑΛΑΙΟ Ι

# ΧΡΟΝΟΣΕΙΡΕΣ

## 1.1 Εισαγωγή για τη πρόβλεψη

Η πρόβλεψη είναι ένα σημαντικό πρόβλημα που εκτείνεται σε πολλούς τομείς, όπως των επιχειρήσεων, της βιομηχανίας, της κυβέρνησης, της οικονομίας, των περιβαλλοντικών επιστημών, της ιατρικής, των κοινωνικών επιστημών, της πολιτικής και των χρηματοοικονομικών. Τα προβλήματα των προβλέψεων μπορούν να ταξινομηθούν ως βραχυπρόθεσμα, μεσοπρόθεσμα και μακροπρόθεσμα. Οι βραχυπρόθεσμες προβλέψεις περιλαμβάνουν την πρόβλεψη γεγονότων για μικρές χρονικές περιόδους στο μέλλον (Ημέρες, εβδομάδες, μήνες). Οι μεσοπρόθεσμες προβλέψεις επεκτείνονται για 1-2 χρόνια στο μέλλον και οι μακροπρόθεσμες προβλέψεις μπορούν να επεκταθούν για πολλά χρόνια. Βραχυπρόθεσμες και μεσοπρόθεσμες προβλέψεις απαιτούνται για δραστηριότητες που κυμαίνονται από τη διαχείριση των εργασιών κατάρτισης του προϋπολογισμού και την επιλογή νέων έργων έρευνας και ανάπτυξης. Οι μακροπρόθεσμες προβλέψεις έχουν επιπτώσεις σε θέματα όπως ο στρατηγικός σχεδιασμός. Βραχυπρόθεσμες και μεσοπρόθεσμες προβλέψεις βασίζονται συνήθως στον εντοπισμό, τη μοντελοποίηση και την επέκταση των σχεδίων που βρίσκονται σε ιστορικά δεδομένα. Επειδή αυτά τα ιστορικά στοιχεία που συνήθως επιδεικνύουν αδράνεια και δεν αλλάζουν δραματικά πολύ γρήγορα, οι στατιστικές μέθοδοι είναι πολύ χρήσιμες για τη βραχυπρόθεσμη και μεσοπρόθεσμη διάρκεια πρόβλεψης (Montgomery et al., 2015). Τα περισσότερα προβλήματα προβλέψεων περιλαμβάνουν τη χρήση των δεδομένων χρονοσειρών. Μια χρονοσειρά είναι μια χρονολογική σειρά παρατηρήσεων μιας μεταβλητής ενδιαφέροντος. Για παράδειγμα, το Σχήμα 1.1 δείχνει την απόδοση της αγοράς ομολόγων στις ΗΠΑ σε 10 χρόνια με σταθερή ημερομηνία λήξης από τον Απρίλιο του 1953, μέχρι το Δεκεμβρίου 2006 (Παράρτημα Β, Πίνακας Β.1, Montgomery et al., 2015). Αυτό το γράφημα ονομάζεται γραφική παράσταση χρονοσειράς.




**Σχήμα 1.1**: Γραφική παράσταση χρονοσειράς της απόδοσης της αγοράς ομολόγων των ΗΠΑ με σταθερή ημερομηνία λήξη 10 χρόνια. Πηγή: Υπουργείο Οικονομικών των ΗΠΑ.

Η τιμή της απόδοσης συλλέγεται σε ισαπέχοντα χρονικά διαστήματα, όπως είναι χαρακτηριστικό στις περισσότερες εφαρμογές χρονοσειρών και προβλέψεων. Πολλές επιχειρήσεις χρησιμοποιούν την εφαρμογή πρόβλεψης καθημερινά, εβδομαδιαία, μηνιαία, τριμηνιαία ή μπορούν να χρησιμοποιηθούν τα ετήσια στοιχεία και όποιο άλλο διάστημα τους είναι χρήσιμο. Επιπλέον, τα δεδομένα μπορεί να είναι στιγμιαία, όπως το ιξώδες ενός χημικού προϊόντος στη χρονική στιγμή όπου μετράται. Μπορεί να είναι και συσωρευτικά, όπως ο συνολικός αριθμός πωλήσεων ενός προϊόντος κατά τη διάρκεια του μήνα. Ακόμα μπορεί να είναι μια στατιστική συνάρτηση η οποία με ορισμένους τρόπους αντικατοπτρίζει την δραστηριότητα της μεταβλητής κατά τη διάρκεια της χρονικής περιόδου, όπως το κλείσιμο μιας ημερήσιας τιμής μιας συγκεκριμένης μετοχής στο Χρηματιστήριο της Νέας Υόρκης.

Η πρόβλεψη μελλοντικών γεγονότων παίζει σπουδαίο σε πολλά είδη σχεδιασμού και διαδικασιών λήψης αποφάσεων, με εφαρμογή σε τομείς όπως οι ακόλουθες:

*1. Διοίκηση Επιχειρήσεων*. Οι επιχειρηματικές οργανώσεις χρησιμοποιούν συνήθως προβλέψεις των πωλήσεων του προϊόντος ή της ζήτησης για τις υπηρεσίες προκειμένου να προγραμματιστεί η παραγωγή, τα αποθέματα έλεγχου, η διαχείριση

της εφοδιαστικής αλυσίδας, ο καθορισμός των απαιτήσεων του προσωπικού, και την ικανότητα σχεδίου. Οι προβλέψεις μπορούν επίσης να χρησιμοποιηθούν για να προσδιοριστούν οι υπηρεσίες που θα προσφέρονται και οι περιοχές στις οποίες τα προϊόντα αυτά πρέπει να παραχθούν.

*2. Marketing*. Η πρόβλεψη είναι σημαντική σε πολλές αποφάσεις μάρκετινγκ. Οι προβλέψεις της ανταπόκρισης των πωλήσεων και για τη δαπάνη των διαφημίσεων, οι νέες προσφορές ή οι αλλαγές στις πολιτικές τιμολόγησης επιτρέπουν στις επιχειρήσεις να αξιολογούν την αποτελεσματικότητά τους, να καθορίζουν εάν οι στόχοι επιτυγχάνονται και να κάνουν προσαρμογές.

*3. Οικονομικά και Διαχείρισης Κινδύνων*. Οι επενδυτές σε χρηματοοικονομικά περιουσιακά στοιχεία ενδιαφέρονται στην πρόβλεψη των αποδόσεων από τις επενδύσεις τους. Αυτά τα περιουσιακά στοιχεία περιλαμβάνουν αλλά δεν περιορίζονται σε μετοχές, ομόλογα και εμπορεύματα. Οι επενδυτικές αποφάσεις μπορούν να γίνουν σε σχέση με τις προβλέψεις της τιμής ενδιαφέροντος, τις επιλογές και τις συναλλαγματικές ισοτιμίες. Η διαχείριση χρηματοοικονομικού κινδύνου απαιτεί προβλέψεις της μεταβλητότητας των αποδόσεων των περιουσιακών στοιχείων, έτσι ώστε οι κίνδυνοι που συνδέονται με επενδυτικά χαρτοφυλάκια να μπορούν να αξιολογηθούν και να ασφαλιστούν, οπότε τα χρηματοοικονομικά παράγωγα να μπορούν να τιμολογηθούν σωστά.

*4. Οικονομικά*. Οι κυβερνήσεις, τα χρηματοπιστωτικά ιδρύματα και οι πολιτικές οργανώσεις απαιτούν προβλέψεις μεγάλων οικονομικών μεταβλητών, όπως το ακαθάριστο εγχώριου προϊόν, η αύξηση του πληθυσμού, η ανεργία, ο πληθωρισμός, η αύξηση των θέσεων εργασίας, η παραγωγή και η κατανάλωση. Οι προβλέψεις αυτές αποτελούν ένα αναπόσπαστο μέρος της καθοδήγησης πίσω από τη νομισματική και δημοσιονομική πολιτική, τα σχέδια του προϋπολογισμού ακόμη και οι αποφάσεις που λαμβάνονται από τις κυβερνήσεις. Επίσης οι προβλέψεις συντελούν στο στρατηγικό σχεδιασμό των αποφάσεων από τις επιχειρηματικές οργανώσεις και τα χρηματοπιστωτικά ιδρύματα.

*5. Έλεγχος βιομηχανικών διεργασιών*. Οι προβλέψεις των μελλοντικών τιμών των κρίσιμων ποιοτικών χαρακτηριστικών της παραγωγικής διαδικασίας μπορεί να βοηθήσει και να την προσδιορίσει, όταν σημαντικές ελεγχόμενες μεταβλητές στη διαδικασία θα πρέπει να αλλάξουν, ή αν η διαδικασία θα πρέπει να κλείσει και να

ανανεωθεί. Η ανατροφοδότηση και τα συστήματα ελέγχου feedforward χρησιμοποιούνται ευρέως για την παρακολούθηση και προσαρμογή των βιομηχανικών διεργασιών. Οι προβλέψεις της διαδικασίας εξόδου αποτελούν αναπόσπαστο μέρος αυτών των συστημάτων.

*6. Δημογραφία*. Οι προβλέψεις του πληθυσμού ανά χώρα και περιφέρειες γίνονται τακτικά, συχνά χωρίζονται από μεταβλητές όπως το φύλο, η ηλικία και η φυλή. Οι δημογράφοι προβλέπουν επίσης τις γεννήσεις, τους θανάτους και τη μετανάστευση των πληθυσμών. Οι κυβερνήσεις χρησιμοποιούν αυτές τις προβλέψεις για το σχεδιασμό πολιτικών και κοινωνικών υπηρεσιών δράσης, όπως οι δαπάνες για την υγειονομική περίθαλψη, προγράμματα συνταξιοδότησης, καθώς και τα προγράμματα για τη μείωση της φτώχειας. Πολλές επιχειρήσεις χρησιμοποιούν τις προβλέψεις των πληθυσμών για να κάνουν στρατηγικά σχέδια σχετικά με την ανάπτυξη νέων γραμμών παραγωγής ή να δημιουργήσουν τους τύπους των υπηρεσιών που θα προσφέρονται.

Αυτές είναι μόνο μερικές από τις πολλές διαφορετικές περιπτώσεις όπου οι προβλέψεις απαιτούνται για να κάνουν καλές αποφάσεις. Παρά το ευρύ φάσμα των καταστάσεων που απαιτούν προβλέψεις, υπάρχουν μόνο δύο μεγάλες κατηγορίες πρόβλεψης τεχνικών-ποιοτικών μεθόδων και ποσοτικών μεθόδων.

*Ποιοτικές τεχνικές πρόβλεψης (Qualitative forecasting)* έχουν συχνά υποκειμενικό χαρακτήρα και απαιτούν κρίση εκ μέρους των εμπειρογνωμόνων. Ποιοτικές προβλέψεις συχνά χρησιμοποιούνται σε περιπτώσεις όπου υπάρχουν λίγα ή καθόλου ιστορικά στοιχεία τα οποία θα βοηθούσαν στη πρόβλεψη. Ένα παράδειγμα θα ήταν η εισαγωγή ενός νέου προϊόντος στην αγορά, η οποία δεν υπάρχει σχετική ιστορία. Σε αυτήν την περίπτωση, η εταιρεία θα μπορούσε να χρησιμοποιήσει τη γνωμοδότηση των πωλήσεων και του προσωπικού μάρκετινγκ ώστε να εκτιμηθούν υποκειμενικά οι πωλήσεις των προϊόντων κατά τη διάρκεια της νέας φάσης εισαγωγής του προϊόντος στο κύκλο ζωής του. Μερικές φορές οι μέθοδοι ποιοτικής πρόβλεψης κάνουν χρήση των δοκιμών μάρκετινγκ, ερευνών των πελατών, και την εμπειρία με την απόδοση των πωλήσεων άλλων προϊόντων (τόσο τη δική τους όσο και των ανταγωνιστών της). Ωστόσο, αν και μπορεί να πραγματοποιηθεί κάποια ανάλυση των δεδομένων, η βάση της πρόβλεψης αποτελεί υποκειμενική κρίση.

*Οι ποσοτικές τεχνικές πρόβλεψης (Quantitative forecasting)* κάνουν χρήση ιστορικών δεδομένων σ' ένα μοντέλο πρόβλεψης. Το μοντέλο συνοψίζει μοτίβα δεδομένων και εκφράζει μια στατιστική σχέση μεταξύ της προηγούμενης και της τρέχουσας τιμής της μεταβλητής. Το μοντέλο πρόβλεψης χρησιμοποιείται για να προεκτείνει το παρελθόν σε συνδυασμό με τις τρέχουσες συμπεριφορές στο μέλλον. Υπάρχουν αρκετοί τύποι των μοντέλων πρόβλεψης σε γενική χρήση. Τα τρία πιο ευρέως χρησιμοποιημένα είναι τα μοντέλα παλινδρόμησης, τα μοντέλα εξομάλυνσης και τα γενικά μοντέλα χρονοσειρών. Τα μοντέλα παλινδρόμησης κάνουν χρήση των σχέσεων μεταξύ της μεταβλητής που μας ενδιαφέρει και μία ή περισσότερες προβλέπουσες ή επεξηγηματικές μεταβλητές. Μερικές φορές τα μοντέλα παλινδρόμησης ονομάζονται αιτιώδη μοντέλα πρόβλεψης (causal forecasting models), επειδή οι μεταβλητές πρόβλεψης ενδέχεται να περιγράψουν τις δυνάμεις που προκαλούν ή οδηγούν στη παρατηρούμενη τιμή της μεταβλητής ενδιαφέροντος. Ένα παράδειγμα θα ήταν η χρήση δεδομένων αγοράς σπιτιού ως δείκτης για την πρόβλεψη των πωλήσεων επίπλων. Η μέθοδος των ελαχίστων τετραγώνων είναι η βάση και τεχνική των περισσότερων μοντέλων παλινδρόμησης. Τα μοντέλα εξομάλυνσης συνήθως χρησιμοποιούν μια απλή συνάρτηση των προηγούμενων παρατηρήσεων και παρέχουν μια πρόβλεψη της μεταβλητής ενδιαφέροντος. Αυτές οι μέθοδοι μπορεί να έχουν μια επίσημη στατιστική βάση, αλλά συχνά χρησιμοποιούνται με βάση το ότι είναι εύκολο στη χρήση και παραγωγή ικανοποιητικών αποτελεσμάτων. Γενικώς τα μοντέλα χρονοσειρών χρησιμοποιούν τις στατιστικές ιδιότητες των ιστορικών δεδομένων για να καθορίσουν ένα τυπικό μοντέλο και τότε εκτιμούν τις άγνωστες παραμέτρους αυτού του μοντέλου (συνήθως) με τη μέθοδο ελαχίστων τετραγώνων. Μια σημειακή πρόβλεψη ακολουθείται πάντα από ένα διάστημα εμπιστοσύνης πρόβλεψης. Έτσι λαμβάνουμε υπόψη το πόσο μεγάλο σφάλμα θα μπορεί να έχει μια σημειακή πρόβλεψη. Το διάστημα εμπιστοσύνης πρόβλεψης (PI) είναι ένα εύρος τιμών για μια μελλοντική παρατήρηση και είναι πιθανό να αποδειχθεί πολύ χρήσιμο στη διαδικασία λήψης αποφάσεων σε αντίθεση με τη σημειακή πρόβλεψη.

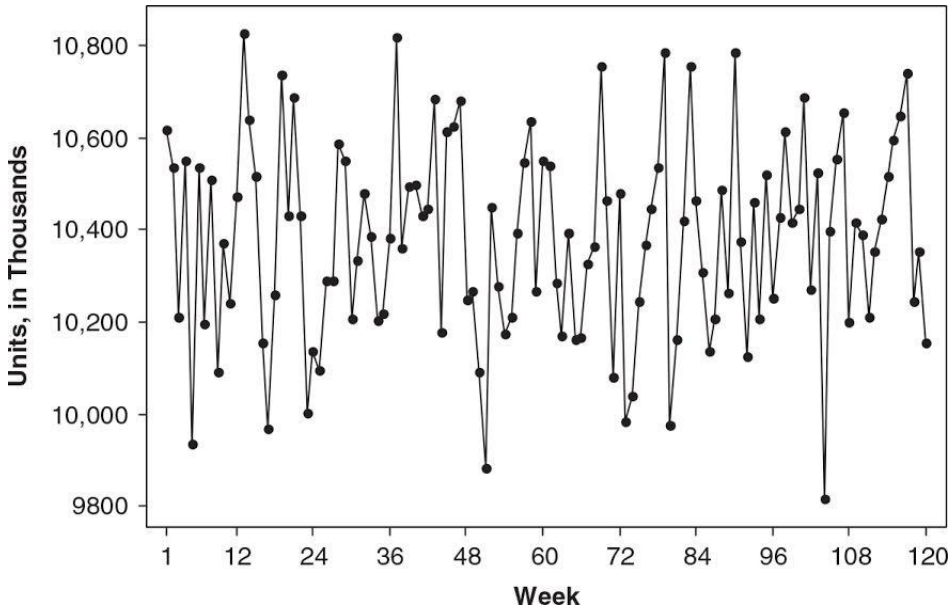
Τώρα αυτές οι προβλέψεις είναι σχεδόν πάντα λάθος διότι υπάρχει το σφάλμα πρόβλεψης. Ως εκ τούτου, είναι συνήθως μια καλή πρακτική το να συνοδεύεται μια πρόβλεψη με μια εκτίμηση του πόσο μεγάλο σφάλμα πρόβλεψης υπάρχει. Ένας τρόπος να γίνει αυτό είναι να παρέχει ένα διάστημα πρόβλεψης (PI). Το διάστημα

πρόβλεψης (PI) είναι ένα εύρος τιμών για μια μελλοντική παρατήρηση και είναι πιθανό να αποδειχθεί πολύ χρήσιμο στη διαδικασία λήψης αποφάσεων σε αντίθεση με τη σημειακή πρόβλεψη.

## 1.2 Παραδείγματα χρονοσειρών

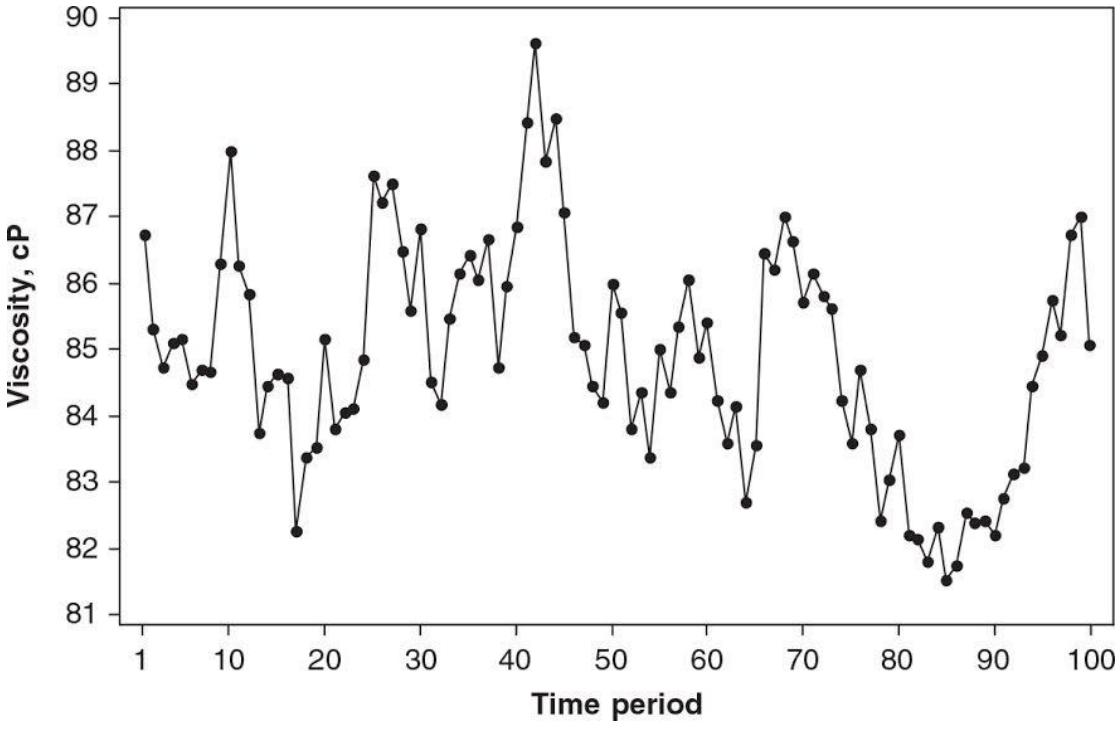
Οι γραφικές παραστάσεις χρονοσειρών μπορεί να αποκαλύψουν μοτίβα (δλδ πληροφορίες ή δομές), όπως τυχαιότητα, τάση, μετατοπίσεις επίπεδου, περιοδικότητα, ασυνήθιστες παρατηρήσεις ή ένα συνδυασμό των προηγούμενων. Τα μοτίβα αυτά βρίσκονται συνήθως στα δεδομένα χρονοσειρών και στη συνέχεια θα αναφέρουμε παραδείγματα αυτών των καταστάσεων που οδηγούν στα παρακάτω σχήματα.

Α) Οι πωλήσεις ενός φαρμακευτικού προϊόντος μπορεί να παραμείνουν σχετικά επίπεδες με την απουσία της αμετάβλητης στρατιγικής για την εμπορίας ή τη παραγωγή. Εβδομαδιαίες πωλήσεις ενός φαρμακευτικού προϊόντος όπως φαίνεται στο Σχήμα 1.2 φαίνεται να είναι μια τυχαία ακολουθία σταθερή στο χρόνο (Παράρτημα Β, Πίνακας Β.2, Montgomery et al., 2015).




**Σχήμα 1.2**: Πωλήσεις φαρμακευτικών προϊόντων.

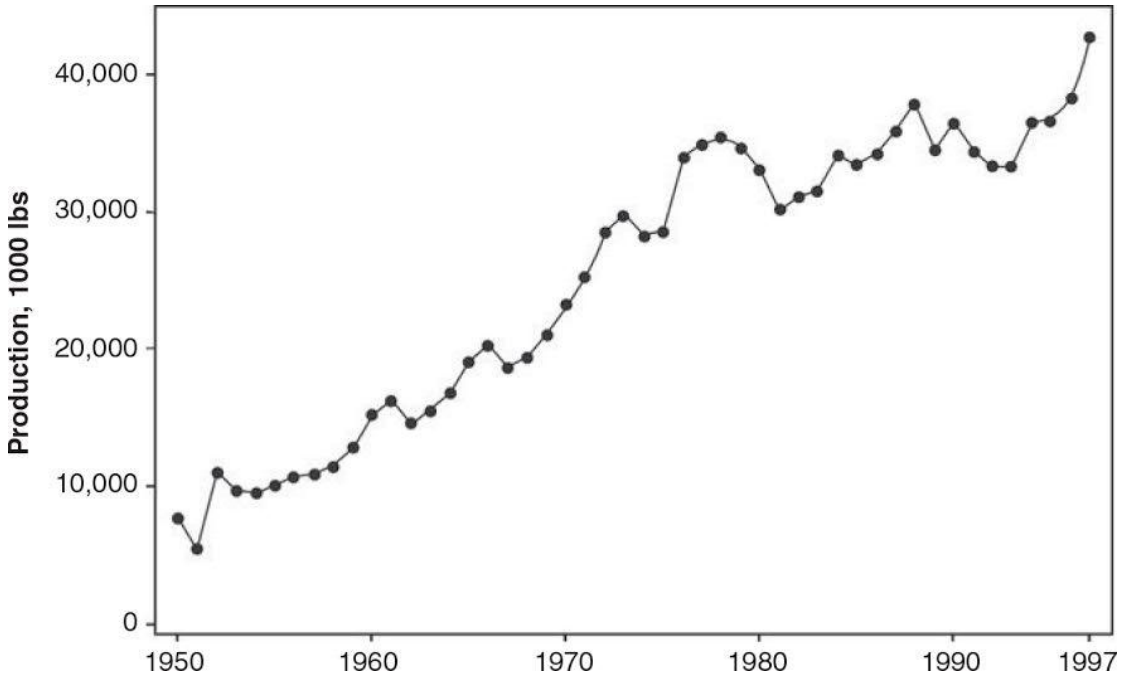
Β) Για να εξασφαλιστεί η συμμόρφωση με τις απαιτήσεις του πελάτη και τις προδιαγραφές των προϊόντων, η παραγωγή χημικών παρακολουθείται ως προς πολλά χαρακτηριστικά. Αυτά μπορεί να είναι μεταβλητές εισόδου, όπως η θερμοκρασία και ο ρυθμός ροής, και η έξοδος των ιδιοτήτων όπως το ιξώδες (viscosity) και την καθαρότητα. Λόγω της συνεχούς φύσης των διεργασιών της χημικής παραγωγής, οι ιδιότητες εξόδου συχνά είναι θετικά αυτοσυσχετισμένες. Δηλαδή αν μια τιμή είναι πάνω από το μακροπρόθεσμο μέσο όρο, τότε τείνει να ακολουθήσουν και άλλες τιμές πάνω από το μέσο όρο, ενώ μια τιμή κάτω του μέσου όρου τείνει να ακολουθείται και από άλλες τιμές κάτω από το μέσο όρο. Οι μετρήσεις ιξώδους απεικονίζονται στο Σχήμα 1.3 παρουσιάζουν αυτοσυσχετισμένη συμπεριφορά, που τείνει σε ένα μακροχρόνιο μέσο όρο περίπου 85 (cP), δηλαδή δεν συμπεριφέρεται εντελώς τυχαία, έχει κάποια δομή (Παράρτημα Β, Πίνακας Β.3, Montgomery et al., 2015).




**Σχήμα 1.3**: Χημικές αναγνώσεις της διαδικασίας του ιξώδους.

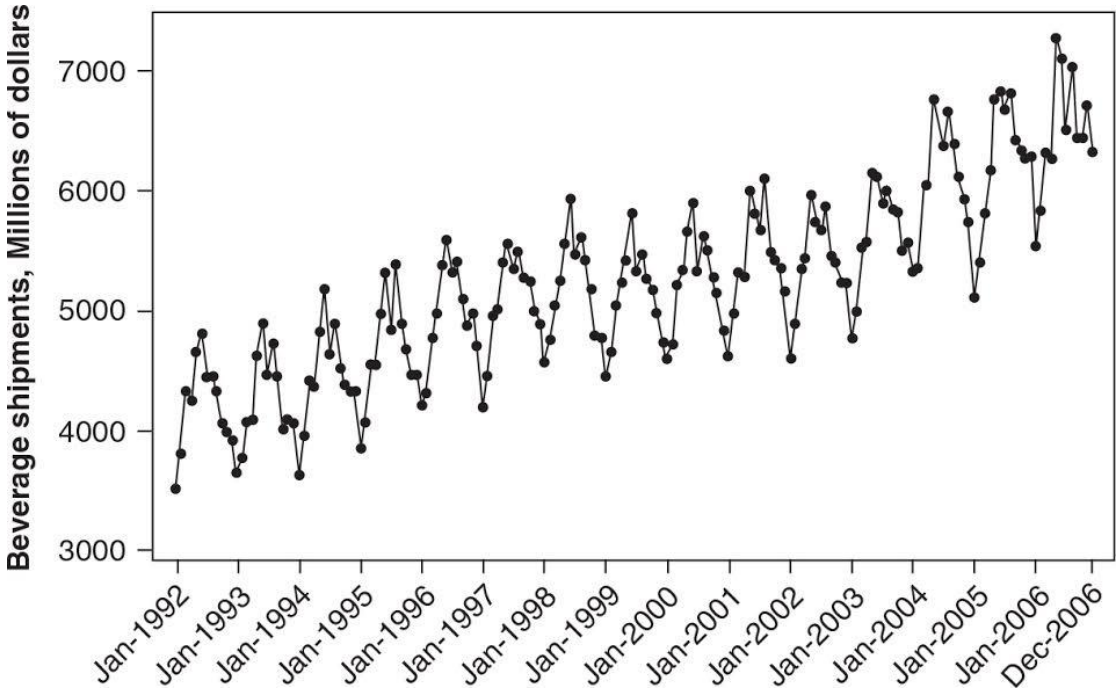
Γ) Το USDA (Εθνική Αγροτική Στατιστική Υπηρεσία) δημοσιεύει στατιστικά στοιχεία για πολλά γεωργικά αγαθά, συμπεριλαμβανομένης της ετήσιας παραγωγής γαλακτοκομικών προϊόντων όπως το βούτυρο, το τυρί, το παγωτό, το γάλα και το γιαούρτι. Αυτά τα στατιστικά στοιχεία χρησιμοποιούνται για την ανάλυση της αγοράς,για δείκτες οικονομίας και για τον προσδιορισμό των αναδυόμενων ζητημάτων.

Το τυρί μπλε γκοργκοντζόλα είναι ένα από τα 32 είδη τυριών τα οποία είναι δημοσιευμένα. Η ετήσια παραγωγή του μπλε γκοργκοντζόλα τυριού στις ΗΠΑ (σε 1000 lb) παρουσιάζεται στο Σχήμα 1.4 (Παράρτημα Β, Πίνακας Β.4, Montgomery et al., 2015). Παρατηρούμε ότι η παραγωγή τετραπλασιάστηκε ανάμεσα στο 1950-1997, και η γραμμική τάση έχει μια σταθερή θετική κλίση με τυχαία, από έτος σε έτος διακύμανση. Η υπηρεσία απογραφής των ΗΠΑ δημοσιεύει ιστορικά στατιστικά στοιχεία για τους κατασκευαστές, τις αποστολές, τα αποθέματα και τις παραγγελίες. Οι στατιστικές αυτές βασίζονται στο Βόρειο Αμερικανικό Σύστημα Βιομηχανικής Ταξινόμησης (NAICS) και χρησιμοποιούνται για τη μέτρηση της παραγωγικότητας και την ανάλυση των σχέσεων μεταξύ της απασχόλησης και της βιομηχανικής παραγωγής.




**Σχήμα 1.4**: Η ετήσια παραγωγή του τυριού μπλε και γκοργκοντζόλα στις ΗΠΑ. Πηγή: USDA-NASS.

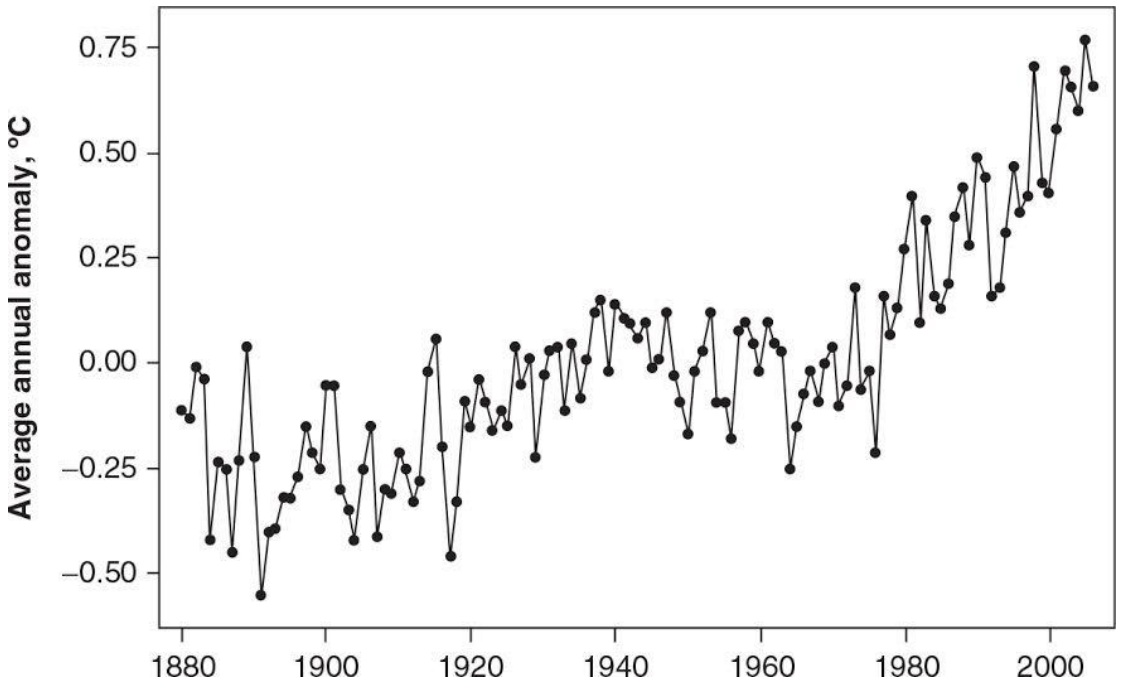
Δ) Η παραγωγή των ποτών και των προϊόντων καπνού ανήκουν σε υποτομέα των αγαθών. Η γραφική παράσταση των μηνιαίων αποστολών των ποτών (Σχήμα 1.5) αποκαλύπτει μια γενική αυξητική τάση, με ένα ξεχωριστό κυκλικό μοτίβο που επαναλαμβάνεται σε κάθε έτος. Οι αποστολές του Ιανουαρίου φαίνεται να είναι οι χαμηλότερες, πιο ψηλά δείχνει ο Μάιος και ο Ιούνιος (Παράρτημα Β, Πίνακας Β.5, Montgomery et al., 2015). Αυτή η μηνιαία ή εποχιακή διακύμανση μπορεί να αποδοθεί σε κάποια αιτία όπως η επίδραση των καιρικών συνθηκών στη ζήτηση των ποτών.




**Σχήμα 1.5**: Η κατασκευή ποτών στις ΗΠΑ σε μηνιαίες αποστολές των προϊόντων. Πηγή: US Census Bureau.

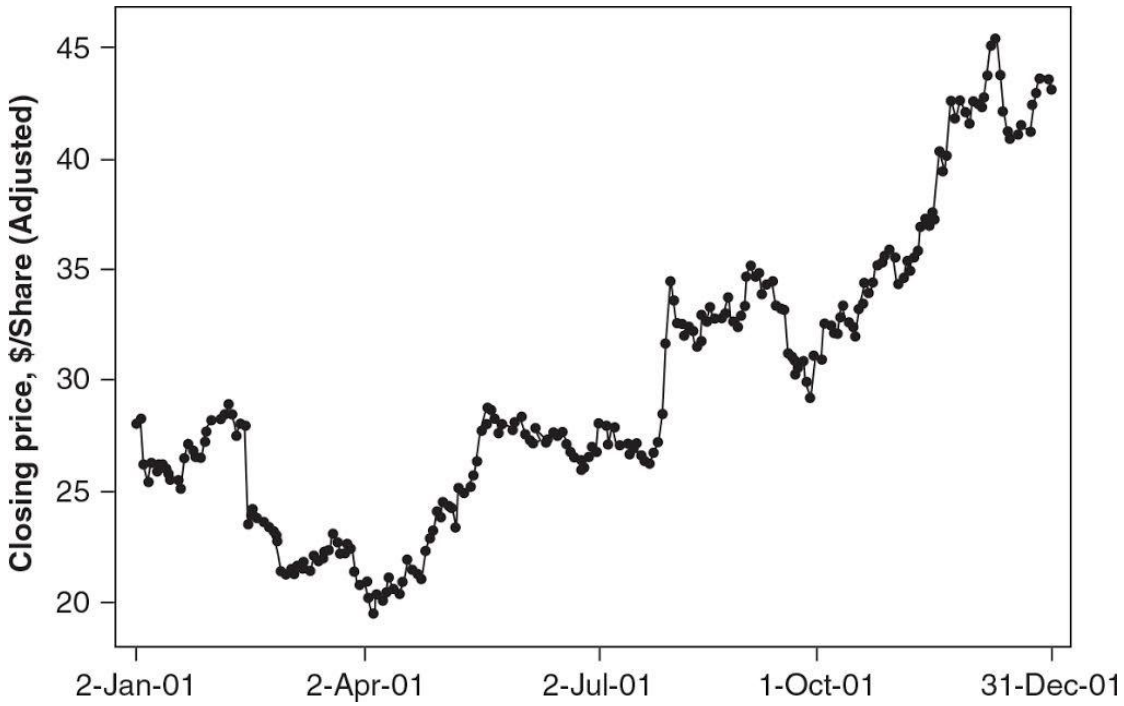
Ε) Για να καθοριστεί η αύξηση ή πτώση της θερμοκρασίας της Γης, οι επιστήμονες παρατηρούν τις ετήσιες μέσες θερμοκρασίες. Ανά σταθμό υπολογίζονται οι ημερίσιοι μέσοι όροι της υψηλότερης και χαμηλότερης θερμοκρασίας. Στη συνέχεια υπολογίζονται οι μέσοι όροι σε σταθμούς σε όλη τη Γη, για έναν ολόκληρο χρόνο. Η αλλαγή στην παγκόσμια ετήσια μέση θερμοκρασία της επιφάνειας του αέρα υπολογίζεται από μια βάση που έχει καθιερωθεί από το 1951 έως 1980, και το αποτέλεσμα αναφέρεται ως "ανωμαλία". Η γραφική παράσταση της ετήσιας μέσης "ανωμαλίας" της παγκόσμιας θερμοκρασίας της επιφάνειας του αέρα (Σχήμα 1.6) παρουσιάζει αυξητική τάση από το 1880. Ωστόσο, η κλίση ή ο ρυθμός μεταβολής ποικίλλει ανάλογα με τις χρονικές περιόδους (Παράρτημα Β, Πίνακας Β.6,

Montgomery et al., 2015). Ενώ η κλίση σε προηγούμενες χρονικές περιόδους φαίνεται να είναι σταθερή υπάρχει ελαφρώς μια αύξηση στην κλίση περίπου από το 1975 μέχρι σήμερα και φαίνεται πολύ πιο απότομη σχετικά με τη περίοδο πριν το 1975.



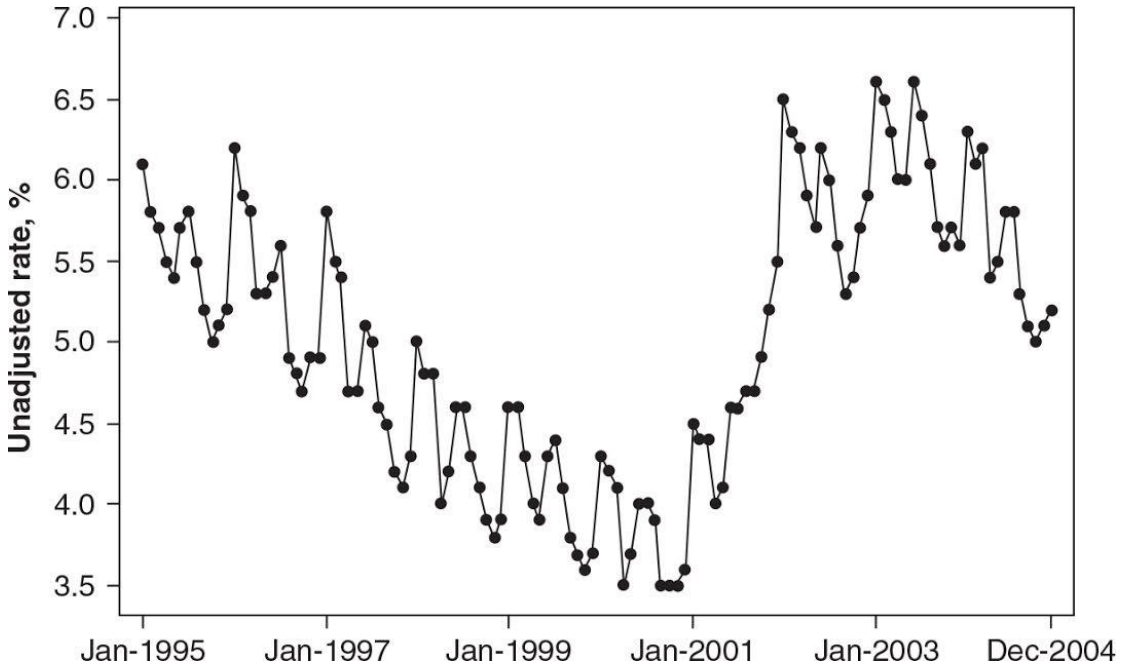

**Σχήμα 1.6**: Η παγκόσμια μέση θερμοκρασία επιφάνειας ετήσια ανωμαλία. Πηγή: NASA-GISS.

Ζ) Τα δεδομένα των επιχειρήσεων, όπως οι τιμές των μετοχών και των επιτοκίων συχνά παρουσιάζουν μη στάσιμες συμπεριφορές. Δηλαδή, η χρονοσειρά δεν έχει καμία φυσική έννοια. Η ημερήσια τιμή κλεισίματος προσαρμόζεται για τις διασπάσεις των μετοχών της Whole Foods Market (WFMI) μετοχής του 2001 (Σχήμα 1.7) παρουσιάζοντας ένα συνδυασμό μοτίβων (patterns) για το επίπεδο του μέσου και της κλίσης (Παράρτημα Β, Πίνακας Β.7, Montgomery et al., 2015). Ενώ η τιμή είναι σταθερή σε μερικές σύντομες χρονικές περιόδους, δεν υπάρχει μια συνεπής μέση στάθμη ως προς το χρόνο. Σε άλλες χρονικές περιόδους, η τιμή μεταβάλλεται με διαφορετικούς ρυθμούς συμπεριλαμβανομένων περιστασιακών απότομων μεταβολών του επιπέδου.



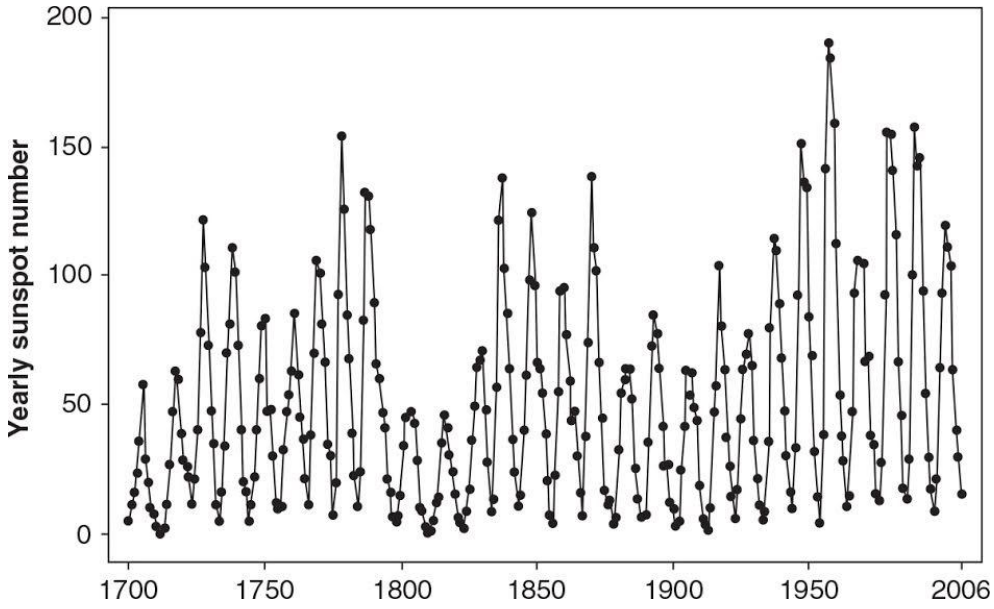

**Σχήμα 1.7**: Η ολόκληρη χρηματιστηριακή τιμή των τροφίμων, το καθημερινό κλείσιμο προσαρμόζεται για τις διασπάσεις.

Η) Η έρευνα του σημερινού πληθυσμού (CPS) ή "έρευνα νοικοκυριών", που εκπονήθηκε από το Υπουργείο Εργασίας των ΗΠΑ, περιέχει εθνικά στοιχεία για την απασχόληση, την ανεργία, τις αποδοχές και άλλα θέματα για την αγορά εργασίας σε σχέση με δημογραφικά χαρακτηριστικά. Τα δεδομένα χρησιμοποιούνται για την έκθεση σχετικά με την κατάσταση της απασχόλησης, για τις προβολές με αντίκτυπο στις προσλήψεις και στην εκπαίδευση, καθώς και για ένα πλήθος άλλων δραστηριοτήτων επιχειρηματικού σχεδιασμού. Τα δεδομένα αναφέρονται με εποχική διόρθωση για να αφαιρεθεί η επίδραση συστηματικών πληροφοριών. Το γράφημα του μηνιαίου ποσοστού ανεργίας χωρίς εποχιακή διόρθωση παρουσιάζεται στο Σχήμα 1.8 και αποτελείται από σχήματα, όμοια με το Σχήμα 1.5 (Παράρτημα Β , Πίνακας Β.8, Montgomery et al., 2015). Υπάρχει ένα ξεχωριστό κυκλικό μοτίβο ότι στη διάρκεια ενός έτους οι μήνες Ιανουάριος, Φεβρουάριος και Μάρτιος γενικά έχουν τα υψηλότερα ποσοστά ανεργίας. Το συνολικό επίπεδο αλλάζει από μια σταδιακή μείωση και μετά βλέπουμε ότι ακολουθεί μια απότομη αύξηση και στη συνέχεια πάλι μια σταδιακή μείωση.




**Σχήμα 1.8**: Μηνιαίο επιτόκιο-πλήρους ανεργίας του εργατικού δυναμικού, αδιόρθωτο. Πηγή: Υπουργείο Εργασίας-BLS ΗΠΑ.

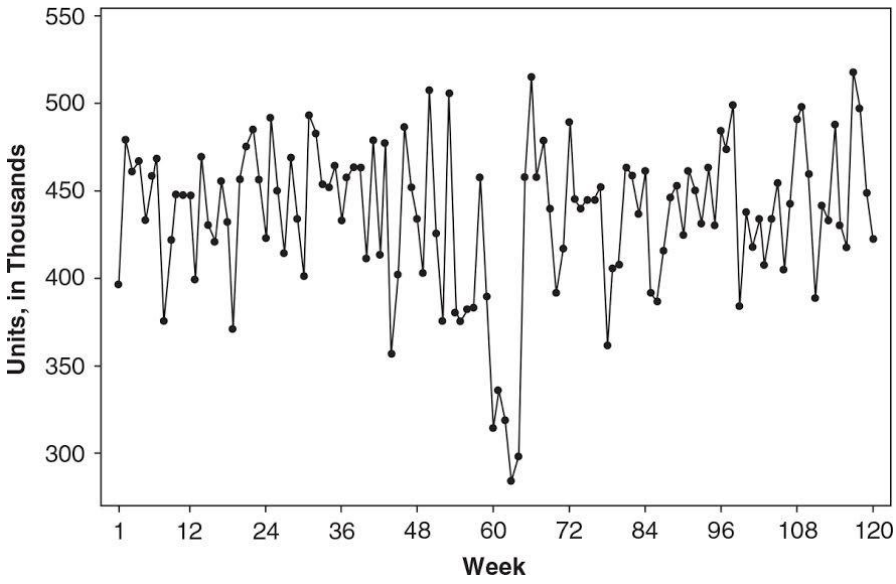
Θ) Η ηλιακή δραστηριότητα έχει από καιρό αναγνωριστεί ως σημαντική πηγή θορύβου επηρεάζοντας τους καταναλωτές και τις στρατιωτικές επικοινωνίες, συμπεριλαμβανομένων των δορυφόρων, των πύργων κινητής τηλεφωνίας και δικτύων ηλεκτρικής ενέργειας. Η ικανότητα πρόβλεψης της ηλιακής δραστηριότητας με ακρίβεια είναι κρίσιμη για διάφορους τομείς. Ο διεθνές αριθμός ηλιακών κηλίδων R είναι ο παλαιότερος δείκτης ηλιακής δραστηριότητας. Αυτός ο αριθμός περιλαμβάνει τον αριθμό των παρατηρούμενων ηλιακών κηλίδων και τον αριθμό των παρατηρούμενων ομάδων των ηλιακών κηλίδων. Στο Σχήμα 1.9, η φραφική παράσταση του ετήσιου αριθμού των ηλιακών κηλίδων αποκαλύπτει κυκλικά σχήματα διαφόρων μεγεθών (Παράρτημα Β, Πίνακας Β.9, Montgomery et al., 2015).




**Σχήμα 1.9**: Ο διεθνής αριθμός των ηλιακών κηλίδων. Πηγή: SIDC.

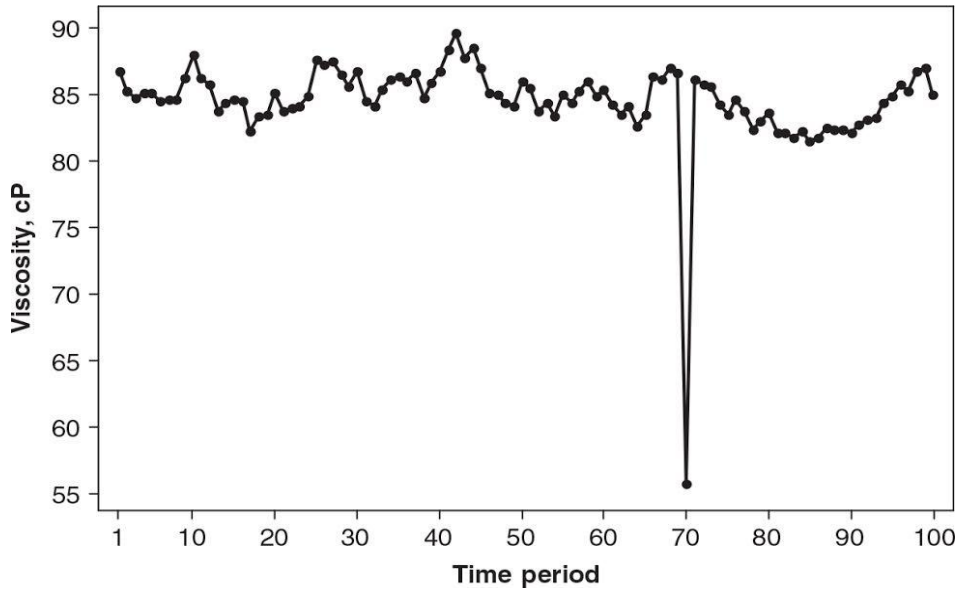
Επιπλέον πέραν του ότι οι γραφικές παραστάσεις των χρονοσειρών μας βοθηθούν να εντοπίσουμε το σωστό μοντέλο, μας βοηθούν να εντοπίσουμε και ακραίες ή ασυνίθιστες παρατηρήσεις. (βλέπε παράγραφο 3.7 καθώς και κεφάλαιο 4)

Ι) Οι εβδομαδιαίες πωλήσεις ενός φαρμακευτικού προϊόντος μειώθηκαν λόγω της περιορισμένης διαθεσιμότητας που προκύπτει από μια πυρκαγιά σε μία από τις τέσσερις εγκαταστάσεις παραγωγής. Η πτώση αυτή είναι εμφανής στο Σχήμα 1.10 της χρονοσειράς των εβδομαδιαίων αυτών πωλήσεων.




**Σχήμα 1.10**: Πωλήσεις φαρμακευτικού προϊόντος

Κ) Ένας άλλο είδος ασυνήθιστων γεγονότων ή σημείων μπορεί να είναι η αποτυχία της μέτρησης δεδομένων ή η λανθασμένη συλλογή δεδομένων. Μετά την εγγραφή έχουμε την μέτρηση ενός πολύ διαφορετικού ιξώδες κατά το χρονικό διάστημα 70 (Σχήμα 1.11), το σύστημα μέτρησης ελέγχθηκε με ένα πρότυπο και προσδιορίστηκε να είναι εκτός βαθμονόμησης. Η αιτία ήταν μια δυσλειτουργία του αισθητήρα.




**Σχήμα 1.11**: Χημικές αναγνώσεις της ιξώδους διαδικασίας, με την δυσλειτουργία του αισθητήρα

## 1.3 Διαδικασία πρόβλεψης

Μια διαδικασία είναι μια σειρά συνδεδεμένων δραστηριοτήτων που μετατρέπουν μια ή περισσότερες εισόδους σε μία ή περισσότερες εξόδους. Όλες οι δραστηριότητες εργασιών που εκτελούνται μέσα στη διαδικασία, και της πρόβλεψης δεν αποτελεί εξαίρεση. Οι δραστηριότητες στη διαδικασία της πρόβλεψης είναι:

1. Ορισμός του προβλήματος
2. Η συλλογή δεδομένων
3. Ανάλυση Δεδομένων
4. Επιλογή μοντέλου και προσαρμογή
5. Επιβεβαίωση εγκυρότητας του μοντέλου
6. Χρήση του μοντέλου πρόβλεψης
7. Παρακολούθηση της επίδοσης του μοντέλου πρόβλεψης

**Ορισμός του προβλήματος** περιλαμβάνει την ανάπτυξη της κατανόησης του πώς η πρόβλεψη θα χρησιμοποιηθεί σε συνδυασμό με τις προσδοκίες του "πελάτη". Οι ερωτήσεις που πρέπει να αντιμετωπιστούν κατά τη διάρκεια αυτής της φάσης περιλαμβάνουν την επιθυμητή μορφή των προβλέψεων (π.χ. απαιτούνται μηνιαίες προβλέψεις), τους ορίζοντες των προβλέψεων ή το χρόνο πόσο συχνά πρέπει να αναθεωρηθούν οι προβλέψεις (το διάστημα πρόβλεψης) και η ακρίβεια προβλέψεων που απαιτείται ώστε να κάνουν καλές επιχειρηματικές αποφάσεις. Αυτό είναι μια ευκαιρία ώστε να εισάγουμε στη λήψη αποφάσεων τα διαστήματα εμπιστοσύνης πρόβλεψης ως μέτρο κινδύνου που συνδέεται με τις προβλέψεις και να εξοικειωθούν οι αναλυτές με αυτή τη προσέγγιση. Συχνά είναι απαραίτητο να πάμε βαθιά σε πολλές πτυχές του συστήματος των επιχειρήσεων ώστε η πρόβλεψη να καθορίζεται σωστά. Για παράδειγμα, στο σχεδιασμό ενός συστήματος πρόβλεψης για τον έλεγχο της απογραφής, μπορεί να απαιτούνται πληροφορίες σχετικά με θέματα όπως η διάρκεια ζωής ενός προϊόντος στο ράφι ή άλλα στοιχεία της γήρανσης, ο χρόνος που απαιτείται για την παραγωγή του προιόντος, καθώς και οικονομικές επιπτώσεις αν πχ είναι διαθέσιμες περισσότερες ή λιγότερες μονάδες του προϊόντος σε σχέση με τη ζήτηση των πελατών. Μεγάλο μέρος στη τελική επιτυχία του μοντέλου πρόβλεψης που να ανταποκρίνεται στη ζήτηση των πελατών καθορίζεται από το σωστό ορισμό του προβλήματος.

*Πρόβλεψη* λοιπόν είναι μια προσπάθεια να γίνει αντιληπτός ο τρόπος με τον οποίο η αγορά και άλλες εξωτερικές και περιβαλλοντικές μεταβλητές θα συμπεριφερθούν κατά τη διάρκεια του προβλεπτικού ορίζοντα μέσα στον οποίο θα δουλεύει μια επιχείρηση. Οι προβλεπτικοί ορίζοντες δείχνουν πόσο συχνά πρέπει να αναθεωρηθούν οι προβλέψεις (το διάστημα πρόβλεψης), και ό, τι το επίπεδο των προβλέψεων απαιτεί ακρίβεια ώστε να παρθούν καλές επιχειρηματικές αποφάσεις.

**Η συλλογή δεδομένων** αποτελείται από τη λήψη του σχετικού ιστορικού για τη μεταβλητή που πρόκειται να προβλέψει, συμπεριλαμβανομένων και των ιστορικών πληροφοριών των πιθανών συμμεταβλητών. Το κλειδί εδώ είναι η συχνά "σχετική" συλλογή και αποθήκευση πληροφοριών, μεθόδων και συστημάτων που αλλάζουν με την πάροδο του χρόνου και δεν είναι όλα τα ιστορικά δεδομένα χρήσιμα για τα τρέχοντα προβλήματα. Συχνά είναι αναγκαία η αντιμετώπιση ελλειπουσών τιμών κάποιων μεταβλητών, πιθανή ύπαρξη ακραίων τιμών, ή άλλα προβλήματα που αφορούν τα δεδομένα που έχουν συμβεί στο παρελθόν. Κατά τη διάρκεια αυτής της

φάσης, είναι επίσης χρήσιμο να αρχίσει η σχεδίαση θεμάτων συλλογής και αποθήκευσης δεδομένων έτσι ώστε η αξιοπιστία και η ακεραιότητα των δεδομένων να διατηρηθεί. Όπως διατυπώνουν οι Montgomery et al. (2015) η ανάπτυξη μοντέλων χρονοσειρών και η χρήση τους για πρόβλεψη απαιτεί στοιχεία των μεταβλητών που παρουσιάζουν ενδιαφέρον για τη λήψη αποφάσεων. Τα δεδομένα αποτελούν τα πρώτα υλικά για τη διαδικασία της μοντελοποίησης και της πρόβλεψης. Οι όροι των δεδομένων και των πληροφοριών συχνά χρησιμοποιούνται εναλλακτικά, αλλά προτιμάμε να χρησιμοποιούμε τον όρο δεδομένα γι αυτά που φαίνονται πιο ακατέργαστα, ενώ θεωρούμε τις πληροφορίες ως κάτι που εξάγεται από τα δεδομένα. Η εξόδος ενός συστήματος πρόβλεψης θα μπορούσε να θεωρηθεί ως πληροφορία και ότι η εξόδος χρησιμοποιεί τα δεδομένα ως είσοδο (Montgomery et al., 2015). Στους περισσότερους σύγχρονους οργανισμούς τα δεδομένα σχετικά με τις πωλήσεις, τις συναλλαγές, τις οικονομικές και επιχειρηματικές επιδόσεις της εταιρείας, τις επιδόσεις του προμηθευτή και τη δραστηριότητα και τη σχέση του πελάτη αποθηκεύονται σε μια αποθήκη γνωστό ως data warehouse. Μερικές φορές αυτό είναι ένα ενιαίο σύστημα αποθήκευσης δεδομένων, αλλά καθώς ο όγκος των δεδομένων που διακινείται από σύγχρονες οργανώσεις αναπτύσσεται με ταχείς ρυθμούς, τα data warehouse έχουν γίνει ένα ολοκληρωμένο σύστημα που αποτελείται από τα συστατικά που είναι φυσικά και συχνά γεωγραφικά κατανεμημένα, όπως τα cloud data storage. Τα data warehouse πρέπει να είναι ικανά να οργανώσουν, να χειριστούν και να ενσωματώσουν τα δεδομένα από πολλαπλές πηγές και διαφορετικά οργανωτικά συστήματα πληροφοριών. Η βασική λειτουργικότητα που απαιτείται περιλαμβάνει την εξαγωγή δεδομένων, το μετασχηματισμό των δεδομένων, καθώς και τη φόρτωση των δεδομένων. Η εξαγωγή δεδομένων αναφέρεται στην απόκτηση δεδομένων από εσωτερικές πηγές και από εξωτερικές πηγές, όπως προμηθευτές ή κρατικούς φορείς και οργανισμούς χρηματοπιστωτικών υπηρεσιών. Τα δεδομένα εξάγονται μια φορά, το στάδιο μετασχηματισμού περιλαμβάνει την εφαρμογή των κανόνων για να αποφεύγεται η επανάληψη των εγγραφών και να αντιμετωπίζονται προβλήματα όπως η έλλειψη πληροφοριών.

**Ανάλυση των δεδομένων** είναι ένα σημαντικό πρώτο βήμα για την επιλογή του μοντέλου πρόβλεψη που θα χρησιμοποιηθεί. Τα διαγράμματα χρονοσειρών των δεδομένων πρέπει να κατασκευαστούν για την αναγνώριση των μοντέλων, όπως η τάση, η εποχικότητα ή άλλες κυκλικές συνιστώσες. Η τάση είναι μια εξελικτική

κίνηση, είτε προς τα πάνω ή προς τα κάτω, στην τιμή της μεταβλητής. Οι τάσεις μπορεί να είναι μακροχρόνιες ή πιο δυναμικές και σχετικά μικρής διάρκειας. Εποχικότητα είναι το στοιχείο της συμπεριφοράς χρονοσειρών που επαναλαμβάνεται σε τακτά χρονικά διαστήματα, όπως κάθε χρόνο. Μερικές φορές εξομαλύνουμε τα δεδομένα ώστε η αναγνώριση από το σχήμα να είναι πιο προφανή. Στοιχεία περιγραφικής στατιστικής όπως ο αριθμητικός μέσος, η τυπική απόκλιση και οι αυτοσυσχετίσεις θα πρέπει να υπολογίζονται και να αξιολογούνται. Άτυπα σημεία πρέπει να επισημαίνονται για πιθανή περαιτέρω μελέτη. Αυτά τα στοιχεία χρειάζονται για να αποκτήσουμε μια αρχική για τα δεδομένα αίσθηση του πόσο ισχυρές είναι οι υποκείμενες μορφές όπως η τάση και η εποχικότητα.

**Επιλογή μοντέλου και προσαρμογή** αποτελείται από την επιλογή ενός ή περισσότερων μοντέλων πρόβλεψης και προσαρμόζοντας το μοντέλο στα δεδομένα. Με την προσαρμογή, εννοούμε την εκτίμηση των άγνωστων παραμέτρων του μοντέλου, αυτό γίνεται συνήθως με τη μέθοδο των ελαχίστων τετραγώνων. Στο επόμενο κεφάλαιο, θα παρουσιαστούν διάφοροι τύποι μοντέλων χρονοσειρών, κυρίως για απαριθμητά δεδομένα όπως θα δούμε στα παραδείγματα πιο κάτω θα παρουσιαστούν και θα σχολιαστούν οι διαδικασίες προσαρμογής του μοντέλου. Θα αναφέρουμε επίσης μεθόδους για την αξιολόγηση της ποιότητας και καταλληλότητας του προσαρμοσμένου μοντέλου, και αν οι βασικές προϋποθέσεις του μοντέλου έχουν παραβιαστεί. Αυτό θα είναι χρήσιμο για να κάνουμε συγκρίσεις μεταξύ των διαφόρων υποψήφιων μοντέλων.

**Η επιβεβαίωση της εγκυρότητας του μοντέλου** αποτελείται από μια αξιολόγηση του μοντέλου πρόβλεψης για να καθορίσει πόσο είναι πιθανό να εκτελέσει την επιδιωκόμενη εφαρμογή. Αυτό πρέπει να πάει πέρα από την απλή αξιολόγηση της "προσαρμογής" του μοντέλου με τα ιστορικά δεδομένα και πρέπει να εξετάσει τι μέγεθος των σφαλμάτων πρόβλεψης θα βιώσει όταν το μοντέλο χρησιμοποιείται για την πρόβλεψη νέων δεδομένων. Τα προσαρμοσμένα σφάλματα θα είναι πάντα μικρότερα από τα σφάλματα πρόβλεψης. Μια ευρέως χρησιμοποιούμενη μέθοδος για την επιβεβαίωση ενός μοντέλου πρόβλεψης είναι η διάσπαση των δεδομένων, όπου τα δεδομένα διαιρούνται σε δύο τμήματα, στο τμήμα προσαρμογής και σ' ένα τμήμα προβλέψεων. Το μοντέλο είναι προσαρμοσμένο ή κατάλληλο μόνο στο τμήμα προσαρμογής δεδομένων, και στη συνέχεια εκτιμούνται οι προβλέψεις από το εν λόγω μοντέλο για τις παρατηρήσεις της κατηγορίας πρόβλεψης. Αυτό μπορεί να

παρέχει χρήσιμες οδηγίες για το πώς το μοντέλο πρόβλεψης θα συμπεριφερθεί όταν εκτίθεται σε νέα δεδομένα. Ακόμα μπορεί να είναι και μια πολύτιμη προσέγγιση για τη διάκριση μεταξύ ανταγωνιστικών μοντέλων πρόβλεψης.

**Η χρήση του μοντέλου πρόβλεψης** από τον πελάτη. Είναι σημαντικό ο πελάτης να κατανοεί τη χρήση του μοντέλου και να παράγει έγκαιρες προβλέψεις από το μοντέλο και αυτό να γίνεται ρουτίνα όσο είναι εφικτό. Η συντήρηση του μοντέλου με παράλληλη συνεχιζόμενη διαθεσιμότητα δεδομένων και άλλων απαιτούμενων πληροφοριών είναι επίσης ένα σημαντικό θέμα για τον πελάτη με επιπτώσεις στην επικαιρότητα και στη τελική χρησιμότητα των προβλέψεων.

**Η παρακολούθηση της απόδοσης του μοντέλου πρόβλεψης** πρέπει να είναι μια συνεχής δραστηριότητα στο μοντέλο που αναπτύχθηκε ώστε να εξακολουθεί να εκτελεί ικανοποιητικά. Οι συνθήκες αλλάζουν με την πάροδο του χρόνου, και ένα μοντέλο που απέδωσε καλά στο παρελθόν ενδέχεται να επιδεινωθεί στην απόδοση. Συνήθως η μείωση απόδοσης του μοντέλου θα έχει ως αποτέλεσμα μεγαλύτερα ή πιο συστηματικά σφάλματα πρόβλεψης. Ως εκ τούτου, η παρακολούθηση των σφαλμάτων πρόβλεψης είναι ένα ουσιαστικό μέρος του ορθού σχεδιασμού του συστήματος πρόβλεψης. Τα διαγράμματα ελέγχου (control charts) των σφαλμάτων στις προβλέψεις είναι ένας απλός αλλά αποτελεσματικός τρόπος για να παρακολουθείται συστηματικά η απόδοση ενός μοντέλου πρόβλεψης.

### 1.4 Παραδείγματα μοντέλων χρονοσειρών

Εδώ θα παρουσιάσουμε τη χρήση μοντέλων παλινδρόμησης για την πρόβλεψη υπό δύο διαφορετικές καταστάσεις. Η πρώτη απ' αυτές είναι η περίπτωση όπου έχουμε μετρήσεις μιας μεταβλητής απόκρισης y και κάποιων επεξηγηματικών μεταβλητών χωρίς εξάρτηση από το χρόνο. Για παράδειγμα, ας υποθέσουμε ότι θέλουμε να αναπτύξουμε ένα μοντέλο παλινδρόμησης για την πρόβλεψη του ποσοστού των καταναλωτών που θα αγοράσουν ένα κουπόνι για την αγορά μιας συγκεκριμένης μάρκας γάλακτος (y) σε συνάρτηση με το ποσό της έκπτωσης ή της ονομαστικής αξίας του κουπονιού (x) καθώς και σε σχέση με άλλες επεξηγηματικές μεταβλητές. Τα δεδομένα αυτά συλλέγονται από μια μελέτη πάνω σε κάποια συγκεκριμένη περίοδο (όπως ένα μήνας) και τα δεδομένα δεν μεταβάλλονται με το χρόνο. Αυτό το είδος των δεδομένων παλινδρόμησης ονομάζονται δεδομένα διατομής (cross-section data). Το μοντέλο παλινδρόμησης για διατομή δεδομένων γράφεται ως

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \,, \mu\varepsilon\ i = 1,2, \ldots, N \qquad (1.1)$$

όπου ο δείκτης $i$ χρησιμοποιείται για να υποδηλώσει κάθε μεμονωμένη παρατήρηση στα δεδομένα και το $N$ αντιπροσωπεύει τον αριθμό των παρατηρήσεων. Στην άλλη κατάσταση η απόκριση και οι επεξηγηματικές μεταβλητές είναι χρονοσειρές, έτσι ώστε το μοντέλο παλινδρόμησης να περιλαμβάνει δεδομένα χρονοσειρών. Για παράδειγμα, η μεταβλητή απόκρισης μπορεί να είναι οι ωριαίες εκπομπές $CO_2$ από ένα χημικό φυτό και οι επεξηγηματικές μεταβλητές μπορεί να είναι ο ωριαίος ρυθμός παραγωγής, οι ωριαίες αλλαγές στην συγκέντρωση ενός ακατέργαστου υλικού εισροής και η θερμοκρασία του περιβάλλοντος μετράται κάθε ώρα. Όλα αυτά είναι στοιχεία δεδομένων των χρονοσειρών.

Το μοντέλο παλινδρόμησης για τα δεδομένα χρονοσειρών γράφεται ως

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t \,, \mu\varepsilon\ t = 1,2, \ldots, T \qquad (1.2)$$

Κατά τη σύγκριση των εξισώσεων (1.1) και (1.2) παρατηρούμε ότι έχουμε αλλάξει την παρατήρηση στο δείκτη από $i$ σε $t$ και τώρα η απόκριση και οι μεταβλητές πρόβλεψης είναι πλέον χρονοσειρές. Επίσης, έχουμε χρησιμοποιήσει T αντί του N που δηλώνει τον αριθμό των παρατηρήσεων και αποτελεί την πιο πρόσφατη ή την τελευταία διαθέσιμη παρατήρηση.

### *1.4.1 Ανίχνευση αυτοσυσχέτισης: Ο έλεγχος Durbin-Watson*

Οι γραφικές παραστάσεις των υπολοίπων μπορεί να είναι χρήσιμες για τον εντοπισμό της αυτοσυσχέτισης. Η πιο χρήσιμη γραφική παράσταση είναι αυτή των υπολοίπων ως προς το χρόνο. Αν υπάρχει θετική αυτοσυσχέτιση, τα υπόλοιπα με το ίδιο πρόσημο εμφανίζονται σε ομάδες. Από την άλλη πλευρά, αν υπάρχει αρνητική αυτοσυσχέτιση τα σημεία των υπολοίπων θα εναλλάσονται πάρα πολύ γρήγορα.

Διάφοροι στατιστικοί έλεγχοι μπορούν να χρησιμοποιηθούν για την ανίχνευση της αυτοσυσχέτισης. Ο έλεγχος που αναπτύχθηκε από τους Durbin και Watson (1950, 1951, 1971) είναι μια πολύ ευρέως χρησιμοποιημένη διαδικασία. Ο έλεγχος αυτός βασίζεται στην υπόθεση ότι τα σφάλματα στο μοντέλο παλινδρόμησης παράγονται από ένα πρώτης τάξης ή πρώτου βαθμού μοντέλου αυτοπαλινδρόμησης που παρατηρήθηκε σε ισαπέχουσες χρονικές περιόδους. Αυτό είναι:

$$\varepsilon_t = \varphi \varepsilon_{t-1} + a_t \qquad (1.3)$$

όπου $\varepsilon_t$ είναι το σφάλμα του μοντέλου στη χρονική περίοδο t, το $a_t$ είναι μια NID(0,$\sigma_\alpha^2$) τυχαία μεταβλητή (που ακολουθεί την κανονική κατανομή και οι τ.μ είναι ανεξάρτητες και ισόνομες) και $\varphi$ είναι μια παράμετρος που καθορίζει τη σχέση μεταξύ των διαδοχικών τιμών των σφαλμάτων $\varepsilon_t$ και $\varepsilon_{t-1}$ του μοντέλου. Απαιτείται το $|\varphi| < 1$, έτσι ώστε το σφάλμα του μοντέλου στο χρονικό διάστημα t να είναι ίσο με ένα κλάσμα του σφάλματος της αμέσως προηγούμενης περιόδου συν μια κανονικά και ανεξάρτητα κατανεμημένη τυχαία διαταραχή μοναδική για τη τρέχουσα περίοδο. Στα μοντέλα παλινδρόμησης χρονοσειρών η $\varphi$ μερικές φορές ονομάζεται παράμετρος αυτοσυσχέτισης. Έτσι, σε ένα απλό γραμμικό μοντέλο παλινδρόμησης τα πρώτης τάξης αυτοπαλινδρομικά σφάλματα θα είναι:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad \text{και} \quad \varepsilon_t = \varphi \varepsilon_{t-1} + a_t \qquad (1.4)$$

όπου $y_t$ και $x_t$ είναι η μεταβλητή απόκρισης και η μεταβλητή πρόβλεψης τη χρονική στιγμή t αντίστοιχα.

Όταν τα σφάλματα του μοντέλου παλινδρόμησης προκύπτουν από το πρώτη τάξης αυτοπαλινδρόμική διαδικασία της Εξίσωσης (1.3), υπάρχουν αρκετές ενδιαφέρουσες ιδιότητες αυτών των σφαλμάτων. Από τη διαδοχική αντικατάσταση για $\varepsilon_t, \varepsilon_{t-1}, \ldots$ στο δεξί μέλος της εξίσωσης (1.3) θα έχουμε:

$$\varepsilon_t = \sum_{j=0}^{\infty} \varphi^j \alpha_{t-j}$$

Με άλλα λόγια, ο όρος του σφάλματος στο μοντέλο παλινδρόμησης για τη χρονική περίοδο t είναι απλώς ένας γραμμικός συνδυασμός όλων των τωρινών και των προηγούμενων υλοποιήσεων των ανεξάρτητων και ισόνομων τυχαίων μεταβλητών $a_t$ του NID(0,$\sigma_\alpha^2$). Επιπλέον, μπορούμε να δείξουμε ότι:

$$E(\varepsilon_t) = 0$$

$$Var(\varepsilon_t) = \sigma^2 = \sigma_\alpha^2 \left(\frac{1}{1-\varphi^2}\right) \qquad (1.5)$$

$$Cov(\varepsilon_t, \varepsilon_{t\pm j}) = \varphi^j \sigma_\alpha^2 \left(\frac{1}{1-\varphi^2}\right)$$

Δηλαδή, τα σφάλματα έχουν μηδενική μέση τιμή και σταθερή διασπορά, αλλά έχουν μη μηδενική συνδιασπορά εκτός αν φ=0. Η αυτοσυσχέτιση μεταξύ δύο σφαλμάτων που απέχουν μια χρονική περίοδο ή η καθυστέρημένη αυτοσυσχέτιση είναι:

$$\rho_1 = \frac{Cov(\varepsilon_t, \varepsilon_{t+1})}{\sqrt{Var(\varepsilon_t)}\sqrt{Var(\varepsilon_t)}} = \frac{\varphi\sigma_\alpha^2\left(\frac{1}{1-\varphi^2}\right)}{\sqrt{\sigma_\alpha^2\left(\frac{1}{1-\varphi^2}\right)}\sqrt{\sigma_\alpha^2\left(\frac{1}{1-\varphi^2}\right)}} = \varphi$$

Η αυτοσυσχέτιση ανάμεσα σε δύο σφάλματα κατά τη κ χρονική περίοδο είναι:

$$\rho_\kappa = \varphi^\kappa, \;\; i = 1,2, \dots$$

Η παραπάνω ονομάζεται συνάρτηση αυτοσυσχέτισης (Brockwell και Davis 2002). Έχουμε απαιτήσει ότι το |φ|<1. Όταν η φ είναι θετική, όλοι οι όροι του σφάλματος έχουν θετική συσχετίση, αλλά το μέγεθος της συσχέτισης μειώνεται καθώς τα σφάλματα απομακρύνονται. Μόνο αν φ = 0 είναι τα σφάλματα του μοντέλου είναι ασυσχέτιστα. Τα περισσότερα προβλήματα στη παλινδρόμηση χρονοσειρών περιλαμβάνουν δεδομένα με θετική αυτοσυσχέτιση. Ο έλεγχος Durbin-Watson εξετάζει την παρουσία θετικής αυτοσυσχέτισης στο μοντέλο παλινδρόμησης των σφαλμάτων. Πιο συγκεκριμένα οι υποθέσεις που εξετάζονται από τον έλεγχο Durbin-Watson είναι:

$$H_0: \varphi = 0$$

$$H_1: \varphi > 0$$

Το στατιστικό του ελέγχου Durbin-Watson δίνεται από το τύπο:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{i=1}^{T} e_t^2} = \frac{\sum_{t=2}^{T} e_t^2 + \sum_{t=2}^{T} e_{t-1}^2 - 2\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{i=1}^{T} \varepsilon_t^2} \cong 2(1 - r_1)$$

όπου $e_t\,, t = 1,2, \dots, T$ είναι υπόλοιπα από την OLS παλινδρόμηση του $y_t$, για $x_t$. Το $r_1$ είναι η καθυστέρηση μιας αυτοσυσχέτισης μεταξύ των υπολοίπων, έτσι για ασυσχέτιστα σφάλματα η τιμή της ελεγχοσυνάρτησης Durbin-Watson πρέπει να είναι περίπου 2. Η ελεγχοσυνάρτηση είναι αναγκαία ώστε να προσδιοριστεί ακριβώς πόσο μακριά από το 2 το Durbin-Watson πρέπει να πέσει για να μπορέσουμε να

συμπεράνουμε αν η υπόθεση των ασυσχέτιστων σφαλμάτων παραβιάζεται. Στη συνέχεια θα παρουσιάσουμε μερικά κλασικά μοντέλα χρονοσειρών.

*1.4.2 Το μοντέλο πεπερασμένης τάξης κινητού μέσου (MA)*

Στα μοντέλα πεπερασμένης τάξης κινητού μέσου ή αλλιώς μοντέλα (MA), η συνθήκη του βάρους $\psi_0$ έχει οριστεί στο 1 και τα βάρη (ή σταθμά) που δεν έχουν οριστεί στο 0 συμβολίζονται από το ελληνικό γράμμα $\theta$ με ένα αρνητικό πρόσημο μπροστά. Ως εκ τούτου, ένα μοντέλο κινητού μέσου τάξης q (MA (q)) δίνεται ως:

$$y_t = \mu + \varepsilon_t - \theta_1\varepsilon_{t-1} - \ldots - \theta_q\varepsilon_{t-q} \qquad (1.6)$$

όπου $\{\varepsilon_t\}$ καλείται λευκός θόρυβος. Το μοντέλο MA(q) είναι πάντα στάσιμο και ανεξάρτητο από τις τιμές των βαρών (weights). Όσο αναφορά τον παράγοντα backward shift B, η διαδικασία MA(q) είναι:

$$y_t = \mu + \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)\varepsilon_t = \mu + \left(1 - \sum_{i=1}^{q} \theta_i B^i\right)\varepsilon_t = \mu + \theta(B)\varepsilon_t \qquad (1.7)$$

όπου $\theta(B) = 1 - \sum_{i=1}^{q} \theta_i B^i$

*1.4.3 Το πρώτης τάξης μοντέλο αυτοπαλινδρόμησης, AR(1)*

Θεωρούμε τη χρονοσειρά:

$$y_t = \mu + \sum_{i=1}^{\infty} \psi_i \varepsilon_{t-i} = \mu + \sum_{i=1}^{\infty} \psi_i B^i \varepsilon_t = \mu + \psi(B)\varepsilon_t \qquad (1.8)$$

όπου $\psi(B) = \sum_{i=1}^{\infty} \psi_i B^i$

Όπως και στα μοντέλα πεπερασμένης τάξης MA, μια προσέγγιση για την μοντελοποίηση αυτής της χρονοσειράς είναι να υποθέσουμε ότι οι διαταραχές ή οι τυχαίες διακυμάνσεις του παρελθόντος θα πρέπει να είναι σε μικρή συσχέτιση με τις πιο πρόσφατες διαταραχές που το μοντέλο έχει βιώσει. Δεδομένου ότι οι διαταραχές είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, μπορούμε να υποθέσουμε μια σειρά από πολλά μη πεπερασμένα βάρη (ή σταθμά) φθινουσών μεγέθων που αντανακλούν τα συνεχώς μειωμένα μεγέθη των συνεισφορών των διαταραχών του παρελθόντος. Μπορεί να δημιουργηθεί μια απλή αλλά και διαισθητική σειρά τέτοιων

βαρών που ακολουθεί ένα φθίνων εκθετικό σχήμα. Γι 'αυτό θα θέσουμε $\psi_i = \varphi^i$, όπου |φ|<1 για να εξασφαλίσουμε την εκθετική "φθορά". Με αυτά τα δεδομένα, τα βάρη στις διαταραχές που ξεκινούν από την τρέχουσα διαταραχή και πηγαίνουν πίσω στο παρελθόν θα είναι $1, \varphi, \varphi^2, \varphi^3, \ldots$ .Ως εκ τούτου, η εξίσωση (1.8) μπορεί να γραφτεί:

$$y_t = \mu + \varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots = \mu + \sum_{i=0}^{\infty} \varphi^i \varepsilon_{t-i} \qquad (1.9)$$

θα θέσουμε όπου $t \to t-1$ οπότε η εξίσωση (1.9) θα γίνει:

$$y_{t-1} = \mu + \varepsilon_{t-1} + \varphi\varepsilon_{t-2} + \varphi^2\varepsilon_{t-3} + \cdots \qquad (1.10)$$

Ο συνδυασμός των εξισώσεων (1.9) και (1.10) θα δώσει:

$$y_t = \mu + \varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots = \mu - \varphi\mu + \varphi y_{t-1} + \varepsilon_t$$

όπου $\varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots = \varphi y_{t-1} - \varphi\mu$ , $\delta = \mu - \varphi\mu$

άρα $$y_t = \delta + \varphi y_{t-1} + \varepsilon_t \qquad (1.11)$$

Η διαδικασία της εξίσωσης (1.11) καλείται πρώτης τάξης διαδικασία αυτοπαλινδρόμησης AR(1), επειδή η εξίσωση (1.11) μπορεί να θεωρηθεί ως παλινδρόμηση του $y_t$ σε σχέση με το $y_{t-1}$ εξού και ο όρος διαδικασία αυτοπαλινδρόμησης. Η υπόθεση |φ|< 1 κάνει τα βάρη να φθίνουν εκθετικά στο χρόνο και εξασφαλίζει ότι $\sum_{i=0}^{\infty}|\psi_i| < \infty$. Η διαδικασία AR(1) είναι στάσιμη αν |φ|<1. Ο μέσος μιας στάσιμης διαδικασίας AR(1) είναι:

$$E(y_t) = \mu = \frac{\delta}{1-\varphi} \qquad (1.12)$$

Η συνάρτηση αυτοσυνδιακύμανσης της στάσιμης διαδικασίας AR(1) μπορεί να υπολογισθεί με τη βοήθεια της εξίσωσης (1.9) ως:

$$\gamma(\kappa) = \sigma^2\varphi^\kappa \tfrac{1}{1-\varphi^2} \; \gamma\iota\alpha \; \kappa = 0,1,2,\ldots \qquad (1.13)$$

Η διασπορά τότε δίνεται από το τύπο:

$$\gamma(0) = \sigma^2 \frac{1}{1-\varphi^2}$$

Αντίστοιχα, η συνάρτηση αυτοσυσχέτισης για μια στάσιμη διαδικασία AR(1) δίνεται ως:

$$\rho(\kappa) = \frac{\gamma(\kappa)}{\gamma(0)} = \varphi^{\kappa} \qquad \text{για κ=0,1,2,…}$$

*1.4.4 Γενικό μοντέλο αυτοπαλινδρόμησης AR(p)*

Το AR μοντέλο p-τάξης δίνεται από το τύπο:

$$y_t = \delta + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t \qquad (1.14)$$

όπου $\varepsilon_t$ καλείται λευκός θόρυβος. Μια άλλη μορφή της (1.13) μπορεί να δοθεί ως:

$$\varphi(B) y_t = \delta + \varepsilon_t \qquad (1.15)$$

όπου $\varphi(B) = 1 - \varphi_1 B - \varphi_1 B^2 - \cdots - \varphi_p B^p$.

*1.4.5 Το ανάμεικτο μοντέλο αυτοπαλινδρόμησης και κινητού μέσου (MIXED-ARMA)*

Τα μοντέλα που παρουσιάσαμε στις προηγούμενες ενότητες αποτελούν ειδικές περιπτώσεις του θεωρήματος του Wold για την αποσύνθεση μιας στάσιμης χρονοσειράς που παριστάνεται από ένα σταθμισμένο άθροισμα μη πεπερασμένων τυχαίων διαταραχών (Bisgaard και Kulahci, 2005). Για παράδειγμα, στο μοντέλο AR(1), τα βάρη (ή σταθμά) στο μη πεπερασμένο άθροισμα υποχρεούνται να ακολουθήσουν μια φθίνουσα εκθετική μορφή όπου φ είναι ο ρυθμός με τον οποίο φθίνει. Δεδομένου ότι δεν υπάρχουν περιορισμοί, πέραν από $\sum_{i=0}^{\infty} |\psi_i| < \infty$ στα βάρη $(\psi_i)$ , ίσως να μην είναι δυνατό να τα προσεγγίσουμε μέσω ενός φθίνοντος εκθετικού σχήματος. Γι 'αυτό, θα πρέπει να αυξήσουμε τη τάξη του μοντέλου AR για την προσέγγιση κάθε μορφής, που αυτά τα βάρη μπορεί στην πραγματικότητα να παρουσιάζουν. Σε ορισμένες περιπτώσεις ωστόσο, είναι δυνατόν να γίνουν απλές προσαρμογές στη φθίνουσα εκθετική μορφή με την προσθήκη μόνο μερικών ορών και ως εκ τούτου να έχουν ένα πιο λιτό (parsimonious) μοντέλο. Ας υποθέσουμε, για παράδειγμα, ότι τα βάρη $\psi_i$ πράγματι εμφανίζουν μια φθίνουσα εκθετική μορφή με ένα σταθερό ρυθμό εκτός από το γεγονός ότι τα $\psi_i$ δεν θα είναι ίσα με αυτό το ρυθμό της φθοράς όπως θα ήταν στην περίπτωση του μοντέλου AR(1). Ως εκ τούτου, αντί να αυξηθεί η τάξη του μοντέλου AR για να εξομαλύνει αυτή την "ανωμαλία", εμείς μπορούμε να προσθέσουμε ένα MA(1) όρο που απλά θα προσαρμόσει το $\psi_1$, καθώς

δεν θα έχει καμία επίδραση στο ρυθμό του φθίνοντος εκθετικού σχήματος των υπόλοιπων βαρών. Αυτό οδηγεί σε ένα ανάμεικτο μοντέλο αυτοπαλινδρόμησης κινητού μέσου ή αλλιώς μοντέλο ARMA(1,1) . Ένα γενικό μοντέλο ARMA(p, q) δίνεται ως:

$$y_t = \delta + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}$$

$$= \delta + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \qquad (1.16)$$

ή

$$\varphi(B) y_t = \delta + \theta(B) \varepsilon_t \qquad (1.17)$$

όπου $\varepsilon_t$ είναι λευκός θόρυβος.

### 1.5 Απαριθμητά δεδομένα

Χρονοσειρές ακολουθώντας τα Γενικευμένα Γραμμικά Μοντέλα

Στην συνήθη γραμμική παλινδρόμηση το πρόβλημα είναι να συσχετιστεί η μέση απόκριση μιας μεταβλητής ενδιαφέροντος με ένα σύνολο επεξηγηματικών μεταβλητών με τη βοήθεια μιας γραμμικής εξίσωσης. Σε πολλές περιπτώσεις αυτό γίνεται με την προϋπόθεση ότι τα δεδομένα είναι της Κανονικής κατανομής είναι ανεξάρτητα μεταξύ τους. Υπάρχουν περιπτώσεις όμως, όσον αφορά μη φυσιολογικές παρατηρήσεις, όπως δυαδικό (binary) και απαριθμητά (counts) δεδομένα που επιλύονται πολύ επιτυχώς από τα γενικευμένα γραμμικά μοντέλα. Στη περίπτωση που οι παρατηρήσεις αυτές είναι χρονικά μεταβαλλόμενες δεν μπορούμε να αξιοποιήσουμε τα μοντέλα που παρουσιάσαμε στη παράγραφο 1.4. Επομένως θα εισάγουμε τις ιδέες των γενικευμένων γραμμικών μοντέλων στα δεδομένα για την μοντελοποίηση χρονοσειρών όπως οι Kedem και Fokianos (2002). Το ερώτημα τότε είναι πώς να επεκτείνουμε τη μεθοδολογία των γενικευμένων γραμμικών μοντέλων στις χρονοσειρές όπου τα δεδομένα είναι εξαρτημένα όπως και οι συμμεταβλητές και ίσως ακόμη και τα βοηθητικά στοιχεία που εξαρτώνται από το χρόνο που είναι επίσης τυχαία. Όπως θα δούμε, με τη χρήση της μερικής πιθανοφάνειας θα μπορούμε να επεκτείνουμε αρκετά εύκολα τα κύρια χαρακτηριστικά κατάλληλα για ανεξάρτητα δεδομένα σε χρονοσειρές με βάση τα γενικευμένα γραμμικά μοντέλα. Ένα βασικό συστατικό αυτού είναι ότι η μερική πιθανοφάνεια επιτρέπει την χρονική ή διαδοχική

υπό συνθήκη συμπερασματολογία χρησιμοποιώντας όλα όσα είναι γνωστά στον παρατηρητή κατά τη στιγμή της παρατήρησης. Αυτό επιτρέπει μια πολύ ευέλικτη υπό συνθήκη συμπερασματολογία που μπορεί εύκολα να λάβει υπόψη συστατικά αυτοπαλινδρόμησης και με συναρτήσεις συμμεταβλητών του παρελθόντος καθώς και όλα τα είδη των αλληλεπιδράσεων μεταξύ των συμμεταβλητών. Στο Κεφάλαιο 3 θα παρουσιαστεί το αναγκαίο υπόβαθρο και μια επισκόπηση των γενικευμένων γραμμικών μοντέλων έχοντας κατά νου εξαρτημένα δεδομένα.

Συγκεκριμένα, α) ορίζουμε τι εννοούμε με τον όρο χρονοσειρές γενικευμένων γραμμικών μοντέλων, θα εισαχθεί η έννοια της μερικής πιθανοφάνειας, και β) οι στατιστικές ιδιότητες, συμπεριλαμβανομένα ασυμπτωτικά αποτελέσματα της εκτιμήτριας μέγιστης μερικής πιθανοφάνειας (βλ. και Karioti και Caroni, 2006).

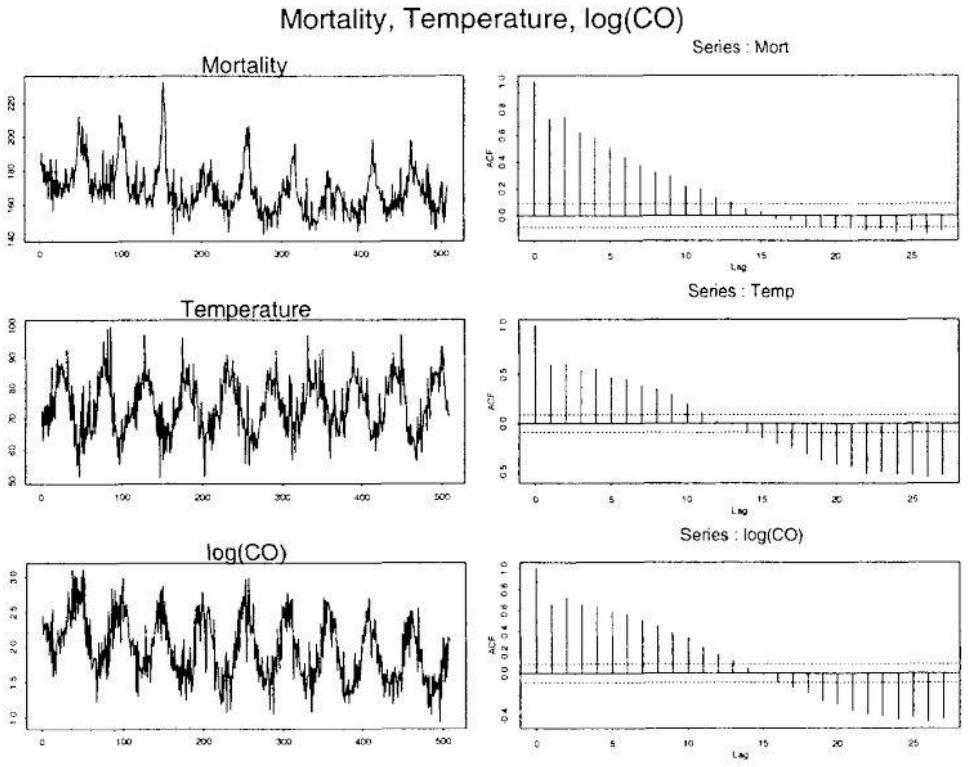
Οι απαριθμητές χρονοσειρές εμφανίζονται σε ποίκιλες περιοχές όπου μια σειρά γεγονότων ανά χρονική περίοδο παρατηρείται στο πέρασμα του χρόνου. Παραδείγματα που δείχνουν το εύρος των εφαρμογών τους είναι ο αριθμός των καθημερινών εισαγωγών στα νοσοκομεία, από τη δημόσια υγεία, ο αριθμός των συναλλαγών στο χρηματιστήριο ανά λεπτό, από την οικονομία ή το πλήθος των ελαττωματικών προϊόντων ανά ώρα, από το βιομηχανικό έλεγχο ποιότητας. Τα μοντέλα των απαριθμητών χρονοσειρών πρέπει να λαμβάνουν υπόψη ότι οι παρατηρήσεις είναι μη αρνητικοί ακέραιοι αριθμοί και να αποδίδουν κατάλληλα την εξάρτηση μεταξύ των παρατηρήσεων.

**Παράδειγμα Απαριθμητών δεδομένων**

Ανάλυση θνησιμότητας απαριθμητών δεδομένων

Το σύνολο των δεδομένων αναφέρεται στο σύνολο της θνησιμότητας και των συναφών συμμεταβλητών στη περιοχή του Los Angeles κατά τη διάρκεια μιας περιόδου 10 ετών από την 1η Ιανουαρίου 1970 έως την 31η Δεκεμβρίου 1979. Τα δεδομένα αποτελούνται από τρείς σειρές θνησιμότητας, τη συνολική θνησιμότητα (Y), δύο καιρικές σειρές τη θερμοκρασία (T) και τη σχετική υγρασία (RH), καθώς και έξι σειρές ρύπανσης όπως το μονοξείδιο του άνθρακα (CO), διοξείδιο του θείου ($SO_2$), διοξείδιο του αζώτου ($NO_2$), υδρογονάνθρακες (HC) , όζον (OZ) και τα σωματίδια (KM). Εδώ θα χρησιμοποιήσουμε την κανονική συνάρτηση σύνδεσης

του μοντέλου της Poisson παλινδρόμηση $\log(\mu_t) = n_t = Z'_{t-1}\beta$ ή $\mu_t(\beta) = \exp(Z'_{t-1}\beta)$ και έχει κάπως διαφορετικό χαρακτήρα. Οι τεχνικές αυτές θα παρουσιαστούν στα κεφάλαια 3,4, λαμβάνοντας υπόψη τις εκτιμήσεις των Shumway και Stoffer (2000) χρησιμοποιούμε ένα γραμμικό φίλτρο της μορφής $\sum_j \alpha_j x_{t-j}$. Οπότε τα δεδομένα εξομάλυνσης αποτελούνται από περιοδικές συνιστώσες με περιόδους μαγαλύτερες από 10 μέρες. Οι εξομαλυσμένες σειρές έγιναν δειγματοληπτικά εβδομαδιαία, έχοντας ως αποτέλεσμα σε λιγότερο από 11 σειρές η κάθε μια να έχει N=508 παρατηρήσεις. Στο Σχήμα 1.12 απεικονίζονται δειγματοληπτικά εβδομαδιαία τα δεδομένα θνησιμότητας, θερμοκρασίας και ο λογάριθμος του μονοξείδιου του άνθρακα. Αυτές οι τρείς σειρές εμφανίζουν μια παρόμοια μορφή ταλάντωσης γεγονός που επαληθεύεται από τις εκτιμημένες αυτοσυσχετίσεις. Για απλούστευση, θα μεταφέρουμε αυτές τις εξομαλυσμένες εβδομαδιαίες σειρές στις μεταβλητές. Οπότε θα είναι η θνησιμότητα $(Y_t)$, η θερμοκρασία $(T_t)$ και ο λογάριθμος του μονοξείδιου του άνθρακα $(log(CO_t))$.




**Σχήμα 1.12**: Τα εβδομαδιαία δεδομένα της φιλτραρισμένης συνολικής θνησιμότητας, θερμοκρασίας και ο λογάριθμος του μονοξείδιου του άνθρακα (CO), και οι αντίστοιχες εκτιμώμενες συναρτήσεις αυτοσυσχέτισης. N = 508.

Το γεγονός ότι τα απαριθμητά δεδομένα για πρώτη φορά εξομαλύνονται δεν αποκλείει την εφαρμογή της Poisson παλινδρόμησης με την αντίστοιχη κανονική συνάρτηση σύνδεσης. Αυτή η μη συνηθισμένη χρήση του μοντέλου της Poisson μπορούμε να την αντιληφθούμε καλύτερα αν σκεφτούμε το πλαίσιο της εκτιμημένης συνάρτησης. Στο παρόν παράδειγμα, το μοντέλο Poisson παρέχει μια χρήσιμη εκτιμημένη συνάρτηση που η ποιότητα προσαρμογής φαίνεται στο Σχήμα 1.13. Ένας άλλος τρόπος για να δικαιολογήσουμε τη διαδικασία είναι να σκεφτείτε τα εξομαλυμένα δεδομένα κατά προσέγγιση ακέραιων δεδομένων.

Το πρώτο μοντέλο ($M_1$) καταγράφεται για λόγους σύγκρισης, ο Πίνακας 1.3 παρέχει στατιστικά στοιχεία διάγνωσης γι αυτά τα μοντέλα. Επιπλέον στον πίνακα περιλαμβάνονται, το άθροισμα των τετραγώνων των working residuals $\left[ w\hat{r}_t = \frac{Y_t - \hat{\mu}_t}{\partial \mu_t / \partial n_t} \right]$ που αξιολογούνται από το β, το μέσο τετραγωνικό σφάλμα της απόκρισης ή των υπολοίπων $Y_t - \hat{\mu}_t$, το άθροισμα των τετραγώνων των υπολοίπων Pearson $x^2$, τη συνάρτηση deviance με N-p βαθμούς ελευθερίας, το κριτήριο $AIC = D + 2p$ και $BIC = D + p * log(N)$ που υπολογίζονται μέσω της συνάρτησης deviance.

| |
|---|
| $M_1: T_t + RH_t + CO_t + S_t + N_t + HC_t + OZ_t + KM_t$ |
| $M_2: Y_{t-1}$ |
| $M_3: Y_{t-1} + Y_{t-2}$ |
| $M_4: Y_{t-1} + Y_{t-2} + T_{t-1}$ |
| $M_5: Y_{t-1} + Y_{t-2} + T_{t-1} + \log(CO_t)$ |
| $M_6: Y_{t-1} + Y_{t-2} + T_{t-1} + T_{t-2} + \log(CO_t)$ |
| $M_7: Y_{t-1} + Y_{t-2} + T_t + T_{t-1} + \log(CO_t)$ |

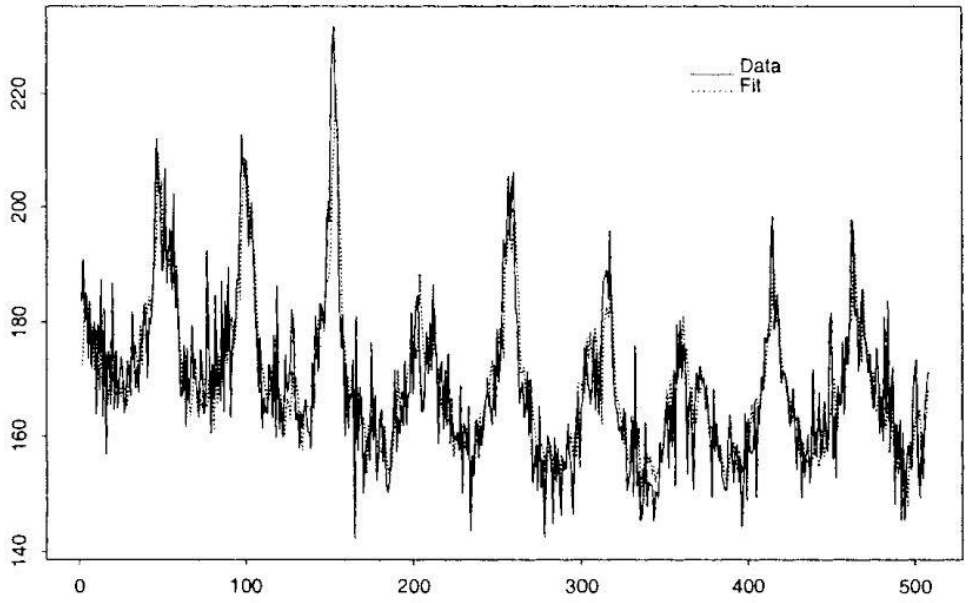
**Πίνακας 1.3**: Τοποθετούμε όπου $N = NO_2$ και $S = SO_2$. Στο τρίτο μοντέλο για παράδειγμα θα έχουμε

$$n_t = \beta_o + \beta_1 Y_{t-1} + \beta_2 Y_{t-2}$$

Με βάση τα αποτελέσματα του Πίνακα 1.4 κρίνεται το μοντέλο 5 ως καταλληλότερο εφόσον έχει το μικρότερο AIC, BIC καθώς και το μικρότερο μέσο τετραγωνικό σφάλμα MSE απ' τα υπόλοιπα (Kedem και Fokianos 2002).

| Μοντελο | p (αριθμός | MSE | D | df | AIC | BIC |
|---|---|---|---|---|---|---|
| 1 | 9 | 108.99 | 315.69 | 499 | 333.69 | 371.76 |
| 2 | 2 | 93.02 | 276.07 | 506 | 280.07 | 288.53 |
| 3 | 3 | 75.52 | 222.23 | 505 | 228.23 | 240.92 |
| 4 | 4 | 69.04 | 203.52 | 504 | 211.52 | 228.44 |
| 5 | 5 | 59.38 | 174.55 | 503 | 184.55 | 205.71 |
| 6 | 6 | 59.38 | 174.53 | 502 | 186.53 | 211.91 |
| 7 | 6 | 58.44 | 171.41 | 502 | 183.41 | 208.79 |

**Πίνακας 1.4**




**Σχήμα 1.13**: Παρατήρηση και πρόβλεψη της εβδομαδιαίας θνησιμότητας από το μοντέλο ($M_5$)

# ΚΕΦΑΛΑΙΟ ΙΙ

# ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

## 2.1 Εκθετική Οικογένεια Κατανομών

Θεωρούμε μια τυχαία απλή μεταβλητή Y της οποίας η συνάρτηση κατανομής εξαρτάται από την απλή παράμετρο θ. Η κατανομή ανήκει στην εκθετική οικογένεια αν μπορεί να γραφτεί στη μορφή

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

όπου α,b,s και t άγνωστες συναρτήσεις. Αξίζει να σημειωθεί η συμμετρία ανάμεσα σε y και θ, άρα η εξίσωση μπορεί να ξαναγραφτεί ως:

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

όπου $s(y) = \exp[d(y)]\,, t(\theta) = \exp[c(\theta)]$

Αν $a(y) = y$ ,τότε λέμε ότι η κατανομή έχει κανονική μορφή και η $b(\theta)$ λέγεται φυσική παράμετρος της κατανομής. Πολλές γνωστές κατανομές ανήκουν στην εκθετική οικογένεια κατανομών όπως η Κανονική, η Διωνυμική, η Poisson καθώς και άλλες που μπορούν να γραφτούν στη παραπάνω κανονική μορφή. Εδώ θα εστιάσουμε το ενδιαφέρον μας στο μοντέλο της Poisson παλινδρόμησης (McCullagh και Nelder, 1989).

Το στήριγμα είναι $S = \{ y \in \mathbb{R} : f(y;\theta) > 0 \}$ το οποίο είναι ανεξάρτητο της παραμέτρου θ και οι συναρτήσεις a(.), b(.), c(.) είναι γνωστές.

Το γενικευμένο γραμμικό μοντέλο ορίζεται με βάση ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών $Y_1, \ldots, Y_N$ με κάθε κατανομή από τη γενική εκθετική οικογένεια κατανομών (Ε.Ο.Κ) και με τις ακόλουθες ιδιότητες:

1. Η κατανομή του κάθε $Y_i$ έχει κανονική μορφή $(\alpha_i(\, y_i\,) = y_i\,)$ και εξαρτάται από μια μόνο παράμετρο $\theta_i$

$$f(y_i\,;\theta_i) = \exp[\; y_i\, b_i(\theta_i) + \; c_i(\theta_i) + d_i(\, y_i\,)\,]$$

2. Η κατανομή όλων των $Y_i$ είναι ίδια, έτσι ώστε οι δείκτες b,c,d να μην χρειάζονται.

Έτσι λοιπόν έχουμε την συνάρτηση πυκνότητας πιθανότητας των $Y_i$ :

$$f(y_1, \dots, y_n;\ \theta_1, \dots \dots, \theta_n) = \prod_{i=1}^{n} \exp[\ y_i\, b_i(\theta_i) + \ c_i(\theta_i) + d_i\ (y_i)\ ]$$

$$= exp[\sum_{i=1}^{n} y_i\, b_i(\theta_i)\ + \sum_{i=1}^{n} c_i(\theta_i)\ + \sum_{i=1}^{n} d_i\ (y_i)\ ]$$

Οι παράμετροι $\theta_i$ δεν είναι άμεσου ενδιαφέροντος. Εμείς συνήθως ενδιαφερόμαστε για ένα μικρότερο σύνολο παραμέτρων $\beta_1,\dots,\beta_p$ (με p<n). Υποθέτουμε ότι $E(Y_i) = \mu_i$, που $\mu_i$ είναι συναρτήσει των $\theta_i$. Για τα γενικευμένα γραμμικά μοντέλα υπάρχει μια μετατροπή των $\mu_i$ :

$$g(\mu_i) = x_i^T \beta$$

**Παρατηρήσεις**

1. Η συνάρτηση $g$ είναι μονότονη, διαφορίσιμη και καλείται συνάρτηση σύνδεσης. Η σχέση της αναμενόμενης τιμής $\mu_i$ και της $x_i^T \beta$ ‹‹γραμμικοποιείται›› μέσω της 1-1 συνάρτησης $g(.)$. Είναι επίπεδη αυξάνοντας ή μειώνοντας με $\mu_i$ , αλλά δεν μπορεί να αυξηθεί για μερικές τιμές των $\mu_i$ και να μειωθεί για άλλες τιμές.
2. Το διάνυσμα $x_i$ είναι ένα p x 1 διάνυσμα επεξηγηματικών μεταβλητών

$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}\ , x_i^T = [x_{i1} \quad \dots \quad x_{ip}]$$

3. Το β είναι ένα διάνυσμα παραμέτρων p x 1

$$\text{Με}\ \ \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Το διάνυσμα $x_i^T$ είναι η i-γραμμή του πίνακα σχεδιασμού X. Τα γενικευμένα γραμμικά μοντέλα έχουν 3 συνιστώσες:

I. Οι μεταβλητές απόκρισης y1,. . . , YN, οι οποίες υποτίθεται ότι μοιράζονται την ίδια κατανομή από την εκθετική οικογένεια.

II. Ένα σύνολο $\beta$ παραμέτρων και επεξηγηματικών μεταβλητών

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

III. Μια μονότονη συνάρτηση $g$ τέτοια ώστε:
$g(\mu_i) = x_i^T \beta$ με $\mu_i = E(Y_i)$

## 2.2 Παλινδρόμηση Poisson

### *2.2.1 Μοντέλο*

Θεωρούμε το μοντέλο $y \sim Po(\mu)$ όπου η εξάρτηση με τις ανεξάρτητες μεταβλητές για μια στατιστική μονάδα εκφράζεται μέσω ενός κατάλληλου μετασχηματισμού $g(.)$ της αναμενόμενης τιμής $y$, οπότε ισχύει η σχέση της μορφής

$$g(\mu_x) = x'\beta$$

Η συνάρτηση $g(\mu_x)$ λέγεται συνάρτηση σύνδεσης. Επίσης θέλουμε να εξασφαλίσουμε το περιορισμό ότι μ>0 άρα η $g(\mu_x)$ δεν μπορεί να είναι της μορφής:

$$g(\mu_x) = \mu_x$$

αφού $\mu_x = x'\beta$ και έτσι δεν τηρείται ο περιορισμός μ>0,άρα εφόσον θέλουμε η $\mu_x$ να είναι μια μη αρνητική συνάρτηση θα την γράψουμε στη παρακάτω μορφή:

$$\mu_i = \mu_{x_i} = n_i e^{x_i'\beta}.$$

Η προσαρμογή του μοντέλου θα γίνει με τη μέθοδο μέγιστης πιθανοφάνειας με συνάρτηση πιθανοφάνειας $L$ δίνεται από τη σχέση:

$$L = \prod_{i=1}^{N} \frac{e^{-\mu_i} {\mu_i}^{y_i}}{y_i!}$$

Η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας δίνεται από τη σχέση:

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

οπότε δεδομένου $\mu_i = n_i e^{x_i'\beta}$ ,θα έχουμε

$$l = \sum_{i=1}^{n} \left[-n_i e^{x_i'\beta} + y_i(ln(n_i) + x_i'\beta) - \ln(y_i!)\right]$$

άρα

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left[-n_i x_{ij} e^{x_i'\beta} + y_i x_{ij}\right].$$

Οι εκτιμήτριες μέγιστης πιθανοφάνειας των $\beta_j$ θα βρεθούν λύνοντας τις εξισώσεις:

$$\sum_{i=1}^{n} \left[-n_i x_{ij} e^{x_i'\widehat{\beta}} + y_i x_{\iota j}\right] = \sum_{i=1}^{n} \left[x_{ij}(y_i - n_i e^{x_i'\widehat{\beta}})\right] = 0, j = 0,1, \ldots \ldots, p$$

αλλιώς, $X(y - \hat{\mu}) = 0$.

Η συνάρτηση σύνδεσης είναι (Dobson και Barnett, 2008): $g_{\widehat{\mu}_i} = \log(\hat{\mu}_i) = n_i + x_i'\hat{\beta}$ και $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots \ldots, \hat{\mu}_n)$.

*2.2.2 Κατανομή Poisson*

Μια εξαρτημένη μεταβλητή Y δεν είναι απαραίτητο να είναι μόνο συνεχής αλλά και απαριθμητή. Μια κατάλληλη κατανομή για την περιγραφή του αριθμού εμφάνισης ενός γεγονότος στο χρόνο ή στο χώρο συχνά θεωρείται η κατανομή Poisson με συνάρτηση πιθανότητας

$$f(y) = \frac{e^{-\theta}\,\theta^{y}}{\mathrm{y!}}, \theta > 0, y = 0,1,2, \ldots \ldots$$

όμως μπορεί να ξαναγραφτεί

$$f(y, \theta) = exp(ylog\theta - \theta - logy!).$$

Η παραπάνω μορφή θεωρείται κανονική επειδή $a(y) = y$ και με φυσική παράμετρο $b(\theta) = log\theta$.

Η κατανομή Poisson συμβολίζεται Y ~ Po(θ) και χρησιμοποιείται συχνά για τη μοντελοποίηση απαριθμητών δεδομένων (counts data). Συνήθως αυτό είναι ο αριθμός εμφανίσεως κάποιου γεγονότος στο χρόνο ή στο χώρο, όταν δηλαδή η πιθανότητα ενός γεγονότος που συνέβη σ' ένα μικρό χρονικό διάστημα είναι μικρή και τα γεγονότα συμβαίνουν ανεξάρτητα. Αν μια τυχαία μεταβλητή ακολουθεί τη κατανομή Poisson τότε η αναμενόμενη τιμή και η διακύμανση είναι ίσες (Οικονόμου και Καρώνη, 2010).

*2.2.3 Ελεγχοσυνάρτηση Deviance*

Γνωρίζουμε τη σχέση $\mu_i = n_i e^{x_{\iota}'\beta}$ και όλα τα $\mu_i$ πρέπει να ικανοποιούν το περιορισμό $\mu_i > 0$ και δεν εξαρτώνται το ένα απ' το άλλο οπότε θα εκτιμούνται από τα $\tilde{\mu}_i = y_i$ οπότε η τιμή της λογαριθμοποιημένης πιθανοφάνειας ισούται με

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

και η αντίστοιχη τιμή για το μοντέλο θα είναι**:**

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

$$\hat{\mu}_i = \hat{y}_i = n_i e^{x_{\iota}'\beta}.$$

Η ελεγχοσυνάρτηση deviance του μοντέλου ορίζεται ως

$$D(\hat{\beta}) = -2\{l(\hat{\beta}) - \tilde{l}\} = 2\left\{\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \sum_{i=1}^{n} (y_i - \hat{\mu}_i)\right\}$$

όμως $\sum_{i=1}^{n} (y_i - \hat{\mu}_i) = 0$ οπότε έχουμε

$$D(\hat{\beta}) = 2\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right).$$

Η ελεγχοσυνάρτηση deviance μετρά την απώλεια προσαρμογής του μοντέλου όταν βάζουμε όπου $\mu_i = n_i e^{x_i'\hat{\beta}}$. Επίσης την χρησιμοποιούμε και για να συγκρίνουμε δύο μοντέλα. Έστω ότι έχουμε τις εκτιμήσεις $\hat{\beta}, \hat{\beta}'$ των μοντέλων Μ,Μ' αντίστοιχα, με Μ' να είναι εμφωλευμένο στο Μ δηλαδή προκύπτει απ το Μ με κάποιους περιορισμούς. Έχουμε λοιπόν τους εξής ελέγχους υποθέσεων:

$$H_0: \text{ισχύει το } M'$$
$$H_1: \text{ισχύει το } M$$

ο λόγος μεγιστοποιημένων πιθανοφανειών θα δώσει

$$-2\{l(\hat{\beta}') - l(\hat{\beta})\} \sim X_q^2$$

q: πλήθος περιορισμών

και με τη βοήθεια της ελεγχοσυνάρτησης deviance η παραπάνω σχέση μπορεί να γραφτεί ως:

$$D(\hat{\beta}') - D(\hat{\beta}) \sim X_q^2$$

### 2.3 Προσαρμογή μοντέλου

Η συνάρτηση πιθανοφάνειας στην γενική περίπτωση της εκθετικής οικογένειας κατανομών γράφεται ως

$$L(\beta) = L(\theta_i, \varphi) = \prod_{i=1}^{N} f(y_i; \theta_i, \varphi) = \prod_{i=1}^{N} exp\left\{\frac{y_i\,\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi)\right\},$$

ενώ η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας θα είναι

$$l = lnL(\beta) = \sum_{i=1}^{n}\left\{\frac{y_i\,\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi)\right\} = \sum_{i=1}^{n} l_i$$

όπου $\quad l_i = \frac{y_i\,\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi).$

Η εκτίμηση των παραμέτρων $\beta$ του μοντέλου $g(\mu_i) = g\big(E(y_i)\big) = x'_i\beta = n_i$ με τη μέθοδο μέγιστης πιθανοφάνειας γίνεται μέσω της επίλυσης των εξισώσεων

$$\frac{\theta l}{\theta \beta_j} = 0, j = 0{,}1 \dots \dots, k.$$

Εφόσον για την εκθετική οικογένεια ισχύουν οι σχέσεις

$$E(y_i) = \mu_i = b'(\theta_i) \text{ και}$$

$$V(y_i) = a(\varphi)b''(\theta_i) = a(\varphi)v(\mu_i)$$

τότε μπορούμε σύμφωνα με το κανόνα αλληλουχίας ή κανόνα αλυσίδας να δείξουμε ότι:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial n_i} \times \frac{\partial n_i}{\partial \beta_j} = \frac{y_i - \mu_i}{V(y_i)} \frac{1}{g'(\mu_i)} x_{ij} \ , i = 1{,}2{,}..{,}n$$

δηλαδή έχουμε

$$u_j = \frac{\theta l}{\theta \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(y_i)} \frac{1}{g'(\mu_i)} x_{ij} \ , j = 0{,}1{,} \dots {,} k.$$

Η οποία καλείται συνάρτηση **score.** Αν θέσουμε όπου $u_j = 0$ ,θα έχουμε ένα σύστημα $k + 1$ εξισώσεων, μια εξίσωση δηλαδή για κάθε παράμετρο του μοντέλου. Αυτές οι εξισώσεις καλούνται εξισώσεις score της μέγιστης πιθανοφάνειας. Στη συνέχεια εκτιμούνται μέσω στης επαναληπτικής μεθόδου Newton-Raphson εκτιμούνται τα $\beta_j$

$$\beta_r = \beta_{r-1} - G_{r-1}^{-1} u_{r-1}$$

όπου $\beta_r$ η νέα εκτίμηση στο r-οστό βήμα της διαδικασίας η οποία υπολογίζεται ως συνάρτηση των τιμών των $\beta$ του προηγούμενου r-1 βήματος, των τιμών των συναρτήσεων score $u' = (u_0, u_1, \dots \dots, u_k)$ του r-1 βήματος καθώς και το πίνακα $G$ των δεύτερων μερικών παραγώγων με jm-οστό στοιχείο $\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_m}\right]_{\beta = \beta_{r-1}}$ και p=k+1.

Στη παραπάνω μέθοδο scoring του *Fisher* μπορούμε να αντικαταστήσουμε το πίνακα $-G$ με το πίνακα πληροφορίας $I(\beta)$. (Οικονόμου και Καρώνη, 2010)

Οπότε θα είναι

$$I_{jm} = E(u_j u_m) = E\left(\sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j}\frac{\partial l_i}{\partial \beta_m}\right) = \sum_{i=1}^{n} \frac{x_{ij}x_{im}}{V(y_i)(g'(\mu_i))^2} = [X'WX]_{jm}\,,$$

$$m = 0,1,2,\dots,k$$

$$W = diag(w_{ii}) \quad και \quad w_{ii} = \frac{1}{V(y_i)(g'(\mu_i))^2}\,, i = 1,2,\dots\dots,n.$$

Ο πίνακας πληροφορίας $I(\beta)$ παίζει σπουδαίο ρόλο στην εκτίμηση του πίνακα διασποράς των εκτιμημένων $\hat{\beta}$. Η εκτιμήτρια $\hat{\beta}$ ακολουθεί ασυμπτωτικά την κανονική κατανομή με αναμενόμενη τιμή β και ο εκτιμηθέν πίνακας διασποράς είναι ο αντίστροφος πίνακας της παρατηρούμενης πληροφορίας

$$\hat{V}(\hat{\beta}) = I^{-1}(\hat{\beta}) = (X'\widehat{W}X)^{-1}.$$

### 2.4 Έλεγχος Wald των συντελεστών β

Η εκτιμήτρια της μέγιστης πιθανοφάνειας $\hat{\beta}$ ακολουθεί ασυμπτωτικά την πολυμεταβλητή Κανονική κατανομή, δηλαδή $\hat{\beta} \sim N_p(\beta, \hat{V}(\hat{\beta}))$, όπου $\hat{V}(\hat{\beta}) = I^{-1}(\hat{\beta})$ η εκτιμήτρια του πίνακα διασποράς της ασυμπτωτικής κατανομής της $\hat{\beta}$ και $I(\hat{\beta})$ ο παρατηρούμενος πίνακας πληροφορίας. Επομένως έχουμε

$$Z = \frac{\widehat{\beta}_j - \beta_j}{(I^{-1}(\widehat{\beta})_{jj})^{1/2}} \sim N(0,1) \;\text{ ,ασυμπτωτικά,}$$

Με $I^{-1}(\hat{\beta})_{jj}$ το j-οστό διαγώνιο στοιχείο του παρατηρούμενου πίνακα πληροφορίας $I(\hat{\beta})$, με $(I^{-1}(\hat{\beta})_{jj})^{1/2}$ το τυπικό σφάλμα $se(\hat{\beta}_j)$.

Η παραπάνω στατιστική συνάρτηση κάνει τον έλεγχο Wald με:

$$H_0: \beta_j = 0, \text{ με εναλλακτική}$$

$$H_1: \beta_j \neq 0$$

και έχουμε και το 100(1-α)% διάστημα εμπιστοσύνης για τη παράμετρο $\beta_j$ που δίνεται από το παρακάτω τύπο:

$$\hat{\beta}_j \pm \mathrm{Z}_{\alpha/2} se(\hat{\beta}_j)$$

## 2.5 Διαγνωστικές μέθοδοι

Οι διαγνωστικές μέθοδοι βασίζονται στα υπόλοιπα, τα οποία χρησιμοποιούνται ως μέτρα συμφωνίας μεταξύ των παρατηρήσεων $y_i$ και των αντίστοιχων προσαρμοσμένων τιμών $\hat{\mu}_i$ ή $\hat{y}_i$ και είναι απαραίτητη η εξέταση τους. Χάρις σε αυτή την εξέταση μπορούμε να πάρουμε αρκετές πληροφορίες για την καταλληλότητα του μοντέλου που δε φαίνονται στον έλεγχο με τη deviance. (Οικονόμου και Καρώνη, 2010)

Παρακάτω αναφέρουμε τα είδη υπολοίπων τα οποία θα χρησιμοποιήσουμε:

- Υπόλοιπα Pearson
- Υπόλοιπα deviance
- Υπόλοιπα πιθανοφάνειας

### *2.5.1 Υπόλοιπα*

1) Υπόλοιπα Pearson

Η γενική τους μορφή είναι:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}, i = 1, \dots, n$$

όπου $v(\hat{\mu}_i)$ είναι η συνάρτηση διασποράς. Το άθροισμα των τετραγώνων των παραπάνω υπολοίπων ισούται με την ελεγχοσυνάρτηση του Pearson. Τα υπόλοιπα Pearson εφόσον διαιρούνται με το $\sqrt{v(\hat{\mu}_i)}$ παρατηρούμε οτί η διασπορά τους είναι μη σταθερή. Άρα θα το διαιρέσουμε με το τυπικό σφάλμα $se(y_i - \hat{\mu}_i) = \sqrt{v(\hat{\mu}_i)(1 - \hat{h}_{ii})}$ , με $\hat{h}_{ii}$ το i-οστό διαγώνιο στοιχείο του πίνακα

$$\hat{H} = \hat{W}^{\frac{1}{2}} X (X' \hat{W} X)^{-1} X' \hat{W}^{\frac{1}{2}}$$

όπου $\hat{W}$ σταθμισμένος πίνακας του προσαρμοσμένου μοντέλου.

Τα τυποποιημένα υπόλοιπα Pearson ως:

$$r_i^{PS} = \frac{r_i^P}{\sqrt{1 - \hat{h}_{ii}}},$$

2) Υπόλοιπα deviance

Πιο συχνά χρησιμοποιούνται τα υπόλοιπα deviance τα οποία είχαμε ορίσει:

$$r_i^D = sgn(y_i - \hat{\mu}_i)\{d_i(\hat{\beta})\}^{\frac{1}{2}}$$

Τα τυποποιημένα υπόλοιπα deviance ορίζονται ως:

$$r_i^{DS} = \frac{r_i^D}{\sqrt{1 - \hat{h}_{ii}}}$$

3) Υπόλοιπα της πιθανοφάνειας (likelihood residuals)

Ορίζονται ως

$$r_i^L = sgn(y_i - \hat{\mu}_i)\sqrt{\hat{h}_{ii}(r_i^{PS})^2 + (1 - \hat{h}_{ii})(r_i^{DS})^2}, i = 1, .., n.$$

Το $r_i^L$ καλείται υπόλοιπο πιθανοφάνειας επειδή μπορεί να ερμηνευτεί ως η τιμή της ελεγχοσυνάρτησης του λόγου πιθανοφανειών για την εισαγωγή στο μοντέλο μιας επιπλέον παραμέτρου που αντιστοιχεί σε μια πιθανή άτυπη παρατήρηση i, ώστε το μοντέλο να έχει τέλεια προσαρμογή σε αυτό το i-οστό σημείο δηλαδή να είναι $y_i - \hat{\mu}_i$.

Χρησιμότητα των υπολοίπων

Τα υπόλοιπα χρησιμοποιούνται συνήθως για να ελέγξουμε την καταλληλότητα του μοντέλου μέσω γραφικών παραστάσεων. Η παρουσία ασυνήθιστα μεγάλων υπολοίπων υποδεικνύει ότι το μοντέλο δεν είναι ικανοποιητικό. Αν οι παρατηρήσεις δίνονται σε χρονική σειρά το γράφημα θα δείξει συσχέτιση μεταξύ των υπολοίπων. Οι γραφικές παραστάσεις των υπολοίπων έναντι κάθε μεταβλητής ή έναντι του εκτιμημένου γραμμικού συνδυασμού $\hat{n}_i = x'_i\hat{\beta}$ (linear predictor) είναι πολύ χρήσιμες για την εξέταση. Κάποιο συστηματικό σχήμα δείχνει πιθανή ύπαρξη προβλήματος στο μοντέλο. Επίσης όλες αυτές οι γραφικές παραστάσεις χρησιμεύουν στον εντοπισμό άτυπων τιμών (outliers) στα δεδομένα (Οικονόμου και Καρώνη, 2010).

*2.5.2 Επιρροή*

Θα εξετάσουμε την επιρροή μιας παρατήρησης στη προσαρμογή ενός μοντέλου. Τα γραφήματα των υπολοίπων πιθανοφάνειας ή deviance ως προς τα $\hat{h}_{ii}$ βοηθούν στην εξέταση σημείων επιρροής στην εκτίμηση του μοντέλου.

*2.5.3 Απόσταση Cook*

Ένα πολύ χρήσιμο μέτρο για τον εντοπισμό σημείων επιρροής, κατά πόσο η αφαίρεση μιας συγκεκριμένης παρατήρησης θα επηρεάσει τις εκτιμήσεις των παραμέτρων ενός μοντέλου, δίνεται από την απόσταση Cook.

Η στατιστική συνάρτηση του Cook είναι

$$CD_i = \frac{1}{p}\left(\hat{\beta}_{(i)} - \hat{\beta}\right)' I\left(\hat{\beta}\right)\left(\hat{\beta}_{(i)} - \hat{\beta}\right), i = 1, \dots, N$$

όπου είναι η παρατηρούμενη πληροφορία κατά Fisher είναι $I\left(\hat{\beta}\right) = (X'\widehat{W}X)$ και $\hat{V}\left(\hat{\beta}\right) = I^{-1}(\hat{\beta})$. Αλλά μπορεί να γραφτεί και στη πιο απλή μορφή

$$CD_i = \frac{\hat{h}_{ii}(r_i^{PS})^2}{p(1-\hat{h}_{ii})},$$

όπου $p = \kappa + 1$ ο αριθμός των παραμέτρων στο μοντέλο.

## 2.6 Επιλογή Μοντέλου

**Δείκτες καλής προσαρμογής AIC και BIC**

*Akaike's information criterion (AIC)*

Το AIC αποτελεί ένα κριτήριο επιλογής βέλτιστου μοντέλου με $l\left(\hat{\beta}\right)$ να είναι η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμηθέν μοντέλο και $p$ το πλήθος των παραμέτρων του μοντέλου.

Οι συναρτήσεις αυτές ορίζονται ως

$$AIC = -2l\left(\hat{\beta}\right) + 2p$$

Το προτιμητέο μοντέλο είναι αυτό με το μικρότερο $AIC$.Η εισαγωγή επιπλέον παραμέτρων στο μοντέλο βελτιώνει τη προσαρμογή του μοντέλου ανεξάρτητα απ' το αν είναι στατιστικά σημαντικές ή όχι.

*Bayesian information criterion (BIC)*

Το BIC αποτελεί ένα κριτήριο επιλογής βέλτιστου μοντέλου μεταξύ μοντέλων με διαφορετικό αριθμό παραμέτρων όπως και στο AIC. Η διαφορά τους είναι ότι η εισαγωγή επιπρόσθετων παραμέτρων αποθαρρύνεται σε μεγαλύτερο βαθμό από το AIC.

$$BIC = -2l(\hat{\beta}) + pln(n)$$

## 2.7 Παράδειγμα στη Poisson Παλινδρόμηση

*Καπνιστές Βρετανοί γιατροί και ο στεφανιαίος θάνατος*

Τα δεδομένα του Πίνακα 2.2 είναι από μια διάσημη μελέτη που διεξήχθη από τον Sir Richard Doll και τους συνεργάτες του. Το 1951, σε όλους τους Βρετανούς γιατρούς είχε σταλεί ένα σύντομο ερωτηματολόγιο σχετικά με το αν κάπνιζαν καπνό. Από τότε συλλέχθηκαν πληροφορίες σχετικά με τους θανάτους. Ο Πίνακας 2.2 παρουσιάζει τον αριθμό των θανάτων από τη στεφανιαία νόσο στους άνδρες γιατρούς 10 χρόνια μετά την έρευνα. Δείχνει επίσης το συνολικό αριθμό ατόμου ανά έτη παρατήρησης κατά τη στιγμή της ανάλυσης (Breslow και Day, 1987). Οι ερωτήσεις ενδιαφέροντος είναι:

Α) Είναι ο ρυθμός θανάτου υψηλότερος για τους καπνιστές σε σχέση με τους μη-καπνιστές;

Β) Αν ναι, κατά πόσο;

Γ) Είναι διαφορετικό το αποτέλεσμα σχετικά με την ηλικία;

| Μεταβλητές | Περιγραφή |
|---|---|
| Y (deaths) | Ο αριθμός των θανάτων από τη στεφανιαία νόσο στους άνδρες γιατρούς 10 χρόνια μετά την έρευνα. |
| X1 (smoke) | Η μεταβλητή smoke είναι κατηγορική, παίρνει τη τιμή 1 αν το άτομο είναι καπνιστής και αν όχι τη τιμή 0 |
| X2 (agecat) | Αποτελεί τα γκρουπ των ηλικιακών ομάδων και παίρνει τη τιμή 1 αν το άτομο είναι ηλικίας από 35-44,<br>τη τιμή 2 αν είναι ηλικίας από 45-54,<br>τη τιμή 3 αν είναι από 55-64,<br>τη τιμή 4 αν είναι από 65-74,<br>τη τιμή 5 αν είναι από 75-84 |
| Ni | Ο αριθμός των γιατρών σε κίνδυνο και τις παρατηρούμενες περιόδους σε κάθε ηλικιακή ομάδα. |

**Πίνακας 2.1**: Περιγραφή των μεταβλητών του Παραδείγματος 2.7

Εδώ αναφέρουμε κάποια πράγματα από θεωρία τα οποία είναι χρήσιμα στο παράδειγμά που ακολουθεί.

$$E(y_i) = \mu_i = n_i \exp(\beta' x_i)$$

$$\ln(\mu_i) = \ln(n_i) + \beta' x_i$$

και ο όρος $\ln(n_i) = offset(\log(n_i))$.

Με τις παρακάτω εντολές στην R διαβάζουμε το αρχείο **c:/doctors.txt** και παίρνουμε τα δεδομένα που φαίνονται στο Πίνακα 2.2.

**>cdat<- read.table("c:/doctors.txt", header=TRUE)**

**>attach(cdat)**

**>cdat**

| | agecat | smoke | deaths | ni |
|---|---|---|---|---|
| 1 | 1 | 1 | 32 | 52407 |
| 2 | 2 | 1 | 104 | 43248 |
| 3 | 3 | 1 | 206 | 28612 |
| 4 | 4 | 1 | 186 | 12663 |
| 5 | 5 | 1 | 102 | 5317 |
| 6 | 1 | 0 | 2 | 18790 |
| 7 | 2 | 0 | 12 | 10673 |
| 8 | 3 | 0 | 28 | 5710 |
| 9 | 4 | 0 | 28 | 2585 |
| 10 | 5 | 0 | 31 | 1462 |

**Πίνακας 2.2**:Πίνακας εισαγωγής των δεδομένων του Παραδείγματος 2.7 στην R

Στην αρχή θα προσαρμόσουμε το μοντέλο μας για να εξετάσουμε αν υπάρχει εξάρτηση του θανάτου με ηλικιακή ομάδα και αν σε κάθε ομάδα οι γιατροί είναι καπνιστές ή όχι. Οπότε το μοντέλο μας θα είναι:

$$\beta' x_i = \beta_0 + \beta_1 smoke + \beta_2 agecat$$

Η αναμενόμενη τιμή θα είναι ίση με:

$$E(y_i) = \mu_i = n_i \exp(\beta' x_i)$$

Ο λογάριθμος θα ισούται με:

$$\ln(\mu_i) = \ln(n_i) + \beta' x_i$$

όπου $\ln(n_i) = offset(\log(n_i))$.

Όπότε το μοντέλο μας θα πάρει τη μορφή:

$$\ln(\mu_i) = \beta_0 + \beta_1 smoke + \beta_2 agecat + offset(\log(n_i))$$

**> mod1<-glm(deaths~smoke+agecat+offset(log(ni)),family=poisson)**

**> summary(mod1)**

Αποτελέσματα 2.1

```
Call:
glm(formula = deaths ~ smoke + agecat + offset(log(ni)), family = poisson)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.5712  -2.7562   0.2857   1.4261   3.7183

Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)  -8.11833    0.13929   -58.282   < 2e-16 ***
smoke         0.40637    0.10720    3.791    0.00015 ***
agecat        0.83583    0.02904   28.777    < 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: 935.067  on 9  degrees of freedom
Residual deviance:  69.182  on 7  degrees of freedom
AIC: 130.25

Number of Fisher Scoring iterations: 4
```

Σχόλια

- Στο επάνω μέρος του πίνακα παίρνουμε πληροφορίες για το εύρος του δείγματος, το ενδοτεταρτομοριακό εύρος και για τη διάμεσο του δείγματος.
- Παρατηρούμε ότι όλοι οι συντελεστές είναι στατιστικά σημαντικοί αφού η p-value είναι πολύ μικρή σ' όλες τις περιπτώσεις.

**> confint(mod1)**

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -8.3965585 | -7.8503272 |
| smoke | 0.2013765 | 0.6220316 |
| agecat | 0.7791144 | 0.8929976 |

- Με την παραπάνω εντολή κατασκευάζονται 95% διαστήματα εμπιστοσύνης για τις παραμέτρους $\beta_j$ του μοντέλου μας όπως φαίνεται στο πάνω πίνακα.

**> exp(confint(mod1))**

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 0.0002256425 | 0.0003896245 |
| smoke | 1.2230851686 | 1.8627085523 |
| agecat | 2.1795412770 | 2.4424402693 |

- Με την παραπάνω εντολή κατασκευάζονται τα αντίστοιχα 95% διαστήματα εμπιστοσύνης για τις παραμέτρους $exp(\beta_j)$ του μοντέλου μας όπως φαίνεται στο πάνω πίνακα.

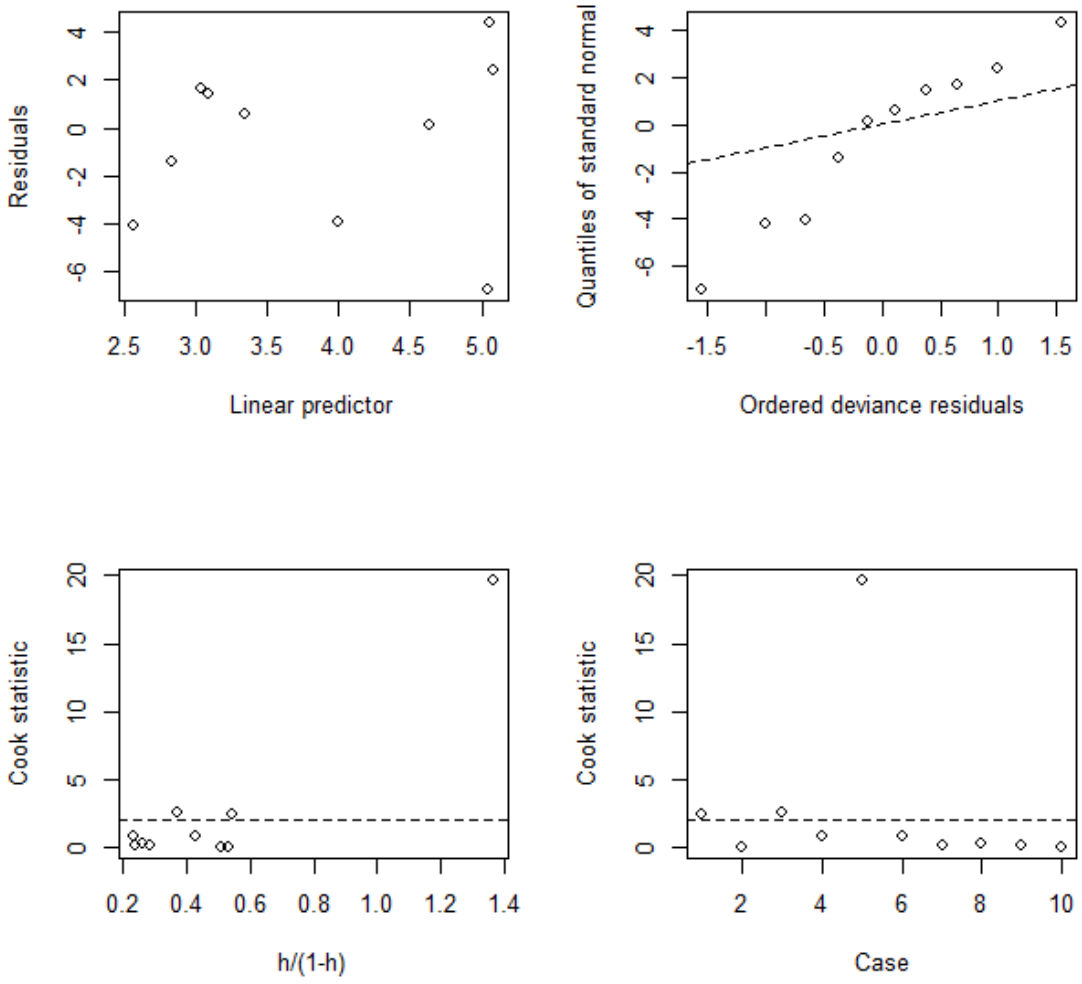
**> logLik(mod1)**

'log Lik.' -62.12501 (df=3)

Οπότε η μεγιστοποιημένη συνάρτηση πιθανοφάνειας του μοντέλου με 3 βαθμούς ελευθερίας είναι:

$$\hat{l}_0 = -62.12501$$

**>glm.diag.plots(mod1)**



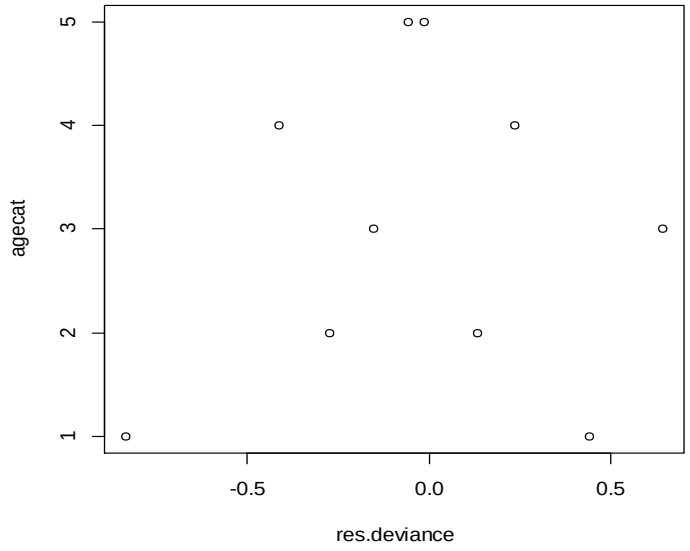

**Σχήμα 2.1**

Παρατηρήσεις για το Σχήμα 2.1

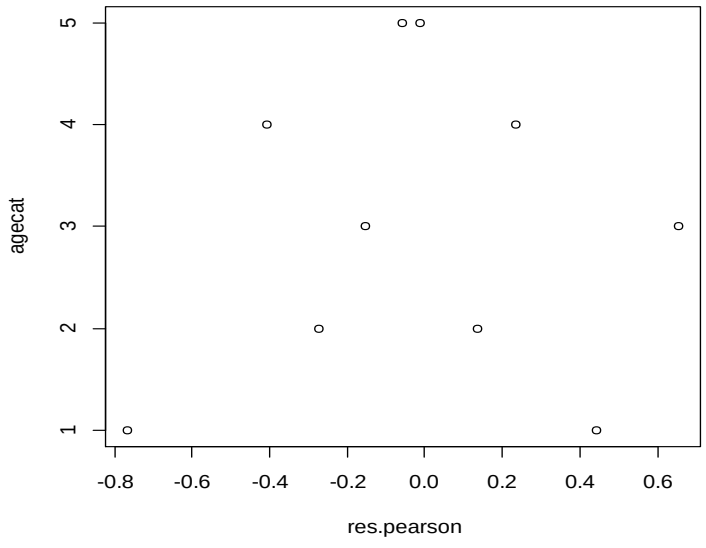
- Στο πάνω αριστερά γράφημα βλέπουμε τα υπόλοιπα έναντι του εκτιμημένου γραμμικού συνδυασμού $\hat{n}_i = x'_i\hat{\beta}$. Εδώ φαίνεται να υπάρχει κάποιο συστηματικό σχήμα. Άρα στο μοντέλο μας πιθανώς να έχει πρόβλημα.
- _Παρατηρούμε ότι η πάνω δεξιά γραφική παράσταση των υπολοίπων deviance δεν είναι ικανοποιητική.
- Στην κάτω δεξιά γραφική παράσταση παρατηρούμε την ύπαρξη ακραίας παρατήρησης και συγκεκριμένα της $5^{ης}$ παρατήρησης (δηλαδή για τους γέρους καπνίζοντες), το ίδιο φαίνεται και στη κάτω δεξιά γραφική παράσταση.

Στη συνέχεια με τις παρακάτω εντολές στην R θα κάνουμε τα γραφήματα των υπολοίπων deviance και pearson έναντι της ηλικιακής ομάδας.

```
>res.deviance<-residuals(mod1,type="deviance")
>res.pearson<-residuals(mod1,type="pearson")
>plot(res.deviance,agecat)
>plot(res.pearson,agecat)
```




**Σχήμα 2.2:** Υπόλοιπα deviance σε σχέση με τη μεταβλητή agecat




**Σχήμα 2.3:** Υπόλοιπα pearson σε σχέση με τη μεταβλητή agecat

Συμπέρασμα

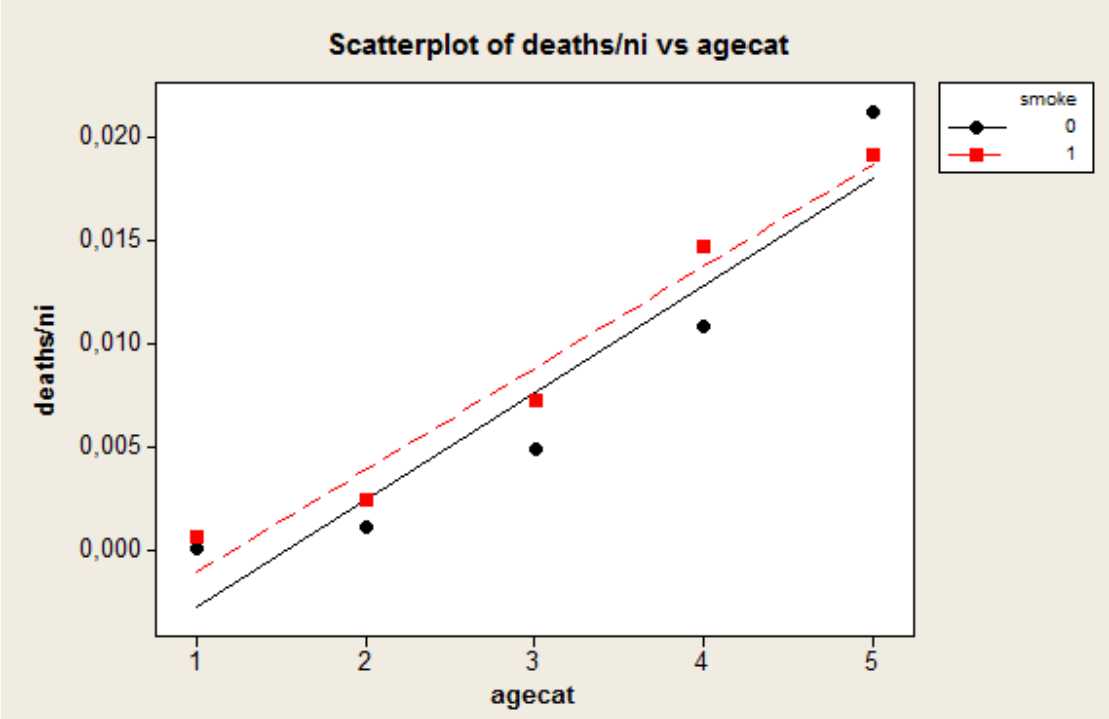
Παρατηρούμε και στα στις δύο γραφικές παραστάσεις τα σημεία να σχηματίζουν καμπύλη γεγονός το οποίο μας οδηγεί στο να εισάγουμε στο μοντέλο το τετράγωνο της agecat. Με τη χρήση της R έχουμε:

**>agesq<-agecat^2**

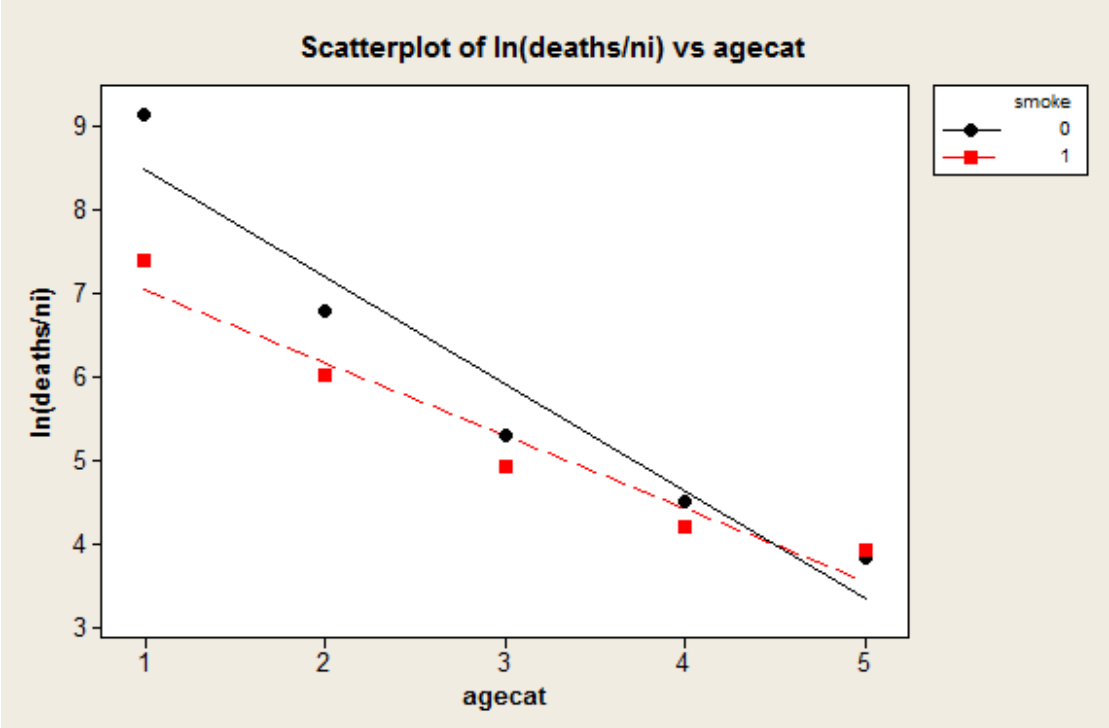
Η μεταβλητή agesq είναι το τετράγωνο της agecat μεταβλητής λαμβάνοντας υπόψη το μη-γραμμικό ρυθμό αύξησης.

Στη συνέχεια θα κάνουμε έναν γραφικό έλεγχο της μεταβλητής deaths/ni και log(deaths/ni) σε σχέση με την επεξηγηματική μεταβλητή agecat προκειμένου να εξετάσουμε τις κλίσεις των ευθειών παλινδόμησης για τους καπνίζοντες και μη καπνίζοντες.

Με τη χρήση του Minitab παίρνουμε τις παρακάτω γραφικές παραστάσεις.




**Σχήμα 2.4**




**Σχήμα 2.5**

Στα Σχήματα 2.4 και 2.5 παρατηρείται ότι η μεταβλητή απόκρισης deaths/ni και ln(deaths/ni) αντίστοιχα σε σχέση με την agecat, οι ευθείες να έχουν διαφορετική κλίση που σημαίνει ότι υπάρχει αλληλεπίδραση μεταξύ των δύο ομάδων των καπνιστών και μη καπνιστών σε σχέση με την ηλικιακή ομάδα. Οπότε θα δημιουργήσουμε μια μεταβλητή η οποία θα εκφράζει την αλληλεπίδραση των μεταβλητών smoke και agecat. Αυτό θα γίνει χρησιμοποιώντας την παρακάτω εντολή στην R:

**>smkage<-smoke*agecat**

Η μεταβλητή smkage εκφράζει την αλληλεπίδραση των μεταβλητών smoke και agecat είναι:

ίση με agecat όταν το άτομο είναι καπνιστής(smoke=1)

ή

ίση με μηδέν(0) όταν το άτομο είναι μη καπνιστής(smoke=0).

Βελτιωμένο Μοντέλο

Προσαρμόζουμε το νέο μας μοντέλο προσθέτοντας τις καινούργιες μεταβλητές agesq και smkage.

**>mod<-glm(deaths~smoke+agecat+agesq+smkage+offset(log(ni)), family=poisson)**

Η παραπάνω εντολή μας βοηθάει να κάνουμε τη προσαρμογή του γενικευμένου γραμμικού μοντέλου μας χρησιμοποιώντας τη κατανομή Poisson.

**>summary(mod)**

Αποτελέσματα 2.2

```
Call:
glm(formula = deaths ~ smoke + agecat + agesq + smkage + offset(log(ni)),
    family = poisson)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.43820 -0.27329 -0.15265  0.23393 -0.05700 -0.83049  0.13404  0.64107
      9       10
-0.41058 -0.01275

Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept) -10.79176   0.45008   -23.978   < 2e-16 ***
smoke         1.44097   0.37220    3.872    0.000108 ***
agecat        2.37648   0.20795   11.428    < 2e-16 ***
agesq        -0.19768   0.02737   -7.223    5.08e-13 ***
smkage       -0.30755   0.09704   -3.169    0.001528 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual deviance:   1.6354  on 5  degrees of freedom
AIC: 66.703
```

Παρατηρήσεις σύμφωνα με τα αποτελέσματα 2.2

- Στο επάνω μέρος του πίνακα πληροφορούμαστε για τις τιμές των υπολοίπων deviance.
- Στη συνέχεια παίρνουμε κάποιες πληροφορίες σχετικά με τους συντελεστές του γενικευμένου γραμμικού μοντέλου μας, στη πρώτη στήλη για τους εκτιμημένους συντελεστές, στη δεύτερη στήλη για τη τυπική απόκλιση, στη Τρίτη στήλη για το z στατιστικό και στη τέταρτη για το p-value.
- Μπορούμε να δούμε ότι σ' όλες τις περιπτώσεις η p-value είναι αρκετά μικρή που σημαίνει ότι όλοι οι συντελεστές είναι στατιστικά σημαντικοί οπότε απορρίπτουμε τις μηδενικές υποθέσεις ότι δηλαδή $\beta_0 = 0, \beta_1 = 0, \beta_2 = 0, \beta_3 = 0\ , \beta_4 = 0$ και δεχόμαστε τις εναλλακτικές υποθέσεις ότι είναι διάφοροι του μηδενός. Άρα η αλληλεπίδραση των μεταβλητών agecat και smoke είναι στατιστικά σημαντική.
- Προτιμούμε το μοντέλο με την αλληλεπίδραση έναντι του μοντέλου χωρίς την αλληλεπίδραση λόγω του κριτηρίου AIC (θέλουμε τη μικρότερη τιμή στο AIC).
- Η αναμενόμενη τιμή της $\beta' x_i$ είναι μεγαλύτερη κατά exp(1.44097) στους καπνιστές από τους μη καπνιστές όταν ανήκουν στην ίδια κατηγορία ηλικίας.
- Η αναμενόμενη τιμή της $\beta' x_i$ θα αυξηθεί κατά exp(2.37648) αν αυξήσουμε κατά μια μονάδα τη κατηγορία ηλικίας με σταθερή όμως κατηγορία καπνίσματος (δηλαδή είτε καπνιστές είτε όχι).

**> confint(mod)**

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -11.7110551 | -9.9453230 |
| smoke | 0.7364876 | 2.1984920 |
| agecat | 1.9782335 | 2.7938787 |
| agesq | -0.2522023 | -0.1448639 |
| smkage | -0.5014372 | -0.1203298 |

- Με την παραπάνω εντολή κατασκευάζονται 95% διαστήματα εμπιστοσύνης για τις παραμέτρους $\beta_j$ του μοντέλου μας όπως φαίνεται στο πάνω πίνακα.

**> exp(confint(mod))**

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 8.202635e-06 | 4.795138e-05 |
| smoke | 2.088587e+00 | 9.011414e+00 |
| agecat | 7.229960e+00 | 1.634429e+01 |
| agesq | 7.770875e-01 | 8.651400e-01 |
| smkage | 6.056596e-01 | 8.866280e-01 |

- Με την παραπάνω εντολή κατασκευάζονται τα αντίστοιχα 95% διαστήματα εμπιστοσύνης για τις παραμέτρους $exp(\beta_j)$ του μοντέλου μας όπως φαίνεται στο πάνω πίνακα.

**>logLik(mod)**

'log Lik.' -28.35166 (df=5)

Οπότε η μεγιστοποιημένη συνάρτηση πιθανοφάνειας του μοντέλου με 5 βαθμούς ελευθερίας είναι:

$$\hat{l}_1 = -28.35166$$

Εδώ θα κάνουμε έναν έλεγχο ανάμεσα στο πρώτο μας μοντέλο $M_0$ και στο βελτιωμένο μας μοντέλο $M_1$ με μεγιστοποιημένες συναρτήσεις πιθανοφάνειας $\hat{l}_0$ και $\hat{l}_1$ αντίστοιχα. Θα κάνουμε τον έλεγχο $-2(\hat{l}_0 - \hat{l}_1) \sim X_2^2$ . Η χρήση της κατανομής $X^2$ δικαιολογείται μόνο ασυμπτωτικά και οι βαθμοί ελευθερίας είναι 2 όση δηλαδή και η διαφορά στο πλήθος των παραμέτρων ανάμεσα στα δύο μοντέλα. Με τη χρήση του στατιστικού πακέτου της R και χρησιμοποιώντας τις παρακάτω εντολές υπολογίζουμε τη p-τιμή του ελέγχου.

```
> ddev ← -2*(logLik(mod1)-logLik(mod))
> ddev
'log Lik.' 67.54671 (df=3)
> pvalue ←1-pchisq(ddev,2)
> pvalue
'log Lik.' 2.109424e-15 (df=3)
```

Για τη σύγκριση αυτών των δύο εμφωλευμένων μοντέλων συγκρίνουμε τις τιμές της ελεγχοσυνάρτησης deviance με τη χρήση της εντολής:

>**anova(mod1,mod,test="Chisq")**

με την οποία λαμβάνουμε τα αποτελέσματα στα οποία παρουσιάζεται ο πίνακας ανάλυσης της deviance για τα δύο μοντέλα. Από τη τιμή του ελέγχου παρατηρούμε ότι απορρίπτεται το πιο απλό μοντέλο ( αυτό χωρίς τις agesq και smkage ) διότι η πρόσθεση των δύο συμμεταβλητών μειώνει σημαντικά τη deviance.

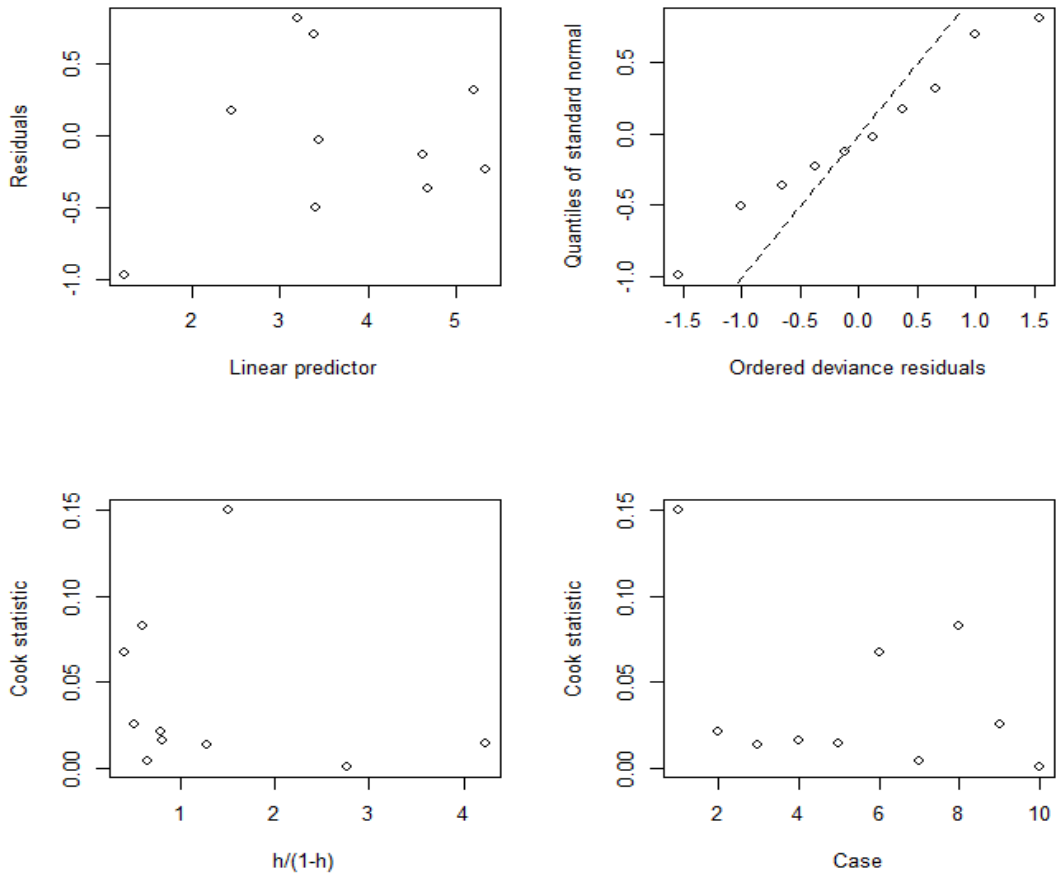
Αποτελέσματα 2.3

```
Analysis of Deviance Table

Model 1: deaths ~ smoke + agecat + offset(log(ni))
Model 2: deaths ~ smoke + agecat + agesq + smkage + offset(log(ni))
  Resid.    Df Resid.    Dev      Df     Deviance      Pr(>Chi)
     1          7         69.182
     2          5         1.635     2     67.547          2.15e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**>library(boot)**

**>glm.diag.plots(mod)**




Παρατηρήσεις

- Στο πάνω αριστερά γράφημα βλέπουμε τα υπόλοιπα έναντι του εκτιμημένου γραμμικού συνδυασμού $\hat{n}_i = x'_i\hat{\beta}$. Δε φαίνεται κάποιο συστηματικό σχήμα να υπάρχει άρα το μοντέλο μας πιθανώς να μην έχει πρόβλημα.
- Στο πάνω δεξιά γράφημα των υπολοίπων deviance βλέπουμε να είναι ικανοποιητικό.
- Στην κάτω δεξιά γραφική παράσταση παρατηρούμε την ύπαρξη ακραίας παρατήρησης και συγκεκριμένα της $1^{ης}$ (δηλαδή για τους καπνίζοντες της πιο νέας ηλικιακής ομάδας) αλλά και δύο παρατηρήσεων που ξεφεύγουν λιγότερο απ την προηγούμενη και αυτές οι παρατηρήσεις είναι η $6^{ης}$ και $8^{ης}$, το ίδιο προκύπτει και απ' τη κάτω δεξιά γραφική παράσταση που παρατηρούνται τρία απόμακρα σημεία.

# ΚΕΦΑΛΑΙΟ ΙΙΙ
# ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΓΙΑ ΑΠΑΡΙΘΜΗΤΕΣ ΧΡΟΝΟΣΕΙΡΕΣ

## 3.1.Γενικευμένα γραμμικά μοντέλα χρονοσειρών για απαριθμητά δεδομένα

Έστω $\{Y_t\}$ η χρονοσειρά ενδιαφέροντος μας και καλείται απόκριση. Κοιτάζοντας μπροστά στη πρόβλεψη θα έχουμε:

$$Z_{t-1} = \left(Z_{(t-1)_1}, \dots \dots .., Z_{(t-1)_p}\right)'.$$

Το οποίο είναι το αντίστοιχο διάνυσμα p-διάστασης των προηγούμενων επεξηγηματικών για $t = 1, .., N$. Οπότε συμβολίζουμε με $F_{t-1}$ το σ-πεδίο που δημιουργείται από $[Y_{t-1}, Y_{t-2}, \dots ., Z_{t-1}, Z_{t-2}, \dots \dots]$. Άρα

$$F_{t-1} = \sigma\{Y_{t-1}, Y_{t-2}, \dots ., Z_{t-1}, Z_{t-2}, \dots \dots\}.$$

Ο συμβολισμός αυτός δίνει έμφαση στο γεγονός ότι οι τιμές του παρελθόντος του $\{Y\}$, χρησιμοποιούνται στη παραγωγή του $F_{t-1}$ μερικές φορές είναι βολικό να σκεφτούμε το $Z_{t-1}$ ως συμπεριλαμβανομένων παρελθόντων τιμών των αποκρίσεων $Y_{t-1}, Y_{t-2}, \dots$ (Kedem και Fokianos, 2002).

Όταν ορισμένες συμμεταβλητές $X_t, W_t, ..$ είναι γνωστοί στο t-1, τότε γράφουμε

$$F_{t-1} = \sigma\{Y_{t-1}, Y_{t-2}, \dots ., X_t, W_t, \dots ., Z_{t-1}, Z_{t-2}, \dots \dots\}.$$

Άρα η $F_{t-1}$ παράγεται από τιμές της σειράς απόκρισης του παρελθόντος και του παρόντος όταν είναι γνωστές οι μεταβλητές. Άρα η παρατήρηση στο χρόνο t-1 ενδεχομένως να συμπεριλαμβάνει τις $X_t, W_t$ όταν είναι γνωστές.

Για παράδειγμα έστω η σχέση μεταξύ της ροής ποταμιού μεταβλητής απόκρισης $(Y)$ και της συμμεταβλητής βροχοπτώσεων $(X)$, όταν η βροχόπτωση $X$ σε συγκεκριμένη τοποθεσία είναι γνωστή τη χρονική στιγμή t και η ροή του ποταμιού προσδιορίζεται πιθανώς σε διαφορετική τοποθεσία και σε μεταγενέστερη χρονική στιγμή. Τότε η $X_t$ εισάγεται στο μοντέλο ως τιμή παρελθόντος.

Ορίζουμε με $\mu_t = E(Y_t|F_{t-1})$ την αναμενόμενη τιμή της μεταβλητής απόκρισης που δίνεται από το παρελθόν. Το πρόβλημα είναι να τη συσχετίσουμε με τις

συμμεταβλητές. Στη κλασική θεωρία των γραμμικών μοντέλων υποτίθενται ότι η συνθήκη αναμονής της μεταβλητής απόκρισης δίνεται στο παρελθόν της διαδικασίας και είναι γραμμική συνάρτηση των συμμεταβλητών της ή των επεξηγηματικών μεταβλητών της. Ωστόσο υπάρχουν προβλήματα με αυτή τη προσέγγιση όταν τα δεδομένα δεν είναι της Κανονικής κατανομής και τότε η σχέση της αναμενόμενης μεταβλητής $\mu_t$ με τις συμμεταβλητές οδηγεί σε παράλογα αποτελέσματα. Για παράδειγμα, για δεδομένα Poisson με μέσο $\mu_t$ η γραμμική παλινδρόμηση μπορεί να οδηγήσει σε αρνητική εκτίμηση του μέσου. Τα γενικευμένα γραμμικά μοντέλα λύνουν αυτά τα προβλήματα όταν οι παρατηρήσεις ακολουθούν μια κατανομή από την εκθετική οικογένεια καθιστώντας το κλασσικό γραμμικό μοντέλο ειδική περίπτωση. Όσον αφορά τις χρονοσειρές, οι βασικές ιδέες των γενικευμένων γραμμικών μοντέλων και οι οικογένειες εκθετικών κατανομών όπως και οι συναρτήσεις σύνδεσης μπορούν εύκολα να επεκταθούν. Ορίζουμε χρονοσειρές που ακολουθούν τα γενικευμένα γραμμικά μοντέλα, τα μοντέλα που αποτελούνται από τις ακολουθίες τυχαίων και συστηματικών συνιστωσών.

1.Τυχαία συνιστώσα

Η συνθήκη κατανομής της μεταβλητής απόκρισης που δόθηκε στο παρελθόν ανήκει στην Ε.Ο.Κ. σε κανονική μορφή.

$$f(y_t; \theta_t, \varphi) = exp\left\{\frac{y_t\,\theta_t - b(\theta_t)}{a_t(\varphi)} + c(y_t, \varphi)\right\}.$$

Η παραμετρική συνάρτηση $a_t(\varphi)$ είναι της μορφής φ/$\omega_t$όπου φ είναι μια παράμετρος διασποράς και $\omega_t$ είναι μια γνωστή παράμετρος που αναφέρεται ως βάρος. Η παράμετρος $\theta_t$ θα καλείται ως φυσική παράμετρος της κατανομής.

2. Συστηματική συνιστώσα

Για t=1,…,N υπάρχει μονότονη συνάρτηση $g(.)$ όπου

$$g(\mu_t) = n_t = \sum_{j=1}^{p} \beta_j\, Z_{(t-1)j} = Z'_{t-1}\beta.$$

Η συνάρτηση $g(.)$ καλείται συνάρτηση σύνδεσης όπου $n_t$ αναφέρεται ως γραμμική πρόβλεψη του μοντέλου.

Τυπικές επιλογές για ${Z'}_{t-1}\beta$ μπορεί να είναι

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

Δίνεται λοιπόν μια δισδιάστατη διαδικασία στην απόκριση $\{Y\}$ και στη συμμεταβλητή ή επεξηγηματική διαδικασία $\{X_t\}$. Μια ενδιαφέρον επιλογή του ${Z'}_{t-1}\beta$ δίνεται από την ακόλουθη έκφραση:

$$ {Z'}_{t-1}\beta = {X'}_t\gamma + \sum_{i=1}^{p} \Phi_i \, H_i(Y_{t-i}) + \sum_{i=1}^{q} \theta_i \, D_i(\mu_{t-i}) $$

όπου $H_i(.)$, $D_i(.)$ γνωστές συναρτήσεις για κάθε i.

Η ειδική περίπτωση κάτω από τον ορισμό

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

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

κατά συνέπεια το συμμεταβλητό διάνυσμα

$$ g(\mu_t) = n_t = {X'}_t\gamma + \textstyle\sum_{i=1}^{p} \Phi_i \, (g(Y_{t-i}) - {X'}_{t-i}\gamma) + \sum_{i=1}^{q} \theta_i \in_{t-i} \qquad (3.1) $$

όπου $\in_{t-i} = g(Y_{t-i}) - n_{t-i}$.

Το μοντέλο (3.1) έχει ονομαστεί μοντέλο γενικευμένου γραμμικού αυτοπαλινδρόμικού κινητού μέσου GLARMA(p,q) ή γενικευμένου αυτοπαλινδρόμικού κινητού μέσου GARMA τάξης (p,y), το συγκεκριμένο GARMA εξαρτάται από την συνθήκη κατανομής. Οπότε όταν οι συνθήκες κατανομής είναι Poisson και εμείς παίρνουμε μια Poisson GARMA(p,q)

$$ g(u) = \begin{cases} \dfrac{u^{\lambda} - 1}{\lambda} \ \gamma\iota\alpha \ \lambda \neq 0 \\ \log(u) \ \ \gamma\iota\alpha \ \lambda = 0 \end{cases} $$

Η παραπάνω είναι μια εύκαμπτη σύνδεση όπου στις περιπτώσεις λ=1 και λ=0 παίρνουμε ένα γραμμικό και ένα λογαριθμογραμμικό μοντέλο αντίστοιχα.

### 3.2 Μοντελοποίηση

Έστω $Y_t, t = 1, \dots., N$ μια απαριθμητή χρονοσειρά που λαμβάνει τιμές μη αρνητικών ακεραίων. Θεωρούμε ότι $Y_t$ είναι η απόκριση και το πιο φυσικό υποψήφιο μοντέλο κατανομής για την διαδικασία της απόκρισης είναι το μοντέλο της κατανομής Poisson. Το μοντέλο του $Y_t$ καθορίζεται από την υπόθεση ότι η υπό συνθήκη πιθανότητα της απόκρισης δοθέντος των παρελθόντων τιμών είναι Poisson με μέση τιμή $\mu_t$,

$$f(y_t; \mu_t | F_{t-1}) = \frac{exp(-\mu_t)\mu_t^{y_t}}{y_t!}, \text{ t=1,……,N} \qquad (3.2)$$

όπου $F_{t-1}$ υποδηλώνει το παρελθόν (διαθέσιμες πληροφορίες για τον παρατηρητή μέχρι το χρόνο t. Για το μοντέλο της Poisson η υπό συνθήκη της αναμενόμενης τιμής είναι ίση με την υπό συνθήκη διασποράς.

$$E(Y_t|F_{t-1}) = Var(Y_t|F_{t-1}) = \mu_t, t = 1, \dots \dots, N$$

με $\{Z_{t-1}\}, t = 1, \dots . N$ δηλώνεται ένα διάνυσμα p-διαστάσεων το οποίο πιθανώς περιλαμβάνει τιμές της διαδικασίας του παρελθόντος και οποιαδήποτε άλλη βοηθητική πληροφορία. Μια τυπική επιλογή του $Z_{t-1}$ είναι $Z_{t-1} = (1, X_t, Y_{t-1})'$

και αν προστεθούν οι αλληλεπιδράσεις τότε:

$Z_{t-1} = (1, X_t, Y_{t-1}, X_t Y_{t-1})'$ και με $\{X_t\}$ μια επιπλέον διαδικασία.

Ένα κατάλληλο μοντέλο για την ανάλυση χρονοσειρών με απαριθμητά δεδομένα λαμβάνετε μέσω του καθορισμού:

$$\mu_t = h(Z'_{t-1}\beta), \ t = 1, \dots., N \qquad (3.3)$$

όπου $\beta$ ένα διάνυσμα άγνωστων παραμέτρων p-διάστασης και με $h(.) = g^{-1}(.)$ να είναι αντίστροφη συνάρτηση σύνδεσης, με $\mu_t > 0$ για κάθε t.

Με μια τροποποίηση του μοντέλου (3.3) θα έχουμε μια παραλλαγή της μορφής

$$\mu_t(\beta) = exp(\beta_1 X_t)[1 + exp(-\beta_0 - \beta_2 Y_{t-1})], t = 1, \dots, N \qquad (3.4)$$

που έχει αναφερθεί από τον Wong (1986), ο Holden (1987) εφάρμοσε την (3.3) σε ένα μοντέλο μετάδοσης για την αεροπειρατεία αεροσκάφους. Το πολλαπλασιαστικό

μοντέλο το οποίο γενικεύει (3.3) την αίσθηση ότι μόνο η πρώτη και η δεύτερη στιγμή που προσδιορίζονται έχει προωθηθεί από τους Zeger και Quash (1988), και αυτό επεκτάθηκε λαμβάνοντας υπόψη τις μη γραμμικές συναρτήσεις των παραμέτρων από τον Efron (1986). Μια ενδιαφέρουσα τροποποίηση προτάθηκε από τον Slud (1995), όπου ένα μοντέλο ARMA έχει προσαρμόστηκε στα υπόλοιπα από την παλινδρόμηση κατά Poisson, χρησιμοποιώντας τα δεδομένα θνησιμότητας.

Προς ένα πιο γενικό μοντέλο, ο Fokianos (2001) υποθέτει ότι η υπό συνθήκη κατανομή της απόκρισης δοθέντος του παρελθόντος είναι η διπλά κολοβή (doubly truncated) κατανομή Poisson. Έστω $\{Y_t\}$, t = 1,. . . , N, μια απαριθμητή χρονοσειρά και υποθέτουμε ότι οι τιμές κάτω από μια γνωστή σταθερά $c_1$ και οι τιμές που υπερβαίνουν μια άλλη γνωστή σταθερά $c_2$ παραλείπονται με $c_1 < c_2$. Στη συνέχεια, η διπλά κολοβή υπό συνθήκη πυκνότητας πιθανότητας της Poisson είναι

$$f(y_t; \mu_t; c_1, c_2 | F_{t-1}) = \frac{{\mu_t}^{y_t}}{y_t! \psi(c_1, c_2, \mu_t)}, \quad y_t = c_1, \dots, c_2 \qquad (3.5)$$

για $t = 1, \dots, N$ , που $\psi$ η συνάρτηση ορίζεται ως

$$\psi(c_1, c_2, \mu) = \begin{cases} \sum_{y=c_1}^{c_2} \frac{\mu^y}{y!}, \text{όπου } 0 \le c_1 < c_2 \\ \psi(0, c_2, \mu), \text{διαφορετικά} \end{cases} \qquad (3.6)$$

όπου $\psi(0, \infty, \mu) = \exp(\mu)$. Η επιλογή $c_1 = 0, c_2 = \infty$ οδηγεί στο κοινό μοντέλο της Poisson (3.2), καθώς η εξίσωση (3.5) μειώνεται καλείται αριστερά κολοβή κατανομή Poisson όταν $c_2 = \infty$. Ειδικότερα, όταν , $c_1 = 0$ , $c_2 = \infty$ τότε η (3.5) γίνεται

$$f(y_t; \mu_t; 1; \infty | F_{t-1}) = \frac{{\mu_t}^{y_t}}{y_t! (exp(\mu_t) - 1)}, t = 1, \dots, N$$

αυτή είναι η συνάρτηση μάζας πιθανότητας της θετικής κατανομής Poisson. Ομοίως, θέτοντας $c_1 = 0$, η εξίσωση (3.5) αποδίδει τη δεξιά κολοβή κατανομή Poisson. Οι Johnson et al. (1992) διατυπώνουν αρκετές πρόσθετες πληροφορίες σχετικά με τις ιδιότητες αυτών των κατανομών.

Οι Kedem-Fokianos (2002) αναφέρουν ότι η υπό συνθήκη μέση τιμή και διασπορά της κολοβής κατανομής Poisson δίνονται από τους παρακάτω τύπους αντίστοιχα.

$$E^{tr}[Y_t; c_1; c_2|F_{t-1}] = \mu_t \frac{\psi(c_1 - 1, c_2 - 1, \mu_t)}{\psi(c_1, c_2, \mu_t)} \qquad (3.7)$$

και $Var^{tr}[Y_t; c_1; c_2|F_{t-1}] = \frac{1}{\psi^2(c_1,c_2,\mu_t)}\{(\mu_t)^2\psi(c_1 - 1, c_2 - 1, \mu_t)\psi(c_1, c_2, \mu_t) +$

$\mu_t\psi(c_1 - 1, c_2 - 1, \mu_t)[\psi(c_1, c_2, \mu_t) - \mu_t\psi(c_1 - 1, c_2 - 1, \mu_t)]\}.$ (3.8)

Το μοντέλο (3.3) μπορεί επίσης να χρησιμοποιηθεί για την ανάλυση παλινδρόμησης της κολοβής απαριθμητής κατανομής. Είναι σημαντικό να σημειωθεί ότι για το μοντέλο της Poisson η υποσυνθήκη μέσης τιμής είναι ίση με την υπό συνθήκη διασποράς, ωστόσο, αυτό το γεγονός αυτό δεν ισχύει για το διπλό κολοβό μοντέλο της Poisson.

Έχουμε θεωρήσει το μοντέλο της Poisson ως υποψήφιο για την ανάλυση παλινδρόμησης απαριθμητών χρονοσειρών και ορισμένων γενικευσεών του διπλού κολοβού μοντέλου της Poisson.

## 3.3 Μοντέλα για απαριθμητές χρονοσειρές

### *3.3.1 Το μοντέλο της Poisson*

Η εξίσωση (3.3) δείχνει ότι η αντίστροφη συνάρτηση σύνδεσης $h(.)$ που χαρτοφραφεί τη πραγματική τιμή στο διάστημα $(0, \infty)$ μπορεί να χρησιμοποιηθεί για την ανάλυση απαριθμητών χρονοσειρών. Ωστόσο, σε εφαρμογές το πιο συνηθισμένο χρησιμοποιημένο μοντέλο δίνεται από την κανονική σύνδεση η οποία για το μοντέλο της Poisson (3.2) αποδεικνύεται ότι είναι λογαριθμική. Ξαναγράφοντας την εξίσωση (3.2) όπως και

$$f(y_t; \mu_t|F_{t-1}) = exp\{(y_t log\mu_t - \mu_t) - logy_t!\},\ t = 1, \dots, N.$$

Έχουμε τη μερική Πιθανοφάνεια για το μοντέλο της Poisson:

$$L(\beta) = \prod_{t=1}^{N} f(y_t; \mu_t|F_{t-1}) = \prod_{t=1}^{N} \frac{\exp(-\mu_t(\beta))\mu_t(\beta)^{y_t}}{y_t}$$

οπότε η πιθανοφάνεια θα είναι:

$$l(\beta) = logL(\beta) = \sum_{t=1}^{N} y_t log\mu_t(\beta) - \sum_{t=1}^{N} \mu_t(\beta) - \sum_{t=1}^{N} \log(y_t!)$$

με $\theta_t = log\mu_t, t = 1, \ldots\ldots, N$ και η αντίστροφη συνάρτηση σύνδεσης θα είναι $h(n_t) = \exp(n_t)\,, t = 1, \ldots\ldots, N$ και $n_t = Z'_{t-1}\beta$ , $Z_{t-1} = (1, X_t, Y_{t-1})'$.

Οπότε έχουμε το λογαριθμογραμμικό μοντέλο:

$$\mu_t(\beta) = \exp(Z'_{t-1}\beta)\,, t = 1, \ldots\ldots, N. \qquad (3.9)$$

Άρα μπορεί η παραπάνω εξίσωση (3.9) να εκφραστεί:

$$\mu_t(\beta) = exp(\beta_0 + \beta_1 X_t + \beta_2 Y_{t-1}),\ \ t = 1, \ldots, N \qquad (3.10)$$

όπου $\beta = (\beta_0, \beta_1, \beta_2)'$. Αν υποθέσουμε ότι $Y_{t-1}$ είναι μη φραγμένη τότε η υπό συνθήκη αναμενόμενη τιμή της μεταβλητής απόκρισης δοθέντος του παρελθόντος τείνει να αυξηθεί μα εκθετικό ρυθμό, όταν $\beta_2 > 0$. Όταν δεν υπάρχει εξάρτηση από τη μεταβλητή $X_t$ το μοντέλο οδηγείται σε μια σταθερή διαδικασία όταν $\beta_2 < 0$. Δηλαδή όταν η $Y_{t-1}$ είναι μη φραγμένη η παραπάνω εξίσωση μπορεί να φιλοξενήσει μόνο αρνητική συσχέτιση χωρίς να αυξάνεται εκθετικά γρήγορα. Στις περισσότερες περιπτώσεις όπου $Y_{t-1}$ ακολουθεί τη κατανομή Poisson, η πιθανότητα οι τιμές είναι πολύ μεγάλες είναι εξαιρετικά μικρή και έτσι μπορούμε να χρησιμοποιήσουμε με σιγουριά τη $Y_{t-1}$ ή την $\log(Y_{t-1})$ και παρόμοιες συμμεταβλητές στην εξίσωση παλινδρόμησης. Κατά την εφαρμογή της θεωρίας θα πρέπει να προχωρήσουμε προσεκτικά και να αποφύγουμε παράλογα αποτελέσματα αλλά ταυτόχρονα να είμαστε σε αρμονία με τις πρακτικές ανάγκες.

*3.3.2 Η διπλά κολοβή (doubly truncated) κατανομή Poisson*

Εκτός από γενικότητά της, αυτό μοντέλο είναι πιο ρεαλιστικό σε εφαρμογές δεδομένου ότι τα δεδομένα περιορίζονται για ένα διάστημα έτσι ώστε να μην παρατηρούνται δεδομένα εκτός του διαστήματος.

Λαμβάνοντας υπόψη την εξίσωση ….μπορεί να ξαναγραφτεί για t=1,…,N ως

$$f((y_t; \mu_t, c_1, c_2|F_{t-1}) = exp\{(y_t log\mu_t - log\psi(c_1, c_2, \mu_t) - \log(y_t)!\} \quad (3.11)$$

για $y_t = c_1, \ldots, c_2$ όπου $c_1$ και $c_2$ υποτίθεται είναι γνωστά. Κατά συνέπεια, η διπλά κατατετμημένη κατανομή Poisson είναι μέλος της εκθετικής οικογένειας κατανομών

όταν είναι γνωστά τα αποκομμένα σημεία, γεγονός που υποδηλώνει ότι το μοντέλο κανονικής σύνδεσης είναι λογάριθμικό και έτσι η αντίστροφη συνάρτηση σύνδεσης h(.) είναι εκθετική. Ως εκ τούτου, παίρνουμε το λογαριθμογραμμικό μοντέλο:

$$\mu_t(\beta) = \exp(Z'_{t-1}\beta)\,, t = 1, \ldots\ldots, N.$$

Από τις εξισώσεις (3.7) και (3.8), η $\mu_t$ δεν είναι ίση με την υπό συνθήκη αναμενόμενη τιμή της μεταβλητής απόκρισης και η δεσμευμένη μέση τιμή δεν χρειάζεται να είναι ίση με την υπό συνθήκη διασποράς το (κολοβό) μοντέλο απόκρισης.

*3.3.3 Το μοντέλο Zeger-Qaqish*

Οι Zeger και Qaqish (1988) εισάγουν το πολλαπλασιαστικό μοντέλο:

$$\mu_t(\beta) = \exp\left(\beta_0 + \beta_1 X_t + \beta_2 log\left(\tilde{Y}_{t-1}\right)\right) = \exp(\beta_0 + \beta_1 X_t)\left(\tilde{Y}_{t-1}\right)^{\beta_2} \qquad (3.12)$$

με $t = 1, \ldots, N$, χωρίς να καθοριστεί καμία υπόθεση κατανομής. Εδώ είναι $Z_{t-1} = (1, X_t, log\left(\tilde{Y}_{t-1}\right))'$, $\beta = (\beta_0, \beta_1, \beta_2)'$ και $\tilde{Y}_{t-1}$ ορίζεται

$$\tilde{Y}_{t-1} = Y_{t-1} + c\,, c > 0 \qquad (3.13)$$

έτσι ώστε $Y_{t-1} = 0$ να μην είναι μια κατάσταση απορρόφησης. Αυτές οι εμπειρικές διορθώσεις ίσως επηρεάσουν τους συντελεστές παλινδρόμησης. Εκτός από τον προσδιορισμό της ροπής πρώτης τάξης στην (3.12), θεωρείται ότι η υπό συνθήκη διασπορά της απόκρισης δίνεται από

$$Var[Y_t|F_{t-1}] = \varphi V(\mu_t), \qquad (3.14)$$

όπου $V(\mu_t)$ είναι η συνάρτηση διασποράς και $\varphi$ είναι μια άγνωστη παράμετρο διασποράς. Οι υποθέσεις γίνονται μόνο για την υπο συνθήκη της πρώτης και δεύτερης τάξης ροπής της απόκρισης. Το μοντέλο (3.12) μπορεί να διευρυνθεί από το ακόλουθο πολλαπλασιαστικό σφάλμα

$$\mu_t(\beta) = \exp(\beta_0 + \beta_1 X_t)\left(\frac{\tilde{Y}_{t-1}}{\exp(\beta_0 + \beta_1 X_t)}\right)^{\beta_2}, t = 1, .., N. \qquad (3.15)$$

Απ' την εξίσωση (3.15) συνεπάγεται ότι, όταν $\beta_2 < 0$ υπάρχει μια αντίστροφη σχέση μεταξύ των $\tilde{Y}_{t-1}$ και $\mu_t(\beta)$ και όταν $\beta_2 > 0$, $\mu_t(\beta)$ μεγαλώνει με $\tilde{Y}_{t-1}$. Όταν $\beta_2 = 0$, το μοντέλο (3.15) μειώνει το λογαριθμογραμμικό μοντέλο. Η αντικατάσταση του (3.13) στην εξίσωση (3.15) δίνει

$$\mu_t(\beta) = \exp(\beta_0 + \beta_1 X_t)\left(\frac{Y_{t-1} + c}{\exp(\beta_0 + \beta_1 X_t)}\right)^{\beta_2} \qquad (3.16)$$

για $c > 0$ και $t = 1, \dots, N$.

Η εξίσωση (3.15) μπορεί να γενικευθεί περαιτέρω από το ακόλουθο μοντέλο

$$\mu_t(\beta) = exp\left[X'_t\gamma + \sum_{i=1}^{q} \theta_i \left(log\tilde{Y}_{t-1} - X'_{t-1}\gamma\right)\right], t = 1, \dots, N, \qquad (3.17)$$

όπου $\beta = \left(\gamma', \theta_1, \dots, \theta_q\right)$ ένα διάνυσμα διάστασης (s+q) που περιέχει ένα το διάνυσμα συμμεταβλητών $\{X_t\},\ t = 1, \dots, N$ και $\{\tilde{Y}_{t-1}\},\ t = 1, \dots, N$ ορίζεται μέσω της (3.13). Για παράδειγμα, όταν s = 2, q =1, $\gamma = (\beta_0, \beta_1)'$ και $\theta_1 = \beta_2$, τότε η εξίσωση (3.17) περιορίζεται στη (3.15). Μοντέλα όπως το (3.12) και παραλλαγές αυτών υποκινούνται από τη θεωρία των εξαρτημένων από το μέγεθος διαδικασιών διακλάδωσης (βλέπε για παράδειγμα Guttorp, 1991). Για να δούμε αυτό, θεωρούμε το (3.12) και για απλότητα υποθέτουμε ότι $\beta_1 = 0$. Θέτουμε $Y_t$ τον αριθμό των μονάδων σε έναν πληθυσμό κατά τη διάρκεια της γενιάς *t*. Τότε:

$$Y_t = \sum_{j=1}^{Y_{t-1}} Z_j(Y_{t-1}), Y_{t-1} > 0$$

όπου $Z_j(Y_{t-1})$ είναι ο αριθμός των απογόνων του j-οστού ατόμου κατά τη χρόνική στιγμή t-1. Αν $Y_{t-1} = 0$, υποθέτουμε ότι γίνεται επανεκκίνηση αύξησης με $Z_0$ τα άτομα του πληθυσμού. Αν $Z_j(Y_{t-1})$ κατανέμεται σύμφωνα με την κατανομή Poisson με μέση τιμή

$$E\left[Z_j(Y_{t-1})|Y_{t-1}\right] = \left(\frac{\mu}{Y_{t-1}}\right)^{1-\beta_2}$$

και

$$E[Z_0] = \left(\frac{\mu}{c}\right)^{1-\beta_2}$$

με $\mu = exp(\beta_2)$ τότε φτάνουμε στο μοντέλο (3.12).

Έχουν προταθεί διάφορες γενικεύσεις του μοντέλου (3.17). Για παράδειγμα οι Albert et al. (1994) επεκτείνουν αυτό το μοντέλο ώστε να αντιπροσωπευτεί για μη γραμμικές παραμέτρους. Για εφαρμογές αυτού του μοντέλου βλέπετε Agresti και Natarajan (2001) και Li (1991).

### 3.4 Εκτίμηση μερικής πιθανοφάνειας για το μοντέλο της Poisson

Η συνάρτηση της μερικής πιθανοφάνειας δίνεται από:

$$L(\beta) = \prod_{t=1}^{N} f(y_t; \mu_t | F_{t-1}) = \prod_{t=1}^{N} \frac{\exp(-\mu_t(\beta))\mu_t(\beta)^{y_t}}{y_t}$$

$$l(\beta) = logL(\beta) = \sum_{t=1}^{N} y_t log\mu_t(\beta) - \sum_{t=1}^{N} \mu_t(\beta) - \sum_{t=1}^{N} \log(y_t!)$$

$$l(\beta) = \sum_{t=1}^{N} y_t log(h(Z'_{t-1}\beta)) - \sum_{t=1}^{N} h(Z'_{t-1}\beta) - \sum_{t=1}^{N} \log(y_t!)$$

παραγωγίζοντας τις μερικές συναρτήσεις score έχουμε:

$$S_N(\beta) = \nabla l(\beta) = \left(\frac{\vartheta l(\beta)}{\vartheta \beta_1}, \dots\dots\dots, \frac{\vartheta l(\beta)}{\vartheta \beta_p}\right)' = \sum_{t=1}^{N} Z_{t-1} \frac{\vartheta h(n_t)}{\vartheta n_t} \frac{1}{\sigma_t^2(\beta)} (Y_t - \mu_t(\beta))$$

όπου $n_t = Z'_{t-1}\beta$ και $\sigma_t^2(\beta) = Var(Y_t|F_{t-1})$.

Τα αποτελέσματα της διαδικασίας ορίζονται από τα μερικά αθροίσματα

$$S_t(\beta) = \sum_{s=1}^{t} Z_{s-1} \frac{\vartheta h(n_s)}{\vartheta n_s} \frac{1}{\sigma_s^2(\beta)} (Y_s - \mu_s(\beta))$$

η λύση της εξίσωσης: $S_N(\beta) = \nabla l(\beta) = logL(\beta) = 0$

θα δίνει την εκτιμήτρια της μέγιστης πιθανοφάνειας και συμβολίζεται με $\hat{\beta}$. Το σύστημα δεν είναι γραμμικό και επιλύεται με τη μέθοδο score του Fisher.

Η αθροιστική υπό συνθήκη του πίνακα πληροφορίας $G_N(\beta)$ ορίζεται:

$$G_N(\beta) = \sum_{t=1}^{N} Cov\left[ Z_{t-1} \frac{\vartheta h(n_t)}{\vartheta n_t} \frac{1}{\sigma_t^2(\beta)} \big(Y_t - \mu_t(\beta)\big) | F_{t-1} \right]$$

$$= \sum_{t=1}^{N} Z_{t-1} (\frac{\vartheta h(n_t)}{\vartheta n_t})^2 \frac{1}{\sigma_t^2(\beta)} Z'_{t-1}$$

και ο μη δεσμευμένος πίνακας πληροφορίας είναι:

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

Πολλές απλουστεύσεις μπορούν να γίνουν όταν χρησιμοποιούμε τη κανονική σύνδεση. Για το μοντέλο της Poisson με την συνάρτηση σύνδεσης έχουμε $\mu_t(\beta) = \exp(Z'_{t-1}\beta)$, και οι ιδιότητες μεταφέρονται με τα καινούργια δεδομένα ως:

$$S_N(\beta) = \sum_{t=1}^{N} (Z_{t-1} \, (Y_t - \mu_t(\beta))$$

$$G_N(\beta) = \sum_{t=1}^{N} Z_{t-1} \, Z'_{t-1} \sigma_t^2$$

στη περίπτωση της κανονικής σύνδεσης αν η εκτιμήτρια μέγιστης μερικής πιθανοφάνειας υπάρχει τότε αυτή είναι μοναδική.

### 3.5 Έλεγχοι υποθέσεων

**Ασυμπτωτική θεωρία**

Εδώ παρουσιάζουμε το θεώρημα που εξασφαλίζει την συνέπεια και την ασυμπτωτική κανονικότητα της εκτιμήτριας μέγιστης μερικής πιθανοφάνειας.

Θεώρημα

Αν μια τοπικά μοναδική εκτιμήτρια μέγιστης μερικής πιθανοφάνειας υπάρχει τότε συγκλίνει προς ένα αριθμό. Οπότε θα υπάρχει μια ακολουθία της εκτιμήτριας μεγιστης πιθανοφάνειας του $\hat{\beta}$ που είναι σταθερή και ασυμπτωτικά κανονική:

όσο $N \to \infty$, όπου $G^{-1}$ είναι ο αντίστροφος ασυμπτωτικός πίνακας πληροφορίας και είναι ένας p x p πίνακας.

$$\sqrt{N}\left(\hat{\beta} - \beta\right) \to N_p(0, G^{-1}(\beta))$$

$$G(\beta) = \int Z(\frac{\vartheta h(n)}{\vartheta n})^2 \frac{1}{h(n)} Z'V(dZ)$$

με $n = Z'\beta$ όσο $N \to \infty$ τότε θα είναι:

$$\frac{G_N(\beta)}{N} \to G(\beta)$$

**Διάστημα εμπιστοσύνης πρόβλεψης**

Θα δώσουμε ένα 100(1-α)% διάστημα για το $\mu_t$ στο μοντέλο της Poisson

$$\mu_t = \mu_t(\hat{\beta}) \pm Z_{a/2} \frac{|h'(Z'_{t-1}\beta)|}{\sqrt{N}} \sqrt{Z'_{t-1} G^{-1}(\beta) Z_{t-1}}$$

Όπου $h'$ είναι η παράγωγος του $h$ και $Z_{\alpha/2}$ είναι το άνω (α/2)100% ποσοστιαίο σημείο της τυποποιημένης Κανονικής κατανομής.

**Έλεγχος υποθέσεων**

Θεωρούμε τις υποθέσεις

$$H_0: C\beta = \beta_0 \text{ ,έναντι}$$

$$H_1: C\beta \neq \beta_0$$

όπου $C$ είναι ο κατάλληλος πίνακας με πλήρη τάξη με r≤p. Συμβολίζουμε με $\tilde{\beta}$ τη περιορισμένη (restricted) εκτιμήτρια μερικής πιθανοφάνειας κάτω από τον έλεγχο της παραπάνω υπόθεσης. Ακολουθώντας τη γενική θεωρία οι πιο γνωστοί στατιστικοί έλεγχοι στο πλαίσιο της Poisson Παλινδρόμησης είναι:

1) Ο λόγος μερικών πιθανοφανειών

$$\lambda_N = 2\{pl(\hat{\beta}) - pl(\tilde{\beta})\}$$

όπου $pl(\hat{\beta})$ είναι η μεγιστοποιημένη συνάρτηση μερικής πιθανοφάνειας.

2) Ο έλεγχος Wald

$$W_N = (C\hat{\beta} - \beta_0)'\left(CG^{-1}(\tilde{\beta})C'\right)^{-1}(C\hat{\beta} - \beta_0)$$

3) Ο στατιστικός έλεγχος με τη συνάρτηση score

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

**<u>Θεώρημα</u>**

Οι παραπάνω στατιστικοί έλεγχοι είναι ασυμπτωτικά ισοδύναμοι. Σύμφωνα λοιπόν με το παραπάνω έλεγχο υποθέσεων η ασυμπτωτική τους κατανομή είναι η $X^2$ με ν βαθμούς ελευθερίας. Όπου $\nu = N - p_0$ με $p_0$ να είναι το πλήθος των παραμέτρων υπό την $H_0$.

### 3.6.Καλή προσαρμογή

*3.6.1 Η ελεγχοσυνάρτηση Deviance*

Για απαριθμητή χρονοσειρά $\{Y_t\}$ υπό το λογαριθμογραμμικό μοντέλο της Poisson η τυποποιημένη συνάρτηση της deviance παίρνει τη μορφή

$$D = 2\sum_{t=1}^{N}\left\{Y_t \log\left(\frac{Y_t}{\hat{\mu}_t}\right) - (Y_t - \hat{\mu}_t)\right\}$$

και ο έλεγχος Pearson γίνεται

$$X^2 = \sum_{t=1}^{N}\sum_{t=1}^{N}\frac{(Y_t - \hat{\mu}_t)^2}{V(\hat{\mu}_t)}$$

με $\hat{\mu}_t = \mu_t(\hat{\beta}), t = 1, \ldots . N.$

Υπό τις κατάλληλες κανονικές συνθήκες(regularity conditions) η ασυμπτωτική κατανομή των δύο παραπάνω ελέγχων προσεγγίζεται με τη $X^2$με N-p βαθμούς ελευθερίας.

*3.6.2 Υπόλοιπα*

Τα υπόλοιπα στο πλαίσιο της Poisson Παλινδρόμησης για απαριθμητές χρονοσειρές ορίζονται ως:

$$\hat{e}_t = Y_t - \hat{\mu}_t, t = 1, \ldots . , N$$

και τα υπόλοιπα Pearson ως:

$$\hat{r}_t = \frac{Y_t - \hat{\mu}_t}{\sqrt{V(\hat{\mu}_t)}} , t = 1, \ldots \ldots , N$$

the working residuals

$$w\hat{r}_t = \frac{Y_t - \hat{\mu}_t}{\partial \mu_t / \partial n_t} , t = 1, \ldots \ldots , N$$

όπου το $\partial \mu_t / \partial n_t$ υπολογίζεται στο $\hat{\beta}$ και η ελεγχοσυνάρτηση deviance για τα υπόλοιπα δίνεται από το παρακάτω τύπο:

$$\hat{d}_t = sign(Y_t - \hat{\mu}_t)\sqrt{2[l_t(Y_t) - l_t(\hat{\mu}_t)]} , t = 1, \ldots \ldots , N$$

## 3.7 Ανάλυση Παρεμβάσεων (Intervention analysis)

*Εισαγωγή*

Οι απαριθμητές χρονοσειρές μετρούνται σε διάφορους τομείς, όταν ένας αριθμός γεγονότων μετρήθηκε κατά τη διάρκεια ορισμένων χρονικών περιόδων. Παραδείγματα υπάρχουν πολλά όπως ο αριθμός τροχαίων ατυχημάτων σε μια περιοχή, ο εβδομαδιαίος αριθμός νέων περιπτώσεων επιδημιολογίας ή ο αριθμός άφιξης των φωτονίων ανά ms (microsecond) σ ’ένα βιολογικό πείραμα. Μια φυσική

παραλλαγή του δημοφιλή αυτοπαλινδρομικού κινητού μέσου όρου ARMA μοντέλου για συνεχείς μεταβλητές βασίζεται στην υπόθεση ότι η παρατήρηση $\{Y_t\}$ κατά τη στιγμή t παράγεται από ένα Γενικευμένο Γραμμικό Μοντέλο δοθέντων των παρελθοντικών τιμών. Οπότε επιλέγουμε μια κατάλληλη κατανομή για απαριθμητά δεδομένα σαν τη Poisson και μια συνάρτηση σύνδεσης g(). Αυτή η προσέγγιση χρονοσειρών ακολουθώντας ένα Γενικευμένο Γραμμικό Μοντέλο επιδιώκεται από τους Kedem και Fokiano (2002). Επικεντρωνόμαστε στις πρώτης τάξης μοντέλα, θεωρούμε τη χρονοσειρά $\big((Y_t): t \epsilon \mathbb{N}_0\big)$ που ακολουθεί το μοντέλο της Poisson

$$Y_t \mid F_{t-1}^{Y} \sim Po(\mu_t)$$

$$g(\mu_t) = \beta_0 + \beta_1 g(Y_{t-1} + c) + \gamma_1 g(\mu_{t-1}), t \geq 1 \qquad (3.18)$$

όπου $F_{t-1}^{Y}$ σημαίνει ότι η σ-άλγεβρα δημιουργείται από $\{Y_t, \dots \dots, Y_0, \mu_0\}$, ακόμα $\beta_0, \beta_1, \beta_2$ είναι άγνωστοι παράμετροι και $c$ είναι μια γνωστή σταθερά. Μοντέλα που απασχολούν άλλες κατανομές όπως η αρνητική διωνυμική μπορούν να αντιμετωπιστούν παρόμοια.

Η φυσική επιλογή για τη $g$ είναι ένας αλγόριθμος και αυτός είναι και ο λόγος προσθήκης της σταθεράς c στο $Y_{t-1}$ στον όρο $g(Y_{t-1} + c)$ εφόσον θα πρέπει να αποφευχθούν οι δυσκολίες που προκύπτουν από παρατηρήσεις που είναι ίσες με μηδέν. Ακολουθώντας τους Fokiano και Tjostheim (2011) οι οποίοι ανέπτυξαν συνθήκες εργοδικότητας για μια υποκατηγορία που προκύπτει από το λογάριθμογραμμικό μοντέλο αν θέσουμε c=1. Θα αναφέρουμε τις πιθανές ερμηνείες του μοντέλου όπως αναφέρονται στη (3.18) όπου με $Y_t$ είναι ο αριθμός των περιπτώσεων που παρατηρήθηκαν στο χρόνο $t$. Για σταθερού μεγέθους πληθυσμό η δεσμευμένη μέση τιμή μετρά το κίνδυνο ενός ατόμου να αρρωστήσει τη χρονική στιγμή $t$. Το μοντέλο μας προϋποθέτει ότι όλες οι επιδράσεις της $\mu_t$ είναι γραμμικές αφού «μεταφερθούν» κατάλληλα από τη συνάρτηση σύνδεσης $g$. Ο όρος $g(Y_{t-1} + c)$ στη δεύτερη εξίσωση του πλαισίου 1 εξαρτάται από τη συνάρτηση σύνδεσης $g(\mu_t)$ όπως και οι παρατηρήσεις $Y_t$, $Y_{t-1}$ και $\beta_1$ η μετράει τη δύναμη αυτής της εξάρτησης. Ένας μεγάλος αριθμός περιπτώσεων $Y_{t-1}$ τη χρονική στιγμή $t-1$ μπορεί να προκαλέσει ένα μεγάλο αριθμό περιπτώσεων κατά τη χρινική στιγμή $t$ επειδή ο κίνδυνος λοίμωξης αυξάνεται. Ακόμα η $g(\mu_{t-1})$ περιγράφει την ύπαρξη περιόδων αυξανόμενου κινδύνου όπως για παράδειγμα λόγω καιρικών συνθηκών και η $\gamma_1$

μετράει το μέγεθος τέτοιων εξαρτήσεων. Λαμβάνοντας υπόψη το μοντέλο (3.18) ένα βασικό ερώτημα είναι αν περιγράφει σωστά όλες τις παρατηρήσεις μιας δοσμένης χρονοσειράς ή αν κατά πόσον ορισμένες παρατηρήσεις έχουν επηρεασθεί από ασυνήθιστες επιδράσεις οι οποίες καλούνται παρεμβάσεις (intervention) (Fried et al. 2014).

**Μοντέλα για την ανάλυση παρεμβάσεων (intervention analysis models)**

Η πιθανότητα να εισάγουμε μια ασυνήθιστη επίδραση στη χρονοσειρά $Y_t$ που παράγεται από το μοντέλο (3.18) είναι η υπόθεση ότι απ' το χρονικό σημείο τ η μέση δεσμευμένη διαδικασία αλλάζει με τη προσθήκη του όρου $\omega\delta^{t-\tau}I(t \geq \tau)$ στη $g(\mu_t)$ ώστε αντί για το $(Y_t)$ εμείς παρατηρούμε τη διαδικασία $(Z_t)$ που παράγεται από το εξής μοντέλο:

$$Z_t \mid F_{t-1}^{Z} \sim Po(\mu_t^c)$$

$$g(\mu_t^c) = \beta_0 + \beta_1 g(Z_{t-1} + c) + \gamma_1 g(\mu_{t-1}^c) + \omega\delta^{t-\tau}I(t \geq \tau), t \geq 1, \quad (3.19)$$

όπου $\mu_t^c$ η διαδικασία μέσης «μόλυνσης» (contamination) η οποία συμπίπτει με $\mu_t$ μέχρι τη χρονική στιγμή τ-1 και μετά επιρεάζεται απο την παρέμβαση. Το $F_{t-1}^{Z}$ υποδηλώνει την σ-άλγεβρα που δίνει τις πληροφορίες σχετικά με το παρελθόν της «μολυσμένης» διαδικασίας και τις αρχικές τιμές που είναι ανάλογες της $F_{t-1}^{Z}$. Η νέα παράμετρος ω καθορίζει το μέγεθος της επίδρασης, αυτή η ποσότητα $I(t \geq \tau)$ εκφράζει αν $t \geq \tau$ ή όχι. Επίσης η δ ϵ[0,1] καθορίζει αν η επίδραση γίνεται στο χρόνο τ όπου δ=0 προκαλώντας ακραίες τιμές, είτε όλο το επίπεδο μετατοπίζεται από τη χρονική στιγμή τ όπου δ=1, είτε μια γεωμετρική φθίνουσα μετατόπιση να συμβαίνει όταν δ ϵ(0,1). Να σημειώσουμε ότι και στη περίπτωση όπου δ=0 ολόκληρη η μελλοντική διαδικασία επηρεάζεται από μια παράμβαση (intervention), η επίδρασηαυτή γίνεται μέσω των $Z_t$ και $g(\mu_t^c), t \geq \tau$. Για παράδειγμα στην επιδημιολογία, μια παρέμβαση σύμφωνα με το (3.19) μπορεί να ερμηνευθεί ως μια εσωτερική αλλαγή της διαδικασίας παραγωγής των δεδομένων. Για κάποιο λόγο, π.χ. λόγω των ιδιαίτερων καιρικών συνθηκών ή άλλες εκθέσεις, η δεσμευμένη μέση διαδικασία του κινδύνου αλλαζει με απρόβλεπτο τρόπο κατά τη χρονική στιγμή τ. Οπότε οι παρατηρήσεις αλλάζουν σ' αυτό το χρονικό σημείο καθώς και στη συνέχεια.

Οι Liboschik et al. (2013) διερευνούν ένα άλλο μοντέλο παραμβάσεων (intervention) στην περίπτωση της ταυτοτικής σύνδεσης. Στη προσέγγιση τους, μια παρέμβαση επηρεάζει την παρατήρηση στο χρόνο τ αλλά όχι την υποβόσκουσα υπό συνθήκη μέση τιμή. Αυτό μπορεί να γίνει κατανοητό με μια εξωτερική αλλαγή, καθώς η παρατήρηση μόλυνσης $Z_\tau$ ισούται με το άθροισμα των μολυσμένων τιμών $Y_\tau$ συν ένα τυχαίο αριθμό $C_\tau$, η οποία προκύπτει εξ αιτίας κάποιων ασυνήθιστων αιτιών και εισέρχεται στη δυναμική της διαδικασίας κατά τον ίδιο τρόπο όπως η $Y_\tau$, ενώ ο βασικός κίνδυνος της $\mu_\tau$ αρχικά δεν επηρεάζεται. Ένα παράδειγμα θα μπορούσε να είναι οι άνθρωποι που έχουν μολυνθεί λόγω εξωτερικών αιτίων, π.χ. σε ταξίδι. Το τροποποιημένο μοντέλο παρέμβασης με μια γενική συνάρτηση σύνδεσης $g$ γράφεται ως

$$Z_t \mid F_{t-1}^Z \sim Po(\mu_t^c)$$

$$g(\mu_t^c) = g(\mu_t) + \omega\delta^{t-\tau} I(t \geq \tau)$$

$$g(\mu_t) = \beta_0 + \beta_1 g(Z_{t-1} + c) + \gamma_1 g(\mu_{t-1}), t \geq 1 \qquad (3.20)$$

Οι δύο τελευταίες εξισώσεις περιγράφουν τη δεσμευμένη μέση διαδικασία που μπορεί να συνοψιστεί ως

$$g(\mu_t^c) = \beta_0 + \beta_1 g(Z_{t-1} + c) + \gamma_1 g(\mu_{t-1}) - \omega\delta^{t-1+\tau} I(t-1 \geq \tau) + \omega\delta^{t-\tau} I(t \geq \tau)$$

έτσι φαίνεται με σαφήνεια η διαφορά με το μοντέλο (3.19).

Αν το χρονικό σημείο τ και το είδος μιας παρέμβασης, για παράδειγμα η τιμή του δ, είναι και τα δύο γνωστά τότε ένα μοντέλο παρέμβασης, όπως διατυπώνεται στη σχέση (3.19) ή (3.20) μπορεί να προσαρμοστεί μεγιστοποιώντας την υπό συνθήκη της πιθανοφάνειας επαναληπτικά ξεκινώντας με κατάλληλες αρχικές τιμές. Η ύπαρξη μιας τέτοιας γνωστής παρέμβασης μπορεί να επιβεβαιωθεί συγκρίνοντας το αντίστοιχο αποτέλεσμα του στατιστικού ελέγχου με τα άνω ποσοστιαία της ασυμπτωτικής $X_1^2$ κατανομής, όπως περιγράφεται παραπάνω. Αν το χρονικό σημείο τ είναι άγνωστο, αλλά ο τύπος είναι γνωστός, τα πειράματα προσομοίωσης δείχνουν ότι παραμετρικές διαδικασίες bootstrap δουλεύουν αρκετά καλά: προσαρμόζεται το μοντέλο χωρίς παρεμβάσεις και υπολογίζεται η τιμή των στατιστικών ελέγχων για όλα τα χρονικά σημεία. Στη συνέχεια χρησιμοποιείται η μέγιστη τιμή όλων των ελεγχοσυναρτήσεων όλων των χρονικών σημείων ως ο τελικός στατιστικός έλεγχος.

Έπειτα παράγουν τεχνητές χρονοσειρές χωρίς παρεμβάσεις από το προσαρμοσμένο μοντέλο και υπολογίζουν την αντίστοιχη μεγαλύτερη τιμή του στατιστικού ελέγχου. Επιλέγουν μια παρέμβαση στο συγκεκριμένο χρονικό σημείο που μεγιστοποιεί το αποτέλεσμα του στατιστικού ελέγχου για τα πραγματικά δεδομένα, αν είναι ανάμεσα στο μεγαλύτερο 100α ποσοστιαίο σημείο όλων των μέγιστων στατιστικών ελέγχων. Αν το είδος της παρέμβασης είναι επίσης άγνωστο τότε οι μέγιστες τιμές των στατιστικών ελέγχων μπορούν να υπολογιστούν για κάθε ένα από τα παραπάνω είδη μοντέλων (3.19) ή (3.20). Οι προσομοιώσεις δείχνουν ότι θα πρέπει να προτιμούνται οι μετατοπίσεις επιπέδων (δ=1) αν αποδειχθούν ότι είναι σημαντικές μιας και μια μετατόπιση επιπέδου συνήθως προκαλεί επιδράσεις στους στατιστικούς ελέγχους άλλων τύπων παρέμβασης, ενώ η αντίστροφη επίδραση τονίζεται λιγότερο. Πολλαπλές παρεμβάσεις μπορούν να αντιμετωπιστούν υπολογίζοντας την επίδραση μιας παρέμβασης και αφαιρώντας την από τις χρονοσειρές, προτού τα δεδομένα αναλυθούν λαμβάνοντας υπόψη επιπλέον παρεμβάσεων.

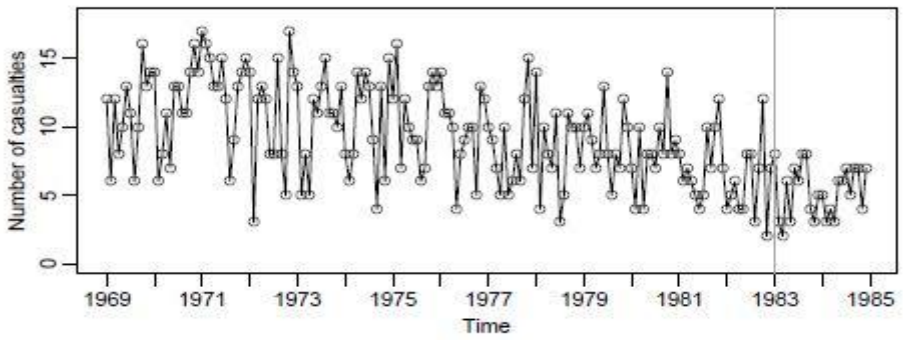
Σημειώνεται ότι τα παραπάνω μοντέλα παρέμβασης δε μπορούν να περιγράψουν προσθετικές ακραίες ή παρεκκλίνουσες τιμές (additive outliers) που αντιπροσωπεύουν για παράδειγμα τη μέτρηση λαθών, όπως στην περίπτωση που μία παρατήρηση έχει αλλάξει χωρίς καμία επίδραση στο μελλοντικό αποτέλεσμα της διαδικασίας. Στην πραγματικότητα τέτοιες προσθετικές ακραίες τιμές (additive outliers) είναι δύσκολο να αντιμετωπιστούν μέσω κλασικών τεχνικών (frequentist approach), αφού θα χρειαζόταν να εξετάσουμε την ύπαρξη τους με υπό-συνθήκη την μη-παρατηρήσιμη τιμή $Y_\tau$ αντί της «μολισμένης» $Z_\tau$. Οι Fried et al (2013) ανέπτυξαν μια μπεϋζιανή προσέγγιση για επιπρόσθετες ακραίες τιμές, εφαρμόζοντας τεχνικές με μαρκοβιανές αλυσίδες. Τα δικά τους αποτελέσματα από την προσομοίωση παρέχουν στοιχεία που μπορούν να αντιμετωπίσουν προσθετικές ακραίες τιμές (additive outliers) αν υπάρχουν αρκετές από αυτές. Μία ή και ελάχιστες προσθετικές ακραίες τιμές θέτουν δυσκολίες σε μια μπεϋζιανή προσέγγιση βασισμένη σε πρότερες κατανομές λιγοστών πληροφοριών, αφού αυτές δεν παρέχουν αρκετές πληροφορίες για τη συγκεκριμένη συνιστώσα της υποβόσκουσας μεικτής κατανομής που προκαλεί τις ακραίες τιμές.

Να σημειωθεί επίσης ότι υποθέτουμε έμμεσα τις επιδράσεις των μοντέλων παραμβάσεων να είναι προσθετικές όταν χρησιμοποιούμε τη ταυτοτική σύνδεση και πολλαπλασιαστικές όταν χρησιμοποιούμε τη λογαριθμική σύνδεση, αφού για λόγους

ευκολίας εισάγουμε τις επιδράσεις των παραμβάσεων με τον ίδιο τρόπο που εισάγουμε τις εξαρτήσεις από το παρελθόν. Ακόμα μία έμμεση υπόθεση που κάνουμε για τα παραπάνω μοντέλα παραμβάσεων, τις κοινές ακραίες τιμές (outliers) και τα μοντέλα παραμβάσεων που έχουν προταθεί ως ARMA-διαδικασίες στη βιβλιογραφία, είναι ότι η δυναμική της διαδικασίας παραμένει ίδια και ακολουθεί το ίδιο μοντέλο πριν και μετά την παρέμβαση (βλ. και Caroni και Karioti, 2004).

# ΚΕΦΑΛΑΙΟ IV
# ΕΦΑΡΜΟΓΗ-ΟΔΙΚΑ ΑΤΥΧΗΜΑΤΑ ΣΤΗ Μ.ΒΡΕΤΑΝΙΑ

Εδώ θα μελετήσουμε το μηνιαίο αριθμό θανάτων των οδηγών ελαφρών οχημάτων στη Μ.Βρετανία ανάμεσα στον Ιανουάριο 1969 και το Δεκέμβριο 1984 σύμφωνα με τους Liboschik et al. (2016) (βλέπε Σχήμα 4.1). Αυτή η χρονοσειρά είναι μέρος του συνόλου δεδομένων το οποίο πρώτα θεωρήθηκε από τους Harvey και Durbin (1986) για το αποτέλεσμα της μελέτης της υποχρεωτικής χρήσης της ζώνης ασφαλείας που εισήχθη στις 31 Ιανουαρίου 1983. Το σύνολο δεδομένων συμπεριλαμβανομένων και των επιπλέον συμμεταβλητών είναι διαθέσιμο στην R στο αρχείο Seatbelts. Στο έγγραφό τους οι Harvey και Durbin αναλύουν τους αριθμούς των θυμάτων για τους οδηγούς και τους επιβάτες των αυτοκινήτων οι οποίοι είναι τόσο μεγάλοι και μπορούν να αντιμετωπιστούν με μεθόδους για συνεχή δεδομένα. Ο μηνιαίος αριθμός νεκρών οδηγών από βαριά οχήματα (φορτηγά κλπ) είναι πολύ μικρότερος και έτσι προτιμάται η μέθοδος για τα απαριθμητά δεδομένα. Για την επιλογή μοντέλου χρησιμοποιούμε μόνο τα στοιχεία μέχρι το Δεκέμβριο του 1981. Επιλέγουμε το λογαριθμογραμμικό μοντέλο με τη λογαριθμική συνάρτηση σύνδεσης επειδή επιτρέπει τις αρνητικές επιδράσεις των συμμεταβλητών. Στοχεύουμε στο να εξετάσουμε τη σειριακή εξάρτηση μικρού εύρους από ένα αυτοπαλινδρομικό μοντέλο πρώτης τάξης (ή πρώτου βαθμού) και την ετήσια εποχικότητα από τη 12 σειρά αυτοπαλινδρόμικών όρων. Και οι δύο αυτοί όροι δηλώθηκαν από το στοιχείο της λίστας που ονομάζεται past_obs του argument model.




**Σχήμα 4.1**: Μηνιαίος αριθμός νεκρών οδηγών βαρέων οχημάτων στη Μεγάλη Βρετανία. Η εισαγωγή της υποχρεωτική χρήσης της ζώνης ασφαλείας στις 31 Ιανουαρίου 1983 χαρακτηρίζεται από μια κάθετη γραμμή.

Ακολουθώντας τους Harvey και Durbin (1986), οι R.Fried et al. (2014) χρησιμοποιούν τη πραγματική τιμή της βενζίνης ως επεξηγηματική μεταβλητή καθώς και μια μεταβλητή για τη γραμμική τάση (linear trend). Θεωρούμε ότι η κατανομή Poisson είναι ένα μοντέλο επαρκές για τα δεδομένα αυτά. Το μοντέλο που έχει προσαρμοστεί καλείται seatbeltsfit και με τις παρακάτω εντολές στην R θα πάρουμε τα Αποτελέσματα 4.1.

```
> library("tscount")
> timeseries <- Seatbelts[, "VanKilled"]
> regressors <- cbind(PetrolPrice = Seatbelts[, c("PetrolPrice")],
+ linearTrend = seq(along = timeseries)/12)
> timeseries_until1981 <- window(timeseries, end = 1981 + 11/12)
> regressors_until1981 <- window(regressors, end = 1981 + 11/12)
> seatbeltsfit <- tsglm(timeseries_until1981,
+ model = list(past_obs = c(1, 12)), link = "log", distr = "poisson",
+ xreg = regressors_until1981)
> summary(seatbeltsfit, B = 500)
```

Με βάση τα Απότελέσματα 4.1 παίρνουμε πληροφορίες για την εκτίμηση των παραμέτρων των μεταβλητών, τα αντίστοιχα τυπικά τους σφάλματα, για τους δέικτες καλής προσαρμογής AIC, BIC, QIC καθώς και τη τιμή της μεγιστοποιημένης συνάρτησης πιθανοφάνειας.

Αποτελέσματα 4.1

Call:
tsglm(ts = timeseries_until1981, model = list(past_obs = c(1,
12)), xreg = regressors_until1981, link = "log", distr = "poisson")

Coefficients:

| | Estimate | Std.Error | CI(lower) | CI(upper) |
|---|---|---|---|---|
| (Intercept) | 1.9037 | 0.4321 | 1.3159 | 2.8570 |
| beta_1 | 0.0856 | 0.0990 | -0.1079 | 0.2313 |
| beta_12 | 0.1509 | 0.0828 | -0.0413 | 0.2778 |
| PetrolPrice | 0.0826 | 2.8072 | -4.8773 | 5.5631 |
| linearTrend | -0.0291 | 0.0341 | -0.0502 | -0.0126 |

Standard errors and confidence intervals (level = 95 %) obtained
by parametric bootstrap with 500 replications.

Link function: log
Distribution family: poisson
Number of coefficients: 5
Log-likelihood: -396.3849
AIC: 802.7697
BIC: 818.019
QIC: 802.7697

Το προσαρμοσμένο μοντέλο για τον αριθμό των οδηγών ελαφρών οχημάτων $(Y_t)$ που σκοτώθηκαν το μήνα t δίνεται από το παρακάτω τύπο:

$$\mathbf{\log(\mu_t) = 1.9 + 0.09Y_{t-1} + 0.15Y_{t-12} + 0.08X_t - \frac{0.03t}{12}, \quad t=1,\ldots\ldots,156}$$

όπου Xt υποδηλώνει την πραγματική τιμή της βενζίνης κατά τον χρόνο t. Ο εκτιμημένος συντελεστής beta_1=0.09 αντιστοιχεί στη πρώτη τάξη του συντελεστή αυτοσυσχέτισης και είναι πολύ μικρή, ακόμα και ελαφρώς κάτω από το μέγεθος του

προσεγγιστικού τυπικού σφάλματος, υποδεικνύοντας ότι δεν υπάρχει καμία αξιοσημείωτη εξάρτηση από τον αριθμό νεκρών οδηγών φορτηγών του προηγούμενου μήνα. Μια εποχιακή επίδραση συλλαμβάνεται από τη δωδέκατη τάξη του συντελεστή αυτοσυσχέτισης beta_12. Ωστόσο σε αντίθεση με το μοντέλο για τους οδηγούς αυτοκινήτων από Harvey και Durbin (1986), η τιμή της βενζίνης δεν φαίνεται να επηρεάζει τον αριθμό οδηγών που σκοτώθηκαν με φορτηγά. Μια εξήγηση μπορεί να είναι ότι τα φορτηγά χρησιμοποιούνται πολύ πιο συχνά για εμπορικούς σκοπούς από τα αυτοκίνητα και ότι η εμπορική κίνηση επηρεάζεται λιγότερο από την τιμή των καυσίμων. Η γραμμική τάση μπορεί να ερμηνευθεί ως η ετήσια μείωση του αριθμού των θυμάτων κατά ένα συντελεστή exp(-0.031)=0.969, δηλαδή κατά μέσο όρο, αναμένουμε περίπου 3% λιγότερους σκοτωμούς οδηγών φορτηγών ανά έτος (το οποίο είναι κάτω από το ένα σε απόλυτους αριθμούς).

Όπως αναφέρουν οι R.Fried et.al (2014), βάση του προσαρμοσμένου μοντέλου στα δεδομένα εκπαίδευσης (training data) μέχρι τον Δεκέμβριο του 1981, μπορούμε να προβλέψουμε τον αριθμό των οδικών ατυχημάτων το 1982 δοσμένης της αντίστοιχης τιμής της βενζίνης. Μία γραφική αναπαράσταση με τις ακόλουθες προβλέψεις δίνεται στο Σχήμα 4.2.

```
> timeseries_1982 <- window(timeseries, start = 1982, end = 1982 + 11/12)
> regressors_1982 <- window(regressors, start = 1982, end = 1982 + 11/12)
> predict(seatbeltsfit, n.ahead = 12, level = 0.9, global = TRUE,
+ B = 2000, newxreg = regressors_1982)$pred
          Jan      Feb      Mar      Apr      May      Jun      Jul
Aug
1982 7.707988 7.454226 7.568846 7.409922 7.210433 6.985772 7.145880 7.826338
          Sep      Oct      Nov      Dec
1982 7.493486 7.816908 8.022388 7.451928
```

Τέλος με την παρακάτω εντολή θα ελέγξουμε αν υπήρχε μια απότομη μετατόπιση του αριθμού των θυμάτων που παρουσιάστηκαν όταν εισήχθη η υποχρεωτική χρήση της ζώνης ασφαλείας στις 31 Ιανουαρίου 1983. Δηλαδή θέλουμε να ελέγξουμε αν υπάρχει σημαντική παρέμβαση (intervention).

Αποτελέσματα 4.2

**> interv_test(seatbeltsfit_alldata, tau = 170, delta = 1, est_interv = TRUE)**

Score test on intervention(s) of given type at given time

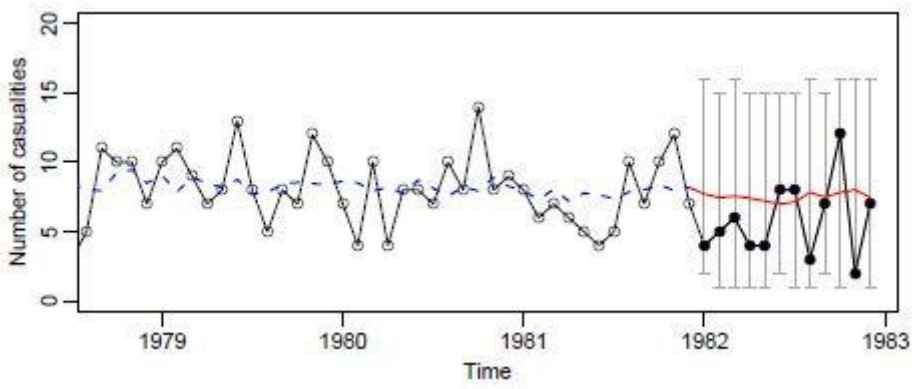
Chisq-Statistic: 1.152953 on 1 degree(s) of freedom, p-value: 0.2829319

Fitted model with the specified intervention:

Call:
tsglm(ts = fit$ts, model = model_extended, xreg = xreg_extended,
link = fit$link, distr = fit$distr)

Coefficients:

| (Intercept) | beta_1 | beta_12 | PetrolPrice | linearTrend | interv_1 |
|---|---|---|---|---|---|
| 0.19508 | 0.08819 | 0.80446 | 3.17408 | -0.04788 | 0.24570 |




**Σχήμα 4.2**: Προσαρμοσμένες τιμές (μπλε διακεκομμένη γραμμή) και προβλεπόμενες τιμές (κόκκινη συνεχής γραμμή), σύμφωνα με το μοντέλο με την κατανομή Poisson. Τα διαστήματα πρόβλεψης (γκρι ράβδοι) έχουν σχεδιαστεί για να εξασφαλίσουν ένα συνολικό ποσοστό κάλυψης 90%. Επιλέγονται για να έχουν ελάχιστο μήκος και βασίζονται σε μια προσομοίωση με 2000 επαναλήψεις.

Με βάση τα Αποτελέσματα 4.2 παίρνουμε πληροφορίες για: 1) Η p-value είναι ίση περίπου με 0.28 το οποίο σημαίνει ότι σε επίπεδο σημαντικότητας 5% δεν μπορούμε να απορρίψουμε τη μηδενική μας υπόθεση ότι δεν υπάρχει παρέμβαση (intervention) (βλέπε παράγραφο 3.7). Σημειώνουμε ότι αυτό το αποτέλεσμα δεν αποκλείει ότι υπάρχει επίδραση του νόμου των ζωνών ασφαλείας που είναι είτε μικρή για να είναι σημαντική ή διαφορετικού τύπου απ' αυτόν που έχει ελεγθεί. Εδώ προσαρμόσαμε το μοντέλο κάτω από την εναλλακτική υπόθεση της μετατόπισης του επιπέδου μετά την εισαγωγή του νόμου για τις ζώνες ασφαλείας. 2) Το πολλαπλασιαστικό μέγεθος της επίδρασης της παρέμβασης είναι $\exp(0.24570) \cong 1.279$. Αυτό δείχνει ότι, σύμφωνα με το προσαρμοσμένο μοντέλο έχουμε -27,9% λιγότεροι οδηγοί βαρέων οχημάτων που σκοτώθηκαν μετά την επιβολή του νόμου. Για λόγους σύγκρισης, οι Harvey και Durbin (1986) εκτιμούν μείωση 18% του αριθμού των νεκρών οδηγών των αυτοκινήτων.

# ΒΙΒΛΙΟΓΡΑΦΙΑ


1. Π. Οικονόμου και Χ. Καρώνη (2010). *Στατιστικά Μοντέλα Παλινδρόμησης*, Εκδόσεις Συμεών.
2. A. Agresti and Natarajan (2001). "Modeling clustered ordered categorical data: A survey." *International Statistical Revie,* **69,** 345-371.
3. P. S. Albert, H. F. McFarland, M. E. Smith, and J .A. Frank (1994). "Time series for modelling counts from a relapsing-remitting disease: Application to modeling disease activity in multiple sclerosis." *Statistics in Medicine,* **13**, 453-466.
4. N. E. Breslow and N. E. Day (1987). "Statistical Methods in Cancer Research." *The Design and Analysis of Cohort Studies*, **2**, Lyon: International Agency for Research on Cancer.
5. S. Bisgaard and M. Kulahci (2005). "Interpretation of time series models." *Qual. Eng.*, **17,** 653-658.
6. P. J. Brockwell and R. A. Davis (2002). *Introduction to Time Series and Forecasting* (2nd edition). Springer, New York.
7. A. C. Cameron and L. Leon (1993). "Markov regression models for time series data." Presented at Western Economics Association Meeting, Lake Tahoe, NV.
8. C. Caroni and V. Karioti (2004). "Detecting an innovative outlier in a set of time series". *Computational Statistics and Data Analysis,* **46**, 561-570.
9. A. J. Dobson and A. G. Barnett (2008). *An Introduction to Generalized Linear Models*. (3rd edition), Chapman and Hall, Boca Raton, Florida.
10. J. Durbin and G.S Watson (1950). "Testing for serial correlation in least squares regression I." *Biometrica,* **37,** 409-438.
11. J. Durbin and G.S Watson (1951). "Testing for serial correlation in least squares regression II." *Biometrica,* **38,** 159-178.
12. J. Durbin and G.S Watson (1971). "Testing for serial correlation in least squares regression III." *Biometrica*, **58,** 1-19.
13. B. Efron (1986)."Double exponential families and their use in generalized linear regression."*Journal of the Ainericarz Statistical Association,* **81,** 709-721.

14. K. Fokianos (2001). “Truncated Poisson regression for time series of counts.” *Scandinavian Journal of Statistics,* **28**, 645-659.
15. K. Fokianos, D. Tjostheim. (2011).” Log-linear Poisson autoregression.” *Journal of Multivariate Analysis*, **102**, 563-578.
16. P. Guttorp (1991). “Statistical Inference for Branching Processes.” Wiley, New York.
17. R. Fried , T.Liboschik, H. Elsaied, S. Kitromilidou and K. Fokianos (2014). “On outliers and interventions in Count Time Series following GLMs.” *Austrian Journal of Statistics,* **43**, 181-193 (http://www.ajs.or.at/)
18. A. C. Harvey, J. Durbin J (1986). “The effects of seat belt legislation on british road casualties: A case study in structural time series modeling.*”* *Journal of the Royal Statistical* Society A, **149**, 187-227.
19. R. T. Holden (1987). “Time series analysis of contagious process.”*Journal of the American Statistical Association,* **82**, 1019-1026.
20. J. E. Hutton and P.I. Nelson (1986). “Quasi-likelihood estimation for semimartingales.” *Stochastic Processes and Their Applications,* **22**, 245-257.
21. N. L. Johnson, S. Kotz, and A. W. Kemp (1992). *Univariate Discrete Distributions.* (2nd edition). Wiley, New York.
22. V. Karioti and C. Caroni (2006). “Properties of the GAR(1) model for time series of counts”. *Journal of Modern Applied Statistical Methods*, **5**, 140-151.
23. L. F. Leon and C. Tsai (1998). “Assessment of model adequacy for Markov regression time series models.” *Biometrics,* **54**, 1165-1175.
24. W. K. Li (1991). ”Testing model adequacy for some Markov regression models for time series.” *Biometrika,* **78,** 83-89.
25. W. K. Li (1994).” Time series models based on generalized linear models: Some further results.” *Biometrics,* **50**, 506-511.
26. T. Liboschik, K. Fokianos and R. Fried (2016). “tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models”. Vignette of package tscount version 1.3.0 (https://cran.rproject.org/web/packages/tscount/vignettes/tsglm.pdf)
27. P. McCullagh and J.A. Nelder (1989). *Generalized Linear Models*, (2nd edition), Chapman and Hall, London.

28. D. C. Montgomery, C. L. Jennings and M.Kulahci (2016). *Introduction to Time Series Analysis and Forecasting.* (2nd edition). Wiley, Hoboken, New Jersey.
29. Z. Ni and B. Kedem (2000). "Normal probabilities in the equicorrelated case." *Journal of Mathematical Analysis and Applications*, **246**, 280-295.
30. R. H. Shumway and D. S. Stoffer (2000). *Time Series Analysis and Its Applications.* Springer, New York.
31. E. V. Slud (1995)." A counting process model for the London mortality data." Technical Report 95-02, Department of Mathematics, University of Maryland at College Park.
32. W. H. Wong (1986). " Theory of partial likelihood." *Annals of Statistics,* **14**, 88-123.
33. S. L. Zeger and B. Qaqish (1988). "Markov regression models for time series: A quasi likelihood approach." *Biometrics,* **44**, 1019-1031.