ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΜΑΘΗΜΑΤΙΚΟΥ ΕΦΑΡΜΟΓΩΝ
ΕΦΑΡΜΟΣΜΕΝΗ ΑΝΑΛΥΣΗ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΣΙΟΥΤΗ ΑΙΚΑΤΕΡΙΝΗ – ge15073

ΤΕΧΝΙΚΕΣ ΠΡΟΒΛΕΨΗΣ
ΘΕΩΡΙΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

Επιβλέπων: Ιωάννης Κολέτσος, Αναπληρωτής Καθηγητής

Η τριμελής εξεταστική επιτροπή

Ι. Κολέτσος	**Μ. Λουλάκης**	**Δ. Φουσκάκης**
Αναπλ.Καθηγητής	**Καθηγητής**	**Καθηγητής**

ΑΘΗΝΑ, ΟΚΤΩΒΡΙΟΣ 2023

ΠΕΡΙΕΧΟΜΕΝΑ

ΓΕΝΙΚΑ ΓΙΑ ΤΗΝ ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ

ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ

Η επιχειρησιακή έρευνα, κάποιες τεχνικές της οποίας θα εφαρμοστούν στην παρούσα διπλωματική, είναι ο επιστημονικός κλάδος που ασχολείται με την βέλτιστη λήψη αποφάσεων με έναν συστηματικό τρόπο. Θεωρείται ότι, επί της ουσίας, «γεννήθηκε» ως κλάδος το 1940, οι ρίζες όμως αυτής της επιστήμης μπορούν να βρεθούν αρκετά πιο πριν. Η αρχή έγινε από τον Charles Babbage (1791-1871), ο οποίος θεωρείται ο «πατέρας της επιχειρησιακής έρευνας», μιας και η έρευνά του για το κόστος μεταφοράς και το κόστος ταξινόμησης της αλληλογραφίας οδήγησε στη δημιουργία του γενικού αγγλικού «Ταχυδρομείου της πένας» το 1840. Το 1917 ο Agner Krarup Erlang (1878-1929) μελέτησε προβλήματα που αφορούσαν το χρόνο απασχόλησης τηλεφωνικών κέντρων και το 1920 ο Horace Levinson μελέτησε προβλήματα σχετικά με τις πωλήσεις και το εμπόριο.

Πάρα αυτές τις πρώτες προσπάθειες, όμως, το 1940 θεωρείται ως η πλέον ξεκάθαρη απαρχή της επιχειρησιακής έρευνας όπως τη γνωρίζουμε σήμερα. Αυτό συμβαίνει λόγω των παρακάτω: την εποχή εκείνη είχε ήδη ξεσπάσει ο Β' Παγκόσμιος Πόλεμος και έτσι, για να φτάσει να λήξει όπως έληξε, χρειάστηκε να επιστρατευθεί πληθώρα επιστημόνων από διάφορους κλάδους, όπως μαθηματικών, φυσικών, στατιστικών, βιολόγων, ψυχιάτρων κ.ά. Έτσι, στο Ηνωμένο Βασίλειο λ.χ. επιστήμονες όπως οι Patrick Blackett, Jonathan Vivian Rosenhead, Conrad Hal Waddington, Owen Wansbrough- Jones και Frank Yates αλλά και ο George Dantzig στις Ηνωμένες Πολιτείες επιδίδονταν εκεί την εποχή στην εύρεση τρόπων λήψης των καλύτερων δυνατών αποφάσεων για τα προβλήματα που προκύπταν σε ζητήματα όπως τα προγράμματα διοικητικής μέριμνας και τα προγράμματα εκπαίδευσης.

Πιο συγκεκριμένα, οι διοικητές των ενόπλων δυνάμεων της Μεγάλης Βρετανίας χρειάζονταν βοήθεια στη μελέτη της νέας για την εποχή τεχνολογίας των ραντάρ και την εξεύρεση πιο αποτελεσματικών μεθόδων για τον εντοπισμό εχθρικών αεροσκαφών. Προς αυτήν την κατεύθυνση, το βρετανικό υπουργείο αεράμυνας ίδρυσε το κέντρο Bawdsey Manor Research Station στο Suffolk με διευθυντή του κέντρου για τα πρώτα χρόνια τον Robert Watson-Watt (ως τότε υπεύθυνο του τμήματος ραδιοκυμάτων του Βρετανικού Εθνικού Εργαστηρίου) και στη συνέχεια τον Albert P.Rowe. Οι πρώτες ασκήσεις που εκτελέστηκαν με χρήση ραντάρ ήταν αποτελεσματικές για την έγκαιρη προειδοποίηση, αλλά απογοητευτικά σε ό,τι αφορούσε τον εντοπισμό μετά το φιλτράρισμα και τη μετάδοση των στοιχείων από το δίκτυο. Ύστερα, έγινε εκ νέου δοκιμή με τέσσερις επιπλέον σταθμούς ραντάρ, το οποίο δημιούργησε μια νέα περιπλοκή, που αφορούσε στο συντονισμό και συνδυασμό συχνά αντικρουόμενων πληροφοριών από τους διαφορετικούς σταθμούς. Έτσι, ενώ τεχνικά τα ραντάρ λειτουργούσαν άρτια, απαιτούνταν έρευνα για τη λειτουργικότητα των συστημάτων αυτών. Έτσι, ένεκα και του πολέμου, αποφασίστηκε να ιδρυθεί το 1939 το Stanmore Research Section και το 1941 το Operation Research Section της βασιλικής αεροπορίας, των οποίων οι αποφάσεις ήταν καθοριστικής σημασίας για την αναχαίτιση των Γερμανών στη μάχη της Βρετανίας και γενικότερα στη μετέπειτα εξέλιξη του πολέμου.

Μετά τον πόλεμο, οι διαθέσιμοι πόροι ήταν κατακερματισμένοι και για αυτό ήταν απαραίτητη η σωστή διαχείρισή τους. Συγχρόνως η ανάπτυξη της βιομηχανίας ήταν ραγδαία, σε σημείο που να επικρατεί σύγχυση λόγω αυξανόμενης πολυπλοκότητας των οργανισμών και της εξειδίκευσης. Οι επιστήμονες και οι ερευνητές της επιχειρησιακής έρευνας που είχαν εργαστεί κατά τη διάρκεια του πολέμου γρήγορα κατάλαβαν ότι τα προβλήματα που τους

απασχόλησαν σχετικά με το στρατό κατά τη διάρκεια του πολέμου δεν απείχαν πολύ από τα νέα «βιομηχανικά» ζητήματα. Έχοντας σαφώς επιτυχημένη πορεία στα χρόνια του πολέμου με τις εφαρμογές της στο στρατό, η επιχειρησιακή έρευνα κατάφερε να κεντρίσει το ενδιαφέρον της βιομηχανίας. Στην Αμερική, λόγου χάρη, ιδρύθηκε το Operations Evaluation Group (OEG) σε συνεργασία με το MIT και υπογράφηκε σύμβαση με την Douglas Aircraft Company για το έργο RAND (Research and Development), το οποίο στην ουσία επέκτεινε τη χρήση των ερευνητών επιχειρησιακής έρευνας για σημαντικό χρονικό διάστημα μετά τη λήξη του πολέμου. Μέχρι το 1951 η επιχειρησιακή έρευνα είχε επικρατήσει στη Μεγάλη Βρετανία και το 1957 ιδρύεται η Διεθνής Συνομοσπονδία Εταιρειών Επιχειρησιακών Ερευνών (IFORS, International Federation of Operations Research Societies). Σύντομα, ο κλάδος της επιχειρησιακής έρευνας εισήχθη και στην Ελλάδα, διευρύνοντας την αντίληψη που υπήρχε μέχρι τότε για το management. Το 1963 ιδρύθηκε, από μια ομάδα πρωτοπόρων επιστημόνων, η Ελληνική Εταιρία Επιχειρησιακών Ερευνών, μια επιστημονική, μη κερδοσκοπική εταιρία που αποσκοπεί στο να προάγει και να διαδώσει την Επιχειρησιακή Έρευνα στην Ελλάδα.

Βασική για την εξέλιξη της επιχειρησιακής έρευνας ήταν η εξέλιξη των τεχνικών που εφαρμόζονταν για την επίλυση προβλημάτων. Εργαλεία όπως ο γραμμικός προγραμματισμός, ο δυναμικός προγραμματισμός και η θεωρία ουρών αναμονής αναπτύχθηκαν πριν το τέλος της δεκαετίας του 1950. Καίριας σημασίας ήταν και η ανάπτυξη της μεθόδου Simplex από τον George Dantzig, που μας δίνει τη λύση σε πληθώρα προβλημάτων του γραμμικού προγραμματισμού και ειδικά στις μέρες μας με πολύ μεγάλη ευκολία, δοθείσης της ραγδαίας ανάπτυξης των ηλεκτρονικών υπολογιστών. Η εξέλιξη των υπολογιστών, πέραν αυτού, συνέβαλε και στην καλύτερη διαχείριση και επεξεργασία των πληροφοριών από τα στελέχη της διοίκησης.

Παρακάτω παραθέτουμε κάποιες σημαντικές εξελίξεις για τον κλάδο της επιχειρησιακής έρευνας και των εφαρμογών του:

- 1943: Νευρωνικά δίκτυα, W.S. McCulloch & W. H. Pitts
- 1944: Εκθετική Εξομάλυνση, R.G. Brown
- 1944:Θεωρία Παιγνίων και Οικονομική Συμπεριφορά, John Von Neumann & Oscar Morgenstern
- 1947: Αλγόριθμος Simplex, George Dantzig
- 1951-1952: Ανάπτυξη της Δυϊκής Θεωρίας.
- 1972: Ο αλγόριθμος Simplex δεν είναι πολυωνυμικός, Victor LaRue Klee- G.J. Minty :Εκθετική συμπεριφορά
- 1982: Υπολογισμός Μέσης Πολυπλοκότητας Αλγορίθμου Simplex, Karl Heinz Borgwardt
- 1984: Ανακάλυψη πολυωνυμικού αλγορίθμου Εσωτερικών Σημείων, Narendra K. Karmarkar
- 1991: Ανακάλυψη Αλγορίθμων Εξωτερικών Σημείων τύπου Simplex, Konstantinos Paparrizos.

ΒΑΣΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ ΤΗΣ ΕΠΙΧΕΙΡΗΣΙΑΚΗΣ ΕΡΕΥΝΑΣ

Τα προβλήματα στην επιχειρησιακή έρευνα δια μέσου των ακόλουθων βημάτων:

- *Ανάλυση του Συστήματος*: Σε αυτό το βήμα, ο σκοπός μας είναι η κατανόηση του συστήματος που πρόκειται να μελετήσουμε. Έτσι, προσπαθούμε να προσδιορίσουμε τη δομή του και τον τρόπο λειτουργίας του. Έπειτα, αναλύουμε τα υποσυστήματά του, εντοπίζουμε παράγοντες που μπορούμε να επηρεάσουμε τη λειτουργία τους και

αναζητούμε τρόπους (στρατηγικές), τους οποίους μπορούμε να εφαρμόσουμε προκειμένου να επηρεάσουμε το υπό μελέτη σύστημα. Με τον τρόπο αυτό αποκτούμε μια σαφή εικόνα του προβλήματος που καλούμαστε να λύσουμε και αναγνωρίζουμε τις μεταβλητές-παραμέτρους του συστήματος, καθώς και τους διάφορους περιορισμούς που επιβάλλονται από τη δομή του, τη λειτουργία του, ή το περιβάλλον.

- *Διατύπωση Στόχων*: Έχοντας αναλύσει το σύστημα, προσδιορίζουμε έπειτα τους στόχους που επιθυμούμε να επιτύχουμε, λ.χ. μεγιστοποίηση κέρδους, ελαχιστοποίηση κόστους, βελτίωση παραγωγικότητας ή άλλο. Είναι ένα πολύ σημαντικό βήμα, μιας και ο στόχος που τίθεται επηρεάζει άμεσα το ποια λύση θα θεωρηθεί η βέλτιστη για το πρόβλημά μας. Δεν είναι πάντα εύκολη η διατύπωση των στόχων, δεδομένου ότι σε αρκετά προβλήματα πρέπει να διαλέξουμε ανάμεσα σε πολλούς στόχους που θα θέλαμε να επιτύχουμε και έτσι αναγκαζόμαστε να θέσουμε μια προτεραιότητα σε κάποιον εξ αυτών. Για παράδειγμα, σε μια εταιρία που παράγει και πουλάει προϊόντα, το τμήμα πωλήσεων έχει ως στόχο την αύξηση κατά το δυνατό περισσότερο των πωλήσεων, το τμήμα παραγωγής έχει ως στόχο την σωστότερη διαχείριση των πόρων, τη μεγιστοποίηση της παραγωγικότητας και την καλύτερη απασχόληση του εργατικού δυναμικού. Από την άλλη, οι μέτοχοι της εταιρίας θέλουν τη μεγιστοποίηση των κερδών και του μερίσματος. Προφανώς, όλοι αυτοί οι στόχοι δεν επιτυγχάνονται με την ίδια στρατηγική και, ως εκ τούτου, πρέπει να γίνει μια ιεράρχηση των προαναφερθέντων στόχων.
- *Διατύπωση του Μοντέλου*: Το επόμενο βήμα είναι ο προσδιορισμός ενός, κατά το δυνατόν, απλού μοντέλου με σκοπό να το μελετήσουμε. να αναλύσουμε την επίδραση διαφόρων παραγόντων στους στόχους που έχουν τεθεί και να επιλέξουμε την καλύτερη στρατηγική. Η κατασκευή του μαθηματικού μοντέλου έχει ως αποτέλεσμα τη μετατροπή του ορισμού του προβλήματος σε μαθηματικές σχέσεις και αποτελεί μια κατά προσέγγιση αναπαράσταση του προβλήματος. Το μοντέλο αυτό, συνήθως, είναι ένα σύνολο ποσοτικών σχέσεων ή εντολών στον υπολογιστή που εκφράζουν τους στόχους που έχουν τεθεί και των περιορισμών που επιβάλλονται. Η διατύπωση του μοντέλου έχει ως αφετηρία την διατύπωση κάποιων υποθέσεων που μπορούν με κάποιο τρόπο να δικαιολογηθούν απ' τη φύση του προβλήματος, ύστερα συνεχίζεται με την έκφραση μαθηματικών σχέσεων ή εντολών στον υπολογιστή που περιγράφουν τις σχέσεις μεταξύ συντελεστών, στόχων, μεταβλητών και του περιβάλλοντος του προβλήματος και ολοκληρώνεται με την επιβεβαίωση του μοντέλου με δοκιμαστική χρήση σε ένα απλό πρόβλημα. Η διαμόρφωση του μοντέλου είναι ένα πολύ σημαντικό και συγχρόνως δύσκολο βήμα, ενώ αν αποτύχει μπορεί να οδηγήσει στη λήψη λανθασμένων αποφάσεων. Ακόμη και αν στη διαδικασία επίλυση του προβλήματος δεν υπάρξει σφάλμα, η απάντηση που θα λάβει ο αναλυτής μπορεί να είναι εσφαλμένη, μιας και ενδέχεται να μην αναφέρεται στο ζητούμενο πρόβλημα. Σημαντικό, επίσης, είναι η κατά το δυνατό ελάττωση της διάστασης του προβλήματος, δηλαδή η αξιοποίηση όσο γίνεται λιγότερων μεταβλητών, παραλείποντας άλλες οι οποίες μπορεί να είναι περιττές και να μην έχουν ουσιαστικά μεγάλη επίδραση στην τελική λύση.
- *Επίλυση του Μοντέλου*: Η επίλυση του μοντέλου χρησιμοποιεί διάφορες τεχνικές για τον εντοπισμό της βέλτιστης λύσης του προβλήματος. Με τη λύση του προβλήματος, εννοούμε τον προσδιορισμό της στρατηγικής που θα ακολουθήσουμε. Οι διάφορες μέθοδοι που χρησιμοποιούνται εδώ στηρίζονται σε Ανώτερα μαθηματικά (όπως το διαφορικό και ολοκληρωτικό λογισμό, την αριθμητική ανάλυση, τη γραμμική άλγεβρα, κλασικές μεθόδους βελτιστοποίησης, λογισμό μεταβολών) στη θεωρία Πιθανοτήτων (κατανομές πιθανοτήτων, διαδικασίες Markov) και τη Στατιστική (περιγραφική στατιστική, στατιστική συμπερασματολογία, εκτίμηση παραμέτρων, ανάλυση

παλινδρόμησης , αλληλοσυσχέτιση, ανάλυση μεταβλητότητας, παραγοντική ανάλυση, χρονοσειρές) ή σε Μεθόδους και Θεωρίες της επιχειρησιακής έρευνας. Οι τελευταίες είναι οι πιο εύκολες και συχνά χρησιμοποιούμενες, αφού είναι κατά κανόνα αριθμητικές, επαναληπτικές μέθοδοι και στηρίζονται στη χρήση αλγορίθμων υλοποιούμενων στον υπολογιστή. Η επιλογή της κατάλληλης μεθόδου ή θεωρίας της επιχειρησιακής έρευνας εξαρτάται από τον τύπο και την πολυπλοκότητα του μαθηματικού μοντέλου. Αν το μοντέλο ταιριάζει σε κάποιο από τα γνωστά μαθηματικά μοντέλα, όπως λ.χ. αυτό του γραμμικού προγραμματισμού, τις περισσότερες φορές είναι εφικτή η εύρεση λύσης με τη χρήση κατάλληλων αλγορίθμων. Εναλλακτικά, αν οι μαθηματικές σχέσεις είναι τόσο περίπλοκες σε σημείο που να καθίσταται αδύνατη η αναλυτική επίλυση, οι ομάδες της επιχειρησιακής έρευνας αναγκάζονται να κάνουν απλοποιήσεις στο μοντέλο και να χρησιμοποιήσουν ευρετικές τεχνικές (heuristics) ή μεθόδους προσομοίωσης (simulation). Σε ορισμένες περιπτώσεις μάλιστα, ίσως χρειαστεί συνδυασμός των προαναφερθέντων μεθόδων.

- *Ανάλυση Ευαισθησίας*: Στο αμέσως προηγούμενο βήμα, η λύση που βρήκαμε είναι η βέλτιστη για τις συγκεκριμένες τιμές των παραμέτρων που εμπλέκονται στο πρόβλημα. Ωστόσο, πριν υλοποιήσουμε την στρατηγική που υποδεικνύει το μοντέλο, είναι χρήσιμο να εξετάσουμε την επίπτωση που θα είχε στη λύση μια ελαφρά αλλαγή στις παραμέτρους. Η διαδικασία αυτή είναι γνωστή ως ανάλυση ευαισθησίας.
- *Υλοποίηση της λύσης*: Έχοντας κάνει όλα τα προαναφερθέντα βήματα, μένει να ληφθούν μέτρα για την εφαρμογή της λύσης. Το στάδιο αυτό είναι συχνά το δυσκολότερο. Η υλοποίηση και διατήρηση της λύσης ενός μοντέλου περιλαμβάνει τη μετατροπή των αποτελεσμάτων σε λειτουργικές οδηγίες, παρουσιασμένες με κατανοητό τρόπο σε άτομα που θα διαχειριστούν το προτεινόμενο σύστημα, ώστε η βελτίωση που επιτεύχθηκε να υλοποιηθεί στο πραγματικό σύστημα και να διατηρηθεί στο μέλλον. Το βάρος αυτού του σταδίου επωμίζεται κυρίως η ομάδα της επιχειρησιακής έρευνας, αφού είναι πιθανό να προκύψουν προβλήματα τα οποία δεν είχαν προβλεφθεί κατά τη διάρκεια της έρευνας.

ΠΡΟΒΛΕΨΗ (FORECASTING)

Σε πάρα πολλές περιπτώσεις πρακτικών εφαρμογών, μας ενδιαφέρει ιδιαίτερα το να μπορούμε να κάνουμε προβλέψεις για μελλοντικές εξελίξεις. Μια εταιρία θα ήθελε να έχει μια ιδέα του τι πορεία θα ακολουθήσει τον προσεχή χρόνο λ.χ. η οικονομία, το χρηματιστήριο, τα επιτόκια ή οι προτιμήσεις των καταναλωτών. Είναι σαφές ότι πολλοί που θα επιχειρήσουν να δώσουν απάντηση σε αυτά τα ερωτήματα θα κάνουν λάθος, μιας και το μέλλον μας παραμένει άγνωστο. Εντούτοις, είναι σημαντικό για την επιβίωση και την εξέλιξη μιας επιχείρησης να διαθέτει προσωπικό που να μπορεί να κάνει κατά το δυνατό σωστές προβλέψεις όσον αφορά, π.χ., τις τάσεις που αναπτύσσονται στην αγορά, ώστε να ληφθεί δράση βάσει αντίστοιχων στρατηγικών.

Από τη στιγμή που υπάρχουν δεδομένα από το ιστορικό των πωλήσεων, μπορούν να αξιοποιηθούν κάποιες στατιστικές μέθοδοι για να προβούμε σε προβλέψεις. Οι μέθοδοι αυτές, βέβαια, στηρίζονται στην υπόθεση ότι η μέχρι τώρα πορεία των δεδομένων θα συνεχιστεί και μελλοντικά. Αν αυτό κρίνεται παράλογο από τον αναλυτή, γίνεται κάποιες απαραίτητες αλλαγές ώστε οι προβλέψεις να συνυπολογίζουν τις τρέχουσες εξελίξεις στην αγορά.

Πέραν αυτών, υπάρχουν και τεχνικές πρόβλεψης που στηρίζονται εξ ολοκλήρου στη γνώμη ενός ειδικού. Τέτοιες μέθοδοι είναι δόκιμες σε περίπτωση που τα διαθέσιμα δεδομένα είναι λίγα ή έχουν λάβει χώρα τεράστιες αλλαγές στην αγορά, σε βαθμό που τα υπάρχοντα δεδομένα να καθίστανται αναξιόπιστα για να χρησιμοποιηθούν σε προβλέψεις.

ΜΕΡΙΚΕΣ ΕΦΑΡΜΟΓΕΣ ΤΟΥ FORECASTING

-Πρόβλεψη πωλήσεων

Κάθε επιχείρηση που προβαίνει στην πώληση αγαθών χρειάζεται να κάνει προβλέψεις για τις πωλήσεις αυτών των αγαθών. Οι κατασκευαστές χρειάζονται να γνωρίζουν την ποσότητα που θα παράξουν. Οι έμποροι χονδρικής και λιανικής αντίστοιχα χρειάζεται να ξέρουν τι απόθεμα πρέπει να κρατήσουν. Αν υποτιμηθεί η ζήτηση, αυτό θα οδηγήσει σε ποικίλα προβλήματα όπως χαμένες πωλήσεις, δυσαρεστημένοι πελάτες και ενδεχομένως να παρουσιαστεί στον ανταγωνισμό η ευκαιρία να επικρατήσει στην αγορά. Από την άλλη, υπερεκτιμώντας τη ζήτηση μπορεί να κοστίσει στην επιχείρηση λόγω (1) υπερβολικών δαπανών για την αποθήκευση, (2) εξαναγκασμένες μειώσεις τιμών, (3) πλεονάζουσα χωρητικότητα αποθηκευτικού χώρου και (4) χαμένες ευκαιρίες προώθησης επικερδών αγαθών.

Οι αεροπορικές εταιρίες πλέον στηρίζονται αρκετά στις υψηλές χρεώσεις που πληρώνονται από επιχειρηματίες που ταξιδεύουν την τελευταία στιγμή, ενώ προσφέρουν εκπτώσεις σε εισιτήρια στους υπόλοιπους επιβάτες για να συμπληρώσουν τις θέσεις. Το ποιες θέσεις θα δοθούν πού είναι ένα σημαντικό ζήτημα που η κατάλληλη επιλογή αναφορικά με αυτό το θέμα μπορεί να μεγιστοποιήσει τα κέρδη. Η American Airlines, για παράδειγμα, χρησιμοποιεί τεχνικές όπως αυτές που θα αναπτύξουμε παρακάτω, για να κάνει προβλέψεις για τη ζήτηση των θέσεων βοηθώντας στην ορθή λήψη απόφασης για το προαναφερθέν πρόβλημα.

-Πρόβλεψη Ανάγκης για ανταλλακτικά

Παρ ’ότι η πρόβλεψη των πωλήσεων επιτυχώς είναι καίριας σημασίας για την οποιαδήποτε επιχείρηση, υπάρχει επίσης μεγάλη ανάγκη να γίνονται σωστές προβλέψεις άλλου τύπου, και μια απ’ τις σημαντικότερες περιπτώσεις αφορά στη διαχείριση των ανταλλακτικών.

Πολλές επιχειρήσεις χρειάζονται να διατηρούν αποθέματα ανταλλακτικών για την έγκαιρη επισκευή εξοπλισμού ή προϊόντων που πωλούνται/δανείζονται στους πελάτες. Το εν λόγω απόθεμα, σε κάποιες περιπτώσεις, ενδέχεται να είναι εξαιρετικά μεγάλο. Για παράδειγμα, η IBM διαθέτει μια αποθήκη που περιέχει χιλιάδες διαφορετικά ανταλλακτικά συνολικής αξίας δισεκατομμυρίων.

Όπως και στην προηγούμενη περίπτωση προβλέψεων σχετικά με τις πωλήσεις, έτσι και εδώ βλέπουμε ότι η διαχείριση της αποθήκης μιας επιχείρησης στηρίζεται κατά μείζονα λόγο στις σωστές προβλέψεις. Υπάρχει διαφορά, βέβαια, στα κόστη που προκύπτουν λόγω υποτίμησης της ζήτησης, αλλά δεν είναι λιγότερο επισφαλείς οι αντίστοιχες συνέπειες. Έτσι, για παράδειγμα, το γεγονός ότι ένα αεροσκάφος δεν μπορεί να πετάξει εξαιτίας της έλλειψης σε κάποιο ανταλλακτικό μπορεί να έχει ως αποτέλεσμα καθυστερήσεις ή ακόμη και αναβολή πτήσης, με επιζήμιο αποτέλεσμα για την αεροπορική εταιρία.

-Πρόβλεψη αναγκών σε προσωπικό

Στις μέρες μας, σε πολλές μεγάλες οικονομίες στηρίζονται ολοένα και περισσότερο στην παροχή υπηρεσιών. Η βιομηχανική παραγωγή πραγματοποιείται σε διαφορετική χώρα, στην οποία ως διαδικασία είναι σαφώς λιγότερο κοστοβόρα, με αποτέλεσμα να υπάρχει το περιθώριο να δοθεί μεγαλύτερη έμφαση σε υπηρεσίες διαφόρων ειδών (π.χ., ταξίδια, τουρισμός, ψυχαγωγία, νομική βοήθεια, υπηρεσίες υγείας, χρηματοοικονομικές, εκπαιδευτικές, σχεδιασμός, συντήρηση κ.λπ.). Έτσι, δεν τίθεται πλέον θέμα πωλήσεων για αυτές τις επιχειρήσεις – και κατ' επέκταση αντίστοιχων προβλέψεων- αλλά περισσότερο θέμα πρόβλεψης της ζήτησης για τις εν λόγω υπηρεσίες. Ως εκ τούτου, ο διευθυντής μιας εταιρείας παροχής κάποιων υπηρεσιών πρέπει να προβλέψει σωστά τις ανάγκες του καταναλωτικού κοινού αυτών των υπηρεσιών. Επομένως, για να μπορεί να ανταποκριθεί σε αυτές, θα πρέπει να μπορεί προβλέψει τον αριθμό υπαλλήλων που θα χρειαστούν για αυτό το σκοπό.

-Πρόβλεψη οικονομικών τάσεων (Forecasting economic trends):

Σαφώς, πέραν των παραγόντων που αφορούν μια επιχείρηση αυτή καθαυτή, υπάρχουν και εξωγενείς παράγοντες της οικονομίας που επιδρούν στην ευημερία μιας επιχείρησης. Έτσι, για παράδειγμα, δεν είναι σημαντικό να βλέπει κανείς μόνο το ιστορικό των πωλήσεων ενός προϊόντος, αλλά συγχρόνως να παρακολουθεί π.χ. το ποσοστό του πληθωρισμού ή το ποσοστό ανεργίας. Τα αντίστοιχα στατιστικά μοντέλα για την πρόβλεψη των οικονομικών τάσεων ονομάζονται οικονομετρικά μοντέλα. Με την αξιοποίηση ιστορικών δεδομένων, αυτά τα οικονομετρικά μοντέλα συνήθως εξετάζουν ένα πολύ μεγάλο αριθμό παραγόντων που συμβάλλουν στην ανάπτυξη της οικονομίας και η ανάπτυξή τους ομοιάζει αρκετά σε αυτήν για αντίστοιχα μοντέλα πρόβλεψης πωλήσεων.

ΣΤΑΔΙΑ ΤΗΣ ΔΙΑΔΙΚΑΣΙΑΣ ΠΡΟΒΛΕΨΗΣ

Η πρόβλεψη ως διεργασία είναι αρκετά σύνθετη και θα μπορούσε να αναλυθεί στα ακόλουθα βήματα:

- Προσδιορισμός του προβλήματος: Πρέπει να προηγηθεί πάντα μια λεπτομερής εξέταση και προσεκτικός καθορισμός του προβλήματος για να έχουν πρακτική αξία στην πραγματικότητα τα παρακάτω βήματα.
- Επιλογή του χρονικού ορίζοντα πρόβλεψης: Για να γίνει σωστά η ανάλυση που θα οδηγήσει σε αποτελεσματικές προβλέψεις, πρέπει να επιλεγεί σωστά ο χρονικός ορίζοντας, δηλαδή πρέπει να είναι σαφές ποια είναι η αρχή και ποιο το τέλος της χρονικής περιόδου που μας ενδιαφέρει να εξετάσουμε και να πραγματοποιήσουμε προβλέψεις.

- Συλλογή πληροφορίας: Αφότου τα παραπάνω προκαταρκτικά στοιχεία έχουν καλυφθεί, πρέπει στη συνέχεια να γίνει συλλογή και οργάνωση της πληροφορίας που θα αξιοποιηθεί για την ανάπτυξη του μοντέλου. Η πηγή αυτής της πληροφορίας μπορεί να είναι είτε στατιστικά δεδομένα είτε η γνώση και η γνώμη διαφόρων εμπειρογνώμων (που αποτελούν μέρος της διαδικασίας συλλογής των δεδομένων) ή και του πελάτη, ο οποίος αποσκοπεί στο να χρησιμοποιήσει τις προβλέψεις.
- Προκαταρκτική (διερευνητική) ανάλυση: Κατόπιν συλλογής των δεδομένων, συνεχίζουμε με τη γραφική παράστασή τους. Από τις γραφικές αυτές, ελέγχουμε κατά πόσο υπάρχουν συστηματικά μοτίβα (patterns), τάση (Trend), σημάδια εποχικότητας (Seasonality). Επιπλέον εξετάζουμε αν υπάρχουν ενδείξεις επιχειρηματικών κύκλων (business cycles).Τέλος, ανιχνεύουμε τυχόν δεδομένα που αποκλίνουν πολύ από τα υπόλοιπα (Outliers).
- Επιλογή και προσαρμογή μοντέλου: Η επιλογή για το μοντέλο στην οποία θα καταλήξουμε επηρεάζεται από τα ιστορικά δεδομένα, το πόσο «στενά» συνδέεται η μεταβλητή απόκρισης (/πρόβλεψης) με τις επεξηγηματικές μεταβλητές, και το τρόπο με τον οποίο η πρόβλεψη θα χρησιμοποιηθεί. Συνήθως καταλήγουμε σε μια επιλογή αφού συγκρίνουμε πολλά διαφορετικά μοντέλα. Θα πρέπει να μας είναι γνωστό, ως εκ τούτου, πότε μπορεί να εφαρμοστεί το εκάστοτε μοντέλο, πόσο αξιόπιστο είναι το καθένα από αυτά και τι είδους δεδομένα απαιτούνται για το καθένα. Μπορεί, ακόμη, πολλά μοντέλα να προσαρμοστούν και να είναι αποδοτικά, οπότε η απόφαση για το ποια θα είναι η πρόβλεψή μας να γίνει βάσει άλλου κριτηρίου.
- Χρήση και αξιολόγηση του μοντέλου πρόβλεψης: Η απόδοση ενός μοντέλου μπορεί να υπολογιστεί επαρκώς όταν τα δεδομένα για την περίοδο πρόβλεψης είναι διαθέσιμα. Αρκετές μέθοδοι έχουν αναπτυχθεί για να βοηθήσουν στην αξιολόγηση της ακρίβειας των προβλέψεων. Στην πράξη όμως, για να αξιολογήσουμε ένα μοντέλο, χωρίζουμε το σύνολο των δεδομένων σε δύο υποσύνολα, το σύνολο εκπαίδευσης (training set) και το σύνολο ελέγχου (test set). Με το πρώτο υποσύνολο, το μοντέλο «εκπαιδεύεται» και στην συνέχεια αξιολογείται με βάση την απόδοση του στο σύνολο ελέγχου. Έτσι με αυτό τον τρόπο μπορούμε να γνωρίζουμε πόσο καλά αναμένουμε να αποδώσει το μοντέλο, που θα επιλέξουμε, στα μελλοντικά άγνωστα δεδομένα. Επιπλέον η πρόβλεψη που λαμβάνεται μέσω οποιουδήποτε μοντέλου θα πρέπει να αξιολογείται από την άποψη του διαστήματος εμπιστοσύνης. Συνήθως όλα τα καλά μοντέλα πρόβλεψης διαθέτουν μεθόδους υπολογισμού της ανώτερης και της κατώτερης τιμής εντός των οποίων αναμένεται να παραμείνει η συγκεκριμένη πρόβλεψη, με προκαθορισμένο επίπεδο σημαντικότητας. Μπορεί επίσης να αξιολογηθεί από λογική άποψη, κατά πόσον η τιμή που λαμβάνεται είναι λογικά εφικτή; Μπορεί επίσης να αξιολογηθεί σε σχέση με κάποια σχετική μεταβλητή ή φαινόμενο. Έτσι, είναι δυνατόν, και μερικές φορές σκόπιμο να τροποποιηθεί η στατιστικώς προβλεπόμενη τιμή με βάση την αξιολόγηση.

5 ΒΑΣΙΚΑ ΕΡΓΑΛΕΙΑ ΠΡΟΒΛΕΨΗΣ, ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΠΡΟΒΛΕΨΕΩΝ

- Τακτοποίηση δεδομένων: Το πρώτο βήμα στην πρόβλεψη είναι η προετοιμασία των δεδομένων ώστε να αποκτήσουν τη σωστή μορφή. Αυτή η διαδικασία μπορεί να περιλαμβάνει την φόρτωση των δεδομένων, τον προσδιορισμό των τιμών που λείπουν, το φιλτράρισμα των χρονοσειρών καθώς και άλλες εργασίες προ-επεξεργασίας.
- Γραφική απεικόνιση δεδομένων (οπτικοποίηση): Ένα καίριο βήμα για την ανάλυση. Μέσω της γραφικής αναπαράστασης των δεδομένων μπορούμε να έχουμε μια πρώτη ιδέα για το ποιο θα μπορούσε να είναι το πλέον κατάλληλο μοντέλο για να χρησιμοποιήσουμε και να εξάγουμε κάποια πρώτα χρήσιμα συμπεράσματα.

- Ορισμός μοντέλου: Έχοντας σαν οδηγό τη γραφική απεικόνιση των δεδομένων (και ενδεχομένως μαζί με ορθή κρίση και εμπειρία) μπορούμε να προβούμε στον καθορισμό της μορφής του μοντέλου που θα έχουμε ως στόχο να υλοποιήσουμε στη συνέχεια.
- Εκπαίδευση του μοντέλου: Έχοντας καθορίσει τη μορφή του μοντέλου μας, χρησιμοποιούμε το επιλεχθέν σύνολο εκπαίδευσης από τα δεδομένα μας για να προβούμε στην εκτίμηση των παραμέτρων του μοντέλου.
- Έλεγχος της απόδοσης του μοντέλου: Μόλις εκτιμηθεί ένα μοντέλο, θα πρέπει με τα δεδομένα που μας απομένουν να εκτιμήσουμε το πόσο αποτελεσματικό είναι όντως αυτό το μοντέλο. Σαφώς δεν έχει νόημα να προσπαθήσουμε να χρησιμοποιήσουμε τα δεδομένα εκπαίδευσης για αυτή τη διεργασία, μιας και η πληροφορία που «κρύβεται» σε αυτά έχει ήδη χρησιμοποιηθεί στο έπακρο για τη διαμόρφωση του μοντέλου, οπότε δεν μπορούμε να τα χρησιμοποιήσουμε συγχρόνως και για τη «διάγνωση» και αξιολόγηση του μοντέλου.
- Δημιουργία προβλέψεων (πρόβλεψη): Έχοντας καλύψει όλα τα παραπάνω, μπορούμε με ασφάλεια να προβούμε στις επιθυμητές προβλέψεις.

Κριτικές Μέθοδοι πρόβλεψης

Αυτές οι μέθοδοι είναι, εκ φύσεως, υποκειμενικές και μπορεί να βασίζονται στην διαίσθηση, τη γνώμη ενός ειδικού και στην εμπειρία. Οδηγούν σε προβλέψεις που κατά βάση στηρίζονται σε ποιοτικά κριτήρια. Αυτές οι μέθοδοι μπορούν να χρησιμοποιηθούν εφόσον δεν γίνεται να χρησιμοποιηθούν ιστορικά δεδομένα για μια στατιστική μέθοδο πρόβλεψης ή ακόμη και αν υπάρχουν δεδομένα διαθέσιμα, εφαρμόζονται συμπληρωματικά σε αυτά. Αυτές οι μέθοδοι αφορούν στη λήψη αποφάσεων βάσει της γνώμης ενός ή περισσοτέρων μελών τις εταιρείας (διευθυντής, άλλα ανώτατα στελέχη, ομάδες πωλητών) ή ακόμη και των καταναλωτών.

ΧΡΟΝΟΣΕΙΡΕΣ

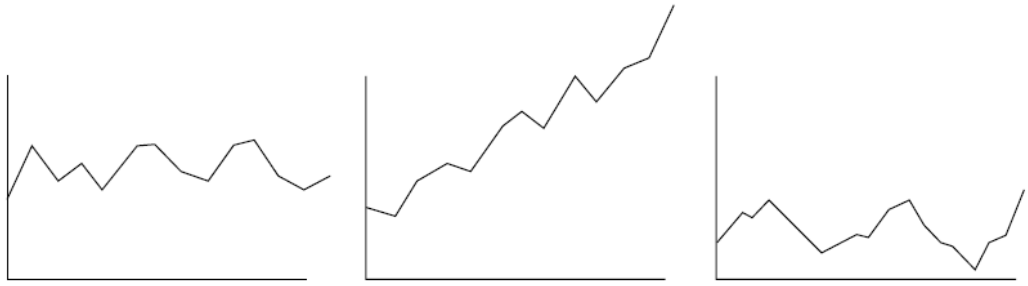
Οι περισσότερες στατιστικές μέθοδοι προβλέψεων αξιοποιούν ιστορικά δεδομένα που έχουν τη μορφή χρονοσειρών. Μια χρονοσειρά είναι μια σειρά παρατηρήσεων μέσα σε κάποιο χρονικό διάστημα για μια ποσότητα που μας ενδιαφέρει (δηλ. μια τυχαία μεταβλητή). Αν, λοιπόν, η X_i είναι η τυχαία μεταβλητή που μας ενδιαφέρει στο χρόνο i και οι παρατηρήσεις λαμβάνονται στους χρόνους $i = 1,2, \dots, t$ τότε οι παρατηρούμενες τιμές

$$X_1 = x_1, X_2 = x_2, \dots, X_t = x_t$$

απαρτίζουν μια χρονοσειρά.

Είναι σαφές ότι επειδή οι χρονοσειρές αποτελούν περιγραφή του παρελθόντος, είναι λογικό να αξιοποιηθούν ώστε να συμπεριλάβουμε τα ιστορικά δεδομένα στην ανάλυσή μας. Αν αυτά μπορούν να μας υποδείξουν τι μπορούμε να αναμένουμε μελλοντικά, μπορούμε να διαμορφώσουμε ένα αντίστοιχο αντιπροσωπευτικό μαθηματικό μοντέλο και ύστερα να κάνουμε τις οποιεσδήποτε προβλέψεις με βάση αυτό.

Στις περισσότερες ρεαλιστικές περιπτώσεις, δεν έχουμε πλήρη γνώση της ακριβούς μορφής του μοντέλου που δίνει τη χρονοσειρά, οπότε αρκούμαστε σε κάποιο προσεγγιστικό μοντέλο. Συχνά, η επιλογή του μοντέλου γίνεται βάσει κάποιου μοτίβου που ενδέχεται να παρατηρήσουμε στη χρονοσειρά. Χαρακτηριστικές τέτοιες περιπτώσεις φαίνονται στο ακόλουθο σχήμα:



Εικόνα 1: Παραδείγματα μοντέλων χρονοσειρών:

Στο 1ο διάγραμμα, παραδείγματος χάριν, βλέπουμε μια τυπική περίπτωση χρονοσειράς που προέρχεται από ένα **μοντέλο σταθερού επιπέδου συνδυασμένο με τυχαίες διακυμάνσεις**. Το 2ο απεικονίζει μια τυπική χρονοσειρά που αναπαρίσταται από μοντέλο που διαθέτει μια **γραμμική (ανοδική) τάση σε συνδυασμό με τυχαίες διακυμάνσεις**. Τέλος, 3ο γράφημα μας δείχνει μια χρονοσειρά που μπορεί να παρατηρηθεί αν η αντίστοιχη διαδικασία που τη δημιουργεί αναπαρίσταται από **ένα σταθερού επιπέδου μοντέλο με εποχιακή επίδραση σε συνδυασμό με τυχαίες διακυμάνσεις**. Παρ 'ότι υπάρχουν και πολλές άλλες περιπτώσεις, θα αρκεστούμε στη μελέτη των τριών παραπάνω.

Αφού έχουμε επιλέξει μορφή για το μοντέλο, δίνεται η μαθηματική αναπαράσταση για τη διαδικασία που γεννά τη χρονοσειρά. Για παράδειγμα, αν θεωρήσουμε ότι το μοντέλο μας είναι αυτό του σταθερού επιπέδου με τυχαίες επιδράσεις, η αντίστοιχη μαθηματική έκφραση θα είναι:

$$X_i = A + e_i, i = 1,2, \dots$$

όπου X_i η τιμή της τυχαίας μεταβλητής που παρατηρείται το χρόνο i, Α το σταθερό επίπεδο του μοντέλου και e_i το αντίστοιχο τυχαίο σφάλμα στη χρονική στιγμή i (το οποίο υποθέτουμε ότι έχει μηδενική μέση τιμή και σταθερή διασπορά). Έστω F_{t+1} η προβλεπόμενη τιμή της μεταβλητής για τη στιγμή $t+1$ δοθέντων όλων των τιμών $X_1 = x_1, \dots, X_t = x_t$. Λόγω του τυχαίου σφάλματος e_{t+1}, είναι αδύνατη η ακριβής πρόβλεψη της $X_{t+1} = x_{t+1}$ με χρήση της F_{t+1}, αλλά ο σκοπός είναι η εκτίμηση του $A = E[X_{t+1}]$ μέσω της F_{t+1} όσο καλύτερα γίνεται. Έτσι, είναι λογικό να περιμένουμε ότι η F_{t+1} θα είναι συνάρτηση κάποιων, έστω, από τις παρατηρηθείσες τιμές της χρονοσειράς.

Στη γενικότερη μελέτη των χρονοσειρών, είναι χρήσιμο να εξετάζουμε τη σχέση που η μεταβλητή που αντιστοιχεί σε μια χρονοσειρά ενδέχεται να έχει με άλλες (επεξηγηματικές) μεταβλητές. Έτσι, είναι χρήσιμο τις μεν και δε να τις απεικονίζουμε από κοινού γραφικά, ώστε να μπορούμε να ανακαλύψουμε αν μια συσχέτιση μεταξύ τους υπάρχει όντως και έπειτα ποια είναι η μορφή αυτής της συσχέτισης.

Επιπλέον, ένα χρήσιμο αριθμητικό μέτρο για τη διερεύνηση αυτού του ζητήματος είναι ο δειγματικός συντελεστής γραμμικής συσχέτισης (κατά Pearson) δύο μεταβλητών, ο οποίος εκφράζει τον "βαθμό" στον οποίο μπορούμε να εκτιμήσουμε γραμμικά την μία τ.μ. όταν γνωρίζουμε
την τιμή της άλλης. Η γραμμική συσχέτιση μεταξύ των μεταβλητών x και y δίνεται από:

$$r = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2}\sqrt{\sum(y_t - \bar{y})^2}}$$

Αρνητική τιμή του r εκφράζει αρνητική γραμμική συσχέτιση, και όσο πιο κοντά στην τιμή -1 είναι τόσο ισχυρότερη είναι. Αντίθετα θετική τιμή εκφράζει θετική γραμμική συσχέτιση των δυο μεταβλητών. Επειδή αυτό το μέγεθος εξετάζει μόνο τη γραμμική συσχέτιση δυο

μεταβλητών, ενδέχεται κάποια αποτελέσματα να είναι παραπλανητικά. Ενδέχεται, για παράδειγμα, το r να είναι πολύ κοντά στο 1 ή -1, αλλά εν τέλει η σχέση των δυο μεταβλητών να είναι πολυωνυμική αντί για γραμμική.

ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ ΓΙΑ ΤΟ ΜΟΝΤΕΛΟ ΣΤΑΘΕΡΟΥ ΕΠΙΠΕΔΟΥ

Θα παρουσιάσουμε στη συνέχεια τέσσερις εναλλακτικές για μοντέλα πρόβλεψης που υπάρχουν για το μοντέλο σταθερού επιπέδου που αναφέραμε παραπάνω. Αυτό το μοντέλο, όπως και κάθε άλλο, είναι ουσιαστικά μια εξιδανικευμένη αναπαράσταση της πραγματικής κατάστασης. Για μια πραγματική χρονοσειρά, μπορούν να συμβαίνουν έστω και μικρές αλλαγές στη τιμή Α. Η κάθε μια από τις παρακάτω μεθόδους αντικατοπτρίζει και μια διαφορετική εκτίμηση του πόσο πρόσφατα μπορεί να έχει συμβεί σημαντική αλλαγή σε αυτήν την τιμή (και αυτό αν έχει συμβεί γενικά μια τέτοια αλλαγή).

ΕΚΤΙΜΗΣΗ ΜΕΣΩ ΤΗΣ ΤΕΛΕΥΤΑΙΑΣ ΤΙΜΗΣ

Αν t η τρέχουσα χρονική στιγμή, η πρόβλεψη μέσω της τελευταίας τιμής χρησιμοποιεί την ήδη γνωστή τιμή x_t για την πρόβλεψη στο χρόνο $t+1$, δηλαδή

$$F_{t+1} = x_t$$

Για παράδειγμα, αν το x_t αναπαριστά τις πωλήσεις κάποιου προϊόντος στο τέλος ενός τριμήνου, αυτή η διαδικασία χρησιμοποιεί τα δεδομένα αυτών των πωλήσεων για να προβλέψει εκείνα του επόμενου τριμήνου. Η μέθοδος αυτή δεν είναι ιδιαίτερα αξιόπιστη για προβλέψεις, βέβαια, μιας και στηριζόμαστε σε δείγμα μεγέθους 1. Αξίζει να εφαρμοστεί μόνο στις εξής περιπτώσεις: (1) δεν ευσταθεί τόσο η υπόθεση του μοντέλου σταθερού επιπέδου και η διαδικασία αλλάζει με τέτοιους ρυθμούς ώστε οποιαδήποτε πληροφορία πριν το χρόνο t να είναι παραπλανητική ή (2) η υπόθεση σταθερής διασποράς στα σφάλματα είναι εσφαλμένη και στην πραγματικότητα η διασπορά στο χρόνο t είναι κατά πολύ μικρότερη από αυτή των προηγούμενων χρονικών στιγμών.

ΜΕΘΟΔΟΣ ΜΕΣΗΣ ΠΡΟΒΛΕΨΗΣ

Αυτή η μέθοδος αποτελεί το άλλο άκρο. Αντί να χρησιμοποιούμε τον μικρότερο δυνατό αριθμό δεδομένων, η συγκεκριμένη μέθοδος πρόβλεψης εκμεταλλεύεται *όλες* τις παρατηρήσεις της χρονοσειράς και λαμβάνει τη μέση τους τιμή. Έτσι, η πρόβλεψη της αμέσως επόμενης τιμής θα είναι:

$$F_{t+1} = \sum_{i=1}^{t} \frac{x_i}{t}.$$

Σε περίπτωση που η διαδικασία είναι απολύτως ευσταθής και άρα είναι εύλογη η υπόθεση του σταθερού μοντέλου, αυτή η εκτίμηση είναι η πλέον κατάλληλη. Ωστόσο, είναι πολλές φορές αμφίβολο αν θα διατηρήσει και μελλοντικά τη μορφή του μοντέλο –είναι πολύ πιθανό, δηλαδή, να εμφανιστούν διακυμάνσεις μελλοντικά. Για αυτό το λόγο, είναι πιο συνήθης η εφαρμογή

αυτής της τακτικής για διαδικασίες που δεν έχουν μεγάλη χρονική διάρκεια.

Η ΜΕΘΟΔΟΣ ΤΟΥ ΚΙΝΟΥΜΕΝΟΥ ΜΕΣΟΥ

Εδώ, αντί να εκμεταλλευόμαστε όλα τα δεδομένα, εξαιρούμε εκείνα τα οποία κρίνεται ότι δεν είναι πλέον συναφή με τα τωρινά δεδομένα. Ειδικότερα, για την εκάστοτε χρονική στιγμή t, λαμβάνουμε υπόψη μόνο τα τελευταία n δεδομένα για να κάνουμε την πρόβλεψή μας:

$$F_{t+1} = \sum_{i=t-n+1}^{t} \frac{x_i}{n}.$$

Παρατηρούμε ότι πρόκειται για μια πρόβλεψη που εύκολα μπορεί να ενημερωθεί από περίοδο σε περίοδο. Το μόνο που χρειάζεται είναι η εξάλειψη της πρώτης παρατήρησης και η υποκατάστασή της από την πιο πρόσφατη.

Παράδειγμα: Έστω ότι τα δεδομένα των μηνιαίων πωλήσεων μιας εταιρείας σε περίοδο 5 μηνών είναι $x_1 = 15, x_2 = 18, x_3 = 12, x_4 = 17, x_5 = 13$. Αν θέλαμε να εφαρμόσουμε την μέθοδο του κινούμενου μέσου λ.χ. για n=2 θα παίρναμε τις προβλέψεις $F_3 = \frac{15+18}{2} = 16.5$, $F_4 = \frac{18+12}{2} = 15, F_5 = \frac{12+17}{2} = 14.5$. Παρακάτω φαίνονται τα δεδομένα μαζί με τις προβλέψεις της μεθόδου για n=2,3,4:

	$F_t\ (n=2)$	F_t (n=3)	F_t(n=4)	x_t
$t=1$	–	–	–	15
2	–	–	–	18
3	16.5	–	–	12
4	15	15	–	17
5	14.5	14	15.5	13
6	15	14	15	–

Πίνακας 1: Παράδειγμα εφαρμογής της μεθόδου κινούμενου μέσου, προβλέψεις για n=2,3,4

Η συγκεκριμένη μέθοδος συνδυάζει τα πλεονεκτήματα του να εστιάσουμε στις πιο πρόσφατες παρατηρήσεις και παράλληλα στο να υπολογίζεται ενός μέσος, μιας και χρησιμοποιεί πιο συναφή δεδομένα και ταυτόχρονα αξιοποιεί πολλαπλές παρατηρήσεις. Ωστόσο, ένα σημαντικό της μειονέκτημα είναι ότι δίνει αρκετά μεγάλο βάρος στις παρατηρήσεις x_{t-n+1}, x_t. Διαισθητικά, θα περιμέναμε ότι μια καλή μέθοδος δίνει περισσότερο βάρος στις πιο πρόσφατες παρατηρήσεις απ' ό,τι τις παλαιότερες, για να αντιπροσωπεύει πιστά τις τρέχουσες συνθήκες. Σε αυτό χρησιμεύει η ακόλουθη μέθοδος.

ΜΕΘΟΔΟΣ ΕΚΘΕΤΙΚΗΣ ΕΞΟΜΑΛΥΝΣΗΣ

Σε αυτή τη μέθοδο χρησιμοποιούμε τον τύπο:

$$F_{t+1} = \alpha x_t + (1-\alpha)F_t,$$

όπου α (0<α<1) η **σταθερά εξομάλυνσης.** Έτσι, η πρόβλεψη είναι ένα σταθμισμένο άθροισμα της πιο πρόσφατης πρόβλεψης και της αμέσως προηγούμενης πρόβλεψης για την περίοδο που

μόλις τελείωσε. Με συγκεκριμένες επιλογές του α βλέπουμε ότι παίρνουμε μεθόδους τις οποίες ήδη έχουμε επισημάνει. Αν, για παράδειγμα, λάβουμε α=1 τότε $F_{t+1} = x_t$, δηλαδή προκύπτει η μέθοδος της τελευταίας τιμής, ενώ αν έχουμε $\alpha = \frac{1}{N}$ με N το μήκος (δηλ. το χρονικό ορίζοντα) της χρονοσειράς, τότε προκύπτει η μέθοδος μέσης πρόβλεψης. Από εκεί και έπειτα, μπορούμε να επιλέξουμε οποιαδήποτε άλλη τιμή μεταξύ του 0 και του 1 για να «τρέξουμε» τη μέθοδο. Συνήθως το α μπορεί να επιλέγει με κάποιο κριτήριο βελτιστοποίησης, συγκεκριμένα έτσι ώστε να ελαχιστοποιεί το αντίστοιχο μέτρο για το σφάλμα (τα οποία θα αναπτύξουμε παρακάτω).

Παρατηρούμε, τώρα, ότι λόγω της αναδρομικής σχέσης μεταξύ των F_t, F_{t+1} ισχύει:

$$F_{t+1} = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha^2)x_{t-2} + \cdots.$$

Σε αυτήν τη μορφή, είναι σαφές ότι πρωτίστως το βάρος «πέφτει» στην x_t και σταδιακά το βάρος μειώνεται καθώς προχωράμε σε παλαιότερες τιμές. Επιπλέον, η πρώτη μορφή έχει το προτέρημα ότι δεν χρειάζεται να διατηρηθούν οι τιμές πριν το χρόνο t, αλλά χρειάζεται μόνο η τρέχουσα τιμή x_t και η προηγούμενη πρόβλεψη F_t.

Μια εναλλακτική εφαρμογή αυτής της τεχνικής είναι μέσω του τύπου:

$$F_{t+1} = F_t + \alpha(x_t - F_t).$$

Η τελευταία σχέση έχει μια διαισθητικά εύκολα αντιληπτή ιδέα: για να κάνουμε την επόμενη πρόβλεψη, χρησιμοποιούμε την αμέσως προηγούμενη και ύστερα διορθώνουμε με το σφάλμα της εκτίμησης στο χρόνο t προσαρμόζοντάς το με κατάλληλο συντελεστή α. Συνήθως αυτή η εναλλακτική μορφή είναι ευκολότερη στη χρήση.

Για να πειστούμε για την αποτελεσματικότητα αυτής της τεχνικής, μπορούμε να δούμε ότι υποθέτοντας την πλήρη σταθερότητα των τιμών, ώστε οι $X_1, X_2, \ldots$ είναι ανεξάρτητες και ισόνομες με διασπορά σ^2, έπεται ότι (για μεγάλο t) :

$$var[F_{t+1}] \approx \frac{\alpha\sigma^2}{2-\alpha},$$

δηλαδή η διασπορά είναι –στατιστικά- ισοδύναμη με αυτή που θα πρόκυπτε απ' τον κινούμενο μέσο με $(2-\alpha)/\alpha$ παρατηρήσεις. Αν, λοιπόν, επιλέξουμε λ.χ. το α να είναι 0.1, τότε (2-α)/α=19. Έτσι, από θέμα διασποράς, η εκθετική εξομάλυνση με τιμή α είναι ισοδύναμη με τη μέθοδο κινούμενη μέσου που χρησιμοποιεί 19 παρατηρήσεις. Ωστόσο, αν υπάρξει μια αλλαγή (π.χ. αύξηση του μέσου), η εκθετική εξομάλυνση θα ανταποκριθεί γρηγορότερα στην ανίχνευσή της απ' ό,τι ο κινούμενος μέσος.

Ένα σημαντικό ελάττωμα της μεθόδου, ωστόσο, είναι ότι καθυστερεί όταν υπάρχει μια συνεχόμενη τάση στις παρατηρήσεις. Αν, δηλαδή, είναι εσφαλμένο το μοντέλο σταθερού επιπέδου και ο μέσος αυξάνει σταθερά, η πρόβλεψη θα μείνει μερικές περιόδους πίσω. Ωστόσο, μπορεί να γίνει κατάλληλη προσαρμογή ώστε να είναι τελικά αποτελεσματική η μέθοδος και σε αυτή την περίπτωση.

Έπειτα, άλλο πρόβλημα της μεθόδου αυτής είναι η επιλογή του α. Η εκθετική εξομάλυνση πρόκειται, ουσιαστικά, για ένα στατιστικό φίλτρο που επεξεργάζεται δεδομένα δίνοντας έναν χρονομεταβλητό μέσο. Αν το α είναι πολύ μικρό, η ανταπόκριση σε αλλαγές είναι αργή, ενώ αν το α είναι υπερβολικά μεγάλο, ναι μεν είναι γρήγορα η ανταπόκριση σε αλλαγές, αλλά είναι ευμετάβλητα τα αποτελέσματα. Πρέπει, λοιπόν, να υπάρξει ένας συμβιβασμός ανάμεσα στα δυο άκρα, ανάλογα με την ευστάθεια της διαδικασίας. Υπάρχει γενική σύσταση το α να μην υπερβαίνει το 0.3 και, συνήθως, μια εύλογη επιλογή είναι κοντά στο 0.1. Αυτή η τιμή μπορεί να αυξηθεί αν αναμένεται μια προσωρινή αλλαγή στη διαδικασία

ή όταν βρισκόμαστε σε αρχικό στάδιο στις πρόβλεψης. Στην αρχή, μια εύλογη προσέγγιση είναι η πρόβλεψη στη $2^{η}$ περίοδο με τη σχέση:

$$F_2 = \alpha x_1 + (1-\alpha)(\alpha\rho\chi\iota\kappa\acute{\eta}\ \epsilon\kappa\tau\acute{\iota}\mu\eta\sigma\eta),$$

όπου χρησιμοποιείται κάποια αρχική εκτίμηση του σταθερού επιπέδου Α. Αν υπάρχουν παλαιότερα δεδομένα διαθέσιμα, μπορεί να χρησιμοποιηθεί ο μέσος τους για την επιθυμητή αρχική εκτίμηση.

Παράδειγμα: Ας υποθέσουμε πάλι τα ίδια δεδομένα με το παράδειγμα στο κινούμενο μέσο. Για μια χρονική στιγμή για την οποία θέλουμε να κάνουμε πρόβλεψη, έχοντας επιλέξει σταθερά εξομάλυνσης α, ο τύπος για την πρόβλεψη –όπως ήδη αναφέραμε- είναι $F_{t+1} = ax_t + a(1-a)x_{t-1} + \cdots + a(1-a)^{t-1}x_1$. Επομένως, για τα εν λόγω δεδομένα και λ.χ. α=0.1 παίρνουμε (αν $F_2 = x_1 = 15$) :

$$F_3 = 0.1\,F_2 + 0.9\,x_2 = 0.1 \cdot 15 + 0.9 \cdot 18 = 17.7,$$
$$F_4 = 0.1 \cdot 17.7 + 0.9 \cdot 12 = 12.57,$$
$$F_5 = 0.1 \cdot 12.57 + 0.9 \cdot 17 = 16.557$$
$$F_6 = 0.1 \cdot 16.557 + 0.9 \cdot 13 = 13.3557$$

Παρακάτω βλέπουμε τα δεδομένα μαζί με τις προβλέψεις τις μεθόδου εκθετικής εξομάλυνσης για $a = 0.1, a = 0.2$ και $a = 0.3$:

	$F_t\ (a = 0.1)$	$F_t\ (a = 0.2)$	$F_t(a = 0.3)$	x_t
t=1	-	-	-	15
2	15	15	15	18
3	17.7	17.4	17.1	12
4	12.57	13.08	13.53	17
5	16.557	16.216	15.959	13
6	13.3557	13.6432	13.8877	-

Πίνακας 2: Παράδειγμα εφαρμογής της μεθόδους εκθετικής εξομάλυνσης για α=0.1, α=0.2 και α=0.3

ΕΠΟΧΙΑΚΕΣ ΕΠΙΔΡΑΣΕΙΣ ΣΕ ΜΟΝΤΕΛΑ ΠΡΟΒΛΕΨΗΣ

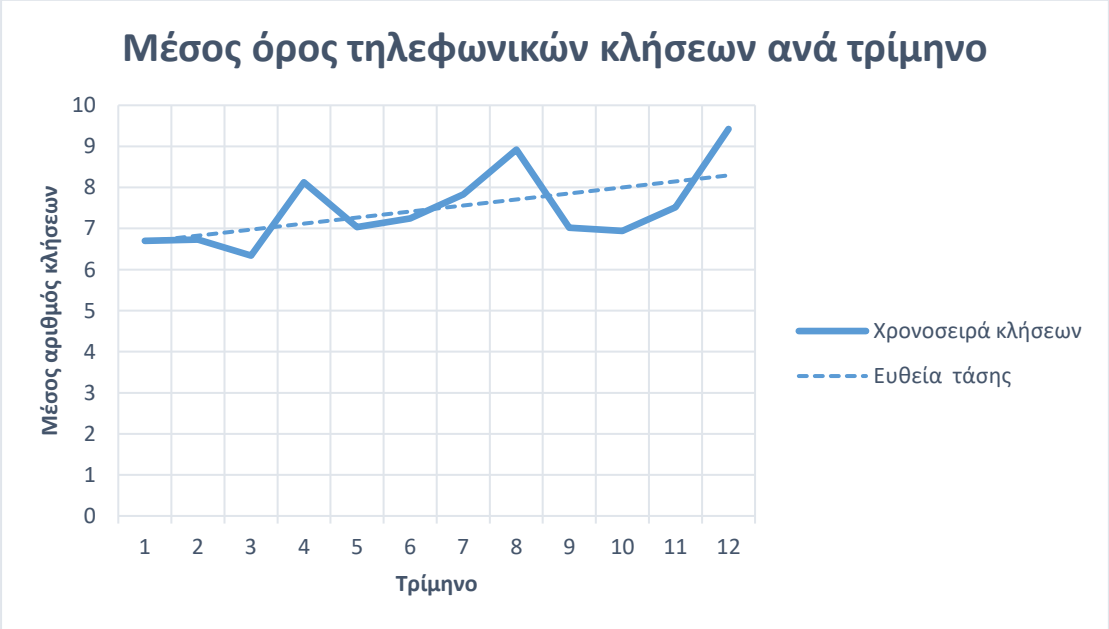
Είναι αρκετά συχνά φαινόμενο μια χρονοσειρά να εμφανίζει ένα εποχιακό μοτίβο με υψηλότερες τιμές κάποια περίοδο ενός έτους σε σχέση με τις υπόλοιπες. Αυτό συμβαίνει, παραδείγματος χάριν, στις πωλήσεις προϊόντων που είναι δημοφιλή ως επιλογές χριστουγεννιάτικων δώρων. Σε μια τέτοια περίπτωση, παραβιάζεται η υπόθεση του μοντέλου σταθερού επιπέδου, οπότε οι μέχρι τώρα παρουσιασθείσες μέθοδοι πρόβλεψης δεν μπορούν να εφαρμοστούν. Είναι δυνατό, ωστόσο, να γίνουν κάποιες προσαρμογές στη χρονοσειρά ώστε να μπορεί, τελικά, να εφαρμοστεί το μοντέλο σταθερού επιπέδου. Θα αξιοποιήσουμε ένα παράδειγμα για να περιγράψουμε τον τρόπο με τον οποίο επιτυγχάνεται αυτό:

Ας υποθέσουμε ότι έχουμε μια εταιρία που πουλάει προϊόντα συναφή με τους Η/Υ σε τιμές ευκαιρίας λαμβάνοντας τηλεφωνικές παραγγελίες απευθείας από πελάτες στο τηλεφωνικό της κέντρο. Τα δεδομένα ακολούθως δίνουν τους μέσους αριθμούς κλήσεων ανά ημέρα στα τέσσερα τρίμηνα μιας περιόδου 3 ετών:

Έτος	Τετράμηνο	Μέσος αριθμός κλήσεων
1	1	6.701
1	2	6.732
1	3	6.340
1	4	8.125
2	1	7.034
2	2	7.245
2	3	7.830
2	4	8.924
3	1	7.021
3	2	6.938
3	3	7.521
3	4	9.423

Πίνακας 3: δεδομένα μέσου αριθμού κλήσεων για περίοδο 3 ετών

Το γράφημα που προκύπτει είναι το ακόλουθο:



Γράφημα 2: Χρονοσειρά μέσων κλήσεων ανά τρίμηνο

Είναι εμφανές ότι υπάρχει σημαντική ανοδική τάση στο τέταρτο τρίμηνο, λόγω της περιόδου των Χριστουγέννων. Επίσης, παρατηρείται μια ελαφρώς υψηλότερη τιμή στο 3ο τρίμηνο εν σχέση με τα τρίμηνα 1, 2 λόγω αντίστοιχων προσφορών για την έναρξη της σχολικής χρονιάς, γεγονός που είναι εμφανές και από την διακεκομμένη ευθεία του παραπάνω διαγράμματος.

Για την ποσοτική ανάλυση των δεδομένων, παρουσιάζουμε τους μέσους στον ακόλουθο πίνακα:

Τρίμηνο	Τριετής Μέσος Όρος	Παράγοντας εποχικότητας
1	6.9187	$\frac{6.9187}{7.486} = 0.924$
2	6.9717	$\frac{6.9717}{7.486} = 0.931$
3	7.23	$\frac{7.23}{7.486} = 0.966$
4	8.824	$\frac{8.824}{7.486} = 1.178$

Πίνακας 4: Μέσος αριθμός κλήσεων, Παράγοντας εποχικότητας

Στην πρώτη στήλη έχουμε το τρίμηνο και στη δεύτερη στήλη το μέσο που προκύπτει για την περίοδο των τριών ετών που έχουμε παρακολουθήσει τις κλήσεις. Υπολογίζουμε, ύστερα, τον συνολικό μέσο που είναι $\frac{6.9187+6.9717+7.23+8.824}{4} = 7.486$ και βρίσκουμε τους εποχιακούς παράγοντες (seasonal factors) ως το πηλίκο του μέσου του αντίστοιχου τριμήνου προς τον συνολικό μέσο. Από τους υπολογισμούς μας επιβεβαιώνονται και τα αντίστοιχα συμπεράσματα που εξάγαμε από το γράφημα: υπάρχει μια ανοδική πορεία στους παράγοντες εποχικότητας αφενός, αφετέρου ο αντίστοιχος παράγοντας για το 4ο τρίμηνο είναι σαφώς μεγαλύτερος των υπολοίπων, ξεπερνώντας τη μονάδα, υποδεικνύοντας ότι υπάρχει σημαντική διαφορά κατά τη συγκεκριμένη περίοδο στον αριθμό των κλήσεων εν σχέσει με τις υπόλοιπες.

Η ΕΠΟΧΙΑΚΑ ΠΡΟΣΑΡΜΟΣΜΕΝΗ ΧΡΟΝΟΣΕΙΡΑ

Είναι ευκολότερο να αναλύσουμε μια χρονοσειρά και να εντοπίσουμε τις τάσεις της αν τα δεδομένα είναι πρώτα προσαρμοσμένα ώστε να εξαλειφθεί το όποιο εποχιακό μοτίβο. Για να επιτευχθεί αυτό, λαμβάνουμε την προσαρμοσμένη χρονοσειρά με τον εξής τρόπο: παίρνουμε τις τιμές της χρονοσειράς (πχ. στο προηγούμενο παράδειγμα, ο όγκος των κλήσεων στο τηλεφωνικό κέντρο) και τις διαιρούμε με τον αντίστοιχο εποχιακό παράγοντα. Τα αντίστοιχα νούμερα φαίνονται και στον ακόλουθο πίνακα:

Παράγοντας εποχικότητας	Εποχιακά προσαρμοσμένες κλήσεις
0.924	7.2521645
0.931	7.23093448
0.966	6.563147
1.178	6.89728353
0.924	7.61255411
0.931	7.78195489
0.966	8.10559006
1.178	7.57555178
0.924	7.59848485
0.931	7.45220193
0.966	7.78571429
1.178	7.9991511

Πίνακας 5: μέσος αριθμός κλήσεων, Δεδομένα εποχιακά προσαρμοσμένης χρονοσειράς

με την προσαρμοσμένη χρονοσειρά να είναι όπως φαίνεται στο ακόλουθο γράφημα:

Γράφημα 2: Εποχιακά Προσαρμοσμένη Χρονοσειρά

Αυτό που πετύχαμε με αυτή την καινούργια χρονοσειρά είναι να εκφράσουμε πώς θα ήταν ο όγκος των τηλεφωνικών κλήσεων σε συγκεκριμένες περιόδους του έτους αν «απλωνόταν» ομοιόμορφα μέσα στο χρόνο. Συγκρίνοντας τα διαγράμματα πριν και μετά την προσαρμογή των δεδομένων, βλέπουμε ότι παρότι ανά έτος έχουμε αποκλείσεις, είναι λιγότερες οι διακυμάνσεις που δημιουργούνται λόγω τριμήνου στο τελικό/προσαρμοσμένο γράφημα. Είναι σαφές, κατά τ ’άλλα, ότι υπάρχουν και τυχαίες επιδράσεις στο μοντέλο, οι οποίες όμως μπορούν κάλλιστα να μελετηθούν μέσα από το τελευταίο γράφημα, έχοντας ξεχάσει πλέον την εποχιακή επίδραση.

Η ΓΕΝΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

Έχοντας κάνει προσαρμογή των δεδομένων αναφορικά με τις εποχιακές επιδράσεις, μπορούμε να εφαρμόσουμε την οποιαδήποτε μέθοδο πρόβλεψης θέλουμε από εκείνες που είδαμε στο μοντέλο σταθερού επιπέδου. Ακολούθως σκιαγραφούμε τη γενική διαδικασία που εφαρμόζουμε:

1. Χρησιμοποιούμε τη σχέση $\frac{\text{πραγματική τιμή}}{\text{εποχιακός παράγοντας}}$ για να φτιάξουμε τις τιμές της προσαρμοσμένης χρονοσειράς.
2. Επιλέγουμε μια μέθοδο πρόβλεψης για τη χρονοσειρά.
3. Εφαρμόζουμε τη μέθοδο για να πάρουμε μια πρόβλεψη για την επόμενη εποχιακά προσαρμοσμένη τιμή (ή τις επόμενες, αντίστοιχα τιμές).
4. Πολλαπλασιάζουμε την πρόβλεψη με τον εποχιακό παράγοντα για να πάρουμε την πρόβλεψη για την πραγματική τιμή.

ΜΙΑ ΜΕΘΟΔΟΣ ΕΚΘΕΤΙΚΗΣ ΛΕΙΑΝΣΗΣ ΓΙΑ ΜΟΝΤΕΛΟ ΜΕ ΓΡΑΜΜΙΚΗ ΤΑΣΗ

Στις μεθόδους πρόβλεψης που αναλύσαμε έως τώρα, υποθέσαμε ότι οι τυχαίες μεταβλητές $\{X_1, X_2, \dots, X_t, X_{t+1}\}$ παράγουν μια χρονοσειρά με σταθερή μέση τιμή ίση με Α, με το βασικό στόχο να είναι η πρόβλεψη F_{t+1} να προσεγγίζει όσα καλύτερα γίνεται το Α. Ωστόσο, είδαμε και νωρίτερα ότι υπάρχουν περιπτώσεις κατά τις οποίες σημειώνεται μια σταθερά ανοδική πορεία στα δεδομένα. Έτσι, το να υποθέσουμε την ισχύ του μοντέλου σταθερού επιπέδου δεν αποδίδει σε μια τέτοια χρονοσειρά, μιας και θα έμενε πίσω από την αντίστοιχη ανοδική τάση της χρονοσειράς.

Ας υποθέσουμε ότι η γενέτειρα διαδικασία τις παρατηρούμενης χρονοσειράς αναπαρίσταται μέσω μια γραμμικής εξάρτησης απ' το χρόνο σε συνδυασμό με τυχαίες επιδράσεις. Θεωρούμε, λοιπόν, την εξής δομή στη χρονοσειρά:

$$X_i = A + Bi + e_i, \qquad i = 1,2, \dots,$$

με X_i η τ.μ. της παρατήρησης στο χρόνο i, Α μια σταθερά, Β ο παράγοντας τάσης και e_i το τυχαίο σφάλμα την αντίστοιχη χρονική στιγμή (για το οποίο υποθέτουμε μηδενική μέση τιμή και σταθερή διασπορά).

Για μια χρονοσειρά που αναπαρίσταται με αυτό το μοντέλο, οι σχετικές υποθέσεις μπορεί να μην επαληθεύονται πλήρως. Συνήθως έχουμε κάποιες μικρές μεταβολές στις τιμές των Α,Β. Είναι σημαντικό, από κει και έπειτα, σε πρακτικό επίπεδο να εντοπίζουμε γρήγορα αυτές τις αλλαγές και να προσαρμόζουμε ανάλογα τις προβλέψεις μας. Έτσι, προτιμάται γενικά μια μέθοδος πρόβλεψης που δίνει μεγαλύτερο βάρος σε πρόσφατες παρατηρήσεις και λίγο ή και καθόλου σε παλαιότερες παρατηρήσεις. Η ακόλουθη μέθοδος εκθετικής εξομάλυνσης είναι σχεδιασμένη για αυτόν ακριβώς το σκοπό.

ΠΡΟΣΑΡΜΟΓΗ ΤΗΣ ΕΚΘΕΤΙΚΗΣ ΕΞΟΜΑΛΥΝΣΗΣ ΣΤΟ ΜΟΝΤΕΛΟ ΜΕ ΓΡΑΜΜΙΚΗ ΤΑΣΗ

Η εκθετική εξομάλυνση δύναται να προσαρμοστεί και σε αυτό το μοντέλο (η αντίστοιχη αυτή προσαρμογή αποκαλείται μέθοδος του Holt (Holt's method)). Έστω

$$T_{t+1}: \eta\ \varepsilon\kappa\theta\varepsilon\tau\iota\kappa\acute{\eta}\ \lambda\varepsilon\acute{\iota}\alpha\nu\sigma\eta\ \sigma\tau\eta\nu\ \varepsilon\kappa\tau\acute{\iota}\mu\eta\sigma\eta\ \tau\sigma\upsilon\ \pi\alpha\rho\acute{\alpha}\gamma\sigma\nu\tau\alpha\ \tau\acute{\alpha}\sigma\eta\varsigma\ B\ \tau\eta\ \chi\rho\sigma\nu\iota\kappa\acute{\eta}\ \sigma\tau\iota\gamma\mu\acute{\eta}$$
$$t\ +1, \delta\sigma\theta\acute{\varepsilon}\nu\tau\omega\nu\ \tau\omega\nu\ X_1 = x_1, X_2 = x_2, \dots, X_t = x_t.$$

Δοθέντος του T_{t+1}, η πρόβλεψη της τιμής της χρονοσειράς το χρόνο $t+1$ (F_{t+1}) προκύπτει απλώς προσθέτοντας το T_{t+1} στον τύπο που ήδη δώσαμε νωρίτερα για την εκθετική εξομάλυνση, δηλαδή:

$$F_{t+1} = \alpha x_t + (1-\alpha)F_t + T_{t+1}$$

όπου, πάλι, το α είναι μια παράμετρος η οποία κυμαίνεται από 0 έως 1.

Για να αποσαφηνίσουμε την ιδέα πίσω από αυτό, παρατηρούμε ότι βάσει των υποθέσεων του μοντέλου είναι $B = E[X_{i+1}] - E[X_i], \forall i = 1,2, \dots$. Έτσι, ο τυπικός στατιστικό εκτιμητής για το Β είναι ο μέσος όρος των διαφορών $x_2 - x_1, x_3 - x_2, \dots, x_t - x_{t-1}$. Ωστόσο, η εκθετική εξομάλυνση αναγνωρίζει ότι οι παράμετροι της στοχαστικής διαδικασίας που γεννά τη χρονοσειρά (μαζί και οι Α,Β) μπορεί να έχουν μια σταδιακή μεταβολή με το πέρασμα του

χρόνου ώστε οι πρόσφατες παρατηρήσεις να είναι οι πιο αξιόπιστες για την εκτίμηση των τρεχουσών παραμέτρων. Έστω

L_{t+1}:η τελευταία τάση στο χρόνο $t+1$ ανάμεσα σε δυο τιμές (x_t, x_{t-1}) και στις τελευταίες 2 προβλέψεις (F_t, F_{t-1}).

Η σχέση για την εκθετική εξομάλυνση στο L_{t+1} είναι

$$L_{t+1} = \alpha(x_t - x_{t-1}) + (1-\alpha)(F_t - F_{t-1}).$$

Τότε:

$$T_{t+1} = \beta L_{t+1} + (1-\beta)T_t,$$

όπου β η σταθερά εξομάλυνσης της τάσης που, όπως και το α, είναι μεταξύ του 0 και του 1. Υπολογίζοντας τα L_{t+1}, T_{t+1} μπορούμε ύστερα να βρούμε την τιμή της F_{t+1}.

Για να μπορεί να ξεκινήσει η εφαρμογή αυτής της μεθόδου πρόβλεψης, πρέπει να έχουμε δυο αρχικές εκτιμήσεις πριν την έναρξη των προβλέψεων. Αυτές οι προβλέψεις είναι:

x_0: η αρχική εκτίμηση της αναμενόμενης τιμής της χρονοσειράς (A) αν οι συνθήκες πριν την έναρξη των προβλέψεων δεν αλλάξουν λόγω κάποιας τάσης των τιμών (ανοδική ή καθοδική)
και
T_1: η αρχική εκτίμηση της τάσης της χρονοσειράς (B) πριν την έναρξη των προβλέψεων.
Οι προβλέψεις που προκύπτουν στην αρχή θα είναι:

$$F_1 = x_0 + T_1$$

$$L_2 = \alpha(x_1 - x_0) + (1-\alpha)(F_1 - x_0)$$

$$T_2 = \beta\, L_2 + (1-\beta)T_1$$

$$F_2 = \alpha x_1 + (1-\alpha)F_1 + T_2.$$

Οι παραπάνω εξισώσεις χρησιμοποιούνται ύστερα για να εξάγουμε τις μεταγενέστερες προβλέψεις.

Παράδειγμα: Ακολουθούν τα δεδομένα που αφορούν τα ετήσια νούμερα επιβατών (σε εκατομμύρια) της Australian airlines τα χρόνια 1998-2016. Έχουμε επιλέξει τις παραμέτρους εξομάλυνσης $\alpha = 0.8321$ και $\beta = 0.0001$. Τα δεδομένα, μαζί με τις προβλέψεις της μεθόδου, φαίνονται στον ακόλουθο πίνακα:

	t	x_t	T_t	L_t	F_t
1989	0		15.57	2.102	
1990	1	17.55	17.57	2.102	17.67
1991	2	21.86	21.49	2.102	19.67
1992	3	23.89	23.84	2.102	23.59
1993	4	26.93	26.76	2.102	25.94
1994	5	26.89	27.22	2.102	28.87
1995	6	28.83	28.91	2.102	29.32
1996	7	30.08	30.24	2.102	31.02
1997	8	30.95	31.18	2.102	32.34
1998	9	30.19	30.71	2.102	33.29

1999	10	31.58	31.79	2.102	32.81
2000	11	32.58	32.80	2.101	33.89
2001	12	33.48	33.72	2.101	34.90
2002	13	39.02	38.48	2.102	35.82
2003	14	41.39	41.25	2.102	40.58
2004	15	41.60	41.89	2.101	43.36
2005	16	44.66	44.55	2.102	44.00
2006	17	46.95	46.90	2.102	46.65
2007	18	48.73	48.78	2.102	49.00
2008	19	51.49	51.39	2.102	50.88
2009	20	50.03	50.61	2.101	53.49
2010	21	60.64	59.31	2.102	52.71
2011	22	63.36	63.03	2.102	61.41
2012	23	66.36	66.15	2.102	65.13
2013	24	68.20	68.21	2.102	68.26
2014	25	68.12	68.49	2.102	70.31
2015	26	69.78	69.92	2.102	70.59
2016	27	72.60	72.50	2.102	72.02

Πίνακες 6: δεδομένα ετήσιου αριθμού επιβατών τα χρόνια 1998-2016 και προβλέψεις με τη μέθοδο Holt

Τα δεδομένα του πίνακα προέκυψαν ως εξής: θεωρήσαμε, αρχικά, ως αρχικές εκτιμήσεις για τα T, L (το επίπεδο και την τάση της χρονοσειράς αντίστοιχα) $T_0 = 15.57$ και $L_0 = 2.102$. Στη συνέχεια, για να προχωρήσουμε στην επόμενη εκτίμηση του επιπέδου και της τάσης κατά την επόμενη χρονική περίοδο, αξιοποιούμε τις προηγούμενες αριθμητικές τιμές μαζί με τη παρατήρηση στο χρόνο $t = 1$. Έτσι, έχουμε

$$T_1 = \alpha\, x_1 + (1-\alpha)(T_0 + L_0) = 0.8321 \cdot 17.55 + (1 - 0.8321) \cdot (15.57 + 2.102) = 17.57$$

και ύστερα

$$L_1 = \beta(T_1 - T_0) + (1-\beta)L_0 = 0.0001(17.57 - 15.57) + 0.9999 \cdot 2.102 = 2.102.$$

Έπειτα, συνεχίζουμε αναδρομικά για να βρούμε τα L, T για τις μετέπειτα χρονικές στιγμές. Οι αντίστοιχες προβλέψεις στην εκάστοτε χρονική στιγμή θα είναι $F_t = T_{t-1} + L_{t-1}$. Έτσι, για παράδειγμα, έχουμε

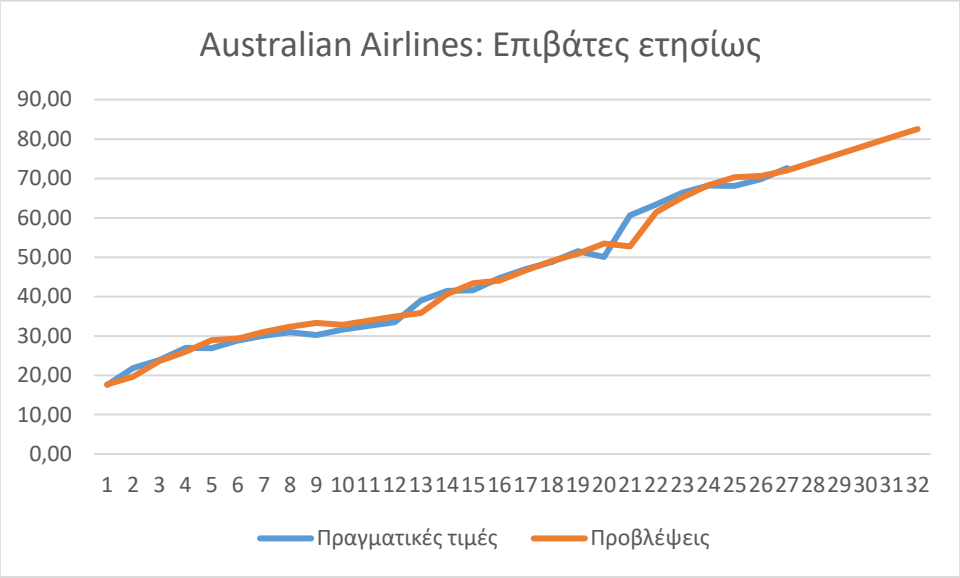
$$F_1 = T_0 + L_0 = 15.57 + 2.102 = 17.67, F_2 = T_1 + L_1 = 17.57 + 2.102 = 19.67$$

κ.ο.κ.

Εάν, τώρα, θέλαμε να κάνουμε σχετικές προβλέψεις για το ετήσιο πλήθος των επιβατών για τα έτη 2017-2021, οι προβλέψεις θα προκύψουν ως $F_{27+h} = F_{27} + hL_{27}$, δηλαδή

$$F_{28} = 72.02 + 2.102 = 74.122\ (2017), F_{29} = 72.02 + 2 \cdot 2.102 = 76.224\ (2018),\ F_{30} = 72.02 + 3 \cdot 2.102 = 78.326 \text{ κ.ο.κ.}$$

Παρακάτω φαίνεται η γραφική απεικόνιση των όσων βλέπουμε στον παραπάνω πίνακα:



Γράφημα 3: Αναπαράσταση δεδομένων αριθμού επιβατών της Australian Airlines ανά έτος και προβλέψεις με βάση τη μέθοδο Holt

Όπως έχουμε ήδη αναφέρει, οι σταθερές εξομάλυνσης έχουν ως σκοπό να δίνουν μεγαλύτερη βαρύτητα στα πιο πρόσφατα δεδομένα, σε περίπτωση που παρουσιάζονται μεγάλες διακυμάνσεις καθώς προχωράμε στο χρόνο. Η παράμετρος α έχει αυτή τη χρησιμότητα σε σχέση με το επίπεδο της χρονοσειράς, ενώ το β σε σχέση με την κλίση της γραμμικής τάσης. Στο παράδειγμα, επιλέξαμε το β να είναι πολύ μικρό, γεγονός που πηγάζει απ' το ότι δε χρειάζεται να παρακολουθούμε τις πιο πρόσφατες αλλαγές στην κλίση, οπότε ουσιαστικά η κλίση των δεδομένων δεν αλλάζει ραγδαία στο χρόνο. Από την άλλη, έχει σημασία να επιλέξουμε σωστά το α για να δώσουμε το βάρος που πρέπει στα πιο πρόσφατα δεδομένα.

ΠΡΟΒΛΕΨΕΙΣ ΓΙΑ (ΠΑΡΑΠΑΝΩ ΤΗΣ ΜΙΑΣ) ΕΠΟΜΕΝΕΣ ΠΕΡΙΟΔΟΥΣ

Μέχρι στιγμής, οι μέθοδοι που αναλύθηκαν παραπάνω αφορούσαν την πρόβλεψη στην αμέσως επόμενη χρονική περίοδο. Ωστόσο, υπάρχουν περιπτώσεις που είναι απαραίτητη η πρόβλεψη παρακάτω στο μέλλον. Είναι σημαντικό, λοιπόν, να απαντήσουμε στο ερώτημα του πώς μπορούμε –ανάλογα στο κάθε μοντέλο- να κάνουμε τέτοιου είδους προβλέψεις.

Στην περίπτωση των μεθόδων για το μοντέλο σταθερού επιπέδου, η πρόβλεψη F_{t+1} για την αμέσως επόμενη περίοδο είναι κατάλληλη και για οποιαδήποτε άλλη μεταγενέστερη περίοδο. Ωστόσο, αν υπάρχει τάση στα δεδομένα, πρέπει να ληφθεί υπόψη στην εκτίμησή μας. Στην περίπτωση της εκθετικής εξομάλυνσης, υπάρχει τρόπος να γίνει άμεσα αυτή η προσαρμογή. Ειδικότερα, αν T_{t+1} η εκτίμηση της τάσης στην περίοδο t+1, τότε η αντίστοιχη πρόβλεψη για n περιόδους αργότερα στο μέλλον θα είναι:

$$F_{t+n} = \alpha x_t + (1-\alpha)F_t + nT_{t+1}.$$

ΕΚΘΕΤΙΚΗ ΕΞΟΜΑΛΥΝΣΗ ΚΑΙ ΕΠΟΧΙΑΚΟΤΗΤΑ – ΜΕΘΟΔΟΣ WINTER

Υπάρχει, εκτός των άλλων, η δυνατότητα να συνδυάσουμε την εποχικότητα με την εκθετική εξομάλυνση. Η μέθοδος αυτή, η οποία καλείται και μέθοδος του **Winter** απαιτεί τον ορισμό δύο ποσοτήτων: η πρώτη είναι το c= ο αριθμός των περιόδων στο μήκος του εποχιακού μοτίβου (για τριμηνιαία δεδομένα λ.χ. c=4 περίοδοι, μία κάθε τρίμηνο και για μηνιαία c=12 περίοδοι, μία για κάθε μήνα) και η δεύτερη είναι η s_t: η εκτίμηση ενός εποχιακού πολλαπλασιαστικού παράγοντα για τον μήνα t, μετά την παρατήρηση της x_t. Για παράδειγμα, έστω ότι ο Ιούλιος είναι ο 7ος μήνας και $s_7 = 2$. Τότε, παρατηρώντας τις τιμές εκείνο το μήνα, εκτιμούμε ότι οι τιμές της χρονοσειράς θα είναι οι διπλάσιες απ' αυτές από το μέσο/τυπικό μήνα. Αν λ.χ. μελετάμε πωλήσεις ενός προϊόντος, ο 24ος μήνας είναι ο Δεκέμβριος και $s_{24} = 0.4$ τότε κατόπιν 24 μηνών παρατηρήσεων των πωλήσεων, προβλέπουμε ότι οι πωλήσεις εκείνο το Δεκέμβριο θα έχουν έρθει στο 40% των αναμένων πωλήσεων σε ένα μέσο μήνα. Στα παρακάτω, οι ποσότητες L_t, T_t έχουν την ίδια έννοια με αυτή που είχαν και στη μέθοδο Holt. Με βάση τις παρακάτω εξισώσεις, ανανεώνονται κάθε φορά οι τιμές των L_t, T_t *και* s_t. Πάλι, οι α,β,γ, είναι σταθερές εξομάλυνσης, κάθε μια εκ των οποίων κυμαίνεται από 0 έως 1. Οι σταθερές αυτές αποσκοπούν στην εξομάλυνση του επιπέδου, της κλίσης και της εποχικότητας αντίστοιχα. Κάθε μια εξ αυτών προσδιορίζει για το αντίστοιχο μέγεθος απ' τα τρία τι βάρος χρειάζεται να δοθεί στις πιο πρόσφατες παρατηρήσεις. Όσο πιο μεγάλη η αντίστοιχη παράμετρος εξομάλυνσης, τόσο πιο σημαντική η «πρόσφατη ιστορία» των δεδομένων για το μέγεθος που της αναλογεί (επίπεδο χρονοσειράς, κλίση ή εποχικότητα) και τόσο πιο ραγδαία η μεταβολή του.

Ειδικότερα, οι μαθηματικές σχέσεις που μας δίνουν το επίπεδο της χρονοσειράς, την κλίση και την εποχικότητα κατά τη χρονική στιγμή t είναι οι εξής:

$$L_t = \alpha \frac{x_t}{s_{t-c}} + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

$$s_t = \frac{\gamma x_t}{L_t} + (1-\gamma)s_{t-c}.$$

Η πρώτη εκ των τριών ανανεώνει την εκτίμηση της βάσης της σειράς παίρνοντας έναν σταθμισμένο μέσο των ακόλουθων δυο ποσοτήτων:

- $L_{t-1} + T_{t-1}$, που είναι η εκτίμηση του βασικού επιπέδου πριν την παρατήρηση της x_t
- Η παρατήρηση της οποίας έχουμε αναιρέσει την εποχικότητα $\frac{x_t}{s_{t-c}}$, που είναι μια εκτίμηση της βάσης που παίρνουμε για την τωρινή περίοδο.

Η δεύτερη σχέση είναι ίδια με αυτή στη μέθοδο του Holt για την ενημέρωση της τάσης. Η τελευταία σχέση αλλάζει την εκτίμηση για τον t-οστό μήνα της εποχικότητας, παίρνοντας έναν σταθμισμένο μέσο των ακόλουθων δυο ποσοτήτων:

-Της πιο πρόσφατης εκτίμησης της εποχικότητας (s_{t-c})

-Της $\frac{x_t}{L_t}$, που είναι η εκτίμηση της εποχιακότητας του t-οστού μήνα, υπολογισμένη μέσα απ' τον τρέχοντα μήνα.

Στο τέλος της περιόδου t, η πρόβλεψη για τον μήνα t+k δίνεται από

$$f_{t,k} = (L_t + kT_t)s_{t+k-c}$$

Έτσι, η τιμή της πρόβλεψης για τη σειρά στην περίοδο t+k προκύπτει πολλαπλασιάζοντας τη βάση της t+k περιόδου, δηλαδή την $(L_t + kT_t)$ με την πιο πρόσφατη εκτίμηση για τον παράγοντα εποχικότητας s_{t+k-c}.

ΕΚΚΙΝΗΣΗ ΤΗΣ ΜΕΘΟΔΟΥ

Για να έχουμε καλή προσέγγιση μέσα απ' τη μέθοδο του Winter, πρέπει να έχουμε καλές εκτιμήσεις της βάσης, της τάσης και όλων των παραγόντων εποχικότητας. Ξεκινώντας τις προβλέψεις μας από τον Ιανουάριο ενός έτους, θα θεωρήσουμε ότι έχουμε δεδομένες εκτιμήσεις των παραγόντων εποχικότητας για τον καθένα από τους προηγούμενους μήνες (θεωρώντας για αυτές τις εκτιμήσεις τον Δεκέμβρη ως μήνα 0, τον Νοέμβρη ως μήνα -1, κ.ο.κ.) Έστω

$$L_0: \eta\ \varepsilon\kappa\tau\acute{\iota}\mu\eta\sigma\eta\ \tau\eta\varsigma\ \beta\acute{\alpha}\sigma\eta\varsigma\ \sigma\tau\eta\nu\ \alpha\rho\chi\acute{\eta}\ \tau o\upsilon\ 1o\upsilon\ \mu\acute{\eta}\nu\alpha$$

$$T_0 = \eta\ \varepsilon\kappa\tau\acute{\iota}\mu\eta\sigma\eta\ \tau\eta\varsigma\ \tau\acute{\alpha}\sigma\eta\varsigma\ \sigma\tau\eta\nu\ \alpha\rho\chi\acute{\eta}\ \tau o\upsilon\ 1o\upsilon\ \mu\acute{\eta}\nu\alpha$$

$$s_{-11} = \eta\ \varepsilon\kappa\tau\acute{\iota}\mu\eta\sigma\eta\ \tau o\upsilon\ \pi\alpha\rho\acute{\alpha}\gamma o\nu\tau\alpha\ \varepsilon\pi o\chi\iota\alpha\kappa\acute{o}\tau\eta\tau\alpha\varsigma\ \tau o\upsilon\ I\alpha\nu o\upsilon\alpha\rho\acute{\iota}o\upsilon\ \sigma\tau\eta\nu\ \alpha\rho\chi\acute{\eta}\ \tau o\upsilon\ 1o\upsilon\ \mu\acute{\eta}\nu\alpha$$

$$s_{-10} = \eta\ \varepsilon\kappa\tau\acute{\iota}\mu\eta\sigma\eta\ \tau o\upsilon\ \pi\alpha\rho\acute{\alpha}\gamma o\nu\tau\alpha\ \varepsilon\pi o\chi\iota\alpha\kappa\acute{o}\tau\eta\tau\alpha\varsigma\ \tau o\upsilon\ \Phi\varepsilon\beta\rho o\upsilon\alpha\rho\acute{\iota}o\upsilon\ \sigma\tau\eta\nu\ \alpha\rho\chi\acute{\eta}\ \tau o\upsilon\ 1o\upsilon\ \mu\acute{\eta}\nu\alpha$$

$$s_0 = \eta\ \varepsilon\kappa\tau\acute{\iota}\mu\eta\sigma\eta\ \tau o\upsilon\ \pi\alpha\rho\acute{\alpha}\gamma o\nu\tau\alpha\ \varepsilon\pi o\chi\iota\alpha\kappa\acute{o}\tau\eta\tau\alpha\varsigma\ \tau o\upsilon\ \Delta\varepsilon\kappa\varepsilon\mu\beta\rho\acute{\iota}o\upsilon\ \sigma\tau\eta\nu\ \alpha\rho\chi\acute{\eta}\ \tau o\upsilon\ 1o\upsilon\ \mu\acute{\eta}\nu\alpha$$

Για τις εκτιμήσεις των τελευταίων υπάρχει πληθώρα μεθόδων. Διαλέγουμε μια απλή μέθοδο η οποία απαιτεί τα δεδομένα δυο ετών. Ας υποθέσουμε ότι τα τελευταία 2 χρόνια οι πωλήσεις κλιματιστικών μιας εταιρίας ανά μήνα είναι οι εξής:

"$\Έτος - 2$"	5	2	13	16	24	28	39	42	26	14	12	13
"$\Έτος - 1$"	7	4	19	25	49	52	74	77	50	31	30	25

Πίνακας 7 Δεδομένα δυο προηγούμενων ετών για τη "ρύθμιση" της μεθόδου Winter

$$\Σύνολο\ πωλήσεων\ έτους - 2: 234, \qquad \Σύνολο\ πωλήσεων\ έτους - 1: 443$$

Εκτιμούμε, έπειτα:

$$T_0 = \frac{(\mu έσες\ πωλήσεις\ ανα\ μήνα\ στο\ έτος - 1) - (\mu έσες\ πωλήσεις\ ανα\ μήνα\ στο\ έτος - 2)}{12}$$

ή

$$T_0 = \frac{\frac{443}{12} - \frac{234}{12}}{12} = 1.4514$$

Για να εκτιμήσουμε το L_0, προσδιορίζουμε πρώτα τη μέση μηνιαία ζήτηση του έτους -1 ($\frac{443}{12}$). Αυτό εκτιμά τη βάση στο μέσο του έτους -1 (δηλ. στο μήνα 6,5 του έτους -1). Για να φέρουμε

αυτή την εκτίμηση στο τέλος του 12ου μήνα του έτους -1, προσθέτουμε $(12-6.5)T_0 = 5.5T_0$. Επομένως εκτιμούμε $L_0 = \frac{443}{12} + 5.5\ (1.4514) = 44.899$.

Για να εκτιμήσουμε τον παράγοντα εποχικότητας ενός δοθέντος μήνα (λ.χ. του Ιανουαρίου, οπότε θέλουμε το s_{-11}) παίρνουμε την εκτίμηση της εποχιακότητας του Ιανουαρίου τα έτη -2 και -1 και λαμβάνουμε το μέσο όρο. Η μέση μηνιαία ζήτηση του Ιανουαρίου το έτος -2 είναι $\frac{234}{12} = 19.5$, ενώ έχουμε 5 πωλήσεις σε κλιματιστικά στο έτος -2 τον Ιανούάριο, οπότε

$$Εκτίμηση\ εποχιακότητας\ του\ Ιανουαρίου\ του\ έτους - 2 = \frac{5}{19.5} = 0.256$$

Ομοίως,

$$Εκτίμηση\ εποχιακότητας\ του\ Ιανουαρίου\ του\ έτους - 1 = \frac{7}{36.917} = 0.190$$

Τέλος, έχουμε

$$s_{-11} = \frac{0.256 + 0.19}{2} = 0.223$$

Με όμοιο τρόπο παίρνουμε:

$$s_{-10} = 0.105, s_{-9} = 0.591, s_{-8} = 0.749, s_{-7} = 1.279, s_{-6} = 1.422, s_{-5} = 2.002,$$

$$s_{-4} = 2.112, s_{-3} = 1.344, s_{-2} = 0.779, s_{-1} = 0.714, s_0 = 0.672$$

Ένας έλεγχος για τους παραπάνω υπολογισμούς είναι ότι η αρχική εκτίμηση του παράγοντα εποχικότητας πρέπει να έχει μέσο όρο 1.

Έτσι, ο υπολογισμός μέσω της μεθόδου μας για α=0.5, β=0.4, γ=0.6 για τους πρώτους 12 μήνες πωλήσεων κλιματιστικών δίνει τα παρακάτω:

Πωλήσεις (πραγματικά δεδομένα)	L_t (ξεκινώντας από t=0)	T_t (ξεκινώντας από t=0)	s_t	$f_{t-1,1}$ (Προβλέψεις)	Σφάλμα	Απόλυτο σφάλμα
11	44.89931	1.451389	0.236201	10.33682	0.663181	0.663181
8	47.83756	2.046135	0.142523	5.260641	2.739359	2.739359
21	62.87159	7.241295	0.436676	41.41356	-20.4136	20.41356
34	52.83288	0.32929	0.685666	39.81085	-5.81085	5.810854
59	49.28235	-1.22264	1.229926	61.47037	-2.47037	2.47037
63	47.094	-1.60892	1.371545	64.6906	-1.6906	1.690598
88	44.89074	-1.84666	1.977092	86.18533	1.814675	1.814675
92	43.49724	-1.6654	2.116971	88.67566	3.324341	3.324341
61	42.61595	-1.35175	1.39638	55.45362	5.546376	5.546376

36	43.32779	-0.52632	0.810061	33.33545	2.664546	2.664546
34	44.51206	0.15792	0.743908	31.89494	2.105059	2.105059
26	46.14409	0.747561	0.606845	31.50808	-5.50808	5.508082
	42.79297	-0.89191				

Πίνακας 8 Υλοποίηση της μεθόδου Winter και προβλέψεις, α=0.5, β=0.4, γ=0.6

ΠΡΟΒΛΕΨΕΙΣ ΚΑΙ ΣΦΑΛΜΑΤΑ

Είδαμε ήδη αρκετές μεθόδους για την πραγματοποίηση προβλέψεων. Προφανώς, με κάθε μια απ' αυτές θα γίνονται κάποια αντίστοιχα σφάλματα. Σκοπός μας, λοιπόν, είναι να ποσοτικοποιήσουμε με κάποιον τρόπο αυτά τα σφάλματα και με βάση αυτό να αποφασίσουμε ποια μέθοδος είναι η κατάλληλη ώστε τα αντίστοιχα σφάλματα – με την έννοια που τα ορίσαμε εμείς- να είναι όσο το δυνατό μικρότερα.

Το *σφάλμα πρόβλεψης ή υπόλοιπο* την περίοδο t είναι η απόλυτη τιμή της απόκλισης της πρόβλεψης εκείνη την περίοδο από την πραγματική τιμή που υπολογίζουμε εκ των υστέρων για εκείνη την περίοδο. Δηλαδή είναι η ποσότητα:

$$E_t = |x_t - F_t|.$$

Δοθέντων των σφαλμάτων για n περιόδους, υπάρχουν δυο μέτρα για την επίδοση μιας μεθόδου. Το ένα είναι η μέση απόλυτη απόκλιση (Mean Absolute Deviation, M.A.D.) και ορίζεται ως ο δειγματικός μέσος των σφαλμάτων, δηλαδή:

$$MAD = \frac{\sum_{t=1}^{n} E_t}{n}.$$

Το άλλο αντίστοιχο μέτρο είναι το μέσο τετραγωνικό σφάλμα (Mean Square Error, MSE) το οποίο αποτελεί το μέσο όρο των τετραγώνων των σφαλμάτων, δηλαδή:

$$MSE = \frac{\sum_{t=1}^{n} E_t^2}{n}.$$

Η μέση απόλυτη απόκλιση έχει το προτέρημα ότι είναι πιο εύκολη στον υπολογισμό. Ωστόσο, το μέσο τετραγωνικό σφάλμα από την άλλη ποινικοποιεί πολύ τα μεγάλα σφάλματα στις προβλέψεις που μπορεί να έχουν επιζήμιες επιπτώσεις αν αγνοηθούν. Στην πράξη, οι διευθυντές συνήθως χρησιμοποιούν το MAD, ενώ οι στατιστικολόγοι προτιμούν το MSE εν γένει.

Αυτά τα μέτρα που μόλις αναφέραμε μπορούν να χρησιμοποιηθούν είτε για να συγκριθούν εναλλακτικές μέθοδοι πρόβλεψης ώστε να αποφασισθεί η καλύτερη δυνατή για το πρόβλημα ή για να παρακολουθούνται οι προβλέψεις που ήδη έχουν ξεκινήσει να γίνονται και να λαμβάνεται η απόφαση αν η αντίστοιχη μέθοδος αποδίδει ή πρέπει να αξιοποιηθεί κάποια άλλη.

Στο παράδειγμα πωλήσεων των κλιματιστικών, αν θέλαμε να υπολογίσουμε το μέσο απόλυτο σφάλμα, θα ήταν:

$$MAD = \frac{0.6631 + \cdots + 5.508}{12} = 3.086.$$

ΠΑΡΑΔΕΙΓΜΑΤΑ

(1) Στη συνέχεια θα εφαρμόσουμε τις παραπάνω μεθόδους που περιγράψαμε σε κάποια αριθμητικά δεδομένα. Στο παρόν παράδειγμα εξετάζουμε τα δεδομένα του Ελληνικού Οργανισμού Φαρμάκων για τις πωλήσεις φαρμάκων ανά μήνα (σε αξίες). Τα δεδομένα που έχουμε διαθέσιμα είναι από τις χρονιές 2019-2020. Στο ακόλουθο γράφημα φαίνεται η σχετική χρονοσειρά:

Γράφημα 4 Δεδομένα μηνιαίων πωλήσεων φαρμάκων στην Ελλάδα 2019-2020 (Πηγή: ΕΟΦ)

Προβλέψεις με κινούμενο μέσο:

Αν θέλαμε, λ.χ., να εφαρμόσουμε την μέθοδο του κινούμενου μέσου θεωρώντας N=3 για τη περίοδο των πρώτων πέντε μηνών, οι προβλέψεις που θα παίρναμε θα ήταν

$$f_4 = \frac{f_1 + f_2 + f_3}{3} = \frac{344089652.56 + 345865984.94 + 348846112.29}{3} = 346267249.9$$

Και

$$f_5 = \frac{f_2 + f_3 + f_4}{3} = \frac{345865984.94 + 348846112.29 + 373890297.51}{3} = 356200798.2$$

(παραπέμπουμε στο παράρτημα, στον πίνακα 3 για τα σχετικά δεδομένα).

Θα εξετάσουμε, στη συνέχεια, το σφάλμα αυτών των προβλέψεων και πώς αυτό επηρεάζεται από την επιλογή του N. Θα χρησιμοποιήσουμε την μέση απόλυτη απόκλιση (MAD). Αν, για παράδειγμα, δουλέψουμε στην ίδια περίοδο με πριν, τα αντίστοιχα υπόλοιπα θα είναι

$$e_4 = x_4 - f_4 = 27623047.58, e_5 = 1909951.09$$

οπότε το MAD θα δίνεται από την:

$$MAD = \frac{|e_4| + |e_5|}{2} = \frac{27623047.58 + 1909951.09}{2} = 14766499.335$$

Επομένως, με βάση αυτό το μέτρο σφάλματος, οι προβλέψεις για τις πωλήσεις φαρμάκων για την εξεταζόμενη περίοδο αποκλίνουν κατά 14766499.335 το μήνα (σαφώς αυτό το αποτέλεσμα είναι απόρροια του ότι το N είναι μικρό).

Εφαρμόζοντας τα παραπάνω δεδομένα για 24 μήνες, έχουμε τα εξής αποτελέσματα:

Μήνας (i)	Φαρμακεία, Φαρμακαποθήκες (Λιανική τιμή) 2019-2020 (x_i)	Εκτιμήσεις κινούμενου μέσου, N=3 (fi)	Σφάλματα (ei)
1	344089652.56		
2	345865984.94		
3	348846112.29		
4	373890297.51	346267249.9	27623047.58
5	358110749.34	356200798.2	1909951.09
6	341219966.94	360282386.4	19062419.44
7	477925386.16	357740337.9	120185048.23
8	239388639.72	392418700.8	153030061.09
9	381696392.68	352844664.3	28851728.41
10	432943230.27	366336806.2	66606424.08
11	373386567.53	351342754.2	22043813.31
12	375676848.64	396008730.2	20331881.52
13	375386551.06	394002215.5	18615664.42
14	371436590.19	374816655.7	3380065.55
15	459088624.13	374166663.3	84921960.83
16	330440325.98	401970588.5	71530262.48
17	324490229.85	386988513.4	62498283.58
18	364730783.19	371339726.7	6608943.46
19	446947280.25	339887113	107060167.24
20	240797550.01	378722764.4	137925214.42
21	429619680.44	350825204.5	78794475.96
22	410026600.78	372454836.9	37571763.88
23	399281965.26	360147943.7	3134021.52
24	405140820.85	412976082.2	7835261.31
		MAD=	**53120021.88**

Πίνακας 9 Ετήσιες πωλήσεις φαρμάκων 2019-2020, δεδομένα ετών 2017-2018 και προβλέψεις της μεθόδου κινούμενου μέσου, N=3

Προβλέψεις με μέθοδο Winter:

Με βάση το παραπάνω γράφημα, φαίνεται ότι το κατάλληλο μοντέλο για εφαρμογή σε αυτό το πρόβλημα είναι το μοντέλο εποχιακών επιδράσεων, μιας και λ.χ. το μήνα Αύγουστο (τους μήνες 8 και 20 στον οριζόντιο άξονα) φαίνεται να έχουμε σημαντική πτώση των πωλήσεων. Θα εφαρμόσουμε, όπως πριν, τη μέθοδο του Winter εκμεταλλευόμενοι τα δεδομένα των πωλήσεων φαρμάκων για τα έτη 2017-2018 (α=0.5, β=0.4, γ=0.6).

Ένας ενδεικτικός υπολογισμός για τους 2 πρώτους μήνες είναι ο εξής:

$$T_1 = \frac{\frac{342557930.74 + 354500479.34}{2} - \frac{338496384.65 + 327965330.86}{2}}{12} = 1274862.274$$

$$L_1 = \frac{342557930.74 + 354500479.34}{2} + 1274862.274 * 5.5 = 355540947.5$$

$$S_{-1} = \frac{\frac{338496384.65}{\frac{338496384.65 + 342557930.74}{2}} + \frac{342557930.74}{\frac{342557930.74 + 354500479.34}{2}}}{2} = 0.999334329$$

$$st_1 = 0.6 * \frac{[Φαρμακεία\ 2019 - 2020]}{L2} + (1 - 0.6) * s(-1) = 0.6 * \frac{344089652.56}{350567332.7} + (1 - 0.6) * 0.999334329 = 0.988647108$$

$$Προβλέψεις, 1 = (T1 + L1) * s(-1) = 356578287.9$$

$$T_2 = 0.4 * (L2 - L1) + (1 - 0.4) * T2 = -1224528.568$$

$$L_2 = 0.5 * \frac{[Φαρμακεία\ 2019 - 2020]}{s(-1)} + (1 - 0.5) * (T1 + L1) = 350567332.7$$

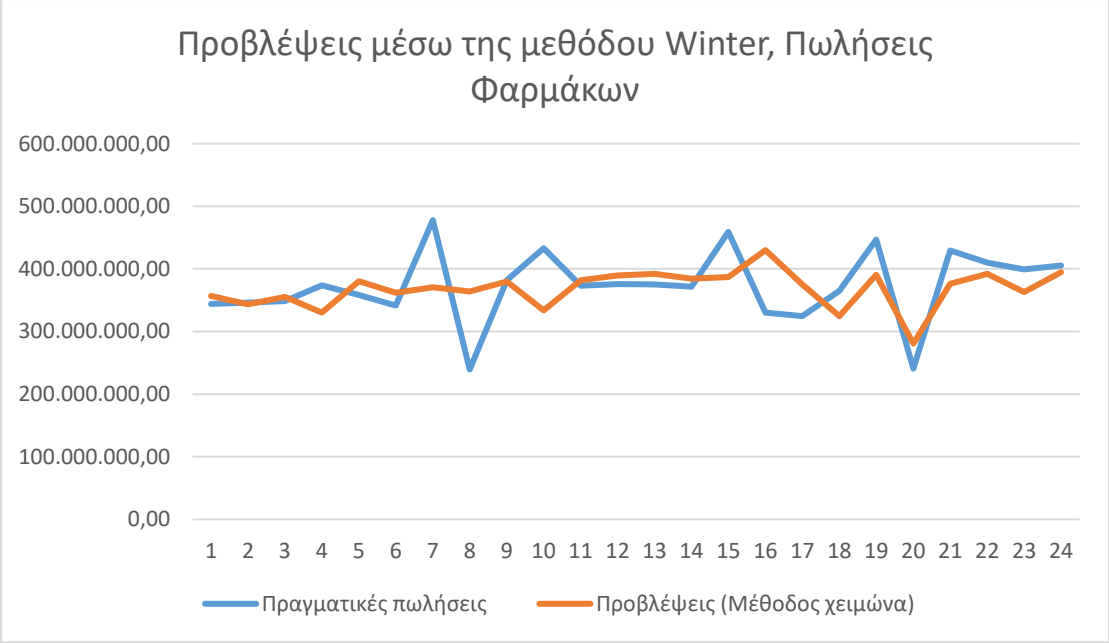
Συνεχίζοντας έτσι για τους υπόλοιπους μήνες των χρόνων 2019,2020 παίρνουμε τα ακόλουθα αποτελέσματα:

Μήνας	Φαρμακεία φαρμακαποθήκες (λιανική τιμή) 2019-2020	Φαρμακεία φαρμακαποθήκες (λιανική τιμή) 2017-2018	Ti	Li	$s(-i)$	st	Προβλέψεις
1	344089652.56	338496384.65	1274862.274	355540947.5	0.999334329	0.988647108	356578287.9
2	345865984.94	320086881.52	-1224528.568	350567332.7	0.983044424	0.985137903	343419495.6
3	348846112.29	346721864.93	-726791.2796	350587147.4	1.01528874	1.009783106	355209280.3
4	373890297.51	307588672.95	-1980260.904	346726682	0.957403927	0.99316926	330061577.4
5	358110749.34	328902979.00	7175481.811	367635777.9	1.014716941	0.996399728	380327334.9
6	341219966.94	323505019.60	2796608.315	363864076	0.987395522	0.969857118	362039117.9
7	477925386.16	395450530.61	-1420374.74	356118226.7	1.04463667	1.123972869	370530383.1
8	239388639.72	231110265.38	19140841.64	406100892.9	0.855375315	0.749559128	363741282.6
9	381696392.68	363907641.12	-9934732.721	352552798.6	1.108045015	1.109844025	379636239.8
10	432943230.27	360807072.03	-9562879.076	343547700	1.000056902	1.077463837	334003825.2

11	373386567.53	332201360.42	10223876.05	383451708.7	0.97024804	0.9636382	381962964.7
12	375676848.64	327965330.86	8455998.801	389255891.7	0.979149341	0.968596984	389419335.6
13	375386551.06	342557930.74	5648973.071	390694326.1			391843656.5
14	371436590.19	326147223.23	2319755.734	388020255.9			384538740.4
15	459088624.13	364877510.36	-340206.9264	383690105			387100250.7
16	330440325.98	310094793.89	13917978.43	418995361.4			429956221.3
17	324490229.85	340237269.40	-6122089.139	382813170.9			375334891.6
18	364730783.19	330630822.39	-16327764.68	351176892.9			324755810.6
19	446947280.25	419217024.18	-8084288.028	355457819.9			390438425.1
20	240797550.01	234580353.07	1970911.118	372511529.7			280696731.8
21	429619680.44	360724959.49	-8675128.746	347867341.2			376450450.2
22	410026600.78	374237224.86	906259.0902	363145682			392252801.2
23	399281965.26	356508803.87	4205450.949	372299920.8			362814958.9
24	405140820.85	354500479.34	11774060.49	395426895.6			394413618

Πίνακας 10 Ετήσιες πωλήσεις φαρμάκων 2019-2020, δεδομένα ετών 2017-2018 και προβλέψεις της μεθόδου Winter

με την αντίστοιχη γραφική παράσταση να είναι



Γράφημα 5: Αναπαράσταση δεδομένων πωλήσεων φαρμάκων και προβλέψεων με τη μέθοδο Winter (α=0.5, β=0.4, γ=0.6)

από όπου φαίνεται να έχουμε μια σχετικά ικανοποιητική εικόνα σε ποιοτικό επίπεδο, αφού οι τάσεις των πωλήσεων ανιχνεύονται από τη μέθοδο, παρότι σε κάποια σημεία έχουμε αποκλείσεις στις τιμές ποσοτικά, κάτι που φαίνεται και από το ότι το μέσο απόλυτο σφάλμα είναι $MAD = \sum_{i=1}^{24} \frac{|xi|-|fi|}{24} = 11607919.11$.

AD HOC ΠΡΟΒΛΕΨΕΙΣ

Ας υποθέσουμε ότι θέλουμε να προσδιορίσουμε πόσοι τραπεζικοί υπάλληλοι χρειάζεται να εργάζονται κάθε μέρα ούτως ώστε να είναι επαρκής η εξυπηρέτηση. Για να το επιτύχουμε αυτό, θα χρησιμοποιούσαμε την αντίστοιχη θεωρία ουρών αναμονής, η οποία χρειάζεται να γνωρίζουμε το πλήθος των πελατών που θα εισέλθουν στην τράπεζα ανά ημέρα. Επομένως, για να εκτιμήσουμε το πλήθος των υπαλλήλων που χρειάζονται, θα πρέπει να μπορούμε να κάνουμε πρόβλεψη για το πλήθος των πελατών. Ας υποθέσουμε, τώρα, στο παράδειγμά μας ότι ο διευθυντής της τράπεζας θεωρεί πως ο μήνας μέσα στο έτος και η μέρα της εβδομάδας επηρεάζουν το πλήθος των πελατών στην τράπεζα (η τράπεζα είναι ανοικτή από Δευτέρα μέχρι Σάββατο, εκτός αργιών). Μπορούμε να αναπτύξουμε ένα απλό μοντέλο πρόβλεψης για να εκτιμήσουμε το πλήθος πελατών που θα εμφανιστούν ανά ημέρα στην τράπεζα;

Για να το κάνουμε αυτό, έχουμε τα δεδομένα ενός έτους από τις αφίξεις πελατών ανά ημέρα στην τράπεζα (το παράδειγμα και οι σχετικοί πίνακες έχουν ληφθεί από το [4]). Χρησιμοποιούμε την κωδικοποίηση 1=Δευτέρα, 2= Τρίτη κ.ο.κ. Τα «Υ» στην «ΑΗ» του παρακάτω πίνακα σημαίνουν ότι πρόκειται για ημέρα που ακολουθεί μιας ημέρας που η τράπεζα ήταν κλειστή λόγω αργίας. Τα δεδομένα δίνονται στο Παράρτημα, στον Πίνακα 3.

Έστω $x_t =$ το πλήθος πελατών στην τράπεζα την ημέρα t. Θεωρούμε (ως, κατά μια έννοια, αξίωμα) ότι $x_t = B \times DW_t \times M_t \times \varepsilon_t$, όπου

$$B = το\ βασικό\ επίπεδο\ διέλευσης\ πελατών\ που\ αντιστοιχεί\ σε\ μια\ μέση\ ημέρα$$

DW_t

$$= ένας\ παράγοντας\ που\ αναλογεί\ στην\ ημέρα\ της\ εβδομάδας\ στην\ οποία\ αντιστοιχεί\ η\ ημέρα\ t$$

$$M_t = μηνιαίος\ παράγοντας\ που\ αναλογεί\ στο\ μήνα\ που\ αντιστοιχεί\ η\ μέρα\ t$$

$$\varepsilon_t = τυχαίο\ σφάλμα\ με\ μέσο\ όρο\ 1$$

Ακολουθεί ο πίνακας με τα σχετικά δεδομένα και τις αντίστοιχες προβλέψεις στην τελευταία στήλη:

Αρχικά, για εκτιμούμε το

$$B = μέσος\ όρος\ αφίξεων\ ανά\ ημέρα\ λειτουργίας\ της\ τράπεζας = 438.8111$$

Για την εκτίμηση του DW_t θα πάρουμε

$$DW_t\ για\ Δευτέρα = μέσος\ αριθμός\ αφίξεων\ τις\ Δευτέρες\ που\ η\ τράπεζα\ είναι\ ανοικτή$$

$$= \frac{495.8431}{438.8111} = 1.1299$$

Ομοίως υπολογίζουμε:

$$DW_t\ για\ Τρίτη = \frac{397.5849}{438.8111} = 0.906$$

$$DW_t\ για\ Τετάρτη = \frac{414.2745}{438.8111} = 0.9441$$

$$DW_t\ για\ Πέμπτη = \frac{414.9}{438.8111} = 0.9455$$

$$DW_t\ για\ Παρασκευή = \frac{557.9412}{438.8111} = 1.2714$$

$$DW_t\ για\ Σάββατο = \frac{353.4706}{438.8111} = 0.8055.$$

Με παρόμοιο τρόπο εκτιμάται και το M_t, δηλαδή λ.χ. για το μήνα Μάιο :

$$M_t\ Μαΐου = μέσος\ αριθμός\ αφίξεων\ τις\ ημέρες\ του\ Μαΐου\ που\ λειτουργεί\ η\ τράπεζα$$
$$= \frac{395}{438.8111} = 0.9002$$

Με τον ίδιο τρόπο βρίσκουμε:

$$M_t\ Ιανουαρίου = \frac{456.9615}{438.8111} = 1.0414$$

$$M_t\ Φεβρουαρίου = \frac{449.36}{438.8111} = 1.024$$

$$M_t\ Μαρτίου = \frac{451.1538}{438.8111} = 1.0281$$

$$M_t\ Απριλίου = \frac{430.2692}{438.8111} = 0.9805$$

$$M_t\ Ιουνίου = \frac{437.28}{438.8111} = 0.9965$$

$$M_t\ Ιουλίου = \frac{420}{438.8111} = 0.9571$$

$$M_t\ Αυγούστου = \frac{441.962963}{438.8111} = 1.0071$$

$$M_t\ Σεπτεμβρίου = \frac{431.9167}{438.8111} = 0.9842$$

$$M_t\ Οκτωβρίου = \frac{430.4074}{438.8111} = 0.9808$$

$$M_t\ Νοεμβρίου = \frac{468.48}{438.8111} = 1.0676$$

$$M_t\ Δεκεμβρίου = \frac{455.375}{438.8111} = 1.0377$$

Για την περιγραφή του πώς διαμορφώθηκε ο παραπάνω πίνακας, ας θεωρήσουμε ότι παράγουμε μια πρόβλεψη για το πλήθος πελατών της τράπεζας την Πέμπτη, 1η Φεβρουαρίου του τρέχοντος έτους. Υποθέτουμε ότι το ε_t ισούται με τη μέση του τιμή 1, οπότε η πρόβλεψή μας θα ήταν

$$B \times (DW_t\ για\ Πέμπτη) \times (M_t\ Φεβρουαρίου) = 438.8111 \cdot (0.9455) \cdot (1.024) = 424.8534$$

πελάτες μέσα στη μέρα. Για πρόβλεψη μιας μελλοντικής ημέρας, λ.χ. Σάββατο 8 Φεβρουαρίου του επόμενου έτους, θα παίρναμε

$$B \times (DW_t\ \gamma\iota\alpha\ \Sigma\acute{\alpha}\beta\beta\alpha\tau o) \times (M_t\ \Phi\epsilon\beta\rho ouαρίου) = 438.8111\ (0.8055) \cdot (1.024) = 361.9454$$

πελάτες.

Για τα δεδομένα του πίνακα που εμφανίσαμε παραπάνω, το απλό αυτό μοντέλο δίνει MAD =103.987. Εάν, όμως, χρησιμοποιούσαμε τη μέθοδο αυτή για να παράξουμε προβλέψεις για το ερχόμενο έτος, το MAD κατά πάσα πιθανότητα θα ξεπερνούσε το 103.987. Αυτό οφείλεται στο ότι έχουμε προσαρμόσει τις παραμέτρους μας σε δεδομένα του παρελθόντος: δεν υπάρχει βεβαιότητα ότι τα μελλοντικά δεδομένα θα ακολουθούν το ίδιο μοτίβο με τα παλαιότερα. Επιπλέον, παραλείψαμε σε αυτήν την ανάλυση τη διερεύνηση για το αν υπάρχει ανοδική τάση στα δεδομένα ή όχι.

Ας υποθέσουμε, τώρα, ότι ο διευθυντής της τράπεζας παρατηρεί σε μια μέρα μετά από αργία ότι η διέλευση πελατών είναι πολύ μεγαλύτερη από αυτή που προβλέπει το μοντέλο. Τα δεδομένα στον ακόλουθο πίνακα το επιβεβαιώνουν:

Μέρα μετά από αργία	Πραγματική τιμή	Πρόβλεψη	Λόγος πραγματικής/πρόβλεψης
2-Ιαν	431	414.0301	1.040987
1-Ιουν	432	412.829	1.046438
5-Ιουλ	615	474.5872	1.295863
3-Σεπ	459	488.0527	0.940472
29-Νοε	701	442.9523	1.582563
26-Δεκ	491	366.8131	1.338556

Πίνακας 11: Διέλευση πελατών μετά από αργία

Πώς θα μπορούσε να χρησιμοποιηθεί αυτή η πληροφορία για να έχουμε καλύτερες προβλέψεις για τις ημέρες μετά από αργίες; Από τον παραπάνω πίνακα, ο μέσος όρος πραγματικού/προβλεπόμενου πλήθους πελατών σε μια ημέρα μετά από αργία είναι 1.20748. Επομένως, αυτό που χρειάζεται να κάνουμε είναι να πολλαπλασιάσουμε τις προβλέψεις που κάναμε νωρίτερα με αυτόν τον παράγοντα, δηλαδή το 1,20748, στις περιπτώσεις που οι αντίστοιχες ημέρες είναι αμέσως μετά από μια αργία.

ΠΡΟΒΛΕΨΕΙΣ ΜΕ ΧΡΗΣΗ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Σε αντίθεση με τις προηγούμενες μεθόδους πρόβλεψης, σε αυτό το σημείο δεν θα ασχοληθούμε με την πρόβλεψη μέσω χρονοσειρών, αλλά θα θεωρήσουμε μια τέτοια σχέση μεταξύ παραγόντων για να κάνουμε τις αντίστοιχες προβλέψεις. Σε αντίστοιχες περιπτώσεις, δηλαδή, θεωρούμε ότι η ποσότητα που μας ενδιαφέρει σχετίζεται άμεσα με κάποιους άλλους παράγοντες, με την παρουσία κάποιων τυχαίων διακυμάνσεων σε αυτή την κατά τα άλλα αιτιατή σχέση.

Εδώ θα εστιάσουμε την προσοχή μας στο μοντέλο της γραμμικής παλινδρόμησης (linear regression). Έχουμε, λοιπόν, την μεταβλητή Y (η οποία καλείται μεταβλητή απόκρισης) την οποία θέλουμε να περιγράψουμε μέσω μιας άλλης X (επεξηγηματική μεταβλητή) και η υπόθεσή μας, στην προκειμένη περίπτωση, είναι ότι υπάρχει μια γραμμική εξάρτηση ανάμεσα στα δυο, με τον εξής τρόπο:

$$E[Y|X = x] = A + Bx.$$

Η παραπάνω έκφραση είναι το λεγόμενο συστηματικό μέρος του μοντέλου. Εναλλακτικά, αν εισάγουμε τις παρατηρήσεις X_t στη συζήτησή μας, η εξίσωση για το μοντέλο περιγράφεται ως εξής:

$$X_t = A + Bt + e_t$$

όπου Α σταθερά, Β η κλίση της ευθείας και e_t το τυχαίο σφάλμα στην t περίοδο. Για τα τυχαία σφάλματα υποθέτουμε ότι η μέση τους τιμή είναι 0 και η διασπορά είναι σταθερή (κοινή για όλες τις περιόδους).

Η μέθοδος ελαχίστων τετραγώνων

Για το παραπάνω μοντέλο, θέλουμε να εντοπίσουμε τα άγνωστα Α,Β ώστε να μπορέσουμε μετά να χρησιμοποιήσουμε το μοντέλο για να κάνουμε προβλέψεις. Πρέπει να γίνει, λοιπόν, η επιλογή μιας ευθείας $\tilde{y} = a + bx$ ώστε μέσω αυτής να προσεγγίζονται όσο γίνεται καλύτερα οι παρατηρήσεις (x_i, y_i) που έχουμε συλλέξει. Η μέθοδος ελαχίστων τετραγώνων εστιάζει στο να βρεθεί η ευθεία εκείνη που ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα, δηλαδή την ποσότητα:

$$\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

Εάν δούμε αυτήν την ποσότητα σαν μια συνάρτηση των δυο μεταβλητών a,b, παραγωγίζοντας και εξισώνοντας τις παραγώγους με μηδέν προκύπτουν οι εξής εκτιμήσεις για τις παραμέτρους αυτές:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}},$$

και

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Στο εξής, λοιπόν, θα θεωρούμε ότι (1) έχουμε ένα δείγμα παρατηρήσεων μεγέθους n $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ και (2) οι Y_i είναι κανονικά κατανεμημένες μεταβλητές με μέσο $A + Bx_i$ και διασπορά σ^2(ανεξάρτητη του i).

Για το σ^2 μια αμερόληπτη εκτιμήτρια είναι η ποσότητα $s_{y|x}^2 = \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-2}$.

Διάστημα εμπιστοσύνης για το E[Y|X=x]

Ένα χρήσιμο χαρακτηριστικό του γραμμικού μοντέλου είναι ότι απ' αυτό μπορούμε να εξάγουμε προβλέψεις για μελλοντικές τιμές, έχοντας την υπόθεση της γραμμικής εξάρτησης μεταξύ των δυο μεταβλητών. Μπορούμε, λοιπόν, απ' αυτή τη γραμμική σχέση να εκτιμήσουμε σημειακά τη E[Y|x] ή να κάνουμε μια εκτίμηση για διάστημα εμπιστοσύνης γύρω από αυτή τη μέση τιμή.

Έτσι, για παράδειγμα, μια σημειακή εκτίμηση για την πρόβλεψη $E[Y|x = x^*]$ είναι η:

$$\tilde{y}^* = a + bx^*,$$

όπου x^* η δοθείσα τιμή της ανεξάρτητης μεταβλητής και η $\tilde{y}^*$ η αντίστοιχη πρόβλεψη. Για το διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης $(1-\alpha)*100\%$ της $E[Y|x = x^*]$ είναι:

$$a + bx^* \pm t_{\frac{\alpha}{2};n-2} s_{y|x} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

όπου $s_{y|x}$ η εκτίμηση της σ^2 που προαναφέραμε και $t_{\frac{\alpha}{2};n-2}$ το $100\alpha/2$ ποσοστιαιο σημείο της Student κατανομής με n-2 βαθμούς ελευθερίας. Παρατηρούμε ότι το αντίστοιχο διάστημα γίνεται ολοένα και μικρότερο όταν το x^* πλησιάζει το $\bar{x}$, ενώ όταν το x^* απομακρύνεται από τη μέση τιμή το διάστημα φαρδαίνει.

Προβλέψεις

Η εκτίμηση της μέσης τιμής της Y είναι χρήσιμη, αλλά ορισμένες φορές ίσως δεν είναι αρκετή. Εκείνες τις φορές, μας χρειάζεται μια εκτίμηση για την πραγματική τιμή και όχι τη μέση για το Y. Έτσι, έχουμε ένα άλλο διάστημα εμπιστοσύνης για την ίδια την πρόβλεψη, τα άκρα του οποίου για συντελεστή εμπιστοσύνης $(1-\alpha)*100\%$ είναι:

$$a + bx^* \pm t_{\frac{\alpha}{2};n-2} s_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \ .$$

Έτσι, είναι $1-\alpha$ πιθανό η μελλοντική τιμή για $x = x^*$ να βρίσκεται εντός του διαστήματος με άκρα τα παραπάνω.

Αυτή η τεχνική πρόβλεψης είναι αποτελεσματική για μια μεμονωμένη πρόβλεψη. Ωστόσο, δεν είναι δυνατό να χρησιμοποιήσουμε τα ίδια δεδομένα για πολλαπλά διαστήματα εμπιστοσύνης διαφόρων προβλέψεων θεωρώντας ότι όλες οι προβλέψεις μας θα είναι σωστές. Η δυσκολία, δηλαδή, έγκειται στο ότι από τη στιγμή που θα χρησιμοποιήσουμε το ίδιο σύνολο δεδομένων για πολλαπλές προβλέψεις, αυτές θα είναι στατιστικά εξαρτημένες.

Αυτό το πρόβλημα μπορεί να ξεπεραστεί χρησιμοποιώντας διαστήματα ταυτόχρονης ανοχής. Τα άκρα αυτού του διαστήματος είναι:

$$a + bx^* \pm c^{**} s_{y|x} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

όπου το c^{**} λαμβάνει τις τιμές που φαίνονται στον ακόλουθο πίνακα:

n	$\alpha = 0.50$	$\alpha = 0.25$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
			$P = 0.90$			
4	7.471	10.160	13.069	14.953	18.663	23.003
6	5.380	7.453	9.698	11.150	14.014	17.363
8	5.037	7.082	9.292	10.722	13.543	16.837
10	4.983	7.093	9.366	10.836	13.733	17.118
12	5.023	7.221	9.586	11.112	14.121	17.634
14	5.101	7.394	9.857	11.447	14.577	18.232
16	5.197	7.586	10.150	11.803	15.057	18.856
18	5.300	7.786	10.449	12.165	15.542	19.484
20	5.408	7.987	10.747	12.526	16.023	20.140
			$P = 0.95$			
4	10.756	14.597	18.751	21.445	26.760	32.982
6	6.652	9.166	11.899	13.669	17.167	21.266
8	5.933	8.281	10.831	12.484	15.750	19.568
10	5.728	8.080	10.632	12.286	15.553	19.369
12	5.684	8.093	10.701	12.391	15.724	19.619
14	5.711	8.194	10.880	12.617	16.045	20.050
16	5.771	8.337	11.107	12.898	16.431	20.559
18	5.848	8.499	11.357	13.204	16.845	21.097
20	5.937	8.672	11.619	13.521	17.272	21.652
			$P = 0.99$			
4	24.466	33.019	42.398	48.620	60.500	74.642
6	10.444	14.285	18.483	21.215	26.606	32.920
8	8.290	11.453	14.918	17.166	21.652	26.860
10	7.567	10.539	13.796	15.911	20.097	24.997
12	7.258	10.182	13.383	15.479	19.579	24.403
14	7.127	10.063	13.267	15.355	19.485	24.316
16	7.079	10.055	13.306	15.410	19.582	24.467
18	7.074	10.111	13.404	15.552	19.794	24.746
20	7.108	10.198	13.566	15.745	20.065	25.122

Εικόνα 2 Πίνακες τιμών του c^{**} (Οι πίνακες προέρχονται από το [2])

Σφάλματα: πόσο αποδοτική είναι μια πρόβλεψη;

Με τη μέθοδο ελαχίστων τετραγώνων είδαμε πώς μπορούμε να κάνουμε προβλέψεις. Πώς μπορούμε, όμως, να πούμε πόσο ακριβείς είναι οι προβλέψεις μας; Πώς μπορούμε, δηλαδή, να ποσοτικοποιήσουμε το σφάλμα λόγω χρήσης αυτού του μοντέλου; Για την απάντηση αυτού του ερωτήματος, θεωρούμε τρεις συνιστώσες μεταβλητότητας στο μοντέλο: το συνολικό άθροισμα τετραγώνων (SST), το συνολικό άθροισμα σφαλμάτων (SSE) και το άθροισμα τετραγώνων παλινδρόμησης (SSR). Το

$$SST = \sum(y_i - \bar{y})^2$$

μετράει τη συνολική μεταβλητότητα του μοντέλου, δηλ. την απόκλιση των παρατηρούμενων τιμών από το μέσο τους όρο. Ύστερα, αυτή η συνολική μεταβλητότητα αναλύεται σε δυο τμήματα: ένα το οποίο είναι η μεταβλητότητα που περιγράφεται από την παλινδρόμηση, δηλ. το

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$

και άλλο ένα το οποίο αποδίδεται καθαρά στις τυχαίες επιδράσεις, δηλ. το

$$SSE = \sum(y_i - \hat{y}_i)^2 = \sum e_i^2 \, .$$

Πρακτικά είναι αδύνατο να πετύχουμε $SSE = 0$, διότι σε αυτήν την περίπτωση όλα τα παρατηρούμενα σημεία θα ήταν ακριβώς πάνω στην αντίστοιχη ευθεία του μοντέλου, κάτι το οποίο δεν μπορεί να γίνει. Ωστόσο, ένα καλό μοντέλο γραμμικής παλινδρόμησης χαρακτηρίζεται απ' το ότι το αντίστοιχό SSE του είναι αρκετά μικρό. Έχουμε, λοιπόν, ότι:

$$SST = SSR + SSE.$$

Με βάση τα όσα είπαμε και πριν, ένα καλό μοντέλο αντιστοιχεί σε μικρό SSE, και άρα λόγω της παραπάνω σχέσης, σε ένα αρκετά μεγάλο SSR. Μπορούμε, επίσης, να ορίσουμε και τον *συντελεστή προσδιορισμού* ως

$$R^2 = \frac{SSR}{SST}$$

που εκφράζει το ποσοστό που η επεξηγηματική μεταβλητή περιγράφει την μεταβλητή απόκρισης. Εναλλακτικά, μπορούμε να γράψουμε $1 - R^2 = \frac{SSE}{SST}$.

Επομένως, αναδιατυπώνοντας, μπορούμε να πούμε ότι ένα καλό μοντέλο παλινδρόμησης – δεδομένου ότι πληρούνται οι προϋποθέσεις που κάνουν τη χρήση του εύλογη- είναι ένα μοντέλο με υψηλό συντελεστή προσδιορισμού.

Όσον αφορά τις προβλέψεις, που είναι το κομμάτι που μας ενδιαφέρει, ένα μέτρο ακρίβειας των προβλέψεων στην παλινδρόμηση είναι το *τυπικό σφάλμα εκτίμησης* (s_e). Αν n είναι ο αριθμός των παρατηρήσεων, το s_e δίνεται απ' τον τύπο:

$$s_e = \sqrt{\frac{SSE}{n-2}}.$$

Ισχύει ότι περίπου το 68% των τιμών του y θα είναι εντός το πολύ s_e μακριά απ'την προβλεπόμενη τιμή $\hat{y}$ και το 95% των τιμών του y θα είναι μέσα σε εύρος $2s_e$ απ' την προβλεπόμενη τιμή $\hat{y}$. Οτιδήποτε εκτός του εύρους $2s_e$ απ' την $\hat{y}$ λέγεται άτυπη παρατήρηση (outlier). Η διαχείριση τέτοιων παρατηρήσεων χρειάζεται ιδιαίτερη προσοχή. Φυσικά, στην περίπτωση που μια τέτοια παρατήρηση έχει προκύψει απλά από λάθος κατά το «πέρασμα» των δεδομένων, χρειάζεται απλά να διορθωθεί και να εισαχθεί το σωστό νούμερο. Εάν, από την άλλη, έχουμε μια τέτοια παρατήρηση η οποία ξεχωρίζει κατά πολύ από τις άλλες, ενδεχομένως η πιο σωστή τακτική είναι να παραλείψουμε τη συγκεκριμένη παρατήρηση και να ξαναπροσαρμόσουμε εκ νέου το μοντέλο.

Προϋποθέσεις για την ισχύ του μοντέλου

Αναφέραμε πριν στο κομμάτι για το συντελεστή προσδιορισμού ότι κάτω από συγκεκριμένες προϋποθέσεις, μπορεί να αποτελέσει ένα μέτρο επεξηγηματικότητας του μοντέλου, μπορεί να μας πει δηλαδή σε τι ποσοστό εξηγείται απ' το μοντέλο μας η μεταβλητότητα της μεταβλητής απόκρισης. Ποιες είναι, όμως, αυτές οι προϋποθέσεις;

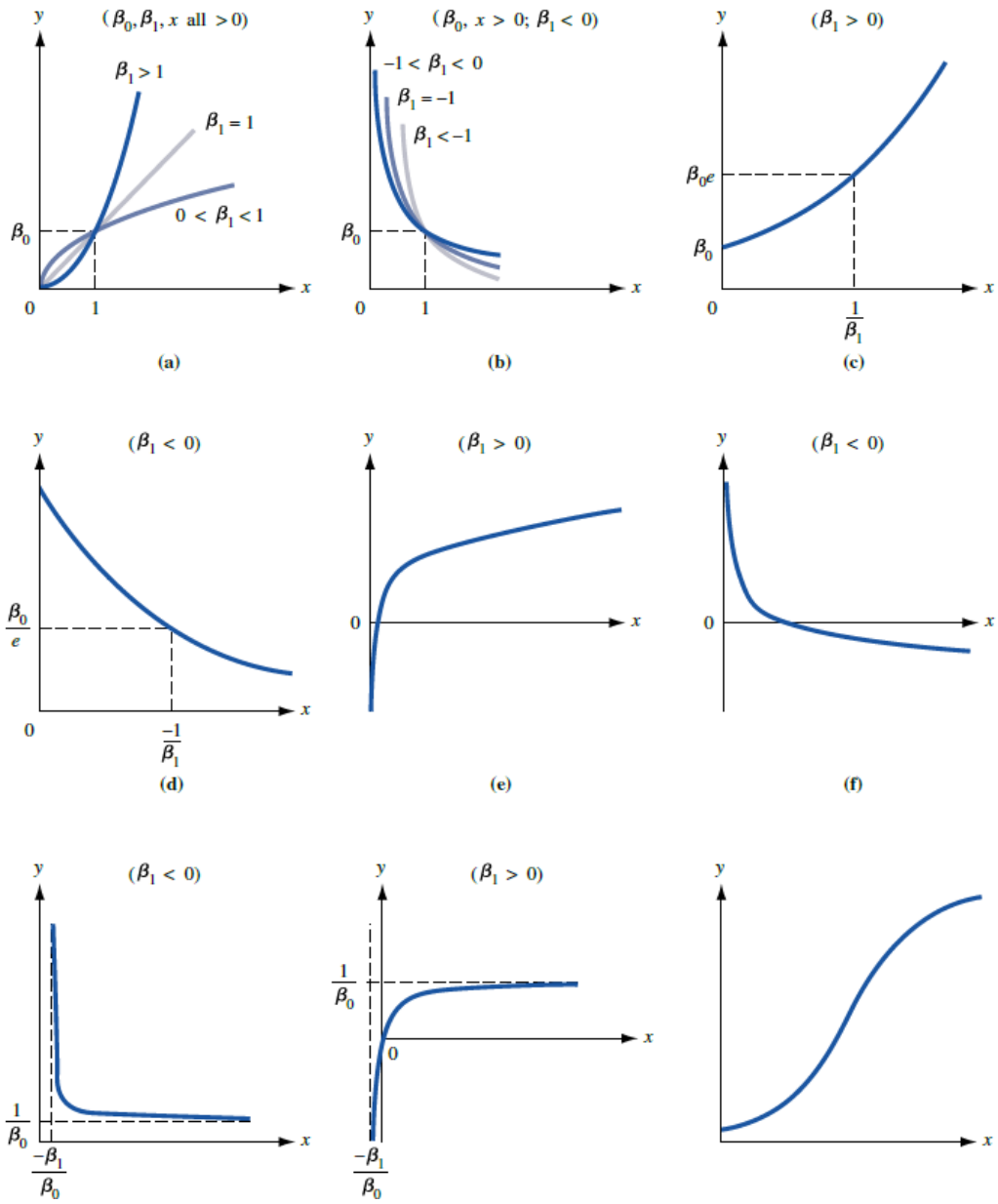
Έχει επικρατήσει οι παραδοχές για το γραμμικό μοντέλο να είναι οι εξής:

- Ομοσκεδαστικότητα: Η διασπορά του όρου του σφάλματος δεν πρέπει να εξαρτάται από την τιμή της επεξηγηματικής μεταβλητής x, αλλά πρέπει να είναι ενιαία για όλες τις τιμές. Σε αντίθεση περίπτωση, κάνουμε λόγο για ετεροσκεδαστικότητα. Ο έλεγχος για αυτήν την υπόθεση γίνεται γραφικά: κάνουμε ένα γράφημα μεταξύ των παρατηρήσεων x και των υπολοίπων e. Αν δεν εμφανίζεται κανένα ξεκάθαρο μοτίβο σε αυτό το γράφημα, θεωρούμε ότι ικανοποιείται η υπόθεση της ομοσκεδαστικότητας.
- Κανονικότητα των Σφαλμάτων: Τα σφάλματα θεωρούμε ότι κατανέμονται κανονικά, έχοντας μέση τιμή 0 και κάποια διασπορά σ^2 (ενιαία, με βάση την υπόθεση της ομοσκεδαστικότητας). Για τον έλεγχο αυτής της υπόθεσης χρησιμοποιούμε τεχνικές ελέγχου για την κατανομή (π.χ. διάγραμμα ποσοστημορίων) στα υπόλοιπα e.
- Ανεξαρτησία των σφαλμάτων: Η συγκεκριμένη υπόθεση μας λέει ότι η γνώση ενός σφάλματος δεν θα πρέπει να μας δίνει πληροφορία για το σφάλμα της επόμενης ή και οποιασδήποτε άλλης παρατήρησης. Για να ελέγξουμε αν ισχύει, πάλι δουλεύουμε γραφικά: κάνοντας ένα γράφημα των e σε σχέση με τη σειρά των παρατηρήσεων, εξετάζουμε αν σχηματίζεται κάποιο μοτίβο στα σημεία του γραφήματος. Αν όχι, μπορούμε να πούμε ότι το μοντέλο πληροί την προϋπόθεση της ανεξαρτησίας των σφαλμάτων.

Παλινδρόμηση με μη γραμμικά μοντέλα

Μέχρι στιγμής εστιάσαμε σε γραμμικά μοντέλα πάνω στο κομμάτι της παλινδρόμησης, θεωρούσαμε δηλαδή ότι η καμπύλη που συνδέει την επεξηγηματική μεταβλητή με τη μεταβλητή απόκρισης είναι μια ευθεία γραμμή. Πώς αντιμετωπίζεται, όμως, η περίπτωση όπου η καμπύλη αυτή δεν είναι ευθεία, αλλά η σχέση που συνδέει αυτές τις δυο μεταβλητές είναι μέσω μιας κάποιας άλλης, μη γραμμικής συνάρτησης;

Στην ακόλουθη εικόνα βλέπουμε διάφορα τέτοια παραδείγματα:



Εικόνα 3 Μοντέλα που μπορούν να μετασχηματιστούν σε γραμμικά (Τα σχήματα προέρχονται από το [4])

Η διαδικασία που ακολουθούμε, λοιπόν, σε αυτές τις περιπτώσεις είναι η εξής:

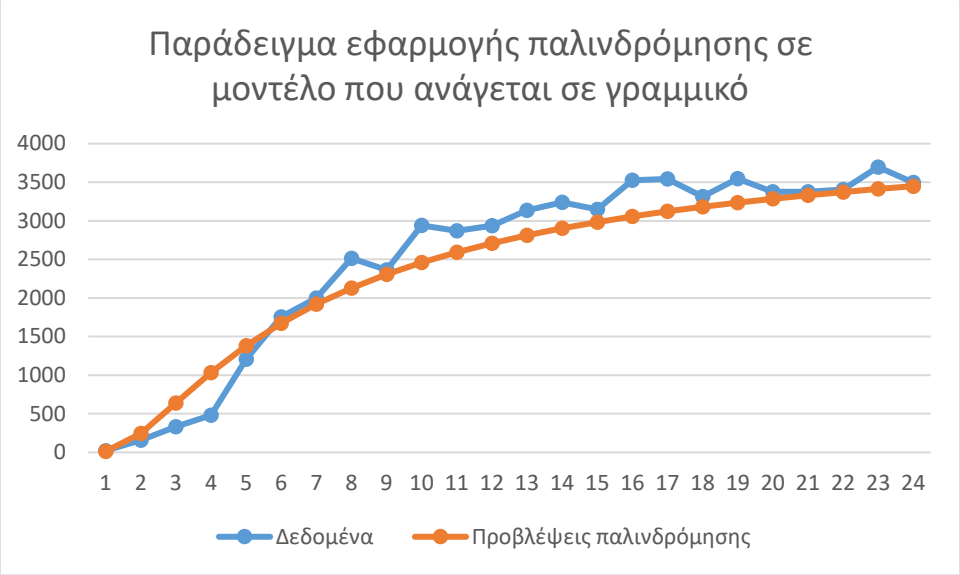
1. Κάνουμε μια γραφική παράσταση των δεδομένων μας και βλέπουμε τι είδους καμπύλη «τείνουν» να διαμορφώσουν (προφανώς τα σημεία δεν θα είναι *ακριβώς* πάνω σε μια καμπύλη λόγω τυχαίων σφαλμάτων) .
2. Ας υποθέσουμε ότι έχουμε προσδιορίσει τη μορφή της καμπύλης. Τότε θεωρούμε την αντίστοιχη συνάρτηση που συνδέει τις x, y, εισάγοντας κάποιες παραμέτρους. Για παράδειγμα, $y = \beta_0 e^{\beta_1 x}$.
3. Σε αυτό το τελευταίο βήμα, αναζητούμε κάποιον μετασχηματισμό ο οποίος να μπορεί αυτό το αρχικά μη-γραμμικό μοντέλο να το ανάγει σε κάποιο αντίστοιχο γραμμικό μοντέλο, ούτως ώστε να εφαρμοστεί σε αυτό το τελευταίο η ανάλυση που περιγράψαμε

νωρίτερα. Στο προαναφερθέν παράδειγμα, μπορούμε να μετασχηματίσουμε το μοντέλο με τη χρήση του φυσικού λογαρίθμου παίρνοντας το αντίστοιχο μη γραμμικό μοντέλο $Y = \ln y = A + BX$, όπου $A = ln\beta_0, B = \beta_1$ και $X = x$.

Ακολούθως δίνουμε και ένα αντίστοιχο παράδειγμα για την υλοποίηση αυτής της διαδικασίας. Ας υποθέσουμε ότι έχουμε δεδομένα για 24 μήνες για τις πωλήσεις ενός προϊόντος, τα οποία παρατίθενται στον παρακάτω πίνακα και απεικονίζονται σχηματικά στο ακόλουθο γράφημα:

Μήνας	Πωλήσεις
1	23
2	156
3	330
4	482
5	1209
6	1756
7	2000
8	2512
9	2366
10	2942
11	2872
12	2937
13	3136
14	3241
15	3149
16	3524
17	3542
18	3312
19	3547
20	3376
21	3375
22	3403
23	3697
24	3495

Πίνακας 12: Μηνιαίες πωλήσεις προϊόντος (σε χιλιάδες)



Γράφημα 6: Μηνιαίες πωλήσεις προϊόντος (σε χιλιάδες), εφαρμογή μη γραμμικού μοντέλου παλινδρόμησης που ανάγεται σε γραμμικό

Οι πωλήσεις εκφράζονται σε χιλιάδες στο γράφημα, ενώ οι παρατηρηθείσες τιμές αναπαρίστανται με κουκίδες και η αντίστοιχη πρόβλεψη με τη συνεχή γραμμή. Βλέποντας αποκλειστικά το μπλε κομμάτι του γραφήματος, που αντιστοιχεί στα δεδομένα μας, είναι εμφανές ότι οι σχέση που αναζητούμε μεταξύ μήνα και αντίστοιχων πωλήσεων δεν είναι γραμμική, αλλά φαίνεται να υπάρχει μια καμπύλη που να περιγράφει την εξάρτηση των δυο.

Η συγκεκριμένη καμπύλη μοιάζει να είναι έχει σχήμα S, οπότε μια πιθανή σχέση μεταξύ x: μήνας και y: πωλήσεις στον αντίστοιχο μήνα, είναι η $y = \exp\left(\beta_0 + \frac{\beta_1}{x}\right)$. Αυτό, λοιπόν, είναι ένα μοντέλο το οποίο εμπίπτει στη περίπτωση που αναφέραμε παραπάνω, οπότε μπορούμε παίρνοντας λογάριθμο να εργαστούμε με το αντίστοιχο μοντέλο

$$Y = \beta_0 + \beta_1 X, \text{ όπου } Y = lny \text{ και } X = \frac{1}{x}.$$

Εφαρμόζοντας τη μέθοδο ελαχίστων τετραγώνων, καταλήγουμε -για τα αρχικά x, y – στη σχέση

$$\hat{y} = \exp\left(\widehat{\beta_0} + \frac{\widehat{\beta_1}}{x} + \frac{s_e^2}{2}\right)$$

όπου $\widehat{\beta_0}, \widehat{\beta_1}$ οι εκτιμήτριες ελαχίστων τετραγώνων για τις παραμέτρους β_0, β_1 και s_e το αντίστοιχο τυπικό σφάλμα. Για το συγκεκριμένο σύνολο δεδομένων, προκύπτει ότι

$\beta_0 = 8.387, s_e = 0.276$

και $\widehat{\beta_1} = -5.788$. Έτσι, ο τύπος για τις αντίστοιχες προβλέψεις είναι:

$$\hat{y} = \exp\left(8.387 + \frac{0.276^2}{2} - \frac{5.788}{x}\right) = \exp\left(8.425 - \frac{5.788}{x}\right)$$

Αν, λοιπόν, επιθυμούμε λ.χ. να προβλέψουμε τις πωλήσεις για το συγκεκριμένο προϊόν στον 26ο μήνα από την έναρξη της περιόδου που εξετάζουμε, η αντίστοιχη πρόβλεψη θα είναι:

$$\hat{y} = \exp\left(8.425 - \frac{5.788}{26}\right) = 3649.3$$

μονάδες προϊόντος το συγκεκριμένο μήνα.

Παρατηρήσεις: 1. Για το αντίστοιχο γραμμικό μοντέλο, μπορεί να υπολογίσει κανείς ότι ο αντίστοιχος συντελεστής προσδιορισμού ισούται με $R^2 = 95\%$. Αυτό σημαίνει, λοιπόν, ότι δεδομένου ότι πληρούνται οι προϋποθέσεις του μοντέλου, η μεταβλητότητα στις πωλήσεις εξηγείται κατά ένα 95% από το συγκεκριμένο μοντέλο. Ωστόσο, αυτό δε μας παρέχει πληροφορία για την ακρίβεια των προβλέψεών μας. Για αυτό το σκοπό, υπολογίζουμε τις προβλέψεις $\hat{y}_i$ που αντιστοιχούν στις παρατηρηθείσες τιμές (x_i, y_i) και τα υπόλοιπα

$$e_i = y_i - \hat{y}_i$$

και ύστερα λαμβάνουμε το μέσο όρο των $|e_i|$ για τους συγκεκριμένους 24 μήνες, δηλαδή το MAD για το αντίστοιχο σύνολο παρατηρήσεων. Η τιμή που προκύπτει είναι $MAD = 170.3$.

Μπορούμε, τότε, να εκτιμήσουμε την τυπική απόκλιση (με βάση προηγούμενο τύπο) για τις παρατηρήσεις ως $1.25 \cdot 170.3 = 212.88$.

Επομένως, περιμένουμε ότι το 95% των φορών θα έχουμε ακρίβεια στις προβλέψεις μας μέσα στο εύρος $2 \cdot 212.88 = 425.76$ μονάδων προϊόντος.

2. Εάν θεωρούσαμε, εσφαλμένα, το αντίστοιχο γραμμικό μοντέλο (δηλαδή γραμμική εξάρτηση μεταξύ των (x, y)) θα παίρναμε $s_e = 546$, γεγονός που υποδεικνύει ότι το μοντέλο που θεωρήσαμε βελτιώνει σημαντικά τις προβλέψεις μας.

ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Σε πολλές περιπτώσεις, η εξαρτημένη μεταβλητή/μεταβλητή απόκρισης δεν εξαρτάται από μια μόνο, αλλά από περισσότερες μεταβλητές. Εάν θέλουμε, λοιπόν, να εφαρμόσουμε παρόμοια ανάλυση σε αυτήν την περίπτωση, τότε κάνουμε λόγο για πολλαπλή παλινδρόμηση. Αν θέλουμε, για παράδειγμα, να εξετάσουμε τις μηνιαίες πωλήσεις σε μια αλυσίδα fast-food εστιατορίων, δεν μπορούμε να περιοριστούμε σε μια παράμετρο, αλλά αντιθέτως είναι αρκετές εκείνες οι παράμετροι που πρέπει να ληφθούν υπόψη: η τιμή των πρώτων υλών, το ποσό χρημάτων που δαπανάται για διαφήμιση, το ποσό που δαπανήθηκε τον προηγούμενο μήνα για διαφήμιση και αρκετές άλλες.

Ας υποθέσουμε, λοιπόν, ότι θεωρούμε πως η y εξαρτάται από k το πλήθος ανεξάρτητες/επεξηγηματικές μεταβλητές $x_1, \dots, x_k$ και έστω ότι έχουμε n το πλήθος παρατηρήσεις της μορφής $(y_j, x_{1j}, \dots, x_{kj}), j = 1,2, \dots, n$ με y_j να είναι η $j-$οστη παρατήρηση για τη μεταβλητή y και x_{ij} η $j-$οστη παρατήρηση για την $i-$οστη μεταβλητή. Ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης, τότε, θα είναι της μορφής:

$$y_j = \beta_0 + \beta_1 x_{1j} + \cdots + \beta_k x_{kj} + \varepsilon_j, j = 1,2, \dots, n$$

όπου τα ε_j έχουν μέση τιμή, ώστε με αυτόν τον τρόπο να εκφράζουμε ότι οι τιμές y_j δεν θα συμπίπτουν απαραίτητα με την έκφραση $\beta_0 + \beta_1 x_{1j} + \cdots + \beta_k x_{kj}$ (το οποίο αναφέρεται και ως το *συστηματικό μέρος* του μοντέλου). Η παράμετρος $\beta_i, i = 1, \dots, k$ εκφράζει την μεταβολή της y εάν η x_i αυξηθεί κατά μια μονάδα, είναι δηλαδή το ανάλογο της $\frac{\partial y}{\partial x_i}$.

Κατά αντίστοιχο τρόπο με το απλό γραμμικό μοντέλο, είναι εφικτή η εφαρμογή της μεθόδου ελαχίστων τετραγώνων για να προσδιοριστούν οι εκτιμήσεις $\widehat{\beta_i}$ των παραμέτρων του μοντέλου. Το μοντέλο που προκύπτει (που προσαρμόζουμε, δηλαδή) έχει τη μορφή:

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \cdots + \hat{\beta}_k x_{kj}.$$

Αντίστοιχα ορίζονται τα υπόλοιπα $e_j = y_j - \hat{y}_j$ και η λογική με την οποία υπολογίζονται τα $\hat{\beta}_i$ είναι ακριβώς η ελαχιστοποίηση του $\sum e_j^2$.

<u>*Παράδειγμα*</u> *(Το παράδειγμα απαντάται στο [5])*: Θέλουμε να μελετήσουμε την απώλεια βάρους γ 10 τύπων ελαστικών κατά τη διάρκεια μιας δοκιμασίας, συσχετίζοντάς τη με τη σκληρότητα x_1 και την αντοχή x_2 αυτών. Ακολουθούν τα σχετικά δεδομένα:

y	x_1	x_2
372	46	162
206	55	233
175	61	232
154	66	232
136	71	231
112	71	237
55	81	224
45	86	219
222	53	203
170	60	188

Εφαρμόζοντας τη μέθοδο ελαχίστων τετραγώνων για τη προσαρμογή απλού γραμμικού μοντέλου σε αυτά τα δεδομένα, προκύπτει η προσέγγιση:

$$\hat{y}_j = 75.234 - 5.954x_1 - 0.956x_2$$

Έτσι, αν θέλουμε, για παράδειγμα, να εκτιμήσουμε την απώλεια βάρους ενός ελαστικού με σκληρότητα 30 μονάδων και αντοχή 200 μονάδων, η αντίστοιχη εκτίμησή μας θα είναι

$$\hat{y} = 758.234 - 5.954 \cdot 30 - 0.956 \cdot 200 = 388.414.$$

Όσον αφορά την ερμηνεία των $\hat{\beta}_1 = -5.954$, $\hat{\beta}_2 = -0.956$, έχουμε τα εξής: για το $\hat{\beta}_1$, ανάμεσα σε ελαστικά που έχουν ίδια την αντοχή του ελαστικού και διαφέρουν κατά μια μονάδα στη σκληρότητα, η απώλεια βάρους είναι κατά 5.954 μονάδες μικρότερη στο ελαστικό με την παραπάνω σκληρότητα. Ομοίως, για το $\hat{\beta}_2$ έχουμε ότι για δυο ελαστικά που έχουν ίδια

σκληρότητα, εκείνο που η αντοχή του είναι μεγαλύτερη κατά μια μονάδα έχει κατά 0.956 μονάδες λιγότερη απώλεια βάρους.

Αναφορικά με την ανάλυση σφάλματος, η διαδικασία είναι η ίδια, με τα SSR, SSE, SST να ορίζονται με ακριβώς τον ίδιο τρόπο με πριν. Επιπλέον, ο συντελεστής προσδιορισμού $R^2 = \frac{SSR}{SST}$ εκφράζει το ποσοστό που πετυχαίνουν οι k μεταβλητές του μοντέλου να εξηγήσουν τη μεταβλητότητα της y. Αν, πάλι, s_e είναι η τυπική απόκλιση για μια παρατήρηση, τότε

$$s_e = \sqrt{\frac{SSE}{n-k-1}}$$

και μπορούμε να περιμένουμε ότι γύρω στο 68% των παρατηρήσεων στον y θα είναι μέσα σε εύρος s_e από τη $\hat{y}$ και περίπου το 95% βρίσκεται μέσα σε εύρος $2s_e$ από τη $\hat{y}$. Στο παράδειγμά μας, είναι $n = 10, k = 2$ και SSE =5839.83, οπότε $s_e = \sqrt{\frac{5839.83}{10-2-1}} = 28.884$, οπότε το 95% των φορών περιμένουμε ότι οι προβλέψεις μας θα είναι ακριβείς μέσα σε εύρος $2s_e = 57.768$.

Ένας βασικός έλεγχος (ο οποίος γίνεται και στην περίπτωση του απλού γραμμικού μοντέλου) είναι ο στατικός έλεγχος t (t-test), το οποίο εξετάζει αν η μεταβλητή x_i είναι αναγκαίο να συμπεριληφθεί στο μοντέλο ή όχι. Ο αντίστοιχος στατιστικός έλεγχος διατυπώνεται ως

$$H_0: \beta_i = 0 \ , \qquad H_1: \beta_i \neq 0$$

Για τον έλεγχο αυτής της υπόθεσης, χρησιμοποιείται η στατιστική συνάρτηση $T_i = \frac{\widehat{\beta}_i}{se(\widehat{\beta}_i)}$. Υπό την υπόθεση H_0, το συγκεκριμένο στατιστικό ακολουθεί την κατανομή Student με $n-k-1$ βαθμούς ελευθερίας. Ένας τρόπος για να αποφανθούμε για το αν η μεταβλητή θα παραμείνει στο μοντέλο ή όχι είναι μέσω της αντίστοιχης p-value=$P[|T_i| > t_i]$, όπου t_i είναι η αντίστοιχη παρατηρούμενη τιμή της έκφρασης $\frac{\widehat{\beta}_i}{se(\widehat{\beta}_i)}$. Αν έχουμε ορίσει ένα επίπεδο σημαντικότητας α και η p-value είναι μικρότερο από αυτό, τότε απορρίπτουμε τη μηδενική υπόθεση και ουσιαστικά συμπεραίνουμε ότι έχει νόημα να παραμείνει η εν λόγω μεταβλητή στο μοντέλο. Σε αντίθετη περίπτωση, λέμε ότι σε αυτό το επίπεδο σημαντικότητας δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση, το οποίο πρακτικά σημαίνει ότι μπορούμε να εξαιρέσουμε τη συγκεκριμένη μεταβλητή από το μοντέλο μας. Συνήθεις επιλογές για το επίπεδο σημαντικότητας α είναι 1%, 5% , 10% .

Επιστρέφοντας για άλλη μια φορά στο παράδειγμά μας, αν εφαρμόσουμε τον έλεγχο αυτό στις x_1, x_2 παίρνουμε τις p-value 0.000335 και 0.080873 αντίστοιχα. Έτσι, λαμβάνοντας- όπως συνηθίζεται στην πράξη- ως επίπεδο σημαντικότητας 5%, βλέπουμε ότι 0.000335<0.05<0.080873, οπότε με βάση αυτό το επίπεδο σημαντικότητας η x_1 αξίζει να διατηρηθεί στο μοντέλο, ενώ η x_2 ενδέχεται να μην είναι κρίσιμο να συμπεριληφθεί, γεγονός που υποδεικνύεται απ' το ότι η δική της p-value υπερβαίνει το επίπεδο σημαντικότητας που θεωρήσαμε.

Ένα βασικό ερώτημα που μας απασχολεί σε αυτές τις περιπτώσεις είναι η επιλογή ανάμεσα σε διαφορετικά μοντέλα παλινδρόμησης, όπου πρέπει να αποφασίσουμε ποιες ανεξάρτητες

μεταβλητές θα χρησιμοποιήσουμε. Ένας τρόπος για να λάβουμε αυτήν την απόφαση είναι εντοπίζοντας τη μικρότερη τιμή για το s_e, μιας και αυτό θα δώσει τις ακριβέστερες προβλέψεις. Από την άλλη, μας ενδιαφέρει οι αντίστοιχες μεταβλητές να προκύπτουν στατιστικά σημαντικές από το t-test. Αυτοί οι δυο στόχοι ενδεχομένως να έρχονται σε σύγκρουση μεταξύ τους. Έτσι, ένας άλλος τρόπος να αποφανθούμε για μια απάντηση σε αυτό το ερώτημα είναι μέσω του στατιστικού C_p, το οποίο, εκτός των άλλων, μας το δίνει το αντίστοιχο υπολογιστικό πρόγραμμα εαν προσαρμόσουμε σε αυτό το μοντέλο παλινδρόμησης. Σε ένα καλό μοντέλο παλινδρόμησης θέλουμε να ισχύει ότι η τιμή του C_p είναι κοντά στο πλήθος των ανεξαρτήτων μεταβλητών του αντίστοιχου μοντέλου+1. Έτσι, αν σε ένα μοντέλο ισχύει για παράδειγμα $C_p = 80$ ενώ το μοντέλο αυτό διαθέτει τρεις μεταβλητές, τότε μπορούμε να είμαστε βέβαιοι ότι δεν πρόκειται για ένα «καλό» μοντέλο. Στην πραγματικότητα, εάν το C_p είναι πολύ μεγαλύτερο από το πλήθος των μεταβλητών, τότε αυτό σημαίνει ότι έχουμε παραλείψει τουλάχιστον μια σημαντική για το μοντέλο μεταβλητή.

Αρκετές φορές υπάρχει το εξής ενδεχόμενο: δυο ή και περισσότερες μεταβλητές στο ίδιο μοντέλο που χρησιμοποιούνται ως επεξηγηματικές ενδέχεται να παρουσιάζουν ισχυρή γραμμική εξάρτηση μεταξύ τους. Αυτό ενδέχεται να καθιστά τις εκτιμήσεις των αντιστοίχων β_i αναξιόπιστες. Το συγκεκριμένο φαινόμενο αναφέρεται ως *πολυσυγγραμμικότητα.* Έτσι, σε κάποιες περιπτώσεις βλέπουμε ότι ενώ για μια μεταβλητή θα είχαμε το β_i θετικό, η αντίστοιχη εκτίμηση $\hat{\beta}_i$ μπορεί να είναι σημαντικά μικρότερη του 0.

Μια πολύ συχνή περίπτωση, επίσης, είναι αυτή κατά την οποία θα συναντήσουμε παραμέτρους οι οποίες δεν είναι ποσοτικές και όχι ποιοτικές. Εάν σε ένα πρόβλημα, για παράδειγμα, μας ενδιαφέρει λ.χ. το φύλο των ανθρώπων ή το αν καπνίζουν ή όχι, αυτά είναι χαρακτηριστικά τα οποία δεν είναι ποσοτικά μετρήσιμα. Τέτοιες μεταβλητές ονομάζονται κατηγορικές. Σε αυτές, έχουμε κάποιες c το πλήθος κατηγορίες στις οποίες κυμαίνονται οι «τιμές» της εκάστοτε μεταβλητής. Είθισται, σε αυτές τις περιπτώσεις, να χρησιμοποιούμε τις λεγόμενες εικονικές μεταβλητές (dummy variables). Για μεταβλητή που αντιστοιχεί σε κατηγορική παράμετρο με c κατηγορίες θεωρούμε c-1 το πλήθος εικονικές μεταβλητές x_i με την x_i ανά επιστρέφει 1 αν η αντίστοιχη τιμή είναι της 1ης κατηγορίας και $x_i = 0$ αλλιώς.

Αναφορικά με την ερμηνεία των συντελεστών σε αυτό το μοντέλο, αυτή γίνεται ως εξής: όταν έχουμε στο μοντέλο μια κατηγορική μεταβλητή ως επεξηγηματική, τότε κατά την προσαρμογή του μοντέλου υπάρχει μια κατηγορία αναφοράς, με τους συντελεστές να αναφέρονται στη μεγαλύτερη/μικρότερη τιμή που δίνουν στη μεταβλητή απόκρισης εν σχέσει με την κατηγορία αναφοράς, ενώ οι υπόλοιπες επεξηγηματικές μεταβλητές του μοντέλου παραμένουν αμετάβλητες. Έτσι, για παράδειγμα, στην απλή περίπτωση που έχουμε το μοντέλο

$$\mathrm{y} = 13.761 - 0.436\mathrm{x}$$

όπου η y είναι ένα μέτρο της εμπιστοσύνης του κόσμου απέναντι σε έναν αστυφύλακα και η μεταβλητή x εκφράζει το φύλο με $x = 0$ εάν ο αστυφύλακας είναι άντρας και $x = 1$ εάν είναι γυναίκα. Έτσι, βάσει των συντελεστών του μοντέλου, το 13.761 εκφράζει τη μέση εμπιστοσύνη απέναντι σε έναν άνδρα αστυνομικό, ενώ το -0.436 εκφράζει το γεγονός ότι η εμπιστοσύνη

απέναντι σε μια γυναίκα αστυνομικό είναι κατά 0.436 μονάδες λιγότερη σε σχέση με αυτή απέναντι σε έναν άνδρα.

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Τα γραμμικά μοντέλα είναι σίγουρα χρήσιμα για τη μελέτη της συσχέτισης μεταξύ μεταβλητών. Ωστόσο, όταν η μεταβλητή απόκρισης δεν είναι συνεχής ενδέχεται το γραμμικό μοντέλο να μην είναι το πλέον κατάλληλο. Όταν η μεταβλητή είναι, παραδείγματος χάριν, ο μισθός των εργαζομένων μιας εταιρίας ή η τιμή μιας μετοχής, τότε υπό τις κατάλληλες προϋποθέσεις είναι εύλογο να χρησιμοποιήσουμε το σύνηθες γραμμικό μοντέλο, όπως το περιγράψαμε ήδη. Λιγότερο αποδοτική, όμως, είναι η εφαρμογή του όταν έχουμε να κάνουμε λ.χ. με τον αριθμό επειγόντων περιστατικών σε ένα νοσοκομείο ή στην απόκριση ασθενών σε μια θεραπεία. Για την αντιμετώπιση αυτής της δυσκολίας, υπάρχει μια διευρυμένη κλάση μοντέλων – στην οποία εντάσσεται και το γραμμικό μοντέλο – τα οποία αποκαλούνται «Γενικευμένα Γραμμικά Μοντέλα» (Γ.Γ.Μ.).

Η ιδέα πίσω από τα γενικευμένα γραμμικά μοντέλα είναι η εξής: θεωρούμε μια κατανομή πιθανότητας για τη μεταβλητή απόκρισης Y που μας ενδιαφέρει να μελετήσουμε. Ανάλογα με το πρόβλημα, είναι εύλογη, αρκετές φορές, η επιλογή μιας συγκεκριμένης κατανομής. Για παράδειγμα, αν η μεταβλητή μας αφορά την έλευση πελατών σε ένα κατάστημα, τότε η κατανομή Poisson αποτελεί φυσιολογική επιλογή για την περιγραφή αυτής της μεταβλητής. Στη συνέχεια, επιθυμούμε τη μέση τιμή αυτής να την συνδέσουμε με την λεγόμενη γραμμική προβλέπουσα του μοντέλου, δηλαδή ένα γραμμικό συνδυασμό των επεξηγηματικών μας μεταβλητών της μορφής

$$a + b_1 x_1 + \cdots + b_k x_k$$

όπου τα $a, b_1, \dots, b_k$ θα προσδιοριστούν κατά την προσαρμογή του μοντέλου. Έτσι, αν g η συνάρτηση που επιτυγχάνει αυτή τη συσχέτιση, η οποία ονομάζεται συνάρτηση σύνδεσης (link function), το αντίστοιχο γ.γ.μ. είναι της μορφής

$$g(\mu) = g(E[Y]) = a + b_1 x_1 + \cdots + b_k x_k.$$

Λογιστική Παλινδρόμηση

Ειδική περίπτωση τέτοιων μοντέλων αποτελεί η λεγόμενη λογιστική παλινδρόμηση. Το εν λόγω μοντέλο αποσκοπεί στη μελέτη μεταβλητών που ως προς την απόκριση είναι δυαδικές, δηλαδή θα μπορούσαμε να τις κωδικοποιήσουμε με χρήση 0 ή 1, «επιτυχία» ή «αποτυχία». Για παράδειγμα, η αποτελεσματικότητα ή όχι μιας θεραπείας, το αν κάποιος καπνίζει ή όχι είτε το αν κάποιος που αναζητά εργασία διαθέτει προϋπηρεσία ή όχι αποτελούν μεταβλητές του τύπου που μόλις περιγράψαμε. Το αντίστοιχο μοντέλο πιθανοτήτων για αυτές τις μεταβλητές αποκαλείται «Δοκιμή Bernoulli» και χαρακτηρίζεται με μια πιθανότητα p η οποία αντιστοιχεί στην τιμή 1 της μεταβλητής (την οποία θα αποκαλούμε «επιτυχία»), ενώ η συμπληρωματική της πιθανότητα αντιστοιχεί στην τιμή 0 («αποτυχία»).

Πέραν αυτού, η λογιστική παλινδρόμηση βρίσκει εφαρμογή και σε δεδομένα στα οποία σε n το πλήθος δοκιμών μετράμε το πόσες «επιτυχίες» είχαμε. Στα συγκεκριμένα δεδομένα, η

κατανομή που θεωρούμε στη μεταβλητή απόκρισης είναι η διωνυμική κατανομή, με τύπο για τη συνάρτηση πιθανότητας:

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

όπου n το πλήθος των δοκιμών και p η πιθανότητα «επιτυχίας».

Σε ό,τι αφορά τη μορφή του μοντέλου, αν θεωρήσουμε π.χ. το απλό μοντέλο με δυαδικά δεδομένα (0 ή 1) η μέση τιμή είναι ίση με την πιθανότητα «επιτυχίας», ενώ συνηθίζεται να λαμβάνεται η λεγόμενη συνάρτηση logit ως συνάρτηση σύνδεσης, όπου:

$$logit(p_x) = \ln\left(\frac{p_x}{1-p_x}\right) = a + b_1 x_1 + \cdots b_k x_k$$

Οι παρατηρήσεις θεωρούνται ανεξάρτητες. Εάν αντιστρέψουμε τη συνάρτηση σύνδεσης, παίρνουμε

$$p_x = \frac{e^{\eta_x}}{1 + e^{\eta_x}}$$

όπου $\eta_x = a + b_1 x_1 + \cdots + b_k x_k$ η γραμμική προβλέπουσα. Η τελευταία σχέση είναι συνεπής με το ότι το p_x εκφράζει πιθανότητα και, ως εκ τούτου, πρέπει να κυμαίνεται από 0 έως 1.

Όσον αφορά την ερμηνεία των συντελεστών του μοντέλου: αρχικά η ουσιαστική πληροφορία αφορά στα b_j. Με βάση την προηγούμενη σχέση βλέπουμε ότι ο λόγος των συμπληρωματικών πιθανοτήτων (ή odds όπως αναφέρεται στην ξένη βιβλιογραφία) σχετίζεται ως εξής με τη γραμμική προβλέπουσα:

$$\frac{\widehat{p_x}}{1-\widehat{p_x}} = e^{\boldsymbol{x}'\hat{\boldsymbol{b}}} = e^{\widehat{b_0} + x_1\widehat{b_1} + \cdots + x_k\widehat{b_k}}$$

όπου το σύμβολο ^ αναφέρεται στην αντίστοιχη εκτίμηση της εκάστοτε παραμέτρου, ενώ $\boldsymbol{x} = (1, x_1, \ldots, x_k)$ και $\boldsymbol{b} = (b_0, b_1, \ldots, b_k)$. Από αυτή τη σχέση φαίνεται, λοιπόν, το αντίκτυπο που έχει στα odds ο κάθε συντελεστής b_j. Ειδικότερα, βλέπουμε ότι αν αυξήσουμε κατά μια μονάδα τη j-οστή μεταβλητή, αφήνοντας αμετάβλητες τις υπόλοιπες, αναμένεται ότι τα odds θα μεταβληθούν κατά e^{β_j}.

Στη συνέχεια, θα επεκταθούμε στον τρόπο μελέτης της λογιστικής παλινδρόμησης μέσω ενός παραδείγματος.

<u>*Παράδειγμα*</u> (Πηγή:[5]): Έστω ότι θέλουμε να μελετήσουμε την αποτελεσματικότητα τριών διαφορετικών τύπων εντομοκτόνων (Ε), για τα οποία μελετήθηκε ο αριθμός εντόμων y που θανατώθηκαν ύστερα από έξι ημέρες έκθεσής τους στο εν λόγω εντομοκτόνο με χρήση διαφορετικής, κάθε φορά, ποσότητας x (σε χιλιοστόγραμμα (mg)). Η έρευνα μας έδωσε τα αντίστοιχα δεδομένα (με n συμβολίζουμε τον αριθμό των δοκιμών που έγινε κάθε φορά):

	y	n	E	x
1	3	50	1	2
2	5	49	1	2.64
3	19	47	1	3.48
4	19	38	1	4.59
5	24	29	1	6.06
6	35	50	1	8
7	2	50	2	2
8	14	49	2	2.64
9	20	50	2	3.48
10	27	50	2	4.59
11	41	50	2	6.06
12	40	50	2	8
13	28	50	3	2
14	37	50	3	2.64
15	46	50	3	3.48
16	48	50	3	4.59
17	48	50	3	6.06
18	50	50	3	8

Πίνακας 13: Δεδομένα για τρία διαφορετικά είδη εντομοκτόνων προς μελέτη με λογιστική παλινδρόμηση

Στη συνέχεια θα χρησιμοποιήσουμε την R για τη μελέτη του συγκεκριμένου μοντέλου. Με χρήση της εντολής glm, συγκεκριμένα, προσαρμόζουμε το μοντέλο της λογιστικής παλινδρόμησης με απόκριση τη μεταβλητή f, η οποία εμπεριέχει τις επιτυχίες και τις αποτυχίες, που στη συγκεκριμένη περίπτωση είναι η θανάτωση των εντόμων, και επεξηγηματικές μεταβλητές τον τύπο του εντομοκτόνου E και τη ποσότητα σε mg του κάθε εντομοκτόνου x. Έτσι, παίρνουμε τα εξής ως σύνοψη του μοντέλου:

```
"Call:

glm(formula = f ~ E + x, family = "binomial", data = d)


Deviance Residuals:

    Min       1Q   Median       3Q      Max

-2.9537  -1.4831   0.6142   1.2207   2.1547


Coefficients:

            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -3.22136    0.27443 -11.739   <2e-16 ***
E2           0.36953    0.20394   1.812     0.07 .
E3           2.68802    0.24067  11.169   <2e-16 ***
x            0.63168    0.05189  12.174   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 413.644  on 17  degrees of freedom
Residual deviance:  48.026  on 14  degrees of freedom
AIC: 118.22

Number of Fisher Scoring iterations: 4”
```

Παρατηρούμε, αρχικά, ότι στο κομμάτι των συντελεστών, η p-τιμές των αντίστοιχων συντελεστών είναι εξαιρετικά μικρές (<2e-16) γεγονός που μας υποδεικνύει ότι είναι στατιστικά σημαντικό το μοντέλο μας, έχει νόημα δηλαδή να αναζητήσουμε σχέση μεταξύ του αριθμού θανόντων εντόμων με το είδος και την ποσότητα του εντομοκτόνου. Επιπλέον, η ερμηνεία του κάθε συντελεστή για τις επεξηγηματικές μεταβλητές είναι η εξής:

- Για χρήση ίδιας ποσότητας εντομοκτόνου, τα odds για τη θνησιμότητα των εντόμων πολλαπλασιάζονται με τον παράγοντα $e^{0.36953} = 1.447$ στην περίπτωση του 2ου εντομοκτόνου σε σχέση με το πρώτο, δηλαδή τα odds σημειώνουν αύξηση 44.7% στο 2ο εντομοκτόνο εν σχέσει με το 1ο.

- Για χρήση ίδιας ποσότητας εντομοκτόνου, τα odds για τη θνησιμότητα των εντόμων πολλαπλασιάζονται με τον παράγοντα $e^{2.68802} = 14.702$, οπότε έχουμε 14πλάσιο odds στην περίπτωση του 3ου εντομοκτόνου σε σχέση με το πρώτο. Δηλαδή, ο λόγος της πιθανότητας επιτυχίας δια της αντίστοιχης πιθανότητας αποτυχίας εκτιμάται ότι είναι 14πλάσιος εκείνου του εντομοκτόνου τύπου 1.

-Για ένα εντομοκτόνο τύπου 1, αν αυξήσουμε κατά 1mg τη δόση του εντομοκτόνου, το odds για τη θνησιμότητα των εντόμων πολλαπλασιάζεται κατά $e^{0.63168} = 1.88$, αυξάνεται δηλαδή κατά ένα ποσοστό περί το 88%.

Αν, τώρα, θέλουμε να κάνουμε πρόβλεψη, θα μπορούσαμε λ.χ. να εκτιμήσουμε ότι για x=2.58mg και με τη χρήση του εντομοκτόνου 2ου τύπου, έχουμε πιθανότητα

$$\hat{p} = \frac{\exp(-3.22136 + 0.36953 + 0.63168 \cdot 2.58)}{1 + \exp(-3.22136 + 0.36953 + 0.63168 \cdot 2.58)} = 0.2276$$

ή αλλιώς εκτιμάται ότι θα είναι αποτελεσματικό κατά 22.76% να σκοτώσει ένα έντομο.

Η καμπύλη Roc

Ένας δείκτης καλής προσαρμογής για την παλινδρόμησή μας είναι η προβλεπτική ικανότητα του μοντέλου: εάν έχουμε $\hat{p}$ την εκτίμηση της πιθανότητας επιτυχίας και p_0 είναι κάποιο όριο που θέτουμε εμείς, τότε

- αν $\hat{p} \geq p_0$, προβλέπουμε ότι $Y = 1$
- αν $\hat{p} < p_0$ προβλέπεται $Y = 0$.

Συγκρίνοντας, λοιπόν, τις προβλέψεις αυτές με τις πραγματικές τιμές της μεταβλητής απόκρισης, σχηματίζεται ο ακόλουθος πίνακας:

Πρόβλεψη	Πραγματική τιμή Y=1	Y=0	
Y=1	a	b	a+b
Y=0	c	d	c+d
	a+c	b+d	n

Πίνακας 14: Πίνακας προβλεπτικής ικανότητας μοντέλου, σύγκριση εκτίμησης πιθανότητας επιτυχίας με δοθέν κατώτατο όριο

οπότε με a, παραδείγματος χάριν, συμβολίζουμε εκείνες τις τιμές που προβλέψαμε σωστά ότι θα ισούνται με 1, b οι τιμές που ενώ ήταν 0 εμείς τις προβλέψαμε ως 1 κ.ο.κ. Βάσει αυτού, ορίζονται οι εξής χρήσιμες ποσότητες:

- Ευαισθησία (sensitivity): Η ευαισθησία ορίζεται ως το πηλίκο $\frac{a}{a+c}$, είναι δηλαδή το ποσοστό ορθής πρόβλεψης της κατάστασης $Y = 1$, με άλλα λόγια είναι δηλαδή το ποσοστό των αληθώς θετικών αποτελεσμάτων (true positive rate).
- Ειδικότητα (specificity): Το συγκεκριμένο μέγεθος είναι ίσο με $\frac{d}{b+d}$, οπότε εκφράζει το ποσοστό ορθής πρόβλεψης της κατάστασης $Y = 0$, δηλαδή το ποσοστό των αληθώς αρνητικών αποτελεσμάτων (true negative rate).

Επίσης χρήσιμο είναι και το μέγεθος $\frac{b}{b+d}$, το οποίο εκφράζει το ποσοστό των ψευδώς θετικών αποτελεσμάτων. Αν οι τιμές της ευαισθησίας και της ειδικότητας βρεθούν για οποιαδήποτε επιλογή του p_0 στο εύρος [0,1], τότε σχηματίζεται η χαρακτηριστική καμπύλη ROC (Receiver operating characteristic curve) του μοντέλου, η οποία απεικονίζει την προβλεπτική ικανότητα του μοντέλου καθώς μεταβάλλουμε το p_0. Όπως είδαμε, απαραίτητη προϋπόθεση για να μπορούμε να σχεδιάσουμε την καμπύλη ROC είναι τα δεδομένα μας να βρίσκονται σε δυαδική μορφή.

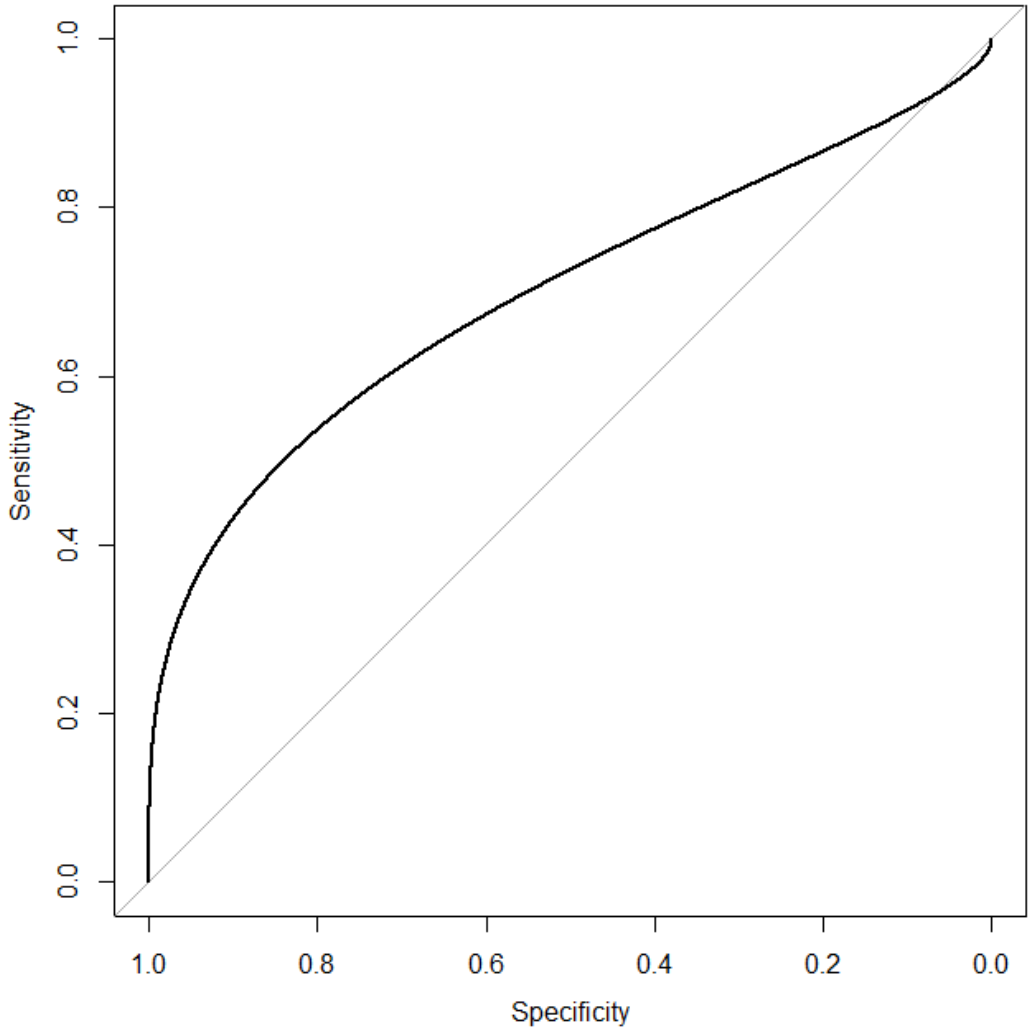
Παράδειγμα :

Θα εξετάσουμε, στο παρόν παράδειγμα, τη σχέση μεταξύ της ηλικίας μιας ομάδας ασθενών και την απόκρισή τους σε μια θεραπεία κατά της λευχαιμίας (τα δεδομένα που θεωρούμε

παρατίθενται στο παράρτημα, στον πίνακα 3). Προσαρμόζοντας το μοντέλο λογιστικής παλινδρόμησης παίρνουμε τα κάτωθι δεδομένα

```
”Call:

glm(formula = d2$response ~ d2$age_1, family = "binomial")


Deviance Residuals:

    Min       1Q   Median       3Q      Max

-1.5910  -0.9679  -0.7056   1.0054   1.8019


Coefficients:

            Estimate Std. Error z value Pr(>|z|)

(Intercept)  2.19678    1.00548   2.185   0.0289 *

d2$age_1    -0.04676    0.01952  -2.395   0.0166 *

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


(Dispersion parameter for binomial family taken to be 1)


    Null deviance: 70.524  on 50  degrees of freedom

Residual deviance: 64.004  on 49  degrees of freedom

AIC: 68.004


Number of Fisher Scoring iterations: 4”
```

ενώ η καμπύλη ROC για το συγκεκριμένο μοντέλο είναι η εξής:



Γράφημα 7: Καμπύλη Roc που αντιστοιχεί στα δεδομένα του παραδείγματος

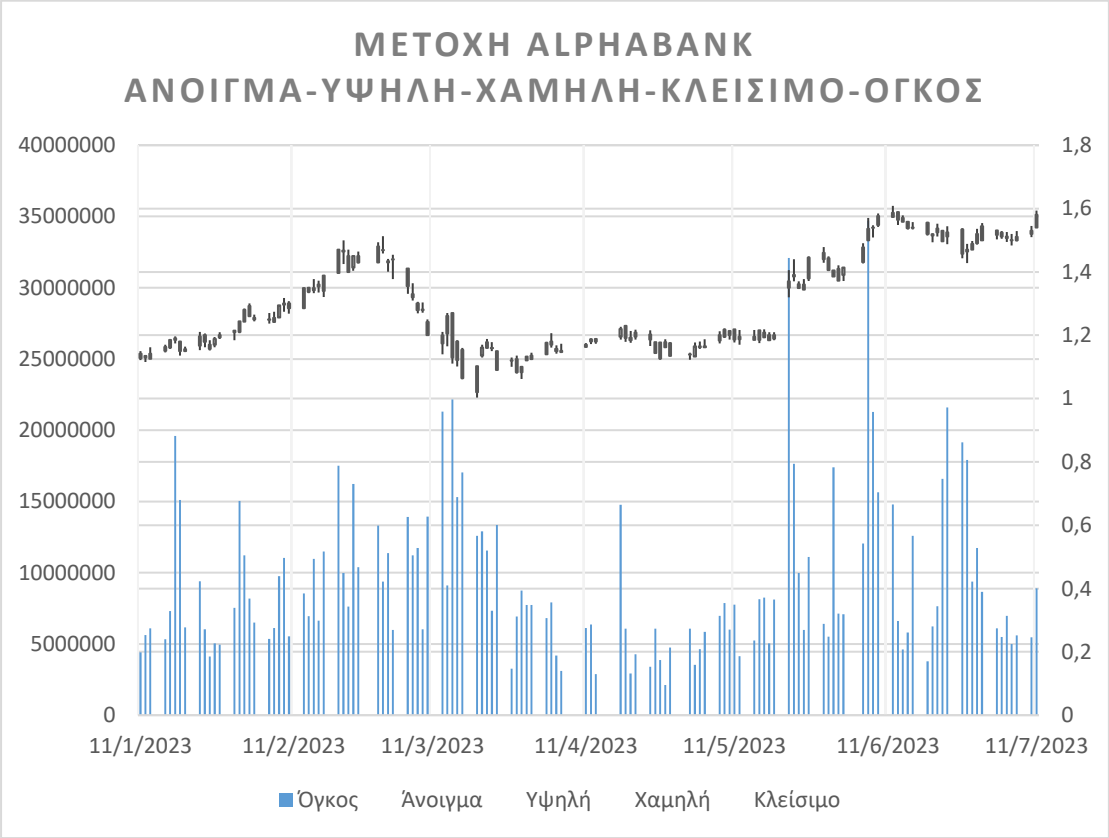
Η R μας δίνει επίσης ότι το εμβαδό κάτω από την καμπύλη ισούται με 0.697 .

Η ερμηνεία των παραπάνω, τώρα, είναι η εξής: ένα ιδανικό μοντέλο έχει εξαιρετικά μεγάλη ευαισθησία και ταυτόχρονα υψηλή ειδικότητα, οπότε η καμπύλη ROC σε αυτό το ιδανικό σενάριο συμπίπτει με την πάνω αριστερή γωνία του ανωτέρω γραφήματος. Η διαγώνια ευθεία γραμμή που φαίνεται στο σχήμα, από την άλλη, εκφράζει το ενδεχόμενο τα αληθώς θετικά και ψευδώς θετικά αποτελέσματα να έχουν τα ίδια ποσοστά, οπότε η πρόβλεψη που θα κάναμε θα ήταν ανεξάρτητη της μεταβλητής απόκρισης Y. Με άλλα λόγια, θα ήταν άχρηστη η εφαρμογή του μοντέλου λογιστικής παλινδρόμησης. Στο συγκεκριμένο παράδειγμα, το εμβαδό κάτω από την καμπύλη (AUC, area under the curve) ισούται με 0,697, με μέγιστη τιμή γενικά να είναι το 1 και στο «χειρότερο» σενάριο να είμαστε στο 0,5, οπότε έχουμε μια σχετικά καλή προσαρμογή μέσω της λογιστικής παλινδρόμησης. Ενδεχομένως αυτό να βελτιωνόταν αν εισάγαμε περαιτέρω μεταβλητές στο μοντέλο.

CASE STUDY: ΠΡΟΒΛΕΨΕΙΣ ΓΙΑ ΔΕΔΟΜΕΝΑ ΜΕΤΟΧΩΝ

Μετά την ανάπτυξη της βασικής θεωρίας γύρω από θέματα πρόβλεψης, θα προχωρήσουμε στην αξιοποίησή της για τη μελέτη της μετοχής της Alpha Bank. Θα εξετάσουμε τα δεδομένα της μετοχής κατά την περίοδο 11.01.23-11.07.23 (πηγή: ιστοσελίδα Euro2day, βλέπε Παράρτημα, πίνακα Π4). Συγκεκριμένα θα εφαρμόσουμε τις διάφορες μεθόδους που έχουμε δει στην εργασία, για να κάνουμε προβλέψεις τόσο για την τιμή κλεισίματος της μετοχής της Alpha Bank (δηλαδή της τιμής που φτάνει η μετοχή κατά το κλείσιμο του χρηματιστηρίου την δεδομένη μέρα), όσο και για τον όγκο της μετοχής (δηλαδή της ποσότητας των μονάδων της μετοχής που είναι διαθέσιμα προς πώληση).

Ξεκινάμε με μια επισκόπηση των δεδομένων με χρήση γραφημάτων και βασικών εργαλείων της περιγραφικής στατιστικής.



Γράφημα 8: Candlestick γράφημα τιμής της μετοχής (με δεδομένα ανοίγματος, χαμηλής, υψηλής και τιμής κλεισίματος) και γράφημα όγκου μετοχής

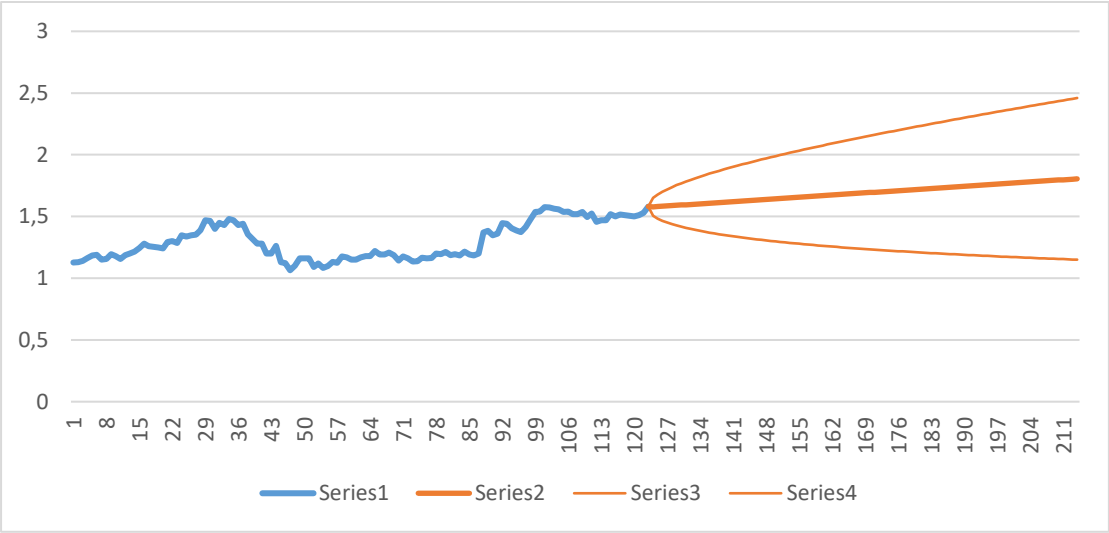
Ήδη μπορούμε να αποκτήσουμε κάποια ιδέα για τη συμπεριφορά των χρονοσειρών που αφορούν τον όγκο αλλά και τα υπόλοιπα στοιχεία της μετοχής (άνοιγμα/τιμή έναρξης, υψηλότερη τιμή, χαμηλότερη τιμή και τιμή κλεισίματος). Το «ιστορικό» της τιμής της μετοχής, λίγο-πολύ- κατά αυτήν την περίοδο θα μπορούσαμε να πούμε ότι είναι μια αύξουσα πορεία μέχρι τις αρχές του Μαρτίου του 2023 (μιας και τόσο η χαμηλή όσο και η υψηλή τιμή έχουν, συνολικά, μια ανοδική πορεία), ύστερα σημειώνεται πτώση και κατόπιν η τιμή

σταθεροποιείται, λίγο-πολύ, γύρω απ' το 1.2 κατά τον Απρίλιο και το Μάιο του 2023. Από τα τέλη Μαΐου και έπειτα, έχουμε εκ νέου μια ανοδική πορεία της μετοχής.

Στατιστικά μεγέθη	Άνοιγμα	Υψηλή	Χαμηλή	Κλείσιμο	Όγκος
Μέσος	1..301463415	1.322813008	1.279658537	1.300239837	9376876.447
Τυπικό σφάλμα	0.013581693	0.013722633	0.013525883	0.013649039	502595.6007
Διάμεσος	1.25	1.27	1.237	1.25	7627576
Επικρατούσα τιμή	1.18	1.27	1.162	1.2	
Μέση απόκλιση τετραγώνου	0.150628263	0.152191362	0.150009294	0.151375163	5574054.858
Διακύμανση	0.022688874	0.023162211	0.022502788	0.02291444	3.10701E+13
Κύρτωση	-1.26263395	-1.37599023	-1.22383963	-1.329289285	4.469815427
Ασυμμετρία	0.424048107	0.405597179	0.421401935	0.399551836	1.797751134
Εύρος	0.569	0.5085	0.5695	0.5155	32567701
Ελάχιστο	1.02	1.1	1.003	1.0645	2121345
Μέγιστο	1.589	1.6085	1.5725	1.58	34689046
Άθροισμα	160.08	162.706	157.398	159.9295	1153355803
Πλήθος	123	123	123	123	123
Βαθμός εμπιστοσύνης(95.0%)	0.026886318	0.027165323	0.026775835	0.027019636	994938.1901

Πίνακας 15 Περιγραφικά στατιστικά μεγέθη των δεδομένων

Ακολουθεί μια γραφική παράσταση που υλοποιεί πρόβλεψη για 3 μήνες ύστερα από τη λήξη (δηλαδή μέχρι 10.10.23) της προαναφερθείσας χρονικής περιόδου: Λωρίδα χρόνου



Γράφημα 9: Τιμή κλεισίματος Alpha Bank, περίοδος 11.01.23-11.07.23 και προβλέψεις (για την τιμή κλεισίματος0 μέχρι τις 10.10.23. (Σειρά 1: Πραγματικές Τιμές, Σειρά 2: Προβλέψεις, Σειρά 3,4: Άνω και Κάτω Άκρα Εμπιστοσύνης)

Ανάμεσα στις τρεις πορτοκαλί γραμμές, η μεσαία είναι η πρόβλεψη χρησιμοποιώντας ένα μοντέλο με γραμμική τάση για τη χρονοσειρά της τιμής κλεισίματος, ενώ οι δυο καμπύλες που την περιβάλλουν αποτελούν τα πάνω και κάτω φράγματα στις τιμές των προβλέψεων από αντίστοιχα 95% διαστήματα εμπιστοσύνης. Επιπλέον στατιστικοί δείκτες είναι οι εξής:

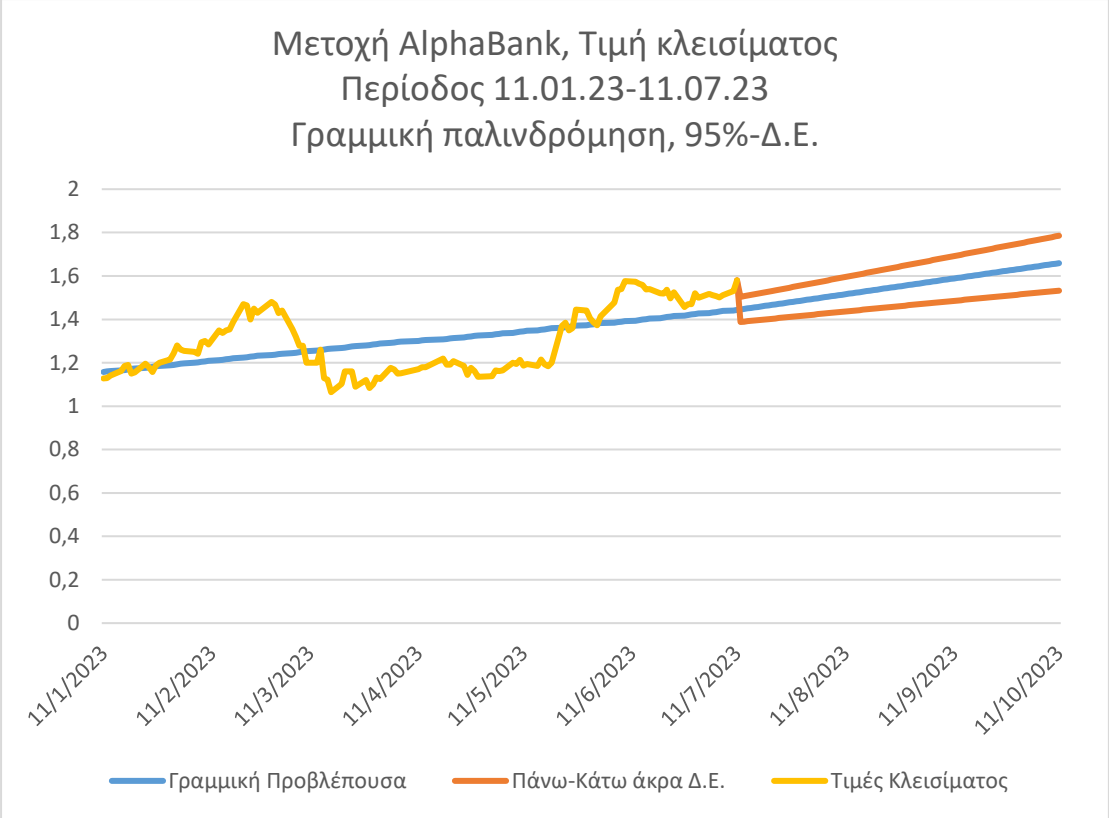
Στατιστικά στοιχεία	Τιμή
α	0.90
β	0.00
γ	0.00
MASE	0.73
SMAPE	0.01
MAE	0.02
RMSE	0.03

Πίνακας 16: Παράμετροι και μέτρα σφάλματος μεθόδου πρόβλεψης για τα δεδομένα 6 μηνών

Τα α,β,γ είναι οι παράμετροι που έχουμε συναντήσει στις αντίστοιχες μεθόδους εκθετικής εξομάλυνσης. Το MASE αποτελεί το μέσο απόλυτο σταθμισμένο σφάλμα (Mean Absolute Scaled Error), το SMAPE το συμμετρικό μέσο απόλυτο ποσοστιαίο σφάλμα (Symmetric Mean Absolute Percentage Error), MAE το μέσο απόλυτο σφάλμα και RMSE η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος. Με εξαίρεση το MASE, τα υπόλοιπα μεγέθη είναι αρκετά μικρά, υποδηλώνοντας μια καλή προσαρμογή για το αντίστοιχο μοντέλο.

Αξίζει να σχολιαστούν, επίσης, οι ενδείξεις που παίρνουμε από το γράφημα. Κρίνοντας από τα γραφήματα τα σημεία των οποίων αποτελούν τα 95% διαστήματα εμπιστοσύνης για την εκάστοτε πρόβλεψη, έχουμε σημαντικές ενδείξεις ότι θα σημειωθεί περαιτέρω αύξηση της τιμής της μετοχής της Alpha Bank, ενώ παράλληλα η μικρότερη τιμή ανάμεσα στα κάτω φράγματα είναι περίπου 1.1504, οπότε έχουμε ενδείξεις υπέρ του ότι η τιμή κλεισίματος της μετοχής δεν πρόκειται να «πέσει» πιο κάτω από αυτό το μέγεθος. Γενικότερα, είμαστε κατά 95% βέβαιοι ότι η εξέλιξη της τιμής κλεισίματος της μετοχής θα πραγματοποιηθεί στην περιοχή που οριοθετείτε από τις καμπύλες των άνω και κάτω φραγμάτων των προβλέψεων.

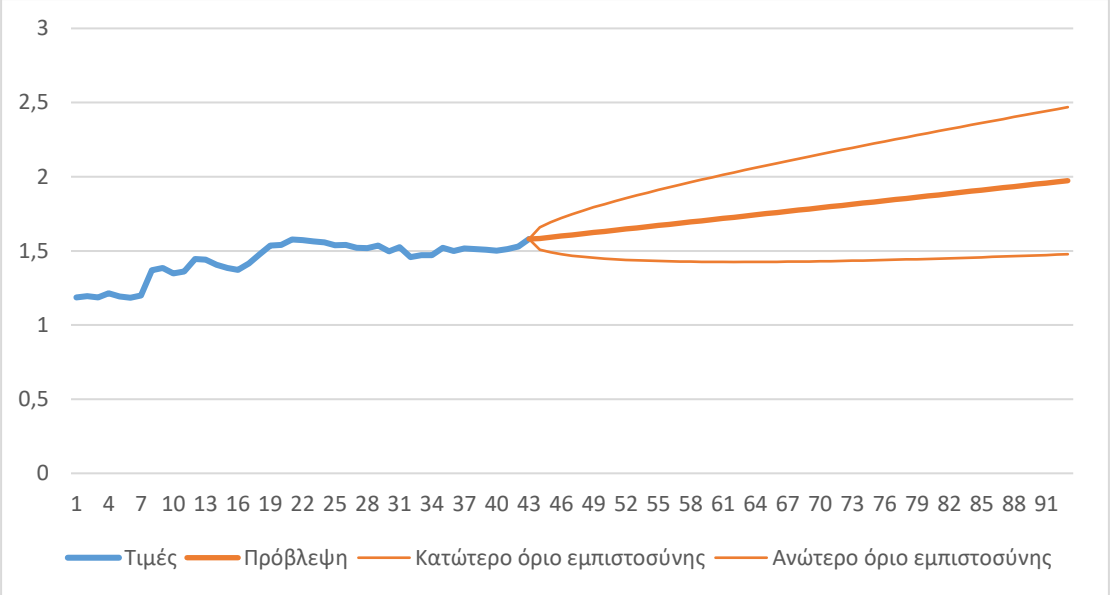
Μπορούμε, επίσης, να εφαρμόσουμε γραμμική παλινδρόμηση για να πραγματοποιήσουμε προβλέψεις. Στο γράφημα που ακολουθεί βλέπουμε τη γραμμική προβλέπουσα για την τιμή κλεισίματος (επεξηγηματική μεταβλητή είναι οι ημέρες που έχουν εκφρασθεί ως ακέραιοι 1,2 κ.ο.κ.) μαζί με ένα 95%-Δ.Ε. γύρω από τις προβλέψεις που γίνονται μέχρι και τις 11 Οκτωβρίου:



Γράφημα 10: Τιμή κλεισίματος μετοχής Alpha Bank για την περίοδο 11.1.23-11.7.23 και προβλέψεις με γραμμική παλινδρόμηση

Εν τοιαύτη περιπτώσει, λαμβάνουμε MAE= 0.105604 και RMSE= 0.12299. Θεωρητικά, θα έπρεπε να εξετάσουμε αν πληρούνται όλες οι προϋποθέσεις για την προσαρμογή του γραμμικού μοντέλου (με τους γραφικούς ελέγχους που επισημάναμε στο σχετικό κεφάλαιο), ωστόσο παρατηρούμε ότι τόσο το μέσο απόλυτο όσο και το μέσο τετραγωνικό σφάλμα είναι αρκετά μεγαλύτερα από το μοντέλο που χρησιμοποιήσαμε νωρίτερα. Συνεπώς, στην προκειμένη περίπτωση, παρ' ότι ενδέχεται να μπορούμε να εφαρμόσουμε γραμμική παλινδρόμηση στα δεδομένα της τιμής κλεισίματος της μετοχής, ενδεχομένως να μην είναι η πιο αποδοτική μέθοδος.

Σε περίπτωση που περιορίσουμε περαιτέρω την χρονική περίοδο που εξετάζουμε, λ.χ. στους 2 μήνες, περιορίζεται επίσης και το χρονικό διάστημα για το οποίο μπορούμε να πραγματοποιήσουμε προβλέψεις. Ακολουθώντας αντίστοιχη διαδικασία με πριν, παίρνουμε πάλι το ακόλουθο γράφημα:



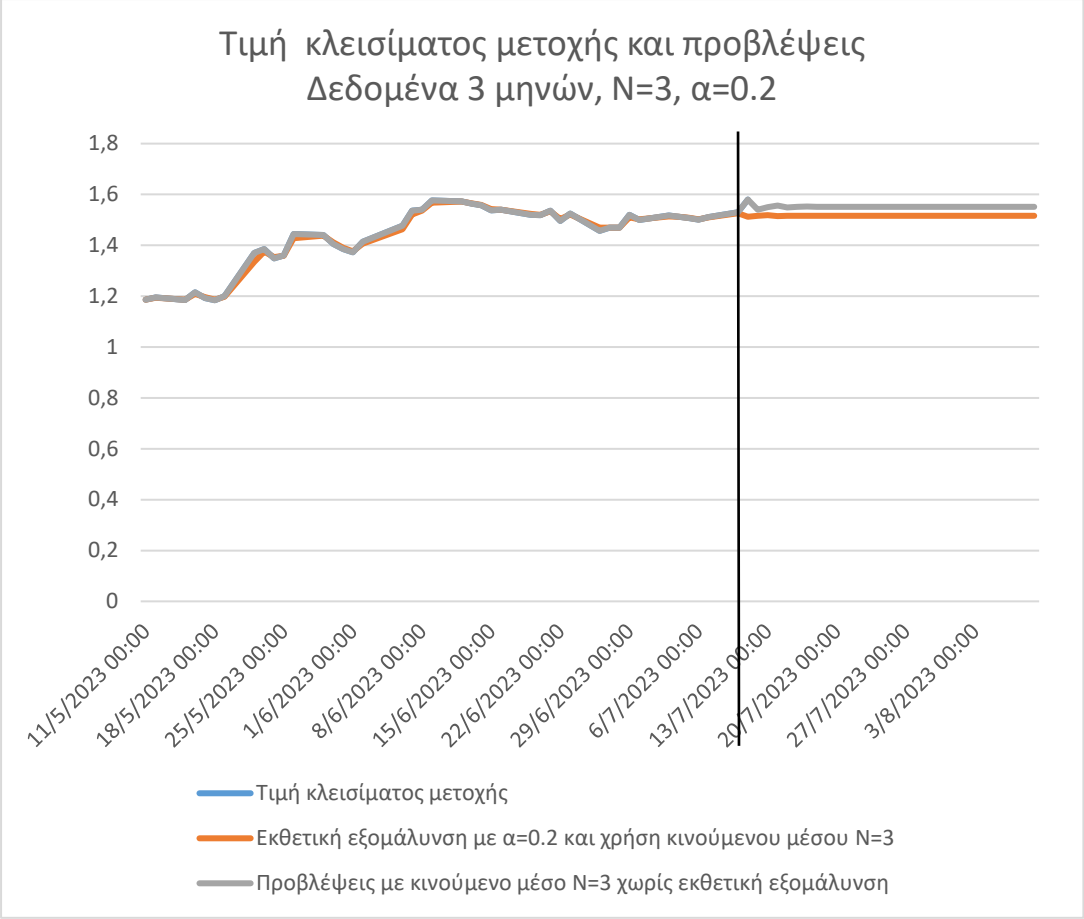
Γράφημα 11: Τιμή κλεισίματος μετοχής Alpha Bank για την περίοδο 11.5.23-11.7.23 , με προβλέψεις μέχρι και 31.8.23 και αντίστοιχα όρια 95%-διαστήματος εμπιστοσύνης

(προβλέψεις μέχρι και τις 31.08.23) με τα αντίστοιχα στατιστικά να είναι τα ακόλουθα:

Στατιστικά στοιχεία	Τιμή
Alpha	0.90
Beta	0.00
Gamma	0.00
MASE	0.62
SMAPE	0.01
MAE	0.02
RMSE	0.02

Πίνακας 17: Παράμετροι και μέτρα σφάλματος της μεθόδου πρόβλεψης για τα δεδομένα τριών μηνών

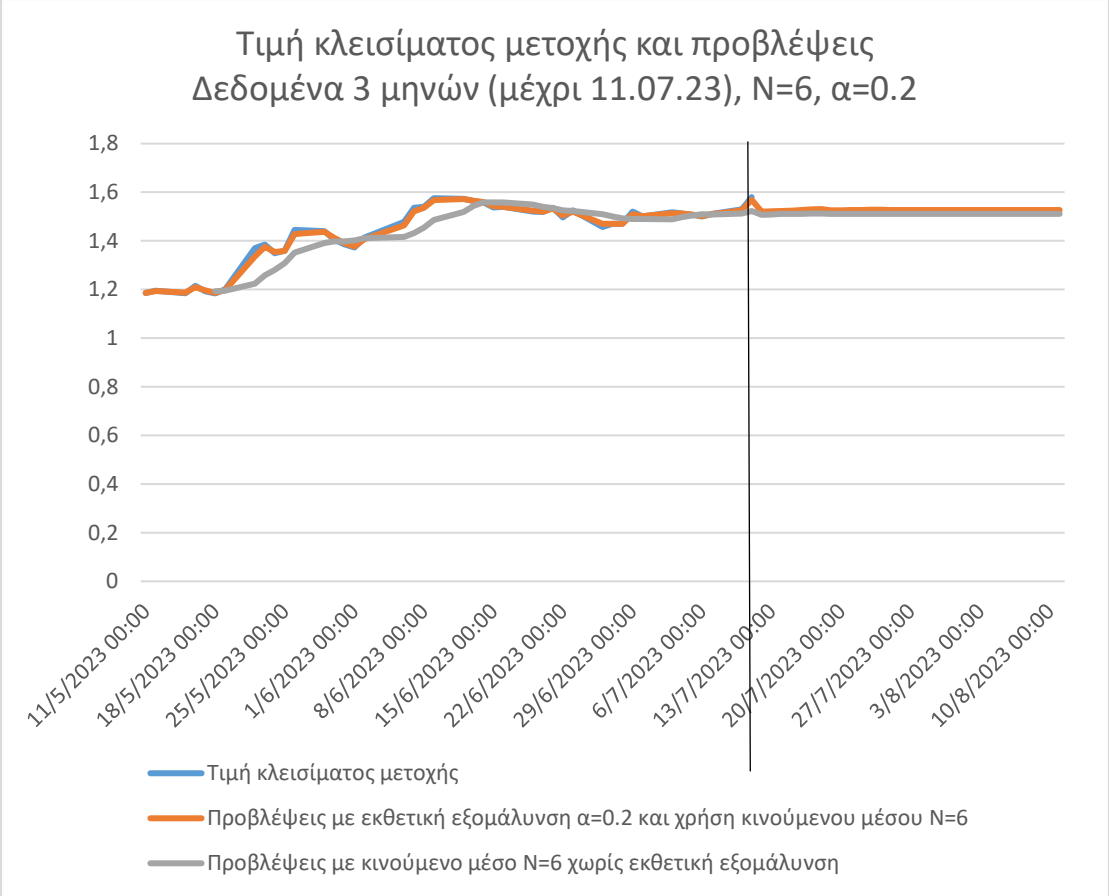
Μπορούμε, έπειτα, να εφαρμόσουμε και άλλες μεθόδους πρόβλεψης από εκείνες που αναφέραμε. Για παράδειγμα, το ακόλουθο γράφημα απεικονίζει τις προβλέψεις με τη μέθοδο του κινούμενου μέσου για N=3, χωρίς τη χρήση εκθετικής εξομάλυνσης αλλά και με τη χρήση της για α=0.2:



Γράφημα 12: Γράφημα σύγκρισης μεθόδων πρόβλεψης κινούμενο μέσου με ή χωρίς εκθετική εξομάλυνση (α=0.2, N=3)

Η κατακόρυφη μαύρη γραμμή αντιστοιχεί στο σημείο απ' το οποίο και έπειτα πραγματοποιούνται προβλέψεις (δηλαδή από την ημερομηνία 11.07.23 και μετά). Οπτικά, η διαφορά στην εφαρμογή των μεθόδων έγκειται στο ότι οι προβλέψεις με τη χρήση της εκθετικής εξομάλυνσης είναι κατά τι πιο συντηρητικές εν σχέσει με εκείνες που έγιναν χωρίς αυτήν την τεχνική.

Επαναλαμβάνοντας τη διαδικασία, θεωρώντας αυτή τη φορά N=6 παίρνουμε το εξής:



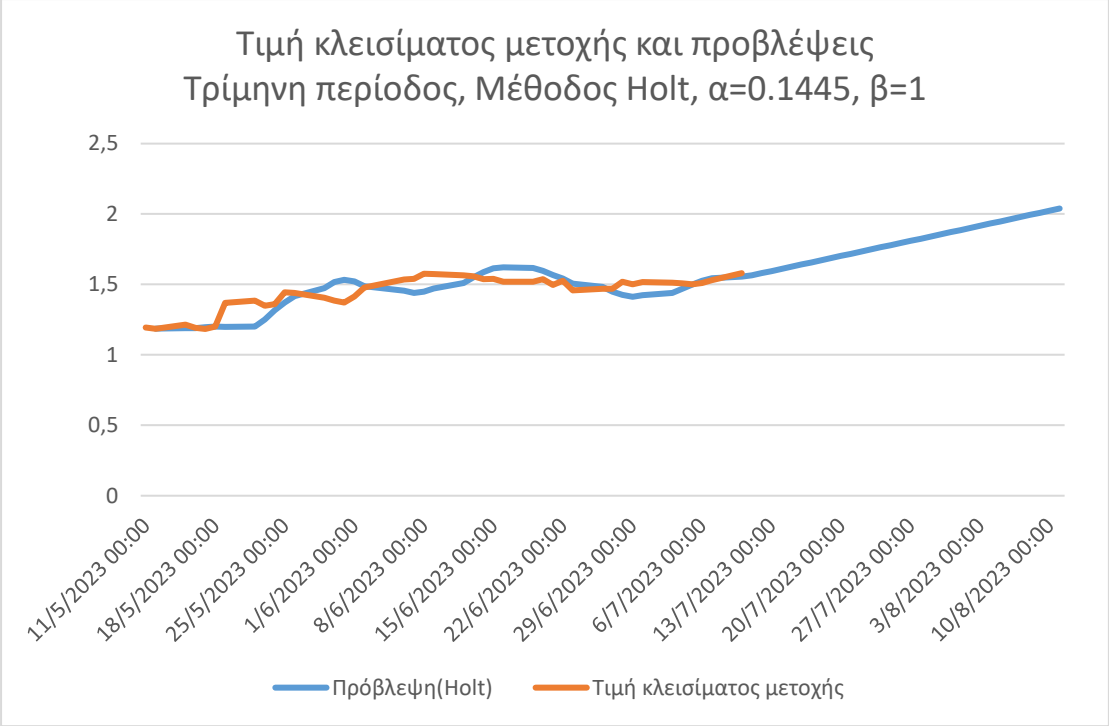
Γράφημα 13: Γράφημα σύγκρισης μεθόδων πρόβλεψης κινούμενο μέσου με ή χωρίς εκθετική εξομάλυνση (α=0.2, N=6)

Τα μέτρα σφάλματος για τις μεθόδους είναι τα εξής:

MAE (εκθ. εξομάλυνση με α=0.2)=	0.033845
RMSE (εκθ. εξομάλυνση με α=0.2)=	0.03947
MAE (κινούμενος μέσος με N=6)=	0.036958
RMSE (κινούμενος μέσος με N=6) =	0.0476

Πίνακας 18 Μέτρα Σφάλματος των μεθόδων για τα δεδομένα τριών μηνών

Ύστερα, θα εφαρμόσουμε τη μέθοδο του Holt για να εξετάσουμε εάν υπάρχει το ενδεχόμενο δημιουργίας γραμμικής τάσης στα δεδομένα. Θα εστιάσουμε, για ακόμη μια φορά, στα δεδομένα της προαναφερθείσας τρίμηνης περιόδου. Εισάγοντας δεδομένα μας στο Excel και βελτιστοποιώντας την επιλογή των α,β με βάση την ελαχιστοποίηση του μέσου απολύτου σφάλματος, παίρνουμε α=0.144549 και β=1. Τα αποτελέσματα απεικονίζονται στο γράφημα που ακολουθεί:



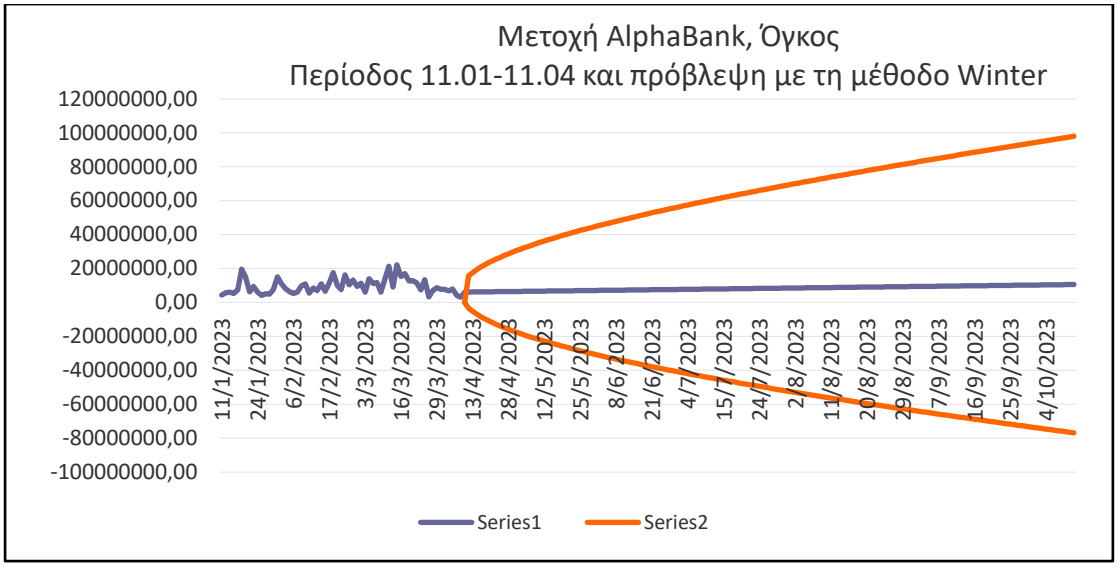
Γράφημα 14: Εφαρμογή της μεθόδου Holt στην τιμή κλεισίματος της μετοχής Alpha Bank (επιλογή παραμέτρων ούτως ώστε να ελαχιστοποιηθεί το μέγιστο απόλυτο σφάλμα)

Τα μέτρα σφάλματος εν τοιαύτη περίπτωση είναι MAE= 0.011879 και RMSE=0.012369.

Μέχρι στιγμής, ανάμεσα στις μεθόδους που χρησιμοποιήσαμε, η τελευταία μοιάζει να είναι οι πλέον αποδοτική, αφού τόσο από την άποψη του μέσου απολύτου σφάλματος όσο και από αυτή του μέσου τετραγωνικού σφάλματος (ή, ισοδύναμα, της τετραγωνικής του ρίζας), οι ποσότητες παρουσιάζουν την ελάχιστη τιμή τους κατά την εφαρμογή της τελευταίας μεθόδου.

Ωστόσο, η ελαχιστοποίηση του μέσου σφάλματος δεν είναι το μοναδικό κριτήριο για την επιλογή ενός καλού μοντέλου πρόβλεψης. Μεγαλύτερη πρακτική σημασία έχει η δημιουργία και ερμηνεία διαστημάτων εμπιστοσύνης γύρω από τις προβλέψεις, ούτως ώστε να μπορούμε να εξασφαλίσουμε σε ένα βαθμό την ασφάλεια απέναντι σε διάφορους κινδύνους. Εάν, για παράδειγμα, ήμασταν αγοραστές της μετοχής, ξεκινώντας με τιμή κλεισίματος λ.χ. 1.84 και έχουμε 95% διαστήματα εμπιστοσύνης γύρω από τις προβλέψεις που μας υποδηλώνουν ότι στο διάστημα πρόβλεψης η τιμή κλεισίματος δεν θα κατέβει λ.χ. από τα 1.85, τότε μπορούμε να είμαστε βέβαιοι- έως ένα βαθμό- ότι η επένδυση σε μετοχές της Alpha Bank είναι μια ασφαλής και κερδοφόρα επένδυση για το διάστημα που ακολουθεί. Ομοίως, με αντίστοιχη εναρκτήρια τιμή στην τιμή κλεισίματος, αν είχαμε διαστήματα εμπιστοσύνης που μας υποδεικνύουν ότι δεν πρόκειται να ξεπεράσουμε το 1.86, αλλά παράλληλα η τιμή μπορεί να κατέβει πολύ περισσότερο, αυτό ίσως δείχνει ότι υπάρχει μεγαλύτερο ρίσκο στο να επενδύσουμε σε αυτή τη μετοχή.

Αντίστοιχη μεθοδολογία μπορεί να εφαρμοστεί και στον όγκο της μετοχής. Λαμβάνοντας από το ίδιο σύνολο δεδομένων τις σχετικές τιμές για τον όγκο, μπορούμε λ.χ. να εφαρμόσουμε τη μέθοδο Winter για να πραγματοποιήσουμε προβλέψεις. Αξιοποιώντας την εντολή forecast.ets στο Excel, προκύπτει η ακόλουθη γραφική παράσταση, η οποία συμπεριλαμβάνει την απεικόνιση των 95% διαστημάτων εμπιστοσύνης γύρω από τις προβλέψεις:

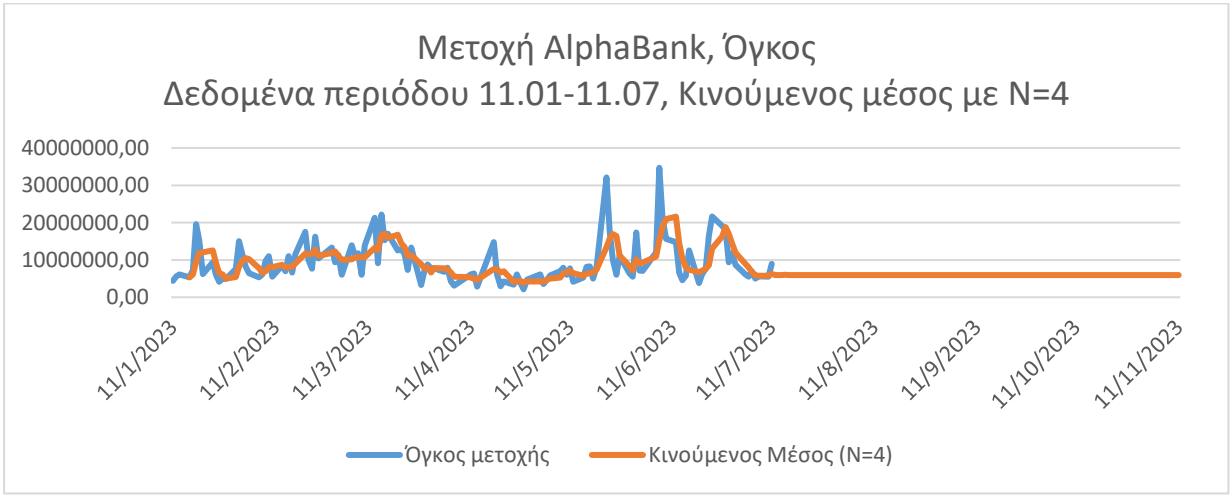


Γράφημα 15: Μέθοδος Winter για τον όγκο της μετοχής της Alpha Bank (11/1-11/4) και 95%-διαστήματα εμπιστοσύνης

Τα δεδομένα που χρησιμοποιούνται είναι για την περίοδο 11.01-11.04 και οι προβλέψεις πραγματοποιούνται μέχρι και την 11.10.23. Βάσει της μεθόδου πρόβλεψης που χρησιμοποιούμε, παρατηρούμε ότι οι πρόβλεψη τείνει να σταθεροποιηθεί, ενώ τα άκρα των αντίστοιχων διαστημάτων εμπιστοσύνης διευρύνονται ολοένα και περισσότερο. Αυτό, φυσικά, είναι αναμενόμενο, αφού όσο πιο πολύ «προχωράμε χρονικά» στο μέλλον, τόσο πιο αβέβαιες γίνονται οι προβλέψεις μας. Όσον αφορά τα μέτρα των σφαλμάτων που έχουμε διαθέσιμα, προκύπτουν τα εξής:

MAE=	2838923
RMSE=	3544339

Ομοίως, παίρνουμε την ακόλουθη εικόνα αν εφαρμόσουμε λ.χ. τη μέθοδο του κινούμενου μέσου με N=4:



Γράφημα 16: Όγκος μετοχής Alpha Bank, προβλέψεις με κινούμενο μέσο (N=4)

Με τα μέτρα σφάλματος να προκύπτουν:

MAE=	3592986.27
RMSE=	4282416.624

Από τις μεθόδους που μόλις εφαρμόσαμε, η μέθοδος Winter να είναι -σχετικά- πιο αποδοτική. Ωστόσο, και οι δυο δεν μας εξασφαλίζουν μεγάλη βεβαιότητα για τις προβλέψεις μας. Σε αντίθεση με την τιμή κλεισίματος, τα δεδομένα του όγκου της μετοχής παρουσιάζουν πολύ μεγαλύτερη μεταβλητότητα, με αποτέλεσμα να μην είναι «ασφαλές» να κάνουμε προβλέψεις για πολύ «μακρινές» χρονικές στιγμές. Αυτό είναι επίσης εμφανές από τα άκρα των διαστημάτων εμπιστοσύνης στις προβλέψεις, κατά την εφαρμογή της μεθόδου Winter: όσο απομακρυνόμαστε από την έναρξη των προβλέψεων, τα άκρα διευρύνονται ολοένα και περισσότερο, με αποτέλεσμα να μην είμαστε σε θέση να εκτελέσουμε προβλέψεις με μεγάλη ακρίβεια για το ποιος πρόκειται να είναι ο όγκος.

ΠΑΡΑΡΤΗΜΑ: ΠΙΝΑΚΕΣ ΔΕΔΟΜΕΝΩΝ

Σε αυτήν την ενότητα συμπεριλαμβάνουμε τους εκτενείς πίνακες δεδομένων που χρησιμοποιούμε στο κυρίως κείμενο της εργασίας.

Πίνακας Π1: Δεδομένα παραδείγματος για την ad hoc μέθοδο

Μήνας	Μέρα Μ	Μέρα W	Πελάτης	ΑΗ	Πρόβλεψη
1	1	1			516.35261
1	2	2	431	Y	414.03014
1	3	3	271		431.41007
1	4	4	362		432.06143
1	5	5	696		581.0192
1	6	6	315		368.09113
1	7	7			516.35261
1	8	1	330		414.03014
1	9	2	352		431.41007
1	10	3	606		432.06143
1	11	4	550		581.0192
1	12	5	626		368.09113
1	13	6	392		516.35261
1	14	7			414.03014
1	15	1	540		431.41007
1	16	2	474		432.06143
1	17	3	457		581.0192
1	18	4	401		368.09113
1	19	5	691		516.35261
1	20	6	388		414.03014
1	21	7			431.41007
1	22	1	533		432.06143
1	23	2	384		581.0192
1	24	3	360		368.09113
1	25	4	515		516.35261
1	26	5	325		414.03014
1	27	6	412		431.41007
1	28	7			432.06143
1	29	1	592		581.0192
1	30	2	366		368.09113
1	31	3	512		516.35261
2	1	4	476		407.14276
2	2	5	531		424.23358
2	3	6	303		424.87411
2	4	7			571.35396
2	5	1	474		361.96795

2	6	2	255		507.7631
2	7	3	282		407.14276
2	8	4	321		424.23358
2	9	5	416		424.87411
2	10	6	257		571.35396
2	11	7			361.96795
2	12	1	638		507.7631
2	13	2	506		407.14276
2	14	3	420		424.23358
2	15	4	459		424.87411
2	16	5	515		571.35396
2	17	6	501		361.96795
2	18	7			507.7631
2	19	1	556		407.14276
2	20	2	510		424.23358
2	21	3	436		424.87411
2	22	4	512		571.35396
2	23	5	547		361.96795
2	24	6	319		507.7631
2	25	7			407.14276
2	26	1	637		424.23358
2	27	2	474		424.87411
2	28	3	487		571.35396
2	29	4	402		361.96795
3	1	5	778		509.79009
3	2	6	374		408.76808
3	3	7			425.92712
3	4	1	544		426.57021
3	5	2	485		573.63481
3	6	3	361		363.41292
3	7	4	315		509.79009
3	8	5	423		408.76808
3	9	6	357		425.92712
3	10	7			426.57021
3	11	1	649		573.63481
3	12	2	351		363.41292
3	13	3	405		509.79009
3	14	4	404		408.76808
3	15	5	483		425.92712
3	16	6	411		426.57021
3	17	7			573.63481
3	18	1	309		363.41292
3	19	2	453		509.79009
3	20	3	515		408.76808

3	21	4	380		425.92712
3	22	5	426		426.57021
3	23	6	427		573.63481
3	24	7			363.41292
3	25	1	489		509.79009
3	26	2	341		408.76808
3	27	3	471		425.92712
3	28	4	517		426.57021
3	29	5	647		573.63481
3	30	6	415		363.41292
3	31	7			509.79009
4	1	1	363		389.84557
4	2	2	337		406.21029
4	3	3	314		406.82361
4	4	4	465		547.08036
4	5	5	584		346.58997
4	6	6	313		486.19111
4	7	7			389.84557
4	8	1	376		406.21029
4	9	2	292		406.82361
4	10	3	484		547.08036
4	11	4	227		346.58997
4	12	5	496		486.19111
4	13	6	395		389.84557
4	14	7			406.21029
4	15	1	625		406.82361
4	16	2	430		547.08036
4	17	3	454		346.58997
4	18	4	372		486.19111
4	19	5	455		389.84557
4	20	6	253		406.21029
4	21	7			406.82361
4	22	1	432		547.08036
4	23	2	469		346.58997
4	24	3	392		486.19111
4	25	4	467		389.84557
4	26	5	684		406.21029
4	27	6	349		406.82361
4	28	7			547.08036
4	29	1	750		346.58997
4	30	2	409		486.19111
5	1	3	348		357.88987
5	2	4	230		372.91318
5	3	5	630		373.47622

5	4	6	358		502.2361
5	5	7			318.17994
5	6	1	269		446.33796
5	7	2	107		357.88987
5	8	3	360		372.91318
5	9	4	208		373.47622
5	10	5	547		502.2361
5	11	6	325		318.17994
5	12	7			446.33796
5	13	1	473		357.88987
5	14	2	337		372.91318
5	15	3	317		373.47622
5	16	4	341		502.2361
5	17	5	338		318.17994
5	18	6	369		446.33796
5	19	7			357.88987
5	20	1	618		372.91318
5	21	2	458		373.47622
5	22	3	457		502.2361
5	23	4	572		318.17994
5	24	5	668		446.33796
5	25	6	318		357.88987
5	26	7			372.91318
5	27	1	300		373.47622
5	28	2	469		502.2361
5	29	3	434		318.17994
5	30	4	419		446.33796
5	31	5			357.88987
6	1	6	432	Y	412.82905
6	2	7			413.45235
6	3	1	463		555.99444
6	4	2	457		352.23728
6	5	3	273		494.11307
6	6	4	327		396.19767
6	7	5	554		412.82905
6	8	6	256		413.45235
6	9	7			555.99444
6	10	1	465		352.23728
6	11	2	479		494.11307
6	12	3	437		396.19767
6	13	4	585		412.82905
6	14	5	616		413.45235
6	15	6	318		555.99444
6	16	7			352.23728

6	17	1	724		494.11307
6	18	2	390		396.19767
6	19	3	550		412.82905
6	20	4	266		413.45235
6	21	5	410		555.99444
6	22	6	303		352.23728
6	23	7			494.11307
6	24	1	514		396.19767
6	25	2	353		412.82905
6	26	3	397		413.45235
6	27	4	539		555.99444
6	28	5	411		352.23728
6	29	6	413		494.11307
6	30	7			396.19767
7	1	1	583		396.51528
7	2	2	477		397.11395
7	3	3	410		534.0232
7	4	4			338.31791
7	5	5	615	Y	474.5872
7	6	6	288		380.54113
7	7	7			396.51528
7	8	1	478		397.11395
7	9	2	298		534.0232
7	10	3	253		338.31791
7	11	4	366		474.5872
7	12	5	410		380.54113
7	13	6	270		396.51528
7	14	7			397.11395
7	15	1	541		534.0232
7	16	2	331		338.31791
7	17	3	318		474.5872
7	18	4	441		380.54113
7	19	5	651		396.51528
7	20	6	300		397.11395
7	21	7			534.0232
7	22	1	308		338.31791
7	23	2	401		474.5872
7	24	3	390		380.54113
7	25	4	391		396.51528
7	26	5	619		397.11395
7	27	6	391		534.0232
7	28	7			338.31791
7	29	1	413		474.5872
7	30	2	474		380.54113

7	31	3	503		396.51528
8	1	4	267		417.88014
8	2	5	619		561.94875
8	3	6	370		356.00949
8	4	7			499.40467
8	5	1	406		400.44067
8	6	2	432		417.25016
8	7	3	333		417.88014
8	8	4	327		561.94875
8	9	5	647		356.00949
8	10	6	407		499.40467
8	11	7			400.44067
8	12	1	396		417.25016
8	13	2	664		417.88014
8	14	3	508		561.94875
8	15	4	519		356.00949
8	16	5	555		499.40467
8	17	6	365		400.44067
8	18	7			417.25016
8	19	1	492		417.88014
8	20	2	420		561.94875
8	21	3	360		356.00949
8	22	4	469		499.40467
8	23	5	488		400.44067
8	24	6	326		417.25016
8	25	7			417.88014
8	26	1	465		561.94875
8	27	2	384		356.00949
8	28	3	280		499.40467
8	29	4	292		400.44067
8	30	5	649		417.25016
8	31	6	493		417.8014
9	1	7			549.17505
9	2	1			347.91701
9	3	2	459	Y	488.05267
9	4	3	353		391.33823
9	5	4	287		407.76561
9	6	5	471		408.38127
9	7	6	266		549.17505
9	8	7			347.91701
9	9	1	505		488.05267
9	10	2	528		391.33823
9	11	3	342		407.76561
9	12	4	551		408.38127

9	13	5	525		549.17505
9	14	6	304		347.91701
9	15	7			488.05267
9	16	1	479		391.33823
9	17	2	258		407.76561
9	18	3	263		408.38127
9	19	4	450		549.17505
9	20	5	540		347.91701
9	21	6	297		488.05267
9	22	7			391.33823
9	23	1	399		407.76561
9	24	2	264		408.38127
9	25	3	479		549.17505
9	26	4	459		347.91701
9	27	5	915		488.05267
9	28	6	247		391.33823
9	29	7			407.76561
9	30	1	725		408.38127
10	1	2	197		547.25605
10	2	3	326		346.70128
10	3	4	374		486.34725
10	4	5	477		389.97076
10	5	6	367		406.34074
10	6	7			406.95425
10	7	1	317		547.25605
10	8	2	205		346.70128
10	9	3	519		486.34725
10	10	4	483		389.97076
10	11	5	489		406.34074
10	12	6	345		406.95425
10	13	7			547.25605
10	14	1	660		346.70128
10	15	2	262		486.34725
10	16	3	395		389.97076
10	17	4	522		406.34074
10	18	5	582		406.95425
10	19	6	335		547.25605
10	20	7			346.70128
10	21	1	503		486.34725
10	22	2	396		389.97076
10	23	3	548		406.34074
10	24	4	471		406.95425
10	25	5	528		547.25605
10	26	6	344		346.70128

10	27	7			486.34725
10	28	1	419		389.97076
10	29	2	429		406.34074
10	30	3	609		406.95425
10	31	4	519		547.25605
11	1	5	674		377.36947
11	2	6	352		529.36812
11	3	7			424.46644
11	4	1	360		442.28447
11	5	2	500		442.95225
11	6	3	339		595.66473
11	7	4	326		377.36947
11	8	5	459		529.36812
11	9	6	255		424.46644
11	10	7			442.28447
11	11	1	432		442.95225
11	12	2	527		595.66473
11	13	3	394		377.36947
11	14	4	424		529.36812
11	15	5	388		424.46644
11	16	6	356		442.28447
11	17	7			442.95225
11	18	1	635		595.66473
11	19	2	309		377.36947
11	20	3	613		529.36812
11	21	4	580		424.46644
11	22	5	627		442.28447
11	23	6	514		442.95225
11	24	7			595.66473
11	25	1	686		377.36947
11	26	2	452		529.36812
11	27	3	384		424.46644
11	28	4			442.28447
11	29	5	701	Y	442.95225
11	30	6	425		595.66473
12	1	7			366.81314
12	2	1	291		514.55987
12	3	2	407		412.59265
12	4	3	458		429.91225
12	5	4	243		430.56135
12	6	5	449		579.00194
12	7	6	315		366.81314
12	8	7			514.55987
12	9	1	633		412.59265

12	10	2	429		429.91225
12	11	3	375		430.56135
12	12	4	540		579.00194
12	13	5	615		366.81314
12	14	6	455		514.55987
12	15	7			412.59265
12	16	1	385		429.91225
12	17	2	472		430.56135
12	18	3	576		579.00194
12	19	4	321		366.81314
12	20	5	679		514.55987
12	22	7			412.59265
12	23	1	407		429.91225
12	24	2	328		430.56135
12	25	3			579.00194
12	26	4	491	Y	366.81314
12	27	5	586		514.55987
12	28	6	367		412.59265
12	29	7			429.91225
12	30	1	707		430.56135
12	31	2	400		579.00194

Πίνακας Π1: Δεδομένα παραδείγματος για την ad hoc μέθοδο

Πίνακας Π2: Δεδομένα Ε.Ο.Φ., Πωλήσεις φαρμάκων ανά μήνα (σε αξίες) τα έτη 2019-2020

Μήνας	Φαρμακεία, φαρμακαποθήκες (λιανική τιμή) 2019-2020
1	344089652.56
2	345865984.94
3	348846112.29
4	373890297.51
5	358110749.34
6	341219966.94
7	477925386.16
8	239388639.72
9	381696392.68
10	432943230.27
11	373386567.53
12	375676848.64
13	375386551.06

14	371436590.19
15	459088624.13
16	330440325.98
17	324490229.85
18	364730783.19
19	446947280.25
20	240797550.01
21	429619680.44
22	410026600.78
23	399281965.26
24	405140820.85

Πίνακας Π2: Δεδομένα Ε.Ο.Φ., Πωλήσεις φαρμάκων ανά μήνα (σε αξίες) τα έτη 2019-2020

Πίνακας Π3: Δεδομένα απόκρισης ασθενών διαφόρων ηλικιών σε θεραπεία κατά της λευχαιμίας

Απόκριση σε θεραπεία	Ηλικία
1	20
1	25
1	26
1	26
1	27
0	27
1	28
1	28
1	31
0	33
1	33
1	33
0	34
1	36
0	37
1	40
1	40
0	43
1	45
1	45
0	45
0	45
0	47
0	48
1	50
1	50
0	51
0	52

1	53
0	53
1	56
0	57
0	59
0	59
0	60
0	60
0	61
0	61
0	61
1	62
0	63
0	65
0	71
1	71
0	73
0	73
1	74
0	74
1	75
1	77
0	80

Πίνακας Π3: Δεδομένα απόκρισης ασθενών διαφόρων ηλικιών σε θεραπεία κατά της λευχαιμίας

Πίνακας Π4: Δεδομένα της μετοχής της Alpha Bank για την περίοδο 11.01.23-11.07.23

Ημερομηνία	Άνοιγμα	Υψηλή	Χαμηλή	Κλείσιμο	Όγκος
11/1/2023	1.14	1.15	1.121	1.128	4398051
12/1/2023	1.14	1.135	1.115	1.13	5619778
13/1/2023	1.13	1.1615	1.126	1.142	6093604
16/1/2023	1.15	1.17	1.146	1.163	5330334
17/1/2023	1.16	1.1855	1.157	1.186	7298724
18/1/2023	1.19	1.196	1.174	1.19	2E+07
19/1/2023	1.18	1.18	1.136	1.15	1.5E+07
20/1/2023	1.15	1.164	1.148	1.156	6162564
23/1/2023	1.17	1.21	1.153	1.196	9387057
24/1/2023	1.2	1.205	1.162	1.18	6026250
25/1/2023	1.17	1.1845	1.152	1.157	4105985
26/1/2023	1.17	1.1945	1.162	1.188	5031047
27/1/2023	1.19	1.2095	1.188	1.2	4947939
30/1/2023	1.21	1.2145	1.185	1.215	7525116

31/1/2023	1.21	1.2435	1.206	1.244	1.5E+07
1/2/2023	1.24	1.2845	1.243	1.28	1.1E+07
2/2/2023	1.29	1.3	1.26	1.26	8185413
3/2/2023	1.25	1.265	1.244	1.255	6493589
6/2/2023	1.25	1.27	1.237	1.25	5349783
7/2/2023	1.26	1.274	1.242	1.242	6120750
8/2/2023	1.26	1.294	1.255	1.294	9753566
9/2/2023	1.3	1.3175	1.274	1.3	1.1E+07
10/2/2023	1.3	1.308	1.268	1.285	5531336
13/2/2023	1.29	1.3485	1.285	1.349	8541419
14/2/2023	1.35	1.352	1.333	1.337	6937676
15/2/2023	1.34	1.377	1.333	1.349	1.1E+07
16/2/2023	1.36	1.372	1.335	1.354	6642624
17/2/2023	1.34	1.388	1.321	1.388	1.1E+07
20/2/2023	1.4	1.47	1.396	1.47	1.8E+07
21/2/2023	1.47	1.5	1.425	1.464	9982254
22/2/2023	1.45	1.47	1.4	1.4	7627576
23/2/2023	1.41	1.452	1.404	1.45	1.6E+07
24/2/2023	1.45	1.464	1.427	1.43	1E+07
28/2/2023	1.43	1.494	1.428	1.48	1.3E+07
1/3/2023	1.47	1.513	1.458	1.469	9383224
2/3/2023	1.43	1.441	1.401	1.43	1.1E+07
3/3/2023	1.44	1.454	1.378	1.44	5991061
6/3/2023	1.41	1.4105	1.331	1.355	1.4E+07
7/3/2023	1.33	1.364	1.31	1.32	1.1E+07
8/3/2023	1.3	1.305	1.269	1.279	1.2E+07
9/3/2023	1.28	1.303	1.272	1.279	6019006
10/3/2023	1.24	1.2495	1.2	1.2	1.4E+07
13/3/2023	1.17	1.21	1.14	1.2	2.1E+07
14/3/2023	1.21	1.27	1.165	1.261	9101177
15/3/2023	1.27	1.27	1.11	1.13	2.2E+07
16/3/2023	1.18	1.1845	1.101	1.122	1.5E+07
17/3/2023	1.15	1.16	1.061	1.065	1.7E+07
20/3/2023	1.02	1.1025	1.003	1.103	1.3E+07
21/3/2023	1.14	1.17	1.13	1.161	1.3E+07
22/3/2023	1.18	1.1885	1.158	1.16	1.2E+07
23/3/2023	1.16	1.1785	1.15	1.16	7328276
24/3/2023	1.15	1.15	1.089	1.09	1.3E+07
27/3/2023	1.12	1.13	1.1	1.12	3252241
28/3/2023	1.12	1.136	1.08	1.083	6929077
29/3/2023	1.08	1.1	1.062	1.1	8742614
30/3/2023	1.12	1.134	1.12	1.133	7742874

31/3/2023	1.14	1.145	1.125	1.125	7734105
3/4/2023	1.14	1.1795	1.14	1.176	6802650
4/4/2023	1.18	1.2065	1.161	1.17	7923794
5/4/2023	1.16	1.1675	1.142	1.15	4192223
6/4/2023	1.15	1.173	1.147	1.151	3113685
11/4/2023	1.16	1.175	1.162	1.17	6111774
12/4/2023	1.19	1.19	1.173	1.18	6362996
13/4/2023	1.19	1.188	1.174	1.18	2888765
18/4/2023	1.2	1.2265	1.186	1.22	1,5E+07
19/4/2023	1.23	1.2305	1.179	1.192	6070131
20/4/2023	1.18	1.213	1.177	1.192	2930753
21/4/2023	1.2	1.2175	1.187	1.208	4264079
24/4/2023	1.2	1.215	1.166	1.186	3393185
25/4/2023	1.18	1.1785	1.142	1.144	6069072
26/4/2023	1.13	1.177	1.122	1.177	3870169
27/4/2023	1.18	1.187	1.162	1.162	2121345
28/4/2023	1.18	1.178	1.135	1.135	4750066
2/5/2023	1.14	1.1445	1.122	1.139	6081279
3/5/2023	1.13	1.177	1.133	1.165	3540337
4/5/2023	1.17	1.1805	1.157	1.162	4636385
5/5/2023	1.16	1.187	1.162	1.165	5851372
8/5/2023	1.18	1.212	1.175	1.2	6975510
9/5/2023	1.22	1.2215	1.191	1.195	7877990
10/5/2023	1.2	1.213	1.186	1.213	6005237
11/5/2023	1.22	1.219	1.178	1.187	7755657
12/5/2023	1.19	1.2165	1.17	1.195	4131777
15/5/2023	1.2	1.217	1.185	1.185	5249789
16/5/2023	1.19	1.215	1.176	1.215	8131578
17/5/2023	1.21	1.219	1.193	1.193	8252957
18/5/2023	1.2	1.209	1.184	1.184	5033050
19/5/2023	1.19	1.2095	1.185	1.2	8105630
22/5/2023	1.35	1.4055	1.32	1.37	3.2E+07
23/5/2023	1.39	1.44	1.371	1.385	1.8E+07
24/5/2023	1.36	1.37	1.349	1.349	9980015
25/5/2023	1.34	1.377	1.344	1.36	5984303
26/5/2023	1.38	1.449	1.371	1.445	1.1E+07
29/5/2023	1.46	1.478	1.429	1.44	6402510
30/5/2023	1.44	1.4495	1.406	1.406	5518121
31/5/2023	1.41	1.4055	1.386	1.386	1.7E+07
1/6/2023	1.41	1.421	1.373	1.373	7119378
2/6/2023	1.39	1.416	1.372	1.414	7081176
6/6/2023	1.43	1.49	1.427	1.478	1.2E+07

7/6/2023	1.5	1.57	1.5	1.536	3.5E+07
8/6/2023	1.54	1.547	1.509	1.54	2.1E+07
9/6/2023	1.55	1.5845	1.542	1.576	1.6E+07
12/6/2023	1.59	1.6085	1.573	1.573	1.5E+07
13/6/2023	1.59	1.59	1.548	1.564	6606539
14/6/2023	1.57	1.5785	1.558	1.558	4605163
15/6/2023	1.56	1.559	1.534	1.538	5792138
16/6/2023	1.54	1.557	1.534	1.54	1,3E+07
19/6/2023	1.56	1.5585	1.515	1.52	3780035
20/6/2023	1.52	1.522	1.494	1.519	6230742
21/6/2023	1.53	1.5525	1.512	1.537	7647312
22/6/2023	1.53	1.53	1.496	1.497	1.7E+07
23/6/2023	1.51	1.544	1.487	1.525	2.2E+07
26/6/2023	1.53	1.5385	1.443	1.458	1.9E+07
27/6/2023	1.46	1.488	1.428	1.47	1.8E+07
28/6/2023	1.49	1.498	1.47	1.47	9373703
29/6/2023	1.49	1.5355	1.487	1.52	1.2E+07
30/6/2023	1.54	1.554	1.5	1.5	8658252
3/7/2023	1.53	1.531	1.503	1.517	6100317
4/7/2023	1.52	1.529	1.505	1.512	5488898
5/7/2023	1.51	1.526	1.495	1.507	6980181
6/7/2023	1.5	1.5195	1.483	1.501	4991152
7/7/2023	1.5	1.529	1.497	1.511	5596859
10/7/2023	1.52	1.545	1.51	1.53	5467191
11/7/2023	1.54	1.5935	1.54	1.58	8903923

Πίνακας Π4: Δεδομένα της μετοχής της Alpha Bank για την περίοδο 11.01.23-11.07.23

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Κολέτσος, Ι. και Στογιάννης, Δ. (2016) Εισαγωγή στην Επιχειρησιακή Έρευνα, 2η Έκδοση, Αθήνα

[2] Hillier, F.S. and Lieberman, G.J. (2010) Introduction to Operations Research. 9th Edition, McGraw-Hill, New York.

[3] Hamdy, T. (2016) Operations Research: An Introduction. 8th Edition, Pearson Publisher, London.

[4] Winston, W.L and Goldberg, J.B. (2004) *Operations Research: Applications and Algorithms*, 4th edition, Belmont CA: Thomson/Brooks/Cole.

[5] Καρώνη, Χ. και Οικονόμου, Π. (2017) Στατιστικά Μοντέλα Παλινδρόμησης: Με χρήση MINITAB και R, 2η έκδοση, Εκδόσεις ΣΥΜΕΩΝ