ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ
ΣΠΟΥΔΩΝ «ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΤΥΠΟΠΟΙΗΣΗ ΣΕ
ΣΥΓΧΡΟΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ»

Μεταπτυχιακή Διπλωματική Εργασία

Predictive analytics models to evaluate the impact of Media on Business
Προγνωστικά μοντέλα για την αξιολόγηση του αντίκτυπου των Media στις επιχειρήσεις

Μετσίνη Ηλέκτρα Φωτεινή

Επιβλέπων Καθηγητής: Κολέτσος Ιωάννης Αναπληρωτής Καθηγητής Ε.Μ.Π.

Μέλη εξεταστικής επιτροπής

Καρώνη Χρυσηίς Καθηγήτρια Ε.Μ.Π.	Κολέτσος Ιωάννης Αναπληρωτής Καθηγητής Ε.Μ.Π.	Στεφανέας Πέτρος Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα , Οκτώβριος 2019

Αφιερώνεται στην κόρη μου, στον άντρα μου,

στη Νατάσσα και στην Αγγέλω μου.

Περίληψη

Ο στόχος της παρούσας διπλωματικής εργασίας είναι να περιγράψει το θεωρητικό και μαθηματικό πλαίσιο της τεχνικής Marketing Mixing Modeling η οποία αποτελείται από μία στατιστική ανάλυση των πωλήσεων και των στοιχείων μάρκετινγκ χρησιμοποιώντας αναδρομικά ιστορικά δεδομένα. Αυτά οδηγούν με τη σειρά τους στην εκτίμηση των επιπτώσεων των διαφόρων τακτικών μάρκετινγκ στις πωλήσεις, με απώτερο και αναγκαίο στόχο την πρόβλεψη των επιπτώσεων, αυτών των μελλοντικών τακτικών, καθώς και τις τάσεις και τα πρότυπα συμπεριφοράς που μπορούν να προκύψουν. Αρχικά στο Κεφάλαιο 1, αναλύονται και αναφέρονται οι Business και Media key performance indicators (KPIs), δηλαδή οι κυριότερες επιχειρησιακές μετρικές που επηρεάζουν την επιχειρηματική απόδοση καθώς και οι αντίστοιχες μετρικές για τα Media, μαζί με μία σύντομη περιγραφή των Business Analytics. Στη συνέχεια στο Κεφάλαιο 2 αναλύεται το πώς μπορούν να συνυπάρξουν και να αλληλοεπιδράσουν μεταξύ τους πολλές μεταβλητές παράλληλα, βοηθώντας συνολικά στη διαμόρφωση ενός τελικού αποτελέσματος σχετικά με τις πωλήσεις μιας μάρκας. Δίνεται η μαθηματική προσέγγιση και τα βήματα για την ανάπτυξη του κατάλληλου μοντέλου ενώ επεξηγείται η αξία της ανάλυσης παλινδρόμησης ως το βασικότερο εργαλείο της Οικονομετρίας.

Στα κεφάλαια που ακολουθούν παρουσιάζεται εκτενώς το μαθηματικό σκέλος της Ανάλυσης Παλινδρόμησης και δίνεται έμφαση στην Απλή και Πολλαπλή Γραμμική Παλινδρόμηση, την Poisson αλλά και στους μετασχηματισμούς γραμμικοποίησης που αφορούν τις περιπτώσεις μη γραμμικότητας, καθώς και στις μεθόδους εκτίμησης των παραμέτρων όπως OLS, ML και Poisson . Ιδιαίτερα στο Κεφάλαιο 4 αναλύονται διεξοδικά οι μέθοδοι της άριστης επιλογής των ανεξάρτητων μεταβλητών και τα διαγνωστικά κριτήρια της εγκυρότητας της Πολλαπλής Παλινδρόμησης, με στόχο ο κάθε αναγνώστης να έχει τελικώς στη διάθεσή του, όλα τα αναγκαία εργαλεία για ένα επιτυχημένο case study. Στο παράρτημα παρατίθενται όλοι οι πίνακες των κατανομών που είναι απαραίτητοι για τη διαμόρφωση των τελικών οικονομετρικών μοντέλων.

Λέξεις και φράσεις κλειδιά: Key performance indicators, marketing mixing modelling, πολυσυγγραμμικότητα , Πολλαπλή Παλινδρόμηση, προγνωστικά μοντέλα ανάλυσης

Abstract

The aim of this thesis is to describe the theoretical and mathematical framework of Marketing Mixing Modeling technique consisting of a statistical analysis of sales and marketing data using retrospective historical data. These in turn lead to an assessment of the impact of the various marketing tactics on sales, with the ultimate and necessary aim of predicting the impact of these future tactics, as well as the trends and patterns of behavior that may arise. First in Chapter 1, Business and Media key performance indicators (KPIs), the main business performance metrics that affect business performance, and the corresponding Media metrics, along with a brief description of Business Analytics, are analyzed and reported. Chapter 2 then discusses how many variables can co-exist and interact with each other, helping to formulate a final result for a brand's sales. The mathematical approach and steps for developing the appropriate model are given while explaining the value of regression analysis as the key tool of Econometrics.

The following chapters give an extensive overview of the mathematical part of Regression Analysis and focus on Simple and Multiple Linear Regression, Poisson but also on linearization transformations for nonlinearity cases, as well as methods for estimating parameters such as OLS, ML and Poisson. Particularly in Chapter 4, we thoroughly analyze the methods of optimal selection of independent variables and the diagnostic criteria for validation of Regression, so that each reader has at his disposal all the tools necessary for a successful case study. In the appendix all the distribution tables necessary for the development of the final econometric models are included.

Keywords: Key performance indicators, marketing mixing modeling, multicollinearity, multiple regression, predictive analytics models

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά όλους τους καθηγητές και τις καθηγήτριες του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών του ΕΜΠ για τις γνώσεις που μου προσέφεραν. Ιδιαίτερα ευχαριστώ τον επιβλέποντα καθηγητή κ. Ιωάννη Κολέτσο για την εμπιστοσύνη, την υπομονή και βοήθειά του στην διάρκεια της εκπόνησης της διπλωματικής μου εργασίας, καθώς και τα μέλη της τριμελούς επιτροπής για τον χρόνο που διέθεσαν στην αξιολόγηση της διπλωματικής μου εργασίας.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου προς την οικογένειά μου, για την αμέριστη συμπαράσταση και στήριξη που μου προσέφεραν καθ΄ όλη τη διάρκεια των σπουδών μου.

Αθήνα , Σεπτέμβριος 2019

Ηλέκτρα Φωτεινή Μετσίνη

Πρόλογος

Η παρούσα διπλωματική εργασία έχει στόχο να δώσει μια συνοπτική περιγραφή της σύνδεσης των Στατιστικών Μοντέλων Παλινδρόμησης και της ευρύτερης θεωρητικής περιοχής της ανάλυσης των δεδομένων. Αποτελεί την ολοκλήρωση των μεταπτυχιακών μου σπουδών στο Διατμηματικό Πρόγραμμα του Εθνικού Μετσόβιου Πολυτεχνείου «Μαθηματική Προτυποποίηση σε Σύγχρονες Τεχνολογίες και στην Οικονομία».

Θέλω να πιστεύω πως στις σελίδες που ακολουθούν παρέχεται μία καλή προσέγγιση και περιγραφή του πως το Marketing Mix Modelling, είναι η διαδικασία που πρακτικά παρέχει τη δυνατότητα στα οικονομετρικά μοντέλα να υπολογίζουν και να ποσοτικοποιούν την αποτελεσματικότητα συγκεκριμένων δραστηριοτήτων στον τομέα του marketing, κατά το επιχειρείν. Ειδικό ενδιαφέρον και ανάλυση παρουσιάζουν τα Media των οποίων ο αντίκτυπος στις επιχειρήσεις μπορεί να αξιολογηθεί. Η εκτίμηση και η περιγραφή της σχέσης εξάρτησης μεταξύ μεταβλητών αποτελεί βασικό στόχο σε πολλές επιστήμες και σε πολλές περιπτώσεις η σχέση δεν είναι συναρτησιακή, αλλά στοχαστική, με συνέπεια η χρήση στατιστικών μοντέλων να είναι απαραίτητη για την περιγραφή των φαινομένων αυτών. Βασικό εργαλείο ανάλυσης αποτελούν τα μοντέλα παλινδρόμησης (regression models), τα οποία συνιστούν και το αντικείμενο μελέτης αυτής της διπλωματικής εργασίας.
Τα στατιστικά μοντέλα γενικά μπορούν να περιγράψουν τις αγορές, να αναβαθμίζουν τη διαφήμιση (πιο στοχευμένη), να καθορίσουν τις τιμές και να ανταποκριθούν στις αλλαγές της καταναλωτικής ζήτησης. Επιπλέον είναι τα βασικά εργαλεία για τη στοιχειοθέτηση των Big Data που διαχειρίζονται οι εταιρείες.

Στις μέρες μας, στην εποχή των μεγάλων αλλαγών που οφείλονται στην οικονομική κρίση και στον έντονο ανταγωνισμό, είναι αναπόφευκτο τόσο για τις πολυεθνικές όσο και για τις εγχώριες επιχειρήσεις όλων των μεγεθών, να ανησυχούν διαρκώς για το αγοραίο τους μερίδιο ενώ παράλληλα να έχουν ως κύρια προτεραιότητά τους την αύξηση των κερδών και των πωλήσεών τους (Lee και Masao, 1988). Άλλες εμπορικές μετρήσεις που διαμορφώνουν ολόκληρη την εμπειρία των καταναλωτών (consumer experience) όπως η αναγνώριση της εμπορικής μάρκας, ο μισθός και η απόλαυση κατανάλωσης βρίσκονται ανελλιπώς υπό μελέτη. Ο αντίκτυπος της διαφήμισης στην οικοδόμηση της ευρύτερη αναγνωρισιμότητας και αξιοπιστίας της επωνυμίας μιας

εταιρείας, δηλαδή αυτό που ονομάζουμε «μάρκα» (brand) εξετάζεται από τους Makasi et al. (2014).

Έτσι, υπάρχει σε μόνιμη εξέλιξη μια επιχειρησιακή και εμπορική έρευνα εντοπισμού ευρύτερων μέσων διαχείρισης της μεταβλητότητας των βασικών δεικτών απόδοσης (key performance indicators (KPIs)) που συχνά καθορίζουν κρίσιμους ετήσιους στόχους της εταιρείας όπως η διείσδυση στην αγορά ή ο όγκος των μηνιαίων πωλήσεων. Επιπλέον, η έντονη διαφοροποίηση που παρουσιάζουν οι προβλέψεις των μετοχικών τιμών υπό διαφορετικές υποθέσεις σε σχέση με το ανταγωνιστικό προφίλ μιας εταιρείας, έχει μελετηθεί από τους Alsem et al. (1989). Τέλος, από τις αρχές της δεκαετίας του '20, οι εφαρμογές των στατιστικών προτύπων σε διάφορους τομείς της πραγματικής ζωής πολλαπλασιάζονται. Έκτοτε ο όρος Business Statistics που στα ελληνικά αποδίδεται ως Στατιστική των Επιχειρήσεων, εισήχθη στον εταιρικό κόσμο και ο ρόλος του στη διαχείριση των επιχειρησιακών δεδομένων έχει καταστεί κεντρικός, με απεριόριστες εφαρμογές.

«Το Μάρκετινγκ εξακολουθεί να είναι μία τέχνη και ο υπεύθυνος προώθησης, ως επικεφαλής, πρέπει να διευθετεί με δημιουργικό τρόπο όλες τις δραστηριότητες του στο χώρο αυτό, με στόχο κάθε φορά να αναβαθμίζει και να εξασφαλίζει τα συμφέροντα της εταιρείας του, βραχυπρόθεσμα αλλά και μακροπρόθεσμα.»

NEIL BORDEN

Περιεχόμενα

Κεφάλαιο 1 : Δομές Δεδομένων σε Επιχειρήσεις και Media

1.1 Σύντομη Εισαγωγική Ιστορική Αναδρομή

Η διαφήμιση θεωρείται ευρέως πως διαδραματίζει καταλυτικό ρόλο στην ενίσχυση των πωλήσεων και στην εμπορική ανάπτυξη, με ρόλο κλειδί στις στρατηγικές μάρκετινγκ. Για τον καθορισμό του βέλτιστου επιπέδου διαφημιστικής δαπάνης για ένα προϊόν, που παρέχεται συσκευασμένο στους καταναλωτές, χρησιμοποιείται μία πειραματική μέθοδος από τους Sunoo και Lin(2013). Στο Danaher et al. (1994) υποστηρίζεται ότι η διαφήμιση πρέπει να θεωρείται ως επένδυση αντί για δαπάνη και αναπτύσσεται μια απλή μέθοδος η οποία καθορίζει το βέλτιστο επίπεδο κόστους που διατίθεται σε διαφήμιση στα Media, το οποίο επιτρέπει στους υπεύθυνους σχεδιασμού να χρησιμοποιήσουν μια προσέγγιση μεγιστοποίησης της απόδοσης αυτής της επένδυσης (Return-on-Investment ROI). Στο Marnik et al (1995) απεικονίζεται το πως η διαφήμιση έχει μεγάλες επιπτώσεις στις πωλήσεις και ο τρόπος με τον οποίο η διαχείριση των μέσων μάρκετινγκ είναι σε θέση να επηρεάσουν τις μακροπρόθεσμες τάσεις στις πωλήσεις και τις άλλες μετρικές απόδοσης. Ωστόσο, από την αρχή της αξιοποίησης της διαφήμισης ως μια επένδυση (ως ένα μέσο αύξησης των πωλήσεων), ένα απλό αλλά συνάμα σύνθετο ως προς τη μαθηματική προτυποποίησή του ερώτημα, εξακολουθεί να παραμένει αναπάντητο : Υπάρχει κάποια μέθοδος ώστε με την εφαρμογή της να ποσοτικοποιηθεί (και κατά συνέπεια να βελτιστοποιηθεί) ο τρόπος με τον οποίο οι διαφημιστικές δαπάνες μπορούν να επηρεάσουν τις πωλήσεις;

Στη βιβλιογραφία υπάρχουν διάφορα άρθρα που ερευνούν προς αυτή την κατεύθυνση. Ο Bendixen (1993) παρουσιάζει τα αποτελέσματα της εφαρμογής ενός οικονομετρικού μοντέλου για την επίδραση της διαφήμισης σε συνάρτηση με την εμπιστοσύνη ως προς την κατανάλωση που φέρει μία εμπορική μάρκα, τα τρέχοντα απλά και σύνθετα αποτελέσματα και τα μεταφορικά αποτελέσματα (carryover effects) . Το μεταφορικό αποτέλεσμα της διαφήμισης δηλώνει τη χρονική υστέρηση μεταξύ των καταναλωτών που εκτίθενται στη διαφήμιση και της απόκρισης τους στην ίδια τη διαφήμιση. Όταν μια εταιρία εγκαινιάζει ένα νέο προϊόν ή μια υπηρεσία και το διαφημίζει , θα υπήρχε

μια ομάδα πελατών που θα ανταποκρίνονταν αμέσως, αλλά θα υπήρχε μια άλλη δεξαμενή δυνητικών πελατών που θα παρατηρούσαν μεν τη διαφήμιση, αλλά η

απόφαση αγοράς θα χρειαζόταν κάποιο χρόνο. Οι εφαρμογές αυτών των οικονομετρικών μοντέλων χρονοσειρών αποδεικνύουν ότι ο εγγενής χαρακτήρας των παραπάνω επιπτώσεων σχετίζεται με τις γνωστικές πτυχές της απόφασης αγοράς. Ο Büschken, J. (2007) εξετάζει το παραπάνω ερώτημα συσχετίζοντας δεδομένα που προκύπτουν από τη Γερμανική αυτοκινητοβιομηχανία. Οι Assmus et al.(1984) προσπάθησαν να αξιολογήσουν μέσω μιας μετα-ανάλυσης όλα τα αποτελέσματα που έχουν προκύψει από τα οικονομετρικά μοντέλα για τις βραχυπρόθεσμες και μακροπρόθεσμες επιπτώσεις των διαφημίσεων επί των πωλήσεων. Ο Clarke (1976) σχολιάζει εκτενώς την οικονομετρική βιβλιογραφία και προσδιορίζει τη διάρκεια της (αθροιστικής) επίδρασης της διαφήμισης, στις πωλήσεις.

Τον τελευταίο καιρό, δεδομένου ότι έχουμε καλύτερη πρόσβαση σε δεδομένα, μαζικά μεγαλύτερη υπολογιστική ισχύ ενώ ταυτόχρονα ασκείται όλο και μεγαλύτερη πίεση προς την ελαχιστοποίηση των εξόδων διαφήμισης, κάποιες προηγμένες τεχνικές μαθηματικής προτυποποίησης όπως τα marketing mix models (MMM) χαίρουν αποδοχής ως τα πλέον αξιόπιστα εργαλεία μάρκετινγκ που μπορούν να έχουν στη διάθεση τους όλες οι μεγάλες εμπορικές εταιρίες (Kucuk, 2016). Με βάση τις μελέτες του Neil H. Borden μας γνωστοποιείται εδώ και πολλά χρόνια το γεγονός ότι η χρήση μεθόδων μάρκετινγκ από τους κατασκευαστές προϊόντων, σε κάθε περίπτωση θα πρέπει να γίνεται μετά από μία ενδελεχή μελέτη, της οποίας η τελική ανάλυση να χρησιμεύει ως ένα στοιχείο-συστατικό του ευρύτερου προγράμματος μάρκετινγκ μιας εταιρείας. Είναι πλέον γεγονός, πώς τα πρακτορεία των ΜΜΕ εργάζονται εντατικά προς την κατεύθυνση της καλύτερης κατανόησης της συμβολής τους στα κέρδη των επιχειρήσεων και στην επίτευξη των ετησίων στόχων των πωλήσεων με μετρήσιμους και ποσοτικοποιημένους τρόπους, έτσι ώστε να μπορούν να αυξήσουν την αξιοπιστία τους και να εξασφαλίσουν μία καίρια θέση στο καθολικό επιχειρησιακό πλαίσιο ως στρατηγικοί εταίροι των μεγάλων εταιριών. Οι Hollis και Nigel (1994) χρησιμοποιούν τα μοντέλα απόκρισης των πωλήσεων και αναπτύσσουν σχέσεις-συνδέσμους μεταξύ της υψηλής αναγνωρισιμότητας του σχετικού λογοτύπου μιας εταιρείας (ad awareness) από τους καταναλωτές και των πωλήσεων για 70 διαφορετικές αμερικάνικες μάρκες. Για την ακρίβεια παρουσιάζουν τη μεγάλη σύνδεση των δύο παραπάνω και εκφράζουν τη διαφωνία τους σχετικά με το πόσο μπορούν να εγγυηθούν πια μία αύξηση στις πωλήσεις, βραχυπρόθεσμα αλλά και μακροπρόθεσμα, οι διαφημίσεις που απλώς προκαλούν μία έντονη εντύπωση. Τέλος,

οι Luo και Donthu (2005), υπολογίζουν τη δυνητική απώλεια επί των πωλήσεων εξαιτίας της ανεπαρκούς διάθεσης κεφαλαίων για τη διαφήμιση.

Σήμερα, οι έννοιες διαφήμιση και διάθεση προϊόντων παντός τύπου είναι άρρηκτα συνδεδεμένες. Με την ευρύτερα διαδεδομένη χρήση των τεχνολογικών μέσων, του ίντερνετ και των μέσων κοινωνικής δικτύωσης, η σύνδεσή τους παγιώνεται και δίνει μία νέα μορφή στην επιστήμη του εμπορίου και της διαφήμισης. Γεννάται συνεπώς η ανάγκη να αναλυθούν εις βάθος οι πληροφορίες που μπορούν να στοιχειοθετήσουν τις δομές όλων των δεδομένων, προς χρήση των επιχειρήσεων και των μέσων ενημέρωσης με στόχο να απαντηθούν τα ακόλουθα ερωτήματα που πρώτος, ο Neil Borden έθεσε:

- Ποια είναι η γενική στρατηγική διαφήμισης η οποία ακολουθείται και πως τροποποιείται αυτή ώστε κάτω από οποιεσδήποτε συνθήκες, ενός συγκεκριμένου οικονομικού πλαισίου, η εταιρία να έχει διαρκώς κέρδος;
- Ποιος είναι ο καταλληλότερος συνδυασμός διαδικασιών, τεχνικών και πολιτικής του μάρκετινγκ ώστε η υιοθέτησή τους να επιφέρει τα επιθυμητά αποτελέσματα στη ροή διακίνησης των προϊόντων και στη συμπεριφορά των καταναλωτών, με κόστη που επιτρέπουν το διαρκές όφελος;

1.2 Τύποι Διαθέσιμων Δεδομένων που οι Μεγάλες Εταιρείες Προτιμούν να Συλλέγουν

Οι έμποροι-επιχειρηματίες έχουν δραστηριοποιηθεί εδώ και πολλά χρόνια και έχουν καταβάλει πολλές προσπάθειες για να καθορίσουν την επιτυχία μιας μάρκας με έναν απλό, μετρήσιμο και αναμφισβήτητο τρόπο. Αυτό το έπραξαν, επειδή θα εκτιμούσαν την ύπαρξη εγγυημένων τύπων, κατευθυντήριων γραμμών και μεθοδολογιών που θα λειτουργούσαν ως μέτρα επιτυχίας. Πολλοί διαφορετικοί βασικοί δείκτες απόδοσης key performance indicators (KPIs) και πολλές διαφορετικές στρατηγικές, τακτικές και προσεγγίσεις έχουν εμπλακεί στη διαδικασία και έχουν αναγνωριστεί ότι παίζουν σημαντικό ρόλο στην επιτυχία μιας εταιρείας . Τέτοιες τακτικές μπορούν να είναι για παράδειγμα οι εκπτώσεις στις τιμές ή οι προωθήσεις προϊόντων προσφορών, με στόχο την οικοδόμηση ενός ευρύτερου δικτύου διανομής και τη διάθεση του προϊόντος σε μεγαλύτερο γεωγραφικό εύρος, ενώ ακόμη μπορεί να περιλαμβάνουν προγράμματα αφοσίωσης και διαδικασίες οικοδόμησης μια αναγνωρίσιμης εταιρικής εικόνας.

Ωστόσο, καθώς το πρόβλημα είναι αρκετά περίπλοκο και δεν μπορεί να λυθεί με ευθύ και άμεσο τρόπο, όλες αυτές οι τακτικές δεν μπορούν να λειτουργήσουν αυτόνομα και πρέπει να συνυπάρξουν και να συνδυαστούν για να εξασφαλίσουν ένα επιτυχημένο αποτέλεσμα.

Τα *προγράμματα αφοσίωσης*, για παράδειγμα, είναι δομημένες στρατηγικές μάρκετινγκ που έχουν σχεδιαστεί για να ενθαρρύνουν τους πελάτες να συνεχίσουν να αγοράζουν την εμπορική μάρκα που σχετίζεται με κάθε πρόγραμμα. Καλύπτουν τα περισσότερα είδη εμπορίου και διαθέτουν διάφορους τρόπους ανταμοιβής πιστών και κατ' επανάληψη πελατών. Η «θεωρία» μάρκετινγκ που βρίσκεται πίσω από τέτοια προγράμματα είναι ότι αν μία γνωστή μάρκα μπορεί να αναπτύξει μια στενότερη σχέση με τους καταναλωτές της, προσφέροντας σημεία και πόντους ανταμοιβής για την κάθε αγορά, τότε θα γίνουν πιο πιστοί στην αντίστοιχη εταιρική επωνυμία. Ωστόσο, αυτή η θεωρία φαίνεται ότι δεν λειτουργεί αποτελεσματικά και οι επενδύσεις σε τέτοια προγράμματα τελευταίως όλο και μειώνονται. Η παγίδα κρύβεται στο γεγονός ότι εκτός των τακτικών χρηστών, δεν μπορούν όλοι άμεσα να ενημερωθούν για το όφελος από τη χρήση των προνομίων που διατίθενται και να εγγραφθούν, ενώ όσοι καταναλώνουν σπανίως το αντίστοιχο προϊόν, είναι λιγότερο πιθανό να ασχοληθούν. Και φυσικά οι μη χρήστες αυτού, σπάνια το γνωρίζουν. Ως αποτέλεσμα, αν και τέτοια προγράμματα μπορούν να λειτουργήσουν ως φορείς διαρροής πληροφοριών για τους υπάρχοντες πελάτες σε σχέση με τους ανταγωνιστές, δεν είναι ικανά να οδηγήσουν στην πραγματική ανάπτυξη.

Είναι γεγονός πως η επιτυχία μιας μάρκας μπορεί εύκολα και απλά να ποσοτικοποιηθεί μέσω του αγοραίου μεριδίου της, το οποίο υπολογίζεται μέσω του όγκου των πωλήσεων του προϊόντος ή των υπηρεσιών που πωλούνται στην αγορά. Τότε το πρόβλημα έγκειται στο να προσδιοριστούν ποιες είναι οι μετρήσεις που επηρεάζουν πραγματικά τις πωλήσεις και ποιο είναι το καλύτερο μίγμα αυτών προκειμένου να διασφαλιστεί η επιτυχία. Μία από τις βασικές αρχές στο μάρκετινγκ είναι ότι η αγορά

ενός προϊόντος οφείλει να είναι πρακτικά εύκολη. Αυτό προϋποθέτει μία φυσική και νοητική διαθεσιμότητα του προϊόντος ενώ η πραγμάτωση της κάθε περίπτωσης συνδέεται με διαφορετικές προϋποθέσεις λειτουργίας και είδους των βασικών δεικτών απόδοσης (KPIs).

Η *φυσική διαθεσιμότητα* περιλαμβάνει όλα τα μέσα και τους τρόπους που καθιστούν μια μάρκα εύκολη στην αγορά, σε πραγματικό χρόνο, όταν ο καταναλωτής είναι

έτοιμος να κάνει την αγορά. Για παράδειγμα, πιθανή πώληση χάνεται εάν το προϊόν είναι εκτός αποθέματος. Η ευκολία στην αγορά είναι ένας άλλος παράγοντας. Η διεύρυνση της διανομής, η απόκτηση νέων καναλιών διανομής, η αύξηση του χώρου αποθήκευσης, η ποικιλία των προϊόντων, η συσκευασία και τα μεγέθη, αποτελούν βασικούς τρόπους για να προσδιοριστεί αυτό που τελικά ονομάζουμε φυσική διαθεσιμότητα της κάθε επωνυμίας.

Η *διανοητική διαθεσιμότητα*, από την άλλη πλευρά, έχει να κάνει με την πιθανότητα ένας καταναλωτής να παρατηρήσει, να αναγνωρίσει και / ή να σκεφτεί ένα εμπορικό σήμα μιας εταιρείας κατά τη στιγμή που πραγματοποιεί την αγορά. Δεδομένου ότι η διανοητική διαθεσιμότητα εξαρτάται από την ποιότητα και την ποσότητα των δομών μνήμης που σχετίζονται με το εμπορικό σήμα, παρουσιάζεται μια περιοχή που μπορεί να ενισχυθεί σημαντικά και να επηρεαστεί από τα μέσα ενημέρωσης και τη διαφήμιση.

Με την έκρηξη του Big Data, έχουμε φτάσει σε ένα σημείο αναφοράς όπου η διαθεσιμότητα των δεδομένων έχει υπερβεί κατά πολύ αυτό που πραγματικά απαιτείται να παρακολουθείται και να καταγράφεται. Ωστόσο, σε πολλές περιπτώσεις, οι διευθυντές και οι ενδιαφερόμενοι φορείς αισθάνονται αρκετά απροστάτευτοι και ανασφαλείς στις επιλογές τους λόγω έλλειψης χρήσιμων μετρήσεων που θα μπορούσαν να λειτουργήσουν ως ασφαλείς πυξίδες στο επιχειρηματικό παιχνίδι. Στο υπόλοιπο του παρόντος κεφαλαίου θα προσπαθήσουμε να παρακολουθήσουμε και να καθορίσουμε τους σημαντικότερους KPIs που υπάρχουν τόσο στον επιχειρηματικό κλάδο όσο και στα ΜΜΕ.

1.3 Οι Σημαντικές Μετρικές.

Οι μετρικές έχουν σημασία στην επιχειρηματική απόδοση. Ωστόσο, από μόνες τους δεν είναι ικανές να τη βελτιώσουν. Είναι η διαδικασία που ακολουθεί ως αποτέλεσμα της ερμηνείας των μετρήσεων που κάνει τη διαφορά. Επιπλέον το πρώτο βήμα έχει πάντα να κάνει με τη δημιουργία ενός ασφαλούς συστήματος μετρήσεων που θα προωθήσει τελικά οποιαδήποτε περαιτέρω δραστηριότητα προς την επιτυχία.

1.3.1. Σημαντικές Επιχειρησιακές Μετρικές

I. Διείσδυση στην αγορά

Η διείσδυση στην αγορά μετρά το σύνολο των πωλήσεων ενός προϊόντος ή μιας υπηρεσίας σε σύγκριση με τη συνολική αγορά-στόχο για το εν λόγω προϊόν (ή υπηρεσία) εντός μιας καθορισμένης περιόδου, συνήθως ενός έτους. Είναι μια πολύ σημαντική μετρική, δεδομένου ότι η ανάπτυξη προέρχεται κυρίως από το κέρδος που οφείλεται σε νέους χρήστες (δηλαδή τη διείσδυση) και όχι από την αύξηση της εμπιστοσύνης στην μάρκα και αποτελεί ένα πρότυπο που συναντάται σε διάφορες κατηγορίες προϊόντων και χώρες. Αυτό έχει άμεσες συνέπειες στους εμπόρους γιατί θα πρέπει να οικοδομήσουν τη φυσική και διανοητική διαθεσιμότητα για την εταιρική τους επωνυμία, μεταξύ όλων των κατηγοριών των αγοραστών, ώστε να αυξηθεί ο αριθμός των χρηστών. Και υπάρχουν δύο τρόποι για να αναπτυχθούν οι εμπορικές μάρκες :

A. Είτε με την απόκτηση των μη αγοραστών.
B. Είτε με τη διατήρηση των σημερινών αγοραστών.

Η αυξανόμενη διείσδυση στην αγορά μπορεί να επιτευχθεί με διάφορους τρόπους, όπως για παράδειγμα με μείωση των τιμών, αύξηση της διάθεσης κεφαλαίου για την προώθηση και τη διανομή ή ακόμα και με την απόκτηση μιας ανταγωνιστικής εταιρείας.

II. Όγκος των Πωλήσεων (volume)

Ο όγκος των πωλήσεων είναι η ποσότητα των αγαθών ή των υπηρεσιών που πωλούνται στο πλαίσιο της κανονικής λειτουργίας μιας επιχείρησης και εντός συγκεκριμένης χρονικής περιόδου (εβδομάδα, μήνας ή ακόμη και ένα έτος).

III. Αξία των πωλήσεων

Η αξία των πωλήσεων είναι το ποσό των χρημάτων που εισπράττονται για τις πραγματοποιηθείσες πωλήσεις εντός μιας συγκεκριμένης περιόδου. Είναι δηλαδή τα έσοδα που αντιστοιχούν στον αντίστοιχο όγκο των πωλήσεων.

IV. Αγοραίο μερίδιο

Το μερίδιο της αγοράς (Share of Market_SOM) είναι το σχετικό επί του συνόλου ποσοστό των συνολικών αγορών ενός προϊόντος ή μιας υπηρεσίας, που αντιστοιχεί σε μια συγκεκριμένη εταιρία που λειτουργεί σε μια αγορά.

V. Δείκτης Τιμών (PI)

Ο δείκτης τιμών (Price index_PI) είναι ο κανονικοποιημένος μέσος όρος της τιμής ενός προϊόντος, σε μια συγκεκριμένη περιοχή, κατά τη διάρκεια ενός συγκεκριμένου χρονικού διαστήματος. Ο Δείκτης Τιμών είναι ένας βασικός δείκτης απόδοσης που έχει σχεδιαστεί για να επιτρέπει συγκρίσεις μεταξύ χρονικών περιόδων, γεωγραφικών τοποθεσιών ή ακόμα και ανταγωνιστών της κύριας αναφορικής εμπορικής μάρκας στην αγορά. Ως εκ τούτου, ένας Δείκτης Τιμών μπορεί να χρησιμοποιηθεί στη διαμόρφωση επιχειρηματικών σχεδίων και στρατηγικών τιμολόγησης και μπορεί να βοηθήσει τους διευθυντές της μάρκας να κατευθύνουν τις επενδύσεις.

VI. Προσφορές στα προϊόντα επί των τιμών (Price promotions)

Οι προσφορές αφορούν τη διάθεση των προϊόντων σε μειωμένη τιμή. Αυτά συνήθως υποδεικνύονται με πινακίδες στους διαδρόμους των καταστημάτων, είτε αναγράφονται με ξεχωριστό τρόπο πάνω στη συσκευασία ενός προϊόντος. Οι προσφορές έχουν άμεσες, αν και βραχυπρόθεσμες επιπτώσεις στις πωλήσεις και χρησιμοποιούνται κυρίως για την ανταμοιβή των αγοραστών ενός προϊόντος. Στη βιβλιογραφία, οι προσφορές επί των τιμών δεν έχουν αποδειχθεί ότι μπορούν να δημιουργήσουν συνεπείς και κατ' επανάληψη καταναλωτές της αντίστοιχης μάρκας.

VII. Διανομή (Distribution)

Η διανομή είναι η διαδικασία κατά την οποία παρέχεται ένα προϊόν (ή υπηρεσία) για κατανάλωση. Αν οι άνθρωποι έχουν τη δυνατότητα να αγοράσουν όλες τις εμπορικές μάρκες στον ίδιο χώρο, τότε θα υπάρξουν πωλήσεις σε όλες τις μάρκες που είναι διαθέσιμες. Έτσι, εάν η διαδρομή για την τελική διάθεση μιας εταιρικής μάρκας είναι ξεχωριστή, τότε είναι πιο πιθανό να έχει διαφοροποιημένους καταναλωτές.

Η διανομή είναι ένα από τα τέσσερα στοιχεία του μίγματος μάρκετινγκ (Marketing mix models_MMM), μαζί με το ίδιο το προϊόν, την τιμολόγησή του και τις προσφορές που επιδέχεται αυτό (επί της τιμής του). Με μία σοφή διαχείριση τεχνικών μάρκετινγκ, η

διανομή μπορεί να κατηγοριοποιηθεί περαιτέρω ως μια *αριθμητική και σταθμισμένη μετρική*. Η αριθμητική διανομή είναι ένα ποσοτικοποιημένο KPI και παρακολουθεί τα συνολικά καταστήματα λιανικής που διαθέτουν τα προϊόντα της αντίστοιχης μάρκας. Η *σταθμισμένη διανομή*, από την άλλη πλευρά, είναι ένας πιο ποιοτικός δείκτης KPI, δεδομένου ότι δεν έχουν όλα τα καταστήματα λιανικής σε μια γεωγραφική περιοχή τον ίδιο αντίκτυπο στην τοπική αγορά, ούτε έχουν τον ίδιο αντίκτυπο στον όγκο των πωλήσεων της εμπορικής μάρκας. Για παράδειγμα ένα κατάστημα εντός ενός μεγάλου εμπορικού κέντρου δεν έχει τον ίδιο αντίκτυπο στις πωλήσεις της μάρκας σε σχέση με ένα μικρό κατάστημα λιανικής στην ίδια περιοχή. Έτσι, η σταθμισμένη διανομή είναι η ποιοτική αξιολόγηση της διανομής.

1.3.2 Οι Σημαντικές Μετρικές για τα Media

Ο κατάλογος των διαθέσιμων μέσων ενημέρωσης (ή σημείων επαφής) σε κάθε διακεκριμένη επιχειρηματική κατηγορία μπορεί να είναι εκπληκτικά μεγάλος (Εικόνα 1.1) και μπορεί να διαφέρει μεταξύ των βιομηχανιών. Οι εταιρείες των μέσων ενημέρωσης θα πρέπει κατά κάποιον τρόπο να μπορούν να μετρήσουν τόσο την απόδοση όσο και την αποτελεσματικότητα κάθε διαφημιστικού καναλιού, όπως η τηλεόραση, το ραδιόφωνο, ο τύπος, το διαδίκτυο κλπ. Αυτό είναι δύσκολο και ως εκ τούτου, για κάθε μέσο διαφήμισης υπάρχουν εταιρείες που ειδικεύονται στη μέτρηση του αντίκτυπου που αυτό έχει στους πιθανούς αγοραστές αλλά και στη συλλογή πολύτιμων δεδομένων για το σκοπό αυτό.

Στις περισσότερες χώρες μια ενιαία εταιρεία έρευνας για τα μέσα ενημέρωσης εγκρίνεται συνήθως μέσω των κορυφαίων σωματείων της διαφημιστικής βιομηχανίας για να ενεργεί ως ο επίσημος πάροχος των μετρήσεων ακροαματικότητας των διαφημίσεων. Τα ίδια τα μέλη της βιομηχανίας χρηματοδοτούν την έρευνα του κοινού και μοιράζονται τα ευρήματα. Η μέτρηση αυτή χρησιμοποιείται σε κάθε χώρα για να αξιολογήσει τα καταναλωτικά πρότυπα που σχηματίζει το διαφημιστικό περιεχόμενο των μέσων ενημέρωσης. Συνήθως αυτό συμβαίνει σε σχέση με την ακρόαση του ραδιοφώνου, την παρακολούθηση της τηλεόρασης, παράλληλα με τα αναγνωσμένα έντυπα (εφημερίδες και περιοδικά) αλλά και την όλο και περισσότερο, διαδικτυακή κίνηση.

Τέσσερις βασικές μέθοδοι συλλογής δεδομένων χρησιμοποιούνται από ερευνητικές εταιρείες για την παρακολούθηση της ακροαματικότητας του κοινού:

Οι ονομαστικές συνεντεύξεις, η διατήρηση ημερολογίου, οι μετρητές ατόμων ή σάρωσης και η μοντελοποίηση.

Οι ερευνητικές εταιρείες προσαρμόζουν τις μεθοδολογίες τους σύμφωνα με τον τύπο των μέσων ενημέρωσης και το κόστος που παρουσιάζει η συλλογή των δεδομένων. Τα αποτελέσματα ταξινομούνται στη συνέχεια βάσει δημογραφικών στοιχείων όπως το φύλο, η ηλικία, η φυλή και το εισόδημα. Λεπτομερή δεδομένα σχετικά με την προσέγγιση (συμπεριλαμβανομένων των ποσοστών των θεατών σε συγκεκριμένες δημογραφικές ομάδες, εβδομαδιαίους και μηνιαίους μέσους όρους, μαζί με τον εκτιμώμενο συνολικό αριθμό θεατών) διανέμονται στη συνέχεια σε τηλεοπτικά δίκτυα και διαφημιζόμενους. Η μέτρηση του κοινού βοηθάει μια ποικιλία επαγγελμάτων, συμπεριλαμβανομένων των ραδιοτηλεοπτικών φορέων και των διαφημιζόμενων, να προσδιορίσουν ποιος λαμβάνει στην πραγματικότητα τα άτομα που επικοινωνούν, αλλά μάλλον παρακολουθεί τον αριθμό των ατόμων που εκτίθενται στο συγκεκριμένο μέσο κάθε φορά. Στη συνέχεια, οι ιδιοκτήτες επιχειρήσεων μπορούν να λάβουν αυτές τις πληροφορίες μέσω των οργανισμών των μέσων ενημέρωσης για να στοχεύσουν τις ομάδες που είναι πιο πιθανό να αγοράσουν το προϊόν ή την υπηρεσία τους.

PAID	OWNED	EARNED
Ads at sports events	After sales service	Awards
Ads near store	Booklets	Blogger review
Ads in waiting rooms	Brand blog	Colleagues reco
Advertorial	Brand event	Comparison websites
Airport ads	Brand store	Cons opinion site/ blogs
Bus shelter ads	Brand video channel	Expert reco
Celebrity endorsement	Brand website	Friend-get-friend schemes
Cinema ads	Branded merchandise	Friends / family reco
Event sponsor	Branded vehicles	Independent review
Facebook ads	Brochure	Print articles
Good causes sponsor	Call centre	Recommendation from a friend
Gym / club ads	Contest / competitions	Relevant magazines
Huge outdoor ads	Coupons	Relevant mobile phone sites
Instant messenger ads	Customer mag	Relevant radio progs
Instore ads	Direct mail	Relevant TV progs
Instore TV ad	Display in public area	Relevant websites
Internet ads	Email	Salesperson reco
Internet display ads	Event demo	Seeing others with brand
Internet print ads	Facebook brand page	Social network comments
Internet radio ads	Free gift with purchase	Specialist one-to-one reco
Internet search	Gym / club sample	Video blogs
Internet video ads	Home visit from salesperson	
Local newspaper ads	Info on packaging	
Magazine ads	Instagram brand page	
Mass transit ads	Instore displays	
Radio ads	Instore flyers	
Radio program sponsor	Instore owned TV	
Retailer customer magazine	Instore promo	
Retailer stores	Instore sample	
Retailer websites	Internet brand video	
Snapchat brand video	Invitation	
Social network ads	Leaflets	
Mobile game ads	Loyalty card	
Newspaper ads	Mobile apps	
Online directories	Mobile SMS	
Outdoor ads	Online promo	
Petrol station pump ads	Packaging	
POC ads	Personal letter	
Print inserts by retailers	Pinterest brand page	
Print inserts / leaflets	Podcasts	
Print promo	Receipts	
Product placement	Retailer catalogues	
Sponsored email	Retailer promotion	
Sports sponsor	Roadshows	
Train ads	Sampling	
TV ads	Seeing product in store	
TV program sponsor	Shelf ads	
TV screens in public areas	Social network page	
Twitter ads	Social networking sites	
Video game ads	Twitter brand tweets	
Website sponsor	Vending machines	

Σχήμα 1.1 Κατάλογος των διαθέσιμων μέσων ενημέρωσης (ή σημείων επαφής) σε κάθε διακεκριμένη επιχειρηματική κατηγορία

I. Media Market

Η Media Market περιγράφει το σύνολο των ατόμων που θα μπορούσαν ενδεχομένως να εκτεθούν σε μια διαφήμιση. Ο πληθυσμός είναι ο συνολικός αριθμός των ατόμων στην αγορά των μέσων μαζικής ενημέρωσης.

II. Κόστος πολυμέσων (Media Cost)

Το Media Cost είναι το συνολικό κόστος που καταβάλλεται για μια διαφήμιση που πρέπει να κοινοποιηθεί στην αγορά των μέσων μαζικής ενημέρωσης. Το Media Cost αποκλείει το κόστος της παραγωγής για τη δημιουργία της διαφήμισης.

III. Spot

Το spot είναι μια ενιαία εκπομπή μιας διαφήμισης.

IV. Rating

Η Rating είναι το ποσοστό των ατόμων (στο πλαίσιο της Media Market) που ενδέχεται να έχουν εκτεθεί σε μια διαφήμιση. Για παράδειγμα, εάν έχουμε ως πιθανή ομάδα-στόχο της επικοινωνίας μας, ενήλικες ηλικίας μεταξύ 25 και 44 ετών, στην τηλεόραση, μία Rating αντιστοιχεί στο 1% των ενηλίκων 25-44 που παρακολούθησαν μια συγκεκριμένη διαφήμιση. Αντίστοιχα, στον Τύπο, μια Rating αντιστοιχεί στο 1% των ενηλίκων 25-44 που διαβάζουν ένα μέσο τεύχος ενός συγκεκριμένου εντύπου και στο ραδιόφωνο, η βαθμολογία 1 ισούται με το 1% των ενηλίκων 25-44 που άκουσαν έναν συγκεκριμένο σταθμό σε μια συγκεκριμένη στιγμή εγκαίρως.

V. Μέσος αριθμός ατόμων (Average Persons)

Ο Μέσος αριθμός ατόμων είναι ο μέσος αριθμός των ατόμων που θα εκτεθούν σε κάθε Spot. Ο αριθμός που αντιστοιχεί στη μετρική Average Persons υπολογίζεται πολλαπλασιάζοντας τον Πληθυσμό με τη Rating και στη συνέχεια διαιρούμε με το 100.

VI. Εμφανίσεις (Impressions)

Οι Εμφανίσεις είναι ο συνολικός αριθμός των εκθέσεων σε μια διαφήμιση. Περιττό να πούμε ότι ένα άτομο μπορεί να εκτίθεται στην ίδια διαφήμιση πολλές φορές. Οι εμφανίσεις υπολογίζονται πολλαπλασιάζοντας τον αριθμό των spot κατά μέσον όρο.

VII. Reach

Η προσέγγιση (RCH) είναι ο αριθμός των πιθανών ατόμων που ενδέχεται να εκτεθούν σε ένα μόνο spot στη Media Market και συνήθως μετράται για συγκεκριμένο χρονικό διάστημα, π.χ. μια εβδομάδα. Η Reach είναι μία από τις βασικές μετρικές για τη γενική αξιολόγηση μιας διαφημιστικής καμπάνιας στα μέσα

ενημέρωσης και πιο συγκεκριμένα στην τηλεόραση. Συνήθως συλλέγονται εκτεταμένα δεδομένα από τους σχεδιαστές των μέσων, μέσω της Reach ενός συγκεκριμένου show ή ενός δικτύου. Αυτά τα δεδομένα χρησιμοποιούνται στη συνέχεια για να καθορίσουν τις αποφάσεις σχετικά με το πότε και πού να εκπέμπουν τα τηλεοπτικά μηνύματα των πελατών τους. Η εκτίμηση της Reach είναι ένα σημαντικό ζήτημα γιατί όταν μια διαφήμιση εκπέμπεται πολλές φορές, το ίδιο άτομο μπορεί να εκτεθεί στη διαφήμιση περισσότερες από μία φορές. Ωστόσο, κατά τη σωστή μέτρηση της κλίμακας Reach, πρέπει να εξαιρεθούν τα διπλά ενδεχόμενα. Υπάρχουν πολλές διαφορετικές μέθοδοι για την εκτίμηση της Reach. Η προσέγγιση μπορεί επίσης να εκφραστεί ως ποσοστό, το οποίο δείχνει το ποσοστό του πληθυσμού που εκτίθεται σε ένα spot τουλάχιστον μία φορά.

VIII. Συχνότητα (Frequency)

Η συχνότητα (ή η μέση συχνότητα) είναι ο μέσος όρος των επαναλήψεων που η διαφήμιση θα παρουσιαστεί στον πληθυσμό της RCH. Ένας τρόπος για να υπολογίσουμε τη συχνότητα είναι να διαιρέσουμε τον αριθμό Impressions με τη Reach. Ένας τεράστιος όγκος ερευνών έχει διεξαχθεί για την εξεύρεση της πιο αποτελεσματικής συχνότητας και του επιπέδου έκθεσης που απαιτείται, σε μία διαφήμιση, ώστε η τελευταία να έχει μια ουσιαστική επίδραση στην αγορά. Αξίζει να σημειωθεί ότι μολονότι μία έκθεση μπορεί να έχει σημαντικό αντίκτυπο στην αγορά, ωστόσο, δύο εκθέσεις μπορεί να μην έχουν διπλάσια επίδραση σε σχέση με την ενιαία έκθεση, αλλά συνήθως κοστίζουν δύο φορές περισσότερο. Αυτή η μειούμενη επιστροφή βαθαίνει όσο περισσότερη συχνότητα προστίθεται στη διαφημιστική εκστρατεία.

IX. Gross Rating Point (GRP)

Το Gross Rating Point (GRP) είναι ένα μέτρο του μεγέθους μιας διαφημιστικής καμπάνιας σε ένα συγκεκριμένο μέσο. Τα GRP υπολογίζονται με τον πολλαπλασιασμό του αριθμού των spot με τη Rating. Εναλλακτικά, τα GRP υπολογίζονται πολλαπλασιάζοντας τη Reach με τη συχνότητα της έκθεσης του συγκεκριμένου κοινού στο διαφημιστικό μήνυμα κατά τη διάρκεια μιας δεδομένης περιόδου. Για παράδειγμα, αν μια τηλεοπτική διαφήμιση φτάσει το 40% του κοινού-στόχου της και η διαφήμιση προβάλλεται 5 φορές, τότε η διαφημιστική καμπάνια έχει συγκεντρώσει 200 μικτούς βαθμούς αξιολόγησης. Ωστόσο, τα GRP δεν μετρούν το

μέγεθος του ακροατηρίου που επιτεύχθηκε. Αντιθέτως, ποσοτικοποιούν τις εντυπώσεις ως ένα ποσοστό του πληθυσμού-στόχου και το ποσοστό αυτό μπορεί να είναι μεγαλύτερο ή, στην πραγματικότητα, πολύ μεγαλύτερο από 100. Το άθροισμα των αξιολογήσεων ενός μηνιαίου ή ετήσιου τηλεοπτικού σχεδίου παράγει τα GRP του σχεδίου των μέσων.

Τα GRP χρησιμοποιούνται κυρίως ως ένα μέτρο των μέσων με υψηλές πιθανές εκθέσεις ή εμφανίσεις (impressions). Ο στόχος της μέτρησης GRP είναι να μετρήσει τις εμφανίσεις σε σχέση με τον αριθμό των ατόμων που βρίσκονται στο στόχο για μια διαφημιστική καμπάνια. Οι διαφημιζόμενοι, οι έμποροι και οι αγοραστές των μέσων ενημέρωσης αξιολογούν τις διαφημιστικές καμπάνιες, παρακολουθώντας τόσο το Reach που προσφέρεται από το μέσο, όσο και την αντίστοιχη συχνότητα με την οποία ο θεατής βλέπει τη διαφήμιση. Οι τιμές GRP χρησιμοποιούνται συνήθως για να συγκρίνουν τη δύναμη της διαφήμισης των επιμέρους στοιχείων ενός διαφημιστικού σχεδίου των μέσων.

X. Cost per Point (CPP)

Το κόστος ανά σημείο Cost per Point (CPP) είναι μια μέτρηση κόστους-αποτελεσματικότητας που επιτρέπει συγκρίσεις μεταξύ των διαφημίσεων:

$$CPP = \frac{Media\ Cost}{GRPs}$$

XI. Cost per Rating Point (CPR)

Το κόστος ανά σημείο βαθμολόγησης Cost per Rating Point (CPR) είναι το κόστος ανά δευτερόλεπτο ώστε να εκτίθεται το μήνυμα σε ένα ποσοστό του κοινού-στόχου:

$$CPR = \frac{\frac{Media\ Cost}{GRPs}}{Average\ spot\ duration}$$

Το CPR θεωρείται μία από τις *σημαντικότερες μετρικές* στον εμπορικό κλάδο, στο χώρο των σχεδιαστών των μέσων και των αγοραστών, δεδομένου ότι όλοι πρέπει να γνωρίζουν πόσοι άνθρωποι αναμένεται να προσεγγιστούν από τις τακτικές μάρκετινγκ που ακολουθούν. Το CPR είναι ιδιαίτερα σημαντικό στα ακριβά μέσα ενημέρωσης όπως η τηλεόραση, όπου η προσδοκία προσέγγισης ενός μεγάλου ακροατηρίου μπορεί να δικαιολογήσει μόνο την τιμή της τοποθέτησης μιας διαφήμισης. Έτσι, είναι μια μέτρηση που χρησιμοποιούν οι έμποροι και οι αναλυτές

για να καθορίσουν πού θα τοποθετήσουν τις διαφημίσεις τους και πόσο θα έπρεπε, μέσα σε ένα λογικό πλαίσιο να ξοδέψουν για αυτές.

Επιπλέον, η μετρική CPR μπορεί να παρέχει μια βασική ακολουθία πληροφοριών για τη σύγκριση του κόστους της διαφήμισης στα διάφορα μέσα ενημέρωσης. Για παράδειγμα, τα εθνικά τηλεοπτικά δίκτυα χρεώνουν συνήθως πολύ πιο ακριβά σε σχέση με τους τοπικούς τηλεοπτικούς ή ραδιοφωνικούς σταθμούς, κυρίως λόγω της δυνατότητας προσέγγισης μεγαλύτερων αγορών. Ωστόσο, με τον υπολογισμό του CPR της κάθε αγοράς, μπορεί κάποιος να επιλέξει την καλύτερη χρονοθυρίδα για την προβολή της διαφήμισης που διαθέτει και η επιλογή αυτή να έχει γίνει με βάση το συνδυασμό της αποτελεσματικότητας και την αξίας της. Τέλος, το CPR είναι το ίδιο με το CPP, όπου κάθε σημείο αναφέρεται στο ένα τοις εκατό μιας συγκεκριμένης αγοράς.

XII. Share

Η κάλυψη (SHR) ενός συμβάντος στην τηλεόραση είναι το ποσοστό των ατόμων που παρακολουθούν τηλεόραση και έχουν παρακολουθήσει μια συγκεκριμένη παρουσίαση. Υπολογίζεται διαιρώντας τον αριθμό των ανθρώπων που έχουν δει τη συγκεκριμένη παρουσίαση με το σύνολο των ανθρώπων του στόχου που παρακολουθούσαν τηλεόραση κατά την ίδια περίοδο.

1.3.3 Οι Σημαντικές Μετρικές για τα Digital Media

Πριν από πολύ καιρό υπήρχαν μόνο μερικά παραδοσιακά μέσα ενημέρωσης (τηλεόραση, ραδιόφωνο, διαφημίσεις στον τύπο) και ήταν αρκετά ευέλικτα στη (λειτουργική) δομή τους. Σήμερα, ζούμε στην ψηφιακή εποχή, με τους καταναλωτές να συνδέονται συνεχώς 24/7/365 στο διαδίκτυο, να περιβάλλονται από επαναφορτιζόμενες συσκευές και να εμπιστεύονται μία ποικιλία έξυπνων συσκευών για να λάβουν σημαντικές αποφάσεις γι' αυτούς. Ειδικά τα κινητά τηλέφωνα νέας τεχνολογίας (smartphones) διαδραματίζουν πρωταγωνιστικό ρόλο στον τρόπο με τον οποίο οι καταναλωτές επιλέγουν να κάνουν τις αγορές τους και δημιουργούν μία νέα αγοραστική εμπειρία για αυτούς. Έρευνες δείχνουν πως πολλές οικονομικές συναλλαγές πραγματοποιούνται πια μέσω του κινητού. Τα όρια μεταξύ των διαφόρων

καναλιών επικοινωνίας είναι πλέον θολά, καθώς το κινητό μπορεί να χρησιμοποιηθεί για ψυχαγωγία, ειδήσεις, πληροφορίες, έρευνα, κοινωνικοποίηση και επικοινωνία.

Αυτό φυσικά οδήγησε τα τελευταία χρόνια στην ανάπτυξη της διαφήμισης μέσω ψηφιακών μέσων, η οποία παραμένει στο προσκήνιο σε ολόκληρη τη διαφημιστική αγορά παγκοσμίως. Και αυτή η ανάπτυξη πραγματοποιείται με τόσο γρήγορο ρυθμό που οι διαφημιστικές δαπάνες για το Διαδίκτυο αυξάνονται ταχύτερα σε σχέση με τις δαπάνες που αφορούν τη διείσδυση (penetration) ή ακόμα και την αύξηση της χρήσης (usage growth) σε όλο τον κόσμο.

Το κύριο πρόβλημα είναι ότι τα ψηφιακά μέσα είναι πολυάριθμα και διαφορετικά όπως είναι άλλωστε και οι διαφορετικές μετρικές και οι δείκτες KPIs ανάμεσά τους. Αναφέρουμε μερικές όπως, οι Impressions, ακόλουθοι (followers), likes (χαρακτηρισμός όταν κάτι είναι αρεστό στο χρήστη όπως συμβαίνει στο Facebook, Instagram) , μέσος χρόνος παραμονής στο site, μετατροπές, CPM, CPA κ.ά. . Παρακάτω, έχουμε μια επιλογή από τους σημαντικότερους δείκτες απόδοσης KPIs των ψηφιακών μέσων.

I. Impressions

Όμοια με τα Media

II. Οι μοναδικές εμφανίσεις (Unique impressions)

Οι μοναδικές εμφανίσεις είναι μία κοινή μέτρηση για την κυκλοφορία στον ιστό (web traffic). Οι μοναδικοί χρήστες υπολογίζονται από τα cookies που εκδίδονται από τα προγράμματα περιήγησης όταν οι επισκέπτες επισκέπτονται έναν συγκεκριμένο ιστότοπο τουλάχιστον μία φορά. Μετράει την εμβέλεια μιας διαφήμισης.

III. Viewability

Η viewability είναι μια μετρική της διαφήμισης στο διαδίκτυο που στοχεύει στην παρακολούθηση μόνο εκείνων των εμφανίσεων που πραγματικά βλέπουν οι χρήστες. Για παράδειγμα, μια διαφήμιση που έχει φορτωθεί στο κάτω μέρος μιας ιστοσελίδας δεν μπορεί να παρατηρηθεί από έναν χρήστη που δεν θα μετακινηθεί αρκετά μακριά, έτσι θα θεωρείται ότι δε θα προβληθεί. Η viewability είναι μια μέτρηση που έχει σχεδιαστεί για να επιτρέπει στους διαφημιζόμενους να πληρώνουν μόνο για τις διαφημίσεις που

θα μπορούσαν ενδεχομένως να βλέπουν οι χρήστες. Βοηθά επίσης τους εμπόρους στο να παρέχουν μετρήσεις σχετικά με το πόσες φορές οι διαφημίσεις τους εμφανίζονται πραγματικά μπροστά από τους χρήστες. Αυτό είναι πολύ σημαντικό, γιατί όταν δεν εμφανίζεται μια διαφήμιση, δεν μπορεί να έχει αντίκτυπο στους καταναλωτές. Επομένως, αυτή η μετρική επιτρέπει την καλύτερη κατανόηση της αποτελεσματικότητας της καμπάνιας και την κατανομή των διαφημιστικών δαπανών στα πιο αξιόλογα μέσα.

IV. Ποσοστό προβολής (Viewability Rate)

Ποσοστό προβολής είναι το ποσοστό των διαφημίσεων που βλέπει στην πραγματικότητα ένας χρήστης, δηλαδή το ποσοστό των εμφανίσεων που θεωρούνται ως προβαλλόμενες. Εμφανίζεται μια εντύπωση περιγράφεται ως τουλάχιστον το 50% στην οθόνη ενός χρήστη για ένα δευτερόλεπτο.

$$Viewability\ Rate = \frac{Total\ measured\ Viewable\ ad\ impressions}{Total\ measured\ ad\ impressions} \times 100\%$$

Για παράδειγμα, εάν ένας ιστότοπος (website) έχει δέκα διαφημίσεις, εκ των οποίων μόνο οι πέντε θεωρούνται ως προβεβλημένες, τότε ο ιστότοπος θα έχει ποσοστό προβολής 50%. Αξίζει να σημειωθεί ότι οι προμηθευτές που ελέγχουν την προβολή δεν μπορούν να μετρήσουν το σύνολο των εμφανίσεων των διαφημίσεων (ad impressions). Ως αποτέλεσμα, οι υπολογίσιμες εμφανίσεις των διαφημίσεων είναι πάντα χαμηλότερες από τις κανονικές εμφανίσεις αυτών.

V. Clicks

Σύμφωνα με την Interactive Advertising Bureau[1], ένα κλικ είναι «μια αλληλεπίδραση μεταξύ του επισκέπτη ενός ιστότοπου και του προγράμματος περιήγησης στο οποίο ο επισκέπτης χρησιμοποιεί μια συσκευή, όπως ένα ποντίκι, για να μετακινήσει τον δρομέα σε μια ενεργή περιοχή της οθόνης και στη συνέχεια να αλληλεπιδράσει εσκεμμένα με αυτήν την περιοχή κάνοντας κλικ σε ένα κουμπί της συσκευής του, ενεργοποιώντας ένα συμβάν». Τα κλικ

[1] The Interactive Advertising Bureau (IAB) is an advertising business organization that develops industry standards, conducts research, and provides legal support for the online advertising industry. The organization represents a large number of prominent media outlets globally.

είναι ο αριθμός των φορών που έχει γίνει κλικ σε μια διαφήμιση και έπειτα έχει συνδεθεί σε μια συγκεκριμένη σελίδα προορισμού ή ιστότοπου.

VI. Click through rate (CTR)

Η αναλογία κλικ / εμφανίσεων (CTR) είναι το ποσοστό των ατόμων που βλέπουν τελικά τη διαφήμισή έχοντας κάνει αρχικά κλικ:

$$CTR = \frac{number\ of\ click - throughs}{number\ of\ ad\ impressions} \times 100\%$$

VII. Unique clicks

Τα Μοναδικά κλικ (unique clicks) είναι ο αριθμός των κλικ που προέρχονται από μοναδικούς θεατές. Είναι ένας αριθμός πάντα μικρότερος από τον αντίστοιχο αριθμό κλικ.

VIII. Cost per Click (CPC)

Το κόστος ανά κλικ (CPC) είναι το κόστος για κάθε εξασφαλισμένο κλικ:

$$CPC = \frac{Cost}{Number\ of\ clicks}$$

Για παράδειγμα, αν δαπανηθούν 1.000 € για τη διαφήμιση σε μια συγκεκριμένη τοποθεσία στο διαδίκτυο και υπάρχουν 500 κλικ πάνω στη διαφήμιση αυτή, τότε $CPC = \frac{1000€}{500\ clicks} = 2€\ per\ click.$

IX. Κόστος ανά απόκτηση (Cost per acquisition)

Το κόστος ανά απόκτηση (CPA) είναι μια διαδικτυακή στρατηγική διαφήμισης που επιτρέπει σε έναν διαφημιζόμενο να πληρώσει έναν έμπορο θυγατρικών για να μετατρέψει ένα άτομο από απλό επισκέπτη ενός site σε έναν πελάτη για την εταιρεία. Μετράει το συνολικό κόστος από την αρχή μέχρι το τέλος για να επιτευχθεί μια τέτοια μετατροπή, δηλαδή από την αρχική ένταξη στα αποτελέσματα της μηχανής αναζήτησης μέχρι τη δημιουργία ενδιαφερόντων σελίδων προορισμού που προσελκύουν την προσοχή των επισκεπτών. Συνήθως μια μετατροπή είναι είτε μια αγορά είτε ένα έντυπο που συμπληρώνεται από έναν επισκέπτη ενός site. Αυτή η μορφή πληρωμής ένταξη συχνά προτιμάται από τους διαφημιζόμενους, αφού πληρώνουν μόνο για το επιτευχθέν αποτέλεσμα. Λάβετε υπόψη ότι τα CPM, το CPC

και το CPA θεωρούνται τα τρία πιο συνηθισμένα μοντέλα τιμολόγησης διαφημίσεων που χρησιμοποιούνται.

1.4 Στατιστική των Επιχειρήσεων (Business Analytics)

Κύριο αντικείμενο έρευνας και μελέτης της Στατιστικής είναι η συλλογή, ταξινόμηση, επεξεργασία, παρουσίαση, ανάλυση και ερμηνεία διαφόρων δεδομένων με απώτερο στόχο την εξαγωγή ασφαλών συμπερασμάτων για τη λήψη ορθών αποφάσεων. Είναι γνωστό πλέον τοις πάσι ότι πρόκειται για μία σημαντική επιστήμη της οποίας οι εφαρμογές έχουν ευρύτατο πεδίο, τόσο στις επιχειρήσεις και στη διοικητική, όσο και στη βιοϊατρική, αλλά και στις θετικές, συμπεριφορικές ή κοινωνικές επιστήμες, ενώ είναι τόσο αναγκαία η γνώση της όσο η ικανότητα που έχει κάποιος να διαβάζει ή να γράφει. Η Στατιστική των Επιχειρήσεων θεωρείται ως η επιστήμη της καλής λήψης αποφάσεων, υπό το πλαίσιο της αβεβαιότητας. Οι εφαρμογές της περιλαμβάνουν πολλούς κλάδους όπως την οικονομική ανάλυση, τη στοιχειοθέτηση των Big Data, τον έλεγχο και τις πρακτικές που εφαρμόζονται για τη βελτίωση των υπηρεσιών και την έρευνα μάρκετινγκ. Με τον όρο Big Data εννοούμε τις τεράστιες ποσότητες δομημένων, ημι-δομημένων και αδόμητων δεδομένων σε συνδυασμό με τις εξελίξεις στην τεχνολογία που συνέβησαν τα τελευταία 50 έτη. Η διαχείριση των Big Data που είναι το νούμερο ένα επιχειρησιακό πρόβλημα περιλαμβάνει:

A. Την αποθήκευση: Καταγραφή και αποθήκευση των δεδομένων .
B. Την επεξεργασία: Αξιολόγηση και ανάλυση των δεδομένων (analytics)
C. Την πρόσβαση: Ανάκληση και οπτικοποίηση των δεδομένων.

Και ας μην ξεχνάμε πως ότι σήμερα θεωρείται «μεγάλος» όγκος δεδομένων (Volume), στο μέλλον θα είναι ακόμα μεγαλύτερος.

Η ανάλυση κατηγοριοποιείται σε τρεις κύριους κλάδους (όπως φαίνεται και στο παρακάτω Σχήμα 1.2) : Περιγραφική Στατιστική (**Descriptive Analytics**), Διαγνωστική Στατιστική (**Predictive Analytics**) και Καθοδηγητική Αναλυτική Στατιστική (**Prescriptive Analytics**) .

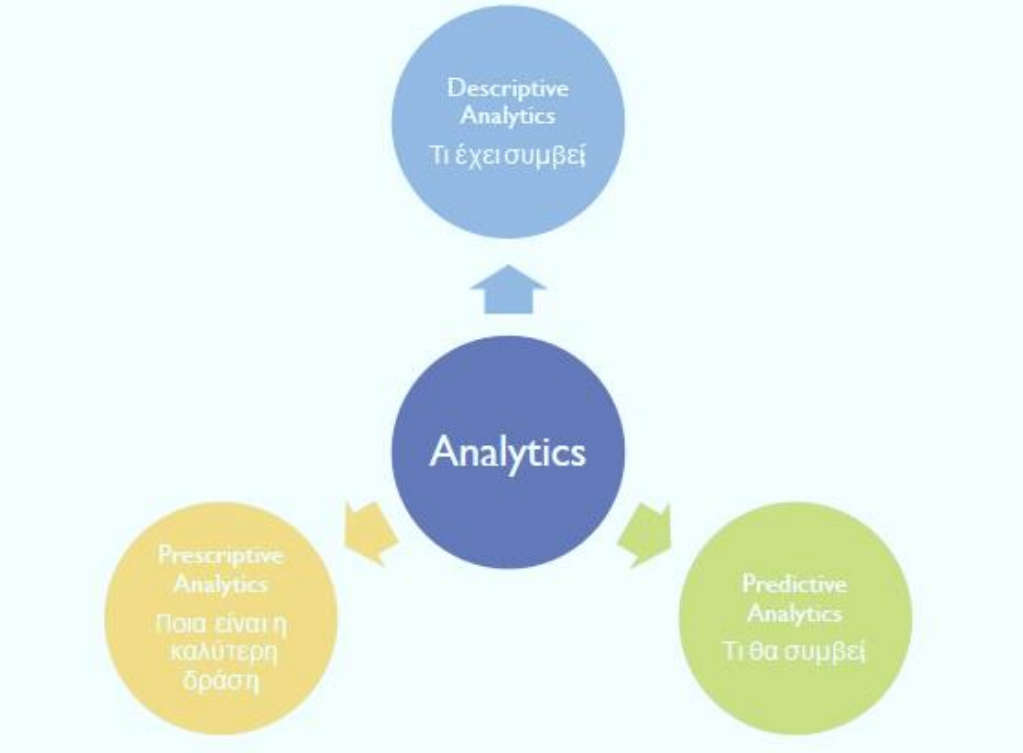
1.4.1.Descriptive Analytics: Μία αναδρομή στο παρελθόν

Κατά τη Περιγραφική Στατιστική Ανάλυση εξετάζουμε τα όσα έχουν ήδη συμβεί και μπορούμε να εξηγήσουμε γιατί συνέβησαν. Με τη χρήση ιστορικών δεδομένων, οι επιστήμονες μπορούν να αναλύσουν επιτυχίες και αποτυχίες του παρελθόντος, διερευνώντας τον αντίκτυπο συγκεκριμένων αποτελεσμάτων. Ορισμένες από τις πιο κοινές εφαρμογές των Descriptive Analytics στην επιχείρηση περιλαμβάνουν τις πωλήσεις, το μάρκετινγκ, τη χρηματοδότηση και τις δραστηριότητες.

Κατά την ανάλυση αυτού του τύπου πρέπει να επιτυγχάνονται τα ακόλουθα:

- Περιγραφή των δεδομένων με στόχο την κατανόησή τους.
- Δημιουργία διαίσθησης για τα δεδομένα.

  Το σύστημα χρησιμοποιεί μια σειρά από «κόλπα» για να εξομοιώσει την ανθρώπινη διαίσθηση όσον αφορά στον εντοπισμό μοτίβων, όπως την αξιοποίηση της δομής των βάσεων δεδομένων που αναλύει για τη δημιουργία νέων συγκριτικών δεδομένων, κάτι που ακολουθείται από μια σειρά διαφορετικών υπολογισμών προκειμένου να βρεθούν συσχετισμοί. Επίσης, ιδιαίτερη προσοχή δίνεται σε συγκεκριμένα δεδομένα (όπως ένας μήνας, ένα όνομα ενός brand ή κάτι ανάλογο), προκειμένου να μελετώνται οι σχέσεις μεταξύ των κατηγοριών τους και των νέων συγκριτικών δεδομένων.
- Δημιουργία αναφορών (reports).
- Εντοπισμός καταστάσεων που χρήζουν προσοχής.
- Ομαδοποίηση αντικειμένων με παρόμοια χαρακτηριστικά (clustering).



Σχήμα 1.2 Ανάλυση μεγάλου όγκου Δεδομένων (Big Data)[2]

1.4.2.Predictive Analytics: Κατανόηση του μέλλοντος

Κατά την Προγνωστική Αναλυτική διαδικασία χρησιμοποιούνται μια ποικιλία προηγμένων, πολύπλοκων στατιστικών τεχνικών όπως η μαθηματική προτυποποίηση και οι αλγόριθμοι εξόρυξης δεδομένων για την πρόβλεψη μελλοντικών αποτελεσμάτων και τάσεων με βάση τα δεδομένα που συλλέγονται. Σε αντίθεση με την Περιγραφική Στατιστική, η Προγνωστική Αναλυτική υπερβαίνει τα όσα αναφέρθηκαν ήδη, για να δημιουργηθούν οι καλύτερες εκτιμήσεις για το τι είναι πιθανό να συμβεί στο εγγύς μέλλον. Υπάρχουν πολλά μοντέλα πρόβλεψης με καλύτερες και χειρότερες επιδόσεις ανά πρόβλημα. Είναι σημαντικό να θυμόμαστε ότι κανένας στατιστικός αλγόριθμος δεν μπορεί να «προβλέψει» το μέλλον με 100% βεβαιότητα. Ορισμένες από τις πιο κοινές εφαρμογές του λογισμικού Predictive Analytics in Business περιλαμβάνουν το μάρκετινγκ και τις επιχειρηματικές δραστηριότητες. Έτσι οι προγνωστικές αναλύσεις μπορούν να χρησιμοποιηθούν σε ολόκληρο τον οργανισμό. Ειδικότερα:

- Στην πρόβλεψη της συμπεριφοράς των πελατών και των προτύπων αγοράς.

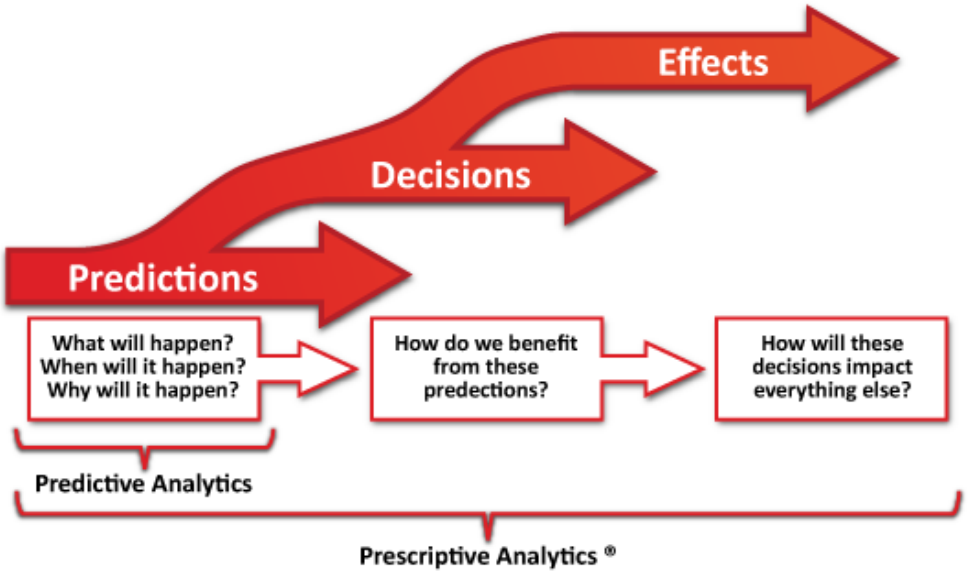
[2] https://slideplayer.gr/slide/2015423/

- Στον εντοπισμό των τάσεων στις δραστηριότητες των πωλήσεων.
- Στην πρόβλεψη της ζήτησης για εισροές από την αλυσίδα εφοδιασμού, τις επιχειρήσεις και το απόθεμα.
- Στη δημιουργία ενός πιστωτικού αποτελέσματος με βαθμολογία. Αυτές οι πιστωτικές βαθμολογίες χρησιμοποιούνται από τις χρηματοπιστωτικές υπηρεσίες για να προσδιοριστεί η πιθανότητα των πελατών να καταβάλλουν έγκαιρα τις μελλοντικές πληρωμές πίστωσης (δάνεια κ.λπ.).

1.4.3. Prescriptive Analytics: Συμβουλές σχετικά με τα πιθανά αποτελέσματα

Η Καθοδηγητική Αναλυτική ξεπερνά το στάδιο της πρόβλεψης των μελλοντικών αποτελεσμάτων και αφορά αποκλειστικά τον καθορισμό της βέλτιστης διαδρομής μιας συγκεκριμένης επιχειρηματικής δραστηριότητας. Αυτό περιλαμβάνει τη γνώση του τι μπορεί να συμβεί και του γιατί μπορεί να συμβεί και επίσης δημιουργεί τα πλαίσια για να μπορεί ο αναλυτής να προτείνει ενέργειες από τις οποίες θα επωφεληθούν οι εταιρείες αφού είναι σε θέση να τους παρουσιάζει και τις συνέπειες της κάθε επιλογής ή απόφασης. Θεωρείται ως ένα νέο πεδίο στην Data Science.

Ένα τέλειο παράδειγμα του Prescriptive Analytics μπορεί να είναι το Waymo One, το αυτόνομα καθοδηγούμενο αυτοκίνητο της Google που αναλύει όλα τα διαθέσιμα δεδομένα σχετικά με το περιβάλλον, την κυκλοφορία, κ.λπ. και αποφασίζει για τη βέλτιστη διαδρομή που θα ακολουθήσει. Το λογισμικό του αυτοκινήτου αποφασίζει για την επιβράδυνση, την αλλαγή λωρίδων, τις επιλογές πλοήγησης ως προς συντομότερες διαδρομές κ.λπ. Είναι εντυπωσιακό πως το Waymo One συμπεριφέρεται ακριβώς όπως ένας άνθρωπος χρησιμοποιώντας όλα τα διαθέσιμα δεδομένα για να αναλύσει την κατάσταση σε πραγματικό χρόνο. Περιττό να πούμε ότι για να εκτελείται μία Καθοδηγητική Ανάλυση και να διατηρούνται αυτόματα τα σχέδια δράσης σε μόνιμα πραγματικό χρόνο, είναι απαραίτητη η συνεχής ενημέρωση των πληροφοριών. Βοηθητικό είναι και το ακόλουθο Σχήμα 1.3.



Σχήμα 1.3 Οι τρεις φάσεις των Analytics[3]

Πρακτικά λοιπόν κατά τη διαδικασία τέτοιων αναλύσεων θα πρέπει ουσιαστικά να έχουμε διαρκώς απαντήσεις στα ακόλουθα ερωτήματα:

- Πώς οι προβλέψεις για το μέλλον μπορούν να αλλάξουν τις αποφάσεις που παίρνουμε έτσι ώστε να αποκτήσουμε πλεονέκτημα στο μέλλον;
- Τι παρενέργειες θα έχουν οι αποφάσεις που θα λάβουμε για το ευρύτερο σύστημα;

Έχοντας μια τέτοια τεράστια και ισχυρή εργαλειοθήκη διαθέσιμη, η μόνη προϋπόθεση για τους αναλυτές είναι η δυνατότητα τους να πάρουν στα χέρια τους τα πραγματικά δεδομένα. Ως εκ τούτου, στη σύγχρονη διαδικασία λήψης αποφάσεων οι πολύπλοκες καταστάσεις που οδηγούν σε σημαντικές αποφάσεις βασίζονται όλο και περισσότερο σε αποδεικτικά στοιχεία και όχι σε αδικαιολόγητες εικασίες.

Στις προηγούμενες ενότητες παρουσιάσαμε μερικούς από τους σημαντικότερους δείκτες KPI που παρακολουθούνται συνήθως από τους εμπόρους, τους αναλυτές και τους υπεύθυνους έρευνας και στρατηγικού σχεδιασμού για τη διαφήμισης στα μέσα

[3] https://en.wikipedia.org/wiki/Prescriptive_analytics

ενημέρωσης και τη διείσδυση στην αγορά στον αγώνα που πραγματοποιούν για την αύξηση της διαθεσιμότητας και της αναγνωρισιμότητας ενός brand name.

Στο επόμενο Κεφάλαιο θα αναλύσουμε τις μεθόδους με τις οποίες πραγματοποιείται μία πρόβλεψη. Για να γίνει αυτό, θα επεξηγήσουμε θεμελιώδεις έννοιες όπως το Marketing Mix Modeling και θα αναλύσουμε τους τρόπους με τους οποίους συνδυάζονται τα μαθηματικά μοντέλα και τα δεδομένα των επιχειρήσεων και των μέσων ενημέρωσης για την καλύτερη πρόβλεψη σε ένα πραγματικό σενάριο.

Κεφάλαιο 2 : Ανάλυση Πρόβλεψης

2.1 Εισαγωγή (Predictive Analytics)

Στο προηγούμενο κεφάλαιο εξετάσαμε τον πιθανό αντίκτυπο που μπορούν να έχουν ξεχωριστά οι διάφοροι Business ή Media KPIs για τα μερίσματα όγκου ή αξίας μιας μάρκας μέσω απλών γραφικών. Σε αυτό το κεφάλαιο θα προχωρήσουμε ένα βήμα παραπέρα, προσπαθώντας να καταλάβουμε πώς μπορούν να συνυπάρξουν και να αλληλοεπιδρούν μεταξύ τους πολλές μεταβλητές παράλληλα, βοηθώντας συνολικά στη διαμόρφωση ενός τελικού αποτελέσματος σχετικά με τις πωλήσεις μιας μάρκας. Και αυτή η μελέτη θα πρέπει να αποτελεί την πιο κατάλληλη προσέγγιση σε ολόκληρο το πρόβλημα, καθώς στην πραγματική ζωή οι επιδόσεις μιας μάρκας δεν κινούνται μόνο σε μία διάσταση και δεν επηρεάζονται ξεχωριστά από έναν ή άλλο παράγοντα και μόνο. Η πραγματικότητα είναι λίγο πιο περίπλοκη από αυτή και υπάρχουν πάντα πολλοί ποσοτικοποιήσιμοι ή ποιοτικοί παράγοντες που μπορούν ταυτόχρονα να επηρεάσουν την απόδοση μιας μάρκας. Για να γίνει αυτό, σε αυτό το κεφάλαιο, πρέπει να κάνουμε μια βαθιά ανάλυση στην έννοια της μαθηματικής μοντελοποίησης και να συζητήσουμε ειδικότερα για τα οικονομετρικά μοντέλα.

2.2 Ιστορική Αναδρομή και τα Διαθέσιμα μοντέλα

Σύμφωνα με τον καθηγητή Neil Borden (1964), η τεχνική MMM (Marketing Mixing Modeling) αποτελείται από μία στατιστική ανάλυση των πωλήσεων και του μάρκετινγκ χρησιμοποιώντας αναδρομικά ιστορικά δεδομένα τα οποία οδηγούν στην εκτίμηση των επιπτώσεων των διαφόρων τακτικών μάρκετινγκ στις πωλήσεις. Στη συνέχεια, προβλέπει τις επιπτώσεις των μελλοντικών τακτικών μάρκετινγκ καθώς και τις τάσεις και τα πρότυπα συμπεριφοράς. Αξίζει βέβαια να αναφερθεί πως υπάρχει μια μεγάλη συζήτηση σχετικά με την αξιοπιστία της από τότε.

Τα τελευταία χρόνια έχουν γίνει εντατικές εργασίες και για τις κατανομές του προϋπολογισμού. Οι Bergera και Bechwatib (2001) προσφέρουν μια γενική προσέγγιση στην οργάνωση της κατανομής του προϋπολογισμού της διαφημιστικής προώθησης, μεγιστοποιώντας των ιδίων κεφαλαίων των πελατών. Όταν πρόκειται για τη βελτιστοποίηση της προσφοράς των μέσων μαζικής ενημέρωσης ή για την εξεύρεση του βέλτιστου συνδυασμού μέσων επικοινωνίας, πρέπει να γνωρίζουμε όχι

μόνο τον συνολικό προϋπολογισμό που χρειάζεται να δαπανηθεί σε μια συγκεκριμένη καμπάνια, αλλά και τον τρόπο με τον οποίο ο προϋπολογισμός αυτός μπορεί να χωριστεί σε κάθε διαθέσιμο μέσο ειδικότερα. Ωστόσο, τέτοιες βελτιστοποιήσεις είναι λίγο βαριές και διφορούμενες και κάθε φορά εξαρτώνται από τα μέσα ενημέρωσης και τους στόχους μάρκετινγκ (δηλαδή πωλήσεις όγκου, μεγιστοποίηση της ευαισθητοποίησης). Ωστόσο, προκειμένου να εξασφαλιστούν τα βέλτιστα επίπεδα προϋπολογισμού, πρέπει να επιτύχουμε τη βελτιστοποίηση της απόδοσης της επένδυσης χρημάτων.

Σήμερα παρότι έχουν παρέλθει περίπου 80 χρόνια από τότε που έγινε η πρώτη χρήση του όρου ΜΜΜ, το θέμα παραμένει ακόμα στο προσκήνιο. Πρόσφατα ερευνήθηκε, η αποτελεσματικότητα του κάθε στοιχείου που συμμετέχει στην τεχνική διαδικασία ΜΜΜ, όταν αυτό χρησιμοποιείται σε συνδυασμό με κάποιο άλλο στοιχείο αυτής , για τη παραγωγή αποτελεσμάτων στις αγορές.

Από τη μία πλευρά, τα οικονομετρικά μοντέλα έχουν αποδειχθεί ότι μπορούν να αντιμετωπίσουν ένα ευρύ φάσμα συγκεκριμένων ερωτημάτων. Ενώ από την άλλη πλευρά οι επιπτώσεις της διαφήμισης μπορούν να είναι βραχυπρόθεσμες, μεσοπρόθεσμες ή μακροπρόθεσμες. Σύμφωνα με τους Cook και Holmes, τα οικονομετρικά μοντέλα μπορούν να χρησιμοποιηθούν στους δύο πρώτους τύπους αποτελεσμάτων ενώ μπορούν να δημιουργηθούν διάφοροι τύποι μοντέλων σύμφωνα με το τελικό αποτέλεσμα:

- Συνολική αποτελεσματικότητα της επικοινωνίας.
- Συγκριτικές επιδράσεις της διαφημιστικής εκστρατείας για να προσδιοριστεί η πιο αποτελεσματική από αυτές με βάση τις πωλήσεις.
- Αποτελεσματικότητα όσον αφορά τα επίπεδα του προϋπολογισμού, το media-mix, τα επίπεδα συχνότητας, κ.λπ.
- Επίδραση cross branding σε περίπτωση εμφάνισης χαρτοφυλακίου προϊόντων. Η cross-branding είναι μια στρατηγική μάρκετινγκ που συνδυάζει δύο προσφορές από ξεχωριστές εταιρείες. Η τεχνική χρησιμοποιείται συνήθως για την πώληση συμπληρωματικών προϊόντων ή υπηρεσιών. Ονομάζεται επίσης πολλαπλή προώθηση ή cross merchandising.
- Ανταγωνιστικές επιπτώσεις στη μάρκα που μας αφορά ή κατά τον ανταγωνισμό.

- Επιπτώσεις εκτός πωλήσεων όπως συμφόρηση τηλεφωνικού κέντρου, awareness κ.τ.λ. Στο μάρκετινγκ, η έννοια awareness αφορά τη μέτρηση του πόσο γνωστή είναι μία μάρκα, μια επιχείρηση ή ένα προϊόν. Οι εταιρείες συνήθως θέτουν έναν στόχο για το βαθμό επίγνωσης που προτίθενται να επιτύχουν και στη συνέχεια σχεδιάζουν μια διαφημιστική εκστρατεία για την επίτευξη αυτού του στόχου.
- Δοκιμές μέσω των μέσων ενημέρωσης για τη διερεύνηση ζητημάτων που προκύπτουν λόγω αλλαγών στη στρατηγική της διαφήμισης.

2.3 Μαθηματική Προσέγγιση και Ανάπτυξη Κατάλληλου Μοντέλου

Η μαθηματική μέτρηση της αποτελεσματικότητας των στοιχείων της τεχνικής marketing-mix επιτυγχάνεται με το να δηλώνεται, ταυτόχρονα με τις πωλήσεις, μια σχέση ανάμεσα στις διάφορες δραστηριότητες μάρκετινγκ, με τη μορφή κατάλληλης γραμμικής ή μη γραμμικής παλινδρόμησης. Το MMM ορίζει την αποτελεσματικότητα καθενός από τα στοιχεία μάρκετινγκ, όσον αφορά τη συμβολή του στον όγκο των πωλήσεων (ή σε παρόμοιου ενδιαφέροντος και σημαντικότητας, επιχειρηματικό KPI). Αυτό μπορεί να επιτευχθεί με τη δημιουργία ενός μοντέλου με τον όγκο πωλήσεων ως εξαρτημένη μεταβλητή και ως ανεξάρτητες μεταβλητές εκείνες που δημιουργούνται από τις διάφορες προσπάθειες μάρκετινγκ.

Η φύση της εξαρτημένης μεταβλητής υπαγορεύει την επιλογή της κατάλληλης μαθηματικής προσέγγισης ως προς τη μοντελοποίηση. Τα μοντέλα πολλαπλής γραμμικής παλινδρόμησης λαμβάνονται υπόψη όταν ο μέσος όρος του επιχειρηματικού KPI που μας ενδιαφέρει εξαρτάται γραμμικά από ένα σύνολο μεταβλητών, τις covariates. Επίσης λαμβάνονται υπόψη για τις μετρήσεις KPI και τα μοντέλα παλινδρόμησης Poisson (π.χ. μηνιαίο αριθμό πακέτων που πωλούνται / παραγγέλλονται). Μια άλλη παρόμοια προσέγγιση αποτελούν οι Μέθοδοι της άριστης επιλογής των ανεξάρτητων μεταβλητών (Best Subsets Regression), καθώς και οι δύο διαδικασίες δημιουργούν μοντέλα από ένα σύνολο προκαθορισμένων προγνωστικών. Ωστόσο, αν και μπορεί να παρέχονται περισσότερες πληροφορίες από μια τυπική παλινδρόμηση, το να συμπεριλαμβάνουμε περισσότερα μοντέλα, είναι συνήθως μια πιο σύνθετη και λιγότερο αποτελεσματική επιλογή. Στην προσέγγισή μας, η απλότητα της μεθόδου που ακολουθείται αποτελεί βασικό κριτήριο, προκειμένου ένα εργαλείο

να υιοθετηθεί εύκολα και να χρησιμοποιηθεί παράλληλα από τις ομάδες της κάθε εταιρείας αλλά και τις αντίστοιχες των Media.

Τα στοιχεία που συνήθως προσμετρώνται στα ΜΜΜ περιλαμβάνουν, μεταξύ άλλων, τα μερίσματα του όγκου των πωλήσεων, τα μέσα ενημέρωσης και διαφήμισης, τις εμπορικές προωθήσεις, τη στρατηγική τιμολόγησης, τη διανομή, τις επιπτώσεις του ανταγωνισμού (δημιουργώντας τις αντίστοιχες μεταβλητές). Μόλις δημιουργηθεί το μοντέλο, εκτελούνται περαιτέρω επικυρώσεις είτε με τη χρήση ενός συνόλου δεδομένων επικύρωσης είτε με τη συνέπεια των αποτελεσμάτων της επιχείρησης.

Για τη μαθηματική μοντελοποίηση, η ισορροπία μεταξύ της χρήσης των αυτοματοποιημένων εργαλείων μοντελοποίησης που γίνονται σε μεγάλες σειρές δεδομένων, σε σχέση με την τεχνική προσέγγιση ενός οικονομολόγου, αποτελεί ένα ανοιχτό θέμα προς συζήτηση. Η μαθηματική μεθοδολογία και η συνολική προσέγγιση υπό την οποία καλείται να εργαστεί ένας αναλυτής, είναι θεμιτό να ακολουθεί το διάσημο απόσπασμα που αποδίδεται στον Albert Einstein, σύμφωνα με το οποίο «Όλα πρέπει να γίνουν όσο το δυνατόν απλούστερα, αλλά όχι απλά».

Μια ποικιλία μοντέλων μάρκετινγκ έχει χρησιμοποιηθεί στη βιβλιογραφία, με διάφορες προσεγγίσεις κατηγοριοποίησης, που προσφέρονται με βάση το σκοπό τους, το βαθμό αβεβαιότητας των μεταβλητών του μοντέλου, το επίπεδο των λεπτομερειών της συμπεριφοράς τους και το επίπεδο ανάλυσης. Τα μοντέλα υποστήριξης των αποφάσεων, τα οποία μπορούν να ταξινομηθούν ως μοντέλα πρόβλεψης (predictive models), αναπτύσσονται για να βοηθούν τους διαχειριστές να διαμορφώνουν καλύτερες αποφάσεις και μελέτες χρονολογικών σειρών και έχουν επανεξεταστεί εκτενώς στη βιβλιογραφία. Για τις μελέτες των περιπτώσεων που θα εξετάσουμε και από τις τρεις κατηγορίες, descriptive, predictive και prescriptive που χρησιμοποιούνται στη βιβλιογραφία για τη διάκριση των μοντέλων μάρκετινγκ, η προσέγγιση της μεθόδου Predictive Analytics, αναπτύσσεται κυρίως για να προβλέψει ή να εκτιμήσει τα μελλοντικά ποσοστά του μεριδίου αγοράς μιας μάρκας ή της διείσδυσης της στην αγορά μέσω πολλαπλής γραμμικής παλινδρόμησης. Η μοντελοποίηση του όγκου των δεδομένων που αφορούν τις πωλήσεις αντί των κυλιόμενου, μεταβλητού όγκου αποτελεί σημείο αναφοράς της μελέτης μας.

Ωστόσο, σε ένα ρεαλιστικό πρόβλημα, διαφορετικά και ανεξάρτητα μέρη πρέπει να ευθυγραμμίσουν τα δεδομένα που παρέχουν. Αυτό όμως δεν είναι πάντα εφικτό λόγω

ασυμβατότητας μεταξύ των παρόχων. Ως εκ τούτου, στην αρκετά τυπική και συνήθη περίπτωσή μας, η έλλειψη διαθέσιμων δεδομένων σε εβδομαδιαία βάση κατέστησε αναγκαία την αλλαγή της ανάλυσης σε μηνιαίο επίπεδο. Αυτό οδήγησε σε αρκετά φτωχά και εύθραυστα αποτελέσματα όσων προσπαθούσαν να διαμορφώσουν και να μοντελοποιήσουν τον όγκο των πωλήσεων και προκειμένου να εξομαλύνουν τα αποτελέσματα στράφηκαν στη μοντελοποίηση των κυλιόμενων όγκων των πωλήσεων (rolling volume sales). Όσον αφορά τη γραμμικότητα της δομής της παλινδρόμησης, αν και τα παραπάνω KPI μετρούνται σε ποσοστά, μια τέτοια προσέγγιση μπορεί να θεωρηθεί εύλογη λόγω του γεγονότος ότι σε όλες τις περιπτώσεις που εξετάστηκαν μέχρι στιγμής, όλες οι τιμές για τους δείκτες πρόβλεψης εμπίπτουν στο 20-70% , υποδεικνύοντας ότι ένα γραμμικό μοντέλο θα μπορούσε να χρησιμοποιηθεί επαρκώς για να περιγράψει το συγκεκριμένο σύνολο δεδομένων, ειδικά έχοντας ως γνώμονα τον αρχικό στόχο του να διαμορφωθεί ένα απλό, αλλά αποτελεσματικό εργαλείο.

Στη βιβλιογραφία υπάρχουν τρεις προσεγγίσεις για την ένταξη των μεταβλητών και εφαρμόζονται, εξ όσων γνωρίζουμε, στο μεγαλύτερο μέρος του στατιστικού λογισμικού:

I. Προοδευτική ή σταδιακή ένταξη των μεταβλητών (***forward selection***).
II. Προοδευτική ή σταδιακή απόρριψη των μεταβλητών (***backward eliminitaion***)
III. Αμφίδρομη επιλογή των μεταβλητών ή μέθοδος επιλογής βήμα προς βήμα (***step by step selection***).

Η τελευταία αποτελεί την πιο διαδεδομένη και συνήθη επιλογή, λόγω του μεγάλου αριθμού των διαθέσιμων πιθανών προγνωστικών. Ωστόσο, ένας αναλυτής θα μπορούσε πολύ προσεκτικά να προβεί σε χειρωνακτική διαχείριση των δυνατών αποτελεσμάτων των δεικτών πρόγνωσης στο τελευταίο βήμα της κάθε ανάλυσης, απλώς και μόνο για να επιτρέψει τη δημιουργία νέων εργαλείων που προσαρμόζονται ανάλογα με τις ξεχωριστές στρατηγικές ανάγκες της κάθε μάρκας.

Προηγουμένως αναφέραμε ότι ένα υποσύνολο των στοιχείων που θα μπορούσαν να συμπεριληφθούν σε μια προσέγγιση MMM θα μπορούσε να είναι ο αντίκτυπος του ανταγωνισμού. Ωστόσο, προτείνεται να πραγματοποιείται μία πρωταρχική προσπάθεια σε κάθε ανάλυση, ώστε να περιλαμβάνονται μόνο μεταβλητές που μπορούν να «επηρεαστούν» από τους υπεύθυνους λήψης των αποφάσεων της μάρκας. Με αυτόν τον τρόπο, το τελικό αποτέλεσμα του μηχανισμού μοντελοποίησης θα μπορούσε να επιτρέψει την εύκολη δημιουργία διαφόρων εφικτών και

ρεαλιστικών σεναρίων "what-if" και τον υπολογισμό των επιπτώσεών τους στο εξαρτώμενο KPI.

Στην ιδανική περίπτωση, αναμένεται ότι μια ανάλυση θα καταλήξει να καλύπτει τρεις διακριτές περιοχές:

1. Κατανόηση του τι έχει συμβεί μέχρι τώρα όσον αφορά την ιστορία της εξαρτημένης μεταβλητής.
2. Κατοχή μιας μεθόδου διερεύνησης του τρόπου με τον οποίο θα επιτευχθούν οι μελλοντικοί στόχοι.
3. Δυνατότητα διεξαγωγής μιας ποικιλίας δοκιμών τύπου "what-if".

2.4 Συνδυάζοντας τα Δεδομένα για Καλύτερη Πρόβλεψη

Όταν το τελικό μοντέλο είναι έτοιμο, τα αποτελέσματα μπορούν να χρησιμοποιηθούν για τη δημιουργία μιας σειράς σεναρίων μάρκετινγκ για μια ανάλυση "What-if". Η ομάδα της εταιρείας και οι διευθυντές μάρκετινγκ μπορούν να ανακατανέμουν τους προϋπολογισμούς του μάρκετινγκ σε διαφορετικές αναλογίες και να εξετάσουν τον άμεσο αντίκτυπο τους στις πωλήσεις. Επιπλέον, οι ομάδες της εταιρείας και των μέσων ενημέρωσης μπορούν να ανακατανέμουν τις μετρικές των media σε διαφορετικές αναλογίες και επίσης να παρακολουθήσουν τον άμεσο εκτιμώμενο αντίκτυπο στις πωλήσεις.

2.4.1. Προτεινόμενη Μεθοδολογική Προσέγγιση

Σε αυτή την ενότητα παρουσιάζουμε μια σύντομη περίληψη των απαραίτητων βημάτων που ενέχονται στη μεθοδολογία μοντελοποίησης. Ακολουθεί και μια σχηματική αναπαράσταση των απαραίτητων βασικών βημάτων Σχήμα 2.1.

Βήμα 0: Σχεδιασμός και οργάνωση της μελέτης project

Το αρχικό βήμα κάθε προσέγγισης οικονομετρικής μοντελοποίησης σχετίζεται πάντοτε με τον προσεκτικό σχεδιασμό του πειράματος και τα αναμενόμενα αποτελέσματα που θα προκύψουν από μια τέτοια ανάλυση. Οι ερωτήσεις στις οποίες παρέχονται όλες οι απαντήσεις στο τέλος της ανάλυσης πρέπει να διατυπώνονται

σαφώς και εξ ’αρχής. Θα πρέπει επιπλέον να δημιουργηθούν διαφορετικά μοντέλα για βραχυπρόθεσμους και μακροπρόθεσμους βασικούς στόχους. Για παράδειγμα αν κάποιος έπρεπε να κατανοήσει την αποτελεσματικότητα μιας συγκεκριμένης δραστηριότητας προώθησης ενός προϊόντος θα λάμβανε υπ’ όψην του διαφορετικούς παράγοντες από ότι στην περίπτωση που η μακροπρόθεσμη πολιτική τιμολόγησης θα ήταν στο προσκήνιο. Όπου θα ενεργούσε πολύ διαφορετικά.

Η οικονομετρία μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση μιας πληθώρας KPI μέσω της κατάλληλης μαθηματικής τεχνικής και δεν θα πρέπει να περιορίζεται μόνο σε μετρήσεις όγκων ή πωλήσεων προϊόντων. Σύμφωνα με τους Cook και Holmes (2017), αυτοί οι δείκτες KPI θα μπορούσαν να περιλαμβάνουν μεταξύ άλλων:

- *Αρχικό ενδιαφέρον* (γενική πληροφόρηση επί του αντικειμένου, ερωτήσεις, δραστηριότητα του τηλεφωνικού κέντρου, αναζήτηση, επισκέψεις σε ιστότοπους, awareness, consideration)
- Απόκριση στο συμπεριφορικό προφίλ των καταναλωτών (μερίδιο, διείσδυση, συναλλαγές, έσοδα από διαφημίσεις)
- Πιθανότητα διατήρησης (ανανέωση κάρτας συνδρομής)
- Διατήρηση της αξιοπιστίας της μάρκας (trust, value for money)

Ένα εξίσου σημαντικό θέμα που πρέπει να έχουμε κατά νου είναι ότι ο αναλυτής θα πρέπει κάθε φορά να έχει μια σαφή και πλήρη εικόνα του πρακτικού και επικοινωνιακού πλαισίου υπό του οποίου πρόκειται να διεξαχθεί μια ανάλυση. Οι ομάδες των αναλυτών τόσο από την ίδια την εταιρεία της μάρκας που εκπροσωπείται, όσο και των μέσων ενημέρωσης και κοινωνικής δικτύωσης, οφείλουν να συνεργαστούν προς την κατεύθυνση αυτή. Ακόμη και μια απλή συνάντηση τύπου brainstorming, θα μπορούσε να βοηθήσει στην σωστή ρύθμιση του πλαισίου. Επίσης, συνιστάται η συμμετοχή όλων των ενδιαφερόμενων μερών σε τέτοιες συνεδρίες, έτσι ώστε όλοι να βρίσκονται στην ίδια νόρμα σχετικά με τις προοπτικές και τις προσδοκίες τους.

Βήμα 1: Συλλογή δεδομένων

Αυτό το βήμα ενεργοποιεί την όλη διαδικασία και είναι εξαιρετικά κρίσιμο για όλη την επιτυχία της ανάλυσης. Μόλις αποφασίσουμε ποιες είναι οι μετρικές (Business και Media) που έχουν τη μεγαλύτερη σημασία, αρχίζουμε να συλλέγουμε όλα τα

αναδρομικά δεδομένα που σχετίζονται με αυτές. Τα επιχειρησιακά δεδομένα προέρχονται συνήθως από τρίτους που ενεργούν ως σύμβουλοι των επιχειρήσεων και παρέχουν τις ανάλογες υπηρεσίες (οι GfK, Nielsen, IRI, eT.c. είναι οι συνήθη πάροχοι δεδομένων). Κατά αντιστοιχία , τα δεδομένα τύπου Media προέρχονται από τις αντίστοιχες εταιρείες μέτρησης αυτών. Τα οικονομετρικά μοντέλα χρησιμοποιούν τέτοια αναδρομικά δεδομένα για να προσδιορίσουν τις σχέσεις μεταξύ των διάφορων KPI (σε οποιαδήποτε κατεύθυνση Business ή Media) με άλλους ανεξάρτητους παράγοντες όπως η εποχικότητα, ο καιρός, η πολιτική κατάσταση κ.α..

Η συλλογή δεδομένων είναι μια εξαιρετικά χρονοβόρα, επώδυνη και κουραστική διαδικασία. Επιπλέον, στις περισσότερες περιπτώσεις ζητείται να διεξαχθεί από ανθρώπους που είτε είναι αδιάφοροι είτε είναι πολύ απασχολημένοι για να κάνουν τον κόπο και να το επιτύχουν. Το μόνο πρόβλημα με αυτό είναι ο γνωστός και απλός κανόνας GIGO (Garbage in – garbage out). Επομένως, εάν τα δεδομένα περιέχουν πολλά λάθη, τότε η αξιοπιστία των αποτελεσμάτων θα είναι υπό αμφισβήτηση και τα συμπεράσματα θα μπορούσαν φυσικά να οδηγήσουν σε εσφαλμένες γνώσεις. Να σημειωθεί ότι η απόκτηση λανθασμένων δεδομένων δεν είναι κάτι που εμφανίζεται σπάνια (ακόμη και μια μόνο στήλη μπορεί να αντιγραφεί-επικολληθεί λανθασμένα). Τελικά, προκειμένου να εξασφαλιστεί η επιτυχής παράδοση ενός οικονομετρικού μοντέλου, όλα τα μέρη, συμπεριλαμβανομένου και του αναλυτή, των ομάδων της εταιρείας και των μέσων ενημέρωσης, έχουν τις ευθύνες που τους αναλογούν για κάθε ανάλυση_project (κατά τη διάρκεια μιας σύντομης ενημέρωσης, σχετικά με την παροχή δεδομένων, την παροχή πληροφοριών, την ερμηνεία των αποτελεσμάτων).

<u>Βήμα 2: Διαχείριση δεδομένων - Περιγραφική στατιστική</u>

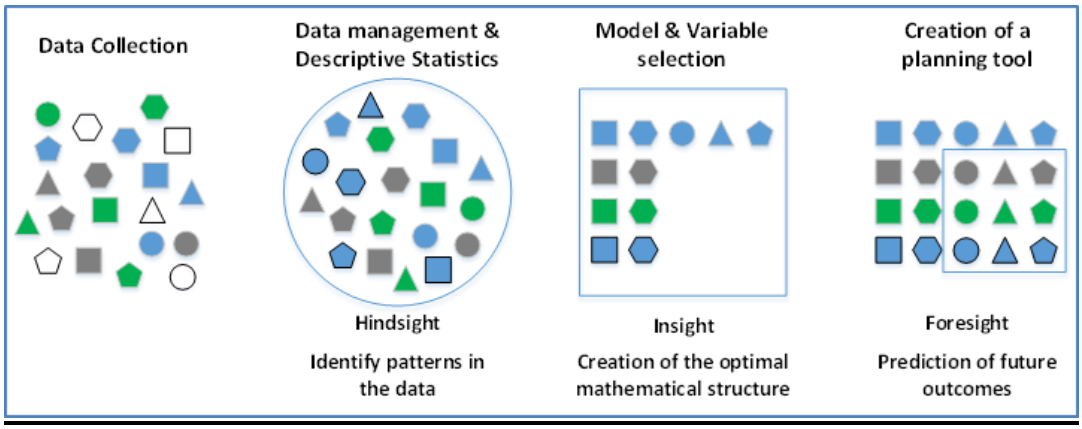
Σε αυτό το βήμα γίνεται η αναγνώριση των μοτίβων που υπάρχουν στα δεδομένα και λειτουργεί ως προϋπόθεση για την κατασκευή ενός προγνωστικού μοντέλου. Σε αυτό το βήμα ο αναλυτής:

- ***Ελέγχει τις συσχετίσεις (correlations) μεταξύ των διαφόρων KPIs, Media και Business.***

  Η συσχέτιση είναι ένα στατιστικό μέτρο που υπολογίζει τον βαθμό στον οποίο δύο μεταβλητές επηρεάζονται μεταξύ τους κατά την εξέλιξη των τιμών τους. Υπολογίζεται ότι ένας συντελεστής συσχέτισης (correlation coefficient) ,

παίρνει τιμές μεταξύ -1 και 1. Όταν η τιμή του συντελεστή συσχέτισης βρίσκεται γύρω στο ± 1, τότε λέμε ότι είναι ένας τέλειος βαθμός συσχέτισης μεταξύ των δύο μεταβλητών. Μια θετική συσχέτιση (ο συντελεστής συσχέτισης είναι περίπου 1) δείχνει ότι οι δύο μεταβλητές κινούνται προς την ίδια κατεύθυνση, με την έννοια από ό, τι για κάθε θετική αύξηση της μίας μεταβλητής, υπάρχει μια αύξηση για την άλλη μεταβλητή. Αντίστοιχα, ένας αρνητικός συσχετισμός σημαίνει ότι οι δύο μεταβλητές κινούνται σε αντίθετες κατευθύνσεις, ενώ μια μηδενική συσχέτιση υποδηλώνει πως δεν υπάρχει σχέση μεταξύ των μεταβλητών. Στη στατιστική, μετρούνται τρεις τύποι συσχετίσεων: οι βαθμοί Pearson, Kendall και Spearman, με τον πρώτο να χρησιμοποιείται ευρέως για να μετρηθεί ο βαθμός της σχέσης μεταξύ γραμμικών συναφών μεταβλητών.

- ***Ελέγχει τους μέσους, διάμεσους, διακυμάνσεις, ακραίες τιμές για κάθε δυνητικό παράγοντα πρόγνωσης που μπορεί να συμπεριληφθεί στο τελικό μοντέλο.***
- ***Εκτελεί απλή ανάλυση παλινδρόμησης και αποδίδει σε διαγράμματα κάθε δυνητικό παράγοντα πρόγνωσης που πρέπει να συμπεριληφθεί στο τελικό μοντέλο.***



Σχήμα 2.1 Κύρια βήματα ως γενική μεθοδολογία, που περιλαμβάνονται σε ένα πρόγραμμα μοντελοποίησης δεδομένων

<u>Βήμα 3: Επιλογή μοντέλου</u>

Σε αυτό το βήμα εντοπίζεται η βέλτιστη μαθηματική δομή για την καλύτερη προσαρμογή και περιγραφή της εξαρτημένης μεταβλητής. Σε αυτό το σημείο ο αναλυτής:

- Ελέγχει διάφορα προγράμματα μοντελοποίησης για να προσδιορίσει το καταλληλότερο για την περιγραφή του συγκεκριμένου συνόλου των δεδομένων. Έτσι αναπόφευκτα δοκιμάζονται πολλοί διαφορετικοί τύποι παλινδρόμησης για τη βέλτιστη προσαρμογή του μοντέλου.
- Αντιμετωπίζει ασυνήθιστες παρατηρήσεις. Οι απομονωμένες τιμές (outliers) και οι ακραίες τιμές (extremes) είναι παρατηρήσεις που απέχουν πολύ από τις υπόλοιπες τιμές. Πρόκειται για παρατηρήσεις που δεν είναι τυπικές των υπόλοιπων στοιχείων, συνεπώς αποτελούν ιδιομορφία και υποδεικνύουν σημεία που δεν είναι αντιπροσωπευτικά του πληθυσμού, απ' όπου προέρχονται τα δεδομένα.

Βήμα 4: Επιλογή της μεταβλητής

Σε αυτό το βήμα παίρνουμε την απόφαση σχετικά με τους στατιστικά σημαντικούς και πιθανούς προγνωστικούς παράγοντες που πρέπει να συμπεριληφθούν στο τελικό μοντέλο. Η προοδευτική ή σταδιακή ένταξη των μεταβλητών (forward selection), η προοδευτική ή σταδιακή απόρριψη των μεταβλητών (backward eliminitaion, η αμφίδρομη επιλογή των μεταβλητών ή μέθοδος επιλογής βήμα προς βήμα (step by step selection) είναι οι τρεις διαθέσιμες μέθοδοι επιλογής (θα τις αναλύσουμε διεξοδικά στο Κεφάλαιο 4), οι οποίες στη συνέχεια θα μπορούσαν να χρησιμοποιηθούν για να αποφασιστούν οι σημαντικές μεταβλητές.

Πολλές φορές, χρησιμοποιείται η μέθοδος επιλογής μοντέλου, βήμα προς βήμα. Η επιλογή αυτή οφείλεται κυρίως σε ένα μεγάλο σύνολο διαθέσιμων δυνητικών προγνωστικών που σχετίζονται με τη σωματική και διανοητική διαθεσιμότητα: διανομή, πολιτική τιμολόγησης, τηλεοπτικά δεδομένα για την περίοδο ανάλυσης για τη μάρκα που μας αφορά και τους ανταγωνιστές, άλλα KPI κατηγορίας Media (π.χ. Ψηφιακά, Ραδιόφωνο, Τύπος) .

Βήμα 5: Έλεγχος της εγκυρότητας του μοντέλου

Αυτό είναι το τελικό βήμα της διαδικασίας και περιλαμβάνει την επιλογή του μοντέλου. Στη βιβλιογραφία υπάρχει μία πληθώρα τεχνικών επικύρωσης του κατάλληλου μοντέλου και παρακάτω, περιγράφονται εν συντομία οι μέθοδοι που χρησιμοποιούνται σε κάθε ανάλυση.

- Στατιστική σημαντικότητα

  Η τυπική διαδικασία είναι να εξασφαλιστεί η στατιστική σημαντικότητα για όλες τις μεταβλητές που περιλαμβάνονται στο τελικό προτεινόμενο μοντέλο μέσω της διαδικασίας της επιλογής της μεταβλητής. Έτσι υπολογίζεται η πιθανότητα ένα στατιστικό αποτέλεσμα να είναι τόσο μεγάλο ή μεγαλύτερο από το παρατηρούμενο και να έχει προκύψει λόγω τυχαίων παραγόντων και όχι λόγω της σχέσης μεταξύ των δύο μεταβλητών. Η πιθανότητα αυτή ονομάζεται «παρατηρούμενο επίπεδο ή στάθμη σημαντικότητας», $p - value$ των στατιστικών ελέγχων.

  Η στατιστική σημαντικότητα (ή ένα στατιστικά σημαντικό αποτέλεσμα) επιτυγχάνεται όταν το $p - value$ είναι μικρότερο από το επίπεδο σημαντικότητας. Η $p - value$ είναι η πιθανότητα απόκτησης ενός αποτελέσματος ακραίας τιμής, δεδομένου ότι η μηδενική υπόθεση (Κεφάλαιο 3) είναι αληθής, ενώ η σημαντικότητα ή επίπεδο σημαντικότητας άλφα (α) είναι η πιθανότητα της απόρριψης της μηδενικής υπόθεσης, δεδομένου ότι είναι αλήθεια. Ως θέμα της ορθής επιστημονικής μεθόδου-πρακτικής, ένα επίπεδο σημαντικότητας επιλέγεται πριν τη συλλογή δεδομένων και συνήθως βρίσκεται στο $0{,}05$ (5%). Άλλα επίπεδα σημαντικότητας (π.χ., $0{,}01$) δύνανται να βρουν εφαρμογή, ανάλογα με τον τομέα μελέτης. Τελικώς, εξασφαλίζουμε ότι ισχύει $p \leq 0{,}05$ (απορρίπτουμε την H_0 και δεχόμαστε την H_1,) κατά τη διαδικασία του ελέγχου των υποθέσεων, που θα αναλυθεί εκτενώς στο Κεφάλαιο 3 . Ωστόσο, λόγω της ανάγκης να συμπεριλαμβάνονται σε ορισμένα μοντέλα, συγκεκριμένοι ανεξάρτητοι παράγοντες πρόβλεψης και να παρακολουθούνται για στρατηγικούς σκοπούς, ορισμένες φορές, η εν λόγω κατευθυντήρια γραμμή μπορεί να μην είναι και τόσο αυστηρή.

- **$R^2, R^2 - adjusted, R^2 - predicted$**

Το R^2 είναι ένας στατιστικός όρος που μετρά πόσο κοντά «πέφτουν» τα δεδομένα στην προσαρμοσμένη γραμμή παλινδρόμησης και καθορίζεται με έναν αρκετά ευθύ τρόπο: είναι το συνολικό ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που μπορεί να εξηγηθεί από το προσαρμοσμένο γραμμικό μοντέλο. Οι τιμές R^2 κυμαίνονται πάντα μεταξύ 0% και 100%, με το 0% να δείχνει ότι το μοντέλο είναι ανεπαρκές να εξηγήσει οποιεσδήποτε από τις διακυμάνσεις της εξαρτημένης μεταβλητής. Αντίστοιχα, το 100% δείχνει ότι το μοντέλο εξηγεί όλη τη μεταβλητότητα της εξαρτημένης μεταβλητής. Συνήθως, όσο μεγαλύτερο είναι το R^2, τόσο καλύτερα το μοντέλο ταιριάζει με τα δεδομένα. Ωστόσο, μια τέτοια κατευθυντήρια γραμμή δεν είναι πανάκεια και υπάρχουν σημαντικές προϋποθέσεις για αυτό. Επομένως, απαιτείται προσοχή όταν οι τιμές R^2 ερμηνεύονται και θα πρέπει πάντα να αξιολογούνται κατά τη διάρκεια γραφικών δοκιμών.

Γενικά, το R^2 παρέχει μια εκτίμηση της ισχύος της σχέσης μεταξύ του προτεινόμενου μοντέλου και της εξαρτώμενης μεταβλητής. Τρεις μορφές του όρου απαντώνται συνήθως στη βιβλιογραφία: R^2, R^2_{adj} προσαρμοσμένος συντελεστής $(R^2 - adjusted)$ και R^2_p συντελεστή πρόβλεψης $(R^2 - predicted)$. Η μεγιστοποίηση αυτών των στατιστικών είναι κρίσιμη για την πρότυπη μοντελοποίηση. Τιμές μεγαλύτερες του 75% ακόμα και για την R^2_p θεωρούνται ως γενικός κανόνας και ως εκ τούτου αποτελούν σημείο αναφοράς για κάθε προτεινόμενο μοντέλο.

- Τυπικό σφάλμα (**S**)

Το τυπικό σφάλμα (S) της παλινδρόμησης αντιπροσωπεύει τη μέση απόσταση των παρατηρούμενων τιμών από τη γραμμή παλινδρόμησης και παρέχει ένα συνολικό μέτρο για το πόσο καλά προσαρμόζεται το προτεινόμενο μοντέλο στα δεδομένα. Το τυπικό σφάλμα επιτρέπει την πρακτική διαισθητικότητα της χρήσης των φυσικών μονάδων της μεταβλητής απόκρισης.

- Γραφικές παραδοχές

Σε κάθε προτεινόμενο μοντέλο σύμφωνα με τη βιβλιογραφία υπάρχουν τέσσερα είδη υπολειμματικών γραφημάτων, ισχύουν για όλο το στατιστικό λογισμικό και συνήθως εφαρμόζονται τόσο για να εξεταστεί η καλή

προσαρμοστικότητα στην παλινδρόμηση όσο και για να καθοριστεί εάν πληρούνται οι παραδοχές της γραμμικότητας.

- Διαγνωστικός Έλεγχος των Durbin–Watson

Το στατιστικό στοιχείο Durbin Watson είναι ένας μοναδικός αριθμός με τιμές μεταξύ 0 και 4. Αυτό το στατιστικό στοιχείο χρησιμοποιείται για την ανίχνευση της αυτοσυσχέτισης στα υπολείμματα. Μια τιμή 2 υποδεικνύει την απουσία αυτοσυσχέτισης στο δείγμα. Οι τιμές από 0 έως 2 δείχνουν θετική αυτοσυσχέτιση και οι τιμές από 2 έως 4 δείχνουν αρνητική αυτοσυσχέτιση. Αυτό το στατιστικό στοιχείο πρέπει πάντα να παρακολουθείται και κατά κανόνα, συνιστάται η τιμή του στατιστικού δείγματος να μην είναι εκτός της κλίμακας $1,5 - 2,5$. Η ύπαρξη ενός αποτελέσματος εντός αυτού του εύρους συνήθως υποδεικνύει σχετικά κανονικές τιμές.

- Εξέταση της πολυσυγγραμμικότητας (multicollinearity)

Η multicollinearity είναι γνωστή και με τα ονόματα ενδοσυσχέτιση και συγγραμμικότητα μεταξύ των ανεξάρτητων μεταβλητών (θα αναλυθεί εκτενώς στην ενότητα 4.8.9.). Η πολυσυγγραμμικότητα προκύπτει όταν τουλάχιστον δύο δείκτες συσχέτισης αξιολογούνται ταυτόχρονα σε ένα μοντέλο παλινδρόμησης. Στη βιβλιογραφία επισημαίνεται πως το να αγνοηθεί ο έλεγχος της πολυσυγγραμμικότητας σε μια ανάλυση παλινδρόμησης μπορεί να έχει σοβαρό αντίκτυπο, με πιο σύνηθες τις παραπλανητικές ερμηνείες των αποτελεσμάτων. Ένα μέτρο του βαθμού της πολυσυγγραμμικότητας αποτελεί ο συντελεστή διογκωμένης διακύμανσης της παλινδρόμησης VIF (Variance Inflation Factor) για κάθε ανεξάρτητη μεταβλητή (Ενότητα 4.8.10). Στη βιβλιογραφία αναφέρεται πως υπάρχει ακόμα και σήμερα ένας ανοιχτός διάλογος σχετικά με το ποιες τιμές του V.I.F. και ποιοι κανόνες σχετικά με αυτόν, αποτελούν σοβαρό κριτήριο πολυσυγγραμμικότητας. Όταν οι τιμές του V.I.F. υποδεικνύουν πολυσυγγραμμικότητα, η συνήθης τακτική είναι να περιοριστούν μία ή περισσότερες μεταβλητές από την ανάλυση, ή να συνδυαστούν δύο ή και περισσότερες μεταβλητές και να χρησιμοποιηθούν ως μία στην ανάλυση. Αυτές όμως οι επιλογές, μπορούν να δημιουργήσουν περισσότερα προβλήματα από όσα τελικά λύνουν. Οι τιμές αναφοράς του V.I.F. πρέπει να αξιολογηθούν στο πλαίσιο πολλών άλλων παραγόντων που

επηρεάζουν τη διακύμανση των συντελεστών παλινδρόμησης. Έτσι τιμές $10, 20, 40$ ή ακόμη και υψηλότερες δεν υποδεικνύουν από μόνες τους την απόκλιση-εξάλειψη μιας ή περισσοτέρων ανεξάρτητων μεταβλητών από την ανάλυση ούτε απαιτούν κατ' ανάγκη το συνδυασμό αυτών σε έναν ενιαίο δείκτη. Γενικά, είναι πιο βολικό, οι τιμές του να κυμαίνονται κάτω του 10. Υπάρχουν όμως και περιπτώσεις που μία μεταβλητή παραμένει στην ανάλυση για παράδειγμα για λόγους στρατηγικής, ακόμα και αν ο V.I.F είναι μεγαλύτερος του 10.

- Συνολική εξομάλυνση καμπύλης του μοντέλου (Overall Goodness of model fit)

  Το $F - Test$ χρησιμοποιείται για τη συνολική εξομάλυνση της καμπύλης του μοντέλου. Αυτή η δοκιμή παρέχει μία επίσημη υπόθεση της ισχύος της σχέσης μεταξύ του προτεινόμενου μοντέλου και της εξαρτημένης μεταβλητής. Μια τιμή $p - value < 0{,}05$ πρέπει να εξασφαλίζεται για κάθε προτεινόμενο μοντέλο, εξασφαλίζοντας έτσι την εξομάλυνση του μοντέλου.

- Μέσο απόλυτο ποσοστιαίο σφάλμα, Mean Absolute Percentage Error (M.A.P.E.)

  Το M.A.P.E. είναι ένα μέτρο ελέγχου της ακρίβειας, ως προς την πρόβλεψη, του κάθε προγνωστικού μοντέλου. Αντιπροσωπεύει τη διαφορά μεταξύ της πραγματικής και της προβλεπόμενης τιμής μιας συγκεκριμένης μονάδας δεδομένων. Σε παρόμοιο πλαίσιο, ένας αναλυτής θα μπορούσε εναλλακτικά να μετρήσει τα στατιστικά στοιχεία MAD (Median Absolute Deviation) και MSD (Mean Squared Deviation). Η μέση απόλυτη απόκλιση (MAD) ενός συνόλου δεδομένων είναι η μέση απόσταση μεταξύ κάθε τιμής δεδομένων και του μέσου όρου. Η μέση απόλυτη απόκλιση είναι ένας τρόπος για να περιγράψουμε τη μεταβολή σε ένα σύνολο δεδομένων.

  Στα στατιστική, το μέσο τετραγωνικό σφάλμα (MSE) ή η μέση τετραγωνική απόκλιση (MSD) μιας εκτιμήτριας μετρά τον μέσο όρο των τετραγώνων των σφαλμάτων, δηλαδή τη μέση τετραγωνική διαφορά μεταξύ των εκτιμώμενων τιμών και των παρατηρήσεων. Ωστόσο στη βιβλιογραφία, το M.A.P.E. είναι το πλέον προτιμώμενο λόγω της απλότητάς του και ιδιαίτερα όταν από τους υπολογισμούς προκύπτουν μικρές τιμές, που δείχνουν καλύτερη προσαρμογή. Οι αποδεκτές τιμές βρίσκονται εντός της κλίμακας $1{,}5 - 2{,}5$.

Είναι γεγονός πως αυτά τα στατιστικά μέτρα δε δίνουν πολλές πληροφορίες από μόνα τους, μπορούν όμως να χρησιμοποιηθούν για να συγκρίνουμε τα αποτελέσματα που προκύπτουν από διαφορετικές μεθόδους.

- Ασυνήθιστες παρατηρήσεις

  Ο τελευταίος και εξίσου σημαντικός έλεγχος είναι αυτός των ασυνήθιστων παρατηρήσεων, οι οποίες πρέπει να ταυτοποιηθούν. Έπειτα ο αναλυτής πρέπει να προσδιορίσει τις επιρροές αυτών καθώς και τις αποκλίσεις ή τις δράσεις που σχηματίζονται μετά την αφαίρεσή τους ή όχι. Για να προσδιοριστεί το πιθανό αποτέλεσμα μιας ασυνήθιστης παρατήρησης, το μοντέλο θα προσαρμοστεί με δύο τρόπους, αρχικά συμπεριλαμβανομένης αυτής και έπειτα χωρίς αυτήν, έτσι ώστε οι παράμετροι να μπορούν απ' ευθείας να συγκριθούν μεταξύ των δύο αυτών μοντέλων.

Η προαναφερθείσα ανάλυση μπορεί να διεξαχθεί ακόμα και από αναλυτές δεδομένων με περιορισμένη εμπειρία στον προγραμματισμό, με τη βοήθεια διαφόρων, αρκετά φιλικών προς το χρήστη, στατιστικών λογισμικών. Ωστόσο, η συγκεκριμένη προσέγγιση μοντελοποίησης που προτείνεται, είναι ουσιώδης και μπορεί να πραγματοποιηθεί μέσω όλων των εφαρμοστέων στατιστικών λογισμικών που διατίθενται στην αγορά ανεξάρτητα από το αν πρόκειται για εμπορεύσιμα προϊόντα ή για λογισμικό ελεύθερης χρήσης. Ως εκ τούτου, λογισμικά όπως τα Minitab, SPSS, STATA, R, SAS κ.ά. μπορεί να λειτουργήσουν ως ιδανικές επιλογές για την εκτέλεση τέτοιων αναλύσεων.

Στα επόμενα κεφάλαια, ακολουθεί μία μαθηματική ανάλυση για το πώς προσεγγίζουμε και οργανώνουμε τα δεδομένα μας μέσω ανάλυσης παλινδρόμησης καθώς και οι μαθηματικές σχέσεις υπολογισμού των στατιστικών δεικτών που προαναφέρθηκαν.

2.5 Οικονομετρία και η Στρατηγική των Πολυεθνικών

Σε μια προσπάθεια να βελτιστοποιήσουν τα ΜΜΜ, οι αναλυτές χρησιμοποιούν συχνά μια ποικιλία εργαλείων και τεχνικών, με πιο διαδεδομένη και ευρέως πιο αποδεκτή αυτή της οικονομετρίας. Φυσικά, δεν πρόκειται για μία πανάκεια τεχνική που μπορεί να δώσει απαντήσεις σε κάθε ερώτημα μάρκετινγκ. Και ας μην ξεχνάμε ότι η διαφήμιση δεν έχει μόνο ποσοτικούς στόχους. Ωστόσο, η οικονομετρία θα

μπορούσε να εφαρμοστεί σε καταστάσεις δυσλειτουργίας των διαφημιστικών ή των επιχειρησιακών δεικτών KPIs. Βασικό εργαλείο για την οικονομετρία είναι το μοντέλο πολλαπλής γραμμικής παλινδρόμησης (Κεφάλαιο 4). Η οικονομετρική θεωρία χρησιμοποιεί στατιστική θεωρία και μαθηματικά στατιστικά για την αξιολόγηση και την ανάπτυξη οικονομετρικών μεθόδων. Οι αναλυτές προσπαθούν να βρουν εκτιμητές που έχουν επιθυμητές στατιστικές ιδιότητες, όπως η αμεροληψία, η αποδοτικότητα και η συνέπεια. Η εφαρμοσμένη οικονομετρία χρησιμοποιεί τη θεωρητική οικονομετρία και δεδομένα του πραγματικού κόσμου για την αξιολόγηση οικονομικών θεωριών, την ανάπτυξη οικονομετρικών μοντέλων, την ανάλυση της οικονομικής ιστορίας και την πρόβλεψη. Η εφαρμογή οικονομετρικών μοντέλων προκειμένου να εκτιμηθεί ο αντίκτυπος των διαφόρων τακτικών μάρκετινγκ στις πωλήσεις, που θα μπορούσε τελικά να οδηγήσει σε καλύτερη πρόβλεψη των τάσεων και των προτύπων συμπεριφοράς, δεν είναι καινούρια. Ως εκ τούτου, οι μεγάλες πολυεθνικές εταιρείες ενδιαφέρονται συνήθως για την εφαρμογή μιας ποικιλίας τέτοιων μοντέλων για τη μέγιστη αποτελεσματικότητα κάθε τεχνικής MMM.

Γι’ αυτό το λόγο έχουν στρατούς επιστημόνων στο τμήμα Έρευνας και Ανάπτυξης (R & D), οι οποίοι διαδραματίζουν το ρόλο των σύγχρονων προφητών, προσπαθώντας να εκτιμήσουν με τον καλύτερο δυνατό τρόπο ποια θα είναι τα μερίσματα αγοράς της μάρκας που αντιπροσωπεύουν και συνεπώς τι θα χρειαστεί να γίνει διαφορετικά προκειμένου να βελτιστοποιηθούν αυτές οι προβλέψεις. Με τα χρόνια τείνουν να έχουν αναπτύξει και εφαρμόσει πρωτόκολλα, μεθοδολογίες και κατευθυντήριες γραμμές («πιλότους»), προκειμένου να δημιουργήσουν βέλτιστα μαθηματικά μοντέλα σύμφωνα με τις συνήθεις ανάγκες τους. Καθώς δεν χρειάζεται να ανακαλύπτουμε κάθε φορά την Αμερική και να δημιουργούμε συμβουλές ή μεθοδολογίες (dos και don'ts) από την αρχή, παρουσιάζεται στη συνέχεια ένα παράδειγμα για το τι θα μπορούσαν να είναι αυτές οι οδηγίες και τα πρωτόκολλα.

I. Προδιαγραφές μοντέλου

Για να είναι ένα οικονομετρικό μοντέλο πιθανό προς εφαρμογή, πρέπει να *περιλαμβάνει όλο το φάσμα των εμπορικών δραστηριοτήτων και των δεδομένων των μέσων μαζικής ενημέρωσης*. Τα μοντέλα που χρησιμοποιούν ως δεδομένα μόνο τα πρώτα θα πρέπει να αποφεύγονται καθώς ερμηνεύουν το πρόβλημα κατά το ήμισυ. Αν προσπαθήσουμε να κατανοήσουμε την αιτιώδη συνάφεια των δεδομένων και των

προβλέψεων, τότε πρέπει να συμπεριληφθούν, όπου είναι δυνατόν, και οι ανταγωνιστικές δραστηριότητες. Ωστόσο, στα μοντέλα πρέπει γενικά να αποφεύγεται να συμπεριλαμβάνονται μεταβλητές που σχετίζονται με τον ανταγωνισμό, όταν ο στόχος είναι αυτά να είναι σε θέση να επηρεάσουν τις τιμές οποιωνδήποτε μεταβλητών του τελικού μοντέλου. Ένα μοντέλο παλινδρόμησης που δεν έχει οριστεί πλήρως μπορεί να οδηγήσει σε παραπλανητικούς συντελεστές για τις αιτιώδεις μεταβλητές που περιλαμβάνονται στο μοντέλο. Και αυτό έπειτα θα οδηγήσει σε παραπλανητικά επίπεδα εισφορών.

II. Μεταβλητές που πρέπει να συμπεριληφθούν

Οι τιμές των μεταβλητών στα μοντέλα παλινδρόμησης συνίσταται να βασίζονται σε δραστηριότητες (όπως σημεία διανομής, GRP) και να μην βασίζονται σε δεδομένα «μεριδίου» (share,όπως τις μετρικές share of voice, share of promotions). Αυτό ισχύει ιδιαίτερα για τις Share of voice και της Share of Shelf, όπου ο όγκος των πωλήσεων μπορεί να μην έχει καταγραφεί με ακρίβεια όταν προκύπτουν σημαντικές αλλαγές στα Media και το χώρο των ραφιών.

III. Τύπος δεδομένων

Τα μοντέλα δε θα πρέπει να βασίζονται σε δεδομένα της μάρκας που προκύπτουν σε εθνικό και μηνιαίο επίπεδο. Το ιδανικό είναι η μοντελοποίηση να προκύπτει σε εβδομαδιαία βάση, σε επίπεδο καταστημάτων με περιφερειακά δεδομένα των Media. Γενικά, τα μοντέλα θα πρέπει να είναι τουλάχιστον εβδομαδιαία τα οποία θα παρακολουθούνται σε μια περίοδο τριών ετών ή εάν είναι μηνιαία, να διαχωρίζονται σε επίπεδο καναλιού ή αγοράς και να διαμορφώνονται μέσω συγκεντρωτικής παλινδρόμησης. Σε ένα πολύ βασικό επίπεδο, τα μοντέλα παλινδρόμησης λειτουργούν όταν συνδυάζουμε τη μεταβολή της εξαρτημένης μεταβλητής (π.χ. στον όγκο των πωλήσεων) με τη μεταβολή στις αιτιώδεις μεταβλητές που χρησιμοποιούνται στο μοντέλο. Όταν ένα μοντέλο κατασκευάζεται με συγκεντρωτικά δεδομένα, ένα μέρος της μεταβολής των μεταβλητών των εισροών και των εξόδων μπορεί να χαθεί και όχι απαραίτητα στον ίδιο βαθμό για κάθε μεταβλητή, με αποτέλεσμα οι συντελεστές που υπολογίζονται σε αυτή την παλινδρόμηση να είναι πιθανότατα αρκετά διαφορετικοί από ένα μοντέλο που βασίζεται σε πιο λεπτομερή δεδομένα. Ωστόσο, στην περίπτωση συνόλων δεδομένων που κατά κανόνα αποτελούνται από εικοσιτέσσερις έως τριάντα δύο μηνιαίες αναδρομικές εγγραφές

ανά μεταβλητή κατά μέσον όρο, η συνήθης προσέγγιση συνίσταται στο διαχωρισμό μόνο ενός μικρού ποσού εγγραφών που θα συμπεριληφθούν τελικά στο σετ των δοκιμών.

IV. Εξαρτημένη μεταβλητή

Συνήθως οι αναλύσεις και τα αντίστοιχα μοντέλα διατίθενται για να λύσουν θέματα που αφορούν τον όγκο των πωλήσεων (volume shares) αντί για τη διείσδυση στην αγορά (market penetration) και αυτό γιατί από την εμπειρία μας αλλά και από μια σύγκριση μεταξύ των αγορών, προκύπτει ότι τα δεδομένα που σχετίζονται με τη διείσδυση στην αγορά έχουν σημαντικά περισσότερα σφάλματα. Αυτό οφείλεται στο γεγονός ότι τα τελευταία προέρχονται πάντα από ένα δείγμα και στο τέλος παρέχονται για τον πληθυσμό. Έτσι, υπάρχει διακύμανση στους βαθμούς συμμόρφωσης των καταναλωτών στην παρακολούθηση του ιστορικού των αγορών τους. Εν τω μεταξύ, ο όγκος των εβδομαδιαίων, μηνιαίων ή περιφερειακών καταγραφών αποδεικνύεται συνήθως ότι παρέχει ένα πιο ισχυρό σύνολο δεδομένων για το μοντέλο. Γενικά οι πιλότοι που μοντελοποιούν δεδομένα που αφορούν τη διείσδυση στην αγορά δεν παρέχουν ούτε πληροφορίες για κάθε στάδιο εξέλιξης ούτε συστάσεις. Αυτό δεν αποτελεί έκπληξη, καθώς τα μερίσματα του όγκου των πωλήσεων και η διείσδυση στην αγορά σχετίζονται άμεσα μεταξύ τους.

V. Κριτήρια για την αξιοπιστία του μοντέλου

Όλα τα μοντέλα πρέπει να ανταποκρίνονται στα ακόλουθα κριτήρια:

- $R^2 > 80\%$
- $MAPE < 10\%$
- $1.4 < DW < 2.2$

Τα «πακέτα» των δοκιμών του μοντέλου θα πρέπει να αποδίδουν τιμές $MAPE$ κάτω από το 10%. Τα συμπεράσματα μιας σύγκρισης μεταξύ των αποτελεσμάτων του συνόλου των δοκιμών που πραγματοποιούνται κατά τη φυσική περίοδο ροής έως τη περίοδο που εμφανίζεται μία μεταβολή στις πωλήσεις, έναντι των μέσων τιμών των πωλήσεων, θα μπορούσαν να χρησιμοποιηθούν για να κατανοηθεί πόσο από τη συνολική μεταβολή εξηγείται από το μοντέλο. Έτσι θα πρέπει να οριστεί ένας στόχος σχετικά με το ποσοστό που εξηγείται από το μοντέλο. Αυτό συνήθως ορίζεται στο 50% ή και περισσότερο. Τα πακέτα των δοκιμών προτείνεται να περιλαμβάνουν το

10% του συνόλου δεδομένων για τα εβδομαδιαία μοντέλα (αντίστοιχα 20% για τα μηνιαία μοντέλα).

VI. Ακεραιότητα και ακρίβεια μοντέλου

Θα πρέπει πάντα να υπάρχουν μερικά, γενικώς αποδεκτά, πρότυπα που αποτελούν σημείο αναφοράς για την ακρίβεια του μοντέλου που χρησιμοποιούμε και προφανώς οφείλουν να διασφαλιστούν:

- Ο VIF γενικά προτείνεται να είναι σε τιμή 10 ή χαμηλότερη. Μεγαλύτερες τιμές του VIF μπορεί να είναι δείκτης υψηλής συγγραμμικότητας με μια άλλη μεταβλητή που μπορεί να διπλασιάζεται. Ως εκ τούτου, συνήθως συνιστάται η λήψη μέτρων για την άρση της συγγραμμικότητας. Μια τέτοια ενέργεια θα μπορούσε να είναι η απόρριψη ή ο συνδυασμός μεταβλητών ή ακόμη και η προσθήκη μιας ακόμα πριν από μια συγκεκριμένη μεταβλητή.(Ενότητα 4.8.10)
- Επίπεδα σημαντικότητας: Τυπικά, μια τιμή $p - value = 0{,}05$ ή μικρότερη είναι αποδεκτή ως στατιστικά σημαντική για να συμπεράνουμε με ασφάλεια ότι τα αποτελέσματα οφείλονται στην επίδραση της ανεξάρτητης μεταβλητής και όχι στην τύχη.
- Τα επίπεδα εμπιστοσύνης πρέπει να είναι περίπου 95%.

VII. Απόδοση του μοντέλου

Όταν πια εξασφαλίσουμε πως το μοντέλο μπορεί να παράγει ακριβή και αξιόπιστα αποτελέσματα, τότε είναι καιρός να σχεδιαστεί το πλήρες πλαίσιο όλων των αποτελεσμάτων ώστε να μπορεί με τη σειρά του ο αναλυτής να το επικοινωνήσει καταλλήλως στο αντίστοιχο δημιουργικό τμήμα της εταιρείας. Μια πρόταση είναι η συγκεντρωτική έκθεση να περιλαμβάνει αλλά όχι απαραίτητα να περιορίζεται στα ακόλουθα:

- Ετήσια αύξηση του όγκου των πωλήσεων που σχετίζεται με τις αλλαγές στις δραστηριότητες της μάρκας.
- Ανάλυση ευαισθησίας για την καταγραφή τυχόν αλλαγών στις δραστηριότητες του μάρκετινγκ ή σε βασικούς επιχειρησιακούς οδηγούς (όπως ο όγκος ανά 1 εκατομμύριο εμφανίσεις _impressions, ο όγκος ανά αλλαγή σημείου διανομής) αλλά και για τη μέτρηση της ελαστικότητας

εξαιτίας των αλλαγών του όγκου των πωλήσεων κατά την εισαγωγή στα δεδομένα διαφορετικών ποσοστών.

- Κάθε τακτική σε επίπεδο μάθησης σχετικά με την αποτελεσματικότητα ή την αποδοτικότητα.

Γενικεύοντας τις παραπάνω οδηγίες καταλήγουμε να έχουμε τρία στάδια της οικονομετρικής ανάλυσης:

1. Η εξειδίκευση του υποδείγματος, δηλαδή ο καθορισμός των μεταβλητών που θα το απαρτίζουν, η καταγραφή αυτών σε εξωγενείς και ενδογενείς, καθώς και στην μαθηματική διατύπωση του υποδείγματος.
2. Η κατάλληλη επιλογή των οικονομετρικών τεχνικών για την εκτίμηση των συντελεστών των μεταβλητών μας. Το στάδιο αυτό ονομάζεται εκτίμηση του υποδείγματος.
3. Ο έλεγχος του υποδείγματος με την παράλληλη εφαρμογή οικονομικών, στατιστικών και οικονομετρικών κριτηρίων για τον έλεγχο των αποτελεσμάτων της εκτιμήσεως.

Η παλινδρόμηση είναι από τα πιο σημαντικά εργαλεία του οικονομέτρη για να αναλύσει τα οικονομικά και χρηματοοικονομικά φαινόμενα. Ασχολείται με ην περιγραφή και αξιολόγηση των σχέσεων μεταξύ μιας μεταβλητής, η οποία καλείται εξαρτημένη (dependent) ή μεταβλητή απόκρισης (response) ή προβλέψιμη (predicted), και μιας ή περισσότερων μεταβλητών οι οποίες ονομάζονται ανεξάρτητες (independent) ή προβλεπτικές (predictive) ή επεξηγηματικές (explanatory). Η ανεξάρτητη μεταβλητή παίρνει το όνομα της καθόσον ελέγχεται με μετρήσεις που διεξάγει ο ερευνητής, το αποτέλεσμα των οποίων αναμένεται να διαπιστωθεί επί της εξαρτημένης μεταβλητής, της οποίας οι τιμές εξαρτώνται άμεσα από τις τιμές της πρώτης. Τέτοια εξαρτημένη σχέση καλείται παλινδρόμηση και πιο συγκεκριμένα όταν εμπλέκονται δύο μόνο μεταβλητές έχουμε την απλή παλινδρόμηση. Ο όρος παλινδρόμηση, που πλέον έχει μόνο ιστορική σημασία, οφείλεται στον Francis Galton και αναφέρεται σε μια μελέτη του αναφορικά με την σχέση ανάμεσα στο ύψος των παιδιών και στο ύψος των γονέων. Ο όρος προήλθε από την παρατήρηση του Galton ότι υπάρχει μια τάση όπου ακραίες ως προς το μέσο τους παρατηρήσεις της ανεξάρτητης τυχαίας μεταβλητής, αντιστοιχούν σε παρατηρήσεις της εξαρτημένης τυχαίας μεταβλητής που δεν είναι το ίδιο ακραίες, αλλά είναι πλησιέστερα προς το μέσο τους. Εναλλακτικά, θα μπορούσε να πει κανείς ότι ακραίες παρατηρήσεις

ακολουθούνται από λιγότερο ακραίες παρατηρήσεις, παρατηρήσεις που είναι πλησιέστερες προς το κέντρο ή το μέσο όρο. Λόγω της τάσης αυτής, η παλινδρόμηση χαρακτηρίστηκε από τον Galton ως "παλινδρόμηση προς την μετριότητα". Συγκεκριμένα, η μελέτη των δεδομένων του Galton κατέληξε πως ασυνήθιστα υψηλοί γονείς τείνουν να έχουν παιδιά χαμηλότερα από τους ίδιους, ενώ ασυνήθιστα χαμηλοί γονείς έχουν συνήθως υψηλότερα παιδιά.

Τα μοντέλα παλινδρομήσεως χρησιμοποιούνται ευρέως σήμερα στη διοίκηση των επιχειρήσεων, στην οικονομία, στη μηχανική, στην υγεία, στη βιολογία και στις κοινωνικές επιστήμες. Στη στατιστική, η ανάλυση παλινδρόμησης είναι μια στατιστική διαδικασία για την εκτίμηση των σχέσεων μεταξύ διαφόρων μεταβλητών. Περιέχει πολλές τεχνικές για τη μοντελοποίηση και την ανάλυση των μεταβλητών αυτών, ενώ επικεντρώνεται συνήθως στη σχέση μεταξύ μιας εξαρτημένης και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Κάποια παραδείγματα απλής παλινδρόμησης θα μπορούσαν να είναι τα έσοδα και τα έξοδα μιας οικογένειας, το ύψος των αποδοχών των υπαλλήλων μιας εταιρείας και ο αριθμός των υπαλλήλων κ.α. Είναι ενδιαφέρον λοιπόν να εξεταστούν οι επιδράσεις που κάποιες μεταβλητές ασκούν σε κάποιες άλλες μεταβλητές.

Η γραμμική παλινδρόμηση αποτελεί μια στατιστική μέθοδο η οποία αποσκοπεί στον προσδιορισμό ενός μαθηματικού μοντέλου για την περιγραφή, ερμηνεία, πρόβλεψη των τιμών ενός χαρακτηριστικού (μεταβλητής) σε σχέση με τις τιμές ενός πλήθους άλλων χαρακτηριστικών (μεταβλητών). Στο Κεφάλαιο 3, αρχικά θα ασχοληθούμε με την απλούστερη περίπτωση παλινδρόμησης που είναι η απλή γραμμική παλινδρόμηση, κατά την οποία υπάρχει μία μόνο ανεξάρτητη μεταβλητή X και η εξαρτημένη μεταβλητή Y που μπορεί να προσεγγιστεί ικανοποιητικά από μια γραμμική συνάρτηση του X. Αξίζει να αναφερθεί, πριν την πλήρη τεχνική ανάλυση που θα ακολουθήσει, πως το πρώτο βήμα για να πραγματοποιηθεί η μελέτη μας, είναι η κατασκευή ενός μαθηματικού μοντέλου που περιγράφει τη φύση της σχέσης που υφίσταται μεταξύ των υπό μελέτη μεταβλητών. Η διαδικασία δημιουργίας μιας μαθηματικής "εξίσωσης" για την περιγραφή ενός φαινομένου μπορεί να είναι ιδιαίτερα πολύπλοκη. Αυτό οφείλεται στο γεγονός ότι για την κατασκευή του μοντέλου απαιτείται κάποια γνώση της φύσης της σχέσης μεταξύ των μεταβλητών. Η σχέση που συνδέει την εξαρτημένη μεταβλητή με τις ανεξάρτητες είναι στατιστική και όχι συναρτησιακή. Στην στατιστική σχέση, για κάθε τιμή της ανεξάρτητης

μεταβλητής υπολογίζεται μια θεωρητική τιμή της εξαρτημένης μεταβλητής, ενώ η πραγματική τιμή της βρίσκεται μέσα σε ένα εύρος τιμών το οποίο περιέχει την θεωρητική τιμή. Στην συναρτησιακή σχέση, δηλαδή σε μια εξίσωση, κάθε τιμή της ανεξάρτητης μεταβλητής δίνει πάντα την ίδια τιμή στην εξαρτημένη μεταβλητή (μορφή $Y = f(X)$). Ωστόσο, για ευκολία χρησιμοποιούμε τον όρο "εξισώσεις παλινδρόμησης", παρόλο που δεν πρόκειται για εξίσωση, αλλά για στατιστικό μοντέλο.

Κεφάλαιο 3 : Ανάλυση Παλινδρόμησης

3.1 Εισαγωγή

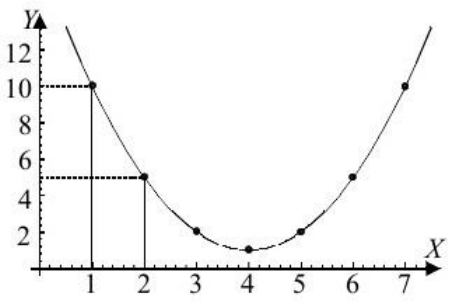
Όπως ήδη αναφέραμε, με την ανάλυση παλινδρόμησης (regression analysis) εξετάζουμε τη σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών με σκοπό την πρόβλεψη των τιμών της μιας, μέσω των τιμών της άλλης (ή των άλλων). Σε κάθε πρόβλημα παλινδρόμησης διακρίνουμε δύο είδη μεταβλητών: τις *ανεξάρτητες* ή ελεγχόμενες ή επεξηγηματικές (independent, predictor, casual, input, explanatory variables) και τις *εξαρτημένες* ή απόκρισης (dependent, response variables). Σε πειραματικές έρευνες, ανεξάρτητη μεταβλητή X είναι εκείνη την οποία μπορούμε να ελέγξουμε, δηλαδή, να καθορίσουμε τις τιμές της. Για παράδειγμα το ύψος της διαφημιστικής δαπάνης ενός προϊόντος ή ο αριθμός των λειτουργούντων ταμείων σε ένα υποκατάστημα τραπέζης. Εξαρτημένη μεταβλητή Y είναι εκείνη στην οποία αντανακλάται το αποτέλεσμα των μεταβολών στις ανεξάρτητες μεταβλητές (π.χ. η ζήτηση ενός προϊόντος, ο χρόνος αναμονής των πελατών ενός υποκαταστήματος τραπέζης, η απόδοση μιας καλλιέργειας, η αντοχή ενός υλικού).

Σε *μη πειραματικές έρευνες* (δειγματοληψίες) η διάκριση μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών δεν είναι πάντοτε σαφής γιατί καμία μεταβλητή δεν είναι ελεγχόμενη αλλά όλες είναι τυχαίες όπως το ύψος και το βάρος των φοιτητών, οι ώρες μελέτης των φοιτητών ενός πανεπιστημιακού τμήματος και η απόδοση τους σε ένα τεστ, οι εβδομάδες εμπειρίας ενός εργάτη σε μια επιχείρηση και ο αριθμός των ελαττωματικών προϊόντων που παράγει, η κατάταξη δέκα προϊόντων από έναν κριτή και η κατάταξη των ιδίων προϊόντων από έναν άλλο κριτή. Ας θεωρήσουμε δύο μεταβλητές X, Y. Αν οι μεταβλητές αυτές συνδέονται με μια σχέση της μορφής $Y = f(X)$ μέσω της οποίας για κάθε τιμή της X μπορούμε να προβλέψουμε ακριβώς την τιμή της Y ,δηλαδή, αν οι τιμές της Y δεν υπόκεινται σε σφάλματα, τότε λέμε ότι οι δύο μεταβλητές συνδέονται με τη συναρτησιακή-προσδιοριστική (deterministic) σχέση $Y = f(X)$. Για παράδειγμα, το ποσό που καταθέτει κάποιος στο Ταμιευτήριο και ο τόκος που παίρνει για το ποσό αυτό, συνδέονται με συναρτησιακή-προσδιοριστική σχέση. Σε αυτές τις περιπτώσεις τα σημεία του διαγράμματος διασποράς βρίσκονται όλα πάνω στην καμπύλη που έχει εξίσωση $Y = f(X)$ και όσες φορές και αν επαναλάβουμε το πείραμα θέτοντας το X στο ίδιο επίπεδο $X = x_i$, θα παίρνουμε πάντα την ίδια τιμή για το Y. Για παράδειγμα, η εξίσωση $Y =$

$(X-4)^2+1$ (που παριστάνει μια παραβολή) περιγράφει προσδιοριστικά τη σχέση μεταξύ των X και Y του παρακάτω πίνακα:

x_i	y_i
1	10
2	5
3	2
4	1
5	2
6	5
7	10

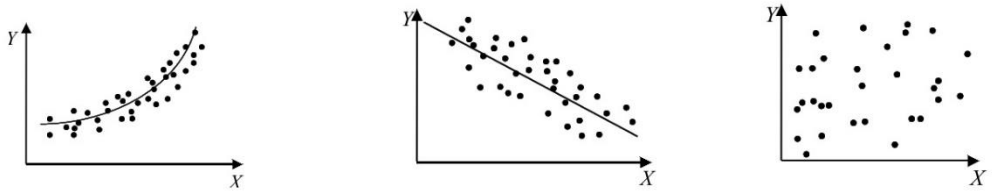
Πίνακας 3.1



Σχήμα 3.1. Γραφική Παράσταση παραβολής.

Οι μη προσδιοριστικές σχέσεις μεταξύ μεταβλητών ονομάζονται *στοχαστικές–στατιστικές (stochastic,probabilistic)* σχέσεις. Στην περίπτωση αυτή, αν επαναλάβουμε το πείραμα πολλές φορές θέτοντας το X στο ίδιο επίπεδο $X = x_i$ τότε στην τιμή x_i της X δεν αντιστοιχεί μια μόνο τιμή y_i της Y αλλά, γενικά, αντιστοιχεί ένα πλήθος διαφορετικών τιμών της Y . Για παράδειγμα, αν X είναι η τιμή ενός προϊόντος και Y είναι η ζήτησή του, η Y βρίσκεται σε στοχαστική σχέση-εξάρτηση από τη X, γιατί η ζήτηση ενός προϊόντος επηρεάζεται και από άλλους παράγοντες όπως είναι το ύψος του εισοδήματος των καταναλωτών, οι τιμές ομοειδών προϊόντων, οι καταναλωτικές συνήθειες, κ.ά.

Σε μια στοχαστική σχέση το διάγραμμα διασποράς είναι, γενικά, ένα *νέφος σημείων* το οποίο πολλές φορές καθορίζει μια ιδεατή γραμμή η οποία δίνει μια πρώτη εικόνα της σχέσης που συνδέει τις δύο μεταβλητές. Η σχέση μάλιστα μεταξύ των δύο μεταβλητών είναι τόσο περισσότερο ισχυρή όσο πιο κοντά στην ιδεατή γραμμή βρίσκονται τα σημεία του διαγράμματος διασποράς. Στην πρώτη από τις παρακάτω γραφικές παραστάσεις έχουμε το διάγραμμα διασποράς μιας ισχυρής σχέσης στην οποία όταν αυξάνουν οι τιμές της X αυξάνουν γενικά και οι τιμές της Y, ενώ στη δεύτερη έχουμε μια λιγότερο ισχυρή σχέση στην οποία όταν αυξάνουν οι τιμές της X ελαττώνονται γενικά και οι τιμές της Y . Τέλος, στην τρίτη περίπτωση, δε φαίνεται να υπάρχει κάποια σχέση μεταξύ των X και Y.



Σχήμα 3.2 Διαγράμματα διασποράς.

Γενικά, δύο μεταβλητές που συνδέονται είτε με συναρτησιακή προσδιοριστική σχέση είτε με στοχαστική σχέση λέγονται «εξαρτημένες». Αν υπάρχει εξάρτηση μεταξύ δύο μεταβλητών, τότε μπορούμε τη μια από αυτές να τη χαρακτηρίσουμε ως «αιτία» και την άλλη ως «αποτέλεσμα». Αυτό όμως, μόνο στην περίπτωση που η εξάρτηση οφείλεται σε σχέση αιτιότητας των δύο μεταβλητών και όχι σε μια απλή συμμεταβολή η οποία μπορεί να οφείλεται σε εξάρτηση των δύο μεταβλητών από μια τρίτη μεταβλητή. Αν, για παράδειγμα, X είναι το ετήσιο εισόδημα μιας οικογένειας και Y, Z είναι τα ποσά που ξοδεύει η οικογένεια αυτή σε ένα έτος για σίτιση και για ένδυση, τότε: αν διαπιστώσουμε σε ένα σύνολο οικογενειών σχέση μεταξύ των X και Y (ή μεταξύ των X και Z) δεχόμαστε ότι υπάρχει εξάρτηση μεταξύ των δύο μεταβλητών και τότε μπορούμε να χαρακτηρίσουμε τη X ως «αιτία» και την Y (ή τη Z) ως «αποτέλεσμα». Αν όμως διαπιστωθεί σχέση μεταξύ των Y και Z (που είναι πολύ πιθανό, αφού και οι δύο μεταβάλλονται με το ετήσιο εισόδημα X) ασφαλώς θα πρόκειται για «νόθα» εξάρτηση.

Για να περιγράψουμε τη στοχαστική εξάρτηση δύο μεταβλητών X και Y προσπαθούμε να βρούμε, όπως και στην προσδιοριστική εξάρτηση, μια σχέση μεταξύ των X και Y η οποία όμως τώρα δε θα δίνει ακριβή αλλά προσεγγιστική μόνο εικόνα της εξάρτησης των X και Y και τα σημεία του διαγράμματος διασποράς των X και Y δε θα βρίσκονται πάνω, αλλά, γύρω από μια καμπύλη.

Ανάλογα με τη σχέση της εξαρτημένης μεταβλητής Y με τις ανεξάρτητες μεταβλητές $X_1, X_2, \ldots$ και τα διαθέσιμα δεδομένα που το ποσοτικοποιούν, υπάρχουν διαφορετικές τεχνικές παλινδρόμησης. Διακρίνουμε τα είδη της παλινδρόμησης σε πολυωνυμική, εκθετική, γραμμική, λογαριθμική. Από αυτές τρεις εντάσσονται στο λεγόμενο ***Γενικευμένο Γραμμικό Μοντέλο (Generalised Linear Model – GLM)*** και είναι οι εξής:

- Γραμμική (linear ή Gaussian) όταν η εξαρτημένη μεταβλητή είναι λόγος (ratio) και ακολουθεί κανονική κατανομή (normal distribution).

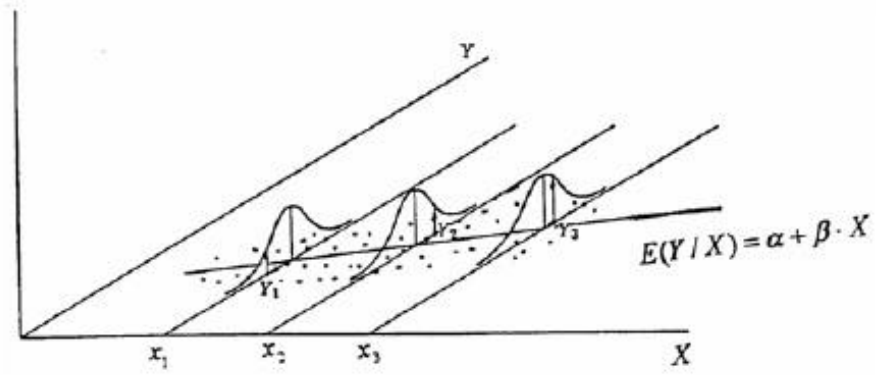
- Poisson όταν η εξαρτημένη μεταβλητή αφορά αριθμό (counts) σπάνιων συμβάντων και ακολουθεί κατανομή Poisson.
- Λογιστική (logistic) όταν η εξαρτημένη μεταβλητή λαμβάνει δύο τιμές (yes/no) και ακολουθεί διωνυμική κατανομή (binomial distribution)

Μια μέθοδος που χρησιμοποιείται για την περιγραφή της στοχαστικής εξάρτησης δύο μεταβλητών είναι η μέθοδος των **ελαχίστων τετραγώνων** και αυτή θα εφαρμόσουμε στη συνέχεια για να μελετήσουμε την πιο απλή μορφή στοχαστικής εξάρτησης, τη **γραμμική**.

3.2 Απλή Γραμμική Παλινδρόμηση

Αν το διάγραμμα διασποράς δύο μεταβλητών X και Y έχει μορφή *επιμήκους κεκλιμένης έλλειψης* ή *πλατυσμένου J* η σχέση των X και Y είναι κατά προσέγγιση γραμμική. Στην περίπτωση αυτή έχουμε την απλούστερη μορφή παλινδρόμησης, την **απλή γραμμική παλινδρόμηση** όπου υπάρχει μόνο μια ανεξάρτητη μεταβλητή X και η εξαρτημένη μεταβλητή Y μπορεί να προσεγγισθεί ικανοποιητικά από μια γραμμική συνάρτηση του X.

Η γραμμική σχέση $Y = \alpha + \beta \cdot X$ δε μπορεί, ασφαλώς, να περιγράψει τη γραμμική εξάρτηση των μεταβλητών X και Y αφού αν, για παράδειγμα, X είναι η τιμή ενός προϊόντος και Y είναι η ζήτηση του προϊόντος αυτού, και διατηρήσουμε τη X στο ίδιο επίπεδο $X = x_1$ τότε οι αντίστοιχες τιμές του Y θα είναι φυσικά διαφορετικές στις διάφορες επαναλήψεις. Για παράδειγμα αν X είναι η ποσότητα ενός λιπάσματος και Y είναι η απόδοση μιας καλλιέργειας, και διατηρήσουμε τη X στο ίδιο επίπεδο $X = x_1$ τότε οι αντίστοιχες τιμές του Y θα είναι διαφορετικές για κάθε επανάληψη αφού παράγοντες όπως, η θερμοκρασία, οι βροχοπτώσεις, η ποιότητα του εδάφους, θα επηρεάζουν, επίσης, την παραγωγή. Επιπλέον, συμβαίνει να παρατηρούνται και σφάλματα μέτρησης των τιμών της Y (λόγω οργάνων ή ελλιπούς πληροφόρησης). Έτσι, για $X = x_1$ το αντίστοιχο Y είναι μια τυχαία μεταβλητή Y_1 που ακολουθεί κάποια κατανομή. Ομοίως, για $X = x_2$ θα έχουμε κάποια άλλη κατανομή Y_2 κ.ό.κ..



Σχήμα 3.3 Πληθυσμιακή ευθεία παλινδρόμησης.

Η ανάλυσή μας λοιπόν είναι εμπειρική (βασίζεται στην υπάρχουσα εμπειρία μας με βάση το τυχαίο δείγμα) και άρα το μοντέλο μας είναι στοχαστικό. Αντιθέτως αν η ανάλυσή μας ήταν θεωρητική, γνωρίζαμε δηλαδή όλον τον πληθυσμό, το μοντέλο θα ήταν προσδιοριστικό. Επομένως, στην εξίσωση $Y = \alpha + \beta \cdot X$, πρέπει να προσθέσουμε έναν ακόμη όρο, το *τυχαίο σφάλμα* ε το οποίο, για δεδομένη τιμή της *X,* να περιγράφει τη διαφορά της παρατηρούμενης από τη θεωρητική $\alpha + \beta \cdot X$ τιμή της Y. Δηλαδή, $\varepsilon = Y - (\alpha + \beta \cdot X)$.

Τότε προκύπτει, το στοχαστικό μοντέλο

$$Y = \alpha + \beta \cdot X + \varepsilon \ .$$

Για λόγους απλούστευσης των υπολογισμών και εφικτής λύσης του προβλήματος, βασιζόμαστε στις υποθέσεις, $E(\varepsilon) = 0$ και $E(Y|X) = \alpha + \beta \cdot X$. Δηλαδή, υποθέτουμε ότι τα σφάλματα έχουν μέση τιμή μηδέν, συνήθως θεωρούμε $\varepsilon \sim N(0, \sigma^2)$ με σ^2 άγνωστο και ότι για τις διάφορες τιμές της *X,* οι αντίστοιχες μέσες τιμές της *Y* βρίσκονται πάνω σε μια ευθεία. *Η ευθεία αυτή* $E(Y|X) = \alpha + \beta \cdot X$ *, ονομάζεται* ***πληθυσμιακή ευθεία παλινδρόμησης*** (Σχήμα 3.3). Πιο συγκεκριμένα:

$$Y_i = a + \beta \cdot X_i + \varepsilon_i \ , \ \varepsilon_i \sim N(0, \sigma^2) \leftrightarrow E(Y_i | X_i = x_i) = a + \beta \cdot x_i \quad (3.1)$$

με ε_i ανεξάρτητες και ισόνομες τυχαίες μεταβλητές .

Οι σταθερές α , β προσδιορίζονται με βάση δείγμα ζευγών παρατηρήσεων $(x_1, y_1), (x_2, y_2), \dots, (x_v, y_v)$.Η σταθερά α εκφράζει τη μέση τιμή της Yόταν το $X = 0$. Η σταθερά β εκφράζει το πόσο αναμένεται να μεταβληθεί η αναμενόμενη τιμή της Y,αν η X αυξηθεί κατά μία μονάδα ενώ η ποσότητα σ^2εκφράζει τη διασπορά των

σφαλμάτων, την οποία θεωρούμε *σταθερή* ανεξάρτητα της τιμής της τυχαίας μεταβλητής X (*υπόθεση ομοσκεδαστικότητας*). Τέλος, επειδή η τυχαιότητα της Y δεδομένης μιας τιμής της $X = x$ οφείλεται στα σφάλματα, το σ^2 εκφράζει και τη διασπορά της δεσμευμένης κατανομής της τυχαίας μεταβλητής $Y|x$.

3.2.1.Μέθοδοι Εκτίμησης Παραμέτρων

Η γραμμή παλινδρομήσεως στον πληθυσμό είναι άγνωστη εφόσον δεν γνωρίζουμε τις τιμές των παραμέτρων α και β. Αν γνωρίζαμε όλες τις δυνατές τιμές που παίρνει η Y για δύο τουλάχιστον τιμές της X, θα μπορούσαμε να υπολογίσουμε τις τιμές των παραμέτρων α και β , αφού σε αυτήν την περίπτωση θα γνωρίζαμε δύο σημεία από τα οποία διέρχεται η γραμμή παλινδρομήσεως. Εφόσον όμως αυτό είναι αδύνατο, εκτιμάμε τις τιμές των συντελεστών α και β από δείγμα παρατηρήσεων για τις μεταβλητές Y και X.

Έτσι, κάνουμε μια εκτίμηση της πληθυσμιακής γραμμής παλινδρομήσεως από την δειγματική εξίσωση παλινδρομήσεως $\hat{Y}$, όπου τα εκτιμώμενα (προβλεπόμενα) λάθη καλούνται υπόλοιπα . Με την γραμμή παλινδρομήσεως του δείγματος, προσπαθούμε να ερμηνεύσουμε τη μεταβλητότητα της Y που εξηγείται από τις μεταβολές στην τιμή της X. Από την άποψη αυτή, η μεταβλητή X είναι η ερμηνευτική μεταβλητή, ενώ η μεταβλητή Y είναι η ερμηνευόμενη μεταβλητή.

3.2.1. Α. Μέθοδος Ελαχίστων Τετραγώνων (Ordinary Least Squares – OLS)

Με τη μέθοδο των ελαχίστων τετραγώνων θα προσδιορίσουμε στη συνέχεια μια εκτίμηση $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$ της ευθείας $E(Y|X) = \alpha + \beta \cdot X$ όπου $\hat{\alpha}$ και $\hat{\beta}$ οι εκτιμήτριες των α , β αντίστοιχα. Η εκτίμηση $\hat{Y} = \hat{\alpha} + \widehat{\beta} \cdot X$ της πληθυσμιακής ευθείας παλινδρόμησης $E(Y|X) = \alpha + \beta \cdot X$,ονομάζεται **ευθεία ελαχίστων τετραγώνων** από τη μέθοδο υπολογισμού των παραμέτρων της.

Θεωρούμε ν ζεύγη παρατηρήσεων $(x_1, y_1), (x_2, y_2), \dots, (x_\nu, y_\nu)$. Αναζητούμε προσέγγιση της μορφής:

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

Όπου τα ε_i τυχαία σφάλματα παριστάνουν τις άγνωστες κατακόρυφες αποκλίσεις της πραγματικής τιμής y_i από την προσαρμοσμένη (θεωρητική ευθεία (3.1)) $\alpha + \beta \cdot x_i$ για δοθείσα x_i. Δηλαδή,

$$\varepsilon_i = y_i - (\alpha + \beta \cdot x_i).$$

Είναι φανερό, ότι η εκλογή (εκτίμηση) των α και β θα πρέπει να γίνει έτσι ώστε να ελαχιστοποιηθούν οι ποσότητες ε_i. Για το σκοπό αυτό, θα αναζητήσουμε τις τιμές των α και β για τις οποίες ελαχιστοποιείται το άθροισμα των τετραγώνων των ε_i. Δηλαδή, η ποσότητα

$$\sum_{i=1}^{\nu} \varepsilon_i^2 = \sum_{i=1}^{\nu} (y_i - \alpha - \beta \cdot x_i)^2 \qquad (3.2)$$

(Η ελαχιστοποίηση του αθροίσματος $\sum \varepsilon_i$ δεν αποτελεί ασφαλές κριτήριο επιλογής διότι κάποια αρνητικά ε_i θα αναιρούν αντίστοιχες θετικές ποσότητες του αθροίσματος).

Παραγωγίζοντας την (3.2) ως προς α και β και εξισώνοντας με μηδέν παίρνουμε τις ακόλουθες δύο εξισώσεις που ονομάζονται **κανονικές εξισώσεις**:

$$\sum_{i=1}^{\nu} y_i = \nu \cdot \alpha + \beta \cdot \sum_{i=1}^{\nu} x_i$$

$$\sum_{i=1}^{\nu} x_i \cdot y_i = \alpha \cdot \sum_{i=1}^{\nu} x_i + \beta \cdot \sum_{\iota=1}^{\nu} x_i^2$$

Λύνοντας το σύστημα των κανονικών εξισώσεων παίρνουμε :

$$\hat{\beta} = \frac{\nu \cdot \sum_{\iota=1}^{\nu} x_i \cdot y_i - (\sum_{i=1}^{\nu} x_i) \cdot (\sum_{i=1}^{\nu} y_i)}{\nu \cdot \sum_{i=1}^{\nu} {x_i}^2 - (\sum_{\iota=1}^{\nu} x_i)^2} = \frac{\sum_{\iota=1}^{\nu} x_i \cdot y_i - \nu \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{\nu} {x_i}^2 - \nu \cdot \bar{x}^2}$$

Ή

$$\hat{\beta} = \frac{\sum_{i=1}^{\nu} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{\nu} (x_i - \bar{x})^2}$$

Ορίζοντας ως

$$c_{xy} = \sum_{i=1}^{\nu} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \qquad (3.3)$$

και $c_{xx} = \sum_{i=1}^{\nu} (x_i - \bar{x})^2 \qquad (3.4)$

τότε :

$$\hat{\beta} = \frac{c_{xy}}{c_{xx}} \text{ και } \hat{a} = \bar{y} - \hat{\beta} \cdot \bar{x}\,, \qquad (3.5)$$

οι οποίες είναι ***τυχαίες μεταβλητές*** (από διαφορετικό δείγμα ενδέχεται να προκύψουν διαφορετικές εκτιμήτριες).

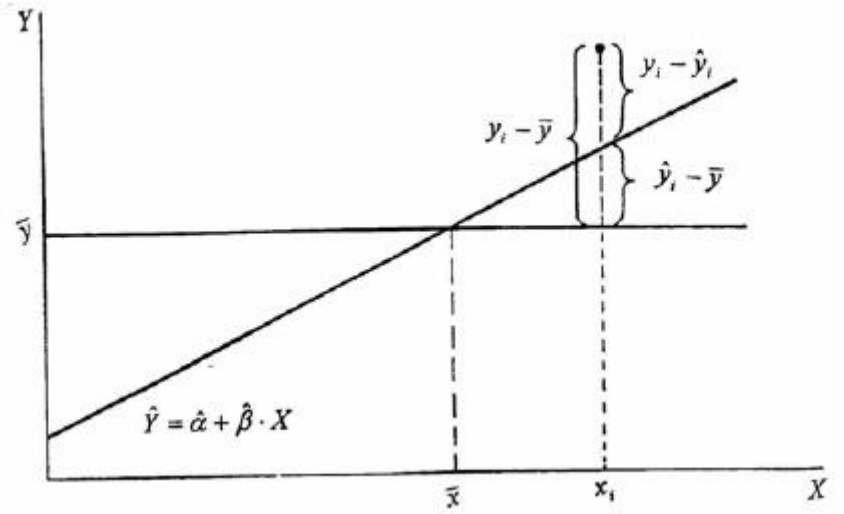
Η **εκτίμηση ελαχίστων τετράγωνων** $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$ της **ευθείας παλινδρόμησης** από

το δείγμα των ν ζευγών παρατηρήσεων είναι, επομένως, η

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X = \bar{y} - \hat{\beta} \cdot \bar{x} + \hat{\beta} \cdot X = \bar{y} + \hat{\beta} \cdot (X - \bar{x})$$

Ή $$\hat{Y} = \bar{y} + \frac{c_{xy}}{c_{xx}} \cdot (X - \bar{x}) \qquad (3.6)$$

Προφανώς, *η ευθεία ελαχίστων τετραγώνων*, διέρχεται από το σημείο $(\bar{x}, \bar{y})$ όπως φαίνεται και στο ακόλουθο σχήμα.



Σχήμα 3.4. Εκτίμηση ελαχίστων τετραγώνων $\hat{Y}$.

Επισημαίνουμε ότι πρέπει να γίνεται διάκριση μεταξύ της παρατηρούμενης τιμής του Y και της $\hat{Y}$ που εκτιμάμε. Η παρατηρούμενη τιμή y_i είναι η πραγματική τιμή της Y, ενώ η τιμή$\hat{y}_i$ της $\hat{Y}$, είναι εκτίμηση της μέσης τιμής $E(Y|X = x_i)$.

Είναι λογικό, στο τέλος των υπολογισμών μας να προκύπτει το ακόλουθο ερώτημα: Πόσο «καλή» είναι η ευθεία ελαχίστων τετραγώνων $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ ως εκτίμηση της ευθείας παλινδρόμησης $E(Y|X) = \alpha + \beta X$; Είναι το μοντέλο σημαντικό; Για την απάντηση σε αυτό το ερώτημα χρειάζεται να βεβαιωθούμε ότι η κλίση της ευθείας β, δεν μπορεί να είναι μηδέν. Αλλιώς, αν η κλίση είναι 0, τότε το μοντέλο της απλής γραμμικής παλινδρόμησης δεν είναι καθόλου χρήσιμο.

I. Παρατηρήσεις και ιδιότητες για την ευθεία ελαχίστων τετραγώνων

1. Οι προβλέψεις που μπορούμε να κάνουμε για την εξαρτημένη μεταβλητή Y από τις τιμές της ανεξάρτητης μεταβλητής X μέσω της *ευθείας ελαχίστων*

τετραγώνων $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ πρέπει να γίνονται μόνο για τις τιμές της ανεξάρτητης μεταβλητής, οι οποίες βρίσκονται στο διάστημα που έχει γίνει η μελέτη ή πολύ κοντά στα άκρα του διαστήματος αυτού.

2. Η εξίσωση της *ευθείας ελαχίστων τετραγώνων* $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ **δε μας επιτρέπει** να κάνουμε προβλέψεις για τις τιμές της X, όταν δίνονται οι τιμές της Y. Για να είναι αυτό δυνατόν, πρέπει να προσδιορίσουμε εξαρχής την *ευθεία ελαχίστων τετραγώνων* **της X πάνω στην Y**, $\hat{X} = \hat{\gamma} + \hat{\delta} \cdot Y$, η οποία γενικά είναι διαφορετική από την

$\hat{Y} = \hat{\alpha} + \hat{\beta}X$. Και στις δύο όμως περιπτώσεις οι ευθείες διέρχονται από το σημείο $(\bar{x}, \bar{y})$, Σχήμα 3.4.

3. Επισημαίνουμε ότι για δοσμένη τιμή x_i της X,η εκτίμηση $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ αφορά **τη μέση τιμή** $E(Y|X = x_i)$ *της* Y και **όχι την πραγματική τιμή** του Y.

4. Το άθροισμα των εκτιμώμενων καταλοίπων είναι μηδέν.

5. Το άθροισμα των τιμών της Y από το δείγμα, είναι ίσο με το άθροισμα των τιμών που υπολογίζουμε από την παλινδρόμηση, δηλαδή: $\sum_{i=1}^{v} y_i = \sum_{i=1}^{v} \hat{y}_i$

6. Το άθροισμα των γινομένων των τιμών της X και των καταλοίπων είναι μηδέν: $\sum_{i=1}^{v} x_i \varepsilon_i = 0$.

7. Το άθροισμα των γινομένων των καταλοίπων και των τιμών της Y που υπολογίζουμε από την παλινδρόμηση του δείγματος είναι μηδέν : $\sum_{i=1}^{v} \hat{y}_i \varepsilon_i = 0$

II. Ιδιότητες των εκτιμητριών ελαχίστων τετραγώνων

Σύμφωνα με το θεώρημα των Gauss-Markov, για το κλασικό γραμμικό υπόδειγμα, οι εκτιμητές που προκύπτουν από τη μέθοδο των ελαχίστων τετραγώνων είναι άριστοι, γραμμικοί και αμερόληπτοι εκτιμητές.

Θεώρημα των Gauss-Markov: Για το απλό γραμμικό μοντέλο, οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\alpha}$ *και* $\hat{\beta}$ είναι:

1. Γραμμικές συναρτήσεις των παρατηρήσεων της εξαρτημένης μεταβλητής Y.
2. Αμερόληπτες.
3. Μεταξύ όλων των γραμμικών αμερόληπτων εκτιμητών, έχουν την μικρότερη διακύμανση.

Πιο αναλυτικά, είναι προφανές ότι οι εκτιμήτριες $\hat{a}, \hat{\beta}$ αλλάζουν τιμή για διαφορετικά δείγματα. Έτσι τα $\hat{a}, \hat{\beta}$ είναι τυχαίες μεταβλητές. Αν $\varepsilon \sim N(0, \sigma^2)$ αποδεικνύεται ότι:

$$E(\hat{\alpha}) = \alpha \ , \qquad E(\hat{\beta}) = \beta$$

(Αυτό σημαίνει ότι οι $\hat{a}, \hat{\beta}$ είναι αμερόληπτες εκτιμήτριες των α, β αντίστοιχα.) Και οι $\hat{a}, \hat{\beta}$ ακολουθούν κανονικές κατανομές με:

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2 (\frac{1}{\nu} + + \frac{\bar{x}^2}{(\nu - 1)s_X^2})\right)$$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{(\nu - 1)s_X^2}\right)$$

Όπου ν, το μέγεθος του δείγματος και

$$s_X^2 = \frac{\sum_{i=1}^{\nu} x_i^2 - \nu \bar{x}^2}{\nu - 1}$$

3.2.1. Β. Μέθοδος Σταθμισμένων Ελαχίστων Τετραγώνων

Η Μέθοδος Σταθμισμένων Ελαχίστων Τετραγώνων (Weighted Least Squares – WLS) είναι μια τροποποίηση της OLS σύμφωνα με την οποία οι εκτιμημένες παράμετροι βασίζονται σε σταθμισμένες τιμές της εξαρτημένης μεταβλητής, που με απλά λόγια σημαίνει ότι κάθε τιμή της εξαρτημένης μεταβλητής πολλαπλασιάζεται με μια τιμή μια νέας μεταβλητής βαρών (weights). Οι σχέσεις μπορούν να διατυπωθούν ως εξής:

$$w_i \hat{y}_i = w_i(\hat{a} + \hat{\beta} x_i) \ \kappa\alpha\iota \sum_{i=1}^{\nu} w_i (y_i - \hat{y}_i)^2$$

3.2.1. Γ. Η Μέθοδος της Μέγιστης Πιθανοφάνειας (Maximum Likelihood – ML)

Η Μέθοδος Μέγιστης Πιθανοφάνειας αποτελεί πιο γενική τεχνική εκτίμησης του μοντέλου σε σχέση με τη μέθοδο OLS. Επιτρέπει την κατασκευή μιας συνάρτησης πιθανοφάνειας η μεγιστοποίηση της οποίας είναι ο τρόπος για να καθοριστούν οι εκτιμήτριες. Σύμφωνα με τους Kleinbaum et al. (1988, σελ. 489) «η μέθοδος ML

παράγει εκτιμήτριες των οποίων οι ιδιότητες είναι ιδανικές για μεγάλα δείγματα (υπό ορισμένες συνθήκες μαθηματικής κανονικότητας), όταν η υποτιθέμενη συνάρτηση πιθανοφάνειας είναι σωστή. Οι εκτιμήτριες ML λέγεται ότι είναι ασυμπτωτικά βέλτιστες, υπό την έννοια ότι επιθυμητές ιδιότητες, όπως η αμεροληψία, η ελάχιστη διακύμανση, και η κανονικότητα ισχύουν ακριβώς στο όριο μόνο όσο η ποσότητα των δεδομένων γίνεται απείρως μεγάλη. Στην πράξη, αυτό σημαίνει ότι είναι λογικό να υποθέσουμε για μεγάλα σύνολα δεδομένων ότι μια εκτιμήτρια ML θα είναι ουσιαστικά αμερόληπτη, θα έχει μικρή διακύμανση και θα ακολουθεί περίπου κανονική κατανομή όταν χρησιμοποιείται η κατάλληλη συνάρτηση μέγιστης Πιθανοφάνειας».

Στο κλασικό κανονικό γραμμικό υπόδειγμα, οι διαταρακτικοί όροι ε_ι για $i = 1,2,\dots,\nu$ είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κανονική κατανομή με μέσο το μηδέν και σταθερή διακύμανση σ^2 . Επομένως, η συνάρτηση πυκνότητας πιθανότητας του διαταρακτικού όρου είναι:

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\varepsilon_i - 0}{\sigma})^2}$$

Δηλαδή, πιο αναλυτικά, υπό τις υποθέσεις ότι: α) η εξαρτημένη μεταβλητή y_i ακολουθεί κανονική κατανομή με μέσο $\mu = E(y_i)$ και διακύμανση $Var(y_i) = \sigma^2$, β) η τιμή x_i είναι μη στοχαστική (η τιμή x_i έχει μετρηθεί χωρίς σφάλμα) και $y_i, i = 1, \dots, n$ είναι ανεξάρτητες μεταξύ τους, η κατανομή του y_ι είναι:

$$f_{Y_i}(y_i; a, \beta, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \alpha - \beta \cdot x_i}{\sigma}\right)^2}$$

Συνεπώς, η συνάρτηση Πιθανοφάνειας του δείγματος είναι:

$$L(y; a, \beta, \sigma^2) = \prod_{i=1}^{\nu} f_{Y_i}\left(y_i; a, \beta, \sigma^2\right)$$

$$= \prod_{i=1}^{\nu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\varepsilon_i}{\sigma})^2} = \frac{1}{(2\pi\sigma^2)^{\frac{\nu}{2}}} e^{-\frac{1\sum \varepsilon_i^2}{2\ \sigma^2}}$$

Αντικαθιστώντας όμως τη σχέση (3.2) στη συνάρτηση L προκύπτει ότι η συνάρτηση Πιθανοφάνειας των παρατηρήσεων y_iτου δείγματος είναι:

$$L(y;a,\beta,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{\frac{\nu}{2}}}e^{-\frac{1}{2}\Sigma\left(\frac{y_i-\alpha-\beta\cdot x_i}{\sigma}\right)^2}$$

Οι εκτιμήτριες ML των $\alpha,\beta\ και\ \sigma^2$ θα είναι εκείνες οι τιμές των $\alpha,\beta\ και\ \sigma^2$, που συμβολίζονται με $\hat{\alpha},\hat{\beta}\ και\ \hat{\sigma}^2$ αντίστοιχα, για τις οποίες το μέγεθος $L(y;a,\beta,\sigma^2)$ φθάνει τη μέγιστη τιμή του ως συνάρτηση των $\alpha,\beta\ και\ \sigma^2$. Αντί να χρησιμοποιήσουμε αυτήν την μορφή της συνάρτησης L, χρησιμοποιούμε τον λογάριθμό της και στην συνέχεια για να βρούμε τις τιμές των παραμέτρων α και β που μεγιστοποιούν την νέα σχέση που προκύπτει, βρίσκουμε τις μερικές παραγώγους και τις εξισώνουμε με το μηδέν. Έχουμε λοιπόν ότι:

$$\log L=-\frac{\nu}{2}\log 2\pi-\frac{\nu}{2}\log\sigma^2-\frac{1}{2}\frac{\sum(y_i-\alpha-\beta\cdot x_i)^2}{\sigma^2}$$

$$\frac{\partial\log L}{\partial\alpha}=0\Rightarrow\sum_{i=1}^{\nu}y_i=\nu\cdot\alpha+\beta\cdot\sum_{i=1}^{\nu}x_i$$

$$\frac{\partial\log L}{\partial\beta}=0\Rightarrow\sum_{i=1}^{\nu}x_i\cdot y_i=\alpha\cdot\sum_{i=1}^{\nu}x_i+\beta\cdot\sum_{\iota=1}^{\nu}x_i^2$$

Παρατηρούμε ότι προκύπτουν οι ίδιες κανονικές εξισώσεις όπως με την μέθοδο των ελαχίστων τετραγώνων. Άλλωστε μεγιστοποίηση της λογαριθμικής σχέσης ως προς α και β , σημαίνει ελαχιστοποίηση του αθροίσματος των τετραγώνων των αποκλίσεων. Μπορεί κανείς να δείξει ότι οι εκτιμήτριες ML, είναι αντίστοιχα:

$$\hat{\beta}=\frac{\sum_{i=1}^{\nu}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sum_{i=1}^{\nu}(x_i-\bar{x})^2}$$

$$\hat{a}=\bar{y}-\hat{\beta}\cdot\bar{x}$$

3.2.2. Ανάλυση Διασποράς και Ολική Μεταβλητότητα (total variation) SST

Από την προφανή σχέση $y_i-\bar{y}=(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})$ μπορεί εύκολα να αποδειχθεί ότι

$$\sum_{i=1}^{\nu}(y_i-\bar{y})^2=\sum_{i=1}^{\nu}(y_i-\hat{y}_i)^2+\sum_{i=1}^{\nu}(\hat{y}_i-\bar{y})^2 \qquad (3.7)$$

Το άθροισμα της σχέσης (3.7) γράφεται και ως εξής:

$$SST = \sum_{i=1}^{\nu}(y_i - \bar{y})^2 \qquad (3.8)$$

Λέγεται **ολικό άθροισμα τετραγώνων (total sum of squares)** ή **ολική μεταβλητότητα (total variation)** των y_iκαι όπως φαίνεται από τη (3.7) αναλύεται σε δύο συνιστώσες: στο **άθροισμα τετραγώνων παλινδρόμησης (regression sum of squares)**

$$SSR = \sum_{i=1}^{\nu}(\hat{y}_i - \bar{y})^2 \qquad (3.9)$$

και στο **άθροισμα τετραγώνων των (εκτιμημένων) σφαλμάτων (error sum of squares)** ή **υπόλοιπο μεταβλητότητας (residual variation)**

$$SSE = \sum_{i=1}^{\nu}\hat{\varepsilon}_i^2 = \sum_{i=1}^{\nu}(y_i - \hat{y}_i)^2 \qquad (3.10)$$

Για το οποίο ισχύουν τα εξής:

- Το άθροισμα τετραγώνων SSE λαμβάνει μη αρνητικές τιμές $SSE \geq 0$ ενώ γίνεται μηδέν αν $\hat{\varepsilon}_i = 0 \Leftrightarrow \hat{y}_i = y_i \, για \, όλα \, τα \, i = 1,2,..\nu$ (η ευθεία παλινδρόμησης διέρχεται από όλα τα σημεία $(x_1, y_1), (x_2, y_2), \dots, (x_\nu, y_\nu)$
- Όταν το SSE λαμβάνει μικρές θετικές τιμές, όλες οι διαφορές $\hat{\varepsilon}_i = y_i - \hat{y}_i$, $i = 1,2, \dots \nu$ θα είναι μικρές (θετικές ή αρνητικές) και επομένως η ευθεία παλινδρόμησης διέρχεται «κοντά» στα σημεία $(x_i, y_i), i = 1,2, \dots, \nu$
- Όταν το SSE λαμβάνει μεγάλες θετικές τιμές, κάποιες διαφορές $\hat{\varepsilon}_i = y_i - \hat{y}_i$, $i = 1,2, \dots \nu$ θα είναι μεγάλες (κατά απόλυτη τιμή) και επομένως η ευθεία παλινδρόμησης δε βρίσκεται «κοντά» σε όλα τα σημεία $(x_i, y_i), i = 1,2, \dots, \nu$

Τελικά ,από τις σχέσεις (3.8) , (3.9) και (3.10) ,η (3.7) γίνεται:

$$\boldsymbol{SST = SSR + SSE}$$

Το SST μετράει τη συνολική μεταβλητότητα των παρατηρήσεων y_i δηλαδή εκφράζει *την αβεβαιότητα στην πρόβλεψη του* Y *όταν δε χρησιμοποιείται το* X. Το SSR εκφράζει το μέρος της μεταβλητότητας που μπορεί να οφείλεται στο X (ερμηνεύεται από την ευθεία παλινδρόμησης) και το $SSE = SST - SSR$ εκφράζει την υπόλοιπη μεταβλητότητα που δεν εξηγείται από την παλινδρόμηση (σφάλμα).

Όταν το άθροισμα τετραγώνων (3.9) διαιρείται με τους βαθμούς ελευθερίας του, το αποτέλεσμα καλείται ***μέσο τετράγωνο (meansquare)*** και συμβολίζεται με $\boldsymbol{MS}$. Ειδικότερα:

$$\frac{SSR}{1} = MSR \, και \, \frac{SSE}{n-2} = MSE \qquad (3.11)$$

Παρομοίως, αν υπολογίζαμε ένα *μέσο του συνολικού αθροίσματος τετραγώνων* (3.8), θα ήταν:

$$MST = \frac{SST}{n-1} = \frac{1}{n-1}\sum_{i=1}^{\nu}(y_i - \bar{y})^2$$

που αποτελεί την γνωστή εκτιμήτρια της διασποράς του δείγματος τιμών y_i. Για τα μέσα τετράγωνα, σε αντίθεση με τα αθροίσματα τετραγώνων:

$$MST \neq MSR + MSE$$

Ο πίνακας ανάλυσης διασποράς παρουσιάζεται παρακάτω:

Πηγή Μεταβλητότητας	**Άθροισμα Τετραγώνων**	**Βαθμοί Ελευθερίας**	**Μέσο άθροισμα Τετραγώνων**	**Έλεγχος F**
Παλινδρόμηση (Regression)	$SSR = \sum_{i=1}^{\nu}(\hat{y}_i - \bar{y})^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Υπόλοιπα (Error)	$SSE = \sum_{i=1}^{\nu}(y_i - \hat{y}_i)^2$	n-2	$MSE = s^2 = \frac{SSE}{n-2}$	
Σύνολο (Total)	$SST = \sum_{i=1}^{\nu}(y_i - \bar{y})^2$	n-1		

Πίνακας 3.2 Ανάλυση Διασποράς.

3.2.3. Συντελεστής Προσδιορισμού R^2 (Coefficient of Determination)

Ο λόγος $R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{\nu}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{\nu}(y_i-\bar{y})^2}$

εκφράζει το ποσοστό της συνολικής διασποράς (μεταβλητότητας) των y_i που εξηγείται (απορροφάται) από την παλινδρόμηση (δηλαδή των ανεξάρτητων x_i). Το

R^2 λέγεται ***συντελεστής προσδιορισμού (coefficient of determination)*** και παίρνει τιμές στο κλειστό διάστημα $[\mathbf{0}, \mathbf{1}]$.

$$0 \leq R^2 \leq 1$$

- Όταν η κλίση της ευθείας ελαχίστων τετραγώνων είναι μηδέν δηλαδή $\hat{\beta} = 0$ θα είναι $R = 0$

$$R^2 = 0 \Leftrightarrow SSR = 0$$

- Όταν όλα τα σημεία (x_i, y_i), $i = 1,2, \dots, \nu$ βρίσκονται πάνω στην ευθεία ελαχίστων τετραγώνων θα έχουμε $y_i = \hat{y_i}$, άρα $SSE = \sum_{i=1}^{\nu}(y_i - \hat{y}_i)^2 = 0$, οπότε $R^2 = 1$, δηλαδή

$$R^2 = 1 \Leftrightarrow SSR = SST \Leftrightarrow SSE = 0$$

- Στις διάφορες πρακτικές εφαρμογές όταν $0 < R^2 < 1$, όσο πλησιέστερα βρίσκεται προς το 1 τόσο *καλύτερη είναι η ευθεία ελαχίστων τετραγώνων ως εκτίμηση της ευθείας παλινδρόμησης*, **υπό την προυπόθεση ότι το γραμμικό μοντέλο είναι το κατάλληλο**. Αξίζει τέλος να σημειωθεί πως το R^2 *δε μετρά* πόσο μεγάλη είναι η κλίση $\hat{\beta}$ της ευθείας παλινδρόμησης.

Το υπόλοιπο ποσοστό $1 - R^2 = \frac{SSE}{SST}$ εκφράζει το ποσοστό της συνολικής διασποράς που οφείλεται στο τυχαίο σφάλμα και παραμένει ανεξήγητο απο τη X και θα πρέπει να αποδοθεί είτε σε άλλες ανεξάρτητες μεταβλητές που επηρεάζουν την τιμή της Y (οι οποίες δε χρησιμοποιήθηκαν στη διαμόρφωση του γραμμικού μοντέλου), είτε σε φυσική μεταβλητότητα (στατιστική τυχαιότητα) των τιμών της Y.

<u>Αξιόλογες παρατηρήσεις αναφορικά με τον Συντελεστή Προσδιορισμού R^2:</u>

I. Το R^2 δεν είναι κατάλληλος δείκτης καλής προσαρμογής ενός μοντέλου. Μπορεί να είναι πολύ μικρό ακόμα και αν το μοντέλο είναι το σωστό. Στα γραμμικά μοντέλα θεωρώντας πως όλες οι προϋποθέσεις ικανοποιούνται, όσο μεγαλώνει το σ^2 τόσο το R^2 μικραίνει.

II. Το R^2 μπορεί να είναι κοντά στο 1 , όταν το μοντέλο είναι εντελώς εσφαλμένο. Το R^2 δεν μας παρέχει γνώση για το σφάλμα πρόβλεψης. Ακόμα και αν το σ^2 είναι ακριβώς το ίδιο και δεν έχουμε διαφορά στις εκτιμήσεις των α και β , η τιμή τουR^2 διαφοροποιείται αλλάζοντας π.χ. το εύρος τιμών του X. Με βάση με τα

παραπάνω, για το σφάλμα πρόβλεψης, είναι προτιμότερο να υπολογίσουμε το μέσο τετραγωνικό σφάλμα MSE (Ενότητα 3.2.4.) .

III. Όπως προαναφέραμε, το R^2 δηλώνει το ποσοστό μεταβλητότητας της Y που εξηγείται από το μοντέλο παλινδρόμησης. Αν αντιστρέψουμε τους ρόλους του Y με το X το R^2 θα παραμείνειτο ίδιο. Άρα υψηλή τιμή του R^2 δεν μας δίνει καμία πληροφορία για το αν μια μεταβλητή εξηγεί μια άλλη(με άλλα λόγια η συσχέτιση δεν σημαίνει κατά ανάγκη και αιτιακή σχέση.

IV. Το R^2 δεν πρέπει να χρησιμοποιείται για να συγκρίνουμε δύο μοντέλα στα οποία στο ένα η μεταβλητή απόκρισης είναι Y και στο άλλο ένας μετασχηματισμός της Y. Αρκετές φορές υπολογίζουμε και τον **διορθωμένο συντελεστή προσδιορισμού** του οποίου η ερμηνεία δίνεται στο πολλαπλό γραμμικό μοντέλο:

$$R^2_{adj} = \frac{\frac{\sum_{i=1}^{\nu}(y_i - \hat{y}_i)^2}{\nu - 2}}{\frac{\sum_{i=1}^{\nu}(y_i - \bar{y}_i)^2}{\nu - 1}}$$

3.2.4. Τυπικό Σφάλμα και Μέσο Τετραγωνικό Σφάλμα (MSE).

Η μέση απόκλιση μεταξύ της πραγματικής και της εκτιμούμενης τιμής της μεταβλητής ονομάζεται ***τυπικό σφάλμα της εκτίμησης (standard error of the estimate)***, ή ***τυπικό σφάλμα της παλινδρόμησης***, συμβολίζεται με s ,είναι η θετική τετραγωνική ρίζα της εκτιμήτριας

$$s^2_{Y|X} = \frac{1}{\nu-2} \cdot \sum_{i=1}^{\nu}(y_i - \widehat{y_i})^2 \quad (MSE) \qquad (3.12)$$

και δίνεται από τον τύπο

$$s = \sqrt{\frac{1}{\nu-2} \cdot \sum_{i=1}^{\nu}(y_i - \widehat{y_i})^2} = \sqrt{\frac{SSE}{\nu-2}} \qquad (3.13)$$

Δηλαδή ισχύει: $$s = \sqrt{s^2_{Y|X}} \qquad (3.14)$$

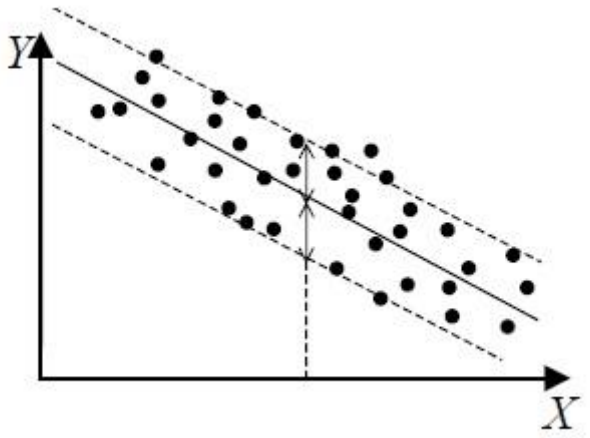
Το $s^2_{Y|X}$ αποτελεί αμερόληπτη εκτιμήτρια του σ^2, υπό την προυπόθεση ότι η τελευταία παραμένει σταθερά για τις διάφορες τιμές του x (ομοσκεδαστικότητα) και καλείται ***μέσο τετραγωνικό σφάλμα*** **(MSE)**.

Εάν το *τυπικό σφάλμα* της εκτίμησης είναι ***μικρό*** τότε οι παρατηρούμενες και οι εκτιμούμενες τιμές δε διαφέρουν πολύ και η ευθεία παλινδρόμησης μας δίνει μια καλή περιγραφή της σχέσης μεταξύ των X και Y. Αν το *τυπικό σφάλμα* της εκτίμησης

είναι μεγάλο τότε δε μπορούμε να ισχυρισθούμε ότι έχουμε μια καλή περιγραφή της σχέσης.

Είναι φανερό, ότι το *τυπικό σφάλμα* της εκτίμησης, είναι ένα μέτρο της *διασποράς* των (x_i, y_i) γύρω από την *ευθεία ελαχίστων τετραγώνων* $\widehat{y_i} = \hat{\alpha} + \hat{\beta}x_i$ (το s^2είναι μια εκτίμηση της διασποράς των σφαλμάτων). Έχει, επομένως, ιδιότητες ανάλογες με αυτές της *τυπικής απόκλισης*. Έτσι, αν φέρουμε δύο ευθείες παράλληλες προς την ευθεία ελαχίστων τετραγώνων και σε κατακόρυφες προς αυτήν αποστάσεις $s, 2s, 3s$ τότε, για μεγάλα ν (μεγαλύτερα του 30), μεταξύ των δύο αυτών ευθειών θα βρίσκεται περίπου το 68%, το 95% και το 99,7% των σημείων του διαγράμματος διασποράς αντίστοιχα.

Στο σχήμα, οι παράλληλες έχουν σχεδιασθεί σε κατακόρυφη απόσταση από την ευθεία ελαχίστων τετραγώνων ίση με $2 \cdot$ s



Σχήμα 3.5 Ευθεία ελαχίστων τετραγώνων και διάγραμμα διασποράς

Κατά αντιστοιχία με τη σχέση (3.4) ορίζουμε ως

$$c_{yy} = \sum_{i=1}^{\nu}(y_i - \bar{y})^2 \qquad (3.15)$$

Εύκολα μπορεί να αποδειχθεί ότι

$$\sum_{i=1}^{\nu}(y_i - \widehat{y_i})^2 = c_{yy} - \frac{{c_{xy}}^2}{c_{xx}} \qquad (3.16)$$

Δηλαδή μπορούμε να έχουμε μια πιο εύχρηστη έκφραση της διασποράς $s^2_{Y|X}$ της σχέσης (3.12) με τη βοήθεια της (3.16),την:

$$s^2_{Y|X} = \frac{1}{\nu-2} \cdot \left\{c_{yy} - \frac{{c_{xy}}^2}{c_{xx}}\right\} = MSE \qquad (3.17)$$

3.2.5. Συντελεστής Συσχέτισης ρ (Correlation Coefficient)

Ο **συντελεστής συσχέτισης (correlation coefficient)** μεταξύ των τυχαίων μεταβλητών X και Y εκφράζει το «βαθμό» στον οποίο μπορούμε να εκτιμήσουμε γραμμικά τη μία τυχαία μεταβλητή όταν γνωρίζουμε την τιμή της άλλης.

$$\rho = {Cov(X,Y)} \Big/ {\{V[X]V[Y]\}^{\frac{1}{2}}} \qquad (3.18)$$

- Όταν $\rho = 0$ οι τυχαίες μεταβλητές είναι ασυσχέτιστες.
- Όταν $\rho = 1$ υπάρχει τέλεια θετική γραμμική συσχέτιση των δύο τυχαίων μεταβλητών.
- Όταν $\rho = -1$ υπάρχει τέλεια αρνητική γραμμική συσχέτιση των δύο τυχαίων μεταβλητών.

Όταν δε γνωρίζουμε το ρ το εκτιμούμε με τη βοήθεια των παρατηρήσεων $(x_i, y_i), i = 1,2, \dots, \nu$ από το **δειγματικό συντελεστή συσχέτισης r**

$$r = \frac{\sum_{i=1}^{\nu}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\{\sum_{i=1}^{\nu}(x_i - \bar{x})^2 \sum_{i=1}^{\nu}(y_i - \bar{y})^2\}^{1/2}} = \frac{c_{xy}}{\sqrt{c_{xx}c_{yy}}} \qquad (3.19)$$

Άρα

$$r^2 = \frac{{c_{xy}}^2}{c_{xx}c_{yy}} \qquad (3.20)$$

Ο δειγματικός συντελεστής συσχέτισης r *εκτιμά το βαθμό στον οποίο οι τυχαίες μεταβλητές* X *και* Y *είναι γραμμικά συσχετισμένες, χωρίς να συνεπάγεται κατά ανάγκη κάποιο είδος αιτιακής σχέσης μεταξύ των* X *και* Y.

Ισχύει ότι $r = \sqrt{R^2}$, ωστόσο υπάρχει μεγάλη διαφορά στην ερμηνεία τους. Ο συντελεστής συσχετίσεως του δείγματος r είναι ένας εκτιμητής του συντελεστή συσχετίσεως στον πληθυσμό ρ και **δεν εξαρτάται από τις μονάδες μέτρησης** των X και Y από την αρχή μέτρησης επάνω στους άξονες, είναι καθαρός αριθμός.

Ισχύει ότι $-1 \leq r \leq 1$ με

- την τιμή -1 να σημαίνει ότι έχουμε πλήρη αρνητική γραμμική συσχέτιση και αντίστοιχα το $+1$ ότι έχουμε πλήρη θετική γραμμική συσχέτιση.
- Όταν $r = \pm 1$ η σχέση είναι *αιτιοκρατική* κι όχι πιθανοκρατική, γιατί γνωρίζοντας την τιμή της μιας τυχαίας μεταβλητής, γνωρίζουμε και την τιμή της άλλης τυχαίας μεταβλητής ακριβώς.

- Η μηδενική τιμή του συντελεστή συσχετίσεως μας δείχνει ότι δεν υπάρχει γραμμική συσχέτιση. Επίσης το ρ είναι μια άγνωστη παράμετρος της συνδυασμένης κατανομής δύο τυχαίων μεταβλητών, ενώ ο συντελεστής προσδιορισμού αναφέρεται στην αναλογία της μεταβλητότητας της Y που ερμηνεύει η μεταβλητή X, η οποία υποθέτουμε ότι δεν είναι τυχαία μεταβλητή. Επιπλέον, ο συντελεστής συσχετίσεως του δείγματος είναι μέτρο μόνο της γραμμικής συσχετίσεως ή εξαρτήσεως δύο μεταβλητών.

Λόγω των ανωτέρω περιορισμών, καθώς και άλλων, η ανάλυση συσχετίσεως έχει περιορισμένη χρήση στην ανάλυση των οικονομικών δεδομένων.

Στο απλό γραμμικό μοντέλο ο συντελεστής προσδιορισμού R^2 συμπίπτει με το τετράγωνο του δειγματικού συντελεστή συσχέτισης r.

$$R^2 = \frac{{c_{xy}}^2}{c_{xx}c_{yy}} = r^2 \qquad (3.21)$$

Η σχέση (3.20) ερμηνεύει το γιατί ο συντελεστής συσχέτισης r χρησιμοποιείται για τον έλεγχο της ύπαρξης ή όχι γραμμικής εξάρτησης μεταξύ των x_i και y_i .

Τέλος εύκολα αποδεικνύεται μέσω των (3.17) και (3.21) , η ακόλουθη χρήσιμη σχέση:

$$s^2_{Y|X} = \frac{c_{yy}}{\nu-2} \cdot \{1 - r^2\} = MSE \qquad (3.22)$$

3.2.6. Έλεγχοι Υποθέσεων

3.2.6. Α Εισαγωγή

Σε κάποιο πληθυσμό, μία παράμετρος θ έχει την τιμή θ_0. Κάποια στιγμή, επιδρούμε με κάποιο τρόπο σε όλα τα στοιχεία του πληθυσμού και το ερώτημα που προκύπτει είναι αν η παράμετρος θ εξακολουθεί (μετά την επίδραση) να έχει την τιμή θ_0 ή αν είναι $\theta \neq \theta_0$. Έτσι έχουμε τις υποθέσεις:

$$H_0: \quad \theta = \theta_0$$
$$H_1: \quad \theta \neq \theta_0$$

Όπου H_0 ,μηδενική υπόθεση , είναι το όνομα της πρώτης υπόθεσης, η οποία αναφέρεται στην κατάσταση του πληθυσμού πριν την επίδραση και H_1, εναλλακτική υπόθεση, είναι το όνομα της δεύτερης υπόθεσης, η οποία αναφέρεται στην κατάσταση μετά την επίδραση.

Γενικά, κατά τον έλεγχο των υποθέσεων χρησιμοποιούμε την έκφραση «Να ελεγχθεί η υπόθεση H_0 έναντι της εναλλακτικής H_1.» Ο έλεγχος υποθέσεων γίνεται χρησιμοποιώντας κάποιο δείγμα από τον πληθυσμό, ενώ όπως και στα διαστήματα εμπιστοσύνης, προκαθορίζεται επίσης, κάποια **στάθμη σημαντικότητας α (σ.σ.).** Το α ονομάζεται και **επίπεδο σημαντικότητας (ε.σ.)**. Αν γ **ο βαθμός εμπιστοσύνης (β.ε.)**, τότε ισχύει: $\alpha = 1 - \gamma$ ή $\gamma = 1 - \alpha$.

3.2.6. Β Έλεγχοι Υποθέσεων για τις παραμέτρους α και β

I. Στατιστικά ελέγχου t-test

Οι εκτιμήσεις των α και β που λαμβάνουμε με την μέθοδο ελαχίστων τετραγώνων βασίζονται στα συγκεκριμένα δεδομένα που διαθέτουμε. Συχνά λοιπόν ενδιαφερόμαστε να ελέγξουμε τις ακόλουθες υποθέσεις, σε ε.σ. έστω a:

$$H_0: \beta = 0 \text{ έναντι της } H_1: \beta \neq 0$$

$$H_0: \alpha = 0 \text{ έναντι της } H_1: \alpha \neq 0$$

- Με τον πρώτο έλεγχο θέλουμε να διαπιστώσουμε αν πράγματι αύξηση κατά μια μονάδα της X σημαίνει και μεταβολή της αναμενόμενης τιμής της Y.
- Στο δεύτερο έλεγχο θέλουμε να δούμε κατά πόσο η αναμενόμενη τιμή της Y είναι 0 όταν $X = 0$. Πολλές φορές η τιμή αυτή δεν έχει ερμηνεία, διότι η τιμή $X = 0$ δεν παρατηρείται ποτέ στην πράξη.

Αν n το μέγεθος του δείγματος τότε τα στατιστικά ελέγχου με βάση τις μηδενικές υποθέσεις είναι:

$$T_2 = \frac{\hat{\beta} - \beta}{se(\hat{\beta})} = \frac{\hat{\beta} - 0}{\sqrt{\sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2}} \approx \frac{\hat{\beta}}{s_\beta} \sim St(n-2) \qquad (3.23)$$

$$T_1 = \frac{\hat{a} - a}{se(\hat{a})} = \frac{\hat{a} - 0}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)}} \approx \frac{\hat{a}}{s_a} \sim St(n-2) \qquad (3.24)$$

Όπου

$$s_{\alpha}^2 = s_{Y|X}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}\right) \quad (3.25)$$

$$s_{\beta}^2 = \frac{s_{Y|X}^2}{(n-1)s_X^2} \quad (3.26)$$

Δηλαδή ακολουθούν την t-κατανομή Student με $n-2$ βαθμούς ελευθερίας.

Υπολογίζουμε λοιπόν τα παραπάνω δύο στατιστικά ελέγχου T_1και T_2και η P-τιμή των ελέγχων είναι δύο φορές η πιθανότητα της περιοχής της t_{n-2}δεξιά από τις τιμές των στατιστικών ελέγχων.

Ισοδύναμα θα μπορούσαμε να είχαμε κατασκευάσει συμμετρικά $(1-\alpha)\%$ Δ.Ε. για τα α και β και να ελέγξουμε αν η τιμή 0 ανήκει σ' αυτά τα Δ.Ε.:

$(\hat{\beta} \pm t_{n-2,a/2} \cdot s_{\beta})$ όπου $(1-\alpha)\%$ Δ.Ε. για το β (3.27)

Και $(\hat{\alpha} \pm t_{n-2,a/2} \cdot s_{\alpha})$ όπου $(1-\alpha)\%$ Δ.Ε. για το α (3.28)

II. Στατιστικά ελέγχου F-test

Ένας άλλος έλεγχος που συνήθως εξετάζουμε στο μοντέλο παλινδρόμησης γνωστός με την ονομασία F-test είναι και ο παρακάτω, ο οποίος ελέγχει κατά πόσο το προτεινόμενο μοντέλο $y = \alpha + \beta x$ διαφέρει από το σταθερό $y=\alpha$.

Στη απλή γραμμική παλινδρόμηση ο εν λόγω έλεγχος είναι ισοδύναμος με τον έλεγχο για το β που είδαμε πριν για $H_0: \beta = 0$.

Δηλαδη αποδεικνύεται ότι:

$$F_{1,n-2} = {t_{n-2}}^2$$

Πιο συγκεκριμένα από τις σχέσεις (3.11) υπολογίζουμε το στατιστικό ελέγχου:

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{\frac{SSR}{\sigma^2}}{\frac{SSE}{\sigma^2(n-2)}} \quad (3.29)$$

$$\text{Ή} \qquad F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

το οποίο κάτω από την μηδενική υπόθεση $H_0: \beta = 0$ έναντι της $H_1: \beta \neq 0$ ακολουθεί την $F_{1,n-2}$ γιατί ισχύει :

$$\frac{SSE}{\sigma^2} \sim X^2(n-2) \ \ και \ \frac{SSR}{\sigma^2} \sim X^2(1)$$

Υπολογίζουμε λοιπόν την τιμή του στατιστικού ελέγχου F και η p-τιμή είναι πιθανότητα της περιοχής της $F_{1,n-2}$ δεξιά από το F που παρατηρούμε. Η μηδενική υπόθεση H_0 απορρίπτεται όταν η $p-$τιμή του ελέγχου, δηλαδή η $P(F_{1,n-2} > F)$ με F την υπολογισμένη τιμή της ελεγχοσυνάρτησης, είναι μικρή.

Η διαφορετική χρησιμότητα των δύο ελέγχων θα φανεί όταν υπάρχουν περισσότερες από μία ανεξάρτητες μεταβλητές στο μοντέλο.

III. Πρόβλεψη και διάστημα εμπιστοσύνης της $E(Y)$ για δοσμένη τιμή του x

Η σημειακή εκτίμηση της $E(Y)$ είναι η $\hat{y} = \hat{\alpha} + \hat{\beta}x$

Αποδεικνύεται ότι

$$E(\hat{Y}|X = x) = a + \beta x \qquad (3.30)$$

$$Var(\hat{Y}) = \sigma^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \qquad (3.31)$$

Αν $\varepsilon \sim N(0, \sigma^2)$ αποδεικνύεται ότι

$$(\hat{Y}|X = x) \sim N\left(a + \beta x \, , \sigma^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\right)$$

Τέλος ένα συμμετρικό $(1-\alpha)\%$ Δ.Ε. της τιμής, έστω y,της δεσμευμένης μέσης τιμής της μεταβλητής απόκρισης Y όταν η επεξηγηματική μεταβλητή X ισούται με x, δηλαδή της τιμής της ποσότητας $E(Y|X = x) = a + \beta x$, είναι το

$$\left(\hat{y} \pm t_{n-2,a/2} \cdot s_{Y|X}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right) \qquad (3.32)$$

Το παραπάνω διάστημα εμπιστοσύνης καλείται και **διάστημα μέσης πρόβλεψης (mean prediction interval)** .

Εναλλακτικά, θα μπορούσαμε να κατασκευάζαμε το παρακάτω συμμετρικό $(1-\alpha)\%$ Δ.Ε. για την τιμή, έστω y, της μεταβλητής απόκρισης Y όταν η επεξηγηματική μεταβλητή X ισούται με x είναι το

$$\left(\hat{y} \pm t_{n-2,a/2} \cdot s_{Y|X}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right) \qquad (3.33)$$

Το εν λόγω διάστημα, καλείται **διάστημα (ατομικής) πρόβλεψης((individual) prediction interval)** και αποτελεί ένα συμμετρικό $(1-\alpha)\%$ Δ.Ε. για την τιμή, έστω y,της τυχαίας μεταβλητής $Y = \alpha + \beta \cdot X + \varepsilon$

Παρατηρήσεις

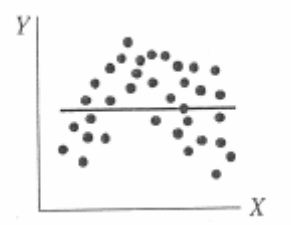
- Το πρώτο Δ.Ε. παρέχει πληροφορία για τον βαθμό αβεβαιότητας που έχουμε για την εκτίμηση της δεσμευμένης μέσης τιμής $E(Y|X=x)$.Το δεύτερο διάστημα παρέχει πληροφορία για τον βαθμό αβεβαιότητας που έχουμε για την τιμή που θα πάρει η τυχαία μεταβλητή Y όταν $X = x$.Το δεύτερο δηλαδή διάστημα λαμβάνει επιπλέον υπόψιν, πέραν της αβεβαιότητας που έχουμε από την εκτίμηση της δεσμευμένης μέσης τιμής και την μεταβλητότητα της δεσμευμένης κατανομής $(Y|(X=x)$. Χρησιμοποιώντας δηλαδή το διάστημα μέσης πρόβλεψης γενικά υποεκτιμούμε την αβεβαιότητά μας για την χρήση της τιμής ως εκτιμήτρια της τιμής που θα πάρει η τυχαία μεταβλητή Y όταν $X = x$.
- Το πρώτο διάστημα (3.32) θεωρείται κατάλληλο και χρησιμοποιείται όταν θέλουμε να κατασκευάσουμε διάστημα εμπιστοσύνης για την τιμή, έστω y,της μεταβλητής απόκρισης Y δοσμένης **μίας εκ των ήδη παρατηρηθέντων** τιμών της επεξηγηματικής μεταβλητής X, για αυτό και λέγεται επίσης και *διάστημα εμπιστοσύνης προσαρμοσμένων (fitted) τιμών*. Αντιθέτως αν θέλουμε να χρησιμοποιήσουμε μια **μελλοντική παρατήρηση**, έστω x, της επεξηγηματικής μεταβλητής X τότε για την κατασκευή του διαστήματος εμπιστοσύνης της τιμής της μεταβλητής απόκρισης Y χρησιμοποιούμε το *διάστημα (ατομικής) πρόβλεψης*.

IV. Ερμηνεία του ελέγχου της υπόθεσης $H_0: \beta = 0$ έναντι της $H_1: \beta \neq 0$ για την κλίση της ευθείας παλινδρόμησης $Y = \alpha + \beta \cdot X + \varepsilon$.

a) Όταν **<u>δεν απορρίπτεται</u>** η μηδενική υπόθεση, τότε συμβαίνει ένα από τα

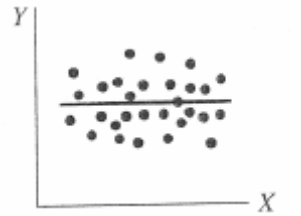
παρακάτω:

- Η σχέση μεταξύ X και Y **δεν είναι** γραμμική



Σχήμα 3.6 Μη γραμμική σχέση X και Y

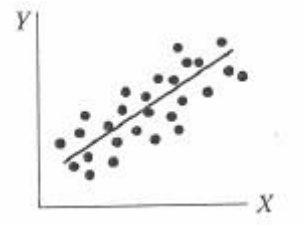
- Πρόκειται για το μοντέλο $E(Y/X) = E(Y) = \alpha$. Δηλαδή, για την περίπτωση όπου **η X δεν συνεισφέρει στην πρόβλεψη της $E(Y/X)$** . Έτσι, η εκτίμηση $\hat{Y} = \bar{y} + \hat{\beta} \cdot (X - \bar{x})$ προβλέπει τη μέση τιμή της Y όσο και η $\hat{Y} = \bar{y}$.



Σχήμα 3.7 Σε γραμμικό μοντέλο η X δε συνεισφέρει στην πρόβλεψη της $E(Y/X)$

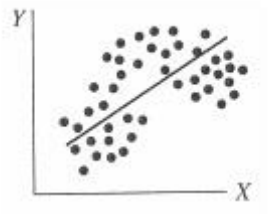
b) Όταν **απορρίπτεται** η μηδενική υπόθεση, τότε συμβαίνει ένα από τα παρακάτω:

- Η X, μέσω του γραμμικού μοντέλου, συνεισφέρει στην πρόβλεψη της $E(Y/X)$ Δηλαδή, η εκτίμηση $\hat{Y} = \bar{y} + \hat{\beta} \cdot (X - \bar{x})$ είναι καλύτερη (στατιστικά πιο σημαντική) από την $\hat{Y} = \bar{y}$.



Σχήμα 3.8 Σε γραμμικό μοντέλο, η X συνεισφέρει στην πρόβλεψη της $E(Y/X)$

- Το γραμμικό μοντέλο είναι μόνο μια καλή γραμμική προσέγγιση, μιας μη γραμμικής, στην πραγματικότητα, σχέσης.



Σχήμα 3.9 Μη γραμμική σχέση

Συνοψίζοντας: Είτε απορρίπτεται η μηδενική υπόθεση είτε όχι, το γραμμικό μοντέλο μπορεί να μην είναι κατάλληλο. Κάποιο άλλο μοντέλο (μη γραμμικό), μπορεί να περιγράφει τη σχέση μεταξύ X και Y καλύτερα

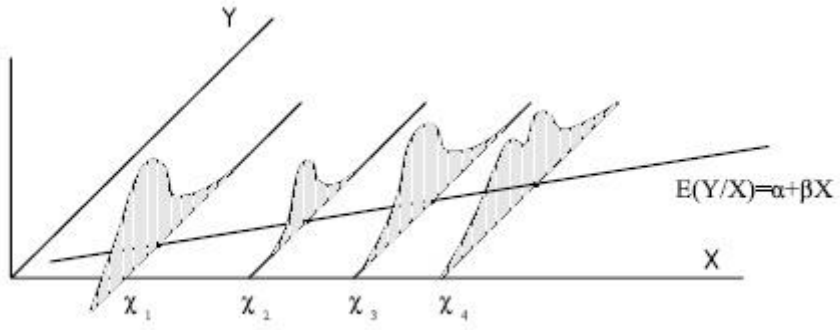
3.3 Προϋποθέσεις-Παραδοχές για την Εφαρμογή του Απλού Γραμμικού Μοντέλου $\boldsymbol{Y = \alpha + \beta X + \varepsilon}$

Η γενική υπόθεση-παραδοχή που κάνουμε για ένα μοντέλο παλινδρόμησης (γραμμικό ή όχι), είναι ότι η μεταβλητή X μετράται χωρίς σφάλμα και ότι η Y, για κάθε επίπεδο x_iτης X, είναι τυχαία μεταβλητή με πεπερασμένη μέση τιμή και διασπορά ενώ οι τιμές της είναι ασυσχέτιστες (uncorrelated).

Για το ***απλό γραμμικό μοντέλο*** κάνουμε επιπλέον τις ακόλουθες υποθέσεις-παραδοχές:

Υπόθεση 1: Γραμμικότητα (Linearity)

Η κατανομή της Y έχει, για τα διάφορα επίπεδα x_i, $i = 1,2,\dots,\nu$ της X, μέση τιμή $E(Y|X = x_i) = \alpha + \beta \cdot x_i$ ή $E(Y|X) = \alpha + \beta X$ όπου, α και β παράμετροι που εκτιμώνται από το δείγμα (x_i, y_i), $i = 1,2,\dots,\nu$. Δηλαδή, υποθέτουμε ότι οι μέσες τιμές της Y, για τα διάφορα επίπεδα της X, είναι γραμμικές συναρτήσεις της X (ότι βρίσκονται δηλαδή σε ευθεία γραμμή).Σημειώνουμε ότι στο μοντέλο
$Y = \alpha + \beta \cdot X + \varepsilon$, τυχαίες μεταβλητές είναι μόνο οι Y και ε .



Σχήμα 3.10

Υπόθεση 2: Ομοσκεδαστικότητα-Σταθερότητα Διασποράς (Homoscedasticity - Variance Stability)

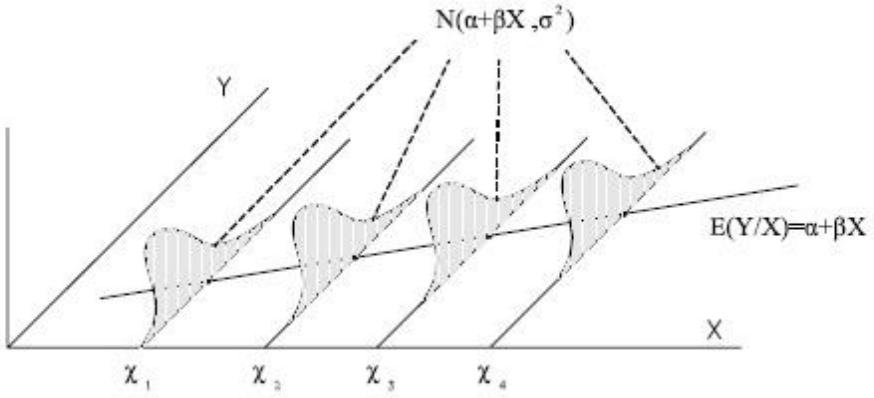
Οι κατανομές της Y έχουν ίδια διασπορά για όλα τα επίπεδα της X, δηλαδή, $Var(Y/X = x_i) = \sigma^2$. Ένα παράδειγμα παραβίασης της υπόθεσης αυτής *(heteroscedasticity)* φαίνεται στο προηγούμενο σχήμα 3.10. όπου η διασπορά της Y, π.χ. στο επίπεδο x_1, είναι μεγαλύτερη από τη διασπορά της Y στο επίπεδο x_2 .

Υπόθεση 3: Ανεξαρτησία (Independence)

Αν, εκτός από την εκτίμηση των συντελεστών παλινδρόμησης, θέλουμε να εκτιμήσουμε διαστήματα εμπιστοσύνης (confidence intervals) ή να κάνουμε ελέγχους στατιστικών υποθέσεων (null hypotheses tests) με το t ή F κριτήριο, τότε οι τιμές της Y που αντιστοιχούν στα διάφορα επίπεδα της X πρέπει επιπλέον να είναι και ανεξάρτητες (independent) μεταξύ τους.

Υπόθεση 4: Κανονικότητα (Normality)

Αν, εκτός από την εκτίμηση των συντελεστών παλινδρόμησης, θέλουμε να εκτιμήσουμε διαστήματα εμπιστοσύνης (confidence intervals) ή να κάνουμε ελέγχους στατιστικών υποθέσεων (null hypotheses tests) με το t ή F κριτήριο, τότε οι τιμές της Y πρέπει επιπλέον να ακολουθούν την κανονική κατανομή.Δηλαδή, η κατανομή της Y για όλα τα επίπεδα της X είναι κανονική.



Σχήμα 3.11

<u>Υποθέσεις ως προς τα σφάλματα</u>

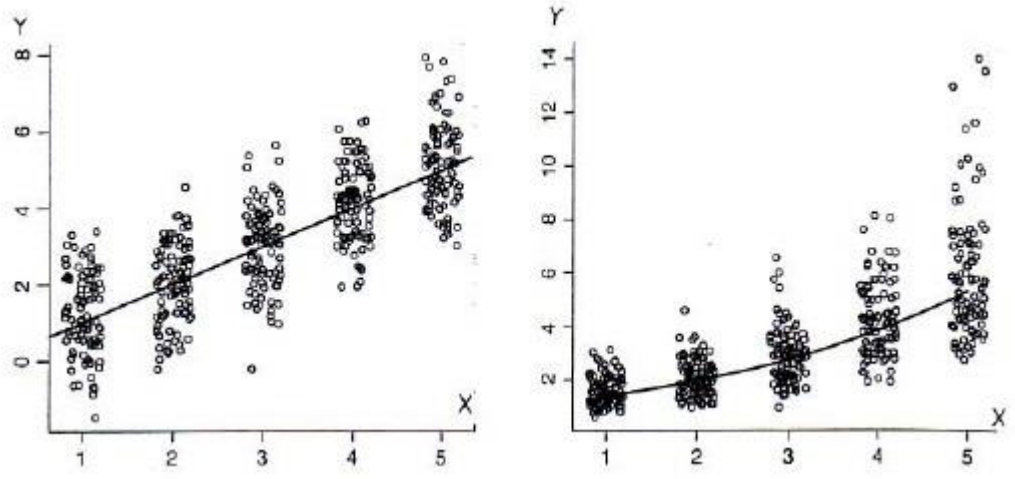
Με βάση τις παραπάνω υποθέσεις για την τυχαία μεταβλητή Y, για την τυχαία μεταβλητή $\varepsilon = Y - (\alpha + \beta X)$ (δηλαδή για τα σφάλματα-residuals) δεχόμαστε ότι:

1. $\varepsilon \sim N(0, \sigma^2)$
2. Τα σφάλματα (υπόλοιπα) πρέπει να είναι τυχαία.
3. Τα σφάλματα πρέπει να είναι ασυσχέτιστα.
4. Οι τιμές της ε που αντιστοιχούν στα διάφορα επίπεδα της X είναι μεταξύ τους ανεξάρτητες.
5. Η διακύμανση των υπολοίπων πρέπει να είναι ομοιογενής (σταθερή)σε όλο το εύρος των πραγματικών τιμών Y. Εναλλακτικά, μπορούμε να εξετάσουμε αν η διακύμανση των υπολοίπων είναι σταθερή σε όλο το εύρος των θεωρητικών τιμών $\hat{Y}$, επειδή οι τιμές της ε και της Y συνήθως συσχετίζονται, ενώ οι τιμές της ε και $\hat{Y}$ της όχι.

Στη συνέχεια, παρουσιάζουμε ορισμένες μεθόδους (γραφικές κυρίως) για τον έλεγχο των παραπάνω προϋποθέσεων-παραδοχών προσαρμογής του απλού γραμμικού μοντέλου.

<u>Οι παραδοχές αυτές αποτελούν την αναγκαία μαθηματική (πιθανοθεωρητική) βάση για την εφαρμογή μεθόδων της στατιστικής συμπερασματολογίας</u> (π.χ. έλεγχοι υποθέσεων, διαστήματα εμπιστοσύνης). Ο έλεγχος επομένως αυτών των παραδοχών είναι αναγκαίος προκειμένου να *<u>αποφεύγουμε λανθασμένες διαδικασίες εξαγωγής συμπερασμάτων για τον πληθυσμό.</u>*

Ένας πρώτος, άμεσος, έλεγχος μπορεί να γίνει με προσεκτική παρατήρηση του ***διάγραμματος διασποράς*** του δείγματος. Ας δούμε δύο παραδείγματα:



Σχήμα 3.12 Διαγράμματα διασποράς για έλεγχο υποθέσεων απλού γραμμικού μοντέλου

Στο πρώτο διάγραμμα διασποράς (αριστερά) φαίνεται ότι για όλα τα επίπεδα της X:

- Οι κατανομές της Y είναι συμμετρικές και έχουν σταθερή διασπορά
- Οι αναμενόμενες μέσες τιμές της Y βρίσκονται σε ευθεία γραμμή.

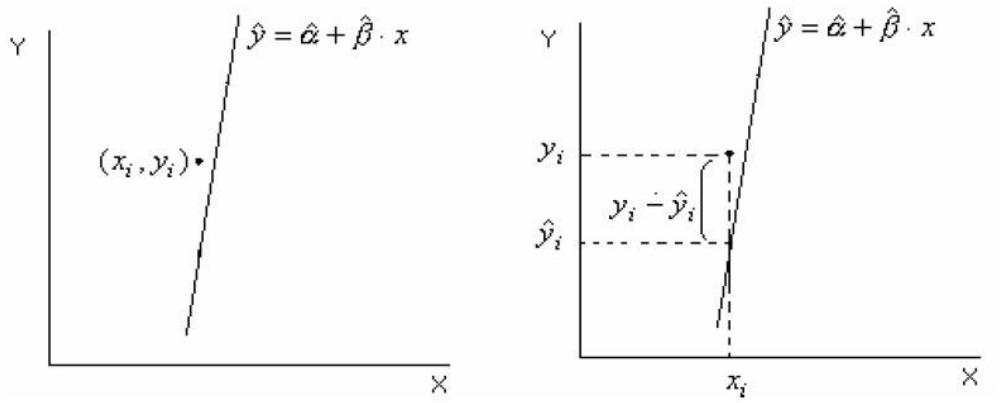
Στο δεύτερο *διάγραμμα διασποράς* (δεξιά) φαίνεται ότι:

- οι κατανομές της Y για τα διάφορα επίπεδα της X δεν είναι συμμετρικές και ούτε έχουν σταθερή διασπορά. Μάλιστα, φαίνεται ότι αυξανομένου του X αυξάνεται η διασπορά καθώς και η ασσυμετρία (θετική) της κατανομής του Y
- οι αναμενόμενες μέσες τιμές της Y για τα διάφορα επίπεδα της X δεν βρίσκονται σε ευθεία γραμμή αλλά σε καμπύλη.

Ας δούμε πιο αναλυτικά, ανά υπόθεση, πώς μπορούμε να διαπιστώσουμε και να αντιμετωπίσουμε πιθανές παραβιάσεις.

3.3.1. Γραμμικότητα (Linearity)

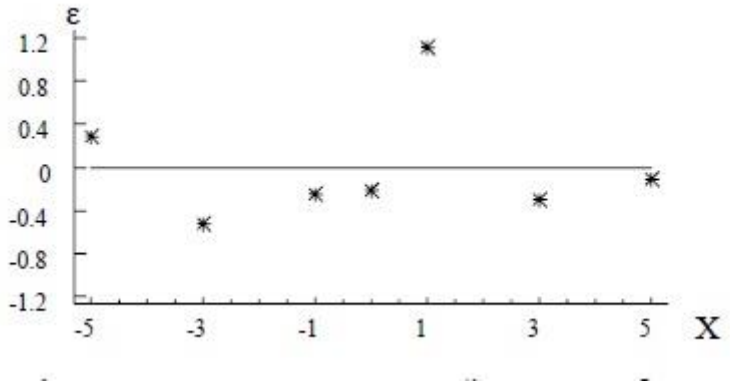
Ένας πρώτος έλεγχος της γραμμικότητας μπορεί να γίνει γραφικά με το *διάγραμμα διασποράς*. Είναι όμως δυνατόν, ιδίως όταν η κλίση της ευθείας παλινδρόμησης που προσεγγίζει τα δεδομένα είναι μεγάλη, να μας δίνεται η εντύπωση ότι τα σημεία $(x_i, y_i),\ i = 1,2,\dots,\nu$ είναι κοντά στην ευθεία παλινδρόμησης ενώ στην πραγματικότητα δεν είναι, όπως φαίνεται και στο παρακάτω σχήμα 3.13.



Σχήμα 3.13 Ευθεία παλινδρόμησης με μεγάλη κλίση.

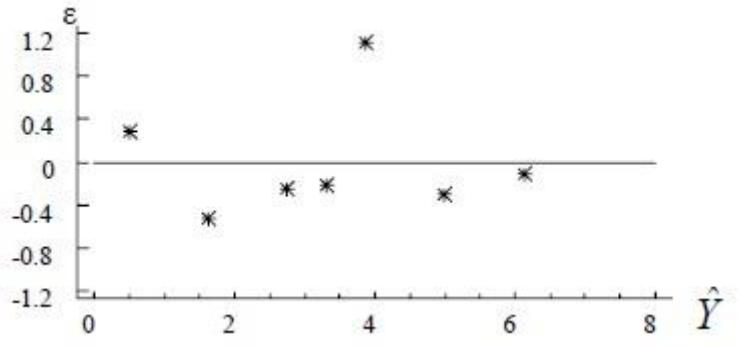
Για το λόγο αυτό, συνήθως, χρησιμοποιούμε τα **διαγράμματα υπολοίπων (residual plots)** όπου, αντί των $(x_i, y_i), i = 1,2, \dots, \nu$ αναπαρίστανται γραφικά τα$(x_i, \hat{\varepsilon}_i)$ ή τα $(y_i, \hat{\varepsilon}_i)$ (όπου $\hat{\varepsilon}_i = y_i - \hat{y}_i$ τα υπόλοιπα-σφάλματα).

Αν στο διάγραμμα υπολοίπων, τα σημεία $(x_i, \hat{\varepsilon}_i)$ ή τα $(y_i, \hat{\varepsilon}_i)$ δεν ακολουθούν κάποιο πρότυπο (κάποια συστηματική τάση) αλλά είναι τυχαία διεσπαρμένα σε μια οριζόντια ζώνη γύρω από την ευθεία $\varepsilon = 0$, τότε η επιλογή γραμμικού μοντέλου δικαιολογείται.



Σχήμα 3.14 Διάγραμμα υπολοίπων $(x_i, \hat{\varepsilon}_i)$.

Τα διαγράμματα υπολοίπων συνήθως παρουσιάζουν την ίδια εικόνα και όταν τα υπόλοιπα $\hat{\varepsilon}_i$ παρασταθούν γραφικά συναρτήσει των προσαρμοσμένων τιμών $\hat{y}_i$.



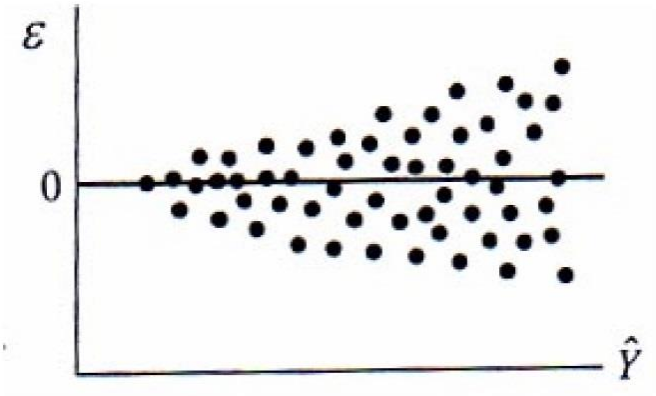
Σχήμα 3.15 Διάγραμμα υπολοίπων $(y_i, \hat{e}_i)$

Όταν διαπιστώνεται ότι η σχέση μεταξύ X και Y είναι μη γραμμική, σε αρκετές περιπτώσεις είναι δυνατόν, με κατάλληλους μετασχηματισμούς στα X ή/και στα Y να προκύψει γραμμική σχέση. Έχουμε έτσι τη δυνατότητα να αξιοποιήσουμε τη στατιστική θεωρία του γραμμικού μοντέλου και σε μη γραμμικά μοντέλα (αφού, αντιστρέφοντας στη συνέχεια τις μετασχηματισμένες μεταβλητές, μπορούμε να πάρουμε τα ζητούμενα συμπεράσματα για τις αρχικές). Στην ενότητα 3.4 δίνουμε παραδείγματα τέτοιων μετασχηματισμών. Γενικά, η στατιστική μελέτη μη γραμμικών μοντέλων, με εξαίρεση τα πολυωνυμικά, παραμένει δύσκολο και ανοικτό πρόβλημα.

3.3.2. Ομοσκεδαστικότητα ή Σταθερότητα Διασποράς (Homoscedasticity-Variance Stability)

Ένας πρώτος έλεγχος της σταθερότητας ή μη της διασποράς της Y (ή της ε) για τα διάφορα επίπεδα της X μπορεί να γίνει με το διάγραμμα διασποράς και τα διαγράμματα υπολοίπων. Τα ζεύγη αυτών των τιμών δεν πρέπει να εμφανίζουν κάποιο συστηματικό τρόπο συμπεριφοράς. Αν για παράδειγμα, το διάγραμμα υπολοίπων έχει μορφή τραπεζίου (ανοιχτής βεντάλιας), όπως το παρακάτω σχήμα 3.16. , η πιο πιθανή αιτία αυτής της διαταραχής είναι η μη σταθερότητα της διασποράς των τυχαίων σφαλμάτων ε. Σε πολλές οικονομικές και εμπορικές εφαρμογές η μεταβολή της διασποράς σ^2 με το X ή με το $\hat{Y}$ δίνει διαγράμματα υπολοίπων μορφής τραπεζίου (αυξανομένου του X ή του $\hat{Y}$, αυξάνει το σ^2 ή αντιστρόφως). Αυτό συμβαίνει διότι τέτοιες εφαρμογές ακολουθούν πολλαπλασιαστικά μοντέλα όπου $\sigma_Y^2 = [E(Y)]^2 \cdot \sigma^2$ και σ^2 η διασπορά των σφαλμάτων ε. Επίσης, ανάλογα διαγράμματα υπολοίπων δίνουν μεταβλητές που

μετρούν αριθμό συμβάντων στη μονάδα χρόνου, χώρου, μήκους, κ.τλ. δηλαδή μεταβλητές που ακολουθούν κατανομή Poisson.



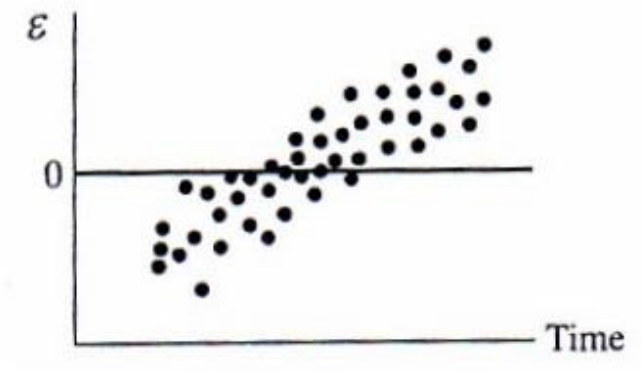
Σχήμα 3.16 Διαγράμματα υπολοίπων μορφής τραπεζίου

Αν από τα *διαγράμματα υπολοίπων* δημιουργούνται υπόνοιες ότι δεν έχουμε σταθερές διασπορές, μπορούμε να ελέγξουμε στατιστικά αν υπάρχει σημαντική διαφορά στις διασπορές ή όχι εφόσον για τα διάφορα επίπεδα της X έχουμε περισσότερες της μιας παρατηρήσεις. Μπορούμε, επίσης, να ταξινομήσουμε τις παρατηρήσεις σε αύξουσα σειρά των X, να τις χωρίσουμε σε δύο ή περισσότερες ομάδες και να ελέγξουμε στατιστικά αν οι ομάδες έχουν σημαντική διαφορά στις διασπορές ή όχι.
Όταν διαπιστώνεται μη σταθερότητα διασπορών μπορούμε, σε αρκετές περιπτώσεις, να αντιμετωπίσουμε το πρόβλημα με κατάλληλους μετασχηματισμούς στις μεταβλητές. Στην ενότητα 3.4 δίνονται παραδείγματα τέτοιων μετασχηματισμών.

3.3.3. Ανεξαρτησία (Independence)

Εξαρτημένα Y εμφανίζονται συνήθως όταν παίρνουμε παρατηρήσεις από την ίδια πειραματική μονάδα σε διαφορετικές χρονικές στιγμές (π.χ. μετράμε την πίεση ή το βάρος του ιδίου ατόμου ανά εβδομάδα). Επίσης, σε περιπτώσεις όπου χρησιμοποιούνται μηχανές (όργανα μέτρησης, κ.τλ) που αλλάζει η απόδοσή τους με τη χρήση ή ο χειριστής βελτιώνεται (ή χειροτερεύει) με την πάροδο του χρόνου. Είναι επομένως χρήσιμο, όταν έχουμε πειραματικά δεδομένα που παίρνονται με χρονική σειρά, να κάνουμε ένα διάγραμμα υπολοίπων ως προς το χρόνο έστω και αν ο χρόνος δεν χρησιμοποιείται ως μεταβλητή στο μοντέλο. Κατασκευάζουμε ένα διάγραμμα υπολοίπων σε σχέση με την σειρά των δεδομένων, στο οποίο δεν πρέπει να

παρουσιάζεται κάποια σχέση και τα υπόλοιπα να συμπεριφέρονται τυχαία. Αν το διάγραμμα υπολοίπων έχει τη μορφή του παρακάτω σχήματος 3.17 τότε είναι πιθανόν να υπάρχει στοχαστική εξάρτηση μεταξύ των σφαλμάτων. Στη συνέχεια, πρέπει να ελέγξουμε στατιστικά την υπόνοια αυτή με το **Durbin-Watson test**. Αν διαπιστωθεί εξάρτηση των τιμών της Y τότε για την προσαρμογή κατάλληλου μοντέλου και την εξαγωγή στατιστικών συμπερασμάτων πρέπει να χρησιμοποιηθούν ειδικές μέθοδοι.



Σχήμα 3.17 Πιθανότητα ύπαρξης στοχαστικής εξάρτησης μεταξύ των σφαλμάτων.

3.3.4. Κανονικότητα (Normality)

Η κανονικότητα μπορεί να ελεγχθεί με διάφορους τρόπους όπως:

- Με *ιστόγραμμα*
- Με *φυλλογράφημα (steam and leaf plot)*
- Με *θηκόγραμμα (box plot)*
- Με *διάγραμμα πιθανοτήτων (normal probability plot)*
- Με στατιστικούς ελέγχους καλής προσαρμογής (*goodness–of-fit test*) όπως *Kolmogorov-Smirnov test* ή X^2*test.*

Όταν διαπιστώνεται παραβίαση της κανονικότητας μπορούμε, σε αρκετές περιπτώσεις, να αντιμετωπίσουμε το πρόβλημα με κατάλληλους μετασχηματισμούς στις μεταβλητές. Στην ενότητα 3.4 δίνονται παραδείγματα τέτοιων μετασχηματισμών.

3.3.5 Έλεγχος Ακραίων Παρατηρήσεων (Outliers)

Πέραν των παραπάνω υποθέσεων-παραδοχών, είναι χρήσιμο να ελέγχουμε την ύπαρξη ή μη *ακραίων παρατηρήσεων* (outliers). Οι *ακραίες παρατηρήσεις* μπορούν να ανιχνευθούν αποτελεσματικά με το *θηκόγραμμα* των παρατηρήσεων ή και με το

διάγραμμα υπολοίπων. Αν διαπιστωθεί ακραία παρατήρηση, πρέπει πρώτα να ερευνηθεί αν οφείλεται σε λανθασμένη παρατήρηση ή πιθανόν σε απότομη στιγμιαία διαταραχή του συστήματος που παρατηρούμε. Αν αυτό συμβαίνει, πρέπει να παραληφθεί από το δείγμα. Αν όμως η ακραία παρατήρηση ανήκει στον πληθυσμό είναι λάθος να παραληφθεί από το δείγμα. Η γενική αρχή που πρέπει να τηρούμε είναι ότι ποτέ δεν απορρίπτουμε μια ακραία παρατήρηση αν δεν είμαστε βέβαιοι ότι πρόκειται για λάθος ή απότομη στιγμιαία διαταραχή. Έγκυρες ακραίες παρατηρήσεις μπορεί να αποδειχθούν οι πλέον ενδιαφέρουσες.

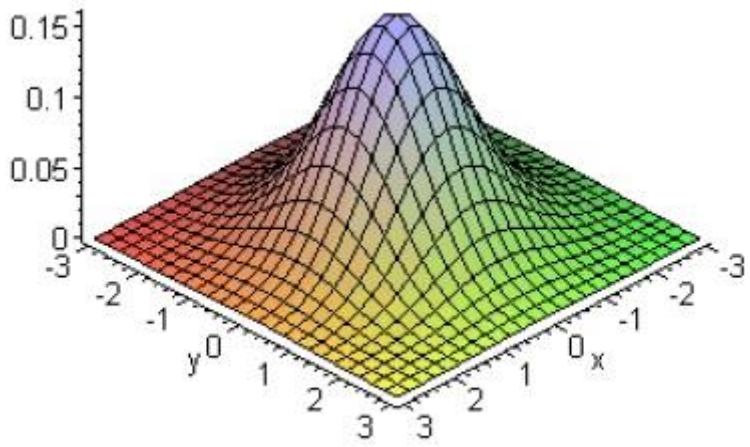
3.3.6. Υπόθεση για την Εφαρμογή του Απλού Γραμμικού Μοντέλου Παλινδρόμησης σε μη Πειραματικά Δεδομένα

Για την ανάπτυξη της στατιστικής θεωρίας του απλού γραμμικού μοντέλου υποθέσαμε ότι η μεταβλητή X **δεν είναι τυχαία** (μετράται χωρίς σφάλμα) και ότι τυχαίες μεταβλητές είναι μόνο οι Y και ε. Αυτή η υπόθεση ικανοποιείται στις **πειραματικές έρευνες** όπου ο ερευνητής ελέγχει (καθορίζει) τις τιμές της X και παρατηρεί πώς οι μεταβολές στις τιμές της X αντανακλώνται στην Y.

Σε **μη πειραματικές έρευνες (δειγματοληψίες)**, όπου ο ερευνητής επιλέγει ένα τυχαίο δείγμα $(x_i, y_i),\ i = 1,2,\dots,\nu$, δηλαδή, όταν όχι μόνο η Y αλλά και η X είναι τυχαία μεταβλητή, τότε με την υπόθεση ότι η από κοινού κατανομή των X και Y είναι **διδιάστατη κανονική κατανομή,** μπορούμε και πάλι να εφαρμόσουμε τη θεωρία του απλού γραμμικού μοντέλου και να υπολογίσουμε την ευθεία ελαχίστων τετραγώνων της Y πάνω στην X ή της X πάνω στην Y διότι από τη θεωρία πιθανοτήτων είναι γνωστό ότι οι δεσμευμένες κατανομές της Y δεδομένης της X και της X δεδομένης της Y είναι κανονικές με

$$\mu_{Y/X} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X) \qquad (3.34)$$

$$\text{και} \quad \mu_{X/Y} = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y) \qquad (3.35) \text{ αντίστοιχα.}$$



Σχήμα 3.18 Διδιάστατη κανονική κατανομή των X και Y

3.4 Μετασχηματισμοί Σταθεροποίησης Διασπορών – Κανονικοποίησης - Γραμμικοποίησης

Κατά τη διερεύνηση της σχέσης μεταξύ δύο μεταβλητών X και Y για την εφαρμογή του γραμμικού μοντέλου παλινδρόμησης, πολλές φορές, διαπιστώνεται παραβίαση μιας ή και περισσότερων εκ των προϋποθέσεων-παραδοχών εφαρμογής της αντίστοιχης στατιστικής θεωρίας. Σε αρκετές περιπτώσεις, μπορούμε να αντιμετωπίσουμε αυτά τα προβλήματα με κατάλληλους μετασχηματισμούς των μεταβλητών.

Πιο συγκεκριμένα, υπάρχουν τρεις βασικοί λόγοι για την αναζήτηση κατάλληλων μετασχηματισμών των μεταβλητών:

1. Για τη **σταθεροποίηση των διασπορών**, όταν παραβιάζεται η παραδοχή της ομοσκεδαστικότητας. Δηλαδή, όταν οι διασπορές της εξαρτημένης μεταβλητής Y δεν είναι ίσες για τα διάφορα επίπεδα της X.
2. Για την **κανονικοποίηση**, όταν οι κατανομές της εξαρτημένης μεταβλητής Y για τα διάφορα επίπεδα της X δεν είναι κανονικές.
3. Για την **γραμμικοποίηση**, όταν τα αρχικά δεδομένα υποδεικνύουν όχι γραμμικό αλλά μη γραμμικό μοντέλο (είτε ως προς τις παραμέτρους παλινδρόμησης είτε ως προς τις μεταβλητές).

Παρότι, για τους ενδεικνυόμενους κατά περίπτωση μετασχηματισμούς, υπάρχει πλούσια βιβλιογραφία, εντούτοις, η αναζήτηση κατάλληλων μετασχηματισμών, για το συγκεκριμένο κάθε φορά πρόβλημα, απαιτεί αρκετή σχετική εμπειρία. Απαιτεί

επίσης καλή γνώση της φύσης του υπό μελέτη προβλήματος, ιδιαίτερα όταν τα δεδομένα παραβιάζουν (δεν υποστηρίζουν) περισσότερες από μία προϋποθέσεις-παραδοχές. Γιατί σε αυτή την περίπτωση, είναι δυνατόν, μετασχηματισμοί που προσφέρονται για την άρση μιας παραβίασης να μην προσφέρονται για την άρση των άλλων ή και να δημιουργούν νέες.

Στη συνέχεια, σταχυολογούμε από τη βιβλιογραφία κάποιες χαρακτηριστικές περιπτώσεις ενδεικνυόμενων μετασχηματισμών.

3.4.1. Λογαριθμικοί Μετασχηματισμοί

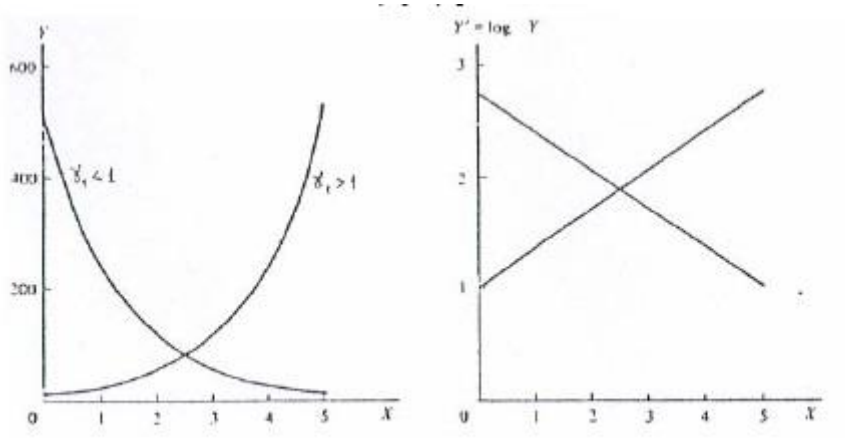
Ο λογαριθμικός μετασχηματισμός $ln(Y) = Y'$ ενδείκνυται:

I. για σταθεροποίηση της διασποράς της Y, όταν αυξάνεται με το Y.

II. για κανονικοποίηση της Y, όταν η κατανομή των υπολοίπων παρουσιάζει θετική ασυμμετρία.

III. για γραμμικοποίηση του μοντέλου όταν τα αρχικά δεδομένα υποδεικνύουν το **πολλαπλασιαστικό μοντέλο**:

1. $Y = \gamma_0 \cdot {\gamma_1}^X \cdot \varepsilon$



Σχήμα 3.19 Λογαριθμικός Μετασχηματισμός

Στην περίπτωση αυτή, το αρχικό μοντέλο (αριστερά) μετασχηματίζεται στο γραμμικό (δεξιά):

$Y' = \ln(\gamma_0) + \gamma_1 X' + \ln(\varepsilon)$, $όπου\ Y' = \ln(Y), \alpha = \ln(\gamma_0), \beta = \ln(\gamma_1), \varepsilon' = \ln(\varepsilon)$

Τότε :

$$Y' = \alpha + \beta \cdot X + \varepsilon'$$

Με $\varepsilon \sim LnNormal(0, \sigma^2) \Rightarrow \ln(\varepsilon) \sim Normal(0, \sigma^2)$

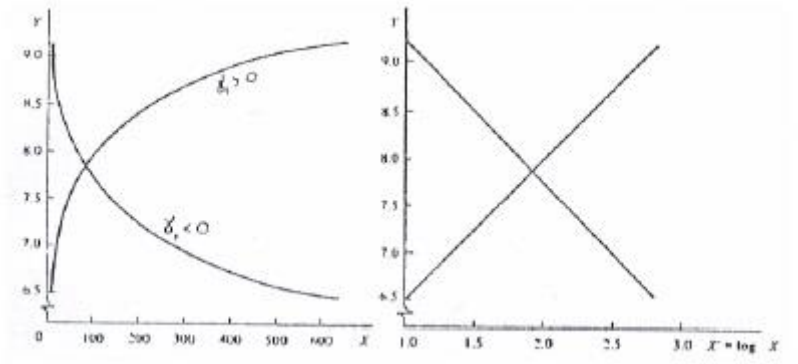
Με λογαριθμικούς μετασχηματισμούς γίνεται, επίσης, **γραμμικοποίηση των** παρακάτω **πολλαπλασιαστικών μοντέλων**:

2. $e^{Y} = \gamma_0 \cdot X^{\gamma_1} \cdot \varepsilon$ με το μετασχηματισμό $\ln(X) = X'$

   Τότε $Y = \ln(\gamma_0) + \gamma_1 \ln(X) + \ln(\varepsilon) \Rightarrow$

   $Y = \ln(\gamma_0) + \gamma_1 X' + \ln(\varepsilon)$ γραμμικό μοντέλο με

   $\varepsilon \sim LnNormal(0, \sigma^2) \Rightarrow \ln(\varepsilon) \sim Normal(0, \sigma^2)$



Σχήμα 3.20 α Λογαριθμικός Μετασχηματισμός προς γραμμικοποίηση των πολλαπλασιαστικών μοντέλων.

3. $Y = \gamma_0 \cdot X^{\gamma_1} \cdot \varepsilon$ με το μετασχηματισμό $\ln(Y) = Y'$ $και$ $\ln(X) = X'$

   Τότε όμοια με 1.

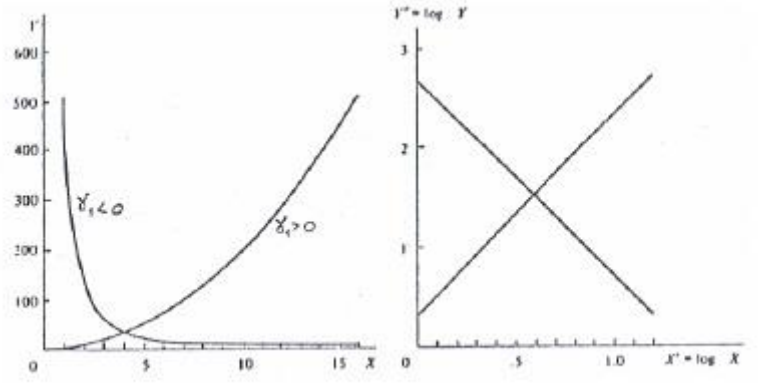
   Έχουμε: $\ln(Y) = \ln(\gamma_0) + \gamma_1 \ln(X) + \ln(\varepsilon) \Rightarrow$ (3.35)

   $Y' = \ln(\gamma_0) + \gamma_1 X' + \ln(\varepsilon)$ γραμμικό μοντέλο με

   $\varepsilon \sim LnNormal(0, \sigma^2) \Rightarrow \ln(\varepsilon) \sim Normal(0, \sigma^2)$

   Ενώ, αν παραγωγίσουμε την Y ως προς X στην (3.35): $\frac{1}{Y}\frac{dY}{dX} = \gamma_1 \frac{dX}{X}$

   1% αύξηση στο X επιφέρει γ_1% μεταβολή στην αναμενόμενη τιμή του Y. Ο συντελεστής γ_1 ελαστικότητα.



Σχήμα 3.20 β Λογαριθμικός Μετασχηματισμός προς γραμμικοποίηση των πολλαπλασιαστικών μοντέλων.

4. $Y = \gamma_0 \cdot + \gamma_1 \log(X) + \varepsilon \Leftrightarrow e^{(Y)} = e^{(\gamma_0)} X^{\gamma_1} e^{(\varepsilon)}$

   Κατά αντιστοιχία με το 2. και με το μετασχηματισμό

   $$e^{(Y)} = Y' \text{ όπου}, e^{(\gamma_0)} = \beta_0, \quad e^{(\varepsilon)} = \varepsilon'$$

   $$Y' = \beta_0 X^{\gamma_1} \varepsilon'$$

   Προκύπτει γραμμικότητα. Όταν το X αυξηθεί κατά 1% η αναμενόμενη τιμή του Y' θα μεταβληθεί κατά $\gamma_1/100$ μονάδες.

5. $\ln(Y) = \gamma_0 + \gamma_1 X + \varepsilon$

   Εκθετίζοντας και κατά αντιστοιχία με τα παραπάνω οδηγούμαστε στο κατάλληλο γραμμικό μοντέλο. Να σημειωθεί πως όταν το X αυξηθεί κατά 1 μονάδα, η αναμενόμενη τιμή του Y θα μεταβληθεί κατα γ_1% .

3.4.2. Αντίστροφοι Μετασχηματισμοί

Ο αντίστροφος μετασχηματισμός $\frac{1}{Y} = Y'$ ενδείκνυται:

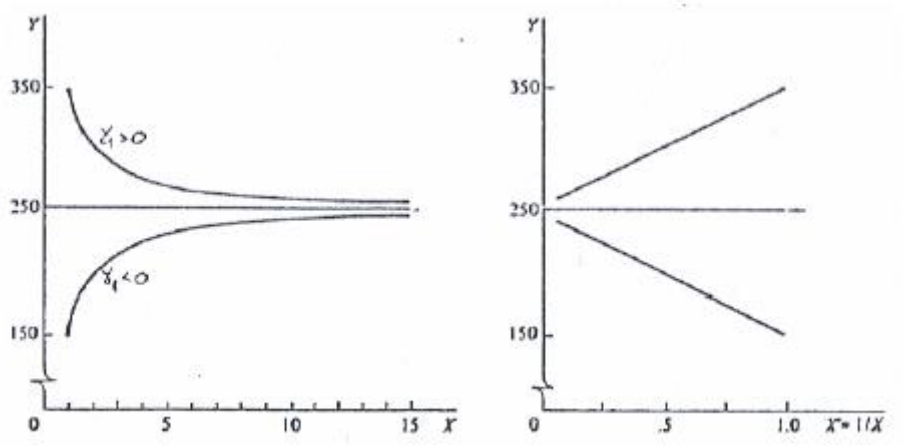
I. για <u>σταθεροποίηση της διασποράς της Y,</u> όταν έχουμε μεγάλη αύξηση της διασποράς πάνω από κάποια τιμή του Y.

II. για <u>γραμμικοποίηση του μοντέλου</u> όταν τα αρχικά δεδομένα υποδεικνύουν το αντίστροφο μοντέλο: $Y = \frac{1}{\gamma_0 + \gamma_1 \cdot X + \varepsilon}$

Στην περίπτωση αυτή, το αρχικό μοντέλο μετασχηματίζεται στο γραμμικό:

$$Y' = \alpha + \beta \cdot X + \varepsilon', \text{όπου } Y' = \frac{1}{Y}, \alpha = \gamma_0, \beta = \gamma_1, \varepsilon' = \varepsilon$$

Με τον αντίστροφο μετασχηματισμό $\frac{1}{X} = X'$ γίνεται, γραμμικοποίηση του αντίστροφου μοντέλου: $Y = \gamma_0 + \gamma_1 \cdot \frac{1}{X} + \varepsilon$



Σχήμα 3.21 Αντίστροφος μετασχηματισμός προς γραμμικοποίηση αντίστροφου μοντέλου.

3.4.3. Μετασχηματισμοί Τετραγωνικής Ρίζας

Ο μετασχηματισμός $\sqrt{Y} = Y'$ ενδείκνυται:

α) για σταθεροποίηση της διασποράς της Y, όταν η διασπορά είναι ανάλογη της μέσης τιμής της Y.

β) για γραμμικοποίηση του μοντέλου όταν τα αρχικά δεδομένα υποδεικνύουν το μοντέλο: $Y = (\gamma_0 + \gamma_1 \cdot X + \varepsilon)^2$

Στην περίπτωση αυτή, το αρχικό μοντέλο μετασχηματίζεται στο γραμμικό:

$Y' = \alpha + \beta \cdot X + \varepsilon'$, όπου $Y' = \sqrt{Y}\,, \alpha = \gamma_0, \beta = \gamma_1, \varepsilon' = \varepsilon$

Με τον μετασχηματισμό $\sqrt{X} = X'$ γίνεται γραμμικοποίηση του μοντέλου:

$$Y = \gamma_0 + \gamma_1 \cdot \sqrt{X} + \varepsilon.$$

3.4.4. Μετασχηματισμός $Y^2 = Y'$

Ο μετασχηματισμός αυτός ενδείκνυται:

α) για σταθεροποίηση της διασποράς της Y, όταν ελαττώνεται με τη μέση τιμή της Y.

β) για κανονικοποίηση της Y, όταν η κατανομή των υπολοίπων παρουσιάζει αρνητική ασυμμετρία.

γ) για γραμμικοποίηση του μοντέλου όταν τα αρχικά δεδομένα υποδεικνύουν

καμπυλόγραμμο μοντέλο π.χ. $Y = \sqrt{\gamma_0 + \gamma_1 \cdot X + \varepsilon}$.

Στην περίπτωση αυτή, το αρχικό μοντέλο μετασχηματίζεται στο γραμμικό:

$$Y' = \alpha + \beta \cdot X + \varepsilon' \text{ , όπου } Y^2 = Y' \text{ , } \alpha = \gamma_0, \beta = \gamma_1, \varepsilon' = \varepsilon$$

Οι παραπάνω μετασχηματισμοί μπορούν φυσικά να συνδυασθούν για την αντιμετώπιση πιο πολύπλοκων περιπτώσεων. Για παράδειγμα το μη γραμμικό μοντέλο

$$Y = \frac{1}{1 + e^{\gamma_0 + \gamma_1 \cdot X + \varepsilon}}$$

εύκολα μετασχηματίζεται σε γραμμικό με το μετασχηματισμό $Y' = \ln(\frac{1}{Y} - 1)$ που είναι ένας αντίστροφος και ένας λογαριθμικός μετασχηματισμός (διαδοχικά).

Κεφάλαιο 4 : Πολλαπλή Παλινδρόμηση

4.1 Εισαγωγή

Η συμμετοχή περισσότερων των δύο μεταβλητών σε μία ανάλυση παλινδρόμησης ή και συσχέτισης αποτελεί ειδικό κεφάλαιο μελέτης, γνωστό ως **πολλαπλή παλινδρόμηση (Multiple Regression Analysis- MRA)** ή και **πολλαπλή συσχέτιση (multiple correlation)**. Αν πραγματοποιήσουμε μετρήσεις ταυτόχρονα για τρεις ή περισσότερες μεταβλητές από τις οποίες η μία θεωρούμε ότι είναι εξαρτημένη (Y) από τη δράση των λοιπών (X_i), π.χ. τις X_1, X_2 και X_3, τότε αναφερόμαστε στην πολλαπλή παλινδρόμηση. Στην περίπτωση αυτή ισχύουν οι εξής προϋποθέσεις για την εξαρτημένη Y: οι τιμές της είναι τυχαίες, έχουν κανονική κατανομή και βρίσκονται σε αντιστοιχία με τους παρατηρούμενους συνδυασμούς των τιμών των ανεξάρτητων

μεταβλητών. Τυχόν επαναληπτικές μετρήσεις της Y σε συνδυασμό πάντοτε με τις τιμές των ανεξάρτητων μεταβλητών, θα πρέπει επίσης, να έχουν κανονική κατανομή και κοινή διακύμανση.

Αν η σχέση των μεταβλητών δεν εκφράζεται με πολλαπλή γραμμική παλινδρόμηση, καθότι λόγω της φύσης και προέλευσης αυτών, δεν υποβάλλονται στο θεσμό της σχέσης εξαρτημένης-ανεξάρτητες, τότε αναφερόμαστε στην πολλαπλή συσχέτιση, με κύρια προϋπόθεση ότι, όλες οι μεταβλητές έχουν κανονική κατανομή. Αυτό σημαίνει για την απλή συσχέτιση ότι οι δύο μεταβλητές προέρχονται από συμμεταβλητό κανονικό πληθυσμό και, για την πολλαπλή συσχέτιση, ότι οι εμπλεκόμενες μεταβλητές προέρχονται από πολυμεταβλητό κανονικό πληθυσμό.

Τα μοντέλα πολλαπλής παλινδρόμησης μπορούν επίσης να διασαφηνιστούν και με δύο διαφορετικές προοπτικές, ως επεξηγηματικά και προβλεπτικά (Pedhazur, 1997). Τα επεξηγηματικά μοντέλα επιδιώκουν την εδραίωση ενός ισχυρού μοντέλου το οποίο οφείλει να επιβεβαιώνει το αποτέλεσμα των προβλέψεων εκείνων μόνο που διαθέτουν ως εφαλτήριο ικανή θεωρητική υπόσταση, απορρίπτοντας άλλες που δεν είναι σχετικές. Δηλαδή, τα μοντέλα οφείλουν να ελέγχουν αν μία σημαντική προβλεπτική μεταβλητή μπορεί, ένεκα του θεωρητικού της υπόβαθρου, να παράγει τη μέγιστη δυνατή διακύμανση οδηγώντας έτσι σε ακριβέστερες προβλέψεις. Τα προβλεπτικά ή διερευνητικά μοντέλα διαθέτουν πιο ελεύθερη θεωρητική βάση και συνεπώς είναι περισσότερο ευέλικτα αφού βασίζονται άμεσα στην απρόσκοπτη

ανάλυση των στοιχείων. Τα μοντέλα αυτά επιχειρούν την ανεύρεση της ομάδας εκείνης των προβλεπουσών μεταβλητών η οποία παρέχει το καλύτερο αποτέλεσμα πρόβλεψης, ανεξαρτήτως αν το μοντέλο προσεγγίζει ή όχι κάποιο ορθό επεξηγηματικό μηχανισμό σε θεωρητικό επίπεδο. Δεν επιδιώκουν ιδιαίτερα να ανιχνεύσουν αν οι προβλέψεις αντανακλούν κάποια πραγματική επιστημονική αιτία υπεύθυνη για την έκβαση του συγκεκριμένου αποτελέσματος.

Σε πολλά πρακτικά προβλήματα είναι απαραίτητο να χρησιμοποιήσουμε δύο ή περισσότερες ανεξάρτητες μεταβλητές προκειμένου να ερμηνεύσουμε με μεγάλη ακρίβεια ένα φυσικό φαινόμενο και να βγάλουμε σωστότερα συμπεράσματα. Για παράδειγμα, προκειμένου να χρησιμοποιηθεί ένα μοντέλο παλινδρόμησης για να προβλεφθεί η ζήτηση ενός προϊόντος μιας εταιρείας σε έναν αριθμό από διαφορετικές πόλεις, είναι ίσως σκόπιμο να χρησιμοποιηθούν κοινωνικοοικονομικές μεταβλητές (μέσο οικογενειακό εισόδημα, μόρφωση), δημογραφικές μεταβλητές (αριθμός μελών οικογένειας, αριθμός συνταξιούχων) και περιβαλλοντολογικές μεταβλητές (μέση ημερήσια θερμοκρασία) κ.α.

Όταν έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές, για να ερμηνεύσουμε τη συμπεριφορά της εξαρτημένης μεταβλητής *Y*, χρησιμοποιούμε το *μοντέλο της πολλαπλής παλινδρόμησης*. Μάλιστα, αν η σχέση της εξαρτημένης μεταβλητής είναι γραμμική συνάρτηση των ανεξάρτητων μεταβλητών, τότε η περιγραφή της σχέσης αυτής γίνεται βάση ενός γραμμικού μοντέλου και έτσι αναφερόμαστε στην *πολλαπλή γραμμική παλινδρόμηση*.

Η πολλαπλή παλινδρόμηση έχει ευρεία επιστημονική αποδοχή διότι θεωρείται ισχυρό και ευέλικτο στατιστικό εργαλείο με πλήθος εφαρμογών σε τελείως διαφορετικά ερευνητικά πεδία (Draper & Smith, 1989, Pedhazur,1997, Weisburg, 1985). Κάποια από αυτά είναι:

- Διοίκηση επιχειρήσεων και έρευνα αγοράς: εκτίμηση του βαθμού επίδοσης του προσωπικού μιας εταιρείας, διαχείριση του αριθμού έκτασης των παραπόνων των πελατών.
- Τρόποι διερεύνησης της συμπεριφοράς του δείκτη νοημοσύνης σε διαγωνιστικό επίπεδο.
- Εκτίμηση της δράσης των χημικών συστατικών ενός τροφίμου στις οργανοληπτικές ιδιότητές του κ.ά. .

Συνοψίζοντας, η πολλαπλή ανάλυση παλινδρόμησης χρησιμοποιείται για

- ✓ την περιγραφή των ειδικών σχέσεων μεταξύ των μεταβλητών
- ✓ τη διακρίβωση θεωρητικών υποθέσεων
- ✓ την πρόβλεψη από λήψεις πειραματικών δεδομένων και
- ✓ τη δημιουργία και επαλήθευση εξισώσεων πολλαπλής παλινδρόμησης.

Το γραμμικό μοντέλο πολλαπλής παλινδρόμησης με p ανεξάρτητες μεταβλητές, είναι της μορφής:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \qquad για\ i = 1,2,\dots,n$$

Ή σε μορφή πινάκων:

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \qquad (4.1)$$

με

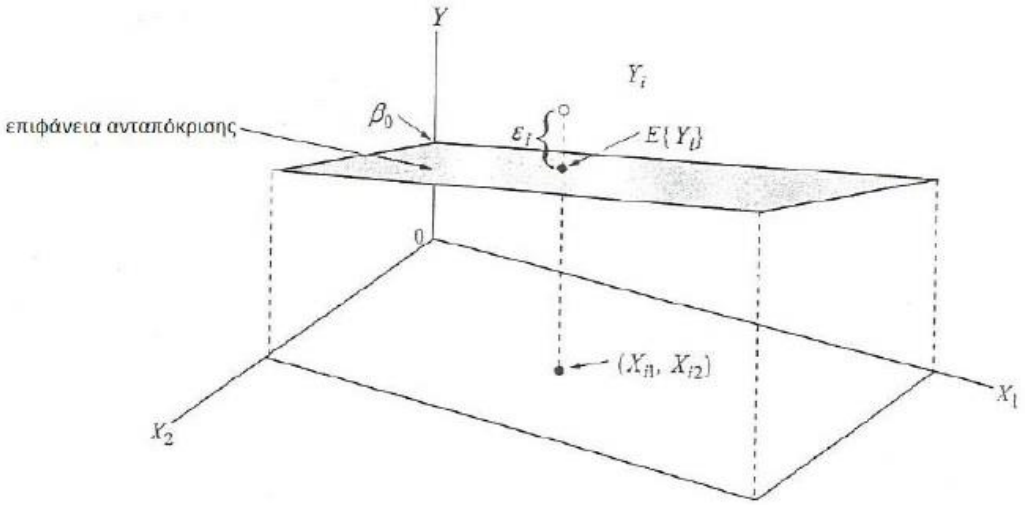
$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1,p-1} \\ 1 & X_{21} & & X_{2,p-1} \\ & \vdots & & \\ 1 & X_{n1} & \dots & X_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

όπου Y_i είναι η τιμή της εξαρτημένης μεταβλητής στην i παρατήρηση. Ανάλογα, η X_{ij}είναι η $i\ (για\ i = 1,2,\dots,n)$,παρατήρηση $της\ j\ (για\ j = 1,\dots,p-1)$ ανεξάρτητης μεταβλητής. *Η β_i αντιπροσωπεύει την μεταβολή στην Y που προέρχεται από μια μεταβολή στην X_i κατά μία μονάδα, <u>όταν όλες οι άλλες ανεξάρτητες μεταβλητές παραμένουν σταθερές</u>*. Τέλος, η ύπαρξη των καταλοίπων ε_i, όπως και στην απλή γραμμική παλινδρόμηση, είναι <u>*απαραίτητη*</u> γιατί στην πράξη κανένα μοντέλο δεν μπορεί να περιγράψει το σύνολο των πληροφοριών ενός σετ δεδομένων. Όσο καλά προσαρμοσμένη και να είναι η γραμμή πολλαπλής παλινδρόμησης στα δεδομένα, πάντα θα υπάρχει ένα μέρος της πληροφορίας που θα εξακολουθεί να μην ερμηνεύεται μέσω του μοντέλου. *Αυτός ο παράγοντας που δεν ερμηνεύεται από το γραμμικό μοντέλο καλείται λάθος της παλινδρόμησης*.

Επομένως, η <u>*πληθυσμιακή εξίσωση παλινδρόμησης ή συνάρτηση παλινδρόμησης ή συνάρτηση ανταπόκρισης*</u> είναι η:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1},\ για\ i = 1,2,\dots,n. \qquad (4.2)$$

Η συνάρτηση αυτή ονομάζεται μερικές φορές και <u>*επιφάνεια παλινδρόμησης ή επιφάνεια ανταπόκρισης*</u>. Μια απεικόνισή της με την χρήση δύο ανεξάρτητων μεταβλητών, θα μπορούσε να είναι η παρακάτω:



Σχήμα 4.1 Επιφάνεια ανταπόκρισης

Η γραμμικότητα των μοντέλων αυτών αναφέρεται στις παραμέτρους και όχι στις ανεξάρτητες μεταβλητές. Έτσι, υπάρχουν και κάποια μοντέλα που είναι γραμμικά ως προς τις παραμέτρους όχι όμως ως προς τις μεταβλητές X. Κάποιες μορφές τέτοιων μοντέλων είναι:

1. Η πολυωνυμική μορφή k βαθμού, όπου:

   $$Y_i = \beta_0 + \beta_1 {X_{i1}}^1 + \beta_2 {X_{i2}}^2 + \cdots + \beta_k {X_{ik}}^k + \varepsilon_i, \qquad \gamma\iota\alpha\ i = 1,2,\dots,n$$

   Το μοντέλο αυτό μπορεί να μετασχηματιστεί σε γραμμικό υπόδειγμα και ως προς τις ανεξάρτητες μεταβλητές θέτοντας $X_{i1} = {X_{i1}}^1$, $X_{i2} = {X_{i2}}^2$, … , $X_{ik} = {X_{ik}}^k$

   Έτσι προκύπτει το μετασχηματισμένο μοντέλο:

   $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{i,k} + \varepsilon_i, \qquad \gamma\iota\alpha\ i = 1,2,\dots,n$$

2. Η αντίστροφη μορφή:

   $$Y_i = \beta_0 + \beta_1 \frac{1}{X_{i1}} + \beta_2 \frac{1}{X_{i2}} + \varepsilon_i, \qquad \gamma\iota\alpha\ i = 1,2,\dots,n$$

   Το μοντέλο αυτό μετασχηματίζεται σε γραμμικό ως προς τις ανεξάρτητες μεταβλητές θεωρώντας

   $$X_{i1}^* = \frac{1}{X_{i1}} \quad \kappa\alpha\iota \quad X_{i2}^* = \frac{1}{X_{i2}}$$

   Οπότε θα έχουμε το μετασχηματισμένο μοντέλο :

   $$Y_i = \beta_0 + \beta_1 X_{i1}^* + \beta_2 X_{i2}^* + \varepsilon_i, \qquad \gamma\iota\alpha\ i = 1,2,\dots,n$$

3. Το υπόδειγμα της μορφής:

$$Y_i = \beta_0 X_{i1}^{\beta_1} X_{i2}^{\beta_2} X_{i,k}{}^{\beta_k} \varepsilon_i, \qquad για\ i = 1,2,\dots,n$$

Σε αυτήν την περίπτωση λογαριθμίζοντας θα έχουμε το μετασχηματισμένο μοντέλο:

$$Y_i{}^* = \beta_0{}^* + \beta_1 X_{i1}^* + \cdots + \beta_k X_{ik}^* + \varepsilon_i{}^*, \qquad για\ i = 1,2,\dots,n$$

4. Η λογαριθμική αντίστροφη μορφή:

$$Y_i = e^{\beta_0 + \beta_1 \frac{1}{X_{i1}} + \beta_2 \frac{1}{X_{i2}} + \varepsilon_i} \quad , \qquad για\ i = 1,2,\dots,n$$

Λογαριθμίζοντας και έπειτα θέτοντας $X_{i1}^* = \frac{1}{X_{i1}}$ $και$ $X_{i2}^* = \frac{1}{X_{i2}}$, παίρνουμε το μετασχηματισμένο μοντέλο:

$$Y_i{}^* = \beta_0 + \beta_1 X_{i1}^* + \beta_2 X_{i2}^* + \varepsilon_i, όπου\ Y_i{}^* = \ln(Y_i) \qquad για\ i = 1,2,\dots,n$$

5. Η ημιλογαριθμική ή γραμμική λογαριθμική μορφή:

$$Y_i = \beta_0 + \beta_1 \ln(X_{i1}) + \beta_2 \ln(X_{i2}) + \varepsilon_i, με\ X_{ij} > 0, \qquad για\ i = 1,2,\dots,n$$

Θεωρώντας $X_{i1}^* = \ln(X_{i1})$ $και$ $X_{i2}^* = \ln(X_{i2})$, παίρνουμε το μετασχηματισμένο μοντέλο:

$$Y_i = \beta_0 + \beta_1 X_{i1}^* + \beta_2 X_{i2}^* + \varepsilon_i, \qquad για\ i = 1,2,\dots,n$$

Ο αριθμός των ανεξάρτητων μεταβλητών στην ανάλυση πολλαπλής παλινδρόμησης έχει σχέση με τον αριθμό των παρατηρήσεων. Στην βιβλιογραφία αναφέρονται εμπειρικοί κανόνες, όπως για παράδειγμα ότι σε κάθε ανεξάρτητη μεταβλητή πρέπει να αντιστοιχούν τουλάχιστον 10 ή 20 παρατηρήσεις. Συνήθως ο αριθμός των ανεξάρτητων μεταβλητών δεν υπερβαίνει τις 10.

4.2 Βασικές Υποθέσεις

Όλες οι βασικές υποθέσεις που αναφέρθηκαν στο κεφάλαιο της απλής γραμμικής παλινδρόμησης ισχύουν και στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης.

Δηλαδή τα κατάλοιπα ε_i:

1. Αποτελούν ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν κανονική κατανομή.
2. Έχουν μηδενική μέση τιμή:

$$E(\vec{\varepsilon}) = E\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix} = \begin{bmatrix}E(\varepsilon_1)\\ E(\varepsilon_2)\\ \vdots\\ E(\varepsilon_n)\end{bmatrix} = \begin{bmatrix}0\\ 0\\ \vdots\\ 0\end{bmatrix} = \vec{0}$$

3. Ικανοποιούν τις συνθήκες ομοσκεδαστικότητας και της μηδενικής συνδιασποράς. Αποδεικνύεται ότι:

$$V(\vec{\varepsilon}) = Cov(\vec{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 \dots & 0 \\ 0 & \sigma^2 \dots & 0 \\ 0 & 0 \dots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

Όσον αφορά τις *ανεξάρτητες μεταβλητές, αυτές*

- *δεν είναι στοχαστικές*. Οι τιμές τους παραμένουν σταθερές, δηλαδή οι μεταβλητές X_i δε συσχετίζονται με το σφάλμα και η συνδιασπορά τους είναι ίση με το μηδέν.
- Οι τιμές δεν είναι όλες ίσες μεταξύ τους, που σημαίνει ότι η διασπορά των X_i είναι διαφορετική από το μηδέν.
- Καμία από τις ανεξάρτητες μεταβλητές δεν μπορεί να εκφραστεί σαν γραμμικός μετασχηματισμός μιας ή περισσότερων από τις υπόλοιπες, πράγμα που αποκλείει τέλεια πολυσυγγραμμικότητα.
- Τέλος, ο αριθμός των ανεξάρτητων μεταβλητών p θα πρέπει να είναι μικρότερος του αριθμού των παρατηρήσεων n.

Υπό την προϋπόθεση ότι $\vec{\varepsilon} \sim N_n(\vec{0}, \sigma^2 I_n)$ και ο γραμμικός μετασχηματισμός $\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$ θα ακολουθεί την πολυμεταβλητή κανονική κατανομή και επιπλέον:

$$E(\vec{Y}) = E(X\vec{\beta} + \vec{\varepsilon}) = \mathrm{E}\left(\begin{bmatrix} 1 & X_{11} \dots X_{1,p-1} \\ 1 & X_{21} \dots X_{2,p-1} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} \dots X_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \right)$$

$$= \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \cdots \beta_{p-1} X_{1,p-1} \\ \beta_0 + \beta_1 X_{21} + \cdots \beta_{p-1} X_{2,p-1} \\ \vdots \\ \beta_0 + \beta_1 X_{n1} + \cdots \beta_{p-1} X_{n,p-1} \end{bmatrix} = X\vec{\beta}$$

Ενώ για τον πίνακα διασπορών-συνδιασπορών έχουμε $Cov(\vec{Y}) = \sigma^2 I_n$.

Τελικά, $\vec{Y} \sim N_n(X\vec{\beta}, \sigma^2 I_n)$

4.3 Μέθοδοι Εκτίμησης Παραμέτρων

Αναφορικά με το θεσμό της απλής γραμμικής παλινδρόμησης για ένα πληθυσμό με ένα ζεύγος μεταβλητών $X - Y$, θα ισχύει η σχέση, $\hat{Y} = a + bX$. Όταν η εξαρτημένη μεταβλητή Y θεωρούμε ότι είναι γραμμικά εξαρτημένη, επιπλέον, και από μία

δεύτερη μεταβλητή (X_2) ή και από μία τρίτη (X_3) ή τελικά από ένα σύνολο m μεταβλητών X, η παραπάνω σχέση διαμορφώνεται σε,

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_mX_m, \text{ δηλαδή σε}$$
$$\hat{Y} = a + \sum_{i=1}^{m} b_mX_m \qquad (4.3)$$

Οι συντελεστές $b_1, \ldots, b_m$ καλούνται **μερικοί συντελεστές παλινδρόμησης** και ο τρόπος υπολογισμού τους αναφέρεται στον Πίνακα 4.1. Ο μερικός συντελεστής b_1 εκφράζει το μέγεθος μεταβολής της Y, όταν μεταβάλλεται η μεταβλητή X_1 κατά μία μονάδα, ενώ παράλληλα οι υπόλοιπες μεταβλητές X_i διατηρούνται σταθερές στην τιμή του μέσου όρου τους. Ή αλλιώς, ο μερικός συντελεστής b_1 εκφράζει τη μέτρηση της σχέσης μεταξύ Y και X_1 , θέτοντας υπό έλεγχο ταυτόχρονα τις λοιπές μεταβλητές X_i ή αλλιώς της σχέσης Y και X_1, αφού προηγούμενα απαλειφθεί (απομακρυνθεί) το αποτέλεσμα των λοιπών μεταβλητών X_i επί της Y και X_1. Παρόμοια, ο συντελεστής b_2, εκφράζει το βαθμό μεταβολής της Y, όταν μεταβάλλεται μόνο η X_2 κοκ. Οι συντελεστές της πολλαπλής παλινδρόμησης καλούνται μερικοί, επειδή εκφράζουν μέρος μόνο της εξαρτημένης σχέσης της Y με τις μεταβλητές X_i. Η παράμετρος a είναι η τιμή της Y, όταν όλες οι μεταβλητές X_i είναι μηδενικές.

Υπολογισμός των Μερικών Συντελεστών Παλινδρόμησης.

Στηρίζεται στη μέθοδο ελαχιστοποίησης του σφάλματος του αθροίσματος των υπολειμμάτων, των αποστάσεων, εφαρμόζοντας τη μέθοδο των ελαχίστων τετραγώνων δηλαδή, από το υπερεπίπεδο προσαρμογής (προσαρμοσμένες τιμές $\hat{Y}$) και υπολογίζεται με την εφαρμογή της άλγεβρας μητρών.

Έστω ότι έχουμε n παρατηρήσεις m μεταβλητών X_i $(X_1, X_2, \ldots, X_m)$, επομένως τα αθροίσματα των τιμών των παρατηρήσεων ανά μεταβλητή θα είναι $\sum X_1$, $\sum X_2, \ldots, \sum X_m$ και τα αθροίσματα των γινομένων μεταξύ αυτών, γνωστών ως σταυροειδών (χιαστί) γινομένων, θα είναι $\sum X_1X_2$, $\sum X_1X_3, \ldots$, $\sum X_1X_m$, $\sum X_2X_3, \ldots, \sum X_iX_j$. Από αυτά προκύπτουν τα αθροίσματα των τετραγώνων των τιμών τους $\sum x_1^2, \sum x_2^2, \ldots, \sum x_m^2$ και των χιαστί γινομένων τους. Επισημαίνεται ότι:

$$\sum x_i^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n} \text{ και } \sum x_ix_j = \sum X_iX_j - \frac{(\sum X_i)(\sum X_j)}{n}.$$

Οι παραπάνω υπολογισμοί παρίστανται ευκρινέστερα με τη δημιουργία μήτρας του αθροίσματος των τετραγώνων των χιαστί γινομένων που συμβολίζεται με S, γνωστής και ως μήτρας διακύμανσης-συνδιακύμανσης:

$$S = \begin{bmatrix} \sum x_1^2 & \sum x_1x_2 & \sum x_1x_3 & \dots & \sum x_1x_m \\ \sum x_2x_1 & \sum x_2^2 & \sum x_2x_3 & \dots & \sum x_2x_m \\ \sum x_3x_1 & \sum x_3x_2 & \sum x_3^2 & \dots & \sum x_3x_m \\ \vdots & \vdots & \vdots & & \vdots \\ \sum x_mx_1 & \sum x_mx_2 & \sum x_mx_3 & \dots & \sum x_m^2 \end{bmatrix}$$

Πίνακας 4.1. Μήτρα διακύμανσης-συνδιακύμανσης

Από τη μήτρα αυτή εύκολα προκύπτει και η αντίστοιχη μήτρα των συντελεστών συσχέτισης, αν αντικαταστήσουμε,

$$r_{ij} = \frac{\sum x_i x_j}{\sqrt{\sum {x_i}^2 \sum x_j^2}},$$

τότε έχουμε:

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1m} \\ r_{21} & r_{22} & r_{23} & \dots & r_{2m} \\ r_{31} & r_{32} & r_{33} & \dots & r_{3m} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \dots & r_{mm} \end{bmatrix}$$

Οι μερικοί συντελεστές της πολλαπλής παλινδρόμησης υπολογίζονται από τη σχέση του γινομένου των μητρών: $S \cdot \boldsymbol{b} = \boldsymbol{y}$, όπου $\boldsymbol{b}$ είναι το διάνυσμα των αγνώστων μερικών συντελεστών παλινδρόμησης και $\boldsymbol{y}$ είναι το διάνυσμα του αθροίσματος των χιαστί γινομένων των m ανεξάρτητων μεταβλητών με την εξαρτημένη Y:

$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_m \end{bmatrix} \quad y = \begin{bmatrix} \sum x_1 y \\ \sum x_2 y \\ \sum x_3 y \\ \vdots \\ \sum x_m y \end{bmatrix}$$

Λύνοντας την εξίσωση ως προς b θα έχουμε: $\boldsymbol{b = S^{-1} \cdot y}$

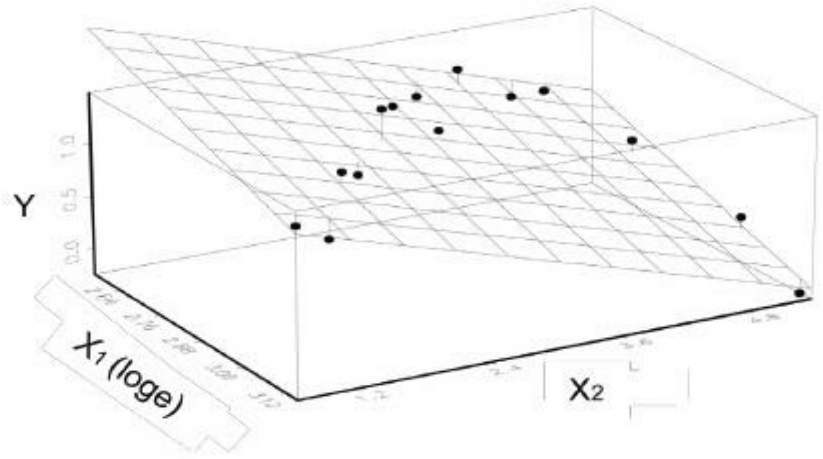
Ο αντίστροφος πίνακας του αθροίσματος των τετραγώνων των χιαστί γινομένων είναι:

$$S^{-1} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots & c_{1m} \\ c_{21} & c_{22} & c_{23} & \dots & c_{2m} \\ c_{31} & c_{32} & c_{33} & \dots & c_{3m} \\ \vdots & \vdots & \vdots & & \vdots \\ c_{m1} & c_{m2} & c_{m3} & \dots & c_{mm} \end{bmatrix}$$

Πίνακας 4.2 Πολλαπλασιαστές Gauss

Τα στοιχεία του καλούνται πολλαπλασιαστές του Gauss και εκείνα που βρίσκονται επί της διαγωνίου της αντίστροφης μήτρας συμβολίζονται με c_{ii} , ενώ εκείνα που βρίσκονται εκτός αυτής με c_{ij} . Οι πολλαπλασιαστές χρησιμοποιούνται για τον υπολογισμό του τυπικού σφάλματος πολλών στατιστικών παραμέτρων της πολλαπλής παλινδρόμησης.

Η εξίσωση της πολλαπλής παλινδρόμησης ορίζει ευθεία γραμμή, όταν μετέχει μία μόνο ανεξάρτητη μεταβλητή, επίπεδο, όταν συμμετέχουν δύο ανεξάρτητες μεταβλητές X_1 και X_2 , Σχήμα 4.2 και υπερεπίπεδο, όταν $X_\iota \geq 3$ με m διαστάσεις, όσες είναι δηλαδή και οι μεταβλητές X_ι. Το υπερεπίπεδο των m διαστάσεων αναφέρεται πολλές φορές και ως επιφάνεια απόκρισης (response surface).



Σχήμα 4.2 Γράφημα πολλαπλής παλινδρόμησης με δύο ανεξάρτητες μεταβλητές X_1και X_2 και μία εξαρτημένη Y.Η επίδραση αυτών επί της εξαρτημένης Y εκφράζεται με

επίπεδο μεταβολής, η κλίση του οποίου επηρεάζεται από τις κλίσεις των πλευρών αυτού, που αντιστοιχούν μία σε κάθε ανεξάρτητη μεταβλητή.

Αντί των μερικών συντελεστών παλινδρόμησης συχνά χρησιμοποιούνται και οι τυποποιημένοι μερικοί συντελεστές β_i (standardized partial regression coefficients) ή βήτα συντελεστές οι οποίοι προκύπτουν ως κανονιστικοί συντελεστές από τον υπολογισμό της εξίσωσης της παλινδρόμησης, αφού προηγουμένως τυποποιηθούν όλες οι συμμετέχουσες μεταβλητές X_i και Y με την αφαίρεση απ` όλες τις τιμές κάθε μεταβλητής του μέσου όρου αυτής και διαιρώντας τη διαφορά με την τυπική απόκλιση των τιμών της μεταβλητής,

$$Y' = \frac{Y-\bar{Y}}{S_Y} \text{ και } X' = \frac{X_i-\overline{X_i}}{S_{(X_i)}}$$

Έτσι, η εξίσωση τροποποιείται σε:

$$\hat{Y}' = \beta_1 X_1' + \beta_2 X_2' + \cdots + \beta_m X_m' \qquad (4.4)$$

Η παράμετρος a απουσιάζει από την εξίσωση, εξαιτίας της φύσης των τυποποιημένων μεταβλητών οι οποίες χαρακτηρίζονται από μέσο όρο ίσο με μηδέν.

Ο τυποποιημένος μερικός συντελεστής βήτα εκφράζει το ρυθμό μεταβολής της Y εξαιτίας της μεταβολής της X_i σε μονάδες τυπικής απόκλισης (οι λοιπές μεταβλητές X_{m-i} διατηρούνται σταθερές ως προς το αποτέλεσμά τους).

Το μεγάλο πλεονέκτημα της χρήσης των συντελεστών αυτών είναι ότι τα μεγέθη τους είναι απευθείας συγκρίσιμα μεταξύ τους ως προς το σχετικό μέγεθος της μεταβολής των ανεξάρτητων μεταβλητών επί της ίδιας εξαρτημένης Y. Εξαλείφεται, δηλαδή, το πρόβλημα της διαφορετικής κλίμακας των μετρήσεων που ελήφθησαν για κάθε μεταβλητή X_i. Έτσι, ένας τυποποιημένος μερικός συντελεστής με υψηλότερη απόλυτη τιμή συγκριτικά με άλλους συντελεστές, δείχνει μεγαλύτερη επίδραση της μεταβλητής X_i την οποία εκπροσωπεί επί της εξαρτημένης Y. Τιμές β μεγαλύτερες της μονάδας ενδεχομένως να εμφανίζονται για μερικές ή και όλες τις ανεξάρτητες μεταβλητές σε υπόδειγμα παλινδρόμησης, που ενδεχομένως να σημαίνουν την παρουσία κάποιου βαθμού πολυσυγγραμμικότητας ή κατασταλτικότητας επίδρασής τους μεταξύ των μεταβλητών στο υπόδειγμα. Οι τυποποιημένοι μερικοί συντελεστές συνδέονται μαθηματικά με τους μερικούς από τη σχέση:

$$\beta_i = b_i \sqrt{\frac{\sum x_i^2}{\sum y^2}} \qquad (4.5)$$

4.4 Ανάλυση της Διακύμανσης στην Πολλαπλή Παλινδρόμηση

Αποτελεί προέκταση της διακύμανσης της απλής παλινδρόμησης, με τις ανάλογες διαφοροποιήσεις, ως προς τον αριθμό m των ανεξάρτητων μεταβλητών που συμμετέχουν με n παρατηρήσεις η καθεμία και τους βαθμούς ελευθερίας των ποσοτήτων που μετρούν τη μεταβλητότητα (Πίνακας 4.3). Εξαιτίας της πολυπλοκότητας των υπολογισμών, θα αναφερθούν μόνο γενικεύσεις των επιμέρους διακυμάνσεων:

Πηγές μεταβλητότητας	df	SS	MS	F
Ολική	n-1	$TSS = \sum (Y_i - \bar{Y})^2$		
Παλινδρόμηση	m	$RSS = \sum (\hat{Y}_i - \bar{Y})^2$	$\frac{RSS}{m}$	
Υπόλειμμα	n-m-1	$ESS = \sum (Y_i - \hat{Y}_i)^2$	$\frac{ESS}{n-m-1}$	$\frac{RMS}{EMS}$

Πίνακας 4.3 Ανάλυση της διακύμανσης στην πολλαπλή παλινδρόμηση

Η ολική πολλαπλή παλινδρόμηση (TSS), όπου το ολικό άθροισμα της διακύμανσης των τετραγώνων των τιμών εκφράζει τη συνολική μεταβλητότητα μεταξύ των Y τιμών: $Y_i - \bar{Y}$

με $n-1$ βαθμούς ελευθερίας, όπου n το πλήθος των παρατηρήσεων (σειρών).Το άθροισμα των τετραγώνων των τιμών της παλινδρόμησης (RSS), το οποίο εκφράζει τη μεταβλητότητα μεταξύ των προσαρμοσμένων τιμών:

$\hat{Y}_i - \bar{Y}$, με m βαθμούς ελευθερίας.

Το άθροισμα των τετραγώνων των τιμών των υπολειμμάτων (ESS), που εκφράζει τη διαφορά της μεταβλητότητας των τιμών Y_i από τις προσαρμοσμένες τιμές:

$Y_i - \hat{Y}_i$, με $n-m-1$ βαθμούς ελευθερίας.

Υπολογισμός των μέσων αθροισμάτων των τετραγώνων της γραμμικής παλινδρόμησης RMS και του υπολείμματος EMS, όπου

$$RMS = \frac{RSS}{m} \quad \text{και} \quad EMS = \frac{ESS}{n-m-2}$$

Η τιμή $F\,(RMS/EMS)$ του ομώνυμου ελέγχου συγκρίνεται με την οριακή $F_{0.05,[m,(n-m-1)]}$ του Πίνακα Π1(Παράρτημα). Αν $\boldsymbol{F \geq F_{0.05}}$, τότε υπάρχει <u>γραμμική εξάρτηση</u> της Y από τις εξεταζόμενες μεταβλητές X_i. Η ποσότητα

$$R^2 = \frac{RSS}{TSS} \qquad (4.6)$$

αποτελεί ***το συντελεστή πολλαπλού προσδιορισμού ή το συντελεστή προσδιορισμού προσαρμογής της πολλαπλής παλινδρόμησης*** (Nagelkerke, 1991). Ο προσδιοριστικός συντελεστής R^2 εκφράζει το ποσοστό της ολικής μεταβλητότητας της Y που εξηγείται από τη συνδυασμένη επίδραση όλων των μεταβλητών X_i που συμμετέχουν στην περιγραφή της εξίσωσης της πολλαπλής γραμμικής παλινδρόμησης επί της εξαρτημένης Y. Ο συντελεστής R^2 λαμβάνει τιμές από μηδέν (κανένα ποσοστό προσαρμοστικότητας) μέχρι 1 (άριστη προσαρμοστικότητα) και επειδή δεν είναι απόλυτα αξιόπιστος, όταν υπάρχουν λίγες παρατηρήσεις στις μεταβλητές, αντικαθίσταται από το διορθωμένο συντελεστή R^2_{adj} :

$$R^2_{adj} = 1 - (1 - R^2)\left(\frac{n-1}{n-m-1}\right) \text{ ή } R^2_{adj} = R^2 - (1 - R^2)\left(\frac{n-1}{n-m-1}\right) \qquad (4.7)$$

Η τιμή του συντελεστή R^2_{adj} είναι <u>πάντα μικρότερη του αντίστοιχου</u> R^2 και παίρνει μερικές φορές αρνητικές τιμές, όταν η αντίστοιχη R^2 είναι αρκετά χαμηλή. Στις περιπτώσεις αυτές συνιστάται η αρνητική τιμή να αντικαθίσταται με το μηδέν, για να μη διακυβεύεται η θετική ποσοτική έννοια του συντελεστή πολλαπλού προσδιορισμού και κατ` επέκταση ο θεσμός της παλινδρόμησης.

Εναλλακτικά, η γραμμικότητα της πολλαπλής παλινδρόμησης μπορεί να εξεταστεί και από το συντελεστή πολλαπλού προσδιορισμού, ελέγχοντας αν αυτός διαφέρει από το μηδέν και χρησιμοποιώντας την ειδική τιμή F που προκύπτει από τη σχέση:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} \qquad (4.8)$$

και τους ίδιους βαθμούς ελευθερίας.

Όταν $F \geq F_{0.05,[m,(n-m-1)]}$ τότε ισχύει η εναλλακτική υπόθεση, $H_A: \rho^2 \neq 0$ που δηλώνει την ύπαρξη γραμμικότητας.

Οι μερικοί συντελεστές της πολλαπλής παλινδρόμησης μπορούν, επίσης, να εξεταστούν ατομικά, αν διαφέρουν από το μηδέν, αν δηλαδή, υπάρχει κλίση και αυτό πραγματοποιείται με τον έλεγχο t,

$$t = \frac{b_i}{S_{(b_i)}} \qquad (4.9) \quad \textit{και} \quad S_{(b_i)} = \sqrt{EMS \cdot c_{ii}} \quad (4.10)$$

όπου c_{ii} είναι ένας διαγώνιος πολλαπλασιαστής του Gauss της αντίστροφης μήτρας (Πίνακας 4.2). Η τιμή t του ελέγχου συγκρίνεται με την θεωρητική $t_{0.05,(n-m-1)}$ του Πίνακα Π2 (Παράρτημα) και αν $t \geq to\rho$, τότε ισχύει $b_i \neq 0$ που δηλώνει ότι η κλίση της μεταβλητής X_i που εκφράζεται από το συντελεστή b_i διαφέρει του μηδενός.

Τα όρια εμπιστοσύνης κάθε συντελεστή b_i προκύπτουν από τη σχέση:

$$b_i \pm t_{0.05,(n-m-1)} \cdot S_{b_i}$$

Οι τυποποιημένοι μερικοί συντελεστές της παλινδρόμησης μπορούν επίσης να ελεγχθούν για τη σημαντικότητά τους με τρόπο παρόμοιο, όπως των απλών μερικών συντελεστών, τροποποιώντας μόνο το τυπικό σφάλμα που δίνεται από τη σχέση:

$$S_{\beta_\iota} = \sqrt{\frac{1-R^2}{n-m-1} \cdot c_{ij} \cdot \sum x_i^2}$$

Η σημαντικότητα ή αλλιώς το μέγεθος της επίδρασης που έχουν οι μερικοί συντελεστές μεταξύ τους και οι οποίοι περιγράφουν μία συγκεκριμένη εξίσωση πολλαπλής παλινδρόμησης ελέγχεται συγκρίνοντας τη σημαντικότητα της διαφοράς δύο μόνο συντελεστών κάθε φορά με τον έλεγχο t,

$$t = \frac{b_A - b_B}{S_{b_A - b_B}} \qquad \text{και} \qquad S_{b_A - b_B} = \sqrt{EMS \cdot (c_{ii} + c_{jj} - 2c_{ij})}$$

όπου b_A $και$ b_B ένα οποιοδήποτε ζεύγος για σύγκριση μερικών συντελεστών με $n-m-1$ βαθμούς ελευθερίας. Τονίζεται ιδιαίτερα ότι, όταν οι μεταβλητές των μερικών συντελεστών δεν λαμβάνονται με τις ίδιες μετρικές μονάδες, το αποτέλεσμα της σύγκρισης των συντελεστών πρέπει να ερμηνεύεται με τη μέγιστη προσοχή, επειδή υπάρχει σοβαρότατος κίνδυνος παρερμηνείας. Η τιμή του ελέγχου συγκρίνεται με τη θεωρητική $t_{0.05,(n-m-1)}$ και αν $t \geq to\rho$, τότε ισχύει η εναλλακτική υπόθεση ότι οι δύο συντελεστές διαφέρουν σημαντικά μεταξύ τους.

Το τυπικό σφάλμα της παραμέτρου a της πολλαπλής παλινδρόμησης δίνεται από τη σχέση:

$$s_a = \sqrt{EMS \cdot \left[\frac{1}{n} + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}\bar{X}_i\bar{X}_j\right]} \qquad (4.11)$$

Γενικά, μία στατιστικά σημαντική τιμή F του ελέγχου της γραμμικής εξάρτησης της Y απ` όλες τις εξεταζόμενες X_i για να έχει υπόσταση θα πρέπει οπωσδήποτε να συνοδεύεται και από ορισμένες στατιστικά σημαντικές τιμές των ελέγχων t της σημαντικότητας των μερικών συντελεστών της πολλαπλής παλινδρόμησης. Οι υπόλοιπες μεταβλητές που έχουν μη στατιστικά σημαντικό το μερικό συντελεστή

τους, θα πρέπει να απομακρύνονται όλες ή κάποιες από αυτές από την εξίσωση της παλινδρόμησης, με ειδική διαδικασία η οποία περιγράφεται, στη συνέχεια, λεπτομερώς. Υπάρχουν όμως περιπτώσεις όπου παρατηρούνται στατιστικά σημαντικές τιμές του ελέγχου t των συντελεστών b_i , χωρίς να συνοδεύονται και από στατιστικά σημαντική τιμή F της εξίσωσης. Αυτές φανερώνουν υψηλό βαθμό συσχέτισης μεταξύ αρκετών από τις ανεξάρτητες μεταβλητές. Κατά κανόνα, *όταν διαπιστώνουμε μη στατιστικά σημαντική τιμή F, καλό θα είναι να μην ελέγχουμε ή να μη λαμβάνουμε υπόψη τις τιμές t της σημαντικότητας των μερικών συντελεστών*.

4.5 Έλεγχος της Σημαντικότητας της Πολλαπλής Συσχέτισης

Ο έλεγχος F της σημαντικότητας της γραμμικότητας της πολλαπλής παλινδρόμησης ισχύει και για την πολλαπλή συσχέτιση, αρκεί να θεωρήσουμε ότι, δεν υπάρχει καμία εξάρτηση κάποιας μεταβλητής από τις άλλες (η μεταβλητή Y παύει να ισχύει) και ότι όλες οι μεταβλητές είναι ανεξάρτητες μεταξύ τους. Άρα, υπάρχουν $m+1$ μεταβλητές Y_i (συνήθως αποφεύγεται η χρήση του όρου X_i στην πολλαπλή συσχέτιση, για να μην προκαλείται σύγχυση). Στην περίπτωση αυτή, εξετάζουμε το συντελεστή πολλαπλής συσχέτισης αν διαφέρει από το μηδέν, υιοθετώντας την τιμή F του ελέγχου της πολλαπλής παλινδρόμησης. Η θετική τετραγωνική ρίζα του συντελεστή πολλαπλού προσδιορισμού

$$R = +\sqrt{R^2} \qquad (4.12)$$

αναφέρεται ως ***συντελεστής πολλαπλής συσχέτισης R*** και λαμβάνει τιμές από 0 μέχρι $+1$, σε αντίθεση με τον απλό συντελεστή συσχέτισης που κυμαίνεται μεταξύ -1 και $+1$. Ο συντελεστής πολλαπλής συσχέτισης μετρά πόσο ικανοποιητικά οι προσαρμοσμένες (προβλεπόμενες) τιμές $\hat{Y}_i$ ταιριάζουν με τις παρατηρούμενες τιμές Y_i. Έτσι, τιμή R ίση με 1 δηλώνει πλήρη ταύτιση των τιμών $\hat{Y}_i$ και Y_i. Αν η τιμή του ελέγχου F είναι μεγαλύτερη ή ίση της θεωρητικής, τότε ισχύει η εναλλακτική υπόθεση $H_A: \rho_0 \neq 0$ που δηλώνει ότι υπάρχει σημαντική γραμμικότητα της πολλαπλής συσχέτισης.

Όταν οι μεταβλητές εξετάζονται κατά ζεύγη, τότε υπολογίζεται ο **συντελεστής συσχέτισης r του Pearson** για κάθε ζεύγος σύμφωνα με τον τύπο:

$$r = \frac{\Sigma y_i y_j}{\sqrt{\Sigma y_i^2 \cdot \Sigma y_j^2}} \qquad (4.13)$$

$$\sum y_i^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

$$\sum y_i y_j = \sum Y_i Y_j - \frac{(\sum Y_i)(\sum Y_j)}{n}$$

Y_i και Y_j οι δύο εξεταζόμενες μεταβλητές και η ποσότητα $\sum y_j^2$ υπολογίζεται κατ' ανάλογο τρόπο με την $\sum y_i^2$.

Ο παρονομαστής της σχέσης αυτής είναι πάντα θετικός, ο αριθμητής όμως μπορεί να είναι θετικός, αρνητικός ή 0. Το εύρος των τιμών του συντελεστή κυμαίνεται μεταξύ -1 και 1. Όταν υπάρχει θετική συσχέτιση μεταξύ δύο μεταβλητών, σημαίνει ότι, αύξηση της τιμής της μίας συνοδεύεται και από αύξηση της τιμής της άλλης. Αρνητική συσχέτιση σημαίνει ότι, αύξηση της τιμής της μίας μεταβλητής επιφέρει μείωση στην τιμή της άλλης και στην περίπτωση που $r = 0$, τότε δεν υπάρχει γραμμική συσχέτιση μεταξύ των μεγεθών των δύο μεταβλητών. Όταν οι τιμές του r πλησιάζουν το $+1$ ή το -1, τότε αναπτύσσεται ισχυρή θετική ή αρνητική συσχέτιση. Αντίθετα, τιμές κοντά στο 0 δηλώνουν ασθενή συσχέτιση (θετική ή αρνητική). Ο συντελεστής συσχέτισης *δεν μετρά την ποσοτική μεταβολή μίας μεταβλητής εξαιτίας μεταβολής της άλλης* (όπως συμβαίνει με τον συντελεστή παλινδρόμησης), **αλλά απλώς την ένταση της σχέσης μεταξύ δύο μεταβλητών**. Ο συντελεστής r υψωμένος στο τετράγωνο μεταπίπτει στο συντελεστή προσδιορισμού της παλινδρόμησης,

$$r = \sqrt{R^2}$$

Οι δύο αυτοί όροι, παρά τη συγγενική μαθηματική σχέση που εμφανίζουν, προσδιορίζουν τελείως διαφορετικές έννοιες και καλό είναι να μην αναφέρονται μαζί, γιατί ανταποκρίνονται σε διαφορετικές τεχνικές ανάλυσης. Οι συντελεστές συσχέτισης όμως δεν λαμβάνουν υπολογιστικά υπόψη την εξάλειψη των αλληλεπιδράσεων των λοιπών μεταβλητών Y_i σε οποιαδήποτε εξεταζόμενο ζεύγος μεταβλητών. Απλά περιγράφουν το μέτρο της γραμμικής έντασης των μεταβλητών, λαμβανομένων υπόψη μόνο ανά δύο κάθε φορά.

Κάθε αντίστοιχο πρόβλημα λύνεται, λαμβάνοντας υπόψη το μερικό συντελεστή συσχέτισης r_{ij} ο οποίος υπολογίζει τη συσχέτιση μεταξύ κάθε ζεύγους μεταβλητών, διατηρώντας, παράλληλα, σταθερό το γραμμικό αποτέλεσμα κάθε άλλου ζεύγους των λοιπών μεταβλητών. Ο συμβολισμός r_{ij} υπονοεί το μερικό συντελεστή συσχέτισης

μεταξύ των μεταβλητών i και j, όταν οι λοιπές μεταβλητές, παραμένουν σταθερές ως προς το αποτέλεσμά τους. Έχουμε εξαλείψει δηλαδή οποιαδήποτε δράση της αλληλεπίδρασης των λοιπών μεταβλητών στη σχέση που διέπει τις μεταβλητές i και j. Έτσι, ο μερικός συντελεστής συσχέτισης $r_{14.235}$ εκφράζει τη μερική συσχέτιση μεταξύ των μεταβλητών Y_1 και Y_4, όταν οι τιμές των λοιπών Y_2 , Y_3και Y_5 κρατούνται σταθερές. Οι μερικοί συντελεστές συσχέτισης υπολογίζονται εύκολα μόνον όταν συμμετέχουν τρεις μεταβλητές, Y_1 , Y_2και Y_3 χρησιμοποιώντας στην εξίσωσή τους απλούς συντελεστές συσχέτισης. Έτσι, για το συντελεστή$r_{12.3}$ (όπου η μεταβλητή Y_3 κρατείται σταθερή και κατά συνέπεια και τα ζεύγη αυτής με τις λοιπές) θα έχουμε:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

όπου r_{12}, r_{13} και r_{23} είναι οι συντελεστές συσχέτισης των τριών μεταβλητών. Οι μερικοί συντελεστές στην πολλαπλή συσχέτιση, όταν γενικά οι μεταβλητές είναι τρεις ή περισσότερες, προκύπτουν εύκολα από τον υπολογισμό πρώτα των στηλών των υπολειμμάτων που προκύπτουν, παλινδρομώντας κάθε μεταβλητή Y_i με μία άλλη ή και περισσότερες, διατηρώντας πάντα τις υπόλοιπες σταθερές. Οι συσχετίσεις των στηλών των υπολειμμάτων μεταξύ τους αποτελούν τους μερικούς συντελεστές πολλαπλής συσχέτισης. Για παράδειγμα, ο υπολογισμός του συντελεστή συσχέτισης $r_{12.3}$ προκύπτει, παλινδρομώντας πρώτα την Y_1επί της Y_3 και μετά την Y_2 επί της Y_3. Από τις δύο παλινδρομήσεις προκύπτουν δύο αντίστοιχες υπολειμματικές στήλες ESS

$$\left(Y_i - \hat{Y}_i\right)$$

των παλινδρομηθεισών μεταβλητών. Η συσχέτιση των στηλών των υπολειμμάτων αποτελεί τον μερικό συντελεστή συσχέτισης. Παρόμοια, αν οι εξεταζόμενες μεταβλητές είναι πέντε και ενδιαφερόμαστε για το μερικό συντελεστή $r_{13.245}$, τότε παλινδρομούμε την Y_1 επί των υπόλοιπων Y_2, Y_4 και Y_5 , εκτελούμε μία δεύτερη πολλαπλή παλινδρόμηση θέτοντας την Y_3 στη θέση της Y_1 και ακολούθως συσχετίζουμε τις δύο στήλες των υπολειμμάτων που προκύπτουν, μία για κάθε παλινδρόμηση.

Η σημαντικότητα του κάθε μερικού συντελεστή συσχέτισης εξετάζεται αν διαφέρει από το μηδέν, H_0: $\rho_{ij} \ldots = 0$ με **τον έλεγχο t**,

$$t = \frac{r_{ij\ldots}}{s_{r_{ij\ldots}}} \quad \text{και} \quad s_{r_{ij}} = \sqrt{\frac{1 - r_{ij\ldots}}{n - M}}$$

όπου M το πλήθος όλων των $(m+1)$ μεταβλητών. Η τιμή του ελέγχου συγκρίνεται με την θεωρητική τιμή $t_{0.05,(n-M)}$ και αν $t \geq t_{o\rho}$, τότε ισχύει η εναλλακτική υπόθεση που δηλώνει ότι, ο μερικός συντελεστής είναι στατιστικά σημαντικός. Η σημαντικότητα κάθε μερικού συντελεστή συσχέτισης μπορεί να ελεγχθεί επίσης συγκρίνοντας απευθείας την τιμή του με την οριακή τιμή $t_{0.05(2)(n-M)}$ που προκύπτει από τον πίνακα Π3 (Παράρτημα). Αν ο συντελεστής είναι μεγαλύτερος ή ίσος με την οριακή τιμή, τότε διαφέρει στατιστικά σημαντικά από το μηδέν.

4.6 Μέθοδοι της Άριστης Επιλογής των Ανεξάρτητων Μεταβλητών

Ας υποθέσουμε ότι μελετούμε την επίδραση ορισμένων οικονομικών και διαφημιστικών μεταβλητών X_i επί της αναγνωρισιμότητας και της κατανάλωσης-αγοράς ενός προϊόντος, η καθεμία των οποίων αποτελεί την εξαρτημένη μεταβλητή Y. Όμως, η θεώρηση όλων των ανεξάρτητων μεταβλητών σε καμία περίπτωση δεν σημαίνει πλήρη αποδοχή τους, ότι δηλαδή έχουν αυτόματα σημαντικό αποτέλεσμα στην απόκριση της Y. Θα πρέπει πρώτα να εξεταστούν ως προς τη σημαντικότητά τους, μία προς μία και μετά οι συνδυασμοί αυτών ανά δύο ή ανά τρεις κοκ. Έτσι, το μείζον θέμα της πολλαπλής παλινδρόμησης έγκειται στον κατάλληλο προσδιορισμό μόνον εκείνων των ανεξάρτητων μεταβλητών που έχουν το ισχυρότερο (καλύτερο) στατιστικά σημαντικό αποτέλεσμα επί της Y (Hocking, 1976). Διάφοροι μέθοδοι έχουν αναπτυχθεί προς την κατεύθυνση αυτή, δηλαδή την άριστη επιλογή του μοντέλου της πολλαπλής παλινδρόμησης, χωρίς να συμφωνούν απαραίτητα μεταξύ τους ως προς:

- Την επιλογή του αριθμού των μεταβλητών.
- Την ταυτότητα ισάριθμων επιλεγμένων μεταβλητών.
- Την ομοφωνία μεταξύ των επιστημόνων ως προς το ποια μέθοδος είναι η πληρέστερη.

Οι μέθοδοι αυτές στηρίζονται αποκλειστικά στη χρήση στατιστικών προγραμμάτων με τη βοήθεια ηλεκτρονικών υπολογιστών, επειδή παρουσιάζουν εξαντλητική διαδικασία ελέγχων, έκδηλη ακόμα και με τη χρήση των προγραμμάτων αυτών.

4.6.1. Επιλογή της Καταλληλότερης Ομάδας Προσαρμογής των Ανεξάρτητων Μεταβλητών (Best Set of Regressions).

Η διαδικασία αυτή περιλαμβάνει την ανάλυση της πολλαπλής παλινδρόμησης με τη συμμετοχή, αρχικά, όλων των ανεξάρτητων μεταβλητών, ακολούθως με τη συμμετοχή $m-1$ μεταβλητών, στη συνέχεια με τη συμμετοχή $m-2$ κοκ., μέχρι το τέλος αυτής της διαδικασίας η οποία θα καταλήγει σε m απλές γραμμικές παλινδρομήσεις. Η διαδικασία μελέτης μπορεί να ακολουθήσει την αντίστροφη κατεύθυνση, αναφορικά με τη συμμετοχή των μεταβλητών, αρχίζοντας από μία τη φορά και αυξάνοντας σταδιακά τον αριθμό συμμετοχής τους. Για παράδειγμα, αν η ανάλυση περιλαμβάνει 4 ανεξάρτητες μεταβλητές, τότε όλοι οι δυνατοί συνδυασμοί θα ανέρχονται σε 15, όπως στον Πίνακα 4.3 παρακάτω.

X_1			
	X_2		
		X_3	
			X_4
X_1	X_2		
X_1		X_3	
X_1			X_4
	X_2	X_3	
	X_2		X_4
		X_3	X_4
X_1	X_2	X_3	
X_1	X_2		X_4
X_1		X_3	X_4
	X_2	X_3	X_4
X_1	X_2	X_3	X_4

Πίνακας 4.3 Δυνατοί συνδυασμοί 4 ανεξάρτητων μεταβλητών.

Η μέθοδος στηρίζεται στην επιλογή εκείνου του συνδυασμού των μεταβλητών που θα πληροί όσο το δυνατόν επαρκέστερα τρία βασικά κριτήρια, στενά συνδεδεμένα μεταξύ τους:

i. Η τιμή του συντελεστή πολλαπλού προσδιορισμού $\mathbf{R^2}$.

Όσο αυξάνει η τιμή του συντελεστή τόσο καλύτερη προσαρμογή δίνει η εξίσωση. Η είσοδος όμως μίας νέας μεταβλητής στην εξίσωση ποτέ δεν μειώνει το R^2. Αντίθετα, προκαλεί πάντα αύξηση, δεν υπάρχει όμως κάποιος συγκεκριμένος έλεγχος για να διαπιστώσουμε τη σημαντικότητα της αύξησης αυτής. Γενικά, μία αύξηση του R^2 τουλάχιστον κατά 5% θεωρείται ικανοποιητική για να κρατηθεί η νέα μεταβλητή στην εξίσωση και να συν υποστεί παραπέρα ελέγχους μέχρι την τελική ένταξή της. Ο συντελεστής αυτός θα πρέπει να χρησιμοποιείται με προσοχή, γιατί σχεδόν πάντα αυξάνει θεαματικά το ποσοστό του όταν προσθέτουμε συνεχώς νέους όρους στην εξίσωση της παλινδρόμησης, είτε αυτοί αποτελούν διαφορετικές δυνάμεις της ίδιας μεταβλητής X, είτε μία νέα μεταβλητή, είτε είναι γινόμενα μεταξύ των μεταβλητών. Η ποσοστιαία αυτή αύξηση παρατηρείται ειδικότερα στις πολυωνυμικές εξισώσεις όπου ο συντελεστής μπορεί να προσεγγίσει ακόμα και τη μονάδα (τέλεια προσαρμογή των στοιχείων), προσθέτοντας απλά διαφορετικές δυνάμεις της ίδιας μεταβλητής X στην εξίσωση (X_2, X_3, X_5 κτλ.). Η αύξηση ή καλύτερα η βελτίωση του συντελεστή πολλαπλού προσδιορισμού θεωρείται στατιστικά έγκυρη όταν μετά την ένταξη ενός νέου όρου στην εξίσωση της παλινδρόμησης η ποσότητα του αθροίσματος των υπολειμμάτων ESS της νέας εξίσωσης μειώνεται σε μέγεθος, τουλάχιστον ίσο με εκείνο της προηγούμενης ποσότητας του μέσου σφάλματος των υπολειμμάτων EMS στην οποία δεν συμμετείχε ο νέος αυτός όρος. Σε αντίθετη περίπτωση, η νέα εξίσωση θα περιέχει συγκριτικά μεγαλύτερο μέσο σφάλμα EMS μη επιθυμητό.

Το κριτήριο R^2 θα πρέπει να χρησιμοποιείται με αποφασιστική βαρύτητα μόνο σε συγκρίσεις εξισώσεων που έχουν ίσο αριθμό μεταβλητών.

ii. Το μέσο σφάλμα των υπολειμμάτων **EMS** ή απλούστερα $\mathbf{s^2}$.

Συνήθως, για λόγους υπολογιστικής ευχέρειας χρησιμοποιείται η τετραγωνική ρίζα του μέσου σφάλματος,

$$s = \sqrt{EMS} \qquad (4.14)$$

γνωστή και ως *τυπικό σφάλμα της πολλαπλής παλινδρόμησης*. Ο αριθμός των μεταβλητών που εισάγονται στην ανάλυση πρέπει να είναι πάντοτε με σειρά προτεραιότητας ως προς τη σημαντικότητά τους που καταδεικνύεται από την τιμή του ελέγχου F. Η τιμή του μέσου σφάλματος EMS στην αρχή, όταν οι εισαγόμενοι όροι

είναι ελάχιστοι (ένας ή δύο) είναι χαμηλή, ακολούθως προσεγγίζει μία ελάχιστη τιμή και μετέπειτα αυξάνει ελαφρώς, καθώς συνεχίζουν να αυξάνουν και οι εισαγόμενοι όροι. Η περιοχή των τιμών EMS γύρω από την ελάχιστη τιμή προσδιορίζουν και τις μεταβλητές που προσαρμόζονται καλύτερα στο μοντέλο. Η προτεραιότητα που δίνουμε στην είσοδο κάποιας μεταβλητής στην ανάλυση, συνήθως καθορίζεται από τη σημαντικότητα που δίνει κάθε μία, όταν παλινδρομείται μόνη της με την εξαρτημένη Y, στην απλή γραμμική της σχέση δηλαδή με την Y.

Τα κριτήρια R^2και EMS, αν δεν υπάρχει συνδυασμός κρίσης που να συνοδεύεται από ομοειδείς ομάδες ένταξης μεταβλητών ή ισάριθμες ομάδες ένταξης, μειονεκτούν στο γεγονός ότι δεν παρέχουν συγκρίσιμα αποτελέσματα μεταξύ των μοντέλων. Για παράδειγμα, ένα μοντέλο με τις μεταβλητές X_1 και X_3 δεν μπορεί να συγκριθεί με ένα άλλο που περιέχει τις μεταβλητές X_1, X_2 και X_4, συγκρίνεται όμως ευχερώς με το μοντέλο που περιέχει τους όρους X_1, X_3 και X_4, δηλαδή δύο ίδιους όρους συν ένα τρίτο. Τα κριτήρια, δηλαδή, είναι αποτελεσματικά όταν συγκρίνουν ομοειδή μοντέλα που διαφέρουν μεταξύ τους ως προς την ένταξη ενός όρου κάθε φορά. Επίσης, δύο μοντέλα με τρεις μεταβλητές το καθένα που διαφέρουν τουλάχιστον ως προς έναν όρο, μπορούν να συγκριθούν με βάση τα δύο κριτήρια, όπως για παράδειγμα το μοντέλο με τους όρους μεταβλητές X_2, X_4 και X_5 και το μοντέλο με τους όρους μεταβλητές X_3, X_4 και X_6.

iii. Το στατιστικό κριτήριο C_p.

Το κριτήριο αυτό στηρίζεται σε δύο δεδομένα: το ένα λαμβάνει υπόψη την πλήρη εξίσωση παλινδρόμησης, δηλαδή με όλους τους όρους μαζί, και το άλλο, στη συνέχεια, επιλέγει το μοντέλο εκείνο που παρουσιάζει το μικρότερο σφάλμα ESS_p με p παρόντες όρους (συμπεριλαμβανομένης και της παραμέτρου a, εφόσον αυτή συμμετέχει στην εξίσωση). Το κριτήριο C_p υπολογίζεται από τη σχέση:

$$C_p = \frac{ESS_p}{\frac{ESS_m}{n-m-1}} - (n - 2p) \qquad (4.15)$$

Η ποσότητα ESS_m είναι το άθροισμα των τετραγώνων των υπολειμμάτων με τη συμμετοχή όλων των m όρων με n πλήθος παρατηρήσεων και $n - m - 1$ βαθμούς ελευθερίας. Το κριτήριο C_p αξιολογείται θετικά, όταν η τιμή του είναι γενικά μικρή και παρουσιάζει στατιστικά σημαντική βαρύτητα, όταν πλησιάζει τον αριθμό p των ενταγμένων όρων ή ακόμα καλύτερα όταν είναι $C_p \leq p$. Έτσι, αν η τιμή C_p ισούται

περίπου με p, σημαίνει ότι το μοντέλο είναι σχετικά ακριβές (έχει δηλαδή μικρή διακύμανση) και επιτρέπει να εκτιμηθούν ορθώς οι μερικοί συντελεστές της παλινδρόμησης όπως και οι τιμές πρόβλεψης αυτής. Αντίθετα, τιμές $\mathrm{C_p} > p$ δηλώνουν σημαντική έλλειψη προσαρμογής του μοντέλου. Στα μειονεκτήματα του κριτηρίου συγκαταλέγεται η έλλειψη συγκεκριμένου ελέγχου για τη σημαντικότητα της σύγκρισης διαφορετικών ομάδων μεταβλητών με παραπλήσιες τιμές $\mathrm{C_p}$ και p. Η μέθοδος αξιολόγησης του καταλληλότερου πολυσυνδυασμού βελτιώνεται σημαντικά, αν ληφθούν υπόψη τα ακόλουθα:

- Όταν συγκρίνουμε μοντέλα με τον ίδιο αριθμό όρων p, η επιλογή εκείνου με τη μεγαλύτερη τιμή R^2ισοδυναμεί με την επιλογή του μοντέλου με την ελάχιστη ποσότητα του τυπικού σφάλματος s.
- Όταν συγκρίνουμε μοντέλα με διαφορετικό αριθμό όρων, η επιλογή του καταλληλότερου μοντέλου συντελείται με το κριτήριο R^2_{adj}, η μεγαλύτερη τιμή του οποίου ισοδυναμεί με την επιλογή του μοντέλου με την ελάχιστη ποσότητα του τυπικού σφάλματος s.
- Όταν συγκρίνουμε ταυτόχρονα μοντέλα με ίδιο ή διαφορετικό αριθμό όρων, η επιλογή των p όρων για την ακριβέστερη περιγραφή της εξίσωσης πολλαπλής παλινδρόμησης θεωρείται άριστη, όταν συνδυάζονται από κοινού: πολύ χαμηλές τιμές των C_p και s και πολύ υψηλές τιμές των R^2 και R^2_{adj}.

Στα μειονεκτήματα της μεθόδου του καταλληλότερου συνδυασμού προσαρμογής των μεταβλητών συγκαταλέγονται:

a) Ο τεράστιος αριθμός ανάπτυξης πολυσυνδυασμών μεταξύ των μεταβλητών που αυξάνεται εκθετικά σύμφωνα με τον τύπο: $2m - 1$. Για παράδειγμα, όταν έχουμε να επιλέξουμε από 6 υποψήφιους όρους, οι συγκρινόμενες παλινδρομήσεις φτάνουν τον αριθμό 63, γεγονός δυσανάλογα χρονοβόρο ακόμα και με τη χρήση ηλεκτρονικών υπολογιστών.

b) Αν κάποιος όρος εμφανίζει ιδιομορφίες ως προς την ποιότητα συλλογής των στοιχείων ή ως προς τον τρόπο αξιολόγησης αυτών, είναι δυνατόν η ένταξή του στο μοντέλο να παρουσιάζει πολύ υψηλό R^2, με αποτέλεσμα να προκαλεί τη διαστρέβλωση της σημαντικότητας της παλινδρόμησης.

   Το μειονέκτημα αυτό εξαλείφεται αν χρησιμοποιήσουμε μία πολύ πρακτική μέθοδο ελέγχου της αξιοπιστίας του επιλεγέντος μοντέλου:

Τα στοιχεία των υποψήφιων μεταβλητών χωρίζονται με τυχαίο τρόπο σε δύο ομάδες. Χρησιμοποιούμε την πρώτη για να βρούμε την άριστη επιλογή πολυσυνδυασμού και έπειτα συγκρίνουμε τις προβλεπόμενες τιμές από την εξίσωση της παλινδρόμησης με τις πραγματικές τιμές της δεύτερης ομάδας κατά πόσο ταιριάζουν επιτυχώς. Η μέθοδος αυτή ελέγχει την «ευρωστία» του επιλεγέντος μοντέλου και είναι εξαιρετικά δημοφιλής μεταξύ των ερευνητών, ιδιαίτερα σε ποικίλες αναλύσεις που στηρίζονται στην απλή και πολλαπλή παλινδρόμηση.

4.6.2. Προοδευτική ή Σταδιακή Ένταξη των Μεταβλητών (Forward Selection).

Με τη μέθοδο αυτή επιλέγουμε πρώτα μία ανεξάρτητη μεταβλητή και ακολούθως εισάγουμε στην εξίσωση νέες μεταβλητές, μία κάθε φορά, μέχρις ότου η παραπέρα ένταξη αυτών να μην αυξάνει σημαντικά το συντελεστή πολλαπλού προσδιορισμού, R^2. Δύο κριτήρια χρησιμοποιούνται στη μέθοδο αυτή:

i. Ο υπολογισμός του $\mathbf{R^2}$ για κάθε νέα ένταξη μεταβλητής.

Είναι φρονιμότερο να εισάγουμε από την αρχή πάντοτε εκείνες τις μεταβλητές που αυξάνουν περισσότερο το R^2.

ii. Ο υπολογισμός του στατιστικού κριτηρίου F

Το οποίο εξετάζει τη σημαντικότητα της αύξησης του R^2 (και κατ' επέκταση τη μείωση της ποσότητας EMS) σε κάθε νέα εισαγωγή μεταβλητής, γνωστός και ως έλεγχος F-ένταξης. Ο έλεγχος αυτός στηρίζεται στην εξής διαδικασία:

Καμία νέα μεταβλητή δεν εισάγεται σε μία ανάλυση πολλαπλής παλινδρόμησης χωρίς προηγούμενο στατιστικό έλεγχο. Για κάθε νέα εισαγόμενη μεταβλητή στην εξίσωση, η σημαντικότητα της μείωσης του EMS ή αύξησης του R^2, ελέγχεται από την τιμή F του ελέγχου που στηρίζεται στην αρχή του πλεονάζοντος αθροίσματος **των τετραγώνων των υπολειμμάτων $\boldsymbol{ESS_a}$.** Το πλεονάζον $\mathrm{ESS_a}$ είναι το τμήμα του σφάλματος των υπολειμμάτων ESS που αφαιρείται (πλεονάζει) όταν ένας νέος όρος εισάγεται προς εξέταση στην εξίσωση, επιπρόσθετα στους ήδη υπάρχοντες:

$$ESS_a = ESS_p - ESS_{p+1}.$$

Η ποσότητα ESS_p δηλώνει το σφάλμα των υπολειμμάτων με p παρόντες όρους συμπεριλαμβανομένης και της παραμέτρου a εφόσον είναι στατιστικά σημαντική ή αν αναγκαστικά υπάρχει. Η ποσότητα ESS_{p+1} δηλώνει πάλι το σφάλμα των υπολειμμάτων αλλά με έναν επιπλέον όρο. Η σημαντικότητα του πλεονάζοντος σφάλματος ελέγχεται από την τιμή F του ελέγχου:

$$F = \frac{ESS_\alpha}{\frac{ESS_{P+1}}{n-p-1}} \qquad (4.17)$$

4.6.3. Προοδευτική ή Σταδιακή Απόρριψη των Μεταβλητών (Backward Eliminitaion).

Η επιλογή των μεταβλητών εδώ γίνεται με ακριβώς αντίθετο τρόπο. Εισάγονται πρώτα όλες οι υποψήφιες μεταβλητές στο μοντέλο (πλήρης πολλαπλή παλινδρόμηση) και μετά απορρίπτουμε σταδιακά, μία κάθε φορά, εκείνες που δεν μειώνουν σημαντικά το συντελεστή R^2. Υπενθυμίζεται ότι, ο συντελεστής R^2 εμφανίζεται πάντα μεγαλύτερος όσον αυξάνει και ο αριθμός των μεταβλητών στην εξίσωση, χωρίς να σημαίνει, στατιστικά, τίποτα το ουσιαστικό. Η μεταβλητή εκείνη που έχει τη μικρότερη τιμή του στατιστικού κριτηρίου F ή t, που στη μέθοδο αυτή για ευνόητους λόγους ονομάζεται ***F*-απόρριψης** ή ***t*-απόρριψης** (εφόσον πρόκειται για το μερικό συντελεστή παλινδρόμησης της μεταβλητής), είναι η πρώτη που θα απαλειφθεί από την εξίσωση. Αμέσως μετά την απόρριψη, υπολογίζονται πάλι οι τιμές F-απόρριψης ή t-απόρριψης των υπόλοιπων υποψήφιων για απόρριψη μεταβλητών και απορρίπτεται εκ νέου εκείνη με τη μικρότερη τιμή, μη στατιστικά σημαντική, σχετικά με το επιλεγμένο επίπεδο σημαντικότητας α. Η διαδικασία της προοδευτικής απόρριψης των μεταβλητών περατώνεται, όταν η τιμή F-απόρριψης είναι στατιστικά σημαντική για όλες τις υπόλοιπες υποψήφιες μεταβλητές, οι οποίες πλέον επιλέγονται ως οι μεταβλητές της τελικής εξίσωσης.

4.6.4. Αμφίδρομη Επιλογή των Μεταβλητών ή Μέθοδος Επιλογής Βήμα προς Βήμα (Step by Step Selection).

Οι προηγούμενες δύο μέθοδοι παρουσιάζουν ένα σοβαρό μειονέκτημα:

- Στην προοδευτική ένταξη, μόλις μία μεταβλητή ενταχθεί στην εξίσωση, αυτή παραμένει για πάντα εκεί χωρίς να έχουμε τη δυνατότητα απομάκρυνσής της,

αν κρίνουμε ότι η συνεισφορά της είναι μικρή, αν αυξάνει δηλαδή, λίγο μόνο η τιμή R^2 σχετικά με τις άλλες ενταχθείσες μεταβλητές.

- Στην προοδευτική απόρριψη μόλις μία μεταβλητή απορριφθεί, δεν έχουμε τη δυνατότητα επαναφοράς της στην εξίσωση, παρόλο που αυτή μπορεί να καταστεί ουσιώδης μεταβλητή μόνο μετά την απόρριψη άλλων μεταβλητών.

Η δυσχέρεια αυτή διορθώνεται με την αμφίδρομη επιλογή των μεταβλητών με την οποία προχωρούμε **βήμα προς βήμα**, εξετάζοντας πάντα μία μεταβλητή κάθε φορά και την οποία μπορούμε να εντάξουμε ή να απορρίψουμε στην εξίσωση κατά βούληση ή να την επαναφέρουμε αργότερα για επανεξέταση. Η αμφίδρομη επιλογή αποτελεί, ουσιαστικά, συνδυασμό των δύο προηγούμενων μεθόδων, επειδή ελέγχει κάθε μεταβλητή με την τιμή $F-ένταξης$ ή $F-απόρριψης$, ανάλογα με την κατεύθυνση επιλογής του χρήστη τη στιγμή εκείνη.

Τα στατιστικά προγράμματα υπολογίζουν, πλην του $ελέγχου\ F$ και του $ελέγχου\ t$ των μερικών συντελεστών παλινδρόμησης, εξίσου ευχερώς και την ακριβή πιθανότητα p στην οποία η τιμή $F-ένταξης$ ή $F-απόρριψης$ ($t-ένταξης$ ή $t-απόρριψης$) δεν είναι στατιστικά σημαντική. Η πιθανότητα αυτή είναι γνωστή και ως $\boldsymbol{p-ένταξης}$ ή $\boldsymbol{p-απόρριψης}$, ανάλογα με το χειρισμό και ουσιαστικά, η τιμή της είναι αυτή που κυριολεκτικά μας ενδιαφέρει απ` όλες τις παραπάνω διαδικασίες επιλογής. Όταν η πιθανότητα αυτή είναι μικρότερη ή ίση της οριακής που επιλέγεται από εμάς π.χ. $p \leq 0{,}05$, ανεξάρτητα αν πρόκειται για ένταξη ή απόρριψη της μεταβλητής, τότε η μεταβλητή αυτή πρέπει να προστεθεί στην εξίσωση, εφόσον πρόκειται για προοδευτική ένταξη ή να παραμείνει σε αυτήν αν πρόκειται για προοδευτική απόρριψη.

Οι μέθοδοι της άριστης επιλογής των μεταβλητών αποτελούν πολύτιμο βοήθημα για το σωστό προσδιορισμό της εξίσωσης της πολλαπλής παλινδρόμησης, ιδιαίτερα όταν συνδυάζονται μεταξύ τους τα ευρήματα της καθεμιάς που δεν είναι απαραίτητο να συμπλέουν πάντοτε. Οι μέθοδοι αυτές παρουσιάζουν όμως τους ακόλουθους κινδύνους που θα πρέπει ο χρήστης να είναι ενήμερος:

a) Επειδή οι διαδικασίες επιλογής των μεταβλητών γίνονται με τη βοήθεια στατιστικών προγραμμάτων, είναι πολύ πιθανόν να γίνει, καταρχήν αυτόματα, τυχαία επιλογή από το σύνολο των μεταβλητών και να παρουσιάζεται έτσι το μοντέλο πολύ ισχυρό μόνο από καθαρή σύμπτωση, χωρίς οι μεταβλητές αυτές να είναι οι πλέον σημαντικές.

b) Μερικές φορές οι αλγόριθμοι πάνω στους οποίους στηρίζονται οι αναλύσεις των μεθόδων επιλογής, αδυνατούν να επιλέξουν τις μεταβλητές εκείνες που η εξίσωσή τους δίνει τη μέγιστη τιμή R^2.

c) Οι αυτόματες επιλογές δεν λαμβάνουν υπόψη και κάποιες άλλες μεταβλητές που μπορεί να μην είναι ιδιαίτερα σημαντικές από στατιστικής πλευράς, είναι όμως εξαιρετικά σημαντικές για τον ερευνητή που τις επέλεξε και που η ένταξή τους στην εξίσωση δίνει μεγάλη πρακτική αξία στο μοντέλο της πολλαπλής παλινδρόμησης.

d) Σε στατιστικές μεθόδους που στηρίζονται αποκλειστικά στην διαδικασία της πολλαπλής παλινδρόμησης, όπως για παράδειγμα η ανάλυση της επιφάνειας απόκρισης ή η ανάλυση των συστατικών μείξης, όλοι οι υποψήφιοι όροι εντάσσονται υποχρεωτικά στο μοντέλο, χωρίς την ανάγκη της άριστης επιλογής των μεταβλητών, αδιάφορα δηλαδή, αν η σημαντικότητα πολλών από αυτές είναι μηδαμινή.

4.7 Εξισώσεις και Σημαντικότητα της Πρόβλεψης

Μετά την προσαρμογή των μεταβλητών στο μοντέλο της πολλαπλής παλινδρόμησης υπολογίζονται οι προβλεπόμενες τιμές της Y, θέτοντας αντίστοιχες τιμές σε όλες τις συμμετέχουσες μεταβλητές X_i.

Η σημαντικότητα των προβλεπόμενων τιμών εκφράζεται με τα 95% όρια εμπιστοσύνης τα οποία προκύπτουν από τη σχέση,

$$\hat{Y}_i \pm t_{0,05,(n-m-1)} \cdot (\tau\upsilon\pi\iota\kappa\acute{o}\ \sigma\phi\acute{a}\lambda\mu\alpha)$$

Το τυπικό σφάλμα εξάγεται με διαφορετικούς υπολογισμούς, εξαρτώμενο πάντοτε από τη φύση του ζητούμενου σε κάθε πρόβλεψη:

- Για το μέσο όρο της εξαρτημένης μεταβλητής Y το τυπικό σφάλμα είναι,

$$s_{\bar{Y}} = \sqrt{\frac{EMS}{n}}$$

- Η εκτιμώμενη μέση τιμή της $\hat{Y}_i$ για συγκεκριμένες τιμές της $X_1, X_2 \dots X_m$ θα προέρχεται από το τυπικό σφάλμα,

$$s_{\hat{Y}} = \sqrt{EMS \cdot \left[\frac{1}{n} + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j\right]}, \text{ με} \quad x_i = X_i - \bar{X}_i \quad , \quad x_j = X_j - \bar{X}_j$$

και c_{ij} είναι οι πολλαπλασιαστές του Gauss επί της διαγωνίου και εκτός αυτής αντίστοιχα του αντίστροφου πίνακα του αθροίσματος των τετραγώνων των χιαστί γινομένων (Πίνακας 4.2).

- Η προβλεπόμενη τιμή $\hat{Y}_i$ από μία νέα πρόσθετη σειρά τιμών των μεταβλητών $X_1, X_2 \dots X_m$ λαμβανόμενη ως αποτέλεσμα επανάληψης του πειράματος, έχει τυπικό σφάλμα,

$$(s_{\hat{Y}})_1 = \sqrt{EMS \cdot \left[1 + \frac{1}{n} + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j\right]}$$

Η προβλεπόμενη μέση τιμή $\hat{Y}$ που προέρχεται από λ πρόσθετες σειρές επαναληπτικών τιμών των μεταβλητών $X_1, X_2 \dots X_m$, έχει τυπικό σφάλμα,

$$(s_{\hat{Y}})_\lambda = \sqrt{EMS \cdot \left[\frac{1}{\lambda} + \frac{1}{n} + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j\right]}$$

4.8 Διαγνωστικά Κριτήρια της Εγκυρότητας της Πολλαπλής Παλινδρόμησης

Προτού περιγραφεί η εξίσωση της πολλαπλής παλινδρόμησης όπως την αναλύσαμε παραπάνω, προηγείται πάντοτε η εξέταση της φύσης των μεταβλητών που συμμετέχουν με διάφορα κριτήρια. Τα κριτήρια αυτά ελέγχουν ορισμένες βασικές προϋποθέσεις που πρέπει να τηρούν τα στοιχεία των μεταβλητών για να διασφαλιστεί έτσι η εγκυρότητα της εξίσωσης της παλινδρόμησης. Ιδιαίτερη αναφορά για το μέγεθος των συνεπειών ως προς την αξιοπιστία προσαρμογής ενός μοντέλου παλινδρόμησης έχει επανειλημμένα δοθεί από τη διεθνή βιβλιογραφία (Besley et al, 1980, Velleman & Welch, 1981,Cook & Weisburg, 1982). Τα κυριότερα διαγνωστικά κριτήρια της πολλαπλής παλινδρόμησης είναι τα ακόλουθα.

4.8.1. Εξέταση των Υπολειμμάτων ως Προς την Κανονικότητα και την Ομοιογένεια της Διασποράς τους.

Η διασπορά εξετάζεται με γραφικές απεικονίσεις των υπολειμμάτων στον άξονα των Y και τις ανεξάρτητες μεταβλητές, καθώς και τις προσαρμοσμένες τιμές στον άξονα

των X. Απαιτείται η λεπτομερής εξέταση των υπολοίπων μέσω των ακόλουθων γραφικών μεθόδων που εξετάζουν τη συμβολή συγκεκριμένων μεταβλητών στο μοντέλο:

- Διάγραμμα πρόσθετων μεταβλητών (Added Variable Plot)
- Διάγραμμα μερικών υπολοίπων (Partial Residual Plot)

Είναι ευκολονόητο πως απαιτείται ένα γράφημα των υπολειμμάτων τη φορά για κάθε ανεξάρτητη μεταβλητή προκειμένου, να εντοπιστεί ποια από τις μεταβλητές αυτές είναι ενδεχομένως υπεύθυνη για την έλλειψη κανονικότητας και ένα μόνο γράφημα για τις προσαρμοσμένες τιμές. Οι κάθετες αποστάσεις των σημείων από την ευθεία προσαρμογής είναι γνωστές με το όνομα υπολείμματα. Ο έλεγχός τους αποτελεί αναπόσπαστο τμήμα της σημαντικότητας της παλινδρόμησης, διότι τεκμηριώνει αποφασιστικά την ποιότητα της σχηματιζόμενης ευθείας προσαρμογής, με ή χωρίς την ανάγκη μετασχηματισμού των μεταβλητών. Η εγκυρότητα της γραμμικής σχέσης της παλινδρόμησης με τη βοήθεια των υπολειμμάτων βασίζεται στην εκπλήρωση δύο προϋποθέσεων:

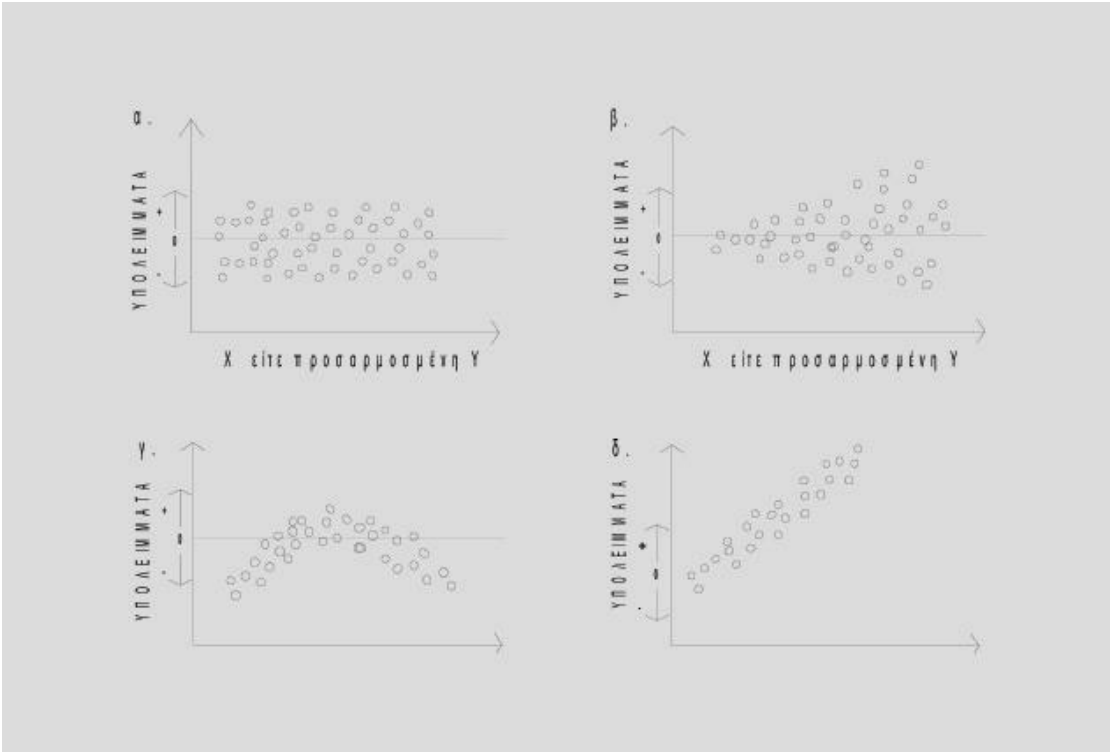
i. Τα υπολείμματα να ακολουθούν κανονική κατανομή η οποία διαπιστώνεται με τη χρήση ενός από τους ελέγχους Kolmogorov-Smirnov, Shapriro-Wilk, D' Agostino, Anderson-Darling ή γραφικών ελέγχων της κανονικότητας.

ii. Τα υπολείμματα να μην εκδηλώνουν τάση μεταβολής ως αποτέλεσμα της δράσης των μεταβλητών X και Y (Σχήμα 4.3). Η δεύτερη αυτή συνθήκη διαπιστώνεται από τη μελέτη δύο γραφημάτων, με τα υπολείμματα $(Y_i - \hat{Y}_i)$ να παρίστανται στον άξονα των Y και στον άξονα των X να παρίστανται, για μεν το ένα γράφημα οι πραγματικές τιμές της μεταβλητής X για δε το δεύτερο οι προσαρμοσμένες τιμές $\hat{Y}_i$. Αν τα υπολείμματα κατανέμονται διάσπαρτα και στα δύο γραφήματα, τότε υπάρχει ομοιομορφία της διασποράς τους πάνω και κάτω του μηδενός (Σχήμα 4.3α). Αναμένεται, δηλαδή, να έχουν μέσο όρο περίπου ίσο με το μηδέν και διακύμανση που είναι ανεξάρτητη από το μέγεθος μεταβολής είτε της ανεξάρτητης τιμής X είτε της προσαρμοσμένης μεταβλητής Ŷ.

   Όταν διαπιστώνεται ετερομορφία, ή πιο εξειδικευμένα ιδιομορφία, στη διασπορά των υπολειμμάτων, στο πρώτο ή και στο δεύτερο γράφημα, αυτή μπορεί να οφείλεται:

a) Στην αυξημένη μεταβλητότητα της Y καθώς αυξάνουν οι τιμές της X (Σχήμα 4.3β), οπότε τα υπολείμματα εμφανίζουν τη μορφή δέσμης. Στην περίπτωση αυτή προτείνεται ο μετασχηματισμός με τους λογάριθμους ή σπανιότερα με την τετραγωνική ρίζα των μεταβλητών X και Y, για να επανέλθει η ομοιοδιασπορά των σημείων στα γραφήματα. Εννοείται ότι, μετά την εφαρμογή των μετασχηματισμών ακολουθεί απαρέγκλιτα η επανεξέταση των υπολειμμάτων όπως παραπάνω.

b) Στην έλλειψη γραμμικής σχέσης μεταξύ των X και Y. Όταν τα υπολείμματα διαγράφουν καμπύλη μεταβολή (Σχήμα 4.3γ) σημαίνει ότι η δευτεροβάθμια εξίσωση $\hat{Y} = a + b_1X + b_2X^2$ συνιστά την καταλληλότερη σχέση, ενώ όταν τα υπολείμματα διαγράφουν οφιοειδές σχήμα η εξίσωση που το περιγράφει έχει τη γενικότερη πολυωνυμική μορφή $\hat{Y} = a + b_1X + b_2X^2 + \ldots + b_nX^n$. Τέλος, όταν τα υπολείμματα διανύουν ανοδική κατεύθυνση (Σχήμα 4.3δ), αυτό σημαίνει την ανάγκη προσθήκης μίας δεύτερης μεταβλητής X στην εξεταζόμενη σχέση X και Y, δηλαδή στην ύπαρξη πολλαπλής παλινδρόμησης $\hat{Y} = a + b_1X + b_2X^2$. Η περίπτωση της ανοδικής τάσης μπορεί να οφείλεται μερικές φορές και σε υπολογιστικό λάθος αποκλειστικά.



Σχήμα 4.3 Γραφική εξέταση των υπολειμμάτων σε σχέση με την ανεξάρτητη μεταβλητή X και την προσαρμοσμένη στήλη Y. Τα στοιχεία εμφανίζουν: α) ομοιοδιασπορά, β) διασπορά δέσμης, γ) καμπυλότητα δ) γραμμική μεταβολή.

4.8.2. Εξέταση των Τυποποιημένων Υπολειμμάτων (Standardized Residuals).

Τα υπολείμματα, όταν διαιρεθούν με την τυπική απόκλισή τους, μετατρέπονται σε τυποποιημένα,

$$r_s = \frac{Y_i - \hat{Y}}{\sqrt{EMS}} \qquad (4.18)$$

τα οποία έχουν μέσο όρο μηδέν και διακύμανση ίση με 1. Τιμές των τυποποιημένων υπολειμμάτων μεγαλύτερες από |2| αλλά συνήθως μεγαλύτερες από |3| αποτελούν ισχυρή ένδειξη σημαντικής απομάκρυνσης των εξεταζόμενων τιμών Y_i από την ευθεία προσαρμογής. Τα τυποποιημένα υπολείμματα έχουν μικρότερη διακύμανση κοντά στο κέντρο των τιμών και μεγαλύτερη όσο οι τιμές απομακρύνονται από το κέντρο. Το μειονέκτημα αυτό λύνεται αν στην εξίσωση προστεθούν και οι συντελεστές μόχλευσης h_i,

$$r_s = \frac{Y_i - \hat{Y}}{\sqrt{EMS \cdot (1 - h_i)}} \qquad (4.19)$$

Οι δύο τύποι των υπολειμμάτων συγκλίνουν στο ίδιο αποτέλεσμα όταν το δειγματοληπτικό μέγεθος είναι μεγάλο.

Καλύτερη προσαρμογή της προηγούμενης εξίσωσης στην εντόπιση ύποπτων τιμών αποτελούν τα ***υπολείμματα τύπου Student***. Αυτά υπολογίζονται αφαιρώντας την παρατήρηση i και υποβάλλοντας τις υπόλοιπες $n-1$ στην ανάλυση της παλινδρόμησης ένεκα της οποίας προκύπτει το αφαιρετικό υπόλειμμα . Η διαδικασία επαναλαμβάνεται n φορές και με τον τρόπο αυτό διαμορφώνεται η αφαιρετική στήλη των υπολειμμάτων και κατ' επέκταση τα υπολείμματα του Student t_i :

$$t_i = \frac{Y_i - Y_{i(i)}}{\sqrt{EMS_{(i)} \cdot (1 - h_{i(i)})}} \qquad (4.20)$$

4.8.3. Εξέταση των Συντελεστών Μόχλευσης ή Επιρροής (Leverage Coefficients) h_i των Τιμών της Ανεξάρτητης Μεταβλητής.

Οι συντελεστές αυτοί, ένας για κάθε τιμή, υπολογίζονται από τη σχέση (η οποία ισχύει ως έχει μόνο στην απλή γραμμική παλινδρόμηση),

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})}{\sum x^2} \qquad (4.21)$$

όπου n είναι το πλήθος των τιμών της ανεξάρτητης μεταβλητής.

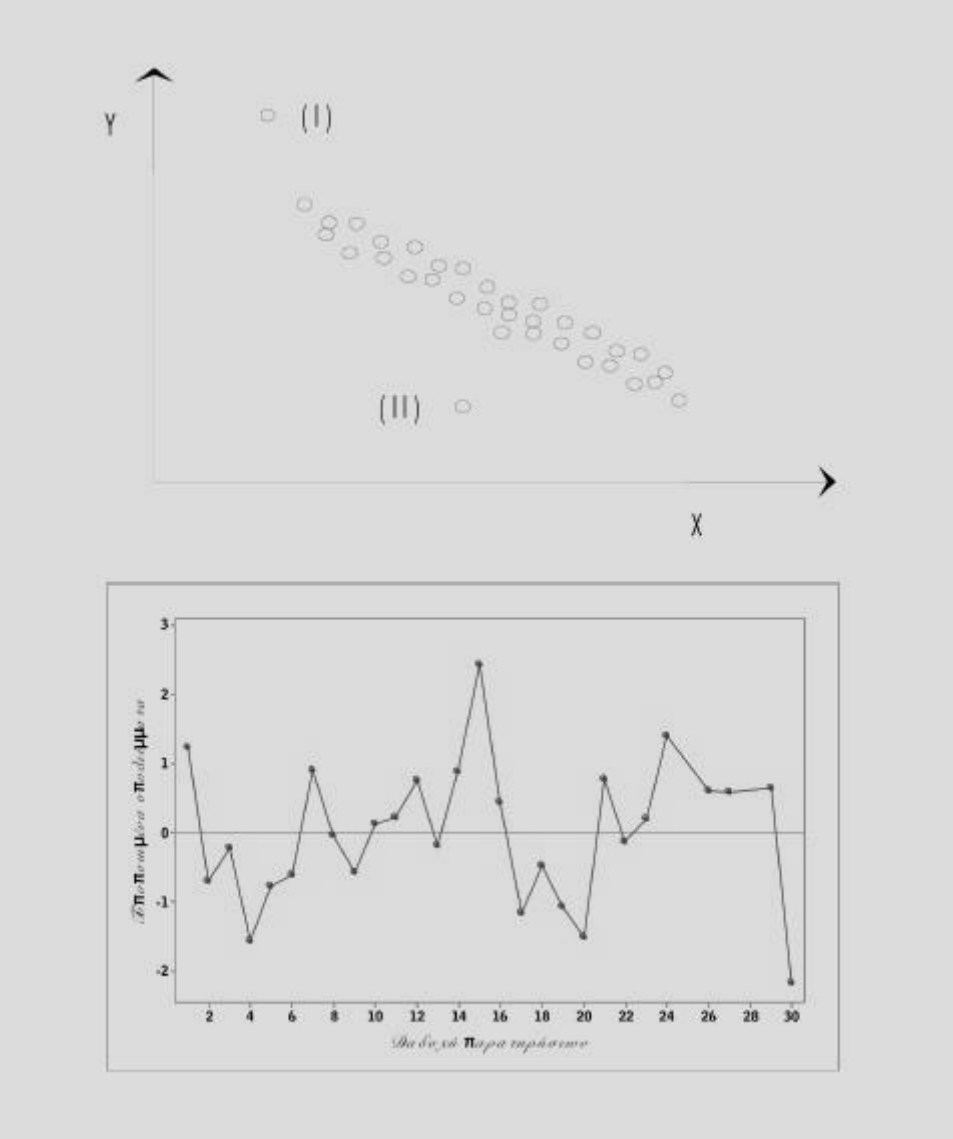
Οι συντελεστές μόχλευσης παίρνουν τιμές από 0 μέχρι 1 και αποτελούν δείκτες της επιρροής των τιμών X_i της μεταβλητής X στον τρόπο προσαρμογής της ευθείας παλινδρόμησης. Με άλλα λόγια, είναι ενδείξεις του μεγέθους της απόκλισης μίας τιμής X_i από το μέσο όρο, εξαρτώμενη αποκλειστικά από την ανεξάρτητη μεταβλητή X. Ουσιαστικά, η εξάρτηση αυτή σημαίνει ότι μελετούμε την οριζόντια απόκλιση των σημείων από την ευθεία προσαρμογής. Τιμές των συντελεστών επιρροής μεγαλύτερες από $2p/n$, κατ` άλλους από $3p/n$, δηλώνουν έντονη απομάκρυνση των αντίστοιχων σημείων X_i από την ευθεία προσαρμογής. Ο όρος p αναφέρεται στον αριθμό συμμετοχής των συντελεστών b (ή αλλιώς στις συμμετέχουσες ανεξάρτητες μεταβλητές) συν 1, εφόσον συμμετέχει και η παράμετρος a. Έτσι, για την απλή γραμμική παλινδρόμηση θα ισχύει $p = 2$ και για την πολλαπλή παλινδρόμηση με 3 συντελεστές θα ισχύει $p = 4$, με παρούσα την παράμετρο a. Πρακτικά, τιμές μόχλευσης μεγαλύτερες του 0,5 πρέπει να αποφεύγονται, ενώ τιμές μικρότερες του 0,2 θεωρούνται ασφαλείς και τιμές μεταξύ 0,2 και 0,5 εμπεριέχουν, αν γίνουν δεκτές, στοιχεία κινδύνου.

4.8.4. Έλεγχος της Σημαντικότητας Επιρροής των Ύποπτων (Εξωκείμενων) Τιμών (Outliers Values).

Ύποπτες τιμές νοούνται εκείνες που απέχουν πολύ από την ευθεία προσαρμογής της παλινδρόμησης και ταυτόχρονα φαίνεται να επηρεάζουν έντονα τον προσανατολισμό της κλίσης της ευθείας. Η οπτική ανίχνευση των ύποπτων τιμών γίνεται σε δύο κατευθύνσεις:

a) Οι ύποπτες τιμές βρίσκονται απόκεντρα και μακράν της ευθείας, δηλαδή στο ανώτερο ή στο κατώτερο άκρο της ευθείας και σε θέση λίγο ή πολύ απομακρυσμένη από τις νοητές προεκτάσεις της ευθείας ,Τύπος I, Σχήμα 4.4α. Η ανίχνευση των τιμών αυτών στατιστικά πραγματοποιείται με τους συντελεστές επιρροής τους, h_i.
b) Οι ύποπτες τιμές απέχουν σημαντικά από την ευθεία προσαρμογής αλλά κεντρικά, όπως φαίνεται στο Σχήμα 4.4α , Τύπος II. Η ανίχνευση των τιμών

αυτών στατιστικά πραγματοποιείται με την εξέταση των τυποποιημένων υπολειμμάτων τους.



Σχήμα 4.4 (α)Τύποι ύποπτων τιμών στην παλινδρόμηση. (β)Τυποποιημένα υπολείμματα σε διαδοχή παρατηρήσεων χωρίς ένδειξη αυτοσυσχέτισης.

Οι τιμές outliers (εξωκείμενες τιμές των παρατηρήσεων) συνιστούν μία μορφή παραβίασης της ομοσκεδαστικότητας και αντιπροσωπεύουν υψηλές τιμές υπολειμμάτων (σφαλμάτων) εκτός βέβαια εκείνων που θα μπορούσαν να δικαιολογηθούν από το υπόδειγμα παλινδρόμησης. Κατά κανόνα επικρατεί η τάση να θεωρείται μία ύποπτη τιμή ως στατιστικά σημαντική, όταν αυτή συνδυάζεται ταυτόχρονα από υψηλό συντελεστή επιρροής, $h_i > {}^{2p}/_{n}$ (ή ${}^{3p}/_{n}$) και από

τυποποιημένο υπόλειμμα μεγαλύτερο του |2|. Οι σημαντικά στατιστικά ύποπτες τιμές επιβάλλεται να απομακρύνονται από το γράφημα $X-Y$, στο οποίο οπτικά ανιχνεύονται εύκολα, ως απόκεντρα σημεία που απέχουν έντονα από την ευθεία προσαρμογής. Ύποπτες τιμές με υψηλό h_i και χαμηλό τυποποιημένο υπόλειμμα θεωρούνται γενικά αποδεκτές, όπως επίσης και ύποπτες τιμές με χαμηλό h_i και μεγάλο τυποποιημένο υπόλειμμα. Όταν τεκμηριώνονται στατιστικά οι ύποπτες τιμές, τότε διαγράφονται πρώτα τα ζεύγη των τιμών που περιέχουν τις ύποπτες τιμές και μετά υπολογίζεται ξανά η γραμμική σχέση των δύο μεταβλητών.

4.8.5. Έλεγχος της Απόστασης του Cook (1977).

Ο δείκτης αυτός ανιχνεύει ταυτόχρονα παρατηρήσεις με ασυνήθιστες τιμές X_i(υψηλός συντελεστής επιρροής) και ασυνήθιστες τιμές Y_i (μεγάλο τυποποιημένο υπόλειμμα), παρέχοντας έτσι μία συνδυασμένη συνολική μέτρηση για κάθε ζεύγος παρατηρήσεων:

$$C_d = \frac{\frac{1}{p}\cdot\frac{h_i}{1-h_i}\cdot(Y_i-\hat{Y})^2}{EMS\cdot(1-h_i)} \qquad (4.22)$$

Τιμές της απόστασης του Cook μεγαλύτερες από την οριακή $F_{0,05,(p,n-p)}$ που βρίσκεται από τον Πίνακα Π1 στο Παράρτημα, δηλώνουν ύποπτες τιμές και συνιστούν την απόρριψή τους από τους υπολογισμούς. Για την απλή γραμμική παλινδρόμηση η θεωρητική τιμή διαμορφώνεται σε $F_{0,05,(2,n-2)}$, όταν υπάρχει η παράμετρος a. Ως διαχωριστικό σημείο εντοπισμού επηρεαστικών τιμών παρατηρήσεων θεωρούνται τιμές μεγαλύτερες της τιμής $4/(n-k-1)$, όπου η =αριθμός παρατηρήσεων και k =αριθμός ανεξάρτητων μεταβλητών (Fox, 1991), ενώ άλλοι συγγραφείς θέτουν $τιμή > 1$ ως κριτήριο σοβαρής ένδειξης προβλήματος εξωκείμενης τιμής και $τιμή > 4/\eta$ ως κριτήριο ένδειξης πιθανού προβλήματος.

Γενικότερα, για την αναγνώριση και απομάκρυνση των outliers των ανεξάρτητων μεταβλητών χρησιμοποιούνται συνδυαστικά η απόσταση Cook, το στατιστικό της τιμής μόχλευσης της παρατήρησης, το τετράγωνο της απόστασης Mahalanobis και, επίσης, με τον έλεγχο των μεταβολών των τυποποιημένων συντελεστών β (beta) ή τον έλεγχο των μεταβολών στις προβλεπόμενες τιμές της πριν από και μετά την

απομάκρυνση των ύποπτων παρατηρήσεων κατά τον υπολογισμό των εξισώσεων παλινδρόμησης.

4.8.6. Έλεγχος της Αυτοσυσχέτισης των Υπολειμμάτων Ανίχνευσης

Στις βασικές υποθέσεις των απλών και πολλαπλών γραμμικών υποδειγμάτων, υποθέσαμε ότι δεν υπάρχει αυτοσυσχέτιση μεταξύ των διαταρακτικών όρων u_t. Αυτή η υπόθεση δηλώνει ότι οι τιμές των διαταρακτικών όρων είναι ανεξάρτητες μεταξύ τους (σειριακή ανεξαρτησία). Δηλαδή για δύο διαφορετικές παρατηρήσεις του διαταρακτικού όρου η συνδιακύμανση τους είναι μηδέν: $Cov(u_t, u_s) = 0,\ t \neq s$.

Αν αυτή η υπόθεση δεν ισχύει, τότε οι διαταραχές δεν είναι ανεξάρτητες κατά ζεύγη, αλλά αυτοσυσχετιζόμενες κατά ζεύγη (ή σειριακά συσχετιζόμενες). Η αυτοσυσχέτιση μπορεί να θεωρηθεί ως ειδική περίπτωση της συσχέτισης δύο μεταβλητών, αλλά αναφέρεται στη συσχέτιση δύο διαδοχικών τιμών της ίδιας μεταβλητής. Αν u_t και u_{t-1} είναι δύο διαδοχικές τιμές των καταλοίπων στην εκτίμηση ενός υποδείγματος παλινδρόμησης, τότε $\bar{u}_t = 0$ *και* $\bar{u}_{t-1} = 0$ και η αυτοσυσχέτισή τους δίνεται από:

$$r_{u_t,u_{t-1}} = \frac{\sum_{t=2}^{T} u_t u_{t-1}}{\sqrt{\sum_{t=1}^{T} u_t^2}\sqrt{\sum_{t=2}^{T} u_{t-1}^2}} \qquad (4.23)$$

Η r, όπως έχουμε προαναφέρει στο Κεφάλαιο 3, παίρνει θετικές, αρνητικές ή μηδέν τιμές, με το τελευταίο να επαληθεύει τη μη ύπαρξη αυτοσυσχέτισης. Ο διαταρακτικός όρος u_t περιλαμβάνει την επίδραση όλων των παραγόντων που δεν μπορούν να συμπεριληφθούν στα υποδείγματα. Πολλές φορές όμως, η επίδραση πολλών παραγόντων μπορεί να μην αναφέρεται στην τρέχουσα χρονική περίοδο, αλλά σε μελλοντικές.

Παράγοντες που προκαλούν αυτοσυσχέτιση:

1. Παραλειπόμενες μεταβλητές.
2. Λανθασμένη εξειδίκευση του υποδείγματος.
3. Συστηματικά σφάλματα μέτρησης.
4. Χρονικές υστερήσεις των μεταβλητών.
5. Η πηγή των στατιστικών στοιχείων και ο τρόπος επεξεργασίας τους.

Η σχέση εξάρτησης ανάμεσα σε διαδοχικές τιμές του διαταρακτικού όρου μπορεί να πάρει διάφορες μορφές:

- *Αυτοσυσχέτιση πρώτης τάξης (first–order autocorrelation)*

  Αν η τιμή του διαταρακτικού όρου της περιόδου t εξαρτάται από την περίοδο $t-1$, είναι της μορφής: $u_t = \rho u_{t-1} + \varepsilon_t$, όπου $-1 \leq \rho \leq 1$ τότε έχουμε την αυτοσυσχέτιση πρώτης τάξης ή το αυτοπαλίνδρομο σχήμα πρώτου βαθμού (first order autoregressive scheme) που συμβολίζεται με AR(1). Το ρ είναι ο συντελεστής της αυτοσυνδιακύμανσης και μετράει τον βαθμό της συσχέτισης μεταξύ δύο διαδοχικών διαταρακτικών όρων, ενώ ο ε_t είναι ο στοχαστικός διαταρακτικός όρος.

  i. Αν $\rho = 0$,τότε δεν υπάρχει αυτοσυσχέτιση.
  ii. Αν ρ πλησιάζει το 1, η τιμή της προηγούμενης παρατήρησης του σφάλματος γίνεται πιο σημαντική στον προσδιορισμό της τιμής του τρέχοντος σφάλματος και συνεπώς υπάρχει υψηλός βαθμός αυτοσυσχέτισης. Στην περίπτωση αυτή η αυτοσυσχέτιση είναι θετική.
  iii. Αν ρ πλησιάζει το 1, τότε έχουμε υψηλό βαθμό αρνητικής αυτοσυσχέτισης.

  Όταν στο στοχαστικό (διαταρακτικό) όρο u_t ισχύουν οι παρακάτω τρεις υποθέσεις: $E(u_t) = 0\,, Var(u_t) = \sigma^2 \; και \; Cov(u_t, u_{t+s}) = 0 \; για \; t \neq s$

  τότε λέμε ότι ο διαταρακτικός όρος είναι λευκός θόρυβος (white noise error term).

- *Δεύτερης τάξης αυτοσυσχέτιση*

  Η τιμή του διαταρακτικού όρου της περιόδου t εξαρτάται από την περίοδο $t-1$ και την περίοδο $t-2$, και είναι της μορφής: $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t$

- *Τρίτης τάξης αυτοσυσχέτιση*

  Η τιμή του διαταρακτικού όρου της περιόδου t εξαρτάται από τις περιόδους $t-1, t-2 \; και \; t-3$, και είναι της μορφής: $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \varepsilon_t$

Ο έλεγχος της αυτοσυσχέτισης των υπολειμμάτων ανίχνευσης η οποία οδηγεί σε συσχετισμένα σφάλματα εξετάζεται:

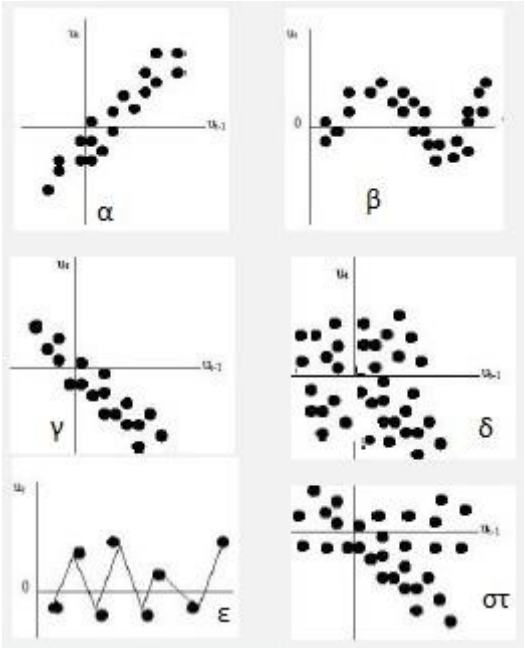
I. Διαγραμματικά με τη διάταξη των τυποποιημένων συνήθως υπολειμμάτων σε σχέση με τη διαδοχική σειρά των παρατηρήσεων του μοντέλου

παλινδρόμησης (και φυσιολογικά αναμένεται μία κατανομή των σημείων χωρίς συγκεκριμένη τάση μεταβολής, ένδειξη ανεξαρτησίας των τυπικών σφαλμάτων. Σχήμα 4.4).

II. Μέσω επίσημων test για αυτοσυσχέτιση, όπως τα παρακάτω:
   - Κριτήριο Durbin Watson (Προτιμώμενη)
   - Έλεγχος με την στατιστική t (Πραγματοποιείται μόνο σε μεγάλα δείγματα για την εύρεση αυτοσυσχέτισης πρώτης τάξης)
   - Με τη μέθοδο Breusch-Godfrey.

Διαγραμματικά

Θετική αυτοσυσχέτιση εμφανίζεται όταν διαμορφώνεται στο διάγραμμα ομάδα υπολειμμάτων με θετικό πρόσημο και αρνητική όταν διαπιστώνονται ταχείες αλλαγές στο πρόσημο διαδοχικών υπολειμμάτων. Πιο βοηθητικά είναι τα ακόλουθα σχήματα.



Σχήμα 4.5 Θετική αυτοσυσχέτιση καταλοίπων (α) και (β), αρνητική αυτοσυσχέτιση (γ) και (ε), δεν υπάρχει αυτοσυσχέτιση (δ) και (στ)

Κριτήριο Durbin Watson

Πρέπει να ικανοποιούνται οι παρακάτω υποθέσεις:

1. Το μοντέλο παλινδρόμησης περιλαμβάνει μια σταθερά.
2. Η αυτοσυσχέτιση πρέπει να είναι μόνο πρώτης τάξης
3. Η εξίσωση δεν περιλαμβάνει μια εξαρτημένη μεταβλητή με χρονική υστέρηση σαν ερμηνευτική μεταβλητή.

Η υπόθεση που θέλουμε να ελέγξουμε είναι:

H_0:Δεν υπάρχει αυτοσυσχέτιση πρώτης τάξης στα κατάλοιπα.

H_1:Υπάρχει αυτοσυσχέτιση πρώτης τάξης στα κατάλοιπα.

Έπειτα ακολουθούμε τα παρακάτω βήματα.

Βήμα 1: Εκτιμούμε το υπόδειγμα παλινδρόμησης με την μέθοδο των ελαχίστων τετραγώνων και παίρνουμε τα κατάλοιπα u_t

Βήμα 2: Υπολογίζουμε το στατιστικό του ελέγχου

$$DW = d = \frac{\sum_{t=2}^{T}(u_t - u_{t-1})^2}{\sum_{t=1}^{T} u_t^2} \qquad (4.24)$$

Οι τιμές του στατιστικού d κυμαίνονται από 0 $έως$ 4

- $d = 0$ πλήρης θετική αυτοσυσχέτιση
- $d = 2$ δεν υπάρχει αυτοσυσχέτιση
- $d = 4$ πλήρης αρνητική αυτοσυσχέτιση

Βήμα 3: Συγκρίνουμε την τιμή του στατιστικού d με την κριτική τιμή d_L (κάτω τιμή) και d_U (άνω τιμή) της αντίστοιχης κατανομής των DW (από πίνακα).

Βήμα 4: Συμπεράσματα

- Ύπαρξη θετικής αυτοσυσχέτισης αν $d < d_L$ ή αν $d > d_U$
- Ύπαρξη αρνητικής αυτοσυσχέτισης αν $(4 - d) < d_L$ ή αν $(4 - d) > d_U$
- Δεν καταλήγουμε σε συμπέρασμα αν $d_L < d < d_U$ ή $αν$ $d_L < (4 - d) < d_U$

Σχήμα 4.6 Έλεγχος αυτοσυσχέτισης με το κριτήριο DW

Συνέπειες της αυτοσυσχέτισης

- Οι εκτιμητές ελαχίστων τετραγώνων είναι αμερόληπτοι και συνεπείς. Η αμεροληψία και η συνέπεια δεν εξαρτώνται από την υπόθεση που παραβιάζεται.
- Οι εκτιμητές ελαχίστων τετραγώνων θα είναι αναποτελεσματικοί (δεν έχουν την μεγαλύτερη δυνατή διακύμανση) και συνεπώς δεν θα είναι πια οι καλύτεροι γραμμικοί εκτιμητές.
- Οι εκτιμημένες διακυμάνσεις των συντελεστών της παλινδρόμησης θα είναι μεροληπτικές και ασυνεπείς, και συνεπώς ο έλεγχος υποθέσεων δεν είναι πια έγκυρος. Στις περισσότερες περιπτώσεις, το R^2 θα είναι υπερεκτιμημένο και τα t-στατιστικά θα τείνουν να είναι υψηλότερα.

4.8.7. Έλεγχος της Έλλειψης Προσαρμογής των Στοιχείων (Lack-of-fit test).

Εφαρμόζεται αποκλειστικά σε μοντέλα πολυωνυμικών εξισώσεων (Ενότητα 4.12) στα οποία σε καθεμία τιμή της X αντιστοιχούν δύο ή και περισσότερες επαναληπτικές τιμές της Y (Burn & Ryan, 1983). Στις περισσότερες περιπτώσεις, η εξίσωση της γραμμικής παλινδρόμησης $X - επαναληπτικές\ Y$ προσαρμόζεται στα στοιχεία, χωρίς να είναι γνωστή από πριν η έκβαση της σχέσης μεταξύ των X και Y, όπως π.χ. θα έδειχνε άμεσα η γραφική απεικόνισή τους. Επομένως, μία επιβεβλημένη ενέργεια είναι ο παράλληλος έλεγχος μίας εξίσωσης που μόλις δημιουργήθηκε, αν πραγματικά περιγράφει σωστά τη σχέση και να αποφεύγονται έτσι τραγικά λάθη, όπως αυτό του Σχήματος 4.7 στο οποίο εφαρμόστηκε η εξίσωση της γραμμικής μορφής αντί της

ορθής πολυωνυμικής μορφής. Τα λάθη αυτά αποφεύγονται με τον έλεγχο της έλλειψης προσαρμογής ο οποίος εξετάζει αν η υπολογισθείσα εξίσωση στα εξεταζόμενα στοιχεία ταιριάζει στατιστικά (τα περιγράφει επαρκώς). Ο έλεγχος αυτός υπολογίζει τις ποσότητες $SS(1)$ και $SS(2)$, δηλαδή τα αθροίσματα των τετραγώνων των τιμών που προέρχονται αντίστοιχα από το **πειραματικό σφάλμα** και από την **έλλειψη προσαρμογής** των στοιχείων.

Το πειραματικό σφάλμα προϋποθέτει την παρουσία επαναληπτικών μετρήσεων της Y για κάθε τιμή της X και υπολογίζεται ως,

$$SS(1) = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2$$

και το σφάλμα της έλλειψης προσαρμογής υπολογίζεται ως

$$SS(2) = \sum_{i=1}^{k}(\bar{Y}_i - \hat{Y}_i)^2$$

όπου k είναι ο αριθμός των ομάδων τιμών της Y που αντιστοιχούν συνολικά σε όλες τις τιμές της X με j παρατηρήσεις ανά ομάδα.

Το ***στατιστικό κριτήριο F*** του ελέγχου προσδιορίζεται από τη διαίρεση των μέσων αθροισμάτων των τετραγώνων $MS(2)$ και $MS(1)$,

$$F = \frac{MS(2)}{MS(1)} = \frac{SS(2)/(N-k)}{SS(1)/(k-1)} \qquad (4.25)$$

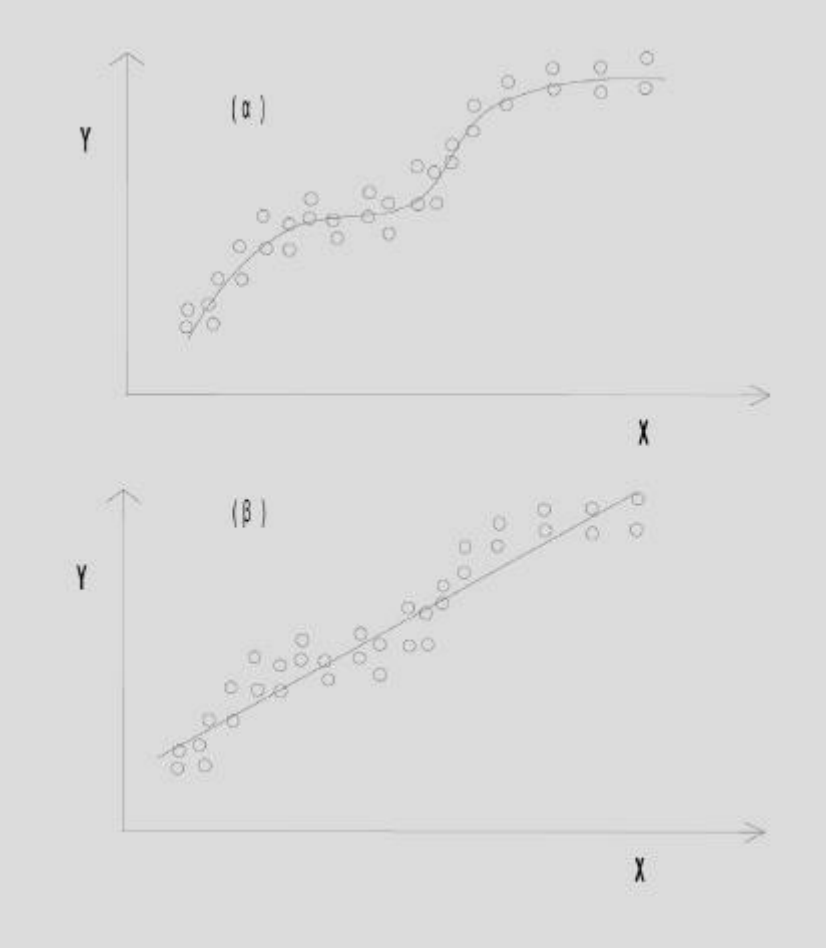
Η τιμή του $\boldsymbol{ελέγχου}\ F$ συγκρίνεται με την οριακή $F_{0.05,(k-2),(N-k)}$ από τον Πίνακα Π1 του Παραρτήματος και αν είναι μεγαλύτερη της οριακής, τότε η γραμμική εξίσωση δεν περιγράφει επαρκώς τη σχέση $X - Y$ και θα πρέπει να αναζητηθούν νέοι τρόποι βελτίωσης της γραμμικής σχέσης, π.χ. μετασχηματισμοί.

4.8.8. Εξέταση του Συντελεστή Πρόβλεψης (Predicted Coefficient) R_p^2 της Παλινδρόμησης (Montgomery et al,2012).

Το προβλεπτικό R_p^2 $(R^2 - predicted)$ υπολογίζεται αφαιρώντας συστηματικά κάθε παρατήρηση από τα στοιχεία (μία τη φορά), εκτιμώντας στη συνέχεια τη νέα εξίσωση παλινδρόμησης και η τιμή του έτσι προσδιορίζει πόσο ικανοποιητικά το μοντέλο

προβλέπει την αφαιρούμενη παρατήρηση. Το προβλεπτικό R_p^2 αποτρέπει την υπερπροσαρμογή των μοντέλων και θεωρείται χρησιμότερο για τη σύγκριση των μοντέλων επειδή υπολογίζεται με τη χρήση νέων παρατηρήσεων που δεν περιλαμβάνονται στην εκτίμηση των μοντέλων. Ο όρος υπερπροσαρμογή αναφέρεται στα μοντέλα εκείνα που εξηγούν τη σχέση που αναπτύσσεται μεταξύ των ανεξάρτητων και της εξαρτημένης μεταβλητής με τα ήδη υπάρχοντα στοιχεία, αποτυγχάνουν όμως να παρέχουν προβλέψεις για νέες παρατηρήσεις. Στην πράξη, το προβλεπτικό R_p^2 δηλώνει πόσο ικανοποιητικά το μοντέλο της παλινδρόμησης προβλέπει τις αποκρίσεις όταν εισάγονται νέες παρατηρήσεις στην εξίσωση σε αντίθεση με το προσδιοριστικό R^2 που δείχνει πόσο καλά το μοντέλο προσαρμόζεται στα υπάρχοντα στοιχεία. Οι τιμές R_p^2 κυμαίνονται μεταξύ 0 και 100%, είναι μικρότερες των αντίστοιχων τιμών R^2 και υπολογίζονται από το **κριτήριο $\boldsymbol{PRESS}$**.

Υψηλές τιμές του R_p^2 υποδηλώνουν μοντέλα με μεγάλη προβλεπτική αξία. Για παράδειγμα έστω ένα μοντέλο με συντελεστή προσδιορισμού $R^2 = 87\%$ και πρόβλεψης $R_p^2 = 52\%$. Η διαφορά αυτή μπορεί να δηλώνει υπερπροσαρμογή του μοντέλου και υπαινίσσεται ότι το μοντέλο δεν θα προβλέπει ικανοποιητικά νέες παρατηρήσεις, δεν θα προβλέπει δηλαδή τιμές παραπλήσιες με αυτές που προκύπτουν από τα υπάρχοντα στοιχεία. Κατά κανόνα, *όταν μεταξύ των δύο συντελεστών η διαφορά υπερβαίνει το 20% τότε θεωρούμε ότι το μοντέλο έχει χαμηλή προβλεπτική αξία.*



Σχήμα 4.7 Προσαρμογή των στοιχείων της σχέσης X-δύο επαναληπτικές τιμές της Y α) σε εξίσωση πολυωνυμικής μορφής (τεταρτοβάθμια εξίσωση), β) λανθασμένα σε εξίσωση γραμμικής μορφής.

Το κριτήριο PRESS είναι το προβλεπτικό άθροισμα των τετραγώνων των υπολειμμάτων και αποτελεί στατιστικό κριτήριο εγκυρότητας που δεν επηρεάζεται από τις τιμές (παρατηρήσεις) του δείγματος.

Το κριτήριο στηρίζεται στην αφαίρεση μίας τιμής κάθε φορά από το δείγμα και την εκτίμηση του υπολείμματος: $Y_i - \hat{Y}_i$. Συγκεκριμένα, μετά την αφαίρεση της τιμής i εφαρμόζεται η παλινδρόμηση σε $n-1$ στοιχεία και ακολούθως υπολογίζεται η προσαρμοσμένη τιμή $\hat{Y}_i$ της αφαιρεθείσας παρατήρησης (εισαγόμενης ως νέα παρατήρηση), την οποία συγκρίνουμε με την τιμή Y_i του στοιχείου από την προηγούμενη εξίσωση των n παρατηρήσεων. Η διαδικασία αυτή επαναλαμβάνεται n φορές όσα και τα στοιχεία της ανάλυσης και το $κριτήριο\ PRESS$ υπολογίζεται ως,

$$PRESS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (4.26)$$

Στην πολλαπλή παλινδρόμηση υπολογίζονται πολλά *κριτήρια PRESS*, όσα και οι εισαγόμενες (ή απορριπτόμενες), μία τη φορά και επιλέγεται εκείνο με τη μικρότερη τιμή. Εναλλακτικά, χρησιμοποιείται η εξίσωση,

$$PRESS = \sum_{i=1}^{n}\left(\frac{\hat{e}_i}{1-h_i}\right)^2$$

όπου ο αριθμητής εκφράζει το i υπόλειμμα και h_i είναι ο συντελεστής επιρροής.

Το προβλεπτικό R_p^2 υπολογίζεται ως,

$$R_p^2 = \frac{PRESS}{1-TSS} = \frac{\sum_{i=1}^{n}\left(\frac{\hat{e}_i}{1-h_i}\right)^2}{1-\sum_{i=1}^{n}(Y_i-\bar{Y}_i)^2} = \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{1-\sum_{i=1}^{n}(Y_i-\bar{Y}_i)^2} \qquad (4.27)$$

4.8.9. Εξέταση της Πολυσυγγραμμικότητας (Multicollinearity)

Η multicollinearity είναι γνωστή και με τα ονόματα ***ενδοσυσχέτιση και συγγραμμικότητα μεταξύ των ανεξάρτητων μεταβλητών***. Αν κάποιες από αυτές συσχετίζονται μεταξύ τους, τότε ενδέχεται οι μερικοί συντελεστές της παλινδρόμησης που αφορούν αυτές τις μεταβλητές να μην αντανακλούν επακριβώς την εξάρτηση της Y από τις συσχετιζόμενες. Η ανίχνευση της έντασης της πολυσυγγραμικότητας πραγματοποιείται, παλινδρομώντας κάθε μεταβλητή X_i που είναι εκτός εξίσωσης (ως Y) με όλες εκείνες που έχουν ήδη εισαχθεί στο μοντέλο.

Η τιμή, $\mathbf{1 - R_i^2}$ που υπολογίζεται για κάθε περίπτωση και καλείται ***ανοχή της πολυσυγραμμικότητας***, *αν είναι* $< 0{,}10$, *τότε υπάρχει σημαντική ενδοσυσχέτιση* μεταξύ της παλινδρομούμενης και κάποιας/ων από τις υπόλοιπες ενταγμένες μεταβλητές. *Αν η ανοχή είναι* $< 0{,}05$, *τότε επιβάλλεται η οριστική απομάκρυνση της παλινδρομηθείσας από την εξίσωση της πολλαπλής παλινδρόμησης.*

Ο ζημιογόνος ρόλος της πολυσυγγραμικότητας εντοπίζεται στα τυπικά σφάλματα των μερικών συντελεστών παλινδρόμησης τα οποία εμφανίζονται μεγάλα, που σημαίνει, ότι οι μερικοί συντελεστές είναι ανακριβείς εκτιμητές της πολλαπλής παλινδρόμησης. Ένας γρήγορος τρόπος ανίχνευσης των ύποπτων μεταβλητών είναι η δημιουργία του πίνακα (μήτρας) των πολλαπλών συσχετίσεων. Από τον πίνακα αυτόν

εύκολα διαπιστώνουμε τις μεταβλητές με υψηλή συσχέτιση, τις οποίες και ακολούθως ελέγχουμε για πολυσυγγραμμικότητα, παλινδρομώντας μία μεταβλητή κάθε φορά με τις υπόλοιπες.

Εμπειρικός κανόνας υποστηρίζει ότι εάν η τιμή του συντελεστή r *είναι πολύ μεγάλη* $(r > 0{,}90)$ *ή για αρκετές τιμές* r *ισχύει* $r > 0{,}70$ *στη μήτρα των συντελεστών συσχέτισης των ανεξάρτητων μεταβλητών, υπάρχει σοβαρή υπόνοια περί πολυσυγγραμμικότητας.*

Στις πολυωνυμικές εξισώσεις, οποιασδήποτε τάξης, παρατηρείται πάντοτε σημαντική πολυσυγγραμμικότητα μεταξύ των διάφορων όρων της μεταβλητής X. Στις περιπτώσεις αυτές, η απομάκρυνση κάποιων όρων της μεταβλητής X δεν θεωρείται αναγκαία και αφήνεται συνήθως στην κρίση του ερευνητή, ανάλογα με τη σκοπιμότητα του πειράματος.

4.8.10. Εξέταση του Συντελεστή Διογκωμένης Διακύμανσης της Παλινδρόμησης VIF (Variance Inflation Factor) για Κάθε Ανεξάρτητη Μεταβλητή

Ο συντελεστής διόγκωσης προκύπτει παλινδρομώντας, όπως και προηγούμενα, κάθε ανεξάρτητη μεταβλητή με τις ήδη ενταχθείσες και ισούται με το αντίστροφο της ανοχής,

$$VIF = \frac{1}{1-R_i^2} \qquad (4.28)$$

Όταν αυξάνει η τιμή VIF*, αυξάνει και η διακύμανση του εκάστοτε συντελεστή παλινδρόμησης άρα και η τυπική του απόκλιση, με αποτέλεσμα να μειώνεται η σημαντικότητά του.*

Γενικά, τιμές του συντελεστή $VIF > 10$ (ή κατ` άλλους > 20), δηλώνουν ότι η εξεταζόμενη ανεξάρτητη μεταβλητή (ως Y) συσχετίζεται ***ισχυρά*** *με μία τουλάχιστον από τις ήδη ενταχθείσες και συνιστά την άμεση απομάκρυνσή της από το μοντέλο της παλινδρόμησης*. Τιμές του συντελεστή διόγκωσης ίσες με 1 δηλώνουν ότι υπάρχει πλήρης έλλειψη συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών.

Γενικά τιμές του $VIF > 5$ θεωρούνται ως ένδειξη πολυσυγραμμικότητας και για να μειωθεί το φαινόμενο αφαιρούμε από το μοντέλο τους μη σημαντικούς παράγοντες με τις μεθόδους:

- Durbin-Watson test
- Casewise Diagnostics: Με την πραγματοποίηση διαγνωστικών ελέγχων με βάση το κριτήριο επιλογής, όπως των ακραίων τιμών (τιμών μεγαλύτερων από προκαθορισμένο μέγεθος που προκύπτει από την προσθαφαίρεση των τυπικών αποκλίσεων). Στόχος είναι να εξεταστούν πιθανές ακραίες τιμές (outliers) και παρατηρήσεις υψηλής επίδρασης (influential-points). Όπως προαναφέραμε στην ενότητα 4.8.4. μια ακραία τιμή (outliers) είναι μια ασυνήθιστη τιμή που δεν συμφωνεί με το pattern των υπόλοιπων δεδομένων ενώ μια παρατήρηση υψηλής επίδρασης μπορεί να είναι μια ακραία τιμή, αλλά, επίσης, μπορεί να είναι μια τιμή των δεδομένων που να έχει μεγάλη συνεισφορά στο σχηματισμό της ευθείας παλινδρόμησης. Το μέτρο που χρησιμοποιείται για την εύρεση της επιρροής μιας παρατήρησης στο μοντέλο είναι η Cook's distance (Ενότητα 4.8.5.)

4.8.11. Έξέταση του Συντελεστή Μεταβλητότητας Κάθε Ανεξάρτητης Μεταβλητής.

Όταν μία μεταβλητή έχει χαμηλό συντελεστή μεταβλητότητας (CV_το πηλίκο της τυπικής της απόκλισης δια του μέσου όρου), τότε αυτή παρουσιάζεται στην εξίσωση σχεδόν σταθερή. Πρακτικά, αυτό σημαίνει ότι οι τιμές των στοιχείων είναι πολύ κοντινές μεταξύ τους. Η διαμόρφωση αυτή των στοιχείων της μεταβλητής μπορεί να προκαλέσει σημαντικά υπολογιστικά προβλήματα ανακρίβειας, διορθώνεται όμως εύκολα, υψώνοντας τη μεταβλητή στο τετράγωνο «αραιώνοντας» έτσι τις τιμές γύρω από το μέσο όρο.

4.8.12. Το Δειγματοληπτικό Μέγεθος

Το δειγματοληπτικό μέγεθος επηρεάζει την εγκυρότητα του τελικού υποδείγματος διότι όταν αυξάνει αμβλύνει την επιρροή των εξωκείμενων τιμών και οφείλει να περιέχει τουλάχιστον περισσότερες από πέντε παρατηρήσεις. Ένας εμπειρικός

κανόνας ορίζει την παρουσία στο τελικό μοντέλο τουλάχιστον $10-15$ παρατηρήσεων ανά ανεξάρτητη μεταβλητή ή 50 παρατηρήσεις κατά βάση με 8 επιπρόσθετες για κάθε νέα εισαγόμενη μεταβλητή τη φορά στην εξίσωση. Όταν εφαρμόζεται η βηματική παλινδρόμηση, αναμένεται ο αριθμός των παρατηρήσεων να είναι 40πλάσιος του αριθμού των ανεξάρτητων μεταβλητών.

4.8.13. Ανάλυση Χρονοσειρών και τα Κριτήρια MAD, MSE, MAPE και MPE

Η ανάλυση χρονοσειρών (time series analysis) ασχολείται αποκλειστικά με τη διερεύνηση της διαχρονικής συμπεριφοράς των τιμών μιας μεταβλητής, οι παρατηρήσεις της οποίας προέρχονται από χρονοσειρά. Η πρόβλεψη των μελλοντικών τιμών της μεταβλητής σύμφωνα με την ανάλυση χρονοσειρών μπορεί να προέλθει από τις παρακάτω κατηγορίες μεθόδων προβλέψεων που θα αναφέρουμε ονομαστικά:

- Μέθοδοι Εξομάλυνσης
- Διάσπαση χρονοσειρών
- Ανάλυση ARIMA

Για την επιλογή της κατάλληλης μεθόδου χρησιμοποιούνται τα κριτήρια αξιολόγησης των μεθόδων προβλέψεων. Τα κριτήρια αυτά βασίζονται στις τιμές των αποκλίσεων των προβλεπόμενων τιμών από τις αντίστοιχες πραγματικές τιμές της χρονοσειράς. Για μία μεταβλητή Y, η απόκλιση της προβλεπόμενης τιμής της $\hat{Y}_t$ από την αντίστοιχη πραγματική τιμή της Y_t για την περίοδο t, όπου $t = 1,2,3,..,n$, ονομάζεται ***σφάλμα της πρόβλεψης (forecast error)***, συμβολίζεται με $\boldsymbol{e_t}$ και ορίζεται ως: $e_t = Y_t - \hat{Y}_t$. Επομένως, για να προσδιορίσουμε την αξιοπιστία μιας συγκεκριμένης μεθόδου πρόβλεψης, θα πρέπει να μελετήσουμε τη διαχρονική συμπεριφορά των τιμών των σφαλμάτων της πρόβλεψης. Αυτό γίνεται με την εφαρμογή διάφορων κριτηρίων, σύμφωνα με τα οποία αξιολογούμε τη χρησιμοποιούμενη μέθοδο πρόβλεψης. Κάθε ένα από τα κριτήρια αυτά ορίζεται από μία συγκεκριμένη συναρτησιακή σχέση των σφαλμάτων της πρόβλεψης και μπορεί να χρησιμοποιηθεί όχι μόνο για την αξιολόγηση μιας μεθόδου πρόβλεψης αλλά και για την επιλογή της " καλύτερης " μεταξύ δύο ή περισσοτέρων εναλλακτικών μεθόδων προβλέψεων. Τα κριτήρια αυτά είναι:

i. Μέση απόλυτη απόκλιση MAD (Mean Absolute Deviation)

Η μέση απόλυτη απόκλιση ορίζεται ως το άθροισμα των απόλυτων τιμών του σφάλματος της πρόβλεψης διαιρούμενο με τον αριθμό των περιόδων n, στις οποίες έγιναν προβλέψεις, δηλαδή:

$$MAD = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right| = \frac{1}{n}\sum_{t=1}^{n}|e_t| \qquad (4.29)$$

Το MAD εκφράζει τη μέση τιμή των απόλυτων αποκλίσεων των προβλεπόμενων τιμών της χρονοσειράς από τις αντίστοιχες πραγματικές και έχει τα ακόλουθα χαρακτηριστικά.

- Η μονάδα μέτρησης του είναι η ίδια με εκείνη των τιμών της χρονοσειράς και έτσι είναι εύκολη η ερμηνεία του.
- Στον υπολογισμό του λαμβάνονται υπ' όψιν μόνο οι απόλυτες τιμές των σφαλμάτων και όχι οι πραγματικές τιμές τους. Αυτό σημαίνει ότι το MAD είναι ανεξάρτητο από θετικές ή αρνητικές τιμές του σφάλματος, δηλαδή είναι ανεξάρτητο από το αν οι τιμές των προβλέψεων είναι μικρότερες (υποεκτίμηση) ή μεγαλύτερες (υπερεκτίμηση) των πραγματικών τιμών.
- Το MAD βασίζεται στην υπόθεση ότι η σοβαρότητα του σφάλματος ή το κόστος που δημιουργείται από το σφάλμα της πρόβλεψης σχετίζεται γραμμικά με το μέγεθος του σφάλματος.

ii. Μέσο σφάλμα τετραγώνου MSE (Mean Squared Error)

Το μέσο σφάλμα τετραγώνου ορίζεται ως το άθροισμα των τετραγώνων των σφαλμάτων διαιρούμενο με τον αριθμό των χρονικών περιόδων n, στις οποίες έγιναν προβλέψεις, δηλαδή:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2 = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \qquad (4.30)$$

Το MSE είναι η μέση τιμή των τετραγώνων των αποκλίσεων των προβλεπόμενων τιμών της χρονοσειράς από τις αντίστοιχες πραγματικές. Η μονάδα μέτρησης του MSE όμως είναι εκφρασμένη στη μονάδα μέτρησης των τιμών των παρατηρήσεων υψωμένη όμως στο τετράγωνο. Για το λόγο αυτό, μερικές φορές χρησιμοποιούμε τη θετική τιμή της τετραγωνικής του ρίζας, που ονομάζεται τετραγωνική ρίζα μέσου σφάλματος τετραγώνου *RMSE (Root Mean Squared Error)* δηλαδή είναι:

$$RMSE = \sqrt{MSE}$$

Το RMSE εκφράζεται στην ίδια μονάδα μέτρησης με εκείνη των τιμών της χρονοσειράς. Η ύπαρξη προβλέψεων που απέχουν πολύ από τις αντίστοιχες πραγματικές τιμές γίνεται πολύ περισσότερο αισθητή με το κριτήριο MSE από ότι με το κριτήριο MAD, επειδή οι τιμές των σφαλμάτων της πρόβλεψης υψώνονται στο τετράγωνο. Συνεπώς ***το κριτήριο MSE είναι στατιστικά περισσότερο αξιόπιστο από το κριτήριο MAD και χρησιμοποιείται συχνότερα για την επιλογή της 'κατάλληλης' μεθόδου πρόβλεψης.***

iii. Μέσο απόλυτο ποσοστιαίο σφάλμα MAPE (Mean Absolute Percentage Error)

Το μέσο απόλυτο ποσοστιαίο σφάλμα εξετάζει τη συμπεριφορά της απόλυτης τιμής του σφάλματος της πρόβλεψης σε σχέση με την πραγματική τιμή της χρονοσειράς. Το MAPE ορίζεται ως το άθροισμα των απόλυτων τιμών των σφαλμάτων της πρόβλεψης προς τις αντίστοιχες πραγματικές τιμές της χρονοσειράς διαιρούμενο με τον αριθμό των χρονικών περιόδων n, στις οποίες έγιναν προβλέψεις, δηλαδή:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \hat{Y}_t|}{Y_t} = \frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{Y_t} \qquad (4.31)$$

Το κριτήριο αυτό είναι απαλλαγμένο από μονάδες μέτρησης και το χρησιμοποιούμε για να συγκρίνουμε την ακρίβεια μιας ή περισσοτέρων μεθόδων προβλέψεων και για περισσότερες από μια χρονοσειρές.

iv. Μέσο ποσοστιαίο σφάλμα MPE (Mean Percentage Error)

Το μέσο ποσοστιαίο σφάλμα το χρησιμοποιούμε όταν ενδιαφερόμαστε να προσδιορίσουμε αν η μέθοδος πρόβλεψης είναι μεροληπτική, δηλαδή αν οι προβλεπόμενες τιμές είναι συστηματικά μεγαλύτερες ή μικρότερες από τις αντίστοιχες πραγματικές.

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{Y_t - \hat{Y}_t}{Y_t} = \frac{1}{n}\sum_{t=1}^{n}\frac{e_t}{Y_t} \qquad (4.32)$$

Αναμφισβήτητα όσο πιο κοντά στο μηδέν είναι η τιμή του MPE, τόσο πιο αμερόληπτη και καλή είναι η μέθοδος πρόβλεψης που χρησιμοποιήθηκε. Αντίθετα, μεγάλες απόλυτες τιμές του MPE φανερώνουν μεγάλη μεροληψία της μεθόδου.

4.9 Εικονικές Μεταβλητές στην Πολλαπλή Παλινδρόμηση

Είναι συχνή τακτική στο θεσμό της παλινδρόμησης, είτε απλής είτε πολλαπλής, να λαμβάνονται αναπόφευκτα υπόψη και μεταβλητές που είναι ονομαστικές (κατηγορικές). Το φύλο των δοκιμαστών, τα είδη της διαφήμισης, οι ποικιλίες του ίδιου του προϊόντος που διατείθεται κτλ., είναι ονομαστικές μεταβλητές με δύο ή περισσότερες κατηγορίες και με γνωστό επιστημονικό υπόβαθρο γνώσης για τη σημασία τους. Σε περίπτωση που ληφθούν υπόψη, μπορεί να επηρεάσουν σημαντικά την έκβαση της εξαρτημένης μεταβλητής ενός μοντέλου παλινδρόμησης. Στις περιπτώσεις αυτές, η συνήθης τακτική περιλαμβάνει χωριστές εξισώσεις παλινδρόμησης, π.χ. δύο, αν αναφερόμαστε στο φύλο των δοκιμαστών, οπότε οι γυναίκες διαχωρίζονται από τους άνδρες και εκτελούμε έτσι δύο εξισώσεις παλινδρόμησης με τις ίδιες μεταβλητές που διαφέρουν μόνο ως προς το φύλο. Κάθε μία εξίσωση περιλαμβάνει τόσες παρατηρήσεις όσες και τα διαχωριζόμενα άτομα του κάθε φύλου. Αν οι κατηγορίες ενός καλλυντικού προϊόντος ήταν τρεις, τρεις θα ήταν και οι εξεταζόμενες παλινδρομήσεις. Ακολούθως, γίνονται συγκρίσεις μεταξύ των εξισώσεων παλινδρόμησης, για να βρεθούν τυχόν διαφορές μεταξύ των μεταβλητών ως προς τις δύο κατηγορίες του φύλου ή τις τρεις κατηγορίες και γενικά ως προς τις κατηγορίες-περιπτώσεις κάθε εξεταζόμενης ονομαστικής μεταβλητής.

Μία εναλλακτική και πολύ δημοφιλής λύση είναι η απευθείας ένταξη μίας ή και περισσότερων ονομαστικών μεταβλητών στο μοντέλο της παλινδρόμησης. Επειδή όμως η ανάλυση της παλινδρόμησης αφορά μόνο την ανάλυση στοιχείων προερχόμενων από συνεχείς μεταβλητές, θα πρέπει οι ονομαστικές μεταβλητές να τροποποιηθούν κατά τέτοιο τρόπο, ώστε να εμφανίζονται ως συνεχείς. *Αυτό επιτυγχάνεται με την ενσωμάτωση εικονικών μεταβλητών ή ψευδομεταβλητών (dummy variables) οι οποίες μετατρέπουν κάθε κατηγορία των κατηγορικών μεταβλητών σε αυτόνομη μεταβλητή με τη χρησιμοποίηση των κωδικών αριθμών* 0 *(απουσία ενδεχομένου) και* 1 *(παρουσία ενδεχομένου)*, π.χ. 0 για τις γυναίκες και 1 για τους άντρες όπως φαίνεται στον ακόλουθο Πίνακα 4.4 .

Α/Α	Φύλο	Επεξεργασία			Ποικιλία	
1	0	0	1	0	0	0
2	1	0	0	0	1	0
3	0	1	0	0	1	0
4	1	0	1	0	0	1
5	0	1	0	0	0	0
6	1	0	0	1	1	0
7	0	0	1	0	0	1
8	0	0	0	0	0	0
9	1	0	0	1	1	0

Πίνακας 4.4. Εικονικές μεταβλητές του φύλου (δύο κατηγορίες), επεξεργασίας υλικού (τέσσερις κατηγορίες) και ποικιλίας οίνου (τρεις κατηγορίες) με 9 παρατηρήσεις η καθεμία.

Έτσι, η εικονική μεταβλητή του φύλου θα περιλαμβάνει τιμές 0 και 1, όμως για μεταβλητές με περισσότερες από δύο κατηγορίες θα απαιτηθούν $k-1$ εικονικές μεταβλητές. Για παράδειγμα, η ονομαστική μεταβλητή της ποικιλίας με $k=3$ κατηγορίες, θα αποτελείται από δύο $(3-1=2)$ εικονικές μεταβλητές οργανωμένες έτσι ώστε, η καθεμιά σειρά τους να ανταποκρίνεται σε τρεις περιπτώσεις κατηγορίας ποικιλιών

Όπως φαίνεται στον Πίνακα 4.4: 0 και 0 (ποικιλία 1), 0 και 1 (ποικιλία 2) και 1 και 0 (ποικιλία 3). Παρόμοια, η μεταβλητή της επεξεργασίας με 4 κατηγορίες θα αποτελέσει 3 εικονικές μεταβλητές, για τις οποίες θα ισχύει για κάθε σειρά: 0,0,0 (κατηγορία 1), 0,1,0 (κατηγορία 2) 0,0,1 (κατηγορία 3) και 1,0,0 (κατηγορία 4).

Ο τρόπος αυτός κωδικοποίησης των κατηγορικών μεταβλητών αποσκοπεί αποκλειστικά και μόνο στη χρησιμοποίησή τους στα στατιστικά προγράμματα. Η ένταξη των εικονικών μεταβλητών στην εξίσωση της παλινδρόμησης, μας επιτρέπει να διαπιστώσουμε τη σημαντικότητα ή μη της επίδρασης των κατηγορικών μεταβλητών στην εξαρτημένη Y. Η ερμηνεία των συντελεστών παλινδρόμησης των ψευδομεταβλητών γίνεται σε σχέση με την παραλειπόμενη κατηγορία αναφοράς (η οποία πρέπει να ορίζεται με σαφήνεια) της ψευδομεταβλητής και αφορά το βαθμό της

μεταβολής της εξαρτημένης μεταβλητής με τη μεταβολή της ψευδομεταβλητής κατά μία μονάδα (π.χ με τη μετακίνηση από το 0 στο 1).

Πληρέστερα, για οποιαδήποτε μεταβολή κάποιας ανεξάρτητης μεταβλητής, το πρόσημο και το μέγεθος του συντελεστή παλινδρόμησης υποδηλώνει την κατεύθυνση (θετική ή αρνητική) και την ένταση του βαθμού μεταβολής της εξαρτημένης μεταβλητής. Επισημαίνεται ότι οι συντελεστές παλινδρόμησης των ψευδομεταβλητών αφορούν όλη την ομάδα των κατηγοριών της μητρικής κατηγορικής μεταβλητής (σε αντίθεση με τους ελέγχους $t-Student$ των συντελεστών b για τις συνεχείς μεταβλητές) και ο $\acute{\varepsilon}\lambda\varepsilon\gamma\chi o\varsigma\ F$ αφορά τη διαφορά των τιμών R^2 με την παρουσία και χωρίς την παρουσία ψευδομεταβλητών στο υπόδειγμα. Αν η επίδραση τεκμηριώνεται στατιστικά, τότε η έκβαση της μεταβλητής Y θα εξηγείται με μεγαλύτερη ακρίβεια από το σύνολο των μεταβλητών, ανεξάρτητων και κατηγορικών. Αρκετά συχνά γίνεται χρήση της αλληλεπίδρασης των όρων, συνήθως ανά δύο, επί της εξαρτημένης, αν και ο αλληλεπιδρών όρος μπορεί να συσχετίζεται σε μεγάλο βαθμό με αντίστοιχη απλή ανεξάρτητη μεταβλητή (πολυσυγγραμμικότητα), δημιουργώντας προβλήματα ως προς την εκτίμηση της επίδρασης των ανεξάρτητων μεταβλητών επί της εξαρτημένης.

4.10 Σύγκριση Πολλαπλών Παλινδρομήσεων

Η σύγκριση πολλαπλών παλιδρομήσεων εφαρμόζεται σε περιπτώσεις που δύο ή περισσότερες εξισώσεις παλινδρόμησης περιέχουν τις ίδιες μεταβλητές και διαφέρουν μεταξύ τους ως προς ένα χαρακτηριστικό π.χ. φύλο δοκιμαστών, διαφορετική επεξεργασία. Η μηδενική υπόθεση που εξετάζεται είναι ότι, όλες οι k εξισώσεις εκτιμούν τον ίδιο πληθυσμό (δεν διαφέρουν μεταξύ τους) ή αλλιώς ότι υπάρχει πλήρης σύμπτωση των κλίσεων των μερικών συντελεστών και των παραμέτρων a των παλινδρομήσεων και ελέγχεται με το $\kappa\rho\iota\tau\acute{\eta}\rho\iota o\ F$:

$$F = \frac{\frac{ESS_t - ESS_p}{(m+1)(k-1)}}{\frac{ESS_p}{N-k(m+1)}} \qquad (4.33)$$

Η ποσότητα $\boldsymbol{ESS_p}$ είναι *το άθροισμα των τετραγώνων όλων των υπολειμμάτων των* k *παλινδρομήσεων* με:

βαθμούς ελευθερίας $N-k(m+1)$, όπου $N=n_1+n_2+\cdots+n_k$ το άθροισμα όλων των παρατηρήσεων και m είναι το σύνολο των μεταβλητών ανά εξίσωση παλινδρόμησης.

Η ποσότητα $\boldsymbol{ESS_t}$ είναι το *άθροισμα των τετραγώνων των υπολειμμάτων που προκύπτει αν ενσωματώσουμε όλες τις μεταβλητές X_1 των διαφορετικών εξισώσεων σε μία νέα X_1, όλες τις μεταβλητές X_2 σε μία νέα X_2 κοκ. και εκτελέσουμε μία μόνο πολλαπλή παλινδρόμηση.*

Αν η τιμή F είναι μεγαλύτερη ή ίση της οριακής $F_{0.05,[(m+1)(k-1)][N-k(m+1)]}$, (Πίνακας Π1 Παράρτημα) $F \geq F_{o\rho}$, τότε οι μεταβλητές X_i δεν προσδιορίζουν την ίδια εξίσωση παλινδρόμησης.

Μία δεύτερη υπόθεση που ελέγχεται στην πολλαπλή παλινδρόμηση, είναι η έννοια της παραλληλίας. Στην απλή γραμμική παλινδρόμηση ο έλεγχος περιορίζεται στη σύγκριση των κλίσεων b_i , k εξισώσεων, αν αυτές δίνουν ευθείες παράλληλες ή όχι. Σε μία πολλαπλή παλινδρόμηση με δύο ανεξάρτητες μεταβλητές η ευθεία γραμμή αντικαθίσταται με ένα επίπεδο και ο έλεγχος προσαρμόζεται στη σύγκριση k επιπέδων k εξισώσεων πολλαπλής παλινδρόμησης. Δύο ή περισσότερα επίπεδα είναι παράλληλα μεταξύ τους, όταν οι μερικοί συντελεστές κάθε παλινδρόμησης είναι ίσοι με τους αντίστοιχους των υπόλοιπων εξισώσεων. Για συγκρίσεις παλινδρομήσεων με περισσότερες μεταβλητές από δύο, ο έλεγχος αφορά τη σύγκριση των k υπερεπιπέδων k εξισώσεων πολλαπλής παλινδρόμησης με $m>2$. Η μηδενική υπόθεση που εξετάζεται είναι ότι, δύο ή περισσότερες πολλαπλές παλινδρομήσεις είναι παράλληλες, όταν έχουν τους ίδιους συντελεστές $\beta_1, \beta_2, \dots, \beta_\kappa$ και ελέγχεται με το $\kappa\rho\iota\tau\acute{\eta}\rho\iota o\ F$:

$$F=\frac{\frac{ESS_c-EES_p}{k-1}}{\frac{ESS_p}{N-k(m+1)}} \qquad (4.34)$$

Η ποσότητα $\boldsymbol{ESS_c}$ είναι *το άθροισμα των τετραγώνων των υπολειμμάτων της κοινής παλινδρόμησης που υπολογίζεται, όταν στους k πίνακες του αθροίσματος των τετραγώνων των χιαστί γινομένων, (όσες δηλαδή και οι συγκρινόμενες παλινδρομήσεις) προστεθούν όλα τα ομοειδή στοιχεία των πινάκων ($\sum x^2, \sum y^2, \sum x_1x_2$ κτλ.) και προκύψει έτσι ένας κοινός πίνακας.* Η τιμή F συγκρίνεται με τη θεωρητική

$F_{0.05,(k-1)[N-k(m+1)]}$ και αν $F \geq F_{o\rho}$, τότε τα υπερεπίπεδα δεν είναι παράλληλα μεταξύ τους, γιατί κάποιοι μερικοί συντελεστές διαφέρουν μεταξύ τους.

Οταν ισχύει η παραλληλία των υπερεπιπέδων, τότε μπορούμε να ελέγξουμε και την ισότητα των παραμέτρων a με το *κριτήριο* F:

$$F = \frac{\frac{ESS_t - ESS_C}{(k-1)}}{\frac{ESS_C}{N-m-1}} \qquad (4.35)$$

Η τιμή F συγκρίνεται με την οριακή $F_{0.05,(k-1)(N-m-1)}$ και αν $F \geq F_{o\rho}$, τότε ισχύει η εναλλακτική υπόθεση που δηλώνει ότι, οι παράμετροι a δεν είναι όλες ίσες μεταξύ τους.

4.11 Κίνδυνοι από την Αλόγιστη Χρήση των Εξισώσεων Παλινδρόμησης

Θεωρητικά, τα μοντέλα παλινδρόμησης εξηγούν μία επιστημονική, οικονομική εξάρτηση δύο ή περισσοτέρων μεταβλητών και η οποία επιβεβαιώνει τη μαθηματική σχέση τους που περιγράφεται από την εξίσωση της παλινδρόμησης. Για να γίνει κατανοητή μία τέτοια περιγραφή ενός φαινομένου, και μάλιστα άγνωστου ή μη εμφανούς, ο ερευνητής θα πρέπει να γνωρίζει εμπεριστατωμένα για τη σχέση των μεταβλητών που περιγράφονται στο μοντέλο.

- Το πρώτο και δυσκολότερο ερώτημα που τίθεται είναι, αν αυτή η εξάρτηση περιγράφεται πραγματικά από μία γραμμική απλή ή και πολλαπλή σχέση ή από μία μη γραμμική σχέση που δεν τέθηκε καν υπόψη της έρευνας.
- Το δεύτερο επίσης βασικό ερώτημα είναι, αν οι επιλεγείσες μεταβλητές επεξηγούν πραγματικά το μοντέλο με κατανόηση και σαφήνεια. Δεν είναι σπάνιο να γίνεται απόπειρα θεμελίωσης μίας σχέσης μεταξύ δύο μεταβλητών, ενώ στην πραγματικότητα υπάρχει μία τρίτη μεταβλητή υπεύθυνη σε μεγάλο βαθμό για τις μεταβολές που προλογίστηκαν για τις δύο πρώτες. Τονίζεται με έμφαση ότι, *είναι εύκολο να αναπτυχθεί μία σχέση μεταξύ δύο μεταβλητών με τη χρήση των στατιστικών προγραμμάτων και η ευκολία αυτή ενθαρρύνει παραπέρα το χρήστη σε εσφαλμένες απόπειρες περιγραφής επιστημονικών φαινομένων.*

Στην πράξη, τα μοντέλα της απλής και πολλαπλής παλινδρόμησης βρίσκουν τεράστια εφαρμογή σε διάφορα επιστημονικά πεδία, όπου χρησιμοποιούνται σε υπερβολικό

βαθμό για λόγους προβλεπτικής αξίας, εφόσον βέβαια είναι γνωστές οι εμπλεκόμενες μεταβλητές στα μοντέλα.

4.12 Μοντέλα Παλινδρόμησης

I. Πολυωνυμική παλινδρόμηση

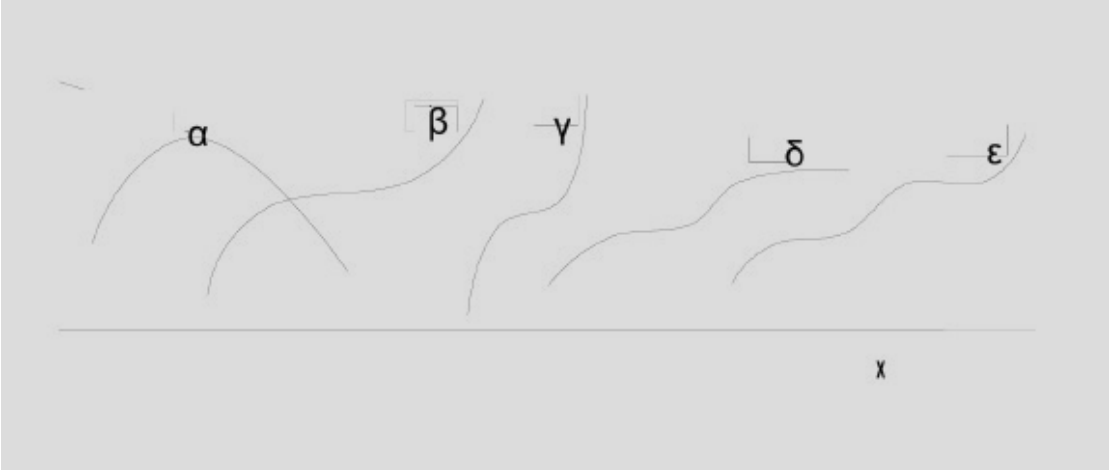
Ειδική μορφή της πολλαπλής γραμικής παλινδρόμησης αποτελεί η πολυωνυμική εξίσωση που εκφράζεται από τη σχέση: $\hat{Y} = a + b_1X + b_2X^2 + b_3X^3 + \cdots + b_mX^m$, η πιο συνεπτυγμένα σε

$$\hat{Y} = a + \sum_{i=1}^{m} b_iX^i \qquad (4.36)$$

όπου m είναι το πλήθος των διαφορετικών δυνάμεων της μεταβλητής X.

Μία τέτοια σχέση παριστάνει αυξανόμενες δυνάμεις της ανεξάρτητης μεταβλητής X, με διαφορετικό μερικό συντελεστή για κάθε δύναμη της X. Με άλλα λόγια, η εξίσωση περιλαμβάνει μία μόνο ανεξάρτητη μεταβλητή X, υψούμενη σε διαφορετικές δυνάμεις, γι`αυτό και η ποσότητα b_iX^i της εξίσωσης αποτελεί τον πολυωνυμικό όρο (***multinomial term***). Για λόγους καθαρά υπολογιστικούς, κάθε δύναμη της X λογίζεται ως μία διαφορετική ανεξάρτητη μεταβλητή, γιατί μόνο με αυτήν την προϋπόθεση γίνεται αποδεκτή η λύση της πολυωνυμικής εξίσωσης από τα στατιστικά προγράμματα.

Οι πολυωνυμικές εξισώσεις, γενικά, έχουν το χαρακτηριστικό της εμπειρικής προσαρμογής και μόνο γιατί οι πολυωνυμικοί όροι, X^2, X^3 κτλ., σε μία εξίσωση δεν παρέχουν ιδιαίτερη επιστημονική πληροφόρηση. Κάθε εξίσωση εκφράζεται γραφικά με καμπύλες διάφορων σχημάτων που εξαρτώνται αποκλειστικά από το πλήθος και τη φύση των εντασσόμενων πολυωνυμικών όρων Σχήμα 4.8.



Σχήμα 4.8 Πολυωνυμικές καμπύλες διαφόρων βαθμών: α) δευτεροβάθμια, β) τριτοβάθμια, γ) ειδική τριτοβάθμια χωρίς το δευτεροβάθμιο όρο, δ) τεταρτοβάθμια, ε) πεμπτοβάθμια.

Όταν ο άξονας X στα γραφήματα αντικατασταθεί με τις τιμές όλων των πολυωνυμικών όρων, τότε οι καμπύλες μετατρέπονται σε ευθεία γραμμή, συνεπώς η επίλυση της πολυωνυμικής εξίσωσης αποτελεί στην πραγματικότητα την περιγραφή μίας γραμμικής σχέσης μεταξύ δύο μεταβλητών X και Y με τη χρήση μετασχηματισμένων τιμών X_ι

Οι γνωστότερες πολυωνυμικές καμπύλες προέρχονται από τη δευτεροβάθμια εξίσωση, $\hat{Y} = a + b_1X + b_2X^2$ και την τριτοβάθμια, $\hat{Y} = a + b_1X + b_2X^2 + b_3X^3$. Το ιδεατό γεωμετρικό σχήμα της δευτεροβάθμιας καμπύλης διαγράφει *κοίλη καμπύλη όταν ο συντελεστής b_2 είναι αρνητικός και κυρτή καμπύλη όταν αυτός είναι θετικός.* Η φύση της καμπύλης παραβολής υπόκειται σε δύο <u>*περιορισμούς*</u>:

a) Είναι συμμετρική, δηλαδή η μεταβλητή Y στην αρχή αυξάνει με την αύξηση της X, προσεγγίζει μία μέγιστη τιμή και ακολούθως ελαττώνεται με τη συνεχιζόμενη αύξηση της X, καταλήγοντας σε μία ελάχιστη τιμή.
b) Η τιμή της Y γίνεται αρνητική στις μέγιστες και ελάχιστες τιμές της X, πράγμα που επιστημονικά φαίνεται παράδοξο. Η τιμή της Y στο μέγιστο και ελάχιστο σημείο της καμπύλης ισούται με,

$$Y_{max/min} = a - \frac{b_1^2}{4b_2}$$

Οι υπόλοιπες πολυωνυμικές εξισώσεις δεν εμφανίζουν συμμετρία, έχουν όμως εξαιρετικά άκαμπτη διαμόρφωση της καμπύλης τους στις μέγιστες και ελάχιστες τιμές.

Ο υπολογισμός της πολυωνυμικής εξίσωσης στηρίζεται και αυτός ***στη μέθοδο της άριστης επιλογής των όρων*** εκείνων της X οι οποίοι εμφανίζουν στατιστική σημαντικότητα και πέρα των οποίων η ένταξη αυξανόμενων δυνάμεων της X στην εξίσωση παρουσιάζεται μη σημαντική. Πρακτικά, σημαίνει ότι χρειάζεται να προσδιορίσουμε τη μέγιστη εκείνη πολυωνυμική δύναμη που τελικά είναι η τελευταία στατιστικά σημαντική, με τον περιορισμό ότι ο μέγιστος ενταχθείς πολυωνυμικός βαθμός θα πρέπει να είναι μικρότερος από $n-2$ παρατηρήσεις της παλινδρόμησης, για να έχει στατιστική αξιοπιστία.

Αν υιοθετήσουμε την προοδευτική απόρριψη των όρων θα πρέπει να συμπεριλάβουμε στην αρχή στο μοντέλο και όρους με μεγαλύτερη δύναμη από εκείνη που υποπτευόμαστε ότι χρειάζεται για την τελική προσαρμογή της εξίσωσης. Η εξέταση για την απόρριψη των υποψήφιων όρων, ξεκινά πάντοτε με τους όρους που έχουν τις υψηλότερες δυνάμεις και φθίνει συνεχώς, μέχρι να βρεθεί μία δύναμη, η μεγαλύτερη στη φθίνουσα πορεία, που εμφανίζει πολυωνυμικό συντελεστή στατιστικά σημαντικό ($b_i \neq 0$). Στο σημείο αυτό, περατώνεται και η διαδικασία της άριστης επιλογής των όρων και το πολυωνυμικό μοντέλο περιγράφεται με όλους τους εναπομείναντες όρους. Η προοδευτική απόρριψη των όρων πραγματοποιείται με τον έλεγχο $F-$ *απόρριψης* ή και με τον *έλεγχο* t του πολυωνυμικού συντελεστή,

$$t = \frac{b_i}{s_{(b_i)}}$$

Αποκλειστικά στην πολυωνυμική παλινδρόμηση και οι δύο αυτοί έλεγχοι είναι ισοδύναμοι.

Καλύτερα αποτελέσματα παρέχουν οι μέθοδοι του καταλληλότερου συνδυασμού προσαρμογής των όρων και της προοδευτικής ένταξης. Η τελευταία είναι ιδιαίτερα δημοφιλής και σύμφωνα με αυτήν ξεκινάμε με το μικρότερο δυνατό μοντέλο που ασφαλώς είναι αυτό της ευθείας γραμμής, $\hat{Y} = a + b_1 X$. Ακολούθως εισάγουμε το δευτεροβάθμιο όρο, τον τριτοβάθμιο κοκ. ενώ παράλληλα σε κάθε νέα ένταξη ελέγχουμε τη σημαντικότητα του όρου, καθώς και τη βελτίωση στην τιμή του πολλαπλού προσδιορισμού R^2 που η ένταξη προκαλεί (***έλεγχος F − ένταξης*** ή

έλεγχος t του πολυωνυμικού συντελεστή). Η προοδευτική ένταξη περατώνεται, όταν βρεθεί όρος με δύναμη μη στατιστικά σημαντική και τότε η εξίσωση της πολυωνυμικής παλινδρόμησης είναι η τελική και περιλαμβάνει όλους τους μέχρι τότε όρους. Επιβάλλεται πάντως, να ελέγχεται ως προς τη σημαντικότητα και ο όρος με την επόμενη μεγαλύτερη δύναμη για καλύτερη επιβεβαίωση της εξίσωσης. Δεν είναι σπάνιο, επίσης, κάποιος ενδιάμεσου βαθμού όρος να απορρίπτεται από την εξίσωση, όταν οι συνθήκες του μοντέλου το απαιτούν. Έτσι, είναι δυνατόν να προσαρμόσουμε στο μοντέλο μας την τριτοβάθμια πολυωνυμική εξίσωση χωρίς την ένταξη του δευτεροβάθμιου όρου, $\hat{Y} = a + b_1X + b_2X^2 + b_3X^3$ και αυτό πιστοποιείται γραφικά με την ανάπτυξη μίας καμπύλης με εξαιρετικά απότομη αλλαγή της καμπής (Σχήμα 4.8.γ). Μετά την προσαρμογή της πολυωνυμικής εξίσωσης, μπορούμε να προβλέψουμε τιμές της Ŷ από συνδυασμένες τιμές των όρων της X με την ίδια στατιστική κάλυψη που αναφέρθηκε στην πολλαπλή παλινδρόμηση.

II. Γενικευμένα γραμμικά μοντέλα

Α. Λογιστική Παλινδρόμηση

Η Λογιστική Παλινδρόμηση (Logistic Regression) συνιστάται στις περιπτώσεις που μελετάται ένα φαινόμενο που λαμβάνει τιμές 0/1 ή ναι/όχι, όταν δηλαδή η εξαρτημένη μεταβλητή είναι δίτιμη (binary variable) και ακολουθεί συνήθως διωνυμική κατανομή. Ωστόσο, συχνά αντί τις ίδιας της πιθανότητας p_i θέτουμε το μετασχηματισμό logit (πιθανότητες της αναλογίας) της πιθανότητας p_i ως εξαρτημένη μεταβλητή (Diez et al, 2012). Το μοντέλο logit περιγράφεται από την εξίσωση:

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \sum_{k=1}^{N} \alpha_k x_{ki} + \varepsilon_i$$

όπου p_i είναι η πιθανότητα η εξαρτημένη μεταβλητή να έχει τιμή 1 στο σημείο i, x_{ki} είναι η τιμή μίας από τις k ανεξάρτητες μεταβλητές στο σημείο i, α_0 είναι η σταθερά, α_k είναι η $k-οστή$ παράμετρος της μεταβλητής x_k και ε_i είναι το σφάλμα. Αν η παραπάνω εξίσωση λυθεί ως προς το p_i τότε προκύπτει:

$$p_i = \frac{e^{\alpha_0 + \sum_{k=1}^{N} \alpha_k x_{ki}}}{1 + e^{\alpha_0 + \sum_{k=1}^{N} \alpha_k x_{ki}}}$$

Β. Παλινδρόμηση Poisson

Η ουσιώδης διαφορά μεταξύ της παλινδρόμησης Poisson και της τυπικής πολλαπλής παλινδρόμησης είναι το γεγονός ότι η πρώτη αφορά την κατανομή Poisson και η δεύτερη την κανονική κατανομή. Η κατανομή πιθανότητας Poisson με παράμετρο μ δίνεται από τον τύπο:

$$p_Y(y;\mu) = pr(Y = y;\mu) = \frac{\mu^y e^{-\mu}}{y!} \quad , \quad y = 0,1,\dots,\infty$$

Μπορεί να αποδειχθεί θεωρητικά ότι $E(Y) = Var(Y) = \mu$. Η ανάλυση παλινδρόμησης Poisson είναι μια διαδικασία βασισμένη στη μέγιστη πιθανοφάνεια που έχει μια πιο σύνθετη συνάρτηση Πιθανοφάνειας σε σχέση με αυτή της γραμμικής παλινδρόμησης. Θα αναλύσουμε την παλινδρόμηση Poisson η οποία βρίσκει πολλαπλές εφαρμογές κατά την ανάλυση δεδομένων. Μια γενική μορφή της συνάρτησης Πιθανοφάνειας της παλινδρόμησης Poisson είναι:

$$L(y;\beta) = \prod_{i=1}^{\nu} p_{Y_i}(y_i;\beta) =$$

$$= \prod_{i=1}^{\nu} \left\{ \frac{[l_i \lambda(x_i,\beta)]^{y_i} e^{-l_i \lambda(x_i,\beta)}}{y_i!} \right\} = \frac{\{\prod_{i=1}^{\nu} [l_i \lambda(x_i,\beta)]^{y_i}\} exp[-\sum_{i=1}^{\nu} l_i \lambda(x_i,\beta)]}{\prod_{i=1}^{\nu} y_i!}$$

Όπου $E(Y_i) = \mu_i = l_i \lambda(x_i,\beta), i = 1,\dots,\nu, Y: εξαρτημένη\ μεταβλητή$. Στην πράξη, μια ιδιαίτερη μορφή της συνάρτησης ποσοστού $\lambda(x_i,\beta)$ θα πρέπει να οριστεί. Οι εκτιμήτριες ML $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ $των$ $\beta_0, \beta_1, \dots, \beta_k$ λαμβάνονται από τις παραπάνω εξισώσεις ως επιλύσεις των k + 1 εξισώσεων :

$$\frac{\partial[\ln L(y;\beta)]}{\partial \beta_i} = 0, \quad i = 0,1,\dots,k$$

Το μοντέλο, έλεγχοι και υπόλοιπα

Θεωρούμε το μοντέλο

$$y \sim Poisson(\mu)$$

Όπου η εξάρτηση από τις ***ανεξάρτητες μεταβλητές ή συμμεταβλητές (covariates)*** για μια στατιστική μονάδα εκφράζεται μέσω ενός κατάλληλου μετασχηματισμού $g(\cdot)$*της αναμενόμενης* τιμής της y, έτσι ώστε να ισχύει μια σχέση της μορφής

$$g(\mu_{\mathbf{x}}) = \mathbf{x}'\beta \qquad (4.37)$$

Η συνάρτηση $g(\cdot)$ καλείται συνάρτηση σύνδεσης (link function). Ο περιορισμός $\mu > 0$ που επιβάλλεται στην κατανομή Poisson σημαίνει ότι η $g(\cdot)$ δεν μπορεί να είναι η ταυτότητα $g(\mu_{\mathbf{x}}) = \mu_{\mathbf{x}}$, αφού στην περίπτωση αυτή $\mu_{\mathbf{x}} = \mathbf{x}'\beta$, η οποία προφανώς δεν τηρεί τον περιορισμό $\mu > 0$ για οποιοδήποτε x. Για να εξασφαλίσουμε τον περιορισμό αυτό, πρέπει η $\mu_{\mathbf{x}}$ να είναι μία μη αρνητική συνάρτηση του $\mathbf{x}'\beta$ και η πιο συνηθισμένη επιλογή είναι η $\mu_{\mathbf{x}} = e^{\boldsymbol{x'\beta}}$. Σε αυτή την περίπτωση επομένως η συνάρτηση σύνδεσης είναι η $g(\mu) = \ln \mu$. Το μοντέλο παλινδρόμησης Poisson έχει τις ακόλουθες προϋποθέσεις:

- $y_{\boldsymbol{x}} \sim Poisson(\mu_{\boldsymbol{x}})$
- $\mu_{\mathbf{x}} = e^{\boldsymbol{x'\beta}}$
- Ανεξαρτησία μεταξύ των $y_{\boldsymbol{x}}$ παρατηρήσεων
- Οι μετρήσεις τόσο της εξαρτημένης όσο και των ανεξάρτητων μεταβλητών δεν υπόκεινται σε σφάλματα μέτρησης.

Οι συμμεταβλητές $\mathbf{x}$ και οι αντίστοιχοι συντελεστές $\boldsymbol{\beta}$ ορίζονται όπως και στο γενικό γραμμικό μοντέλο, δηλαδή $\mathbf{x}' = (x_0, x_1, \dots, x_k)$ *και* $\boldsymbol{\beta}' = (\beta_0, \beta_1, \dots, \beta_k)$, όπου $x_0 \equiv 1$. Το μοντέλο $g(\mu_{\mathbf{x}}) = \mathbf{x}'\beta$ θυμίζει γραμμικό αλλά προφανώς είναι μη γραμμικό σε ό,τι αφορά τις παραμέτρους και γι'αυτό η προσαρμογή του απαιτεί ειδικές μεθόδους υπολογισμών.

Η προσαρμογή του μοντέλου στα δεδομένα γίνεται με τη μέθοδο μέγιστης Πιθανοφάνειας. Η συνάρτηση L ενός δείγματος τιμών $y_1, y_2, .., y_v$ με συμμεταβλητές $\mathbf{x}_i' = (x_{i0}, x_{i1}, \dots, x_{ik})$ που αντιστοιχούν στις y_i, όπου $x_{i0} \equiv 1$, δίνεται από τη σχέση

$$L = \prod_{i=1}^{v} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \qquad (4.38)$$

Η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας δίνεται από τη σχέση

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

η οποία λαμβάνοντας υπόψη ότι $\mu_\iota = \mu_{x_i} = e^{x_i'\beta}$, λαμβάνει τη μορφή

$$l = \sum_{i=1}^{\nu}[- e^{x_i'\beta} + y_i x_i'\beta - \ln(y_i!)] \qquad (4.39)$$

Και άρα

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{\nu}[x_{ij}(y_i - e^{x_i'\beta})], \qquad j = 0,1, \dots, k$$

Οι εκτιμήτριες μέγιστης πιθανοφάνειας των β_j προκύπτουν από την επίλυση των εξισώσεων

$$\sum_{i=1}^{\nu}[x_{ij}(y_i - e^{x_i'\widehat{\beta}})] = 0, \qquad j = 0,1, \dots, k$$

οι οποίες μπορούν να εκφραστούν ως $\mathbf{X}'(\mathbf{y} - \widehat{\boldsymbol{\mu}}) = \mathbf{0}$ όπου

$$\widehat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_\nu) \mu\varepsilon \ \ln \hat{\mu}_i = x_i'\hat{\beta}$$

Η εξίσωση είναι μη γραμμική ως προς τα $\hat{\beta}$, επειδή $\hat{\mu}_i = e^{x_i'\widehat{\beta}}$ και λύνεται μόνο με επαναληπτικές μεθόδους.

Καθεμία από τις ποσότητες $e^{\widehat{\beta}_j}$, $j = 0,1, \dots, k$, εκφράζει την αναμενόμενη πολλαπλασιαστική μεταβολή της y για μια μονάδα αύξησης της αντίστοιχής συμμεταβλητής x_j, δεδομένου ότι οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές.

Στη συνέχεια ο αναλυτής πραγματοποιεί ελέγχους για τους συντελεστές β, ενώ κρίνεται απαραίτητος και ο προσδιορισμός της ελεγχοσυνάρτησης «Deviance», η οποία μετρά την απώλεια προσαρμογής, όταν επιβάλλουμε τη δομή $\mu_i = e^{x_i'\widehat{\beta}}$ και δίνεται από τη σχέση :

$$D(\hat{\beta}) = 2\sum_{i=1}^{\nu} y_i \ln(y_i/\hat{\mu}_i) \qquad (4.40)$$

Όπως και στην περίπτωση του γενικού γραμμικού μοντέλου, έτσι και εδώ είναι απαραίτητη η εξέταση των υπολοίπων που εκφράζουν τη συμφωνία μεταξύ των παρατηρήσεων y_i και των αντίστοιχων προσαρμοσμένων τιμών $\hat{y}_i$ ή $\hat{\mu}_i$. Αυτή η εξέταση προσφέρει λεπτομέρειες για την καταλληλότητα του μοντέλου που δεν

φαίνονται στον ολικό έλεγχο με τη deviance. Χρησιμοποιούνται κυρίως δύο είδη υπολοίπων:

- Τα υπόλοιπα Pearson: $r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$ (4.41)
- Τα υπόλοιπα deviance: $r_i^D = sgn(y_i - \hat{\mu}_i)\{d_i(\hat{\beta})\}^{\frac{1}{2}},\ i = 1, \dots, \nu$ (4.42)

  όπου $sgn(y_i - \hat{\mu}_i)$ *είναι* το πρόσημο της διαφοράς $y_i - \hat{\mu}_i$, δηλαδή είναι η συνάρτηση που δίνει στην ποσότητα $\{d_i(\hat{\beta})\}^{\frac{1}{2}}$ θετικό πρόσημο όταν $y_i \geq \hat{\mu}_i$. Η ποσότητα $d_i(\hat{\beta}) = 2[y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$ αποτελεί τη συμβολή της παρατήρησης i στην ελεγχοσυνάρτηση deviance

  $$D(\hat{\beta}) = \sum_{i=1}^{\nu} (r_i^D)^2 \qquad (4.43)$$

Κεφάλαιο 5 : **Συμπεράσματα και το Μέλλον της Οικονομετρίας στα Media**

Η τεχνική MMM είναι μια διαδικασία που επιτρέπει στα οικονομετρικά μοντέλα να μετρήσουν και να ποσοτικοποιήσουν την αποτελεσματικότητα συγκεκριμένων δραστηριοτήτων μάρκετινγκ κατά τις επιχειρηματικές επιδόσεις. Στη συνέχεια τα τελικώς διαμορφωμένα μοντέλα χρησιμοποιούνται για να αξιολογήσουμε την «επιστροφή επί της επένδυσης», γνωστό και ως RoI (Return on Investment), για τη βέλτιστη κατανομή των πόρων μάρκετινγκ ώστε να ευνοηθεί η ανάπτυξη και τέλος για να προβλέψουμε τη διακύμανση της καταναλωτικής ζήτησης σε επερχόμενες χρονικές περιόδους.

Καθώς κατασκευάζεται ένα μοντέλο, ταυτοποιούνται οι οδηγοί «κλειδιά» και υπολογίζονται οι ποσοτικές επιπτώσεις τους τόσο βραχυπρόθεσμα όσο και μακροπρόθεσμα. Αυτό εξυπηρετεί πολύ τον επικοινωνιακό σχεδιασμό, καθώς επιτρέπει τον προσδιορισμό του τρόπου με τον οποίο κάθε κανάλι πολυμέσων πρέπει να συνεισφέρει στο συνολικό μίγμα. Επιπλέον όταν η επικοινωνιακή μέθοδος σταματά να είναι αποτελεσματική, επιτρέπει τον προσδιορισμό των κατωτάτων ορίων στα επίπεδα των δαπανών για τη διαφήμιση.

Ωστόσο, παρά τα οφέλη που αποφέρουν, τα οικονομετρικά μοντέλα συχνά αντιμετωπίζονται με σκεπτικισμό ή παρερμηνεύονται. Ο λόγος είναι ότι ένας αναλυτής θα μπορούσε να χρησιμοποιήσει το ίδιο σύνολο δεδομένων και να δημιουργήσει μια ποικιλία μαθηματικών μοντέλων που θα μπορούσαν να καταλήξουν να είναι ακόμη και αντιφατικά μεταξύ τους. Επιπλέον, καθώς για τη βέλτιστη πρόβλεψη και μοντελοποίηση απαιτείται αρκετή εμπειρία, η κρίση και επιδεξιότητα του αναλυτή, δεν είναι τόσο επιστημονικά βασισμένη όσο φαίνεται. Και αυτό είναι κάτι που δίνει υπερβολική δύναμη στα χέρια ενός αναλυτή γιατί είναι αυτός που αποφασίζει εάν ένας ανταγωνιστής είναι πιο σημαντικός από έναν άλλο. Η προστιθέμενη αξία συνήθως προκύπτει στις περιπτώσεις όπου τα αποτελέσματα της διαδικασίας της μοντελοποίησης θέτουν μία λογική βάση πίσω από μια συγκεκριμένη στρατηγική, υποστηρίζουν συγκεκριμένες υποθέσεις ή ακόμα και ελαχιστοποιούν τις απώλειες μιας τακτικής χωρίς ιδιαίτερες προοπτικές.

Το Mix Marketing ως εργαλείο διαχείρισης έχει σημειωθεί ότι έχει δύο κύριους περιορισμούς: τον εσωτερικό προσανατολισμό του μοντέλου και την έλλειψη

εξατομίκευσης. Έτσι προτείνεται όποια προσπάθεια γίνεται, να επικεντρώνεται στη διαμόρφωση των εννοιολογικών θεμελίων και των μεθοδολογιών μάρκετινγκ που ανταποκρίνονται καλύτερα στις εμπορικές ανάγκες του σήμερα και του αύριο. Γι' αυτό και σε αυτή την εργασία παραθέσαμε μεθοδολογικά εργαλεία με στόχο να εξυπηρετήσουμε την προαναφερθείσα νοοτροπία.

Η δημιουργία μεταβλητών για το Marketing Mix Modeling είναι μια περίπλοκη υπόθεση και είναι τόσο τέχνη όσο και επιστήμη. Έτσι η συζήτηση για το ποιος έχει τη θέση υπεροχής μεταξύ των αυτοματοποιημένων εργαλείων μοντελοποίησης που διαχειρίζονται μεγάλα σύνολα δεδομένων και του άρτια ειδικευμένου οικονομολόγου, εξακολουθεί να υφίσταται στη θεωρία του MMM. Ωστόσο, είναι γεγονός ότι η MMM ως τεχνική έχει καταφέρει να εξελιχθεί από μια ελάχιστα χρησιμοποιούμενη μεθοδολογία με ακαδημαϊκό προσανατολισμό και πλαίσιο, σε ένα ισχυρό, ευρέως διαδεδομένο και κοινό εργαλείο μάρκετινγκ. Η μεθοδολογία που παρουσιάστηκε σε αυτή την εργασία μαζί με τα εικονικά γραφήματα που συνοδεύουν πολλές φορές τα δεδομένα μπορεί να δοκιμαστεί σε ένα case study για διάφορες κατηγορίες αγορών, μεταξύ των οποίων της Fast-Moving Consumer Goods (FMCG), της αυτοκινητοβιομηχανίας και της βιομηχανία των fast food, καθώς και σε πολλές πολυεθνικές μάρκες.

Η προσέγγιση των Predictive Analytics που έχουμε χρησιμοποιήσει, διαμορφώνει έναν τύπο μοντέλων μάρκετινγκ μικρής κλίμακας, εστιάζοντας σε μεμονωμένες μάρκες ή προϊόντα, αντί να εστιάζει σε συγκεντρωτικές τιμές. Με άλλα λόγια, αξιολογούνται και συγκρίνονται τα αποτελέσματα των εναλλακτικών στρατηγικών και αναπτύσσονται τεχνικές επίλυσης για την επίτευξη ενός επιθυμητού στόχου στην περίπτωση ενός case study. Επομένως, το γενικό πεδίο εφαρμογής ήταν να αναλυθούν μεταβλητές που μεσολαβούν στην επίδραση της διαφήμισης στις πωλήσεις, με μετρήσιμο τρόπο. Η προτεινόμενη μεθοδολογία μπορεί να μας δώσει πολύ ακριβή αποτελέσματα πρόβλεψης, ακόμη και σε μια εκτίμηση επιπέδου σημείου. Επιπλέον, σχηματίζει μια νοοτροπία που περικλείεται στο πλαίσιο που δημιουργείται και είναι βασισμένο στους επιχειρηματικούς KPIs. Ακόμα, μας επιτρέπει τη μέτρηση της αποτελεσματικότητας με διάφορους τρόπους: από τη βελτιστοποίηση της στόχευσης συγκεκριμένων ακροατηρίων μέχρι την επίτευξη προβολής σε συγκεκριμένες ζώνες ώρας. Τελευταίο αλλά εξίσου σημαντικό, είναι το γεγονός πως καταδεικνύει την ανάπτυξη εργαλείων που μπορούν να

χρησιμοποιηθούν παράλληλα από τα Media και τις ομάδες της μάρκας ώστε να υλοποιηθούν οι αμοιβαίοι στόχοι τους για την ενδυνάμωση της εκάστοτε εταιρείας.

Αλλά το ταξίδι δεν περιορίζεται στην παραδοσιακή απόδοση των Media, όπου οι επενδύσεις στο μάρκετινγκ εκτός σύνδεσης επηρεάζουν τις πωλήσεις εκτός σύνδεσης. Τα σύγχρονα οικονομετρικά μοντέλα έχουν εξελιχθεί ώστε να λαμβάνουν υπόψη τον τρόπο με τον οποίο επηρεάζουν τα ψηφιακά μέσα τις πωλήσεις, τη διείσδυση και οποιοδήποτε άλλο εξαρτώμενο KPI του ενδιαφέροντος μας. Αυτή η προσέγγιση μοντελοποίησης συναντάται συνήθως στη βιβλιογραφία με τον τίτλο «Μοντελοποίηση ψηφιακής απόδοσης» (Digital attribution modeling) και στοχεύει στον εντοπισμό του συνδυασμού των δραστηριοτήτων ηλεκτρονικού μάρκετινγκ και των σημείων επαφής που συμβάλλουν στη μετατροπή των online πωλήσεων. Και λόγω του γεγονότος ότι υπάρχουν πολλές μετρήσεις που σχετίζονται με την αναγνώριση μοναδικών χρηστών, μια τέτοια ανάλυση μπορεί φυσικά να εντοπίσει τις ενέργειες μεμονωμένων ατόμων. Τέλος, στις μέρες μας είναι εφικτό να συνδυάσουμε τις επενδύσεις στον τομέα του μάρκετινγκ που πραγματοποιούνται ταυτόχρονα offline και online σε ένα δυναμικό σύστημα για να αποκτήσουμε μια πλήρη, συγκεντρωτική άποψη της κατανομής των πωλήσεων των προϊόντων.

Συμπεραίνοντας, έχει υπάρξει εκτενής ανάλυση στη βιβλιογραφία σχετικά με την προσαρμογή του κάθε μοντέλου και την ικανότητα πρόβλεψης, ενώ διάφορες μετρήσεις, μετρικές, δοκιμές και δείκτες φαίνεται να ενισχύουν τους αναλυτές στην σωστή απόφαση για το βέλτιστο προτεινόμενο μοντέλο. Σε κάθε βήμα των αναλύσεων ακολουθείται ένας μεγάλος αριθμός υπολογισμών που εμφανίζονται στις προτεινόμενες κατευθυντήριες γραμμές της βιβλιογραφίας. Όλα τα αποτελέσματα αυτών των μετρήσεων αντισταθμίστηκαν, τροφοδοτήθηκαν και καθοδηγούνται από τις στρατηγικές ανάγκες και τις αποφάσεις της μάρκας κατά τον αποκλεισμό ή όχι των δυνητικών ανεξάρτητων προγνωστικών. Τέτοιες αποφάσεις, που λαμβάνονται με βάση την πιθανή επίδρασή τους στην προγνωστική ικανότητα του μοντέλου, ανυψώνουν το μαθηματικό μοντέλο σε μια περίπλοκη υπόθεση, εξισορροπώντας την τέχνη με την επιστήμη. Επομένως, είτε μας αρέσει είτε όχι, ζούμε σε ενδιαφέρουσες στιγμές σχετικά με το παρόν και το μέλλον της οικονομετρίας. Φαίνεται ότι όλα τα κομμάτια του παζλ είναι πια στη θέση τους:

- Επαρκείς μαθηματικές γνώσεις.
- Συνεχής αύξηση της υπολογιστικής ισχύος.

- Υπάρχει μία επείγουσα ανάγκη για συνεχή βελτιστοποίηση από πλευράς απόδοσης, αποτελεσματικότητας, ROI.
- Οι έμπειροι μαθηματικοί επαγγελματίες μοντελοποίησης έχουν αρχίσει να φεύγουν από τα πανεπιστήμια και έχουν αρχίσει να προσεγγίζουν τη βιομηχανία για μια πιο πολλά υποσχόμενη καριέρα.
- Οι υπεύθυνοι λήψης αποφάσεων θέτουν τα σωστά ερωτήματα (απαντήσεις ποσοτικά μετρήσιμες).
- Επάρκεια των δεδομένων.

Σε αυτή την περίπτωση, διαφαίνεται πως μπαίνουμε σε μία εντελώς νέα και συναρπαστική εποχή.

Βιβλιογραφικές Αναφορές

1. J.Coletsos, D. Stogiannis. Evaluating the impact of Media Key Performance Indicators on Business Key Performance Indicators
2. Π.Οικονόμου, Χ.Καρώνη Στατιστικά Μοντέλα Παλινδρόμησης
3. Dr.Matthew Ryan Lavery, Dr. Parul Acharya, Dr. Stephen A. Sivo & Dr. Lihua Xu. Number of Predictors and Multicollinearity: What Are Their Effects on Error and Bias in Regression? Communications in Statistics - Simulation and Computation
4. Canan Eryigit. Hacettepe University, Turkey. Marketing models. International Journal of Market Research Vol.59 Issue 3.
5. Neil H.Borden. Harvard Business School. The Concept of the Marketing Mix.
6. Besley D.A., Kuh E. and Welsch R.E. (1980). Regression Diagnostics. John Wiley & Sons, N.Jersey 310 p.
7. Burn D.A. & Ryan T.A. (1983). A Diagnostic Test for Lack of Fit in Regression Models. Proceedings of the Statistical Computing Section, 286-290.
8. Cook R.D. (1977). Detection of Influential Observations in Linea r Regression. Technometrics, 19, 15-18.
9. Cook R.D. & Weisberg S. (1982). Residuals and Influence in Regression. Chapman and Hall, London, 229p.
10. Draper N.R. & Smith H. (1981). Applied regression analysis, 2nd ed. John Wiley and Sons, New York, 709p.
11. Hocking R.R. (1976). A Biometrics Invited Paper: The Analysis and Selection of Variables in Linear Regression, Biometrics, 32, 1-49.
12. Kleinbaum D.G, Kupper L.L., Muller K.E. and Nizam A. (1998). Applied regression analysis and other multivariate methods. Duxbury Press, N. York, 489p,798 p.
13. Montgomery D.C., Peck E.A. and Vining G.G. (2012). Introduction to Linear Regression Analysis. 5th ed. John Wiley & Sons, N. Jersey, 672 p.
14. Nagelkerke N.J.D. (1991). A note on the general definition of the coefficient of determination. Biometrika, 78(3), 691-692.
15. Pedhazur E.J. (1997). Multiple regression in behavioral research. Explanation and prediction, 3rd ed. Orlando FL: Harcourt Brace, 1098 p.
16. Petridis D. & Rogdakis I. (1996). The development of growth and feeding equations for sea bream (Sparus auratus) culture. Aquaculture Research, 27, 413-419.
17. Velleman P. & Welsch R. (1981). Efficient Computation of Regression Diagnostics. The American Statistician, 35, 234-242.
18. Weisberg S. (1985). Applied Linear Regression, 2nd ed. John Wiley and Sons, New York, 310 p.

19. Alsem, K.J., Leeflang, P.S.H. and Reuyl, J.C. (1989). The forecasting accuracy of market share models using predicted values of competitive marketing behavior. International Journal of Research in Marketing. 6:3, pp. 183-198.
20. Assmus, G., Farley, J.U. and Lehmann, D.R. (1984). How Advertising Affects Sales: Meta-Analysis of Econometric Results. Journal of Marketing Research. 21:1, 65-74.
21. Austina, P.C., Steyerbergd, E.W. (2015). The number of subjects per variable required in linear regression analyses. Journal of Clinical Epidemiology. 68, pp. 627-636.
22. Andreyevaa, T., Kelly, I.R. and Harris, J.L. (2011). Exposure to food advertising on television: Associations with children's fast food and soft drink consumption and obesity. Economics & Human Biology. 9: 3. pp. 221-233.
23. Bendixen, M.T. (1993). Advertising Effects and Effectiveness. European Journal of Marketing. 27:10, pp. 19-32.
24. Bergera, P.D., Bechwatib, N.N. (2001). The allocation of promotion budget to maximize customer equity. Omega. 29:1, pp. 49-61.
25. Bodenlos, J.S. and Wormuth, B.M. (2013). Watching a food-related television show and caloric intake. A laboratory study. Appetite. 61, pp. 8-12.
26. Borden, N.H. (1964). The Concept of the Marketing Mix. Journal of Advertising Research. pp. 1-7.
27. Borden, N.H. (1942). The Economic Effects of Advertising. The ANNALS of the American Academy of Political and Social Science. 221: 11, pp 218-219.
28. Boyland, E.J. and Halford, J.C.G. (2013). Television advertising and branding. Effects on eating behaviour and food preferences in children. Appetite. 62:1, pp. 236-241.
29. Brown, S. (1996). Art or science? Fifty years of marketing debate. Journal of Marketing Management. 12:4, pp. 243-267.
30. Büschken, J. (2007). Determinants of Brand Advertising Efficiency: Evidence from the German Car Market. Journal of Advertising. pp. 51-73.
31. Cain. P. (2014). Econometrics: Digital media attribution. Admap. pp 1-6.
32. Canan E. (2017). Marketing models: A review of the literature. International Journal of Market Research, 59:3, pp. 355-381.
33. Carpenter, G.S. and Lehmann, D.R. (1985). A Model of Marketing Mix, Brand Switching, and Competition. Journal of Marketing Research. 22: 3, pp. 318-329.
34. Clarke, D.G. (1976). Econometric Measurement of the Duration of Advertising Effect on Sales. Journal of Marketing Research. 13:4, pp. 345-357.
35. Constantinides, E. (2006). The Marketing Mix Revisited: Towards the 21st Century Marketing. Journal of Marketing Management. 22:3-4, pp. 407-438.
36. Cook, L. and Holmes, M. (2004). Econometrics explained. IPA.
37. Cook, L. (2017). Econometrics explained 2. IPA.

38. Cook, L. (2014). Econometrics: Get the best from econometric modelling. WARC. pp. 1-6.
39. Cooper. L.G. and Nakanishi. M. (1988). Market-Share Analysis: Evaluating Competitive Marketing Effectiveness. Kluwer Academic Publishers. Boston Dordrecht London. pp 290.
40. Crespo-Cuaresma, J. and Stoeckl, M. (2012). The Effect of Marketing Spending on Sales in the Premium Car Segment: New Evidence from Germany. Economics and Finance. 2. pp. 1-26.
41. Danaher, P.J. and Rust, R.T. (1994). Determining the optimal level of media spending. Journal of Advertising Research. 34:1, 28-34.
42. Dekimpe, M.G. and Hanssens, D.M. (1995). The Persistence of Marketing Effects on Sales. Marketing Science. 14:1.
43. Denning, S. (2014). "The Best Of Peter Drucker". Forbes.
44. Donatos, G.S. and Kioulafas, K.E. (1990). A quantitative analysis of New Car Sales and Advertising in Greece, European Journal of Operational Research, 48:3, pp. 311-317.
45. Dughill, C. (2014). The Power of Habit: Why We Do What We Do in Life and Business. Random House Trade Paperbacks.
46. Harari, Y.N. (2016). Homo Deus: A Brief History of Tomorrow. Harvill Secker.
47. Hoffer, G., Marchand, J. and Albertine, J. (1976). Pricing in the Automobile Industry: A Simple Econometric Model. Southern Economic Journal, 43: 1, pp. 948-951.
48. Interactive Advertising Bureau (2018). Artificial Intelligence: Myth versus reality in the digital advertising world.
49. Jayson, R. (2016). Effectiveness: The impact of paid and digital owned media on Sales. Warc. pp. 1-6.
50. Jones, C.I. (1995). R&D-based models of economic growth. Journal of Political Economy. 103:4, pp. 759–784.
51. Kehrer. D. (2015). Precise attribution fuels marketing effectiveness. Admap. pp 1-6.
52. Kucuk, S.U. (2016). Marketing-Mix Modeling. Visualizing Marketing. pp. 83-93.
53. Little, J.D.C. (1975). Brandaid: A Marketing-Mix Model, Part 1: Structure. Operations Research. pp. 628-655.
54. Little, J.D.C. (1975). Brandaid: A Marketing-Mix Model, Part 2: Implementation, Calibration, and Case Study. Operations Research. pp. 656-673.
55. Lavery, M.R., Acharya, P., Sivo, S.A., Xu, L. (2017). Number of Predictors and Multicollinearity: What Are Their Effects on Error and Bias in Regression?. Communications in Statistics - Simulation and Computation. pp 1-21.

56. Leeflang, P.S. & Wittink, D.R. (2000). Building models for marketing decisions: past, present and future. International Journal of Research in Marketing, 17:2, pp. 105–126.
57. Leeflang, P.S., Wittink, D.R., Wedel, M. & Naert, P.A. (2000). Building Models for Marketing Decisions. Boston, MA. Kluwer Academic Publishers.
58. Luo, X. and Donthu, N. (2005). Assessing advertising media spending inefficiencies in generating sales. Journal of Business Research. 58:1, pp. 28-36.
59. Mariel, P., Lopez, C. and Fernandez, K. (2006). Sales-advertising relationship: an application of panel data from the German automobile industry. Prague economic papers. 1, pp. 29-43.
60. McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey & Company.
61. McKinsey Global Institute. (2016). The age of analytics: Competing in a data-driven world. McKinsey & Company.
62. Makasi, A., Govender, K., Rukweza, C. (2014). Building Brand Equity through Advertising. Mediterranean Journal of Social Sciences. 5:20, pp. 2613-2624.
63. Nigel, H. (1994). The link between TV ad awareness and sales. New evidence from sales response modelling. Journal of the Market Research Society. pp. 41.
64. O'Brien, R.M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors, Quality & Quantity. 41:5, 673–690.
65. Ofir, C. Khuri, A. (1986). Multicollinearity in marketing models: Diagnostics and remedial measures. International Journal of Research in Marketing. 3:3, pp 181-205.
66. Pesaran, M.H. (1987). The New Palgrave: A Dictionary of Economics, Econometrics. 2, pp. 8–22.
67. Rothman K.J., Greenland S. and Lash T. (2008). Modern Epidemiology. Lippincott, Williams and Wilkins, third edition.
68. Samuelson, P.A., Koopmans, T.C. and Stone J.R.N. (1954). Report of the Evaluative Committee for Econometrica, Econometrica. 22(2), pp. 141-146.
69. Samuelson P.A. and Nordhaus, W.D. (2004). Economics. 18th ed., McGraw-Hill.
70. Sethuraman, R., Tellis, G. J. and Briesch, R.A. (2011). How well does advertising work? Generalizations From A Meta-Analysis of Brand Advertising Elasticity. Journal of Marketing Research. 48:3. pp. 457-471.
71. Sharp. B. (2010). How brands grow: what marketers don't know. Oxford University Press. pp 246.
72. Shaw, R. (2015). Marketing's magic metric. WARC. pp. 1-7.
73. Sloane, C. and Fellows, S. (2014). Econometrics for media optimisation. WARC. pp. 1-6.

74. Snee, R.D. and Marquardt, D.W. (1984). Collinearity Diagnostics Depend on the Domain of Prediction, the Model, and the Data, The American Statistician, 38:2, 83-87.
75. Sundsoy, P., Bjelland, J., Iqbal, A.M., Pentland, A.S. and Montjoye, Y.A. (2014). Big Data-Driven Marketing: How machine learning outperforms marketers' gut-feeling. Social Computing, Behavioral-Cultural Modeling & Prediction, LNCS. Volume 8393, pp 367-374.
76. Sunoo, D.H. and Lin, L.Y.S. (2013). A Search for Optimal Advertising Spending Level. Optimal Advertising Expenditures. pp. 25-28.
77. Tellis, G. J. and Weiss, D. L. (1995). Does TV advertising really affect sales? The role of measures, models and data aggregation. Journal of Advertising. 24;3. pp. 1-12.
78. The global media intelligence report 2018: A Reference Guide to Consumers' Media Use in 40 Countries, eMarketer.
79. Vakratsas, D. and Ambler. T. (1999). How Advertising Works: What Do We Really Know?. Journal of Marketing. 63:1, pp. 26-43.
80. Vatcheva, K.P., Lee, M., McCormick, J.B. and Rahbar, M.H. (2016).
81. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. Epidemiology. 6:2, pp 1-20.
82. Velilla, S. (2018). A note on collinearity diagnostics and centering, The American Statistician. 72:2, pp 140-146.
83. Winer, R.S. and Neslin, S.A. (2014). The History of Marketing Science. Singapore: World Scientific Pub. Co.; Hanover, MA: Now Publishers Inc.
84. Wilton, D.A. (1972). An Econometric Model of the Canadian Automotive Manufacturing Industry and the 1965 Automotive Agreement. The Canadian Journal of Economics. 5:2, pp. 157-181.
85. Young, B. (2003). Does food advertising influence children's food choices? A critical review of some of the recent literature. International Journal of Advertising - The Review of Marketing Communications. 22:4. pp. 441-459.

Σύνδεσμοι

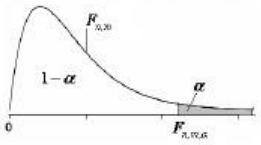
1. https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/
2. https://en.wikipedia.org/wiki/Prescriptive_analytics
3. https://el.wikipedia.org/wiki/%CE%A3%CF%84%CE%B1%CF%84%CE%B9%CF%83%CF%84%CE%B9%CE%BA%CE%AE
4. https://slideplayer.gr/slide/2015423/
5. http://www.math.ntua.gr/~fouskakis/Data_Analysis/07.pdf

Παράρτημα

Τιμές $F_{n,m;\alpha}$ της κατανομής $F_{n,m}$

Οι Πίνακες δίνουν τα άνω α -ποσοστιαία σημεία της κατανομής F με n και m βαθμούς ελευθερίας, για $\alpha = 0.05$ και $\alpha = 0.01$, αντίστοιχα.

Αν $X \sim F_{n,m}$, ισχύει, $P(X > F_{n,m;\alpha}) = a$. Επίσης ισχύει, $F_{n,m;1-\alpha} = 1/F_{m,n;\alpha}$



n = Βαθμοί ελευθερίας για τον αριθμητή

$\alpha = 0.05$

m \ n	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

n = Βαθμοί ελευθερίας για τον αριθμητή

$\alpha = 0.01$

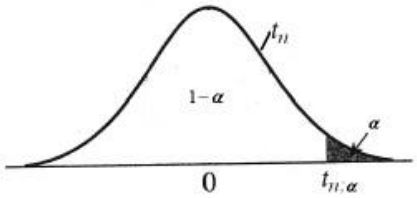
m \ n	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Πίνακας Π1 της κατανομής $F_{n,m}$

Τιμές $t_{n;\alpha}$ της κατανομής t_n

Ο Πίνακας δίνει τα άνω α -ποσοστιαία σημεία της κατανομής t με n βαθμούς ελευθερίας

Αν $T \sim t_n$, ισχύει, $P(T > t_{n;\alpha}) = \alpha$. Επίσης, ισχύει, $t_{n;1-\alpha} = -t_{n;\alpha}$



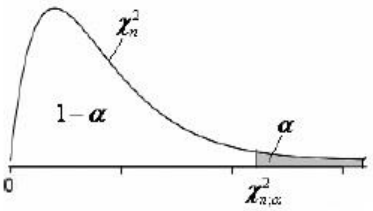
n	*α = 0.1*	*α = 0.05*	*α = 0.025*	*α = 0.01*	*α = 0.005*
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

Πίνακας Π2 της κατανομής t_n

Τιμές $\chi^2_{n;\alpha}$ της κατανομής χ^2_n

Ο Πίνακας δίνει τα άνω α -ποσοστιαία σημεία της κατανομής χ^2 με n βαθμούς ελευθερίας

Αν $X \sim \chi^2_n$, ***ισχύει,*** $P(X > \chi^2_{n;\alpha}) = \alpha$.



n	*α = 0.995*	*α = 0.99*	*α = 0.975*	*α = 0.95*	*α = 0.05*	*α = 0.025*	*α = 0.01*	*α = 0.005*
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21,920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.414	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.878	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.335
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.706	22.164	24.4331	26.509	55.756	59.342	63.691	66.766
50	27.991	29.708	32.3574	34.764	67.505	71.420	76.154	79.490
60	35.535	37.485	40.4817	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.7576	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.1532	60.392	101.879	106.629	112.329	116.321
90	59.196	61.754	65.6466	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.2219	77.930	124.342	129.561	135.807	140.169

Πίνακας Π3 της κατανομής χ^2_n

Πίνακας Π4 Durbin Watson $\alpha = 1\%$ $και$ k =αριθμός των ερμηνευτικών μεταβλητών

n\k	1		2		3		4		5		6		7		8		9		10	
6	0.390	1.142																		
7	0.435	1.036	0.294	1.676																
8	0.497	1.003	0.345	1.489	0.229	2.102														
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433												
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690										
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453	0.124	2.892								
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.665	0.105	3.053						
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838	0.090	3.182				
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	0.183	2.667	0.122	2.981	0.078	3.287		
15	0.811	1.070	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101	0.068	3.374
16	0.844	1.086	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944	0.094	3.201
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811	0.127	3.053
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.697	0.160	2.925
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597	0.196	2.813
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510	0.232	2.174
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434	0.268	2.625
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691	0.587	1.849	0.510	2.015	0.437	2.188	0.368	2.367	0.304	2.548
23	1.017	1.186	0.938	1.290	0.858	1.407	0.777	1.535	0.699	1.674	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308	0.340	2.479
24	1.037	1.199	0.959	1.298	0.881	1.407	0.805	1.527	0.728	1.659	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255	0.375	2.417
25	1.055	1.210	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645	0.682	1.776	0.610	1.915	0.540	2.059	0.473	2.209	0.409	2.362
26	1.072	1.222	1.000	1.311	0.928	1.410	0.855	1.517	0.782	1.635	0.711	1.759	0.640	1.889	0.572	2.026	0.505	2.168	0.441	2.313
27	1.088	1.232	1.019	1.318	0.948	1.413	0.878	1.514	0.808	1.625	0.738	1.743	0.669	1.867	0.602	1.997	0.536	2.131	0.473	2.269
28	1.104	1.244	1.036	1.325	0.969	1.414	0.901	1.512	0.832	1.618	0.764	1.729	0.696	1.847	0.630	1.970	0.566	2.098	0.504	2.229
29	1.119	1.254	1.053	1.332	0.988	1.418	0.921	1.511	0.855	1.611	0.788	1.718	0.723	1.830	0.658	1.947	0.595	2.068	0.533	2.193
30	1.134	1.264	1.070	1.339	1.006	1.421	0.941	1.510	0.877	1.606	0.812	1.707	0.748	1.814	0.684	1.925	0.622	2.041	0.562	2.160

n\k	1		2		3		4		5		6		7		8		9		10	
31	1.147	1.274	1.085	1.345	1.022	1.425	0.960	1.509	0.897	1.601	0.834	1.698	0.772	1.800	0.710	1.906	0.649	2.017	0.589	2.131
32	1.160	1.283	1.100	1.351	1.039	1.428	0.978	1.509	0.917	1.597	0.856	1.690	0.794	1.788	0.734	1.889	0.674	1.995	0.615	2.104
33	1.171	1.291	1.114	1.358	1.055	1.432	0.995	1.510	0.935	1.594	0.876	1.683	0.816	1.776	0.757	1.874	0.698	1.975	0.641	2.080
34	1.184	1.298	1.128	1.364	1.070	1.436	1.012	1.511	0.954	1.591	0.896	1.677	0.837	1.766	0.779	1.860	0.722	1.957	0.665	2.057
35	1.195	1.307	1.141	1.370	1.085	1.439	1.028	1.512	0.971	1.589	0.914	1.671	0.857	1.757	0.800	1.847	0.744	1.940	0.689	2.037
36	1.205	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.987	1.587	0.932	1.666	0.877	1.749	0.821	1.836	0.766	1.925	0.711	2.018
37	1.217	1.322	1.164	1.383	1.112	1.446	1.058	1.514	1.004	1.585	0.950	1.662	0.895	1.742	0.841	1.825	0.787	1.911	0.733	2.001
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.584	0.966	1.658	0.913	1.735	0.860	1.816	0.807	1.899	0.754	1.985
39	1.237	1.337	1.187	1.392	1.137	1.452	1.085	1.517	1.033	1.583	0.982	1.655	0.930	1.729	0.878	1.807	0.826	1.887	0.774	1.970
40	1.246	1.344	1.197	1.398	1.149	1.456	1.098	1.518	1.047	1.583	0.997	1.652	0.946	1.724	0.895	1.799	0.844	1.876	0.749	1.956
45	1.288	1.376	1.245	1.424	1.201	1.474	1.156	1.528	1.111	1.583	1.065	1.643	1.019	1.704	0.974	1.768	0.927	1.834	0.881	1.902
50	1.324	1.403	1.285	1.445	1.245	1.491	1.206	1.537	1.164	1.587	1.123	1.639	1.081	1.692	1.039	1.748	0.997	1.805	0.955	1.864
55	1.356	1.428	1.320	1.466	1.284	1.505	1.246	1.548	1.209	1.592	1.172	1.638	1.134	1.685	1.095	1.734	1.057	1.785	1.018	1.837
60	1.382	1.449	1.351	1.484	1.317	1.520	1.283	1.559	1.248	1.598	1.214	1.639	1.179	1.682	1.144	1.726	1.108	1.771	1.072	1.817
65	1.407	1.467	1.377	1.500	1.346	1.534	1.314	1.568	1.283	1.604	1.251	1.642	1.218	1.680	1.186	1.720	1.153	1.761	1.120	1.802
70	1.429	1.485	1.400	1.514	1.372	1.546	1.343	1.577	1.313	1.611	1.283	1.645	1.253	1.680	1.223	1.716	1.192	1.754	1.162	1.792
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.586	1.340	1.617	1.313	1.649	1.284	1.682	1.256	1.714	1.227	1.748	1.199	1.783
80	1.465	1.514	1.440	1.541	1.416	1.568	1.390	1.595	1.364	1.624	1.338	1.653	1.312	1.683	1.285	1.714	1.259	1.745	1.232	1.777
85	1.481	1.529	1.458	1.553	1.434	1.577	1.411	1.603	1.386	1.630	1.362	1.657	1.337	1.685	1.312	1.714	1.287	1.743	1.262	1.773
90	1.496	1.541	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636	1.383	1.661	1.360	1.687	1.336	1.714	1.312	1.741	1.288	1.769
95	1.510	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.641	1.403	1.666	1.381	1.690	1.358	1.715	1.336	1.741	1.313	1.767
100	1.522	1.562	1.502	1.582	1.482	1.604	1.461	1.625	1.441	1.647	1.421	1.670	1.400	1.693	1.378	1.717	1.357	1.741	1.335	1.765
150	1.611	1.637	1.598	1.651	1.584	1.665	1.571	1.679	1.557	1.693	1.543	1.708	1.530	1.722	1.515	1.737	1.501	1.752	1.486	1.767
200	1.664	1.684	1.653	1.693	1.643	1.704	1.633	1.715	1.623	1.725	1.613	1.735	1.603	1.746	1.592	1.757	1.582	1.768	1.571	1.779

n\k	11		12		13		14		15		16		17		18		19		20	
6																				
7																				
8																				
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16	0.098	3.503																		
17	0.138	3.378	0.087	3.557																
18	0.177	3.265	0.123	3.441	0.078	3.603														
19	0.220	3.159	0.160	3.335	0.111	3.496	0.070	3.642												
20	0.263	3.063	0.200	3.234	0.145	3.395	0.100	3.542	0.063	3.676										
21	0.307	2.976	0.240	3.141	0.182	3.300	0.132	3.448	0.091	3.583	0.058	3.705								
22	0.349	2.897	0.281	3.057	0.220	3.211	0.166	3.358	0.120	3.495	0.083	3.619	0.052	3.731						
23	0.391	2.826	0.322	2.979	0.259	3.128	0.202	3.272	0.153	3.409	0.110	3.535	0.076	3.650	0.048	3.753				
24	0.431	2.761	0.362	2.908	0.297	3.053	0.239	3.193	0.186	3.327	0.141	3.454	0.101	3.572	0.070	3.678	0.044	3.773		
25	0.470	2.702	0.400	2.844	0.335	2.983	0.275	3.119	0.221	3.251	0.172	3.376	0.130	3.494	0.094	3.604	0.065	3.702	0.041	3.790
26	0.508	2.649	0.438	2.784	0.373	2.919	0.312	3.051	0.256	3.179	0.205	3.303	0.160	3.420	0.120	3.531	0.087	3.632	0.060	3.724
27	0.544	2.600	0.475	2.730	0.409	2.859	0.348	2.987	0.291	3.112	0.238	3.233	0.191	3.349	0.149	3.460	0.112	3.563	0.081	3.658
28	0.578	2.555	0.510	2.680	0.445	2.805	0.383	2.928	0.325	3.050	0.271	3.168	0.222	3.283	0.178	3.392	0.138	3.495	0.104	3.592
29	0.612	2.515	0.544	2.634	0.479	2.755	0.418	2.874	0.359	2.992	0.305	3.107	0.254	3.219	0.208	3.327	0.166	3.431	0.129	3.528
30	0.643	2.477	0.577	2.592	0.512	2.708	0.451	2.823	0.392	2.937	0.337	3.050	0.286	3.160	0.238	3.266	0.195	3.368	0.156	3.465

n\k	11		12		13		14		15		16		17		18		19		20	
31	0.531	2.248	0.475	2.367	0.422	2.487	0.371	2.609	0.322	2.730	0.277	2.851	0.234	2.970	0.193	3.087	0.160	3.201	0.128	3.311
32	0.558	2.216	0.503	2.330	0.450	2.446	0.399	2.563	0.350	2.680	0.304	2.797	0.261	2.912	0.221	3.026	0.184	3.137	0.151	3.246
33	0.585	2.187	0.530	2.296	0.477	2.408	0.426	2.520	0.377	2.633	0.331	2.746	0.287	2.858	0.246	2.969	0.209	3.078	0.174	3.184
34	0.610	2.160	0.556	2.266	0.503	2.373	0.452	2.481	0.404	2.590	0.357	2.699	0.313	2.808	0.272	2.915	0.233	3.022	0.197	3.126
35	0.634	2.136	0.581	2.237	0.529	2.340	0.478	2.444	0.430	2.550	0.383	2.655	0.339	2.761	0.297	2.865	0.257	2.969	0.221	3.071
36	0.658	2.113	0.605	2.210	0.554	2.310	0.504	2.410	0.455	2.512	0.409	2.614	0.364	2.717	0.322	2.818	0.282	2.919	0.244	3.019
37	0.680	2.092	0.628	2.186	0.578	2.282	0.528	2.379	0.480	2.477	0.434	2.576	0.389	2.675	0.347	2.774	0.306	2.872	0.268	2.969
38	0.702	2.073	0.651	2.164	0.601	2.256	0.552	2.350	0.504	2.445	0.458	2.540	0.414	2.637	0.371	2.733	0.330	2.828	0.291	2.923
39	0.723	2.055	0.673	2.143	0.623	2.232	0.575	2.323	0.528	2.414	0.482	2.507	0.438	2.600	0.395	2.694	0.354	2.787	0.315	2.879
40	0.744	2.039	0.694	2.123	0.645	2.210	0.597	2.297	0.551	2.386	0.505	2.476	0.461	2.566	0.418	2.657	0.377	2.748	0.338	2.838
45	0.835	1.972	0.790	2.044	0.744	2.118	0.700	2.193	0.655	2.269	0.612	2.346	0.570	2.424	0.528	2.503	0.488	2.582	0.448	2.661
50	0.913	1.925	0.871	1.987	0.829	2.051	0.787	2.116	0.746	2.182	0.705	2.250	0.665	2.318	0.625	2.387	0.586	2.456	0.548	2.526
55	0.979	1.891	0.940	1.945	0.902	2.002	0.863	2.059	0.825	2.117	0.786	2.176	0.748	2.237	0.711	2.298	0.674	2.359	0.637	2.421
60	1.037	1.865	1.001	1.914	0.965	1.964	0.929	2.015	0.893	2.067	0.857	2.120	0.822	2.173	0.786	2.227	0.751	2.283	0.716	2.338
65	1.087	1.845	1.053	1.889	1.020	1.934	0.986	1.980	0.953	2.027	0.919	2.075	0.886	2.123	0.852	2.172	0.819	2.221	0.789	2.272
70	1.131	1.831	1.099	1.870	1.068	1.911	1.037	1.953	1.005	1.995	0.974	2.038	0.943	2.082	0.911	2.127	0.880	2.172	0.849	2.217
75	1.170	1.819	1.141	1.856	1.111	1.893	1.082	1.931	1.052	1.970	1.023	2.009	0.993	2.049	0.964	2.090	0.934	2.131	0.905	2.172
80	1.205	1.810	1.177	1.844	1.150	1.878	1.122	1.913	1.094	1.949	1.066	1.984	1.039	2.022	1.011	2.059	0.983	2.097	0.955	2.135
85	1.236	1.803	1.210	1.834	1.184	1.866	1.158	1.898	1.132	1.931	1.106	1.965	1.080	1.999	1.053	2.033	1.027	2.068	1.000	2.104
90	1.264	1.798	1.240	1.827	1.215	1.856	1.191	1.886	1.166	1.917	1.141	1.948	1.116	1.979	1.091	2.012	1.066	2.044	1.041	2.077
95	1.290	1.793	1.267	1.821	1.244	1.848	1.221	1.876	1.197	1.905	1.174	1.943	1.150	1.963	1.126	1.993	1.102	2.023	1.079	2.054
100	1.314	1.790	1.292	1.816	1.270	1.841	1.248	1.868	1.225	1.895	1.203	1.922	1.181	1.949	1.158	1.977	1.136	2.006	1.113	2.034
150	1.473	1.783	1.458	1.799	1.444	1.814	1.429	1.830	1.414	1.847	1.400	1.863	1.385	1.880	1.370	1.897	1.355	1.913	1.340	1.931
200	1.561	1.791	1.550	1.801	1.539	1.813	1.528	1.824	1.518	1.836	1.507	1.847	1.495	1.860	1.484	1.871	1.474	1.883	1.462	1.896

Πίνακας Π5 Durbin Watson $\alpha = 5\%$ $και$ k =αριθμός των ερμηνευτικών μεταβλητών

n\k	1		2		3		4		5		6		7		8		9		10	
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.796	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.079	1.891	1.015	1.978	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.876	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.197
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.838	1.369	1.874	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.747	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.846	1.608	1.862	1.593	1.877
200	1.758	1.779	1.748	1.789	1.738	1.799	1.728	1.809	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

n\k	1		2		3		4		5		6		7		8		9		10	
6	0.610	1.400																		
7	0.700	1.356	0.467	1.896																
8	0.763	1.332	0.559	1.777	0.367	2.287														
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588												
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822										
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004								
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149						
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266				
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360		
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.750	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.013	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.959	0.874	2.071	0.798	2.188	0.723	2.309	0.649	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.681	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363

n\k	11		12		13		14		15		16		17		18		19		20	
6																				
7																				
8																				
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16	0.060	3.446																		
17	0.084	3.286	0.053	3.506																
18	0.113	3.146	0.075	3.358	0.047	3.557														
19	0.145	3.023	0.102	3.227	0.067	3.420	0.043	3.601												
20	0.178	2.914	0.131	3.109	0.092	3.297	0.061	3.474	0.038	3.639										
21	0.212	2.817	0.162	3.004	0.119	3.185	0.084	3.358	0.055	3.521	0.035	3.671								
22	0.246	2.729	0.194	2.909	0.148	3.084	0.109	3.252	0.077	3.412	0.050	3.562	0.032	3.700						
23	0.281	2.651	0.227	2.822	0.178	2.991	0.136	3.155	0.100	3.311	0.070	3.459	0.046	3.597	0.029	3.725				
24	0.315	2.580	0.260	2.744	0.209	2.906	0.165	3.065	0.125	3.218	0.092	3.363	0.065	3.501	0.043	3.629	0.027	3.747		
25	0.348	2.517	0.292	2.674	0.240	2.829	0.194	2.982	0.152	3.131	0.116	3.274	0.085	3.410	0.060	3.538	0.039	3.657	0.025	3.766
26	0.381	2.460	0.324	2.610	0.272	2.758	0.224	2.906	0.180	3.050	0.141	3.191	0.107	3.325	0.079	3.452	0.055	3.572	0.036	3.682
27	0.413	2.409	0.356	2.552	0.303	2.694	0.253	2.836	0.208	2.976	0.167	3.113	0.131	3.245	0.100	3.371	0.073	3.490	0.051	3.602
28	0.444	2.363	0.387	2.499	0.333	2.635	0.283	2.772	0.237	2.907	0.194	3.040	0.156	3.169	0.122	3.294	0.093	3.412	0.068	3.524
29	0.474	2.321	0.417	2.451	0.363	2.582	0.313	2.713	0.266	2.843	0.222	2.972	0.182	3.098	0.146	3.220	0.114	3.338	0.087	3.450
30	0.503	2.283	0.447	2.407	0.393	2.533	0.342	2.659	0.294	2.785	0.249	2.909	0.208	3.032	0.171	3.152	0.137	3.267	0.107	3.379
n\k	11		12		13		14		15		16		17		18		19		20	
31	0.531	2.248	0.475	2.367	0.422	2.487	0.371	2.609	0.322	2.730	0.277	2.851	0.234	2.970	0.193	3.087	0.160	3.201	0.128	3.311
32	0.558	2.216	0.503	2.330	0.450	2.446	0.399	2.563	0.350	2.680	0.304	2.797	0.261	2.912	0.221	3.026	0.184	3.137	0.151	3.246
33	0.585	2.187	0.530	2.296	0.477	2.408	0.426	2.520	0.377	2.633	0.331	2.746	0.287	2.858	0.246	2.969	0.209	3.078	0.174	3.184
34	0.610	2.160	0.556	2.266	0.503	2.373	0.452	2.481	0.404	2.590	0.357	2.699	0.313	2.808	0.272	2.915	0.233	3.022	0.197	3.126
35	0.634	2.136	0.581	2.237	0.529	2.340	0.478	2.444	0.430	2.550	0.383	2.655	0.339	2.761	0.297	2.865	0.257	2.969	0.221	3.071
36	0.658	2.113	0.605	2.210	0.554	2.310	0.504	2.410	0.455	2.512	0.409	2.614	0.364	2.717	0.322	2.818	0.282	2.919	0.244	3.019
37	0.680	2.092	0.628	2.186	0.578	2.282	0.528	2.379	0.480	2.477	0.434	2.576	0.389	2.675	0.347	2.774	0.306	2.872	0.268	2.969
38	0.702	2.073	0.651	2.164	0.601	2.256	0.552	2.350	0.504	2.445	0.458	2.540	0.414	2.637	0.371	2.733	0.330	2.828	0.291	2.923
39	0.723	2.055	0.673	2.143	0.623	2.232	0.575	2.323	0.528	2.414	0.482	2.507	0.438	2.600	0.395	2.694	0.354	2.787	0.315	2.879
40	0.744	2.039	0.694	2.123	0.645	2.210	0.597	2.297	0.551	2.386	0.505	2.476	0.461	2.566	0.418	2.657	0.377	2.748	0.338	2.838
45	0.835	1.972	0.790	2.044	0.744	2.118	0.700	2.193	0.655	2.269	0.612	2.346	0.570	2.424	0.528	2.503	0.488	2.582	0.448	2.661
50	0.913	1.925	0.871	1.987	0.829	2.051	0.787	2.116	0.746	2.182	0.705	2.250	0.665	2.318	0.625	2.387	0.586	2.456	0.548	2.526
55	0.979	1.891	0.940	1.945	0.902	2.002	0.863	2.059	0.825	2.117	0.786	2.176	0.748	2.237	0.711	2.298	0.674	2.359	0.637	2.421
60	1.037	1.865	1.001	1.914	0.965	1.964	0.929	2.015	0.893	2.067	0.857	2.120	0.822	2.173	0.786	2.227	0.751	2.283	0.716	2.338
65	1.087	1.845	1.053	1.889	1.020	1.934	0.986	1.980	0.953	2.027	0.919	2.075	0.886	2.123	0.852	2.172	0.819	2.221	0.789	2.272
70	1.131	1.831	1.099	1.870	1.068	1.911	1.037	1.953	1.005	1.995	0.974	2.038	0.943	2.082	0.911	2.127	0.880	2.172	0.849	2.217
75	1.170	1.819	1.141	1.856	1.111	1.893	1.082	1.931	1.052	1.970	1.023	2.009	0.993	2.049	0.964	2.090	0.934	2.131	0.905	2.172
80	1.205	1.810	1.177	1.844	1.150	1.878	1.122	1.913	1.094	1.949	1.066	1.984	1.039	2.022	1.011	2.059	0.983	2.097	0.955	2.135
85	1.236	1.803	1.210	1.834	1.184	1.866	1.158	1.898	1.132	1.931	1.106	1.965	1.080	1.999	1.053	2.033	1.027	2.068	1.000	2.104
90	1.264	1.798	1.240	1.827	1.215	1.856	1.191	1.886	1.166	1.917	1.141	1.948	1.116	1.979	1.091	2.012	1.066	2.044	1.041	2.077
95	1.290	1.793	1.267	1.821	1.244	1.848	1.221	1.876	1.197	1.905	1.174	1.943	1.150	1.963	1.126	1.993	1.102	2.023	1.079	2.054
100	1.314	1.790	1.292	1.816	1.270	1.841	1.248	1.868	1.225	1.895	1.203	1.922	1.181	1.949	1.158	1.977	1.136	2.006	1.113	2.034
150	1.473	1.783	1.458	1.799	1.444	1.814	1.429	1.830	1.414	1.847	1.400	1.863	1.385	1.880	1.370	1.897	1.355	1.913	1.340	1.931
200	1.561	1.791	1.550	1.801	1.539	1.813	1.528	1.824	1.518	1.836	1.507	1.847	1.495	1.860	1.484	1.871	1.474	1.883	1.462	1.896