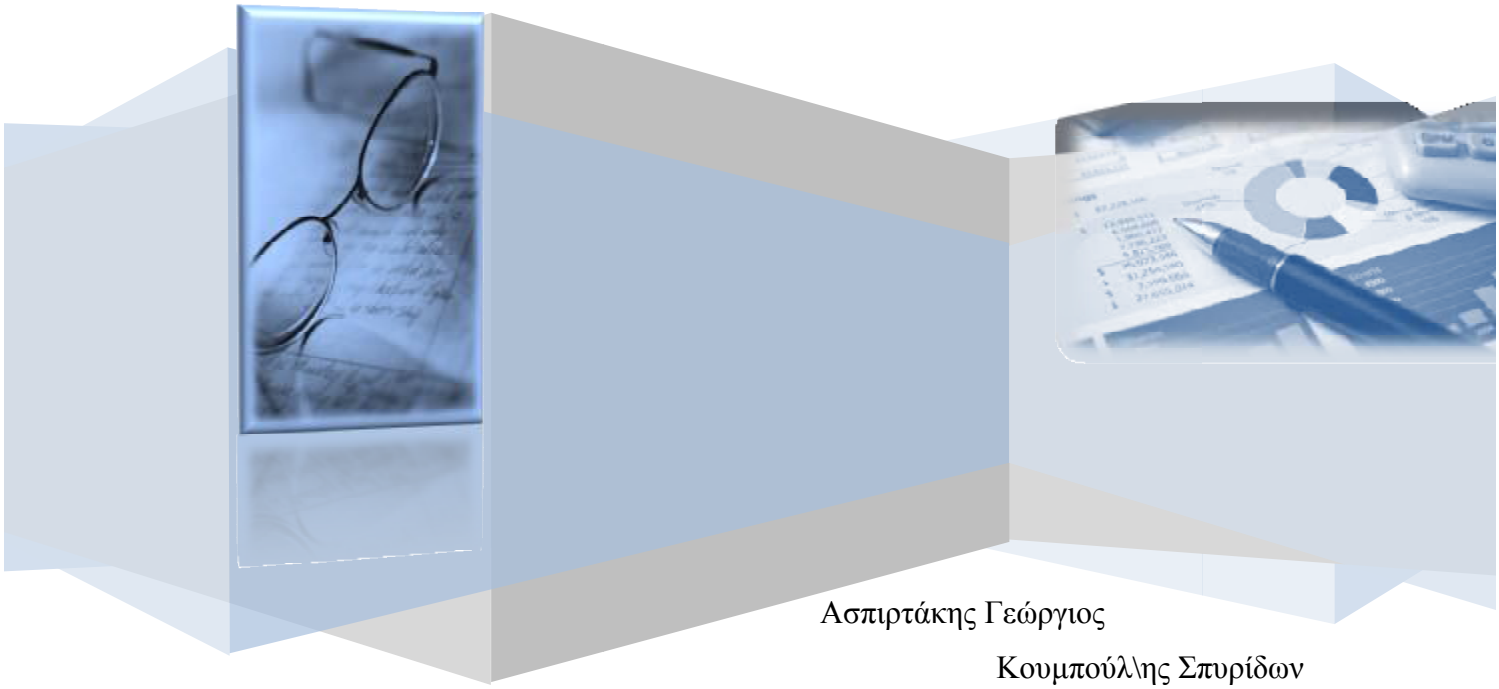
ΤΕΙ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

Σχολή Διοίκησης και Οικονομίας

Τμήμα Διοίκηση Επιχειρήσεων

Απλή Γραμμική Παλινδρόμηση και εφαρμογές της

Πτυχιακή Εργασία



Ασπιρτάκης Γεώργιος

Κουμπούλ\ης Σπυρίδων

Πετράκη Ελευθερία

Πάτρα 2016

Περίληψη

Είναι κοινά αποδεκτό πως το στατιστικό μοντέλο αποτελεί μία τυποποίηση στοχαστικών σχέσεων μεταξύ μεταβλητών, όπως αυτές παρουσιάζονται με τη μορφή μαθηματικών σχέσεων και εξισώσεων. Ο απώτερος στόχος από μια στατιστική ανάλυση είναι να επιτυγχάνεται η πιο ακριβή περιγραφή ενός συστήματος (φαινομένου ή γεγονότος). Σχεδόν σε κάθε φυσικό ή τεχνητό σύστημα, υπάρχουν μεταβλητές των οποίων οι ποσότητες συνεχώς αλλάζουν. Ανέκαθεν εύλογο ήταν εκείνο το ερώτημα για την μελέτη του βαθμού επίδρασης που οι περισσότερες ενεργούν και με ποιόν τρόπο πάνω σε άλλες. Η μελέτη αυτή είναι το αντικείμενο της ανάλυσης παλινδρόμησης, μίας ευρέως χρησιμοποιούμενης στατιστικής τεχνικής, η οποία χρησιμοποιείται για να την δημιουργία των κατάλληλων μοντέλων σχέσεων και εξαρτήσεων μεταξύ μεταβλητών. Τα διάφορα στατιστικά μοντέλα παλινδρόμησης, βασίζονται σε κάποιες βασικές υποθέσεις, τις οποίες οι ερευνητές υποχρεούνται να ελέγχουν πριν την ανάλυση του μοντέλου, δημιουργώντας τις κατάλληλες συνθήκες για την επίλυση τους. Η πεποίθηση του ερευνητή ότι δεν υπάρχει διαφορά μεταξύ των δύο πληθυσμών ή μεταξύ ενός δείγματος που έχει επιλεγεί από έναν πληθυσμό αξιολογείται από την εφαρμογή ενός ελέγχου υπόθεσης. Εντούτοις, οι υποθέσεις αυτές συχνά παραβιάζονται και ειδικότερα όταν τα δεδομένα συλλέγονται από τον πραγματικό κόσμο ενώ υπάρχουν και εξωγενείς παράγοντες που δεν μπορούν να συνεκτιμηθούν.

Στην παρούσα διπλωματική εργασία γίνεται μια συνολική προσπάθεια αποτύπωσης των τεχνικών ανάλυσης της παλινδρόμησης, ξεκινώντας από την παρουσίαση των βασικών εννοιών των μεταβλητών και των μεταξύ τους συσχετίσεων και καταλήγοντας στην ερμηνεία των αποτελεσμάτων που προκύπτουν από αυτήν. Προκειμένου να επιτευχθεί αυτό, παρουσιάζονται παράλληλα και κατάλληλα παραδείγματα που βοηθούν στην κατανόηση των ανωτέρω εννοιών. Επιπρόσθετα, καταγράφονται και οι τρόποι αντιμετώπισης στις περιπτώσεις αδυναμίας επιλογής του κατάλληλου μοντέλου καθώς και την τεκμηρίωση του τρόπου αντιμετώπισης των προβληματικών καταστάσεων. Είναι σημαντικό να αναφερθεί πως για κάποιον που αναζητά να αποκτήσει μια ολοκληρωμένη εικόνα για τα στατιστικά μοντέλα, είναι αναγκαίο να τα εξετάζει πάντα στον πραγματικό κόσμο για να έχει την δυνατότητα να μειώνει τις τυχόν παραβιάσεις των κανόνων στο μοντέλο, και παράλληλα να προβλέπει το βαθμό στατιστικής σημαντικότητας.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ ΠΡΩΤΟ
ΕΙΣΑΓΩΓΗ

1.1 Η Στατιστική και οι Κατηγοριοποιήσεις της

Η στατιστική αποτελεί ένα ισχυρό εργαλείο στην υπηρεσία οποιασδήποτε επιστήμης παρέχοντας αμέτρητες δυνατότητες, όσον αφορά στον προσδιορισμό της μεταβλητότητας, στην αντιμετώπιση, στην πρόβλεψη, στο σχεδιασμό και λήψη αποφάσεων (Σιώμκος, 2005). Είναι κοινά αποδεκτό, πως τις τελευταίες δεκαετίες η στατιστική έχει αναλάβει καταλυτικό ρόλο σε όλους σχεδόν τους τομείς της ανθρώπινης δραστηριότητας (Λαζαρίδης & Lazaridou, 2008). Επιπρόσθετα, ως εξελισσόμενη επιστήμη λαμβάνει και ανάλογα την αντίστοιχη ονομασία για κάθε ερευνητικό αντικείμενο με το οποίο ασχολείται, όπως για παράδειγμα «κοινωνική στατιστική», «στατιστική επιχειρήσεων» και «οικονομική στατιστική» (Δημητριάδης, 2002).

Γενικά, ο όρος «Στατιστική» σημαίνει συστηματική απαρίθμηση και παρουσίαση αριθμητικών δεδομένων ή στοιχείων, τα οποία προέρχονται από πολλές μετρήσεις ή παρατηρήσεις. Παράλληλα μέσα από αυτήν, προσδιορίζονται οι διαδικασίες συλλογής και περιγραφής δεδομένων, αξιολόγησης γεγονότων και πιθανοτήτων με απώτερο σκοπό τη δημιουργία τελικών αναλύσεων χρήσιμων στην λήψη αποφάσεων (Δημητριάδης, 2002). Με άλλα λόγια, η στατιστική είναι η επιστήμη που έχει σαν σκοπό τη συλλογή και την ανάλυση ποσοτικών και ποιοτικών δεδομένων για την εξαγωγή συμπερασμάτων (Λαζαρίδης & Lazaridou, 2008). Στην βάση αυτή, δημιουργούνται και τα απαραίτητα στατιστικά μοντέλα που αφορούν τη θεμελίωση σχέσεων μεταξύ των εμπλεκόμενων μεταβλητών για την παρατηρούμενη συμπεριφορά των δεδομένων της έρευνας που πραγματοποιείται (Ζαχαροπούλου, 1998) και μέσω των απαιτούμενων πληροφοριών που δόθηκαν από ένα δείγμα του πληθυσμού (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα, 1998).

Συνεπώς, ανάλογα με το βαθμό ανάλυσης των δεδομένων η στατιστική διαχωρίζεται σε τρία στάδια: α) την περιγραφική στατιστική, β) την εκτιμητική στατιστική, και γ) την συμπερασματολογία (Λαζαρίδης & Lazaridou, 2008). Η

περιγραφική στατιστική αφορά τη συλλογή μεγάλου όγκου ποσοτικών ή και ποιοτικών δεδομένων όπου μέσω της απλής επεξεργασίας τους, εξάγονται κυρίως γραφικές παραστάσεις (Δημητριάδης, 2002). Χαρακτηριστικό παράδειγμα της περιγραφικής στατιστικής είναι η αποτύπωση των αποτελεσμάτων των εθνικών εκλογών, όπου πληθυσμιακές μονάδες ερωτώνται για την επιθυμητή κυβέρνηση και από την συλλογή των δεδομένων γίνεται συνολική παρουσίαση των αποτελεσμάτων μέσω στατιστικών πινάκων, διαγραμμάτων και γραφημάτων.

Αναφορικά με την εκτιμητική στατιστική αυτή αποτελεί επέκταση της περιγραφικής, για την οποία ο ερευνητής κάνοντας χρήση των δεδομένων από ένα υποσύνολο του γενικού πληθυσμού δηλ. ένα δείγμα του πληθυσμού, «εκτιμά» την επίδραση των αποτελεσμάτων σε ολόκληρο τον πληθυσμό. Χαρακτηριστικό παράδειγμα εκτιμητικής στατιστικής είναι και οι προεκλογικές προβλέψεις που δύναται να αποτελέσουν εκτιμήσεις για τις εθνικές εκλογές, όταν γίνει η κατάλληλη επεξεργασία των απαντήσεων από ένα δείγμα του πληθυσμού.

Τέλος, η συμπερασματολογία αφορά τους ελέγχους των διάφορων στατιστικών στοιχείων (μεταβλητών) που προκύπτουν από το πέρας της εκτιμητικής στατιστικής (δηλαδή την μελέτη του/των δείγματος/ων του πληθυσμού), και αποσκοπεί στην εύρεση πιθανών σχέσεων αιτίας και αιτιατού μεταξύ αυτών (Λαζαρίδης & Lazaridou, 2008). Αναφορικά με το προηγούμενο παράδειγμα των εθνικών εκλογών, μέσω της συμπερασματολογίας οι ερευνητές προσπαθούν να βρουν σχέσεις μεταξύ των μεταβλητών επιρροής στην διαδικασία λήψης απόφασης ψήφου των πολιτών και την τελική εκλογική τους συμπεριφορά.

Να επισημανθεί πως σύμφωνα με τους Κλολυβά-Μαχαιρά και Μπόρα-Σέντα (1998) καθώς και τον Δημητριάδη (2002), η στατιστική χωρίζεται σε δύο μεγάλες κατηγορίες: την περιγραφική στατιστική, και την επαγωγική (ή αναλυτική) στατιστική, η οποία μπορεί να θεωρηθεί πως αφορά στα δυο τελευταία στάδια στατιστικής που αναφέρθηκαν παραπάνω. Είναι σαφές πως η αναλυτική-συμπερασματολογική στατιστική είναι απαραίτητη, εφόσον καμία από τις δύο κατηγοριοποιήσεις δεν αναιρείται από την ύπαρξη άλλης και η περιγραφική στατιστική χαρακτηρίζει κυρίως το ερευνητικό περιβάλλον του 17^{ου} αιώνα για τη δημιουργία έγκυρων στατιστικών μοντέλων (Δημητριάδης 2002).

Επιπλέον να αναφερθεί πως κατά την Ζαχαροπούλου (1998) όλες οι στατιστικές τεχνικές που στοχεύουν στη δημιουργία ενός αξιόπιστου στατιστικού μοντέλου σχετίζονται με τον όρο «Ανάλυση Παλινδρόμησης» και εμπεριέχουν δύο

ομάδες προβλημάτων: την Στατιστική Εκτίμηση *(statistical estimation),* και τον Έλεγχο Στατιστικών υποθέσεων *(test of hypotheses)* (Δημητριάδης 2002).

1.2 Η Έννοια της Παλινδρόμησης

Η παλινδρόμηση είναι εκείνη η στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών (Κόκκινος, 2011). Στην ελληνική βιβλιογραφία ο όρος παλινδρόμηση ορίζεται ως η «συναγωγή σχέσεων από τα δεδομένα» (Ζαχαροπούλου 1998). Τον 19^{ο} αιώνα ο Francis Galton αναφέρει στα συγγράμματα του τον όρο παλινδρόμηση για την περιγραφή ενός βιολογικού φαινομένου. Συγκεκριμένα η περιγραφή του φαινομένου αφορούσε το ύψος των απογόνων των ψηλών προγόνων τείνουν να υποχωρούν προς το φυσιολογικό μέσο όρο, φαινόμενο γνωστό και ως «*οπισθοδρόμηση προς το μέσο*» (Galton, 1885). Αν και για τον Galton ο όρος παλινδρόμηση είχε μόνο αυτή τη βιολογική έννοια, αργότερα το έργο του εξετάστηκε και συμπληρώθηκε από τους Yule και Pearson (1903) σε ένα πιο γενικό στατιστικό πλαίσιο για να εφαρμοστεί σε ένα φαινόμενο που παρουσιάζεται συχνά κατά τη διάρκεια ερευνών που αφορούσε την εμφάνιση σχέσεων αλληλεπίδρασης και αλληλεξάρτησης ανάμεσα σε (δύο ή και περισσότερες) συγκεκριμένες μεταβλητές.

Αυτή η σχέση αλληλεπίδρασης και αλληλεξάρτησης των μεταβλητών σε ένα ερευνητικό/στατιστικό πλαίσιο ονομάζεται «πρόβλημα παλινδρόμησης» (Λαζαρίδης & Lazaridou, 2008). Για παράδειγμα αν μια φαρμακευτική εταιρία θέλει να ερευνήσει την επιρροή ενός φαρμάκου στην αρτηριακή πίεση των ατόμων που το λαμβάνουν, χορηγεί σε ομάδες ανθρώπων με παρόμοια στατιστικά χαρακτηριστικά το ίδιο φάρμακο αλλά σε διαφορετικές δοσολογίες. Αν τα αποτελέσματα που εξαχθούν δείξουν πως η δοσολογία του φαρμάκου (ανεξάρτητη μεταβλητή) επηρεάζει το επίπεδο πίεσης των ανθρώπων (εξαρτημένη μεταβλητή), τότε υπάρχει το λεγόμενο πρόβλημα παλινδρόμησης δηλαδή του βαθμού επίδρασης της μεταβλητής «επίπεδο αρτηριακής πίεσης» επί της μεταβλητής «δοσολογία φαρμάκου».

Συνεπώς για την ανάδειξη της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής και ενός ή περισσότερων ανεξάρτητων μεταβλητών, η μέθοδος της ανάλυσης παλινδρόμησης συμπεριλαμβανομένης και των τεχνικών για την μοντελοποίηση και την ανάλυση των μεταβλητών, βοηθά τον ερευνητή να εξετάσει την επίδραση της εξαρτημένης μεταβλητής όταν κάποιες από τις ανεξάρτητες μεταβλητές παραμένουν σταθερές ενώ άλλες μεταβάλλονται.

1.3 Η ανάλυση της Παλινδρόμησης

Η ανάλυση παλινδρόμησης είναι μια διαδικασία που εφαρμόζεται για την ανάλυση των σχέσεων μεταξύ μιας εξαρτημένης μεταβλητής και μιας η περισσοτέρων ανεξάρτητων μεταβλητών (Σιώμκος, 2005). Έτσι, χρησιμοποιείται για να κατανοηθεί ποιες από τις ανεξάρτητες μεταβλητές αλληλοσυσχετίζονται με την εξαρτημένη μεταβλητή, να διερευνηθεί η ποικιλομορφία των πιθανών σχέσεων, και σε περιορισμένες περιπτώσεις, να συναχθούν αιτιώδεις σχέσεις μεταξύ της ανεξάρτητης και εξαρτημένης μεταβλητής.

Για τη διενέργεια της ανάλυσης παλινδρόμησης έχει αναπτυχθεί ένα μεγάλο πεδίο τεχνικών το οποίο συμπεριλαμβάνει και τις παραμετρικές και τις μη-παραμετρικές μεθόδους. Στις παραμετρικές μεθόδους, όπως η γραμμική παλινδρόμηση και η απλή παλινδρόμηση ελαχίστων τετραγώνων, η συνάρτηση της παλινδρόμησης ορίζεται από έναν πεπερασμένο αριθμό άγνωστων παραμέτρων που εκτιμώνται από τα δεδομένα (Freedman, 2005). Αντίθετα, η μη-παραμετρική παλινδρόμηση αναφέρεται σε τεχνικές που επιτρέπουν τη λειτουργία της παλινδρόμησης να εκτείνεται σε ένα καθορισμένο σύνολο λειτουργιών, οι οποίες όμως μπορεί να είναι απείρων διαστάσεων (Freedman, 2005).

Στην πραγματικότητα, η επίδοση των μεθόδων ανάλυσης παλινδρόμησης εξαρτάται από τη μορφή της διαδικασίας παραγωγής δεδομένων, και από το πως αυτή σχετίζεται με την προσέγγιση της παλινδρόμησης που χρησιμοποιείται. Δεδομένου ότι η αληθινή μορφή της διαδικασίας παραγωγής δεδομένων είναι σε γενικές γραμμές άγνωστη, συχνά η ανάλυση παλινδρόμησης εξαρτάται σε κάποιο βαθμό από την δημιουργία υποθέσεων γύρω από τη διαδικασία αυτή. Αυτές οι υποθέσεις τις περισσότερες φορές είναι ελέγξιμες, αν και αυτό ισχύει μόνο εφόσον υπάρχει διαθέσιμη μεγάλη ποσότητα πρωτογενών δεδομένων (Mogul, 2005).

Έχει ενδιαφέρον να σημειωθεί πως τα μοντέλα παλινδρόμησης θεωρούνται χρήσιμα εργαλεία πρόβλεψης αν και τις περισσότερες φορές δεν μπορούν να εκτελεστούν άριστα στις περιπτώσεις όπου οι υποθέσεις είναι μη ελέγξιμες. Τις τελευταίες δεκαετίες, νέες μέθοδοι παλινδρόμησης έχουν αναπτυχθεί για καλύτερα αποτελέσματα όπως τεχνικές μη παραμετρικής παλινδρόμησης, η μέθοδος Bayesian, μέθοδοι παλινδρόμησης στις οποίες οι μεταβλητές πρόβλεψης μετρώνται με σφάλμα, καθώς και μοντέλα παλινδρόμησης με περισσότερες μεταβλητές πρόβλεψης από παρατηρήσεις για τη δημιουργία αιτιωδών συμπερασμάτων (Good & Hardin, 2009). Ωστόσο, σε κάποιες εφαρμογές, όπως στις μικρές επιπτώσεις (small effects) ή στις

ερωτήσεις αιτιότητας που βασίζονται σε δεδομένα παρατηρήσεων, οι τεχνικές παλινδρόμησης μπορεί να δώσουν παραπλανητικά ή αποτελέσματα παρεμηνείας. Συνεπώς οι τεχνικές και μέθοδοι παλινδρόμησης συνεχίζουν να αποτελούν ένα ακμαίο πεδίο δραστήριας έρευνας (Σιώμκος, 2005).

1.4 Μοντέλα και Μεταβλητές Παλινδρόμησης

Τα μοντέλα παλινδρόμησης περιλαμβάνουν τρεις κύριες κατηγορίες μεταβλητών που είναι: α) οι άγνωστες παράμετροι συσχέτισης που συμβολίζονται με ß (διάνυσμα), β) οι ανεξάρτητες μεταβλητές που συμβολίζονται με X (διάνυσμα), και γ) η εξαρτώμενη μεταβλητή που συμβολίζονται με Y (Ravishankar & Dey, 2002). Αν και σε διαφορετικά πεδία εφαρμογής μπορεί να χρησιμοποιούνται διαφορετικοί συμβολισμοί για την εξαρτημένη και τις ανεξάρτητες μεταβλητές, η ιδέα πίσω από όλα τα μοντέλα παλινδρόμησης αφορά την εξαρτημένη μεταβλητή σε συνάρτηση με τις ανεξάρτητες μεταβλητές και τις άγνωστες παραμέτρους ß, και προσεγγίζεται μέσω της τυποποιημένης αριθμητικής παράστασης E(Y | X) = *f*(X, ß). Η προηγούμενη μαθηματική σχέση αποτελεί ένα μοντέλο παλινδρόμησης που συσχετίζει το Y σε μία συνάρτηση παλινδρόμησης (***regression)*** των X και ß.

Η *f* είναι η συνάρτηση παλινδρόμησης της Y επί των $X_1,...,X_k$ (Ζαχαροπούλου, 1998) και για τη διεξαγωγή ανάλυσης παλινδρόμησης ο λειτουργικός τύπος της *f* πρέπει να διευκρινίζεται εξ' αρχής. Μερικές φορές η μορφή της λειτουργίας της *f* βασίζεται σε προϋπάρχουσα γνώση για τη σχέση μεταξύ της Y και της X που δεν βασίζεται στα δεδομένα, όταν όμως δεν υπάρχει διαθέσιμη τέτοια γνώση, επιλέγεται μια πιο ευέλικτη ή κατάλληλη μορφή για την *f* η οποία καθορίζεται από το περιεχόμενο της κάθε έρευνας. Για παράδειγμα, σύμφωνα με τους Kutner et. al. (2004), αν υποτεθεί πως το διάνυσμα των αγνώστων παραμέτρων ß είναι μήκους ***k,*** για την εκτέλεση μιας ανάλυσης παλινδρόμησης, ο χρήστης πρέπει να παρέχει πληροφορίες σχετικά με την εξαρτημένη μεταβλητή Y.

Εάν παρατηρηθούν N σημεία δεδομένων του εντύπου (Y, X), όπου N<k, οι κλασικότερες προσεγγίσεις στην ανάλυση παλινδρόμησης δεν μπορούν να εκτελεστούν, διότι δεν υπάρχουν αρκετά δεδομένα για να την ανάκτηση των ß στο σύστημα εξισώσεων του μοντέλου παλινδρόμησης. Εάν παρατηρηθούν ακριβώς N=k σημεία δεδομένων και η συνάρτηση *f* είναι γραμμική, η εξίσωση Y = *f* (X, ß) μπορεί να λυθεί επακριβώς. Αυτό μειώνει την επίλυση μιας σειράς N εξισώσεων με N αγνώστους (τα στοιχεία ß), η οποία έχει μια μοναδική λύση εφόσον οι μεταβλητές X

είναι γραμμικά ανεξάρτητες. Αν η f είναι μη γραμμική, τότε είτε δεν υπάρχει λύση, είτε μπορούν να υπάρξουν πολλές λύσεις. Παρόλο αυτά, στις περισσότερες καταστάσεις παρατηρούνται N>k σημεία δεδομένων, όπου υπάρχουν επαρκείς πληροφορίες στα δεδομένα για τον προσδιορισμό μιας μοναδικής τιμή για την ß που να ταιριάζει με τα δεδομένα και το μοντέλο παλινδρόμησης.

Αναφερόμενοι στην τελευταία περίπτωση, η ανάλυση παλινδρόμησης παρέχει τα εργαλεία για την εξεύρεση λύσης για τις άγνωστες παραμέτρους ß που, για παράδειγμα, θα ελαχιστοποιήσει την απόσταση μεταξύ των μετρούμενων και προβλεπόμενων τιμών της εξαρτημένης μεταβλητής Y (επίσης γνωστή ως μέθοδος των ελαχίστων τετραγώνων). Επιπλέον, σύμφωνα με ορισμένες στατιστικές υποθέσεις, η ανάλυση παλινδρόμησης χρησιμοποιεί το πλεόνασμα των πληροφοριών για να παρέχει στατιστικές πληροφορίες σχετικά με τις άγνωστες παραμέτρους ß και τις προβλεπόμενες τιμές της εξαρτημένης μεταβλητής Y. Τέλος, είναι σημαντικό να σημειωθεί πως στις περιπτώσεις όπου k>1 το μοντέλο παλινδρόμησης θεωρείται πολυμεταβλητό, ενώ αντίθετα όταν k=1 το μοντέλο παλινδρόμησης θεωρείται απλό (Ζαχαροπούλου, 1998).

1.4.1 Μελέτη Συμπεριφοράς μοντέλων παλινδρόμησης

Κατά την δημιουργία του μοντέλου παλινδρόμησης σημαντικό είναι να αποφανθεί κάποιος για την επιβεβαίωση της καλής προσαρμογής του μοντέλου και της στατιστικής σημαντικότητας των εκτιμώμενων παραμέτρων. Οι ανεξάρτητες μεταβλητές σχετίζονται με στο μοντέλο παλινδρόμησης. Εξάλλου αυτό είναι επιθυμητό έτσι ώστε να «εξηγηθεί» η εξαρτημένη μεταβλητή με βάση τις ανεξάρτητες (Σιώμκος, 2005). Εκτός από την συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτημένη, οι ανεξάρτητες μεταβλητές ίσως να συσχετίζονται και μεταξύ τους. Όταν οι ανεξάρτητες μεταβλητές αλληλοσυσχετίζονται, τότε υπάρχει η λεγόμενη πολυσυγγραμμικότητα (multicollinearity). Οι έλεγχοι προσαρμοστικότητας του μοντέλου καθώς και των συσχετίσεων που παρουσιάζονται συνήθως περιλαμβάνουν ανάλυση του R-squared, αναλύσεις του τρόπου διεξαγωγής των καταλοίπων (residuals) ή σφαλμάτων εκτίμησης που αποτελούν το μέρος των δεδομένων που δε μπορούν να επεξηγηθούν και εξαρτώνται από εξωγενείς ή άλλου είδους παράγοντες οι παραμέτρους λόγω της τυχαίας επιλογής της μεταβλητής Y (Ζαχαροπούλου, 1998), καθώς και με έλεγχο υποθέσεων. Επιπλέον, η στατιστική

σημαντικότητα μπορεί να ελεγχθεί αρχικά από ένα F-test για τη συνολική καταλληλότητα του μοντέλου και έπειτα από t-tests για τις επιμέρους παραμέτρους.

Οι ερμηνείες αυτών των διαγνωστικών τεστ βασίζονται σε μεγάλο βαθμό στις παραδοχές και υποθέσεις των μοντέλων. Αν και η εξέταση των καταλοίπων μπορεί να χρησιμοποιηθεί για να απορρίψει ένα μοντέλο, τα αποτελέσματα του t-test ή F-test είναι μερικές φορές πιο δύσκολο να ερμηνευτούν σε περιπτώσεις που παραβιάζονται οι υποθέσεις του μοντέλου. Για παράδειγμα, αν το περιθώριο σφάλματος δεν ακολουθεί κανονική κατανομή, σε μικρά δείγματα οι εκτιμώμενες παράμετροι δεν θα ακολουθήσουν τη συνήθη κατανομή και έτσι θα περιπλέξουν την εξαγωγή συμπερασμάτων. Ωστόσο, τα σχετικά μεγάλα δείγματα ακολουθώντας το κεντρικό οριακό θεώρημα επιτρέπουν να γίνεται έλεγχος των υποθέσεων χρησιμοποιώντας ασυμπτωτικές προσεγγίσεις (Fox, 2000).

1.5 Σύνοψη

Αναφέρθηκε, λοιπόν, ότι τα στατιστικά μοντέλα είναι κάθε ομάδα μαθηματικών και πιθανοθεωρητικών εξισώσεων που χρησιμοποιούνται για να περιγράψουν, να συνοψίσουν, και να ερμηνεύσουν καταστάσεις και φαινόμενα. Στο παρόν κεφάλαιο περιγράφεται συνοπτικά η χρήση της στατιστικής και των εννοιών που διέπουν από αυτήν ενώ αναφέρθηκαν γενικά οι έννοιες των γραμμικών μοντέλων, τα οποία είναι κανονικά και αναφέρονται, συνήθως, ως παλινδρομικά γραμμικά μοντέλα. Η στατιστική μοντελοποίηση αποτελεί ένα χρήσιμο μαθηματικό εργαλείο, το οποίο επιτρέπει στον ερευνητή μέσω μίας ενοποιημένης διαδικασίας, να απεικονίσει και να δώσει ερμηνεία σε απλές ή σύνθετες σχέσεις μεταξύ μεταβλητών που χαρακτηρίζουν καταστάσεις και φαινόμενα.

ΚΕΦΑΛΑΙΟ ΔΕΥΤΕΡΟ
ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ

2.1 Εισαγωγή

Οι χαρακτηριστικές ιδιότητες των στατιστικών μονάδων ενός πληθυσμού, με την μελέτη των οποίων ασχολείται η Στατιστική, ονομάζονται μεταβλητές. Οι αριθμοί ή οι άλλες συμβολικές εκφράσεις που αντιπροσωπεύουν τις διάφορες καταστάσεις μιας μεταβλητής ονομάζονται *τιμές μεταβλητής*. Ο συμβολισμός κάθε μεταβλητής γίνεται με την χρήση κεφαλαίων γραμμάτων X, Y, Z, ενώ οι τιμές αν η μεταβλητή είναι ποσοτική συμβολίζονται με αντίστοιχα μικρά γράμματα: $x_1,x_2,x_3,...,x_n$ ή $y_1,y_2,y_3,\ldots,y_k$. Να αναφερθεί σε αυτό το σημείο ότι οι μεταβλητές διακρίνονται σε *ποιοτικές* και *ποσοτικές* καθώς και *εξαρτημένες* και *ανεξάρτητες* (Κιόχος, 1993) .

- *Ποιοτικές* καλούνται οι μεταβλητές οι οποίες δεν επιδέχονται αριθμητική μέτρηση και επομένως οι τιμές τους εκφράζονται με λέξεις. Για παράδειγμα, ποιοτικές μεταβλητές αποτελούν το φύλο ενός ατόμου, η οικογενειακή κατάσταση ή το επάγγελμα ενός ατόμου.
- *Ποσοτικές μεταβλητές* χαρακτηρίζονται οι μεταβλητές εκείνες που επιδέχονται μέτρηση και οι τιμές τους είναι αριθμοί που αναφέρονται σε συγκεκριμένες μονάδες μέτρησης. Έτσι για παράδειγμα, ως ποσοτικές μεταβλητές μπορούμε να χαρακτηρίσουμε το ύψος ή το βάρος ενός μαθητή, την ηλικία ή το εισόδημα ενός εργαζόμενου, τη θερμοκρασία και το ύψος των εξαγωγών μιας χώρας.

Αναφορικά με τον διαχωρισμό μεταξύ *εξαρτημένων* και *ανεξάρτητων μεταβλητών* οι *ανεξάρτητες μεταβλητές* είναι εκείνες που μπορούν να αναλυθούν ενώ οι *εξαρτημένες μεταβλητές* είναι μόνο μετρήσιμες ή διαθέτουν την δυνατότητα καταχώρησης. Όπως πολλοί ερευνητές υποστηρίζουν η εν λόγω διάκριση εμφανίζεται ως ιδιαίτερη και ιδιάζουσα καθώς δημιουργεί σύγχυση σε πολλούς, δεδομένου του ισχυρισμού ότι όλες οι μεταβλητές εξαρτώνται από κάτι. Οι όροι εξαρτημένη και ανεξάρτητη μεταβλητή σχετίζονται συνήθως με την πειραματική έρευνα, όπου μερικές μεταβλητές κατόπιν ανάλυσης και επεξεργασίας δύναται να γίνουν ανεξάρτητες κατά τα αρχικά σχέδια αντίδρασης, τα χαρακτηριστικά γνωρίσματα, τις προθέσεις, κ.λπ. των θεμάτων. Μερικές άλλες μεταβλητές θεωρούνται εξαρτημένες

λόγω συγκεκριμένης επεξεργασία, το χειρισμό τους ή τις έκαστοτε πειραματικές συνθήκες, συχνό φαινόμενο που αναδεικνύεται μέσω της απάντησης που πρέπει να δοθεί στο ερώτημα «τι θα κάνει το υποκείμενο;».

Σε πειραματικούς σχεδιασμούς ο ερευνητής χειρίζεται (μεταβάλλει) τις τιμές της ανεξάρτητης μεταβλητής προκειμένου να προκαλέσει αλλαγές στην εξαρτημένη μεταβλητή. Ο ορισμός των μεταβλητών σε ανεξάρτητες και εξαρτημένες καθορίζεται από το θεωρητικό υπόβαθρο της έρευνας, και δεν αντικατοπτρίζει διακρίσεις περιεχομένου ή μορφής, αλλά έχει λειτουργικό ρόλο. Σε κάποιες έρευνες μία μεταβλητή μπορεί να αντιμετωπίζεται ως ανεξάρτητη, ενώ σε άλλη έρευνα να ορίζεται ως εξαρτημένη, και αντίστροφα. Για παράδειγμα, εάν σε ένα πείραμα, τα αρσενικά συγκρίνονται με τα θηλυκά σχετικά με τα χρωμοσώματα τους, το φύλο θα μπορούσε να κληθεί ως ανεξάρτητη μεταβλητή και το χρωμόσωμα ως εξαρτημένη μεταβλητή.

2.2 Πειραματική και μη Πειραματική Έρευνα

Η πειραματική έρευνα επιχειρεί να προσδιορίσει σχέση αιτίας και αποτελέσματος μεταξύ δύο ή περισσοτέρων μεταβλητών. Μέσα από την εν λόγω έρευνα επηρεάζονται κάποιες μεταβλητές και έπειτα ακολουθεί μέτρηση της επίδρασης του χειρισμού. Για παράδειγμα ανάλυση της συμπεριφοράς μικρών παιδιών τα οποία πρωτίστως έχουν παρακολουθήσει σκηνές βίας. Συνεπώς, στην πειραματική έρευνα ο/η ερευνητής/τρια διατυπώνει μία ή περισσότερες υποθέσεις (hypothesis) και μέσα από το πείραμα προσπαθεί να τις επαληθεύσει ή να τις απορρίψει. Ο έλεγχος των υποθέσεων ο/η ερευνητής το επιτυγχάνει μέσα από στατιστικές αναλύσεις.

Η ανάλυση στοιχείων στην πειραματική έρευνα οδηγεί στον υπολογισμό των συσχετισμών μεταξύ των μεταβλητών, συγκεκριμένα, εκείνων που χειρίζονται και εκείνων που επηρεάζονται από το χειρισμό. Εντούτοις, τα πειραματικά στοιχεία μπορούν ενδεχομένως να παρέχουν τις ποιοτικά καλύτερες πληροφορίες: Μόνο τα πειραματικά στοιχεία μπορούν αποφασιστικά να καταδείξουν τις αιτιώδεις σχέσεις μεταξύ των μεταβλητών. Παραδείγματος χάριν, εάν διαπιστωθεί ότι αλλαγές στις μεταβλητές Α επιφέρουν και μεταβολές στις μεταβλητές Β, κατόπιν προκύπτει το συμπέρασμα ότι το Α επηρεάζει το Β.

Η μη πειραματική έρευνα αποτελεί και αυτή είδος της ποσοτικής έρευνας και περιλαμβάνει τη Περιγραφική/Descriptive έρευνα, την έρευνα σύγκρισης/Casual-

comparative, την έρευνα συσχέτισης/Correlational), την έρευνα αντιστροφής, την έρευνα μη ισοδύναμων ομάδων και την έρευνα διαδοχικών μετρήσεων. Όπως είναι κατανοητό από τις παραπάνω η έρευνα συσχέτισης είναι εκείνη που αντιδρά καλύτερα στην σχέση μεταξύ των μεταβλητών. Έτσι μέσα από αυτήν διερευνάται καλύτερα οι σχέσεις που υπάρχουν μεταξύ δύο μεταβλητών. Για παράδειγμα πως συσχετίζεται η μεταβλητή, απόδοση των μαθητών, με τη μεταβλητή, μορφωτικό επίπεδο των γονέων τους Ο βαθμός σχέσης μεταξύ δύο μεταβλητών εκφράζεται με έναν συντελεστή συσχέτισης (*correlation coefficient*). Ως απώτερος στόχος της έρευνας συσχέτισης είναι να γίνουν προβλέψεις, ενώ η ύπαρξη σχέσης μεταξύ δύο μεταβλητών δεν σημαίνει και ύπαρξη σχέσης αιτίας – αποτελέσματος μεταξύ τους. Τα στοιχεία από τη συσχετιστική έρευνα μπορούν μόνο να ερμηνεύσουν τους αιτιώδεις όρους βασισμένους σε μερικές θεωρίες που έχουμε, αλλά τα συσχετιστικά στοιχεία δεν μπορούν αποφασιστικά να αποδείξουν την αιτιότητα.

2.3 Σχέσεις μεταξύ των μεταβλητών

Ανεξάρτητα από τον τύπο τους, δύο ή περισσότερες μεταβλητές συσχετίζονται εάν σε ένα δείγμα των παρατηρήσεων, οι τιμές εκείνων των μεταβλητών κατανέμονται κατά τρόπο συνεπή. Με άλλα λόγια, οι μεταβλητές συσχετίζονται εάν οι τιμές τους αντιστοιχούν συστηματικά η μια στην άλλη για αυτές τις παρατηρήσεις. Παραδείγματος χάριν, το ύψος συσχετίζεται με το βάρος επειδή χαρακτηριστικά τα ψηλά άτομα είναι βαρύτερα από τα κοντά, ενώ ο δείκτης νοημοσύνης συσχετίζεται με τον αριθμό λαθών σε μια δοκιμή, αφού οι άνθρωποι με τον υψηλότερο δείκτη νοημοσύνης κάνουν λιγότερα λάθη.

2.3.1 Ο βαθμός σημαντικότητας της σχέσης μεταξύ των μεταβλητών

Γενικά, ο απώτερος σκοπός κάθε ερευνητικής προσπάθειας ή επιστημονικής ανάλυσης είναι να ανακαλύπτει τις σχέσεις μεταξύ των μεταβλητών. Η φιλοσοφία της επιστήμης διδάσκει ότι δεν υπάρχει κανένας άλλος τρόπος για να αναζητηθεί η σημασία ενός φαινομένου εκτός από την συσχέτιση ποσοτικών και ποιοτικών χαρακτηριστικών, επομένως την αναζήτηση σχέσεων μεταξύ των μεταβλητών. Κατά συνέπεια, η πρόοδος της επιστήμης εξαρτάται από την ικανή εύρεση των νέων σχέσεων μεταξύ των μεταβλητών. Η έρευνα συσχέτισης περιλαμβάνει τη μέτρηση τέτοιων σχέσεων με τον απλούστερο τρόπο. Εντούτοις, η πειραματική έρευνα δεν είναι καθόλου διαφορετική από αυτή την άποψη. Στο προηγούμενο παράδειγμα όπου

συγκρίνονται χρωμοσώματα στα αρσενικά και τα θηλυκά μπορεί να προκύψει ένας ακόμη συσχετισμός μεταξύ δύο μεταβλητών: φύλο και χρωμοσώματα. Οι στατιστικές μελέτες έγκεινται στον αποτελούν αρωγούς για την αξιολόγηση των σχέσεων μεταξύ των μεταβλητών.

2.4 Χαρακτηριστικά γνωρίσματα μεταξύ των μεταβλητών

Οι βασικές ιδιότητες κάθε σχέσης μεταξύ των μεταβλητών είναι δύο και είναι το *μέγεθος* και η *αξιοπιστία* .

· *Μέγεθος:* Το μέγεθος είναι ευκολότερο να κατανοηθεί και να μετρηθεί σε σχέση με την αξιοπιστία. Για παράδειγμα, εάν κάθε αρσενικό στο δείγμα βρέθηκε να είναι ψηλότερο από οποιοδήποτε θηλυκό, θα μπορούσε να ειπωθεί ότι το μέγεθος της σχέσης μεταξύ των δύο μεταβλητών (φύλο και ύψος) είναι υψηλότερο στο δείγμα. Συνεπώς, θα υπήρχε κατάλληλη πρόβλεψη το ενός βασισμένου στο άλλο (τουλάχιστον μεταξύ των μελών του δείγματος).

· *Αξιοπιστία:* Η αξιοπιστία μιας σχέσης είναι μια πολύ λιγότερο διαισθητική έννοια, αλλά ακόμα εξαιρετικά σημαντική. Αναφέρεται στη "αντιπροσωπευτικότητα" του αποτελέσματος που βρίσκεται στο συγκεκριμένο δείγμα για ολόκληρο τον πληθυσμό. Με άλλα λόγια, λέει πόσο πιθανό είναι μια παρόμοια σχέση να βρισκόταν εάν το πείραμα ξαναγινόταν με άλλα δείγματα που προήλθαν από τον ίδιο πληθυσμό. Να γίνει κατανοητό ότι ο ερευνητής δεν ενδιαφέρεται σχεδόν ποτέ «τελικά» μόνο για αυτό που γίνεται στο δείγμα αλλά ενδιαφέρεται για το δείγμα μόνο στην έκταση που μπορεί να παρέχει τις πληροφορίες για τον πληθυσμό. Εάν η μελέτη ικανοποιεί μερικά ειδικά κριτήρια, η αξιοπιστία μιας σχέσης μεταξύ των μεταβλητών που παρατηρούνται στο δείγμα μπορεί να υπολογιστεί ποσοτικά και να αντιπροσωπευθεί χρησιμοποιώντας ένα τυποποιημένο μέτρο (τεχνικά αποκαλούμενο ως p-value ή στατιστικό επίπεδο σημαντικότητας).

2.5 Συσχέτιση (Correlation)

Η συσχέτιση μεταξύ δύο ή περισσότερων μεταβλητών μετρά το βαθμό συνάφειας-αλληλεπίδρασης ανάμεσα σε δύο ή περισσότερες μεταβλητές. Πρακτικά σημαίνει, ότι από την τιμή ενός δείκτη (συντελεστή συσχέτισης) κατανοείται πόσο έντονη ή χαλαρή είναι η συσχέτιση δύο μεταβλητών. Η διαδικασία συσχέτισης παρουσιάζεται όχι μόνο σε ποσοτικές μεταβλητές (συντελεστής Pearson) αλλά και σε ποιοτικές ή κατηγορικές μεταβλητές. Εδώ θα πρέπει να γίνει μια σαφής διάκριση. Το γεγονός της

ύπαρξης ή μη έντονης συνάφειας-συσχέτισης ανάμεσα σε δύο μεταβλητές, δεν συνεπάγεται απαραίτητα και την ύπαρξη μίας συναρτησιακής σχέσης μεταξύ αυτών.

Οι συντελεστές συσχέτισης που θα αναφερθούν χωρίζονται σε δύο κατηγορίες: η πρώτη εξεταζόμενη στο δεύτερο κεφάλαιο αφορά το συντελεστή γραμμικής συσχέτισης του Pearson και αναφέρεται σε ποσοτικές μεταβλητές και η δεύτερη κατηγορία που θα εξεταστεί αναλυτικότερα στο τρίτο κεφάλαιο αφορά τους συντελεστές Spearman και Kendall, οι οποίοι χρησιμοποιούνται σε ποιοτικές μεταβλητές και κατηγορικές μεταβλητές δηλαδή μεταβλητές των οποίων οι τιμές δεν επιδέχονται ιεράρχηση.

2.6 Η έννοια της Διασποράς

Στατιστικά μέτρα όπως ο αριθμητικός μέσος (mean), η διάμεσος (median) καθώς και η επικρατούσα τιμή (mode) και τα υπόλοιπα μέτρα έχουν ως αντικειμενικό σκοπό να αντιπροσωπεύσουν έναν πληθυσμό με μία μόνο παράμετρο η οποία δίνει το σημείο στο οποίο παρουσιάζουν την τάση να συγκεντρώνονται οι τιμές μιας μεταβλητής του πληθυσμού που ερευνάται (Κιόχος, 1993).

Όμως η αντιπροσώπευση ενός πληθυσμού με μία από τις παραπάνω παραμέτρους έχει νόημα εφόσον ο πληθυσμός προς εξέταση παρουσιάζει μεγάλη ομοιογένεια. Στην περίπτωση που ο πληθυσμός παρουσιάζει ανομοιογένεια, τότε τα παραπάνω μέτρα κεντρικής τάσης και θέσης δεν μπορούν να χρησιμοποιηθούν ως αντιπροσωπευτικά του συγκεκριμένου πληθυσμού. Ο βαθμός κατά τον οποίο οι διάφορες τιμές ενός πληθυσμού τείνουν να είναι διεσπαρμένες γύρω από τον μέσο αριθμητικό όρο ονομάζεται *διασπορά (variance)*.

Θεωρώντας το παρακάτω παράδειγμα με δύο μεταβλητές x και y και υποθέτοντας ότι παίρνουν τις ακόλουθες τιμές όπως αναγράφονται στον Πίνακα 1:

x_1	10	40	43	46	47	48	50	50	52	52	54
y_2	7	14	15	23	38	48	50	50	75	85	90

Πίνακας 1. Αναγραφόμενες τιμές των μεταβλητών x & y

Μπορεί λοιπόν να παρατηρηθεί ότι και για τις δύο παραπάνω μεταβλητές ο μέσος αριθμητικός όρος είναι 45 και η διάμεσος 48. Η μία όμως μεταβλητή παρουσιάζει εμφανώς διαφορετική μορφή από την άλλη. Συγκεκριμένα στην μεταβλητή x οι τιμές των παρατηρήσεων κυμαίνονται ανάμεσα στους αριθμούς 10

και 54 ενώ οι τιμές για τη μεταβλητή y κυμαίνονται μεταξύ των αριθμών 7 και 90. Συνεπώς, οι πληροφορίες οι πληροφορίες που δίνουν αυτές οι παράμετροι που χαρακτηρίζουν την τάση και την θέση μια κατανομής παρουσιάζονται ανεπαρκείς καθώς δεν δίνουν ενδείξεις για τον τρόπο συγκέντρωσης γύρω από τους κεντρικούς και μέσους όρους. Ως εκ τούτου είναι αναγκαία η χρησιμοποίηση ενός δείκτη που να δίνει τον *βαθμό συγκέντρωσης ή διασποράς* των τιμών μιας μεταβλητής από τον μέσο αριθμητικό όρο. Η παράμετρος που χρησιμοποιείται προκειμένου να πληροφορηθεί κανείς για το βαθμό που οι τιμές των παρατηρήσεων για μία μεταβλητή είναι συγκεντρωμένες ή διασκορπισμένες (σε σχέση με τον μέσο αριθμητικό) καλείται *διασπορά* ή *διακύμανση*.

Το εύρος και ο βαθμός της διασποράς ή συγκέντρωσης καθώς και ο βαθμός συσχέτισης δύο μεταβλητών μπορεί να αντικατοπτριστεί με τα λεγόμενα διαγράμματα διασποράς (dispersion diagrammes). Ένα διάγραμμα είναι μια απλοποιημένη και δομημένη οπτική παρουσίαση εννοιών, ιδεών, κατασκευών, σχέσεων, στατιστικών δεδομένων, ανατομίας κ.λπ. Χρησιμοποιείται σε όλες τις ανθρώπινες δραστηριότητες για να παρουσιάσει, απλοποιήσει και γενικά να κάνει κατανοητό το θέμα με το οποίο σχετίζεται. Έτσι λοιπόν και το διάγραμμα διασποράς αποτελεί μια οπτικοποιημένη παρουσίαση του βαθμού συγκέντρωσης ή διασποράς των τιμών μίας μεταβλητής καθώς και του βαθμού συσχέτισης δύο μεταβλητών όπως θα συζητηθεί εκτενέστερα παρακάτω.

2.7 Μέτρα διασποράς

Τα κύρια μέτρα διασποράς που χρησιμοποιούνται συνήθως στη στατιστική περιλαμβάνουν: α) το εύρος μεταβολής (range), β) το ημιενδοτεταρτημοριακό εύρος, γ) τη μέση απόκλιση και δ) την τυπική απόκλιση.

Α) Εύρος απόκλισης: Το εύρος απόκλισης αποτελεί μία παράμετρο που δίνει τη διαφορά ανάμεσα στη μεγαλύτερη και την μικρότερη τιμή μιας σειράς παρατηρήσεων. Για παράδειγμα η μέση ημερήσια θερμοκρασία σε 18 πόλεις της χώρας στις 30 Ιουλίου είχε ως εξής: 19,20,20,23,24,24,25,25,26,28,29,30, 31,32,33,34,35. Το εύρος μεταβολής των συγκεκριμένων παρατηρήσεων θα προκύψει από τη διαφορά ανάμεσα στη μεγαλύτερη και την μικρότερη τιμή θερμοκρασίας δηλαδή: 35-19=16.

Η παράμετρος όμως αυτή δεν είναι ικανοποιητική γιατί επηρεάζεται και εξαρτάται από τις ακραίες τιμές των παρατηρήσεων. Έτσι αν οι ακραίες τιμές είναι αρκετά

απομακρυσμένες από τα εκάστοτε σύνολα παρατηρήσεων τότε δίνουν μια λάθος και ψευδή εικόνα της έντασης της διασποράς. Ως εκ τούτου το συγκεκριμένο μέτρο χρησιμοποιείται σε τομείς όπου οι ακραίες τιμές είναι σημαντικές, όπως το χρηματιστήριο και τη μετεωρολογία. Αναφορικά με τα πλεονεκτήματα του εν λόγω μέτρου διασποράς, είναι πολύ εύκολο στον υπολογισμό του και περιλαμβάνει και τις ακραίες τιμές της κατανομής. Από την άλλη πλευρά στα μειονεκτήματα περιλαμβάνονται λόγοι που οφείλονται σε αλλοίωση από τις ακραίες τιμές με αποτέλεσμα, σε πολλές περιπτώσεις, να μην παρουσιάζει μια αντιπροσωπευτική εικόνα της διασποράς της κατανομής. Επιπρόσθετα, δεν παρέχει καμιά πληροφορία σχετικά με τη διασπορά των τιμών μεταξύ των άκρων της κατανομής. Για παράδειγμα, δεν λέει τίποτα για τη διασπορά των τιμών της κατανομής γύρω από το μέσο όρο.

Β) Ενδοτεταρτημοριακό Εύρος: Το Ενδοτεταρτημοριακό εύρος είναι το εύρος του κεντρικού 50% των τιμών μιας κατανομής. Το ενδοτεταρτημοριακό εύρος ισούται με τη διαφορά μεταξύ του 1^{ου} και του 3^{ου} τεταρτημορίου τα οποία και είναι τα σημεία που χωρίζουν την κατανομή σε τέσσερα ίσα μέρη:

· 1ο τεταρτημόριο ή 25ο

· 2ο τεταρτημόριο ή 50ο ή διάμεσος

· 3ο τεταρτημόριο ή 75ο

Τα πλεονεκτήματα και μειονεκτήματα του ενδοτεταρτημοριακού εύρους παρουσιάζονται στον κάτωθι Πίνακα 2.

Πλεονεκτήματα	**Μειονεκτήματα**
· Δεν επηρεάζεται από τις ακραίες τιμές	· Δεν λαμβάνει υπόψη τις ακραίες τιμές της κατανομής
· Είναι σχετικά εύκολο στον υπολογισμό του	· Όπως και το εύρος δεν επιτρέπει την ακριβή ερμηνεία μιας συγκεκριμένης τιμής της κατανομής
· Είναι αντιπροσωπευτικό των κεντρικών τιμών της κατανομής	· Δεν είναι ακριβές όταν τα δεδομένα είναι ομαδοποιημένα κατά μεγάλα διαστήματα τιμών
	· Όπως και η διάμεσος, δεν περιγράφει καμιά από τις παραμέτρους, οι οποίες είναι βασικές για την επαγωγική στατιστική

Πίνακας 2. Πλεονεκτήματα & Μειονεκτήματα ενδοτεταρτημοριακού εύρους

δ) Τυπική Απόκλιση: Δείκτης διασποράς αντιπροσωπευτικός των αποκλίσεων μιας ομάδας τιμών από το μέσο όρο. Πρέπει να ειπωθεί ότι ο στατιστικός τύπος για τη

Τυπική απόκλιση διαφέρει ανάλογα αν αφορά όλο τον πληθυσμό μίας αριθμητικής σειράς ή για ένα μέρος της, ένα δείγμα της. Η αλλαγή είναι μικρή, αντί για διαίρεση με N που γίνεται στον στατιστικό τύπο για τη τυπική απόκλιση για τον πληθυσμό, στο δείγμα γίνεται διαίρεση με n-1. Οι τύποι υπολογισμού της τυπικής απόκλισης έχουν ως ακολούθως:

για τον πληθυσμό είναι: $\sigma = \sqrt{\frac{1}{N} * \sum_{i=1}^{N}(x_i - \mu)^2}$

Εξίσωση 1. Τύπος υπολογισμού τυπικής απόκλισης (πληθυσμός)

για το δείγμα είναι: $s = \sqrt{\frac{1}{n-1} * \sum_{i=1}^{n}(x_i - \bar{x})^2}$

Εξίσωση 2. Τύπος υπολογισμού τυπικής απόκλισης (δείγμα)

Τα παραπάνω σύμβολα των τύπων επεξηγούνται όπως: (σ) = Τυπική Απόκλιση Πληθυσμού και (s) = Τυπική Απόκλιση Δείγματος. Ο συνολικός αριθμός πληθυσμού (N) & Δείγματος (n). Μέσος όρος Πληθυσμού (μ) & Δείγματος ($\bar{x}$), ενώ το $\sum$ = δείχνει ότι πρέπει να γίνει πρόσθεση των μελών που ακολουθούν αφού τελειώσουν οι αριθμητικές πράξεις που υποδηλώνονται. Για το x_i= το «i» δείχνει τη θέση του κάθε αριθμητικού μέλους σε ένα αριθμητικό σύνολο, αν αυτό αντικατασταθεί από τον αριθμό θέσης του μέλους αυτού ή αλλιώς το x_5 δείχνει τον 5^{ο} αριθμό σε μία αριθμητική σειρά και ο x_7 τον 7^{ο}. Μπορεί να αντικατασταθεί και να επαναληφθεί τόσες φορές όσες είναι και ο συνολικός αριθμός μελών της αριθμητικής σειράς. Αναφορικά με τα πλεονεκτήματα η τυπική απόκλιση α) μπορεί να χρησιμοποιηθεί για τον υπολογισμό των παραμέτρων του πληθυσμού, β) λαμβάνει υπόψη όλες τις τιμές της κατανομής και γ) είναι ο πιο ευαίσθητος από τους δείκτες διασποράς. Στα μειονεκτήματα συγκαταλέγεται ο υπολογισμός της που είναι σχετικά πιο περίπλοκος σε σχέση με τους υπόλοιπους δείκτες διασποράς και το ότι είναι πολύ ευαίσθητη στις ακραίες τιμές της κατανομής.

2.8 Επιλογή του κατάλληλου δείκτη διασποράς

Μεγάλη σημασία έχουν οι ακραίες τιμές στην επιλογή του κατάλληλου δείκτη. Ωστόσο πρέπει να εξετάζεται και το επίπεδο μέτρησης που έχει προκύψει επιτυχώς. Έτσι, στην ιεραρχική κλίμακα μέτρησης, μεγαλύτεροι αριθμοί δείχνουν μεγαλύτερη ποσότητα από οτιδήποτε μετριέται, αλλά μεγαλύτερες και μικρότερες διαφορές μεταξύ των αριθμών μπορεί να μη δείχνουν μεγαλύτερες και μικρότερες διαφορές

ανάμεσα στα πράγματα που μετρώνται. Σε μια τέτοια περίπτωση αρκεί ο υπολογισμός του εύρους.

Σε κλίμακα ίσων διαστημάτων ή αναλογική κλίμακα, μεγάλες διαφορές στις μετρήσεις αντιστοιχούν πράγματι σε μεγάλες διαφορές στα πράγματα που μετρώνται. Σε αυτή την περίπτωση, και εφόσον δεν αναμένεται να υπάρχουν ακραίες τιμές, επιλέγεται ο μέσος όρος και η τυπική απόκλιση. Αν είναι πιθανό να υπάρχουν ακραίες τιμές ή αν η μέτρηση είναι σε ιεραρχική κλίμακα τότε πρέπει να χρησιμοποιούνται η διάμεσος και το ενδοτεταρτημοριακό εύρος. Η επικρατούσα τιμή και το εύρος μπορούν να χρησιμοποιηθούν αν επαρκεί μία κατά προσέγγιση εικόνα των τιμών του δείγματος. Αναφορικά με την επικρατούσα τιμή ή κορυφή (mode) ορίζεται ως η παρατήρηση με τη μεγαλύτερη συχνότητα. Είναι προφανές ότι η επικρατούσα τιμή μπορεί να οριστεί και στην περίπτωση ποιοτικών δεδομένων. Ο συνηθέστερος τρόπος περιγραφής της διασποράς των τιμών μιας μεταβλητής είναι μέσω της τυπικής απόκλισης. Ο σημαντικότερος λόγος για τον οποίο προτιμάται η τυπική απόκλιση από τους υπόλοιπους δείκτες διασποράς είναι η δυνατότητα που προσφέρει να υπολογίζονται παράμετροι του πληθυσμού. Ακολουθώντας το επόμενο παράδειγμα, έστω ένας καθηγητής της φυσικής που εξέτασε για δύο τμήματα (Τμήμα Α και Β) μιας τάξης του σχολείου το ίδιο τεστ. Η επίδοση των μαθητών του κάθε τμήματος (όπως μετρήθηκε με τη χρήση μιας εικοσαβάθμιας κλίμακας) παρουσιάζεται στον Πίνακα 3:

ΤΜΗΜΑ Α				**ΤΜΗΜΑ Β**			
17	**18**	**19**	**14**	**20**	**16**	**14**	**17**
18	**17**	**13**	**16**	**17**	**20**	**9**	**16**
20	**13**	**17**	**17**	**19**	**16**	**17**	**20**
14	**17**	**15**	**18**	**15**	**17**	**9**	**16**
16	**12**	**19**	**17**	**17**	**10**	**20**	**17**
14				**19**			

Πίνακας 3. Βαθμοί των μαθητών του Τμήματος Α και Β αντίστοιχα

Ένα εύλογο ερώτημα που προκύπτει είναι ποιο από τα δύο τμήματα σημείωσε καλύτερη επίδοση. Υπολογίζοντας των τριών δεικτών κεντρικής τάσης (mean, median, mode δηλαδή μέσο όρο, διάμεσο και κυρίαρχη τιμή αντίστοιχα) βλέπουμε ότι: ο μέσος όρος (mean) των επιδόσεων των μαθητών και για τα δύο τμήματα είναι ο ίδιος: 16,24. Η διάμεσος (median) και στις δύο ομάδες είναι: 16,75 και η κυρίαρχη ή

επικρατούσα τιμή (mode) είναι: 17 (με ίδια συχνότητα 6 και για τα δύο τμήματα). Ωστόσο εντοπίζοντας με μία άμεση ματιά την μικρότερη και τη μεγαλύτερη τιμή για το κάθε ένα τμήμα (12 – 20 για το Τμήμα Α και 9-20 για το Τμήμα Β) γίνεται εμφανές ότι οι επιδόσεις των δύο τμημάτων διαφέρουν. Το εύρος μεταβολής (range) ως ένα πρώτο δείγμα της διασποράς είναι διαφορετικό και επομένως οι επιδόσεις των μαθητών των δύο τμημάτων δεν είναι οι ίδιες.

Αν και όπως ειπώθηκε παραπάνω ο μέσος όρος είναι ίδιος και για τις δύο κατανομές-τιμές επιδόσεων των δύο ομάδων μαθητών στο τεστ, σημαντικό είναι να υπολογιστεί και η τυπική απόκλιση για κάθε μία από τις κατανομές η οποία για το Τμήμα Α είναι 2,211 και για το Τμήμα Β είναι 3,36. Με βάση το μέσο όρο που είναι κοινός καθώς και με την εύρεση της τυπικής απόκλισης κάθε τμήματος ξεχωριστά, θα εφαρμοστεί ένα μέτρο το οποίο βοηθά στη σύγκριση ομάδων τιμών, που εκφράζονται είτε σε διαφορετικές μονάδες μέτρησης, είτε στην ίδια μονάδα μέτρησης, αλλά έχουν σημαντικά διαφορετικές μέσες τιμές. Αυτός είναι ο συντελεστής μεταβολής ή συντελεστής μεταβλητότητας (coefficient of variation), ο οποίος ορίζεται από το λόγο: $CV = \frac{\tau\upsilon\pi\iota\kappa\acute{\eta}\ \alpha\pi\acute{o}\kappa\lambda\iota\sigma\eta}{\mu\acute{\epsilon}\sigma\eta\ \tau\iota\mu\acute{\eta}} * 100\% = \frac{s}{\bar{x}} * 100\%$. Ο συντελεστής μεταβολής εκφράζεται επί τοις εκατό, είναι συνεπώς ανεξάρτητος από τις μονάδες μέτρησης και παριστάνει ένα μέτρο σχετικής διασποράς **τ**ων τιμών και όχι της απόλυτης διασποράς Εκφράζει, δηλαδή, τη μεταβλητότητα των δεδομένων απαλλαγμένη από την επίδραση της μέσης τιμής. Έτσι ο συντελεστής μεταβολής του Τμήματος Α είναι 13,61% ενώ του Τμήματος Β είναι 20,6% δίνοντας μεγαλύτερη σχετική διασπορά στο Τμήμα Β. Αυτό μεταφράζεται στο ότι υπάρχει μεγαλύτερη ομοιογένεια βαθμών στο Τμήμα Α παρά στο Β. Υπάρχουν πολλές οι έρευνες για τις οποίες υπάρχει ενδιαφέρον να γίνει μελέτη της αλληλεξάρτησης μεταξύ των μεταβλητών. Δηλαδή, να διαπιστωθεί κατά πόσο οι τιμές που παίρνει μια μεταβλητή επηρεάζονται από τις τιμές που παίρνει η άλλη μεταβλητή.

Δίνοντας ένα παράδειγμα, για την μελέτη της σχέσης ανάμεσα στη νοημοσύνη και την ικανότητα του ατόμου να επεξεργάζεται αριθμητικές έννοιες (αριθμητικός συλλογισμός), ζητείται να διαπιστωθεί κατά πόσο η μεταβολή στις τιμές της νοημοσύνης επιφέρουν αλλαγές στην επίδραση της αριθμητικής ικανότητας. Το στατιστικό κριτήριο που χρησιμοποιείται για να διαπιστωθεί αν υπάρχει αλληλεξάρτηση μεταξύ δύο μεταβλητών, ονομάζεται δείκτης συσχέτισης **(**r**)** (Correlation Coefficient). Οι πληροφορίες από την εφαρμογή του δείκτη συσχέτισης

περιλαμβάνουν το α) αν υπάρχει αλληλεξάρτηση (συσχέτιση), β) το είδος της συσχέτισης και γ) το βαθμό της συσχέτισης.

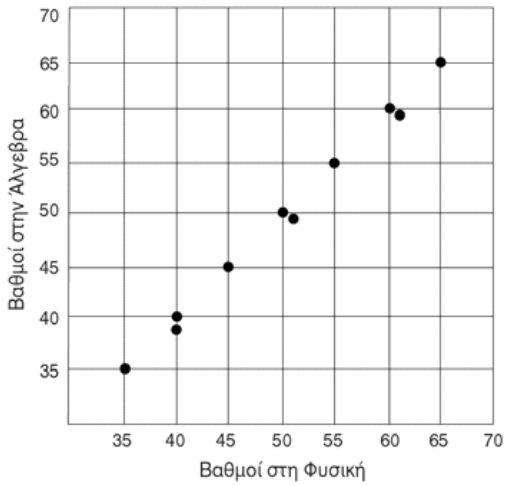
2.9 Απόδοση του δείκτη συσχέτισης

Η συσχέτιση μεταξύ δύο μεταβλητών μπορεί να αποδοθεί με δύο τρόπους, α) την αριθμητική τιμή και β) τη γραφική αναπαράσταση. Τα είδη συσχέτισης που υπάρχουν αφορούν την α) **Θετική Συσχέτιση (positive correlation)** και β) **Αρνητική Συσχέτιση (negative correlation).** Για παράδειγμα**,** η αύξηση των δαπανών για μια διαφήμιση μπορεί να συνεπάγεται την αύξηση των πωλήσεων ενός προϊόντος (θετική συσχέτιση) ή η αύξηση των δαπανών για την προώθηση ενός προϊόντος μπορεί να συνεπάγεται τη μείωση της αντιλαμβανόμενης από τους καταναλωτές ποιότητας του προϊόντος (αρνητική συσχέτιση). Προκειμένου να αποδοθεί ο δείκτης συσχέτισης με γραφικό τρόπο χρησιμοποιούνται τα διάγραμμα διασποράς ή διασκόρπισης (scatter plot). Ο παρακάτω πίνακας 4 αντικατοπτρίζει τις επιδόσεις των δέκα μαθητών στα τέσσερα μαθήματα Άλγεβρα, Φυσική, Νέα Ελληνικά και Μουσική, ζητώντας την ύπαρξη τυχόν συσχέτισης.

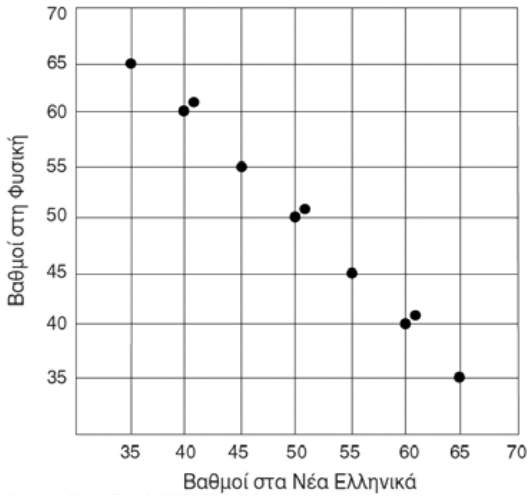
Μαθητής	Άλγεβρα	Φυσική	Ν. Ελληνικά	Μουσική
Α	65	65	35	61
Β	60	60	40	35
Γ	60	60	40	46
Δ	55	55	45	40
Ε	50	50	50	50
ΣΤ	50	50	50	60
Ζ	45	45	55	60
Η	40	40	60	40
Θ	40	40	60	55
Ι	35	35	66	41

Πίνακας 4. Αριθμητικά στοιχεία μαθητών σε τέσσερα μαθήματα

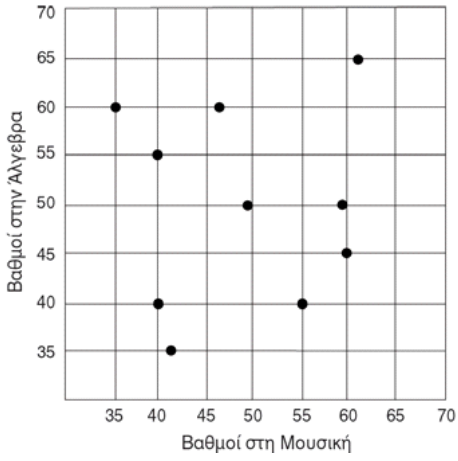
Με βάση τα στοιχεία που αντιστοιχούν στις επιδόσεις καθ ενός από τους δέκα μαθητές, ενδεικτικά παρουσιάζονται τα διαγράμματα διασποράς στα συγκεκριμένα μαθήματα και προκύπτουν τα παρακάτω διαγράμματα διασποράς (Εικόνα 1, 2 και 3) με βάση τις τιμές των εμπλεκόμενων μεταβλητών.



Εικόνα 1. Διάγραμμα Διασποράς για τους βαθμούς στην Άλγεβρα και την Φυσική[1]



Εικόνα 2. Διάγραμμα Διασποράς για τους βαθμούς στην Φυσική και τα Νέα Ελληνικά[2]

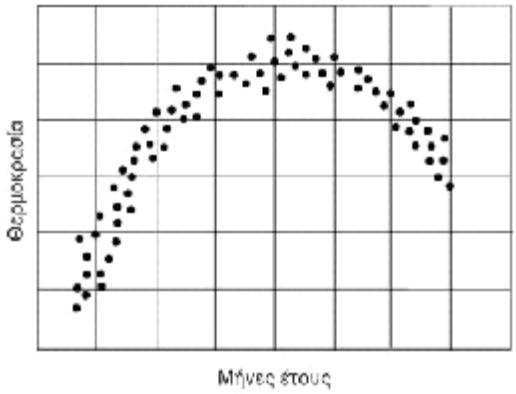


Εικόνα 3. Διάγραμμα Διασποράς για τους βαθμούς στην Άλγεβρα και Μουσική[3]

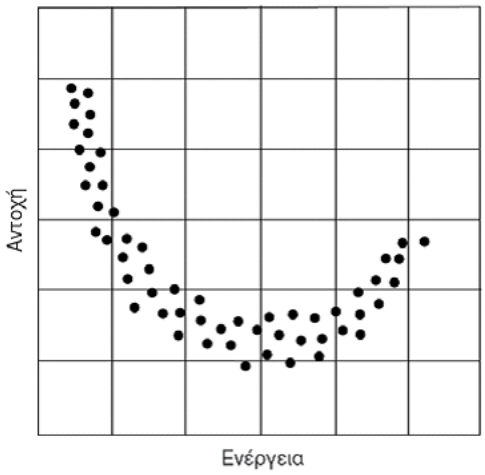
[1]**Πηγή:** Ρούσσος, Π., Λ., Τσαούσης, Γ., (2006). Στατιστική Εφαρμοσμένη στις Κοινωνικές Επιστήμες. Αθήνα: Ελληνικά Γράμματα.

[2,3] **Πηγή:** Ρούσσος, Π., Λ., Τσαούσης, Γ., (2006). Στατιστική Εφαρμοσμένη στις Κοινωνικές Επιστήμες. Αθήνα: Ελληνικά Γράμματα.

Εάν υπάρχει καμπυλόγραμμη αντί για ευθύγραμμη σχέση ανάμεσα στις δύο μεταβλητές που μελετώνται είναι πιθανόν να βρεθεί χαμηλή ή και καθόλου συσχέτιση αλλά αυτό να μην ισχύει για όλο το μήκος της σχέσης (Εικόνα 4 και 5).



Εικόνα 4. Διάγραμμα Διασποράς για θετική καμπυλόγραμμη συσχέτιση ανάμεσα στη θερμοκρασία και τους μήνες του έτους[4]



Εικόνα 5. Διάγραμμα Σκεδασμού: Αρνητική καμπυλόγραμμη Συσχέτιση ανάμεσα στη αντοχή και την ενέργεια[5]

Ο δείκτης συσχέτισης αποτελεί ένα στατιστικό κριτήριο που πληροφορεί μόνο για τη συμμεταβολή των δύο μεταβλητών που μελετώνται και όχι για το εάν υπάρχει αιτιώδης σχέση μεταξύ τους. Η υψηλή συσχέτιση δεν δηλώνει κατ ανάγκη σχέσεις αιτίου και αποτελέσματος. Μπορεί να οφείλεται σε μια τρίτη μεταβλητή, την οποία δεν έχει συμπεριληφθεί στην έρευνα και η οποία να λειτουργεί ως αίτιο.

[4,5]**Πηγή:** Ρούσσος, Π., Λ., Τσαούσης, Γ., (2006). Στατιστική Εφαρμοσμένη στις Κοινωνικές Επιστήμες. Αθήνα: Ελληνικά Γράμματα.

2.10 Συντελεστής γραμμικής συσχέτισης του Pearson

Θεωρώντας δύο τυχαίες μεταβλητές *X*, *Y* και n ζεύγη παρατηρήσεων (x_1,y_1), (x_2,y_2),..., (x_n,y_n) από *τυχαίο δείγμα* μεγέθους n. ***Ο δειγματικός συντελεστής γραμμικής συσχέτισης (r)*** δύο ποσοτικών μεταβλητών ορίζεται από το πηλίκο:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

Εξίσωση 3. Εξίσωση δειγματικού συντελεστή γραμμικής συσχέτισης

όπου:

$$s_{xy} = Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{n-1}$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{και} \quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Επομένως:

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n \cdot \bar{y}^2}}$$

Η τελευταία σχέση αντικατοπτρίζει την διακύμανση μεταξύ των μεταβλητών x, y και s_x, s_y οι τυπικές αποκλίσεις (standard deviation) αυτών. ***Ο πληθυσμιακός συντελεστής γραμμικής συσχέτισης του Pearson ορίζεται ανάλογα και συμβολίζεται με ρ*** ενώ δίνεται από τον τύπο: $\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$.

Ο πληθυσμιακός όπως και ο δειγματικός συντελεστής γραμμικής συσχέτισης αποτελείται από δύο στοιχεία: α) **ένα πρόσημο** (πληροφορίες για την κατεύθυνση) και β) **μια αριθμητική τιμή από 0 ως 1** (πληροφορίες για το βαθμό). Οι ιδιότητες του πληθυσμιακού συντελεστή (ρ) αναγράφονται αναλυτικά στα παρακάτω:

- Ο συντελεστής γραμμικής συσχέτισης είναι καθαρός αριθμός και δεν έχει μονάδες μέτρησης.
- **-1 <ρ < 1.**Για το διάστημα των τιμών μεταξύ των 0 και 1 ισχύει:
  1. Αν ο συντελεστής είναι μικρότερος του ± 0.30 δεν **υπάρχει συσχέτιση**
  2. Αν ο συντελεστής κυμαίνεται μεταξύ ± 0.30 – 0.49 υπάρχει **Χαμηλή συσχέτιση**
  3. Αν ο συντελεστής κυμαίνεται μεταξύ ± 0.50 – 0.69 υπάρχει **Μέτρια συσχέτιση**

4. Αν ο συντελεστής κυμαίνεται μεταξύ ± 0.70 – 0.79 υπάρχει **Υψηλή συσχέτιση**
5. Αν ο συντελεστής κυμαίνεται μεταξύ ± 0.80 – 0.99 υπάρχει **Πολύ υψηλή συσχέτιση**

· Όταν ρ= -1, σημαίνει ότι υπάρχει πλήρης (τέλεια) συσχέτιση και μάλιστα οι τιμές της μιας μεταβλητής αυξάνουν, ενώ οι τιμές της άλλης μεταβλητής μειώνονται.

· Όταν το ρ= +1 σημαίνει πλήρης (τέλεια) συσχέτιση των δύο μεταβλητών και μάλιστα οι τιμές και των δύο βαίνουν αύξουσες ή φθίνουσες.

Και στις δύο αυτές ακραίες τιμές του συντελεστή γραμμικής συσχέτισης ισχύει ανάμεσα στις δύο μεταβλητές Χ και Υ η ποσοτική (συναρτησιακή, μαθηματική σχέση Υ= α+ β·Χ)

· Αν r=0 τότε οι μεταβλητές Χ και Υ λέγονται ασυσχέτιστες.

Εδώ θα πρέπει να γίνει υπενθύμιση ότι άλλο πράγμα νοείται με τον όρο ανεξάρτητες μεταβλητές και άλλο με τον όρο ασυσχέτιστες. Στις προηγούμενες παραγράφους εξετάστηκε το πρόβλημα της παλινδρόμησης δηλ. το πρόβλημα της εύρεσης γραμμικής σχέσης ανάμεσα σε δύο μεταβλητές x και y. Σ ένα πρόβλημα παλινδρόμησης η ανεξάρτητη μεταβλητή x δεν είναι τυχαία μεταβλητή (τ.μ.) αφού οι τιμές της είναι σταθερές ή προκαθορισμένες ενώ η εξαρτημένη μεταβλητή y είναι τ.μ. γιατί η παρατήρησή της συλλέγεται τυχαία από κατανομή πιθανότητας όταν η Χ έχει συμβεί. Στη συσχέτιση δεν υπάρχει διάκριση μεταξύ των δύο μεταβλητών καθώς και οι δύο μεταβλητές x και y είναι τ.μ. αφού ούτε οι τιμές της x ούτε της y είναι ορισμένες εκ των προτέρων. Για παράδειγμα:

· Το ύψος του άνδρα και της γυναίκας σ ένα ζευγάρι

· Οι βαθμοί των φοιτητών σε διάφορα μαθήματα

· Το ύψος παραγωγής των προϊόντων και το ύψος των τιμών τους

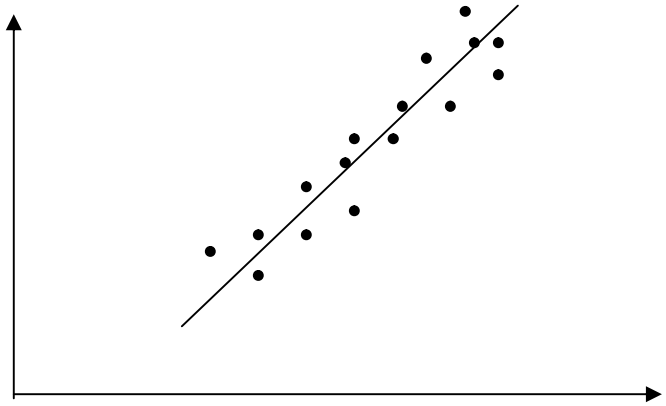
· Ο αριθμός των γάμων και τα ύψος των συναλλαγών με επιπλοποιούς

· Ο αριθμός των ελαττωματικών προϊόντων και το ύψος των παραγόμενων προϊόντων

Παρ όλο που ο εντοπισμός της ανεξάρτητης και της εξαρτημένης μεταβλητής ταλαιπωρεί συχνά, ωστόσο στα προβλήματα της συσχέτισης αυτό που ενδιαφέρει είναι η διευκρίνιση της ύπαρξης ή όχι συσχέτισης ανάμεσα σε δύο τυχαίες

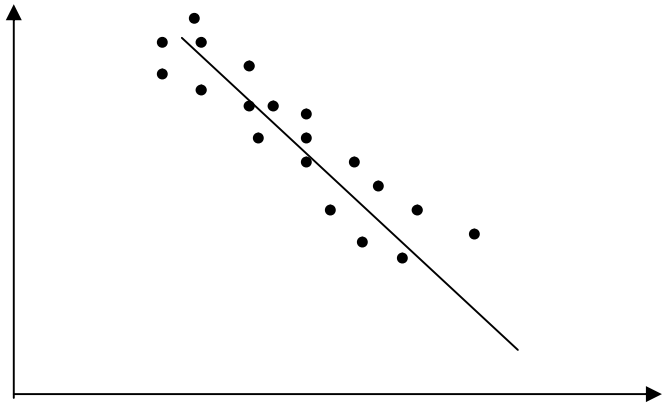
μεταβλητές. Έτσι υποθέτοντας ότι εξετάζονται τα Ν άτομα ενός πληθυσμού ή n άτομα ενός δείγματος ως προς τις μεταβλητές ιδιότητές τους Χ και Υ. Αν τα ζευγάρια των παρατηρήσεων του δείγματος: (x_1, y_1), (x_2, y_2) ,..... (x_n, y_n) βρίσκονται σε μια ευθεία (ή προσεγγίζονται από μια ευθεία) τότε θα ισχύει ότι οι δύο μεταβλητές είναι συσχετισμένες. Ακόμα δύο τυχαίες μεταβλητές (τ.μ.). θα λέγονται:

· *Θετικά συσχετισμένες* αν η αύξηση των τιμών της μιας τ.μ. έχει σαν συνέπεια και την αύξηση των τιμών της άλλης (Εικ. 6)



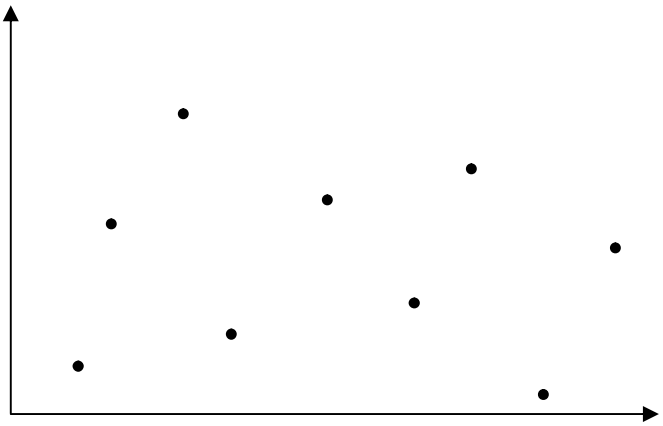
Εικόνα 6. Θετικά Συσχετισμένες μεταβλητές

· *Αρνητικά συσχετισμένες* αν η αύξηση των τιμών της μιας τ.μ. έχει σαν συνέπεια τη μείωση των τιμών της άλλης (Εικ. 7)



Εικόνα 7. Αρνητικά Συσχετισμένες μεταβλητές

· *Ασυσχέτιστες ή μη συσχετισμένες γραμμικά* αν το διάγραμμα των (x_1, y_1), δεν δείχνει κάποια συστηματική διάταξη ώστε να προσαρμόζονται σε μία ευθεία (Εικ. 8).



Εικόνα 8. Ασυσχέτιστες ή μη συσχετισμένες γραμμικά μεταβλητές

Εδώ θα πρέπει να τονιστεί ότι η ύπαρξη συσχέτισης μεταξύ δυο τυχαίων μεταβλητών δεν σημαίνει ότι οι τιμές της μιας επηρεάζονται από τις τιμές της άλλης. Η συσχέτιση αυτή πιθανόν να οφείλεται στην επίδραση άλλων παραγόντων σε αυτές. Έτσι η συσχέτιση πληροφορεί για την ένταση της (γραμμικής σχέσης) όταν αυτή υπάρχει χωρίς να διερευνά τις αιτίες που την προκαλούν.

2.11 Συσχέτιση δε σημαίνει αιτιότητα

Όταν σε μια μη πειραματική έρευνα (δειγματοληψία) δύο μεταβλητές X και Y βρίσκονται συσχετισμένες αυτό σημαίνει μόνο ότι οι μεταβλητές αυτές συνδέονται με κάποια σχέση. Δε συνεπάγεται, κατ' ανάγκη, αιτιότητα. Οι δύο μεταβλητές μπορεί βεβαία να συνδέονται με σχέση αιτιότητας, μπορεί όμως, όχι. Για παράδειγμα, μπορεί και οι δύο να επηρεάζονται από μια τρίτη μεταβλητή. Ας αναφερθούν δύο (2) παραδείγματα:

1. Παρατηρήθηκε ότι το *ύψος των μαθητών* ενός σχολείου, ηλικίας 6 έως 13 ετών, έχει ισχυρή θετική γραμμική συσχέτιση με την *αντιληπτική ικανότητα των μαθητών*. Προφανώς η αντιληπτική ικανότητα των μαθητών δεν επηρεάζεται από το ύψος τους. Απλώς τόσο η πνευματική όσο και η φυσική ανάπτυξη των μικρών μαθητών επηρεάζονται παράλληλα από άλλους παράγοντες.

2. Παρατηρήθηκε ότι οι πωλήσεις ταχύπλοων στο Sidney είχαν, για μια μακράν περίοδο, ισχυρή θετική συσχέτιση με τις πωλήσεις έγχρωμων τηλεοράσεων στη

Melbourne. Προφανώς, τόσο οι πωλήσεις ταχύπλοων όσο και οι πωλήσεις έγχρωμων τηλεοράσεων ήταν συνάρτηση γενικότερων ευνοϊκών οικονομικών παραγόντων.
Είναι, κατά συνέπεια, φανερό ότι η πρόχειρη ή επιπόλαιη ερμηνεία και χρήση του συντελεστή συσχέτισης (ρ) οδηγεί πολλές φορές σε παρερμηνείες ή και σε λανθασμένα συμπεράσματα. Για αιτιολογικά συμπεράσματα, σχεδόν πάντοτε, απαιτείται πειραματισμός.

2.12 Πίνακας Συσχετίσεων

Το διάγραμμα μήτρας συσχετίσεων είναι ένας πολύ παραστατικός τρόπος οπτικοποίησης των συντελεστών συσχέτισης πολλών μεταβλητών. Ο πίνακας συσχετίσεων είναι ο πίνακας που περιέχει σαν στοιχεία του τους συντελεστές συσχέτισης του Pearson για κάθε ζευγάρι μεταβλητών. Ο συντελεστής συσχέτισης του Pearson μετράει μόνο τη γραμμική συσχέτιση ανάμεσα στις μεταβλητές και επομένως δεν μπορεί να δώσει πληροφορία για άλλης μορφής συσχέτιση. Ο συντελεστής συσχέτισης του Pearson είναι κατάλληλος μόνο για ζεύγη ποσοτικών μεταβλητών.

Ο πίνακας έχει απαραίτητα τιμές ίσες με τη μονάδα στη διαγώνιο, είναι συμμετρικός και κανένα στοιχείο του δεν μπορεί να πάρει τιμή μεγαλύτερη σε απόλυτη τιμή από το 1. Όπως προαναφέρθηκε, τιμές –1 και 1 σημαίνουν απόλυτα γραμμική σχέση των δύο μεταβλητών, το πρόσημο υποδηλώνει την ύπαρξη θετικής ή αρνητικής σχέσης. Η θετική σχέση ερμηνεύεται πως όσο αυξάνει η τιμή της μιας μεταβλητής τόσο αυξάνει και η τιμή της άλλης ενώ η αρνητική σχέση ερμηνεύεται πως όσο αυξάνει η τιμή της μιας μεταβλητής μειώνεται η τιμής της άλλης.

Εξετάζοντας το παρακάτω χαρακτηριστικό παράδειγμα, έστω έρευνα που διεξήχθη σε ένα δείγμα 30 μαθητών με σκοπό να αποτυπώσουν τις απόψεις τους επί της χρήσης και της αποτελεσματικότητάς στη μαθησιακή τους εμπειρία ενός διαδικτυακού εργαλείου αξιολόγησης. Μετά την ανάλυση των ερωτηματολογίων τα αποτελέσματα που προέκυψαν αντικατοπτρίζονται στον ακόλουθο Πίνακα 5 όπου δίνονται στοιχεία περιγραφικής στατιστικής για τις τέσσερις κατηγορίες: μέγεθος δείγματος (n), εύρος (=μέγιστο-ελάχιστο), ελάχιστο, μέγιστο, μέση τιμή, τυπική απόκλιση.

	n	Εύρος	Ελάχιστο	Μέγιστο	Μέση Τιμή	Τυπική Απόκλιση
Χρησιμότητα	40	4,55	2,17	6,71	5,46	0,921
Ευχρηστία	40	3,13	3,88	7,00	5,84	0,809
Ευκολία στη μάθηση	40	3,67	3,33	7,00	6,31	0,920
Ικανοποίηση	40	3,00	4,00	7,00	5,89	0,773

Πίνακας 5. Περιγραφική Στατιστική για τις Τέσσερις Διαστάσεις[6]

Από τις μέσες τιμές που παρατηρούνται από τα παραπάνω, προκύπτει ότι το μέσο σκορ για την κατηγορία χρησιμότητα είναι 5,46, το μέσο σκορ για την κατηγορία ευχρηστία είναι 5,84, το μέσο σκορ για την κατηγορία ευκολία στη μάθηση είναι 6,31 και το μέσο σκορ για την κατηγορία ικανοποίηση είναι 5,89.

2.12.1 Πίνακας Συσχετίσεων για τις τέσσερεις Κατηγορίες

Ο πίνακας συσχετίσεων (Pearson correlation matrix) για τις τέσσερις κατηγορίες (χρησιμότητα, ευχρηστία, ευκολία στη μάθηση και ικανοποίηση) δίνεται στον Πίνακα 6, όπου προκύπτει ότι υπάρχουν συσχετίσεις μεταξύ όλων των κατηγοριών του ερωτηματολογίου αξιολόγησης ανά δύο. Συγκεκριμένα, παρατηρείται ότι για τις κατηγορίες χρησιμότητα και ευχρηστία να έχουν την ισχυρότερη από τις γραμμικές σχέσεις στον πίνακα συσχετίσεων (ρ = 0,786), ενώ οι κατηγορίες ευκολία στη μάθηση και ικανοποίηση έχουν την ασθενέστερη από τις σχέσεις του πίνακα συσχετίσεων (ρ = 0,509). Η διαγώνιος του πίνακα (από την άνω αριστερή γωνία μέχρι την κάτω δεξιά) αποτελείται από τη συσχέτιση της κάθε μιας μεταβλητής με τον εαυτό της, πράγμα που όπως είναι φυσικό δίνει συντελεστή συσχέτισης 1.000.

	Χρησιμότητα	Ευχρηστία	Ευκολία στη μάθηση	Ικανότητα
Χρησιμότητα	1			
Ευχρηστία	0,786	1		
Ευκολία στη μάθηση	0,607	0,585	1	
Ικανοποίηση	0,568	3,00	0,509	1

Πίνακας 6. Πίνακας Συσχετίσεων των τεσσάρων κατηγοριών[7]

[6,7] Πηγή: ΕΡΚΥΝΑ: Επιθεώρηση Εκπαιδευτικών-Επιστημονικών Θεμάτων (Τεύχος 1ο, 2014: ISSN:2241-8393), http://erkyna.gr/e_docs/periodiko/teyxos/teyxos-1-%281_2014%29.pdf

2.13 Έλεγχος στατιστικής σημαντικότητας του ρ

Η στατιστική σημαντικότητα είναι μια δήλωση της πιθανότητας να προκύψει ένας συγκεκριμένος συντελεστής συσχέτισης για ένα δείγμα δεδομένων αν δεν υπάρχει συσχέτιση (δηλαδή αν η συσχέτιση είναι 0.00) στον πληθυσμό από τον οποίο λήφθηκε το δείγμα.

Όπως αναλύθηκε παραπάνω ο συντελεστής συσχέτισης Pearson ρ αποτελεί ένα συντελεστή γραμμικής σχέσης μεταξύ δύο μεταβλητών με δύο κύριες ιδιότητες: το μέγεθος και την κατεύθυνση. Όταν είναι κοντά στο μηδέν, δεν υπάρχει συσχέτιση, αλλά όπως αναφέρθηκε όταν πλησιάζει τις τιμές -1 ή +1 υπάρχει ισχυρή αρνητική ή θετική σχέση αντίστοιχα μεταξύ των μεταβλητών που εξετάζονται. Αλλά πώς κάποιος γνωρίζει εάν μια συσχέτιση απέχει αρκετά από το μηδέν προκειμένου να εξασφαλίζει το γεγονός ότι υφίσταται σχέση μεταξύ των εξεταζόμενων μεταβλητών;

Στην περίπτωση που κάποιος ενδιαφέρεται για το είδος της γραμμικής σχέσης μεταξύ δύο μεταβλητών, είναι απαραίτητο να χρησιμοποιήσει ανάλυση παλινδρόμησης και να πραγματοποιήσει τον έλεγχο της κλίσης, κάτι το οποίο θα αναλυθεί και εκτενέστερα σε επόμενο κεφάλαιο. Προϋπόθεση για τον έλεγχο αυτό είναι για κάθε τιμή της ανεξάρτητης μεταβλητής ο πληθυσμός των τιμών της εξαρτημένης μεταβλητής να έχει κανονική κατανομή με σταθερή τυπική απόκλιση. Η προϋπόθεση αυτή απαιτείται είτε τα δεδομένα είναι παρατηρούμενα είτε πειραματικά. Αν το μόνο που ενδιαφέρει είναι η ύπαρξη και όχι το είδος της γραμμικής σχέσης μεταξύ δύο μεταβλητών, δύναται να χρησιμοποιηθεί ο συντελεστής συσχέτισης ρ. Αυτό προϋποθέτει ότι τα δεδομένα είναι παρατηρούμενα και ότι οι δύο μεταβλητές σχηματίζουν μια διμεταβλητή κανονική κατανομή. Όπως προαναφέρθηκε ο συντελεστής συσχέτισης για το σύνολο ενός πληθυσμού συμβολίζεται με το ελληνικό γράμμα ρ. επειδή πρόκειται για παράμετρο του πληθυσμού, η τιμή του συντελεστή συσχέτισης ρ είναι κατά κανόνα άγνωστη και πρέπει να εκτιμηθεί από τα δεδομένα του δείγματος. Ο συντελεστής συσχέτισης δείγματος σύμφωνα με προηγούμενο τύπο υπολογίζεται ως εξής: $r = \frac{s_{xy}}{s_x \cdot s_y}$. Αν δεν υπάρχει γραμμική σχέση μεταξύ των δύο μεταβλητών, ο συντελεστής συσχέτισης του πληθυσμού θα είναι ρ=0. Συνεπώς πρέπει να ελεγχθεί η υπόθεση:

$H_0: \rho = 0$

$H_1: \rho \neq 0$

Ο έλεγχος γίνεται βάσει του τύπου: $t = r\sqrt{\frac{n-2}{1-r^2}}$, όπου η κατανομή δειγματοληψίας είναι η κατανομή student t[8] με βαθμούς ελευθερίας : ν = n-2, με την προϋπόθεση ότι οι δύο μεταβλητές σχηματίζουν μια διμεταβλητή κανονική κατανομή. Δεδομένου της πλήρης ανάλυσης που θα γίνει σε επόμενο κεφάλαιο για την ανάλυση της απλής γραμμικής παλινδρόμησης και συσχέτισης, για την οποία θα παρουσιαστούν και παραδείγματα, εδώ θα παρατεθεί ένα απλό παράδειγμα ελέγχου γραμμικής σχέσης. Οπότε, έστω τα παρακάτω δεδομένα που παρουσιάζονται στον Πίνακα 7 και αφορούν ένα τυχαίο δείγμα 100 ατόμων που αγόρασαν πρόσφατα σε χιλιάδες ευρώ ένα σπίτι (τιμή πώλησης) στο κέντρο της Αθήνας και στην διπλανή στήλη παρουσιάζονται οι μετρήσεις των αποστάσεων τους (σε χιλιάδες μίλια) από τον Παρθενώνα. Το ζητούμενο είναι να βρεθεί εάν υπάρχει γραμμική σχέση μεταξύ των δύο μεταβλητών. Η υπόθεση είναι ότι οι δύο μεταβλητές σχηματίζουν μια διμεταβλητή κανονική κατανομή, ενώ γίνεται χρήση της στάθμης σημαντικότητας α=5%[9].

Α/Α	Τιμή πώλησης	Μετρήσεις αποστάσεων
1	14,6	37,4
2	14,1	44,8
3	14,0	45,8
…	…	…
99	14,7	39,2
100	14,3	36,4

Πίνακας 7. Δεδομένα για τυχαίο δείγμα 100 ατόμων

Ξεκινώντας την επίλυση του ανωτέρω παραδείγματος, πρέπει να ελεγχθεί η συνθήκη:

$H_0: \rho = 0$

[8] Ο μέσος και η διασπορά της κατανομής student t δίνονται από τους τύπους E(t)=0 και V(t) = ν/ν-2 για ν>2. Η καμπύλη της κατανομής έχει αρκετή ομοιότητα με την καμπύλη της τυποποιημένης κανονικής κατανομής, καθώς είναι και οι δύο συμμετρικές ως προς το μηδέν, ενώ στο σχήμα τους έχει πολύ μικρές διαφορές. Όσο οι βαθμοί ελευθερίας αυξάνονται η διασπορά της κατανομής student t (σ^2 = ν/ν-2) τείνει στην μονάδα και η καμπύλη τείνει να ταυτιστεί με την καμπύλη της τυποποιημένης κανονικής κατανομής (Σιώμκος, 2005).

[9] Κατά τον έλεγχο υποθέσεων υπάρχουν δύο δυνατοί τύποι σφάλματος. Το σφάλμα τύπου Ι συμβαίνει όταν απορρίπτουμε μια αληθινή μηδενική υπόθεση, ενώ το σφάλμα τύπου ΙΙ συμβαίνει όταν δεν απορρίπτουμε μια ψευδή μηδενική υπόθεση. Για παράδειγμα σε ένα δικαστήριο το σφάλμα τύπου Ι σημαίνει ότι ένας αθώος καταδικάστηκε ενώ το σφάλμα τύπου ΙΙ σημαίνει ότι ένας ένοχος αθωώθηκε. Η πιθανότητα σφάλματος τύπου Ι συμβολίζεται με ελληνικό γράμμα α και ονομάζεται στάθμη σημαντικότητας (significance level) (Σιώμκος, 2005).

$H_1: \rho \neq 0$

Υπολογίζοντας τα s_{xy} = -2,909, s_x^2 = 43,509 και s_y^2 = 0,300 από όπου: $s_x = \sqrt{43,509}$ = 6,596 και $s_y = \sqrt{0,300}$ = 0,5477. Αντικαθιστώντας αυτές τις τιμές προκύπτει ότι: $r = \frac{s_{xy}}{s_x \cdot s_y}$ = -0,8052. Ο έλεγχος είναι: $t = r\sqrt{\frac{n-2}{1-r^2}} = -13,44$ και η περιοχή απόρριψης είναι: $t < -t_{\alpha/2.\nu} = -t_{0,025.98}$ @-1,984 και $t > t_{\alpha/2.\nu} = t_{0,025.98}$ @1,984. Ο έλεγχος t = -13,44 βρίσκεται στην περιοχή απόρριψης και η τιμή-p είναι πρακτικά ίση με μηδέν. Συνεπώς μπορεί να απορριφθεί η μηδενική υπόθεση, ερμηνεύοντας το αποτέλεσμα ότι η απόσταση από τον Παρθενώνα μπορεί να επηρεάσει την τιμή πώλησης.

2.14 Σύνοψη

Η βασική μέθοδος παρουσίασης δύο ποιοτικών χαρακτηριστικών είναι η κατασκευή της κοινής κατανομής συχνοτήτων (πίνακας συνάφειας) και ο υπολογισμός των αντίστοιχων ποσοστών. Ο υπολογισμός των περιγραφικών στατιστικών μέτρων για τον εντοπισμό της φύσης και της έντασης της σχέσης μεταξύ δύο ποσοτικών μεταβλητών πραγματοποιείται με τον υπολογισμό του συντελεστή γραμμικής συσχέτισης του Pearson (διαδικασία Correlate) και με την κατασκευή του διαγράμματος διασποράς (διαδικασία Scatter). Η ανάλυση συσχέτισης δίνει πληροφορίες στον ερευνητή για την κατεύθυνση και την ένταση της σχέσης μεταξύ των μεταβλητών της έρευνας. Συνήθως οι συντελεστές συσχέτισης λαμβάνουν τιμές μεταξύ του -1 και του +1, αποτυπώνοντας έτσι αν υπάρχει θετική ή αρνητική ή δεν υπάρχει καθόλου συσχέτιση των μεταβλητών. Υπενθυμίζεται ότι, αν και η ανάλυση συσχέτισης αντικατοπτρίζει την ένταση της σχέσης μεταξύ των μεταβλητών, δεν εξηγεί τη σχέση αιτίας – αιτιατού.

ΚΕΦΑΛΑΙΟ ΤΡΙΤΟ
ΛΟΙΠΟΙ ΣΥΝΤΕΛΕΣΤΕΣ ΣΥΣΧΕΤΙΣΗΣ

3.1 Εισαγωγή

Αυτό που στην ουσία επιτυγχάνει η συσχέτιση είναι να μετρά το βαθμό συνάφειας-αλληλεπίδρασης ανάμεσα σε δύο ή περισσότερες μεταβλητές. Αυτό, με τη σειρά του μεταφράζεται στο ότι από τη τιμή ενός δείκτη (συντελεστή συσχέτισης) μπορεί να κατανοηθεί πόσο έντονη ή χαλαρή είναι η συσχέτιση δύο μεταβλητών. Οι μεταβλητές είναι: *"μαθηματικά μεγέθη τα οποία αναπαριστούν μεγέθη παραγόντων του περιβάλλοντος που διερευνούμε (π.χ. φύλο, ηλικία, χρόνος, κ.α). Μεταξύ των μεγεθών που μελετούμε γίνεται ένα βασικός διαχωρισμός ο οποίος αντικατοπτρίζεται και στις μεταβλητές που τα αναπαριστούν: υπάρχουν μεγέθη τα οποία επηρεάζουν και μεγέθη τα οποία επηρεάζονται, δηλαδή μια σχέση αιτίου και αιτιατού. Οι μεταβλητές που αναπαριστούν τα αίτια συχνά ονομάζονται ανεξάρτητες ενώ οι μεταβλητές οι οποίες επηρεάζονται ονομάζονται εξαρτημένες, επειδή ακριβώς οι τιμές τους εξαρτώνται από τις τιμές των πρώτων."* (Βερονίκης, 2011).

Η διαδικασία συσχέτισης εμφανίζεται σε ποσοτικές μεταβλητές αλλά και σε ποιοτικές ή κατηγορικές μεταβλητές. Ωστόσο, θα έπρεπε να τονιστεί μία σημαντική διαφορά. Το γεγονός εάν υπάρχει ή όχι έντονη συνάφεια - συσχέτιση ανάμεσα σε δύο μεταβλητές, δεν συνεπάγεται απαραίτητα και την ύπαρξη μίας συναρτησιακής σχέσης μεταξύ αυτών. Όπως λοιπόν προελέχθη, οι συντελεστές συσχέτισης χωρίζονται σε δύο κατηγορίες. Η πρώτη αναφέρεται σε ποσοτικές μεταβλητές και αφορά το συντελεστή γραμμικής συσχέτισης του Pearson και η δεύτερη αναφέρεται σε ποιοτικές και κατηγορικές μεταβλητές (μεταβλητές των οποίων οι τιμές δεν επιδέχονται ιεράρχηση) και περιλαμβάνει τους συντελεστές Spearman και Kendall (Εθνική Αθλητική Ακαδημία Σόφιας, 2010).

Από την άλλη, ο συντελεστής Point Biserial r_{bis} (***biserial coefficient of correlation***) θεωρείται ως η πιο σύγχρονη μέθοδος ανάλυσης ερωτημάτων που προβλέπει υπολογισμό δεικτών συνάφειας, που φανερώνει το βαθμό συσχέτισης δύο μεταβλητών, στις οποίες η μία έχει δύο μόνο βαθμίδες (διχοτομημένη), και η άλλη είναι συνεχής και αριθμητική. Τόσο αυτός όσο και κάθε άλλος δείκτης της σύγχρονης ανάλυσης ερωτημάτων συγκρίνει την επίδοση των αξιολογούμενων σε κάθε ερώτημα

με κάτι άλλο, το οποίο θεωρείται ως σταθερό σημείο αναφοράς (Τσοπάνογλου, 2008).

Έστω ότι υπάρχουν δύο τυχαίες μεταβλητές ***X, Y*** και ***n*** ζεύγη παρατηρήσεων (x_1, y_1), $(x_2, y_2),\ldots,(x_n, y_n)$ από τυχαίο δείγμα μεγέθους ***n***. Αναφερόμενοι, δηλαδή, σε μη πειραματικά δεδομένα που σημαίνει ότι ο ερευνητής δε μπορεί να προκαθορίσει τις τιμές καμιάς από τις δύο μεταβλητές. Για παράδειγμα:

· Χ το ύψος των φοιτητών ενός πανεπιστημιακού τμήματος και Υ το βάρος τους

· Χ οι ώρες μελέτης των φοιτητών ενός πανεπιστημιακού τμήματος και Υ η απόδοση τους σε ένα τεστ, και

· Χ οι εβδομάδες εμπειρίας ενός εργάτη σε μια επιχείρηση και Υ ο αριθμός των ελαττωματικών προϊόντων που παράγει.

Αντίθετα δεν μπορούν να αναφερθούν περιπτώσεις όπως οι ακόλουθες:

· Χ ο αριθμός των ανοιχτών ταμείων ενός υποκαταστήματος τραπέζης (από τη στιγμή που αυτό το καθορίζει ο διευθυντής και αλλάζει σύμφωνα με τις ανάγκες και επιθυμίες του) και Υ ο χρόνος αναμονής των πελατών (αφού είναι άμεσα συνυφασμένο με το αποτέλεσμα της πρώτης μεταβλητής), και

· Χ το ύψος της διαφημιστικής δαπάνης ενός προϊόντος (που καθορίζεται από μια επιχείρηση ανάλογα πάλι με τις ανάγκες και τον προϋπολογισμό της) και Υ το ύψος των πωλήσεων του προϊόντος.

Σε περιπτώσεις βέβαια όπου υπάρχει τυχαία δειγματοληψία από τον πληθυσμό και σε κάθε μονάδα του δείγματος μελετώνται δύο ή περισσότερα χαρακτηριστικά, είναι λογικό να αναζητηθούν μέτρα τα οποία να μπορούν να παρουσιάσουν και να ποσοτικοποιήσουν τη πιθανή συμμεταβολή-συσχέτιση των χαρακτηριστικών αυτών (Παπαδόπουλος, 2011).

Η επιλογή του συντελεστή συσχέτισης εξαρτάται *α) από το είδος των μεταβλητών* και *β) από τις παραμετρικές τους ιδιότητες*. Τις περισσότερες φορές χρειάζεται να είναι γνωστό τι είδους σχέση υπάρχει ανάμεσα στα δύο μεγέθη. Η συσχέτιση είναι λοιπόν «*ένα μέτρο περιγραφής της γραμμικής εξάρτησης μεταξύ δύο μεταβλητών*» (Σιώμκος, 2005). Στατιστικώς η σχέση μεταξύ δύο μεταβλητών εκφράζεται κυρίως χρησιμοποιώντας το συντελεστή συσχέτισης ***(correlation coefficient)*** ο οποίος υποδεικνύει εάν, με ποιο τρόπο και σε τι βαθμό σχετίζονται δύο μεγέθη, δηλαδή εάν οι τιμές ενός μεγέθους επηρεάζονται από τις τιμές ενός άλλου μεγέθους και πόσο. Θα πρέπει να τονιστεί ωστόσο ότι δεν παρέχει πληροφορία για τη φορά επίδρασης

μεταξύ των δύο μεγεθών. Δηλαδή σε ένα παράδειγμα που βασίζεται στο πόσο μπορεί το αυξημένο άγχος να επηρεάσει την επίδοση ενός μαθητή – φοιτητή και να τη χαμηλώσει βαθμολογικά δεν δύναται λαμβάνοντας υπόψη μόνο αυτές τις δύο μεταβλητές (αυξημένο άγχος – χαμηλή βαθμολογική επίδοση) να προκύψει το εν λόγω συμπέρασμα και άρα δεν είναι εφικτό να ισχυριστεί κανείς πως «*τα υψηλά επίπεδα άγχους προκαλούν χαμηλές βαθμολογικές επιδόσεις*». Αυτό βασικά συμβαίνει επειδή ενδέχεται να υπάρχει και άλλη μεταβλητή που μπορεί να είναι καταγεγραμμένη, μπορεί και όχι, η οποία να επιδρά άμεσα στη σχέση μεταξύ των δύο συγκεκριμένων μεταβλητών και ν' αλλάζει την ισορροπία και το αποτέλεσμά τους. (Βερονίκης, 2011)

Όπως προαναφέρθηκε τα είδη συσχετίσεων είναι τα εξής:

· *Θετική Συσχέτιση (Positive Correlation).* Αυτή η συσχέτιση υφίσταται όταν σχετικά υψηλές τιμές σχετίζονται με σχετικά υψηλές τιμές και σχετικά χαμηλές τιμές σχετίζονται με σχετικά χαμηλές τιμές. Ένα πολύ καλό παράδειγμα για τη συγκεκριμένη περίπτωση είναι η σχέση του μισθού με τα χρόνια της εκπαίδευσης και μόρφωσης. Τα μεγέθη δηλαδή είναι ανάλογα. Όσο αυξάνεται το ένα, αυξάνεται και το άλλο και όσο μειώνεται το ένα, μειώνεται και το άλλο.

· *Αρνητική Συσχέτιση (Negative Correlation).* Αυτή η συσχέτιση υφίσταται όταν σχετικά χαμηλές τιμές σχετίζονται με σχετικά υψηλές τιμές και σχετικά υψηλές τιμές σχετίζονται με σχετικά χαμηλές τιμές. Ένα παράδειγμα αρνητικής συσχέτισης είναι τα χρόνια που κάποιος καπνίζει με το προσδόκιμο ζωής του. Τα μεγέθη δηλαδή είναι αντιστρόφως ανάλογα. Όσο αυξάνεται το ένα, μειώνεται το άλλο και το αντίστροφο.

· *Μηδενική Συσχέτιση (Zero Correlation).* Αυτή η συσχέτιση υφίσταται όταν δεν υπάρχει βασικά καθόλου σχέση ανάμεσα στα μεγέθη και στα δεδομένα. Τα παραδείγματα μπορούν να είναι άπειρα, όπως η σχέση ανάμεσα στα χρόνια της εκπαίδευσης και στο προσδόκιμο ζωής. Κανείς δεν μπορεί να συσχετίσει πως όσο αυξάνεται η μόρφωση, αυξάνεται ή μειώνεται το προσδόκιμο ζωής με αποτέλεσμα να υπάρχει μηδενική συσχέτιση (Koltko-Rivera, 2012).

Υπάρχουν κάποια βασικά χαρακτηριστικά του συντελεστή συσχέτισης και της μέτρησης των εκάστοτε μεταβλητών παραθέτοντας τα επακριβώς για αποφυγή παρανοήσεων και λαθών «…

1. *Παίρνει τιμές συγκεκριμένες,*
2. *Είναι 1 μόνο όταν όλες οι τιμές «πέφτουν» επάνω στην ευθεία γραμμή,*

3. *Αν προστεθεί μια σταθερά σε όλες τις τιμές X ή σε όλες τις τιμές Y ο συντελεστής δεν αλλάζει,*
4. *Αν όλες οι τιμές του X ή του Y πολλαπλασιαστούν με μια σταθερά, ο συντελεστής δεν αλλάζει (εκτός από το πρόσημό του αν η σταθερά έχει διαφορετικό πρόσημο από τις τιμές X ή Y),*
5. *Επηρεάζεται από την αξιοπιστία των μετρήσεων,*
6. *Για ιεραρχικά δεδομένα χρησιμοποιείται κυρίως η φόρμουλα του Spearman;*
7. *Ο συντελεστής Point Biserial μετριέται με μία κατηγορική και μία συνεχή μεταβλητή, και*
8. *Οι μεταβλητές μπορούν επίσης να εκφράζονται ως σταθερές τιμές (π.χ. Standard scores, normal curve equivalent scores, etc.)…»* (Σιδερίδης, 2012).

Επειδή, τόσο η στατιστική συνάρτηση *rho* (Spearman) όσο και η *τ* (Kendall) αποτελούν αθροίσματα τυχαίων μεταβλητών, μπορεί κανείς να χρησιμοποιήσει μια μορφή του κεντρικού οριακού θεωρήματος[10] για να προσεγγίσει τις κατανομές τους στην περίπτωση μεγάλων δειγμάτων. Και οι δύο συντελεστές έχουν συμμετρικές κατανομές γύρω από το μηδέν και, συνεπώς, έχουν και οι δύο μέση τιμή ίση με το μηδέν. Από την άλλη, οι διασπορές των στατιστικών αυτών συναρτήσεων είναι πιο δύσκολο να προσδιορισθούν. Έτσι τυχόν διαίρεσή τους με τις αντίστοιχες διασπορές τους οδηγεί σε τυχαίες μεταβλητές (Stuart, 1954), οι οποίες έχουν την κανονική κατανομή για μεγάλες τιμές του n (Οικονομικό Πανεπιστήμιο Αθηνών, 2012).

3.2 Συντελεστής του Spearman

Ο συντελεστής συσχέτισης Spearman, πήρε το όνομά του από τον Charles Spearman και συχνά συμβολίζεται με το ελληνικό γράμμα ρ (rho) ή ως r_s. Ο συγκεκριμένος συντελεστής είναι ένα μη-παραμετρικό μέτρο της στατιστικής εξάρτησης μεταξύ δύο μεταβλητών (X, Y) και στην πραγματικότητα αξιολογεί το πόσο καλά περιγράφεται η σχέση μεταξύ των δύο αυτών μεταβλητών χρησιμοποιώντας μια μονότονη συνάρτηση. Σε περίπτωση που δεν υπάρχουν επαναλαμβανόμενες τιμές των

[10] Το θεώρημα αυτό αποδεικνύει ότι η κατανομή των μέσων τυχαίων δειγμάτων με μέγεθος σχετικά μεγάλο, πρακτικά μεγαλύτερο ή ίσο του τριάντα, δεν εξαρτάται από την κατανομή του αρχικού πληθυσμού αλλά ακολουθεί, κάτω από ορισμένες προϋποθέσεις, οι οποίες συνήθως πληρούνται, την κανονική κατανομή. Όταν το μέγεθος του δείγματος είναι μικρότερο του τριάντα, για να ακολουθεί η κατανομή δειγματοληψίας των μέσων την κανονική κατανομή, θα πρέπει και ο πληθυσμός από τον οποίο προέρχεται το δείγμα να ακολουθεί την κανονική κατανομή. **Κεντρικό Οριακό Θεώρημα:** Αν X_1, X_2,...,Xn είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές (δηλ. έχουν την ίδια κατανομή πιθανοτήτων) με μέση τιμή μ και πεπερασμένη διακύμανση σ^2, τότε, όσο αυξάνει το n, η κατανομή της τυχαίας μεταβλητής: $Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ προσεγγίζει την τυπική κανονική κατανομή (Keller, 2009)

δεδομένων, μια τέλεια συσχέτιση Spearman κατά +1 ή -1 λαμβάνει χώρα όταν κάθε μία από τις μεταβλητές είναι μια τέλεια μονότονη συνάρτηση της άλλης. Στο ερώτημα «*όταν υπάρχουν κατηγορικές μεταβλητές και θέλουμε να ελέγξουμε την συσχέτιση μεταξύ τους, ποιο συντελεστή είναι σοφότερο να χρησιμοποιήσουμε;*» η απάντηση είναι μία «*τη γραμμική συσχέτιση μεταξύ μεταβλητών κατά Spearman*».

Όταν οι παραμετρικές προϋποθέσεις (η κανονικότητα, η γραμμικότητα, το εύρος των παρατηρήσεων και η ύπαρξη ισοδιαστημικής κλίμακας) δεν ικανοποιούνται, τότε είναι ανάγκη να χρησιμοποιηθούν εναλλακτικοί στατιστικοί δείκτες για την ανίχνευση σχέσεων μεταξύ μεταβλητών. Ένας από αυτούς είναι και ο δείκτης συσχέτισης του Spearman, ο οποίος υπολογίζεται μετατρέποντας τα δεδομένα σε «σειρές» με βάση το μέγεθος τους (οι αρχικές τιμές μπαίνουν σε σειρά με βάση το μέγεθός τους, π.χ. πρώτος, δεύτερος, κλπ.). Με αυτόν τον τρόπο, οι αποστάσεις μεταξύ των παρατηρήσεων χάνουν τη σημασία τους και αξιολογείται η σειρά των συμμετεχόντων στην πρώτη μεταβλητή σε σχέση με τη σειρά που αυτοί έχουν στην δεύτερη μεταβλητή και τα λοιπά. Το πρόσημο αλλά και το μέγεθος της σχέσης εκφράζονται από το μέγεθος της συμφωνίας ή όχι της σειράς στις δύο μεταβλητές. Τις περισσότερες φορές οι μη-παραμετρικοί δείκτες είναι πιο «ακριβείς» στα αποτελέσματά τους αφού στοχεύουν στο να διορθώσουν πιθανά προβλήματα που προκαλούνται από τις καταπατήσεις των προϋποθέσεων. (Εμβαλωτής, Κατσής & Σιδερίδης, 2006).

Ο συντελεστής Spearman στην ουσία είναι ο συντελεστής Pearson και για το δείγμα του πληθυσμού δίνεται από τον τύπο: $r_s = \frac{s_{ab}}{s_a \cdot s_b}$, όπου a και b είναι οι τάξεις των μεταβλητών x και y αντίστοιχα στον πίνακα κατάταξης, s_{ab} είναι η συμμεταβλητότητα των τάξεων, και s_a, s_b είναι οι τυπικές αποκλίσεις των τάξεων. Συγκεκριμένα, ο συντελεστής συσχέτισης ρ (rho) δίνεται και από τον τύπο rho = 1- $\frac{6\sum d_i^2}{n(n^2-1)}$ υπολογιζόμενος όμως στις σχετικές θέσεις/τάξεις (ranks) των τιμών x_i και y_i , όπου n το μέγεθος του δείγματος και $d_i = x_i - y_i$, i = 1,2,...., n. Για παράδειγμα υπάρχουν οκτώ φοιτητές (n=8) και η σειρά κατάταξής τους όσον αφορά την απόδοσή τους στα μαθηματικά. Για να υπολογιστεί ο συντελεστής συσχέτισης Spearman δημιουργείται ο παρακάτω πίνακας 8, όπου εμφανίζεται ο αριθμός κατάταξης τους αλλά και σύμφωνα με την απόδοσή τους στο συγκεκριμένο μάθημα. Εν συνεχεία υπολογίζεται η διαφορά και τελικά το d_i^2. Αναλυτικότερα:

ΦΟΙΤΗΤΕΣ i	Σειρά κατάταξης στη Στατιστική (X)	Σειρά κατάταξης στα Μαθηματικά (Y)	d_i	d_i^2
Α	1ος	4ος	-3	9
Β	2ος	2ος	0	0
Γ	3ος	3ος	0	0
Δ	4ος	5ος	-1	1
Ε	5ος	1ος	4	16
ΣΤ	6ος	6ος	0	0
Ζ	7ος	8ος	-1	1
Η	8ος	7ος	1	1
				$\Sigma d_i^2 = 28$

Πίνακας 8. Δεδομένα φοιτητών

Από τον συνολικό υπολογισμό προκύπτει ότι: rho = 1- $\frac{6\sum d_i^2}{n(n^2-1)}$ = 0,667. Το πρόσημο της συσχέτισης Spearman δείχνει την κατεύθυνση της σχέσης μεταξύ της *X* (ανεξάρτητη μεταβλητή) και της *Y* (εξαρτημένη μεταβλητή). Ο συντελεστής συσχέτισης Spearman είναι θετικός εάν, όταν η *X* αυξάνει, τείνει να αυξάνεται και η *Y*. Από την άλλη, ο συντελεστής συσχέτισης Spearman είναι αρνητικός εάν, όταν η *X* αυξάνει, η *Y* τείνει να μειώνεται. Τέλος, όταν η *X* αυξάνει αλλά δεν υπάρχει τάση για την *Y* είτε να αυξηθεί ή να μειωθεί, έχουμε μια μηδενική συσχέτιση Spearman. Όταν η *X* και η *Y* είναι πιο κοντά στο να είναι τέλειες μονότονες συναρτήσεις η μία της άλλης, η συσχέτιση Spearman αυξάνει σε μέγεθος ενώ ο συντελεστής γίνεται 1 όταν η *X* και η *Y* έχουν απόλυτη μονοτονική σχέση (Παπαδόπουλος, 2011). Ο συντελεστής συσχέτισης Spearman συνήθως περιγράφεται ως «μη παραμετρικός» αποκτώντας διττή σημασία. Πρώτον, το γεγονός ότι μια τέλεια συσχέτιση Spearman προκύπτει όταν *X* και *Y* σχετίζονται με οποιαδήποτε μονότονη συνάρτηση, που μπορεί να αντιπαραβληθεί με τη συσχέτιση Pearson, η οποία δίνει μόνο μια τέλεια τιμή όταν *X* και *Y* σχετίζονται με μια γραμμική συνάρτηση. Δεύτερον, ο συσχετισμός Spearman είναι μη παραμετρικός καθότι η ακριβής κατανομή της δειγματοληψίας του μπορεί να ληφθεί χωρίς να απαιτείται γνώση της κοινής κατανομής πιθανότητας της *X* και *Y*. (Γναρδέλλης, 2003)

Ο συντελεστής συσχέτισης rho του Spearman χρησιμοποιείται ως μέτρο του βαθμού συμφωνίας της σειράς κατά την οποία διατάσσονται οι τιμές ενός χαρακτηριστικού - όταν αυτές δεν ακολουθούν την κανονική κατανομή - μετά από επαναληπτικές μετρήσεις. Πρόκειται στην ουσία (όπως προαναφέρθηκε) για μια γενίκευση του συντελεστή Pearson ο οποίος εκφράζει τη συσχέτιση μεταξύ των διατεταγμένων θέσεων μιας παρατήρησης βάσει δύο χαρακτηριστικών. Οι παρατηρήσεις θα πρέπει ωστόσο να διατάσσονται είτε με αύξουσα είτε με φθίνουσα

σειρά, κατά τον ίδιο τρόπο, ώστε να μην αλλάζει το πρόσημο της συσχέτισης. Ο συντελεστής συσχέτισης Spearman λαμβάνει τιμές στο διάστημα [-1,+1]. Με αυτόν τον τρόπο οι τιμές κοντά στο -1 δείχνουν τέλεια αρνητική συσχέτιση, οι τιμές κοντά στο +1 δείχνουν τέλεια θετική συσχέτιση και οι τιμές κοντά στο 0 φανερώνουν την απουσία σχέσης. Ουσιαστικά, ο συντελεστής συσχέτισης Spearman χρησιμοποιείται για τον προσδιορισμό της συνάφειας μεταξύ δύο μεταβλητών οι οποίες εκφράζουν τη σειρά κατάταξης σε δύο αντίστοιχες μεταβλητές. Οι συντελεστές αυτοί όμως μειονεκτούν στη διερεύνηση του βαθμού συμφωνίας των μετρήσεων μεταξύ δύο χαρακτηριστικών.

Για την κατανόηση των συντελεστών συσχέτισης συνίσταται τις περισσότερες φορές να δίνονται κατάλληλα παραδείγματα, όπως αυτό που αφορά σε κλινικές διατροφικές έρευνες. Συγκεκριμένα, τα τελευταία χρόνια, εστιάζεται η έρευνα στην αποτίμηση της σχέσης της διατροφής με την εμφάνιση χρόνιων νοσημάτων και το πιο σύνηθες εργαλείο για την επίτευξη της αξιολόγησης των διατροφικών συνηθειών, σε ατομικό επίπεδο, είναι το ημι-ποσοτικοποιημένο ερωτηματολόγιο συχνότητας κατανάλωσης τροφίμων (ΕΣΚΤ). Η αποτελεσματικότητά του ερωτηματολογίου αυτού έγκειται στη δυνατότητα εξαγωγής αποτελεσμάτων με εσωτερική και εξωτερική ακρίβεια. Ο έλεγχος του βαθμού επαναληψιμότητας, ώστε να αποδειχθεί και η εγκυρότητα του διατροφικού εργαλείου, γίνεται με διάφορες στατιστικές μεθόδους και κυρίως με τη χρήση των συντελεστών συσχέτισης (π.χ. Pearson και Spearman) καθώς και έλεγχοι μέσω κριτηρίων ελέγχου συμφωνίας των κατανομών, όπως το στατιστικό κριτήριο *τ* του Kendall. Βέβαια οι συντελεστές συσχέτισης δεν αποτελούν πάντα την κατάλληλη μεθοδολογική προσέγγιση για τον έλεγχο της επαναληψιμότητας, από τη στιγμή που μπορεί μεν να δείχνουν υψηλή γραμμική συσχέτιση, αλλά ο βαθμός συμφωνίας μεταξύ των δύο μετρήσεων μπορεί να είναι χαμηλός. Τέλος, τα κριτήρια συμφωνίας μειονεκτούν στο να μπορέσουν να αναδείξουν πιθανή μεροληψία στις καταγραφές των ατόμων (Μπουντζιούκα & Παναγιωτάκος, 2009).

3.3 Συντελεστής Συσχέτισης του Kendall

Υπάρχουν δύο βασικά σημεία κατά τα οποία ο συντελεστής συσχέτισης *τ* του Kendall μοιάζει με τον συντελεστή ρ του Spearman. Πρώτον, υπολογίζεται με βάση την τάξη μεγέθους των παρατηρήσεων και όχι με βάση τις παρατηρήσεις αυτές καθαυτές και, δεύτερον, η κατανομή του δεν εξαρτάται από την κατανομή των μεταβλητών *X* και *Y*, όταν αυτές είναι ανεξάρτητες και συνεχείς. Αναφορικά με τον εν λόγω συντελεστή

ορίζεται ως εξής: «*Δύο παρατηρήσεις, έστω* (x_j, y_j) *και* (x_k, y_k)*, ονομάζονται εναρμονισμένες ή συσχετισμένες (concordant), αν και τα δύο μέλη της μίας παρατήρησης είναι μεγαλύτερα (ή μικρότερα) από τα αντίστοιχα μέλη της άλλης παρατήρησης. Δηλαδή, αν* $x_j > x_k$ *(αντίστοιχα,* $x_j < x_k$ *), τότε* $y_j > y_k$ *(αντίστοιχα,* $y_j < y_k$*). Οι παρατηρήσεις* (x_j , y_j) *και* (x_k , y_k) *θα ονομάζονται μη εναρμονισμένες ή μη συσχετισμένες (discordant), αν η διάταξη των πρώτων μελών τους είναι αντίθετη από την διάταξη των δεύτερων μελών τους, δηλαδή, αν* $x_j > x_k$ *(αντίστοιχα,* $x_j < x_k$*), τότε* $y_j < y_k$ *(αντίστοιχα,* $y_j > y_k$*). Ισοδύναμα, δύο ζεύγη παρατηρήσεων* (x_j , y_j) *και* (x_k, y_k) *θα ονομάζονται εναρμονισμένα αν οι διαφορές* $x_j - x_k$ *και* $y_j - y_k$ *έχουν το ίδιο πρόσημο (αν* $(x_j - x_k)(y_j - y_k) > 0$*). Τα ζεύγη* (x_j, y_j) *και* (x_k, y_k) *θα ονομάζονται μη εναρμονισμένα αν οι διαφορές* $x_j - x_k$ *και* $y_j - y_k$ *έχουν αντίθετο πρόσημο [αν* $(x_j - x_k)(y_j - y_k) < 0$*]. Έστω* n_c *και* n_d *οι αριθμοί των εναρμονισμένων και μη εναρμονισμένων ζευγών παρατηρήσεων, αντίστοιχα. Τα ζεύγη των παρατηρήσεων* (x_j , y_j) *και* (x_k, y_k)*, για τα οποία ισχύει ότι* $x_j = x_k$ *ή/και* $y_j = y_k$ *, δεν είναι ούτε εναρμονισμένα ούτε μη εναρμονισμένα. Τα ζεύγη αυτά ονομάζονται ισοβαθμούντα (tied)*» (Οικονομικό Πανεπιστήμιο Αθηνών, 2012).

Με περισσότερες λεπτομέρειες ο συντελεστής τ παριστάνει την διαφορά μεταξύ των ποσοστών των εναρμονισμένων και μη εναρμονισμένων ζευγών παρατηρήσεων. Ο συντελεστής τ είναι ίσος με 1 αν όλα τα ζεύγη παρατηρήσεων είναι εναρμονισμένα. Αντίθετα, η τιμή του συντελεστή τ είναι –1 αν όλα τα ζεύγη είναι μη εναρμονισμένα. Οι τιμές δηλαδή του συντελεστή τ μεταξύ –1 και 1. Επιπρόσθετα, ο συντελεστής *τ* ικανοποιεί όλες τις προϋποθέσεις που αναφέρθηκαν στον παραπάνω ορισμό. (Οικονομικό Πανεπιστήμιο Αθηνών, 2012)

Ένα ακόμη χαρακτηριστικό του συντελεστή τ είναι ότι αξιολογεί το βαθμό συμφωνίας μεταξύ δύο διατεταγμένων ομάδων παρατηρήσεων οι οποίες προέρχονται από τον ίδιο πληθυσμό ο οποίος εξαρτάται από τον αριθμό των διαφωνούντων ζευγών στη σειρά της κατάταξης. Έτσι, ο συντελεστής τ του Kendall μπορεί επίσης να ερμηνευθεί ως: «*ο συντελεστής συσχέτισης μεταξύ δύο συνόλων* $n(n-1)$ *δίτιμων παρατηρήσεων όπου κάθε σύνολο αντιστοιχεί σε όλους τους πιθανούς συνδυασμούς της διάταξης των* n *παρατηρήσεων, λαμβάνοντας την τιμή 1 όταν ένα ζεύγος τιμών εμφανίζεται στη σειρά και την τιμή 0 αν δεν εμφανίζεται. Ο συντελεστής τ του Kendall βασίζεται στην απόσταση των ζευγών μεταξύ τους και εκφράζει τη διαφορά στη πιθανότητα τα διατεταγμένα ζεύγη να συμφωνούν από την πιθανότητα τα διατεταγμένα*

ζεύγη να διαφωνούν» (Μπουντζιούκα & Παναγιωτάκος, 2009). Ο βασικός τύπος του συντελεστή συσχέτισης *τ* του Kendall είναι ο: $\tau = \frac{2 \cdot k}{n \cdot (n-1)}$.

Το κύριο πλεονέκτημα του εν λόγω συντελεστή σε σχέση με το συντελεστή ρ του Spearman είναι ότι τείνει στην κανονική κατανομή σχετικά γρήγορα και η προσέγγισή του είναι καλύτερη από την αντίστοιχη προσέγγιση της κατανομής του συντελεστή ρ του Spearman, με βασική προϋπόθεση να αληθεύει η μηδενική υπόθεση της ανεξαρτησίας μεταξύ των μεταβλητών *X* και *Y*. Τέλος, ένα ακόμη σημαντικό πλεονέκτημα του συντελεστή *τ* του Kendall είναι ότι μπορεί άμεσα να ερμηνευθεί μέσω των πιθανοτήτων με τις οποίες παρατηρούνται *εναρμονισμένα ή συσχετισμένα (concordant)* ζεύγη τιμών και *μη εναρμονισμένα ή μη συσχετισμένα (discordant)* ζεύγη τιμών (Οικονομικό Πανεπιστήμιο Αθηνών, 2012).

Ένα εξαιρετικό παράδειγμα για τους δύο συντελεστές συσχέτισης Spearman και Kendall είναι ο παρακάτω συγκριτικός πίνακας (Πίνακας 9) στον οποίο παρουσιάζονται σε επιλεγμένα τρόφιμα (αντιπροσωπευτικά της διατροφής του ελληνικού πληθυσμού) οι συντελεστές συσχέτισης κατά Spearman και Kendall. Παρατηρείται λοιπόν ότι ο συντελεστής Spearman κυμαίνεται (στο συνολικό πληθυσμό) από 0,54 για την κατανάλωση πουλερικών έως 0,83 για την κατανάλωση δημητριακών πρωινού. Ο συντελεστής Kendall για τα ίδια τρόφιμα είναι 0,49 και 0,75 αντίστοιχα. Έτσι παρατηρείται η διαφορά στη σειρά κατάταξης των τροφίμων αλλά και η γενικότερη διαφορά μεταξύ των δύο καταγραφών για τα τρόφιμα αυτά, τόσο για το σύνολο του δείγματος όσο και μεταξύ των δύο φύλων. Βέβαια από τον πίνακα καθίσταται εμφανές το ότι οι δύο συντελεστές ενίοτε συσχετίζονται μα δε σημαίνει απαραίτητα ότι πρέπει και να συμφωνούν.

ΤΡΟΦΙΜΑ	**SPEARMAN'S RHO**			**KENDALL'S TAU**		
	Α	**Γ**	**Σ**	**Α**	**Γ**	**Σ**
Γάλα / Γιαούρτι	0,76	0,72	0,73	0,69	0,65	0,67
Τυρί φέτα	0,69	0,76	0,73	0,61	0,67	0,65
Αυγό	0,68	0,63	0,65	0,63	0,57	0,59
Άσπρο Ψωμί	0,72	0,76	0,75	0,63	0,67	0,66
Δημητριακά πρωινού	0,85	0,81	0,83	0,78	0,72	0,75
Λευκό ρύζι	0,52	0,61	0,58	0,49	0,57	0,54
Ζυμαρικά	0,54	0,61	0,59	0,5	0,57	0,55
Μοσχάρι	0,64	0,65	0,65	0,59	0,61	0,53
Πουλερικά	0,61	0,5	0,54	0,56	0,46	0,49
Χοιρινό	0,66	0,61	0,66	0,62	0,56	0,61
Ψάρια	0,6	0,61	0,65	0,56	0,57	0,57
Όσπρια	0,54	0,66	0,61	0,51	0,62	0,58
Τομάτα, Αγγούρι	0,57	0,64	0,62	0,51	0,57	0,55
Πράσινα φυλλώδη Λαχανικά	0,6	0,62	0,63	0,53	0,55	0,55
Βραστά λαχανικά	0,57	0,57	0,57	0,51	0,51	0,51

Πορτοκάλι	0,76	0,72	0,73	0,68	0,65	0,66
Μήλο	0,79	0,75	0,77	0,7	0,67	0,68
Μπανάνα	0,79	0,67	0,72	0,71	0,59	0,63
Καλοκαιρινά φρούτα	0,62	0,54	0,57	0,55	0,47	0,5
Κρασί	0,68	0,79	0,76	0,62	0,66	0,7

Πίνακας 9. Συγκριτικός πίνακας τροφίμων και αντίστοιχων συντελεστών, όπου Α= άνδρες, Γ= γυναίκες, Σ= συνολικός πληθυσμός (Μπουντζιούκα & Παναγιωτάκος, 2012)

3.4 Ο Συντελεστής Biserial

Ο τύπος εύρεσης του συντελεστή Biserial (r_{pbi}) είναι ο: $r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq}$. Ο συντελεστής point-biserial ή αλλιώς και συντελεστής σημειακής δισειριακής συσχέτισης, είναι ένας ειδικός τύπος που συσχετίζει την παρατηρούμενη αντίδραση ενός αντικειμένου με το τελικό αποτέλεσμα ενός τεστ. Χρησιμοποιείται ειδικότερα όταν μια ομάδα δεδομένων είναι εκ φύσεως διχοτομημένη. Σε αυτή την περίπτωση, τα στοιχεία που βρίσκονται στο αποτέλεσμα ενός τεστ πολλαπλής επιλογής είναι διχοτομημένα δεδομένα, για παράδειγμα μπορούν να πάρουν αξία 1 (για κάθε σωστή απάντηση) και αξία 0 (για κάθε λάθος απάντηση). Ο υπολογισμός του συντελεστή point-biserial είναι απλούστερος του συντελεστή του Pearson αλλά τα αποτελέσματα της συσχέτισης και η ερμηνεία τους είναι τα ίδια. Με άλλα λόγια, ένας θετικός συντελεστής point-biserial θα σήμαινε ότι οι υψηλές τιμές στα διχοτομημένα δεδομένα έχουν άμεση σχέση με τις υψηλές τιμές στο τελικό αποτέλεσμα του τεστ. Για αυτό και ονομάζεται και συντελεστής *«συσχέτισης αντικειμένου-τελικού αποτελέσματος (item-total correlation)»* (Keller, 2009).

Ο συντελεστής point-biserial, υπολογιζόμενος για κάθε αντικείμενο ενός τεστ πολλαπλής επιλογής, θεωρείται πολύ χρήσιμος καθότι αντικατοπτρίζει το πόσο καλά ένα αντικείμενο «ξεχωρίζει» από τα υπόλοιπα. Μία υψηλή συσχέτιση point-biserial σημαίνει ότι για παράδειγμα οι μαθητές οι οποίοι επιλέγουν σε ένα τεστ τη σωστή απάντηση είναι και αυτοί που έχουν υψηλότερα τελικά αποτελέσματα. Αντίθετα, οι μαθητές που επιλέγουν τις λάθος απαντήσεις σε μία ερώτηση σχετίζονται με χαμηλότερα τελικά αποτελέσματα. Δεδομένου των προαναφερθέντων, το αντικείμενο «ξεχωρίζει» τους εξεταζόμενους χαμηλής απόδοσης από τους εξεταζόμενους υψηλής απόδοσης. Και αυτό φυσικά είναι και ένα επιθυμητό χαρακτηριστικό των τεστ με ερωτήσεις.

Τέλος, τα αντικείμενα τα οποία θεωρήθηκαν εσφαλμένα μπορούν να αναγνωριστούν υπολογίζοντας ξανά με νέα δεδομένα πολύ χαμηλές ή αρνητικές συσχετίσεις point-biserial. Όπως προαναφέρθηκε στον υπολογισμό του point-biserial

υπάρχει μια διχοτομημένη μεταβλητή και μία συνεχής και αριθμητική. Συνεχής είναι στο παράδειγμά μας ο αριθμός των ωρών που ένας μαθητής διαβάζει για ένα διαγώνισμα και μπορούν να εκτείνονται από 0 έως 90 ώρες την εβδομάδα. Διχοτομημένη είναι μια ερώτηση του τύπου «πέρασε ο μαθητής το διαγώνισμα ή όχι;» οπότε και υπάρχουν δύο βαθμίδες που χρήζουν αναλύσεως. Πρέπει να τονιστεί βέβαια ότι σε κάθε στατιστική μέθοδο υπάρχουν περιορισμοί. Για παράδειγμα, ο συντελεστής point-biserial είναι άμεσα εξαρτώμενος από το δείγμα που χρησιμοποιείται κάθε φορά. Τοποθετώντας ένα αντικείμενο σε διαφορετική ομάδα αλλάζουν οι αξίες αυτόματα οπότε πρέπει να γίνει εκ νέου υπολογισμός του συντελεστή (Measured Progress Assessment Firm's official site, 2012).

Ένα επίσης αξιόλογο παράδειγμα για τη χρήση του συντελεστή point-biserial είναι η συσχέτιση μεταξύ των αναγνωστικών ενδιαφερόντων των γονέων και της βαθμολογίας του παιδιού στο σχολείο. Προκειμένου να διερευνηθεί η συσχέτιση αυτή μπορεί να χρησιμοποιηθεί ο δείκτης συνάφειας Point-Biserial. Παρατηρούνται λοιπόν σποραδικές συνάφειες, οι οποίες αποδεικνύουν ότι η επίδοση των παιδιών στο πρώτο τρίμηνο αυξάνεται γενικότερα (point-biserial = 0,22), επίσης αυξάνεται στα Νέα Ελληνικά (0,21) και στην Έκθεση (0,22) αλλά και στη γενική βαθμολογία του δευτέρου τριμήνου (0,24) όσο περισσότερο οι γονείς διαβάζουν επιστημονικά συγγράμματα. Επιπρόσθετα, όσο οι γονείς διαβάζουν λογοτεχνικά βιβλία, τόσο αυξάνεται και η επίδοση των παιδιών στα Αρχαία Ελληνικά (point-biserial = 0,22) αλλά και στα Νέα Ελληνικά (0,21). Βέβαια, αν και οι δείκτες αυτοί συνάφειας είναι σχετικά μικροί, το συστηματικό μοτίβο που παρατηρείται ως προς το συνδυασμό της επίδοσης του παιδιού στο σχολείο και των αναγνωστικών ενδιαφερόντων των γονέων του, επιτείνεται από τους αντίστροφους δείκτες συνάφειας point-biserial που παρατηρούνται όταν οι γονείς δηλώνουν ότι δεν διαβάζουν τίποτε. Οι αντίστοιχοι δείκτες για την επίδοση πρώτου και δευτέρου τριμήνου, τα Αρχαία και τα Νέα Ελληνικά καθώς και την Έκθεση κυμαίνονται μεταξύ (-0,21 και -0,27), αποδεικνύοντας ότι όντως μειώνεται η επίδοση των παιδιών όσο οι γονείς δηλώνουν ότι δεν διαβάζουν τίποτε. Μόνο στα μαθηματικά δεν παρατηρούνται αντίστοιχες συνάφειες πράγμα που σημαίνει ότι τα αναγνωστικά ενδιαφέροντα των γονέων δεν φαίνεται να επηρεάζουν τη βαθμολογία των παιδιών στα μαθηματικά (Κατσαμάγκου, 2003).

3.5 Μειονεκτήματα Των Συντελεστών Συσχέτισης

Οι συντελεστές συσχέτισης είναι μέτρα που δηλώνουν αν τα αποτελέσματα δύο μετρήσεων που πραγματοποιήθηκαν με δύο διαφορετικές μεθόδους ή από δύο διαφορετικά άτομα, σχετίζονται. Δηλαδή, αν πιθανές αλλαγές στη μία μέτρηση επηρεάζουν την άλλη. Αυτό όμως δεν αποδεικνύει και το βαθμό της ομοιότητας των διανυσμάτων των αποτελεσμάτων τους.

Ο βασικός ρόλος των συντελεστών συσχέτισης είναι να μετράνε το πόσο ισχυρή είναι η σχέση δύο μεταβλητών. Γραφικά, η απόδειξη μεγάλης συσχέτισης είναι τα δεδομένα να βρίσκονται πάνω σε οποιαδήποτε ευθεία. Αυτό όμως δε σημαίνει ότι τα ζεύγη των τιμών είναι ζεύγη της μορφής (x,x). Κάτι τέτοιο θα συγκέντρωνε τις τιμές γύρω μόνο από την ευθεία y=x. Επιπροσθέτως, μπορεί μια αλλαγή στη μονάδα μέτρησης να μη μεταβάλλει το συντελεστή συσχέτισης, επηρεάζει όμως τη συμφωνία. Άρα, παρότι μπορεί δύο μεταβλητές να φαίνονται ότι είναι ισχυρά συσχετισμένες, εντούτοις, μιας και έχουν μετρηθεί με διαφορετικές μονάδες, δεν είναι εφικτό να ελεγχθεί η συμφωνία τους.

Επίσης, ως γνωστόν, ο συντελεστής συσχέτισης εξαρτάται και από το εύρος του δείγματος. Μεγάλες τιμές συσχέτισης τείνουν να προέρχονται από μεταβλητές μεγάλου μεγέθους. Στις ιατρικές έρευνες όμως για παράδειγμα, είναι κάτι το σύνηθες να υπάρχουν μεγάλα δείγματα αφού οι επιστήμονες επιλέγουν να καλύπτουν όσο το δυνατόν μεγαλύτερο φάσμα περιπτώσεων. Έτσι, μπορεί η σχέση των μεθόδων να φαίνεται ισχυρή, αυτό όμως δεν υποδηλώνει αναγκαστικά και τη συμφωνία τους.

Η διακύμανση των τιμών των μεταβλητών και το σφάλμα των μετρήσεων είναι ένας ακόμη παράγοντας που αναμφίβολα μεταβάλλει την τιμή των συντελεστών συσχέτισης. Όταν η διακύμανση των μετρήσεων είναι υψηλή σε σύγκριση με το στατιστικό λάθος τότε και ο συντελεστής θα έχει μεγάλη τιμή. Αντιθέτως, όταν η διακύμανση είναι μικρή σε σύγκριση με το στατιστικό λάθος τότε η συσχέτιση μεταξύ των μεταβλητών θα είναι μικρή. Τέλος, η τιμή της συσχέτισης επηρεάζεται και από τον τρόπο επιλογής του δείγματος και από την κατανομή των μεταβλητών. Θα πρέπει να τονιστεί δε, ότι στον υπολογισμό της τιμής δεν λαμβάνεται υπ' όψιν η ροπή μεταξύ των μεταβλητών. Όταν όμως συγκρίνονται δύο διαφορετικοί τρόποι μέτρησης, αυξάνεται η πιθανότητα να υπάρχει ροπή και έτσι ο συντελεστής συσχέτισης θα οδηγήσει σε λάθος συμπεράσματα.

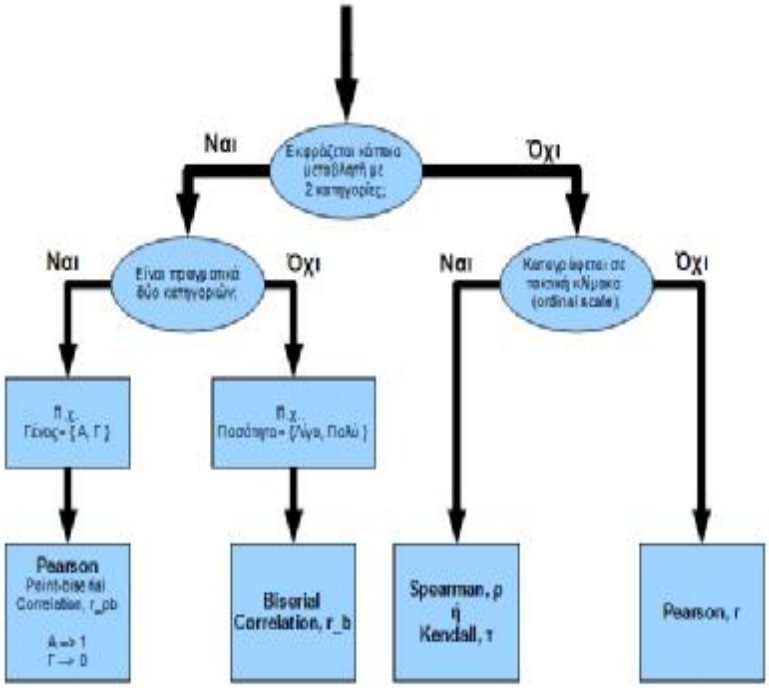
Αναμφίβολα η συσχέτιση μπορεί να κρίνει την εγκυρότητα δύο μεθόδων μέτρησης, αλλά δεν μπορεί να απαντήσει στο αν οι μέθοδοι αυτοί είναι το ίδιο αποδοτικοί ή στο αν θα μπορούσαν να χρησιμοποιηθούν εναλλακτικά. Σύμφωνα μάλιστα με τους Hallman και Teramo (1981) «…*ένας συντελεστής συσχέτισης κοντά στο 1 δεν θα μπορούσε να μας πει ότι η καινούρια μέθοδος θα μπορούσε να αντικαταστήσει την παλιά δίνοντας αξιόπιστα αποτελέσματα…*» (Ευαγγέλου, 2008).

3.6 Σύνοψη

Έχοντας αποσαφηνίσει τις έννοιες περί συσχετίσεων δύναται να συνοψιστούν κάποιες βασικές αρχές σύμφωνα με τον Βερονίκη (2011):

1. Ο συντελεστής συσχέτισης βοηθάει να μελετηθεί και να υπολογιστεί στατιστικώς το μέγεθος αλληλεξάρτησης μεταξύ δύο μεταβλητών,
2. Δεδομένου των αρκετών συντελεστών συσχέτισης που εμφανίζονται βιβλιογραφικά, το κριτήριο επιλογής του εκάστοτε εξαρτάται από το είδος των μεταβλητών, την κλίμακα αποτίμησης και το είδος της κατανομής του,
3. Οι τιμές που λαμβάνει ο συντελεστής συσχέτισης βρίσκονται στο διάστημα [-1,1],
4. Για τον υπολογισμό της τιμής του συντελεστή αναφέρεται και το επίπεδο στατιστικής σημαντικότητας, και
5. Ο βαθμός συσχέτισης μεταξύ δύο μεταβλητών μπορεί να επηρεάζεται από μία ή και περισσότερες άλλες μεταβλητές που έχουν χρησιμοποιηθεί για τον σκοπό αυτό.

Επιπρόσθετα, σύμφωνα με τον Βερονίκη (2011) παρουσιάζεται στην Εικόνα 9, ένα κατάλληλο πρότυπο γράφημα για τα κριτήρια επιλογής του συντελεστή συσχέτισης.



Εικόνα 9. Διαγραμματική Αναπαράσταση για τα αλγοριθμικά κριτήρια επιλογής του κατάλληλου συντελεστή συσχέτισης

Τέλος, σύμφωνα με τον Σιδερίδη (2012) ο συντελεστής συσχέτισης πρέπει να χρησιμοποιείται με δέουσα προσοχή ή κάποιες φορές καλό είναι να μην χρησιμοποιείται καν. Οι λόγοι που πρέπει να συνεκτιμώνται στην απόφαση αυτή είναι όταν:

1. Οι μεταβλητές έχουν περιορισμένο εύρος (μειώνεται η τιμή του συντελεστή),
2. Είναι διαθέσιμα μικρά δείγματα, συνεπώς υπάρχει και μικρή δύναμη ανίχνευσης στατιστικά σημαντικών σχέσεων αλλά και λίγη αυτοπεποίθηση για εύρεση της πραγματικής σχέσης,
3. Είναι διαθέσιμα πολύ μεγάλα δείγματα οπότε και η δύναμη του τεστ είναι πολύ μεγάλη,
4. Λείπουν πολλές παρατηρήσεις,
5. Η σχέση δεν είναι γραμμική, και
6. Δεν γίνεται κατανοητή η ανάλυση γενικότερα.

ΚΕΦΑΛΑΙΟ ΤΕΤΑΡΤΟ
ΣΤΑΤΙΣΤΙΚΟΙ ΕΛΕΓΧΟΙ

4.1 Εισαγωγή

Όπως προαναφέρθηκε και στην εισαγωγή αυτής της μελέτης σε κάθε αναλυτική/συμπερασματολογική στατιστική, ή αλλιώς ανάλυση παλινδρόμησης, υπάρχουν δύο ομάδες προβλημάτων, η Στατιστική Εκτίμηση (statistical estimation), και ο Έλεγχος Στατιστικών υποθέσεων(test of hypotheses) (Δημητριάδης 2002).

4.2 Στατιστική Εκτίμηση (Statistical Estimation)

Η στατιστική εκτίμηση χαρακτηρίζει τη διαδικασία κατά την οποία οι τιμές του πληθυσμού εκτιμώνται με βάση τα δεδομένα ενός δείγματος και χωρίζετε σε δυο ειδών εκτίμησης και «εκτιμητές»: την Εκτίμηση Σημείου (Point Estimation) ή αλλιώς τον εκτιμητή σε σημείο, και την Εκτίμηση Διαστήματος (Interval Estimation) ή αλλιώς εκτιμητής σε διάστημα.

4.2.1 Εκτίμηση Σημείου

Με τη μέθοδο εκτίμησης σημείου, η τιμή μιας παραμέτρου *θ* υπολογίζετε απο μία αριθμητική συνάρτηση η οποία χρησιμοποιώντας τις μετρήσεις του δείγματος δίνει ένα αποτέλεσμα με τη μορφή απλού αριθμού, ο οποίος με τη σειρά του ονομάζετε «εκτιμητής σε σημείο της παραμέτρου» (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα 1998). Με άλλα λόγια με τη μέθοδο εκτίμησης σημείου η τιμή του πληθυσμού (παράμετρος *θ*) καθορίζετε από μόνο μία τιμή του δείγματος (Δημητριάδης 2002). Έστω για παράδειγμα πως η μέση τιμή ενός δείγματος 30 προϊόντων την εταιρίας Johnson & Johnson είναι 7 ευρώ, τότε ο εκτιμητής σε σημείο της παραμέτρου θ (δηλαδή της τιμής του προϊόντος) είναι 7 και έτσι εκτιμάτε πως η μέση τιμή όλων των προϊόντων την εταιρίας Johnson & Johnson είναι περίπου 7 ευρώ.

Σύμφωνα με τους Λαζαρίδης και Lazaridou (2008, σελ. 171), η αριθμητική παράσταση που ορίζει τη τιμή της παραμέτρου θ είναι: $\theta_n = \varphi (x_1, x_2, \ldots, x_n)$ και ο εκτιμητής σε σημείο συμβολίζεται με την παράσταση $T=\varphi (X_1, X_2, \ldots X_n, \theta)$ όπου το σύμβολο φ υποδεικνύει πως η συγκεκριμένη παράσταση αφορά τον εκτιμητή θ_n της παραμέτρου θ.

Επιπλέον, σύμφωνα με τους ίδιους ερευνητές, ένα εκτιμητής $\theta_n = \varphi (X_1, X_2, \ldots, X_n, \theta)$ μπορεί να είναι είτε *αμερόληπτος* είτε *ασυμπτωματικά αμερόληπτος*, αναλόγως *με* την τιμή της μαθηματικής του ελπίδας (Λαζαρίδης και Lazaridou (2008, σελ. 172). Όταν για παράδειγμα η τιμή της μαθηματικής ελπίδας ενός εκτιμητή είναι ίση με την τιμή της μεταβλητής θ που μετράει τότε αυτός ο εκτιμητής λέγετε αμερόληπτος, ειδάλλως ονομάζετε ασυμπτωματικά αμερόληπτος. Είναι σημαντικό να ειπωθεί σε αυτό το σημείο πως οι τεχνικές εκτίμησης σημείου χωρίζονται σε δύο κύριες κατηγορίες, στις αναλυτικές τεχνικές και στις πιθανοθεωρητικές τεχνικές. Η ομάδα των αναλυτικών μεθόδων συμπεριλαμβάνει τη μέθοδο των ροπών, τη μέθοδο της μέγιστης πιθανοφάνειας και τη μέθοδο των ελαχίστων τετραγώνων. Η πιθανοθεωρητική ομάδα τεχνικών εκτίμησης σημείου συμπεριλαμβάνει μεθόδους στις οποίες οι τιμές των παραμέτρων εκτιμώνται από «στατιστικές συναρτήσεις που ικανοποιούν κάποιες βέλτιστες ιδιότητες όπως της αμεροληψίας, της ελάχιστης διασποράς, και άλλες» (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα, 1998). Πιο συγκεκριμένα, σύμφωνα με τους Λαζαρίδης και Lazaridou (2008), στις πιθανοθεωρητικές σημειακές εκτιμήσεις «ο εκτιμητής χαρακτηρίζεται από το νόμο της πιθανότητας που περιγράφει το αποτέλεσμα της παραμέτρου στις παρατηρήσεις» και, ενώ όταν σχετίζεται με μια συνεχή τυχαία μεταβλητή X η συνάρτηση είναι: $f(x_1, x_2, \ldots, x_n, \theta)$ ενώ όταν αφορά μια διακριτή τυχαία μεταβλητή X τότε η συνάρτηση γίνεται: $P(X_1=x_1, X_2=x_2, \ldots X_n=x_n \ \theta)$.

Όπως προκύπτει, οι πιθανοθεωρητικές μέθοδοι εκτίμησης παραμέτρων σε σημείο απαιτούν άριστη γνώση της θεωρίας των πιθανοτήτων (Κλολυβά-Μαχαιρά και Μπόρα-Σέντα, 1998) και συνεπώς είναι λογικό να θεωρούνται δυσκολότερες από ότι οι αναλυτικές μέθοδοι.

4.2.2 Εκτίμηση διαστήματος (Διάστημα εμπιστοσύνης)

Κατά τη μέθοδο εκτίμησης διαστήματος η τιμή του πληθυσμού (ή μιας παραμέτρου) βασίζετε σε ένα τυχαίο διάστημα τιμών το οποίο ονομάζετε διάστημα εμπιστοσύνης (confidence interval) (Δημητριάδης, 2002) και καθορίζετε από έναν αριθμητικό τύπο (τον εκτιμητή σε διάστημα) ο οποίος χρησιμοποιώντας τις μετρήσεις των τιμών του δείγματος υπολογίζει τα άκρα του διαστήματος που εμπεριέχει τις τιμές της μεταβλητής (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα 1998). Σε προηγούμενο παράδειγμα αν η μέση τιμή ενός δείγματος 30 προϊόντων την εταιρίας Johnson & Johnson είναι 7 ευρώ, με βάση το δείγμα των τριάντα προϊόντων μπορεί να δημιουργηθεί ένα

διάστημα εμπιστοσύνης, για παράδειγμα 5-9 μέσα στο οποίο κυμαίνονται οι τιμές όλων των προϊόντων την εταιρίας Johnson & Johnson. Κατά συνέπεια, ο σημαντικότερος παράγοντας στην εκτίμηση/υπολογισμό των διαστημάτων εμπιστοσύνης είναι το μέγεθος του δείγματος καθώς «όσο μεγαλύτερο είναι το δείγμα τόσο μικρότερο είναι το εύρος του διαστήματος εμπιστοσύνης που εκτιμάται» (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα 1998). Έτσι, εάν στο παραπάνω παράδειγμα το δείγμα αποτελούνταν από εξήντα προϊόντα έναντι των τριάντα, τότε θα μπορούσε να εκτιμηθεί ένα διάστημα εμπιστοσύνης 6-8 αντί του 5-9.

Παρόλα αυτά, έστω και αν η αληθινή τιμή της μεταβλητής θ που μετράται βρίσκεται μέσα στο εκτιμώμενο διάστημα εμπιστοσύνης, η εκτίμηση του έχει πάντα το χαρακτηριστικό randomness (τυχαίας επιλογής) (Λαζαρίδης & Lazaridou, 2008). Αυτό συμβαίνει καθώς ενώ η εύρεση και ο καθορισμός ενός τέτοιου διαστήματος αποσκοπεί στο να εμπεριέχει την αληθινή τιμή της παραμέτρου, τα άκρα του αποτελούν τυχαίες μεταβλητές (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα, 1998) και έτσι υπάρχει πάντα η πιθανότητα αυτό να μην συμβεί. Συνεπώς, όλα τα διαστήματα εμπιστοσύνης εμπεριέχουν ένα συντελεστή εμπιστοσύνης (γνωστός και ως «επίπεδο στατιστικής σημαντικότητας») ο οποίος καθορίζει το ποσοστό επιτυχίας της εκτίμησης και παίρνει τιμές από 0 έως 1 (Δημητριάδης, 2002). Έτσι, σε περίπτωση που ο συντελεστής εμπιστοσύνης στο προηγούμενο παράδειγμα είναι 0.85, τότε το ενδεικνυόμενο ποσοστό επιτυχίας της εκτίμησης είναι 85%, και άρα υπάρχουν 85% πιθανότητες το διάστημα εμπιστοσύνης 5-9 για τις τιμές των προϊόντων της εταιρίας Johnson &Johnson να είναι σωστό. Έχει ενδιαφέρον να σημειωθεί πως στις περιπτώσεις όπου ο συντελεστής εμπιστοσύνης είναι 1 (δηλαδή 1-a) τότε το διάστημα εμπιστοσύνης ονομάζεται «100(1- a)% διάστημα εμπιστοσύνης» και ορίζεται ως: $P(A<\theta<\Delta) = 1- a$, όπου Α και Δ είναι τα άκρα του διαστήματος (Α= αριστερό άκρο, Β=δεξί άκρο) και θ η άγνωστη παράμετρος του πληθυσμού (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα, 1998). Προδήλως, πιστεύεται πως όσο μικρότερο είναι το διάστημα εμπιστοσύνης τόσο μεγαλύτερος είναι ο συντελεστής εμπιστοσύνης και η ορθότητα της εκτίμησης (Ζαχαροπούλου, 1998).

Αυτή τη λογική ακολουθούν οι μέθοδοι εκτίμησης όλων των ειδών διαστημάτων εμπιστοσύνης όπως το διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού (ανεξάρτητα από το μέγεθος του δείγματος και τη διασπορά), το διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών δύο πληθυσμών

(ανεξάρτητα από το εύρος εξάρτησης των δειγμάτων και το είδος των παρατηρήσεων), διάστημα εμπιστοσύνης για την αναλογία p στοιχείων ενός πληθυσμού, διάστημα εμπιστοσύνης για τη διαφορά p_1-p_2 των αναλογιών δύο πληθυσμών, διάστημα εμπιστοσύνης για τη διασπορά ενός πληθυσμού, κ.α.

4.3 Έλεγχος Στατιστικών Υποθέσεων

Δεδομένου ότι όλες οι έρευνες, συνεπώς και οι στατιστικές, βασίζονται σε υποθέσεις οι οποίες με τη σειρά τους επηρεάζουν τη διαδικασία λήψης αποφάσεων των ανθρώπων σε σχέση με το συγκεκριμένο ερευνητικό θέμα, ο έλεγχος ορθότητας της κάθε υπόθεσης (δηλαδή εναλλακτικής απόφασης) είναι απαραίτητος ώστε να ελαχιστοποιηθεί ο κίνδυνος λανθασμένης απόφασης (Λαζαρίδης & Lazaridou, 2008). Έστω, για παράδειγμα, πως μια εταιρία αντιμετωπίζει σημαντική μείωση στο ποσοστό επενδυτικών προτάσεων που δέχεται και αποφασίζει να πάρει μέτρα βελτίωσης της επιχειρησιακής της εικόνας και ελκυστικότητας. Έτσι ο υπεύθυνος επιχειρησιακής ανάπτυξης πρέπει να πάρει μια απόφαση ανάμεσα σε τουλάχιστον δύο εναλλακτικά αναπτυξιακά σχέδια που το κάθε ένα περιέχει μια διαφορετική υπόθεση γύρω από το χρονικό διάστημα αποδοτικότητας και αποτελεσματικότητας του (έστω πως η αποτελεσματικότητα του σχεδίου Α εκτιμάται πως θα είναι εμφανή εντός τεσσάρων εβδομάδων από την ημέρα εφαρμογής του, και το σχέδιο Β «υπόσχεται» εμφανή αποτελέσματα εντός δύο εβδομάδων).

Η διαδικασία ελέγχου των υποθέσεων ονομάζεται «έλεγχος υποθέσεων» ή «στατιστικό τεστ» και ορίζεται από τα δεδομένα της έρευνας (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα, 1998). Για να διεξαχθεί κατάλληλα ένα στατιστικό τεστ η μία από τις υποθέσεις ονομάζεται «μηδενική» και συμβολίζεται H_0 ενώ η άλλη ονομάζεται «εναλλακτική» και συμβολίζεται H_1 (Λαζαρίδης & Lazaridou, 2008). Πιο συγκεκριμένα, η H_0 ενός τεστ συνήθως είναι «$\theta = \theta_0$ όπου θ είναι η ελεγχόμενη παράμετρος του πληθυσμού και θ_0 μια συγκεκριμένη τιμή της» (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα, 1998). Στη περίπτωση του παραπάνω παραδείγματος $\theta = \alpha$ και $\theta_0=4$ (δεδομένου ότι H_0: $\alpha = 4$) καθόσον η υπόθεση αποδοτικότητας του σχεδίου Α (όπου $\alpha=4$) είναι η μηδενική υπόθεση H_o και η υπόθεση του σχεδίου Β (όπου $\alpha<4$) είναι η εναλλακτική H_1. Αν και η τιμή της μηδενικής υπόθεσης είναι συγκεκριμένη ($\theta = \theta_0$) «η εναλλακτική υπόθεση μπορεί να είναι $\theta > \theta_0$, $\theta < \theta_0$, ή και $\theta \neq \theta_0$: στις δύο πρώτες περιπτώσεις το τεστ λέγεται μονόπλευρο ενώ στην τρίτη δίπλευρο» (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα, 1998).

Είναι σημαντικό να σημειωθεί πως ένα από τα κυριότερα στοιχεία ενός στατιστικού τεστ (ανεξάρτητα από το αν είναι μονόπλευρο ή δίπλευρο), είναι ο ορισμός της απορριπτικής περιοχής της μηδενικής υπόθεσης H_0 του τεστ. Η απορριπτική περιοχή της H_0 συμβολίζεται με R και χαρακτηρίζει την περιοχή στα σημεία της οποίας η H_0 απορρίπτεται. Οι κυριότεροι λόγοι που ο ορισμός της απορριπτικής περιοχής της H_0 είναι από τα σημαντικότερα στοιχεία ενός στατιστικού τεστ είναι όχι μόνο το ότι ο ορισμός της απορριπτικής περιοχής βοηθά τον ερευνητή να κάνει την ορθότερη επιλογή υποθέσεως αλλά και το ότι έχει άμεση σχέση με τους κινδύνους που συντρέχουν οι στατιστικοί έλεγχοι και τον τύπο σφαλμάτων αυτών, οι οποίοι θα αναλυθούν παρακάτω. Επιπροσθέτως, σύμφωνα με την Ζαχαροπούλου (1998), ένας επιπλέον λόγος για την διεξαγωγή ελέγχου υποθέσεων σε ένα μοντέλο στατιστικής ανάλυσης είναι ο εντοπισμός πλεονάζουσων (redundant) ερμηνευτικών μεταβλητών, δηλαδή μεταβλητών των οποίων η αφαίρεση δε μειώνει σημαντικά την συνολική ερμηνευτική ικανότητα του μοντέλου.

Συνοψίζοντας, σύμφωνα με τις Κλολυβά-Μαχαιρά & Μπόρα-Σέντα (1998) τα στοιχεία ενός στατιστικού τεστ, ή ελέγχου υπόθεσης, είναι τα πέντε ακόλουθα: (1) ο ορισμός της μηδενικής υπόθεσης H_0, (2) ο ορισμός της εναλλακτικής υπόθεσης H_1, (3) ο ορισμός του στατιστικού τεστ από το δείγμα, (4) ο ορισμός της απορριπτικής περιοχής R της υπόθεσης H_0, και (5) η εξαγωγή συμπερασμάτων.

4.4 Κίνδυνοι Ελέγχου / Τύποι Σφαλμάτων και Διαδικασία Ελέγχου

Όπως είναι λογικό, κάθε έλεγχος υποθέσεων διατρέχεται και από τον κίνδυνο να μην παρθεί η σωστή απόφαση. Αυτός ο κίνδυνος διαχωρίζετε σε δύο συγκεκριμένες υποκατηγορίες κινδύνου ή σφάλματος, τον κίνδυνο I είδους, και τον κίνδυνο II είδους. Ο κίνδυνος I είδους υφίσταται εφόσον ληφθεί η εναλλακτική υπόθεση H_1 ως σωστή ενώ στην πραγματικότητα αληθές είναι η μηδενική υπόθεση H_0. Ο κίνδυνος I είδους είναι επίσης γνωστός και με την ονομασία *επίπεδο σημαντικότητας* ή, αν η έρευνα είναι ποιοτική, με τον ορισμό *κίνδυνος του προμηθευτή* (Λαζαρίδης & Lazaridou, 2008). Επιπλέον, η πιθανότητα να απορριφθεί η H_0 ενώ στην πραγματικότητα ισχύει (σφάλμα τύπου I) συμβολίζεται με το γράμμα ***a*** και είναι a=P (Κλολυβά-Μαχαιρά & Μπόρα-Σέντα, 1998) και καθορίζεται πριν από κάθε έλεγχο με τιμές όπως 0.10, 0.05, 0.01, 0.002 και ούτω καθεξής. Το επίπεδο εμπιστοσύνης του ελέγχου καθορίζεται από την πιθανότητα P= 1-α, όπου P είναι το κάθε σημείο

πραγματοποίησης (x_1, x_2,…x_n) ενός δείγματος μεγέθους n (X_1, X_2, …, X_n) (Λαζαρίδης & Lazaridou, 2008).

Ο κίνδυνος ΙΙ είδους υφίσταται όταν ο ερευνητής επιλέγει την μηδενική υπόθεση H_0 ενώ αληθές είναι η εναλλακτική υπόθεση και στην ποιοτική έρευνα είναι επίσης γνωστός με την ονομασία *κίνδυνος του αγοραστή* και συμβολίζεται με το γράμμα *β*. Έχει ενδιαφέρον να σημειωθεί πως ο κίνδυνος *β* είναι πολύ δύσκολο, έως και αδύνατο, να ορισθεί δεδομένου ότι δεν είναι ποτέ γνωστός εκ των προτέρων. Έτσι, η πιθανότητα 1-β χαρακτηρίζει τη ισχύ του στατιστικού ελέγχου (Λαζαρίδης & Lazaridou, 2008), δηλαδή την «πιθανότητα απόρριψης της H_0 . Η σημαντικότητα του κάθε είδους κινδύνου είναι σχετική και εξαρτάται από το περιεχόμενο της κάθε έρευνας, τις συνέπειες του κάθε σφάλματος, καθώς και τα διαφορετικά ενδιαφέροντα των ατόμων/εταιριών κλπ. Συνεπώς υπάρχουν πολλοί έλεγχοι υποθέσεων και η εκλογή του σημαντικότερου καθορίζεται από τον τύπο και το είδος του δείγματος, αν είναι δηλαδή μικρό ή μεγάλο, με γνωστή ή άγνωστη διασπορά πληθυσμού και τα λοιπά. Ονομαστικά, σύμφωνα με τις Κλολυβά-Μαχαιρά & Μπόρα-Σέντα (1998) οι διάφοροι έλεγχοι υπόθεσης είναι:

1. Έλεγχος υπόθεσης για τη μέση τιμή μ του πληθυσμού
2. Έλεγχος υπόθεσης για τη διαφορά μ1-μ2 των μέσων τιμών δύο πληθυσμών
3. Έλεγχος υπόθεσης για την αναλογία στοιχείων ενός πληθυσμού
4. Έλεγχος υπόθεσης για τη διαφορά p1-p2 των αναλογιών δύο πληθυσμών
5. Έλεγχος υπόθεσης για τη διασπορά ενός πληθυσμού
6. Έλεγχος υπόθεσης για το λόγο σ^2_1 / σ^2_2 των διαπορών δύο πληθυσμών

4.5 Έλεγχος Στατιστικής Σημαντικότητας

Δεδομένου πως και η στατιστική εκτίμηση και ο έλεγχος των υποθέσεων διατρέχουν ένα ποσοστό κινδύνου για λάθος, τα αποτελέσματα μιας έρευνας κρίνονται κατάλληλα μόνο αν βρεθεί στατιστική σημαντικότητα. Συνεπώς ο έλεγχος στατιστικής σημαντικότητας είναι απαραίτητος και αφορά την ταυτολογία και την εύρεση (πιθανής) παραβίασης των στατιστικών υποθέσεων που απαιτούν συμπληρωματική ανάλυση (Higgins, 2009).

Στον πυρήνα της, ο έλεγχος στατιστικής σημαντικότητας είναι μια διαδικασία για τον προσδιορισμό της πιθανότητας ενός αποτελέσματος αν υποτεθεί πως η μηδενική υπόθεση είναι πραγματική. Πιο συγκεκριμένα, οι τεχνικές που χρησιμοποιούνται συνηθέστερα για τον έλεγχο της στατιστικής σημαντικότητας (όπως τα t-ratios και ανάλυση παλινδρόμησης που θα αναλυθεί σε επόμενο κεφάλαιο) είναι διαδικασίες για τον προσδιορισμό της πιθανότητας ενός αποτελέσματος (η οποία είναι συνήθως κάποιο προκαθορισμένο επίπεδο που αναφέρεται ως άλφα) υποθέτοντας ότι η μηδενική υπόθεση είναι αληθές με δεδομένο ένα τυχαίο δείγμα και ένα μέγεθος δείγματος n (Carver, 1978).

Όπως προκύπτει λογικά, ο έλεγχος στατιστικής σημαντικότητας εξαρτάται από το μέγεθος του δείγματος και τις υποθέσεις δοκιμής (Thompson, 1987· Fan & Jacoby, 1995). Είναι σημαντικό να ειπωθεί πως η εξάρτηση αυτή από το μέγεθος του δείγματος δείχνει πως έλεγχος στατιστικής σημαντικότητας παρέχει μια εκτίμηση μόνο για το μέγεθος του δείγματος. Με άλλα λόγια, ο έλεγχος στατιστικής σημαντικότητας μπορεί να είναι μια ταυτολογική προσπάθεια που δεν είναι σε θέση να παράσχει καθαρά αναπαράγονται αποτελέσματα (Thompson, 1987).

Περαιτέρω, οι τεχνικές ελέγχου στατιστικής σημαντικότητας, όπως η παλινδρόμηση για παράδειγμα, λειτουργούν με βάση σημαντικές υποθέσεις που όμως παραβιάζονται σε συνεχή βάση, γεγονός που εμποδίζει την ικανότητα των ερευνητών να διεξάγουν εκτενή και αλάθητα συμπεράσματα (Thompson, 1992· Fan & Jacoby, 1995). Ως εκ τούτου, πολλοί ερευνητές θεωρούν πως ο έλεγχος στατιστικής σημαντικότητας από μόνος του μπορεί να είναι μια φτωχή μέθοδος για την επανα-εξακρίβωση των στατιστικών αποτελεσμάτων μίας έρευνας. Παρόλα αυτά δε μπορεί κάποιος παρά να συμφωνήσει πως μία φτωχή μέθοδος ελέγχου στατιστικής σημαντικότητας μέσω της παλινδρόμησης είναι πάντα καλύτερη από το να μην εφαρμοστεί κανένας απολύτως έλεγχος. Έτσι, σύμφωνα με την Ζαχαροπούλου (1998) μπορεί να γίνει ένας ολικός έλεγχος της σημαντικότητας του μοντέλου συγκρίνοντας «το πλήρες μοντέλο των k ερμηνευτικών μεταβλητών με το μειωμένο μοντέλο στο οποίο υπάρχει μόνο ο σταθερός όρος». Τέλος, είναι σημαντικό να ειπωθεί πως σπανίως είναι ανάγκη ένας ερευνητής να χρησιμοποιήσει υπολογισμούς αριθμητικών παραστάσεων μιας και όλα τα προγράμματα στατιστικής ηλεκτρονικών υπολογιστών, δίνουν αυτόματα την F_{π} ώς F-value μαζί με την p-value (δηλαδή τη τιμή πιθανότητας μια τιμή της F να είναι ίση

ή μεγαλύτερη της F_{π} όταν ισχύει η H_0). Αν η p-value είναι μικρότερη απο το α (το οποίο συνήθως ισούται με 0.05), τότε η H_0 απορρίπτεται (Ζαχαροπούλου, 1998).

4.5.1 Έλεγχος στατιστικής σημαντικότητας των συντελεστών μερικής συσχέτισης

Πολλοί ερευνητές θεωρούν την έννοια της μερικής συσχέτισης μια χρήσιμη προσέγγιση για τη μελέτη της σχέσης μεταξύ δύο μεταβλητών x και y με την παρουσία μιας τρίτης μεταβλητής z (η σχέση αυτή ονομάζεται μερική συσχέτιση). Έτσι, μερική συσχέτιση είναι στην ουσία η συσχέτιση δύο μεταβλητών, ενώ μια τρίτη, ή περισσότερες μεταβλητές είναι υπό έλεγχο. Με άλλα λόγια μερική συσχέτιση ονομάζεται η διαδικασία που επιτρέπει τον ερευνητή να μετρήσει την περιοχή οπού τρεις μεταβλητές συμπίπτουν και να καθορίσει ποια είναι η σχέση μεταξύ των δύο από αυτές τις μεταβλητές όταν η τρίτη (μεταβλητή) παραμένει σταθερή (Lowry, 2010).

Πιο συγκεκριμένα, σύμφωνα με την Ζαχαροπούλου (1998), ο συντελεστής μερικής συσχέτισης «μετρά την ένταση της γραμμικής συμμεταβολής» ανάμεσα στην εξαρτημένη μεταβλητή Y και μια συγκεκριμένη ερμηνευτική μεταβλητή X_{κ} μετά την αφαίρεση των γραμμικών επιδράσεων των υπόλοιπων ερμηνευτικών μεταβλητών. Τέλος, η στατιστική σημαντικότητα των συντελεστών μερικής συσχέτισης μπορεί να ελεγχθεί με τη χρήση ενός ελέγχου στατιστικού αποτελέσματος παρόμοιο με εκείνο που χρησιμοποιείται για τον συντελεστή απλής συσχέτισης:

$$t = \frac{(Partialr)\sqrt{n-q-2}}{\sqrt{1-(partialr^2)}}$$

όπου q είναι ο αριθμός των μεταβλητών που παραμένουν σταθερές. Η τιμή του t συγκρίνεται με πινακοποιημένες τιμές t για n-q-2 βαθμούς ελευθερίας.

4.6 Σύνοψη

Οι έλεγχοι υποθέσεων μαζί με τα διαστήματα εμπιστοσύνης είναι τα δύο σημαντικότερα εργαλεία της στατιστικής συμπερασματολογίας. Αντικείμενο του κλάδου αυτού της στατιστικής είναι η εξαγωγή συμπερασμάτων σχετικά με τα χαρακτηριστικά ενός πληθυσμού από πληροφορίες που αναφέρονται σε ένα μόνο δείγμα. Σύμφωνα με την πρακτική των διαστημάτων εμπιστοσύνης, υπολογίζεται μια

σημειακή εκτίμηση μιας παραμέτρου του πληθυσμού (π.χ. μέση τιμή) και στη συνέχεια σχηματίζεται ένα διάστημα εμπιστοσύνης γύρω από την εκτίμηση αυτή, για το οποίο υπάρχει βεβαιότητα (εμπιστοσύνη) κατά ένα ποσοστό ότι βρίσκεται η ζητούμενη παράμετρος του πληθυσμού. Αυτή η ανάλυση δε λαμβάνει υπόψη οποιαδήποτε πεποίθηση υπάρχει σχετικά με τον πληθυσμό.

Στην άλλη πλευρά βρίσκονται οι **έλεγχοι υποθέσεων** (ή **έλεγχοι σημαντικότητας** ή **κανόνες αποφάσεων**). Η πρακτική αυτή λαμβάνει υπόψη την πεποίθηση που υπάρχει σχετικά με την παράμετρο ενός πληθυσμού, η οποία οδηγεί στην κατάστρωση μιας υπόθεσης. Σκοπός της ανάλυσης είναι η αποδοχή ή η απόρριψη της υπόθεσης χρησιμοποιώντας τις πληροφορίες που παρέχει το δείγμα του πληθυσμού. Επισημαίνεται ότι θέτοντας μικρότερο επίπεδο σημαντικότητας, απαιτούνται πιο «σημαντικές αποδείξεις» για την απόρριψη της H_o και τον χαρακτηρισμό των ευρημάτων στο δείγμα ως στατιστικά σημαντικών. Έτσι, μπορεί, σε κάποιο επίπεδο σημαντικότητας α , π.χ. $\alpha = 0.05$, να απορρίπτεται η H_o και σε κάποιο μικρότερο, π.χ. $\alpha = 0.01$, να μην απορρίπτεται γιατί απαιτούνται σημαντικότερες αποδείξεις. Όσο πιο μικρό είναι το επίπεδο σημαντικότητας στο οποίο μπορεί να απορριφθεί η H_o, τόσο πιο σημαντική είναι η τιμή της στατιστικής συνάρτησης ελέγχου που παρατηρείται στο δείγμα, με την έννοια ότι δίνει πιο ισχυρές αποδείξεις εναντίον της H_o. Άρα, όσο **πιο μικρό** είναι το επίπεδο σημαντικότητας στο οποίο μπορεί να απορριφθεί η H_o, τόσο **πιο σημαντικό**, στατιστικά, είναι το αποτέλεσμα του ελέγχου. Τέλος, είναι προφανές, ότι αν η H_o απορρίπτεται σε κάποιο επίπεδο σημαντικότητας α, τότε επίσης απορρίπτεται σε οποιοδήποτε μεγαλύτερο, ενώ αν δεν απορρίπτεται σε κάποιο επίπεδο σημαντικότητας α, τότε επίσης δεν απορρίπτεται σε οποιοδήποτε μικρότερο.

ΚΕΦΑΛΑΙΟ ΠΕΜΠΤΟ
ΠΑΛΙΝΔΡΟΜΗΣΗ

5.1 Εισαγωγή

Στα διάφορα προβλήματα που εμφανίζονται τυχαίες μεταβλητές συνήθως δεν αρκεί η μελέτη των βασικών χαρακτηριστικών (μέση τιμή, διασπορά κ.λπ.) της κάθε μιας μεμονωμένης μεταβλητής αλλά επιδιώκεται, αν είναι δυνατό, να προσδιοριστεί με ποιο τρόπο σχετίζονται αυτές μεταξύ τους. Για παράδειγμα:

1. Η ηλικία και το βάρος ενός παιδιού έχουν κάποια θετική συσχέτιση μεταξύ τους με την έννοια ότι όσο πιο μεγάλο είναι το παιδί τόσο μεγαλύτερο βάρος θα έχει.
2. Η διάρκεια ζωής των διαφόρων οργανισμών σε μία περιοχή και το ποσοστό μόλυνσης της περιοχής έχουν αρνητική συσχέτιση.
3. Η μέση θερμοκρασία μιας ημέρας σε ένα τόπο σχετίζεται με την ηλιοφάνεια, τη σχετική υγρασία, το ύψος βροχής κ.λπ.

Ο γενικός τομέας της στατιστικής που εξετάζει τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών με απώτερο σκοπό την πρόβλεψη μιας από αυτές μέσω των άλλων, χαρακτηρίζεται με την ονομασία *ανάλυση παλινδρόμησης* (regression analysis). Ιστορικά, ο όρος «regression» χρησιμοποιήθηκε για πρώτη φορά από το Βρετανό ανθρωπολόγο Sir Francis Galton στην εργασία του «Regression Towards Mediocrity in Hereditary Stature» (1885) όπου, μελετώντας τα ύψη των παιδιών σε σχέση με το μέσο ύψος των γονέων διαπιστώθηκε ότι αυτά είχαν την τάση να παλινδρομούν (regress) προς το μέσο γενικό ύψος αντί να στρέφονται προς ακραίες τιμές. Παρότι όμως αρχικά ο όρος παλινδρόμηση χρησιμοποιήθηκε για να περιγράψει τη συγκεκριμένη αυτή διαπίστωση, με την πάροδο του χρόνου επεκτάθηκε η χρήση του και σήμερα έχει γίνει συνώνυμος με τη στατιστική μελέτη της σχέσης μεταξύ μεταβλητών.

Σε κάθε πρόβλημα παλινδρόμησης διακρίνουμε συνήθως δύο είδη μεταβλητών: τις ανεξάρτητες (independent) και τις εξαρτημένες (dependent). Όπως θα συζητηθεί και παρακάτω, οι ανεξάρτητες μεταβλητές (οι οποίες συμβολίζονται συνήθως με X) είναι εκείνες στις οποίες μπορούμε να δίνουμε μία συγκεκριμένη τιμή (π.χ. θερμοκρασία επεξεργασίας ενός προϊόντος, ρυθμός τροφοδοσίας με καταλύτη) ή παίρνουν τιμές που μπορούμε να παρατηρήσουμε αλλά όχι να ελέγξουμε (π.χ. θερμοκρασία περιβάλλοντος, ηλιοφάνεια). Η εξαρτημένη μεταβλητή (Y) αντανακλά το

αποτέλεσμα μεταβολών στις ανεξάρτητες μεταβλητές (π.χ. χρώμα-καθαρότητα ενός προϊόντος). Ο προσδιορισμός της εξαρτημένης μεταβλητής (Y) μονοδρομεί και τον ορισμό των ανεξαρτήτων μεταβλητών (X), υποδεικνύοντας το γεγονός πως είναι πιθανό η διάκριση μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών να μην είναι σαφής εξαρχής, ενώ αυτό επιτυγχάνεται μέσα από την μαθηματική εξίσωση που συσχετίζει την εξαρτημένη με τις ανεξάρτητες μεταβλητές όταν αυτές οριστούν.

5.2 Μοντέλα Παλινδρόμησης

Για την εύρεση του κατάλληλου μοντέλου για την περιγραφή της σχέσης μεταξύ δύο ή περισσοτέρων μεταβλητών (ανεξάρτητων και εξαρτημένων) έχουν αναπτυχθεί κατάλληλα μαθηματικά πρότυπα και στατιστικά υποδείγματα που βοηθούν στην επεξεργασία και ανάλυση τους. Για το λόγο αυτό, παρακάτω θα γίνει μια πλήρη ανάλυση του κάθε μοντέλου ξεχωριστά.

5.2.1 Απλή Γραμμική Παλινδρόμηση

Η απλούστερη περίπτωση παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση όπου υπάρχει μία μόνο ανεξάρτητη μεταβλητή X και μια εξαρτημένη μεταβλητή Y που μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του X. Η σχέση μεταξύ των δύο μεταβλητών χαρακτηρίζεται ως αιτιώδης διότι οι τιμές των ερμηνευτικών μεταβλητών, ερμηνεύουν την τιμή της εξαρτημένης. Η απλή σχέση παλινδρόμησης μπορεί να παρουσιαστεί με την εξής απλή μαθηματική μορφή:

$$Y= f(X)$$

Η οικονομική θεωρία ενδεχομένως να παρέχει πληροφόρηση για το πρόσημο της σχέσης μεταξύ Y και X και για τη συναρτησιακή μορφή της σχέσης. Η βασική υπόθεση είναι ότι η συναρτησιακή σχέση μεταξύ Y και X είναι γραμμική καταλήγοντας σε μία απλή γραμμική σχέση παλινδρόμησης (simple linear regression relationship) της μορφής:

$$Y= \beta_o + \beta_1 X$$

Όπου β_0 και β_1 είναι οι παράμετροι της σχέσης. Ειδικότερα, ο συντελεστής β_0 είναι ο σταθερός όρος (intercept ή constant term), δηλαδή το σημείο από το οποίο ξεκινά η ευθεία που διέρχεται ανάμεσα από τα σημεία των συντεταγμένων των δύο μεταβλητών, όταν η τιμή της μεταβλητής X είναι 0. Είναι δηλαδή το αυτόνομο τμήμα της τιμής της Y. Ο όρος αυτός είναι η μαθηματική ελπίδα της εξαρτημένης μεταβλητής όταν η τιμή της ανεξάρτητης μεταβλητής είναι μηδέν και άρα είναι ο σταθερός όρος της πληθυσμιακής εξίσωσης παλινδρόμησης. Ο συντελεστής β_1

αντιπροσωπεύει την κλίση (slope coefficient), δηλαδή δείχνει τη μεταβολή της Y για μια μοναδιαία μεταβολή της X. Μια αρνητική κλίση δείχνει την ύπαρξη αρνητικής σχέσης μεταξύ των μεταβλητών. Η κλίση β_1 μετρά τη μεταβολή στη μαθηματική ελπίδα της εξαρτημένης μεταβλητής όταν μεταβάλλεται η ανεξάρτητη μεταβλητή.

Γενικά, η κλίση β_1 ενδιαφέρει σε μεγαλύτερο βαθμό τους οικονομολόγους και ειδικότερα όσους ασχολούνται με τις στατιστικές μεθόδους οικονομικών προτύπων. Για παράδειγμα, αν Y είναι το επίπεδο κατανάλωσης ενός συγκεκριμένου προϊόντος, και x είναι η τιμή ανά μονάδα του συγκεκριμένου προϊόντος, τότε για ένα κανονικό προϊόν η οικονομική θεωρία προτείνει ότι το πρόσημο του β_1 είναι αρνητικό. Η απλή γραμμική παλινδρόμηση είναι ένας χρήσιμος προσδιορισμός, καθώς η εκτίμηση του εν λόγω μοντέλου μπορεί να πραγματοποιηθεί, χρησιμοποιώντας τη σχετικά απλή διαδικασία της εκτίμησης των ελαχίστων τετραγώνων (Least Squares Estimation). Να αναφερθεί, ότι η γραμμικότητα της συναρτησιακής σχέσης δεν είναι αναγκαία, αν και μία μη γραμμική σχέση απαιτεί πολυπλοκότερες διαδικασίες εκτίμησης. Η υπόθεση της γραμμικότητας δεν είναι τόσο περιοριστική όσο ενδεχομένως να φαίνεται με μία πρώτη ματιά. Η απλή γραμμική παλινδρόμηση είναι γραμμική ως προς τις παραμέτρους και όχι κατ' ανάγκη και ως προς τις μεταβλητές, αφήνοντας έτσι αρκετό περιθώριο ευελιξίας στον προσδιορισμό του μοντέλου.

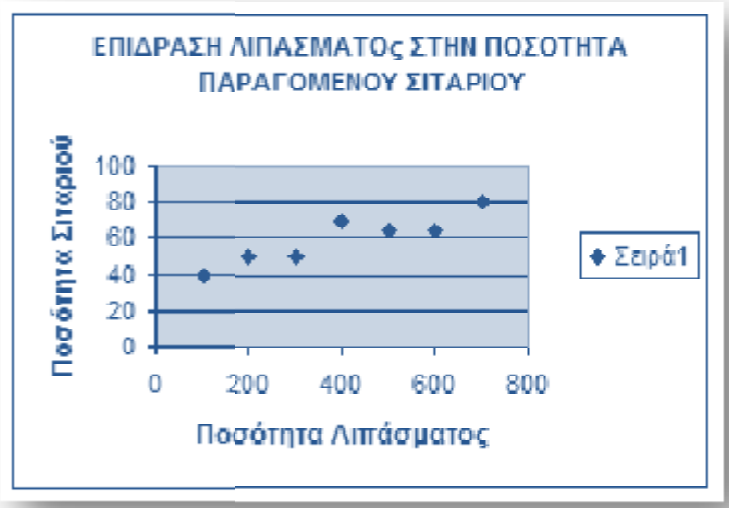
Παράδειγμα:

Για τη μελέτη της επίδρασης ενός λιπάσματος στην ποσότητα Y σιταριού που παράγεται, συγκεντρώθηκαν δεδομένα που αφορούν παρατηρήσεις σε 7 διαφορετικούς αγρούς (με παρόμοια απόδοση) (Πίνακας 10).

i	**Ποσότητα λιπάσματος (x_i)**	**Ποσότητα σιταριού (y_i)**
1	100	40
2	200	50
3	300	50
4	400	70
5	500	65
6	600	65
7	700	80

Πίνακας 10. Δεδομένα παρατηρήσεων για (7) διαφορετικούς αγρούς

Σχηματίζοντας το διάγραμμα διασποράς (scatter diagram, scatter plot) των δεδομένων (Εικόνα 10), παρατηρεί κανείς ότι τα ζεύγη (x_i,y_i) όπου i=1,2,….,7, είναι συγκεντρωμένα περίπου γύρω από ευθεία, δηλαδή η σχέση μεταξύ του Y και του X είναι κατά προσέγγιση γραμμική.



Εικόνα 10. Διάγραμμα Διασποράς των υπό επεξεργασία δοθέντων δεδομένων

Μια τέτοια ευθεία είναι και η ευθεία της ανωτέρω εικόνας. Θεωρώντας, λοιπόν, προσεγγίσεις της μορφής:

$y_i = \beta_o + \beta_1 x_i$, όπου i=1,2,…..,7

ή ισοδύναμα

$y_i = \beta_o + \beta_1 x_i + \varepsilon_i$, όπου i=1,2,….,7

(τα ε_i παριστάνουν τις αποκλίσεις της πραγματικής τιμής y_i από την προσαρμοσμένη ευθεία $\beta_0+\beta_1 x_i$). Ιδιαίτερο ενδιαφέρον είναι η εκλογή (εκτίμηση) των παραμέτρων β_0 και β_1 για την οποία να ελαχιστοποιούνται οι ποσότητες ε_i. Ο Gauss πρότεινε σαν κριτήριο επιλογής των β_0 και β_1 την ελαχιστοποίηση του αθροίσματος τετραγώνων, η οποία μέθοδος θα αναλυθεί στη συνέχεια επαρκώς.

5.2.2 Μη Γραμμική Παλινδρόμηση

Είναι κοινά αποδεκτό (Μαυρωτάς, Τσιαφογιάννη, Διακουλάκη & Σαρίμβεης, 2011) πως τα βασικά βήματα για την ανάλυση παλινδρόμησης είναι τα ακόλουθα:

1. Ορισμός του προβλήματος
2. Επιλογή των δυναμικών σχετικών μεταβλητών
3. Συλλογή δεδομένων
4. Προδιαγραφή του μοντέλου
5. Επιλογή της μεθόδου προσαρμογής
6. Προσαρμογή μοντέλου
7. Επικύρωση και κριτική μοντέλου

Παρόλο αυτά είναι πολύ πιθανό σε κάποια προβλήματα να υπάρχουν θεωρητικές ενδείξεις, ότι η εξάρτηση μιας εξαρτημένης τυχαίας μεταβλητής Y από μια ανεξάρτητη μεταβλητή X να είναι κάποιας συγκεκριμένης μη-γραμμικής μορφής. Σε κάποιες μάλιστα περιπτώσεις, μπορεί η μη-γραμμική μορφή να υποδειχθεί απο το διαγνωστικό γράφημα για ένα μοντέλο απλής γραμμικής παλινδρόμησης. Στις περιπτώσεις λοιπόν αυτές είναι ανάγκη να εκτιμηθούν οι παράμετροι μιας μη γραμμικής συνάρτησης ο βασικός τύπος της οποίας είναι ο εξής:

$$Y = \frac{\beta_o}{1 + e^{\beta_1 X}} + \varepsilon_i$$

Εξίσωση 4. Γενική μορφή μοντέλου με ανεξάρτητη μεταβλητή και κανονικά κατανεμημένα σφάλματα

Συνηθέστερα μοντέλα μη γραμμικής παλινδρόμησης με εξίσου μεγάλο δείκτη σημαντικότητας είναι η Εγγενής Γραμμική Συνάρτηση Παλινδρόμησης και η Πολυωνυμική Παλινδρόμηση.

5.2.2.1 Εγγενής Γραμμική Συνάρτηση Παλινδρόμησης

Με τον κατάλληλο μετασχηματισμό είναι δυνατόν κάποιες μη-γραμμικές συναρτήσεις να γίνουν γραμμικές, και μια τέτοια συνάρτηση ονομάζεται τότε *εγγενής γραμμική* (intrinsically linear). Στον Πίνακα 11 παρακάτω δίνονται οι τέσσερις πιο γνωστές εγγενείς γραμμικές συναρτήσεις και για κάθε μια δίνονται ο καταλληλότερος μετασχηματισμός και η καταλληλότερη γραμμική μορφή που δίνει ο μετασχηματισμός αυτός. Για την κατανόηση των συμβόλων όπου α νοείται η σταθερά β_o και όπου β η σταθερά β_1 όπως αυτά αναφέρθηκαν στους τύπους της γραμμικής παλινδρόμησης. Επίσης αξίζει να σημειωθεί ότι όπου δίνεται ο δεκαδικός λογάριθμος μπορεί ισοδύναμα να χρησιμοποιηθεί ο νεπέρειος λογάριθμος.

Εγγενής συνάρτηση	Μετασχηματισμός	Γραμμική συνάρτηση
1. Εκθετική: $y = \alpha e^{\beta x}$	$y' = 1n(y)$	$y' = 1n(\alpha) + \beta x$
2. Δύναμης: $y = \alpha x^{\beta}$	$y' = \log(y)$, $x' = \log(x)$	$y' = \log(\alpha) + \beta x'$
3. $y = \alpha + \beta\log(x)$	$x' = \log(x)$	$y = \alpha + \beta x'$
4. Αντίστροφη: $y = \alpha + \beta\frac{1}{x}$	$x' = \frac{1}{x}$	$y = \alpha + \beta x'$

Πίνακας 11. Εγγενείς γραμμικές συναρτήσεις, οι κατάλληλοι μετασχηματισμοί και οι γραμμικές συναρτήσεις που προκύπτουν από τους μετασχηματισμούς

Το βασικότερο πλεονέκτημα όταν είναι γνωστό πως η μορφή της συνάρτησης παλινδρόμησης είναι εγγενής γραμμική είναι πως υπάρχει η δυνατότητα να

εκτιμηθούν οι παράμετροι της συνάρτησης (το ίδιο εύκολα όπως στη γραμμική παλινδρόμηση) με τη μέθοδο των ελαχίστων τετραγώνων. Αυτό συμβαίνει γιατί η συνάρτηση του αθροίσματος των τετραγώνων των σφαλμάτων παραμένει γραμμική ως προς τις παραμέτρους. Για κάθε εγγενή γραμμική συνάρτηση, η αντίστοιχη στοχαστική συνάρτηση που σχηματίζεται προσθέτοντας θόρυβο[11] (ε_i) δεν είναι πάντα εγγενής γραμμική. Παραδείγματος χάριν, το εκθετικό στοχαστικό μοντέλο $y = \alpha e^{\beta x} + \varepsilon_i$ και το στοχαστικό μοντέλο δύναμης $y = \alpha x^{\beta} + \varepsilon_i$ (θεωρώντας προσθετικό θόρυβο ε_i στα μοντέλα), δεν είναι εγγενείς στοχαστικές συναρτήσεις γιατί ο μετασχηματισμός της λογαρίθμησης εφαρμόζεται σε άθροισμα έχοντας δυο όρους: την ανεξάρτητη μεταβλητή x και το θόρυβο ε_i και άρα δε μπορούν αυτά να διαχωριστούν.

Αν όμως θεωρηθεί ότι ο θόρυβος ε_i πολλαπλασιάζεται στα μοντέλα, δηλαδή $y = \alpha e^{\beta x} \cdot \varepsilon_i$ ή $y = \alpha x^{\beta} \cdot \varepsilon_i$, τότε ο διαχωρισμός μπορεί να συμβεί. Μάλιστα αν ο θόρυβος ε_i έχει λογαριθμική κανονική κατανομή (lognormal distribution) τότε ο μετασχηματισμός δίνει θόρυβο $\varepsilon_i' = \ln\varepsilon_i$ με κανονική κατανομή. Για τις δύο άλλες εγγενείς γραμμικές συναρτήσεις (δηλαδή τις δύο τελευταίες στον παραπάνω Πίνακα 11), ο θόρυβος είναι προσθετικός και τότε οι στοχαστικές συναρτήσεις $y = \alpha + \beta\log(x) + \varepsilon_i$ και $y = \alpha + \beta\frac{1}{x} + \varepsilon_i$ είναι εγγενείς γραμμικές, δηλαδή ο μετασχηματισμός δίνει ισοδύναμο γραμμικό μοντέλο παλινδρόμησης.

Θα πρέπει να σημειωθεί ιδιαίτερα πως οι εκτιμήσεις των παραμέτρων στο μετασχηματισμένο γραμμικό μοντέλο παλινδρόμησης είναι οι καλύτερες (στην περίπτωση εγγενούς γραμμικής συνάρτησης παλινδρόμησης) καθότι από αυτές μπορούν να εκτιμηθούν και οι παράμετροι του αρχικού μη-γραμμικού μοντέλου παλινδρόμησης, μόνο που αυτές οι εκτιμήσεις δεν είναι και οι καλύτερες (με την έννοια της ελαχιστοποίησης των σφαλμάτων). Για να επιτευχθεί κάτι τέτοιο θα πρέπει να εφαρμοστεί η μέθοδος των ελαχίστων τετραγώνων απευθείας στο αρχικό μη-γραμμικό σύστημα. Η μέθοδος αυτή βέβαια απαιτεί τη λύση ενός μη-γραμμικού συστήματος εξισώσεων ως προς τις παραμέτρους, η οποία μπορεί να είναι αρκετά σύνθετη ανάλογα με τη μορφή της μη-γραμμικής συνάρτησης παλινδρόμησης.

[11] Το ε_i ονομάζεται σφάλμα παλινδρόμησης (regression error), λανθασμένος όρος, διαταρακτικός όρος, ή θόρυβος. Αυτή η μεταβλητή ε_i καλύπτει όλους τους άλλους παράγοντες που επηρεάζουν την εξαρτημένη μεταβλητή (Y), εκτός από τις ερμηνευτικές μεταβλητές x_i. Η σχέση μεταξύ του όρου σφάλματος και των μεταβλητών, για παράδειγμα αν συσχετίζονται, αποτελεί κρίσιμο βήμα για τη διαμόρφωση ενός μοντέλου γραμμικής παλινδρόμησης, καθώς θα καθορίσει τη μέθοδο που θα χρησιμοποιηθεί για την εκτίμηση του.

5.2.2.2 Πολυωνυμική Παλινδρόμηση

Το κοινό χαρακτηριστικό των μη-γραμμικών μοντέλων παλινδρόμησης που δίνονται από εγγενείς γραμμικές συναρτήσεις της εξαρτημένης μεταβλητής Y προς την ανεξάρτητη μεταβλητή X είναι ότι οι συναρτήσεις είναι μονότονες, αύξουσες ή φθίνουσες. Η θεωρητική τους προσέγγιση ή το διάγραμμα διασποράς συνιστά ότι η συνάρτηση έχει ένα ή περισσότερα σημεία καμπής. Σε παρόμοιες περιπτώσεις η πολυωνυμική συνάρτηση κάποιου βαθμού k μπορεί να αποτελεί ικανοποιητική προσέγγιση της πραγματικής συνάρτησης παλινδρόμησης. Το μοντέλο πολυωνυμικής γραμμικής παλινδρόμησης βαθμού k (k-th degree polynomial regression model) δίνεται από την παρακάτω εξίσωση, όπου υπάρχει η υπόθεση πως τα σφάλματα παλινδρόμησης ακολουθούν κανονική κατανομή με μέση τιμή 0 και διασπορά σ_ε^2.

$$y = \beta_0 + \beta_1\chi + \beta_2\chi^2 + ...+ \beta_k\chi^k + \varepsilon_i$$

Εξίσωση 5. μοντέλο πολυωνυμικής γραμμικής παλινδρόμησης βαθμού k

Η υπόθεση αυτή επιτρέπει αρχικά να εκτιμηθούν τα διαστήματα εμπιστοσύνης και να γίνουν οι έλεγχοι για τις παραμέτρους του μοντέλου, και στη συνέχεια να εκτιμηθούν τα κατάλληλα διαστήματα πρόβλεψης. Η μέθοδος ελαχίστων τετραγώνων όμως δεν προϋποθέτει κανονικότητα των σφαλμάτων για να δώσει τις καλύτερες (σημειακές) εκτιμήσεις των παραμέτρων. Η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο ελαχίστων τετραγώνων όπως και για το γραμμικό μοντέλο παλινδρόμησης γιατί ενώ η πολυωνυμική συνάρτηση παλινδρόμησης είναι μη-γραμμική ως προς την ανεξάρτητη μεταβλητή x, είναι γραμμική ως προς τους συντελεστές $\beta_0, \beta_1, \ldots, \beta_k$. Το άθροισμα των τετραγώνων των σφαλμάτων για κάποιο διμεταβλητό δείγμα μεγέθους n των (X,Y), είναι $f(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (yi - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ...+ \beta_k x_i^k))^2$. Τα σφάλματα του μοντέλου πολυωνυμικής παλινδρόμησης που εκτιμήθηκε με τη μέθοδο ελαχίστων τετραγώνων είναι $e_i = y_i - \hat{y}_i$,όπου $\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ...+ \beta_k x_i^k$. Η εκτίμηση της διασποράς των σφαλμάτων e_i ορίζεται ως: $\sigma_e^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Επιπρόσθετα όπως και για την απλή γραμμική παλινδρόμηση που αναφέρθηκε σε προηγούμενη ενότητα ορίζεται ο συντελεστής του πολλαπλού προσδιορισμού (coefficient of multiple determination) R^2 ή αλλιώς και συντελεστής συσχέτισης R^2 που δείχνει το ποσοστό της διακύμανσης της εκάστοτε εξαρτημένης μεταβλητής που «εξηγείται» από την υπό μελέτη ανεξάρτητη μεταβλητή και δίνεται από τον τύπο:

$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$ που δηλώνει την αναλογία της μεταβλητότητας που εξηγείται από το μοντέλο. Είναι φυσικό λοιπόν πως αν προστεθούν περισσότεροι μη-γραμμικοί όροι (δυνάμεις) της ανεξάρτητης μεταβλητής X στο πολυωνυμικό μοντέλο παλινδρόμησης θα βελτιωθεί η προσαρμογή του στις n ζευγαρωτές παρατηρήσεις, χωρίς βέβαια αυτό να σημαίνει ότι ένας μεγάλος βαθμός k είναι πάντα ο πιο κατάλληλος. Γι αυτό το λόγο χρησιμοποιείται ο προσαρμοσμένος συντελεστή του πολλαπλού προσδιορισμού (adjusted coefficient of multiple determination), ο οποίος δίνει μικρότερες τιμές από το R^2 της σχέσης που το ποσό μείωσης προσαρμόζεται στο πλήθος των μη-γραμμικών όρων k και δίνεται από τον τύπο: $adjR^2 = 1 - \frac{n-1}{n-(k+1)} \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$

Για τις παραμέτρους $\beta_0,\beta_1,\ldots,\beta_k$, όπως και για τις παραμέτρους του γραμμικού μοντέλου παλινδρόμησης, μπορούν να εκτιμηθούν τα ανάλογα διαστήματα εμπιστοσύνης και να γίνουν οι στατιστικοί έλεγχοι, αφού πρωτίστως γίνει η υπόθεση πως τα σφάλματα ακολουθούν κανονική κατανομή με σταθερή διασπορά για κάθε τιμή του x. Απο την άλλη, η διασπορά της εκτίμησης για κάθε μια από τις $\beta_0,\beta_1,\ldots,\beta_k$, και τα διαστήματα πρόβλεψης για την πολυωνυμική παλινδρόμηση δίνονται από σχετικά πολύπλοκους τύπους που εξαρτώνται και από το βαθμό του πολυώνυμου.

Για παράδειγμα στον παρακάτω Πίνακα 12, δίνονται τα δεδομένα για την ημέρα της συγκομιδής (αριθμός ημερών αφού ανθίσει) και του μεγέθους σοδειάς (σε kg/ha) ενός είδους Ινδικού ρυζιού που λέγεται paddy.

A/A	Ημέρες	Σοδειά
1	16	2508
2	18	2518
3	20	3304
4	22	3423
5	24	3057
6	26	3190
7	28	3590
8	30	3883
9	32	3823
10	34	3646
11	36	3708
12	38	3333
13	40	3517
14	42	3241
15	44	3103
16	46	2776

Πίνακας 12. Τιμές ημερών για τη συγκομιδή και μεγέθους σοδειάς για το Ινδικό ρύζι

Υπολογίζοντας τα σφάλματα προσαρμογής του κάθε μοντέλου και αντίστοιχα τον συντελεστή προσδιορισμού R^2 και τον προσαρμοσμένο συντελεστή προσδιορισμού

adjR^2 διαφαίνεται πως το γραμμικό μοντέλο παλινδρόμησης δεν είναι κατάλληλο. Επίσης ο συντελεστής προσδιορισμού είναι πολύ κοντά στο 0, που σημαίνει ότι το μοντέλο αδυνατεί να ερμηνεύσει τις παρατηρήσεις. Η πρόσθεση του όρου του τετραγώνου των ημερών για τη συγκομιδή (ανεξάρτητη μεταβλητή) δίνει την παραβολή που προσαρμόζεται πολύ καλά στα συγκεκριμένα δεδομένα. Ο συντελεστής προσδιορισμού R^2, όπως και ο προσαρμοσμένος συντελεστής προσδιορισμού adjR^2, υπολογίζονται άνετα που σημαίνει ότι με το πολυωνυμικό μοντέλο δευτέρου βαθμού μπορεί να εκτιμηθεί η σοδειά του Ινδικού ρυζιού (paddy) όταν δίνεται ο αριθμός των ημερών για τη συγκομιδή. Η παραπέρα αύξηση του βαθμού του πολυωνύμου δε φαίνεται να επιφέρει βελτίωση στην παλινδρόμηση του μεγέθους σοδειάς του paddy προς τον αριθμό ημερών για τη συγκομιδή. Έτσι, το R^2 παραμένει το ίδιο ενώ το adjR^2 μειώνεται. Το πολυωνυμικό μοντέλο δευτέρου βαθμού εκτιμάται να είναι: $y = -1.1242 + 0.2979x - 0.0046x^2$ και υπάρχει η δυνατότητα χρήσης του σε προβλέψεις του μεγέθους της σοδειάς για κάθε δεδομένη χρονική περίοδο μέχρι τη συγκομιδή.

5.2.3 Πολλαπλή Γραμμική Παλινδρόμηση

Μη-γραμμικοί όροι (όπως δυνάμεις των (υποθετικά) ανεξάρτητων μεταβλητών $x_1, x_2, \ldots, x_k$ και όροι αλληλεπίδρασης τους) είναι δυνατόν να εμπεριέχονται σε ένα μοντέλο πολλαπλής παλινδρόμησης. Στη γενική του μορφή ένα τέτοιο μοντέλο λέγεται μοντέλο προσθετικής πολλαπλής παλινδρόμησης (additive multiple regression model), και λέγεται «προσθετικής» καθότι όλοι οι όροι του μοντέλου περιλαμβάνονται σ' αυτό αθροιστικά. Δηλαδή, έχουμε μια εξαρτημένη μεταβλητή και περισσότερες απο μία ανεξάρτητες (Μαυρωτάς, Τσιαφογιάννη, Διακουλάκη, Σαρίμβεης, 2011). Τα δυνατά μοντέλα προσθετικής πολλαπλής παλινδρόμησης είναι τα ακόλουθα και αποτελούν ένα παράδειγμα όπου χρησιμοποιούνται δύο ανεξάρτητες μεταβλητές x_1 και x_2:

1. Το μοντέλο πρώτου πολυωνυμικού βαθμού *(γραμμικής πολλαπλής παλινδρόμησης)*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$$

2. Το μοντέλο δεύτερου πολυωνυμικού βαθμού χωρίς αλληλεπίδραση

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \varepsilon_i$$

3. Το μοντέλο πρώτου πολυωνυμικού βαθμού με αλληλεπίδραση

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_i$$

4. Το πλήρες μοντέλο δευτέρου πολυωνυμικού βαθμού (με αλληλεπίδραση)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon_i$$

Βεβαίως και υπάρχουν διαφορές ανάμεσα στα τέσσερα παραπάνω μοντέλα (ως προς τις μεταβλητές x_1 και x_2) αλλά όλα είναι γραμμικά ως προς τις παραμέτρους που η εκτίμηση τους μπορεί να γίνει με την κλασσική μέθοδο των ελαχίστων τετραγώνων. Το ίδιο βέβαια ισχύει και όταν οι μεταβλητές είναι περισσότερες των δύο. Επίσης είναι σημαντικό να αναφερθούν δύο λόγια για τις γραφικές εκτιμήσεις των μοντέλων. Όταν οι ανεξάρτητες μεταβλητές είναι δύο, η γραφική εκτίμηση του κατάλληλου μοντέλου μπορεί να γίνει από το γράφημα της (x_1, y) για διαφορετικές τιμές της x_2 (ή ισοδύναμα αντιστρέφοντας τις θέσεις των x_1 και x_2). Για τα τέσσερα μοντέλα δύο ανεξάρτητων μεταβλητών που αναφέρθηκαν προηγουμένως θα περίμενε κανείς τα εξής (αγνοώντας την ύπαρξη του θορύβου ε_i):

7. Το γράφημα των σημείων (x_1, y) είναι σε παράλληλες ευθείες για κάθε τιμή του x_2 γιατί η μεταβολή της y είναι ανεξάρτητη της x_2, για κάποια μεταβολή της x_1 (π.χ. κατά μια μονάδα)

8. Το γράφημα των σημείων (x_1, y) είναι σε παράλληλες καμπύλες παραβολής για κάθε τιμή του x_2 εξαιτίας της παρουσίας των τετραγωνικών όρων

9. Το γράφημα των σημείων (x_1, y) είναι σε ευθείες για κάθε τιμή του x_2 οι οποίες τέμνονται γιατί η μεταβολή της y ως προς τη x_1 δεν είναι τώρα ανεξάρτητη της x_2 επειδή είναι παρών ο όρος αλληλεπίδρασης

10. Το γράφημα των σημείων (x_1, y) είναι σε καμπύλες παραβολής για κάθε τιμή του x_2 που δεν είναι παράλληλες εξαιτίας της παρουσίας του όρου αλληλεπίδρασης.

Πρέπει επίσης να σημειωθεί ότι όταν η παλινδρόμηση αφορά περισσότερες από δύο μεταβλητές δεν υπάρχουν γραφικά εργαλεία για να καθοριστεί η μορφή του μοντέλου προσθετικής πολλαπλής παλινδρόμησης και γι' αυτό το λόγο πρέπει να γίνουν δοκιμές διαφορετικών μοντέλων και η κατάλληλη επιλογή ενός εξ αυτών.

Ένα παράδειγμα πάνω στην πολλαπλή γραμμική παλινδρόμηση είναι το ακόλουθο: *Σε μελέτη της επίδρασης γεωργικών χημικών στην προσρόφηση ιζημάτων και εδάφους, υπάρχουν 13 δεδομένα για το δείκτη προσρόφησης φωσφορικού άλατος (y), για το εξαγώγιμο σίδηρο (x_1) και το εξαγώγιμο αργίλιο (x_2) όπως παρουσιάζονται στον Πίνακα 13. Το μοντέλο που θα εκτιμηθεί είναι:* $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$.

Α/Α	Εξαγώγιμο Σίδηρο	Εξαγώγιμο Αργίλλιο	Δείκτης Προσρόφησης
1	61	13	4
2	175	21	18
3	111	24	14
4	124	23	18
5	130	64	26
6	173	38	26
7	169	33	21
8	169	61	30
9	160	39	28
10	244	71	36
11	257	112	65
12	333	88	62
13	199	54	40

Πίνακας 13. Τιμές του δείκτη προσρόφησης, του εξαγώγιμου σιδήρου και εξαγώγιμου αργιλίου για τη μελέτη της επίδρασης γεωργικών χημικών στο έδαφος.

Οι εκτιμήσεις των συντελεστών του μοντέλου με τη μέθοδο των ελαχίστων τετραγώνων και η εκτίμηση της τυπικής τους απόκλισης δίνονται στον παρακάτω πίνακα και είναι οι εξής:

ΠΑΡΑΜΕΤΡΟΣ	ΕΚΤΙΜΗΤΗΣ b_i	ΕΚΤΙΜΗΣΗ SD s_{bi}
β_0	-7351	3485
β_1	0.11273	0.02969
β_2	0.34900	0.07131

Πίνακας 14. Εκτίμηση παραμέτρων και τυπική απόκλισης τους

Το 95% διάστημα εμπιστοσύνης[12] για το συντελεστή του εξαγώγιμου σιδήρου β_1 είναι ($t_{10,0.975} = 2.228$)

$$0.11273 \pm 2.228 * 0.02969 = [0.0466, 0.1789]$$

και αντίστοιχα για το συντελεστή του εξαγώγιμου αργιλλίου β_2 είναι

$$0.34900 \pm 2.228 * 0.07131 = [0.1901, 0.5079].$$

Παρατηρείται λοιπόν πως και τα δύο διαστήματα εμπιστοσύνης δεν περιέχουν το 0, αλλά συμπεραίνεται με 95% βεβαιότητα πως το εξαγώγιμο σίδηρο και αργίλιο επηρεάζουν σημαντικά το δείκτη προσρόφησης φωσφορικού άλατος και καλώς

[12] Ο υπολογισμός γίνεται για τα παραμετρικά διαστήματα εμπιστοσύνης για τις παραμέτρους $\beta_0,\beta_1,\ldots,\beta_k$ για τις οποίες όμως η εκτίμηση της διασποράς είναι σύνθετη. Γενικά αν η εκτίμηση της διασποράς είναι $s^2_{b_j}$ για $j = 0, 1, \ldots, k$ τότε το $(1-\alpha)$% διάστημα εμπιστοσύνης για το συντελεστή bj είναι: $b_j \pm t_{n-(k+1),1-a/2} s_{b_j}$. Ο έλεγχος για την τιμή β_j^o της β_j, $H_o: \beta_j = \beta_j^o$ γίνεται με το στατιστικό $t = \frac{\beta_j - \beta_j^o}{s_{b_j}} \sim t_{n-(k+1)}$. Το $(1-\alpha)$% διάστημα εμπιστοσύνης για τη μέση τιμή της y όταν δίνονται τα $x_1, \ldots, x_k$ είναι: $\hat{y} \pm t_{n-(k+1),1-a/2} s_{\hat{y}}$ όπου η διασπορά της εκτίμησης $\hat{y}$, $s^2_{\hat{y}}$, δίνεται από επίσης σύνθετη έκφραση. Αντίστοιχα το $(1-\alpha)$% διάστημα πρόβλεψης μιας (μελλοντικής) τιμής της y είναι: $\hat{y} \pm t_{n-(k+1),1-\frac{a}{2}} \sqrt{s^2_\varepsilon + s^2_{\hat{y}}}$.

συμπεριλαμβάνονται στο μοντέλο. Πρέπει επίσης να αναφερθεί πως η τυπική απόκλιση των σφαλμάτων είναι s_e = 4.616, ο συντελεστής του πολλαπλού προσδιορισμού είναι R^2 = 0.948 ή 94,8% και ο προσαρμοσμένος συντελεστής είναι $adjR^2$ = 0.931 ή 93,1%. Αν υποτεθεί η πρόβλεψη του δείκτη προσρόφησης *y* όταν ο εξαγώγιμος σίδηρος είναι x_1 = 160 και ο εξαγώγιμος αργίλλιος είναι x_2 = 39, τότε προκύπτει : $\hat{y} = -7.351 + 0.11273 * 160 + 0.34900 * 39 = 24.30$.

Η εκτίμηση της τυπικής απόκλισης για αυτήν την πρόβλεψη ŷ βρέθηκε να είναι $s_{\hat{y}}$ = 1.30. Το 95% διάστημα εμπιστοσύνης για το μέσο δείκτη προσρόφησης *y* όταν ο εξαγώγιμος σίδηρος είναι x_1 = 160 και ο εξαγώγιμος αργίλιος είναι x_2 = 39 βρίσκεται ως 24.30 ± 2.228 * 1.30 = [21.40, 27.20] και το αντίστοιχο 95% διάστημα πρόβλεψης για μια μελλοντική τιμή του *y* (για x_1 = 160 και x_2 = 39) είναι: $24.30 \pm 2.228 \sqrt{4.616^2 + 1.30^2}$ = [13.62, 34.98]

5.3 Μελέτη του Μοντέλου της Απλής Γραμμικής Παλινδρόμησης

Όπως προαναφέρθηκε σε προηγούμενη ενότητα το μοντέλο της απλής γραμμικής παλινδρόμησης περιλαμβάνει μια ανεξάρτητη μεταβλητή και τη γραμμική συνάρτηση παλινδρόμησης. Τα β_0 και β_1 αποτελούν τις άγνωστες παραμέτρους του μοντέλου, τα y_i, i=1,2,….,n είναι οι τιμές (συνήθως λέγονται και αποκρίσεις) της εξαρτημένης μεταβλητής Y, τα x_i είναι οι τιμές της ανεξάρτητης (ελεγχόμενης) μεταβλητής X που θεωρούνται γνωστές σταθερές και n είναι το πλήθος των δεδομένων (x_i,y_i), που πρόκειται να αναλυθούν. Τα ε_i θεωρούνται ασυσχέτιστες ανά δύο τυχαίες μεταβλητές με $E(\varepsilon_i)=0$ και $Var(\varepsilon_i)=\sigma^2$. Όπως ειπώθηκε σε προηγούμενη ενότητα το ε_i λέγεται *τυχαίο σφάλμα ή σφάλμα παλινδρόμησης* και παριστά, για δοθείσα τιμή x_i της X, την απόκλιση της αντίστοιχης τιμής y_i της Y από την άγνωστη γραμμή παλινδρόμησης, δηλαδή $\varepsilon_i=y_i-E(y_i)$.

Το παραπάνω απλό μοντέλο παλινδρόμησης, περιέχει μια ανεξάρτητη μεταβλητή και είναι γραμμικό ως προς τις παραμέτρους, γιατί καμιά παράμετρος δεν εμφανίζεται σε εκθέτη ή πολλαπλασιάζεται ή διαιρείται με άλλη παράμετρο και επιπλέον είναι γραμμικό και ως προς την ανεξάρτητη μεταβλητή, γιατί το X εμφανίζεται μόνο στην πρώτη δύναμη. Ένα μοντέλο παλινδρόμησης με τις παραπάνω ιδιότητες λέγεται και *μοντέλο πρώτης τάξης*. Σύμφωνα με τις υποθέσεις του μοντέλου τα y_i αποτελούνται από τον σταθερό όρο $\beta_0+\beta_1x_i$ και τον τυχαίο όρο ε_i. Είναι δηλαδή, τα Y ασυσχέτιστες τυχαίες μεταβλητές με $E(y_i) = \beta_0=\beta_1x_i$ και $Var(Y_i)=\sigma^2$. Για κάποια τιμή της X, έστω x_i, η τυχαία μεταβλητή Y έχει κατανομή με μέση τιμή

$E(Y)=\beta_0+\beta_1 X$, που αποτελεί και τη *συνάρτηση παλινδρόμησης* του παραπάνω μοντέλου.

Το τυχαίο σφάλμα ε_i, για δοθείσα τιμή x_i, ισούται με την απόκλιση της συγκεκριμένης τιμής της y_i από την αντίστοιχη τιμή της συνάρτησης παλινδρόμησης. Οι παράμετροι β_0 και β_1 του μοντέλου λέγονται και *συντελεστές παλινδρόμησης*. Το β_1 είναι η κλίση της γραμμής παλινδρόμησης και δείχνει τη μεταβολή της μέσης τιμής του Y για κάθε μοναδιαία αύξηση στο X. Το β_0 είναι το σημείο που η γραμμή παλινδρόμησης τέμνει τον άξονα των Y και δίνει τη μέση τιμή της Y για κάθε X=0. Βασική υπόθεση του μοντέλου είναι ότι η διακύμανση των y_i είναι σταθερή για τις διάφορες τιμές της ανεξάρτητης μεταβλητής, είναι δηλαδή ανεξάρτητη του X.

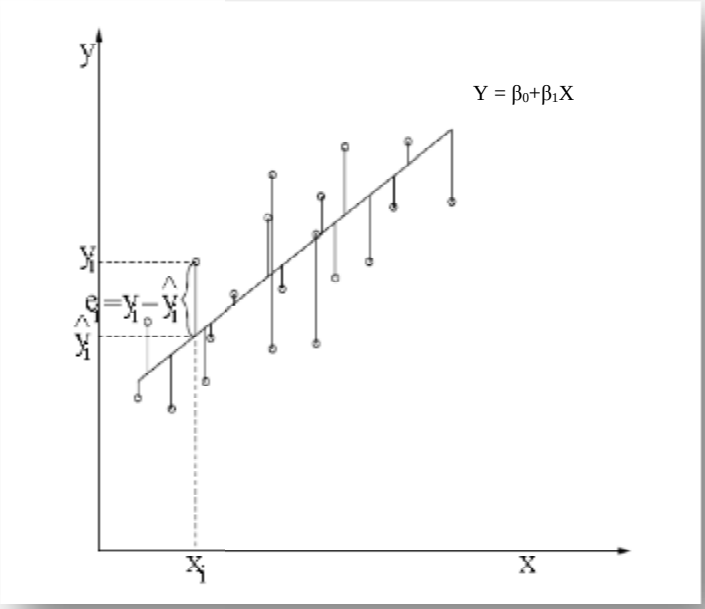
5.3.1 Εκτίμηση της Συνάρτησης Παλινδρόμησης

Υποθέτοντας ότι το μοντέλο της απλής γραμμικής παλινδρόμησης είναι κατ' αρχήν κατάλληλο για να εκφράσει τη σχέση μεταξύ της εξαρτημένης μεταβλητής Y και της ανεξάρτητης μεταβλητής X, που πρόκειται να μελετηθεί, και έστω ότι υπάρχουν στην διάθεση του ερευνητή – μελετητή οι παρατηρήσεις (x_i,y_i), i=1,2,….,n, που πάρθηκαν μετά από κάποιο πείραμα. Οι τιμές των παραμέτρων β_0 και β_1 είναι άγνωστες και θα πρέπει να εκτιμηθούν βάσει της τεχνικής που είναι γνωστή ως μέθοδος ελαχίστων τετραγώνων. Για κάθε παρατήρηση (X_i,Y_i) το σφάλμα ελαχίστων τετραγώνων ή απλά υπόλοιπο (residual), δίνει την απόκλιση της y_i από την (άγνωστη) αναμενόμενη τιμή της (Εικ. 11), αφού $e_i = y_i - \hat{y}_i = y_i - (\beta_0+\beta_1 x_i)$. Έστω $Q(\beta_o,\beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \beta_o - \beta_1 x_i)^2$, το άθροισμα των τετραγώνων των n αποκλίσεων. Με τη μέθοδο των ελαχίστων τετραγώνων, οι εκτιμητές των β_0 και β_1 είναι αντίστοιχα εκείνες οι τιμές $\hat{\beta}_o$ και $\hat{\beta}_1$, που ελαχιστοποιούν την συνάρτηση $Q(\beta_0,\beta_1)$ ως προς β_0 και β_1 και εξισώνοντας το αποτέλεσμα με το μηδέν. Τελικά βρίσκουμε ότι ισχύει:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}},$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n}$$

Εξίσωση 6. Τύποι εξισώσεων για την εύρεση των εκτιμητών $\hat{\boldsymbol{\beta}}_{\mathbf{o}}$ και $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$



Εικόνα 11. Ευθεία Ελαχίστων Τετραγώνων και Κατάλοιπα

Τα $\hat{\beta}_o$ και $\hat{\beta}_1$ λέγονται και εκτιμητές ελαχίστων τετραγώνων των β_0 και β_1 αντίστοιχα και έχουν πολύ καλές στατιστικές ιδιότητες, για παράδειγμα είναι αμερόληπτοι εκτιμητές ελάχιστης διακύμανσης. Γνωρίζοντας τώρα τους εκτιμητές των παραμέτρων της συνάρτησης παλινδρόμησης $E(Y)=\beta_0+\beta_1X$, είναι φυσικό να την εκτιμηθεί με την $\hat{Y}= \hat{\beta}_o + \hat{\beta}_1 X$, όπου $\hat{Y}$ είναι η τιμή της εκτιμώμενης συνάρτησης παλινδρόμησης. Είναι δηλαδή το $\hat{Y}$ ένας εκτιμητής της μέσης τιμής της εξαρτημένης μεταβλητής Y, όταν δίνεται ότι η ανεξάρτητη μεταβλητή παίρνει την τιμή X. χρησιμοποιώντας την παραπάνω σχέση είναι κάποιος στη θέση να προβλέψει, δοθείσης μιας τιμής x_0 της X, την αντίστοιχη τιμή της Y. Συνεπώς, $\hat{y}_o= \hat{\beta}_o + \hat{\beta}_1 \hat{x}_o$ είναι ο εκτιμητής της μέσης τιμής της Y, δηλαδή το $\hat{y}_o$ είναι η προβλεπόμενη τιμή, που αντιστοιχεί στο $\hat{x}_o$.

Στο παρακάτω παράδειγμα δίνονται τα δεδομένα σπουδών πατέρα και παιδιού για 10 οικογένειες (Πίνακας 15).

Αύξων αριθμός οικογένειας (i)	1	2	3	4	5	6	7	8	9	10
Διάρκεια σπουδών πατέρα (X_i)	3	6	6	8	2	8	6	8	3	10
Διάρκεια σπουδών παιδιού (Y_i)	5	9	13	14	4	16	11	12	8	18

Πίνακας 15. Παρατηρήσεις για τις σπουδές πατέρα και παιδιού

Για να βρούμε τους εκτιμητές $\hat{\beta}_o$ και $\hat{\beta}_1$ στην περίπτωση των δεδομένων του παραπάνω πίνακα, υπολογίζονται οι ποσότητες που εμφανίζονται στις σχέσεις της εξίσωσης 6 (μαθηματικοί τύποι των εκτιμητών). Οπότε αντικαθιστώντας προκύπτουν τα εξής αποτελέσματα:

$$\sum X_i = 3 + 6 + \cdots + 10 = \mathbf{60}$$

$$\bar{X} = \frac{\mathbf{60}}{\mathbf{10}} = \mathbf{6}$$

$$\sum X_i^{\mathbf{2}} = \mathbf{3^2} + \mathbf{6^2} + \cdots + \mathbf{10^2} = \mathbf{422}$$

$$\sum X_i Y_i = 3 * 5 + 6 * 9 + \cdots + 10 * 18 = \mathbf{761}$$

$$\sum Y_i = 5 + 9 + \cdots + 18 = \mathbf{110}$$

$$\bar{Y} = \frac{\mathbf{110}}{\mathbf{10}} = \mathbf{11}$$

Μετά από αντικατάσταση στους τύπους των αντίστοιχων τιμών από τις παραπάνω σχέσεις προκύπτει ότι:

$$\widehat{\beta_{\mathbf{1}}} = \frac{\sum_{i=\mathbf{1}}^{n} X_i Y_i - \left(\sum_{i=\mathbf{1}}^{n} X_i\right)\left(\sum_{i=\mathbf{1}}^{n} Y_i\right)/_n}{\sum_{i=\mathbf{1}}^{n} X_i^{\mathbf{2}} - \left(\sum_{i=\mathbf{1}}^{n} X_i\right)^{\mathbf{2}}/_n} = \frac{761 - 60 * \frac{\mathbf{110}}{\mathbf{10}}}{422 - \frac{\mathbf{60^2}}{\mathbf{10}}} = 1.62\mathbf{9}$$

$$\widehat{\beta_{\mathbf{0}}} = \bar{Y} - \widehat{\beta_{\mathbf{1}}}\bar{X} = 11 - 1.629 * 6 = 1.22\mathbf{6}$$

Εκτιμάμε δηλαδή ότι θα έχουμε, κατά μέσο όρο, μια αύξηση 1.629 έτη στις σπουδές του παιδιού για κάθε μοναδιαία αύξηση στα έτη σπουδών του πατέρα. Η εκτιμώμενη συνάρτηση παλινδρόμησης είναι: $\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 X$ =1.226 + 1.629X.

Έτσι, για παράδειγμα δύναται να προβλεφθεί κατά μέσο όρο, ότι αν η διάρκεια σπουδών κάποιου πατέρα είναι X=7 έτη, η διάρκεια σπουδών του παιδιού θα είναι $\hat{Y}$=1.226+1.629*7=12.629 έτη.

Για την τιμή που βρέθηκε, μπορεί να ειπωθεί ότι αν βρεθούν πολλές οικογένειες στις οποίες οι σπουδές για τον πατέρα διήρκησαν 7 έτη, τότε κατά μέσο όρο η διάρκεια σπουδών για τα παιδιά τους θα ήταν 12.629 έτη. Βέβαια για κάποια μεμονωμένη περίπτωση είναι πολύ πιθανόν η διάρκεια σπουδών να είναι μικρότερη ή μεγαλύτερη από 12.629 έτη, λόγω της μεταβλητότητας που υπάρχει στο σύστημα και εκφράζεται

στο μοντέλο με τα σφάλματα ε_i. Χρησιμοποιώντας την εκτιμώμενη συνάρτηση παλινδρόμησης μπορεί να υπολογιστεί για κάθε μία από τις τιμές της ανεξάρτητης μεταβλητής X των δεδομένων, την εκτιμώμενη τιμή της εξαρτημένης μεταβλητής Y. Για παράδειγμα, για την τιμή $X_1=3$ των δεδομένων του πίνακα, προκύπτει $\hat{Y}$ =1.226+1.629*3=6.113, όπου διαφέρει από την παρατηρηθείσα τιμή $Y_1=6$. Η διαφορά αυτή ονομάζεται χαρακτηριστικά υπόλοιπο (residual). Για τα δοθέντα δεδομένα τα υπόλοιπα (residuals) συμβολίζονται με e_i και υπολογίζονται ως οι διαφορές $e_i = y_i - \hat{y}_i$, i=1,2,…,n και αποτελούν τις κατακόρυφες αποκλίσεις των παρατηρήσεων Y_i από την ευθεία της εκτιμώμενης συνάρτησης Παλινδρόμησης.

Τα υπόλοιπα e_i δεν πρέπει να συγχέονται με τα τυχαία σφάλματα $\varepsilon_i=Y_i-E(Y_i)$, τα οποία ως γνωστόν αποτελούν τις (άγνωστες) κατακόρυφες αποκλίσεις των Y_i από την ευθεία της συνάρτησης παλινδρόμησης, που περιλαμβάνει τις άγνωστες παραμέτρους β_0 και β_1 και είναι φυσικά άγνωστη. Τα υπόλοιπα e_i είναι πολύ χρήσιμα στη μελέτη της καταλληλότητας ενός μοντέλου παλινδρόμησης.

Μπορεί τέλος να αποδειχθεί, ότι ένας αμερόληπτος εκτιμητής της άγνωστης διακύμανσης σ^2 των σφαλμάτων ε_i, άρα και των Y_i, δίνεται από τη στατιστική συνάρτηση (μέσο τετραγωνικό σφάλμα):

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

Εξίσωση 7. Τύπος υπολογισμού μέσου τετραγωνικού σφάλματος

Το μέσο τετραγωνικό σφάλμα υπολογίζεται ευκολότερα από την ισοδύναμη σχέση

$$S^2 = \frac{1}{n-2}\left[\left(\sum Y_i^2 - n\bar{Y}^2\right) - \hat{\beta}_1^2\left(\sum X_i^2 - n\bar{X}^2\right)\right].$$

Εξίσωση 8. Εναλλακτικός τύπος υπολογισμού μέσου τετραγωνικού σφάλματος

Σύμφωνα με τις δοθέντες παρατηρήσεις του Πίνακα 16 αν υπολογιστούν οι απαραίτητες ποσότητες των παραπάνω εξισώσεων ξεχωριστά δύναται να βρεθεί το μέσο τετραγωνικό σφάλμα. Οπότε προκύπτει ότι:

$\mathbf{Y_i}$	5	9	13	14	4	16	11	12	8	18
$\hat{\mathbf{Y}}_\mathbf{i}$	6.113	11	11	14.258	4.484	14.258	11	14.258	6.113	17.516

Πίνακας 16. Παρατηρήσεις για τον υπολογισμό του μέσου τετραγωνικού σφάλματος

$$S^2 = \frac{1}{10-2}[(5-6.113)^2 + (9-11)^2 + \cdots + (18-17.516)^2] = \frac{21.47}{8} = 2.68$$

που είναι ένας εκτιμητής της άγνωστης κοινής διακύμανσης σ^2 των τυχαίων μεταβλητών Y_i. Αν γίνει η χρήση της δεύτερης εξίσωσης με $\bar{X} = 6, \bar{Y} = 11$,

$$\sum X_i^2 = 422, \sum Y_i^2 = 1396$$

,

και $\hat{\beta}_1$=1.629, οπότε

$$S^2 = \frac{1}{10-2}[(1396 - 10 * 11^2) - 1.629^2 * (422 - 10 * 6^2)] = \frac{21.47}{8} = 2.68$$

5.3.2 Συντελεστής Προσδιορισμού

Όλη η ανάλυση που θα ακολουθήσει βασίζεται στα σφάλματα της ευθείας ελαχίστων τετραγώνων ή υπόλοιπα (e_i). Όσο μεγαλύτερη είναι η επίδραση της x επί της y τόσο μικρότερα είναι τα σφάλματα και αντίστροφα. Πρώτα θα γίνει εκτίμηση της συνολικής διασποράς γύρω από τη γραμμή της παλινδρόμησης, δηλαδή το άθροισμα των τετραγώνων των αποκλίσεων των πραγματικών τιμών της y από τις αντίστοιχες τιμές ŷ του υποδείγματος της παλινδρόμησης. Επομένως, θα υπολογιστεί το $\Sigma e^2=\Sigma(y-\hat{y})^2$, που επίσης ονομάζεται **άθροισμα των τετραγώνων των σφαλμάτων** (sum of squared errors) και συμβολίζεται με SSE που δίνεται από τον παρακάτω τύπο:

$$SSE = \sum e^2 = \sum (y-\hat{y})^2 = \sum (y-\alpha-\beta x)^2 = \sum (y-\alpha-\beta x)(y-\alpha-\beta x) = \sum y^2 - a\sum y - \beta \sum yx$$

Εξίσωση 9. Τύπος για το άθροισμα των τετραγώνων των σφαλμάτων (e_i)

Παράδειγμα[13]:

Το διοικητικό συμβούλιο μιας ασφαλιστικής εταιρίας προβληματίζεται αν θα πρέπει να αυξήσει τις δαπάνες διαφήμισης ή να προσλάβει νέους πωλητές. Στον παρακάτω Πίνακα 17 φαίνεται η εξέλιξη των πωλήσεων, των δαπανών για διαφήμιση και του αριθμού των πωλητών.

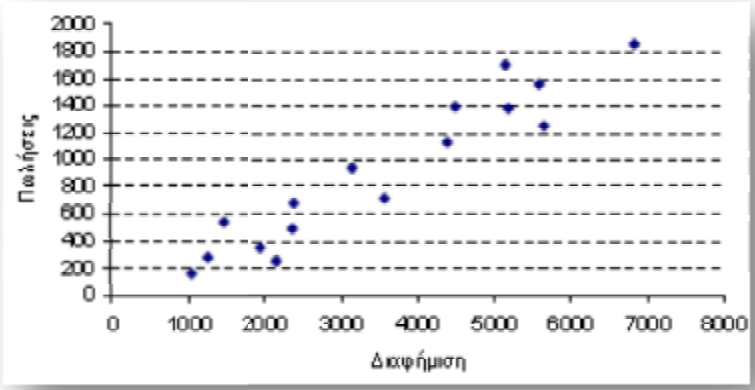
[13] Πηγή: http://androulakis.bma.upatras.gr/mediawiki/index.php/Εκτίμηση_διακύμανσης_της_ευθείας_παλινδρόμησης

Έτος	Πωλήσεις	Διαφήμιση	Πωλητές	Έτος	Πωλήσεις	Διαφήμιση	Πωλητές
1985	1050	162	32	1993	3570	720	98
1986	1260	285	47	1994	4410	1140	43
1987	1470	540	23	1995	4500	1395	76
1988	2160	261	68	1996	5610	1560	89
1989	1950	360	32	1997	5190	1380	108
1990	2400	690	17	1998	5670	1260	76
1991	2370	495	58	1999	5160	1710	65
1992	3150	948	75	2000	6840	1860	93

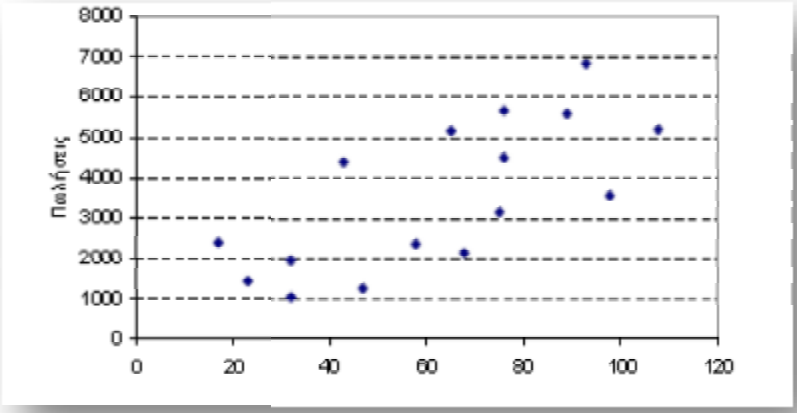
Πίνακας 17. Δεδομένα παρατηρήσεων πωλήσεων και διαφημίσεων ανά έτος 1985-2000

Ο πιο απλός τρόπος για να διαπιστωθεί αν υπάρχει συσχέτιση μεταξύ δύο μεταβλητών είναι η κατασκευή του **διαγράμματος διασποράς**

1. Το διάγραμμα διασποράς μεταξύ των πωλήσεων και διαφήμισης αποκαλύπτει τη θετική σχέση μεταξύ των 2 μεταβλητών (Εικ. 12).
2. Το διάγραμμα διασποράς μεταξύ των πωλητών και διαφήμισης αποκαλύπτει τη θετική σχέση μεταξύ των 2 μεταβλητών, η οποία δεν είναι τόσο έντονη όσο στη προηγούμενη περίπτωση (Εικ.13).



Εικόνα 12. διάγραμμα διασποράς μεταξύ των πωλήσεων και διαφήμισης



Εικόνα 13. διάγραμμα διασποράς μεταξύ των πωλητών και διαφήμισης

Η εκτίμηση του συντελεστή συσχέτισης, που συμβολίζεται με **r** προκύπτει ως εξής βάσει των υπολογισμών που παρατίθενται στους Πίνακες 18 και 19 :

$$r = \frac{16\,(66.883.410) - (56.760)\,(14.766)}{\sqrt{\left(16\,(18.289.944) - (14.766)^2\right)\left(16\,(251.029.800) - (56.760)^2\right)}} = 0.95$$

Έτος	Πωλήσεις (Υ)	Διαφήμιση (Χ)	ΥΧ	Y^2	X^2
1985	1050	162	170100	1102500	26244
1986	1260	285	359100	1587600	81225
1987	1470	540	793800	2160900	291600
1988	2160	261	563760	4665600	68121
1989	1950	360	702000	3802500	129600
1990	2400	690	1656000	5760000	476100
1991	2370	495	1173150	5616900	245025
1992	3150	948	2986200	9922500	898704
1993	3570	720	2570400	12744900	518400
1994	4410	1140	5027400	19448100	1299600
1995	4500	1395	6277500	20250000	1946025
1996	5610	1560	8751600	31472100	2433600
1997	5190	1380	7162200	26936100	1904400
1998	5670	1260	7144200	32148900	1587600
1999	5160	1710	8823600	26625600	2924100
2000	6840	1860	12722400	46785600	3459600
Αθροισμα	**56760**	**14766**	**66883410**	**251029800**	**18289944**

Πίνακας 18. Υπολογισμός παρατηρήσεων για την εύρεση συντελεστή συσχέτισης πωλήσεων και διαφήμισης

$$r = \frac{16\,(4.094.160) - (56.760)\,(1.000)}{\sqrt{\left(16\,(74.152) - (1.000)^2\right)\left(16\,(251.029.800) - (56.760)^2\right)}} = 0.72$$

Έτος	Πωλήσεις (Υ)	Πωλητές (Χ)	ΥΧ	Y^2	X^2
1985	1050	32	33600	1102500	1024
1986	1260	47	59220	1587600	2209
1987	1470	23	33810	2160900	529
1988	2160	68	146880	4665600	4624
1989	1950	32	62400	3802500	1024
1990	2400	17	40800	5760000	289
1991	2370	58	137460	5616900	3364
1992	3150	75	236250	9922500	5625
1993	3570	98	349860	12744900	9604
1994	4410	43	189630	19448100	1849
1995	4500	76	342000	20250000	5776
1996	5610	89	499290	31472100	7921
1997	5190	108	560520	26936100	11664
1998	5670	76	430920	32148900	5776
1999	5160	65	335400	26625600	4225
2000	6840	93	636120	46785600	8649
Άθροισμα	**56760**	**1000**	**4094160**	**251029800**	**74152**

Πίνακας 19. Υπολογισμός παρατηρήσεων για την εύρεση συντελεστή συσχέτισης πωλήσεων και πωλητών

Ο συντελεστής συσχέτισης μεταξύ διαφήμισης και πωλήσεων είναι 0.95 και μεταξύ πωλητών και διαφήμισης είναι 0.72. Σε προηγούμενη ενότητα αναφέρθηκε πως η συσχέτιση μεταξύ δύο μεταβλητών (x,y) μετράει τον βαθμό αλληλεξάρτησης τους. Αν θεωρηθεί ότι οι μεταβλητές *y* και *x* είναι συσχετισμένες, σημαίνει ότι υπάρχουν ενδείξεις γραμμικής σχέσης μεταξύ των δύο μεταβλητών. Οι μεταβολές των δύο μεταβλητών, κατά μέσο όρο, συνδέονται με το συντελεστή συσχέτισης. Επομένως, **δεν** υπονοείται ότι οι μεταβολές της y οφείλονται από τις μεταβολές της x και αντίστροφα. Από την ανάλυση της συσχέτισης το συμπέρασμα είναι ότι η διαφήμιση επιδρά περισσότερο στις πωλήσεις παρά ο αριθμός των πωλητών. Κάνοντας χρήση τα αποτελέσματα για τις πωλήσεις και τη διαφήμιση, μπορεί να εκτιμηθεί η εξίσωση παλινδρόμησης των πωλήσεων ως προς τη διαφήμιση. Έτσι:

$$\hat{\beta}_1 = \frac{n\sum X_i Y_i - \sum Y_i \sum X_i}{n\sum X_i^2 - (\sum (X_i))^2} = \frac{16 \times 66883410 - 14766 \times 56760}{16 \times 18289944 - 14766^2} = 3,11$$

$$\hat{\beta}_0 = \overline{Y} - \beta_1 \overline{X} = \frac{56760}{16} - 3,11\frac{14766}{16} = 677,4$$

ενώ η εξίσωση παλινδρόμησης είναι η παρακάτω προκύπτουσα:

$$\hat{Y} = 677,4 + 3,11 \times X$$

όπου Y = πωλήσεις και X = διαφήμιση. Προκύπτει ότι αν η εταιρία δεν κάνει καμιά διαφήμιση (X=0), οι μέσες πωλήσεις θα διαμορφωθούν σε 677,4 €. Για κάθε ένα € που θα δαπανώνται σε διαφήμιση, οι πωλήσεις θα αυξάνουν κατά 3,11€. Η παλινδρόμηση μεταξύ πωλήσεων και πωλητών υπολογίζεται αντίστοιχα:

$$\hat{Y} = 615,3,4 + 46,92 \times X$$

Το πιο ουσιαστικό ερώτημα που θα πρέπει να απαντηθεί με όσα προαναφέρθηκαν πριν γίνει χρήση της εξίσωσης παλινδρόμησης είναι: ποια είναι η προβλεπτική ικανότητα της εξίσωσης ή τι ποσοστό των μεταβολών της εξαρτημένης μεταβλητής Y οφείλεται στις επιδράσεις της X. Όσο μεγαλύτερη είναι η επίδραση της X επί της Y τόσο μικρότερα είναι τα υπόλοιπα (e_i) και αντίστροφα.

Αρχικά θα εκτιμηθεί η συνολική διασπορά γύρω από τη γραμμή της παλινδρόμησης, δηλαδή το άθροισμα των τετραγώνων των αποκλίσεων των πραγματικών τιμών της Y από τις αντίστοιχες τιμές της. Ουσιαστικά θα υπολογιστεί το **άθροισμα των τετραγώνων των σφαλμάτων (sum of squared errors) που συμβολίζεται με SSE.**

Για τα 2 παραδείγματα, οι υπολογισμοί εμφανίζονται στους παρακάτω Πίνακες 20 και 21.

$\hat{Y} = 677.4 + 3.11X$

Έτος	Πωλήσεις (Y)	Διαφήμιση (X)	$\hat{Y}$	$SSE = (Y - \hat{Y})^2$
1985	1050	162	1181	17219
1986	1260	285	1564	92264
1987	1470	540	2357	786414
1988	2160	261	1489	450093
1989	1950	360	1797	23409
1990	2400	690	2823	179183
1991	2370	495	2217	23455
1992	3150	948	3626	226271
1993	3670	720	2917	428932
1994	4410	1140	4223	35044
1995	4500	1395	5016	266101
1996	5610	1560	5529	6561
1997	5190	1380	4969	48753
1998	5670	1260	4595	1153476
1999	5160	1710	5996	698060
2000	6840	1560	6462	142884
Άθροισμα				**4576119**

Πίνακας 20. Εκτίμηση του SSE της εξίσωσης παλινδρόμησης των πωλήσεων ως προς τη Διαφήμιση

$\hat{Y} = 615,3.4 + 46,92 \times X$

Έτος	Πωλήσεις (Y)	Πωλητές (X)	$\hat{Y}$	$SSE = (Y - \hat{Y})^2$
1985	1050	32	2117	1137584
1986	1260	47	2820	2434564
1987	1470	23	1694	50327
1988	2160	68	3806	2707787
1989	1950	32	2117	27747
1990	2400	17	1413	974480
1991	2370	58	3336	933890
1992	3150	75	4134	968147
1993	3570	98	5213	2899456
1994	4410	43	2633	3158985
1995	4500	76	4181	101860
1996	5610	89	4791	671151
1997	5190	108	5682	242219
1998	5670	76	4181	2217538
1999	5160	65	3665	2236666
2000	6340	93	4978	3465464
Άθροισμα				**24026845**

Πίνακας 21. Εκτίμηση του SSE της εξίσωσης παλινδρόμησης των πωλήσεων ως προς τους Πωλητές

Από τις προηγούμενες εξισώσεις παλινδρόμησης που βρέθηκαν είναι φανερό πως το υπόδειγμα της παλινδρόμησης στον πληθυσμό ορίζεται από την εξίσωση y=E(y)+e=α+βx+e. Θα πρέπει να διευκρινίσουμε ότι για την πλήρη περιγραφή του υποδείγματος, εκτός από τους συντελεστές παλινδρόμησης α και β πρέπει να γνωρίζουμε και τη διακύμανση του σφάλματος e, δηλαδή τη σ_e^2. Η διακύμανση του σφάλματος είναι η παράμετρος που καθορίζει την ένταση της εξάρτησης της y από την x. Η εκτίμηση της θα βασιστεί στο άθροισμα των τετραγώνων των σφαλμάτων γύρω από τη γραμμή παλινδρόμησης, δηλαδή το SSE. Συμβολίζοντας την εκτίμηση του σ_e^2 με s_e^2, προκύπτει:

$$s_{\hat{e}}^2 = \sum \frac{(y - \hat{y})^2}{n - 2} = \frac{SSE}{n - 2}$$

Όπου n-2 είναι οι βαθμοί ελευθερίας[14]. Χάνονται δύο βαθμοί διότι η εκτίμηση του $s_{\hat{e}}^2$ βασίζεται στην εκτίμηση δύο παραμέτρων: των α και β. Εξετάζοντας τώρα τη συνολική μεταβλητότητα (διασπορά) της εξαρτημένης μεταβλητής y, η διασπορά μιας μεταβλητής ορίζεται από το άθροισμα των τετραγώνων των αποκλίσεων των τιμών από το μέσο τους όρο, δηλαδή $\sum(y - \bar{y})^2$. Αυτό το άθροισμα ονομάζεται συνολικό άθροισμα τετραγώνων (total sum of squares) και συμβολίζεται με TSS. Επομένως: SST = $\sum(y - \bar{y})^2$. Το SSE αντιπροσωπεύει το μέρος της συνολικής μεταβλητότητας της y που δεν εξηγείται από την εξίσωση παλινδρόμησης.

Το υπόλοιπο, δηλαδή SST-SSE, αποτελεί το μέρος της διασποράς της SST που οφείλεται στις επιδράσεις της x. με άλλα λόγια, η συνολική μεταβλητότητα της y χωρίζεται σε δύο μέρη (συνιστώσες): στην «εξηγημένη» από την εξίσωση παλινδρόμησης και στην «ανεξήγητη», δηλαδή εκείνη που οφείλεται στην επίδραση όλων των άλλων παραμέτρων εκτός της x. Το άθροισμα $\sum(\hat{y} - \bar{y})^2$ αποτελεί το μέρος της διασποράς της y που οφείλεται στις επιδράσεις της x. Ως εκ τούτου εξηγείται (περιγράφεται) από την εξίσωση παλινδρόμησης και ονομάζεται *άθροισμα των τετραγώνων των τιμών της παλινδρόμησης* (sum of squared regression) και συμβολίζεται με SSR. Δηλαδή: SSR= $\sum(\hat{y} - \bar{y})^2$. Επομένως ισχύει:
SST=SSR+SSE δηλ. $\sum(y - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum(y - \hat{y})^2$ ενώ γραφικά απεικονίζεται η εξίσωση στην Εικόνα 14[15].

Το ποσοστό της συνολικής μεταβλητότητας του y που εξηγείται από την εξίσωση παλινδρόμησης, δηλαδή οφείλεται στις επιδράσεις της x, ονομάζεται *συντελεστής προσδιορισμού* και συμβολίζεται με R^2. Έτσι, έχουμε:

$$R^2 = \frac{SSR}{TSS} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

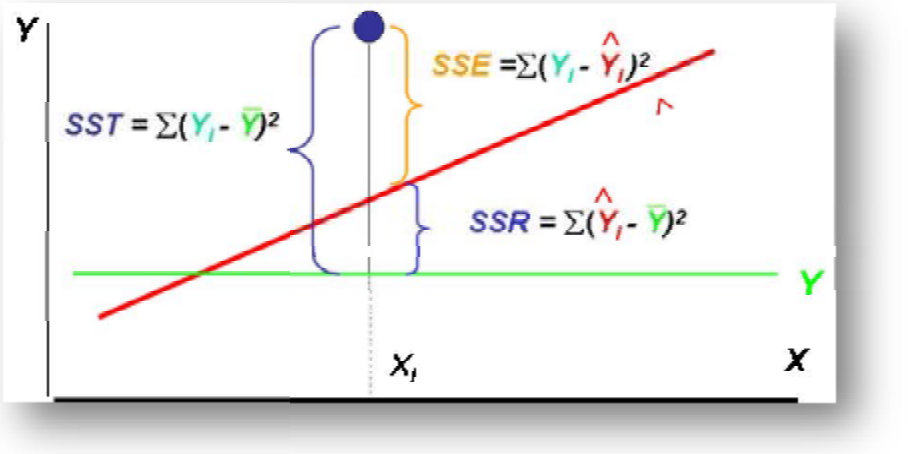
ή

[14] Σε πολλά επιστημονικά πεδία, οι βαθμοί ελευθερίας ενός συστήματος είναι ο αριθμός των παραμέτρων του συστήματος, ο οποίος ενδέχεται να ποικίλει ανεξαρτήτως. Για παράδειγμα, η θέση ενός σχήματος πάνω σε ένα επίπεδο έχει τρεις βαθμούς ελευθερίας, τον προσανατολισμό και τις δύο συντεταγμένες του οποιουδήποτε σταθερού σημείου του σχήματος. Στην Στατιστική οι βαθμοί ελευθερίας είναι ο αριθμός των τιμών κατά τον τελικό υπολογισμό μιας στατιστικής, οι οποίες είναι ελεύθερες να ποικίλλουν. Πηγή: https://el.wikipedia.org/wiki/Βαθμοί_ελευθερίας

[15] Πηγή: http://www.trizsigma.com/regression.html

$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{\Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2} = 1 - \frac{\Sigma y^2 - a\Sigma y - \beta \Sigma yx}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}$$

Από τα παραπάνω προκύπτει ότι ο συντελεστής προσδιορισμού R^2 παίρνει μόνο θετικές τιμές στο διάστημα [0,1].



Εικόνα 14. Διαγραμματική αποτύπωση της σχέσης **SST=SSR+SSE**

Με βάση την παραπάνω σχέση οι εκτιμήσεις των συντελεστών προσδιορισμού των δύο εξισώσεων είναι:

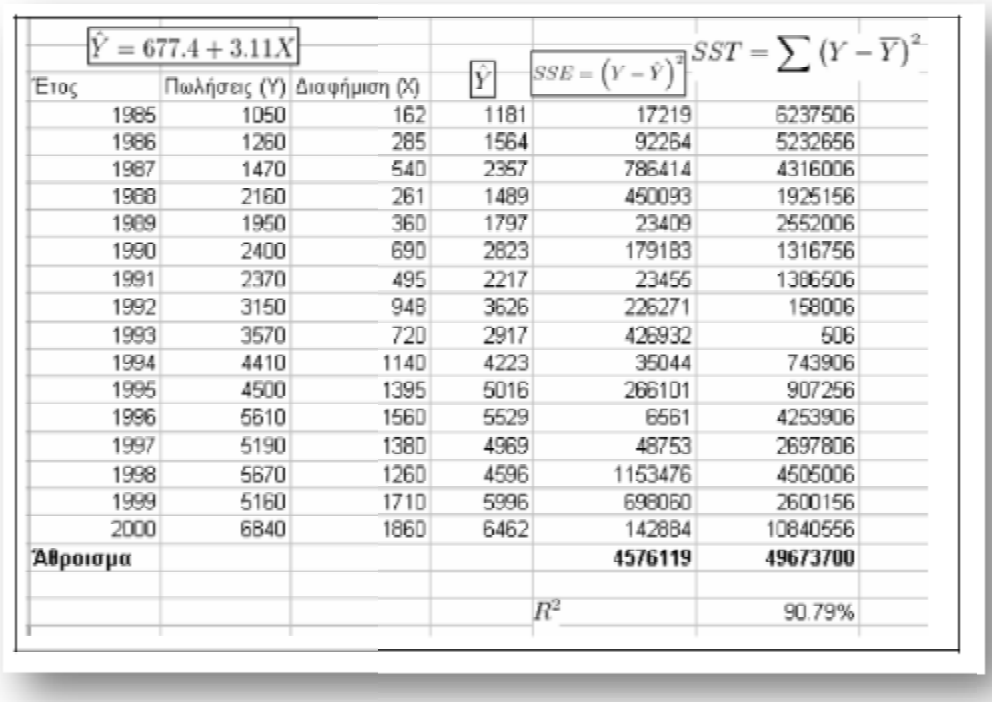
Για την παλινδρόμηση των πωλήσεων ως προς τη διαφήμιση:

$$R^2 = 1 - \frac{\Sigma y^2 - a\Sigma y - \beta \Sigma yx}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} = 1 - \frac{508458}{27892200 - \frac{18920^2}{16}} = 1 - 0{,}092 = 0{,}908$$

Για την παλινδρόμηση των πωλήσεων ως προς τους πωλητές:

$$R^2 = 1 - \frac{\Sigma y^2 - a\Sigma y - \beta \Sigma yx}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} = 1 - \frac{2669649}{27892200 - \frac{18920^2}{16}} = 1 - 0{,}484 = 0{,}516$$

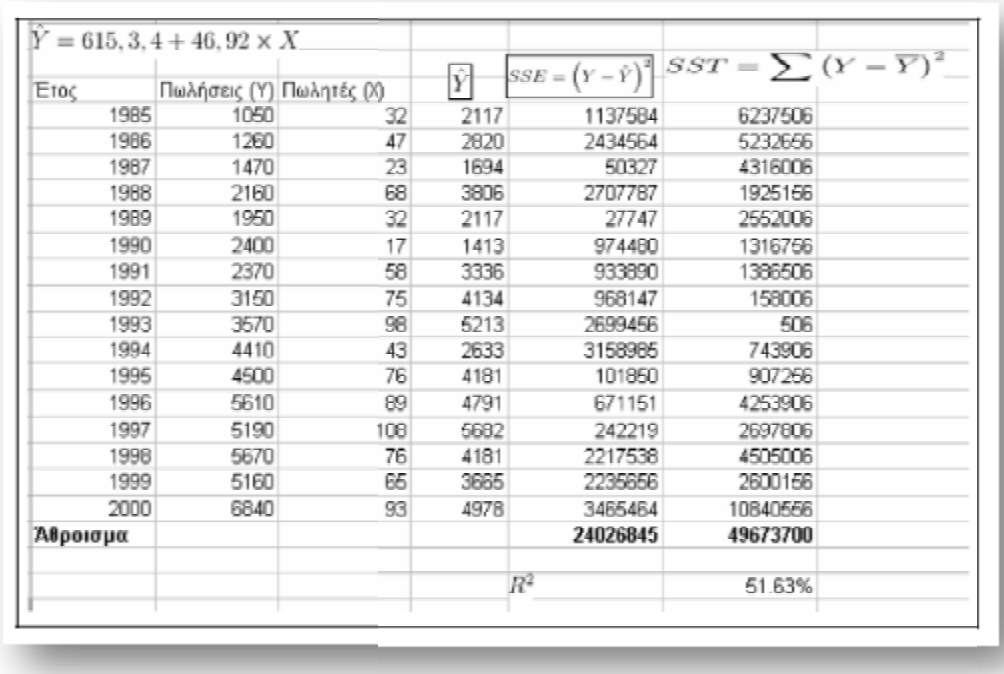
Οι υπολογισμοί εμφανίζονται και στους παρακάτω πίνακες 22 και 23 αντίστοιχα. Από τα παραπάνω συμπεραίνεται ότι εάν χρησιμοποιηθούν ως ανεξάρτητη μεταβλητή οι δαπάνες για διαφήμιση, τότε το 90,8% της μεταβλητότητας των πωλήσεων οφείλεται στην επίδραση της διαφήμισης και μόνο το 9,2% (=100%-90,8%) οφείλεται σε όλους τους άλλους παράγοντες.



$\hat{Y} = 677.4 + 3.11X$

Έτος	Πωλήσεις (Υ)	Διαφήμιση (Χ)	$\hat{Y}$	$SSE = (Y - \hat{Y})^2$	$SST = \sum (Y - \overline{Y})^2$
1985	1050	162	1181	17219	6237506
1986	1260	285	1564	92264	5232656
1987	1470	540	2357	786414	4316006
1988	2160	261	1489	450093	1925156
1989	1950	360	1797	23409	2552006
1990	2400	690	2823	179183	1316756
1991	2370	495	2217	23455	1386506
1992	3150	948	3626	226271	158006
1993	3570	720	2917	426932	506
1994	4410	1140	4223	35044	743906
1995	4500	1395	5016	266101	907256
1996	5610	1560	5529	6561	4253906
1997	5190	1380	4969	48753	2697806
1998	5670	1260	4596	1153476	4505006
1999	5160	1710	5996	698060	2600156
2000	6840	1860	6462	142884	10840556
Άθροισμα				**4576119**	**49673700**
			R^2		90.79%

Πίνακας 22. Υπολογισμός του Συντελεστή Προσδιορισμού διαφήμισης και πωλήσεων



$\hat{Y} = 615,3,4 + 46,92 \times X$

Έτος	Πωλήσεις (Υ)	Πωλητές (Χ)	$\hat{Y}$	$SSE = (Y - \hat{Y})^2$	$SST = \sum (Y - \overline{Y})^2$
1985	1050	32	2117	1137584	6237506
1986	1260	47	2820	2434564	5232656
1987	1470	23	1694	50327	4316006
1988	2160	68	3806	2707787	1925156
1989	1950	32	2117	27747	2552006
1990	2400	17	1413	974480	1316756
1991	2370	58	3336	933890	1386506
1992	3150	75	4134	968147	158006
1993	3570	98	5213	2699456	506
1994	4410	43	2633	3158985	743906
1995	4500	76	4181	101850	907256
1996	5610	89	4791	671151	4253906
1997	5190	108	5682	242219	2697806
1998	5670	76	4181	2217538	4505006
1999	5160	65	3665	2235656	2600156
2000	6840	93	4978	3465464	10840556
Άθροισμα				**24026845**	**49673700**
			R^2		51.63%

Πίνακας 23. Υπολογισμός του Συντελεστή Προσδιορισμού πωλητών και πωλήσεων

Αντίθετα, εάν οι πωλήσεις εκφραστούν ως γραμμική συνάρτηση, ως προς τον αριθμό των πωλητών, η εξίσωση της παλινδρόμησης εξηγεί μόνο το 51,6% της διασποράς

των πωλήσεων. Επομένως η εταιρία μπορεί να επηρεάσει με μεγαλύτερη βεβαιότητα τις πωλήσεις μέσω της διαφήμισης παρά μέσω των πωλητών.

Στην απλή γραμμική παλινδρόμηση λαμβάνεται υπόψη μόνο η επίδραση της ανεξάρτητης μεταβλητής. Όλες οι άλλες επιδράσεις περιλαμβάνονται στην κατάλοιπο συνιστώσα e. Έτσι, οι διάφορες ανεξάρτητες μεταβλητές δίνουν διαφορετικά ποσοστά επίδρασης (R^2) που δεν έχουν καμία σχέση μεταξύ τους και επομένως δεν μπορούν να αθροιστούν. Τέλος, υπάρχει εναλλακτικός τρόπος υπολογισμού του συντελεστή προσδιορισμού. Αποδεικνύεται ότι ο συντελεστής προσδιορισμού R^2 ισούται με το τετράγωνο του συντελεστή συσχέτισης r. Δηλαδή: $R^2=r^2$

Η σχέση αυτή ισχύει μόνο για την απλή γραμμική παλινδρόμηση όπου υπάρχει μία ανεξάρτητη μεταβλητή. Εφαρμόζοντας την παραπάνω σχέση στα παραδείγματα, έχουμε:

Για την παλινδρόμηση των πωλήσεων ως προς τη διαφήμιση:

$$R^2 = r^2 = (0.952)^2 = 0.908$$

Για την παλινδρόμηση των πωλήσεων ως προς τους πωλητές:

$$R^2 = r^2 = (0,718)^2 = 0,516$$

Επομένως συμπεραίνεται ότι το 90.8% της μεταβλητότητας των πωλήσεων οφείλεται στην επίδραση της διαφήμισης και μόνο το 9.2%(1-0.908) οφείλεται σε άλλους παράγοντες. Αντίθετα, εάν οι πωλήσεις εκφραστούν ως γραμμική συνάρτηση ως προς τον αριθμό των πωλητών, η εξίσωση της παλινδρόμησης εξηγεί μόνο το 51.6% της διασποράς των πωλήσεων. Επομένως η εταιρεία μπορεί να επηρεάσει με μεγαλύτερη βεβαιότητα τις πωλήσεις μέσω της διαφήμισης παρά μέσω πωλητών

5.3.3 Η Μέθοδος των Ελαχίστων Τετραγώνων

Όταν η συνάρτηση παλινδρόμησης είναι γραμμική και η συμπεριφορά του τυχαίου όρου ικανοποιεί ορισμένες συνθήκες τότε η άριστη μέθοδος εκτίμησης του μοντέλου είναι η μέθοδος των ελαχίστων τετραγώνων (Least Squares Method). Η μέθοδος αυτή εξακολουθεί να είναι άριστη και σε ορισμένες περιπτώσεις που οι επιθυμητές συνθήκες δεν ικανοποιούνται αρκεί να γίνει ο κατάλληλος μετασχηματισμός των δεδομένων.

Ας υποτεθεί ότι η σχέση που συνδέει δύο ιδιότητες ενός ατόμου που αντιστοιχούν σε ποσοτικές τυχαίες μεταβλητές (τ.μ) είναι γραμμική, της μορφής y=α+βx, όπου οι συντελεστές α και β θα πρέπει να εκτιμηθούν. Δεδομένου n παρατηρήσεων και λαμβάνονται τα ζεύγη (x_i,y_i) όπου i=1,2,…,n. Αν το μοντέλο είναι προσδιοριστικό τότε ισχύει ότι $y_i=\alpha+\beta x_i$ όπου i=1,2,…,n. Επειδή όμως οι παρατηρήσεις υπόκεινται σε σφάλματα (διαφόρων τύπων), η σχέση που συνδέει τα x_i με τα y_i είναι της μορφής $y_i=\alpha+\beta x_i+\varepsilon_i$ όπου i=1,2,…,n και ε_i είναι τυχαία σφάλματα. Είναι φυσικό να αναζητείται όσο το δυνατόν μικρότερα σφάλματα, δηλαδή η ελαχιστοποίηση τους: $\varepsilon_i=y_i-(\alpha+\beta x_i)$ όπου i=1,2,…,n ή ισοδύναμα τη σχέση:

$$\sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n} (y_i - a - \beta x_i)^2$$

Παραγωγίζοντας τη σχέση αυτή ως προς τους προσδιοριστικούς συντελεστές α και β, προκύπτει το παρακάτω σύστημα δύο εξισώσεων με δύο αγνώστους που καλούνται κανονικές εξισώσεις:

$$\sum_{i=0}^{n} y_i = na + \beta \sum_{i=0}^{n} x_i$$

$$\sum_{i=0}^{n} x_i y_i = a \sum_{i=0}^{n} x_i + \beta \sum_{i=0}^{n} x_i^2$$

Η λύση αυτού του συστήματος ως προς α και β δίνει:

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}$$

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

Οι $\hat{a}$ και $\hat{\beta}$ ονομάζονται **εκτιμητές ελαχίστων τετραγώνων** των παραμέτρων α και β όπου:

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=0}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=0}^{n} x_i^2 - n\bar{x}^2\right)$$

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=0}^{n} y_i$$

$$s_{\bar{y}}^2 = \frac{1}{n-1}\left(\sum_{i=0}^{n} y_i^2 - n\bar{y}^2\right)$$

Η εκτίμηση $\hat{y}=\hat{\alpha} + \hat{\beta}x$, της ευθείας παλινδρόμησης $E(y)=\alpha+\beta x$, καλείται *ευθεία ελαχίστων τετραγώνων* (από τον τρόπο υπολογισμού των συντελεστών της). Σε αυτό το σημείο θα πρέπει να γίνει διάκριση μεταξύ της παρατηρούμενης τιμής του y και της τιμής $\hat{y}$ που εκτιμώνται. Η παρατηρούμενη τιμή είναι η πραγματική τιμή του y, ενώ η τιμή $\hat{y}$ είναι αυτή που υπολογίζουμε για το y, όταν δοθεί το x, με τη βοήθεια της ευθείας που εκτιμήθηκαν. Αυτές οι δύο τιμές μπορεί να μη συμπίπτουν και φυσικά τόσο καλύτερο είναι το μοντέλο μας, όσο η διαφορά $y-\hat{y}$ είναι μικρότερη.

Η μέση απόκλιση μεταξύ της πραγματικής και της εκτιμούμενης τιμής της μεταβλητής y, καλείται τυπικό σφάλμα της εκτίμησης (standard error of the estimate) συμβολίζεται με s και ισχύει:

$$s = \sqrt{\frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Ένα το s είναι μικρό, τότε τα y και $\hat{y}$ δεν θα διαφέρουν πολύ και η ευθεία γραμμικής παλινδρόμησης, δίνει μια καλή περιγραφή της σχέσης μεταξύ των x και y. Αν το s είναι μεγάλο, τότε δεν μπορεί να υπάρχει καλή περιγραφή της σχέσης. Με τη βοήθεια των παραπάνω σχέσεων αποδεικνύεται ότι:

$$s^2 = \frac{1}{n-2}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2 = \frac{n-1}{n-2}\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right) = \frac{n-1}{n-2}\left(s_y^2 - \hat{\beta}^2 s_x^2\right) = \frac{n-1}{n-2}\left(s_y^2 - \hat{\beta} s_{xy}\right)$$

υποθέτοντας ότι τα σφάλματα έχουν μέση τιμή $E(e_i)=0$ και διασπορά $Var(e_i)=\sigma^2$ η οποία συνήθως είναι άγνωστη και θα πρέπει να εκτιμηθεί. Η ποσότητα s^2 είναι η εκτίμηση της διασποράς των σφαλμάτων $\sigma^2=Var(e)$.

Παράδειγμα:

Μία εταιρία πετρελαιοειδών για να βελτιώσει την ποιότητα των καυσίμων της έκανε το εξής πείραμα: Χρησιμοποιήθηκαν 10 αυτοκίνητα της ίδιας μάρκας στα οποία

έβαλαν βενζίνη με κάποια βελτιωτική ουσία σε 7 διαφορετικά επίπεδα. Στη συνέχεια, μετρήθηκε η μείωση των εκπεμπόμενων καυσαερίων και οι τιμές δίνονται στον παρακάτω Πίνακα 24:

Επίπεδα βελτιωτικής ουσίας	**Μείωση καυσαερίων**
1	2,1
1	2,5
2	3,1
3	3,0
4	3,8
4	3,2
5	4,3
6	3,9
6	4,4
7	4,8

Πίνακας 24. Παρατηρήσεις καυσαερίων και επιπέδων βελτιωτικής ουσίας

Το ερώτημα που ανακύπτει είναι αν υπάρχει σχέση μεταξύ των επιπέδων της βελτιωτικής ουσίας και της μείωσης των καυσαερίων; Θα πρέπει πρώτα να γίνει το διάγραμμα διασποράς. Η ελεγχόμενη μεταβλητή x είναι τα επίπεδα της βελτιωτικής ουσίας, ενώ η μεταβλητή y είναι η μείωση των καυσαερίων .

Εικόνα 15. Διάγραμμα διασποράς καυσαερίων και βελτιωτικής ουσίας

Παρατηρώντας, τα δεδομένα μπορεί να έχουν μία γραμμική σχέση μεταξύ τους, της μορφής: $y_i = a + \beta x_i + e_i$ όπου i=1,2,….,10. Προκύπτει ότι:

$$\bar{x} = 3.9 \quad s_x^2 = \frac{1}{9}(193 - 10(3.9)^2) = 4.54$$

$$\bar{y} = 3.51 \quad s_y^2 = \frac{1}{9}(130.05 - 10(3.51)^2) = 0.61$$

$$s_{xy} = \frac{1}{9}(152.7 - 10(3.9)(3.51)) = 1.76$$

Οι εκτιμητές ελαχίστων τετραγώνων των παραμέτρων α και β είναι:

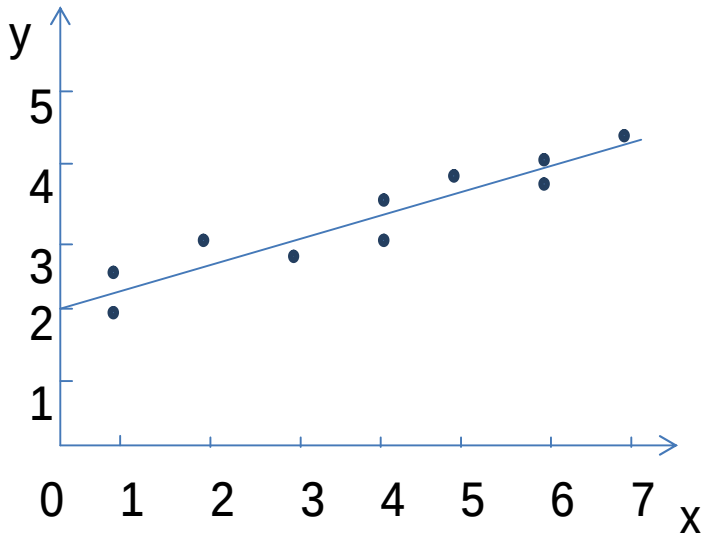
$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = 0.387$$

$$\hat{a} = \bar{y} - \hat{\beta}\bar{x} = 2$$

Η ευθεία που προσαρμόζεται καλύτερα στα δεδομένα είναι η:

$$\hat{y} = 2 + 0.387x$$

και η γραφική της παράσταση αποτυπώνεται στην Εικόνα 16.



Εικόνα 16. Γραφική παράσταση της ευθείας παλινδρόμησης

Η εκτίμηση της διασποράς των σφαλμάτων είναι:

$$s^2 = \frac{9}{8}((0.61)^2 - (0.387)^2(4.54)^2) = 0.0925$$

Ενώ το τυπικό σφάλμα της εκτίμησης είναι $s = \sqrt{0.0925} \approx 0.304$ το οποίο είναι μικρό και μπορεί να ειπωθεί ότι η προσαρμογή των δεδομένων στην ευθεία είναι «καλή» δηλαδή, για δεδομένη τιμή της βελτιωτικής ουσίας μπορεί να προβλεφθεί η μείωση των καυσαερίων, χωρίς μεγάλο σφάλμα.

5.3.4 Διαστήματα Εμπιστοσύνης για τις Παραμέτρους β_0 και β_1

Η κλίση μιας ευθείας γραμμής μας δείχνει πόσο αλλάζει η εξαρτημένη μεταβλητή αν αλλάξει η ανεξάρτητη κατά μία μονάδα. Η πληθυσμιακή γραμμή παλινδρόμησης με κλίση β_1 δίνει τον τρόπο αντιστοιχίας των πληθυσμιακών τιμών της Y με διάφορες τιμές της X. δηλαδή, αν β_1=0,18 τότε αν η ανεξάρτητη μεταβλητή αυξηθεί κατά μία μονάδα τότε η εξαρτημένη θα αυξηθεί κατά 0,18. Αυτή η τιμή της κλίσης είναι μια σημειακή εκτίμηση που δεν μας δίνει κάποια πληροφόρηση για την ακρίβεια της εκτίμησης αυτής. Ο σημειακός εκτιμητής της πληθυσμιακής κλίσης β_1 δίνεται από την τιμή της εκτίμησης $\hat{\beta}$ ή σε άλλους τύπος b_1. Μπορούμε με τη χρήση των αποτελεσμάτων της παλινδρόμησης να κατασκευάσουμε ένα διάστημα εμπιστοσύνης για την πληθυσμιακή κλίση σύμφωνα με τον τύπο:

$$\Pr\left(b_1 - t_{n-2,\frac{\alpha}{2}} s_{b1} \le \beta_1 \le b_1 + t_{n-2,\frac{\alpha}{2}} s_{b1}\right) = (100 - a)\%$$

Ομοίως για τον πληθυσμιακό σταθερό όρο:

$$\Pr\left(b_0 - t_{n-2,\frac{\alpha}{2}} s_{b0} \le \beta_0 \le b_0 + t_{n-2,\frac{\alpha}{2}} s_{b0}\right) = (100 - a)\%$$

Για ένα επίπεδο στατιστικής σημαντικότητας α βρίσκουμε την κριτική τιμή της $t_{\alpha/2}$ με n-2 βαθμούς ελευθερίας. Για παράδειγμα για α=5%, το διάστημα εμπιστοσύνης για τις πληθυσμιακές παραμέτρους β_1 και β_2 θα είναι:

$$\Pr\left(b_1 - t_{n-2,0.025} s_{b1} \le \beta_1 \le b_1 + t_{n-2,0.025} s_{b1}\right) = (100 - 5)\% = 95\%$$

$$\Pr\left(b_0 - t_{n-2,0.025} s_{b0} \le \beta_0 \le b_0 + t_{n-2,0.025} s_{b0}\right) = (100 - 5)\% = 95\%$$

Παράδειγμα:

Τα παρακάτω δεδομένα (Πίν. 25) αναφέρονται στον αριθμό των αυτοκινητικών ατυχημάτων και τον αριθμό των οχημάτων σε κυκλοφορία (σε χιλιάδες οχήματα).

ΕΤΗ	ΑΡΙΘΜΟΣ ΑΤΥΧΗΜΑΤΩΝ	ΑΡΙΘΜΟΣ ΑΥΤΟΚΙΝΗΤΩΝ
1980	112	332
1981	113	353
1982	137	391
1983	161	421
1984	176	442
1985	168	470
1986	187	509
1987	188	557
1988	228	621
1989	228	672
1990	234	723

Πίνακας 25. Δοθέντες παρατηρήσεις αριθμού ατυχημάτων και αυτοκινήτων

Από τα αποτελέσματα αυτά εύκολα υπολογίζονται τα διαστήματα εμπιστοσύνης για τον σταθερό όρο και την κλίση. Αυτό απαιτεί την εύρεση των παρακάτω στοιχείων:

1. Της κριτικής τιμής από τους πίνακες της t για n-2 βαθμούς ελευθερίας και α/2 επίπεδο στατιστικής σημαντικότητας. Η κριτική τιμή για $t_{0.025,9}$=2,262.
2. Τις τιμές των συντελεστών. Από τον πίνακα των αποτελεσμάτων $\hat{\beta}$ = b_1=0.32236 και $\hat{\alpha}$ = b_0=14.72.
3. Τις τιμές των τυπικών σφαλμάτων των συντελεστών. Από τον πίνακα των αποτελεσμάτων s_{b1}=0.02916 και s_{b0}=15.

Κατόπιν αντικαθιστώντας τα παραπάνω στοιχεία στους ανωτέρω αναφερόμενους προκύπτουν:

$$\Pr\left(b_1 - t_{n-2,0.025}s_{b1} \le \beta_1 \le b_1 + t_{n-2,0.025}s_{b1}\right) = (100-5)\% = 95\% \Leftrightarrow$$

$$\Pr\left(0{,}32236 - 2{,}262(0{,}02916) \le \beta_1 \le 0{,}32236 + 2{,}262(0{,}02916)\right) = 95\% \Leftrightarrow$$

$$\Pr(0.254 \le \beta_1 \le 0{,}3883) = 95\%$$

Δηλαδή, με πιθανότητα 95% ο πληθυσμιακός σταθερός όρος θα βρίσκεται περίπου μεταξύ -19 και 49. Πρέπει να σχολιαστεί ότι το διάστημα αυτό είναι πολύ μεγάλο και αυτό οφείλεται στο μεγάλο τυπικό σφάλμα. Αυτό συνδυάζεται και με το γεγονός ότι ο σταθερός όρος δεν είναι στατιστικά σημαντικός.

5.3.5 Ιδιότητες Εκτιμητών Ελάχιστων Τετραγώνων

Οι εκτιμητές β_o και β_1 των ελαχίστων τετραγώνων έχουν τις εξής ιδιότητες:

1. Το άθροισμα των καταλοίπων e_i γύρω από τη γραμμή παλινδρόμησης ισούται με το μηδέν, δηλαδή:

$$\sum e_i = \sum (y_i - \hat{y}_i) = 0$$

διότι:

$$\sum e = \sum (y - \beta_o - \beta_1 x) = \sum y - n\beta_o - \beta_1 \sum x = 0$$

σύμφωνα με το ότι:

$$\sum_{i=0}^{n} x_i y_i = a \sum_{i=0}^{n} x_i + \beta \sum_{i=0}^{n} x_i^2$$

2. Η γραμμή παλινδρόμησης περνάει από το σημείο $(\bar{x},\bar{y})$ που αντιστοιχεί στους μέσους των μεταβλητών x και y. Δηλαδή, ισχύει η σχέση: $\bar{y}=\alpha+\beta\bar{x}$ που προκύπτει από την παρακάτω εξίσωση αν διαιρέσουμε και τα δύο μέλη με n:

$$\sum_{i=0}^{n} y_i = na + \beta \sum_{i=0}^{n} x_i$$

3. Οι συντελεστές των ελαχίστων τετραγώνων α και β είναι αμερόληπτες εκτιμήσεις των συντελεστών παλινδρόμησης του πληθυσμού α και β αντίστοιχα. Επομένως:

$$E(b_o) = \beta_o \text{ και } E(b_1) = \beta_1$$

4. Επίσης, οι συντελεστές b_0 και b_1 είναι αποτελεσματικές εκτιμήσεις των β_0 και β_1, δηλαδή έχουν το μικρότερο τυπικό σφάλμα εκτίμησης. Τα τυπικά σφάλματα εκτίμησης των b_0 και b_1 είναι συναρτήσεις του αθροίσματος των τετραγώνων των καταλοίπων (Σe^2), που είναι ελάχιστο όπως προκύπτει από τη μέθοδο των ελαχίστων τετραγώνων.

Θεωρώντας το γραμμικό μοντέλο: $y_i = \alpha + \beta x_i + e_i$, όπου i=1,2,…,n και e_i είναι τυχαία σφάλματα (ανεξάρτητα μεταξύ τους) με μέση τιμή $E(e_i)=0$ και διασπορά $var(e_i)=\sigma^2$. Οι συντελεστές των ελαχίστων τετραγώνων α και β είναι αμερόληπτες εκτιμήσεις των συντελεστών παλινδρόμησης του πληθυσμού α και β αντίστοιχα. Επομένως:

$$E(\hat{\alpha}) = \alpha \text{ και } E(\hat{\beta}) = \beta$$

και αντίστοιχα ότι ισχύει:

$$var(\hat{\alpha}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)$$

$$var(\hat{\beta}) = \frac{\sigma^2}{(n-1)s_x^2}$$

Με την υπόθεση επιπλέον ότι τα σφάλματα ακολουθούν κανονική κατανομή $N(0,\sigma^2)$ αποδεικνύεται ότι:

1. Το $\hat{\alpha}$ ακολουθεί κανονική κατανομή $N(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right))$
2. Το $\hat{\beta}$ ακολουθεί κανονική κατανομή $N(\beta, \frac{\sigma^2}{(n-1)s_x^2})$

$$Οι\ τ.μ. T_\alpha = \frac{\hat{\alpha} - \alpha}{s_a}, \qquad T_\beta = \frac{\hat{\beta} - \beta}{s_\beta},$$

$$s_{\hat{\alpha}}^2 = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right), \qquad s_{\hat{\beta}}^2 = \frac{s^2}{(n-1)s_x^2}$$

Ακολουθούν την t_{n-2} κατανομή Student. Έτσι μπορούν να κατασκευαστούν τα διαστήματα εμπιστοσύνης (δ.ε.) για τις παραμέτρους α και β του γραμμικού μοντέλου E(y)=α+βx.

1. 100(1-a)% δ.ε. για το α: ($\hat{\alpha}$ ±s_α*$t_{n-2,a/2}$)
2. 100(1-a)% δ.ε. για το β: ($\hat{\beta}$ ±s_β*$t_{n-2,a/2}$)

<u>Έλεγχοι υποθέσεων για την παράμετρο α:</u>

Η αρχική υπόθεση H_0: α=α_0 απορρίπτεται σε σ.σ. a όταν:

1. $\frac{\hat{\alpha}-\alpha_0}{s_{\hat{\alpha}}} > t_{n-2,\alpha}$ με εναλλακτική υπόθεση H_1: α>α_0.
2. $\frac{\hat{\alpha}-\alpha_0}{s_{\hat{\alpha}}} < -t_{n-2,\alpha}$ με εναλλακτική υπόθεση H_1: α<α_0.
3. $\frac{|\hat{\alpha}-\alpha_0|}{s_{\hat{\alpha}}} > t_{n-2,\alpha/2}$ με εναλλακτική υπόθεση H_1: α≠α_0.

<u>Έλεγχοι υποθέσεων για την παράμετρο β:</u>

Η αρχική υπόθεση H_0: β=β_0 απορρίπτεται σε σ.σ. a όταν:

1. $\frac{\hat{\beta}-\beta_0}{s_{\hat{\beta}}} > t_{n-2,\alpha}$ με εναλλακτική υπόθεση H_1: β>β_0.
2. $\frac{\hat{\beta}-\beta_0}{s_{\hat{\beta}}} < -t_{n-2,\alpha}$ με εναλλακτική υπόθεση H_1: β<β_0.
3. $\frac{|\hat{\beta}-\beta_0|}{s_{\hat{\beta}}} > t_{n-2,\alpha/2}$ με εναλλακτική υπόθεση H_1: β≠β_0.

Συνήθως ενδιαφέρει να ελέγχεται η υπόθεση H_0: β=0 διότι αν ισχύει αυτή η υπόθεση, τότε το μοντέλο γίνεται E(y)=α που σημαίνει ότι η τιμή της μεταβλητής y είναι ανεξάρτητη από την τιμή της μεταβλητής x.

<u>Υποθέσεις που αφορούν την E(Y):</u>

Οι μέθοδοι που αναπτύχθηκαν μπορούν να επεκταθούν και στην περίπτωση ελέγχου υποθέσεων που αφορά στην αναμενόμενη τιμή του y, για μια δεδομένη τιμή του x. Δεδομένου ότι η εκτίμηση του E(y) δίνεται από την σχέση: $\hat{y}=\hat{\alpha}+\hat{\beta}x$

Μπορεί να αποδειχθεί ότι:

$$E(\hat{y}) = a + \beta x, \qquad var(\hat{y}) = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)$$

Με την υπόθεση ότι τα σφάλματα e_i ακολουθούν κανονική κατανομή $N(0,\sigma^2)$, αποδεικνύεται ότι το $\hat{y}$, ακολουθεί κανονική κατανομή $N\left(a + \beta x, \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)\right)$. Το 100(1-a)% δ.ε. για την αναμενόμενη τιμή του E(Y) στο σημείο x, δίνεται από τον τύπο:

$$\left(\hat{y} \pm s * t_{n-2;\frac{a}{2}}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}\right)$$

Όπου το $s^2 = \frac{n-1}{n-2}\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right)$ είναι μία εκτίμηση του σ^2.

5.3.6 Τα Σφάλματα Εκτίμησης ή Κατάλοιπα

Έχοντας δημιουργήσει την γραφική παράσταση των μεταβλητών, η Μέθοδος Ελαχίστων Τετραγώνων (Ordinary Least Squares, OLS), χρησιμοποιείται για να επιλεχθεί η κατάλληλη γραμμή (από τις άπειρες) που περνάει από τα σημεία της γραφικής παράστασης των δύο αυτών μεταβλητών. Οι σχέσεις μεταξύ των μεταβλητών δεν είναι πάντα ακριβείς. Μη παρατηρήσιμες ή τυχαίες διακυμάνσεις στα παρατηρηθέντα στοιχεία αναγκάζουν την αυστηρή μαθηματική σχέση μεταξύ των μεταβλητών να μην επαληθεύεται πάντα στην πράξη. Για να συμπεριληφθούν και οι συγκεκριμένες διακυμάνσεις, ένα στοχαστικό-τυχαίο τμήμα προστίθεται στο μοντέλο παλινδρόμησης. Αν γίνει χρήση της X για την επεξήγηση της συμπεριφοράς της Y, οποιαδήποτε ευθεία γραμμή μπορεί να αποδοθεί με τη μορφή:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Το $\beta_0+\beta_1 X$ είναι το συστηματικό τμήμα της εξίσωσης, ενώ το ε είναι το τυχαίο τμήμα, το οποίο ονομάζεται διαταρακτικός όρος (disturbance term) ή σφάλμα (error). Τα σφάλματα παρουσιάζονται στα πειράματα επειδή γίνονται λάθη κατά τη διαδικασία της μέτρησης της εξαρτημένης μεταβλητής ή επειδή το μοντέλο είναι ελλιπώς προσδιορισμένο. Η πρώτη αιτία είναι εύκολα κατανοητή. Η δεύτερη αιτία μπορεί να εξηγηθεί μέσω ενός παραδείγματος. Όπως είναι γνωστό, η κατανάλωση ενός αγαθού εξαρτάται όχι μόνο από την τιμή του συγκεκριμένου αγαθού αλλά και από πολλούς άλλους παράγοντες, όπως τις τιμές των υποκατάστατων και των συμπληρωματικών

αγαθών, το εισόδημα, τα επιτόκια, το εισόδημα παρελθόντων χρόνων, το προσδοκώμενο μελλοντικό εισόδημα, την ηλικία του πληθυσμού, τη διαφήμιση, κλπ. Έτσι, αν προσδιοριστεί ένα υπόδειγμα κατανάλωσης ως γραμμική συνάρτηση της τιμής ή του εισοδήματος, αυτό δεν είναι επαρκές. Δύναται να υπάρχει η αντιμετώπιση ενός προβλήματος ελλιπούς προσδιορισμού. Όλοι οι παράγοντες, οι οποίοι επηρεάζουν την κατανάλωση και δεν συμπεριλήφθηκαν μέσα από το υπόδειγμα, θα αντιπροσωπεύονται από τον διαταρακτικό όρο ε_i.

Για να εκτιμηθεί ένα υπόδειγμα, πρέπει πρώτα να συλλεχθεί ένα δείγμα στοιχείων για την εξαρτημένη και την ανεξάρτητη μεταβλητή που ενδιαφέρει. Αν $Y_1,Y_2,\ldots,Y_n$ και $X_1,X_2,\ldots,X_n$ αντιπροσωπεύουν ένα τυχαίο δείγμα n ανεξάρτητων παρατηρήσεων ενός πληθυσμού Y_i και X_i αντιπροσωπεύουν τις i^{th} τυχαίες παρατηρήσεις του δείγματος, τότε με δεδομένα τα n ζεύγη παρατηρήσεων Y_i και X_i, ο στόχος της ανάλυσης παλινδρόμησης είναι να αποκτηθούν εκτιμήσεις για τις άγνωστες πληθυσμιακές παραμέτρους β_0 και β_1. Πρακτικά όμως οι επιδράσεις στο τυχαίο τμήμα της παραπάνω εξίσωσης δεν μπορούν να προβλεφθούν. Είναι απαραίτητο να προσδιοριστεί μια κατανομή για τον διαταρακτικό όρο και να γίνει η υπόθεση για τα εξής:

1. Σε οποιαδήποτε τιμή της X, ο διαταρακτικός όρος είναι μια τυχαία μεταβλητή, η οποία κατανέμεται με μέσο 0 και διακύμανση σ^2. Δηλαδή: $E(\varepsilon_i)=0$ και $Var(\varepsilon_i)=\sigma^2$ για κάθε i. Δηλαδή ε_i είναι μια τυχαία μεταβλητή που παίρνει τιμές θετικές και αρνητικές έτσι ώστε η κατά μέσο όρο της να είναι μηδέν.
2. Οι δειγματικές τιμές του ε_i κατανέμονται ανεξάρτητα, δηλαδή τα σφάλματα δεν συσχετίζονται μεταξύ τους. Αυτό σημαίνει ότι για δύο διαφορετικές παρατηρήσεις του διαταρακτικού όρου ε_i και ε_j με $i \neq j$ η αναμενόμενη τιμή $E(\varepsilon_i, \varepsilon_j)=0$ και η συνδιακύμανση τους (Cov) θα είναι μηδέν: $Cov(\varepsilon_i, \varepsilon_j)=E(\varepsilon_i-E\varepsilon_j)(\varepsilon_j-E\varepsilon_i)=E\varepsilon_i, \varepsilon_j=0$ καθώς $(E\varepsilon_i)=0$ και $(E\varepsilon_j)=0$. Όπως αναφέρθηκε στην 1^{η} υπόθεση, κάθε δειγματικό σφάλμα ε_i κατανέμεται με την ίδια διακύμανση σ^2. Η διακύμανση της τυχαίας μεταβλητής είναι σταθερή για όλες τις τιμές της ανεξάρτητης μεταβλητής. Δηλαδή, η διασπορά των τιμών της ανεξάρτητης μεταβλητής. Σε αυτή την περίπτωση τονίζεται ότι ο όρος συμπεριφέρεται ομοσκεδαστικά.
3. Κάθε δειγματικό σφάλμα κατανέμεται κανονικά για κάθε i.
4. Οι τιμές της ανεξάρτητης μεταβλητής X λαμβάνονται ως σταθερές και για μία συγκεκριμένη τιμή της X αντιστοιχεί μια ολόκληρη κατανομή της Y. Έτσι

κάθε διαφοροποίηση της Y οφείλεται στους παράγοντες που συμπεριλαμβάνονται στον διαταρακτικό όρο.

Οι υποθέσεις αυτές απαιτούν σωστό προσδιορισμό του υποδείγματος αναφορικά με τη συναρτησιακή μορφή και τις μεταβλητές που έχουν συμπεριληφθεί. Ο στόχος της ανάλυσης παλινδρόμησης είναι να εκτιμήσει τις παραμέτρους του υποδείγματος, ώστε η ανερμήνευτη μεταβολή της Y οριζόμενη ως το κατάλοιπο (residual), να είναι μικρή και μη συστηματική. Η παραβίαση των συγκεκριμένων υποθέσεων οδηγεί σε προβλήματα αυτοσυσχέτισης, ετεροσκεδαστικότητας κλπ. Τα συγκεκριμένα προβλήματα αποτελούν σημαντικά θέματα και χρήζουν ιδιαίτερης ανάλυσης αν αποσκοπούν σε σωστή μοντελοποίηση κάποιων οικονομικών και κοινωνικών φαινομένων με σκοπό τις προβλέψεις.

Οι παραπάνω υποθέσεις υποδηλώνουν ότι τα σφάλματα είναι ανεξάρτητες μεταβλητές που κατανέμονται κανονικά ως $N(0,\sigma^2)$. Το άθροισμα των τετραγώνων των σφαλμάτων (Sum of Squared Errors) όπως προαναφέρθηκε έχει καθιερωθεί να συμβολίζεται με SSE Μια ισοδύναμη έκφραση του SSE που διευκολύνει σημαντικά τους υπολογισμούς είναι η:

$$SSE = \sum \varepsilon_i^2 = \sum Y_i^2 - \hat{b}_0 \sum Y_i - \hat{b}_1 \sum X_i Y_i$$

Βασικές ιδιότητες των σφαλμάτων είναι οι εξής:

1. Αποδεικνύεται ότι το SSE είναι το μικρότερο δυνατό (ελάχιστο).
2. Το άθροισμα των σφαλμάτων ισούται με μηδέν. Πράγματι:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i = \bar{Y} - \hat{b}_1 \bar{X} + b_1 X_i = \bar{Y} + \hat{b}_1 (X_i - \bar{X})$$

Γνωρίζοντας ότι οι διαφορές των εκτιμήσεων από τις παρατηρήσεις συμβολίζονται με $\hat{\varepsilon}_i$, προκύπτει ότι:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \bar{Y} + \hat{b}_1 (X_i - \bar{X}), \ i=1,2,\ldots,n$$

Αθροίζοντας για i=1,2…,n προκύπτει[16]:

$$\sum \hat{\varepsilon}_i = \sum (Y_i - \bar{Y}) - \hat{b}_1 \sum (X_i - \bar{X}) = 0$$

3. $\sum X_i \hat{\varepsilon}_i = 0$

Πράγματι, ισχύει:

[16] ***Σημείωση:*** η ιδιότητα αυτή δεν ορίζει την ευθεία ελαχίστων τετραγώνων αφού ικανοποιεί από κάθε ευθεία που περνάει από το σημείο $(\bar{X},\bar{Y})$. από τις ευθείες όμως αυτές η ευθεία ελαχίστων τετραγώνων είναι η μόνη που ικανοποιεί και την ιδιότητα:

$$\sum X_i \hat{\varepsilon}_i = \sum X_i \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right) = \sum X_i Y_i - \hat{b}_0 \sum X_i - \hat{b}_1 \sum X_i^2 = 0$$

4. $\sum \hat{Y}_i \hat{\varepsilon}_i = 0$

Πράγματι ισχύει:

$$\hat{Y}_i \hat{\varepsilon}_i = \left(\hat{b}_0 + \hat{b}_1 X_i\right)\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right) = \hat{b}_0 \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right) + \hat{b}_1 X_i \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)$$

Αθροίζοντας για i=1,2,….,n προκύπτει:

$$\sum \hat{Y}_i \hat{\varepsilon}_i = \hat{b}_0 \left(\sum Y_i - n\hat{b}_0 - \hat{b}_1 \sum X_i\right) + \hat{b}_1 \left(\sum X_i Y_i - \hat{b}_0 \sum X_i - \hat{b}_1^2 \sum X_i^2\right) = 0$$

αφού από τις δύο κανονικές εξισώσεις οι παραστάσεις στις δύο παρενθέσεις μηδενίζονται.

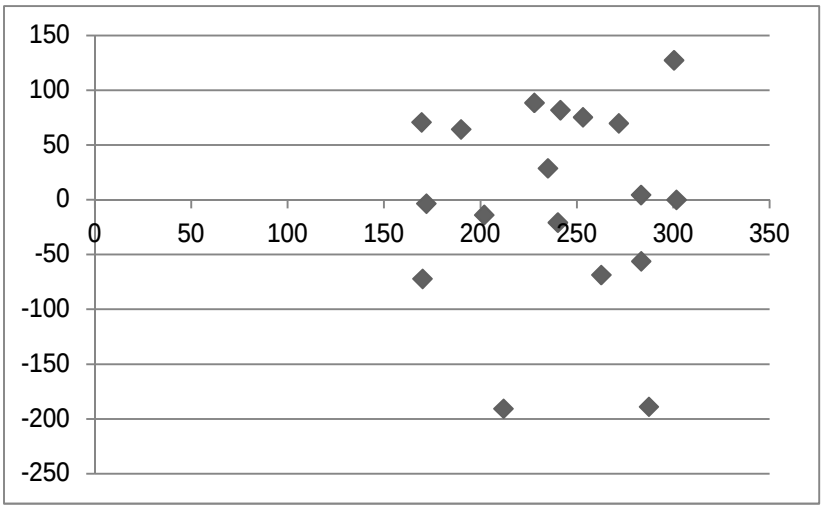
Παράδειγμα

Σε 18 διαδοχικές εβδομάδες παρατηρήθηκε η αξία των πωλήσεων Y σε επιλεγμένα επίπεδα διαφημιστικής δαπάνης X (Πίνακας 26), από όπου προέκυψαν τα ακόλουθα 18 ζεύγη:

Y_i	X_i	$\hat{Y}_i$	$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$
1100	170.0	1171.911	-71.91
1180	172.0	1183.269	-3.27
1240	169.5	1169.071	70.93
1350	190.0	1285.497	64.50
1220	212.0	1410.441	-190.44
1590	228.0	1501.310	88.69
1340	202.0	1353.648	-13.65
1570	235.0	1541.065	28.94
1660	241.5	1577.980	82.02
1550	240.2	1570.597	-20.60
1820	283.3	1815.374	4.63
2040	300.4	1912.490	127.51
1760	283.4	1815.942	-55.94
1630	262.7	1698.381	-68.38
1920	301.7	1919.873	0.13
1720	253.2	1644.428	75.57
1650	287.4	1838.660	-188.66
1820	271.8	1750.063	69.94

Πίνακας 26. Η αξία των πωλήσεων Y σε επιλεγμένα επίπεδα διαφημιστικής δαπάνης X

Η γραφική παράσταση των καταλοίπων ως προς τις τιμές X_i δίνεται στην παρακάτω Εικόνα 17, από όπου φαίνεται ότι τα κατάλοιπα κατανέμονται τυχαία γύρω από το μηδέν χωρίς καμιά σημαντικότητα.



Εικόνα 17. Γραφική παράσταση καταλοίπων ως προς τις τιμές X_i

5.4 Εφαρμογή την ανάλυσης της παλινδρόμησης στον επενδυτικό κίνδυνο

Η Γραμμική Παλινδρόμηση χρησιμοποιείται για τη δημιουργία γραμμών τάσεως, χρησιμοποιώντας δεδομένα του παρελθόντος για να προβλέψει τις μελλοντικές αποδόσεις ή «τάσεις». Σε ένα επιχειρησιακό περιβάλλον η ανάλυση της παλινδρόμησης, χρησιμοποιείται συνήθως για να υποδειχθεί η κίνηση των οικονομικών ή τα χαρακτηριστικά ενός προϊόντος στη πάροδο του χρόνου. Σε ένα επενδυτικό πλαίσιο η ανάλυση της παλινδρόμησης χρησιμοποιείται ώστε να αναλυθούν οι τιμές των μετοχών, οι τιμές του πετρελαίου, ή οι προδιαγραφές ενός προϊόντος και έτσι να ορισθεί ένα σχετικός βαθμός επικινδυνότητας μιας επένδυσης. Έτσι η γραμμική παλινδρόμηση και η σωστή χρήση της θεωρείτε το κλειδί για την εκτίμηση του κινδύνου που συνδέεται με τις περισσότερες επενδυτικές κινήσεις (Keener, 2011).

Για αυτό το λόγο με τη βοήθεια της ανάλυσης της γραμμικής παλινδρόμησης αναπτύχθηκε το μοντέλο Capital Asset Pricing το οποίο υπολογίζει ένα κοινό μέτρο της μεταβλητότητας μιας μετοχής ή επένδυσης το οποίο ονομάζετε beta (και καθορίζεται με τη βοήθεια της γραμμικής παλινδρόμησης). Στο χρηματοοικονομικό τομέα, το beta μιας μετοχής ή του χαρτοφυλακίου είναι ένας αριθμός που περιγράφει τη σχέση της απόδοσης της με συγκεκριμένης μετοχής με αυτές της χρηματοπιστωτικής αγοράς στο σύνολό της (Levinson, 2006).

Ένα θετικό beta σημαίνει ότι οι αποδόσεις των περιουσιακών στοιχείων (ή μετοχών) ακολουθούν σε γενικές γραμμές τις αποδόσεις της σχετικής αγοράς, με την έννοια ότι είτε και οι δύο τείνουν να είναι ανώτερες των αντίστοιχων μέσων όρων τους μαζί, ή και τα δύο τείνουν να είναι κάτω των αντίστοιχων μέσων όρων τους από κοινού. Με την ίδια λογική ένα αρνητικό beta σημαίνει ότι οι αποδόσεις των περιουσιακών στοιχείων έχουν εν γένει αντίθετη κίνηση από τις αποδόσεις της αγοράς: όταν η τιμή της μίας είναι κάτω του μέσου όρου της, η άλλη θα τείνει να είναι άνω του μέσου όρου της (Myron & Joseph, 1977).

Σύμφωνα με τον McAlpine (2010) ο συντελεστής beta είναι μια βασική παράμετρος της στατιστικής διακύμανσης ενός περιουσιακού στοιχείου στο μοντέλο τιμολόγησης κεφαλαίου (CAPM) το οποίο δεν μπορεί να αφαιρεθεί από τη διαφοροποίηση που παρέχεται από τα χαρτοφυλάκια πολλών υψηλού κινδύνου περιουσιακών στοιχείων, λόγω της συσχέτισης των αποδόσεων του με τις αποδόσεις των άλλων περιουσιακών στοιχείων που βρίσκονται στο χαρτοφυλάκιο. Ο συντελεστής beta μπορεί να εκτιμηθεί για μεμονωμένες εταιρείες χρησιμοποιώντας την ανάλυση παλινδρόμησης έναντι ενός χρηματιστηριακού δείκτη. Ο τύπος για την beta ενός περιουσιακού στοιχείου σε ένα χαρτοφυλάκιο είναι :

$$\beta_a = \frac{\mathrm{Cov}(r_a, r_p)}{\mathrm{Var}(r_p)},$$

όπου το r_a μετρά το ποσοστό απόδοσης του περιουσιακού στοιχείου, r_p μέτρα το ποσοστό απόδοσης του χαρτοφυλακίου, και Cov (r_a,r_p) είναι η συνδιακύμανση μεταξύ των ποσοστών της επιστροφής. Το χαρτοφυλάκιο του ενδιαφέροντος (portfolio of interest) στις διατυπώσεις του CAPM είναι το χαρτοφυλάκιο της αγοράς που περιλαμβάνει όλα τα επικίνδυνα στοιχεία ενεργητικών μετοχών, και έτσι οι όροι r_p του άνωθεν τύπου αντικαθίσταται από r_m (δηλαδή το ποσοστό απόδοσης της αγοράς).

Εξ ορισμού, ο συντελεστής beta της επενδυτικής αγοράς είναι 1.0 και οι μεμονωμένες μετοχές κατατάσσονται ανάλογα με το πόσο αποκλίνουν από τη μακροοικονομική της αγοράς (για λόγους απλούστευσης, η S&P 500 μερικές φορές χρησιμοποιείται ως υποκατάστατο για την αγορά στο σύνολό της). Μία μετοχή της οποίας η απόδοση είναι υψηλότερη από το μέσο όρο των αποδόσεων της αγοράς στην πάροδο του χρόνου μπορεί να έχει μια beta του οποίου η απόλυτη τιμή είναι μεγαλύτερη από 1.0 (αν είναι στην πραγματικότητα μεγαλύτερη από 1.0 θα εξαρτηθεί από το συσχετισμό των αποδόσεων της μετοχής με τις αποδόσεις της αγοράς).

Ομοίως, μία μετοχή της οποίας η απόδοση είναι χαμηλότερη από το μέσο όρο των αποδόσεων της αγοράς έχει έναν beta με απόλυτη τιμή μικρότερη από 1.0 (Σαββίδης 1994). Μια μετοχή με beta 2 έχει επιστροφή (απόδοση) που αλλάζει κατά μέσο όρο δύο φορές συχνότερα απο ότι οι αποδόσεις του συνόλου της αγοράς, έτσι για παράδειγμα όταν η απόδοση της αγοράς πέφτει ή ανεβαίνει κατά 3%, η απόδοση της μετοχής θα πέσει ή θα αυξηθεί (αντίστοιχα) κατά 6% σε μέσο όρο. Ωστόσο, επειδή ο συντελεστής beta εξαρτάται επίσης από τη συσχέτιση των αποδόσεων, μπορεί να υπάρχει σημαντική διακύμανση στο μέσο όρο: όσο υψηλότερος είναι ο συσχετισμός, τόσο μικρότερη η διακύμανση, και όσο χαμηλότερος είναι ο συσχετισμός, τόσο μεγαλύτερη είναι η διακύμανση). Επιπλέον, ο beta μπορεί επίσης να είναι και αρνητικός, που σημαίνει ότι η αποδόσεις των μετοχών τείνουν να κινούνται προς την αντίθετη κατεύθυνση των αποδόσεων της αγοράς. Με αυτό το τρόπο, μια μετοχή με beta -3 θα αντιμετωπίζει μείωση των αποδόσεων της 9% (κατά μέσο όρο) όταν η απόδοση της αγοράς αυξάνεται κατά 3%, και αντίστροφα, θα έχει άνοδο στις αποδόσεις της 9% (κατά μέσο όρο) εάν η αποδόσεις της αγοράς μειωθούν κατά 3% (Tofallis, 2008).

Είναι ευρέως αποδεχτό πως, αν και οι μετοχές υψηλότερων beta παρέχουν τη δυνατότητα υψηλότερων αποδόσεων, τείνουν να είναι πιο ασταθής και ως εκ τούτου πιο ριψοκίνδυνες. Συνεπώς, οι μετοχές με χαμηλότερους συντελεστές beta είναι λιγότερο επιβλαβή αλλά γενικά προσφέρουν χαμηλότερες αποδόσεις. Παρόλα αυτά η ιδέα αυτή έχει αμφισβητηθεί απο τον McAlpin (2010) ο οποίος υποστηρίζει ότι τα στοιχεία δείχνουν μικρή σχέση μεταξύ του beta και της πιθανής απόδοσης, και ότι τις περισσότερες φορές οι μετοχές χαμηλότερων beta είναι λιγότερο επικίνδυνες και περισσότερο κερδοφόρες. Τέλος, σύμφωνα με τον Klarman (1991), με τον ίδιο τρόπο που ο συντελεστής beta μιας μετοχής δείχνει τη σχέση της με αλλαγές της αγοράς, είναι επίσης και ένας δείκτης για την απαιτούμενη απόδοση των επενδύσεων (ROI). Λαμβάνοντας υπόψη ένα επιτόκιο μηδενικού κινδύνου ύψους 2%, για παράδειγμα, αν η αγορά (με beta 1) έχει αναμενόμενη απόδοση 8%, μία μετοχή με συντελεστή beta της τάξης του 1,5 θα πρέπει να έχει απόδοση 11% (= 2% + 1,5 (8 % - 2%)).

5.5 Σύνοψη

Σχεδόν όλα τα μοντέλα παλινδρόμησης του πραγματικού κόσμου περιλαμβάνουν πολλαπλούς προγνωστικούς παράγοντες, και οι βασικές περιγραφές της γραμμικής παλινδρόμησης συχνά διατυπώνονται όσον αφορά το μοντέλο πολλαπλής

παλινδρόμησης. Ένας μεγάλος αριθμός διαδικασιών έχουν αναπτυχθεί για την παράμετρο εκτίμηση και το συμπέρασμα σε γραμμική παλινδρόμηση. Στη στατιστική και την αριθμητική ανάλυση, το πρόβλημα των αριθμητικών μεθόδων για γραμμικές μεθόδους ελαχίστων τετραγώνων είναι σημαντικό επειδή τα γραμμικά μοντέλα παλινδρόμησης είναι ένας από τους πιο σημαντικούς τύπους μοντέλου, τόσο ως επίσημο στατιστικό μοντέλο όσο και για την εξερεύνηση των συνολικών δεδομένων.

Η πλειοψηφία των στατιστικών πακέτων για υπολογιστή περιέχουν μοντέλα για την ανάλυση παλινδρόμησης που κάνουν χρήση της γραμμικής μεθόδου υπολογισμού των ελαχίστων τετραγώνων. Ως εκ τούτου, η σημαντική προσπάθεια που έχει αφιερωθεί στο έργο της διασφάλισης, είναι ότι οι εν λόγω υπολογισμοί πραγματοποιούνται αποτελεσματικά και λαμβάνουν δεόντως υπόψη την αριθμητική ακρίβεια. Μεμονωμένες στατιστικές αναλύσεις σπάνια αναλαμβάνονται μεμονωμένα, αλλά μάλλον αποτελούν μέρος μιας ακολουθίας ερευνητικών βημάτων. Η διαρρύθμιση των γραμμικών μοντέλων των ελαχίστων τετραγώνων συχνά, αλλά όχι πάντα, τίθεται στο πλαίσιο της στατιστικής ανάλυσης.

ΚΕΦΑΛΑΙΟ ΕΚΤΟ

ΣΥΝΟΨΗ - ΣΥΜΠΕΡΑΣΜΑΤΑ

Είναι κοινά αποδεκτό πως τα μοντέλα παλινδρόμησης χρησιμοποιούνται ευρέως σήμερα στη διοίκηση των επιχειρήσεων, στην οικονομία, στη μηχανική, στην υγεία, τη βιολογία και τις κοινωνικές επιστήμες. Στη στατιστική, η ανάλυση παλινδρόμησης είναι μία 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 περισσοτέρων ανεξαρτήτων μεταβλητών. Στα βασικά μοντέλα παλινδρόμησης υποθέτουμε ότι οι ερμηνευτικές μεταβλητές X_1..... , X_p είναι καθορισμένες σταθερές, και το κύριο ενδιαφέρον βρίσκεται στα συμπεράσματα για την εξαρτημένη μεταβλητή Y βάσει των ερμηνευτικών μεταβλητών.

Επίσης για την περίπτωση μίας μόνο ερμηνευτικής μεταβλητής, η ανάλυση παλινδρόμησης για ένα κανονικό μοντέλο παλινδρόμησης, ισχύει ακόμα και όταν το Χ είναι μια τυχαία μεταβλητή, με την προϋπόθεση ότι η μεταβλητή Υ ακολουθεί ορισμένες προϋποθέσεις και ότι η περιθώρια κατανομή της Χ δεν περιλαμβάνει τις παραμέτρους του μοντέλου παλινδρόμησης. Με την ανάλυση παλινδρόμησης (regression analysis) εξετάζεται η 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 διακρίνονται δύο είδη μεταβλητών: οι ανεξάρτητες ή ελεγχόμενες ή επεξηγηματικές ***(independent, predictor, casual, input, explanatory variables)*** και τις εξαρτημένες ή απόκρισης ***(dependent, response variables)***. Σε πειραματικές έρευνες, ανεξάρτητη μεταβλητή Χ είναι εκείνη την οποία μπορεί να ελεγχθεί, δηλαδή, να καθοριστούν οι τις τιμές της ***(π.χ. το ύψος της διαφημιστικής δαπάνης ενός προϊόντος, ο αριθμός των λειτουργούντων ταμείων σε ένα υποκατάστημα τραπέζης, η ποσότητα λιπάσματος, η θερμοκρασία επεξεργασίας ενός προϊόντος)***. Εξαρτημένη μεταβλητή Υ 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 ***(π.χ. 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, ο αριθμός των πωλήσεων μουσικών CD σε μια περιοχή και ο αριθμός των νέων στην ίδια περιοχή).

Κατά τη διερεύνηση της σχέσης μεταξύ δύο μεταβλητών Χ και Υ 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, δύναται να αντιμετωπιστούν 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 Υ δεν είναι ίσες για τα διάφορα επίπεδα της Χ.
2. Για την κανονικοποίηση, όταν οι κατανομές της εξαρτημένης μεταβλητής Υ για τα διάφορα επίπεδα της Χ 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.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική Βιβλιογραφία

[1] Ανδρουλάκης, Γ., (2008), *Στοιχειώδεις Έννοιες της Στατιστικής*, Διαθέσιμο στην ηλεκτρονική διεύθυνση: http://androulakis.bma.upatras.gr/mediawiki/index.php

[2] Βερονίκης, Σ., (2010), *Πίσω στα Βασικά, Μέρος 3ο: Βασικές Αρχές Στατιστικής για Κοινωνιολογικές Έρευνες: Συσχέτιση Μεταβλητών*, Τμήμα Αρχειονομίας – Βιβλιοθηκονομίας, Ιόνιο Πανεπιστήμιο. Διαθέσιμο στην ηλεκτρονική διεύθυνση: http://dlib.ionio.gr/~spver/seminars/statistics/

[3] Βερονίκης, Σ., (2010), *Πίσω στα Βασικά: Βασικές Αρχές Στατιστικής για Κοινωνιολογικές Έρευνες*. Διαθέσιμο στην ηλεκτρονική διεύθυνση: http://dlib.ionio.gr/~spver/seminars/statistics/

[4] Γναρδέλλης, Χ., (2003), *Εφαρμοσμένη Στατιστική*, Εκδόσεις Παπαζήση, Αθήνα

[5] Δαμιανού, Χ., & Κούτρας, Μ., (1998), *Εισαγωγική στη Στατιστική, Μέρος ΙΙ*, Εκδόσεις: Συμμετρία, Αθήνα

[6] Δημητριάδης, Ε., (2002), *Περιγραφική Στατιστική*, Εκδόσεις: Κριτική ΑΕ

[7] Εθνική Αθλητική Ακαδημία Σόφιας, (2006), *Στατιστικές μέθοδοι επεξεργασίας και ανάλυσης δεδομένων: Συσχέτιση – Correlation*, Διαθέσιμο στην ηλεκτρονική διεύθυνση: 7nsa-virtualeducation.com/images/corri.pdf

[8] Εμβαλωτής, Α., Κατσής, Α., Σιδερίδης, Γ., (2006), *Στατιστική Μεθοδολογία Εκπαιδευτικής Έρευνας. Α' Έκδοση, Ιωάννινα*. Διαθέσιμο στην ηλεκτρονική διεύθυνση: ftp://ftp.soc.uoc.gr/Psycho/Sideridis/.../Stat_Notes%20ISBN.pdf

[9] Ευαγγέλου, Α., (2008), *Μέτρα Συσχέτισης και Μέτρα Συμφωνίας στη Βιοστατιστική*, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

[10] Ζαχαροπούλου, Χ., (1998), *Στατιστική Μέθοδοι – Εφαρμογές – Παλινδρόμηση & Συσχέτιση. Τόμος Β*, Εκδόσεις Σοφία. Θεσσαλονίκη

[11] Κατσαμάγκου, Μ., (2003), *Η σχέση των αναγνωστικών ενδιαφερόντων των μαθητών και των δυσκολιών που αντιμετωπίζουν στο σχολείο*, Διαθέσιμο στην ηλεκτρονική διεύθυνση: http://www.pee.gr/wpcontent/uploads/_praktika_synedrion_files/e21_11_03/sin_ath_mer_c/them_enot_vi/katsamagoy.htm

[12] Κιόχος, Π., (1993), *Περιγραφική Στατιστική*, Εκδόσεις INTERBOOKS Αθήνα

[13] Κιόχος, Π., (1998), *Επαγωγική Στατιστική*, Εκδόσεις INTERBOOKS Αθήνα

[14] Κολυβά-Μαχαίρα, Φ. και Μπόρα-Σέντα, Ε., (1998), *Στατιστική: Θεωρία – Εφαρμογές*, Εκδόσεις Ζήτη, Θεσσαλονίκη

[15] Λαζαρίδης, Α., και Lazaridou, N., (2008), *Στατιστική Πλήρης ανάπτυξη της θεωρίας: Λυμένα παραδείγματα και προβλήματα*, Εκδόσεις Δίαυλος

[16] Λουκάς, Σ., (2003), *Στατιστική*, Εκδόσεις: Κριτική

[17] Μαυρωτάς, Γ., Τσιαφογιάννη, Στ., Διακουλάκη, Δ., Σαρίμβεης, Χαρ. Πολυκριτηριακή (2011), *Ανάλυση Παλινδρόμησης & Επιλογή Μεταβλητών*, Διαθέσιμο στην ηλεκτρονική διεύθυνση: www.openarchives.gr

[18] Μπουντζιούκα, Β., Παναγιωτάκος, Δ. (2209), *Στατιστικές Μέθοδοι Για Τον Έλεγχο Της Επαναληψιμότητας Ερωτηματολογίων Για Την Αποτίμηση Της Διατροφικής Πρόσληψης. Ελληνικό Στατιστικό Ινστιτούτο, Πρακτικά 22ου Πανελληνίου Συνεδρίου Στατιστικής (2009), σελ 139-148.* Ανάκτηση απο: http://www.esi-stat.gr/drastiriotites/TOMOS%20PRAKTIKON%20CHANION/pdf/139-148.pdf

[19] Οικονομικό Πανεπιστήμιο Αθηνών. *Σημειώσεις Τμήματος Στατιστικής: Ο Συντελεστής_Συσχέτισης τ Του Kendall.* Ανάκτηση από: www.statathens.aueb.gr/gr/prop/notes/np342.pdf

[20] Παπαδόπουλος Γ. Εργαστήριο Μαθηματικών & Στατιστικής, *Συσχέτιση Δύο Μεταβλητών*, Διαθέσιμο στην ηλεκτρονική διεύθυνση: www.aua.gr/gpapadopoulos/files/sisxetisi091.pdf

[21] Ρούσσος, Π., Λ., Τσαούσης, Γ., (2006), *Στατιστική Εφαρμοσμένη στις Κοινωνικές Επιστήμες*, Αθήνα: Ελληνικά Γράμματα

[22] Σιδερίδης, Γ., *Διάλεξη πάνω στη Συσχέτιση*, Πανεπιστήμιο Κρήτης. Διαθέσιμο στην ηλεκτρονική διεύθυνση: ftp://filer.soc.uoc.gr/Psycho/...II/.../lecture%202%20correlation.ppt

[23] Τσοπάνογλου, Α., (2008), *Σύστημα Αξιολόγησης και Πιστοποίησης Γλωσσομάθειας Στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης. Υπουργείο Εθνικής Παιδείας και Θρησκευμάτων, 2008*, Διαθέσιμο στην ηλεκτρονική διεύθυνση: http://repository.edulll.gr/edulll/retrieve/4496/1289_02_02_ΠΕ3_ΔημιουργίαΤράπεζαςΔοκιμασιών_Παραδοτέο6.pdf

[24] Χαλικιάς, Ι., (2003), *Μέθοδοι Ανάλυσης για Επιχειρηματικές Αποφάσεις*, Εκδόσεις ROSILI

[25] Χάλκος, Ε., (2000), *Θεωρία Εφαρμογές και Χρήση Στατιστικών Προγραμμάτων σε Η/Υ*, Αθήνα

Ξενόγλωσση Βιβλιογραφία

[1] Fox, J. (2000), *Nonparametric Simple Regression,* Thousand Oaks, CA: Sage Publications

[2] Fan & Jacoby (1995), *Regression Analysis, University of Texas.* Koltko-Rivera, M. How to Calculate Point Biserial Correlation, Διαθέσιμο στην ηλεκτρονική διεύθυνση: http://www.ehow.com/how_7303303_calculate-point-biserial-correlation.html

[3] Measured Progress Assessment Firm's official site, *Discovering the Point Biserial* Διαθέσιμο στην ηλεκτρονική διεύθυνση: http://www.measuredprogress.org/learning-tools-statistical-analysis-the-point-biserial

[4] Lund, A.M. (2001), *Measuring Usability with the USE Questionnaire. Usability Interface, 8(2), STC Usability and User Experience Community.* Διαθέσιμο στην ηλεκτρονική διεύθυνση: http://www.stcsig.org/usability/newsletter/ 0110_measuring_with_use

[5] Kutner,M., Nachtsheim, C., Neter, J., Li, W., (2004), *Applied Linear Statistical Models* McGraw-Hill/Irwin; 4 edition

[6] Higgins, J., (2009), *A re-evaluation of random-effects meta-analysis,* Journal of the Royal Statistical Society Series A 2009; 172: 137-159

[7] Carver, R., (1978), *The Case Against Statistical Significance Testing,* Harvard Educational Review, Vol 48, No 3, August 1978, 378-399

[8] Savvides, C., (1994), *Risk Analysis in Investment Appraisal. MPRA Paper 10035,* University Library of Munich, Germany

[9] Schildcrout, J. S., & Heagerty, P. J., (2005), *Regression analysis of longitudinal binary data*

[10] Thompson, E.L.; Shumann, E. L., (1987), *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy,* Law and Human Behavior (Springer) II (3): 167

[11] Levinson, Mark, (2006), *Guide to Financial Markets.* London: The Economist (Profile Books). pp. 145–6

[12] Lowry, R., (2010), *Significance between two correlation coefficients.* Retrieved from http://daculty.vascar.edu/lowry/rdiff.html

[13] Keener, (2011), *Statistical Inference.* Διαθέσιμο στην ηλεκτρονική διεύθυνση: http://www.stat.lsa.umich.edu/%18keener/

[14] Klarman, Seth; Williams, Joseph, (1991), *Beta,* Journal of Financial Economics 5 (3): 117

[15] McAlpine, Chad, (2010), *Low-risk TSX stocks have out earned riskiest peers over 30-year period,* The Financial Post Trading Desk, June 22

[16] Tofallis, C., (2008), *Forecasting using percentage least squares regression,* Golden Anniversary conference of the Operational Research Society