ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ ΜΕΣΟΛΟΓΓΙΟΥ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ

Π Τ Υ Χ Ι Α Κ Η Ε Ρ Γ Α Σ Ι Α

ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΟΙ ΕΦΑΡΜΟΓΕΣ ΤΗΣ ΣΤΙΣ ΕΠΙΧΕΙΡΗΣΗΣ ΚΑΙ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ

ΠΑΝΑΓΟΥΛΙΑΣ ΠΟΛ. ΕΥΑΓΓΕΛΟΣ

ΕΙΣΗΓΗΤΗΣ
ΜΕΓΑΡΙΤΗΣ ΑΘΑΝΑΣΙΟΣ

Μ Ε Σ Ο Λ Ο Γ Γ Ι 2 0 1 3

ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΜΕΣΟΛΟΓΓΙΟΥ

ΣΧΟΛΗ ΔΙΟΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ

Π Τ Υ Χ Ι Α Κ Η Ε Ρ Γ Α Σ Ι Α

ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΟΙ ΕΦΑΡΜΟΓΕΣ ΤΗΣ ΣΤΙΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ

ΠΑΝΑΓΟΥΛΙΑΣ ΠΟΛ. ΕΥΑΓΓΕΛΟΣ (Α.Μ 15079)

ΕΙΣΗΓΗΤΗΣ
ΜΕΓΑΡΙΤΗΣ ΑΘΑΝΑΣΙΟΣ

Μ Ε Σ Ο Λ Ο Γ Γ Ι 2 0 1 3

ΠΡΟΛΟΓΟΣ

Η πτυχιακή μου εργασία καλύπτει το μάθημα της «Στατιστικής των επιχειρήσεων» και αφορά κυρίως τη γραμμική παλινδρόμηση και τις εφαρμογές της στις επιχειρήσεις και στην οικονομία.

Βασικό χαρακτηριστικό της εργασίας μου αυτής, είναι να παρουσιάσω έναν αριθμό μεθόδων που αναπτύχθηκαν για τα σύγχρονα προβλήματα της παλινδρόμησης. Αυτή είναι η δύναμη, και χωρίς αμφιβολία, η ιδέα να ασχοληθώ με την παλινδρόμηση και κυρίως με τη γραμμική.

Η γραμμική παλινδρόμηση δεν αποτελεί μια απλή θεωρία. Οι διάφορες υποθεωρίες, όπου περιλαμβάνονται οι πρακτικές εφαρμογές της παλινδρόμησης γίνεται με την χρήση του κατάλληλου στατιστικού λογισμικού ως αναγκαία προϋπόθεση για την εφαρμογή των μεθόδων. Προς το σκοπό αυτόν, στα αντίστοιχα κεφάλαια η επίλυση των παραδειγμάτων γίνεται με τη χρήση με τη χρήση του SPSS, συνοδευόμενη και από το σχολιασμό των παραγόμενων αποτελεσμάτων, έτσι ώστε να γίνεται καλύτερη η κατανόηση της παλινδρόμησης. Γι' αυτό συμπεριλαμβάνονται στην πτυχιακή διάφοροι κλάδοι της παλινδρόμησης, ώστε να καλυφθεί όλο και περισσότερη η ύλη τους.

Η δομή της πτυχιακής χωρίζεται σε επιμέρους πέντε κεφάλαια. Το πρώτο κεφάλαιο αναφέρεται στη στατιστική δηλ. από πού ξεκίνησε η ιστορία της να προέρχεται και ποιες είναι οι διάφορες βασικές έννοιες της που θα αναφερθούν παρακάτω.

Στο δεύτερο κεφάλαιο αναφέρεται η παλινδρόμηση και οι διάφορες κατηγορίες της που διακρίνεται η παλινδρόμηση.

Το τρίτο κεφάλαιο περιλαμβάνει την πρώτη διάκριση της παλινδρόμησης στην οποία θα εμπεριέχονται η βασική έννοια, πως παρουσιάζονται το διάγραμμα διασποράς καθώς επίσης και οι κατηγορίες συντελεστών που υπάρχουν.

Στο τέταρτο κεφάλαιο αναπτύσεται η δεύτερη διάκριση της παλινδρόμησης η οποία θα περιλαμβάνει ομοίως τη βασική έννοια, καθώς και πόσα είδη συντελεστών υπάρχουν.

Στο πέμπτο κεφάλαιο και το πιο βασικό συστατικό της εργασίας αυτής θα είναι οι **εφαρμογές της γραμμικής παλινδρόμησης στο χώρο των επιχειρήσεων και την οικονομία**, στην οποία θα εμπεριέχονται οι εφαρμογές της με παραδείγματα από ασκήσεις.

Στόχος της εργασίας μου, είναι να έχω δώσει όσο το δυνατόν καλύτερες βασικές γνώσεις στατιστικής και πλήρες από άποψη υλικού στο κοινό όπως αντιστοιχεί σ' ένα εισαγωγικό μάθημα στατιστικής.

Τέλος, θα ήθελα να ευχαριστήσω τον καθηγητή μου που μου έδωσε την δυνατότητα και με τις συμβουλές του να ασχοληθώ με την εργασία για ένα τόσο σημαντικό μάθημα όπως είναι η στατιστική των επιχειρήσεων στην οποία απέκτησα μέσα από αυτήν χρήσιμα πράγματα που θα με βοηθήσουν και στο μέλλον.

Σεπτέμβριος 2013

ΠΑΝΑΓΟΥΛΙΑΣ ΕΥΑΓΓΕΛΟΣ

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ (ΣΥΝΕΧΕΙΑ)

ΕΙΣΑΓΩΓΗ

Στα κεφάλαια που θα ακολουθήσουν στην πτυχιακή μου εργασία, θα αναφερθώ κυρίως στη γραμμική παλινδρόμηση όπου είναι ένα από τα πιο σημαντικά θέματα της Στατιστική των επιχειρήσεων. Η γραμμική παλινδρόμηση εξετάζει τη γραμμική σχέση μεταξύ δύο ή περισσότερων ποσοτικών μεταβλητών με απώτερο σκοπό την πρόβλεψη της τιμής μιας μεταβλητής με βάση την τιμή άλλων μεταβλητών. Η ανάλυση της γραμμικής παλινδρόμησης γίνεται με την κατασκευή μιας μαθηματικής εξίσωσης, που περιγράφει τη σχέση μεταξύ της μεταβλητής που πρόκειται να προβλεφθεί, η οποία ονομάζεται εξαρτημένη μεταβλητή (dependent variable) και συμβολίζεται ως Y, και μιας ή περισσότερων άλλων μεταβλητών, οι οποίες ονομάζονται ανεξάρτητες μεταβλητές (independent variables) και συμβολίζονται ως $X_1, X_2, ..., X_\kappa$ (όπου κ το πλήθος των ανεξάρτητων μεταβλητών).

Τέλος, επειδή η ανάλυση της γραμμικής παλινδρόμησης περιλαμβάνει αρκετές νέες έννοιες και τεχνικές, η παρουσίαση της έχει χωριστεί στα επιμέρους κεφάλαια που θα αναλύονται παρακάτω.

ΚΕΦΑΛΑΙΟ 1 ΣΤΑΤΙΣΤΙΚΗ

1.1 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

Ο όρος **στατιστική**, κατά μία άποψη, προέρχεται από τη λατινική λέξη « status»[1] η οποία, χρησιμοποιήθηκε αρχικά για το χαρακτηρισμό αριθμητικών δεδομένων που αναφέρονται κυρίως στον πληθυσμό μιας χώρας. Μπορεί όμως να προέρχεται και από την αρχαία ελληνική λέξη στατίζω[2]. Κάνοντας μία σύντομη ιστορική αναδρομή αξίζει να αναφέρουμε ότι οι πρώτες προσπάθειες συλλογής στατιστικών στοιχείων εμφανίστηκαν στους αρχαιότατους χρόνους. Από την ιστορία πληροφορούμαστε ότι σε απογραφές πληθυσμού, γης, κ.τ.λ. προέβαιναν και λαοί της αρχαιότητας (Έλληνες, Κινέζοι, Αιγύπτιοι, κ.τ.λ). Η Στατιστική ως αυτοτελής επιστήμη εφαρμόζεται από τον 17^{ο} αιώνα. Τότε, άρχισε να διαμορφώνεται ένας νέος κλάδος, που προήλθε από τη μελέτη των τυχερών παιχνιδιών, η θεωρία των πιθανοτήτων, η οποία προάγεται και συμπληρώνεται κυρίως από τους *Bernoulli, Gauss, Laplace*. Η στατιστική έχει πλέον εισαχθεί (18^{ος} αιώνας) στις ακαδημαϊκές σπουδές, και η συστηματική οργάνωση των κρατικών στατιστικών υπηρεσιών αλλά και επιστημονικών εταιρειών είναι πλέον γεγονός (19^{ος} αιώνας).

Αξίζει να αναφερθεί ότι ο πρώτος Κυβερνήτης της Ελλάδας, *Ι. Καποδίστριας*, είχε ενδιαφερθεί σοβαρότατα για τη δημιουργία στατιστικής υπηρεσίας στην Ελλάδα διαβλέποντας τον κρίσιμο ρόλο που θα έπαιζε στη δημιουργία του νέου κράτους.

Έτσι αρκετά νωρίς για την ελληνική και διεθνή πραγματικότητα, το 1833, ιδρύθηκε η Υπηρεσία Γενικής Στατιστικής του Κράτους, η οποία τέθηκε στην αρμοδιότητα του Υπουργείου των Εσωτερικών. Με το Κανονιστικό Διάταγμα (Κ.Δ.) της 29/4 του 1834, πάλι στο Υπουργείο Εσωτερικών, ιδρύθηκε το «Γραφείο Δημοσίας Οικονομίας», το οποίο συμπεριέλαβε και την Υπηρεσία Γενικής Στατιστικής του Κράτους.

Περί τα τέλη του 19^{ου} αιώνα η Στατιστική έχει το κατάλληλο επιστημονικό υπόβαθρο για την ανάπτυξη στατιστικών μεθόδων.

Ενώ παλαιότερα η Στατιστική ασχολείται μόνο με την παράθεση τεράστιων πινάκων με δεδομένα και αναρίθμητων διαγραμμάτων, σήμερα μπορούμε να διακρίνουμε σε μία στατιστική έρευνα τρία στάδια: Τη συλλογή του στατιστικού υλικού, την επεξεργασία και παρουσίαση του και τέλος την ανάλυση αυτού του υλικού και την εξαγωγή χρήσιμων

[1] status: κράτος, κατάσταση
[2] στατίζω: τοποθετώ, συμπεραίνω

συμπερασμάτων. Τα τρία αυτά στάδια επιτυγχάνονται με την εφαρμογή κατάλληλων για κάθε περίπτωση στατιστικών μεθόδων, όπως και με την βοήθεια των Υπολογιστών, οι οποίοι σημείωσαν τεράστια ανάπτυξη στις μέρες μας.

Συμπερασματικά λοιπόν μπορούμε δώσουμε ως ορισμό της Στατιστικής το συνηθέστερο και πλέον γνωστό ορισμό του R.A. Fisher (1890-1962), πατέρα της σύγχρονης Στατιστικής:

Στατιστική είναι ένα σύνολο αρχών και μεθοδολογιών για:

- Το σχεδιασμό της διαδικασίας συλλογής δεδομένων.
- Τη συνοπτική και αποτελεσματική παρουσίασή τους.
- Την ανάλυση και εξαγωγή αντίστοιχων συμπερασμάτων.

1.2 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΣΤΑΤΙΣΤΙΚΗΣ

Μια βασική έννοια που εξετάζουμε πρώτα στη στατιστική των επιχειρήσεων είναι ο πληθυσμός-μεταβλητές.

Για την εξαγωγή χρήσιμων συμπερασμάτων είναι απαραίτητο σε κάθε στατιστική έρευνα να εξετάζονται τα στοιχεία ενός συνόλου ως προς ένα ή περισσότερα χαρακτηριστικά του. Η αρχή αυτή ακολουθείται όταν, για παράδειγμα, ενδιαφερόμαστε να βγάλουμε συμπεράσματα για:

- Τον αριθμό των επισκεπτών των μουσείων κατά φύλο και ηλικία.
- Τα ποσοστά της ανεργίας κατά ομάδα ηλικιών, φύλο και επίπεδο εκπαίδευσης.
- Την τουριστική κίνηση στα ξενοδοχειακά καταλύματα.
- Τις συνέπειες του καπνίσματος στην υγεία των καπνιστών, κ.τ.λ.
- Τον αριθμό υπαλλήλων μιας επιχείρησης.

Σε κάθε ένα από τα παραπάνω παραδείγματα έχουμε ένα σύνολο και θέλουμε να εξετάσουμε τα στοιχεία του ως προς ένα ή περισσότερα χαρακτηριστικά τους. Ένα τέτοιο σύνολο ονομάζεται **πληθυσμός** (population). Τα στοιχεία του πληθυσμού συχνά αναφέρονται και ως μονάδες ή άτομα του πληθυσμού. Για να πάρουμε μια απόφαση για ένα πληθυσμό (για παράδειγμα., αν είναι αποδεκτό το ποσοστό των ελαττωματικών λαμπτήρων που φτιάχνει ένα εργοστάσιο) συνήθως εξετάζουμε ένα μόνο μέρος (δηλ. μερικούς μόνο λαμπτήρες) από το σύνολο αυτό. Επιλέγουμε δηλαδή μια μικρή ομάδα ή ένα υποσύνολο του πληθυσμού, το οποίο καλείται **δείγμα**, απ' όπου αντλούνται πληροφορίες για τα χαρακτηριστικά του συγκεκριμένου πληθυσμού, δηλαδή δείγμα είναι κάθε τμήμα του πληθυσμού.

Τα χαρακτηριστικά τα οποία εξετάσουμε ένα πληθυσμό λέγονται **μεταβλητές** (variables) και τις συμβολίζουμε συνήθως με κεφαλαία γράμματα X, Y, Z, A, B,… . Οι δυνατές τιμές που μπορεί να πάρει μια μεταβλητή λέγονται **τιμές της μεταβλητής.** Από την διαδοχική εξέταση των ατόμων του πληθυσμού ως προς ένα χαρακτηριστικό τους προκύπτει μια σειρά από δεδομένα, που λέγονται στατιστικά δεδομένα ή παρατηρήσεις.

Οι μεταβλητές χωρίζονται σε δύο βασικές κατηγορίες:

- Σε **ποιοτικές** μεταβλητές, των οποίων οι τιμές τους δεν είναι αριθμοί. Τέτοιες είναι για παράδειγμα, η ομάδα αίματος (με τιμές A, B, AB, O), το φύλο (με τιμές αγόρι, κορίτσι), οι συνέπειες του καπνίσματος (με τιμές καρδιακά νοσήματα, καρκίνος κτλ), καθώς επίσης και το ενδιαφέρον των μαθητών ή φοιτητών για τη στατιστική, που μπορεί να χαρακτηριστεί ως υψηλό, μέτριο ή χαμηλό.
- Σε **ποσοτικές** μεταβλητές, των οποίων οι τιμές είναι αριθμοί και διακρίνονται σε:

i. **Διακριτές** μεταβλητές, που παίρνουν μόνο μεμονωμένες[3] τιμές. Τέτοιες μεταβλητές είναι για παράδειγμα, ο αριθμός υπαλλήλων μιας επιχείρησης (με τιμές 1, 2, …), το αποτέλεσμα της ρίψης ενός ζαριού (με τιμές 1, 2, …, 6) κτλ.

ii. **Συνεχείς** μεταβλητές, που μπορούν να πάρουν οποιαδήποτε τιμή ενός διαστήματος πραγματικών αριθμών (α, β). Τέτοιες μεταβλητές είναι το ύψος και το βάρος των υπαλλήλων σε μια οικονομική εταιρία, η διάρκεια μιας τηλεφωνικής κλήσης κτλ.

Επίσης ένας τρόπος για να πάρουμε τις απαραίτητες πληροφορίες που χρειαζόμαστε για κάποιο πληθυσμό είναι να εξετάσουμε όλα τα άτομα (στοιχεία) του πληθυσμού ως προς το χαρακτηριστικό που μας ενδιαφέρει. Η μέθοδος αυτή συλλογής των δεδομένων καλείται **απογραφή**.

1.3 ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ – ΕΠΑΓΩΓΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Ανάλογα με τον τρόπο που χρησιμοποιούμε τη Στατιστική διακρίνουμε σήμερα δύο μεγάλους τομείς της: την **περιγραφική στατιστική** και την **επαγωγική στατιστική ή στατιστική συμπερασματολογία.** Ο πρώτος τομέας η περιγραφική στατιστική περιλαμβάνει τις μεθόδους για τη συλλογή, ταξινόμηση, περιγραφή, παρουσίαση δεδομένων με πίνακες και γραφήματα και στα οποία υπολογίζουμε τα στατιστικά μέτρα (μέση τιμή, εύρος).

[3] μεμονωμένες: αριθμήσιμο πλήθος τιμών

Ενώ ο δεύτερος τομέας, η επαγωγική στατιστική ή στατιστική συμπερασματολογία, που ασχολείται με την εξαγωγή συμπερασμάτων που αφορούν το δείγμα, από το δείγμα στο πληθυσμό.

Οι βασικές τεχνικές σύνοψης και περιγραφής αριθμητικών δεδομένων είναι οι πίνακες, τα διαγράμματα και τα αριθμητικά μέτρα που θα αναλυθούν παρακάτω.

1.4 ΠΙΝΑΚΕΣ

Οι πίνακες είναι οι πλέον απλές τεχνικές σύνοψης και παρουσίασης δεδομένων και χρησιμοποιούνται για την περιγραφή και κατανόηση μεταβλητών διαφόρων τύπων. Από απόψεως κατασκευής, ένας πίνακας ορίζεται από ένα νοητό πλέγμα γραμμών και στηλών, στο εσωτερικό του οποίου οργανώνονται τα δεδομένα με τρόπο συστηματικό. Επομένως παράδειγμα ενός πίνακα σύμφωνα με τον παραπάνω ορισμό έχουμε ως εξής:

ΠΙΝΑΚΑΣ 1

Τομέας αρμοδιοτήτων του κλάδου επαγγέλματός τους 10 υπαλλήλων σε ένα κατάστημα της ΕΘΝΙΚΗΣ ΤΡΑΠΕΖΑΣ ΕΛΛΑΔΟΣ Α.Ε.

α/α	**ΚΛΑΔΟΣ ΕΠΑΓΓΕΛΜΑΤΟΣ ΥΠΑΛΛΗΛΩΝ**
1	ΟΙΚΟΝΟΜΟΛΟΓΟΣ
2	ΤΑΜΕΙΑΣ
3	ΛΟΓΙΣΤΗΣ
4	ΤΑΜΕΙΑΣ
5	ΤΑΜΕΙΑΣ
6	ΟΙΚΟΝΟΜΟΛΟΓΟΣ
7	ΛΟΓΙΣΤΗΣ
8	ΟΙΚΟΝΟΜΟΛΟΓΟΣ
9	ΟΙΚΟΝΟΜΟΛΟΓΟΣ
10	ΤΑΜΕΙΑΣ

1.4.1 ΠΙΝΑΚΕΣ ΣΥΧΝΟΤΗΤΩΝ

Οι πίνακες συχνοτήτων χρησιμοποιούνται για την παρουσίαση κατανομών συχνοτήτων μεταβλητών όλων των τύπων. Αν πρόκειται για κατανομές κατηγορικών ή διατεταγμένων μεταβλητών, οι πίνακες αυτοί αποτελούνται από το σύνολο των επιμέρους κατηγοριών ή τάξεων που περιλαμβάνει η μεταβλητή, μαζί με τον αριθμό των παρατηρήσεων που αντιστοιχούν σε κάθε κατηγορία ή τάξη. Ο αριθμός των παρατηρήσεων που αντιστοιχούν σε κάθε κατηγορία ή τάξη μιας μεταβλητής, λέγεται απόλυτη συχνότητα ή απλά **συχνότητα** (frequency) και ορίζεται ως εξής:

Ας υποθέσουμε ότι $x_1, x_2, ..., x_\kappa$ είναι οι τιμές μιας μεταβλητής X, που αφορά ένα δείγμα μεγέθους ν, $\kappa \le \nu$. Στην τιμή x_i αντιστοιχεί η απόλυτη **συχνότητα** ν_i, δηλαδή ο φυσικός αριθμός που δείχνει πόσες φορές εμφανίζεται η τιμή x_i της εξεταζόμενης μεταβλητής X στο σύνολο των παρατηρήσεων και έχουμε:

$$\nu_1 + \nu_2 + ... + \nu_\kappa = \nu.$$

Εάν διαιρέσουμε την συχνότητα ν_i με το μέγεθος ν του δείγματος, προκύπτει η **σχετική συχνότητα** f_i της τιμής x_i, δηλαδή

$$f_i = \frac{\nu_i}{\nu}, \ i = 1,2,...,\kappa.$$

Για την σχετική συχνότητα ισχύουν οι ιδιότητες:

i) $0 \le f_i \le 1$ για $i = 1,2,...,\kappa$ αφού $0 \le \nu_i \le \nu$.

ii) $f_1 + f_2 + ... + f_\kappa = 1$, αφού

$$f_1 + f_2 + ... + f_\kappa = \frac{\nu_1}{\nu} + \frac{\nu_2}{\nu} + ... + \frac{\nu_\kappa}{\nu} = \frac{\nu_1 + \nu_2 + ... + \nu_\kappa}{\nu} = \frac{\nu}{\nu} = 1.$$

Επομένως σύμφωνα με τα παραπάνω και με τον πίνακα 1 έχουμε:

Για παράδειγμα για την μεταβλητή X: ‘‘κλάδος επαγγέλματος υπαλλήλων’’ οι συχνότητες για τις τιμές x_1 = ΟΙΚΟΝΟΜΟΛΟΓΟΣ, x_2 = ΤΑΜΕΙΕΣ,

x_3 = ΛΟΓΙΣΤΗΣ είναι, αντίστοιχα:

$\nu_1 = 4, \ \nu_2 = 4, \ \nu_3 = 2$ με $\nu_1 + \nu_2 + \nu_3 = 10$.

Άρα η λύση του παραδείγματος αυτού φαίνεται στον παρακάτω πίνακα 2.

ΛΥΣΗ: Οπότε φτιάχνουμε έναν πίνακα συχνοτήτων

ΠΙΝΑΚΑΣ 2

ΥΠΑΛΛΗΛΟΙ x_i	ΣΥΧΝΟΤΗΤΑ ν_i	ΣΧΕΤΙΚΗ ΣΥΧΝΟΤΗΤΑ f_i	ΣΧΕΤΙΚΗ ΣΥΧΝΟΤΗΤΑ $f_i\%$
ΟΙΚΟΝΟΜΟΛΟΓΟΣ	4	$\frac{4}{10}=0,4$	40
ΤΑΜΕΙΑΣ	4	$\frac{4}{10}=0,4$	40
ΛΟΓΙΣΤΗΣ	2	$\frac{2}{10}=0,2$	20
ΣΥΝΟΛΟ	10	1	100

Δηλαδή έχουμε: ν = $\nu_1+\nu_2+\nu_3=10$

Για τους οικονομολόγους έχουμε:

$$f_1=\frac{\nu_1}{\nu}=\frac{4}{10}=0,4$$

Για τους ταμείες έχουμε:

$$f_2=\frac{\nu_2}{\nu}=\frac{4}{10}=0,4$$

Ενώ για τους λογιστές έχουμε:

$$f_3=\frac{\nu_3}{\nu}=\frac{2}{10}=0,2$$

Άρα:

$$f_1+f_2+f_3=0,4+0,4+0,2=1.$$

Συνήθως τις σχετικές συχνότητες f_i τις εκφράζουμε επί τοις εκατό, οπότε συμβολίζονται με $f_i\%$, δηλαδή $f_i\%=100\times f_i$.

Συνεπώς

$f_1\%=100\times 0,4=40\%$, $f_2\%=40\%$ και $f_3\%=20\%$ με $f_1\%+f_2\%+f_3\%=100\%$.

Ερμηνεία $f_i\%$**:** Από τους 10 υπαλλήλους της ΕΘΝΙΚΗΣ ΤΡΑΠΕΖΑΣ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. το 40% των αρμοδιοτήτων τους, στον κλάδο επαγγέλματός τους είναι οικονομολόγοι και ταμείες , ενώ το 20% είναι λογιστές.

1.4.2 ΑΘΡΟΙΣΤΙΚΕΣ ΣΥΧΝΟΤΗΤΕΣ

Στην περίπτωση των **ποσοτικών μεταβλητών** εκτός από τις συχνότητες ν_i και f_i χρησιμοποιούνται συνήθως και οι λεγόμενες **αθροιστικές συχνότητες** (cumulative frequencies) N_i και οι **αθροιστικές σχετικές συχνότητες** (cumulative relative frequencies) F_i, οι οποίες εκφράζουν το πλήθος και το ποσοστό αντίστοιχα των παρατηρήσεων που είναι μικρότερες ή ίσες της τιμής x_i. Συχνά οι F_i πολλαπλασιάζονται επί 100 εκφραζόμενες έτσι επί τοις εκατό, δηλαδή $F_i\% = 100F_i$, (πίνακας 3). Αν οι τιμές $x_1, x_2, ..., x_\kappa$ μιας ποσοτικής μεταβλητής Χ είναι σε αύξουσα διάταξη, τότε η αθροιστική συχνότητα της τιμής x_i είναι $N_i = \nu_1 + \nu_2 + ... + \nu_i$. Ομοίως, η αθροιστική σχετική συχνότητα είναι $F_i = f_1 + f_2 + ... + f_i$, για $i = 1,2,..., \kappa$.

Για παράδειγμα, για την μεταβλητή Χ ‘‘ΥΠΑΛΛΗΛΟΙ’’ του πίνακα 2 είναι:

$N_1 = \nu_1 = 4$, $N_2 = \nu_1 + \nu_2 = 4 + 4 = 8$ και $N_3 = \nu_1 + \nu_2 + \nu_3 = \nu = 10$, οπότε

$F_1 = f_1 = 0{,}4$, $F_2 = f_1 + f_2 = 0{,}8$ και $F_3 = f_1 + f_2 + f_3 = 1$, οπότε

$F_1\% = 40\%$, $F_2\% = 80\%$ και $F_3\% = 100\%$. Άρα ισχύουν οι σχέσεις:

$\nu_1 = N_1$, $\nu_2 = N_2 - N_1, ..., \nu_\kappa = N_\kappa - N_{\kappa-1}$

και

$f_1 = F_1$, $f_2 = F_2 - F_1, ..., f_\kappa = F_\kappa - F_{\kappa-1}$.

ΠΙΝΑΚΑΣ 3

Κατανομή συχνοτήτων και αθροιστικών συχνοτήτων της μεταβλητής ‘‘ΥΠΑΛΛΗΛΟΙ’’ της ΕΘΝΙΚΗΣ ΤΡΑΠΕΖΑΣ του πίνακα 2.

Υπάλληλοι x_i	Συχνότητα ν_i	Σχετ. Συχν. f_i	Σχετ. Συχν. $f_i\%$	Αθροιστ. Συχνότ. N_i	Αθροιστ. Σχετ. Συχν. F_i	Αθροιστ. Σχετ. Συχν. $F_i\%$
Οικονομολόγοι	4	0,4	40	4	0,4	40
Ταμείας	4	0,4	40	8	0,8	80
Λογιστής	2	0,2	20	10	1	100
Σύνολο	10	1	100	_	_	_

1.4.3 ΠΙΝΑΚΕΣ ΣΥΧΝΟΤΗΤΩΝ ΜΕ ΟΜΑΔΟΠΟΙΗΜΕΝΑ ΔΕΔΟΜΕΝΑ

Οι πίνακες συχνοτήτων και κατ' αναλογίαν τα αντίστοιχα διαγράμματα είναι δύσκολο να κατασκευαστούν, όταν το πλήθος των τιμών μιας μεταβλητής είναι αρκετά μεγάλο. Αυτό μπορεί να συμβεί είτε στην περίπτωση μιας διακριτής μεταβλητής είτε, πολύ περισσότερο, στην περίπτωση μιας συνεχούς μεταβλητής, όπου αυτή μπορεί να πάρει οποιαδήποτε τιμή στο διάστημα ορισμού της. Σ' αυτές τις περιπτώσεις είναι απαραίτητο να ταξινομηθούν (ομαδοποιηθούν) τα δεδομένα σε μικρό πλήθος ομάδων, που ονομάζονται **κλάσεις** (class intervals), έτσι ώστε κάθε τιμή να ανήκει μόνο σε μία κλάση. Τα άκρα των κλάσεων καλούνται **όρια των κλάσεων** (class boundaries). Συνήθως υιοθετούμε την περίπτωση που μια κλάση περιέχει το κάτω άκρο της (κλειστή αριστερά) αλλά όχι το άνω άκρο της (ανοικτή δεξιά), δηλαδή που οι κλάσεις είναι της μορφής [,). Οι παρατηρήσεις κάθε κλάσης θεωρούνται όμοιες, οπότε μπορούν να "αντιπροσωπευθούν" από τις **κεντρικές τιμές**, τα κέντρα δηλαδή κάθε κλάσης.

- Το πρώτο βήμα στην ομαδοποίηση των δεδομένων είναι η εκλογή του αριθμού κ των ομάδων ή κλάσεων. Ο αριθμός αυτός συνήθως ορίζεται αυθαίρετα από τον ερευνητή σύμφωνα με την πείρα του. Γενικά όμως μπορεί να χρησιμοποιηθεί ως οδηγός ο παρακάτω πίνακας:

ΠΙΝΑΚΑΣ 4

Μέγεθος δείγματος ν	Αριθμός κλάσεων κ
< 20	5
$20 - 50$	6
$50 - 100$	7
$100 - 200$	8
$200 - 400$	9
$400 - 700$	10
$700 - 1000$	11
≥ 1000	12

- Το δεύτερο βήμα είναι ο προσδιορισμός του πλάτους των κλάσεων. **Πλάτος μιας κλάσης** ονομάζεται η διαφορά του κατωτέρου από το ανώτερο όριο της κλάσης. Στην πλειονότητα των πρακτικών εφαρμογών οι κλάσεις έχουν το ίδιο πλάτος. Φυσικά υπάρχουν και περιπτώσεις όπου επιβάλλεται οι κλάσεις να έχουν άνισο[4] πλάτος, όπως, για παράδειγμα, στις κατανομές εισοδήματος, ημερών απεργίας κτλ. Για να κατασκευάσουμε ισοπλατείς κλάσεις, χρησιμοποιούμε το **εύρος** (range) R του δείγματος, δηλαδή τη διαφορά της μικρότερης παρατήρησης από τη μεγαλύτερη παρατήρηση του συνολικού δείγματος. Τότε υπολογίζουμε το πλάτος c των κλάσεων διαιρώντας το εύρος R διά του αριθμού των κλάσεων κ, στρογγυλεύοντας, αν χρειαστεί για λόγους διευκόλυνσης, πάντα προς τα πάνω.
- Το επόμενο βήμα είναι η κατασκευή των κλάσεων. Ξεκινώντας από την μικρότερη παρατήρηση, ή για πρακτικούς λόγους λίγο πιο κάτω από την μικρότερη παρατήρηση, και προσθέτοντας κάθε φορά το πλάτος c δημιουργούμε τις κ κλάσεις. Αυτονόητο είναι ότι η μεγαλύτερη τιμή του δείγματος θα ανήκει οπωσδήποτε στην τελευταία κλάση.
- Τέλος, γίνεται η **διαλογή** των παρατηρήσεων. Το πλήθος των παρατηρήσεων ν_i που προκύπτουν από τη διαλογή για την κλάση i καλείται **συχνότητα της κλάσης** αυτής ή **συχνότητα της κεντρικής τιμής** x_i, $i = 1,2,...,\kappa$.

ΠΑΡΑΔΕΙΓΜΑ: Έστω ο παρακάτω πίνακας ο οποίος δείχνει το ύψος (cm) των 40 εργαζομένων της επιχείρησης singularlogic.

ΠΙΝΑΚΑΣ 5

170	180	178	165	170	168	175	175	173	162
160	170	167	177	180	170	182	178	165	178
[156]	175	172	173	167	187	170	180	178	[191]
176	169	167	166	179	178	180	164	170	173

ΛΥΣΗ:

Άρα R = μεγαλύτερη – μικρότερη παρατήρηση = 191 – 156 = 35.

[4] άνισο: διαφορετικό

Επειδή έχουμε $\nu = 40$ παρατηρήσεις, χρησιμοποιούμε $\kappa = 6$ κλάσεις. Το πλάτος των κλάσεων είναι:

$$c = \frac{R}{\kappa} = \frac{35}{6} = 5{,}83 \approx 6.$$

Αν θεωρήσουμε ως αρχή της πρώτης κλάσης το 156, θα έχουμε τον επόμενο πίνακα 6.
Πρέπει να προσεχτεί ότι:

- Καμία παρατήρηση δεν μπορεί να μείνει έξω από κάποια κλάση.

- Οι κεντρικές τιμές διαφέρουν μεταξύ τους όσο και το πλάτος των κλάσεων, που εδώ είναι ίσο με 6.

- Μία παρατήρηση που συμπίπτει με το άνω άκρο μιας κλάσης θα τοποθετηθεί κατά τη διαλογή στην αμέσως επόμενη κλάση. Για παράδειγμα, ο εργαζόμενος με ύψος 180 θα τοποθετηθεί στην πέμπτη κλάση [180, 186).

ΠΙΝΑΚΑΣ 6

Κλάσεις [-)	Κεντρικές τιμές x_i	Συχνότητα ν_i	Σχετική Συχνότητα $f_i\%$	Αθροιστική συχνότητα N_i	Αθροιστική σχετ. συχν. $F_i\%$
156-162	159	2	5	2	5
162-168	165	8	20	10	25
168-174	171	12	30	22	55
174-180	177	11	27,5	33	82,5
180-186	183	5	12,5	38	95
186-192	189	2	5	40	100
	Σύνολο	40	100	_	_

1.5 ΔΙΑΓΡΑΜΜΑΤΑ

Τα διαγράμματα είναι γραφικές κατασκευές, οι οποίες όπως και οι πίνακες στοχεύουν στην παρουσίαση των αριθμητικών δεδομένων. Είναι ευκολότερα στην ανάγνωσή τους σε σχέση με τους πίνακες, υστερούν όμως έναντι αυτών ως προς το βαθμό λεπτομέρειας που διασφαλίζουν κατά την παρουσίαση των δεδομένων. Η υστέρηση αυτή των διαγραμμάτων έναντι των πινάκων αντισταθμίζεται[5] από την αμεσότητα που έχουν τα διαγράμματα ως προς τη γραφική απεικόνιση της πληροφορίας που περιέχουν τα δεδομένα.

1.5.1 ΡΑΒΔΟΓΡΑΜΜΑΤΑ

Τα ραβδογράμματα (bar charts) χρησιμοποιούνται για την γραφική παράσταση των τιμών μιας ποιοτικής μεταβλητής. Τα ραβδογράμματα αποτελούνται από ορθογώνιες στήλες που οι βάσεις τους βρίσκονται πάνω στον οριζόντιο ή τον κατακόρυφο άξονα. Σε κάθε τιμή της μεταβλητής Χ αντιστοιχεί μια ορθογώνια στήλη της οποίας το ύψος είναι ίσο με την αντίστοιχη συχνότητα ή σχετική συχνότητα. Έτσι έχουμε το **ραβδόγραμμα συχνοτήτων** και το **ραβδόγραμμα σχετικών συχνοτήτων.** Τόσο η απόσταση μεταξύ των στηλών όσο και το μήκος των βάσεων τους καθορίζονται αυθαίρετα.[6] Στον πίνακα 7 έχουμε την κατανομή συχνοτήτων της μεταβλητής Χ: ‘‘απασχόληση ωρών λειτουργίας διάφορων οικονομικών επιχειρήσεων’’ και στα διαγράμματα 1 (α), (β) τα αντίστοιχα ραβδογράμματα συχνοτήτων και σχετικών συχνοτήτων.

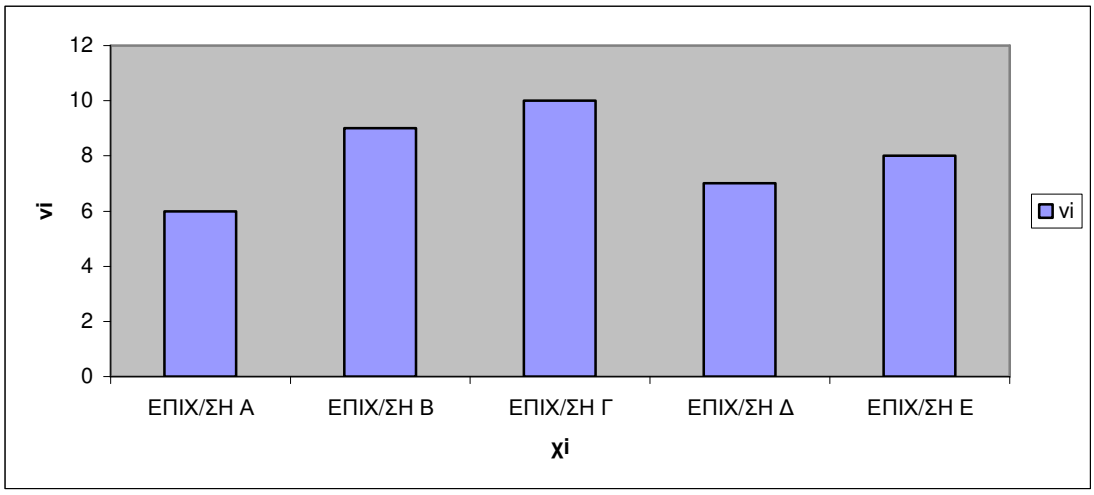
[5] αντισταθμίζω: ισορροπώ, εξισώνω.
[6] αυθαίρετα: οικοδόμημα

ΠΙΝΑΚΑΣ 7

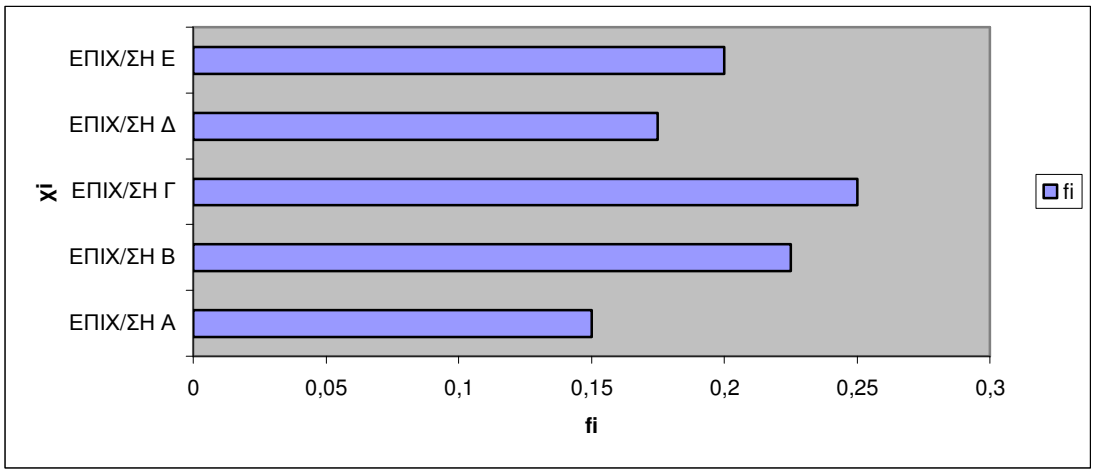
Κατανομή συχνοτήτων για την απασχόληση ωρών λειτουργίας διάφορων επιχειρήσεων

i	Απασχόληση x_i	Συχνότητα ν_i	Σχετική συχνότητα f_i	Σχετική συχνότητα $f_i\,\%$
1	ΕΠΙΧΕΙΡΗΣΗ Α	6	0,15	15
2	ΕΠΙΧΕΙΡΗΣΗ Β	9	0,225	22,5
3	ΕΠΙΧΕΙΡΗΣΗ Γ	10	0,25	25
4	ΕΠΙΧΕΙΡΗΣΗ Δ	7	0,175	17,5
5	ΕΠΙΧΕΙΡΗΣΗ Ε	8	0,2	20
ΣΥΝΟΛΟ		40	1	100

ΔΙΑΓΡΑΜΜΑ 1 (α)



ΔΙΑΓΡΑΜΜΑ 1 (β)



1.5.2 ΙΣΤΟΓΡΑΜΜΑ

Η αντίστοιχη γραφική παράσταση ενός πίνακα συχνοτήτων με ομαδοποιημένα δεδομένα γίνεται με το λεγόμενο ιστόγραμμα (histogram) συχνοτήτων. Στον οριζόντιο άξονα ενός συστήματος ορθογωνίων αξόνων σημειώνουμε, με κατάλληλη κλίμακα, τα όρια των κλάσεων. Στη συνέχεια, κατασκευάζουμε διαδοχικά ορθογώνια (ιστούς), από καθένα από τα οποία έχει βάση ίση με το πλάτος της κλάσης και ύψος τέτοιο, ώστε το εμβαδόν του ορθογωνίου να ισούται με την συχνότητα της κλάσης αυτής.

1.5.3 ΠΟΛΥΓΩΝΟ ΣΥΧΝΟΤΗΤΩΝ

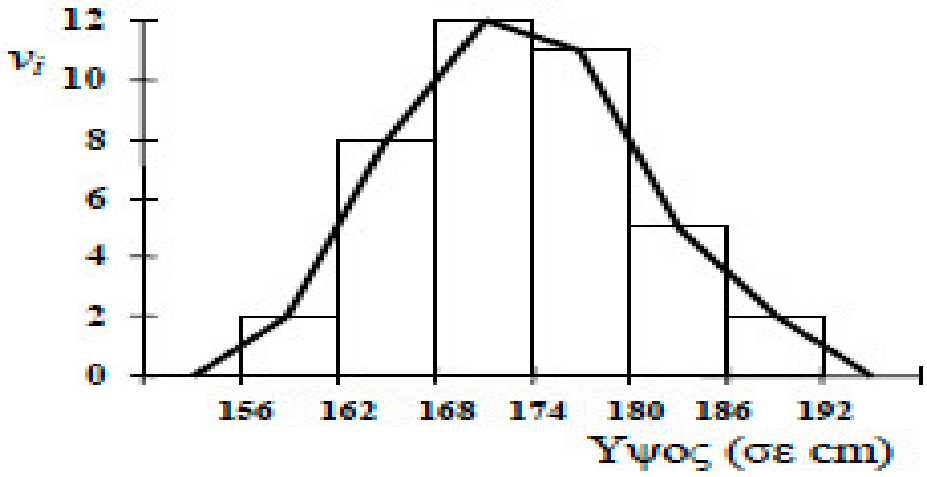
Το πολύγωνο συχνοτήτων (frequency polygon) είναι ένα παράγωγο διάγραμμα, το οποίο απορρέει[7] από την κατασκευή ενός ιστογράμματος. Κατασκευάζεται, αν σε ένα ιστόγραμμα (συχνοτήτων ή σχετικών συχνοτήτων) ενώσουμε τα κέντρα των κορυφών των στηλών του. Η πολυγωνική γραμμή που θα προκύψει, ορίζει το αντίστοιχο πολύγωνο συχνοτήτων ή το πολύγωνο σχετικών συχνοτήτων. Θεωρητικά, όταν ο αριθμός των παρατηρήσεων αυξάνει απεριόριστα και το εύρος των διαστημάτων ελαττώνεται, τότε το ιστόγραμμα και το πολύγωνο των συχνοτήτων ή σχετικών συχνοτήτων τείνουν να συμπέσουν σε μια συνεχή

[7] απορρέει: προκύπτει

καμπύλη, η οποία ονομάζεται καμπύλη συχνοτήτων (frequency curve). Τόσο το ιστόγραμμα όσο και το πολύγωνο συχνοτήτων, ουσιαστικά μεταφέρουν την ίδια πληροφορία για την κατανομή της μεταβλητής στην οποία αναφέρονται.

Άρα το πολύγωνο συχνοτήτων με βάση του παραδείγματος από το ύψος των εργαζομένων της singularlogic και από τα δεδομένα του πίνακα 6 θα είναι το εξής διάγραμμα:

ΔΙΑΓΡΑΜΜΑ 2



1.6 ΑΡΙΘΜΗΤΙΚΑ ΠΕΡΙΓΡΑΦΙΚΑ ΜΕΤΡΑ

Τα αριθμητικά περιγραφικά μέτρα είναι αντιπροσωπευτικές τιμές, οι οποίες περιγράφουν με τρόπο ποσοτικό την κατανομή μιας μεταβλητής. Τα μέτρα αυτά διακρίνονται σε μέτρα κεντρικής τάσης ή θέσης και μέτρα διασποράς.

1.6.1 ΜΕΤΡΑ ΚΕΝΤΡΙΚΗΣ ΤΑΣΗΣ Η ΘΕΣΗΣ

Κεντρική τάση ή θέση μιας ομάδας δεδομένων είναι η ιδιότητα κατά την οποία οι τιμές μιας ομάδας τείνουν να συγκεντρωθούν γύρω από μια ορισμένη τιμή (μέση τιμή), η τιμή αυτή

είναι το «κέντρο» της κατανομής συχνοτήτων. Οι παράμετροι που συνοψίζουν[8] τα δεδομένα της παρατήρησης με ένα αντιπροσωπευτικό αριθμό και μετράνε την κεντρική τάση των μετρήσεων που ονομάζονται μέσοι όροι ή απλώς μέσοι. Οι μέσοι διακρίνονται σε δύο κατηγορίες:

- Μέση κεντρικής τάσης στους οποίους περιλαμβάνονται ο αριθμητικός, ο γεωμετρικός και ο αρμονικός μέσος.
- Μέση (παράμετροι) θέσης στους οποίους περιλαμβάνονται η διάμεσος, τα τεταρτημόρια, τα δεκατημόρια, τα εκατοστημόρια και η επικρατούσα τιμή.

1.6.1.1 ΜΕΣΗ ΤΙΜΗ

Το πλέον γνωστό και ευρύτερα χρησιμοποιούμενο μέτρο κεντρικής τάσης ή θέσης είναι η μέση τιμή. Η μέση τιμή ενός συνόλου ν παρατηρήσεων ορίζεται ως το άθροισμα των παρατηρήσεων διά του πλήθους των παρατηρήσεων. Όταν σε ένα δείγμα μεγέθους ν οι παρατηρήσεις μιας μεταβλητής Χ είναι $t_1, t_2, ..., t_\nu$, τότε η μέση τιμή συμβολίζεται με $\bar{x}$ και δίνεται από την σχέση:

$$\bar{x} = \frac{t_1 + t_2 + ... + t_\nu}{\nu} = \frac{\sum_{i=1}^{\nu} t_i}{\nu} = \frac{1}{\nu}\sum_{i=1}^{\nu} t_i \qquad (1)$$

όπου το σύμβολο $\sum_{i=1}^{\nu} t_i$ παριστάνει μια συντομογραφία του αθροίσματος $t_1 + t_2 + ... + t_\nu$ και διαβάζεται "άθροισμα των t_i από $i = 1$ έως ν". Συχνά, όταν δεν υπάρχει πρόβλημα σύγχυσης, συμβολίζεται και ως $\sum t_i$ ή ακόμα πιο απλά με $\sum t$.

Σε μια κατανομή συχνοτήτων, αν $x_1, x_2, ..., x_\kappa$ είναι οι τιμές της μεταβλητής Χ με συχνότητες $\nu_1, \nu_2, ..., \nu_\kappa$ αντίστοιχα, η μέση τιμή ορίζεται ισοδύναμα από την σχέση:

$$\bar{x} = \frac{x_1\nu_1 + x_2\nu_2 + ...x_\kappa\nu_\kappa}{\nu_1 + \nu_2 + ... + \nu_\kappa} = \frac{\sum_{i=1}^{\kappa} x_i\nu_i}{\sum_{i=1}^{\kappa} \nu_i} = \frac{1}{\nu}\sum_{i=1}^{\kappa} x_i\nu_i \qquad (2)$$

Η παραπάνω σχέση ισοδύναμα γράφεται:

[8] συνοψίζουν: συντομεύουν, συγκεφαλαιώνουν

$$\bar{x} = \sum_{i=1}^{\kappa} x_i \frac{\nu_i}{\nu} = \sum_{i=1}^{\kappa} x_i f_i$$

όπου f_i οι σχετικές συχνότητες.

1.6.1.2 ΣΤΑΘΜΙΚΟΣ ΜΕΣΟΣ

Αν έχουμε μια σειρά ν τιμών $x_1, x_2, ..., x_\nu$ και σε κάθε τιμή δώσουμε διαφορετική σημασία που εκφράζονται από αριθμούς που λέγονται συντελεστές στάθμισης (βαρύτητας), $w_1, w_2, ...w_\nu$ τότε ο μέσος όρος υπολογίζεται βάση του τύπου:

$$\bar{x} = \frac{x_1 w_1 + x_2 w_2 + ... + x_\nu w_\nu}{w_1 + w_2 + ... + w_\nu} = \frac{\sum_{i=1}^{\nu} x_i w_i}{\sum_{i=1}^{\nu} w_i}.$$

1.6.1.3 ΔΙΑΜΕΣΟΣ

Όταν οι τιμές μιας μεταβλητής x_i ταξινομηθούν κατά την φυσική τους τάξη από την μικρότερη προς την μεγαλύτερη, τότε η διάμεσος είναι η τιμή εκείνη της μεταβλητής η οποία κατέχει την κεντρική θέση. Με άλλα λόγια διάμεσος είναι η τιμή της μεταβλητής η οποία χωρίζει το σύνολο τιμών της μεταβλητής σε δύο ισοπληθείς ομάδες. Συνεπώς κατά τις διαμέσου βρίσκονται το 50% των μεταβλητών και άνω τις διαμέσου τα υπόλοιπα 50% αυτών. Επομένως η διάμεσος υπολογίζεται με δύο τρόπους:

1ος τρόπος: **<u>ΥΠΟΛΟΓΙΣΜΟΣ ΔΙΑΜΕΣΟΥ ΑΤΑΞΙΝΟΜΗΤΩΝ ΔΕΔΟΜΕΝΩΝ</u>**

- Όταν το πλήθος των τιμών είναι μονός αριθμός. Στην περίπτωση αυτή, υπάρχει μόνο μια τιμή της μεταβλητής που κατέχει την κεντρική θέση. Η κεντρική αυτή τιμή είναι και η διάμεσος.

π.χ.: $n = 9$

164, 174, 166, 171, 159, 169, 186, 176, 178

ΛΥΣΗ: Η διάμεσος υπολογίζεται εφόσον διαταχθούν οι τιμές κατ' αύξουσα σειρά:

159, 164, 166, 169, 171, 174, 176, 178, 186

Άρα:

$$\frac{n+1}{2}=\frac{9+1}{2}=5$$, δηλαδή M = 171

- Όταν το πλήθος των τιμών είναι ζυγός αριθμός. Στην περίπτωση αυτή δεν υπάρχει μόνο μια τιμή, η οποία κατέχει την κεντρική θέση, αλλά δύο τιμές ως τιμή της διαμέσου θεωρούμε τον μέσο όρο των τιμών, των δύο κεντρικών τιμών.

π.χ.: $n=8$

159, 165, 166, 169, 171, 174, 176, 178

ΛΥΣΗ: Η διάμεσος εδώ υπολογίζεται μεταξύ του 169 και 171

Άρα:

$$\frac{n+1}{2}=\frac{8+1}{2}=4{,}5$$

Επομένως:

$$\mathrm{M}=\frac{169+171}{2}=170\ .$$

2ος τρόπος: **<u>ΥΠΟΛΟΓΙΣΜΟΣ ΔΙΑΜΕΣΟΥ ΤΑΞΙΝΟΜΗΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΚΑΤΑΝΟΜΗ ΣΥΧΝΟΤΗΤΩΝ</u>**

Έχουμε:

$$\mathrm{M}=x_{\lambda}+\frac{\delta}{\nu_i}\left[\frac{\mathrm{N}}{2}-\Phi_i\right]$$

Όπου x_{λ} : το κατώτερο όριο της τάξης που εντοπίζεται η διάμεσος

δ: το πλάτος της τάξης που εντοπίζεται η διάμεσος

ν_i : η συχνότητα της τάξης που εντοπίζεται η διάμεσος

N_i : το σύνολο συχνοτήτων της κατανομής

Φ_i : η αμέσως μικρότερη του N / 2, δεξιόστροφη αθροιστική συχνότητα δηλαδή

$\Phi_i \le \frac{\mathrm{N}}{2}$.

ΠΑΡΑΔΕΙΓΜΑ: Δίνεται ο πίνακας της κατανομής 200 επιχειρήσεων ανάλογα με το ύψος μηνιαίων πωλήσεων τους.

ΠΙΝΑΚΑΣ 8

Τάξεις αξία πωλήσεων (εκατ. ευρώ)	Αριθμός επιχειρήσεων ν_i	Φ_i
4-6	10	10
6-8	20	30
8-10	30	60
10-12	80	140 ←
12-14	30	170
14-16	20	190
16-18	10	200
ΣΥΝΟΛΟ	200	_

ΛΥΣΗ:

$$\frac{\mathrm{N}}{2}=\frac{200}{2}=100$$

Έπειτα πρέπει να εντοπίσω την τάξη στην οποία ανήκει η διάμεσος. Αυτό γίνεται ως εξής:

Πηγαίνουμε στην στήλη Φ_i και βρίσκουμε την 100η παρατήρηση που είναι στο 140 ανάμεσα στις τάξεις 10-12 και έχουμε:

$x_\lambda = 10$. Δηλαδή που είναι η μικρότερη τιμή της τάξης την οποία και παίρνουμε

$\nu_i = 80$

$\delta = 2$

$\Phi_i = 60$. Δηλαδή παίρνουμε την αμέσως πάνω τιμή.

Οπότε

$$\mathrm{M} = x_\lambda + \frac{\delta}{\nu_i}\left[\frac{\mathrm{N}}{2} - \Phi_i\right]$$

$$M = 10 + \frac{2}{80}\left[\frac{200}{2} - 60\right]$$

$$M = 10 + \frac{1}{40}(100 - 60)$$

$$M = 10 + \frac{1}{40}40 = 11 \text{ εκατ. €}$$

Σημαίνει ότι το 50% των επιχειρήσεων έχει αξία πωλήσεων μέχρι 11 εκατ. €, ενώ το υπόλοιπο των 50% έχει αξία πωλήσεων πάνω από 11 εκατ. €.

1.6.1.4 ΕΠΙΚΡΑΤΟΥΣΑ ΤΙΜΗ

Είναι η τιμή της μεταβλητής που αντιστοιχεί στην μεγαλύτερη συχνότητα της κατανομής γι' αυτό ονομάζεται σημείο μεγαλύτερης συχνότητας και συμβολίζεται με το M_0. Ο τύπος είναι ο εξής:

$$M_0 = x_\lambda + \delta\frac{\Delta_1}{\Delta_1 + \Delta_2}$$

x_λ : το κατώτερο όριο της τάξης στην οποία αντιστοιχεί μεγαλύτερη συχνότητα

δ: το πλάτος της τάξης με την μεγαλύτερη συχνότητα

Δ_1 : Διαφορά: Μεγαλύτερη συχνότητα – προηγούμενη

Δ_2 : Διαφορά: Μεγαλύτερη συχνότητα – επόμενη

ΠΑΡΑΔΕΙΓΜΑ: Δίνεται η κατανομή 100 εργατών ανάλογα με τα ημερομίσθια

ΠΙΝΑΚΑΣ 9

ΗΜΕΡΟΜΙΣΘΙΑ ΕΡΓΑΤΩΝ ΤΑΞΕΙΣ	ΑΠΟΛΥΤΗ ΣΥΧΝΟΤΗΤΑ ν_i	Φ_i
15-25	5	5
25-35	13	18
35-45	20	38
45-55	35	73
55-65	18	91
65-75	7	98
75-85	2	100
ΣΥΝΟΛΟ	100	

ΛΥΣΗ:

Εντοπίζουμε την μεγαλύτερη συχνότητα που αντιστοιχεί στο 73 και έχουμε:

$x_\lambda = 45,\ \delta = 10,\ \Delta_1 = 35 - 20 = 15,\ \ \Delta_2 = 35 - 18 = 17.$

Άρα:

$$\mathrm{M}_0 = x_\lambda + \delta \frac{\Delta_1}{\Delta_1 + \Delta_2} =$$

$$= 45 + 10 \frac{15}{15 + 17} =$$

$$= 45 + 10 \frac{15}{32} =$$

$$= 45 + 4{,}68 = 49{,}6\,€$$

Επομένως το συνηθέστερο ημερομίσθιο είναι περίπου 49,6 €.

1.6.2 ΜΕΤΡΑ ΔΙΑΣΠΟΡΑΣ

Οι παράμετροι που μετράνε πόσο συγκεντρωμένα είναι τα δεδομένα γύρω από τον αριθμητικό μέσο, ονομάζονται μέτρα διασποράς. Αυτά είναι:

α) το **εύρος μεταβολής:** Είναι το απλούστερο μέτρο διασποράς και είναι η διαφορά ανάμεσα στην μεγαλύτερη (M) και στην μικρότερη (m) τιμή των δεδομένων, δηλαδή R = M – m. Το R δεν θεωρείται αξιόπιστο μέτρο διασποράς, γιατί εξαρτάται μόνο από τις δύο ακραίες τιμές των στατιστικών δεδομένων. Αυτό είναι το μεγάλο μειονέκτημα του εύρους ως μέτρο διασποράς, διότι αν μια τιμή στα δεδομένα μας είναι πολύ χαμηλή ή πολύ υψηλή τότε η τιμή αυτή θα επηρεάσει σημαντικά το μέγεθος του εύρους.

πλεονεκτήματα: Το R μπορεί να χρησιμοποιηθεί σαν μέτρο διασποράς ορισμένων μεγεθών, τα οποία εμφανίζονται με μορφή ανώτερων και κατώτερων τιμών όπως στα χρηματιστήρια για τις τιμές των μετοχών κ.τ.λ.

β) το **ημιενδοτεταρτημοριακό εύρος:** Στο διάστημα ανάμεσα στα δύο τεταρτημόρια Q_1 και Q_3 περιλαμβάνονται τα 50% των τιμών της μεταβλητής. Τα υπόλοιπα 50% των τιμών βρίσκονται κάτω του Q_1 και πάνω του Q_3 . Επομένως όσο μεγαλύτερη είναι η απόσταση ανάμεσα στο Q_1 και Q_3 τόσο μεγαλύτερη θα είναι και η διασπορά των τιμών της εξεταζόμενης μεταβλητής και αντίστροφα. Το μισό της διαφοράς $Q_3 - Q_1$ αποτελεί ένα ακόμα μέτρο διασποράς που ονομάζεται ημιενδοτεταρτημοριακό εύρος και υπολογίζεται βάση του τύπου:

$$Q = \frac{Q_3 - Q_1}{2}.$$

Το Q ως μέτρο διασποράς πλεονεκτεί του εύρους, γιατί δεν επηρεάζεται από τις ακραίες τιμές της μεταβλητής και επιπλέον μπορεί να υπολογιστεί και στις ανοικτές κατανομές συχνοτήτων. Το Q ως μέτρο διασποράς έχει το μειονέκτημα ότι δεν λαμβάνει υπόψη όλες τις τιμές της μεταβλητής.

γ) Η **μέση απόκλιση:** Η μέση απόκλιση ορίζεται ως ο αριθμητικός μέσος των απολύτων αποκλίσεων (διαφορά) των τιμών μιας μεταβλητής x_i από τον μέσο αριθμητικό τους. Ο τύπος είναι:

$$\mathrm{M_A} = \frac{\Sigma|x_i - \bar{x}|}{n}.$$

ΠΑΡΑΔΕΙΓΜΑ: Έστω ότι έχουμε τα ημερομίσθια 10 εργατών:

x_i : 40, 42, 43,5 , 45, 47, 48, 50, 51,5 , 52, 53

Να βρεθεί η μέση απόκλιση.

ΛΥΣΗ:

$$\bar{x} = \frac{40+42+43{,}5+45+47+48+50+51{,}5+52+53}{10} = \frac{472}{10} = 47{,}2\,.$$

$$\mathrm{M_A} = \frac{|40-47{,}2|+|42-47{,}2|+|43{,}5-47{,}2|+\ldots+|53-47{,}2|}{10} = 3{,}7\,€\,.$$

Η διαφορά 3,7 € σημαίνει ότι το ημερομίσθιο κάθε εργάτη διαφέρει κατά μέσο όρο, από το μέσο ημερομίσθιο κατά 3,7 €. Η μέση απόκλιση πλεονεκτεί από τα δύο επόμενα μέτρα διασποράς R και Q γιατί λαμβάνει υπόψη όλες τις τιμές της μεταβλητής.

δ) η **διακύμανση:** Ένας άλλος τρόπος για να υπολογίσουμε τη διασπορά των παρατηρήσεων $t_1, t_2, ..., t_\nu$ μιας μεταβλητής X θα ήταν να αφαιρέσουμε τη μέση τιμή $\bar{x}$ από κάθε παρατήρηση και να βρούμε τον αριθμητικό μέσο των διαφορών αυτών, δηλαδή τον αριθμό:

$$\frac{(t_1-\bar{x})+(t_2-\bar{x})+\ldots+(t_\nu-\bar{x})}{\nu} = \frac{\sum_{i=1}^{\nu}(t_i-\bar{x})}{\nu}.$$

Ο αριθμός όμως αυτός είναι ίσος με μηδέν αφού:

$$\frac{(t_1-\bar{x})+(t_2-\bar{x})+\ldots+(t_\nu-\bar{x})}{\nu} = \frac{t_1+t_2+\ldots+t_\nu}{\nu} - \frac{\nu\bar{x}}{\nu} = \bar{x}-\bar{x} = 0\,.$$

Γι' αυτό, ως ένα μέτρο διασποράς παίρνουμε τον μέσο όρο των τετραγώνων των αποκλίσεων των t_i από τη μέση τιμή τους $\bar{x}$. Το μέτρο αυτό καλείται **διακύμανση** ή **διασπορά** (variance) και ορίζεται από τη σχέση:

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(t_i-\bar{x})^2 \qquad (1)$$

Ο τύπος αυτός αποδεικνύεται ότι μπορεί να πάρει την ισοδύναμη μορφή:

$$s^2 = \frac{1}{\nu}\left\{\sum_{i=1}^{\nu}t_i^{\,2} - \frac{\left(\sum_{i=1}^{\nu}t_i\right)^2}{\nu}\right\} \qquad (2)$$

η οποία διευκολύνει σημαντικά τους υπολογισμούς κυρίως όταν η μέση τιμή $\bar{x}$ δεν είναι ακέραιος αριθμός.

Όταν έχουμε πίνακα συχνοτήτων ή ομαδοποιημένα δεδομένα, η διακύμανση ορίζεται από την σχέση:

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{\kappa}\left(x_i - \bar{x}\right)^2 \nu_i \qquad (3)$$

ή την ισοδύναμη μορφή:

$$s^2 = \frac{1}{\nu}\left\{\sum_{i=1}^{\kappa} x_i{}^2 \nu_i - \frac{\left(\sum_{i=1}^{\kappa} x_i \nu_i\right)^2}{\nu}\right\}. \qquad (4)$$

όπου $x_1, x_2, ..., x_\kappa$ οι τιμές της μεταβλητής (ή τα κέντρα των κλάσεων) με αντίστοιχες συχνότητες $\nu_1, \nu_2, ..., \nu_\kappa$.

ε) η **τυπική απόκλιση**: Αν πάρουμε τη θετική τετραγωνική ρίζα της διακύμανσης, θα έχουμε ένα μέτρο διασποράς που θα εκφράζεται με την ίδια μονάδα μέτρησης του χαρακτηριστικού, όπως ακριβώς είναι και όλα τα άλλα μέτρα θέσης, που εξετάσαμε έως τώρα. Η ποσότητα αυτή λέγεται **τυπική απόκλιση** (standard deviation), συμβολίζεται με s και δίνεται από τη σχέση:

$$s = \sqrt{s^2}\,.$$

στ) ο **συντελεστής μεταβλητότητας:** Ένα μέτρο το οποίο μας βοηθά στη σύγκριση ομάδων τιμών, που είτε εκφράζονται σε διαφορετικές μονάδες μέτρησης είτε εκφράζονται στην ίδια μονάδα μέτρησης, αλλά έχουν σημαντικά διαφορετικές μέσες τιμές, είναι ο συντελεστής μεταβολής ή συντελεστής μεταβλητότητας (coefficient of variation), ο οποίος ορίζεται από το λόγο:

$$CV = \frac{s}{\bar{x}}.$$

Ο συντελεστής μεταβολής εκφράζεται επί τοις εκατό, είναι συνεπώς ανεξάρτητος από τις μονάδες μέτρησης και παριστάνει ένα μέτρο **σχετικής διασποράς** των τιμών και όχι της απόλυτης διασποράς, εκφράζει δηλαδή τη μεταβλητότητα των δεδομένων απαλλαγμένη από την επίδραση της μέσης τιμής.

ΚΕΦΑΛΑΙΟ 2 ΠΑΛΙΝΔΡΟΜΗΣΗ

2.1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΠΑΛΙΝΔΡΟΜΗΣΗ

Σε διάφορα προβλήματα της Στατιστικής το ενδιαφέρον μας εστιάζεται στην ταυτόχρονη μελέτη **δύο ή περισσότερων** μεταβλητών, για να προσδιορίσουμε με ποιο τρόπο οι μεταβλητές αυτές σχετίζονται μεταξύ τους. Για παράδειγμα:

- Η ηλικία και το βάρος ενός παιδιού έχουν κάποια θετική εξάρτηση (συσχέτιση) μεταξύ τους με την έννοια ότι όσο πιο μεγάλο είναι το παιδί τόσο μεγαλύτερο βάρος θα έχει.
- Η διάρκεια ζωής των ζώντων οργανισμών σε μια περιοχή και το ποσοστό μόλυνσης της περιοχής έχουν αρνητική εξάρτηση μεταξύ τους, με την έννοια ότι όσο πιο μεγάλο είναι το ποσοστό μόλυνσης της περιοχής τόσο μικρότερη είναι η διάρκεια ζωής των οργανισμών που ζουν στην περιοχή.
- Η συνολική παραγωγή ενός αγρού εξαρτάται από την ποσότητα λιπάσματος που χρησιμοποιήθηκε, από τη θερμοκρασία και την υγρασία της περιοχής κτλ.
- Το ύψος των αποδοχών των υπαλλήλων μιας εταιρίας εξαρτάται από τα χρόνια υπηρεσίας στην εταιρία, από το επίπεδο μόρφωση τους κτλ.

Έτσι λοιπόν είναι ενδιαφέρον να εξεταστούν οι επιδράσεις που κάποιες μεταβλητές ασκούν σε κάποιες άλλες μεταβλητές. Η ύπαρξη μιας συναρτησιακής σχέσης (εξίσωσης) μεταξύ των μεταβλητών μπορεί να είναι εξαιρετικά πολύτιμη για την πρόβλεψη των τιμών μιας μεταβλητής από τις γνώσεις που διαθέτουμε για τις άλλες μεταβλητές, όταν ισχύουν κάποιες συγκεκριμένες συνθήκες.

Ο κλάδος της Στατιστικής που εξετάζει τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών με απώτερο σκοπό την πρόβλεψη μιας απ'αυτές μέσω των άλλων χαρακτηρίζεται με την ονομασία **ανάλυση παλινδρόμησης** (regression analysis). Ιστορικά, ο όρος "regression" χρησιμοποιήθηκε για πρώτη φορά από τον Άγγλο ανθρωπολόγο Galton (1822-1911) το 1885. Ο Galton, κατά την μελέτη του ύψους των παιδιών σε σχέση με το ύψος των γονέων διαπιστώθηκε ότι παιδιά ψηλών γονέων τείνουν, κατά μέσο όρο, να είναι κοντύτερα των γονιών τους, ενώ παιδιά κοντών γονέων τείνουν, κατά μέσο όρο, να γίνονται ψηλότερα των γονιών τους.

2.2 ΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Στην απλή παλινδρόμηση, χρησιμοποιούμε μόνο μια μεταβλητή Χ, και μια δεύτερη μεταβλητή Υ η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία συνάρτηση του Χ πχ. Υ να εκφράζεται μέσω της Χ ως $Y \approx 3X + 5$

Χ: ανεξάρτητη μεταβλητή

Υ: εξαρτημένη μεταβλητή

Επίσης, η παλινδρόμηση στην οποία υπάρχει μόνο μία ανεξάρτητη μεταβλητή καλείται **απλή παλινδρόμηση.** Για παράδειγμα:

Η εύρεση της σχέσης μεταξύ της συνολικής παραγωγής ενός αγρού και της ποσότητας λιπάσματος που χρησιμοποιήθηκε.

Τέλος για την εύρεση του κατάλληλου μοντέλου για την περιγραφή της σχέσης μεταξύ δύο μεταβλητών που μας ενδιαφέρουν, συνήθως ξεκινάμε κατασκευάζοντας το διάγραμμα διασποράς στο επίπεδο των παρατηρήσεων που διαθέτουμε. Σε ένα τέτοιο διάγραμμα οι τιμές της μεταβλητής Χ τοποθετούνται στον οριζόντιο άξονα και της μεταβλητής Υ στον κατακόρυφο άξονα.

2.3 ΠΟΛΛΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Είναι η ανάλυση παλινδρόμησης με μία εξαρτημένη μεταβλητή και δύο ή περισσότερες ανεξάρτητες μεταβλητές. Για παράδειγμα:

Η εύρεση της σχέσης μεταξύ της συνολικής παραγωγής ενός αγρού και της ποσότητας λιπάσματος που χρησιμοποιήθηκε, της θερμοκρασίας της περιοχής και της υγρασίας της περιοχής.

2.3.1 ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

1. Το **πιθανοθεωρητικό μοντέλο πολλαπλής παλινδρόμησης** είναι:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots + \beta_\kappa x_\kappa + \varepsilon$$

όπου:

y: η τιμή της εξαρτημένης μεταβλητής

α : η σταθερά της παλινδρόμησης

β_1 :ο συντελεστής ευαισθησίας της πρώτης ανεξάρτητης μεταβλητής

β_2 : ο συντελεστής ευαισθησίας της δεύτερης ανεξάρτητης μεταβλητής

β_κ : ο συντελεστής ευαισθησίας της κ-ανεξάρτητης μεταβλητής

κ: ο αριθμός των ανεξάρτητων μεταβλητών

ε: το σφάλμα της πρόβλεψης.

2. Το **εκτιμώμενο μοντέλο παλινδρόμησης** έχει την μορφή:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + ... + \hat{\beta}_\kappa x_\kappa$$

όπου:

$\hat{y}$: η προβλεφθείσα τιμή της y

$\hat{\alpha}$: η εκτίμηση της σταθεράς της παλινδρόμησης

$\hat{\beta}_1$: η εκτίμηση του συντελεστή της πρώτης παλινδρόμησης

$\hat{\beta}_2$: η εκτίμηση του συντελεστή της δεύτερης παλινδρόμησης

$\hat{\beta}_\kappa$: η εκτίμηση του συντελεστή κ της παλινδρόμησης

κ: ο αριθμός των ανεξάρτητων μεταβλητών.

3. Το **πολλαπλό μοντέλο παλινδρόμησης με δύο ανεξάρτητες μεταβλητές (πρώτης τάξης)**

Εδώ έχουμε:

α) Μοντέλο Πληθυσμού:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

όπου:

α : η σταθερά της παλινδρόμησης

β_1 :ο συντελεστής ευαισθησίας της πρώτης ανεξάρτητης μεταβλητής

β_2 : ο συντελεστής ευαισθησίας της δεύτερης ανεξάρτητης μεταβλητής

ε: το σφάλμα πρόβλεψης.

β) Εκτιμώμενο μοντέλο:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

όπου:

ŷ: η προβλεπόμενη τιμή της y

$\hat{\alpha}$: η εκτίμηση της σταθεράς της παλινδρόμησης

$\hat{\beta}_1$: η εκτίμηση του συντελεστή ευαισθησίας της πρώτης παλινδρόμησης

$\hat{\beta}_2$: η εκτίμηση του συντελεστή ευαισθησίας της δεύτερης παλινδρόμησης.

2.3.2 ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΠΟΛΛΑΠΛΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Τέλος, εδώ έχουμε της εξής περιπτώσεις:

α) Συνολικός έλεγχος του μοντέλου:

Υπόθεση μηδέν (H_0): $\beta_1 = \beta_2 = \beta_3 = ... = \beta_\kappa = 0$.

Εναλλακτική Υπόθεση (H_1): Τουλάχιστον ένας από τους συντελεστές παλινδρόμησης είναι $\neq 0$.

β) Έλεγχος σημαντικότητας μεμονωμένων συντελεστών της παλινδρόμησης:

$H_0: \beta_1 = 0$ $H_0: \beta_2 = 0$ $H_0: \beta_3 = 0$ και $H_0: \beta_\kappa = 0$

$H_1: \beta_1 \neq 0$ $H_1: \beta_2 \neq 0$ $H_1: \beta_3 \neq 0$ $H_1: \beta_\kappa \neq 0$.

2.4 ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Είναι μία τεχνική προσδιορισμού μίας ποσοτικής (μαθηματικής) έκφρασης, για την περιγραφή του τρόπου αλληλοσυσχέτισης μεταξύ δύο ή περισσοτέρων μεταβλητών.

$$y = \alpha + \beta x$$

Η παραπάνω σχέση αποτελεί το ‘‘γραμμικό μοντέλο’’ στο οποίο θεωρούμε ότι μεταξύ των x και y υπάρχει σχέση αιτίου ή αποτελέσματος, η οποία προσδιορίζεται με τον υπολογισμό των τιμών των σταθερών ποσοτήτων α και β. Η μεταβλητή x λέγεται **ανεξάρτητη** (αίτιο) και η y **εξαρτημένη** (αποτέλεσμα), διότι προσδιορίζεται έμμεσα βάσει των τιμών της x , από την σχέση:

$$y = \alpha + \beta x$$

Επίσης η χρησιμότητα της γραμμικής παλινδρόμησης είναι εξαιρετικά χρήσιμη στην Διοίκηση των Επιχειρήσεων γιατί υπάρχει πληθώρα καταστάσεων συμμεταβαλλομένων καταστάσεων, που μπορούν να μελετηθούν.

Π. χ. Οι πωλήσεις σχετίζονται άρα και συμμεταβάλλονται με τα πάγια και τα μεταβλητά έξοδα παραγωγής καθώς και με τα έξοδα διαφήμισης, τον αριθμό των πωλητών κτλ.

Επιπρόσθετα πρέπει να αναφέρουμε και τη **μελέτη** της γραμμικής παλινδρόμησης η οποία μπορεί να περιλάβει τις ακόλουθες ενότητες:

- Διάγραμμα διασποράς.
- Μέθοδος ελαχίστων τετραγώνων.
- Συντελεστής γραμμικής συσχέτισης του Pearson.
- Συντελεστής προσδιορισμού.
- Έλεγχος καταλληλότητας μοντέλου – ανάλυση υπολοίπων.

Όλα αυτά τα παραπάνω που αναφέρουμε συνοψίζονται στο επόμενο κεφάλαιο που θα ακολουθήσει.

Τέλος, υπάρχουν αρκετά αυστηρές **προϋποθέσεις** για να χρησιμοποιηθεί η γραμμική παλινδρόμηση που είναι τα εξής:

- **Γραμμικότητα:** Η εξαρτημένη μεταβλητή y πρέπει να συσχετίζεται κατά τρόπο γραμμικό με την ανεξάρτητη μεταβλητή x (ή με κάθε μία από τις ανεξάρτητες μεταβλητές, σε πολλαπλή παλινδρόμηση).

Αυτή ή απαίτηση, που εμφανίζεται πολύ περιοριστική, πολλές φορές μπορεί να παρακαμφθεί σε περιπτώσεις που μη γραμμικές συσχετίσεις μπορούν να μετασχηματισθούν σε γραμμικές (Π. χ. η σχέση $y = 3 + 6x^2$ μετασχηματίζεται σε γραμμική αν θέσουμε $x^2 = z$).

Το διάγραμμα διασποράς, είναι ιδιαίτερα βοηθητικό στην (οπτική) αναγνώριση της συσχέτισης των x και y.

- **Ομοσκεδασμός:** Ονομάζεται η απαίτηση να υπάρχει σταθερή διακύμανση μεταξύ των πραγματικών τιμών y_i και αυτών που υπολογίζονται από την παλινδρόμηση $\hat{y}_i$.
- **Ανεξαρτησία σφαλμάτων:** Τα κατάλοιπα, ή ανερμήνευτα σφάλματα της γραμμικής παλινδρόμησης (δηλαδή τα $y_i - \hat{y}_i$), πρέπει να είναι ανεξάρτητα μεταξύ τους. Σε αντίθετη περίπτωση λέμε ότι υπάρχει ‘‘αυτοσυσχέτιση’’ μεταξύ διαδοχικών τιμών καταλοίπων. Η ύπαρξη αυτοσυσχέτισης σημαίνει ότι έχει παραληφθεί κάποια σημαντική ανεξάρτητη μεταβλητή από τη συνάρτηση παλινδρόμησης (και έτσι τα σφάλματα δεν έχουν τυχαία διακύμανση).

- **Κανονική κατανομή σφαλμάτων:** Η μη ύπαρξη συσχέτισης στα σφάλματα δημιουργεί συνθήκες κανονικής κατανομής. Συνήθως, όταν έχουμε πάνω από 30 παρατηρήσεις υπάρχει μεγάλη πιθανότητα τα σφάλματα να ακολουθούν την κανονική κατανομή.
- **Ανυπαρξία Πολυσυγγραμμικότητας:** Η πολυσυγγραμμικότητα είναι ένα φαινόμενο, που συμβαίνει στην πολλαπλή γραμμική παλινδρόμηση, όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές έχουν μεταξύ τους υψηλό βαθμό συσχέτισης, με αποτέλεσμα να υπάρχει επίδραση στη διαμόρφωση των τιμών της εξαρτημένης μεταβλητής.

2.5 ΜΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η ύπαρξη γραμμικής σχέσης μεταξύ της μέσης τιμής Y και των τιμών της ανεξάρτητης μεταβλητής X (δηλαδή σχέση της μορφής $E(Y) = \alpha + \beta_1 x$), ελέγχεται γραφικά αν κατασκευάσουμε το διάγραμμα διασποράς των διαθέσιμων δεδομένων (x_i, y_i), $i = 1,2,...,\nu$ και παρατηρήσουμε κατά πόσο τα απεικονιζόμενα σημεία βρίσκονται «περίπου» γύρω από μια ευθεία. Αν διαπιστώσουμε ότι η μεταβλητή απόκρισης Y δεν προσεγγίζεται ικανοποιητικά μέσω ενός γραμμικού συνδυασμού της ανεξάρτητης μεταβλητής τότε προσπαθούμε, είτε να εκμεταλλευτούμε τη φυσική περιγραφή του προβλήματος για τον προσδιορισμό της σχέσης των X και Y, ή να παρατηρήσουμε το διάγραμμα διασποράς, ώστε να αναγνωρίσουμε τη γραφική παράσταση κάποιας γνωστής μαθηματικής συνάρτησης την οποία χρησιμοποιούμε στη συνέχεια για να προσαρμόσουμε τα δεδομένα μας.

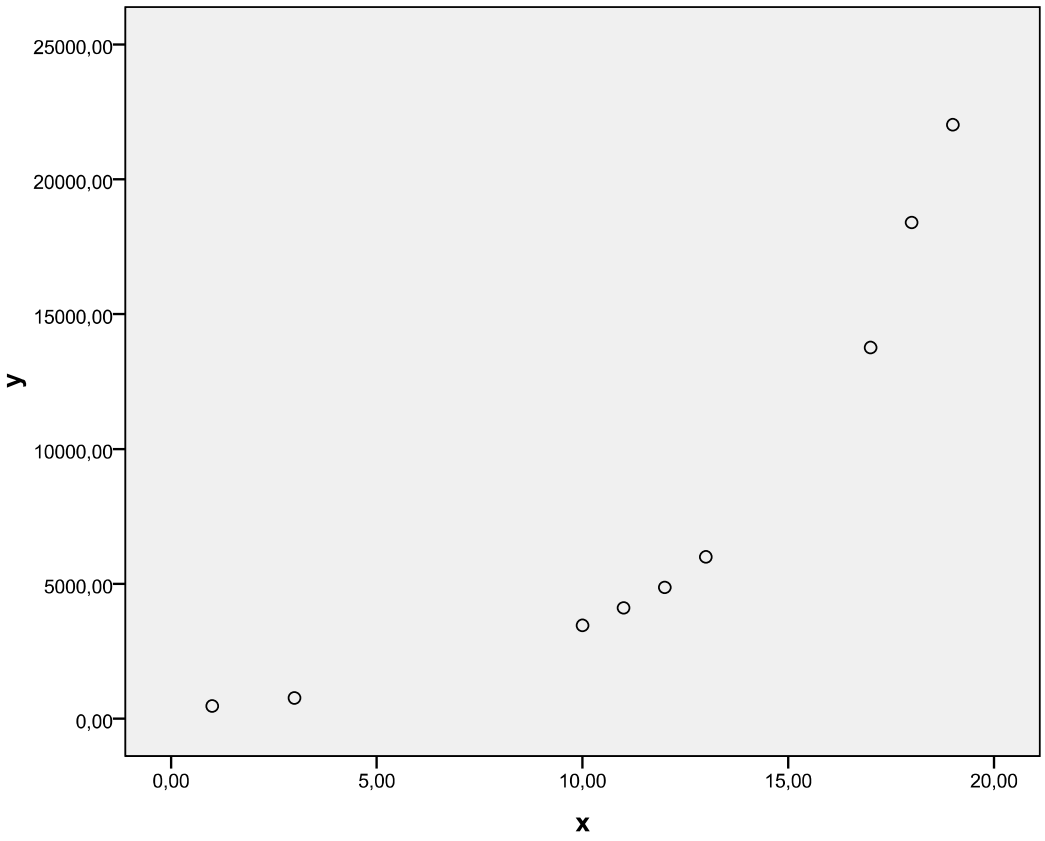
ΠΑΡΑΔΕΙΓΜΑ: Ο παρακάτω πίνακας δίνει τον πληθυσμό μιας περιοχής για διάστημα μιας εικοσαετίας. Στον πίνακα έχει καταγραφεί η ανεξάρτητη μεταβλητή X η οποία μετράει έτη, χρησιμοποιώντας ως έτος αναφοράς το 1990 (η τιμή $x_1 = 1$ αντιστοιχεί στο έτος 1990 + 1 = 1991 κ.ο.κ.) και η εξαρτημένη μεταβλητή Y που δίνει τον πληθυσμό της περιοχής στο αντίστοιχο έτος.

ΠΙΝΑΚΑΣ 1

i	1	2	3	4	5	6	7	8	9
Χρόνος (X)	1	3	10	11	12	13	17	18	19
Πληθυσμός (Y)	470	770	3460	4110	4870	6000	13760	18400	22020

ΛΥΣΗ:

ΔΙΑΓΡΑΜΜΑ 1



Τα σημεία που απεικονίζονται στο διάγραμμα διασποράς φαίνεται να μην προσεγγίζονται ικανοποιητικά από μια ευθεία, οπότε η προσαρμογή ενός γραμμικού μοντέλου της μορφής $y_i = \alpha + \beta_1 x_i$ δεν αναμένεται να είναι καλή.

ΠΙΝΑΚΑΣ 2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4548,011	2905,252		-1,565	,161
	x	1103,770	223,702	,881	4,934	,002

a. Dependent Variable: y

Από τον παραπάνω πίνακα βλέπουμε ότι η εξίσωση παλινδρόμησης είναι:

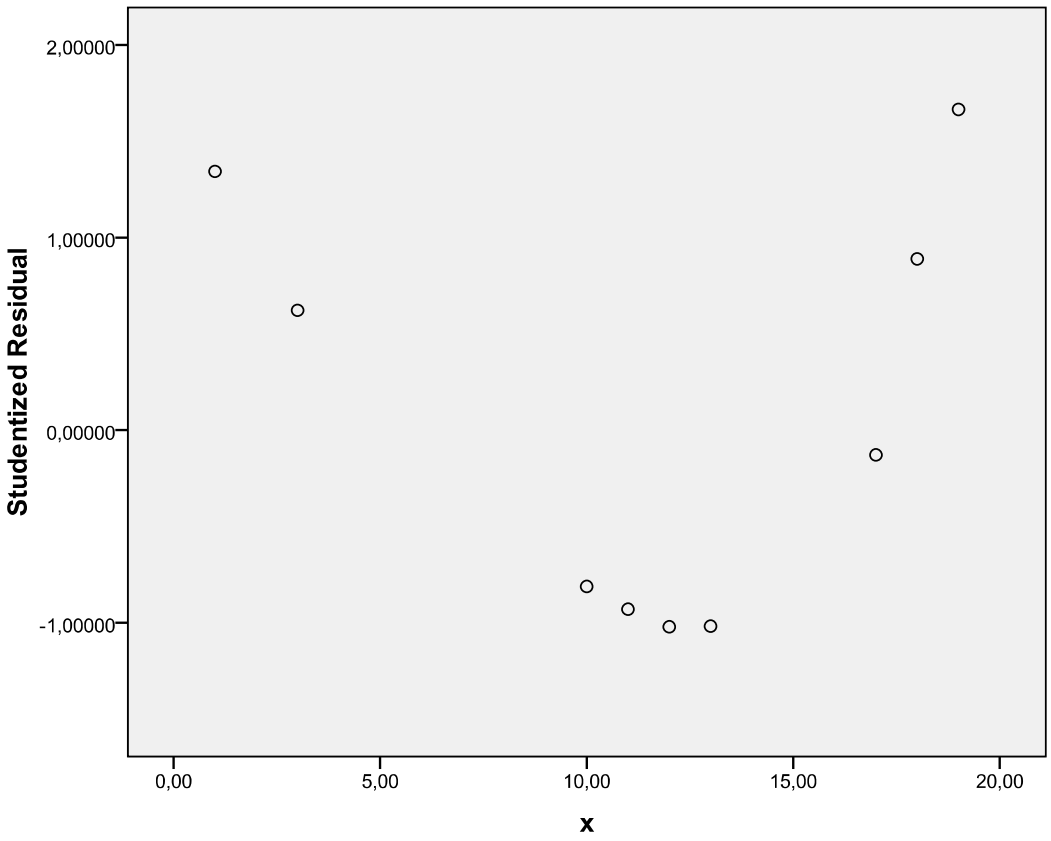
$$\hat{y} = -4548{,}011 + 1103{,}770x$$

Στην συνέχεια κατασκευάζουμε τον παρακάτω πίνακα για να βρούμε την διαφορά των σφαλμάτων που υπάρχει μεταξύ της εξαρτημένης μεταβλητής Υ και της ανεξάρτητης μεταβλητής x.

ΠΙΝΑΚΑΣ 3

x_i	y_i	$\hat{\varepsilon}_i^*$
1	470	1,34332
3	770	0,62215
10	3460	–0,81131
11	4110	–0,92931
12	4870	–1,02082
13	6000	–1,01725
17	13760	–0,12857
18	18400	0,88961
19	22020	1,66534

ΔΙΑΓΡΑΜΜΑ 3



Η μη καταλληλότητα του απλού γραμμικού μοντέλου για τα δεδομένα του παραδείγματος, φαίνεται ξεκάθαρα από το παραπάνω **διάγραμμα υπολοίπων.**

Από την επιστήμη της Δημογραφίας είναι γνωστό ότι, ένα κατάλληλο μοντέλο για την περιγραφή της διαχρονικής εξέλιξης των πληθυσμών είναι το λεγόμενο **εκθετικό μοντέλο,** σύμφωνα με το οποίο το μέγεθος Υ του πληθυσμού μετά από πάροδο Χ ετών (σε σχέση με ένα συγκεκριμένο έτος αναφοράς) δίνεται προσεκτικά από τον τύπο:

$$Y \cong \mathrm{A} \cdot e^{\beta \cdot x}.$$

Επομένως, αυτό που ήδη παρατηρήσαμε, δηλαδή ότι η σχέση μεταξύ της μεταβλητής απόκρισης Υ (μέγεθος πληθυσμού) και της ανεξάρτητης μεταβλητής Χ (χρονική απόκλιση από το έτος αναφοράς) δεν είναι γραμμική, επιβεβαιώνεται και από την Δημογραφική θεωρία.

Αν λογαριθμίσουμε την προσεγγιστική σχέση $Y \cong \mathrm{A} \cdot e^{\beta \cdot x}$ θα πάρουμε

$$\ln Y \cong \ln \mathrm{A} + \beta \cdot x$$

και μπορούμε να γράψουμε την τελευταία στη μορφή:

$$Y' = \alpha + \beta_1 \cdot x' + \varepsilon$$

όπου

$$Y' = \ln Y,\ \ x' = x,\ \ \alpha = \ln \mathrm{A},\ \text{και}\ \ \beta_1 = \beta\ .$$

Ο νέος όρος που προστέθηκε, δίνει την απόκλιση μεταξύ του $Y' = \ln Y$ και του γραμμικού τμήματος $\alpha + \beta_1 \cdot x' = \ln \mathrm{A} + \beta \cdot x$ ώστε να αποκατασταθεί η ισότητα μεταξύ των Y' και $\alpha + \beta_1 \cdot x'$. Έτσι φτάνουμε σε ένα πρότυπο όμοιο με το απλό γραμμικό μοντέλο και μπορούμε να το μελετήσουμε με τις τεχνικές που έχουν αναπτυχθεί για αυτό.

Η ανάλυση του μοντέλου γίνεται χρησιμοποιώντας, αντί των αρχικών δεδομένων x_i, y_i, $i = 1,2,...,\nu$ τα λεγόμενα **μετασχηματισμένα** δεδομένα x_i', y_i', $i = 1,2,...,\nu$ τα οποία προκύπτουν τροποποιώντας τα αρχικά σύμφωνα με τους τύπους που μας βοήθησαν να καταλήξουμε στο γραμμικό μοντέλο.

ΚΕΦΑΛΑΙΟ 3 ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

3.1 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Η απλούστερη περίπτωση παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση (simple linear regression), κατά την οποία υπάρχει μόνο μια **ανεξάρτητη μεταβλητή** X (independent or input variable), και η εξαρτημένη μεταβλητή Y (dependent or response variable), η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του X.

Η εκτίμηση των τιμών της μεταβλητής Y από τις τιμές της X, δια μέσου του υποδείγματος της απλής γραμμικής παλινδρόμησης, μπορεί να γίνει όταν διασφαλίζονται οι εξής προϋποθέσεις:

1. Ο προσδιορισμός των τιμών της μεταβλητής X γίνεται χωρίς σφάλμα. Επειδή στην πραγματικότητα καμία συνεχής μέτρηση δεν είναι απαλλαγμένη σφαλμάτων, η παραδοχή αυτή υπονοεί ότι το μέγεθος του σφάλματος κατά τη μέτρηση της X είναι αμελητέο.

2. Σε κάθε τιμή της X αντιστοιχεί ένας υπο-πληθυσμός τιμών της Y, ο οποίος ακολουθεί την κανονική κατανομή.

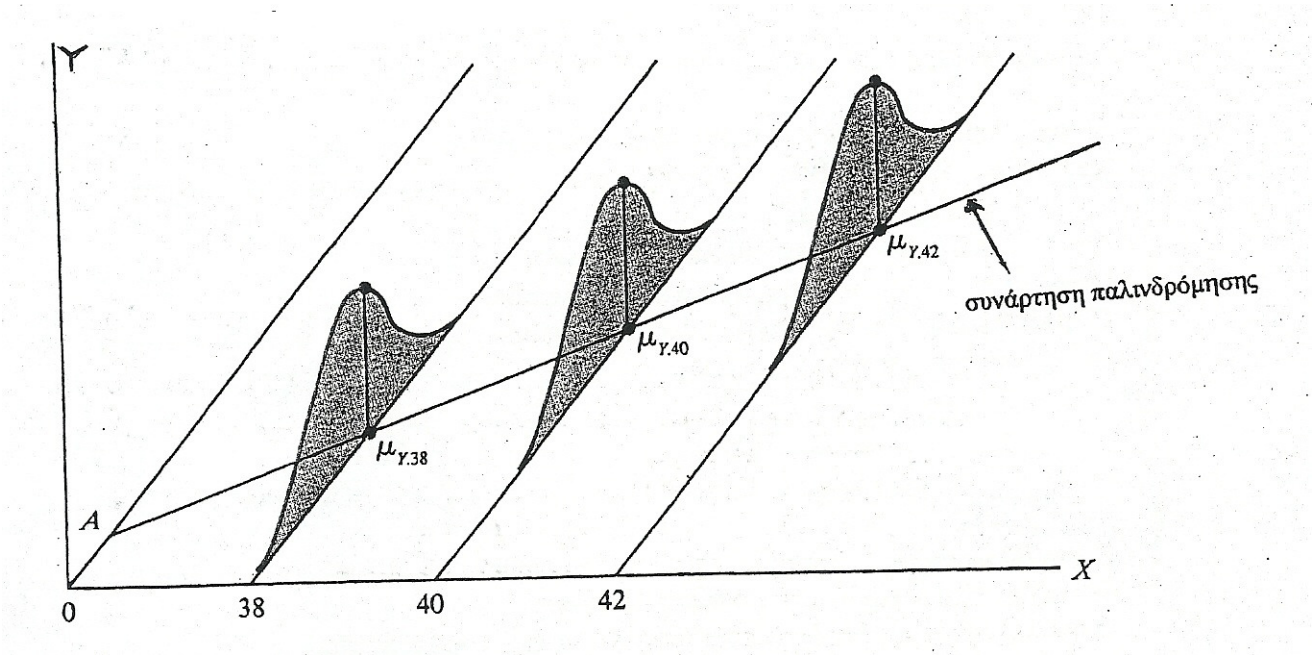
3. Οι διακυμάνσεις των υπο-πληθυσμών της Y που ορίζονται για τις διάφορες τιμές της X, είναι ίσες. Η κοινή διακύμανση των υπο-πληθυσμών της Y συμβολίζεται με $\sigma^2_{y|x}$. Η παραδοχή της ισότητας των διακυμάνσεων των τιμών της Y, ονομάζεται **ομοσκεδαστικότητα** (homoscedasticity) και είναι ανάλογη με την παραδοχή της ισότητας των διακυμάνσεων.

4. Οι μέσες τιμές των υπο-πληθυσμών της Y συνδέονται με τις αντίστοιχες τιμές της X, δια μέσου μιας γραμμικής σχέσης της μορφής:

$$\mu_{y|x} = \alpha + \beta x ,$$

όπου $\mu_{y|x}$ είναι η μέση τιμή του υπο-πληθυσμού της Y που αντιστοιχεί σε μια συγκεκριμένη τιμή x της μεταβλητής X. Οι ποσότητες α και β ονομάζονται **πληθυσμιακοί συντελεστές της παλινδρόμησης**. Το παραπάνω υπόδειγμα ορίζει μια ευθεία γραμμή, επί της ουσίας είναι τοποθετημένες οι μέσες τιμές $\mu_{y|x}$ των διάφορων υπο-πληθυσμών της Y. Η ευθεία αυτή ονομάζεται **πληθυσμιακή ευθεία της παλινδρόμησης**. Γεωμετρικά οι συντελεστές α και β αναπαριστούν αντίστοιχα την τεταγμένη στο σημείο 0 και την κλίση της ευθείας της παλινδρόμησης (σχήμα 1).

ΣΧΗΜΑ 1



5. Οι τιμές της Y είναι ανεξάρτητες η μία της άλλης.

Όλες οι προηγούμενες προϋποθέσεις συνοψίζονται στην παρακάτω εξίσωση, η οποία ονομάζεται **υπόδειγμα της απλής γραμμικής παλινδρόμησης:**

$$y = \alpha + \beta x + \varepsilon,$$

όπου y είναι μια οποιαδήποτε τιμή του υπο-πληθυσμού των τιμών της Y που αντιστοιχεί στην τιμή x, και α, β είναι οι πληθυσμιακοί συντελεστές της παλινδρόμησης.

Αν επιλύσουμε την προηγούμενη εξίσωση ως προς ε, έχουμε:

$$\varepsilon = y - (\alpha + \beta x) = y - \mu_{y|x}.$$

Η ποσότητα ε, η οποία ονομάζεται **σφάλμα**, εκφράζει τη διαφοροποίηση της y από τη μέση τιμή του υπο-πληθυσμού της Y στον οποίο αυτή ανήκει, ή, αλλιώς εκφράζει την απόκλιση[9] της y από την ευθεία της παλινδρόμησης. Ως συνέπεια της παραδοχής ότι οι διάφοροι υπο-πληθυσμοί της Y ακολουθούν κανονική κατανομή με κοινή διακύμανση, οι ποσότητες ε για κάθε τιμή X ακολουθούν επίσης κανονική κατανομή με διακύμανση ίση με την κοινή διακύμανση $\sigma^2_{y|x}$ των αντίστοιχων υπο-πληθυσμών της Y.

Επιπλέον, από τον ορισμό των σφαλμάτων προκύπτει ότι η μέση τιμή τους είναι ίση με 0.

Επομένως οι κυριότερες βασικές έννοιες που θα ασχοληθούμε σε αυτό το κεφάλαιο και οι πιο σημαντικές που αναλύονται παρακάτω είναι οι έξης:

---

[9] απόκλιση: διαφοροποίηση ως προς αυτό που θεωρείται δεδομένο

Το διάγραμμα διασποράς, η μέθοδος ελαχίστων τετραγώνων, ο συντελεστής γραμμικής συσχέτισης του Pearson και ο συντελεστής προσδιορισμού.

3.2 ΔΙΑΓΡΑΜΜΑ ΔΙΑΣΠΟΡΑΣ

Ο προσδιορισμός της σχέσης μεταξύ δύο συνεχών μεταβλητών ονομάζεται **διάγραμμα διασποράς** (scatter plots). Τα διαγράμματα αυτά μπορούν να κατασκευαστούν για κάθε ζεύγος συνεχών τυχαίων μεταβλητών οι οποίες υπολογίζονται στο ίδιο σύνολο παρατηρήσεων.

ΠΑΡΑΔΕΙΓΜΑ: Ο παρακάτω πίνακας 1 δίνει τα ύψη Χ (σε cm) και τα βάρη Υ (σε kg) των 18 μαθητών. Οι τιμές του ύψους δίνονται σε αύξουσα σειρά.

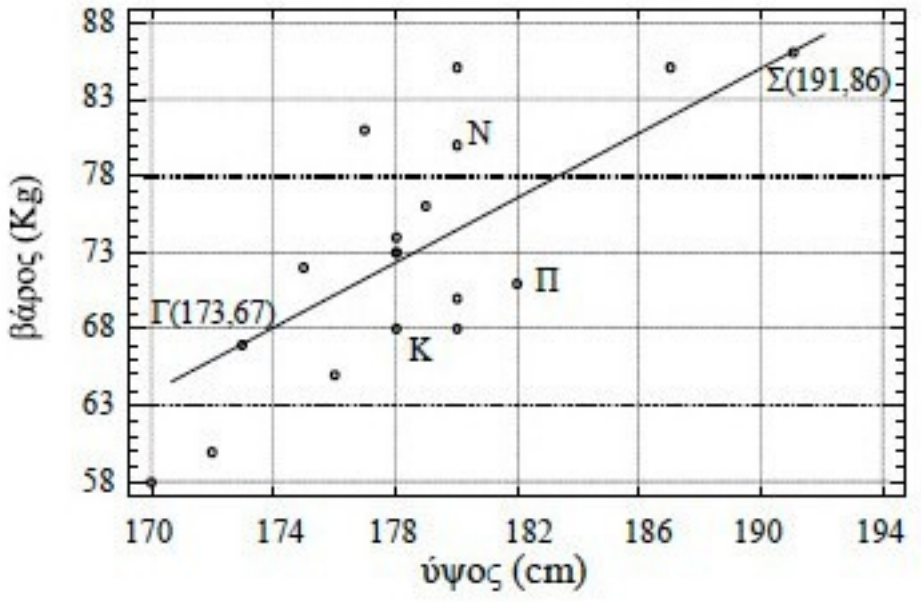
ΠΙΝΑΚΑΣ 1

ΜΑΘΗΤΗΣ	ΥΨΟΣ Χ	ΒΑΡΟΣ Υ
Α	170	58
Β	172	60
Γ	173	67
Δ	175	72
Ε	176	65
Ζ	177	81
Η	178	73
Θ	178	74
Ι	178	73
Κ	178	68

Λ	179	76
Μ	180	68
Ν	180	80
Ξ	180	70
Ο	180	85
Π	182	71
Ρ	187	85
Σ	191	86

Στο παράδειγμα αυτό έχουμε την περίπτωση όπου σε κάθε άτομο (μαθητή) γίνονται δύο μετρήσεις. Δηλαδή το δείγμα αποτελείται από τα ζεύγη τιμών των συνεχών μεταβλητών Χ (ύψος) και Υ (βάρος).
Αν παραστήσουμε τα ζεύγη (x, y) των παρατηρήσεων σε ένα σύστημα ορθογώνιων αξόνων, παρατηρούμε ότι προκύπτει μία "διασπορά" των σημείων που αντιστοιχούν στους μαθητές που εξετάζουμε. Η παράσταση αυτή των σημείων καλείται **διάγραμμα διασποράς**.

ΔΙΑΓΡΑΜΜΑ 1



Στο παράδειγμα αυτό το διάγραμμα διασποράς δείχνει, γενικά, ότι οι ψηλοί μαθητές είναι συνήθως και πιο βαρείς. Για παράδειγμα, ο Ν είναι ψηλότερος και βαρύτερος από τον Κ, ο Π είναι ψηλότερος και βαρύτερος από τον Κ, αλλά υπάρχουν και εξαιρέσεις, όπως ο Π είναι ψηλότερος από τον Ν αλλά ο Ν είναι βαρύτερος από τον Π.

Τέλος από το διάγραμμα διασποράς του παραδείγματος φαίνεται καθαρά ότι υπάρχει μία σχέση ανάμεσα στο ύψος Χ και το βάρος Υ των 18 μαθητών. Τα σημεία (x, y) είναι συγκεντρωμένα περίπου γύρω από μία ευθεία, δηλαδή η σχέση μεταξύ των Χ και Υ είναι κατά προσέγγιση γραμμική. Από τον ορισμό της απλής γραμμική παλινδρόμησης, μπορούμε να θεωρήσουμε τη μία μεταβλητή ως ανεξάρτητη μεταβλητή και την άλλη ως εξαρτημένη. Εδώ θεωρούμε ως ανεξάρτητη μεταβλητή το ύψος Χ και ως εξαρτημένη μεταβλητή το βάρος Υ, οπότε η ευθεία που θα προσαρμόζεται καλύτερα στα σημεία αυτά καλείται **ευθεία παλινδρόμησης** της Υ πάνω στη Χ.

Οπότε η εξίσωση μιας ευθείας δίνεται από τη σχέση:

$$y = \alpha + \beta x \qquad (1)$$

όπου α και β είναι παράμετροι τις οποίες θέλουμε να υπολογίσουμε, έτσι ώστε η ευθεία που θα προκύψει να μας δίνει όσο το δυνατόν την καλύτερη περιγραφή της σχέσης (εξάρτησης) που υπάρχει μεταξύ των μεταβλητών Χ και Υ.

Η παράμετρος α μας δίνει το σημείο $(0,\alpha)$, όπου η ευθεία αυτή τέμνει τον άξονα $y'y$ ενώ η παράμετρος β παριστάνει το συντελεστή διεύθυνσης της ευθείας.

Για να βρούμε τα α και β, εργαζόμαστε ως εξής:

- Επιλέγουμε δύο σημεία, έστω τα Γ(173, 67) και Σ(191, 86)
- Αντικαθιστούμε τις συντεταγμένες (x, y) των σημείων αυτών στην (1), οπότε προκύπτει το σύστημα:

$$\begin{cases} y_1 = \alpha + \beta\chi_1 \\ y_2 = \alpha + \beta\chi_2 \end{cases} \Leftrightarrow \begin{cases} 67 = \alpha + 173\beta \\ 86 = \alpha + 191\beta \end{cases}$$

- Επιλύοντας το σύστημα αυτό βρίσκουμε $\alpha = -115{,}6$ και $\beta = 1{,}06$ οπότε η εξίσωση της ευθείας (1) γίνεται:

$$y = -115{,}6 + 1{,}06x\,. \qquad (2)$$

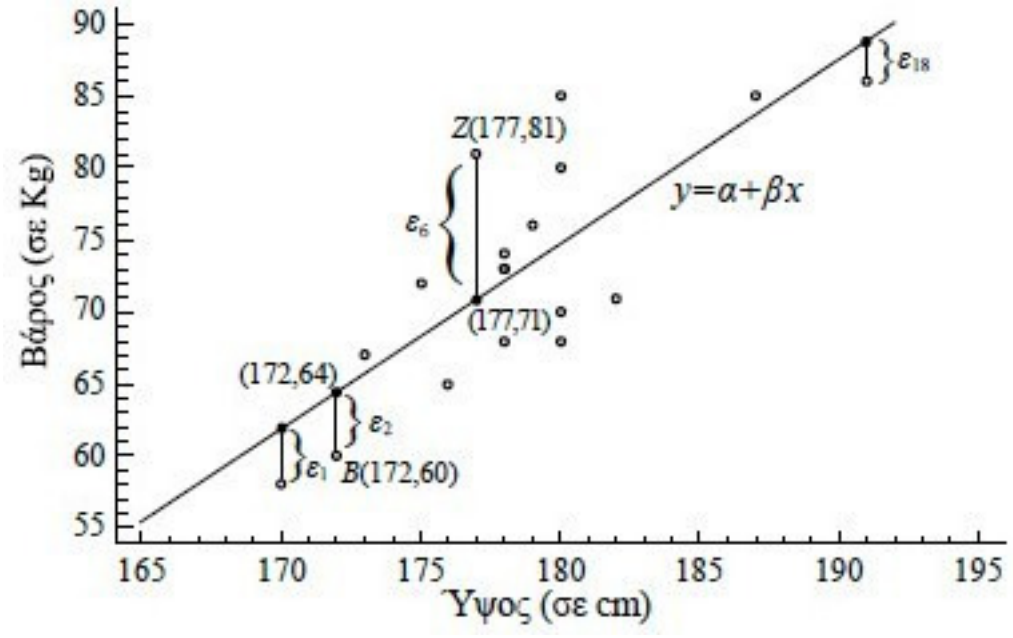
Επομένως, η ευθεία που προσαρμόζεται στα σημεία του διαγράμματος διασποράς διέρχεται από το σημείο $(0,\ -115{,}6)$ και έχει συντελεστή διεύθυνσης 1,06.

3.3 ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Μια μέθοδος που χρησιμοποιείται για την εκτίμηση των παραμέτρων α και β, άρα και για την εύρεση της εξίσωσης της καλύτερης ευθείας είναι η **μέθοδος ελαχίστων τετραγώνων**. Ο λόγος για τον οποίο χρησιμοποιείται η συγκεκριμένη μέθοδος, προκύπτει από την παρακάτω διαδικασία προσδιορισμού της ευθείας.

Έστω το διάγραμμα διασποράς του προηγουμένου παραδείγματος για τα ύψη X και τα βάρη Y των 18 μαθητών. Στο διάγραμμα αυτό έχουμε μια ευθεία $y = \alpha + \beta x$, που προσαρμόζεται καλύτερα στα σημεία (x_i, y_i) για τις ν = 18 συνολικά μετρήσεις των μεταβλητών X και Y.

ΔΙΑΓΡΑΜΜΑ 2



Προσαρμογή ευθείας ελαχίστων τετραγώνων στο διάγραμμα διασποράς των δεδομένων του πίνακα 1.

Έτσι, για παράδειγμα, για το μαθητή B, σημείο B(172, 60), με ύψος $x_2 = 172\,\text{cm}$ έχουμε βρει, όπως φαίνεται στον πίνακα 1, βάρος $y_2 = 60kg$, ενώ, σύμφωνα με την ευθεία που έχουμε φέρει, το βάρος του αναμένεται να είναι (περίπου) $64kg$, έχουμε δηλαδή ένα σφάλμα $\varepsilon_2 = 60 - 64 = -4$, δηλαδή βάρος $4kg$ λιγότερο από το αναμενόμενο. Ομοίως για το μαθητή Z, σημείο Z(177, 81), το βάρος του που μετρήθηκε ήταν $y_6 = 81kg$, ενώ το

αναμενόμενο βάρος του σύμφωνα με την ευθεία που έχουμε φέρει είναι $71kg$, έχουμε δηλαδή ένα σφάλμα $\varepsilon_6 = 81 - 71 = 10$, δηλαδή βάρος $10kg$ περισσότερο από το αναμενόμενο. Ανάλογα σφάλματα υπολογίζονται και για τους άλλους μαθητές. Θα θέλαμε λοιπόν να βρούμε με κάποια μέθοδο εκείνη την ευθεία $y = \alpha + \beta x$, έτσι ώστε τα σφάλματα που προκύπτουν να είναι όσο το δυνατόν μικρότερα.

Η μέθοδος των ελαχίστων τετραγώνων συνίσταται στον προσδιορισμό των παραμέτρων α, β , έτσι ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων (x_i, y_i) από την ευθεία $y = \alpha + \beta x$, δηλαδή το

$$\sum_{i=1}^{\nu} \varepsilon_i{}^2 = \sum_{i=1}^{\nu} (y_i - \alpha - \beta x_i)^2 \qquad (4)$$

να γίνεται ελάχιστο.

Οι τιμές των παραμέτρων α και β, που ελαχιστοποιούν την (4), καλούνται **εκτιμήτριες ελαχίστων τετραγώνων** (least square estimators), συμβολίζονται με $\hat{a}$ (‘‘α καπέλο’’) και $\hat{\beta}$ (‘‘β καπέλο’’), αντιστοίχως και αποδεικνύεται ότι δίνονται από τις σχέσεις:

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

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad (5)$$

όπου

$$\bar{y} = \frac{1}{\nu}\sum_{i=1}^{\nu} y_i \,, \qquad \bar{x} = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i$$

Η ευθεία

$$\hat{y} = \hat{\alpha} + \hat{\beta} x \qquad (6)$$

καλείται **ευθεία ελαχίστων τετραγώνων** ή **ευθεία παλινδρόμησης** της Y (πάνω) στη X. Αντικαθιστώντας το $\hat{a} = \bar{y} - \hat{\beta}\,\bar{x}$ στη σχέση (6) βρίσκουμε την

$$\hat{y} - \bar{y} = \hat{\beta}(x - \bar{x}) \,,$$

η οποία φανερώνει ότι η ευθεία ελαχίστων τετραγώνων $\hat{y} = \hat{\alpha} + \hat{\beta} x$ διέρχεται από το σημείο με συντεταγμένες ($\bar{x}$, $\bar{y}$) και έχει συντελεστή διεύθυνσης το $\hat{\beta}$. Αντικαθιστώντας τις τιμές x_i και y_i από τον πίνακα 1 στις σχέσεις (5) βρίσκουμε:

$$\hat{\beta} = 1{,}28 \text{ και } \hat{a} = -156{,}1$$

οπότε η ευθεία ελαχίστων τετραγώνων που προσαρμόζεται καλύτερα στα δεδομένα είναι από τη σχέση (6), η

$$\hat{y} = -156{,}1 + 1{,}28x\,.$$

Δηλαδή παρατηρούμε ότι υπάρχει σημαντική διαφορά από την ευθεία

$y = -115{,}6 + 1{,}06x$.

ΕΡΜΗΝΕΙΑ ΤΩΝ ΕΚΤΙΜΗΤΡΙΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Στην εξίσωση ελαχίστων τετραγώνων $\hat{y} = \hat{\alpha} + \hat{\beta}x$ η τιμή της εκτιμήτριας $\hat{\alpha}$ της παραμέτρου α παριστάνει την τεταγμένη του σημείου στο οποίο η ευθεία τέμνει τον άξονα $y'y$, δηλαδή την τιμή της εξαρτημένης μεταβλητής Υ όταν $x = 0$. Όταν το $\hat{\alpha} = 0$ τότε η ευθεία διέρχεται από την αρχή των αξόνων.

Έστω δύο τιμές x_1 και $x_2 = x_1 + 1$ της ανεξάρτητης μεταβλητής. Τότε λαμβάνοντας τη διαφορά των αντίστοιχων προβλεπόμενων τιμών της εξαρτημένης μεταβλητής έχουμε:

$$\hat{y}_2 - \hat{y}_1 = \left(\hat{\alpha} + \hat{\beta}x_2\right) - \left(\hat{\alpha} + \hat{\beta}x_1\right) = \hat{\alpha} + \hat{\beta}(x_1 + 1) - \left(\hat{\alpha} + \hat{\beta}x_1\right) = \hat{\beta}$$

δηλαδή $\hat{y}_2 = \hat{y}_1 + \hat{\beta}$. Συνεπώς ο συντελεστής διεύθυνσης $\hat{\beta}$ της ευθείας $\hat{y} = \hat{\alpha} + \hat{\beta}x$ παριστά τη μεταβολή της εξαρτημένης μεταβλητής Υ όταν το Χ μεταβληθεί κατά μια μονάδα. Έτσι, όταν το x αυξηθεί κατά μια μονάδα τότε το $\hat{y}$ αυξάνεται κατά $\hat{\beta}$ μονάδες όταν $\hat{\beta} > 0$ ή ελαττώνεται κατά $\hat{\beta}$ μονάδες όταν $\hat{\beta} < 0$.

ΠΑΡΑΔΕΙΓΜΑ ΓΙΑ ΤΗΝ ΕΥΘΕΙΑ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Ο παρακάτω πίνακας δίνει την ζήτηση ενός προϊόντος (Y) για διάφορα επίπεδα διαφημιστικής δαπάνης (X).

ΠΙΝΑΚΑΣ 2

X (σε χιλιάδες ευρώ)	Y (σε χιλιάδες τεμάχια)
15	5
13	6
11	8
9	10
9	9
6	12
5	15
4	11

α) Να βρεθεί η ευθεία ελαχίστων τετραγώνων $\hat{y} = \hat{\alpha} + \hat{\beta}\, x$ και να χαραχτεί το αντίστοιχο διάγραμμα διασποράς.

β) Να ερμηνευθεί η έννοια των εκτιμήτριων $\hat{a}$ και $\hat{\beta}$.

ΛΥΣΗ:

α) Για τον προσδιορισμό της εξίσωσης της ευθείας ελάχιστων τετραγώνων κατασκευάζουμε τον παρακάτω πίνακα με τις απαραίτητες πράξεις.

Επομένως, έχουμε:

ΠΙΝΑΚΑΣ 3

x	y	x^2	xy
15	5	225	75
13	6	169	78
11	8	121	88
9	10	81	90
9	9	81	81
6	12	36	72
5	15	25	75
4	11	16	44
$\Sigma x = 72$	$\Sigma y = 76$	$\Sigma x^2 = 754$	$\Sigma xy = 603$

$\nu = 8$

$$\bar{x} = \frac{\Sigma x}{\nu} = \frac{72}{8} = 9$$

$$\bar{y} = \frac{\Sigma y}{\nu} = \frac{76}{8} = 9{,}5$$

$$\hat{\beta} = \frac{\nu\Sigma xy - (\Sigma x)(\Sigma y)}{\nu\Sigma x^2 - (\Sigma x)^2} = \frac{8(603) - (72)(76)}{8(754) - (72)^2} =$$

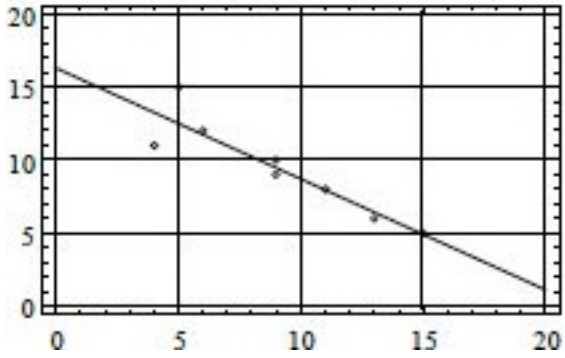
$$= \frac{-648}{848} = -0{,}76$$

$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = 9{,}5 + (0{,}76)\cdot(9) = 16{,}34.$

Άρα, η ευθεία ελαχίστων τετραγώνων είναι $\hat{y} = 16{,}34 - 0{,}76x$.

Επειδή γνωρίζουμε ότι η ευθεία που διέρχεται από τα σημεία $(0, \hat{\alpha})$ και $(\bar{x}, \bar{y})$, είναι εύκολο να χαραχτεί στο διάγραμμα διασποράς, όπως φαίνεται παρακάτω:

ΔΙΑΓΡΑΜΜΑ 3



β) Το $\hat{\alpha} = 16{,}34$ προσδιορίζει την προβλεπόμενη ζήτηση του προϊόντος, όταν η τιμή του είναι 0€. Προφανώς εδώ τέτοια περίπτωση δεν είναι ρεαλιστική.

Το $\hat{\beta}$ προσδιορίζει τη μεταβολή που επέρχεται στην εξαρτημένη μεταβλητή Y, όταν η X μεταβληθεί κατά μία μονάδα. Δηλαδή, αν οι διαφημιστικές δαπάνες αυξηθούν κατά 1000€ τότε η ζήτηση του προϊόντος θα μειωθεί κατά 760 τεμάχια.

3.4 ΣΥΝΤΕΛΕΣΤΗΣ ΓΡΑΜΜΙΚΗΣ ΣΥΣΧΕΤΙΣΗΣ ΤΟΥ PEARSON

Έστω X και Y ένα ζεύγος συνεχών τυχαίων μεταβλητών με μέσες τιμές μ_X και μ_Y και διακυμάνσεις $\sigma^2{}_X$ και $\sigma^2{}_Y$, οι οποίες σχετίζονται μεταξύ τους με τρόπο γραμμικό. Ως μέτρο της υφιστάμενης σχέσης δύο μεταβλητών ορίζεται η ποσότητα:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)],$$

δηλαδή η αναμενόμενη τιμή του γινομένου των διάφορων $(X-\mu_X)$ και $(Y-\mu_Y)$. Η ποσότητα αυτή ονομάζεται συνδιακύμανση (covariance) των X και Y.

Αν υπάρχει θετική συσχέτιση μεταξύ των δύο τυχαίων μεταβλητών, δηλαδή αν οι υψηλές τιμές της X τείνουν να εμφανίζονται με υψηλές τιμές της Y και οι χαμηλές τιμές της X να εμφανίζονται με χαμηλές τιμές της Y, τότε η συνδιακύμανση είναι θετική.

Αν υπάρχει αρνητική σχέση μεταξύ των δύο μεταβλητών, δηλαδή αν οι υψηλές τιμές της X τείνουν να εμφανίζονται με χαμηλές τιμές της Y και οι χαμηλές τιμές της X με υψηλές τιμές

της Y, τότε η συνδιακύμανση είναι αρνητική. Αν δεν υπάρχει γραμμική σχέση μεταξύ των X και Y, τότε η συνδιακύμανση τους είναι 0.

Η συνδιακύμανση εξαρτάται από τις μονάδες των δύο συγκρινόμενων μεταβλητών και επομένως, αποτελεί απόλυτο μέτρο του βαθμού της συσχέτισης που υπάρχει μεταξύ τους. Προκειμένου να οριστεί ένα μέτρο ανεξάρτητο μονάδων, η συνδιακύμαση διαιρείται με το γινόμενο των δύο τυπικών αποκλίσεων σ_X και σ_Y. Η ποσότητα που προκύπτει με αυτόν τον τρόπο ονομάζεται **συντελεστής συσχέτισης** (correlation coefficient) και συμβολίζεται με r(X , Y) ή απλά με r οπότε:

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}.$$

Ο παραπάνω ορισμός του συντελεστή συσχέτισης αφορά τη σχέση δύο μεταβλητών που ορίζονται σε πληθυσμιακό επίπεδο. Ένα τυχαίο δείγμα ν ζευγών παρατηρήσεων (x_i, y_i) $i = 1,2,...,\nu$, προερχόμενο από τον αντίστοιχο πληθυσμό, μπορεί να χρησιμοποιηθεί για την εκτίμηση του πληθυσμιακού συντελεστή συσχέτισης.

Εκτίμηση του πληθυσμιακού συντελεστή είναι η ποσότητα:

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

η οποία ονομάζεται συντελεστής συσχέτισης του Pearson.

Ο r μετράει την ένταση της εξάρτησης μεταξύ των μεταβλητών X και Y με την προϋπόθεση ότι η σχέση εξάρτησης είναι γραμμικής μορφής, συνεπώς ο r δεν είναι κατάλληλο στατιστικό μέτρο συσχέτισης όταν η σχέση εξάρτησης είναι καμπυλόγραμμη. Συνεπώς ο r έχει τις ακόλουθες χαρακτηριστικές ιδιότητες:

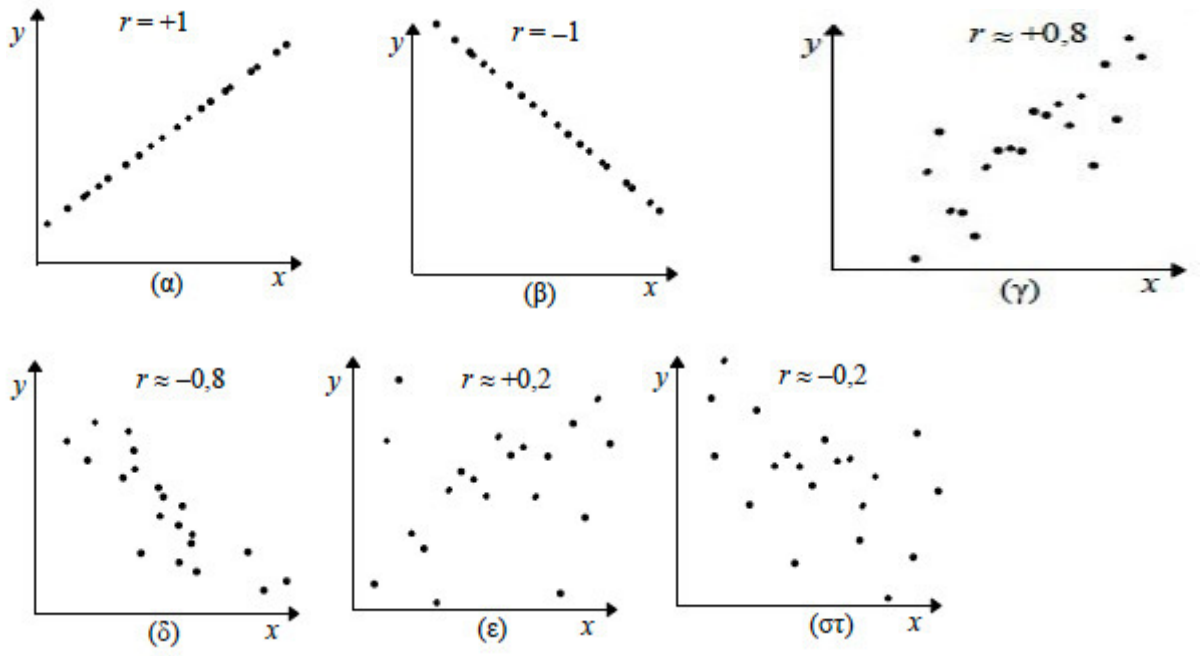
- Είναι ένα στατιστικό μέτρο ανεξάρτητο των μονάδων μέτρησης των μεταβλητών X και Y δηλαδή αν οι τιμές των μεταβλητών X και Y πολλαπλασιαστούν ή διαιρεθούν με μια σταθερή ποσότητα λ τότε ο ν παραμένει αμετάβλητος.
- Η τιμή του συντελεστή συσχέτισης κυμαίνεται ανάμεσα στο -1 και στο +1 δηλαδή $-1 \leq r \leq 1$.

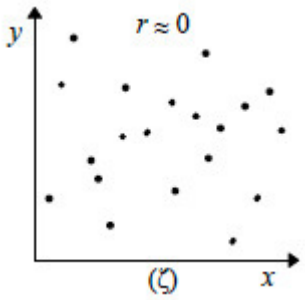
Πιο συγκεκριμένα όταν:

- Αν ο r είναι **θετικός** (r > 0) τότε η συσχέτιση μεταξύ των X και Y είναι θετική δηλαδή σε κάθε αύξηση (μείωση) της μιας μεταβλητής αντιστοιχεί αύξηση (μείωση) και της άλλης μεταβλητής.

- Αν ο r < 0 τότε η συσχέτιση μεταξύ των X και Y είναι **αρνητική** δηλαδή σε κάθε αύξηση της μιας μεταβλητής αντιστοιχεί μείωση της άλλης μεταβλητής και αντίστροφα.
- Αν r = +1 τότε έχουμε **τέλεια θετική γραμμική συσχέτιση** δηλαδή σε κάθε μεταβολή της μιας μεταβλητής κατά ορισμένη έννοια ακολουθεί ανάλογη μεταβολή και της άλλης μεταβλητής κατά την ίδια έννοια ή όλα τα σημεία βρίσκονται πάνω σε μια ευθεία με θετική κλίση, δηλαδή y = α + βx, β > 0, (σχήμα α).
- Αν r = -1 τότε έχουμε **τέλεια αρνητική συσχέτιση** δηλαδή σε κάθε μεταβολή της μιας μεταβλητής κατά ορισμένη έννοια ακολουθεί ανάλογη μεταβολή κατά την αντίθετη έννοια ή όλα τα σημεία βρίσκονται πάνω σε μια ευθεία αρνητική κλίση y = α + βx, β < 0, (σχήμα β).
- Αν r = 0 τότε **δεν υπάρχει γραμμική συσχέτιση** μεταξύ των μεταβλητών, δηλαδή οι μεταβλητές X και Y είναι **γραμμικά ασυσχέτιστες,** (σχήμα ζ).

Επομένως σύμφωνα με αυτά κατασκευάζουμε και τα αντίστοιχα διαγράμματα διασποράς και συντελεστές συσχέτισης για διάφορα ζεύγη παρατηρήσεων (x_i, y_i).





Αποδεικνύεται ότι ο συντελεστής γραμμικής συσχέτισης του Pearson (r) δίνεται ισοδύναμα και από τον παρακάτω τύπο, η χρήση του οποίου διευκολύνει συχνά τους υπολογισμούς κυρίως στην περίπτωση που οι $\bar{x}$, $\bar{y}$ δεν είναι ακέραιοι:

$$r = \frac{\nu\sum_{i=1}^{\nu} x_i y_i - (\sum_{i=1}^{\nu} x_i)(\sum_{i=1}^{\nu} y_i)}{\sqrt{\nu\sum_{i=1}^{\nu} x_i{}^2 - (\sum_{i=1}^{\nu} x_i)^2}\sqrt{\nu\sum_{i=1}^{\nu} y_i{}^2 - (\sum_{i=1}^{\nu} y_i)^2}}$$

Επίσης, είναι σημαντικό να αναφέρουμε και τον έλεγχο σημαντικότητας για τον συντελεστή συσχέτισης, ο οποίος αναφέρει ότι ο δειγματικός συντελεστής συσχέτισης μεταξύ δύο μεταβλητών X και Y, αποτελεί μία εκτίμηση του συντελεστή συσχέτισης r στο στατιστικό πληθυσμό. Ελέγχουμε την υπόθεση r = 0, δηλαδή ότι δεν υπάρχει γραμμική σχέση μεταξύ των X και Y (οι μεταβλητές X και Y είναι ασυσχέτιστες) ως προς την εναλλακτική $r \neq 0$. Δηλαδή:

H_0: r = 0 και

H_1: $r \neq 0$

το κριτήριο του ελέγχου είναι: $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$

ΠΑΡΑΔΕΙΓΜΑ: Να υπολογιστεί και να ερμηνευτεί ο συντελεστής γραμμικής συσχέτισης r (του Pearson) μεταξύ των μεταβλητών X και Y με βάση τις παρακάτω τιμές:

ΠΙΝΑΚΑΣ 4

x	y
10	21
13	24
17	29
21	25
25	36
28	33
30	40

ΛΥΣΗ: Για τον υπολογισμό του συντελεστή συσχέτισης μεταξύ των X και Y φτιάχνουμε τον παρακάτω πίνακα:

ΠΙΝΑΚΑΣ 5

x	y	x^2	y^2	xy
10	21	100	441	210
13	24	169	576	312
17	29	289	841	493
21	25	441	625	525
25	36	625	1296	900
28	33	784	1089	924
30	40	900	1600	1200
$\Sigma x = 144$	$\Sigma y = 208$	$\Sigma x^2 = 3308$	$\Sigma y^2 = 6468$	$\Sigma xy = 4564$

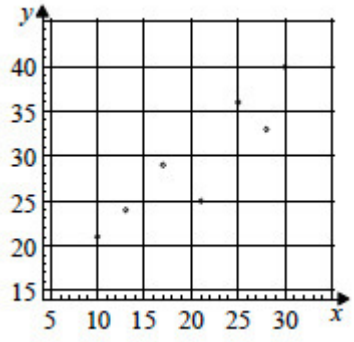
Ο συντελεστής συσχέτισης υπολογίζεται από τη σχέση:

$$r = \frac{\nu \sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\nu \sum x^2 - \left(\sum x\right)^2}\sqrt{\nu \sum y^2 - \left(\sum y\right)^2}}$$

$$= \frac{7(4564)-(144)(208)}{\sqrt{7(3308)-(144)^2}\sqrt{7(6468)-(208)^2}} \approx 0{,}9 .$$

Η υψηλή τιμή του r μας δείχνει ότι υπάρχει πολύ έντονη θετική γραμμική συσχέτιση μεταξύ των μεταβλητών X και Y, όπως εξάλλου μπορούμε να το διαπιστώσουμε και από το αντίστοιχο διάγραμμα διασποράς.

ΔΙΑΓΡΑΜΜΑ 4



3.5 ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ ΓΙΑ ΤΗΝ ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η διασπορά της μεταβλητής Y εκφράζεται με τις αποκλίσεις $y_i - \bar{y}$ των διαφόρων τιμών από τη μέση τιμή τους. Αν όλες οι τιμές ήταν ίσες μεταξύ τους δεν θα υπήρχε μεταβλητότητα στα δεδομένα και κάθε απόκλιση $y_i - \bar{y}$ θα ήταν ίση με το μηδέν. Όσο μεγαλύτερες είναι οι αποκλίσεις ($y_i - \bar{y}$), τόσο μεγαλύτερη θα είναι και η διασπορά των δεδομένων. Η ολική μεταβλητότητα (διασπορά) των παρατηρήσεων εκφράζεται σαν το άθροισμα των τετραγώνων των αποκλίσεων ($y_i - y$) και συμβολίζεται με

SST = Σ $(y_i - \bar{y})^2$ που ονομάζεται **Ολικό άθροισμα τετραγώνων** (Total Sum of Squares).

Ένα νέο μέτρο της μεταβλητότητας των y_i γύρω από την ευθεία παλινδρόμησης, είναι το **άθροισμα τετραγώνων των σφαλμάτων** (Error Sum of Squares) Σ e_i^2 και συμβολίζεται με

SSE = Σ $(y_i - \hat{y}_i)^2$

Επομένως η διαφορά των SST και SSE συμβολίζεται με SSR = SST – SSE.

Το SSR = Σ $(\hat{y}_i - \bar{y})^2$ καλείται **άθροισμα τετραγώνων παλινδρόμησης** (Regression Sum of Squares) και εκφράζει την επίδραση της σχέσης παλινδρόμησης των δύο μεταβλητών στη μείωση της μεταβλητότητας των παρατηρήσεων y_i.

Γενικά αποδεικνύεται ότι ισχύει η σχέση

Σ $(y_i - \bar{y})^2$ = Σ $(\hat{y}_i - \bar{y})^2$ + Σ $(y_i - \hat{y}_i)^2$ ή

SST = SSR + SSE

δηλαδή η συνολική μεταβλητότητα των τιμών εκφράζεται σαν άθροισμα δύο όρων, της μεταβλητότητας που ερμηνεύεται από την παλινδρόμηση (SSR) και της μεταβλητότητας που παραμένει ανερμήνευτη από την μεταβλητή X, σαν το υπόλοιπο ή σφάλμα (SSE).

Οι βαθμοί ελευθερίας που αντιστοιχούν σε κάθε άθροισμα τετραγώνων είναι:

n – 1 βαθμοί ελευθερίας για το SST διότι υπάρχουν n παρατηρήσεις και ο περιορισμός Σ ($y_i - \bar{y}$) = 0 και

n – 2 βαθμοί ελευθερίας για το SSE διότι υπάρχουν n υπόλοιπα και δύο περιορισμοί στα υπόλοιπα, e_i εκτιμώντας τις παραμέτρους β_0 και β_1 από τις κανονικές εξισώσεις.

Το SSR έχει 1 βαθμό ελευθερίας, διότι υπάρχουν δύο παράμετροι στη συνάρτηση παλινδρόμησης και οι αποκλίσεις $(\hat{y}_i - \bar{y})$ υπόκεινται στον περιορισμό Σ $(\hat{y}_i - \bar{y}) = 0$.

Κάθε άθροισμα τετραγώνων όταν διαιρείται με τους αντίστοιχους βαθμούς ελευθερίας, καλείται μέσο άθροισμα τετραγώνων.

Μέσο άθροισμα τετραγώνων παλινδρόμησης (*Regression* Mean Squares): $MSR = \frac{SSR}{1}$

Μέσο άθροισμα τετραγώνων σφαλμάτων (Error Mean Squares): $MSE = \frac{SSE}{n-2}$

Τα αθροίσματα τετραγώνων και τα αντίστοιχα μέσα αθροίσματα τετραγώνων συνοψίζονται στον παρακάτω πίνακα ανάλυσης διασποράς που ακολουθεί.

ΠΙΝΑΚΑΣ 6

Πηγή μεταβολής (source of varation)	**Βαθμοί Ελευθερίας d.f**	**Άθροισμα τετραγώνων SS**	**Μέσο άθροισμα τετραγώνων MSS**
Παλινδρόμηση (Regression)	1	$SSR = \Sigma\,(\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{1}$
Σφάλμα (Error)	n – 2	$SSE = \Sigma\,(y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-2}$
Ολικό (Total)	n – 1	$SST = \Sigma\,(y_i - \bar{y})^2$	

..

3.6 ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ

Μετά τον προσδιορισμό της ευθείας των ελαχίστων τετραγώνων, διά μέσου της εξίσωσης

$$\hat{y} = \hat{\alpha} + \hat{\beta}x\,,$$

εναπομένει η αξιολόγηση της προσαρμογής της ευθείας αυτής επί των δειγματικών τιμών. Ένας τρόπος για να αξιολογήσουμε την προσαρμογή της ευθείας των ελαχίστων τετραγώνων είναι να υπολογίσουμε το συντελεστή προσδιορισμού. Ο συντελεστής προσδιορισμού της δειγματικής ευθείας της παλινδρόμησης, συμβολιζόμενος με R^2, ορίζεται ως το τετράγωνο του δειγματικού συντελεστή συσχέτισης, δηλαδή $R^2 = r^2$.

Επειδή ο δειγματικός συντελεστής συσχέτισης παίρνει τιμές στο διάστημα [-1, 1], ο συντελεστής προσδιορισμού παίρνει τιμές στο διάστημα [0,1]. Όταν $R^2 = 1$, όλα τα σημεία που αναπαριστούν τις δειγματικές τιμές των X και Y βρίσκονται τοποθετημένα επί της ευθείας των ελαχίστων τετραγώνων. Όταν $R^2 = 0$, δεν υπάρχει γραμμική σχέση μεταξύ των δειγματικών τιμών των X και Y.

Ο συντελεστής προσδιορισμού, ως μέτρο της προσαρμογής της ευθείας των ελαχίστων τετραγώνων επί των δειγματικών τιμών, ορίζεται πρωτογενώς από την ανάλυση της συνολικής διασποράς της εξαρτημένης μεταβλητής Y σε επιμέρους συνιστώσες. Χρησιμοποιώντας την ταυτότητα

$$\left(y_i - \hat{y}_i\right) = \left(y_i - \bar{y}\right) - \left(\hat{y}_i - \bar{y}\right),\ i = 1,2,...,n\,,$$

η οποία ισχύει για τις δειγματικές τιμές της μεταβλητής Y, υψώνοντας και τα δύο μέλη της στο τετράγωνο και αθροίζοντας παίρνουμε:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[(y_i-\bar{y})-(\hat{y}_i-\bar{y})]^2=$$

$$=\sum_{i=1}^{n}[(y_i-\bar{y})^2+(\hat{y}_i-\bar{y})^2-2(y_i-\bar{y})(\hat{y}_i-\bar{y})]=$$

$$=\sum_{i=1}^{n}(y_i-\bar{y})^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2-2\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y}).$$

Επειδή

$$-2\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y})=-2\sum_{i=1}^{n}(y_i-\bar{y})(\hat{\alpha}+\hat{\beta}x_i-\hat{\alpha}-\hat{\beta}\bar{x})=$$

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

θέτοντας όπου

$$\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})=\hat{\beta}\sum_{i=1}^{n}(x_i-\bar{x})^2$$, (από τον υπολογισμό του $\hat{\beta}$)

και

$$x_i-\bar{x}=\frac{\hat{y}_i-\bar{y}}{\hat{\beta}}$$

(από την αντικατάσταση του $\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$ στην εξίσωση παλινδρόμησης), προκύπτει

$$-2\hat{\beta}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})=-2\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2=-2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2,$$

και τελικά

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

$$=\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

ή

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

Στην παραπάνω εξίσωση, η ποσότητα $\sum_{i=1}^{n}(y_i-\bar{y})^2$ ονομάζεται **συνολικό άθροισμα τετραγώνων** και αποτελεί μέτρο της συνολικής διασποράς των δειγματικών τιμών της Y

γύρω από τη μέση τιμή τους $\bar{y}$. Η ποσότητα $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ ονομάζεται **άθροισμα τετραγώνων επεξηγούμενο από την γραμμική παλινδρόμηση** και εκφράζει τη διασπορά των εκτιμώμενων τιμών της Υ γύρω από τη δειγματική μέση τιμή $\bar{y}$. Η ποσότητα αυτή αποτελεί μέτρο της διασποράς των δειγματικών τιμών της Υ, που ερμηνεύεται από το υπόδειγμα της γραμμικής παλινδρόμησης (της διασποράς δηλαδή που οφείλεται στη γραμμική επίδραση της Χ επί της Υ). Τέλος η ποσότητα

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ είναι το γνωστό **άθροισμα τετραγώνων των σφαλμάτων** και εκφράζει τη διασπορά των δειγματικών τιμών της Υ γύρω από την εκτιμώμενη ευθεία της παλινδρόμησης. Όσο μικρότερο είναι το άθροισμα των τετραγώνων των σφαλμάτων τόσο πλησιέστερα βρίσκονται οι δειγματικές τιμές της εξαρτημένης μεταβλητής Υ στην ευθεία των ελαχίστων τετραγώνων. Ισχύει επομένως ότι:

Συνολικό άθροισμα τετραγώνων = Άθροισμα τετραγώνων επεξηγούμενο από τη γραμμική παλινδρόμηση + Άθροισμα τετραγώνων των σφαλμάτων.

Για να είναι η προσαρμογή της ευθείας των ελαχίστων τετραγώνων επί των δειγματικών δεδομένων όσο το δυνατόν καλύτερη, θα πρέπει το άθροισμα των τετραγώνων των σφαλμάτων να είναι όσο το δυνατόν μικρότερο και, επομένως, σύμφωνα με την προηγούμενη εξίσωση, το άθροισμα τετραγώνων το επεξηγούμενο από τη γραμμική παλινδρόμηση να είναι όσο το δυνατόν μεγαλύτερο. Το ποσοστό, επομένως, του συνολικού αθροίσματος τετραγώνων που επεξηγείται από τη γραμμική παλινδρόμηση, υπολογίζεται από το λόγο:

$$R^2 = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2},$$

αποτελεί μέτρο της προσαρμογής της ευθείας των ελαχίστων τετραγώνων επί των δειγματικών τιμών, και ορίζει το συντελεστή προσδιορισμού. Ο συντελεστής προσδιορισμού, επομένως, μπορεί να ερμηνευθεί ως το ποσοστό της μεταβλητότητας των τιμών της Υ που επεξηγείται από τη γραμμική παλινδρόμηση. Με απλούς αλγεβρικούς μετασχηματισμούς αποδεικνύεται ότι:

$$R^2 = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{\Sigma(\hat{\alpha} + \hat{\beta}x_i - \hat{\alpha} - \hat{\beta}\bar{x})^2}{\Sigma(y_i - \bar{y})^2} =$$

$$= \beta^2 \frac{\Sigma(x_i - \bar{x})^2}{\Sigma(y_i - \bar{y})^2} =$$

$$= \left[\frac{\Sigma (\mathrm{x_i} - \bar{x})(y_i - \bar{y})}{\Sigma (\mathrm{x_i} - \bar{x})^2} \right]^2 \left[\frac{\Sigma (\mathrm{x_i} - \bar{x})^2}{\Sigma (\mathrm{y_i} - \bar{y})^2} \right] =$$

$$= \frac{\left[\Sigma (\mathrm{x_i} - \bar{x})(y_i - \bar{y}\right]^2}{\left[\Sigma (\mathrm{x_i} - \bar{x})^2\right]\left[\Sigma (\mathrm{y_i} - \bar{y})^2\right]} = r^2$$

δηλαδή ότι $R^2 = r^2$.

Επίσης ο έλεγχος της προσαρμογής της ευθείας της παλινδρόμησης επί των πληθυσμιακών τιμών των μεταβλητών X και Y, δηλαδή ο έλεγχος της ύπαρξης γραμμικής σχέσης μεταξύ των μεταβλητών X και Y, μπορεί να γίνει με τη βοήθεια της ανάλυσης διακύμανσης της μεταβλητής Y.

Έστω ότι ισχύουν οι προϋποθέσεις της απλής γραμμικής παλινδρόμησης, ο έλεγχος της ισότητας:

$\mathrm{H_0} : \beta = 0$

έναντι της εναλλακτικής

$\mathrm{H_A} : \beta \neq 0,$

γίνεται με την βοήθεια του λόγου

$$F = \frac{MSR}{MSE}.$$

Η ποσότητα MSR ονομάζεται **μέσο τετράγωνο της παλινδρόμησης** και ισούται με το άθροισμα των τετραγώνων, το επεξηγούμενο από την παλινδρόμηση διαιρούμενο με τους αντίστοιχους βαθμούς ελευθερίας. Οι βαθμοί ελευθερίας που αντιστοιχούν στο συγκεκριμένο άθροισμα τετραγώνων ορίζονται από τον αριθμό των συντελεστών του υποδείγματος ελαττωμένο κατά 1. Στην προκειμένη περίπτωση, οι συντελεστές του υποδείγματος αυτού είναι 2, άρα οι βαθμοί ελευθερίας είναι 2 – 1 = 1. Επομένως,

$$MSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

Η ποσότητα MSE ονομάζεται **μέσο τετράγωνο των σφαλμάτων** και ισούται με το άθροισμα τετραγώνων των σφαλμάτων, διαιρούμενο επίσης με τους αντίστοιχους βαθμούς ελευθερίας, οι οποίοι είναι ίσοι με n – 2. Δηλαδή

$MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ (n-2)$. Τέλος όταν ισχύει η μηδενική υπόθεση $\mathrm{H_0} : \beta = 0$, ο λόγος

$$F = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} (\hat{\mathrm{y}}_\mathrm{i} - \bar{y})^2}{\sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{y_i} - \hat{y}_i)^2 \Big/ (n-2)}.$$

ακολουθεί την κατανομή F με 1 και n – 2 βαθμούς ελευθερίας.

3.7 ΕΛΕΓΧΟΣ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΜΟΝΤΕΛΟΥ-ΑΝΑΛΥΣΗ ΥΠΟΛΟΙΠΩΝ

Μια βασική τεχνική για το έλεγχο της καταλληλότητας ενός υποδείγματος παλινδρόμησης είναι η ανάλυση υπολοίπων. Με τις γραφικές παραστάσεις των υπολοίπων μπορούμε να ελέγξουμε εάν οι όροι e_i είναι τυχαίες μεταβλητές στατιστικά ανεξάρτητες, που έχουν κατανομή με μέση τιμή 0 και σταθερή διακύμανση σ^2.

Επομένως, οι γραφικές παραστάσεις των υπολοίπων, με τις οποίες ελέγχουμε την καταλληλότητα του υποδείγματος είναι οι εξής:

- Διαγράμματα κανονικότητας

  Αν τα υπόλοιπα έχουν κανονική κατανομή, θα σχηματίζουν στο διάγραμμα μία ευθεία με κλίση 45^o. Αν τα σημεία στη γραφική παράσταση αποκλίνουν από την ευθεία, δεν ισχύει η υπόθεση της κανονικότητας.
- Ιστόγραμμα των υπολοίπων.
- Υπόλοιπα ως προς τις εκτιμήσεις $\hat{y}_i$.
- Υπόλοιπα ως προς την ανεξάρτητη μεταβλητή X.
- Υπόλοιπα ως προς τη σειρά των αντίστοιχων παρατηρήσεων.

Αυτή η γραφική παράσταση χρησιμοποιείται για τον εντοπισμό μη τυχαίων σφαλμάτων (συστηματικά σφάλματα), κυρίως για επιδράσεις συσχέτισης των υπολοίπων.

Τέλος, θα εξετάσουμε τις εξής αποκλίσεις από τις υποθέσεις του γραμμικού υποδείγματος.

1. Η συνάρτηση παλινδρόμησης δεν είναι γραμμική.

2. Οι κατανομές των τιμών της μεταβλητής Y, ισοδύναμα οι όροι σφάλματος, δεν έχουν σταθερή διακύμανση για όλα τα επίπεδα τιμών της X.

(Υπόθεση ομοσκεδαστικότητας).

3. Οι κατανομές των τιμών της μεταβλητής Y, ισοδύναμα οι όροι σφάλματος, δεν ακολουθούν κανονική κατανομή.

4. Οι όροι σφάλματος, δεν είναι στατιστικά ανεξάρτητοι.

ΚΕΦΑΛΑΙΟ 4 ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

4.1 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Στο υπόδειγμα της πολλαπλής γραμμικής παλινδρόμησης υποθέτουμε ότι η σχέση μιας συνεχούς μεταβλητής Y και μιας σειράς κ ανεξάρτητων μεταβλητών X_1, X_2, …, $X_κ$ είναι γραμμική. Η εκτίμηση των τιμών της Y από τις τιμές των ανεξάρτητων μεταβλητών κατ' αναλογία με το υπόδειγμα της απλής γραμμικής παλινδρόμησης είναι εφικτή, όταν ισχύουν οι εξής προϋποθέσεις:

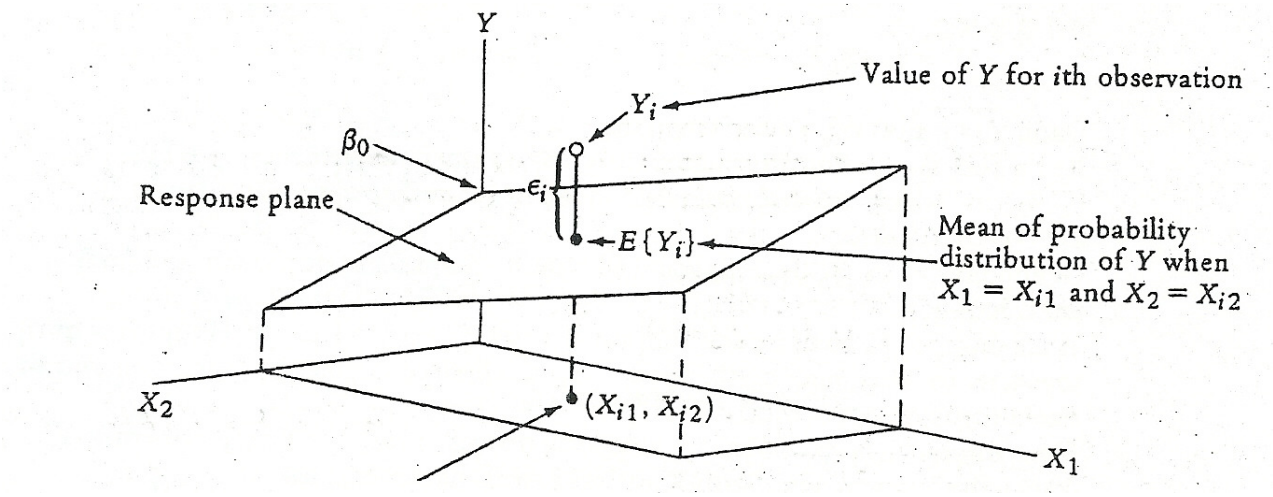
1. Ο προσδιορισμός των τιμών των μεταβλητών X_1, X_2, …, $X_κ$ γίνεται χωρίς σφάλμα.

2. Σε κάθε σύνολο τιμών $x_1, x_2, ..., x_κ$ των μεταβλητών X_1, X_2, …, $X_κ$ αντιστοιχεί ένας υπο-πληθυσμός τιμών της Y, ο οποίος ακολουθεί την κανονική κατανομή.

3. Οι μέσες τιμές των υπο-πληθυσμών της Y συνδέονται με τις αντίστοιχες τιμές των μεταβλητών X_1, X_2, …, $X_κ$, διά μέσου μιας γραμμικής σχέσης της μορφής:

$$\mu_{y|x_1,x_2,...,x_κ} = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_κ x_κ + \varepsilon ,$$

ΣΧΗΜΑ 1



4. Οι διακυμάνσεις των υπο-πληθυσμών της Y που ορίζονται για κάθε σύνολο τιμών $x_1, x_2, ..., x_κ$ είναι ίσες. Η κοινή διακύμανση των υπο-πληθυσμών της Y συμβολίζεται με $σ^2$. Η παραδοχή της ισότητας των διακυμάνσεων των υπο-πληθυσμών της Y, όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης, ονομάζεται **ομασκεδαστικότητα.**

5. Οι τιμές της Y είναι ανεξάρτητες η μια της άλλης.

Όλες οι προηγούμενες προϋποθέσεις συνοψίζονται στην παρακάτω εξίσωση, η οποία ονομάζεται **υπόδειγμα της πολλαπλής παλινδρόμησης:**

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_\kappa x_\kappa + \varepsilon , \qquad (1)$$

όπου

y: είναι μία οποιαδήποτε τιμή του υπο-πληθυσμού της Y που αντιστοιχεί στις τιμές των ανεξάρτητων μεταβλητών $X_1, X_2, \ldots, X_\kappa$,

$\alpha, \beta_1, \beta_2, ..., \beta_\kappa$: είναι σταθερές, οι οποίες ονομάζονται **μερικοί συντελεστές της παλινδρόμησης** και

ε : είναι η τιμή μιας τυχαίας μεταβλητής η οποία έχει μέση τιμή 0 και διακύμανση ίση με την κοινή διακύμανση των διάφορων υπο-πληθυσμών της Y, σ^2. Η κατανομή της τυχαίας αυτής μεταβλητής είναι κανονική, ενώ οι επιμέρους τιμές της είναι ανεξάρτητες η μια της άλλης.

Η ερμηνεία των συντελεστών $\alpha, \beta_1, \beta_2, ..., \beta_\kappa$ του υποδείγματος της πολλαπλής παλινδρόμησης είναι αντίστοιχη με την ερμηνεία των συντελεστών της απλής παλινδρόμησης. Ο σταθερός όρος α είναι η τιμή της εξαρτημένης μεταβλητής Y, όταν όλες οι ανεξάρτητες μεταβλητές παίρνουν τιμή 0, ενώ ένας οποιασδήποτε συντελεστή β_i είναι η μεταβολή της μέσης τιμής της εξαρτημένης μεταβλητής Y για μία μονάδα αύξησης της ανεξάρτητης μεταβλητής X_i, εφόσον οι τιμές των άλλων ανεξάρτητων μεταβλητών παραμένουν σταθερές.

Τέλος, οι πραγματικές τιμές της Y αποτελούνται από δύο συνιστώσες: τη συνιστώσα της Y, την E(Y), που οφείλεται στις συστηματικές επιδράσεις των $X_1, X_2, \ldots, X_\kappa$, και την τυχαία (σφάλμα) συνιστώσα (ε) που ενσωματώνει όλους τους άλλους (εκτός των $X_1, X_2, \ldots, X_\kappa$) παράγοντες που επηρεάζουν τη διαμόρφωση της τιμής της Y. Δηλαδή:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_\kappa x_\kappa + \varepsilon = E(Y) + \varepsilon$$

όπου:

$$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_\kappa x_\kappa \qquad (2).$$

4.2 Η ΕΞΙΣΩΣΗ ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Σκοπός στο κεφάλαιο αυτό είναι να εκτιμήσουμε τις παραμέτρους του υποδείγματος της πολλαπλής παλινδρόμησης, δηλαδή τους συντελεστές $\alpha, \beta_1, \beta_2, ..., \beta_\kappa$. Οι εκτιμήσεις από τα δεδομένα του δείγματος των συντελεστών πολλαπλής παλινδρόμησης του πληθυσμού

($\alpha, \beta_1, \beta_2, ..., \beta_\kappa$) συμβολίζονται με $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, ...,$ και $\hat{\beta}_\kappa$ αντίστοιχα. Έτσι, η εξίσωση που θα προκύψει από την εκτίμηση των **συντελεστών πολλαπλής παλινδρόμησης** είναι η:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_\kappa x_\kappa \qquad (3)$$

Η εξίσωση (3) σχηματίζει ένα υπερεπίπεδο στο χώρο κ + 1 διαστάσεων, που βέβαια δεν μπορεί να απεικονιστεί στο χαρτί δύο διαστάσεων. Μόνο στην περίπτωση των τριών μεταβλητών (δύο ανεξάρτητες μεταβλητές) αντιστοιχεί στην εξίσωση (3) ένα επίπεδο στο χώρο των τριών διαστάσεων. Οι συντελεστές $\hat{\beta}_1, \hat{\beta}_2, ...,$ και $\hat{\beta}_\kappa$ δείχνουν τη μερική επίδραση που ασκούν οι ανεξάρτητες μεταβλητές στην εξαρτημένη μεταβλητή. Για παράδειγμα, ο συντελεστής $\hat{\beta}_2$ υποδηλώνει τη μεταβολή της Y, που θα προκύψει εάν η μεταβλητή X_2 μεταβληθεί κατά μία μονάδα μέτρησής της, και οι άλλες ανεξάρτητες μεταβλητές (X_1, X_3, …, X_κ) παραμένουν σταθερές. Επομένως, ο συντελεστής $\hat{\beta}_2$ μετράει τη μερική επίδραση της ανεξάρτητης μεταβλητής X_2. Γι' αυτό το λόγο οι συντελεστές $\hat{\beta}_1, \hat{\beta}_2, ...,$ και $\hat{\beta}_\kappa$ ονομάζονται και **συντελεστές μερικής παλινδρόμησης**.

Η εκτίμηση της εξίσωσης (3) θα προκύψει από τη μέθοδο των ελαχίστων τετραγώνων, δηλαδή θα αναζητήσουμε εκείνες τις τιμές των $\hat{\alpha}, \hat{\beta}_1, ..., \hat{\beta}_\kappa$ που ελαχιστοποιούν το άθροισμα των τετραγώνων των αποκλίσεων μεταξύ των πραγματικών τιμών της Y και των θεωρητικών τιμών $\hat{y}$ οι οποίες προκύπτουν από την εξίσωση παλινδρόμησης. Ας εξετάσουμε την περίπτωση με δύο ανεξάρτητες μεταβλητές x_1, και x_2. Το υπόδειγμα που θα εκτιμήσουμε είναι:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \qquad (4)$$

Η $\hat{y}$ είναι η εκτίμηση της E(Y) και κατά αναλογία με την εξίσωση (1), που αναφέρονται στην εξίσωση παλινδρόμησης του πληθυσμού, οι αποκλίσεις μεταξύ των πραγματικών τιμών της y και των τιμών $\hat{y}$ της (4) συμβολίζονται με e, δηλαδή:

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

ή

$$e_i = y_i - (\hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}), \text{ για } i = 1, ..., n$$

όπου n είναι το μέγεθος του δείγματος.

Το άθροισμα των τετραγώνων των αποκλίσεων για τις n τριάδες των παρατηρήσεων ισούται με:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \Sigma[y_i - (\hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})]^2$$

και ας συμβολίσουμε το άθροισμα αυτό με Q. Αναζητούμε εκείνες τις τιμές $\hat{\alpha}, \hat{\beta}_1$ και $\hat{\beta}_2$ που ελαχιστοποιούν την έκφραση:

$$Q = \Sigma[y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}]^2$$

που σημαίνει ότι πρέπει να παραγωγίσουμε τη Q ως προς $\hat{\alpha}, \hat{\beta}_1$ και $\hat{\beta}_2$, να εξισώσουμε τις παραγώγους με το μηδέν και να λύσουμε ως προς τις άγνωστες παραμέτρους. Έτσι θα προκύψει ένα σύστημα τριών εξισώσεων με τρεις αγνώστους. Δηλαδή:

$$\partial Q/\partial \hat{\alpha} = -2\,\Sigma[y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}] = 0$$

$$\partial Q/\partial \hat{\beta}_1 = -2\,\Sigma x_{1i}[y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}] = 0$$

$$\partial Q/\partial \hat{\beta}_2 = -2\,\Sigma x_{2i}[y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}] = 0$$

ή

$$\Sigma y_i - n\,\hat{\alpha} - \hat{\beta}_1\,\Sigma x_{1i} - \hat{\beta}_2\,\Sigma x_{2i} = 0$$

$$\Sigma y_i x_{1i} - \hat{\alpha}\,\Sigma x_{1i} - \hat{\beta}_1\,\Sigma x_{1i}^{\,2} - \hat{\beta}_2\,\Sigma x_{1i} x_{2i} = 0$$

$$\Sigma y_i x_{2i} - \hat{\alpha}\,\Sigma x_{2i} - \hat{\beta}_1\,\Sigma x_{1i} x_{2i} - \hat{\beta}_2\,\Sigma x_{2i}^{\,2} = 0$$

ή

$$\Sigma y_i = n\,\hat{\alpha} + \hat{\beta}_1\,\Sigma x_{1i} + \hat{\beta}_2\,\Sigma x_{2i}$$

$$\Sigma y_i x_{1i} = \hat{\alpha}\,\Sigma x_{1i} + \hat{\beta}_1\,\Sigma x_{1i}^{\,2} + \hat{\beta}_2\,\Sigma x_{1i} x_{2i} \qquad (5)$$

$$\Sigma y_i x_{2i} = \hat{\alpha}\,\Sigma x_{2i} + \hat{\beta}_1\,\Sigma x_{1i} x_{2i} + \hat{\beta}_2\,\Sigma x_{2i}^{\,2}$$

Η λύση του συστήματος των εξισώσεων (5) με την μέθοδο των οριζουσών δίνει τις τιμές των $\hat{\beta}_1$ και $\hat{\beta}_2$. Ο συντελεστής $\hat{\alpha}$ προκύπτει από την πρώτη εξίσωση του συστήματος εξισώσεων (5) με αντικατάσταση των τιμών των $\hat{\beta}_1$ και $\hat{\beta}_2$ και διαιρώντας δια n έχουμε:

$$\hat{\alpha} = (\Sigma y)/n - \hat{\beta}_1(\Sigma x_1)/n - \hat{\beta}_2(\Sigma x_2)/n \qquad (6)$$

ΠΑΡΑΔΕΙΓΜΑ: Σύμφωνα με μια έρευνα αγοράς που έγινε από μια εταιρία με σκοπό να εξετάσει τη ζήτηση ενός τυποποιημένου είδους γιαουρτιού. Πιο συγκεκριμένα, από ένα τυχαίο δείγμα 14 (τετραμελών) οικογενειών συλλέγονται και η μέση τιμή τριών ανταγωνιστικών προϊόντων. Σκοπός της έρευνας είναι να διαπιστώσουμε την επίδραση που έχουν στη ζήτηση του προϊόντος τόσο η τιμή πώλησης όσο και το μέσο επίπεδο τιμών των ανταγωνιστικών προϊόντων.

Ο πίνακας 1 δίνει τα στοιχεία που συλλέχθηκαν.

Στον επόμενο πίνακα 2 παρουσιάζονται όλοι οι απαραίτητοι υπολογισμοί για την εκτίμηση των γνωστών όρων του συστήματος των τριών εξισώσεων.

Έτσι, το σύστημα (5) έχει ως εξής:

ΛΥΣΗ: $306 = 14\,\hat{\alpha} + 2.167\,\hat{\beta}_1 + 2.313\,\hat{\beta}_2$

$46.597 = 2.167\,\hat{\alpha} + 338.913\,\hat{\beta}_1 + 355.307\,\hat{\beta}_2$

$51.304 = 2.313\,\hat{\alpha} + 355.307\,\hat{\beta}_1 + 385.423\,\hat{\beta}_2$

ΠΙΝΑΚΑΣ 1 Ποσότητα και Τιμή Αγοράς και Μέση Τιμή Ανταγωνισμού

Οικογένεια	**Ποσότητα (Μονάδες)** (y)	**Μέση Τιμή Αγοράς (Λεπτά)** (x_1)	**Μέση Τιμή Αντ/σμού (Λεπτά)** (x_2)
1	23	141	174
2	26	145	175
3	17	167	139
4	29	133	183
5	18	162	156
6	22	158	180
7	16	178	143
8	24	151	183
9	19	173	155
10	18	185	148
11	21	148	171
12	20	153	150
13	25	135	178
14	28	138	178

Η λύση του συστήματος με την μέθοδο των οριζουσών δίνει τις παρακάτω τιμές των συντελεστών μερικής παλινδρόμησης $\hat{\beta}_1$ και $\hat{\beta}_2$.

$$\hat{\beta}_1 = \frac{\begin{vmatrix} 14 & 306 & 2.313 \\ 2.167 & 46.597 & 355.307 \\ 2.313 & 51.304 & 385.423 \end{vmatrix}}{\begin{vmatrix} 14 & 2.167 & 2.313 \\ 2.167 & 338.913 & 355.307 \\ 2.313 & 355.307 & 385.423 \end{vmatrix}} = \frac{-6.845.567}{57.487.950} = \text{-0,119.}$$

και

$$\hat{\beta}_2 = \frac{\begin{vmatrix} 14 & 2.167 & 306 \\ 2.167 & 338.913 & 46.597 \\ 2.313 & 355.307 & 51.304 \end{vmatrix}}{\begin{vmatrix} 14 & 2.167 & 2.313 \\ 2.167 & 338.913 & 355.307 \\ 2.313 & 355.307 & 385.423 \end{vmatrix}} = \frac{7.451.353}{57.487.950} = 0,130$$

ΠΙΝΑΚΑΣ 2 Υπολογισμοί για την Επίλυση του Συστήματος των Εξισώσεων

	y	x_1	x_2	x_1^2	x_2^2	$x_1 \cdot x_2$	$y \cdot x_1$	$y \cdot x_2$
	23	141	174	19.881	30.276	24.534	3.243	4.002
	26	145	175	21.025	30.625	25.375	3.770	4.550
	17	167	139	27.889	19.321	23.213	2.839	2.363
	29	133	183	17.689	33.489	24.339	3.857	5.307
	18	162	156	26.244	24.336	25.272	2.916	2.808
	22	158	180	24.964	32.400	28.440	3.476	3.960
	16	178	143	31.684	20.449	25.454	2.848	2.288
	24	151	183	22.801	33.489	27.633	3.624	4.392
	19	173	155	29.929	24.025	26.815	3.287	2.945
	18	185	148	34.225	21.904	27.380	3.330	2.664
	21	148	171	21.904	29.241	25.308	3.108	3.591
	20	153	150	23.409	22.500	22.950	3.060	3.000
	25	135	178	18.225	31.684	24.030	3.375	4.450
	28	138	178	19.044	31.684	24.564	3.864	4.984
Σ =	**306**	**2.167**	**2.313**	**338.913**	**385.423**	**355.307**	**46.597**	**51.304**

Ο συντελεστής $\hat{\alpha}$ προκύπτει από την πρώτη εξίσωση του συστήματος εξισώσεων (5) με αντικατάσταση των τιμών των $\hat{\beta}_1$ και $\hat{\beta}_2$. Δηλαδή:

$\hat{\alpha} = (\Sigma y)/n - \hat{\beta}_1(\Sigma x_1)/n - \hat{\beta}_2(\Sigma x_2)/n =$

= (306) / 14 – (– 0119) (2.167) / 14 – (0,130) (2.313) / 14 = 18,874

Άρα η εξίσωση πολλαπλής παλινδρόμησης είναι:

$$\hat{y} = 18{,}874 - 0{,}119x_1 + 0{,}130x_2$$

όπου:

y : Μηνιαία Κατανάλωση (μονάδες προϊόντος)

x_1 : Μέση τιμή αγοράς (λεπτά)

x_2 : Μέση τιμή ανταγωνιστών (λεπτά)

Η ερμηνεία των συντελεστών μερικής παλινδρόμησης είναι ανάλογη με την ερμηνεία του συντελεστή της απλής παλινδρόμησης, με τη διαφορά ότι τώρα έχουμε την επίδραση περισσότερων της μιας ανεξάρτητων μεταβλητών. Έτσι, από την παραπάνω εξίσωση προκύπτει ότι για κάθε αύξηση της τιμής πώλησης του προϊόντος κατά ένα λεπτό και με σταθερή τη μέση τιμή του ανταγωνισμού, η μηνιαία ζήτηση μειώνεται κατά μέσο όρο κατά 0,12 μονάδες ανά οικογένεια (ή για αύξηση της τιμής κατά 10 λεπτά η μέση μείωση ανά οικογένεια είναι 1,2 μονάδες). Αντίθετα, με σταθερή την τιμή πώλησης του προϊόντος, για κάθε αύξηση των τιμών των ανταγωνιστικών κατά ένα λεπτό, η μηνιαία ζήτηση αυξάνεται κατά μέσο όρο κατά 0,13 μονάδες ανά οικογένεια (ή για αύξηση της τιμής των ανταγωνιστών κατά 10 λεπτά η μέση αύξηση ανά οικογένεια είναι 1,3 μονάδες). Τέλος, ο σταθερός όρος ($\hat{\alpha}$ = 18,874) σημαίνει ότι εάν το συγκεκριμένο είδος γιαουρτιού προσφερθεί από όλους τους παραγωγούς δωρεάν (πχ., στα πλαίσια μιας διαφημιστικής εκστρατείας), η μέση μηνιαία κατανάλωση ανά οικογένεια αναμένεται να διαμορφωθεί σε 19 περίπου μονάδες του προϊόντος.

Με βάση τους παραπάνω συντελεστές μερικής παλινδρόμησης υπολογίζουμε και τους **συντελεστές μερικής ελαστικότητας** της ζήτησης του προϊόντος σε σχέση με την τιμή των ανταγωνιστικών προϊόντων (για διάφορα επίπεδα τιμών). Πιο συγκεκριμένα, οι μέσες ελαστικότητες είναι:

$$\overline{\eta}_{y/x_1} = \hat{\beta}_1 \frac{\overline{x}_1}{\overline{y}} = \hat{\beta}_1 \frac{\Sigma x_1 / n}{\Sigma y / n} =$$

$$= -0{,}119 \frac{2.167/14}{306/14} = -0{,}843$$

και

$$\overline{\eta}_{y/x_2} = \hat{\beta}_2 \frac{\overline{x}_2}{\overline{y}} = \hat{\beta}_2 \frac{\Sigma x_2 / n}{\Sigma y / n} =$$

$$= 0{,}130 \frac{2.313/14}{306/14} = 0{,}983$$

Η ερμηνεία των συντελεστών μερικής ελαστικότητας είναι ανάλογη με την ερμηνεία των συντελεστών μερικής παλινδρόμησης, με τη διαφορά ότι τώρα μετράμε τη σχέση μεταξύ των ποσοστιαίων μεταβολών των μεταβλητών. Έτσι, από τις παραπάνω μέσες ελαστικότητες προκύπτει ότι για κάθε αύξηση της τιμής πώλησης του προϊόντος κατά 1%, και με σταθερή τη μέση τιμή του ανταγωνισμού, η μηνιαία ζήτηση μειώνεται κατά μέσο όρο κατά 0,84% μονάδες ανά οικογένεια. Αντίθετα, με σταθερή την τιμή πώλησης του προϊόντος, για κάθε αύξηση των τιμών των ανταγωνιστών κατά 1%, η μηνιαία ζήτηση αυξάνεται κατά μέσο όρο κατά 0,98% ανά οικογένεια. Δηλαδή, σε απόλυτους όρους, η ελαστικότητα της ζήτησης ως προς τις τιμές των ανταγωνιστών είναι μεγαλύτερη από την ελαστικότητα της ζήτησης ως προς την τιμή του προϊόντος.

Αυτό αποτελεί πλεονέκτημα για τον κατασκευαστή του συγκεκριμένου τύπου γιαουρτιού, αφού οι μεταβολές στην κατανάλωση του είναι λιγότερο ευαίσθητες στις μεταβολές της τιμής πώλησης σε σχέση με τις μεταβολές των τιμών των ανταγωνιστικών προϊόντων. Εάν, για παράδειγμα, μία αύξηση της τιμής του γάλακτος προκαλέσει αύξηση των τιμών σε όλους τους παραγωγούς γιαουρτιού κατά 5%, τότε η ζήτηση θα μειωθεί κατά 4,2% (= – 0,84 · 5%) λόγω της αύξησης της τιμής πώλησης , αλλά ταυτόχρονα θα κερδίσει μέρος της αγοράς γιαουρτιού που έχουν τα ανταγωνιστικά προϊόντα που θα αντισταθμίσει τις απώλειες, και συγκεκριμένα η αύξηση θα ανέλθει σε 4,9% (= 0,98 · 5%). Έτσι, η τελική μεταβολή της ζήτησης μετά από την αύξηση των τιμών στον κλάδο γιαουρτιού θα είναι – 4,2% + 4,9% =
= + 0,7%.

4.3 ΣΥΝΤΕΛΕΣΤΗΣ ΠΟΛΛΑΠΛΗΣ ΣΥΣΧΕΤΙΣΗΣ

Ο συντελεστής πολλαπλής συσχέτισης συμβολιζόμενος $R_{y \cdot 12...\kappa}$, είναι ένα μέτρο της συνολικής γραμμικής σχέσης που υπάρχει μεταξύ μιας (εξαρτημένης) μεταβλητής Y και ενός συνόλου άλλων (ανεξάρτητων) μεταβλητών $X_1, X_2, \ldots, X_\kappa$. Με τον όρο
‘‘γραμμική σχέση’’, στην περίπτωση του συντελεστή πολλαπλής συσχέτισης εννοείται η συσχέτιση της y με τις εκτιμώμενες από το υπόδειγμα τιμές $\hat{y}$. Επειδή η εξίσωση των ελαχίστων τετραγώνων ορίζει το υπερεπίπεδο με την καλύτερη προσαρμογή επί των τιμών

της y, η συσχέτιση της y με τις εκτιμώμενες τιμές $\hat{y}$ είναι η μέγιστη που μπορεί να έχει η y με οποιοδήποτε άλλο γραμμικό συνδυασμό των μεταβλητών $X_1, X_2, \ldots, X_\kappa$. Επιπλέον η τιμή του συντελεστή συσχέτισης της y με τις εκτιμώμενες τιμές ŷ είναι πάντα θετική (ή μηδέν).

Ο συντελεστής πολλαπλής συσχέτισης είναι μια γενίκευση του απλού συντελεστή συσχέτισης r, για την περίπτωση ενός υποδείγματος με πολλές ανεξάρτητες μεταβλητές. Η τιμή του μπορεί να οριστεί είτε ευθέως, ως συντελεστής συσχέτισης της y με τις εκτιμώμενες από το υπόδειγμα τιμές $\hat{y}$, επομένως:

$$R_{y\cdot 12\ldots\kappa} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]\left[\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right]}}$$

είτε έπειτα από μετασχηματισμό της προηγούμενης σχέσης ως η θετική τετραγωνική ρίζα του συντελεστή πολλαπλού προσδιορισμού

$R_{y\cdot 12\ldots\kappa} = \sqrt{R^2_{y\cdot 12\ldots\kappa}}$.

4.4 ΣΥΝΤΕΛΕΣΤΗΣ ΜΕΡΙΚΗΣ ΣΥΣΧΕΤΙΣΗΣ

Ο συντελεστής μερικής συσχέτισης της εξαρτημένης μεταβλητής Y και μιας ανεξάρτητης μεταβλητής X_i εκφράζει τη συσχέτιση μεταξύ της Y και της X_i, όταν οι επιδράσεις των υπολοίπων ανεξάρτητων μεταβλητών επί της Y και της X_i έχουν απομακρυνθεί. Ο συντελεστής μερικής συσχέτισης εκτιμάται από τα δείγματα δεδομένων ως εξής:

1. Υπολογίζεται η εξίσωση της πολλαπλής παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής Y και των ανεξάρτητων μεταβλητών $X_1, X_2, X_{i-1}, X_{i+1}, \ldots, X_\kappa$.

2. Υπολογίζεται η εξίσωση της πολλαπλής παλινδρόμησης μεταξύ της μεταβλητής X_i και των ανεξάρτητων μεταβλητών $X_1, X_2, X_{i-1}, X_{i+1}, \ldots, X_\kappa$.

3. Υπολογίζεται ο δειγματικός συντελεστής συσχέτισης ο οποίος αυτός συντελεστής είναι ο συντελεστής μερικής συσχέτισης της εξαρτημένης μεταβλητής Y και της ανεξάρτητης μεταβλητής X_i.

Σε ένα υπόδειγμα με δύο ανεξάρτητες μεταβλητές X_1 και X_2, ο συντελεστής μερικής συσχέτισης της εξαρτημένης μεταβλητής Y με τη X_1, συμβολιζόμενος $r_{y1\cdot 2}$,
υπολογίζεται από την εξίσωση

$$r_{y1\cdot 2} = \frac{(r_{y1} - r_{y2}r_{12})}{\sqrt{(1-r^2{}_{y2})(1-r^2{}_{12})}},$$

ενώ ο συντελεστής μερικής συσχέτισης της Y με τη X_2, $r_{y2\cdot 1,}$ υπολογίζεται από την εξίσωση

$$r_{y2\cdot 1} = \frac{(r_{y2} - r_{y1}r_{12})}{\sqrt{(1-r^2{}_{y1})(1-r^2{}_{12})}},$$

όπου r_{y1} ο δειγματικός συντελεστής συσχέτισης της Y με τη X_1, r_{y2} ο δειγματικός συντελεστής συσχέτισης της Y με τη X_2.

Στο γενικό υπόδειγμα της πολλαπλής παλινδρόμησης , το τετράγωνο του συντελεστή μερικής συσχέτισης της Y με μια ανεξάρτητη μεταβλητή X_i κατ' αντιστοιχία με την ερμηνεία του συντελεστή πολλαπλού προσδιορισμού που θα αναφερθούμε παρακάτω, εκφράζει το ποσοστό της μεταβλητότητας της Y που ερμηνεύεται από τη X_i, εφόσον οι γραμμικές σχέσεις των υπολοίπων μεταβλητών έχουν αφαιρεθεί τόσο από την Y όσο και από τη X_i. Στο βαθμό που υπάρχουν πολλαπλές επιδράσεις στις τιμές μιας εξαρτημένης μεταβλητής και η χρήση ενός πολλαπλού υποδείγματος παλινδρόμησης είναι πιο κατάλληλη για τη μελέτη της μεταβλητής αυτής από ότι η χρήση ενός απλού υποδείγματος, ο συντελεστής μερικής συσχέτισης μπορεί να οδηγήσει σε εσφαλμένα[10] συμπεράσματα.

Έστω ότι η μεταβλητή X_2 συσχετίζεται (στην πραγματικότητα) θετικά με τη μεταβλητή Y και με τη μεταβλητή X_1. Λόγω της συσχέτισης αυτής, οι μεταβλητές Y και X_1, εμφανίζονται επίσης να συσχετίζονται θετικά μεταξύ τους με έναν υψηλό συντελεστή συσχέτισης r_{y1}. Αν ερμηνεύσουμε τη σχέση της Y με τη X_1, αγνοώντας την επίδραση της X_2 χρησιμοποιώντας δηλαδή το δειγματικό συντελεστή συσχέτισης r_{y1}, κινδυνεύουμε, το συμπέρασμα στο οποίο θα καταλήξουμε να είναι εντελώς εσφαλμένο. Μέτρο της πραγματικής γραμμικής σχέσης που υπάρχει μεταξύ της Y και X_1 είναι ο μερικός συντελεστής συσχέτισης $r_{y1\cdot 2}$, ο οποίος συνοψίζει τη σχέση των δύο μεταβλητών απαλλαγμένη από την γραμμική επίδραση που ασκεί σε αυτήν η X_2.

Τέλος η επίδραση που ασκεί η μεταβλητή X_2 στη σχέση των Y και X_1 ονομάζεται **συγχυτική επίδραση**, η δε μεταβλητή X_2 ονομάζεται **συγχυτικός παράγοντας** στη σχέση των Y και X_1.

[10] εσφαλμένα: λανθασμένα

4.5 ΣΥΝΤΕΛΕΣΤΗΣ ΠΟΛΛΑΠΛΟΥ ΠΡΟΣΔΙΟΡΙΣΜΟΥ

Στην απλή γραμμική παλινδρόμηση περιγράψαμε το συντελεστή προσδιορισμού που μετρά το ποσοστό της μεταβλητικότητας της Y που οφείλονται στις επιδράσεις της ανεξάρτητης μεταβλητής X. Στην πολλαπλή παλινδρόμηση χρησιμοποιούμε επίσης τον ανάλογο συντελεστή για να μετρήσουμε το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής Y που οφείλονται στις επιδράσεις όλων μαζί των ανεξάρτητων μεταβλητών X_1, X_2, …, $X_κ$. Επειδή στο υπόδειγμα της πολλαπλής παλινδρόμησης περιλαμβάνονται περισσότερες από μία ανεξάρτητες μεταβλητές, ο συντελεστής πολλαπλού προσδιορισμού μετράει τη συνολική επίδραση που δέχεται η Y από τις X_1, X_2, …, $X_κ$. Ο συντελεστής πολλαπλού προσδιορισμού ισούται με:

$R^2 = SSR / SST = 1 – SSE / SST$ (ή

$$R^2_{y\cdot 12\ldots\kappa} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad) \qquad (7)$$

όπου:

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

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

$SSE = \Sigma e^2 = \Sigma (y - \hat{y})^2$

επομένως ισχύει η σχέση:

$SST = SSR + SSE$

Από τον ορισμό του πολλαπλού συντελεστή προσδιορισμού, προκύπτει ότι η προσθήκη μιας νέας ανεξάρτητης μεταβλητής θα οδηγήσει σε μείωση της ανερμήνευτης συνιστώσας (αποκλίσεις μεταξύ y και ŷ) και επομένως σε αύξηση της τιμής του συντελεστή R^2. Όμως, κάθε νέα ανεξάρτητη μεταβλητή στοιχίζει και ένα βαθμό ελευθερίας. Οι βαθμοί ελευθερίας ισούται με n – κ – 1, όπου κ είναι ο αριθμός των ανεξάρτητων μεταβλητών. Το ερώτημα είναι εάν η αύξηση αυτή του R^2 είναι τόσο σημαντική, ώστε να αξίζει την απώλεια ενός βαθμού ελευθερίας. Η προσθήκη πολλών ανεξάρτητων μεταβλητών μπορεί να οδηγήσει σε «τεχνητή»[11] αύξηση της τιμής του R^2 που δεν θα έχει καμία αξία, όταν μάλιστα ο αριθμός των ανεξάρτητων μεταβλητών (κ) είναι υψηλός σε σχέση με το μέγεθος του δείγματος (n).

[11] τεχνητή: ψεύτικη, πλασματική

Το πρόβλημα αυτό αντιμετωπίζεται με το «διορθωμένο» (ή «προσαρμοσμένο») συντελεστή πολλαπλού προσδιορισμού που λαμβάνει υπόψη του την απώλεια των βαθμών ελευθερίας. Ο διορθωμένος συντελεστής πολλαπλού προσδιορισμού R_a^2 (ή $\overline{R}^2_{y\cdot 12\ldots\kappa}$) ισούται με:

$$R_a^2 = 1 - (1 - R^2) \cdot \left(\frac{n-1}{n-\kappa-1} \right) \quad (\text{ή}$$

$$\overline{R}^2_{y\cdot 12\ldots\kappa} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (\text{n} - \kappa - 1)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (\text{n} - 1)}. \;) \qquad (8)$$

Στο προηγούμενο παράδειγμα από την εξίσωση ελαχίστων τετραγώνων, εάν περιλάβουμε μία μόνο ανεξάρτητη μεταβλητή έστω τη X_1 ο R^2 χωρίς τη διόρθωση ισούται με 0,85. Ο διορθωμένος συντελεστής R_a^2 είναι:

$$R_a^2 = 1 - (1 - 0{,}85) \cdot \left(\frac{14-1}{14-2-1} \right) =$$

$$= 1 - (0{,}15) \cdot \left(\frac{13}{11} \right) = 1 - (0{,}15) \cdot 1{,}18 =$$

$$= 1 - 0{,}177 = 0{,}82$$

Δηλαδή, 3 ποσοστιαίες μονάδες μικρότερος από την αρχική του τιμή. Εάν η συνεισφορά της νέας μεταβλητής είναι αμελητέα[12], ο διορθωμένος συντελεστής αντί να αυξηθεί θα μειωθεί. Για παράδειγμα, εάν προσθέσουμε δύο νέες μεταβλητές (το δείκτη τιμών καταναλωτή και το συνολικό διαθέσιμο εισόδημα), μία αύξηση του R^2 κατά 2 ποσοστιαίες μονάδες είναι παραπλανητική. Διότι ο R_a^2 είναι:

$$R_a^2 = 1 - (1 - 0{,}87) \cdot \left(\frac{14-1}{14-4-1} \right) = 0{,}81.$$

Στην πραγματικότητα, δηλαδή, έχουμε μείωση του συντελεστή προσδιορισμού, παρά αύξηση. Αυτό οφείλεται στο γεγονός ότι η μικρή αύξηση του R^2 δεν αντισταθμίζει τη μείωση του αριθμού των βαθμών ελευθερίας. Ο R_a^2 έχει σημαντικό ρόλο στις περιπτώσεις εκείνες που ο αριθμός των ανεξάρτητων μεταβλητών είναι αρκετά μεγάλος σε σχέση με το μέγεθος του δείγματος. Όπως προκύπτει από τον τύπο (8) για μεγάλο (σε σχέση με το κ) μέγεθος δείγματος, ο R_a^2 διαφέρει ελάχιστα από τον R^2.

Το άθροισμα των τετραγώνων των σφαλμάτων SSE υπολογίζεται ως εξής:

$$\text{SSE} = \Sigma e^2 = \Sigma (y - \hat{y})$$

[12] αμελητέα: ασήμαντα

$$= \Sigma[(y - (\hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2)]^2$$

$$= \Sigma y^2 - \hat{\alpha}\,\Sigma y - \hat{\beta}_1\,\Sigma y x_1 - \hat{\beta}_2\,\Sigma y x_2 \qquad (9)$$

και το συνολικό άθροισμα των τετραγώνων με:

$$\text{SST} = \Sigma y^2 - (\Sigma y)^2 / n$$

Επομένως, με βάση τα αποτελέσματα του πίνακα 2 και λαμβάνοντας υπόψη ότι για τα παραπάνω δεδομένα $\Sigma y^2 = 6.910$ έχουμε:

$$\text{SSE} = \Sigma y^2 - \hat{\alpha}\,\Sigma y - \hat{\beta}_1\,\Sigma y x_1 - \hat{\beta}_2\,\Sigma y x_2$$

$$= 6.910 - (\ 18{,}874\ )\ (\ 306\ ) - (-0{,}119\ )\ (\ 46.597\ ) - (\ 0{,}13\ )\ (\ 51.304\ )$$

$$= 33{,}322$$

και

$$\text{SST} = \Sigma y^2 - (\Sigma y)^2 / n = 6.910 - (\ 306\ )^2 / 14 = 221{,}714$$

δηλαδή:

$$R^2 = 1 - \text{SSE} / \text{SST}$$

$$= 1 - (\ 33{,}322 / 221{,}714\ ) = 0{,}85.$$

4.6 ΕΠΑΓΩΓΙΚΟΙ ΕΛΕΓΧΟΙ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΤΗΣ ΠΟΛΛΑΠΛΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Οι έλεγχοι της στατιστικής σημαντικότητας στην ανάλυση της πολλαπλής παλινδρόμησης έχουν σκοπό πρώτα να ελέγξουν εάν η εξίσωση της παλινδρόμησης, στο σύνολό της, εξηγεί ένα σημαντικό μέρος των μεταβολών της εξαρτημένης μεταβλητής Y και στη συνέχεια, εφόσον έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές, να ελέγξουμε τη σημαντικότητα των συντελεστών παλινδρόμησης. Έτσι, θα διαπιστώσουμε ποιες μεταβλητές ασκούν σημαντική επίδραση στη Y και ποιες όχι.

Ο έλεγχος της στατιστικής σημαντικότητας της εξίσωσης παλινδρόμησης ταυτίζεται με τον έλεγχο της στατιστικής σημαντικότητας του συντελεστή πολλαπλού προσδιορισμού R^2. Δηλαδή θα ελέγξουμε αυτό που μετρά ο R^2, εάν το ποσοστό των μεταβολών της Y που οφείλονται στις επιδράσεις των ανεξάρτητων μεταβολών X_1, X_2, …, X_κ, και ως εκ τούτο εξηγείται από την εξίσωση πολλαπλής παλινδρόμησης, είναι διάφορο του μηδενός. Έτσι, η υπόθεση μηδέν (H_0) και η εναλλακτική υπόθεση (H_1) διατυπώνονται ως εξής:

Υπόθεση μηδέν (H_0): Η εξίσωση της παλινδρόμησης δεν εξηγεί καθόλου τις μεταβολές της Y και επομένως $\beta_1 = \beta_2 = \ldots = \beta_\kappa = 0$.

Εναλλακτική Υπόθεση (H_1): Η εξίσωση της παλινδρόμησης εξηγεί ένα μέρος των μεταβολών της Y και τουλάχιστον ένας συντελεστής $\beta_i \neq 0$.

Επομένως, θα συγκρίνουμε τις δύο συνιστώσες της SST, την εξηγημένη (SSR) και την ανεξήγητη (SSE). Εάν η πρώτη είναι σημαντικά μεγαλύτερη της δεύτερης, σημαίνει ότι η επίδραση της εξίσωση παλινδρόμησης είναι σημαντική. Ενώ στην αντίθετη περίπτωση που η ανεξήγητη (SSE) είναι σημαντικά μεγαλύτερη από την εξηγημένη (SSR), σημαίνει ότι το ποσοστό της SST που περιγράφεται από την εξίσωση είναι αμελητέο.

Τα SSR και SSE είναι αθροίσματα τετραγώνων αποκλίσεων, που όμως βασίζονται σε διαφορετικό αριθμό βαθμών ελευθερίας. Επομένως, η σύγκριση μεταξύ τους θα γίνει αφού διαιρεθούν με τους αντίστοιχους βαθμούς ελευθερίας, οι οποίοι οι λόγοι που θα προκύψουν ονομάζονται μέσα τετράγωνα και ο έλεγχος μεταξύ τους βασίζεται στην κατανομή F. Ο πίνακας 3 δείχνει όλη τη διαδικασία του ελέγχου.

Εάν η τιμή $F_{\kappa,n-\kappa-1}$ είναι μεγαλύτερη της κριτικής τιμής $F_{(\kappa,n-\kappa-1),\alpha}$ (όπου α: επίπεδο σημαντικότητας), απορρίπτεται η υπόθεση μηδέν και αντίστροφα. Ο λόγος SSE / (n – κ – 1) ονομάζεται και **μέσο τετραγωνικό σφάλμα** και συμβολίζεται με MSE. Με άλλα λόγια, το μέσο τετραγωνικό σφάλμα είναι το τετράγωνο του τυπικού σφάλματος εκτίμησης και ισούται με s_e^2.

ΠΙΝΑΚΑΣ 3 έλεγχος της στατιστικής σημαντικότητας της εξίσωσης πολλαπλής παλινδρόμησης

Πηγή Μεταβλητικότητας	Αθροίσματα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσα Τετράγωνα	Λόγος $F_{\kappa,n-\kappa-1}$
Παλινδρόμηση	SSR	κ	SSR / κ	[SSR / κ] / [SSE /(n-κ-1)]
Σφάλμα (ή κατάλοιπος	SSE	n – κ – 1	SSE /(n-κ-1)	
Σύνολο	SST	n – 1		

Έτσι, η στατιστική F με βαθμούς ελευθερίας κ και n – κ – 1, ισούται με:

$F_{(\kappa, n-\kappa-1)} = [\ \Sigma (\ \hat{y} - \bar{y}\ )^2 / \kappa\ ] / [\ \Sigma (\ y - \hat{y}\ )^2 / (\ n - \kappa - 1\ )\ ]$

$= [SSR / \kappa] / [\ SSE /(n-\kappa-1)]$

$= [SSR / \kappa] / MSE$ (10)

Επιπλέον, εάν ο έλεγχος της στατιστικής σημαντικότητας του συντελεστή πολλαπλού προσδιορισμού R^2 δείξει ότι η εξίσωση παλινδρόμησης, στο σύνολό της, εξηγεί ένα σημαντικό μέρος των μεταβολών της Y, το επόμενο βήμα είναι να ελέγξουμε τη σημαντικότητα των συντελεστών μερικής παλινδρόμησης $\hat{\beta}_i$, $i = 1,2,...,\kappa$. Όπως συμβαίνει με όλες τις παραμέτρους που η εκτίμησή τους βασίζεται σε δείγμα παρατηρήσεων, έτσι και οι συντελεστές $\hat{\beta}_i$ υπόκεινται στα σφάλματα της δειγματοληψίας. Αυτό σημαίνει ότι πρέπει να γνωρίζουμε όχι μόνο εάν οι β_i είναι διάφοροι του μηδενός, αλλά και σε ποιο διάστημα εμπιστοσύνης βρίσκονται οι τιμές των συντελεστών μερικής παλινδρόμησης του πληθυσμού. Άλλωστε δεν πρέπει να ξεχνάμε ότι οι συντελεστές παλινδρόμησης β_i , $i = 1,2,...,\kappa$ είναι εκείνοι που έχουν όλη την ευθύνη της περιγραφής μεταξύ της σχέσης εξάρτησης της Y από τις μεταβλητές της X.

Το τυπικό σφάλμα της κατανομής δειγματοληψίας του συντελεστή $\hat{\beta}_i$ συμβολίζεται με $\sigma_{\hat{\beta}_i}$ και έτσι, η υπόθεση μηδέν (H_0) και η εναλλακτική υπόθεση (H_1) διατυπώνονται ως εξής:

Υπόθεση μηδέν (H_0): $\beta_i = \beta_i^*$

(Ο συντελεστής μερικής παλινδρόμησης i του πληθυσμού ισούται με β_i^*)

Εναλλακτική Υπόθεση (H_1): $\beta_i \neq \beta_i^*$

(Ο συντελεστής μερικής παλινδρόμησης i του πληθυσμού είναι διάφορος του β_i^*)

Ο έλεγχος γίνεται με το γνωστό κριτήριο t και n – κ – 1 βαθμούς ελευθερίας, δηλαδή:

$$\left| t_{n-\kappa-1} \right| = \frac{\left| \hat{\beta}_i - \beta_i^* \right|}{s_{\hat{\beta}i}} \qquad (11)$$

Εάν η τιμή $\left| t_{n-\kappa-1} \right|$ είναι μεγαλύτερη της κριτικής τιμής $\left| t_{n-\kappa-1,\alpha/2} \right|$, απορρίπτεται η υπόθεση μηδέν και αντίστροφα. Για επίπεδο σημαντικότητας α, η κριτική τιμή αντιστοιχεί στο κάτω ή άνω $\frac{\alpha}{2}$ της αντίστοιχης κατανομής t με n – κ – 1 βαθμούς ελευθερίας. Με τον τύπο (11) ελέγχουμε και την υπόθεση μηδέν (H_0) με $\beta_i = 0$. Θα πρέπει να διευκρινίσουμε ότι οι έλεγχοι του τύπου αυτού δεν είναι ανεξάρτητοι μεταξύ τους, αλλά ο έλεγχος για κάθε

συντελεστή β_i γίνεται υπό προϋπόθεση ότι υπάρχουν και άλλες ανεξάρτητες μεταβλητές στο υπόδειγμα, δηλαδή οι έλεγχοι του τύπου (11) είναι υπό συνθήκη έλεγχοι της στατιστικής σημαντικότητας των β_i . Έτσι, η σωστή διατύπωση της υπόθεσης H_0 και της εναλλακτικής H_1 για τον έλεγχο της στατιστικής σημαντικότητας των β_i είναι:

Υπόθεση μηδέν (H_0): $\beta_i = 0$, δεδομένου ότι όλες οι ανεξάρτητες μεταβλητές περιλαμβάνονται στο υπόδειγμα.

Εναλλακτική Υπόθεση (H_1): $\beta_i \neq 0$, δεδομένου ότι όλες οι ανεξάρτητες μεταβλητές περιλαμβάνονται στο υπόδειγμα.

Τέλος για το επίπεδο σημαντικότητας α, το διάστημα εμπιστοσύνης του i συντελεστή παλινδρόμησης ισούται με:

$$\hat{\beta}_i - s_{\hat{\beta}i} \cdot t_{n-\kappa-1,a/2} < \beta_i < \hat{\beta}_i + s_{\hat{\beta}i} \cdot t_{n-\kappa-1,a/2} \qquad (12)$$

Συμπερασματικά, σύμφωνα με τα παραπάνω κεφάλαια θα αναφερθούμε, πως λειτουργούν οι εφαρμογές της γραμμικής παλινδρόμησης στο χώρο των επιχειρήσεων και την οικονομία στο εξής παρακάτω κεφάλαιο μέσω του στατιστικού προγράμματος (SPSS).

ΚΕΦΑΛΑΙΟ 5 ΕΦΑΡΜΟΓΕΣ ΤΗΣ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΤΟ ΧΩΡΟ ΤΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ ΚΑΙ ΤΗΝ ΟΙΚΟΝΟΜΙΑ

ΑΣΚΗΣΗ 1: Τα ετήσια έξοδα διατροφής (σε χιλιάδες €) που αντιστοιχεί στην εξαρτημένη μεταβλητή Y και το ετήσιο οικογενειακό εισόδημα που αντιστοιχεί στην ανεξάρτητη μεταβλητή X έχουν ως εξής:

x_i	y_i
5,2	28
5,1	26
5,6	32
4,6	24
11,3	54
8,1	58
7,8	44
5,8	30
6,1	40
16,0	82
4,9	42
11,8	55
5,2	28
4,8	20
7,9	42
6,4	46
15,8	110
14,0	85
5,1	31
2,9	26
12,0	70
13,6	75
17,2	80

12,8	82
15,0	90

Α) Να κατασκευαστεί το διάγραμμα διασποράς των μεταβλητών Χ και Υ.

Β) Να ερμηνευθεί ο συντελεστής του Pearson.

Γ) Να βρεθεί ο συντελεστής προσδιορισμού.

Δ) Να βρεθεί η ευθεία παλινδρόμησης και να σχολιαστεί.

Ε) Να βρεθεί ο πίνακας ανάλυσης διακύμανσης (ANOVA) και να κατασκευαστεί το διάγραμμα διασποράς p-p plot (κανονικότητα των σφαλμάτων).

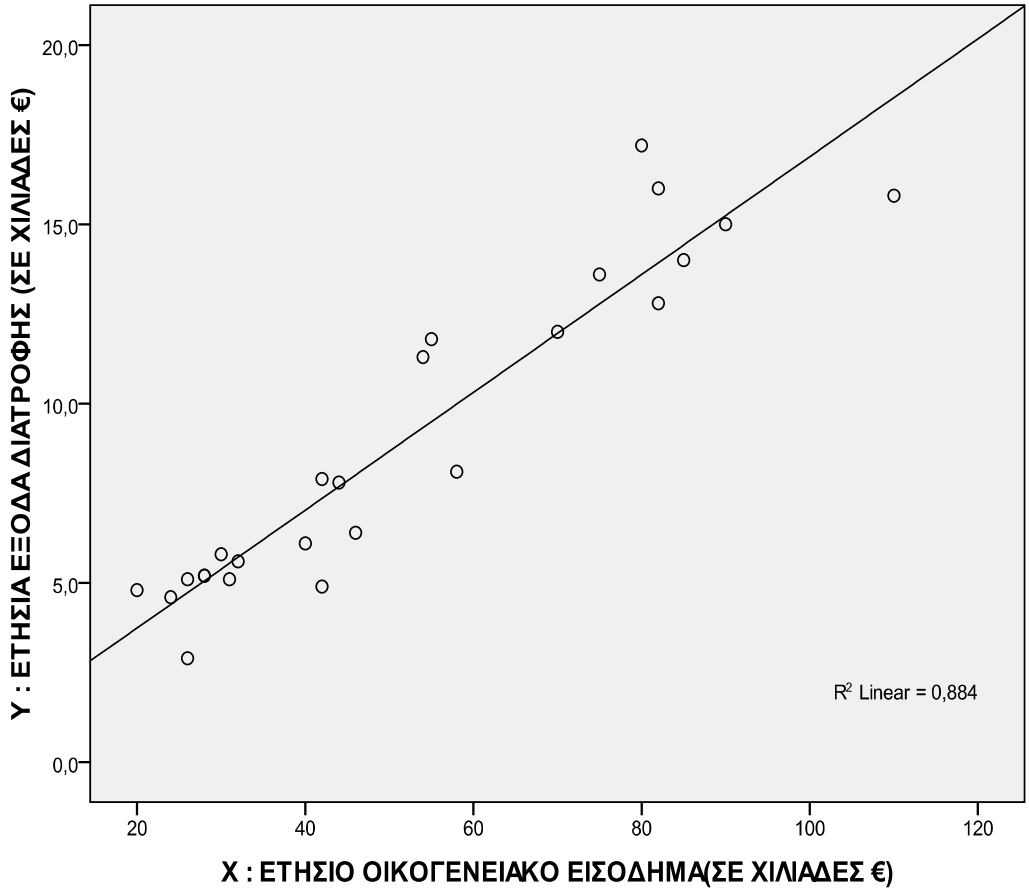
ΣΤ) Να ερμηνευθεί ο έλεγχος του Kolmogorov-Smirnov.

Ζ) Να βρεθεί η ανεξαρτησία των καταλοίπων μέσα από τον έλεγχο ροών και να σχολιαστεί.

ΛΥΣΗ:

Α) Με βάση την **απλή γραμμική παλινδρόμηση** έχουμε:

ΔΙΑΓΡΑΜΜΑ 1



Στο παραπάνω γράφημα δίνεται το διάγραμμα διασποράς της μεταβλητής X που αντιστοιχεί στο ετήσιο οικογενειακό εισόδημα και της μεταβλητής Y που αντιστοιχεί στα ετήσια έξοδα διατροφής. Από το γράφημα παρατηρούμε ότι οι μεταβλητές είναι θετικά συσχετισμένες δηλαδή η αύξηση των τιμών της μεταβλητής X συνεπάγεται και αύξηση της μεταβλητής Y.

Β)

ΠΙΝΑΚΑΣ 1

Correlations

		Y	X :
Pearson Correlation	Y :	1,000	,940
	X	**,940**	1,000
Sig. (1-tailed)	Y	.	,000
	X	,000	.
N	Y	25	25
	X	25	25

Ο συντελεστής Pearson λαμβάνει την τιμή 0,94, τιμή που μας υποδηλώνει ότι οι μεταβλητές είναι έντονα θετικά συσχετισμένες, όπως είχαμε διαπιστώσει στο παραπάνω διάγραμμα διασποράς.

Το αποτέλεσμα αυτό επιβεβαιώνεται στατιστικά και από τον έλεγχο:

H_0: $r_{x,y} = 0$ vs H_1: $r_{x,y} \neq 0$

όπου μηδενική υπόθεση είναι ότι οι μεταβλητές x, y είναι ασυσχέτιστες έναντι της εναλλακτικής ότι συσχετίζονται.

Από το p-value του ελέγχου που δίνεται από τον πίνακα ισχύει ότι

p-value = 0,000 < α (όπου α = 5%, το επίπεδο σημαντικότητας του ελέγχου) και επομένως απορρίπτουμε τη μηδενική υπόθεση.

Γ)

ΠΙΝΑΚΑΣ 2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,940[a]	**,884**	,879	1,5397	,884	176,071	1	23	,000

a. Predictors: (Constant), X : ΕΤΗΣΙΟ ΟΙΚΟΓΕΝΕΙΑΚΟ ΕΙΣΟΔΗΜΑ(ΣΕ ΧΙΛΙΑΔΕΣ €)

Από τον παραπάνω πίνακα βλέπουμε ότι ο συντελεστής προσδιορισμού R^2= 0,884 τιμή που αποτελεί ένδειξη της καταλληλότητας του μοντέλου καθώς προσεγγίζει το 1.

Δ)

ΠΙΝΑΚΑΣ 3

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	**,457**	,714		,640	**,529**	-1,020	1,933
	X	**,164**	,012	,940	13,269	**,000**	,139	,190

a. Dependent VariableQ Y Q ETHSIA EJODA DIATROFHS (SE XILIADES €)

Στο παραπάνω πίνακα δίνονται οι συντελεστές της ευθείας παλινδρόμησης η οποία δίνεται από την σχέση

$\hat{y} = \hat{\alpha} + \hat{\beta}x \Rightarrow \hat{y} = 0{,}457 + 0{,}164x$

Όσον αφορά τον σταθερό όρο (α) , παρατηρούμε μέσω του ελέγχου :

H_0: α = 0 vs H_1: α≠0

ότι το p-value = 0,529 > 0,05 = α και επομένως αποδεχόμαστε την μηδενική υπόθεση συνεπώς ο σταθερός όρος δεν είναι στατιστικά σημαντικός.

Επίσης από τον πίνακα μπορούμε να πραγματοποιήσουμε και τον αντίστοιχο έλεγχο για την κλίση της ευθείας που είναι ο ακόλουθος τύπος

H_0: β = 0 vs H_1: β≠0

διακρίνουμε ότι p-value = 0,000 < 0,05 = α και επομένως απορρίπτεται η μηδενική υπόθεση συνεπώς η παράμετρος β είναι στατιστικά σημαντική.

Τους παραπάνω ελέγχους μπορούμε εναλλακτικά να τους πραγματοποιήσουμε μέσω των διαστημάτων εμπιστοσύνης. Συγκεκριμένα από τον πίνακα βλέπουμε ότι το 95% διάστημα εμπιστοσύνης για τον σταθερό όρο είναι το [-1,020 , 1,933] διαπιστώνουμε ότι το 0 ανήκει στο διάστημα και συνεπώς αποδεχόμαστε την H_0. Αντίστοιχα το 95% διάστημα εμπιστοσύνης για την κλίση είναι το [0,139 , 0,190] και αφού το 0 δεν ανήκει στο διάστημα εμπιστοσύνης η H_0 απορρίπτεται.

Στην συνέχεια δίνεται ο πίνακας της ανάλυσης διακύμανσης (ANOVA)

Ε)

ΠΙΝΑΚΑΣ 4

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	417,431	1	417,431	176,071	**,000**[a]
	Residual	54,529	23	2,371		
	Total	471,960	24			

a. Predictors: (Constant), X : ΕΤΗΣΙΟ ΟΙΚΟΓΕΝΕΙΑΚΟ ΕΙΣΟΔΗΜΑ(ΣΕ ΧΙΛΙΑΔΕΣ €)

β. Δεπενδεντ Ωαριαβλε: Y : ΕΤΗΣΙΑ ΕΞΟΔΑ ΔΙΑΤΡΟΦΗΣ (ΣΕ ΧΙΛΙΑΔΕΣ €)

Από τον πίνακα μπορούμε να λάβουμε το συνολικό άθροισμα των τετραγώνων δηλαδή (SST) SST = 471,960 το οποίο εκφράζει την συνολική διασπορά των τιμών της μεταβλητής Y. Επίσης λαμβάνουμε το SSR = 417,431 το οποίο αποτελεί τη διασπορά των τιμών της Y κάτω από την ευθεία παλινδρόμησης, καθώς και SSE =

= 54,529 το οποίο αποτελεί το άθροισμα τετραγώνων των σφαλμάτων

Ακόμα δίνεται η τιμή της στατιστικής συνάρτησης F = MSR / MSE = 176,071 που αντιστοιχεί στον έλεγχο H_0: β = 0 vs H_1: β≠0

όπου η p-value = 0,000 < 0,05 = α και επομένως απορρίπτεται η μηδενική υπόθεση άρα η παράμετρος β είναι στατιστικά σημαντική.

Εδώ πρέπει να αναφέρουμε ότι ο παραπάνω έλεγχος από τον πίνακα ANOVA είναι ισοδύναμος με τον αντίστοιχο που λάβαμε από τον πίνακα με τους συντελεστές και αυτό οφείλεται στο γεγονός ότι έχουμε μόνο μια ανεξάρτητη μεταβλητή.

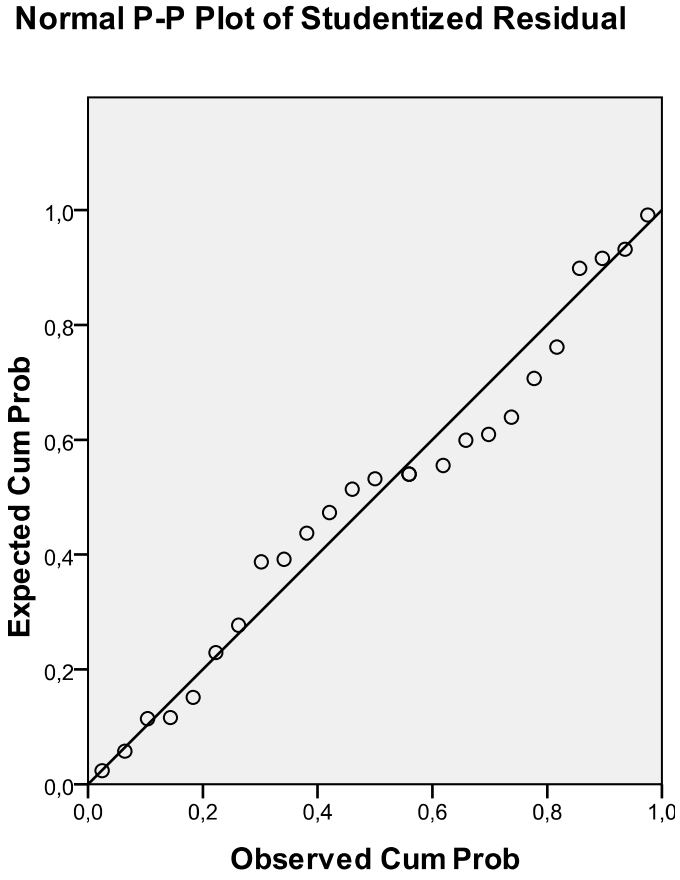
Απαραίτητες προϋποθέσεις για την εφαρμογή και την καταλληλότητα του μοντέλου είναι οι ακόλουθες:

1. τα σφάλματα ακολουθούν κανονική κατανομή
2. ομασκεδαστικότητα των σφαλμάτων
3. ανεξαρτησία των σφαλμάτων

Αρχικά θα ελέγξουμε αν τα σφάλματα ακολουθούν κανονική κατανομή.

Γραφικά την κανονικότητα των καταλοίπων θα την ελέγξουμε μέσω του διαγράμματος p-p plot.

ΔΙΑΓΡΑΜΜΑ 2



Από το παραπάνω γράφημα διακρίνουμε ότι τα κατάλοιπα προσεγγίζουν ικανοποιητικά την ευθεία γεγονός που αποτελεί ένδειξη ότι ακολουθούν κανονική κατανομή. Την ένδειξη αυτή μπορούμε να αποδείξουμε και στατιστικά μέσω του ελέγχου Kolmogorov-Smirnov

ΣΤ)

ΠΙΝΑΚΑΣ 5

One-Sample Kolmogorov-Smirnov Test

		Studentized Residual
N		25
Normal Parameters[a,,b]	Mean	-,0068690
	Std. Deviation	1,03196383
Most Extreme Differences	Absolute	,121
	Positive	,121
	Negative	-,107
Kolmogorov-Smirnov Z		,603
Asymp. Sig. (2-tailed)		**,860**

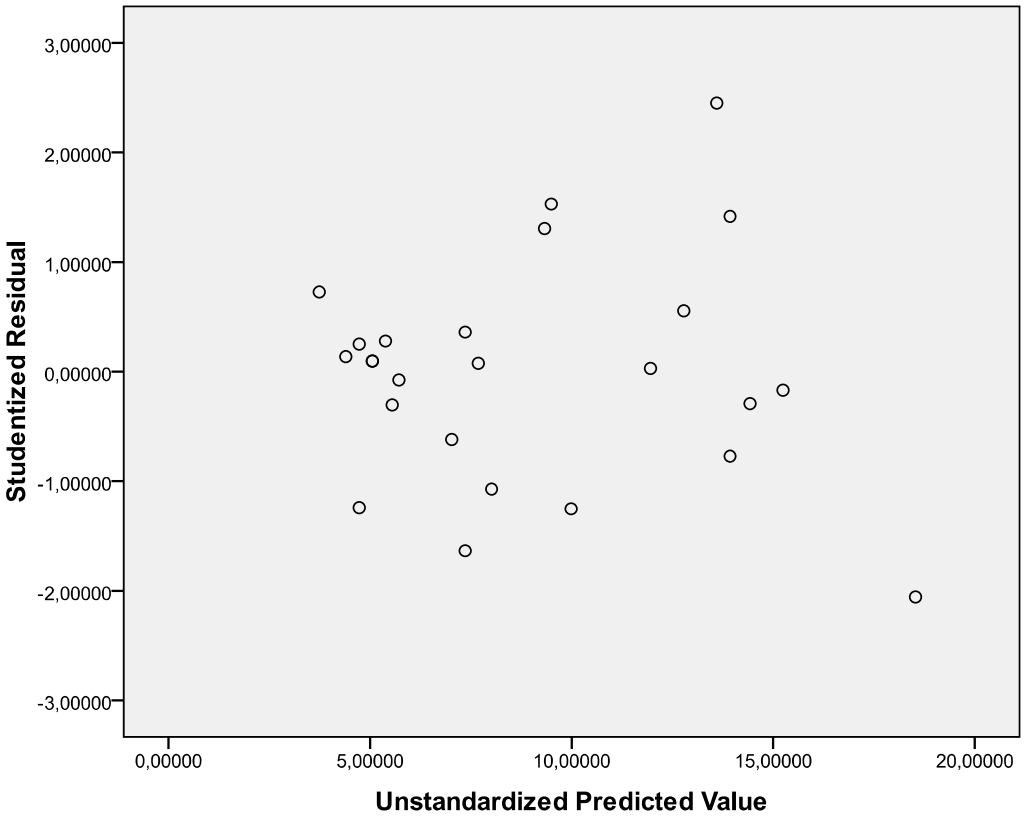
a. Test distribution is Normal.

b. Calculated from data.

Από τον πίνακα έχουμε ότι η p-value του ελέγχου είναι 0,860 > 0,05 και επομένως αποδεχόμαστε την μηδενική υπόθεση δηλ. την υπόθεση της κανονικότητας των σφαλμάτων

Στη συνέχεια δίνεται το διάγραμμα διασποράς όπου στον κατακόρυφο άξονα έχουμε τις εκτιμώμενες τιμές και στον οριζόντιο τα κατάλοιπα.

ΔΙΑΓΡΑΜΜΑ 3



Μέσω του διαγράμματος μπορούμε να ελέγξουμε την ομασκεδαστικότητα και ανεξαρτησία των σφαλμάτων. Επομένως οι τιμές είναι τυχαία διασκορπισμένες και άρα οι προϋποθέσεις πληρούνται.

Z)

Την ανεξαρτησία των καταλοίπων μπορούμε να την επιβεβαιώσουμε και στατιστικά από τον έλεγχο ροών.

ΠΙΝΑΚΑΣ 6

Runs Test

	Studentized Residual
Test Value[a]	,07598
Cases < Test Value	12
Cases >= Test Value	13
Total Cases	25
Number of Runs	12
Z	-,401
Asymp. Sig. (2-tailed)	**,688**

a. Median

Από τον πίνακα έχουμε ότι η p-value του ελέγχου είναι 0,688 > 0,05 και επομένως αποδεχόμαστε την μηδενική υπόθεση δηλ. την υπόθεση της ανεξαρτησίας των σφαλμάτων

Συμπεραίνουμε με βάση τα παραπάνω ότι οι υποθέσεις τις καταλληλότητας του γραμμικού υποδείγματος ικανοποιούνται που μας οδηγεί στο γεγονός ότι το μοντέλο που έχουμε εξάγει είναι ικανοποιητικό.

Διαπιστώσαμε ότι η εξίσωση παλινδρόμησης είναι οι εξής:

$\hat{y} = 0{,}457 + 0{,}164x$

Η κλίση της ευθείας είναι 0,164 και ουσιαστικά μας δείχνει ότι όταν η τιμή του χ μεταβληθεί κατά μία μονάδα (που στο παράδειγμα αντιστοιχεί σε 1000 ευρώ) τότε το $\hat{y}$ θα αυξηθεί κατά 0,164.

Ο σταθερός όρος 0,457 υποδηλώνει την τιμή που θα λάβει το $\hat{y}$ όταν το $x = 0$.

ΑΣΚΗΣΗ 2: Τα επενδυμένα κεφάλαια Χ και τα πραγματοποιηθέντα κέρδη Υ δεκαπέντε επιχειρήσεων κατά το έτος 2013 είχαν ως εξής (σε χιλιάδες ευρώ):

x_i	100	60	30	50	120	200	40	70	170	230	190	250	80	210	140
y_i	15	11	6	11	17	21	5	9	18	30	19	32	12	22	20

Α) Να κατασκευαστεί το διάγραμμα διασποράς των μεταβλητών Χ και Υ.

Β) Να προσδιορισθεί η ευθεία γραμμικής παλινδρόμησης $\hat{y} = \hat{\alpha} + \hat{\beta} \cdot x$.

Γ) Να ερμηνευθούν οι τιμές των συντελεστών παλινδρόμησης $\hat{\alpha}$ και $\hat{\beta}$.

Δ) Να βρεθεί ο συντελεστής γραμμικής συσχέτισης *r* και να σχολιασθεί το αποτέλεσμα.

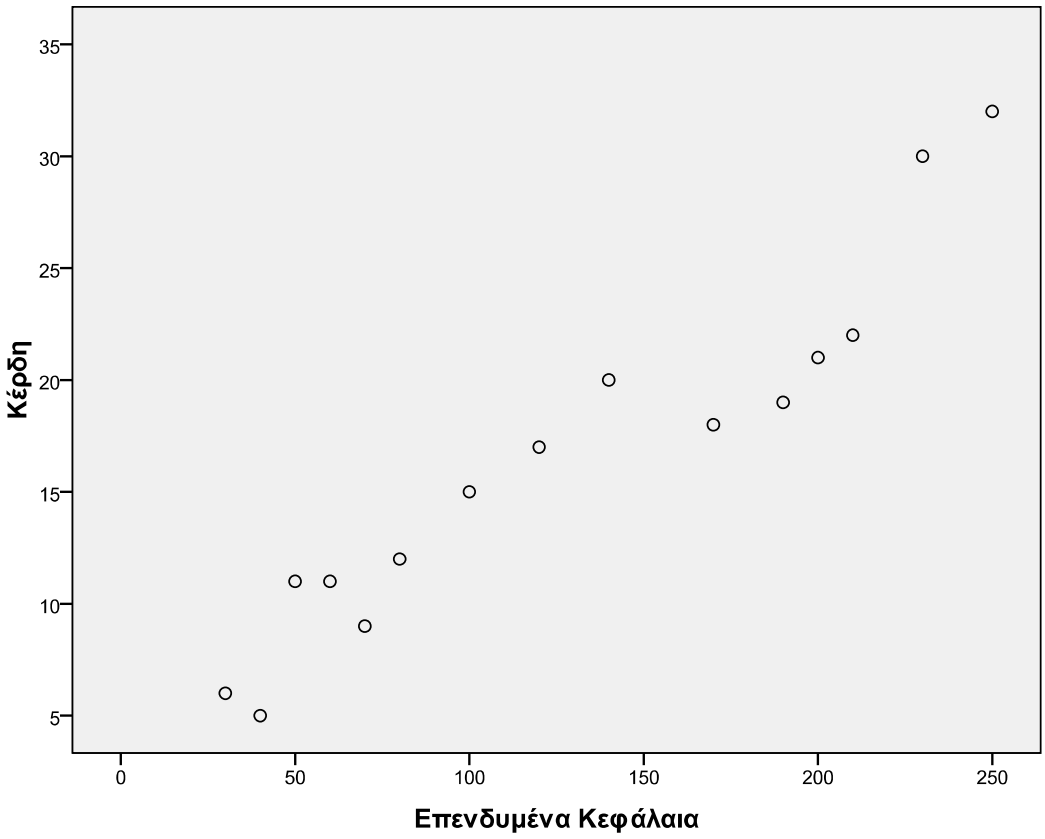
Ε) Να βρεθούν οι διαφορές μεταξύ των πραγματικών τιμών της εξαρτημένης μεταβλητής Υ και των αντίστοιχων εκτιμώμενων τιμών (residuals).

Ζ) Ποιο είναι το κέρδος που περιμένουμε να έχει μια επιχείρηση, εάν είναι γνωστό ότι έχει επενδύσει 160 χιλιάδες ευρώ;

ΛΥΣΗ:

Α) Το διάγραμμα διασποράς των μεταβλητών Χ και Υ φαίνεται στο παρακάτω σχήμα:

ΔΙΑΓΡΑΜΜΑ 4



Β) Κατασκευάζουμε τον παρακάτω πίνακα:

ΠΙΝΑΚΑΣ 7

x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
100	15	1500	10000	225
60	11	660	3600	121
30	6	180	900	36
50	11	550	2500	121
120	17	2040	14400	289
200	21	4200	40000	441
40	5	200	1600	25
70	9	630	4900	81
170	18	3060	28900	324
230	30	6900	52900	900
190	19	3610	36100	361
250	32	8000	62500	1024
80	12	960	6400	144
210	22	4620	44100	484
140	20	2800	19600	400
$\sum_{i=1}^{15} x_i = 1940$	$\sum_{i=1}^{15} y_i = 248$	$\sum_{i=1}^{15} x_i y_i = 39910$	$\sum_{i=1}^{15} x_i^2 = 328400$	$\sum_{i=1}^{15} y_i^2 = 4976$

Έχουμε:

$$\bar{x} = \frac{\sum_{i=1}^{15} x_i}{15} = \frac{1940}{15} = 129{,}33$$

και

$$\bar{y} = \frac{\sum_{i=1}^{15} y_i}{15} = \frac{248}{15} = 16{,}53\,.$$

Η εξίσωση παλινδρόμησης είναι της μορφής:

$$\hat{y} = \hat{\alpha} + \hat{\beta} \cdot x$$

Οι συντελεστές παλινδρόμησης $\hat{\alpha}$ και $\hat{\beta}$ υπολογίζονται με τους παρακάτω τύπους:

$$\hat{\beta} = \frac{15\sum_{i=1}^{15} x_i y_i - \left(\sum_{i=1}^{15} x_i\right) \cdot \left(\sum_{i=1}^{15} y_i\right)}{15\sum_{i=1}^{15} {x_i}^2 - \left(\sum_{i=1}^{15} x_i\right)^2} = \frac{15 \cdot 39910 - 1940 \cdot 248}{15 \cdot 328400 - 1940^2} = 0{,}101$$

και

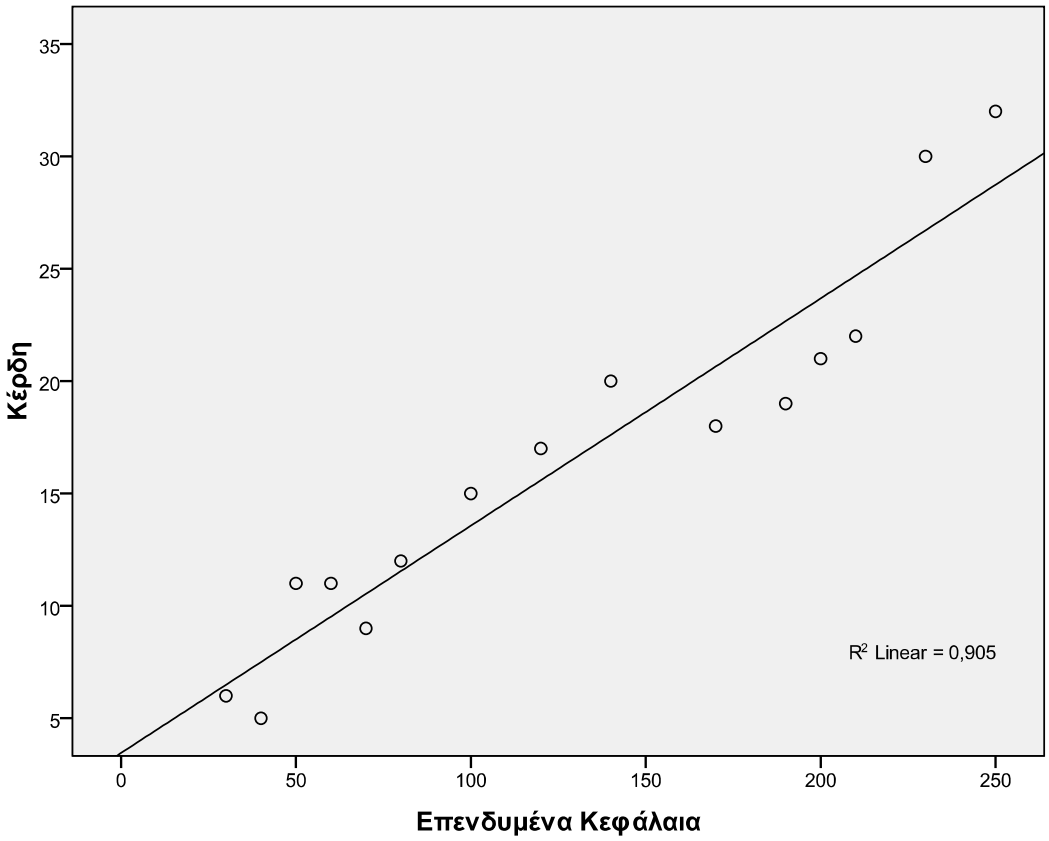
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 16{,}53 - 0{,}101 \cdot 129{,}33 = 3{,}46.$$

Συνεπώς, η ζητούμενη εξίσωση παλινδρόμησης είναι:

$\hat{y} = 3.46 + 0.101 \cdot x$.

Η ευθεία παλινδρόμησης φαίνεται στο παρακάτω σχήμα:

ΔΙΑΓΡΑΜΜΑ 5



Γ) Η σταθερά $\hat{\alpha} = 3{,}46$ προσδιορίζει (θεωρητικά) τα πραγματοποιηθέντα κέρδη που αντιστοιχούν στην τιμή X = 0. Η σταθερά $\hat{\beta} = 0{,}101$ δείχνει ότι όταν τα επενδυμένα κεφάλαια αυξηθούν κατά 1000 ευρώ, τότε τα πραγματοποιηθέντα κέρδη θα αυξηθούν κατά 101 ευρώ.

Δ) Ο συντελεστής γραμμικής συσχέτισης είναι:

$$r = \frac{15\sum_{i=1}^{15} x_i y_i - \left(\sum_{i=1}^{15} x_i\right)\cdot\left(\sum_{i=1}^{15} y_i\right)}{\sqrt{15\sum_{i=1}^{15} {x_i}^2 - \left(\sum_{i=1}^{15} x_i\right)^2}\sqrt{15\sum_{i=1}^{15} {y_i}^2 - \left(\sum_{i=1}^{15} y_i\right)^2}}$$

$$= \frac{15\cdot 39910 - 1940\cdot 248}{\sqrt{15\cdot 328400 - 1940^2}\sqrt{15\cdot 4976 - 248^2}} = 0{,}951$$

Επειδή η τιμή του συντελεστή συσχέτισης είναι πολύ κοντά στη μονάδα, υπάρχει ισχυρή θετική συσχέτιση μεταξύ των μεταβλητών X και Y.

Ε) Οι διαφορές $\hat{u}_i = y_i - \hat{y}_i$ μεταξύ των πραγματικών τιμών της εξαρτημένης μεταβλητής Y και των αντίστοιχων εκτιμώμενων τιμών δίνονται στον παρακάτω πίνακα:

ΠΙΝΑΚΑΣ 8

x_i	y_i	$\hat{y}_i$	$\hat{u}_i = y_i - \hat{y}_i$
100	15	13,56745	1,43255
60	11	9,52306	1,47694
30	6	6,48976	-0,48976
50	11	8,51196	2,48804
120	17	15,58964	1,41036
200	21	23,67842	-2,67842
40	5	7,50086	-2,50086
70	9	10,53415	-1,53415
170	18	20,64513	-2,64513
230	30	26,71172	3,28828
190	19	22,66733	-3,66733
250	32	28,73391	3,26609
80	12	11,54525	0,45475
210	22	24,68952	-2,68952
140	20	17,61184	2,38816

Ζ) Το κέρδος που περιμένουμε να έχει μια επιχείρηση, εάν είναι γνωστό ότι έχει επενδύσει 160 χιλιάδες ευρώ θα βρεθεί από την εξίσωση παλινδρόμησης για $x = 160$. Έχουμε:

$$\hat{y} = 3,46 + 0,101 \cdot 160 = 19,62.$$

Συνεπώς, το κέρδος που περιμένουμε να έχει μια επιχείρηση, εάν είναι γνωστό ότι έχει επενδύσει 160 χιλιάδες ευρώ είναι 19620 ευρώ.

ΑΣΚΗΣΗ 3: Στον παρακάτω πίνακα δίνονται οι πωλήσεις 10 οικοπέδων που έλαβαν χώρα σε μια περιοχή.

ΠΙΝΑΚΑΣ 9

Τιμή πώλησης ανά m^2 (Y)	2500	2630	2400	2700	2570	2100	2780	2230	2490	2340
Εμβαδόν οικοπέδου σε m^2 (X_1)	390	430	385	435	410	300	440	370	400	350
Συντελεστής δόμησης (X_2)	2,6	2,8	2,6	3	2,8	2,3	3,1	2,4	2,7	2,5
Μήκος πρόσοψης σε m (X_3)	21	22	20	21	19	17	22	19	20	18
Εμπορικότητα (X_4)	2	2,1	1,6	2,3	2	1,4	2,5	1,5	1,8	1,6

Να εκτιμηθεί η αξία οικοπέδου $420\,m^2$ της περιοχής που έχει συντελεστή δόμησης 2,9, μήκος πρόσοψης $18\,m$ και εμπορικότητα 1,9.

ΛΥΣΗ:

Με τη βοήθεια του λογισμικού στατιστικής επεξεργασίας ερευνητικών δεδομένων SPSS 17.0 βρίσκουμε την εξίσωση παλινδρόμησης:

$$\hat{y} = 618{,}72 + 1{,}456 \cdot x_1 + 314{,}372 \cdot x_2 + 5{,}234 \cdot x_3 + 180{,}393 \cdot x_4$$

ΠΙΝΑΚΑΣ 10

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	618,872	370,151		1,672	,155
	Εμβαδόν οικοπέδου	1,456	1,096	,297	1,329	,241
	Συντελεστής δόμησης	314,372	234,955	,376	1,338	,239
	Μήκος πρόσοψης	5,234	20,015	,041	,262	,804
	Εμπορικότητα	180,393	138,005	,309	1,307	,248

a. Dependent Variable: Τιμή πώλησης

Συνεπώς, η αξία οικοπέδου $420\,m^2$ της περιοχής που έχει συντελεστή δόμησης 2,9, μήκος πρόσοψης $18\,m$ και εμπορικότητα 1,9 είναι:

$$618{,}872 + 1{,}456 \cdot 420 + 314{,}372 \cdot 2{,}9 + 5{,}234 \cdot 18 + 180{,}393 \cdot 1{,}9 = 2579{,}0295.$$

ΑΣΚΗΣΗ 4: Στον παρακάτω πίνακα παρατίθενται οι πωλήσεις ενός προϊόντος (σε χιλιάδες ευρώ) σε μια συγκεκριμένη περίοδο (Y), οι δαπάνες διαφήμισης του προϊόντος (σε χιλιάδες ευρώ) κατά τη διάρκεια της ίδιας περιόδου (X_1) και ο αριθμός των ανταγωνιστικών προϊόντων που πωλούνται σε κάθε περιοχή (X_2).

ΠΙΝΑΚΑΣ 11

Περιοχή	**Πωλήσεις** (Y)	**Διαφημιστικές Δαπάνες** (X_1)	**Αριθμός Ανταγωνιστών** (X_2)
1	5	4	15
2	10	7	16
3	9	6	10
4	20	15	7
5	17	12	9
6	8	5	11
7	22	15	6

8	13	10	8
9	16	11	10
10	18	13	5
Σύνολο	**138**	**98**	**97**

Α) Να κατασκευαστεί το τρισδιάστατο διάγραμμα διασποράς των μεταβλητών X_1, X_2, και Y.

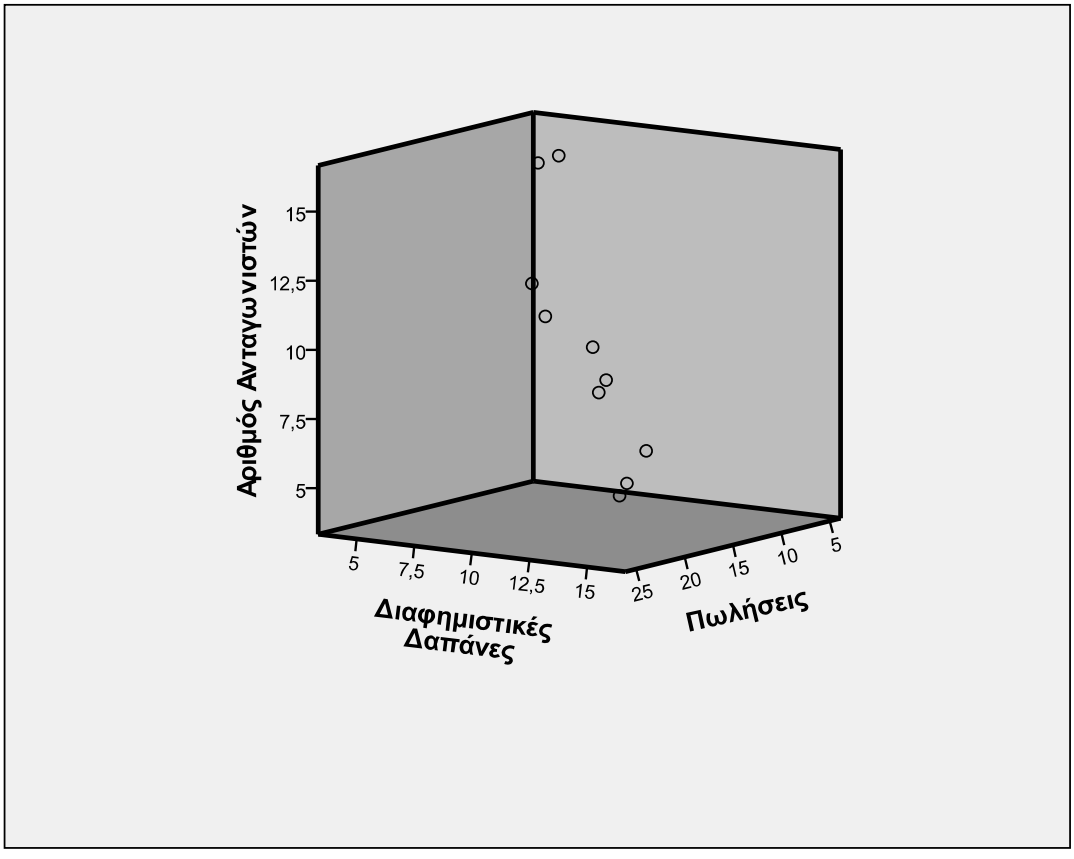
Β) Να βρεθεί το επίπεδο παλινδρόμησης της μεταβλητής Y ως προς τις μεταβλητές X_1 και X_2.

Γ) Να βρεθούν ο συντελεστής πολλαπλής συσχέτισης (R), ο συντελεστής πολλαπλού προσδιορισμού (R^2) και ο προσαρμοσμένος συντελεστής πολλαπλού προσδιορισμού (R^2_{adj}).

ΛΥΣΗ:

Α) Το τρισδιάστατο διάγραμμα διασποράς των μεταβλητών X_1, X_2 και Y φαίνεται στο παρακάτω σχήμα:

ΔΙΑΓΡΑΜΜΑ 6



Β) Με τη βοήθεια του λογισμικού στατιστικής επεξεργασίας ερευνητικών δεδομένων SPSS 17.0 βρίσκουμε την εξίσωση παλινδρόμησης:

$$\hat{y} = 0{,}920 + 1{,}349 \cdot x_1 - 0{,}035 \cdot x_2$$

ΠΙΝΑΚΑΣ 12

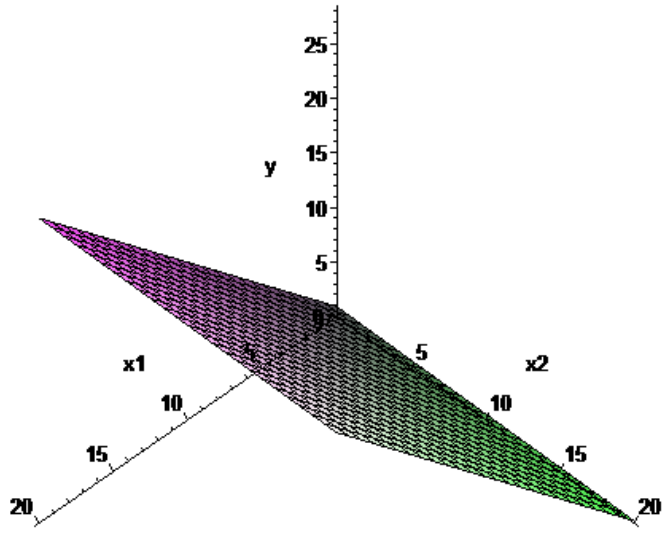
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,920	2,294		,401	,700
	Διαφημιστικές Δαπάνες	1,349	,115	,973	11,701	,000
	Αριθμός Ανταγωνιστών	-,035	,131	-,022	-,270	,795

a. Dependent Variable: Πωλήσεις

Το επίπεδο παλινδρόμησης της μεταβλητής Y ως προς τις μεταβλητές X_1 και X_2 φαίνεται στο παρακάτω σχήμα:

ΔΙΑΓΡΑΜΜΑ 7



Γ) Στον παρακάτω πίνακα δίνεται ο συντελεστής πολλαπλής συσχέτισης ($R = 0{,}991$) μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής του υποδείγματος. Στον ίδιο πίνακα, αναφέρεται ο συντελεστής πολλαπλού προσδιορισμού ($R^2 = 0{,}983$), ο οποίος ορίζει το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής που ερμηνεύεται από τις ανεξάρτητες. Επίσης, στον πίνακα δίνεται η τιμή του προσαρμοσμένου συντελεστή πολλαπλού προσδιορισμού:

$$R^2_{\text{adj}} = 0{,}978.$$

ΠΙΝΑΚΑΣ 13

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,991[a]	,983	,978	,844

a. Predictors: (Constant), Αριθμός Ανταγωνιστών, Διαφημιστικές Δαπάνες

ΑΣΚΗΣΗ 5: Ένας ερευνητής θέλει να καθορίσει την καλύτερη τοποθεσία για το επόμενο εστιατόριο μιας αλυσίδας εστιατορίων. Μια προσεκτική μελέτη δείχνει ότι θα πρέπει να εξετάσει τις παρακάτω τρεις μεταβλητές:

X_1: Το μέσο εισόδημα των νοικοκυριών του πληθυσμού σε ευρώ.

X_2: Ο αριθμός των ανθρώπων που ζουν σε ακτίνα πέντε χιλιομέτρων από τη θέση του εστιατορίου.

X_3: Ο αριθμός των άμεσων ανταγωνιστών στην αγορά μέσα σε μια ακτίνα τριών χιλιομέτρων από τη θέση του εστιατορίου.

Στον παρακάτω πίνακα δίνονται οι πωλήσεις σε ευρώ 10 εστιατορίων της αλυσίδας εστιατορίων.

ΠΙΝΑΚΑΣ 14

Εστιατόρια	Πωλήσεις (Υ)	Μέσο Εισόδημα (X_1)	Πληθυσμός (X_2)	Ανταγωνιστές (X_3)
1	150000	19000	8000	4
2	113000	15000	6250	9
3	134000	17000	7350	5
4	95000	13000	5900	7
5	162000	20000	8600	4
6	146000	18500	7120	5
7	105000	12000	6360	8
8	170000	23000	8500	3
9	125000	16500	5900	5
10	152000	20000	7200	4

Α) Να βρεθεί η εξίσωση παλινδρόμησης $\hat{y} = \hat{\alpha} + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \hat{\beta}_3 \cdot x_3$.

Β) Να βρεθούν ο συντελεστής πολλαπλής συσχέτισης (R), ο συντελεστής πολλαπλού προσδιορισμού (R^2) και ο προσαρμοσμένος συντελεστής πολλαπλού προσδιορισμού (R^2_{adj}).

ΛΥΣΗ:

Α) Με τη βοήθεια του λογισμικού στατιστικής επεξεργασίας ερευνητικών δεδομένων SPSS 17.0 βρίσκουμε τον παρακάτω πίνακα:

ΠΙΝΑΚΑΣ 15

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3764,809	29219,235		,129	,902
	Μέσο Εισόδημα	4,966	1,247	,672	3,982	,007
	Πληθυσμός	7,038	3,011	,284	2,338	,058
	Ανταγωνιστές	-939,672	1824,707	-,073	-,515	,625

a. Dependent Variable: Πωλήσεις

Από τον παραπάνω πίνακα έχουμε:

$$\hat{\alpha} = 3764{,}809,\ \hat{\beta}_1 = 4{,}966,\ \hat{\beta}_2 = 7{,}038,\ \hat{\beta}_3 = -939{,}672.$$

Για παράδειγμα, ο συντελεστής $\hat{\beta}_1$ του μέσου εισοδήματος είναι η αύξηση των πωλήσεων σε ευρώ για κάθε αύξηση του μέσου εισοδήματος κατά ένα ευρώ, εφόσον ο πληθυσμός και ο αριθμός των ανταγωνιστών παραμένουν αμετάβλητοι.

Συνεπώς, η εξίσωση παλινδρόμησης είναι:

$$\hat{y} = 3764{,}809 + 4{,}966 \cdot x_1 + 7{,}038 \cdot x_2 - 939{,}672 \cdot x_3$$

Β) Στον παρακάτω πίνακα δίνεται ο συντελεστής πολλαπλής συσχέτισης ($R = 0{,}986$) μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής του υποδείγματος. Ο συντελεστής πολλαπλής συσχέτισης αποτελεί μέτρο της συνολικής γραμμικής σχέσης που υπάρχει μεταξύ της εξαρτημένης μεταβλητής, δηλαδή των πωλήσεων και των ανεξάρτητων μεταβλητών, δηλαδή του μέσου εισοδήματος, του πληθυσμού και του αριθμού των ανταγωνιστών. Στον ίδιο πίνακα, αναφέρεται ο συντελεστής πολλαπλού προσδιορισμού ($R^2 = 0{,}973$), ο οποίος ορίζει το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής που ερμηνεύεται από τις ανεξάρτητες. Από την τιμή του συγκεκριμένου συντελεστή προκύπτει ότι το 97,3% της μεταβλητότητας των πωλήσεων ερμηνεύεται από το μέσο εισόδημα, τον πληθυσμό και τον αριθμό των ανταγωνιστών. Επίσης, στον πίνακα δίνεται η τιμή του προσαρμοσμένου συντελεστή πολλαπλού προσδιορισμού:

$$R^2_{\text{adj}} = 0{,}959.$$

ΠΙΝΑΚΑΣ 16

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,986[a]	,973	,959	5072,409

a. Predictors: (Constant), Ανταγωνιστές, Πληθυσμός, Μέσο Εισόδημα

ΣΥΜΠΕΡΑΣΜΑΤΑ

Συνοψίζοντας, αντικείμενο της διπλωματικής αυτής εργασίας είναι η γραμμική παλινδρόμηση καθώς και οι εφαρμογές αυτής στις επιχειρήσεις και την οικονομία. Όπως αναφέρθηκε, η παλινδρόμηση αποτελεί μια ευρέως χρησιμοποιημένη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μιας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Τέτοια παραδείγματα εφαρμογής της παλινδρόμησης αποτελεί η πρόβλεψη της ζήτησης για ένα νέο προϊόν ή υπηρεσία συναρτήσει των δαπανών διαφήμισης ή ο υπολογισμός της ταχύτητας του ανέμου σε σχέση με την θερμοκρασία, την υγρασία και την ατμοσφαιρική πίεση του περιβάλλοντος.

Στα πλαίσια της εργασίας αρχικά δόθηκε μία σύντομη αναδρομή της στατιστικής καθώς και στις βασικές έννοιες αυτής, στη συνέχεια τα κεφάλαια της εργασία ήταν επικεντρωμένα στην παλινδρόμηση. Συγκεκριμένα, δόθηκε αναλυτική περιγραφή της απλής γραμμικής παλινδρόμησης, όπου μελετήθηκε η μέθοδος εξαγωγής της ευθείας γραμμικής παλινδρόμησης (μέθοδος ελαχίστων τετραγώνων), οι μέθοδοι ελέγχου της στατιστικής σημαντικότητας του γραμμικού υποδείγματος στο οποίο έχουμε καταλήξει αλλά και οι βασικές υποθέσεις καταλληλότητας του μοντέλου. Αντίστοιχη ανάλυση πραγματοποιήθηκε και για την πολλαπλή γραμμική παλινδρόμηση καθώς, και για λόγους πληρότητας της εργασίας και έρευνας, ιδιαίτερη μνεία έγινε και για τη μη γραμμική παλινδρόμηση.

Τέλος, βασικό τμήμα της εργασίας αποτέλεσε και η πρακτική εφαρμογή της θεωρίας με τη χρήση παραδειγμάτων-ασκήσεων που πηγάζουν από το χώρο των επιχειρήσεων και της οικονομίας όπου η επίλυση των ασκήσεων αυτών επετεύχθη με τη χρήση του στατιστικού πακέτου SPSS.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- ΥΠΟΥΡΓΕΙΟ ΕΘΝΙΚΗΣ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ ΠΑΙΔΑΓΩΓΙΚΟ ΙΝΣΤΙΤΟΥΤΟ: ΑΔΑΜΟΠΟΥΛΟΣ Λ. , ΔΑΜΙΑΝΟΥ Χ. , ΣΒΕΡΚΟΣ Α. (1999) *ΜΑΘΗΜΑΤΙΚΑ ΚΑΙ ΣΤΟΙΧΕΙΑ ΣΤΑΤΙΣΤΙΚΗΣ Γ΄ ΓΕΝΙΚΟΥ ΛΥΚΕΙΟΥ*. ΟΡΓΑΝΙΣΜΟΣ ΕΚΔΟΣΕΩΝ ΔΙΔΑΚΤΙΚΩΝ ΒΙΒΛΙΩΝ, ΑΘΗΝΑ.
- ΓΝΑΡΔΕΛΗΣ Χ. (2003) *ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ*. ΕΚΔΟΣΕΙΣ ΠΑΠΑΖΗΣΗ, Αθήνα
- Χαλικιάς Ι. (2003) *Στατιστική: Μέθοδοι Ανάλυσης για Επιχειρηματικές Αποφάσεις*. Εκδοτικός Οίκος Rosili και Ιωάννης Γ. Χαλικιάς, Γέρακας.
- Norman D. – Harry S (1997) *ΕΦΑΡΜΟΣΜΕΝΗ ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ* ΔΕΥΤΕΡΗ ΑΓΓΛΙΚΗ ΕΚΔΟΣΗ Μετάφραση – Επιμέλεια: Χατζηκωνσταντινίδης Ε. , Καλαματιανού Α. ΕΚΔΟΣΕΙΣ ΠΑΠΑΖΗΣΗ Α.Ε.Β.Ε. , Αθήνα

ΗΛΕΚΤΡΟΝΙΚΕΣ ΠΗΓΕΣ

- www.unipi.gr/faculty/mkoutras/regress.htm
  ΚΟΥΤΡΑΣ Μ. ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ, Διαφάνειες Μαθημάτων 2010-2011, αντλήθηκε στις 12.10.2011 .
- http: //www.pi-schools.gr/lessons/tee/economic/
  Παιδαγωγικό Ι. Στατιστική Επιχειρήσεων για την Α΄ Τάξη των Τ.Ε.Ε. του τομέα Οικονομίας και Διοίκησης, 2013, αντλήθηκε στις 25.5.2013